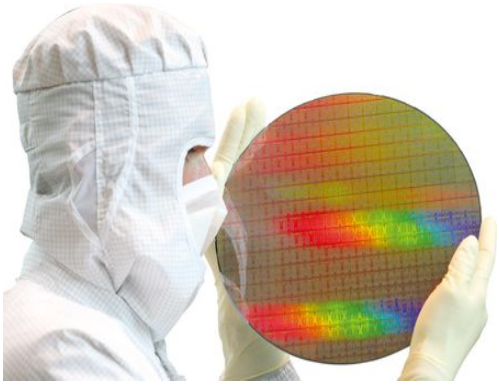
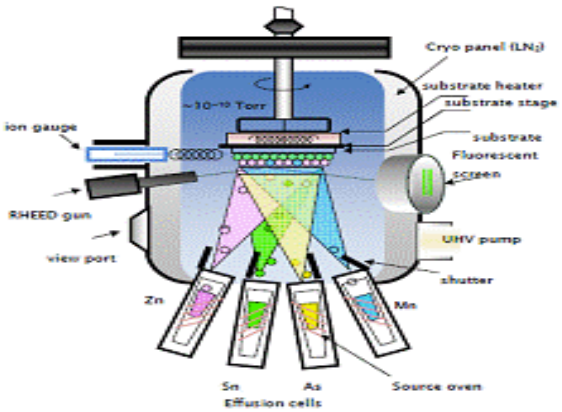
№530785-TEMPUS-1-2012-1-PL-TEMPUS-JPCR

Institute of High Technology, Taras Shevchenko National University of Kyiv

Microfabrication of IC and Microsystem Devices

Valeriy Skryshevsky, Anatoliy Evtukh, Volodymyr Ilchenko, Anatoliy Shkavro, Volodymyr Verbitskiy





Textbook

2016

Textbook " **Microfabrication of IC and Microsystem Devices**" developed to help higher education institutions in Ukraine to introduce new master's educational program "Designing microsystems".

Textbook " **Microfabrication of IC and Microsystem Devices**" was created with the support of the European Union within the Joint European Project "Curricula Development for New Specialization: Master of Engineering in Microsystems Design" (MastMST), identification number 530785-TEMPUS-1-2012-1-PL-TEMPUS-JPCR.

Project Coordinator prof. Zbigniew Lisik, Lodz University of Technology, Lodz, Poland.

TEMPUS teams:

- LvivPolitechnicalNationalUniversity, Lviv, Ukraine ,
  Coordinator prof. Mykhailo Lobur.
- Taras Shevchenko National University of Kyiv, Ukraine,
  Coordinator prof. ValeriySkryshevsky.
- Kharkiv National University of Radioelectronics, Ukraine ,
  Coordinator prof. Vladimir Hahanov.
- Donetsk National Technical University, Krasnoarmiysk,
  Coordinator prof. VolodymyrSviatny.
- Ilmenau University of Technology, Germany,
  Coordinator prof. Ivo Rangelow.
- Lyon Institute of Applied Sciences, France,
  Coordinator prof. Alexandra Apostoluk.
- University of Pavia, Italy,
  Coordinator prof. Paolo Di Barba.

The Textbook was approved by Editorial Committee (prof. Paolo Di Barba (University of Pavia) - Co-Chair, prof. Alexandra Apostoluk (LyonInstitute of Applied Sciences) – Co-Chair, members: prof. Zbigniew Lisik (Lodz University of Technology), Dr.Jacek Podgorski (Lodz University of Technology), Dr.Janusz Wozny (Lodz University of Technology), Dr.Valentyn lshchuk (Ilmenau University of Technology), Dr.Maria Evelina Mognaschi (University of Pavia), Dr.Roberto Galdi (University of Pavia) May 6, 2016, Pavia, Italy.

The authors express their deep gratitude to the aforementioned universities for full support of the project.

Content

Chapter 4. Dielectric and polycrystalline silicon deposition (A. Evtukh)
4.1. Introduction
4.2. Silicon Dioxide Deposition
4.3. Silicon Nitride Deposition
4.4. Low-dielectric Constant Materials Deposition
4.5. High-dielectric Constant Materials Deposition
4.6. Polysilicon Deposition
4.6.1. Gas Dynamic
4.6.2. Wafer-to-Wafer Uniformity
4.6.3. Silicon Gas Sources
4.6.4. Doping During Deposition
4.6.5. Polysilicon Deposition Process
4.7. Conclusions

Chapter 5. Silicon oxidation (A. Evtukh)
5.1. Introduction
5.2. Thermal Oxidation Process
5.2.1. Growth Kinetics
5.2.2. Thin Oxide Growth
5.3. Impurity Redistribution during Oxidation
5.4. Masking Properties of Silicon Dioxide
5.5. Silicon Oxide Quality
5.6. Silicon Oxide Structure
5.7. Oxidation of Polycrystalline Silicon
5.71. Oxide Growth on Polysilicon
5.72. Oxide-Thickness Evaluation
5.8. Conclusions

Chapter 6. Microlithography (V.Verbitsky)
6.1 Properties of microlithography
6.2. Types of lithography
6.3. Photolithography
6.3.1. Technology of photolithography
6.3.2. Photoresists
6.3.3 Exposure methods
6.4. Models exhibiting
6.5. Exposure of negative photoresist
6.6. Advances photolithography

Chapter 7. Etching (V.Skryshevsky)
7.1 Introduction
7.2 Wet chemical etching.
7.3 Isotropic wet etching
7.3.1 Silicon dioxide
7.3.2 Silicon isotropic etching
7.3.3 Siliconnitride
7.3.4 Aluminum and other materials
7.3 Anisotropic silicon etching
7.4 Etch stop process
7.5 Physical dry etching

Chapter 8. Diffusion (V.Ilchenko)
8.1. Basic diffusion process
8.1.1. Diffusion Equation
8.1.1.1.Constant-source diffusion: predeposition
8.1.1.2. Limited-source diffusion: drive-in.
8.1.2. Diffusion profiles
8.2. Extrinsic diffusion
8.3. Lateral diffusion

Chapter 9. Ion implantation (V.Skryshevsky)
9.1 Introduction
9.2 Set up and work of ion implanters
9.3 Ion range
9.3.1. Binary collision and stopping power
9.3.2.Profile of the implanted ions
9.4. Backscattering, surface sputter and channeling
9.5.Implantation through a mask
9.6. Creation and healing of the defects
9.6.1.Primary collision and cascade
9.6.2 Point defects
9.6.3. Accumulation of damages, amorphization
9.6.4 Damage healing and dopant activation
9.7. Applications of ion implantation in traditional technologies CMOS

Chapter 10 Method of thin film deposition. Metallization. (V.Skryshevsky, V.Verbitsky)
10.1. Introduction
10.2 Physical Vapor Deposition
10.2.1 Evaporation
10.2.2 Sputtering
10.3 Chemical Vapor Deposition
10.3.1 PECVD: Plasma-Enhanced CVD
10.4 ALD: Atomic Layer Deposition
10.5 Electrochemical Deposition (ECD)
10.5.1 Electroplating/galvanic deposition
10.5.2 Plating on structured wafer
10.5.3 Electroless deposition
10.6 Application of Metallic Thin Films in MEMS and IC
10.6.1 Properties of metallic thin films
10.7Multilevel metallization
10.8 Polymer Films
10.8.1 Spin coating
10.8.2 Self-limiting methods
10.8.3 Properties of polymers

Chapter 11. Silicon-on-Insulator technology (V.Skryshevsky)
11.1 Introduction
11.2 Properties and advantages of SOI devices
11.2.1 The summary of Advantages of SOI technology:
11.3 Heteroepitaxial techniques
11.3.1 Silicon-on-Sapphire (SOS)
11.3.2 Silicon-on-Zirconia (SOZ)

13.2.2 Etch Stops
13.3 Surface Micromachining
13.3.1. The base of Surface Micromachining
13.3.2 Micromachining fabrication of the polysilicon thin film membrane
13.3.3 Sacrificial and Structural Materials
13.4 LIGA
13.5 Microelectrodischarge Machining
13.5.1General description
13.5.2 EDM Die-Sinking
13.5.3 Wire EDM (WEDM)
13.5.4 Electrodischarge Grinding (EDG)
13.5.5 Application of μ-EDM
13.6 Nanofabrication by Focused-ion-beam technique
13.6.1 Nanoscale stack fabrication by focused-ion-beam
13.7 Laser micromashining
13.7 .1 General remarks
13.7.2 Principles of laser material removal
13.7.3 Typical examples of Laser micromashiningApplications
13.8 Femtosecond Laser Processing
13.8.1 Peculiarities of Femtosecond Laser irradiation
13.8.2 Femtosecond-Laser-Assisted Wet Chemical Etching

Chapter 14. Nanotechnologies. (V.Ilchenko)
14.1. What is nanotechnologies?
14.1.1. Characterization of the nanostructures.
14.2. Fabrication methods.
14.2.1 Top-down process.
14.2.2 Bottom-up process.
14.3. Ordering of nanosystems.
14.4. Method for templating the growth of nanomaterials.
14.5. Nanocrystalline semiconductors.

Chapter 15. Methods for control (V.Ilchenko)
15.1. General classification of characterization methods.
15.2. Microscopy techniques.
15.3. Electron microscopy.
15.4. Field ion microscopy.
15.5. Diffraction techniques.
15.6. Spectroscopy techniques.
15.7. Scanning probe techniques.
15.8. Surface analysis and depth profiling.
15.9. Summary of techniques for property measurements.

Chapter 16. Process monitoring (A.Shkavro, A.Evtukh)
16.1. Process Flow and Key Measurement Points
16.2. Wafer State Measurements
16.2.1. Blanket Thin Film
16.2.2. Patterned Thin Film
16.2.3. Particle and Defect Inspection
16.2.4. Electrical Testing
16.3. Techniques for Characterization and Failure Analysis of Integrated Circuits
16.3.1. In-circuit Measurements

16.3.2. In-circuit Excitation
16.3.3. Repair Techniques
16.3.4. Comparison and Outlook

Chapter 17. Processes to produce integrated circuits (V.Verbitsky)
17.1. The technological features of production of integrated circuits
17.2. Technological processes of manufacturing a bipolar IC
17.3. Processes to produce bipolar circuits with isolation by back-biased p-n junction
17.3.1 Standard planar- epitaxial technology with grooves n+-layer andisolation by backward displaced p-n junction
17.3.2 The planar-epitaxial technology with deepen n+-layer and collector insulating diffusion (KID-technology).
17.3.3. Planar-epitaxial technology with base insulated diffusion (BID technology)
17.3.4. Planar-epitaxial technology with buried p-layer and insulation with double diffusion
17.4. Technological manufacturing processes of bipolar circuits with dielectric insulation elements
17.4.1. Microplanar epitaxial (planar) technology with dielectric isolation and poly-silicon application (EPIC-technology);
17.4.2. Microplanar epitaxial (planar) technology with glass isolation, signals or ceramics.
17.4.3. Microplanar epitaxial (planar) technology with insulated V-shaped grooves created by anisotropic etching of silicon (VIP-technology)
17.4.4. Microplanar epitaxial (planar) technology “silicon on sapphire” (SOS)
17.4.5. Microplanar epitaxial (planar) technology “silicon on dielectric” (SOD)

Chapter 18.Reliability of IC and microsystem devices (A.Shkavro)
18.1 Introduction
18.2 Factors that influence on the yield of serviceable IC
18.2.1. Technological factors
18.2.2. Factors of scheme projection (design factors)
18.2.3.The Point Defects
18.2.4. Ways to increase the yield of appropriate crystals
18.3. Characteristics of IC reliability
18.3.1. General concepts and terms of reliability
18.3.2. Quantitative indicators of reliability
18.3.3. Laws of distribution of random variables
18.3.4. Methods to forecast and assess the reliability.
18.4. Failure types and insecurity factors.
18.4.1. Failure types
18.4.2. Insecurity factors and causes to failure
18.4.3. Types of culling tests
18.5. Experimental methods to analyze the items’ quality, defectiveness, and malfunctions.
18.5.1 Testing structures
18.6. Causes and mechanisms of failures of discrete devices and IC
18.6.1. Failures of interconnections in ICs
18.6.2. Failures of IC metallization
18.6.3. Degradation and failure of contact metal-semiconductor in silicon IC with single-layer metallization
18.6.4. Processes of degradation in IC with multilayer metallization

Laboratory works
1.1 Determination of surface resistivity of semiconductor wafers and technological layers using four-probe method

1.2 Investigation of MIS structures electrical parameters in microelectronics technology by high-frequency C-V measurements
1.3 Research of dopant concentration profiles at surface of semiconductor in microelectronics technology
1.4 Investigation of electrical characteristics and parameters of the MIS transistors in microelectronics technology
1.5 Study of efficient lasing lifetime of minority carriers in a semiconductor MIS structures in microelectronics technology by dynamic nonequilibrium I-V characteristics

References

Chapter 1. Introduction to the microfabrication of IC and MEMS

V.Skryshevky

1.1 What is it microfabrication?

Concept of *Microfabrication* or *Microsystems technology* (MST) includes the technologies of *integrated circuits (*IC), *microelectromechanical systems* (MEMS), *microfluidics*, *nanotechnolo*gy, solar cells, micro-optics, flat-panel displays and countless others. Microfabrication can be applied by different way in all of these technologies. For example, the electroplating and photolytography is essential for deep submicron IC metallization and for LIGA-microstructures; the etch process is a key technology in surface micromashining of MEMS; imprint lithography is utilized in microfluidics where typical dimensions are 100 μm, as well as in nanotechnology, where feature sizes are down to 10 nm. In general, microfabrication is the collection of techniques used to fabricate devices in the micrometer range. Typical dimensions of microsystems are around 1 micrometer in the plane of the wafer (the range is rather wide, from 0.02 to 100 μm). Vertical dimensions range from atomic layer thickness (0.1 nm) to hundreds of micrometers, but thicknesses from 0.01 to 10 μm are most typical.

Well-known , the invention of the transistor in 1947 sparked a revolution. The transistor was born out of the fusion of radar technology (fast crystal detectors for electromagnetic radiation) and solid state physics. Developments of microfabrication methods enabled the fabrication of many transistors on a single piece of semiconductor and, a few years later, the fabrication of integrated circuits; that is, transistors were connected to each other on the wafer, rather than separated from each other and reconnected on the circuit board.

In many applications, the *microelectronic*s use of the semiconductor properties of silicon, but it is also important that silicon dioxide is such a useful material, for passivating silicon surfaces and protecting silicon during wafer processing. Silicon dioxide is readily formed on silicon, and it is high-quality electrical insulator. In addition to silicon transistors, integrated circuits require multiple levels of metal wiring, to route signals.

Micromechanics makes use of the mechanical properties of silicon. Silicon is extremely strong, and flexible beams, cantilevers and membranes can be made from it. Pressure sensors, resonators, gyroscopes, switches and other mechanical and electromechanical devices utilize the excellent mechanical properties of silicon. Microelectromechanical systems (MEMS) or microsystems, as they are also called, have expanded in every possible direction: microfluidics, microacoustics, biomedical microdevices, DNA microarrays, microreactors and microrockets to name a few. New subfields have emerged: BioMEMS, PowerMEMS, RF MEMS, as shown in Figures 1.1.

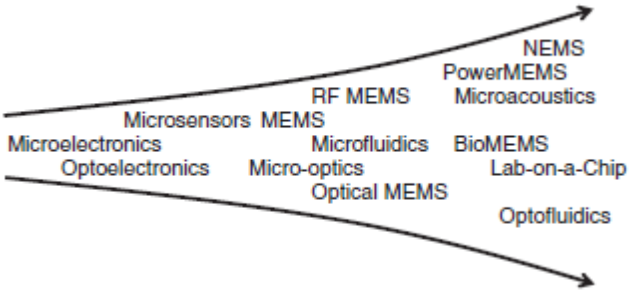


Figure 1.1 Evolution of microtechnology subfields from the 1960s onwards

Silicon optoelectronic devices can be used as light detectors like diodes and solar cells, but light emitters like lasers and LEDs are made of gallium arsenide and A^3B^5 semiconductors. Micro-optics makes use of silicon in another way: silicon, silicon dioxide and silicon nitride are used as waveguides and mirrors. MOEMS, or optical MEMS, utilize silicon in yet another way: silicon can be machined to make tilting mirrors, adjustable gratings and adaptive optical elements. The micromirror of Figure 1.2 takes advantage of silicon's smoothness and flatness for optics and its mechanical strength for tilting.

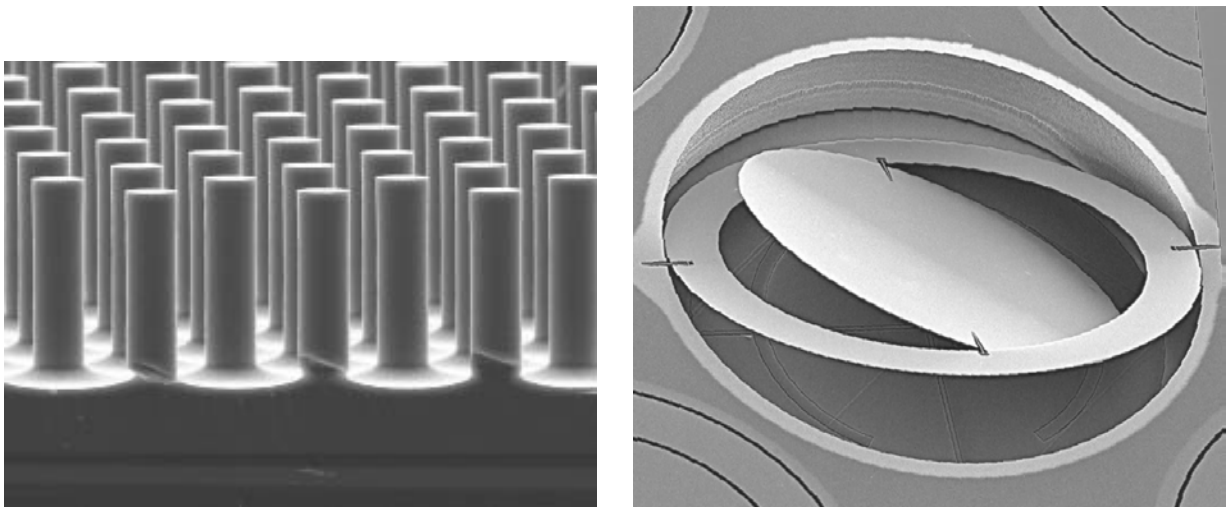


Figure 1.2. Silicon mictopillars for bioreactors (few μm) and b) silicon micromirror, 1mm in diameter, is supported by torsion bars 1.2 μm wide and 4μm thick.

Microtechnology has evolved into *nanotechnology* in many respects. Some of the tools are common, like electron beam lithography machines, which were used to draw nanometer-sized structures long before the term nanotechnology was coined. Electron beam and ion beam defined nanostructures are shown in Figure 3. Thin films down to atomic layer thicknesses have been grown and deposited by different methods. The tools of nanotechnology, such as the atomic force microscope (AFM), have been adopted both for microfabrication and characterization of microstructures.

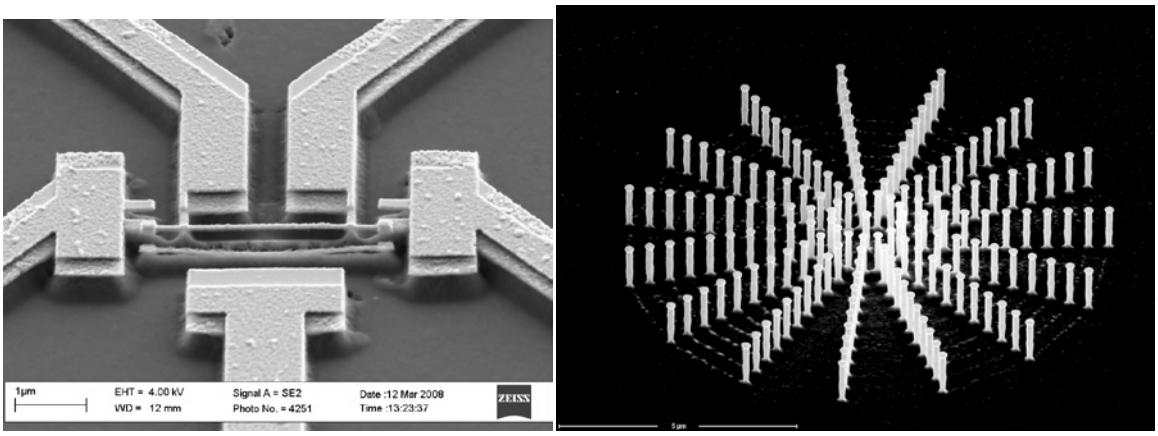


Figure 1.3. Electron microscope image of an electron beam defined gold–palladium horizontal nanobridge and vertical ion beam patterned nanopillars; 100 nm minimum dimension in both.

1.2 Materials for microfabrications

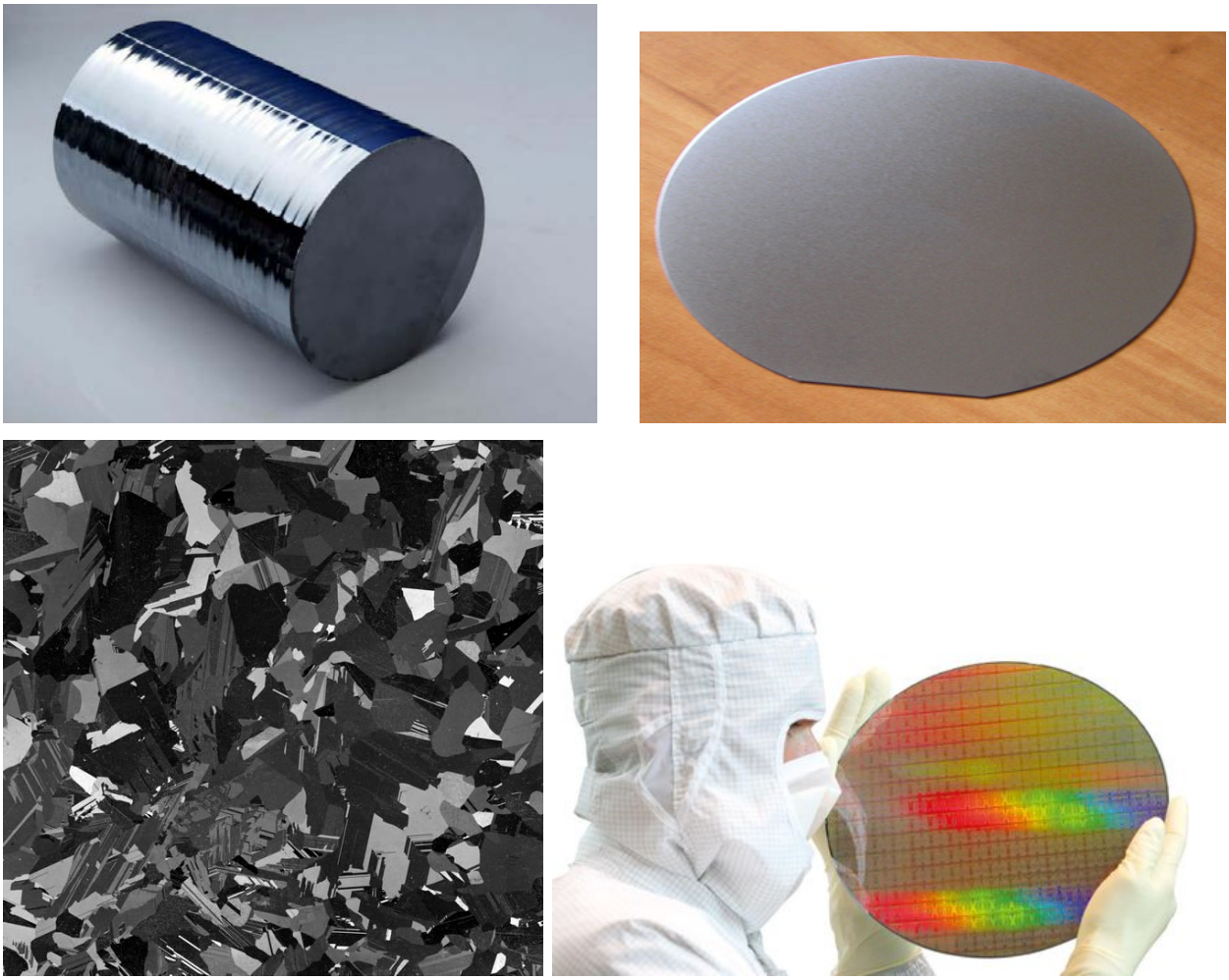


Figure 1.4. Silicon ingot (a), monocrystalline wafer (b), multicrystalline silicon (c) and 200 mm wafer with ICs (d).

As was note before, semiconductor silicon is the basic material of microfabrication. Silicon is available in both p-type (holes as charge carriers) and n-type (electrons as charge carriers), and its resistivity can be tailored over a wide range, from 0.001 to 20 000 Ω.cm. Silicon wafers are available in 100, 125, 150, 200 and 300mm diameters and various thicknesses. Silicon is available in different crystal orientations, and the control of its crystal quality is very advanced.

Bulk silicon wafers (Figure 1.4) are single crystal pieces cut from larger single crystal ingots and polished. Silicon is extremely strong, on a par with steel, and it also retains its elasticity to much higher temperatures than metals. However, single crystalline (also known as monocrystalline) silicon wafers are fragile: once a fracture starts, it immediately develops across the wafer because covalent bonds do not allow dislocation movements. Many microfabrication disciplines use silicon for convenience: it is available in a wide variety of sizes and resistivities; it is smooth, flat, mechanically strong and fairly cheap. Most of the machinery for microfabrication was originally developed for silicon ICs and newer technologies ride on those developments.

Single crystalline substrates include silicon, quartz (crystalline SiO_2), gallium arsenide (GaAs), silicon carbide (SiC), lithium niobate ($LiNbO_3$) and sapphire (Al_2O_3). Polycrystalline silicon is widely used in solar cell production. Amorphous substrates are also common: glass (which is SiO_2 mixed with metal oxides like Na_2O), fused silica (pure SiO_2; chemically it is identical to quartz) and alumina (Al_2O_3) are used in microfluidics, optics and microwave circuits, respectively. Sheets of polyimides, acrylates and many other polymers are also used as substrates. Substrates must be evaluated for available sizes, purities, smoothness, thermal stability, mechanical strength, etc. Round substrates are compatible with silicon, but square and rectangular ones need special processing because tools for microfabrication are geared for round silicon wafers.

More functionality is built on the substrates by deposition (and further processing) of thin films: various conducting, semiconducting, insulating, transparent, superconducting,catalytic, piezoelectric and other layers are deposited on the substrates. Thin films for microfabrication include a wide variety of elements: metals of common usage include aluminum, copper, tungsten, titanium, nickel, gold and platinum. Metallic alloys and compounds commonly encountered include Al–0.5% Cu, TiW, titanium silicide ($TiSi_2$), tungsten silicide (WSi_2) and titanium nitride (TiN). The most common dielectric thin films are silicon dioxide (SiO_2) and silicon nitride (Si_3N_4). Other dielectrics include aluminum oxide (Al_2O_3), hafnium dioxide (HfO_2), diamond, aluminum nitride (AlN) and many polymers.

A special case of thin-film deposition is epitaxy: the deposited film registers the crystalline structure of the underlying substrate, and, for example, more single crystal silicon can be deposited on a silicon wafer but with different dopant atoms and different dopant concentration. The general material structure of a microfabricated devices is shown in Figure 1.5. Interfaces between the thin film and bulk, and between films, are important for the stability of structures. Wafers experience a number of thermal treatments during their fabrication, and various chemical and physical processes are operative at interfaces, for example chemical reactions and diffusion. Sometimes reactions between films are desired, but most often they should be prevented. This can be achieved by adding extra films, known as barriers, in between films (fig.1.5,b).

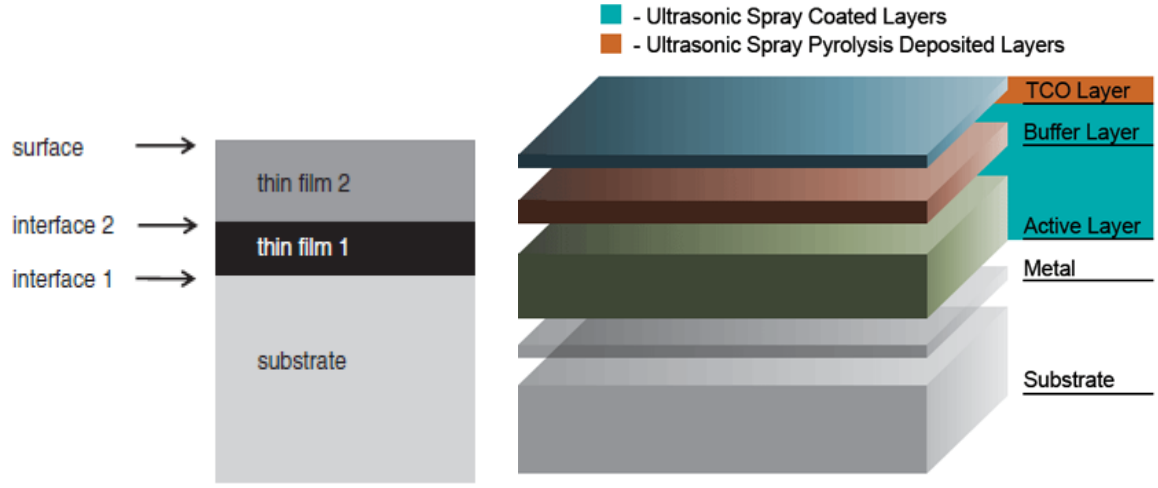


Figure 1.5 Materials and interfaces in a schematic microstructure (a) and thin film solar cels based on CIGS (b)

For example, thin film 1 might present an aluminum conductor and thin film 2 the passivation layer of silicon nitride; or films 1 and 2 are antireflective and scratchresistant coatings in optics; or film 1 is thin tunnel oxide and film 2 a charge storage layer (as in memory cards). Surface physical properties like roughness and reflectivity are material and fabrication process dependent. The chemical nature of the surface is important: some surfaces are reactive, others passive. Many surfaces will be covered by native oxide films if left unattended for some time: for example, silicon, aluminum and titanium form surface oxides over a time scale of hours. Water vapor adsorbed on surfaces must be eliminated before the wafers are processed further. Thin films can be deposited both on flat (planar wafers) and over 3D strips (Figure 1.6).

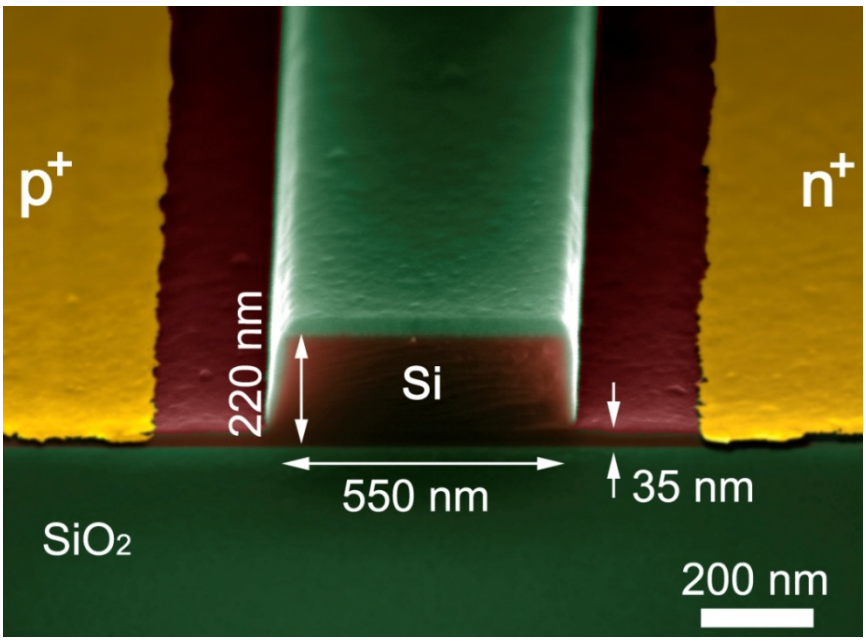


Figure 1.6. The optical modulator uses “silicon nanophotonic waveguides,” to control the flow of light on a silicon chip. The waveguides are made of tiny silicon strips

1.3 Processes of microfabrications

Microfabrication processes consist of four basic operations:

1. Surface preparation and wafer cleaning.

2. High-temperature processes to modify the substrate.

3. Thin-film deposition on the substrate.

4. Patterning of thin films and the substrate.

5. Bonding and layer transfer.

Under each basic operation there are many specific technologies, which are suitable for certain devices, substrates, linewidths or cost levels. Surfaces are modified by etching away a few atomic layers, or by depositing one molecular layer. Surface preparation requirements are widely different in different process steps: in wafer bonding it is paramount to eliminate particles that would create voids if left between the wafers, while in oxidation it is important to eliminate metallic contamination and in epitaxy to ensure that native oxides are removed.

High-temperature steps are used to oxidize silicon and to dope silicon by diffusion, and they are crucial for making transistor, diodes and other electronic devices. Devices like piezoresistive pressure sensors also rely on high-temperature steps, with epitaxy and resistor diffusion as the key processes. The high-temperature regime in microfabrication is typically in 900 ^{o}C to 1200^{o}C, temperatures where dopants readily diffuse and the silicon oxidation rate is technically relevant. Many chemical and physical processes are exponentially temperature dependent. The Arrhenius equation rate ~ $e(-E_a/kT)$ is a very general and very useful description

of the rates of thermally activated processes. Activation energy E_a can be illustrated as a jumping process over a barrier. According to the Boltzmann distribution, an atom at temperature T has an excess of energy E_a with a probability exp $(-E_a/kT)$.

In etching reactions, the activation energy is below 1 eV, in polysilicon chemical vapor deposition E_a is 1.7 eV, in substitutional dopant diffusion it is 3.5–4 eV and in silicon self-diffusion 5 eV. For a silicon etching process with 0.7 eV activation energy, raising the temperature from 20 to 40 ^oC results in a rate six times higher. A great many microfabrication processes show Arrhenius-type dependence: etching, resist development, oxidation, epitaxy, chemical vapor deposition (which are chemical processes) are all governed by exponential temperature dependencies, as are diffusion, electromigration and grain growth (which are physical processes). Low-temperature processes leave metal-to-silicon interfaces stable, and generally 450 ^oC is regarded as the upper limit for low temperatures. Between 450 and 900 ^oC there is a middle range which must be discussed with specific materials and interfaces in mind.

Thin-film steps do not affect the dopant distribution inside silicon; that is,diodes and transistors are unaffected by them. Processes act on whole wafers – this is the basic premise. The whole wafer is subject to, for instance, diffusion from the gas phase, and metal is evaporated everywhere. Either selected areas must be protected by masks before the process, or else the material must be removed from selected areas afterward, by etching or polishing. Patterning processes define structures usually in two steps: polymer processing to form an intermediate pattern which then acts as a mask for etching, deposition, ion implantation or other modification of the underlying material; and after the pattern has been transferred to solid material, the intermittent polymer mask is removed.

The main patterning technique in microfabrication is optical lithography, also known as photolithography. In Figure 1.7 photolithography is shown side by side with the thermal imprint/embossing process. In both processes a polymer film is modified locally to create patterns. In lithography, photosensitive polymer film is exposed to UV light, which hardens the polymer by crosslinking (so-called negative resists). In imprinting, a thermoplastic polymer softens upon heating, and a master stamp is pressed against it. The system is allowed to cool down before the stamp is released, and then the polymer retains its imprinted shape. Many old methods have been successfully scaled down to micrometer and nanometer scales. For example, the metal etching with similar acidic solutions can make aluminum patterns in the micrometer range. Once an original microstamp or nanostamp has been made, its replication into polymers is fairly easy. Electroplating is likewise easily applicable to nanometer structures. Casting polymers into micromolds is also popular in microfabrication: the elastomeric (rubber-like) material PDMS (poly(dimethyl)siloxane) is a favorite material for simple microfluidic devices.

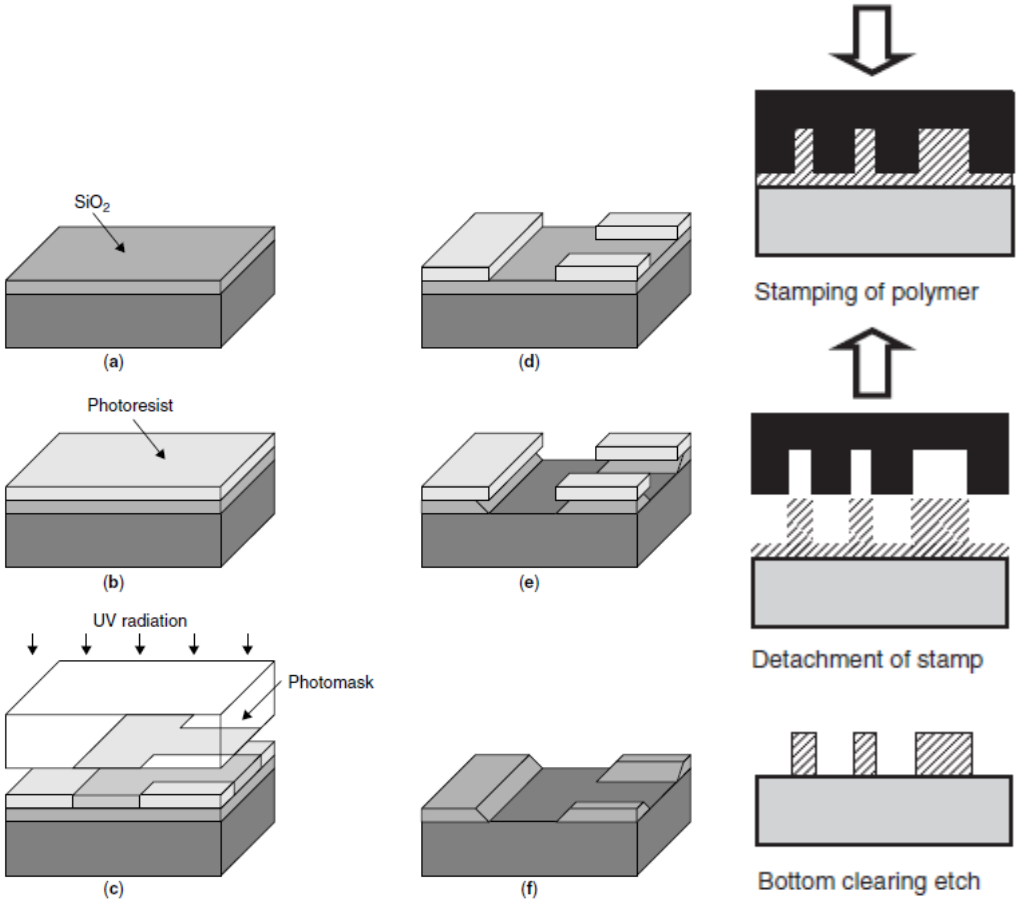


Figure 1.7 (left): Optical lithography patterning process: (a) oxide-film deposition; (b) photoresist application; (c) UV exposure through a photomask; (d) development of resist image; (e) etching of oxide and (f) photoresist removal; and thermal imprint (right): the softened polymer is forced to shape, and after cooling the shape is retained even though the masteris removed. In imprinting, some material remains at the bottom and must be cleared by etching.

Wafer bonding and layer transfer enable more complex structures to be made. Bonding a wafer on top of a trench turns it into a channel, useful for microfluidics. Bonding more wafers can lead to elaborate fluidic channel patterns, as in the burner of a flame ionization detector, Figure 1.8. Bonding two wafers with electrodes creates a capacitor, for instance for pressure sensing. Bonding two different wafers can also be used simply as a method to create a new kind of a starting wafer, with the best properties of the two wafers combined. These elementary operations of patterning, modification, deposition and bonding are combined many times over to create devices. Process complexity is often discussed in terms of the number of lithography steps (the term mask levels is also used): five lithography steps are enough for a simple MOS transistor and many MEMS, flat-panel display devices can be made with two to six photolithography steps, but 32 nm linewidth microprocessors and logic circuits require over 30 patterning steps.

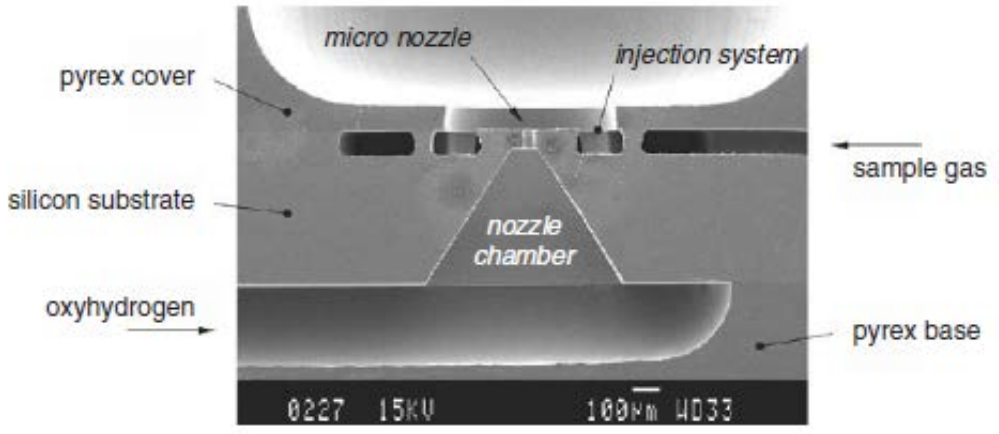


Figure 1.8 Oxyhydrogen burner of a flame ionization detector by Pyrex–glass/silicon/Pyrex–glass bonding.

Microfabricated systems have minimum dimensions from few nm to 50 μm, depending on the device types. Advanced microprocessors and memories and the read/write heads of hard disk drives must have features <100 nm to be competitive. In Figure 1.9 the SEM micrograph shows the cross-section of a 50 nm MOS gate. Many other electronic devices like RF and power transistors make do with 100 nm to 1 μm dimensions. MEMS devices typically have 1–10 μm minimum lines and microfluidic devices might have 50 μm as thesmallest feature.

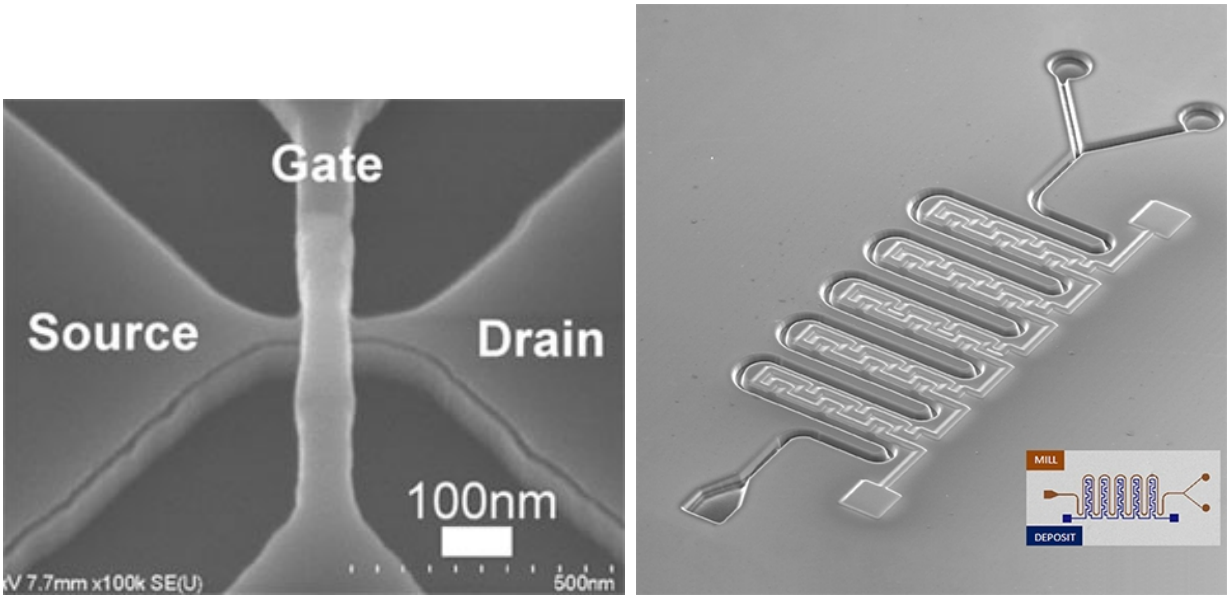
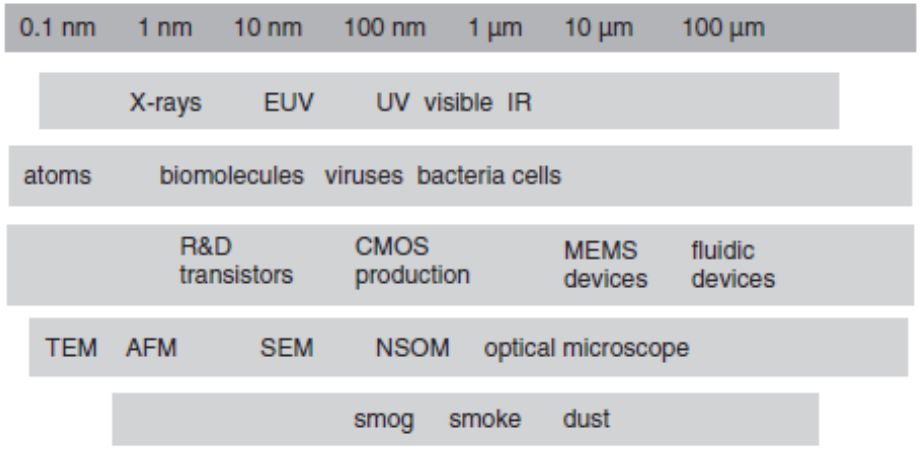


Fig. 1.9 SEM image of an upright-type double-gate MOS transistor (a) and 50 μm microfludic device.

Microfabricated device sizes are compared to physical, chemical and biological small objects in Figure 1.10, with microscopy methods capable of observing them rate. Chemical vapor deposition (CVD) can be used for anything from a few nanometers to a few micrometers. Sputtering also produces films from 0.5 nm to 5 μm. Spin coating is able to produce films as thin as 100 nm, or as thick as 100 μm. Typical applications include polymer spinning. Electroplating (galvanic deposition) can produce metal layers of almost any thickness, from a few nanometers up to hundreds of micrometers.



.

Figure 1.10. Dimensions in the microworld: electromagnetic radiation, natural objects, humanmade devices, microscopy methods and dirt

But almost every device includes structures with dimensions of about 100 μm. These are needed to interface the microdevices to the outside world: most devices need electrical connections (by a wire-bonding or bumping process); microfluidic devices must be connected to capillaries or liquid reservoirs; solar cells and power semiconductors must have thick and large metal areas to bring in and take out the high currents involved; and connections to and from optical fibers require structures about the size of fibers, which is also on the order of 100 μm.

1.4 Devices

Microfabricated device can be classified on device material or functionality:

• material: silicon, III–V, wide band gap (SiC, diamond), polymer, glass

• integration: monolithic integration, hybrid integration, discrete devices

• active vs. passive: transistor vs. resistor, valve vs. sieve

• interfacing: externally (e.g., sensor) vs. Internally (e.g., processor).

Microfabricated device can be classified on fabrication technologies:

• volume (bulk) devices

• surface devices

• thin-film devices

• stacked devices.

Power transistors, thyristors, radiation detectors and solar cells are volume devices (fig.1.11): currents are generated and transported (vertically) through the wafer, or, alternatively, device structures extend through the wafer, as in many bulk micromechanical devices. The starting wafers for volume devices need to be uniform throughout. Patterns are often made on both sides of the wafer and it is important to note that some processes affect both sides of the wafer and some are one sided.

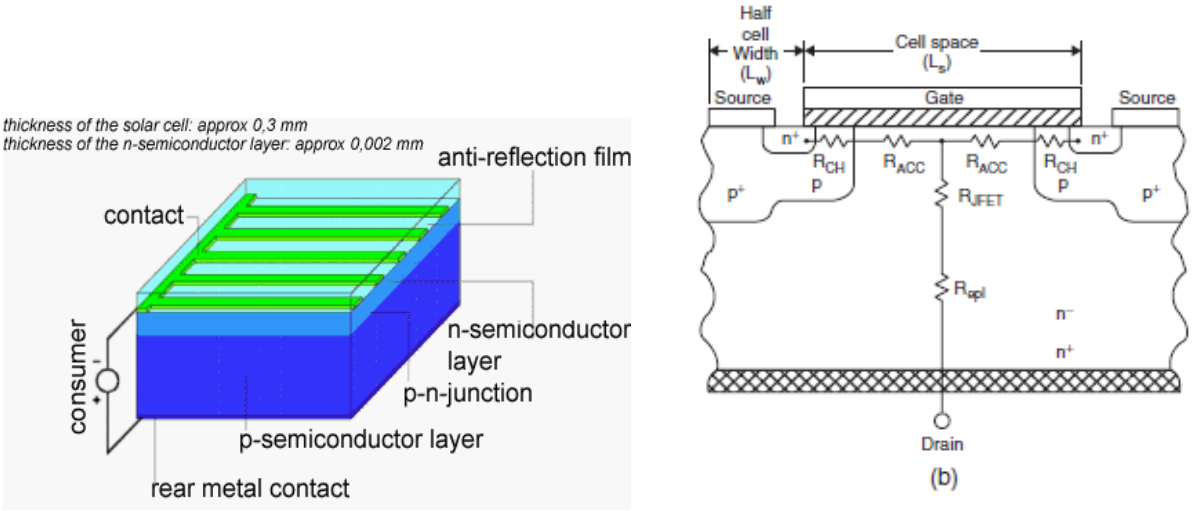


Figure 1.11 Volume devices: (a) passivated emitter, rear locally diffused solar cell (b) n channel power MOSFET cross-section

Surface devices make use of the material properties of the substrate but generally only a fraction of wafer thickness is utilized in making the devices. However, device structure or operation is connected with the properties of the substrate. Most ICs fall under this category: namely, MOS and bipolar transistors, photodiodes, CCD image sensors as well as III–V optoelectronic devices. In silicon CMOS, only the top 5 μm layer of the wafer is used in making the active devices, the remaining 500 μm of wafer thickness being for support: that is, mechanical strength and impurity control. Shown in Figure 1.12,a are CMOS polysilicon gates of 0.5 μm width and 0.25 μm height. Surface devices can have very elaborate 3D structures, like multilevel metallization in logic circuits, which can be 10 μm thick, but this is still only a fraction of wafer thickness; therefore the term surface device applies. Devices can be built by depositing and patterning thin films on the wafers, where the wafer has no role in device operation. Thin-film transistors (TFTs) are most often fabricated on non-semiconductor substrates of glass, plastic or steel. Devices like RF switches and relays, optical modulators often fabricated on silicon wafers for convenience, but they could be fabricated on glass or polymer

substrates as well. Figure 1.12,b shows a RF switch: the silicon nitride/gold thin film flap curls up because of film stresses, but can be forced flat by electrostatic actuation.

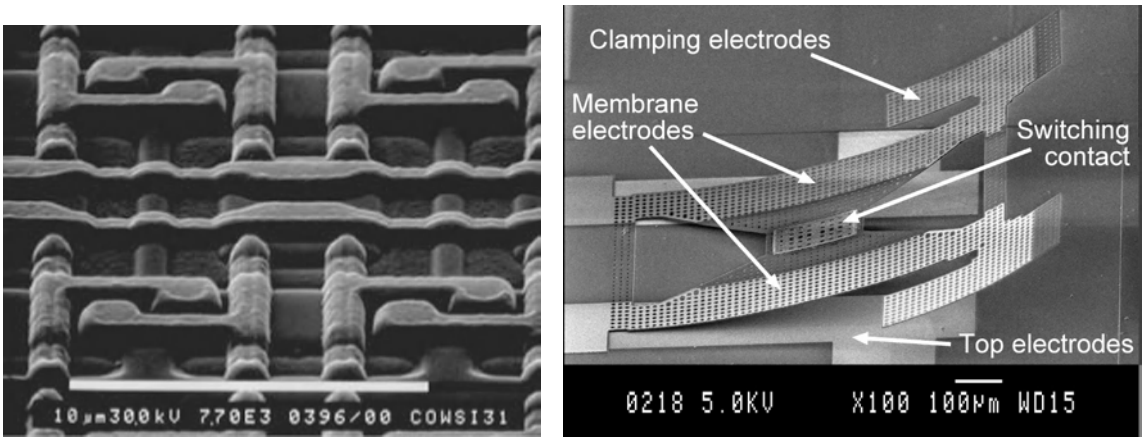


Figure 1.12 Surface devices: 0.5 μm minimum linewidth MOS in a scanning electron microscope (SEM) view (a) and RF switch (b).

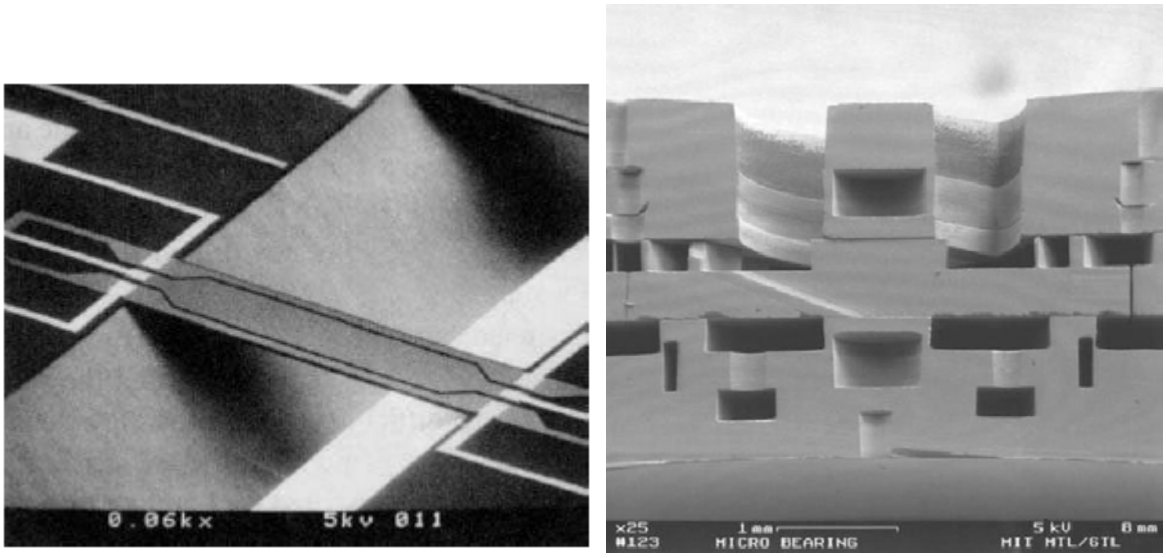


Figure 13 (a) Mass flow sensor: a resonating bridge over an etched channel, (b) A microturbine by five-wafer silicon-tosilicon bonding.

Membrane devices are a sub-class of thin-film devices: again, all functionality is in the thin top layer, but instead of full wafer mechanical support, only a thin membrane supports the structures. Many thermal devices are membrane devices for thermal isolation: thermopiles, bolometers, chemical microreactors and mass flow meters (Figure 1.13,a). Many acoustic devices also utilize bulk removal. Optical paths can be opened by removing the bulk semiconductor. X-ray lithography masks are gold or tungsten microstructures on a micrometrethick membrane. Stacked devices are made by layer transfer and bonding techniques. Two or more wafers are joined together permanently. Devices with vacuum cavities, for example absolute pressure sensors, accelerometers and gyroscopes, are stacked devices made of bonded silicon/glass wafer pairs. Micropumps and valves are typically stacks of many wafers. Figure

1.13,b shows a microturbine. It is made by bonding together five wafers. More and more layer transfer and wafer bonding techniques are being developed, and stacked devices of various sorts are expected to be appear, for example GaAs optical devices bonded to Si based electronics, or MEMS devices bonded to ICs.

The MOS transistor is a capacitor with a silicon substrate as the bottom electrode, the gate oxide as the capacitor dielectric and the gate metal as the top electrode (Figure 1.14). The MOS transistor has been the driving force of the microfabrication industries. It is the top device by all measures: number of devices sold, the narrowest linewidths and the thinnest oxides in mass production, as well as dollar value of production. Most equipment for microfabrication was originally designed for MOS IC fabrication, and later adapted to other applications.

Despite the name MOS, the gate electrode is usually made of phosphorus-doped polycrystalline silicon, not of metal. The basic function of a MOS transistor is to control the flow of electrons from the source to drain by the gate voltage and the field it generates in the channel. In a NMOS transistor, a positive voltage on the gate pulls electrons from the p-type channel to the Si/SiO_2 interface where an overabundance of electrons inverts the region under the gate to n-type, enabling electrons to flow from the n+ source to the n+ drain. Transistors are isolated electrically from neighboring transistors by SiO_2 field oxide areas. This isolation takes up a lot of area, and therefore the transistor packing density on a chip does not depend on transistor dimensions alone.

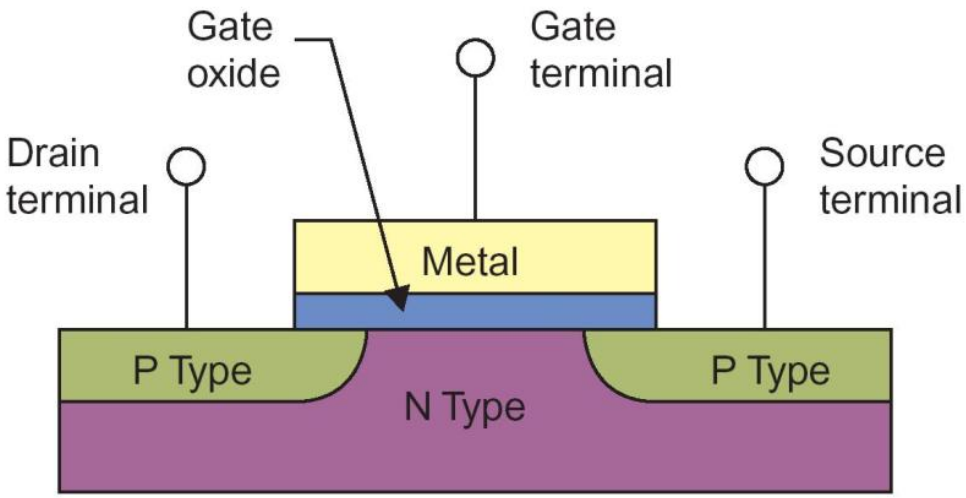


Figure 1.14 Schematic of a MOS transistor: gate, source (S) and drain (D) in an active area defined by thick isolation oxide

1.5 Main tendencies of microfabrication development

Nowadays the Microsystems technology focuses on the *miniaturization* of engineering systems to accommodate design specifications of small space, light weight and enhanced

portability. An additional advantage of such portable systems is their wide-scale utility in distributed transducer networks. The importance of MST lies, for a large part, in the economical and technical development of innovative systems that it makes possible.

The evolution of microelectronic devices is influenced by factors such as growing demands in memory capacity, high transmission data speed, optical communications, etc. This requires electronic devices with faster speed operation and smaller size. The first object of miniaturization was the integrated transistor, the workhorse device by means of which major new markets were created. For example, information and communication technology (ICT) relies on the technical principles of miniaturization by integrating more and more electronic functional elements into the same restricted area of a silicon die, the chip. Complementing this chip with a large data storage capacity that has fast read/write access and a high-definition display has given rise to systems which have penetrated all layers of personal and professional human lives. These types of devices are a smart combination of millions of transistors on a single chip, produced on dedicated microelectronic production lines.

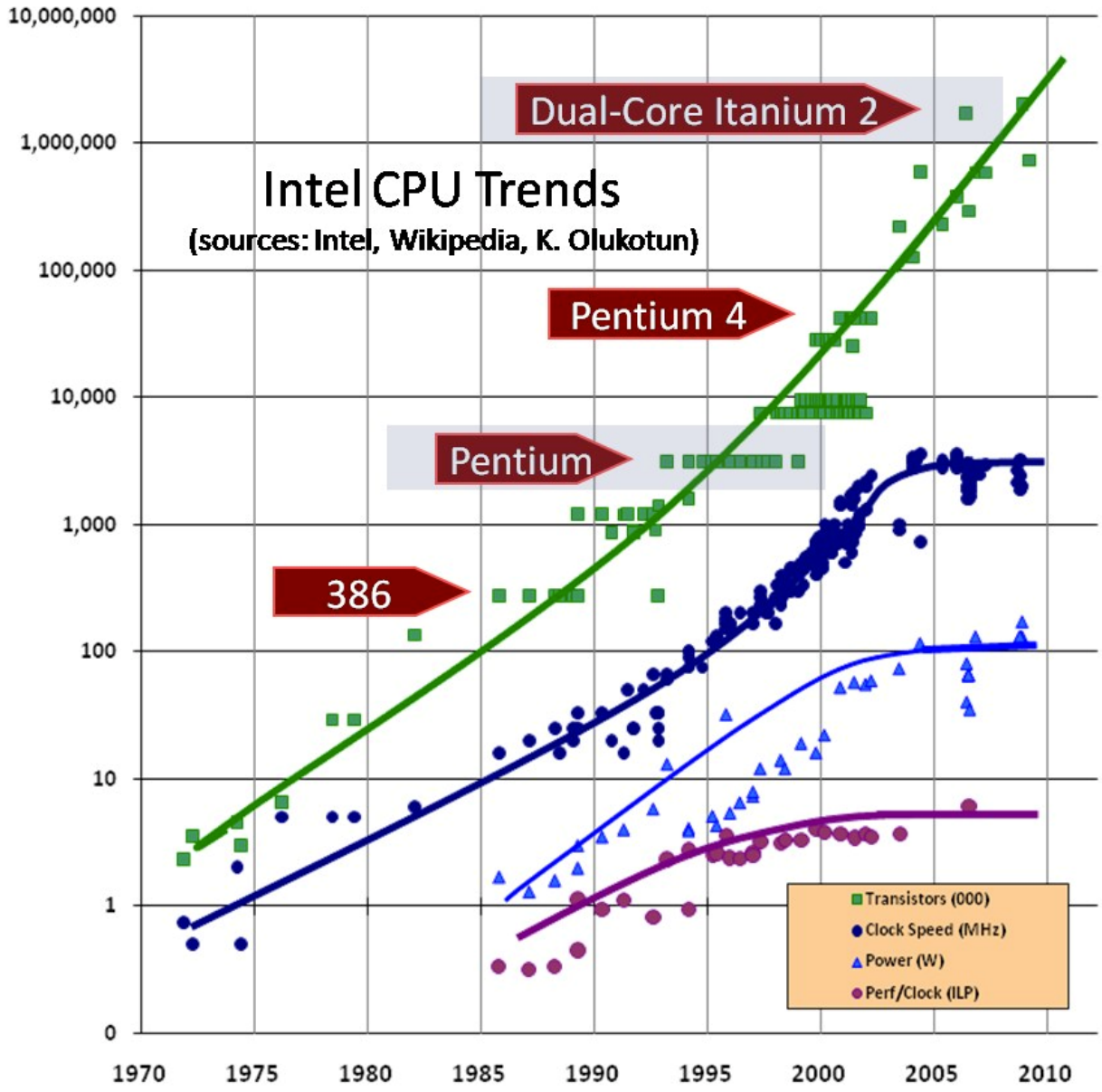


Figure 1.15. Plot of CPU transistor counts against dates of introduction. The line corresponds to exponential growth, with transistor count doubling every 2 years

Silicon-based semiconductor device dimensions have been scaled continuously over the last 40 years. Current CMOS-based technologies have device dimension in the sub-100-nm range with gate dielectric thicknesses in the 1- to 2-nm range. Such advances largely followed the industry's governing tenet, i.e., Moore's law, which states that the number of transistors on a chip doubles every 2 years, as shown in Figure 1.15. To meet these needs, device dimension have shrunk 0.7 times per generation to improve performance by doubling frequency and reducing gate delay. Table 1 shows the scaling of typical device dimensions for different CMOS Generations. For the 70-nm technology node, the typical gate length is only 35 nm while

the electrical gate oxide thickness (t_{OX}) is 1.6 nm and the source drain extension (SDE) depth is 17 nm.

Table 1.1

Scaling Projection of Transistor Parameters for Different Technology Generation Levels

Parameter	Generation Level (nm)				Scaling Factor
	180	130	100	70	
L_{GATE} (nm)	100	70	50	35	0.7×
V_{DD} (V)	1.5	1.2	1.0	0.8	0.8×
t_{ox} (e) (nm), t_{ox} (phys) (nm)	3.1, 2.1	2.5, 1.5	2.0, 1.0	1.6, 0.6	0.8×
SDE depth (nm)	50	35	24	17	0.7×
SDE under diff (nm)	23	16	11	8	0.7×
L_{MET} (nm)	55	40	27	20	0.7×
Channel doping ($\times 10^{18}$ cm^{-3})	1	1.6	2.6	4	$1/(0.8)^2 = 1.6\times$
I_{DSAT} (relative)	1	1	1	1	1×
I_{OFF} (nA/μm), 25°C	20	40	80	160	2×

Figure 1.16 shows the reduction of feature size of metal-oxide-semiconductor (MOS) transistors for dynamic random access memories (DRAMs) , as well as the number of bits per chip for the period 1970–2000. For example, a 256 M-bit DRAM contains about 10^9 transistors with a feature size L close to 100 nm. For structures with these dimensions, transport can still be treated classically, but we are already at the transition regime to quantum transport. Today it is believed that present silicon technology will evolve towards feature sizes still one order of magnitude lower, i.e. L~10 nm; but below this size, transistors based on new concepts like single electron transistors, resonant tunnelling devices, etc. will have to be developed. The operation of this new kind of devices has to be described by the concepts of mesoscopic and quantum physics.

Modern microprocessors are manufactured with billions of transistors. Keeping power dissipation, variability, and reliability under control is therefore critical. Just, we briefly consider the *Technology Scaling Challenges* as main problem at fabrications of IC.

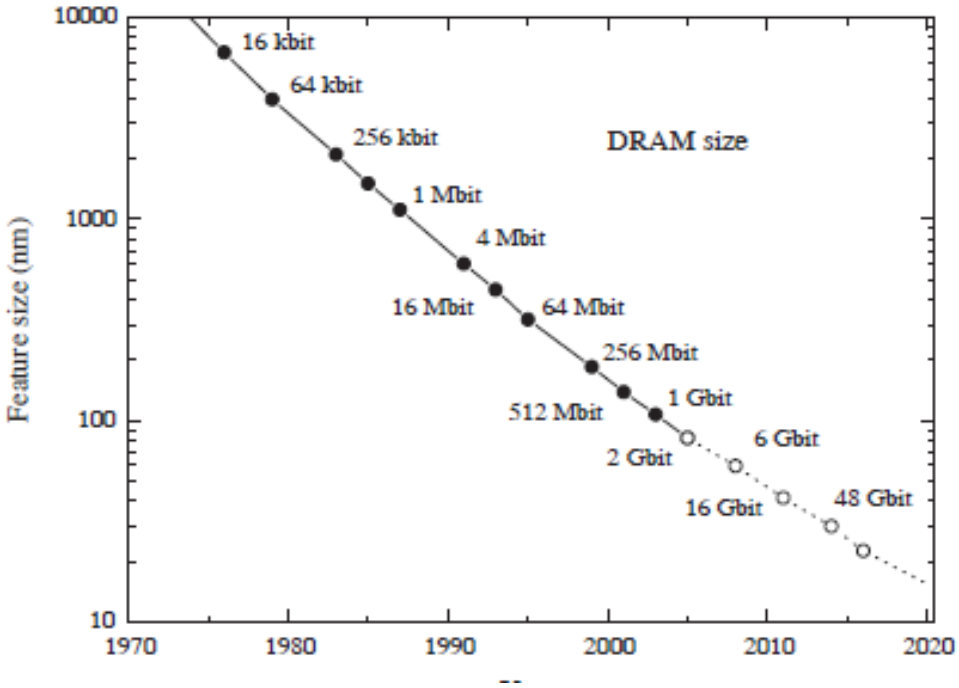


Figure 1.16. Evolution of the minimum feature size of a Si DRAM

1.6 Main problems at fabrications of IC

Power Dissipation

The active power dissipated in a CMOS chip is given by eq.1

$$Power = C \cdot V_{DD}^2 \cdot f \qquad (1)$$

where, C is the capacitance, V_{DD} is the supply voltage, and f is the frequency of the circuit. There are two approaches to technology scaling: (1) constant power scaling where V_{DD} is not scaled and a reduction in capacitance is negated by an increase in f, and (2) reducing power where dissipation by scaling V_{DD} by 0.7 times, leading to a 50% reduction in active power for the scaled technology. V_{DD} scaling directly impacts the gate delay, and increases sub-threshold leakage current, which in turn increases (static) power dissipation due to leakage. On the other hand, constant power scaling leads to gate leakage increase as the physical gate oxide thickness is scaled continuously to meet the performance requirements.

Sub-Threshold Leakage

The sub-threshold current flows from the source to the drain of a transistor due to the diffusion of the minority carriers for gate-to-source voltages (V_{GS}) below the threshold voltage (V_{TH}). It depends exponentially on both V_{GS} and V_{TH} and is a strong function of temperature. Ideally, the ratio of V_{TH}/V_{DD} is kept below 0.25 so that the gate overdrive capability of the scaled device can be maintained and CMOS circuit performance is not compromised. In short-channel devices, source and drain depletion regions penetrate significantly into the channel and control the potential and the field inside the channel. This is known as the short channel effect (SCE).

As a result of SCEs, V_{TH} reduces via (1) a reduction in channel length, and (2) an increase in drain bias (drain induced barrier lowering). This results in increased sub-threshold currents in short-channel devices. In order to keep SCEs under control, both the gate oxide thickness and the depletion width of the transistor must be reduced. The latter requires tailoring of the channel doping profile by implanting retrograde wells while the former directly leads to reliability challenges associated with gate oxide thickness scaling.

Gate Leakage

Gate leakage increases exponentially with a decrease in the gate oxide thickness and an increase in the potential drop across oxide. It exhibits a weak temperature dependence. Gate current is primarily due to the tunneling of electrons (or holes) from the silicon bulk through the gate oxide potential barrier into the gate (or vice versa). Figure 1.17 shows how reducing the gate oxide thickness leads to an increase in tunneling current. Gate leakage is critical during the off-state of the devices and results in the standby power dissipation of the chip. One possible solution is to use dielectric films with higher dielectric constants (known as high-K materials, e.g., HfO_2 or Al_2O_3 etc.) such that the physical gate stack thickness can be increased, leading to a lower gate leakage current.

Transistor Reliability

Reliability is the probability that a product will perform a required function under stated conditions for a stated period of time. A typical example of reliability is gate oxide integrity. An oxide is defined as reliable if it maintains its insulating properties for 10 or 25 years at a specified bias, temperature, chip area, and failure fraction. Reliability studies typically require accelerated testing conditions such that the physical mechanism responsible for breakdown can be studied in a time frame much shorter than the targeted lifetime. Intrinsic reliability studies revolve around generation of material defects that lead to product failure. Since defect generation is random in nature, the statistical nature of defect generation and its impact on reliability must be understood. Time-dependent dielectric breakdown (TDDB) occurs during the off-state when the voltage across the gate dielectric is high. TDDB failure was traditionally catastrophic and caused the gate dielectric to lose its insulating properties after the breakdown event, leading to a functional failure of the chip. As technology is scaled downward, TDDB is no longer automatically considered catastrophic since the dielectric does not fully lose its insulating properties for sufficiently thin gate oxides. Bias temperature instability (BTI) occurs during an off-state condition with a uniform field across the oxide. It causes a shift in FET parameters such as threshold voltage (V_{TH}), saturation regime drain current (IDSAT). BTI is a major challenge as it occurs at low fields and is enhanced at higher temperatures.

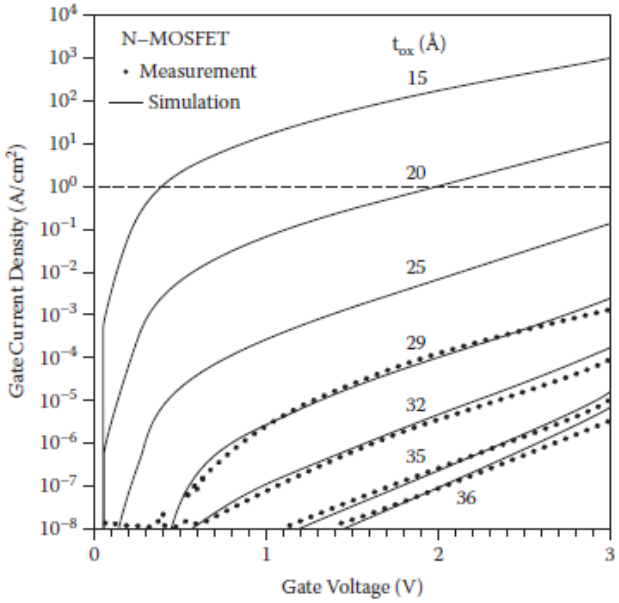


Figure 1.17. Measured and simulated I_G–V_G characteristics under inversion conditions for different oxide thicknesses.

Hot Carrier Injection

Hot carrier injection (HCI) occurs during the on-state condition with a high voltage on the drain that leads to a non-uniform field across the oxide. As with BTI, HCI also causes a shift in FET parameters such as threshold voltage (V_{TH}), saturation regime drain current (IDSAT). HCI modeling and data have changed as the technology has scaled. Hot carriers are generated by a high lateral electric field in the channel. When the mean kinetic energy of the carrier is higher than the lattice temperature, a carrier is "hot." The generated hot carriers can be injected into the oxide causing bulk defect generation or charge trapping. Typically the damage due to HCI is highly localized. HCI increases as the channel length (LG) is reduced. To reduce the HCI-induced device parameter shift, the lightly doped drain (LDD) was introduced. The main goal was to reduce the peak lateral electric field as the technology was scaled.

Back End of Line: Interconnect Technology

Advanced integrated circuits (ICs) require elaborate wiring (or interconnect) systems to distribute power, grounding, and various clock and input and output (I/O) signals to and from transistor devices. To maintain the cost and performance benefits associated with reduced transistor feature size and higher on-chip device density, the interconnect architecture must correspondingly increase in complexity and density; this is achieved by reducing the geometrical dimensions of the wirings and increasing the number of interconnect layers. The downsides to the "interconnect scaling," however, are an increase in wiring resistance (due to the smaller cross-sectional area) and the parasitic capacitance of wires that exert serious impacts on dynamic power consumption, self-heating, and signal propagation speeds in the form of increased resistance–capacitance (RC) delay.

Chapter 2. Crystal Growth

A. Evtukh

2.1. Introduction

The two most important semiconductors for discrete devices and integrated circuits are silicon and gallium arsenide. The common techniques for growing single crystals of these two semiconductors will be described in this chapter. The basic process flow from starting materials to polished wafers is shown in Fig. 2.1 [1,2].

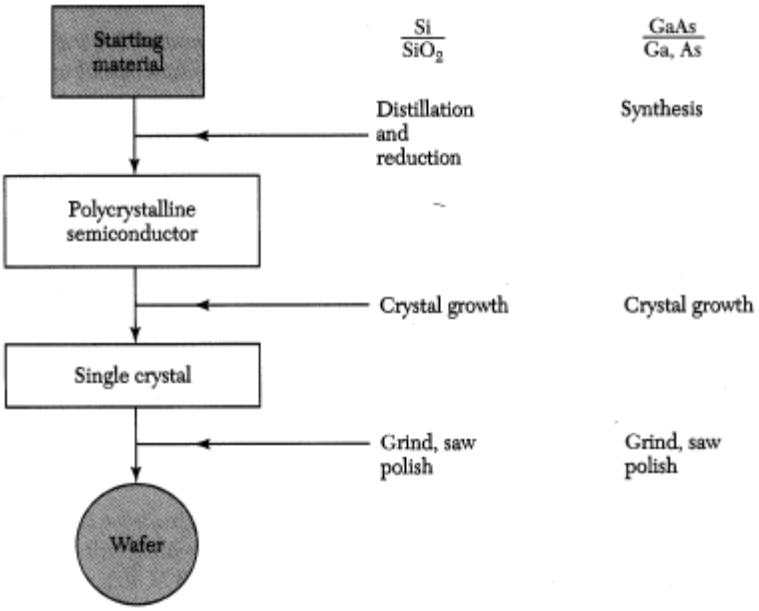


Fig. 2.1. Czochralski crystal puller.

The starting materials, silicon dioxide for a silicon wafer and gallium and arsenic for a gallium arsenide wafer, are chemically processed to form a high-purity polycrystalline semiconductor from which single crystals are grown. The single-crystal ingots are shaped to define the diameter of the material and sawed into wafers. These wafers are etched and polished to provide smooth, seculars surfaces on which devices will be made. Specifically, the following topics we cover will be covered: Basic techniques to grow silicon and GaAs single-crystal ingots. Wafer-shaping steps from ingots to polished wafers. Wafer characterization in term of its electrical and mechanical properties.

2.2. Silicon Crystal Growth from the Melt

The basic technique for silicon crystal growth from the melt, which is material in liquid form, is the Czochralski technique. A substantial percentage (> 90%) of the silicon crystals for

the semiconductor industry is prepared by the Czochralski technique, and virtually all the silicon used for fabricating integrated circuits is prepared by this technique.

2.2.1. Starting material

The starting material for silicon is a relatively pure form of sand (SiO_2) called quartzite. This is placed in a furnace with various forms of carbon (coal, coke, and wood chips). Although a number of reactions take place in the furnace, the overall reaction is

$$SiC\ (solid) + SiO_2\ (solid) \rightarrow Si\ (solid) + SiO\ (gas) + CO\ (gas). \quad (2.1)$$

This process produces metallurgical-grade silicon with a purity of about 98%. Next, the silicon is pulverized and treated with hydrogen chloride (HCI) to form trichlorosilane ($SiHC1_3$):

$$Si\ (solid) + 3HC1\ (gas) \rightarrow^{300°C} SiHC1_3\ (gas) + H_2\ (gas). \quad (2.2)$$

The trichlorosilane is a liquid at room temperature (boiling point 32°C); Fractional distillation of the liquid removes the unwanted impurities. The purified $SiHCl_3$ is then used in a hydrogen reduction reaction to prepare the electronic-grade silicon (EGS):

$$SiHC1_3\ (gas) + H_2\ (gas) \rightarrow Si\ (solid) + 3HC1\ (gas). \quad (2.3)$$

This reaction takes place in a reactor containing a resistance-heated silicon rod, which serves as the nucleation point for the deposition of silicon. The EGS, a polycrystalline material of high purity, is the raw material used to prepare device-quality, single-crystal silicon. Pure EGS generally has impurity concentrations in the parts-per-billion range [3].

2.2.2. The Czochralski technique

The Czochralski technique uses an apparatus called a crystal puller. A simplified version is shown in Fig. 2.2. The puller has three main components: (*a*) a furnace, which includes a fused-silicon (SiO_2) crucible, a graphite susceptor, a rotation mechanism (clockwise as shown), a heating element, and a power supply; (*b*) a crystal-pulling mechanism, which includes a seed holder and a rotation mechanism (counter-clockwise); and (*c*) an ambient control, which includes a gas source (such as argon), a flow control, and a exhaust system. In addition, the puller has an overall microprocessor-based control system to control process parameters such as temperature, crystal diameter, pull rate, and rotation speeds, as well as to permit programmed process steps. Also, various sensors and feedback loops allow the control system to respond automatically, reducing operator intervention.

In the crystal-growing process, polycrystalline silicon (EGS) is placed in the crucible and the furnace is heated above the melting temperature of silicon. A suitably oriented seed crystal (e.g., <111>) is suspended over the crucible in a seed holder. The seed is inserted into the melt. Part of it melts, but the tip of the remaining seed crystal still touches the liquid surface. It is then

slowly withdrawn. Progressive freezing at the solid-liquid interface yields a large, single crystal. A typical pull rate is a few millimeters per minute. For large-diameter silicon ingots, an external magnetic field is applied to the basic Czochialski puller. The purpose of the external magnetic field is to control the concentration of defects, impurities, and oxygen content [4]. Figure 2.3 shows a 300 mm (12 in.) and a 400 mm (16 in.) Czochralski grown silicon ingots.

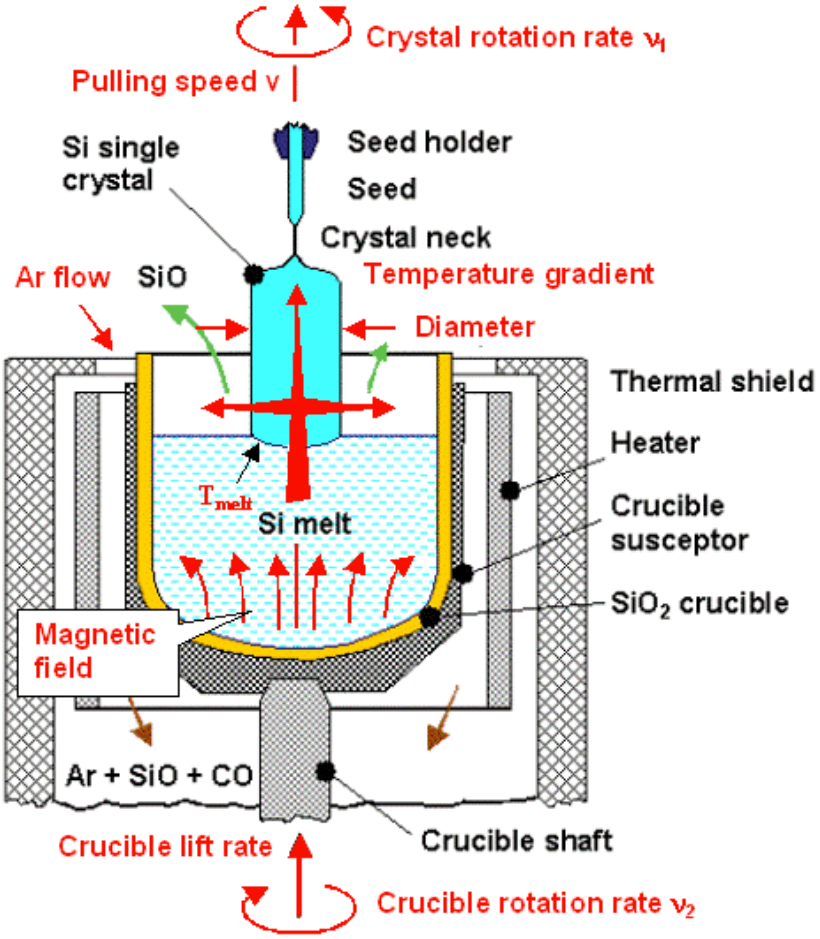


Fig. 2.2. Czochralski crystal puller. CW, clockwise; CCW, counter clockwise.

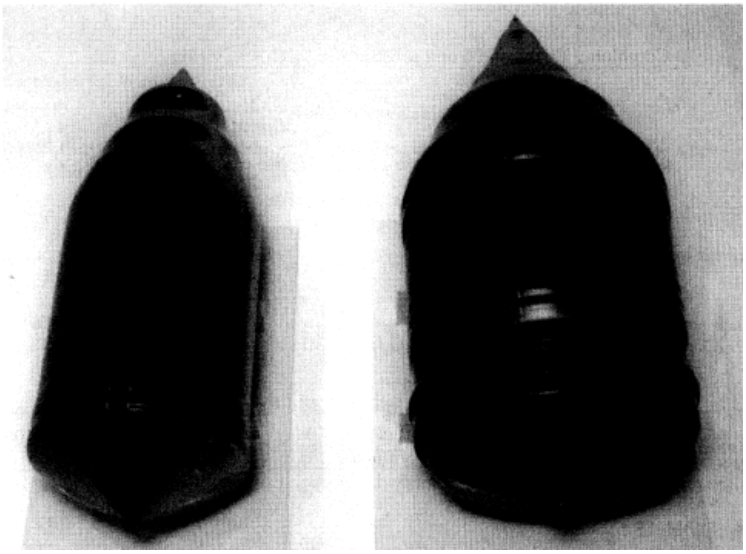


Fig. 2.3. 300 mm (12 in.) and 400 mm (16 in.) Czochralski-grown silicon ingots.

2.2.3. Distribution of Dopant

In crystal growth, a known amount of dopant is added to the melt to obtain the desired doping concentration in the grown crystal. For silicon, boron and phosphorus are the most common dopants for *p*- and *n*-type materials, respectively.

As a crystal is pulled from the melt, the doping concentration incorporated into the crystal (solid) is usually different from the doping concentration of the melt (liquid) at the interface. The ratio of these two concentrations is defined as the equilibrium segregation coefficient k_0**:**

$$k_0 \equiv \frac{C_s}{C_l}, \tag{2.4}$$

where C_s and C_l are, respectively, the equilibrium concentrations of the dopant in the solid and liquid near the interface. Table 2.1 lists values of k, for the commonly used dopants for silicon. Note that most values are below 1, which means that during growth the dopants are rejected into the melt. Consequently, the melt becomes progressively enriched with the dopant as the crystal grows.

Table 2.1. Equilibrium segregation coefficients for dopants in Si

Dopant	k_0	Type	Dopant	k_0	Type
B	8×10^{-1}	*p*	As	3×10^{-1}	*n*
Al	2×10^{-3}	*p*	Sb	2.3×10^{-2}	*n*
Ga	8×10^{-3}	*p*	Te	2×10^{-4}	*n*
In	4×10^{-4}	*p*	Li	1×10^{-2}	*n*
O	1.25	*n*	Cu	4×10^{-4}	-*
C	7×10^{-2}	*n*	Au	2.5×10^{-5}	-*
P	0.35	*n*			

*Deep-lying impurity level.

Consider a crystal being grown from a melt having an initial weight M_0 with an initial doping concentration C_0 in the melt (i.e., the weight of the dopant per 1 g of melt). At a given point of growth when a crystal of weight M has been grown, the amount of dopant remaining in the melt (by weight) is S. For an incremental amount of the crystal with weight dM, the corresponding reduction of the dopant ($-dS$) from the melt is $C_s\,dM$, where C_s is the doping concentration in the crystal (by weight):

$$-dS = C_s dM. \tag{2.5}$$

Now, the remaining weight of the melt is M_0 - M, and the doping concentration in the liquid (by weight), C_l , is given by

$$C_l = \frac{S}{M_0 - M}. \tag{2.6}$$

Combining Eqs. 2.5 and 2.6 and substituting $C_s / C_1 = k_0$ yields

$$\frac{dS}{S} = -k_0 \left(\frac{dM}{M_0 - M}\right). \tag{2.7}$$

Given the initial weight of the dopant, $C_0 M_0$, we can integrate Eq. 2.7:

$$\int_{C_0 M_0}^{S} \frac{dS}{S} = k_0 \int_0^M \frac{-dM}{M_0 - M}. \tag{2.8}$$

Solving Eq. 2.8 and combining with Eq. 2.6 gives

$$C_s = k_0 C_0 \left(1 - \frac{M}{M_0}\right)^{k_0 - 1}. \tag{2.9}$$

Figure 2.4 illustrates the doping distribution as a function of the fraction solidified (M/M_0) for several segregation coefficient [5, 6]. As crystal growth progresses, the composition initially at $k_0 C_0$ will increase continually for $k_0 < 1$ and decrease continually for $k_0 > 1$. When $k_0 \approx 1$, a uniform impurity distribution can be obtained.

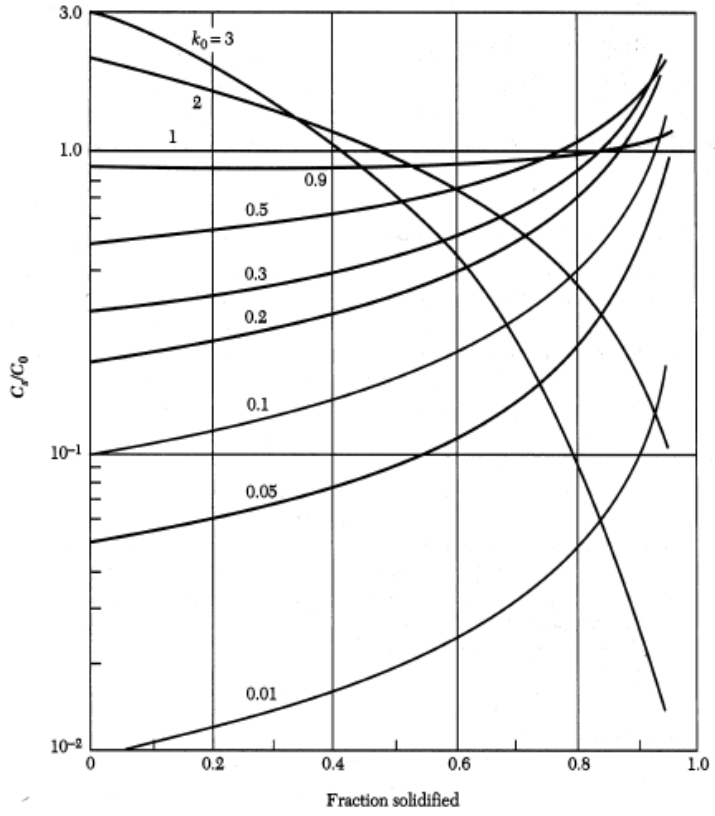


Fig. 2.4. Curves of growth from the melt showing the doping concentration in a solid as a function of the fraction solidified [6].

2.2.4. Effective Segregation Coefficient

While the crystal is growing, dopants are constantly being rejected into the melt (for $k_0 < 1$). If the rejection rate is higher than the rate of which the dopant can be transported away by diffusion or stirring, then a concentration gradient will develop at the interface, as illustrated in Fig. 2.5. The segregation coefficient is $k_0 = C_s/C_l(0)$.We can define an effective segregation coefficient k_e, which is the ratio of C_s and the impurity concentration far away from the interface:

$$k_e \equiv \frac{C_s}{C_l}. \tag{2.10}$$

Consider a small, virtually stagnant layer of melt with width δ in which the only flow is that required to replace the crystal being withdrawn from the melt. Outside this stagnant layer, the doping concentration has a constant value C_l. Inside the layer, the doping concentration can be described by the continuity equation. At steady state, the only significant terms are the second and third terms on the righthand side (we replace n_p by C and $\mu_n E$ by v):

$$0 = v\frac{dC}{dx} + D\frac{d^2C}{dx^2}, \tag{2.11}$$

where D is the dopant diffusion coefficient in the melt, v is the crystal growth velocity, and C is the doping concentration in the melt.

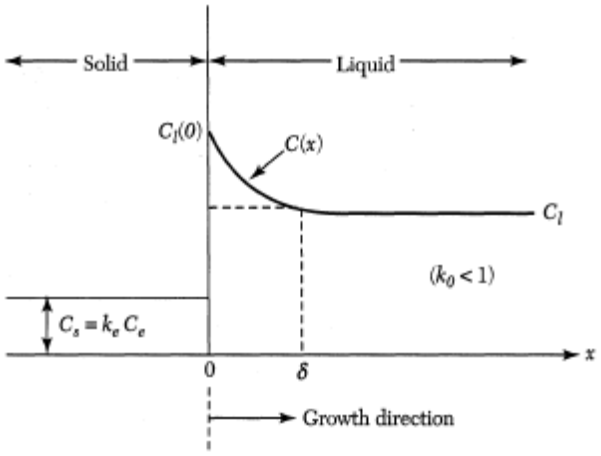


Fig. 2.5. Doping distribution near the solid-melt interface.

The solution of Eq. 2.11 is

$$C = A_1 \exp(-vx/D) + A_2 \tag{2.12}$$

where A_1 and A_2 are constants to be determined by the boundary conditions. The first boundary condition is that $C = C_l$ (0) at $x = 0$. The second boundary condition is the conservation of the total number of dopants; that is, the sum of the dopant fluxes at the interface must be zero. By considering the diffusion of dopant atoms in the melt (neglecting diffusion in the solid), we have

$$D(\frac{dC}{dx})_{x=0} + [C_l(0) - C_s]v = 0. \tag{2.13}$$

Substituting these boundary conditions into Eq. 2.12 and noting that $C = C_l$ at $x = \delta$ gives

$$\exp(-v\delta / D) = \frac{C_l - C_s}{C_l(0) - C_s}. \tag{2.14}$$

Therefore,

$$k_e \equiv \frac{C_s}{C_l} = \frac{k_0}{k_0 + (1-k_0)\exp(-v\delta / D)}. \tag{2.15}$$

The doping distribution in the crystal is given by the same expression as in Eq. 2.9, except that k_0 is replaced by k_e. Values of k_e are larger than those of k_0 and can approach 1 for large values of the growth parameter $v\delta/D$. Uniform doping distribution ($k_e \rightarrow 1$) in the crystal can be obtained by employing a high pull rate and a low rotation speed (since δ is inversely proportional to the rotation speed). Another approach to achieve uniform doping is to add ultra pure polycrystalline silicon continuously to the melt so that the initial doping concentration is maintained.

2.3. Novel Czochralski Crystal Growth

2.3.1. Semicontinuous and Continuous Cz

The conventional Czochralski crystal growth is a batch process and includes the following pitfalls: (*a*) It consumes one quartz crucible per run since the crucible cracks during cooling off the furnace for crystal harvesting, (*b*) It produces crystals with a large difference in doping concentration between the seed and tang ends of ingots, (*c*) It requires a long machine idle time to dismantle and set-up of furnace for each crystal growth run. These pitfalls can be reduced or eliminated if the crystal growth is converted to a semicontinuous or continuous process [7]. The work in these areas had been investigated by the authors [8-10]. These new processes are particularly appealing to the photovoltaic industry. It requires low cost silicon wafers for the fabrication of terrestrial solar cells.

The semicontinuous process can use a conventional Czochralski puller. However, a gate valve is required between the growth and harvesting chambers as shown in Fig. 2.6. The growth

of the first ingot by this process is identical to that by the batch process. After the grown ingot is separated from the melt and raised to the harvest chamber, the melt should be kept molten and the gate valve is closed. The crystal is then removed from the puller and is replaced with a polysilicon charge. After a few minutes of purging, the gate valve can be open and the polysilicon is loaded into the crucible. After the recharging of the crucible with the polysilicon, the growth of the second ingot can be initiated. The process can continue to alternate growth and recharging for several times from a single crucible without cooling the furnace. It has been used an atmospheric crystal puller and reproduciblely demonstrated the growth of three dislocation free ingots with total weight of 31-32 kg from a 8" diameter crucible by two rechargings. A reduced pressure puller that contained a vacuum-tight gate valve has been used [9]. It have been demonstrated the growth of 5 single crystal ingots from a 12" diameter crucible by four rechargings. The total weight of five crystals were approximately 100 kg.

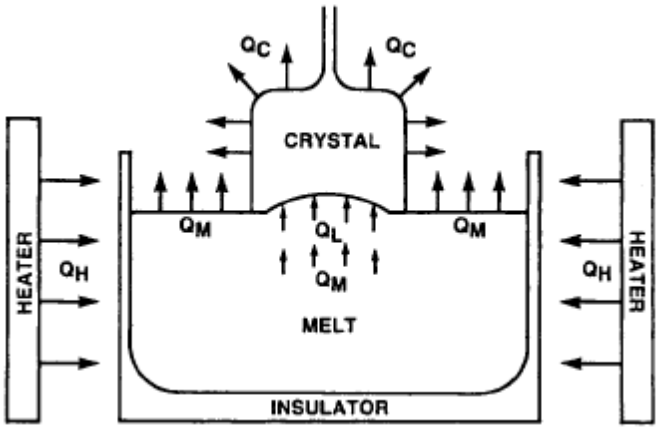


Fig. 2.6. A heat flow pattern in the furnace during crystal pulling.

Several methods have been developed for the recharging of the polysilicon. A long polysilicon rod which was about a one-half section of the U-shaped polysilicon rod harvested from a Siemens reactor was used [9]. The polysilicon rod is attached to the recharging mechanism which they added to the puller. The recharging mechanism was incorporated with a weighting device so that the amount of each recharging from this rod could be controlled. A charge container to hold the preweighed polysilicon cylinders was used in [8]. The dopant can be placed between two poly cylinders for addition into the melt. The bottom of the container consists of heat deformable support members. At low temperatures the support members rigidly hold the polysilicon cylinders in the container. When the container is lowered to about 1-2 inches above the melt, the heat of the furnace and the weight of the polysilicon force the support member to deform, as shown in Fig. 2.7. This opens the bottom of the container and allows the

polysilicon to descend into the crucible. During the melting of the polysilicon, the container is gradually withdrawn from the furnace and removed from the puller.

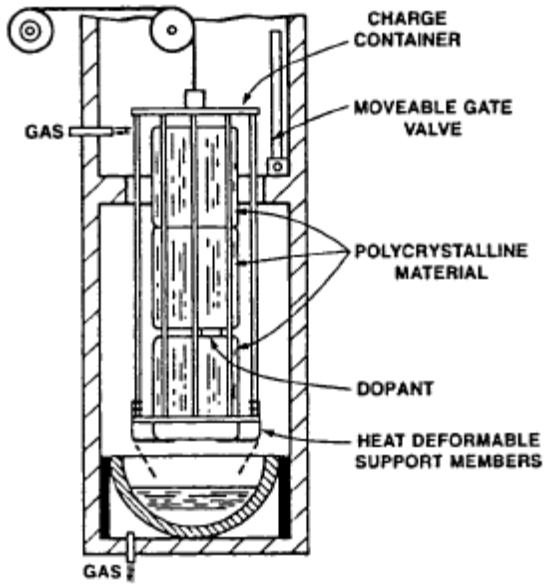


Fig. 2.7. A device used for recharging polysilicon into the melt for the multiple ingot growth from a single crucible.

One concern about the recharging technique is increase of impurity concentration in the melts with the number of rechargings. The impurity concentrations can be calculated by the repeated application of Eq. (2.16).

$$C_s = kC_0(1-g)^{k-1} \tag{2.16}$$

where C_0 is the initial impurity concentration in the melt, k is the segregation coefficient, C_s is the concentration of the impurity in the crystal, and g is the fraction of the melt pulled.

Let us assume that the concentration of an impurity in the polysilicon or the initial melt is C_0, and that the g fraction of the melt is pulled and an equal weight of polysilicon is recharged in each cycle. The impurity concentration in the recharged melt at the beginning of the n-th pull, ($C^i_L(n)$) has been deduced [11]**:**

$$C^i_L(n) = C_0 p^{n-1} + C_0 g(p^{n-2} + p^{n-3} \_ ... + 1) = C_0[p^{n-1} + g(p^{n-1} - 1)/(p-1)] \tag{2.17}$$

where $p=(1-g)^k$. At the end of the n-th growth run, the impurity concentration of the melt left in the crucible before the recharge is

$$C^f_L(n) = (1-g)^{k-1} C^i_L(n) = (p/(1-g))^I_L(n) \tag{2.18}$$

If k<<l, then p=l and Eqs. (2.17) and (2.18) can be approximated by

$$C^i_L(n) = C_0[1 + g(n-1)] \tag{2.19}$$

$$C_L^f(n) = C_0[1 + ng/(1-g)] \quad (2.20)$$

The impurity concentration in the seed and tang ends of the *n*-th pull ingot can be calculated from Eqs. (2.19) and (2.20) respectively by multiplying them by ***k*.**

To visualize the build-up of impurities in the multiple recharging processes, let us assume that *g* equals 0.9 and *n* equals 5 as an example. The build-up of an impurity in the melt at the beginning and the end of the 5th pull are respectively 4.6 and 46 times of the initial value. This seems to be a considerable increase in concentration. However, its impact on the formation of constitutional supercooling and acceptable impurity concentration in the crystals is still negligible. This is because the increase of the concentration in the melt is still below the critical concentration for onset of the constitutional supercooling [9]. Therefore, stable growth of the crystals may still be achieved. This has been verified by experiments that have obtained dislocation-free single crystals from the first to fifth pulls. In addition, the very low distribution coefficient of metallic impurities in silicon also makes the build-up of the impurity in the melt of little concern. For example, experimental results have shown that no detectable increase in the concentration of most metallic impurities could be found from the first to fourth pulled ingots. One exception is aluminum, which has a higher distribution coefficient than other metallic impurities. The increase in aluminum concentration with the number of pulls is shown in Table 2.2. The concentration of aluminum at 1.7×10^{15} atoms / cm^3 does not affect the electrical property of silicon as measured by the minority carrier lifetime or solar cell efficiency [11]. The concentration of carbon and oxygen in the multiple-pulled ingots are also listed in Table 2.2. It shows that the carbon concentration in the seed end of the first three pull ingots which are less than 2×10^{16} atoms / cm^3 are not detectable. However, the carbon concentration in the tang end of the ingots are slightly increased with the number of pulls. This can be understood from the fact that the chance of the graphite parts being exposed to trace amount of air increases with the number of recharging. Table 2.2 also shows that the variation of carbon concentration in the crystals is affected to a greater extent by air-tightness of the gate valve than by the number of pulls. The oxygen concentration shown in Table 2.2 does not follow a clear trend although one of the experiments indicates a decrease with the number of pulls.

Table 2.2. Concentration (atom/cm^3) of Al[a], C[b], and O[b] in multiple-pulled silicon ingots.

	first pull		second pull	
	seed	tang	seed	tang
Al	-	-	3.03×10^{14}	10.3×10^{14}

C(a)	$<2.0\times10^{16}$	2.2×10^{16}	$<2.0\times10^{16}$	4.84×1016
C(b)	-	2.7×10^{17}	-	4.60×10^{17}
O(a)	1.99×10^{18}	1.43×10^{l8}	1.88×10^{l8}	1.37×10^{l8}
O(b)	-	1.30×10^{18}	-	1.10×10^{18}
	third pull		fourth pull	
	seed	tang	seed	tang
Al	1.54×10^{14}	-	16.3×10^{14}	17.7×10^{14}
C(a)	$<2\times10^{16}$	8.6×10^{16}	-	-
C(b)	-	4.7×10^{17}	-	-
O(a)	1.42×10^{18}	1.33×10^{l8}	-	-
O(b)	-	1.40×10^{l8}	-	-

[a]From [9], measured by neutron activation and spark source spectrometry.

[b]From [8], measured by infrared absorption.

Continuous Czochralski growth of silicon crystals has been developed in [10, 12]. The puller that was used is shown in Fig. 2.8. This continuous crystal puller consists of two separated furnaces connected by a continuous liquid feed quartz tube. One furnace is for crystal pulling and the other for the melting of polysilicon. The silicon melt is transferred by siphon action. The crucible in the growth chamber contains a quartz baffle which dampens the melt vibration caused by the melt feeding. The melt is fed at such a rate that a constant melt level in the growth chamber is maintained.

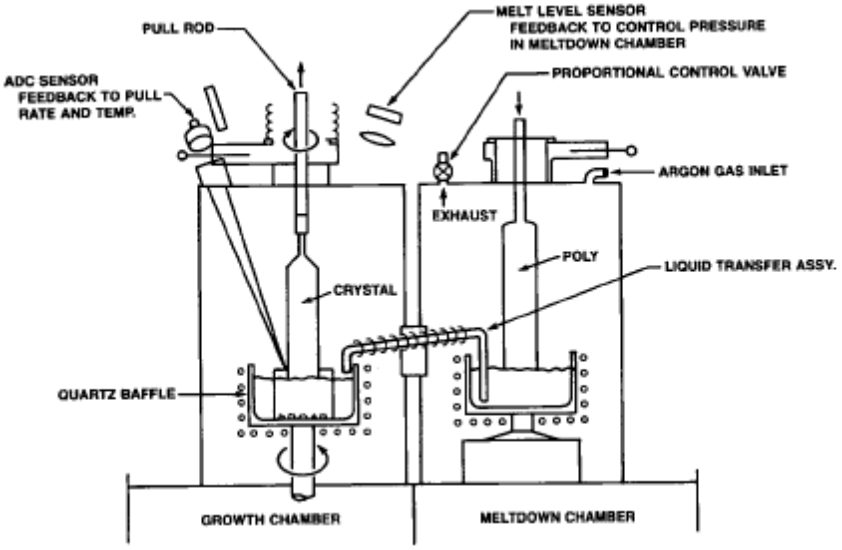


Fig. 2.8. An arrangement of two crystal pullers for continuous growth of silicon crystals [10].

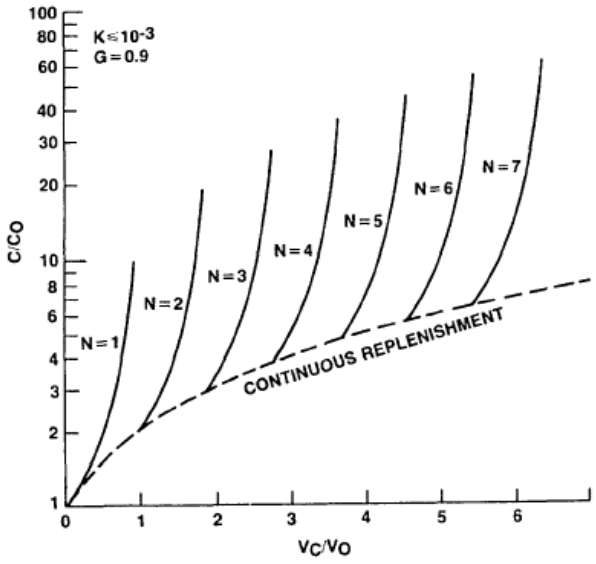


Fig. 2.9. Plot of C_L/C_0 vs V_c/V_0.

One advantage of this method is the uniform impurity distribution in the axial direction of the grown crystals. The distribution of the impurity can be derived from the following differential equation which describes the overall conservation of solute in a system [11].

$$V_0 dC_L = (C_0 - kC_L)dV_c \tag{2.21}$$

where V_0 is the melt volume and V_c is the volume of crystal which has been grown. With the boundary condition of $C_L=C_0$, when $V_c = 0$, the solution of Equation (2.21) is

$$\frac{C_L}{C_0} = \frac{1}{k}[1-(1-k)\exp(-\frac{kV_c}{V_0})] \tag{2.22}$$

The plot of C_L/C_0 versus V_c/V_0 is shown in Fig. 2.9. The C_L/C_0 from semicontinuous processes for various number of pulls (n) is also shown in this figure for comparison. This graph clearly shows that the axial distribution of impurities in the continuous process is much more uniform than by that from the semicontinuous process. The axial impurity distribution in the ingots grown by the semicontinuous process can be improved if only a small fraction of the melt is pulled in each recharged cycle (except the last cycle). This will not increase the silicon loss since the leftover melt is re-used by adding new polysilicon charge. Drawbacks of the continuous growth process are the complexities in equipment and processing. The major process problems are the transfer of the melt from one chamber to another and the control of equality between the feed and pull rates. Nevertheless, crystal sizes up to 65 kg have been grown by this process. It has also demonstrated the capability for growing dislocation-free ingots.

2.3.2. Magnetic Czochralski (MCz) Crystal Growth

The electrical conductivity of silicon increases with the temperature. The conductivity further increases to 12300 $ohm^{-l}cm^{-1}$ when it transforms into a molten state at its melting temperature [13]**.** This number is within the same range of conductivity values for many metals. Electrically, the molten silicon can be considered as a metal. We have pointed out that the molten silicon flows in the crucible due to the thermal-driven convection. In other words, a "moving metal" is confined to circulate inside the crucible during the crystal growth. Therefore, application of a magnetic field into the silicon melt can result with a force that retards its flow (i.e. Lenz law). The distribution of impurities into the crystal can also be altered by the magnetic field.

Two types of magnetic fields have been applied to the Cz growth of silicon crystals. They are the transverse (horizontal) and axial (vertical) fields. Figures 2.10(a) and 2.10(b) show distributions of the magnetic flux generated by the axial and transverse superconductive magnets respectively [14]. Few detailed studies have been made on the effect of a magnetic field on heat and mass transfer of solutes in a conductive solution during freezing. Therefore, understanding of impurity distribution in MCZ crystal growth is still lacking. Experimentally, it has been found that the silicon crystals grown under the influence of a transverse magnetic field provide better physical characteristics. For example, the presence of a transverse field (> 100 gauss) considerably decreases incorporation of oxygen in the crystals and improves the uniformity of axial and radial distributions of dopant and oxygen. On the contrary, the presence of an axial field (also > 100 gauss) increases incorporation of oxygen, carbon and phosphorus in the crystals, and increases the non-uniformity of the impurity distributions [15].

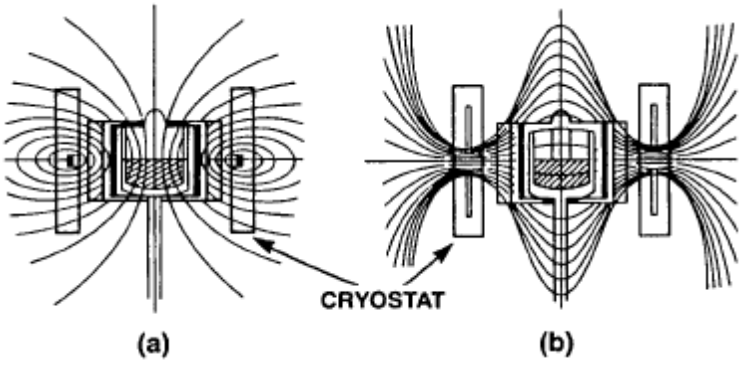


Fig. 2.10. Distribution of magnetic flux in a crystal puller. (*a*) An axial magnetic field. (*b*) **A** transverse magnetic field [14].

Interpretation of these results can partially be made by the observation of the change in melt temperature after the application of a magnetic field. Temperature of the melt at the center surface of the crucible is decreased by the axial field but not by the transverse field although both magnetic fields can reduce the temperature fluctuation. In order to maintain the temperature of the melt at the melting point, the heater temperature has to be increased when an axial field is used. This also increases the crucible temperature and increases the dissolution rate of oxygen from quartz crucible. This explains the high oxygen content in crystals grown under the influence of an axial magnetic field. It has also been suggested [15] that the axial magnetic field strongly suppresses the radial outflow of melt at the interface. This creates non-mixing cells in the melt and causes a higher radial non-uniformity in the crystal.

It has been shown that crystals grown under a transverse field result in very uniformly impurity distribution both in the longitudinal and radial directions [16, 17]. Micro-inhomogeneities such as striations have been eliminated. Bulk stacking faults in the crystal have also been eliminated by applying high pull rates.

2.3.3. Square Ingot Growth

Two approaches have been used to grow square silicon ingots from a Czochralski puller. One approach is to enhance the formation of natural crystal habits. This requires a melt with an extremely good radial symmetry and stable temperatures. Under these conditions, the fast growth portions (or directions) of the ingot will not be melted back during the crystal rotation and the ingot will maintain a natural crystal habit. The growth of (100) square ingots from charge sizes of 1 to 1.5 kg melt has been demonstrated [18]. Continuous seed and crucible rotations both at 10 rpm were applied in opposite directions. The temperature fluctuation at a given point was kept below ±2°C. They have found that a square ingot was obtained when the temperature variations along a circular contour in the melt were less than ±2.5°C and that a circular ingot was grown when the variations were ±10 to 15°C. The square ingots exhibited such a crystal habit that the diagonals of the square were along <100> directions and four edges of the square were perpendicular to <110>. They have been able to maintain square cross-sections throughout the length of the ingots. The ingots were single crystals with etch pit density variations from zero at the ingot center to $1\text{-}2\times10^3/cm^3$ along the crystal periphery. Resistivity measurements on the wafers showed that the iso-resistivity contours were parallel to the edges of the wafers.

Another approach is to shape the temperature profile of the melt into a square configuration. A thermal insulation plate suspended above the melt surface has been used [19]. The gap between the melt and the plate was approximately 2 cm or less. The center of the plate was cut in a square opening. A single crystal seed was dipped into the center of melt surface

through this opening. During the crown portion of the crystal growth, a high supercooling was applied to the melt. The high supercooling forces the crystal to grow faster in <110> direction than in <l00>. Thus the crystal crown will grow into a square with diagonals in <110> and edges in <l00>. Once the size of the crystal crown was approaching that of the opening, the crystal rotation was paused in such a way that the four edges of the crystal crown were parallel to the four sides of the opening. Then the crystal was pulled with a discontinuous rotation during the growth of main crystal body. Each rotation applies a 90 degree turn of the crystal. The purpose of pause rotation is to keep the ingot growing straight since the temperature of the melt was not perfectly symmetrical with respect to the pull axis. The time interval between each rotation is determined by the symmetry of the melt. The poorer symmetry in the melt, the shorter the pause time is needed between each rotation and the less square will be in the ingot. Figure 2.11(*a*) shows a square ingot grown by this technique and Fig. 2.11(*b*) shows wafers cut from a square ingot.

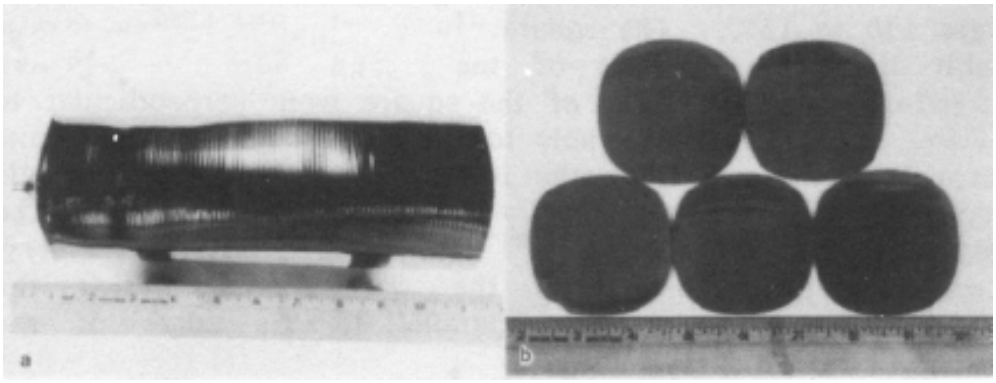


Fig. 2.11. (*a*) **A "**square" silicon ingot. (*b*) Wafers cut from the "square" ingot.

2.3.4. Web and EFG Techniques

Several techniques have been investigated for the growth of silicon crystals in a flat sheet form. Two of the most well developed techniques are the dendritic web growth and edge-defined, film-fed growth (EFG). The web technique was first developed in [20] for Ge growth. The EFG technique was first applied to grow sapphire filaments, tubes, and ribbons [21, 22]), then applied to grow silicon ribbons [23]. Since the mid 1970's these two techniques, particularly EFG, have produced a considerable amount of silicon sheets for solar cell applications.

Web Technique. **A** silicon web can be pulled from a melt by surface tension between two coplanar dendrites of the same seed as shown in Fig. 2.12 [24]. The two parallel dendrites act as a solid frame which holds the web until it solidifies. The frozen web exhibits the same crystallographic orientation as the dendrites except that it is thinner and smoother on he surfaces.

The main surfaces of the web consist of (111) planes. The preferred propagation directions of the dendrites are <211>. During pulling, the dendritic tips extend below the melt surface while the web/melt interface is on or above the melt surface. The growth mechanisms of semiconductor webs have been studied extensively [25, 26]. Growth is initiated by the dendritic growth at both edges of the web. The dendritic growth is controlled by the twin plan reentrant growth mechanism. Figure 2.13 shows sketches of twin planes in a platelet habit that is often observed in diamond cubic materials [27]. The main surfaces of the platelet are bounded by (111) planes on which nucleation is difficult. However, the twin planes provide 141° reentrant grooves for easy nucleation. The nucleation is followed by rapid lateral growth that terminates when it has reached the main (111) surfaces. The repeated nucleation at the reentrant grooves and lateral growth cause the platelet to become a dendrite which advances in a <211> direction. If two or more twin planes are present, the reentrant grooves will serve as perpetual nucleation sites. Then the remain as a flat strip rather than a rod [27]. For web growth, it is also required that the two bounding dendrites contain no branches. Branching of dendrites can be avoided by using a three-twin seed that keeps the branching directions from emerging into the melt [24].

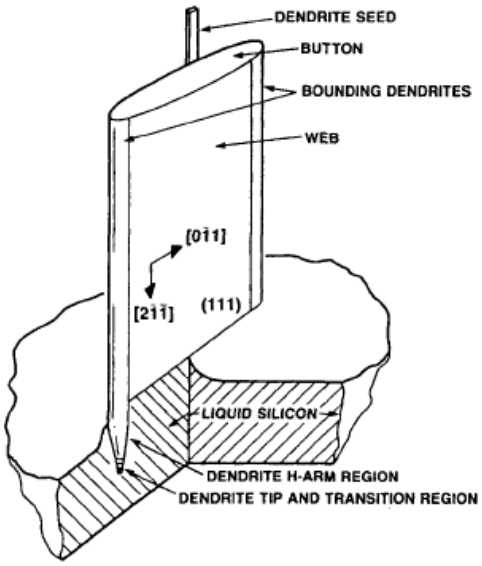


Fig. 2.12. A silicon web grown by freezing a thin sheet of melt supported by two coplanar silicon dendrites [24].

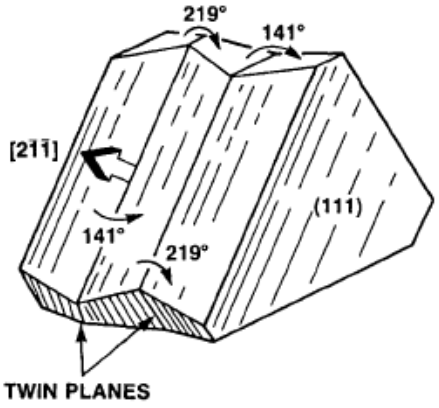


Fig. 2.13. Dendritic growth by the twin plane reentrant growth mechanism [25].

A web crystal grown in the early stages of development always contained twin planes which were extended from the bounding dendrites (Fig. 2.14*a*). The twin planes in the dendrites are located at the center and parallel to the main surfaces. The web can be free of twin planes as shown in Fig. 2.14*b* if the web is not coinciding with the center plane of two bounding dendrites. This can be achieved by pulling the web from a melt of asymmetrical temperature distribution with respect to the pulling plane. Web crystals can also be grown free of dislocations. This is achieved by using a flat interface [28]

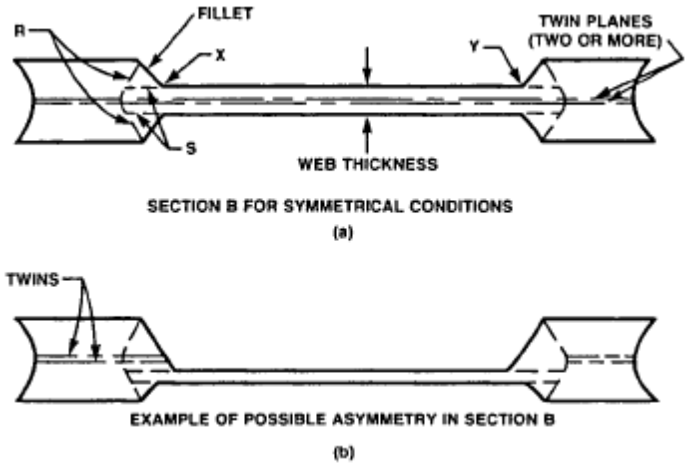


Fig. 2.14. (*a*) Web contains twin planes when the thermal symmetric plane is coincided with a twin plane in dendrites. (*b*) Free of twin planes in the web when the plane of thermal symmetry is not overlaid with any twin plane between two dendrites [26].

EFG Technique. The EFG technique [29] utilizes a die that is inserted into the melt. The construction of the die is the predominate entity that governs the growth of ribbon crystals. The die parameters include the shape, temperature distribution, and the material of construction.

Figures 2.15(*a*) and (*b*) show examples of two types of die designs. Figure 2.15(a) is a die which contains a narrow slot, while Fig. 2.15(b) is another die with multiple holes in it. When a die is made of a material such as graphite that can be wetted by silicon melt, the melt will flow into the holes or slot and rise above the melt surface by means of capillary action. The height, *h*, of the capillary rise can be calculated from:

$$h = (2\gamma \cos\theta)/(\rho g d) \qquad (2.23)$$

where γ is the interfacial surface tension, θ is the wetting angle of the liquid silicon on the die, ρ is the density of silicon, g is the gravitational constant, and d is the diameter of the capillary, or the distance of the capillary spacing of the slot in the die. Figure 2.16 shows the plot of theoretical capillary rise versus capillary spacing for dies made of fused silica and graphite. Although low capillary rise in a silica die can be improved by using extremely thin capillaries, it is extremely difficult to control the dimensions in quartzware. The high plasticity of fused silica at the melting temperature of silicon can induce both stress on the ribbon and momentary freezing of the ribbon to the die during the ribbon pulling. Therefore, the use of a fused silica die has never been popular.

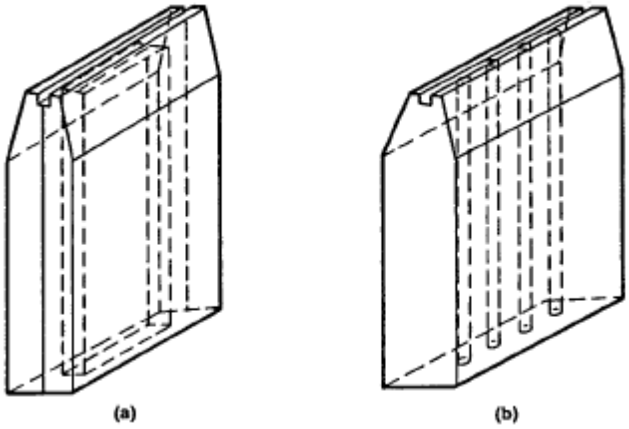


Fig. 2.15. Examples of die design. (*a*) A die contains a narrow slot. (*b*) A die contains multiple holes [29].

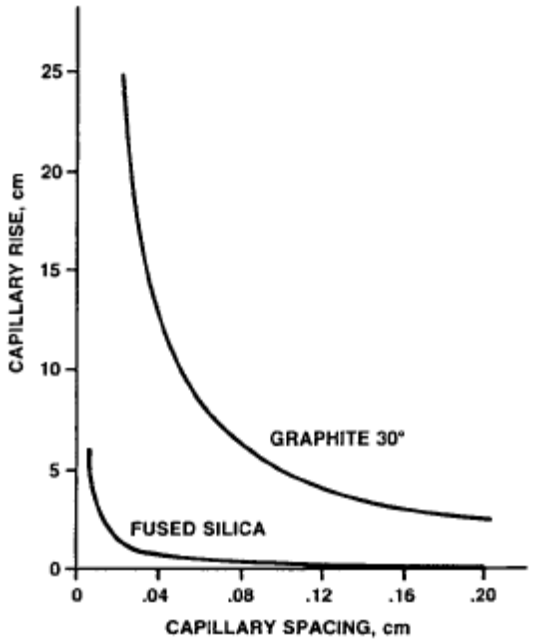


Fig. 2.16. Plot of Capillary rise vs Capillary space in a die [29].

A die should be so designed that when it is placed in the melt, the vertical distance of the die that floats above the melt surface should be less than the theoretical capillary rise. When this condition is met, the melt will feed to the top of the die. The top of a die is shaped with a shallow groove for melt confinement. If the temperature of the die is kept above the melting point of silicon, the silicon in the groove will remain as liquid film. A flat thin silicon seed can be made to contact the liquid film in the groove and a silicon ribbon can be pulled from it. The width of the groove and other factors such as pull rate and melt temperature can affect the thickness of the ribbon. During the ribbon pulling process, the liquid silicon film in the groove is continuously fed from the melt outside the die through capillary holes. Multiple silicon ribbons have been pulled from the melt in a single crucible that contained multiple dies. Continuous ribbon growth process has been demonstrated. This is achieved by (*a*) the use of a spool for winding of the pulled ribbon, and (*b*) continuous melt replenishment of the crucible.

EFG ribbons prefer to grow in a <211> direction. The main surfaces are bounded by the planes approaching {110}. This growth orientation is naturally formed and maintained regardless of the seed orientation. Dislocation-free ribbons are difficult to obtain. The magnitude of dislocation densities is typically at $10^5/cm^2$. The ribbons also contain multiple close-spaced twin planes that are parallel to the growth directions. Approximately 10 ppma of carbon is also contained in the ribbons because of the use of a graphite die. Typical minority carrier diffusion lengths in the EGF ribbon range from 30 to 80 μm as compared to 100 or higher in the

Czochralski wafers. However, solar cells with efficiency in excess of 10% have been fabricated from EFG ribbons.

2.4. Silicon Float-zone Process

The float-zone process can be used to grow silicon that has lower contaminations than that normally obtained from the Czochralski technique. A schematic setup of the float zone process is shown in Fig. 2.17*a*. The procedure used in float zone growth is shown in Fig. 2.18. A high-purity polycrystalline rod with a seed crystal at the bottom is held in a vertical position and rotated. The rod is enclosed in a quartz envelope within which an inert atmosphere (argon) is maintained. During the operation, a small zone (a few centimeters in length) of the crystal is kept molten by a radio-frequency heater, which is moved from the seed upward so that this *floating zone* traverses the length of the rod. The molten silicon is retained by surface tension between the melting and growing solid-silicon faces. As the floating zone moves upward, a single-crystal silicon freezes at the zone's retreating end and grows as an extension of the seed crystal. Materials with higher resistivities can be obtained from the float-zone process than from the Czochralski process because it can be used to purify the crystal more easily. Furthermore, since no crucible is used in the float-zone process, there is no contamination from the crucible (as with Czochralski growth). At the present time, float-zone crystals are used mainly for high-power, high-voltage devices, where high-resistivity materials are required.

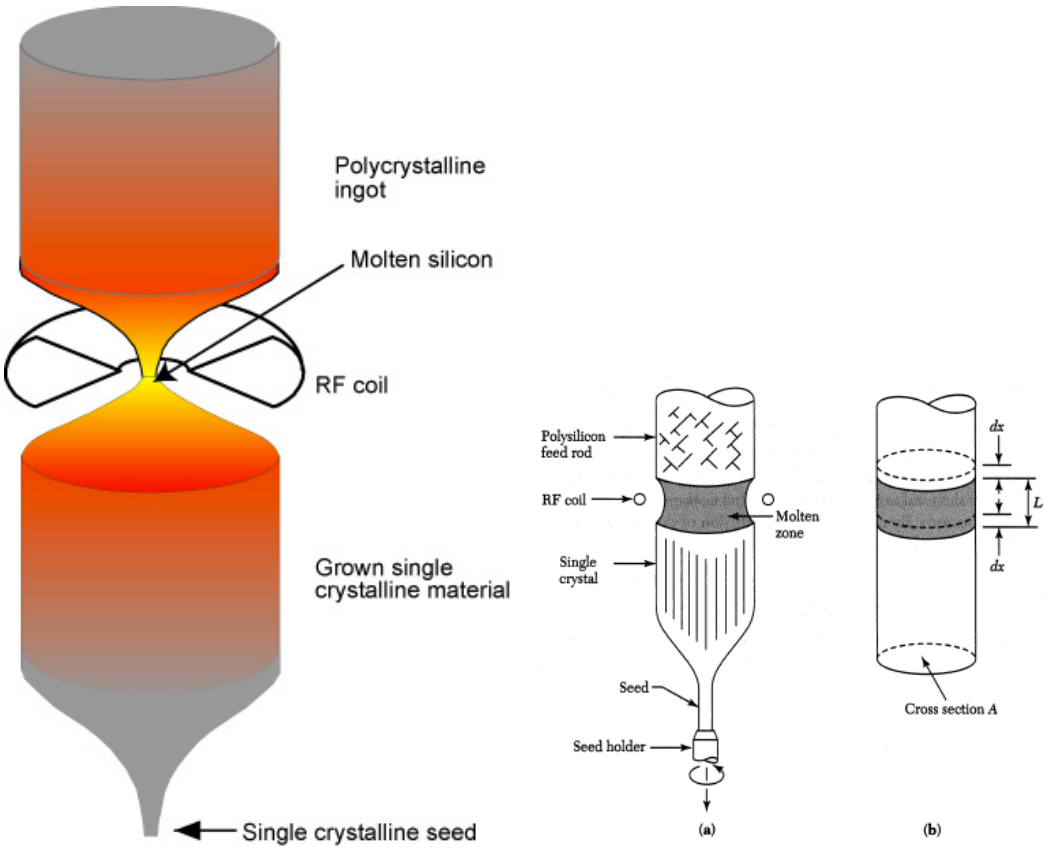


Fig. 2.17. Float-zone process. (*a*) Schematic setup. (*b*) Simple model for doping evaluation.

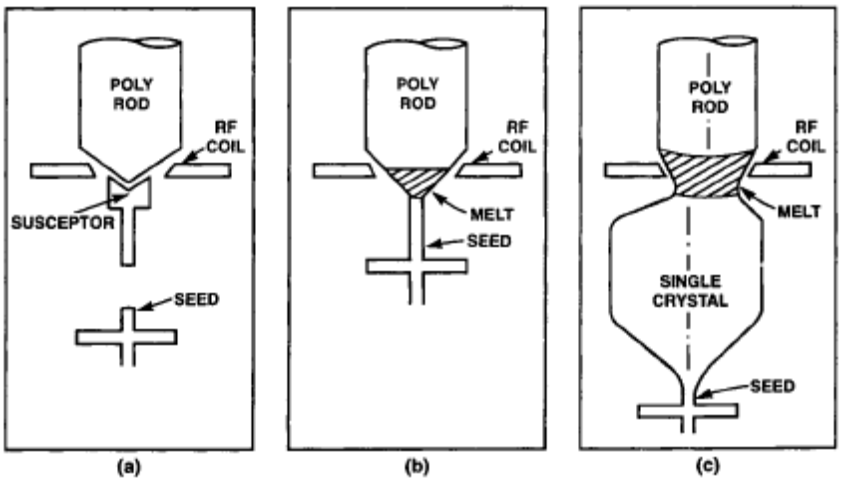


Fig. 2.18. A procedure used in float zone growth of a silicon crystal. (*a*) Preheat of polysilicon rod by a graphite susceptor. (*b*) Initiate growth by dipping single crystal seed. (*c*) Growth of an ingot with diameter greater than that of molten silicon by off-setting the pull axis.

To evaluate the doping distribution of a float-zone process, consider a simplified model, as shown in Fig. 2.17*b*. The initial, uniform doping concentration in the rod is C_0 (by weight). L is the length of the molten zone at a distance x along the rod, A the cross-sectional area of the rod, ρ_d the specific density of silicon, and S the amount of dopant present in the molten zone. As the

zone traverses a distance *dx*, the amount of dopant added to it at its advancing end is $C_0\,\rho_d\,A\,dx$, whereas the amount of dopant removed from it at the retreating end is $k_e(S\,dx/L)$, where k_e, is the effective segregation coefficient. Thus,

$$dS = C_0\rho_d A dx - \frac{k_e S}{L}dx = (C_0\rho_d A - \frac{k_e S}{L})dx, \tag{2.24}$$

so that

$$\int_0^x dx = \int_{S_0}^{S} \frac{dS}{C_0\rho_d A - k_e S/L}, \tag{2.24a}$$

where $S_0 = C_0\rho_d AL$ is the amount of dopant in the zone when it was first formed at the front end of the rod. From Eq. 2.24a we obtain

$$\exp(\frac{k_e x}{L}) = \frac{C_0\rho_d A - k_e S_0/L}{C_0\rho_d A - k_e S/L} \tag{2.25}$$

or

$$S = \frac{C_0\rho_d AL}{k_e}[1-(1-k_e)^{-k_e x/L}]. \tag{2.25a}$$

Since C_s (the doping concentration in the crystal at the retreating end) is given by $C_s = k_e(S/A\rho_d L)$, then

$$C_s = C_0[1-(1-k_e)^{-k_e x/L}]. \tag{2.26}$$

Figure 2.19 shows the doping concentration versus the solidified zone length for various values of k_e.

These two crystal growth techniques can also be used to remove impurities. A comparison of Fig. 2.19 with Fig. 2.4 shows that a single pass in the float-zone process does not produce as much purification as a single Czochralski growth. For example, for $k_0 = k_e = 0.1$, C_s/C_0 is smaller over most of the solidified ingot made by the Czochralski growth. However, multiple float-zone passes can be performed on a rod much more easily than a crystal can be grown, the end region cropped off, and regrown from the melt. Figure 2.20 shows the impurity distribution for an element with $k_e = 0.1$ after a number of successive passes of the zone along the length of the rod [6]. Note that there is a substantial reduction of impurity concentration in the rod each pass. Therefore, the float-zone process is ideally suited for crystal purification. This process is also called the zone-refining technique, which can provide a very-high-purity level of the raw material.

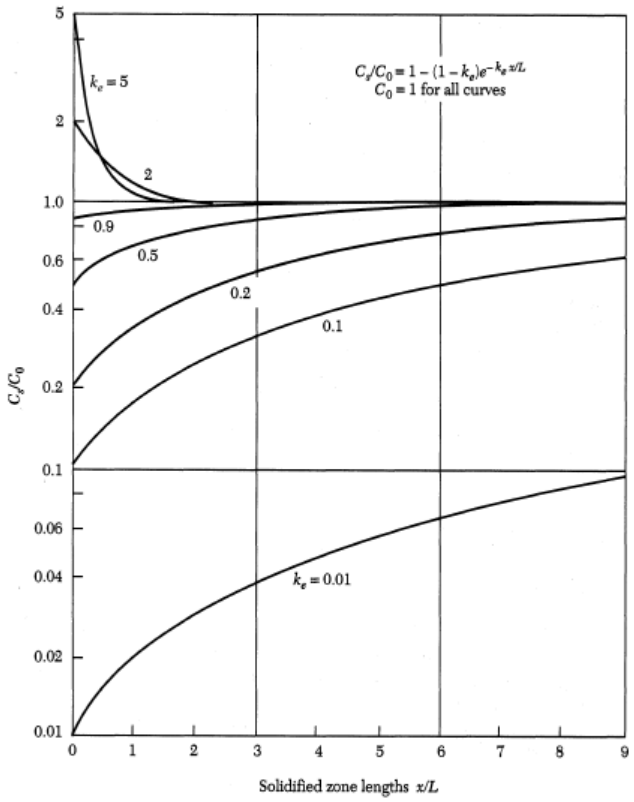


Fig. 2.19. Curves for the float-zone process showing doping concentration in the solid as a function of solidified zone lengths [6].

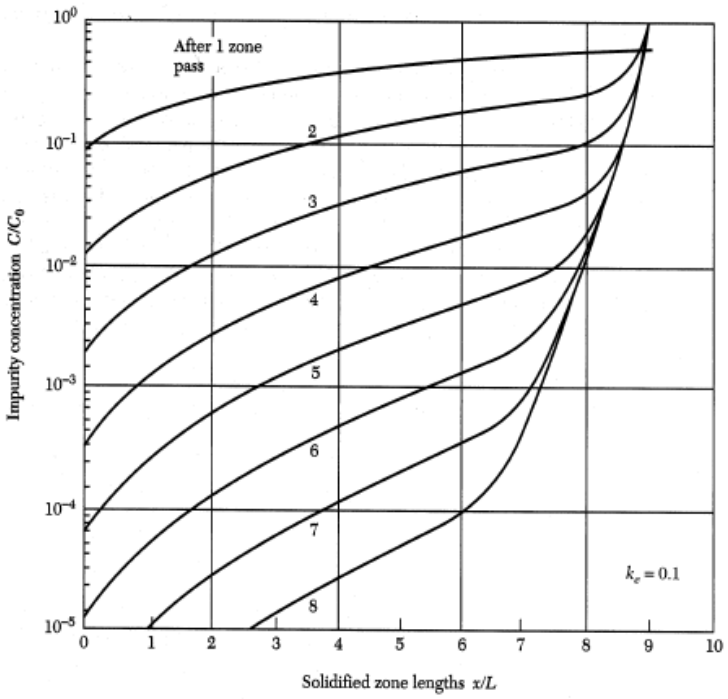


Fig. 2.20. Relative impurity concentration versus zone length for a number of passes. *L* denotes the zone length [6].

If it is desirable to dope the rod rather than purity it, consider the case in which all the dopants are introduced in the first zone ($S_0 = C_1 A \rho_d L$) and the initial concentration C_0 is negligibly small. Equation 2.25 gives

$$S_0 = S \exp(\frac{k_e x}{L}). \tag{2.27}$$

Since $C_s = k_e(S/A\rho_d L)$, we obtained from Eq. 2.19

$$C_s = k_e C_1 \exp(-k_e x / L). \tag{2.28}$$

Therefore, if $k_e x/L$ is small, C_s will remain nearly constant with distance except at the end that is last to solidify.

For certain switching devices, such as high-voltage thyristors, large chip areas are used frequently an entire wafer for a single device. This size imposes stringent requirements on the uniformity of the starting material. To obtain homogeneous distribution of dopants, we use a float-zone silicon slice that has an average doping concentration well below the required amount. The slice is then irradiated with thermal neutrons. This process, called neutron irradiation, gives rise to fractional transmutation of silicon into phosphorus and dopes the silicon *n*-type:

$$Si_{14}^{30} + \text{neutron} \rightarrow Si_{14}^{31} + \gamma\,\text{ray} \rightarrow^{2.62\text{hr}} P_{15}^{31} + \beta\,\text{ray}. \tag{2.29}$$

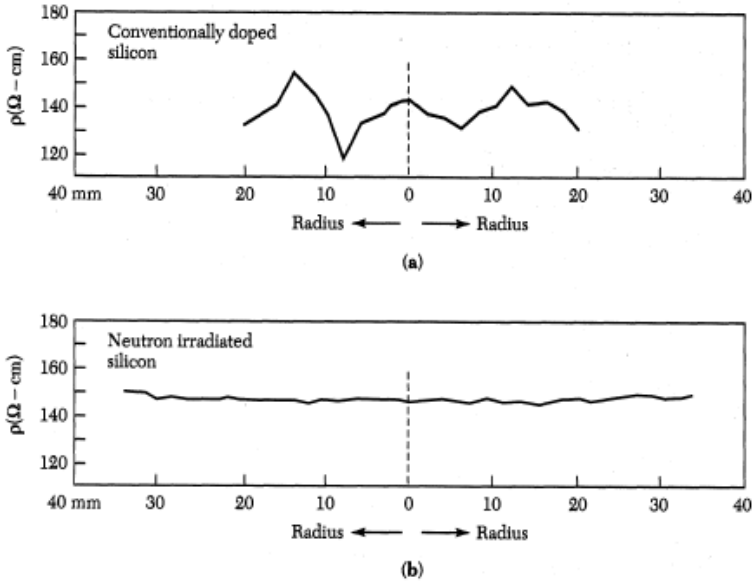


Fig. 2.21. (*a*) Typical lateral resistivity distribution in conventionally doped silicon. (*b*) Silicon doped by neutron irradiation [30].

The half-life of the intermediate element Si_{14}^{31} is 2.62 hours. Because the penetration depth of neutrons in silicon is about 100 cm, doping is very uniform throughout the slice. Figure 2.21

compares the lateral resistivity distributions in conventionally doped silicon and in silicon doped by neutron irradiation [30]. Note that the resistivity variations for the neutron-irradiated silicon are much smaller than that for the conventionally doped silicon.

2.5. Trends in Silicon Crystal Growth

Trends in silicon crystal growth technology are dictated by the evolution of semiconductor devices. The semiconductor devices are heading toward (*a*) higher operation speeds, (*b*) smaller sizes of each individual device, (*c*) a larger scale of integration (i.e. toward ULSI), and (*d*) lower cost of manufacturing. High speed operations of devices require silicon with low capacitance (i.e. high resistivity), although the speed can also be improved significantly by shrinkage of device dimensions. Resistivity in the range of 15 to 30 ohm-cm has been widely used for both MOS and Bipolar integrated circuits. These IC devices may not require wafer resistivity greater than 100 ohm-cm. This is because the advantage of capacitance reduction is traded off for adverse factors such as the increase in substrate currents (noise) resulting from higher minority carrier lifetimes in the extrinsic regions of the devices. However, high resistivity wafers will be continuously needed by power devices. Fabrications of these devices are primarily provided by the FZ crystals. Cz crystals have not been able to meet a high resistivity specification because of the presence of oxygen donors. Recent development of the transverse magnetic Cz technique may potentially produce wafers with stable resistivities of up to 200 ohm-cm.

Shrinkage of device dimensions requires lower leakage currents in the devices. Electrical leakages, particularly the junction leakages, are due to the existence of crystallographic defects in the active regions of devices. A trend in silicon wafer technology is to continuously reduce the defect densities in these areas. One of the current methods is to form a denuded zone based on controlled oxygen precipitation. Better control of oxygen precipitation will be needed for the fabrication of smaller devices. To meet this need, crystal growth techniques such as transverse MCz growth and oxygen doping in FZ growth have to be further developed.

Greater integration of devices requires smaller power dissipation per device. CMOS circuits consume the least amount of power and are the best candidate for ULSI. One of the disadvantages of the CMOS circuits is the ease in latchup of the adjacent devices through the formation of parasitic transistors. Latchup can be reduced or eliminated by using a lightly doped epitaxial layer on heavily-doped substrates. Demand for heavily-doped crystals will increase. However, research and development on heavily doped crystals has not been very active since these materials were not used for IC manufacturing. Quality improvement on heavily doped

crystals is needed particularly when Sb is used as a dopant. Recently, it has been found that defect densities in the epitaxial layers are always high when a heavily doped Sb wafer is used as the substrate. Several causes have been suggested for the result of this finding. (*a*) Heavily-doped Sb crystals contain less oxygen, and, therefore, lack intrinsic gettering capability. (*b*) Sb dopant is less pure than other dopants so that the greater amount of metallic impurities can diffuse to the epitaxial layer to form epitaxial haze (i.e. shallow etch pits). Further understanding of heavily doped Sb crystals is needed in order to solve this problem.

Lowering of the manufacturing cost of devices requires use of larger diameter silicon wafers. Although the cost saved due to the increase in diameter from 6" to 8" is not as dramatic as from 2" to 3", the trend still heads for larger diameter wafers. Problems arising from large diameter wafers are (*a*) increase of wafer warpage, and (*b*) increase of wafer batch weight, which can exceed the yield point of quartz tubes used for high temperature processing of silicon wafers. The first problem can be solved if the hardness of wafer is increased. The solution to the latter problem is to reduce the thickness of the wafers. This in turn also requires improvement of wafer hardness to prevent bending and warpage. The improvement of silicon hardness has been unintentionally carried out by incorporation of oxygen into silicon lattice during crystal growth. However, oxygen can easily form SiO_2 precipitates which, contrary to oxygen in a solid solution, can enhance the wafer warpage. Therefore, we need a reliable wafer hardening technique. Nitrogen dopings in FZ crystals have been considered as an alternative to oxygen in Cz crystals. Further development is needed in this area.

The size of crystals will continue to increase in the next few years. The growth of dislocation-free Cz ingots of up to 10" in diameter can be obtained by the extension of current growth technology. However, the growth processes require further automation which may include control of ingot diameter, dislocation-free structures, and oxygen and dopant concentration targets. The batch growth process should also be converted to a continuous process in order to reduce the cost as well as improve axial resistivity uniformity.

2.6. GaAs Crystal-growth Techniques

2.6.1. Starting Materials

The starting materials for the synthesis of polycrystalline gallium arsenide are the elemental, chemically pure gallium and arsenic. Because gallium arsenide is a combination of two materials, its behavior is different from that of a single material such as silicon. The behavior of a combination can be described by a phase *diagram.* A phase is a state (e.g., solid,

liquid, or gaseous) in which a material may exist. A phase diagram shows the relationship between the two components, gallium and arsenic, as a function of temperature.

Figure 2.22 shows the phase diagram of the gallium-arsenic system. The abscissa represents various compositions of the two components in terms of atomic percent (lower scale) or weight percent (upper scale) [31, 32]. Consider a melt that is initially of composition *x* (e.g., 85 atomic percent arsenic shown in Fig. 2.22). When the temperature is lowered, its composition will remain fixed until the *liquidus* line is reached. At the point (T_l, *x*), material of 50 atomic percent arsenic (i,e., gallium arsenide) will begin to solidify.

Unlike silicon, which has a relatively low vapor pressure at its melting point (~10^{-6} atm at 1412°C), arsenic has much higher vapor pressures at the melting point of gallium arsenide (1240°C). In its vapor phase, arsenic has As_2 and As_4 as its major species. Figure 2.23 shows the vapor pressures of gallium and arsenic along the liquidus curve [33]. Also shown for comparison is the vapor pressure of silicon. The vapor pressure curves for gallium arsenide are double valued. The dashed curves are for arsenic-rich gallium arsenide melt (right side of liquidus line in Fig. 2.22), and the solid curves are for gallium-rich gallium arsenide melt (left side of liquidus line in Fig. 2.22). Because there is a larger amount of arsenic in an arsenic-rich melt than in a gallium-rich melt, more arsenic (As_2 and As_4) will be vaporized from the arsenic-rich melt, thus resulting in a higher vapor pressure. A similar argument can explain the higher vapor pressure of gallium in a gallium-rich melt. Note that long before the melting point is reached, the surface layers of liquid gallium arsenide may decompose into gallium and arsenic. Since the vapor pressure of gallium and arsenic is different, there is a preferential loss of the more volatile arsenic species, and the liquid becomes gallium rich.

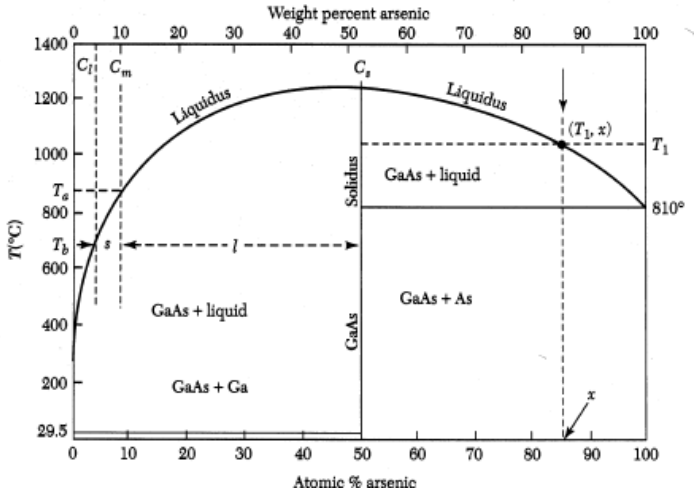


Fig. 2.22. Phase diagram for the gallium-arsenic system [31].

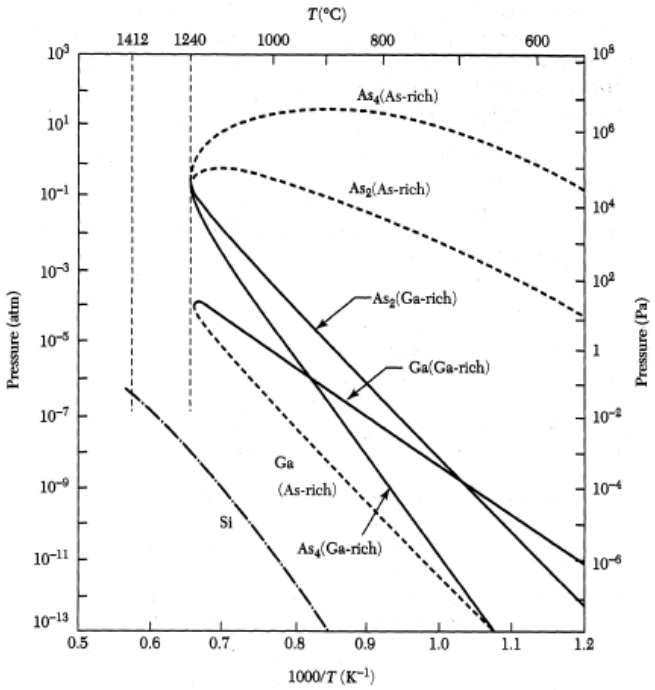


Fig. 2.23. Partial pressure of gallium and arsenic over gallium arsenide as a function of temperature [33]. Also shown is the partial pressure of silicon.

To synthesize gallium arsenide, an evacuated, sealed quartz tube system with a two temperature furnace is commonly used. The high-purity arsenic is placed in a graphite boat and heated to 610°-620°C, whereas the high-purity gallium is placed in another graphite boat and heated to slightly above the gallium arsenide melting temperature (1240°-1260°C). Under these conditions, an overpressure of arsenic is established (*a*) to cause the transport of arsenic vapor to the gallium melt, converting it into gallium arsenide, and (*b*) to prevent decomposition of the gallium arsenide while it is being formed in the furnace. When the melt cools, a high-purity polycrystalline gallium arsenide results. This serves as the raw material to grow single-crystal gallium arsenide [32].

2.6.2. Crystal Growth Techniques

There are two techniques for GaAs crystal growth: the Czochralski technique and the Bridgman technique. Most gallium arsenide is grown by the Bridgman technique. However, the Czochralski technique is more popular for the growth of larger-diameter GaAs ingots.

For Czochralski growth of gallium arsenide, the basic puller is identical to that for silicon. However, to prevent decomposition of the melt during crystal growth, a liquid encapsulation method is employed. The liquid encapsulant is a molten boron trioxide (B_2O_3) layer about 1 cm thick. Molten boron trioxide is inert to the gallium arsenide surface and serves as a cap to cover the melt. This cap prevents decomposition of the gallium arsenide as long as the pressure on its

surface is higher than l atm (760 Torr). Because boron trioxide can dissolve silicon dioxide, the fused-silica crucible is replaced with a graphite crucible.

To obtain the desired doping concentration in the grown crystal of GaAs, cadmium and zinc are commonly used for *p*-type materials, whereas selenium, silicon, and tellurium are used for *n*-type materials. For semiinsulating GaAs, the material is undoped. The equilibrium segregation coefficients for dopants in GaAs are listed in Table 2.3. Similar to those in Si, most of the segregation coefficients are less than 1. The expressions derived previously for Si are equally applicable to GaAs (Eqs. 2.4 to 2.15).

Table 2.3. Equilibrium segregation coefficients for dopants in GaAs.

Dopant	k_0	Type
Be	3	*p*
Mg	0.1	*p*
Zn	4×10^{-1}	*p*
C	0.8	*n/p*
Si	1.85×10^{-1}	*n/p*
Ge	2.8×10^{-2}	*n/p*
S	0.5	*n*
Se	5×10^{-1}	*n*
Sn	5.24×10^{-2}	*n*
Te	6.8×10^{-2}	*n*
Cr	1.03×10^{-4}	Semiinsulating
Fe	1.0×10^{-3}	Semiinsulating

Figure 2.24 shows a Bridgman system in which a two-zone furnace is used for growing single-crystal gallium arsenide. The left-hand zone is held at a temperature (~610°C) to maintain the required overpressure of arsenic, whereas the right-hand zone is held just above the melting point of gallium arsenide (1240°C). The sealed tube is made of quartz and the boat is made of graphite. In operation, the boat is loaded with a charge of polycrystalline gallium arsenide, with the arsenic kept at the other end of the tube.

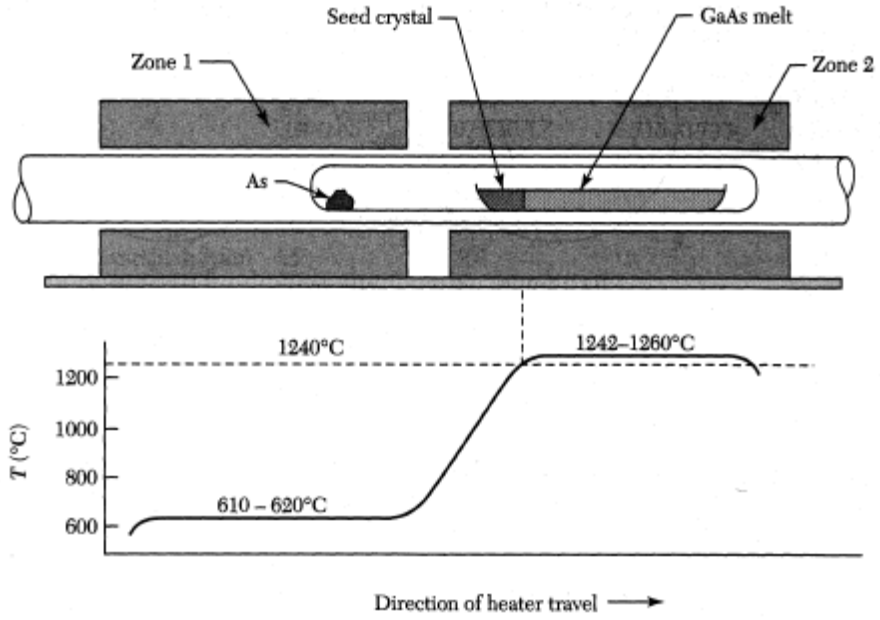


Fig. 2.24. Bridgman technique for growing single-crystal gallium arsenide and a temperature profile of the furnace.

As the furnace is moved toward the right, the melt cools at one end. Usually, there is a seed placed at the left end of the boat to establish a specific crystal orientation. The gradual freezing (solidification) of the melt allows a single crystal to propagate at the liquid- solid interface. Eventually, a single crystal of gallium arsenide is grown. The impurity distribution can be described essentially by Eqs. 2.9 and 2.15, where the growth rate is given by the traversing speed of the furnace.

2.7. Material Characterization

2.7.1. Wafer Shaping

After a crystal is grown, the first shaping operation is to remove the seed and the other end of the ingot, which is last to solidify [3]. The next operation is to grind the surface so that the diameter of the material is defined. After that, one or more flat regions are ground along the length of the ingot. These regions, or *flats,* mark the specific crystal orientation of the ingot and the conductivity type of the material. The largest flat, the *primary flat*, allows a mechanical locator in automatic processing equipment to position the wafer and to orient the devices relative to the crystal. Other smaller flats, called *secondary flats,* are ground to identify the orientation and conductivity type of the crystal, as shown in Fig. 2.25. For crystals with diameters equal or larger than 200 mm, no flats are ground. Instead, a small groove is ground along the length of the ingot.

The ingot is ready to be sliced by diamond saw into wafers. Slicing determines four wafer parameters: *surface orientation* (e.g., <111> or <l00>), *thickness* (e.g., 0.5-0.7 mm, depending on wafer diameter); *taper*, which is the wafer thickness variations from one end to another; and *bow*, which is the surface curvature of the wafer, measured from the center of the wafer to its edge.

After slicing, both sides of the wafer are lapped using a mixture of Al_2O_3 and glycerine to produce typical flatness uniformity within 2 μm. The lapping operation usually leaves the surface and edges of the wafer damaged and contaminated. The damaged and contaminated regions can be removed by chemical etching. The final step of wafer shaping is polishing. Its purpose is to provide a smooth, specular surface where device features can be defined by lithographic processes. Figure 2.26 shows 200 mm (8 in.) and 400 mm (16 in.) polished silicon wafers in cassettes.

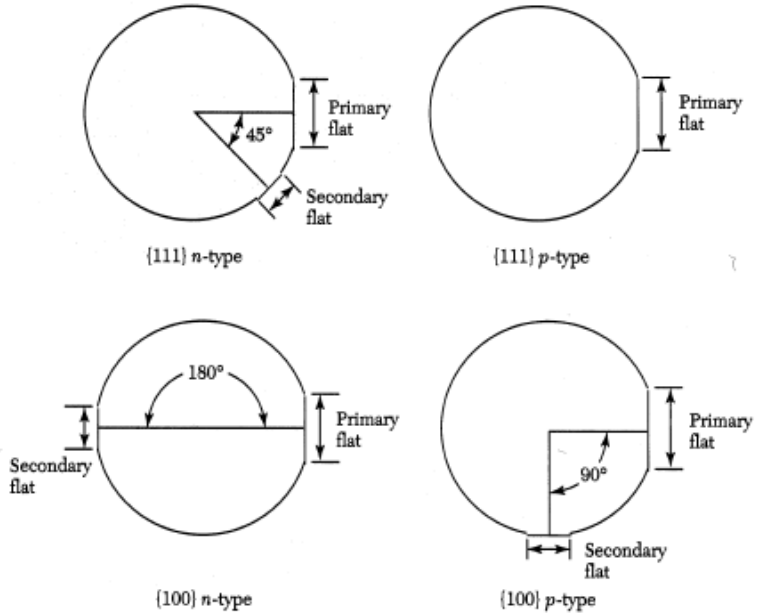


Fig. 2.25. Identifying flats on a semiconductor wafer.

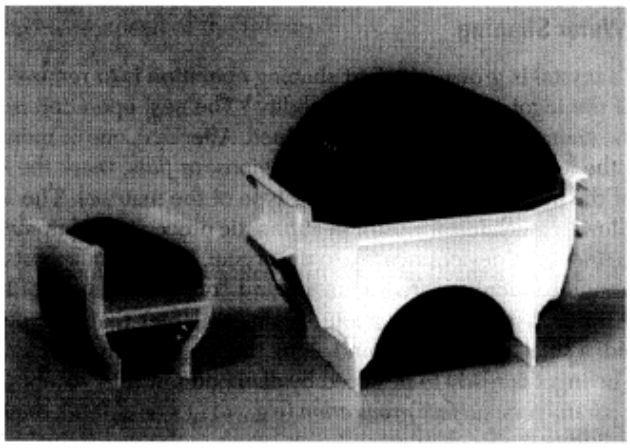


Fig. 2.26. 200 mm (8 in.) and 400 mm (16 in.) polished silicon wafers in cassettes.

Table 2.4 shows the specifications for 125, 150, 200, and 300 mm diameter polished silicon wafers from the Semiconductor Equipment and Materials Institute (SEMI). As mentioned previously, for large crystals (≥ 200 mm diameter) no flats are ground; instead, a groove is made on the edge of the wafer for positioning and orientation purpose.

Gallium arsenide is a more fragile material than silicon. Although the basic shaping operation of gallium arsenide is essentially the same as that for silicon, greater care must be exercised in gallium arsenide wafer preparation. The state of gallium arsenide technology is relatively primitive compared with that of silicon. However, the technology of group III-V compounds has advanced partly because of the advances in silicon technology.

Table 2.4. Specification for polished monocrystalline silicon wafers.

Parameter	125 mm	150 mm	200 mm	300 mm
Diameter (mm)	125±1	150±1	200±1	300±1
Thickness (mm)	0.6-0.65	0.65-0.7	0.715-0.735	0.755-0.775
Primary flat length (mm)	40-45	55-60	NA	NA
Secondary flat length (mm)	25-30	35-40	NA	NA
Bow (μm)	70	60	30	<30
Total thickness variation (μm)	65	50	10	<10
Surface orientation	(100)±1°	Same	Same	Same
	(111)±1°	Same	Same	Same

NA: not available.

2.7.2. Crystal Characterization

Crystal Defects

A real crystal (such as a silicon wafer) differs from the ideal crystal in important ways. It is finite; thus, surface atoms are incompletely bonded. Furthermore, it has defects, which strongly influence the electrical, mechanical, and optical properties of the semiconductor. There are four categories of defects: point defects, line defects, area defects, and volume defects.

Figure 2.27 shows several forms of *point defects* [3, 34]. Any foreign atom incorporated into the lattice at either a substitutional site [i.e., at a regular lattice site (Fig. 2.27*a*)] or interstitial site [i.e., between regular lattice sites (Fig.2.27*b*)] is a point defect. A missing atom in the lattice creates a vacancy, also considered a point defect (Fig. 2.27*c*). A host atom that is situated between regular lattice sites and adjacent to a vacancy is called a *Frenkel defect* (Fig. 2.27*d*). Point defects are particularly important subjects in the kinetics of diffusion and oxidation processes.

The next class of defects is the *line defect*, also called a disloation [35]. There are two types of dislocations: the edge and screw types. Figure 2.28*a* is a schematic representation of an edge dislocation in a cubic lattice. There is an extra plane of atoms AB inserted into the lattice. The line of the dislocation would be perpendicular to the plane of the page. The screw dislocation may be considered as being produced by cutting the crystal partway through and pushing the upper part one lattice spacing over, as show in Fig. 2.28*b*. Line defects in devices are undesirable because they act as precipitation sites for metallic impurities, which may degrade device performance.

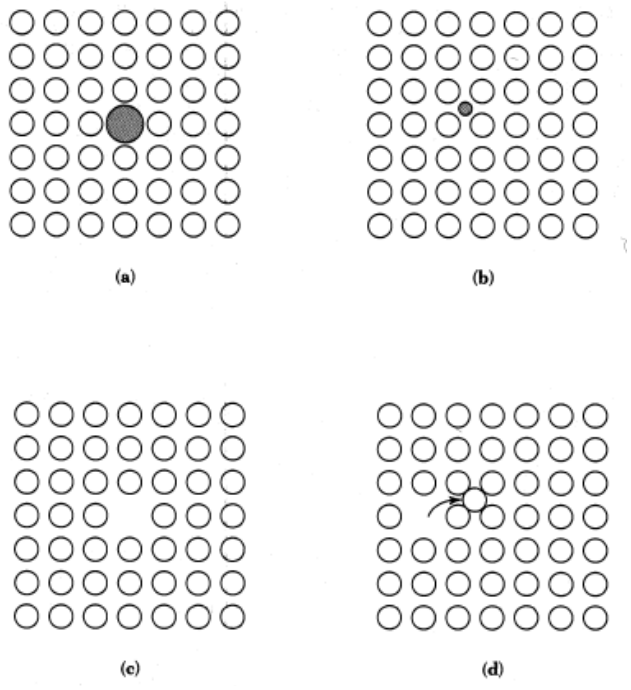


Fig. 2.27. Point defects. *(a)* Substitutional impurity. *(b)* Interstitial impurity. *(c)* Lattice vacancy. *(d)* Frenkel-type defects [34].

Area defects represent a large area discontinuity in the lattice. Typical defects are twins and grain boundaries. Twinning represents a change in the crystal orientation across a plane. A grain

boundary is a transition between crystals having no particular orientational relationship to one another. Such defects appear during crystal growth. Another area defect is the stacking fault [11]. In this defect, the stacking sequence of atomic layer is interrupted. In Fig. 2.29 the sequence of atoms in a stack is ABCABC . . . . When a part of layer C is missing, this is called the intrinsic stacking fault (Fig. 2.29*a*). If an extra plane A is inserted between layers B and C, this is an extrinsic stacking fault (Fig. 2.29*b*). Such defects may appear during crystal growth. Crystals having these area defects are not usable for integrated-circuit manufacture and are discarded.

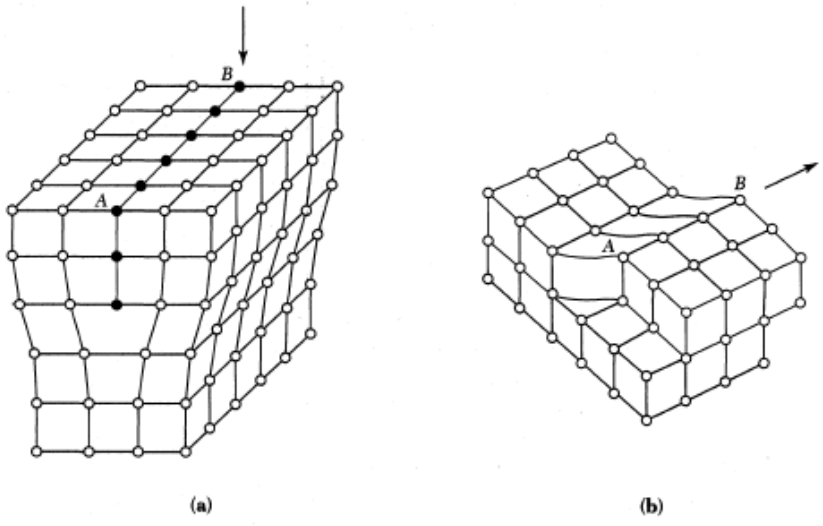


Fig. 2.28. *(a)* Edge and *(b)* screw dislocation formation in cubic crystals [35].

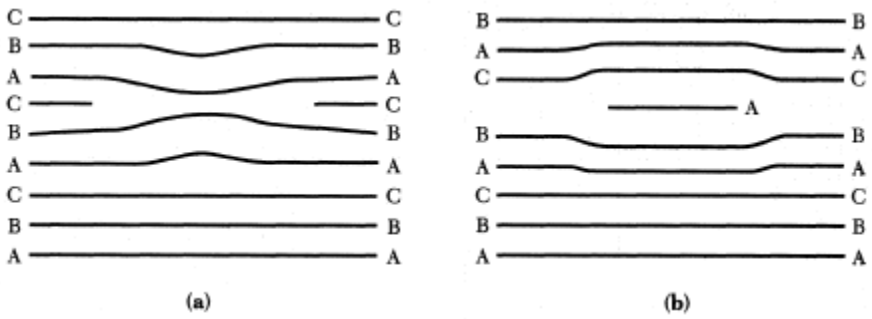


Fig. 2.29. Stacking fault in semiconductor. *(a)* Intrinsic stacking fault. *(b)* Extrinsic stacking fault [34].

Precipitates of impurities or dopant atoms make up the fourth class of defects, the volume defects. These defects arise because of the inherent solubility of the impurity in the host lattice. There is a specific concentration of impurity that the host lattice can accept in a solid solution of itself and the impurity. Figure 2.30 shows solubility versus temperature for a variety of elements in silicon [36]. The solubility of most impurities decreases with decreasing temperature. Thus, at a given temperature, if an impurity is introduced to the maximum concentration allowed by its solubility and the crystal is then cooled to a lower temperature, the crystal can only achieve an

equilibrium state by precipitating the impurity atoms in excess of the solubility level. However, the volume mismatch between the host lattice and the precipitates results in dislocations.

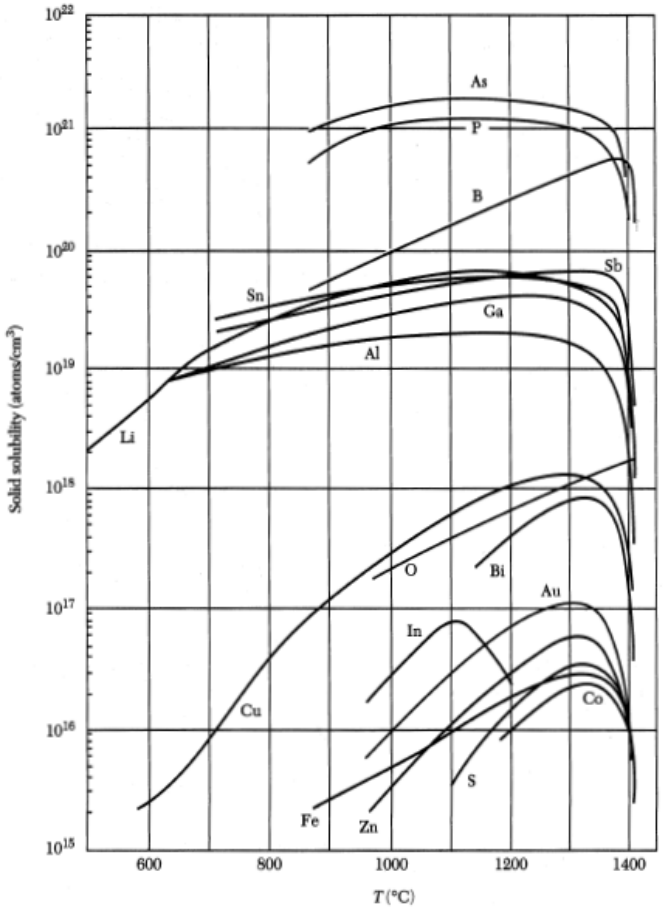


Fig. 2.30. Solid solubilities of impurity elements in silicon [34].

Material Properties

Table 2.5 compares silicon characteristics and the requirements for ultralarge-scale integration (ULSI) [37,38]. The semiconductor material properties listed in Table 2.4 can be measured by various methods. The resistivity is measured by the four-point probe method and the minority-carrier lifetime can be measured by the photoconductivity method. The trace impurities such as oxygen and carbon in silicon can be analyzed by the secondary-ion-mass spectroscope (SIMS) techniques. Note that although the current capabilities can meet most of the wafer specifications listed in Table 2.4, many improvements are needed to satisfy the stringent requirements for ULSI technology [38].

The oxygen and carbon concentrations are substantially higher in Czochralski crystals than in float-zone crystals because of to the dissolution of oxygen from the silica crucible and transport of carbon to the melt from the graphite susceptor during crystal growth. Typical carbon

concentrations range from 10^{16} to about 10^{17} atoms/cm^3 and carbon atoms in silicon occupy substitutional lattice sites. The presence of carbon is undesirable because it aids the formation of defects. Typical oxygen concentrations range from 10^{17} to 10^{18} atoms/cm^3. Oxygen, however, has both deleterious and beneficial effects. It can act as a donor, distorting the resistivity of the crystal caused by intentional doping. On the other hand, oxygen in an interstitial lattice site can increase the yield strength of silicon.

In addition, the precipitates of oxygen due to the solubility effect can be used for *gettering.* Gettering is a general term meaning a process that removes harmful impurities or defects from the region in a wafer where devices are fabricated. When the wafer is subjected to high-temperature treatment (e.g., 1050°C in N_2), oxygen evaporates from the surface. This lowers the oxygen content near the surface. The treatment creates defect-free (or *denuded*) zone for device fabrication, as shown in the inset of Fig. 2.31. Additional thermal cycles can be used to promote the formation of oxygen precipitate in the interior of the wafer for gettering of impurities. The depth of the defect-free zone depends on the time and temperature of the thermal cycle and on the diffusivity of oxygen in silicon. Measured results for the denuded zone are shown in Fig. 2.31. It is possible to obtain Czochralski crystals of silicon that are virtually free of dislocations.

Table 2.5. Comparison of silicon material characteristics and requirements for ULSI.

	Characteristics		
Property*	Czochralski	Float zone	Requirements for ULSI
Resistivity (phosphorus) *n*-type (ohm-cm)	1-50	1-300 and up	5-50 and up
Resistivity (antimony) *n*-type (ohm-cm)	0.005-10	-	0.001-0.02
Resistivity (boron) *p*-type (ohm-cm)	0.005-50	1-300	5-50 and up
Resistivity gradient (four-point probe) (%)	5-10	20	< 1
Minority carrier lifetime (μs)	30-300	50-500	300-1000
Oxygen (ppma)	5-25	Not detected	Uniform and controlled

Carbon (ppma)	1-5	0.1-1	< 0.1
Dislocation (before processing) (per cm^2)	≤500	≤500	≤1
Diameter (mm)	Up to 200	Up to 100	Up to 300
Slice bow (μm)	≤ 25	≤ 25	< 5
Slice taper (μm)	≤ 15	≤ 15	< 5
Surface flatness (μm)	≤ 5	≤ 5	< 1
Heavy-metal impurities (ppba)	≤1	≤0.01	0.001

*ppma, parts per million atoms; ppba, parts per billion atoms.

Commercial melt-grown materials of gallium arsenide are heavily contaminated the crucible. However, for photonic applications, most requirements call for heavily doped materials (between 10^{17} and $10^{18}cm^{-3}$). For integrated circuits or for discrete MESFET (metal-semiconductor field-effect transistor) devices, undoped gallium arsenide can be used as the starting material with a resistivity of $l0^9$ Ω-cm. Oxygen is an undesirable impurity in GaAs because it can form a deep donor level, which contributes to a trapping charge in the bulk of the substrate and increases its resistivity. Oxygen contamination can be minimized by using graphite crucibles for melt growth. The dislocation content for Czochralski grown gallium arsenide crystals is about two orders of magnitude higher than that for silicon. For Bridgman GaAs crystals, the dislocation density is about an order of magnitude lower than that for Czochralski-grown GaAs crystals.

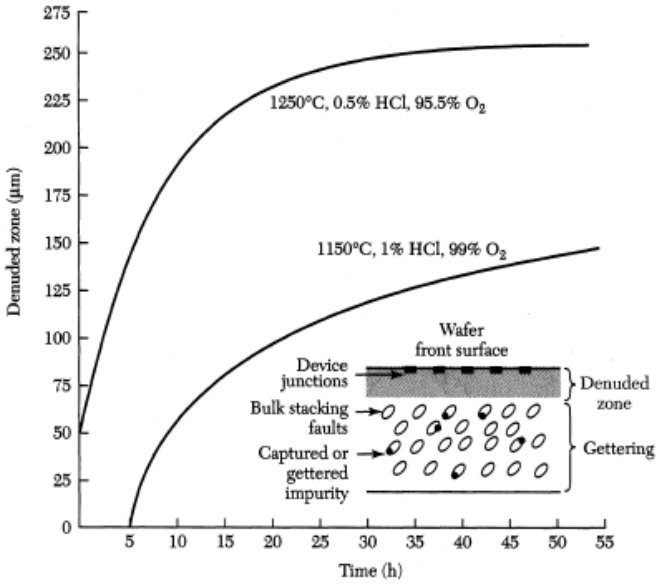


Fig. 2.31. Denuded zone width for two sets of processing conditions. Inset shows a schematic of the denuded zone and gettering sites in a wafer cross section [3].

2.8. Conclusions

Several techniques are available to grow single crystals of silicon and gallium arsenide. For silicon crystals, we use sand (SiO_2) to produce polycrystalline silicon, which then serves as the raw material in a Czochralski puller. A seed crystal with the desired orientation is used to grow a large ingot from the melt. Over 90% of silicon crystals are prepared by this technique. During crystal growth, the dopant in the crystal will redistribute. A key parameter is the segregation coefficient, i.e., the ratio of the dopant concentration in the solid to that in the melt. Since most of the coefficients are less than 1, the melt becomes progressively enriched with the dopant as the crystal grows.

Another growth technique for silicon is the float-zone process. It offers lower contamination than that normally obtained from the Czochralski technique. Float-zone crystals are used mainly for high-power, high-voltage devices where high-resistivity materials are required.

To make GaAs, we use chemically pure gallium and arsenic as the starting materials that are synthesized to form polycrystalline GaAs. Single crystals of GaAs can be grown by the Czochralski technique. However, a liquid encapsulant (e.g., B_2O_3) is required to prevent decomposition of GaAs at the growth temperature. Another technique is the Bridgman process, which uses a two-zone furnace for gradual solidification of the melt.

After a crystal is grown, it usually goes through wafer-shaping operations to give an end product of highly polished wafers with a specified diameter, thickness, and surface orientation. For example, 200 mm silicon wafers for a MOSFET (metal-oxide-semiconductor field-effect transistor) fabrication line should have a diameter of 200 ± 1 mm, a thickness of 0.725 ± 0.01 mm, and a surface orientation of (100) ± l°. Wafers with diameters larger than 200 mm are being manufactured for future integrated circuits. Their specifications are listed in Table 2.4.

A real crystal has defects that influence the electrical, mechanical, and optical properties of the semiconductor. These defects are point defects, line defects, area defects, and volume defects. We also discussed means to minimize such defects. For the more demanding ULSI applications, the dislocation density must be less than 1 per square centimeter. Other important requirements are listed in Table 2.5.

Chapter 3. Epitaxy

A. Evtukh

3.1. Introduction

When a material is grown epitaxially upon a substrate of the same material, the process is called homoepitaxy. That is the case for growth of silicon upon a silicon substrate. If, however, the layer is grown upon a chemically different substrate the process is termed heteroepitaxy. An example of heteroepitaxy is the epitaxial deposition of silicon on sapphire (SOS) [1-3]. Even when the crystal structures of the layer and substrate are basically the same, shifts of composition result in differences in lattice parameters. The resultant mismatch of lattice parameters at the film/substrate interface limit the technologist's ability to produce epitaxial layers of dissimilar materials.

In an epitaxial process, the substrate wafer acts as the seed crystal. Epitaxial processes are differentiated from the melt-growth processes in that the epitaxial layer can be grown at a temperature substantially below the melting point, typically 30-50% lower. The common techniques for epitaxial growth are chemical-vapor deposition (CVD) and molecular-beam epitaxy (MBE).

3.2. Chemical Vapor Deposition

CVD is also known as vapor-phase epitaxy (VPE). CVD is a process whereby an epitaxial layer is formed by a chemical reaction between gaseous compounds. CVD can be performed at atmospheric pressure (APCVD) or at low pressure (LPCVD) [3].

Fig. 3.1 shows three common susceptors for epitaxial growth. Note that the metric shape of the susceptor provides the name for the reactor: horizontal, pancake, barrel susceptors-all made from graphite blocks. Susceptors in the epitaxial reactors are analogues to the crucible in the crystal growth furnace. Not only do they mechanically support the wafer, but in induction-heated reactors they also serve as the source of thermal energy for the reaction. The mechanism of CVD involves a number of steps: (*a*) the reactants such as the gases and dopants are transported to the substrate region, (*b*) they are transferred to the substrate surface where they are adsorbed, (*c*) a chemical reaction occurs, catalyzed at the surface, followed by growth of the epitaxial layer, (*d*) the gaseous products are desorbed into the main gas stream, and (*e*) the reaction products are transported out of the reaction chamber

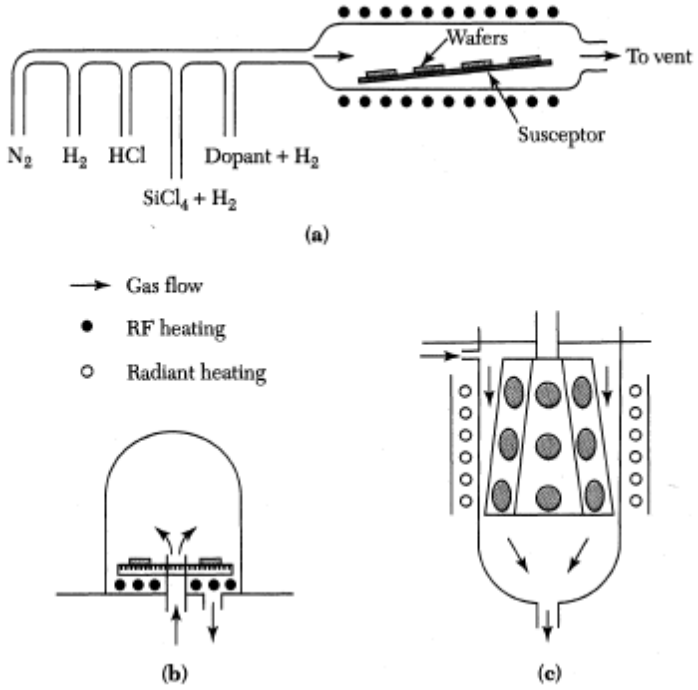


Fig. 3.1. Three common susceptors for chemical vapor deposition: (a) horizontal, (b) pancake, and (c) barrel susceptor.

3.2.1. Epitaxy of Silicon by CVD

Four silicon sources have been used for VPE growth. They are silicon tetrachloride ($SiC1_4$), dichlorosilane (SiH_2C1_2) trichlorosilane ($SiHCl_3$), and silane (SiH_4). Silicon tetrachloride has been the most studied and has the widest industrial use. The typical reaction temperatures is 1200°C. Other silicon sources are used because of lower reaction temperatures. The substitution of a hydrogen atom for each chlorine atom from silicon tetrachloride permits about a 50°C reduction in the reaction temperature. The overall reaction of silicon tetrachloride that results in the growth of silicon layers is

$$SiCl_4\,(gas) + 2H_2\,(gas) \leftrightarrow Si\,(solid) + 4HC1\,(gas). \qquad (3.1)$$

An additional competing reaction is taking place along with that given in Eq. 3.1:

$$SiC1_4\,(gas) + Si\,(solid) \leftrightarrow 2SiC1_2\,(gas). \qquad (3.2)$$

As a result, if the silicon tetrachloride concentration is too high, etching rather than growth of silicon will take place. Fig. 3.2 shows the effect of the concentration of silicon tetrachloride in the gas on the reaction, where the mole fraction is defined as the ratio of the number of molecules of a given species to the total number of molecules [4]. Note that initially the growth rate increases linearly with an increasing concentration of silicon tetrachloride. As the concentration of silicon tetrachloride is increased, a maximum growth rate is reached. Beyond that, the growth rate starts to decrease and eventually etching of the silicon will occur. Silicon is usually grown in the low-concentration region, as indicated in Fig. 3.2.

The reaction of Eq. 3.1 is reversible, that is, it can take place in either direction. If the carrier gas entering the reactor contains hydrochloric acid, removal or etching will take place. Actually, this etching operation is used for in-situ cleaning of the silicon wafer prior to epitaxial growth.

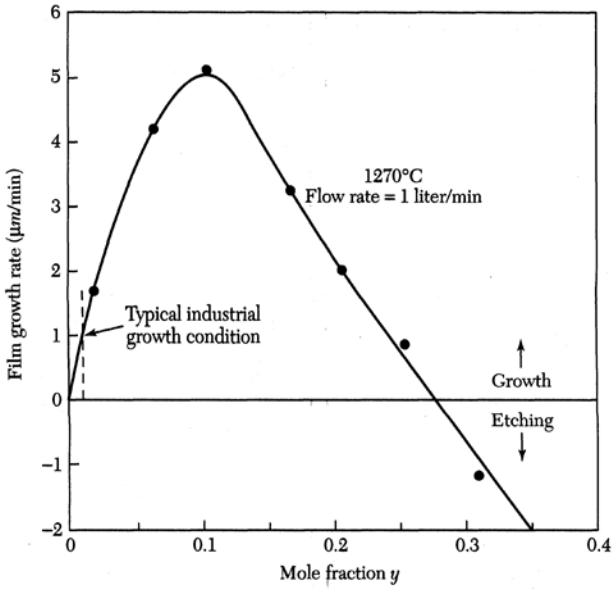


Fig. 3.2. Effect of $SiCl_4$ concentration on silicon epitaxial growth [5].

The dopant is introduced at the same time as the silicon tetrachloride during epitaxial growth (Fig. 3.1*a*). Gaseous diborane (B_2H_6) is used as the p-type dopant, whereas phosphine (PH_3) and arsine (AsH_3) are used as n-type dopants. Gas mixtures are ordinarily used with hydrogen as the diluent to allow reasonable control of flow rates for the desired doping concentration. The dopant chemistry for arsine is illustrated in Fig. 3.3, which shows arsine being adsorbed on the surface, decomposing, and being incorporated into the growing layer.

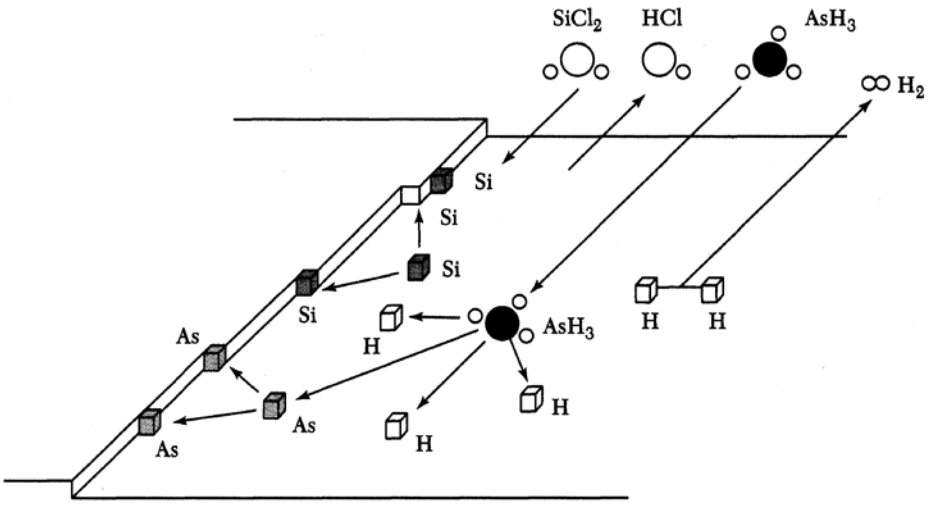


Fig. 3.3. Schematic representation of arsenic doping and the growing processes [6].

Fig. 3.3 also shows the growth mechanisms at the surface, which are based on the surface adsorption of host atoms (silicon) as well as the dopant atom (e.g., arsenic) and the movement of these atoms toward the ledge sites [6]. To give these adsorbed atoms sufficient mobility for finding their proper positions within the crystal lattice, epitaxial growth needs relatively high temperatures.

Growth Kinetics and Mechanisms. The most common model for addition of atoms to the growing epitaxial layer is the so-called plateau-ledge-kink model shown in Figure 3.4. In this model, the incoming silicon atom reacts on the plateau region. However, in that type site it forms only one or two bonds, and will surface diffuse to a “better fitting site”. The atom surface diffuses to a growth ledge which is a site of higher binding and then along the ledge to a kink site. There the adatom makes several near neighbor bonds, which reduces its ability to diffuse along the surface. Adatoms which fail to reach such low energy sites may react with species in the gas and leave the surface. In molecular beam epitaxy, the probability of an atom returning to the vapor phase is modeled as a sticking coefficient which is temperature and species dependent. A similar concept of a surface reaction probability can be used in vapor phase deposition.

Impurity atoms contained in surface precipitates may react with components of the gas (especially HC1) to remove impurity from the surface locally, resulting in a pit. Or alternatively, growth may be accelerated by the impurity resulting in a mound on the surface. Particles which are on the surface disrupt the crystal growth and result in various types of growth spikes.

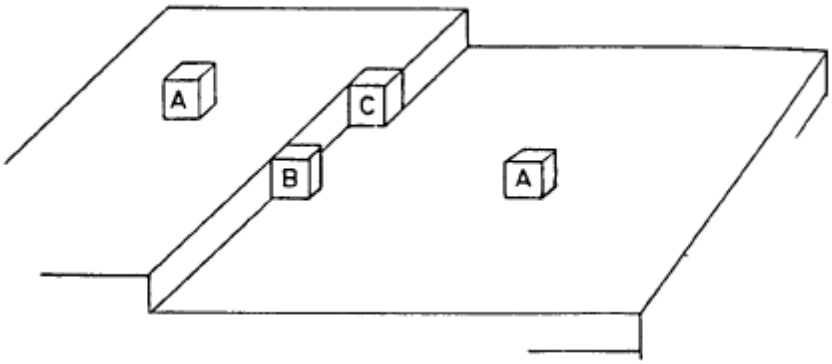


Fig. 3.4. Schematic of the ledge-kink growth model showing adatoms surface diffusing from positions A to B to C. Lower energy, more stable sites being characterized by an increasing number of near neighbor bonds.

Kinetics of Growth from Silane. Early models of growth from silane included only a single step of surface reaction for the pyrolysis reaction. More recently, however, it has suggested a more complex model [7]. In this model, the species adsorbed is SiH_2 instead of the SiH_4. The following

sequence has been suggested: (*a*) diffusion of silane through a boundary layer to near the surface, (*b*) dissociation to SiH_2 plus an H_2 molecule at or near the surface, (*c*) adsorption of the SiH_2, (*d*) surface diffusion to a kink site, (*e*) incorporation of silicon into the crystal lattice, and finally (*f*) desorption of the H_2. At temperatures above 1000°C, steps (*b*)-(*f*) occur rapidly and the overall process becomes limited by step (a) which is a diffusion limited supply of reactant to the surface.

It has been have shown [8] that for higher temperatures, the reaction rate is directly proportional to silane partial pressure until limited by supply kinetics (see Fig. 3.5). When the silane is diluted with hydrogen, the growth rate remains proportional to silane partial pressure but inversely proportional to the square root of hydrogen partial pressure.

Studies at atmospheric pressure at lower temperatures [7, 9] suggest that at the lower temperatures the silicon surface is covered by adsorbed hydrogen. The reaction then becomes limited by a surface reaction, probably the adsorption of the SiH_2.

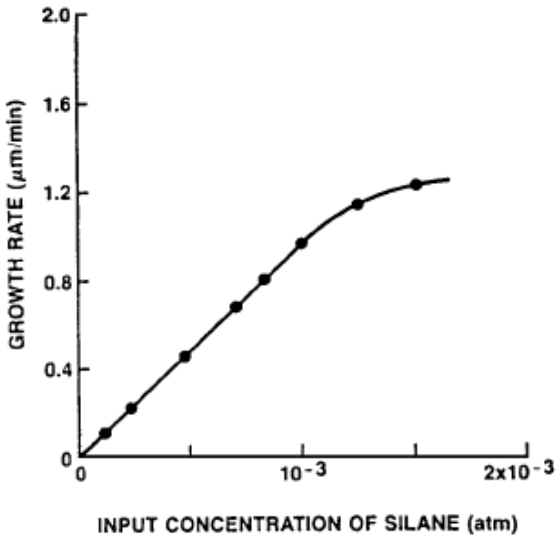


Fig. 3.5. Growth rate for epi increasing linearly with silane concentration until limited by kinetic effects [8, 10], however, found no such effect of hydrogen partial pressure at pressures below 1 Torr. Thus the surface site blocking action of adsorbed hydrogen may be dependent on total pressure.

Nucleation.

Homogeneous Nucleation. Nucleation of solid silicon from the silicon source gas can occur either in the gas phase (homonucleation) or upon a solid surface (heteronucleation). Nuclei forming in the gas phase are undesirable since they may fall upon the surface and nucleate defects in the growing layer. Since homonucleation requires a higher degree of supersaturation, prevention of gas phase nucleation consists of keeping the input source concentrations below critical levels and by careful ramping of the source gases during start-up of the reaction.

In homogeneous nucleation theory, a particle will only continue to grow if a nuclei reaches a critical size. When particles are below the critical size, the energy for creation of additional surface

is greater than the negative energy of formation of additional solid and the particle is unstable. Fluctuations may, however, allow the particle to attain the critical size beyond which is grows spontaneously with a decrease in net energy. Low supersaturation increases the size of the critical radius and makes gas phase nucleation less likely. The rate of nucleation is related to temperature and the change in free energy for formation of the critical radius particle, ΔE. It has been shown [11] that the nucleation rate, dN/dt, for homogeneous nucleation can be described by:

$$dN/dt = C \exp(-\Delta E/k_B T) \qquad (3.3)$$

where C is a constant, k_B is Boltzman's constant, and T is the absolute temperature.

The critical concentration of silane for gas phase (homogeneous) nucleation with the results shown in Fig. 3.6 [12,13]. The critical concentration to avoid gas phase nuclei for silane decreases with increasing temperature (decreasing values of reciprocal temperature). Studies have also shown that addition of some HCl to the gas stream decreases the gas phase nucleation rate [14].

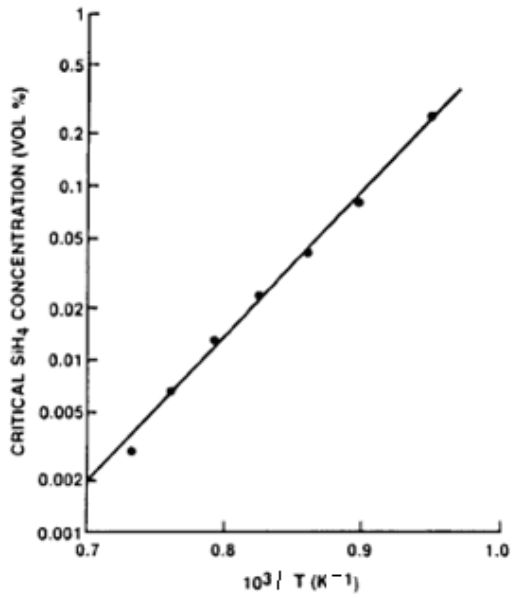


Fig. 3.6. Critical silane concentrations for homogeneous nucleation. Concentrations above the line produce gas phase nucleation (smoke) and highly defective epitaxial growth [12, 13].

Heterogeneous Nucleation. Below the supersaturations leading to homogeneous nucleation, nucleation upon a solid surface may still be energetically favored (heterogeneous nucleation). In that heterogeneous nucleation region, silicon is more likely to grow upon a silicon substrate at high temperature than upon a foreign substrate since no nucleation is involved and growth proceeds by the ledge-kink mechanism. Silicon surfaces near (111) contain few ledges and the growth near the isolated ledge areas develops large growth facets and leads to a rough surface. By cutting the substrate slightly off the (111) plane, (usually 3 degrees), a high density of growth ledges are

created and smooth epitaxial growth is obtained. Orientations near (100) have ample surface ledges and those substrates are generally cut on the (100) plane within 0.5 degree. It has been found [15] that the heterogeneous nucleation rate is strongly dependent upon the nature of the foreign substrate. At a given temperature and input concentration of the source gas, the nucleation rate decreases in the following order: (*i*) single crystal silicon, (*ii*) polysilicon, (*iii*) silicon nitride, (*iv*) aluminum oxide, (*v*) silicon oxide.

The nucleation is further suppressed by increasing the growth temperature and by additions of HCl to the gas stream as shown by data for the saturation nucleus density on SiO_2 in Fig. 3.7 [16]. By taking advantage of those nucleation trends, epitaxial layers may be deposited selectively upon patterned substrates masked with silicon oxide or nitride with little or no deposition upon the mask material [15].

3.2.2. Epitaxy of GaAs by CVD

There are four main techniques by which GaAs and AlGaAs epitaxial films are grown: chloride transport vapor phase epitaxy (VPE) [17], liquid phase epitaxy (LPE) [18], molecular beam epitaxy (MBE) [19], and metalorganic chemical vapor deposition (MOCVD) [20]. These epitaxial growth techniques are compared in Table 3.1 [21].

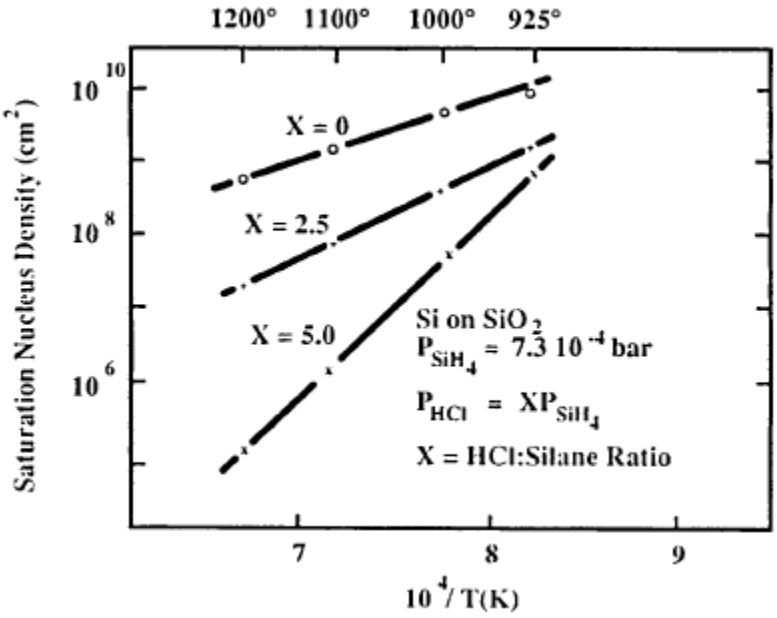


Fig. 3.7. Nucleation density on silicon dioxide versus reciprical temperature for silane growth with and without additions of HCl [16].

In the chloride transport VPE growth system, any silicon contamination is serious for it creates unfavorable thermodynamics in the compound and alloy containing Al, making it very

difficult for growth to occur. The LPE technique, through it has been successfully used in compound semiconductors, is not suitable for mass production by substrate limitation. The abruptness of the interface in LPE growth is unsatisfactory for GaAs high-speed device fabrication.

Table 3.1. Comparison of epitaxial growth techniques

	LPE	Hydride VPE	MOCVD	MBE
Al alloys	Capable	Difficult	Capable	Capable
Range of growth rate (μm/min)	0.1-10	0.01-0.5	0.005-1.5	Few-0.05
Minimum thickness (nm)	50	25	2	0.5
Homogeneity	good	good	good	good
Surface morphology	bad	good	good	good
Abruptness of interface	bad	good	good	excellent
Doping level (cm^{-3})	10^{14}-10^{19}	10^{14}-10^{19}	10^{14}-10^{19}	10^{14}-10^{19}
Number of heating point	1	2	1	3
Productivity	low	high	high	very low

For gallium arsenide, the basic setup is similar to that shown in Fig. 3.3. Since gallium arsenide decomposes into gallium and arsenic upon evaporation, its direct transport in the vapor phase is not possible. One approach is the use of As_4 for the arsenic component and gallium chloride ($GaCl_3$) for the gallium component. The overall reaction leading to epitaxial growth of gallium arsenide is

$$As_4 + 4GaC1_3 + 6H_2 \rightarrow 4GaAs + 12HC1. \quad (3.4)$$

The As_4 is generated by thermal decomposition of arsine (AsH_3):

$$4AsH_3 \rightarrow As_4 + 6H_2, \quad (3.4a)$$

and the gallium chloride is generated by the reaction

$$6HCl + 2Ga \rightarrow 2GaCl_3 + 3H_2. \quad (3.4b)$$

The reactants are introduced into a reactor with a carrier gas (e.g., H_2). The gallium arsenide wafers are typically held within the 650-850°C temperature range. There must be sufficient arsenic overpressure to prevent thermal decomposition of the substrate and the growing layer.

3.3. Metalorganic CVD

Metal-organic chemical vapor deposition (MOCVD) technology represents the fastest growing and most promising technology for the new compound semiconductor industry. The technique involves the reaction, at a temperature well below the melting point of the resultant solid,

of two or more chemically reactive gases at atmospheric or reduced pressure. The MOCVD is also a VPE process based on pyrolytic reactions. Unlike the conventional CVD, MOCVD is distinguished by the chemical nature of the precursor. It is important for those elements that do not form stable hydrides or halides but that form stable metalorganic compounds with reasonable vapor pressure. MOCVD has been extensibly applied in the heteroepitaxial growth of III-V, III-N, and II-VI semiconductor compounds.

3.3.1. Metalorganic CVD of III-V Semiconductors

The MOCVD technique has demonstrated its effectiveness for growing the widest variety of III-V semiconductor materials. As pointed out in Table 3.1, the MOCVD technique has several advantages over other growth techniques [22, 23]: (1) The formation of the desired compound occurs via the pyrolysis of the metalorganics and hydrides, and the subsequent recombination of the atomic or molecular species occur at or near the substrate surface. (2) The composition and impurity concentration can be controlled well by fixing the flow rates of the various reactants with electronic mass flow controllers. (3) Complex multilayer epitaxial structures are readily formed by exchanging one gas composition for another gas using automatic gas mixing system. (4) This technique is suitable for mass production based on its similarity to silicon CVD process.

Let’s consider mainly pyrolysis reactions where one or more of the sources involved is a metal alkyl. In a typical reaction of this type, a lower-order metal alkyl, such as trimethylgallium (TMGa), is mixed in the vapor phase with a hydride, such as arsine. As a result of pyrolysis, atomic or molecular species combine at or near a heated substrate. The deposition results in epitaxy on a single-crystal substrate such as GaAs or InP. The emphasis is on epitaxial growth, and on the fact that the metal alkyls in typical use belong to a broader class of compounds known as metal organic chemical vapor deposition (MOCVD) [24]. There are several reasons for MOCVD to become an important epitaxial growth technology in a wide variety of III-V, III-N, and II VI materials. For example, all constituents are in the vapor phase, which allows for accurate electronic control of such important system parameters as gas flow rates and hence partial pressures. The pyrolysis reaction is relatively insensitive to growth temperature, allowing for efficient and reproducible deposition of thin layers and abrupt interfaces between deposited layers. Complex multiple layer heterostructures can be grown utilizing computer-controlled automatic gas-exchange systems.

To grow GaAs, it is possible to use the metalorganic compounds such as trimethylgallium $Ga(CH_3)_3$ for the gallium component and aresine AsH_3 for the arsenic component. Both chemicals can be transported in vapor form into the reactor. The overall reaction is

$$AsH_3 + Ga(CH_3)_3 \rightarrow GaAs + 3CH_4. \qquad (3.4)$$

For Al-containing compounds, such as AlAs, it is possible to use trimethylaluminum $A1(CH_3)_3$. During epitaxy, the GaAs is doped by introducting dopants in vapor form. Diethylzine $Zn(C_2H_5)_2$ and diethylcadmium $Cd(C_2H_5)_2$ are typical *p*-type dopants and silane SiH_4 is an n-type dopant for III-V compounds. The hydrides of sulfur and selenium or tetramethyltin are also used for *n*-type dopants; and chromyl chloride is used to dope chromium into GaAs to form semiinsulating layers. Since these compounds are highly poisonous and often spontaneously inflammable in air, rigorous safety precautions are necessary in the MOCVD process.

A schematic diagram of an MOCVD reactor is shown in Fig. 3.8 [25]. Typically, the metalorganic compound is transported to the quartz reaction vessel by hydrogen carrier gas, where it is mixed with AsH_3 in the case of GaAs growth. The chemical reaction is induced by heating the gases to 600-800°C above a substrate placed on a graphite susceptor using radiofrequency heating. A pyrolytic reaction forms the GaAs layer. The advantages of using metalorganics are that they are volatile at moderately low temperatures and there are no trouble some liquid Ga or In sources in the reactor.

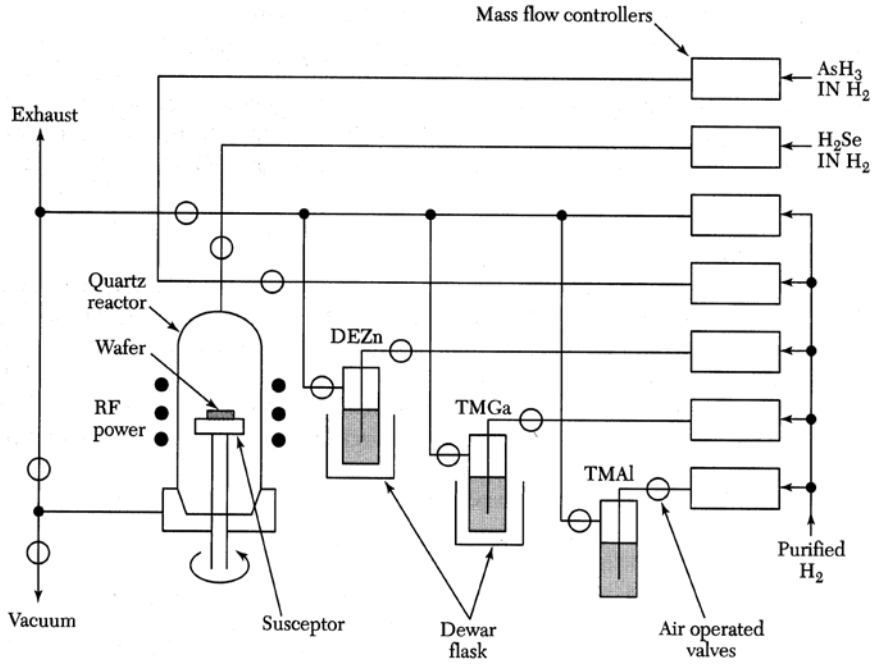


Fig. 3.8. Schematic diagram of a vertical atmospheric-pressure metalorganic chemical-vapor deposition (MOCVD) reactor [25]. DEZn is diethylozinc $Zn(C_2H_5)_2$, TMGa is trimethylgallium $Ga(CH_3)_3$ and TMAl is trimethylaluminum $A1(CH_3)_3$.

3.3.1.1. Components Sources

Metal-organic sources. The result metal-organic sources normally have two basic characteristics: (1) They must have suitable vapor pressures (~10 torr) at reasonable temperature (-20° to + 20°C), and (2) they must thermally decompose at typical growth temperatures to yield the

desired group-III or group-V element for the growth process [24]. The vapor pressure of several OM sources used for III V growth are plotted versus temperature in Figs. 3.9 and 3.10 [26].

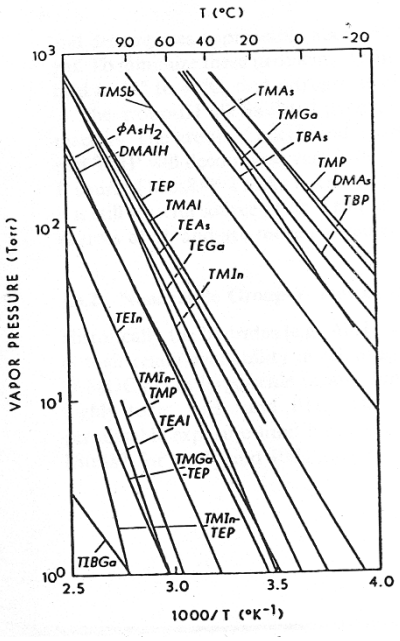


Fig. 3.9. Temperature dependence of vapor pressures for common group III and group V organometallic sources [26].

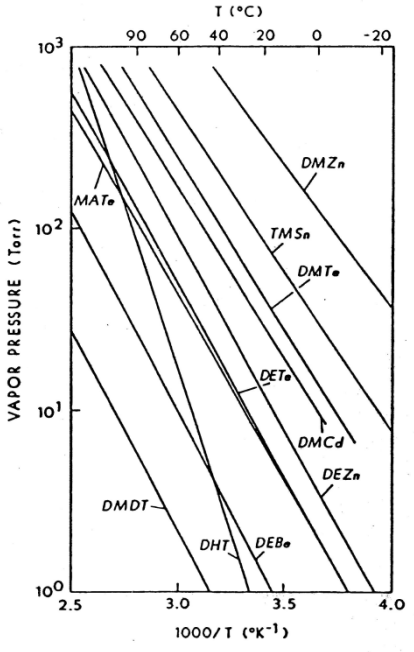


Fig. 3.10. Temperature dependence of vapor pressures for common group II and group VI organometallic sources [26].

In general, alkyls with the highest vapor pressure, usually the lowest molecular preferred. Thus TMGa and TMA1 are used whenever possible for GaAs and AlGaAs growth. TMIn is a solid

at room temperature due to its unique tetrameric nature; hence it is either sublimed or heated to above its melting point. This necessitates heating all lines between the TMIn bubbler and the reactor, or dilution with large quantity of H_2 before leaving the heated part of the bubbler. TMIn has the highest vapor pressure of all the indium alkyls. The triethyl-III (TE-III) compounds are also used, but their lower vapor pressures usually yield lower growth rates. They are even less stable, tending to form polymers in atmospheric pressure reactors, which further reduces the growth rate [27, 28] and may lead to inhomogeneous growth. However the less carbon contamination may be expected in the GaAs epilayer when it is deposited by the TMGa method [29, 30].

Triethylindium (TEIn) is a liquid source. Due to its loose polymer structure, it has a very low vapor pressure. It is also much less thermally stable than the trimethyl and decomposes appreciably above 40°C, as well as on exposure to hydrogen carrier gas. To minimize these problems, TEIn is used in reduced pressure deposition systems held at 35° to 40°C, and nitrogen is used as a carrier gas.

The group-III alkyls listed pyrolize efficiently in normal low pressure, or 1-atm, reactors. The group-V alkyls and hydrides are known to pyrolize more slowly. TEP and TMP will decompose very little, even less than PH_3, and are thus an ineffective P source [31]. TMAs is known to pyrolize less rapidly than AsH_3 [32], but it is still a useful source of As. Trimethyl and triethyl antimony are both very effective sources of Sb and are more convenient to use than SbH_3 [33, 34].

Nonhydride Group-V Sources for MOCVD. Historically the hydrides (e.g., AsH_3 and PH_3) have been used for MOCVD because of their ready availability in relatively high pure form. A major obstacle to the use of MOCVD in large-scale production operations is in the use of large quantities of highly toxic AsH_3 and PH_3. The threshold limit values (TLVs) the maximum permissable exposure limit based on a time weighted average for an eight hour day [35], for AsH_3 and PH_3 are 0.05 and 0.3 ppm, respectively [36]. The value of LC_{50} (rats), defined as the lethal concentration for 50% of the population in rat testing, necessarily involves both concentration level and length of exposure. Typically only the concentration is given. The exposure time, which is assumed to be on the order of four hours, to be 11 ppm for PH_3 [37]. No such value is listed for arsine. In recent years the general public has become increasingly aware to alert to the dangers associated with the use of toxic materials near residential neighborhoods. A large fraction of the expense of both purchasing and operating a reactor are deemed safe by today's standards is devoted to safety features. Fortunately a number of organometallic group-V sources are much less toxic than the hydrides. The ideal group-V source would be a nontoxic liquid with a moderate vapor pressure *P* to 500 torr.

Requirements for group- V sources. The requirements for group-V sources for MOCVD are stringent. They include high vapor pressure, low-temperature stability, pyrolysis at temperatures at and above 400°C, no inherent purity limitations such as excess carbon contamination, and no interaction with the group-III sources producing parasitic reactions.

The commonly used nonhydride sources are low vapor pressure liquids or solids at room temperature, which is advantageous from a safety viewpoint. However, to useful for MOCVD, at room temperature the vapor pressures should be greater than 50 torr, in order to avoid using extremely high carrier gas flow rates through the bubblers in which the sources are contained or heating of the bubbler and downstream lines. The group-V sources that have been successfully used for MOCVD are listed in Table 3.2 [35] along with their vapor pressures. Clearly the elemental sources must be heated to temperatures well above 300 K. Di- and triethylarsine also have vapor pressures too low to be conveniently used without heating the bubblers.

Table 3.2. Vapor pressure of group V sources for OM VPE.

Compound	Vapor pressure (torr)	Temperature (°C)
Phosphorus		
P	1	260
PH_3	760	-87.8
TMP	381	20
TEP	46.5	50
IBP	122	23
TBP	286	23
Arsenic		
As	1	370
AsH_3	760	-55
TMAs	238	20
TEAs	15.5	37
DEAs	0.6	18
TBAs	96	-10

Two factors are important when considering the stability of the group V sources. First, the materials must be relatively easy to synthesize and purify. Second, the materials suitable for sources should have shelf lifetimes at room temperature measured in years.

The group-V source molecules must pyrolyze at the relatively low temperatures used for MOCVD growth. At present, GaAs growth temperatures as low as 550 °C are common. For smaller band-gap materials, with lower bond strengths giving lower melting points, the optimum growth temperature may be considerably lower. For example, InSb melts at 530°C, so the growth temperatures must be somewhat lower. Only the slower pyrolysis of common sources such as AsH_3 and TMSb prevents the use of even lower growth temperatures. AsH_3 is 50% pyrolyzed only at temperatures of greater than 757°C [38]. Even higher temperatures of greater than 850°C are required for PH_3 pyrolysis [39].

Another practical requirement is the absence of reactions of the group V sources with the group-Ill organometallic sources, leading to depletion of the source materials from the vapor phase upstream from the substrate. Naturally a reaction that yields the III-V semiconductor at high temperatures is desired. However, parasitic reactions frequently lead to decreased growth rates due to deposition of undesirable nonvolatile material on the reactor walls. An example is the interaction of both TEGa and TEIn with AsH_3 and PH_3 to form adducts that subsequently eliminate methane forming nonvolatile polymers [40]. Fortunately the more commonly used TMGa and TMIn do not interact with the group-V hydrides in parasitic reactions at ordinary growth temperatures [40].

3.3.1.2. Basic Reaction

During growth by CVD a number of processes take place, partly in series and partly in parallel. Their relative importance depends on both the chemical nature of the species and the design of the reactor. Reactor chamber designs are divided into between horizontal types [41] and vertical types [42]. The growth rate is determined by the slowest process in series of events needed to come to deposition. In the case of MOCVD the growth rate in epitaxial growth is controlled by the diffusion of the metal organic components through the boundary layers [43, 44]. The epitaxial layers of GaAs can be grown at atmospheric pressure or under reduced pressure (LP MOCVD) system.

The fundamental aspects of the CVD process have been investigated, and the models developed for the process are based on the assumption that the arrival of reactant species on the growth surface is limited by diffusion through a boundary layer [45, 46].

Basic chemical reactions in MOCVD. In the discussion of basic MOCVD reactions for the growth of compound semiconductor epitaxial layers, let’s consider to those reactions involving organometallic compounds or mixed organometallic compounds and hydrides. The exact chemical decomposition pathways in MOCVD are not yet clearly understood. The nature of the reactions are in part determined by the dynamic of the gases, that is, by the velocity and temperature profiles in

the vicinity of the susceptor and the subsequent concentration and thermal gradients that are established [47-49].

The reaction pathways of course are also strongly influenced by the choice of precursor chemicals. Alkyls of the group-III metals and hydrides of the group-V elements are usually used as precursor species in MOCVD. Dilute vapor of these chemicals are transported at or near room temperature to a hot zone where a pyrolysis reaction occurs. The most commonly used reactions for the growth of compound semiconductor MOCVD layers is given in the general form

$$R_nM + XH_n \rightarrow MX + nRH, \tag{3.5}$$

where *R* is the organic radical, *M* is one component of the resulting semiconductor layers, *X* is the other component, *n* is the integer.

In the reaction of Eq. (3.5) the simple organometallic species represented by R_n-M reacts with X-H_n to generate the compound semiconductor MX and a residual organic compound RH. Important examples of these reactions are

$$Ga(CH_3)_3 + AsH_3 \rightarrow GaAs + 3CH_4. \tag{3.6}$$

$$x(CH_3)_3Al + (1 - x)(CH_3)_3Ga + AsH_3 \rightarrow Al_xGa_{1-x}As + 3CH_4. \tag{3.7}$$

Although relatively simple to describe in an equation, the actual kinetics and mechanism associated with the heterogeneous reaction between the materials at the surface of the growing interface and the gas phase reactions are quite complex. The right-hand side of Eq. (3.7) is particularly interesting because it reflects impact of alloy layers and heterostructures on compound semiconductor devices. In all of these reactions, the metal alkyl is stored as a liquid at low temperature (-10°C for TMGa) and the hydrides are gaseous. The metal alkyl is transported to the quartz reaction cell (termed a Bass cell) by hydrogen carrier gas where it is intimately mixed with arsine and brought into the vicinity of a heated susceptor. The susceptor can be heated using an RF coil that surrounds the cell. The reactants are pyrolyzed by the heat of the susceptor and fragment into atomic or molecular forms of the component species, which then combine to form the deposited semiconductor.

Another alternative is the substitution of a related organometallic compound for the hydride source. This is described by a reaction of the general form [50]

$$R_nM + R^*_mX \rightarrow MX + nRH + mR^*H, \tag{3.8}$$

where *R, R** are the organic radicals, *M, X* are the compound semiconductor components.

The H_2 indicates that these reactions typically take place in a reducing atmosphere. Some typical examples of these types of reaction are

$$(CH_3)_3Ga + (CH_3)_3As \rightarrow GaAs + 6CH_4 \tag{3.9}$$

and

$$(1 - x)(CH_3)_3In + x(CH_3)_3Ga + (CH_3)_3Sb \rightarrow In_{1-x}Ga_xSb + 6CH_4. \quad (3.10)$$

Reactions involving only organometallic sources have not been studied to the extent that the metal alkyl-hydride systems have been studies. There is the possibility that organometallic compounds are easier to purify than are gaseous hydrides [51]. Background doping levels are also lower in epitaxial layers grown only with organometallic compounds.

Some metal alkyls, particularly indium compounds, react rapidly at room temperature with hydrides, such as PH_3, to form adducts of the form

$$R_nM{:}XH_n. \quad (3.11)$$

An example of this is

$$(CH_3)_3In{:}PH_3, \quad (3.12)$$

which is thought to be polymeric at room temperature and a simple 1:1 adduct at lower temperature [52]. The group III metal alkyls are strong Lewis acids and hydrides are Lewis bases. It is thus not surprising that these adducts are formed [53]. No such intermediate reactions are observed in the TMGa – AsH_3 system.

Several empirical approaches have been used to minimize the formation of intermediate reaction products or parasitic polymers that can deleteriously affect the growth rate or composition of an MOCVD-grown epitaxial layer. The first approach is the *physical solution*. It is based on the recognition that the rate of formation of the adduct and the rates of the elimination reactions may be rather slow, possibly on the order of a fraction of a second [54]. Through the use of low pressure, the residence time of the source molecules, together in the gas phase before reacting with the hot substrate, is reduced by at least one order of magnitude. This has allowed to grow layers of InP and GaInAsP alloys using TEIn and PH_3 [55-57]. One approach includes prepyrolysis the PH_3 prior to mixing it with TEIn in a low pressure (0.1 atm) MOCVD system [58]. It was produced GalnAs layers using TEIn, TEGa, and AsH_3 but observed no intermediate reaction product or parasitic polymer formation problems [52]. It was tried the additional step of passing the PH_3 through a 760 C furnace to partially crack the PH_3, using a carrier gas of 50% N_2 and 50°/ H_2 [52, 55].

The second approach might be termed the *chemical solution*. The adduct $(C_2H_5)_3In\text{-}PH_3$ probably decomposes by the spontaneous elimination of C_2H_6 molecules formed from one C_2H_5 of In and one H of P. Simply cracking the PH_3 before H enters the reactor could eliminate this process, especially at low H_2 pressure [59]. Alternatively, we could substitute TEP or TMP for PH_3. The elimination reaction is clearly less favorable in this case, partly because no stable organic byproduct is produced and partly because of the tight binding of the ethyl radical to the phosphorus. It was reported that the P from TEP is not incorporated into the solids [53]; GalnAs can be grown from TMIn -TEP, TMGa, and AsH_3. It was used a similar approach using a TMIn-TMP adduct as the In

source and PCl_3 as the P source [60]. It was also have substituted TMAs for AsH_3 in the growth of In- and As-containing alloys [61].

Basic growth mechanisms. The basic growth mechanisms as a function of growth temperature is shown in Fig. 3.11 [62]. At lower temperatures (< 550°C) the growth mechanism is surface kinetic limited. Between 550°C and 750°C it is mass transport limited, which is nearly temperature independent. At higher temperature (>800°C) it is thermodynamics limited, where temperature increase leads to growth rate decrease. With the onset of kinetic growth, the region shifts to higher temperatures and the bond strength of the used As compound (AsH_3 < TEAs < TMAs) increases [63].

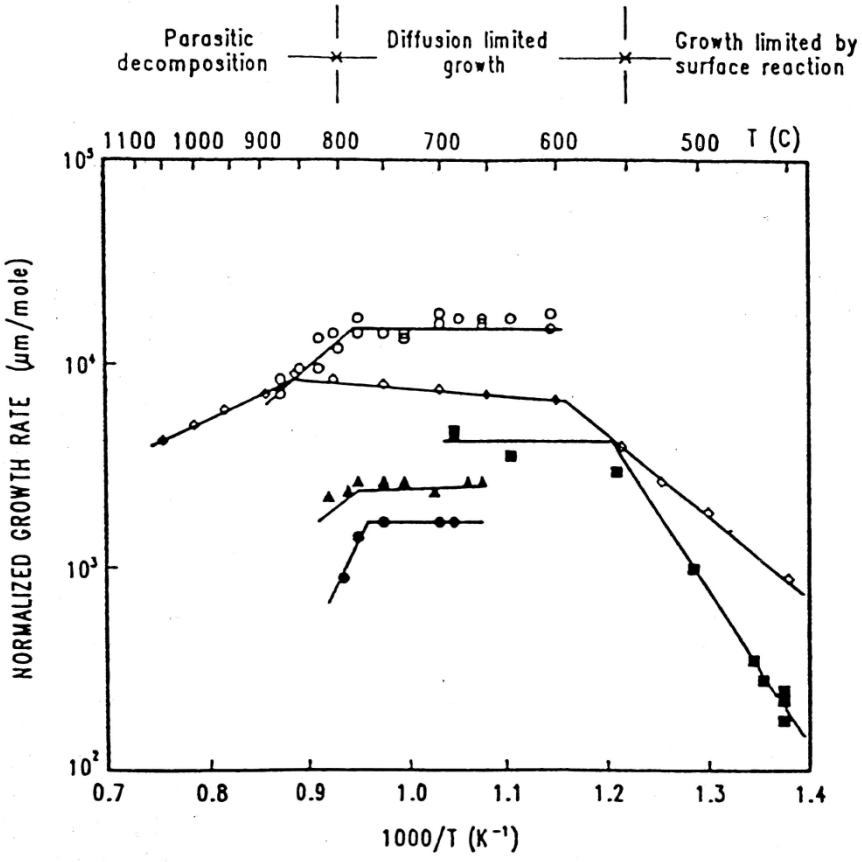


Fig. 3.11. Normalized growth rate as a function of inverse of temperature [62].

3.3.1.3. Purity and Dopants

The purity of GaAs grown by MOCVD is controlled by the growth temperature, the purity of the starting materials, and the arsenic-to-gallium ratio in the reactor. The purity of GaAs grown by TMGa and AsH_3 increases monotonically with decreasing growth temperature [64, 65], It was indicated that this increase in purity is due to a monotonic decrease in the incorporation of carbon and silicon in GaAs grown by MOCVD [64]. Similar studies indicated a less strong, but important,

dependence of the total, electrically activated, impurity concentration in GaAs on the arsenic-to-gallium ratio in the reactor [65]. The purity of both the TMGa and AsH_3 is vitally important to obtaining high-purity GaAs.

Since the initial work on MOCVD of GaAs, researchers have been concerned with the incorporation of carbon as a background impurity [66-68]. This concern was accentuated by early studies that showed high background C levels [66]. The acceptor behavior of C in GaAs is well known, and C acceptors can be readily distinguished by using low-temperature PL techniques [69, 70]. Carbon acceptors are seen in virtually all GaAs grown using TMGa [71]. It was shown that an increase in growth temperature results in large increase in carbon incorporation [35, 72].

In the MOCVD process the carbon (the alkyl radicals resulting from the surface decomposition reaction) is removed from the growing surface much more easily [73]. The interpretation for this effect is the possibility of β-elimination processes in the decomposition of TEGa [74]:

$$(C_2H_5)_3Ga \rightarrow (C_2H_5)_2GaH + C_2H_4 \ (\beta\text{-elimination}). \qquad (3.13)$$

The $Ga(C_2H_5)_3$ (TEGa) decomposes stepwise into components $GaH_n(C_2H_5)_{3-n}$ with simultaneous formation of C_2H_4. The stable C_2H_4 molecules formed from TEGa does not lead to the carbon incorporation [75, 76]. The C_2H_4 molecule has a very low probability of sticking on the GaAs surface even at room temperature.

β-elimination is not possible in the TMGa decomposition; the split-off of Ga leaves a very reactive CH_3 radical, which sticks to the surface quite readily. The CH_3 can be transformed into the little-reactive CH_4 by adding atomic H, the result of AsH_3 decomposition.

Carbon contamination in $Al_xGa_{1-x}As$ is a more severe problem. Electron mobility is dramatically reduced with increasing *x*. This is believed to be due to C being incorporated into the strong Al-C bond, which decreases the activity coefficient of C in the solid. The C concentration determined from the temperature dependence of electron mobility is approximately $10^{17}cm^{-3}$ for $Al_xGa_{1-x}As$ grown with TMGa, TMAl, and AsH_3.

Another major residual impurity is oxygen. For GaAs a few parts per million (ppm) of O_2 and/or H_2O in the gas stream has little effect on the properties. In fact ^{18}O-doping studies indicate that when a few ppm of O_2 are added to the gas stream, less than $10^{16}cm^{-3}$ is incorporated into the solid [77]. For $Al_xGa_{1-x}As$, on the other hand, ^{18}O-doping studies clearly show that 1 ppm of $^{18}O_2$ in the vapor gives 10^{20} cm^{-3} of ^{18}O in the solid. Assuming that the $P(O_2)$ at the growing interface is zero, this is exactly what would be predicted for a total flow rate of 2 liters/min if the reactor is about 10% efficient.

In the solid $10^{19}cm^{-3}$ of oxygen is sufficient to affect adversely the PL efficiency. This is a serious problem. It is very difficult to maintain a gas stream with <0.1 ppm O_2 from leaks, desorption from walls, outgassing from the susceptor, and perhaps most important, an impurity in the AsH_3 [40]. Studies by several groups have shown that PL efficiency [78] and mobility [79] are strongly affected by the AsH_3 source. Several techniques have been described for purification of the gas stream just above the substrate.

The use of graphite baffles to catalyze the reaction between TMAl and any O_2 or H_2O in the gas stream is now well established. The effect was first observed as an increase in PL efficiency of MOCVD AlGaAs when graphite baffles were introduced into the gas stream [80]. If the SiC-coated graphite baffles are replaced by simple graphite baffles, there is a reduction of ^{l8}O in the epitaxial layer by a factor greater than 10^3.

Let's next consider the *n*- and *p*-type dopants. A number of donor impurities have been used in MOCVD, including the group VI elements S, Se, and Te and the group-IV elements Si and Ge. There are fewer acceptor impurities to choose from, since the group-IV elements are not amphoteric. Thus we are left with Zn, the major *p*-type dopant, and Cd, Be, and Mg. These some dopants have been used for the entire aims of III-V compounds and alloys.

Selenium is probably the most widely used *n*-type dopant in MOCVD material. Hydrogen selenide (H_2Se) is used for the Se source. It has a distribution coefficient of somewhat less than unity, (the distribution coefficient can be defined by the fraction of the group-III or V lattice sites filled by the dopant atoms, divided by the ratio of input partial pressure) depending on the growth temperature and the III/V ratio. It is expected to have a lower distribution coefficient with increasing temperature because of its increased volatility [40], as is generally seen to be the case for the volatile dopants Se, S, and Zn. There is a decrease in electron carrier concentration when the growth temperature increases, which is due to surface desorption of selenium. The use of Se apparently produces good PL efficiencies and minority carrier lifetimes in GaAs and $Al_xGa_{1-x}As$ [80].

Tellurium is obtained from DETe, which may be purchased diluted in H_2 in high- pressure cylinders. In the MOCVD growth of GaAs, Te can be used to produce doping levels as high as $10^{18}cm^{-3}$ with only the expected increase in the PL efficiency [81]. At high doping levels, all impurities produce nonradiative recombination centers either through complex formation or precipitation. However, low-threshold current density lasers with very good reliability can be fabricated using Te as the *n*-type dopant.

Silicon acts as donor in MOCVD GaAs and AlGaAs. The effective distribution coefficient and PL efficiency of the resulting material both increase strongly with increasing substrate temperature. The electron carrier concentration increases with growth temperature for a fixed silane

molefraction. Silicon is the *n*-type dopant used for injection-laser devices. In summary, the best donor impurities for MOCVD growth appear to be Se, from an H_2Se source, Te using DETe, and Si from SiH_4.

The standard *p*-type dopant for MOCVD is Zn from either DEZn or DMZn. Although DMZn and DEZn sources are suitable for the *p*-type dopant over the entire doping range anticipated for most usages, DEZn appears to be the most controllable and provides the widest range of doping for all alloy compositions. Zinc has a small distribution coefficient especially at high substrate temperatures. The effect of the V/III ratio is the same as that expected for a group-II impurity substituting on the group-III sublattices. That is, as the V/III ratio increases, so does the Zn concentration in the solid.

3.3.2. Epitaxy of III-N by MO CVD

High-quality epitaxial III-N (GaN, InN, AlN) films and heterostructures for devices have been accomplished by organometallic vapor phase epitaxy (OMVPE) technique [82]. This technique has been applied to the deposition of GaN and AlN [83]. Using triethylgallium (TEG) and ammonia (NH_3) as source gases for group III and V species, respectively, it was obtained c-axis oriented films on sapphire (0 0 0 1) and on 6H-SiC(0 0 0 1) substrates. MIS-like LEDs followed, albeit they relied on deep states induced by Zn and suffered from very low efficiencies because of their poor crystalline quality. The development of low temperature (LT) buffer layers addressed the quality issue some [84]. The technique improved over the years to the point that undoped GaN films with a low background carrier concentration of 5×10^{16} cm^{-3} and with an X-ray symmetric peak FWHM of ~30 arcsec have been grown [85]. The X-ray data should be treated with caution, as the symmetric peak is not as sensitive to the edge dislocation as the asymmetric peak. OMVPE has been used for the development of LEDs [86], lasers [87], transistors [88], and detectors [89].

The best OMVPE reactors for group III nitride film growth incorporate laminar flow at high operating pressures and separate inlets for the nitride precursors and ammonia to minimize predeposition reactions. A successful, two-flow OMVPE reactor is shown in Fig. 3.12 [90]. The main flow composed of reactant gases with a high velocity is directed through the nozzle parallel to a rotating substrate. The subflow gas composed of nitrogen and hydrogen is directed perpendicular to the substrate. The purpose of the flow normal to the substrate surface is to bring the reactant gases in contact with the substrate and to suppress thermal convection effects. Hydrogen is the carrier gas of choice. A rotating susceptor was used to enhance uniformity of the deposited films. If one goes with the premise that smallest rocking curve half width implies an all-around good quality, GaN films can claim this quality. These films with one of the narrowest rocking curves with

FWHM values of 37 arcsec (values even under 30 arsec have been obtained) were grown with a modified EMCORE GS 3200 UTM reactor. It should be stated, however, that the X-ray data based on the symmetric diffraction peak are not a critical measure of sample quality necessarily. For a more complete analysis, one should also inspect the asymmetric peak, which is sensitive to edge dislocations. This reactor generally incorporates separate inlets for ammonia and the nitride precursor, all are normal to the substrate surface, which rotates at speeds over 1000 rpm, and a laminar flow cell to assure a uniform growth [91].

OMVPE reactors incorporating new concepts have been designed to grow layers at lower temperatures. (It should be mentioned that the motivation for lower temperature growth spawned from the perceived need to minimize the loss of nitrogen from the surface.

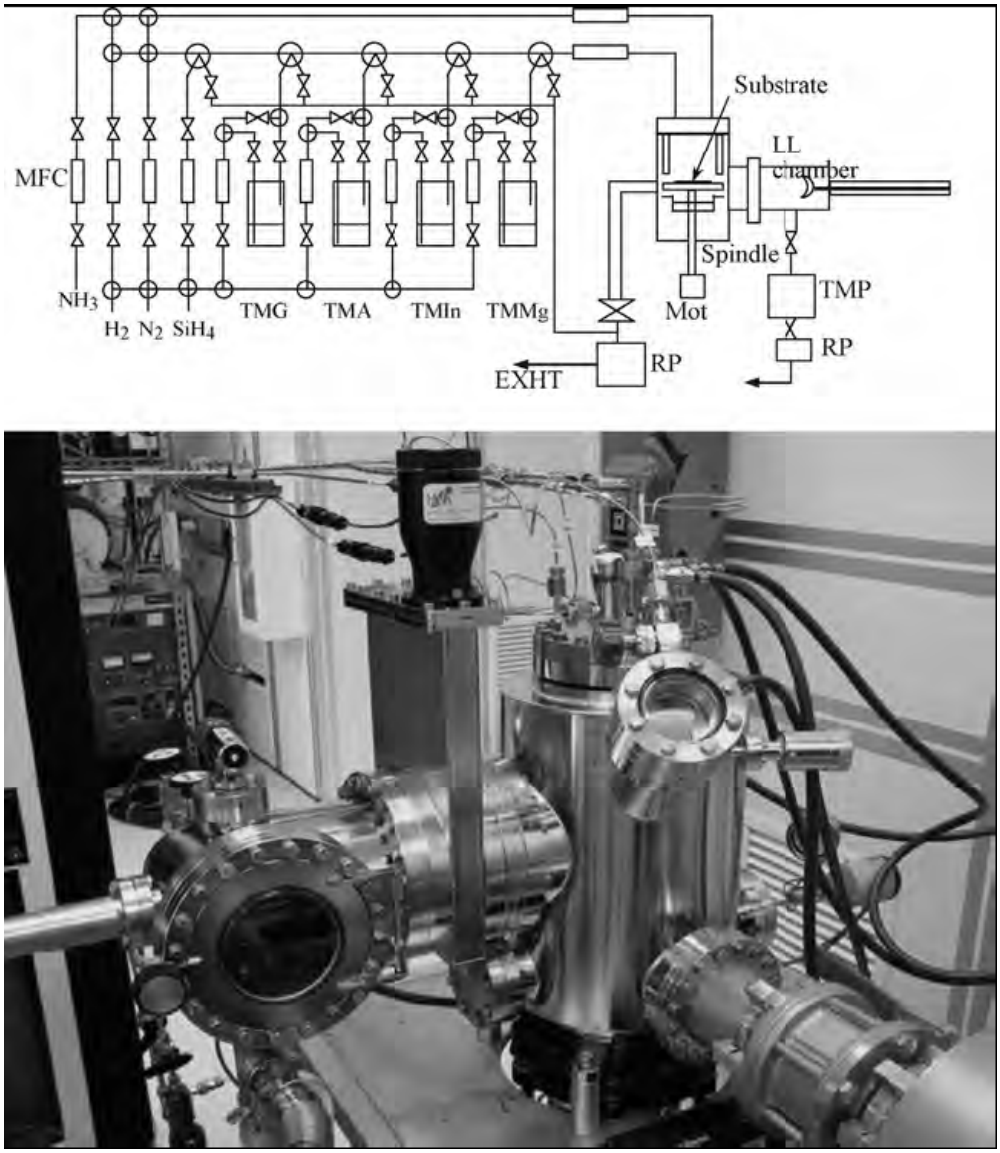


Fig. 3.12. A schematic representation of a vertical OMVPE system employed at Virginia Commonwealth University along with a picture of the deposition chamber (*a*); a photograph of the reactor chamber of the same (*b*).

However, later it became evident that high temperatures are needed to mobilize treading dislocations, as they are useful for reducing dislocation density and facilitating lateral growth. These technologies utilize an activated form of nitrogen to lower deposition temperatures of group III nitrides. That these technologies are interesting is apparent, for example, from the deposition of polycrystalline and amorphous GaN films at temperatures lower than 350 °C by plasma-enhanced CVD. Epitaxial GaN and AlN have been grown by variants of methods activating nitrogen, such as laser-assisted CVD, remote plasma enhanced CVD, atomic layer epitaxy with NH_3 cracked by a hot filament, with ammonia catalytically decomposed, photoassisted CVD, and ECR plasma-assisted CVD. However, none of these approaches has been able to produce material comparable in quality with the standard OMVPE systems and, consequently, they did not really become players in the field.

As for the mechanism involved, growth of nitride semiconductors by OMVPE relies on the transport of organometallic precursor gases, hydrides for the nitrogen source, and reacting them on or near the surface of a heated substrate. The deposition is through pyrolysis. The underlying chemical mechanisms are complex and involve a set of gas phase and surface reactions. Although OMVPE has long been assumed to be a thermodynamically equilibrium process, nitride OMVPE processes may involve kinetics as well. The fundamental understanding of the processes involved is still evolving and, as such, the reaction mechanism and the related kinetic rate parameters are poorly understood.

The deposition of epitaxial nitride layers by OMVPE involves the reaction of metalcontaining In, Ga, or Al gases with ammonia, NH_3. Commonly, the metal-containing gases are trimethylgallium ($(CH_3)_3Ga$), trimethylindium ($(CH_3)_3In$), or trimethylaluminum ($(CH_3)_3Al$). Radicals, reactive by most definitions, react in the gas phase with donors containing acidic hydrogen, such as NH_3, and form adducts. The key here is to eliminate the unwanted radicals by forming stable molecules followed by their removal from the reaction region.

The analysis of the mechanisms involved in the OMVPE process clearly indicates that any precursor must balance the requirements of volatility and stability, which often counter each other, to be transported to the surface and decomposed for deposition. To put it another way, these precursors must have appropriate reactivity to decompose thermally into the desired solid and to generate readily removable gaseous side products. Ideally, the precursors should be nonpyrophoric, water and oxygen insensitive, noncorrosive, and nontoxic. The trialkyls, trimethylgallium (TMG) [92-94] and triethylgallium (TEG), trimethylaluminum (TMA) [95], trimethylindium (TMI) [96], and others are usually used as III metal precursors. Ammonia (NH_3) [97], hydrazine (N_2H_4) [98-100], monomethylhydrazine $(CH_3)N_2H_2$ [101, 102], and dimethylhydrazine $(CH_3)_2N_2H_2$ have all

been used as nitrogen precursors with varying degrees of success. Although trialkyl compounds (TMA, TMG, TMI, etc.) are pyrophoric and extremely water and oxygen sensitive, and ammonia is highly corrosive, much of the best material grown today is produced by conventional OMVPE by reacting these compounds with NH_3 at substrate temperatures close to 1000 °C [103–114].

Investigators have reacted TMG [115, 116], TEG, and GaCl [117, 118] with NH_3 plasma. It was also reacted TMA and NH_3 in the presence of hydrogen plasma [119] and was grown InN by reacting TMI with microwave-activated N_2 [120]. The nitrogen was used to transport metallic Ga to the reaction zone where it was reacted with active nitrogen [121]. These commonly employed precursors at least satisfy the criteria of sufficient volatility and appropriate reactivity. During the growth of nitrides by employing trialkyl precursors, adduct formation between ammonia and TMA and TMG is well documented. Usually mixing at room temperature, adduct formation between TMG or TEG and ammonia is complete in less than 0.2 s. The resulting adduct $Ga(CH_3)_3$-NH_3 has a vapor pressure of 0.92 Torr at room temperature, while the vapor pressure of $Ga(C_2H_5)_3$: NH_3 is much lower.

To counter early beliefs that stability of ammonia and required relatively high growth temperature the use of other more volatile nitrogen sources were explored. For example, NH_3 was replaced with N_2H_4 and observed that a significantly smaller amount of N_2H_4 was required to maintain the same growth rate [100]. However, they also noted that the CVD growth rate was limited by the decomposition of TMG, thus limiting the benefits of N_2H_4. TMA and NH_3 was used to grow nitrides at a substrate temperature range of 673–1473 K [122]. Single-crystal AlN films were obtained only at 1473 K. The presence of a hot filament near the substrate increased the growth rate of AlN grown with TMA and NH_3 by two orders of magnitude [123]. However, the use of a hot filament immediately raises concerns about residual contamination, most prominently oxygen.

A case in point illustrating the elimination of unwanted radicals by forming stable molecules followed by their removal from the reaction region is AlN growth from mixtures of methyl alkyls that may proceed by the formation of an intermediate gas phase adduct (CH_3AlNH_3), followed by the elimination of CH_4. The exact path may be that coadsorption of $(CH_3)_3Al$ and NH_3 at room temperature generates surface adduct species such as ($(CH_3)_2AlNH_3$) and adsorbed NH_3 [124]. As the substrate temperature is raised above 320 °C, the appearance of vibrational bands corresponding to AlN indicates the formation of extended (Al–N) networks on the surface. These $Al(NH_2)_2Al$ species finally eliminate H_2 at the surface to form AlN [125]. The possible chemical reactions in the process are (*a* stands for adsorbate on the surface and *g* stands for gas-phase product):

$$2(CH_3)_2Al{:}NH_3(a) \rightarrow CH_3Al(NH_2)_2AlCH_3(a) + 2CH_4(g),$$
$$CH_3Al(NH_2)_2AlCH_3(a) \rightarrow Al(NH)_2Al+2CH_4(g), \qquad (3.14)$$

and

$$Al(NH)_2Al(a) \rightarrow 2AlN(a) + H_2(g).$$

As for GaN, investigations are relatively limited, but it would be fair to assume that processes similar to that with AlN growth are most likely in place. Adducts of Ga compounds are weaker electron acceptors than the corresponding Al adducts, and therefore these adducts may not be abundant owing to redissociation in the hot zone. It may be because of this that successful GaN growth by OMVPE requires very large V/III ratios, which favor adduct formation. Thermal stability of NH_3, although low compared to that of N_2, could be partially responsible for the use of high substrate temperatures, typically above 550 °C for InN and above 1000 °C for GaN and AlN. The high growth temperature necessitated by the process itself, associated with high nitrogen vapor pressure over GaN, lead to the inevitable nitrogen loss from the nitride film. This may also be the path to carbon contamination from the decomposition of the organic radical during metalorganic pyrolysis. The loss of nitrogen can be alleviated by using high V/III gas ratios during the deposition, particularly for InGaN (e.g., >2000 : 1).

Assuming that high substrate temperatures represent a problem in relation to ammonia, which seems reasonable particularly in early days, various alternative approaches can be and have been taken. One approach is to use alternative nitrogen precursors that are thermally less stable than NH_3. Hydrazine (N_2H_4), which is a larger and less stable molecule, has been used in combination with TMA to deposit AlN at temperatures as low as 220 °C [99]. However, hydrazine is toxic, unstable, and not as pure as NH_3. Consequently, a compromise between quality and substrate temperature must be made. Researchers took the quality/purity as the primary parameter and stayed with NH_3. More recently, other nitrogen sources such as tertbutylamine (t-$BuNH_2$) [126], isopropylamine (i-$PrNH_2$), and trimethylsilylazide ($TMeSiN_3$) have been used with TMA or t-Bu_3Al to deposit AlN films at lower substrate temperatures (400–600 °C) and reduced V/III gas ratios (5 : 1–70 : 1) [127]. However, the deposited films were invariably contaminated with high levels of residual carbon (up to 11 at.%).

Hydrogen and, to a lesser extent, nitrogen are predominantly used as the transport gas. They can influence the chemical reaction mechanism of Et_3Mor $(CH_3)_3M$ in the gas phase by changing the reaction temperature of the metalorganic compounds or the concentration of reaction products. Hydrogen at the surface of the growing film can influence the growth rate and the structural properties [128, 129]. To obtain a basic understanding of the role of hydrogen in GaN growth, the possible sources of hydrogen and the influence of hydrogen on the chemical reaction mechanism in the gas phase and at the surface are unique and important issues in the context of OMVPE.

Pyrolysis of highly concentrated NH_3 in the presence of H_2 as the carrier gas results in a high concentration of molecular and atomic hydrogen near the substrate surface. Because the

growth temperatures above 900 °C are employed, which are higher than the decomposition temperature of the metalorganic compounds and their hydrocarbon ligands, the conditions for the desired bond breaking between the metal atomand the methyl or ethyl groups of the precursors are in place. However, the same can also lead to pyrolysis of the hydrocarbons with incorporation of hydrogen and carbon into the films. In this vein, decomposition of $(CH_3)_3Ga$ (TMG) and Et_3Ga in hydrogen and nitrogen atmospheres using a quadrupole mass analyzer has been investigated [130]. The decomposition reaction of the metalorganic precursors was found to be strongly affected by the presence of molecular hydrogen. The decomposition of $(CH_3)_3Ga$ occurs at 400 and 500 °C in H_2 and N_2, respectively. Similarly, the decomposition of Et_3M occurs at 260 and 300 °C in H_2 and N_2, respectively. Clearly, molecular hydrogen reduces the reaction temperature. The reaction mechanisms involve hydrolysis for $(CH_3)_3Ga$ in H_2, homolytic fission for $(CH_3)_3Ga$ in N_2, and β-elimination for Et_3M in both H_2 and N_2. By changing the reaction temperature and the reaction mechanism of the metalorganic precursor, the partial pressure of hydrogen affects the deposition rate of GaN and therefore the structural properties of the resulting film, especially at low growth temperatures [131].

3.4. Molecular-beam Epitaxy

Molecular beam epitaxy (MBE) is considered to be the most promising future growth technique because it allows for precise thickness control, dopant control, and pattern drawing. Growth is performed under an ultra-high vacuum chamber with a low growth rate (0.1 to 10 μm/h). This permits one to accurately control the impinging atoms or molecules and thus the thickness of the film. The system has multiple sources that allow accurate stoichiometric growth. The straight-line beam impinging on the substrate also prevents collision, scattering, or diffusion during beam flight.

MBE is a versatile technique for epitaxial growth of semiconductor, metal, and insulator thin films [132-139]. It is an epitaxial process involving the reaction of one or more thermal beams of atoms or molecules with a crystalline surface under ultrahigh-vacuum conditions (~10^{-8} Pa) [140]. MBE is a highly refined form of vacuum deposition with precise control of the beam fluxes and deposition conditions and as a result can achieve precise control in both chemical compositions and doping profiles. Single-crystal multilayer structures with dimensions on the order of atomic layers can be made using MBE. Thus, the MBE method enables the precise fabrication of semiconductor heterostructures having thin layers from a fraction of a micron down to a monolayer.

Compounds or elements are evaporated from heated crucibles, called *Kniudsen cells*, onto clean ordered substrates. The deposition rate of the components and the temperature of the substrate must ee carefully chosen and controlled, and the substrate surface must be clean and as free of defects as possible. The need for extreme cleanliness means that the substrates and the Knudsen cells must be enclosed in an ultra-high-vacuum chamber, so the MBE process is generally carried out in large stainless steel high-vacuum systems [141]. However, provided that all the relevant parameters are carefully controlled it is possible to grow good quality single crystal materials of large area in this way (>50 cm^2). The technique thus offers the flexibility to prepare elemental semiconductors, compound semiconductors, semiconductor alloys, and heterostructures that involve alternating layers of different materials. The thickness of these layers can be accurately controlled, and the interfaces between them may be made remarkedly abrupt [142-144]. The substrate holder rotates continuously to achieve uniform epitaxial layers (e.g., ±1% in doping variations and ±0.5% in thickness variations).

MBE uses an evaporation method in the vacuum systems. An important parameter for vacuum technology is the molecular impingement rate, that is, how many molecules impinge on a unit area of the substrate per unit time. The impingement rate ϕ is a function of the molecular weight, temperature, and pressure. The rate can be expressed as [145]

$$\phi = P(2\pi m k_B T)^{-1/2} \tag{3.15}$$

or

$$\phi = 2.64\times 10^{20}(\frac{P}{\sqrt{MT}}) molecules/cm^2 s, \tag{3.16}$$

where P is the pressure in Pa, m is the mass of a molecule in kg , k_B is Boltzmams constant in J/K, T is the temperature in Kelvin, and M is the molecular weight. Therefore, at 300 K and $l0^{-4}$ Pa pressure, the impingement rate is 2.7×10^{14} molecules/cm^2-s for oxygen (M = 32).

There are two ways to clean a surface in situ for MBE. High-temperature baking can decompose native oxide and remove other adsorbed species by evaporation or diffusion into the wafer. Another approach is to use a low-energy ion beam of an inert gas to sputter-clean the surface, followed by a low-temperature annealing to reorder the surface lattice structure.

MBE can use a wide variety of dopants (compared with CVD and MOCVD), and the doping profile can be exactly controlled. However, the doping process is similar to the vapor-phase growth process: a flux of evaporated dopant atoms arrives at a favorable lattice site and is incorporated along the growing interface. Fine control of the doping profile is achieved by adjusting the dopant flux relative to the flux of silicon atoms (for silicon epitaxial films) or gallium atoms (for gallium arsenide epitaxial films). It is also possible to dope the epitaxial film using a low-current, low-energy ion beam to implant the dopant.

The substrate temperatures for MBE range from 400-900°C; and the growth rates range from 0.001 to 0.3 μm/min. Because of the low-temperature process and low-growth rate, many unique doping profiles and alloy compositions not obtainable from conventional CVD can be produced in MBE. Many novel structures have been made using MBE. These include the superlattice, which is a periodic structure consisting of alternating ultrathin layers with its period less than the electron mean free path (e.g., GaAs/$Al_xGa_{1-x}As$, with each layer 10 nm or less in thickness), and the heterojunction field-effect transistors.

The III-V compound is a new class of semiconductors for microwave devices, high-speed digital integrated circuits, and highly efficient optoelectronic devices. These compound semiconductors usually consist of the group-III elements, Ga, Al, and In, and the group-V elements, As, P and Sb. Several compounds such as GaAs [142, 146], GaP [147], $Al_xGa_{1-x}As$ [147], $GaAs_xSb_{1-x}$ [148], and $Ga_xIn_{1-x}As_yP_{1-y}$ [149] were first studied. The potential for excellent thickness control of MBE was first demonstrated by the growth of GaAs/$Al_xGa_{1-x}As$ periodic structures [150], In the process of evaluating device performance, it was found that the photoluminescent intensity increased more than an order of magnitude when the substrate temperature was increased from 540° to 650°C during growth [151]. Excellent results with double-heterostructure lasers [152-156], microwave field-effect transistors [157-162], pseudomorphic HEMT [136], heterojunction bipolar transistors, hot-electron transistors, and resonant-tunneling transistors, coupled with the high throughput and highly uniform growth with rotating sample holders [163, 164] made MBE an important thin film technology. A further development in MBE has replaced the group III elemental sources by metalorganic compounds such as trimethygallium (TMG) or triethylegallium (TEG). This approach is called metalorganic molecular-beam epitaxy (MOMBE) and is also referred to as chemical-beam epitaxy (CBE). Although closely related to MOCVD, it is considered a special form of MBE. The metalorganics are sufficiently volatile that they can be admitted directly into the MBE growth chamber as a beam and are not decomposed before forming the beam. The dopants are generally elemental sources, typically Be for *p*-type and Si or Sn for *n*-type GaAs epitaxial layers.

3.4.1. MBE Growth Systems and Deposition Sources

The rapid development of MBE system in a relatively short period has changed from custom-designed special ultra-high-vacuum (UHV) evaporators to dedicate high-throughput complete MBE instruments with proven ability to fabricate high-quality material.

System Configuration

Nowadays ultra-thin films and superlattice structure with arbitrary number of layers and layer thicknesses are grown by MBE. The defect density during crystal growth should be reduced to

a minimum. The study of monoatomic layer control in MBE becomes important. The multiple sources should be kept clean.

The art of UHV technology usually consists of two aspects. The first is about pumping, and the second is about materials used in a chamber and their outgassing characteristics. In an MBE chamber the vacuum condition is more determined by the outgassing characteristics of various contributing materials than by the pumping capability with which the chamber is equipped. An MBE chamber is also less forgiving in terms of background contaminants.

A typical commercial MBE system is shown in Fig. 3.13. It has three vacuum chambers: a growth chamber (left), an analysis chamber (center), and a small chamber for sample load-lock (right). The three chambers are vacuum-isolated from each other by either metal-sealed or viton-sealed gate valves. The function of the sample load- lock is to facilitate transfer of wafers in and out of the MBE system with minimum disturbance of the vaccum.

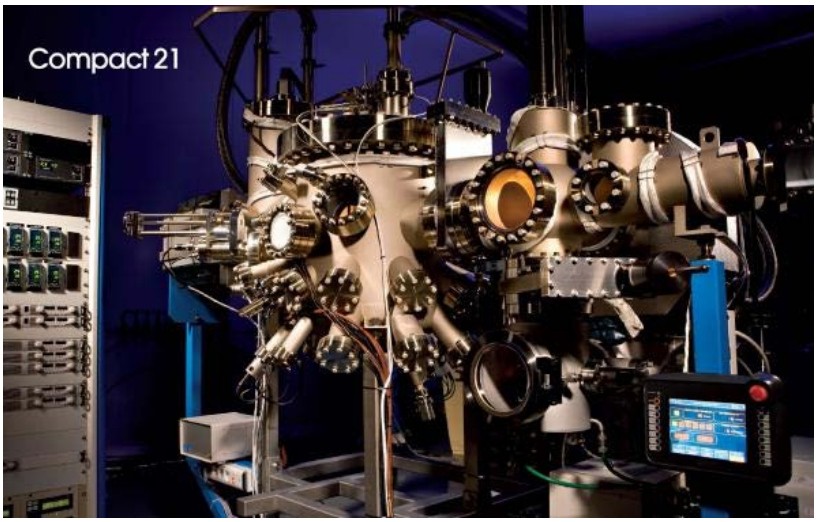


Fig. 3.13. Commercial gas source MBE machine.

The analysis chamber is where most of the surface-analysis instruments are housed. Auger electron spectrometer (AES) and a secondary ion mass spectrometer (SIMS) are used to analyze the sample.

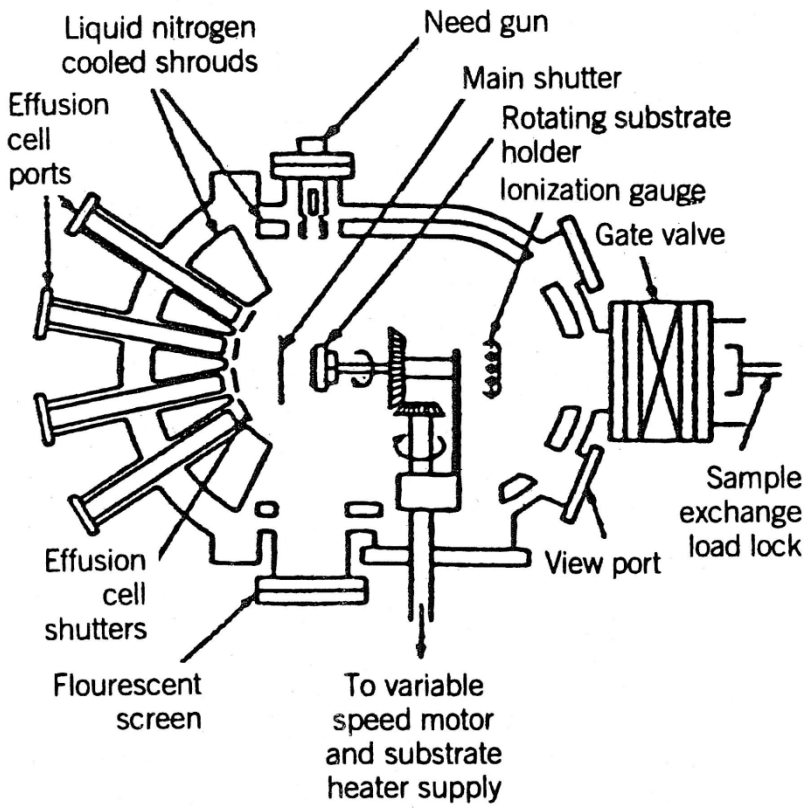


Fig. 3.14. Schematic of the MBE system viewed from the top. The rotating sample holder has a variable speed from 0.1 to 5 rpm [166].

The growth chamber is the most important component in an MBE system A growth chamber usually consists of source furnaces, shutters, a substrate manipulator, cryoshrouds, and some surface-analysis equipment. The cryoshrouds, high-quality source furnaces, and sample load-locks have been most instrumental to the development of MBE as a material preparative technique. A schematic of the MBE system viewed from the top is shown in Fig. 3.14. The sample exchange load-lock permits the maintenance of an ultra-high vacuum while changing substrates. The cryoshroud is used to enclose the entire growth area in order to minimize the residual water vapor and carbon-containing gases in the vacuum chamber during epitaxy. The modern commercial MBE growth chamber is often equipped with a rotary substrate manipulator capable of turning the substrate azimuthally during growth at a speed of about 3-5 rpm. With this feature the substrate can be heated more uniformly, resulting in epilayers of very good thickness and doping uniformity [163, 164]. To increase the throughput and yield of GaAs ICs, commercial MBE systems also have the capability of handling 4-in. diameter substrates, for indium-free mounting of substrates and for loading a large number of substrates (10 to 20) in a cassette [165].

The effusion cells are generally 2.5 cm in diameter and 7.5-12 cm in length and they are made of pyrolytic boron nitride (PBN). The PBN crucible is usually selected for work with reactive materials at high temperatures. These large-capacity crucibles are offered in two configurations: an "upward-looking" version and a "downward- looking" reverse-insert version. The upward-looking crucible kit is designed for furnaces mounted on the lower half of the MBE GeN II source flange. The downward-looking crucible kit is for furnaces mounted on the upper half of the source flange. The volume of the PBN reverse-insert crucible is 16 cc, and the volume of the standard PBN crucible is 40 cc. The growth of uniform epitaxial films from multiple effusion cells requires special effusion cell geometry and continuous rotation of the substrate around an axis normal to the substrate surface. The substrate holder can feature rotation speeds up to 125 rpm. The control unit remotely orients the sample holder into any of hour positions: growth, transfer, E-beam, and auxiliary. The controller also allows remote continual adjustment of rotation speed.

Deposition sources

The molecular beams are usually generated in Knudsen cells. These cells may be made of graphite or boron nitride. A water-cooled enclosure usually surrounds each Knudsen cell, and the cell arrangement is surrounded by surfaces cooled to liquid nitrogen temperatures, since it is most important to avoid cross contamination of the cells [141]. The mean free path of the atoms or molecules is much greater than the cell orifice, so the molecular rather than the hydrodynamic flow pattern of pressure is a concern.

In practice the conditions are rather far removed from those of the ideal Knudsen because large orifices are usually employed to obtain faster growth rates, as well as to improve the uniformity of films. The flux emanating from the Knudsen cells is controlled by accurate control of the temperature, and the flux arriving at the sample may be regulated by shutters in front of the Knudsen cells.

The flux arriving in the sample position is monitored by using an ion gauge attached to the reverse side of the sample holder. To calibrate the flux, the ion gauge is turned into the molecular beam by rotating the sample holder through 180°. Several Knudsen cells may be incorporated in the growth chamber in order to dope the semiconductor or to grow multicomponent compounds with alloys.

Sometimes additional cells called cracker cells are inserted between a Knudsen cell and the substrate. The effusion beams are directed from a conventional Knudsen crucible enclosure via a high-temperature (cracker) region onto the substrate. At an elevated temperature this region will provide a multiple collision path that will dissociate the molecular species emanating from the Knudsen cell. Thus the heating of solid arsenic and phosphorous in a Knudsen cell generates As_4

and P_4 tetramers. By allowing these molecules through the cracker cell at a temperature between 800° and 1000°C, it is possible to generate a beam of dimers, As_2 and P_2. The dimer sources offer the same advantages as the group-V MBE source and are likely to be accepted as the standard form of the source. From a practical point of view the sticking coefficient of As_2 has been shown to be twice that of As_4 [167] so only half the arsenic flux should be needed for each growth run. It is also found that use of the dimmer source reduces deep levels in the resulting gallium arsenide thin films [168].

In situ analysis. In the initial development of MBE, surface analysis performed during deposition played a major role in the understanding of the growth process. It has a reflective high-energy electron diffraction (RHEED) apparatus and an ion gauge in the growth chamber. A modern MBE system often includes other surface diagnostic facilities such as Auger electron spectroscopy (AES), secondary ion mass spectroscopy (SIMS), X-ray photoelectron spectroscopy (XPS), and scanning electron microscopy (SEM).

3.4.2. Growth of III-V Compounds

Epitaxial growth of III-V compounds by MBE involves a series of events: (1) adsorption of the constituent atoms and molecules, (2) surface migration and dissociation of the adsorbed molecules, and (3) incorporation of the atoms to the substrate resulting in nucleation and growth. An important understanding of the epitaxial growth process for GaAs was obtained by the kinetic studies of gallium and arsenic atoms on GaAs surfaces [169]. Arsenic has a very low sticking coefficient above 500°C unless it is combined with a gallium atom to form GaAs. Stoichiometric GaAs is formed as long as an excess of arsenic is supplied at the growing surface. The arsenic that does not form a bond to gallium will reevaporate from the surface. However, in the case of growing ternary and quaternary compounds such as $Ga_xIn_{1-x}As$ and $Al_xIn_{1-x}As_yP_{1-y}$, precise ratios of the beam fluxes are required for the compounds to grow with the desired mole fractions.

Processes of thin film formation

The chemical reactions involved in the molecular beam epitaxial growth processes are governed by thermodynamic considerations. However, the rate at which a system moves toward thermodynamic equilibrium is generally governed by kinetic considerations. Let’s consider growth from Ga + As_4 and Ga + As_2 and their reactions on GaAs (001) surfaces [170, 171]. To grow GaAs, an overpressure of As is maintained, since the sticking coefficient of Ga to GaAs is unity, whereas

that for As is zero, unless there is a previously deposited Ga layer. For a silicon MBE system, an electron gun is used to evaporate silicon. One or more effusion ovens are used for the dopants. Effusion ovens behave like small-area sources and exhibit a $cos\theta$ emission, where θ is the angle between the direction of the source and the normal to the substrate surface.

A beam of neutral atoms or molecules, having thermal velocities and with an intensity in the range $l0^{11}$-10^{16} atoms (or molecules) $cm^{-2}s^{-1}$ is directed at a substrate surface and the desorbing flux detected mass spectrometrically. The experiment is performed under UHV conditions ($<10^{-10}$ torr) to minimize surface impurities effects. Some provision may be made for structural and compositional analysis of the surface, most frequently by RHEED and AES.

Consider the interaction of arsenic and arsenic + gallium of a GaAs surface. Gallium is always monatomic, but the arsenic flux comprises either As_2 or As_4 molecules. The reaction processes of As_2 and As_4 on GaAs substrate are quite different.

The Reaction Process of As_4. By measuring the desorption rate of As_4 as a function of the incident As_4 flux at a fixed Ga flux, it was obtained the result shown in Fig. 3.15 [171]. At low fluxes (low surface concentrations) the desorption rate is second order with respect to the incident rate but becomes first order as the incident flux is increased. This does not imply a change of mechanism; it is the rate controlling step of the reaction that changes as the As_4 surface population increases.

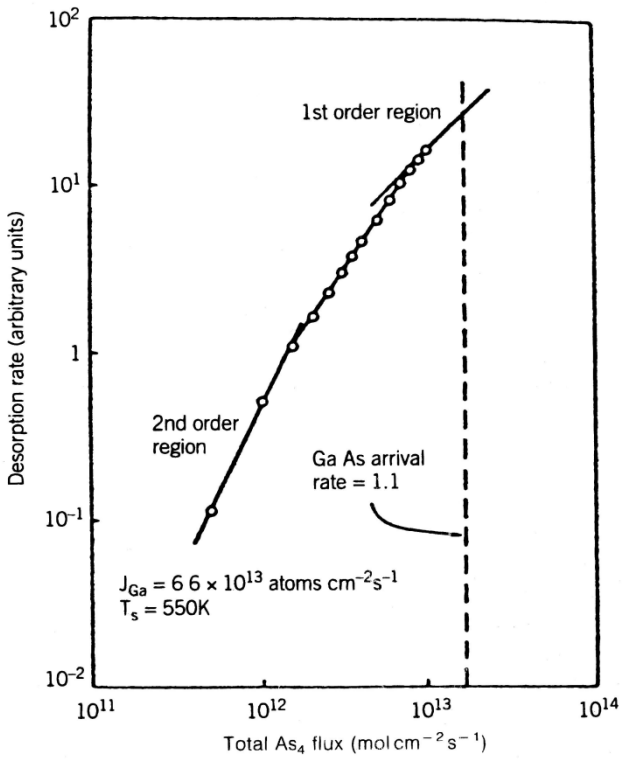


Fig. 3.15. Desorption rate of As_4 as a function of As_4 adsorption rate under Ga-rich surface conditions. The desorption rate is second order with respect to the incident flux at low fluxes, gradually becoming first order as the flux increases [171].

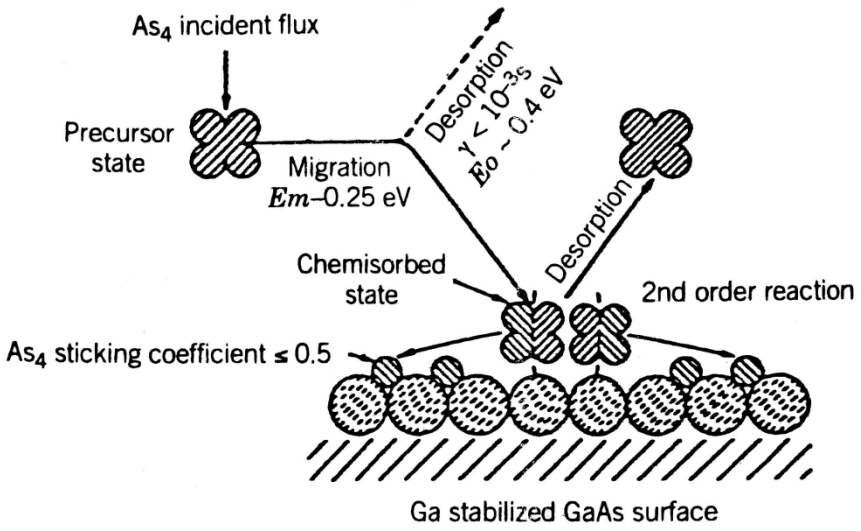


Fig. 3.16. Model of the growth of GaAs from molecular beams of Ga and As_4 [171].

The model for the growth mechanism of GaAs from Ga and As_4 beams that we have constructed from these results is shown in Fig. 3.16 [171]. The critical feature is the pairwise dissociation of As_4 molecules adsorbed on adjacent Ga atoms. From any two As_4 molecules four As atoms are incorporated in the GaAs lattice and the other four desorbed as an As_4 molecule. This is consistent with the observed relationship $S_{As4} = J_{As4} / 4J_{Ga}$ when $J_{Ga} << J_{As4}$ that is, when the surface population of Ga is large. The model also explains the second-order As_4 desorption rate at low relative incident rates, that is, low surface concentration of As_4 molecules, since under these conditions the desorption rate will be determined by probability of pairs of As_4 molecules being adsorbed on adjacent sites, which is simply proportional to the square of the adsorption rate. The change to first order at higher coverages reflects the change to the supply rate limited desorption rate.

The Reaction Process of As_2. It was shown that the sticking coefficient of As_2 (S_{As2}) increases linearly with the Ga adatom population of one monolayer during the raction pocess of As_2 As-Ga-GaAs interactions [172, 173]. Above ~600K the additional surface processes become significant [171]. Below 600 K there is no measurable dissociation of GaAs, but some incident As_2 molecules may associate on the surface to form As_4 before desorbing [173], as shown in Fig. 3.17. The desorbing As_2 flux decreases monotonically with decreasing temperature, but the desorption rate of As_4 reaches a maximum at 450 K. The decrease at lower temperatures is an artefact arising from the use of a GaAs source to produce the As_2. It leads to the nondissociative adsorption of some As_4 molecules on the Ga atoms that arrive with the As_2 flux.

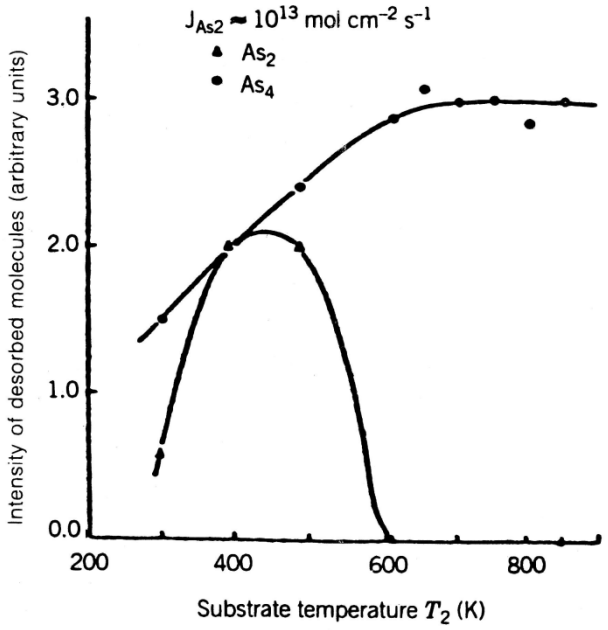


Fig. 3.17. Relative desorption rates of As_2 and As_4 for an incident As_2 flux as a function of temperature. Note the surface association of As_2 to form As_4 below 600°K [171].

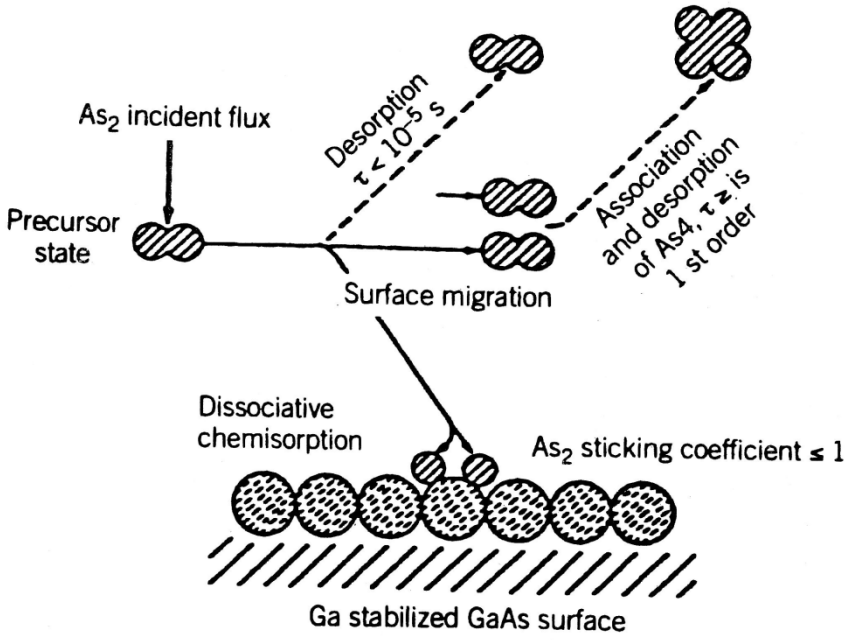


Fig. 3.18. Model of the growth of GaAs from molecular beams of Ga and As_2 [171].

The Ga-As_2 interactions on GaAs are summarized in the growth model shown in Fig. 3.18. According to this model As_2 molecules are first adsorbed into a mobile, weakly bound precursor state. Dissociation of adsorbed As_2 can only occur when the molecules encounter paired Ga lattice sites while migrating on the surface. In the absence of free surface Ga adatoms, As_2 has a measurable surface lifetime, but no permanent condensation occurs. Thus As has a very low sticking coefficient above 773 K unless it is combined with a Ga atom to form GaAs. Stoichiometric GaAs is formed as long as an excess As is supplied at the growing surface. At lower

substrate temperatures (<600 K), a pairwise association of adsorbed As molecules, followed by desorption as As_4 molecules, commences and becomes increasing dominant.

Effects of Asenic Species on Film Properties. The different growth mechanisms of GaAs films prepared from As_2 or As_4 might be expected to influence film properties. The crucial difference between the two reaction systems is that As_2 chemisorption involves a single Ga surface atom, while with As_4 there is an interaction involving adjacent pairs of Ga surface atoms. The steady state arsenic surface population should therefore be higher with As_2; with As_4 the maximum coverage will be 100%. Some proportion of single sites (10%) will always remain unoccupied. The nonoccupied surface sites leads to (As) vacancy introduction, which will influence the deep level concentrations. It has been conclusively demonstrated that the concentrations of three characteristics deep states (Ml, M3, and M4) in MBE-grown GaAs are substantially lower in films grown from As_2 than from As_4, all other conditions being identical [168, 174]. The states involved are all electron traps in *n*-type material, and they cannot unequivocally be related to native defects.

Let's now consider the influence of the arsenic species, and the associated surface chemistry, on the incorporation of an amphoteric dopant, germanium, in GaAs films. On the basis of these models the relatively higher arsenic surface population obtained with As_2 should favor the incorporation of Ge as a donor, and the degree of auto- compensation should consequently be less with As_2 than As_4-grown films. It was observed almost a decade difference in the ratio at a growth temperature of 820 K [168, 174]. If the system is grown under gallium-stabilized conditions, a stoichiometric GaAs composition cannot be maintained, and the surface morphology degrades rapidly.

Growth conditions

Nucleus Formation. For III-V compounds the passivated oxide layer serves as a protection for the freshly chemical-etched substrate from atmospheric contamination before epitaxial growth. After the MBE system is pumped down, the liquid nitrogen shroud cooled, and the effusion cells brought up to the desired temperatures, one begins to heat the substrate. In the case of GaAs, the oxide on the substrate desorbed between 580° and 600°C, and for InP, the oxide desorbed at about 520°C [175]. At this point the substrate is almost atomically clean and ready for epitaxial growth. Assuming that the substrate is properly prepared and atomically clean, the epitaxial layer will be mirror shiny if the group-V to group-III ratio in the molecular beam is above a certain value, giving an As-stabilized surface structure [176].

This value is also a function of the substrate temperature. The approximate relationship, also referred as the "MBE phase diagram," is shown in Fig. 3.19. In commercial MBE systems a GaAs growth rate of l0 μm/h may be achieved with a substrate temperature of 620°C.

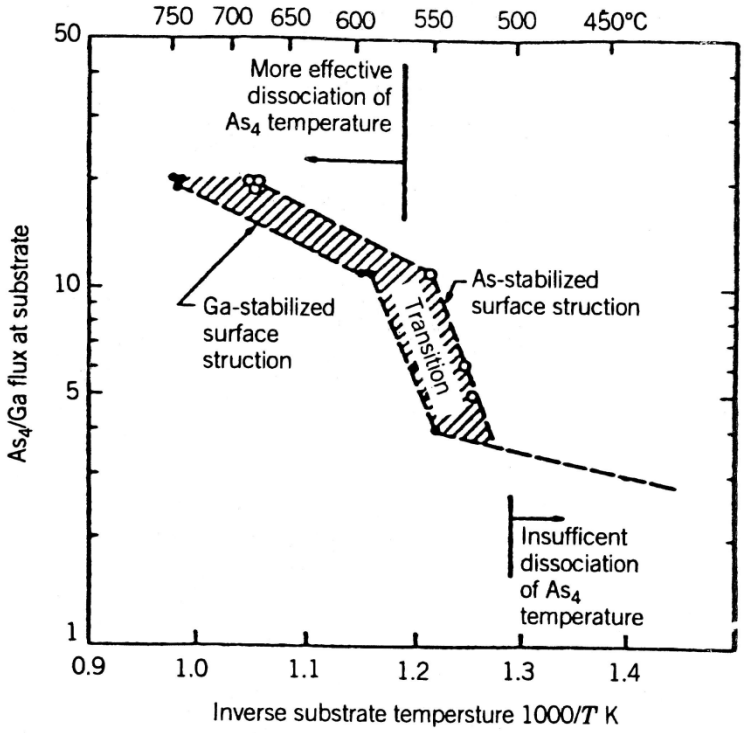


Fig. 3.19. As_4/Ga molecular beam flux ratio as a function of substrate temperature when the transition between As-stabilized and Ga-stabilized structures occurs on the (001) GaAs surface. The beam flux was measured by an ion gauge at the substrate position with Ga flux equal to 8×10^{-7} torr, giving a growth rate of about 1 μm/h [166].

The construction of this phase diagram is made possible from the knowledge gained about surface atom structures by the use of RHEED [176, 177]. In the case of GaAs in the <100> and <111> directions, the crystal is formed with alternate layers of Ga and As atoms. The terms Ga-rich and As-rich surface structures are used to describe the growth conditions where the top layer is terminated with Ga or As, respectively [176, 177]. On the (100) surface the Ga-stabilized surface structure is C(8 x 2), and the As-stabilized surface structure is C(2 x 8). These results were later confirmed with mass spectrometry [178, 179] and Auger studies [180, 181]. There are many more surface structures reported on (100) [180] and (111) [177] surfaces. The ratio of As/Ga in the molecular beam to form an As-stabilized surface on a (100) surface is different from that on a (111) surface for a given substrate temperature; the latter requires a higher As/Ga ratio.

Higher growth temperature resulted in higher-quality epitaxial layers, and it was related to the efficiency of photoluminescence [182]. This luminescence result has been proved to be directly related to the performance of double-heterostructure (DH) lasers [182].

The upper limit of the growth temperature is controlled by the availability of group V over pressure, or the group-V arrival rate that prevents the noncongruent evaporation of the compound.

A higher growth temperature requires a larger As consumption in growing GaAs or $Al_xGa_{1-x}As$. Furthermore above 640°C the Ga adsorption lifetime becomes sufficiently short to affect the growth rate [183]. The growth rates of GaAs, AlAs, and the GaAs fraction in $Al_xGa_{1-x}As$ as a function of the substrate temperature are shown in Fig. 3.20. Below 640°C the growth rates are nearly independent of the substrate temperature, implying that the sticking coefficient of Ga is nearly unity. Note that only a small percent of A1 in the beam can significantly increase the Ga sticking coefficient at high temperatures [169].

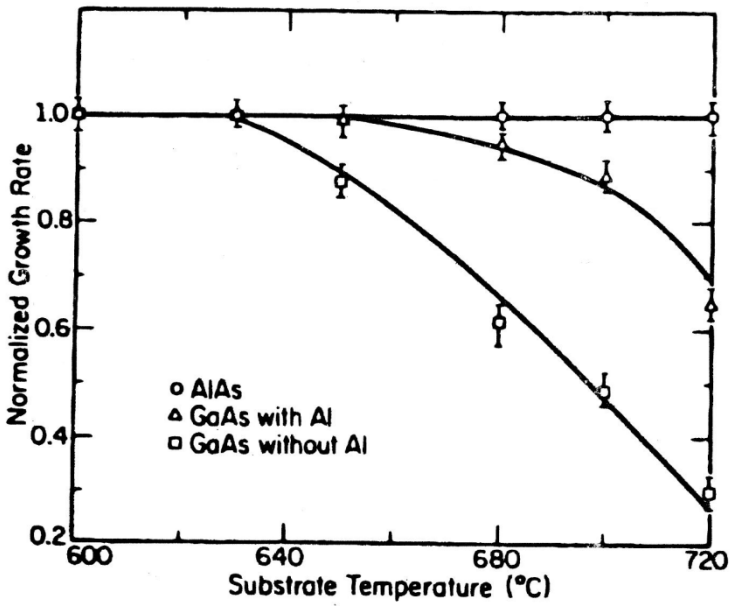


Fig. 3.20. Growth rates normalized to low temperature (≤620°C) values as a function of substrate temperature. The curve labeled "GaAs with Al" represents the Ga fraction of the growth rate [166].

3.4.3. MBE Growth of III-N Compounds

In this section, a succinct overall view of growth by MBE is given followed by an indepth discussion of physical processes that take place in MBE growth environments. In MBE technique, thin films are formed in vacuum on a heated substrate through various reactions between thermal molecular beams of the constituent elements and the surface species on the substrate [184]. The composition of the epilayer and its doping level depend on the arrival rates of the constituent elements and dopants [185]. Therefore, MBE growth is carried out under conditions governed primarily by the kinetics, rather than by mass transfer [186]. A thorough understanding of the growth kinetics, especially the surface processes of growth, is therefore critical. MBE is an

extremely versatile technique for preparing thin semiconductor heterostructures owing to the control over the growth parameters that it offers, and the inherent in-situ monitoring capability. As mentioned earlier, thin films are formed on a heated substrate through various reactions between thermal molecular beams (atomic beams in the case of RF-activated nitrogen) of the constituent elements participated by the surface species on the substrate originating from the substrate itself by surface or bulk contamination. The composition of the epilayer and its doping level depend on the arrival rates of the constituent elements and dopants, respectively. The typical growth rate of 1 μmh^{-1}, or slightly more than one monolayer per second (ML s^{-1}), is sufficiently low to allow for surface migration of the impinging species on the surface. In the case of growth along the <111> for cubic and *c*-directions for wurtzitic systems, one monolayer constitutes a bilayer. As MBE growth occurs under conditions that are governed primarily by the kinetics, rather than by mass transfer, this allows the preparation of many different structures that are otherwise not possible to attain. This is in contrast to OMVPE of conventional compound semiconductors, such as GaAs. However, in the case of GaN, even the OMVPE has elements of kinetics [187]. In nitride growth by MBE, the metal species are provided by Ga, In, and Al metal sources, the dopants are provided by pure Si for *n*-type and Mg for *p*-type using conventional Knudsen effusion cells that are heated to sufficient temperatures for the desired growth rate, composition, and doping levels. On the contrary, nitrogen gas is one of the least reactive gases because of its large molecular cohesive energy (946.04 kJ mol = 9.8 eV per N_2 molecule). Because of the triple bond between the two nitrogen atoms, dissociation of one molecule into two reactive nitrogen atoms requires a large amount of energy, which cannot be provided by thermal means.

In a plasma environment, however, and at reduced pressures, a significant dissociation of the nitrogen molecules takes place. Atomic nitrogen is reactive even at room temperature and bonds with many metals. Consequently, group III nitrides can be grown by plasma-assisted molecular beam epitaxy, where the plasma-induced fragmentation of nitrogen molecules is combined with the evaporation of metal atoms from effusion cells. In this vein, MBE growth of GaN has been reported by electron cyclotron resonance microwave plasma assisted molecular beam epitaxy (ECR-MBE). Several laboratories in the past have attempted RMBE growth in which N_2 or NH_3 was decomposed on the substrate surface [188-192].

The sequence of processes taking place during growth by MBE is adsorption, desorption, surface diffusion, incorporation and decomposition. All of these processes are in effect in many ways compete with each other during growth by MBE. Adsorption can be summed as the atoms or molecules impinging on the substrate surface and sticking by overcoming an activation barrier. Desorption, on the contrary, is a process in which the species that are not incorporated into the crystal lattice leave the substrate surface by thermal vibration. During surface diffusion, imperative

for growth, the constituent atoms or molecules diffuse on the substrate surface to find the low-energy crystal sites for incorporation. During the incorporation phase, the constituent atoms or molecules enter the crystal lattice of the substrate or the epilayer already grown by attaching to a dangling bond, vacancy, step edge, and so on. Owing to high temperatures involved, albeit much lower than those employed in OMVPE, during decomposition, the atoms in the crystal lattice leave the surface by breaking the bond. These processes are schematically shown in Fig. 3.21. Let us discuss these five processes that govern growth by MBE.

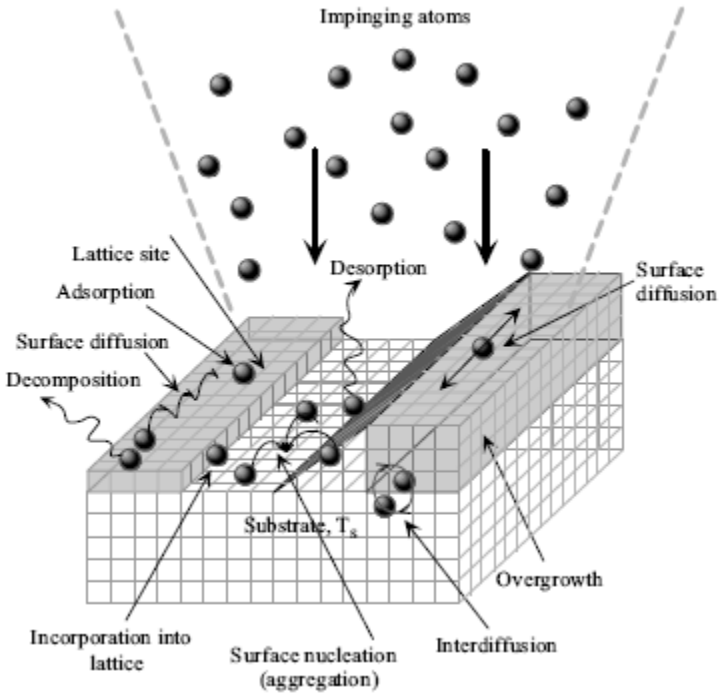


Fig. 3.21. Surface processes during the MBE growth: adsorption, desorption, surface diffusion, lattice incorporation, and decomposition.

Adsorption. Impinging gas molecules (atoms) condense on the surface and, depending on the strength of the interaction between the adsorbate and the surface, can be adsorped by physisorption (weak) or chemisorption (strong) [193]. Physisorption represents weak adsorbate–surface interaction because of van der Waals forces with typical binding energies on the order of 10–100 meV. Owing to the weak interaction, physisorbed atoms or molecules do not disturb the structural environment near the adsorption sites to a significant degree. In chemisorption, on the contrary, an adsorbate forms strong chemical bonds with the substrate atoms with typical binding energies on the order of 1–10 eV, thus changing the adsorbate.s chemical state.

The adsorbate coverage characterizes the surface concentration of adsorbed species expressed in monolayer units. The coverage is a relative value associated with a given substrate. It

can be converted to an absolute surface density of atoms. Considering the kinetic approach in the case of a uniform solid surface exposed to an adsorbing gas, the adsorption rate is defined as [193]

$$r_a = \sigma f(\Theta)\exp(-E_{act}/k_B T_s)\frac{p}{\sqrt{2\pi m k_B T_s}}, \qquad (3.17)$$

where p is the partial pressure of the adsorbing gas; σ is the condensation coefficient responsible for the effects of the orientation and the energy accommodation of the adsorbed molecules; $f(\Theta)$ is a coverage-dependent function that describes the probability of finding an adsorption site; $exp(-E_{act}/k_BT)$ is the temperature-dependent Boltzmann term associated with the energetics of the activated adsorption.

Finite equilibrium Ga adlayer coverage has been reported for typical substrate temperatures and Ga fluxes [194]. For large Ga fluxes, up to 2.5±0.2 monolayers of Ga are adsorbed on the GaN surface. For higher Ga fluxes, Ga droplets are formed [195]. At typical growth temperatures used in MBE, Ga adlayer does not condense into a reconstruction but behaves like a liquid film [196]. The reported height of the Ga adlayer is about 0.388 nm, as measured by scanning tunneling microscopy (STM), which corresponds to 1.9 ML [197]. For lower Ga fluxes, a discontinuous transition to Ga monolayer equilibrium coverage is found, followed by a continuous decrease toward zero coverage.

Desorption. Desorption is a process in which the adsorbate species gain sufficient thermal energy to escape from the adsorption well and leave the surface. The probability of desorption depends on the bonding strength of the particular atom to the surface. Bonding energies are different for different materials and the strength of a bond is expressed in terms of the amount of energy needed to break it. In the kinetic approach, desorption is described in terms of a desorption rate, r_{des}, which represents the number of species desorbed from unit surface area per unit time, and can be expressed as [193]

$$r_{des} = \sigma^* f^*(\Theta)\exp(-E_{des}/k_B T_s), \qquad (3.18)$$

where $f^*(\Theta)$ describes the coverage dependence and σ^* is the desorption coefficient standing for steric and surface mobility factors. For desorption to occur, the adsorbed species must overcome a barrier called the desorption energy, E_{des}. In case of activated chemisorption, the desorption energy is the sum of the binding energy in the chemisorbed state and the activation energy for adsorption, $E_{des}=E_{ads}+E_{act}$. In case of nonactivated chemisorption, the desorption energy is simply the binding energy in the chemisorbed state, $E_{des}=E_{ads}$.

The lifetime of an adsorbate as a function of temperature is needed for studying the desorption energy. The lifetime τ is defined as the average time spent by the adsorbate on the

surface marked from the time of adsorption to the time of desorption and obeys an Arrhenius dependence in the form [195]:

$$\tau = \tau_0 \exp(\frac{E_{des}}{k_B T_s}). \quad (3.19)$$

Equations 3.18 and 3.19 escribe the desorption process to be quite sensitive to temperature. For low T_s, the lifetime of adsorbates is sufficient to point that the desorption process can be neglected. For intermediate temperatures, the growth rate is determined through the competing processes of adsorption and desorption. However, for sufficiently high T_s, the desorption rate can be greater than the deposition rate, in which case evaporation rather than deposition would take place.

Specific to the case of GaN, Ga desorption has been investigated by mass spectrometry [195,198] and RHEED [199,200] with consistency lacking. The measured activation energies reported for Ga desorption are in the range of 0.4–5.1 eV [195–200]. For N, which is usually desorbed as a dimer (N_2), the desorption rate is limited by the surface diffusion of two N atoms. Some reports indicate this to be a first-order process [201], meaning that the surface diffusion may play less of a role in the desorption mechanism than the surface structure. Another plausible mechanism is that once an N atom diffuses near another N atom, the N_2 molecule would immediately form owing in part to the large N–N binding energy and desorbs from the surface because of the highly exothermic nature of the N_2 formation [202].

Surface Diffusion. As mentioned earlier, surface diffusion describes the motion of adsorbates (atoms or molecules) on the substrate (film) surface. The motion of adsorbates that become mobile because of thermal activation is described as a random walk. When a concentration gradient is present, this random walk motion of many particles results in their net diffusion being opposite to the gradient direction in a macroscopic sense. The microscopic view of surface diffusion is an activated process that is also affected by factors such as interaction between diffusing adsorbates, formation of surface phases, and defects. For an atom on the surface to diffuse to the next lattice position, it must overcome the lattice potential between the two neighboring positions. This activation energy required for diffusion, E_0, is the microscopic origin of the lattice potential.

The average length of diffusion, λ_s, in a unit time interval is a very important parameter to characterize the diffusion process and has an exponential temperature dependence given by Eq. 3.20, called the Einstein equation [203]:

$$\lambda_S = \sqrt{D_S \tau_S} = \sqrt{D_{S0} \exp(-E_{SD} / k_B T) \tau_S}, \quad (3.20)$$

where D_S is the diffusion coefficient, τ_S is the lifetime of the diffusion event, D_{S0} is the temperature-independent diffusion coefficient, E_{SD} is the diffusion activation energy, and k_B is the Boltzmann's constant.

At high growth temperatures where the surface diffusion length is larger than the terrace width, atoms move by surface diffusion from the terrace to the step edge to be incorporated into the growing lattice, which is called the step-flow growth. As the diffusion length decreases, for example, by reduced temperature, atoms meet and nucleate new islands on the terraces before reaching the edge of an existing island or a step edge. Moreover, islands can nucleate on top of the existing islands, a process that leads to rough surface formation and 3D growth. In the intermediate region, the islands do form on terraces, but the diffusion length is sufficiently long for adsorbates to diffuse to the edges of these islands and incorporate there, which leads to smooth surface also. In this mode of growth, the RHEED intensity oscillations are observed in that the intensity is maximum when the islands expand to cover the terrace and the intensity is minimum when the coverage is 50%.

Specific to the case of GaN, diffusivity for Ga and N adatoms on GaN surface is different. It was calculated the surface potential energy for Ga and N adatoms on GaN(0 0 0 1) surface [201]. The simulations led to the presence of two transition sites, as shown in Fig. 3.22. For Ga diffusion, the lower energy transition site is the bridge position (0.4 eV) and the higher energy site is the "on-top" position (>3 eV). However, for N adatoms, the barrier for bridge diffusion is 1.4 eV and the barrier for the on-top position is similar to Ga. Significantly lower diffusion barrier for Ga relative to N results in Ga being very mobile at typical growth temperatures, whereas the diffusion of N is slower by orders of magnitude. Further, the presence of excess N strongly increases the Ga diffusion barrier from 0.4 to 1.8 eV.

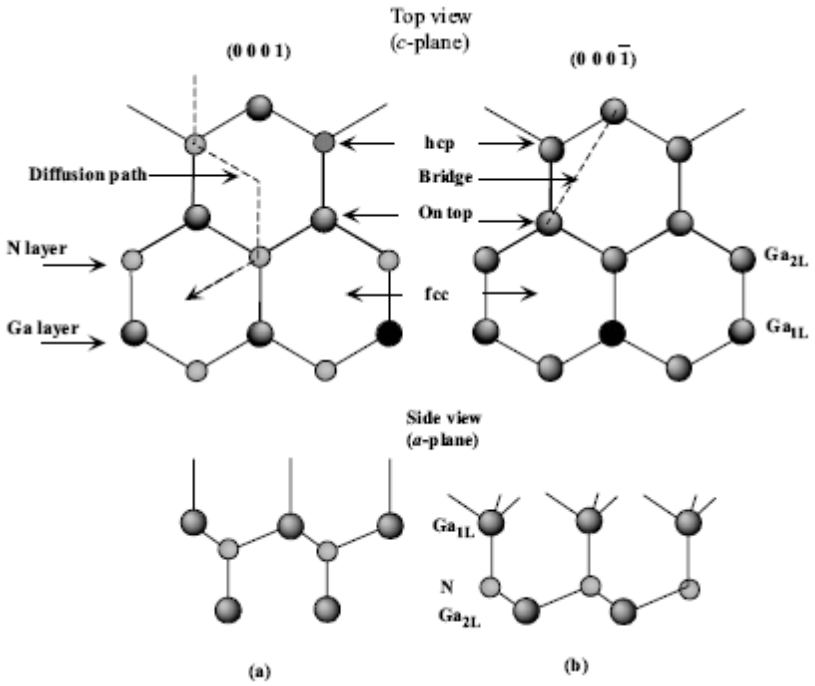


Fig. 3.22. Schematic diagram of diffusion paths on the GaN (0 0 0 1) (a) and (0 0 0 1^{-}) (b) surfaces. The side views as in the *a*-plane for each are also showed [201].

The very divergent surface mobilities of Ga and N adatoms have serious consequences. In the Ga-rich regime, the Ga adatoms are highly mobile and a step-flow mode results in 2D growth. Furthermore, if excess Ga adatoms are present on the surface, N adatoms can be efficiently incorporated because the probability of fast-moving Ga adatoms capturing N atoms is high. The presence of a Ga bilayer on the surface in Ga-rich growth conditions reduces the lateral diffusion barrier from about 1.3 to about 0.5 eV, paving the way for growth of smooth layers at low temperatures [203]. However, N-rich growth conditions show roughly a five times higher diffusion barrier for Ga. A rough surface can, therefore, be expected for N-rich conditions, consistent with experimental observations.

Incorporation. In the incorporation process, the molecules or atoms bond to the crystal through various reactions between the constituent elements and the surface species on the substrate. This process is controlled by the interplay of thermodynamics and kinetics. The general trends in film growth are understood within the thermodynamic approach in terms of the relative surface and interface energies. However, film growth by MBE is a nonequilibrium kinetic process in which rate-limiting steps affect the growth mode [186]. There are three modes of growth/incorporation that come to bear in general, namely, Frank–van der Merve (FM), Stranski–Krastanov (SK), and Volmer–Weber (VW) (Fig. 3.23). Suffice it to say, the Frank–van der Merve mode is a layer-by-layer process. Each layer is fully completed before the next layer starts to grow, and it is strictly a

two-dimensional growth mode. The Volmer–Weber mode is an island growth mode. Three-dimensional islands nucleate and grow directly on the substrate surface, which is typically seen in metals unless very low – even below room temperature – deposition conditions are employed. The Stranski–Krastanov mode is a layer-plus-island growth mode and represents the intermediate case between the FM and VW growth modes. After the formation of a complete two-dimensional layer of a few monolayer thickness (the exact value of which depends on the local strain), the growth of a three-dimensional layer (islands) takes place. The occurrences of growth have their genesis in the competition between the surface desorption and the surface diffusion. Because the desorption and diffusion processes are noticeably affected by the deposition rate, the surface condition and temperature, the growth mode, and the epilayer surface can be controlled by choosing a proper III/V ratio and a substrate temperature.

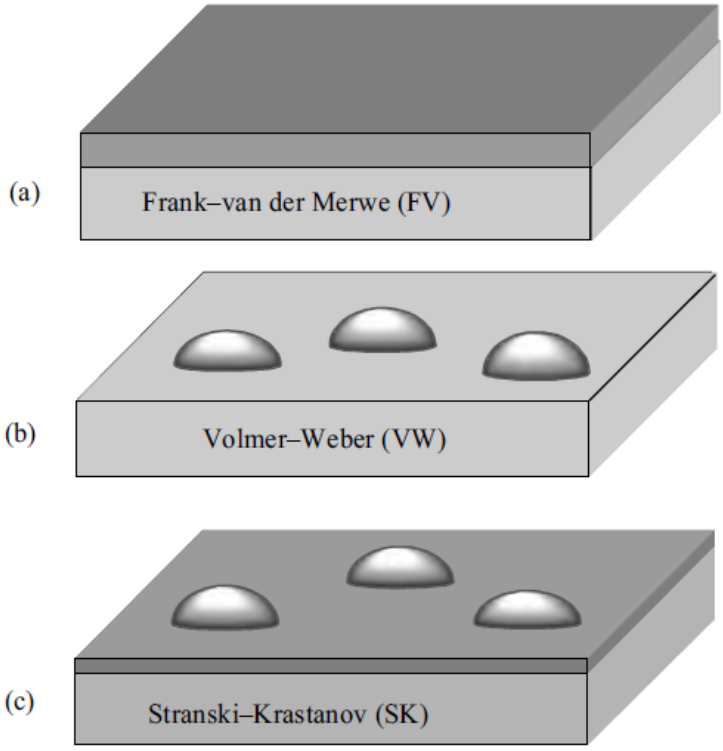


Fig. 3.23. Various growth modes occurring in general on exactly oriented templates. (a) FV mode, layer-by-layer 2D growth mode; (b) Volmer–Weber mode applicable mostly to metals and leads to 3D growth, (c) Stranski–Krastanov mode, which is driven by strain and is typical of semiconductors.

Several studies [204,205] have revealed that not all incident Ga atoms are incorporated into the growing GaN at the usual growth temperature in PAMBE (650–750 °C), even when an excess of N flux is present (III/V flux ratio <1). The incorporation ratio of Ga during GaN growth under N-rich conditions by monitoring the reflected Ga signals detected by the mass spectrometer and found that the Ga incorporation rate is sensitive to the growth temperature [195]. At low growth temperatures, the lifetime of Ga adatoms is sufficiently long so that almost all Ga atoms encounter

N adatoms and are incorporated into the crystal, resulting in a near unity incorporation ratio. As the growth temperature increases, the residence time of a Ga adatom on the surface decreases. Thus, the probability of Ga adatoms encountering N adatoms decreases and their incorporation ratio drops.

Decomposition. As mentioned previously, when the desorption rate is greater than the incorporation rate, the film would begin to decompose. In compound semiconductors such as GaN, one compound splits into two (or more pieces) as the bonds between the constituents is broken. The III nitrides have chemical bonds with the Ga–N bond strength estimated near 2.2 eV [206]. In addition, III nitride bonds have a sizeable ionic bond (31–40%) nature (excluding BN) than other III–V semiconductors (<8%) [207]. A lower ionic bond would have resulted in larger covalent bond with much larger bond strength. As in the case of other III–Vsemiconductors, the heat of formation of GaN is small and similar to that of InP and AlAs, indicating that GaN easily decomposes. In this vein, GaN does not melt congruently at pressures typically used in MBE growth, but decomposes above 800 °C at atmospheric pressure and at lower temperatures in vacuum. It was reported a decomposition rate of three to four monolayers per minute at 830 °C [208]. The decomposition rate is nearly zero below 750 °C, but increases rapidly above 800 °C, and reaches 1 μmh^{-1} at 850 °C [209]. This means that it may be impossible to grow GaN at high temperatures and that the growth temperature should be kept below ~800 °C in relation to growth in vacuum. Several different mechanisms have been proposed to explain GaN decomposition. These include decomposition into gaseous Ga and nitrogen, liquid Ga and nitrogen, and sublimation of GaN into a diatomic or polymeric product. One would surmise that these critical temperatures could be altered to some extent by the overpressure of N or N-containing reactive species used for growth, even though reports indicate that, to a large extent, GaN decomposition rate is similar in vacuum and at OMVPE pressures. However, a good deal of dispersion in the data allow us to assume that GaN is no different and that group V overpressure should have some affect on the rate of decomposition. In fact, GaN growth temperature with OMVPE is well in excess of 1050 °C, and if the vacuum experiments regarding decomposition rates were to hold, it would be nearly impossible to achieve deposition.

To make some sense of all these sometime competing processes, involving adsorption, desorption, surface diffusion, incorporation, and decomposition, it is compelling to summarize.

The conventional III–V semiconductors, and nitrides are no exception, are grown under conditions (substrate temperature and system pressure) in which the III–V compounds and the gaseous environment with which they are in contact are thermodynamically stable. For high-quality compound semiconductor growth, the optimum substrate temperatures used are in the range of ~1/2 to 2/3 of the melting temperature of the semiconductor. In the case of III nitrides, however,

synthesis occurs, particularly with MBE, at temperatures significantly below ~1/2 of the predicted melting temperature. Specific to GaN, the substrate temperature during MBE growth is about ~20 to 35% (600–900 °C) of the GaN melting temperature (T_M ~2500 °C). Thus, GaN growth takes place in kinetic regime and the surface processes are critical in determining the quality. The synthesis of GaN involves a metastable growth process, which is controlled by a competition between the forward reaction (incorporation of Ga and N into the film and GaN epilayer forms) and the reverse reaction (decomposition of GaN). The forward reaction depends on the arrival rates of Ga atoms and activated nitrogen species on the surface, as well as the substrate temperature, whereas the reverse reaction is strongly affected only by the substrate temperature. For a net growth to take place, the rate of GaN formation must be larger than the rate of decomposition.

The remaining three additional processes determining the eventual interface morphology of a growing film include deposition, desorption, and surface diffusion. Their relative importance depends on the microscopic properties of the interface, the bonding energies, and the diffusion barriers. The experimentally controllable parameters are the arrival rates of group III and V compounds and their ratios and substrate temperature. By tuning these parameters, a variety of morphologies can be achieved, ranging from layer-by-layer growth with an essentially smooth interface to a rough self-affined surface.

3.5. Structures and Defects in Epitaxial Layers

3.5.1. Lattice-matched and Strained-layer Epitaxy

For conventional homoepitaxial growth, a single-crystal semiconductor layer is grown on a single-crystal semiconductor substrate. The semiconductor layer and the substrate are the same material having the same lattice constant. Therefore, homoepitaxy is, by definition, a lattice-matched epitaxial process. The homoepitaxial process offers one important means of controlling the doping profiles so that device and circuit performance can be optimized. For example, an *n*-type silicon layer with a relatively low doping concentration can be grown epitaxially on an n^+-silicon substrate. This structure substantically reduces the series resistance associated with the substrate.

For heteroepitaxy, the epitaxial layer and the substrate are two different semiconductors, and the epitaxial layer must be grown in such a way that an idealized interfacial structure is maintained. This implies that atomic bonding across the interface must be continuous without interruption. Therefore, the two semiconductor must either have the same lattice spacing or be able to deform to adopt a common spacing. These two cases are referred to as lattice-matched epitaxy and strained-layer epitaxy.

Fig. 3.24 shows a lattice-matched epitaxy where the substrate and the film have the same lattice constant. An important example is the epitaxial growth of $Al_xGa_{1-x}As$ on a GaAs substrate where for any *x* between 0 and 1, the lattice constant of $Al_xGa_{1-x}As$ differs from that of GaAs by less than 0.13%.

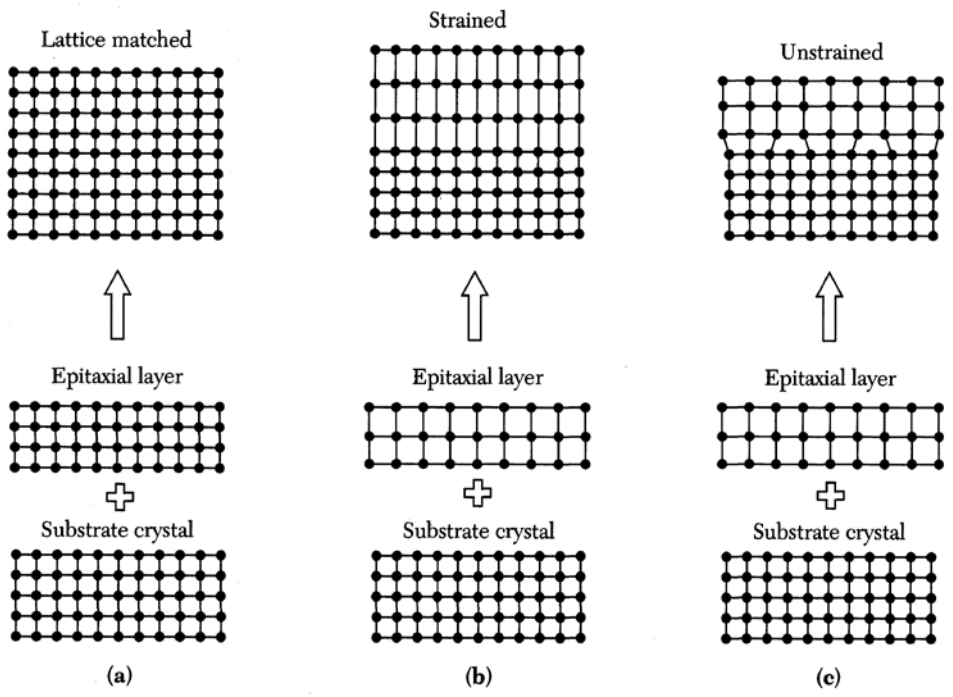


Fig. 3.24. Schematic illustration of (a) lattice-matched, (b) strained, and (c) relaxed heteroepitaxial structures [210]. Homoepitaxy is structurally identical to the lattice-matched heteroepitaxy.

For the lattice-mismatched case, if the epitaxial layer has a larger lattice constant and is flexible, it will be compressed in the plane of growth to conform to the substrate spacing. Elastic forces then compel it to dilate in a direction perpendicular to the interface. This type of structure is called strained-layer epitaxy and is illustrated in Fig. 3.24*b* [210]. On the other hand, if the epitaxial layer has a smaller lattice constant, it will be dilated in the plane of growth and compressed in a direction perpendicular to the interface. In the above strained-layer epitaxy, as the strained-layer thickness increases, the total number of atoms under strain or the distorted atomic bonds grows, and at some point misfit dislocations are nucleated to relieve the homogeneous strain energy. This thickness is referred to as the critical layer thickness for the system. Fig. 3.24*c* shows the case in which there are edge dislocations at the interface.

The critical layer thicknesses for two material systems are shown in Fig. 3.25 [211]. The upper curve is for the strained-layer epitaxy of a Ge_xSi_{1-x} layer on a silicon substrate, and the lower curve is for a $Ga_{1-x}In_xAs$ layer on a GaAs substrate, For example, for $Ge_{0.3}Si_{0.7}$ on silicon, the maximum epitaxial thickness is about 70 nm. For thicker films, edge dislocations will occur.

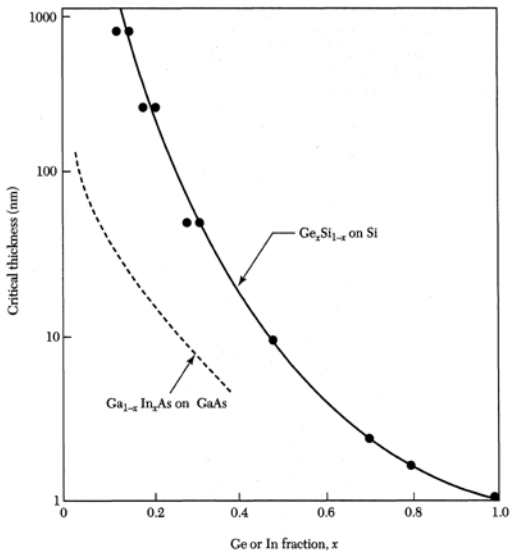


Fig. 3.25. Experimentally determined critical layer thickness for defect-free, strained-layer epitaxy of Ge_xSi_{1-x} on Si, and $Ga_{1-x}In_xAs$ on GaAs [211].

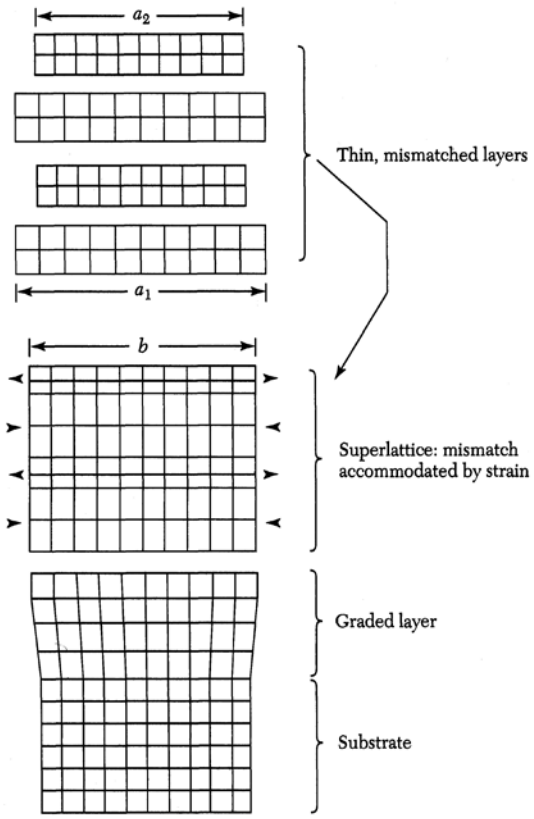


Fig. 3.26. Illustration of the elements and formation of an strained-layer superlattice. Arrows show the direction of the strain [140].

A related heteroepitaxial structure is the strained-layer superlattice (SLS). A superlattice is an artificial one-dimensional periodic structure constituted by different materials with a period of about 10 nm. Fig. 3.26 shows a SLS having two semiconductors with different equilibrium lattice constants $a_1 > a_2$ grown in a structure with a common inplane lattice constant b, where $a_1 > b > a_2$.

For sufficiently thin layers, the lattice mismatch is accommodated by uniform strains in the layers. Under these conditions, no misfit dislocations are generated at the interfaces, so high-quality crystalline materials can be obtained. These artificially structured materials can be grown by MBE. These materials provide a new area in semiconductor research and permit new solid-state devices, especially for high-speed and photonic applications,

3.8.2. Defects in Epitaxial Layers

Defects in epitaxial layers will degrade device properties. For example, defects can result in reduced mobility or increased leakage current. The defects in epitaxial layers can be categorized into five groups.

1. Defects from the substrates. These defects may propagate from the substrate into the epitaxial layer. To avoid these defects, dislocation-free semiconductor substrates are required.

2. Defects from the interface. The oxide precipitates or any contamination at the interface of the epitaxial layer and substrate may cause the formation of misoriented clusters or nuclei containing stacking faults. These clusters and stacking faults may coalesce with normal nuclei and grow into the film in the shape of an inverted pyramid. To avoid these defects, the surface of the substrate must be thoroughly cleaned. In addition, an in-situ etch back may be used such as the reversable reaction of Eq. 3.1.

3. Precipitates or dislocation loops. Their formation is due to supersaturation of impurities or dopants. Epitaxial layers containing very high intentional or unintentional dopants or impurity concentrations are susceptible to such defects.

4. Low-angle grain boundaries and twins. Any misoriented areas of an epitaxial film during growth may meet and coalesce to form these defects.

5. Edge dislocations. These are formed in the heteroepitaxy of two lattice-mismatched semiconductors. If both lattices are rigid, they will retain their fundamental lattice spacings, and the interface will contain rows of misbonded atoms described as misfit or edge dislocations. The edge dislocations can also form in a strained layer when the layer thickness becomes larger than the critical layer thickness.

Morphological defects. In the MBE growth system there are several surface-morphological problems related to growth conditions [208,212-219]. Among the reported defects on the MBE-grown layers are oval defects [208, 212-214], polycrystallites [208,212-213] whiskers [208,213-214], stacking faults [216], and dislocations [216]. Some of the observed defects are easily removed by preparing the substrate properly. Others like oval defects, however, are difficult to eliminate. These are oval- shaped hillocks, oriented along the [110] direction on a (001) substrate [220]. The

size is about 100 μm in length, 4 to 5 μm in width and 0.1 to 0.2 μm in height. At present the typical oval defect density is 500 to 2000 per square centimeter.

The morphologically defects in MBE-grown layers have studied in [221]. Fig. 3.27 deals with the oval defect density as a function of substrate temperature. The measured samples were consecutively grown under the same growth conditions. The average oval defect density measured by phase contrast microscopy was obtained from five regions of 100-μm diameter. Two sets of epilayers were investigated; one with a growth rate of 1.7 μm/h and the other with approximately 1 μm/h. The oval defect density decreases monotonically with increasing substrate temperature.

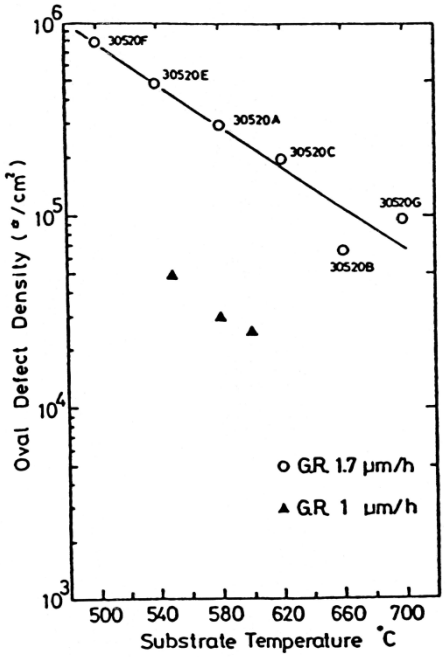


Fig. 3.27. Oval defect density as a function of substrate temperature. The growth conditions are identical except for the sample growth with higher As pressure to maintain a mirrorlike surface [221].

The origins of these oval defects are not fully understood. Several attempts were made to uncover their origins. Gallium oxide in the Ga flux is the major contaminant. It has been shown conclusively that the oval defect density can be reduced from 6×10^5 to 2×10^3 by careful elimination of gallium oxide [212]. The reaction of Ga and As fluxes can be described as [212]

$$4GaAs(s) + Ga_2O_3(s) \rightarrow 3Ga_2O(v) + As_4(v), \qquad (3.21)$$

$$4Ga(s) + Ga_2O_3(s) \rightarrow 3Ga_2O(v). \qquad (3.22)$$

Since the Ga_2O_3 is a nonvolatile oxide on the GaAs substrate, it becomes a nucleation center for the oval defect in the epilayers. Raising the substrate temperature will increase the maximum allowable

Ga_2O pressure, leading the reaction Eq. (3.21) at the right-hand side. Therefore the amount of Ga_2O_3 on the surface decreases, and the number of nucleation centers for the oval defects is reduced. The observed Ga droplets on the growth layers can be interpreted by Eq. (3.22). The increase of Ga on the surface arises from the gallium oxide. It has also been reported [222] that no oval defects were observed when growth took place under the influence a strong Mg flux, which is a getter of oxides.

Fig. 3.28 shows the relation between RHEED patterns as functions of substrate temperature and As pressure. The growth rate is 1 μm/h. The transition temperature for the change from the As-stabilized condition to the Ga-stabilized condition at the same As pressure is about 50°C in the study [221]. Fig. 3.29 illustrates the effect of growth rates on the oval defect density. The epilayers were grown at 580°C. Oval defect density decreases with decreasing growth rate. The surface morphology of the high growth rate is inferior to those of the low growth rate, even at the same thickness.

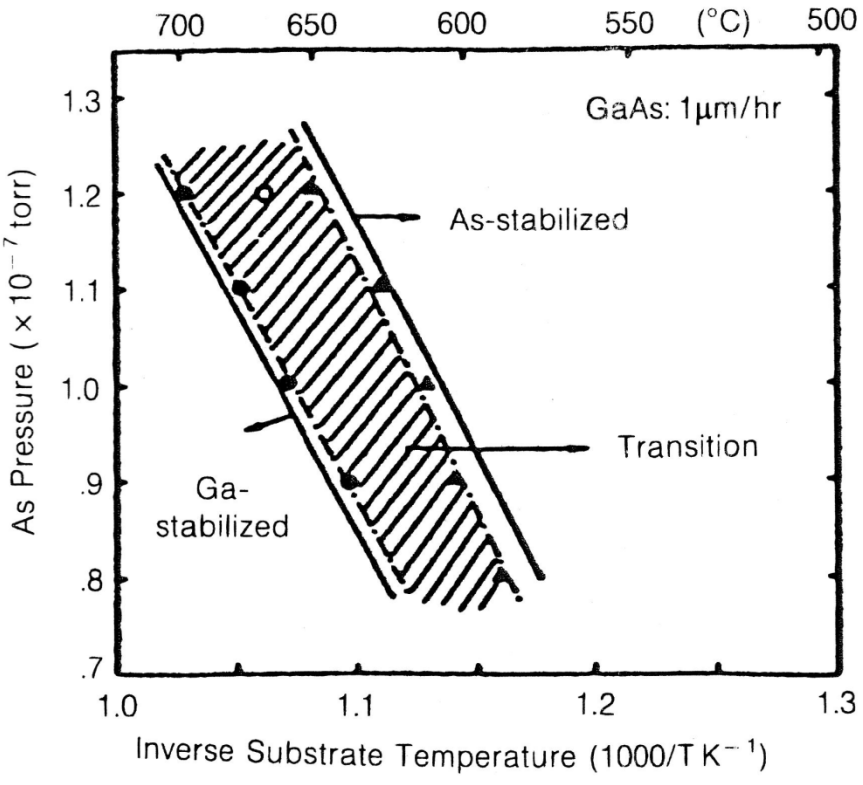


Fig. 3.28. RHEED pattern as a function of substrate temperature and As pressure [221].

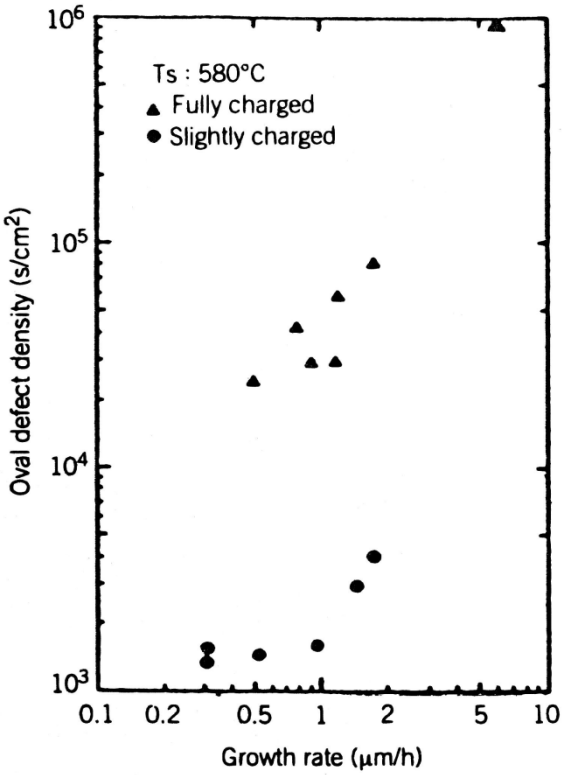


Fig. 3.29. Oval defect density as a function of the growth rate. Two sets of samples were investigated: one with almost fully charged Ga sources and the other with only a little Ga melt in the crucible [221].

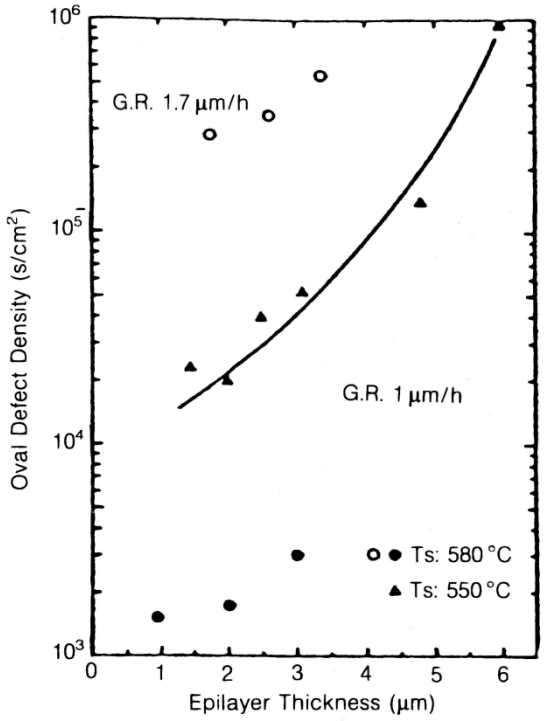


Fig. 3.30. Oval defect density as a function of epitaxal thickness. Two sets of samples at various growth rates of 1 and 1.7 μm/h were considered; • shows the samples grown at little Ga melt in the crucible [221].

Fig. 3.30 demonstrates the dependence of the oval defect density on the epilayer thickness. The oval defect density monotonically increases with increasing epilayer thickness. It means that not all of the oval defects are formed in the initial starting growth (i.e., epilayer-substrate interface), although most defects arise from here. The relation between the oval defect density and background As pressure, or As/Ga ratios, is shown in Fig. 3.31 [221]. The thickness of inspected undoped GaAs layers is approximately 2 μm. It indicates that the oval defect density slightly increases with increasing As/Ga ratios consistent with the trend of Eq. (3.20). As the As/Ga ratio increases, the nonvolatile oxide for nucleated sites is enhanced to form the oval defects. It is also found that very large As/Ga ratios show large whisker density which is believed to be vapor-liquid-solid (VLS) mechanism, on account of Ga droplets [213, 215].

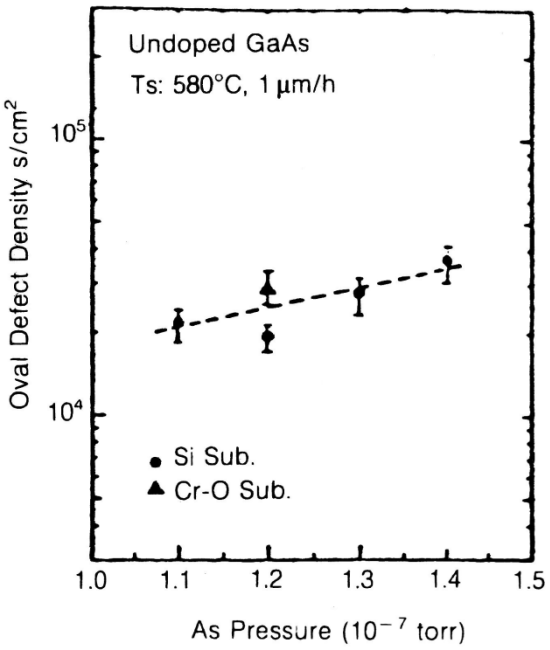


Fig. 3.31. Oval defect as a function of background As pressure. The growth rate was kept at 1 μm/h [221].

The conclusions on the observed oval defects are drawn as follows:

(*i*) The oval defect density increases with increasing growth rates and epitaxial thickness. Most of the observed oval defects are formed during growth. The low growth rate at 0.3 μm/h can reduce the oval defect density from 10^6 to 10^3 cm^{-2} or less.

(*ii*) The effects of substrate types, doping concentrations, and the compositions of etching solution are minor.

(*iii*) The oval defect density decreases with increasing substrate temperature and decreasing As pressure, as is the case with the model proposed in [223].

It was suggested [216] that oval defects have a higher resistivity. An increase in oval defect area inside the source-to-drain region will decrease the source-to-drain channel conductance. The breakdown voltage of the MESFET will degrade when the core of the oval defects is at or near the gate edge [219].

A higher growth temperature will increase the surface mobility of Ga atoms, and a lower growth rate [224] will increase the residence time for Ga atoms to find a favorite site. Both of these conditions will minimize Ga segregation. Lowering the Ga flux while maintaining the same growth rate [225], decreasing the distance between the substrate and the Ga cell, or increasing the Ga crucible diameter will minimize Ga spitting. Careful preparation and choice of the substrates, as well as proper outgassing of a fresh Ga source after the growth chamber is exposed to air, will also minimize the oval defect density from Ga_2O. Since MOCVD-grown high-quality selectively doped heterostructures have been reported without oval defects, MOMBE (CBE) may be a feasible approach to grow high-quality films.

3.6. Summary

A technology closely related to crystal growth is the epitaxial process. In this process, the substrate wafer is the seed. High-quality, single-crystal films can be grown at a temperature 30%-50% lower than the melting point. The common techniques for epitaxial growth are chemical-vapor deposition (CVD), metalorganic CVD (MOCVD), and molecular-beam epitaxy (MBE). CVD and MOCVD are chemical deposition processes. Gases and dopants are transported in vapor form to the substrate, where a chemical reaction occurs that results in the deposition of the epitaxial layer. Inorganic compounds are used for CVD, whereas metalorganic compounds are used for MOCVD. MBE, on the other hand, is a physical deposition process. It is done by the evaporation of a species in an ultrahigh vacuum system. Because it is a low-temperature process that has a low growth rate, MBE can grow single-crystal, multilayer structures with dimensions on the order of atomic layers.

In addition to conventional homoepitaxy, such as *n*-type silicon on an n^+-silicon substrate, the heteroepitaxy that includes lattice-matched and strained layer structures has also been considered. For strained-layer epitaxy, there is a critical layer thickness above which edge dislocations will nucleate to relieve the strain energy.

Besides the edge dislocations in an epitaxial layer, there are defects from the substrate, defects from the interface, precipitates, and low-angle grain boundaries and twins. These defects degrade device performance. Various means have been presented to minimize or even to eliminate these defects so that a defect-free semiconductor layer can be grown either homoepitaxially or heteroepitaxially.

Chapter 4. Dielectric and polycrystalline silicon deposition

A. Evtukh

4.1. Introduction

Deposited dielectric films are used mainly for insulation and passivation of discrete devices and integrated circuits. There are three commonly used deposition methods: atmospheric pressure CVD (chemical vapor deposition), low-pressure CVD (LPCVD), and plasma-enhanced chemical vapor deposition (PECVD, or plasma deposition). PECVD is an energy-enhanced CVD method, in which plasma energy is added to the thermal energy of a conventional CVD system. Considerations in selecting a deposition process are the substrate temperature, the deposition rate and film uniformity, the morphology, the electrical and mechanical properties, and the chemical composition of the dielectric films [1,2].

The reactor for atmospheric-pressure CVD is similar to the one shown in Fig. 4.1**,** except that different gases are used at the gas inlet. In a hot-wall, reduced-pressure reactor as shown in Fig. 4.2*a*, the quartz tube is heated by a three-zone furnace, and gas is introduced at one end and pumped out at the opposite end. The semiconductor wafers are held vertically in a slotted quartz boats [3]. The quartz tube wall is hot because it is adjacent to the furnace, in contrast to a cold-wall reactor such as the horizontal epitaxial reactor that uses radiofrequency (rf) heating.

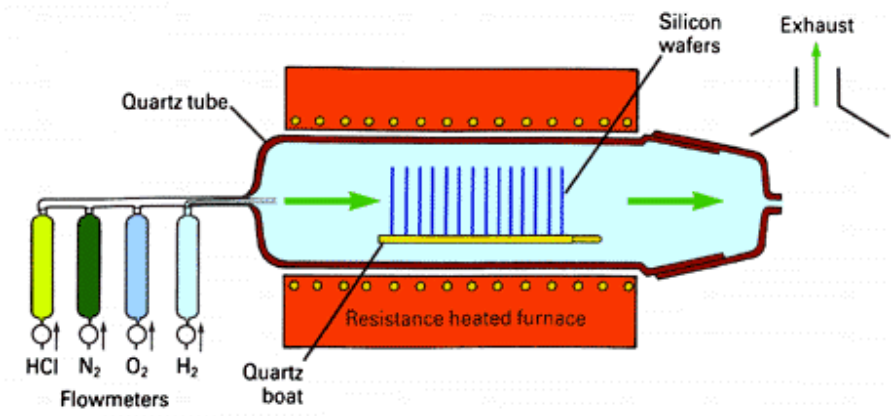


Fig. 4.1. Schematic cross section of a resistance-heated oxidation furnace.

The parallel-plate, radial-flow, PECVD reactor shown in Fig. 4.2*b* consists of a cylindrical glass or aluminum chamber sealed with aluminum endplates. Inside are two parallel aluminum electrodes. An rf voltage is applied to the upper electrode, whereas the lower electrode is grounded. The rf voltage causes a plasma discharge between the electrodes. Wafers are placed

on the lower electrode, which is heated between 100°C and 400°C by resistance heaters. The reaction gases flow through the discharge from outlets located along the circumference of the lower electrode. The main advantage of this reactor is its low deposition temperature. However, its capacity is limited, especially for large-diameter wafers, and the wafers may become contaminated if loosely adhering deposits fall on them.

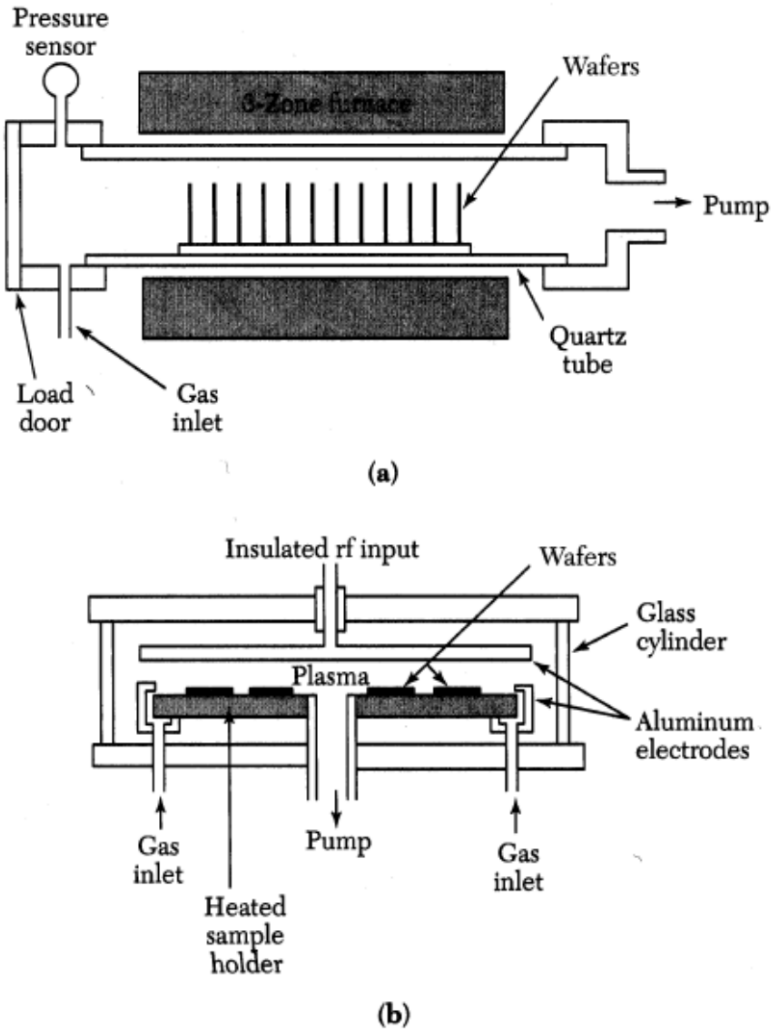


Fig. 4.2. Schematic diagrams of chemical-vapor deposition reactors. (*a*) Hot-wall, reduced-pressure reactor. (*b*) Parallel-plate plasma deposition reactor, (rf is radio frequency) [3].

4.2. Silicon Dioxide Deposition

CVD silicon dioxide does not replace thermally grown oxides because the best electrical properties are obtained with thermally grown films. CVD oxides are used instead to complement the thermal oxides. A layer of undoped silicon dioxide is used to insulate multilevel metallization,

to mask ion implantation and diffusion, and to increase the thickness of thermally grown field oxides. Phosphorus-doped silicon dioxide is used both as an insulator between metal layers and as a final passivation layer over devices. Oxides doped with phosphorus, arsenic, or boron are used occasionally as diffusion sources.

Deposition Methods

Silicon dioxide films can be deposited by several methods. For low-temperature deposition (300°C -500°C), the films are formed by reacting silane, dopant, and oxygen. The chemical reactions for phosphorus-doped oxides are

$$SiH_4 + O_2 \rightarrow^{(450°C)} SiO_2 + 2H_2, \quad (4.1)$$

$$4PH_3 + 5O_2 \rightarrow^{(450°C)} 2P_2O_5 + 6H_2. \quad (4.2)$$

The deposition process can be performed either at atmospheric pressure in a CVD reactor or at reduced pressure in an LPCVD reactor (Fig. 4.2*a*). The low deposition temperature of the silane-oxygen reaction makes it a suitable process when films must be deposited over a layer of aluminum.

For intermediate-temperature deposition (500°-800°C), silicon dioxide can be formed by decomposing tetraethylorthosilicate, $Si(OC_2H_5)_4$ in an LPCVD reactor. The compound, abbreviated TEOS, is vaporized from a liquid source. The TEOS compound decomposes as follows:

$$Si(OC_2H_5)_4 \rightarrow^{(700°C)} SiO_2 + \text{by-products}, \quad (4.3)$$

forming both SiO_2 and a mixture of organic and organosilicon by-products. Although the higher temperature required for the reaction prevents its use over aluminum, it is suitable for polysilicon gates requiring a uniform insulating layer with good step coverage. The good step coverage is a result of enhanced surface mobility at higher temperatures. The oxides can be doped by adding small amounts of the dopant hydrides (phosphines, arsine, or diborane), similar to the process in epitaxial growth.

The deposition rate as a function of temperature varies as $exp(-E_a/kT)$**,** here E_a is the activation energy. The E_a of the silane-oxygen reaction is quite low: about 0.6 eV for undoped oxides and almost zero for phosphorus doped oxide. In contrast, E_a for the TEOS reaction is much higher: about 1.9 eV for undoped oxide and 1.4 eV when phosphorus doping compounds are present. The dependence of the deposition rate on TEOS partial pressure is proportional to $[(1- exp(-P/P_0)]$, here P is the TEOS partial pressure and P_0 is about 30 Pa. At low TEOS partial pressures, the deposition rate is determined by the rate of the surface reaction. At high partial pressures, the surface becomes nearly saturated with adsorbed TEOS and the deposition rate becomes essentially independent of TEOS pressure [3].

Recently, atmospheric-pressure and low-temperature CVD processes using TEOS and ozone (O_3) have been proposed [4], as shown in Fig. 10. This CVD technology produces oxide films with high conformality and low viscosity under low deposition temperature. In addition, the shrinkage of oxide film during annealing is also a function of ozone concentration, as shown in Fig. 4.4. Because of their porosity, O_3-TEOS CVD oxides are often accompanied by plasma-assisted oxides to permit planarization in ULSI processing.

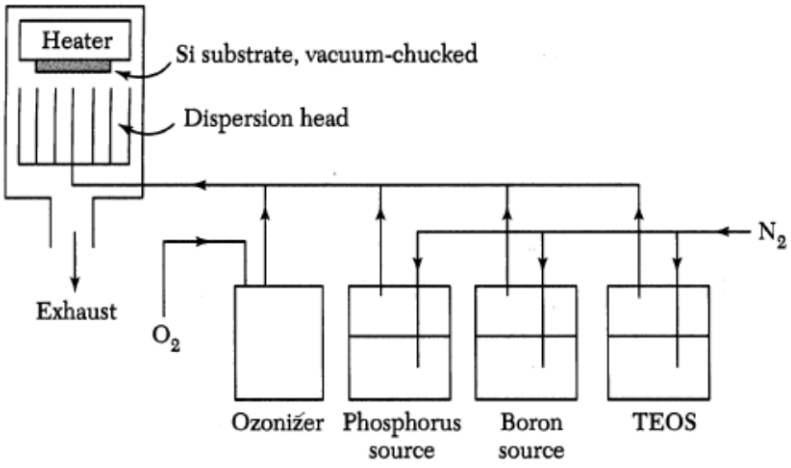


Fig. 4.3. Experimental apparatus for the O_3-TEOS chemical-vapor deposition (CVD) system.

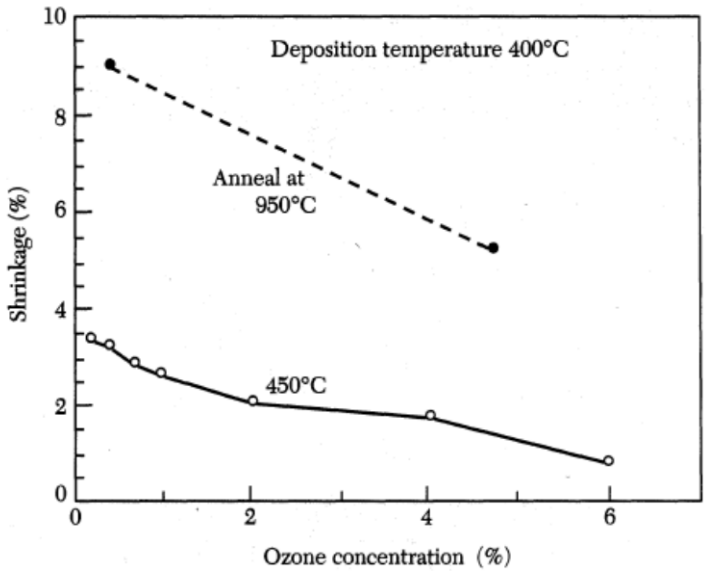


Fig. 4.4. Dependence of the shrinkage of the O3-TEOS CVD film on ozone concentration using annealing.

For high-temperature deposition (900°C), silicon dioxide is formed by reacting dichlorosilane, $SiC1_2H_2$, with nitrous oxide at reduced pressure:

$$SiC1_2H_2 + 2N_2O \rightarrow^{(900°C)} SiO_2 + 2N_2 + 2HC1. \quad (4.4)$$

This deposition gives excellent film uniformity and is sometimes used to deposit insulating layers over polysilicon.

Properties of Silicon Dioxide

Deposition methods and properties of silicon dioxide films are listed [3] in Table 4.1. In general, there is a direct correlation between deposition temperature and film quality. At higher temperatures, deposited oxide films are structurally similar to silicon dioxide that has been thermally grown.

The lower densities occur in films deposited below 500°C. Heating deposited silicon dioxide at temperatures between 600°C and 1000°C causes densification, during which the oxide thickness decreases, whereas the density increases to 2.2 g/cm^3. The refractive index of silicon dioxide is 1.46 at a wavelength of 0.6328 nm. Oxides with lower indices are porous, such as the oxide from the silane-oxygen deposition, which has a refract index of 1.44. The porous nature of the oxide also is responsible for the lower dielectric strength, which is the applied electric field that will cause a high current to flow in the oxide film. The etch rates of oxides in a hydrofluoric acid solution depend on deposition temperature, annealing history, and dopant concentration. Usually higher-quality oxides are etched at lower rates.

Table 4.1. Properties of SiO_2 films.

Property	Thermally grown at 1000°C	SiH_4+O_2 at 450°C	TEOS at 700°C	$SiCl_2H_2$ + N_2O at 900°C
Composition	SiO_2	SiO_2 (H)	SiO_2	SiO_2 (Cl)
Density (g/cm^3)	2.2	2.1	2.2	2.2
Refractive index	1.46	1.44	1.46	1.46
Dielectric strength (10^6 V/cm)	>10	8	10	10
Etch rate (nm/min) (100:1 H_2O:HF)	3	6	3	3
Etch rate (nm /min) (buffered HF)	44	120	45	45
Step coverage	-	Nonconformal	Conformal	Conformal

Step Coverage

Step coverage relates the surface topography of a deposited film to the various steps on the semiconductor substrate. In the illustration of ideal, or conformal, step coverage shown in Fig. 4.5*a*, film thickness is uniform along all surfaces of the step. The uniformity of the film thickness, regardless of topography, is due to the rapid migration of reactants after adsorption on the step surfaces [5].

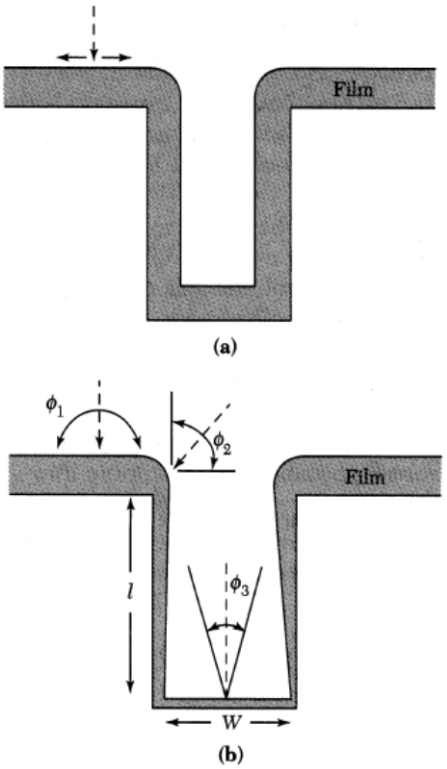


Fig. 4.5. Step coverage of deposited films. (*a*) Conformal step coverage. *(b)* Nonconformal step coverages [3].

Figure 4.5*b* shows an example of nonconformal step coverage, which results when the reactants adsorb and react without significant surface migration. In this instance, the deposition rate is proportional to the arrival angle of the gas molecules. Reactants arriving along the top horizontal surface come from many different angles and ϕ_1, the arrival angle, varies in two dimensions, from 0 to 180°, whereas reactants arriving at the top of a vertical wall have an arrival angle ϕ_2 that varies from 0° to 90°. Thus, the film thickness on the top surface is double that of a wall surface. Further down the wall, ϕ_3 is related to the width of the opening, and the film thickness is proportional to

$$\phi_3 \cong arctan(W/l), \tag{4.5}$$

where l is the distance from the top surface and W is the width of the opening. This type of step coverage is thin along the vertical walls, with a possible crack at the bottom of step caused by self-shadowing.

Silicon dioxide formed by TEOS decomposition at reduced pressure gives a nearly conformal coverage due to rapid surface migration. Similarly, the high-temperature dichlorosilane-nitrous oxide reaction also results in conformal coverage. However, during silane-oxygen deposition, no surface migration takes place and the step coverage is determined by the arrival angle. Most evaporated or sputtered materials have a step coverage similar to that in Fig. 4.5*b*.

P-Glass Flow

A smooth topography is usually required for the deposited silicon dioxide used as an insulator between metal layers. If the oxide used to cover the lower metal layer is concave, circuit failure may result from an opening that may occur in the upper metal layer during deposition. Because phosphorus-doped silicon dioxide (P-glass) deposited at low temperatures becomes soft and flows upon heating, it provides a smooth surface and is often used to insulate adjacent metal layers. This process is called P-glass flow.

Figure 4.6 shows four cross sections of scanning electron microscope photographs of P-glass covering a polysilicon step [5]. All samples are heated in steam at 1100°C for 20 min. Figure 4.6*a* shows a sample of glass that contains a negligibly small amount of phosphorus and does not flow. Note the concavity of the film and that the corresponding angle θ is about $120°$. Figures 4.6*b*, 4.6*c*, and 4.6*d* show samples of P-glass with progressively higher phosphorus contents up to 7.2 wt% (weight percent). In these samples the decreasing step angles of the P-glass layer indicate how flow increases with phosphorus concentration. P-glass flow depends on annealing time, temperature, phosphorus concentration, and the annealing ambient [5].

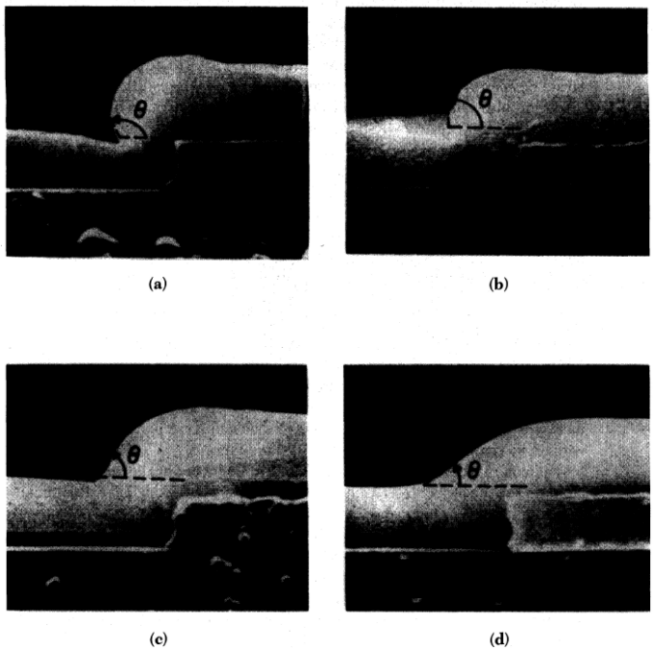


Fig. 4.6. Scanning-electron micrographs (10,000×) of samples annealed in steam at 1100°C for 20 minutes for the following weight percent of phosphorus [5]. (*a*) 0 wt%; (*b*) 2.2 wt%; (*c*) 4.6 wt%; and (*d*) 7.2 wt%.

The angle θ as a function of weight percent of phosphorus as shown in Fig. 4.6 can be approximated by

$$\theta \cong 120°((10\text{-wt\%})/10) \qquad (4.6)$$

If we want an angle smaller than 45° we require a phosphorus concentration larger than 6 wt%. However, at concentrations above 8 wt%, the metal film (e.g., aluminum) may be corroded by the acid products formed during the reaction between the phosphorus in the oxide and atmospheric moisture. Therefore, the P-glass flow process uses phosphorus concentrations of 6-8 wt%.

4.3. Silicon Nitride Deposition

It is difficult to grow silicon nitride by thermal nitridation (e.g., with ammonia, NH_3) because of its low growth rate and high growth temperature. However, silicon nitride films can be deposited by an intermediate-temperature (750°C) LPCVD process or a low-temperature (300°C) plasma-assisted CVD process [6,7]. The LPCVD films are of stoichiometric composition (Si_3N_4) with high density (2.9-3.1 g/cm^3). These films can be used to passivate devices because they serve as good barriers to the diffusion of water and sodium. The films also can be used as masks for the selective oxidation of silicon because silicon nitride oxidizes very

slowly and prevents the underlying silicon from oxidizing. The films deposited by plasma-assisted CVD are not stoichiometric and have a lower density (2.6-2.8 g/cm^3). Because of the low deposition temperature, silicon nitride films can be deposited over fabricated devices and serve as their final passivation. The plasma-deposited nitride provides excellent scratch protection, serves as a moisture barrier, and prevents sodium diffusion.

In the LPCVD process, dichlorosilane and ammonia react at reduced pressure to deposit silicon nitride at temperatures between 700° and 800°C. The reaction is

$$3SiCl_2H_2 + 4NH_3 \rightarrow^{750°C} Si_3N_4 + 6HC1 + 6H_2. \quad (4.7)$$

Good film uniformity and high wafer throughout (the number of wafers processed per hour) are advantages of the reduced-pressure process. As in the case of oxide deposition, silicon nitride deposition is controlled by temperature, pressure, and reactant concentration. The activation energy for deposition is about 1.8 eV. The deposition increases with increasing total pressure or dichlorosilane partial pressure and decreases with an increasing ammonia-to-dichlorosilane ratio.

Silicon nitride deposited by LPCVD is an amorphous dielectric containing atomic percent (at %) hydrogen. The etch rate in buffered HF is less than 1 nm/min. The film has a very high tensile stress of approximately 10^{10} dynes/cm^2, which is nearly 10 times that of TEOS-deposited SiO_2. Films thicker than 200 nm may crack because the very high stress. The resistivity of silicon nitride at room temperature is about 10^{16} Ω-cm. Its dielectric constant is 7 and its dielectric strength is 10^7 V/cm.

In the plasma-assisted CVD process, silicon nitride is formed either by reacting silane and ammonia in an argon plasma or by reacting silane in a nitrogen discharge. The reactions are as follows:

$$SiH_4 + NH_3 \rightarrow^{300°C} SiNH + 3H_2, \quad (4.8)$$

$$2SiH_4 + N_2 \rightarrow^{300°C} 2SiNH + 3H_2. \quad (4.9)$$

The products depend strongly on deposition conditions. The radial-flow, parallel-plate reactor (Fig. 4.2*b*) is used to deposit the films. The deposition rate generally increases with increasing temperature, power input, and reactant gas pressure.

Large concentrations of hydrogen are contained in plasma-deposited films. The plasma nitride (also referred to as SiN) used in semiconductor processing generally contains 20-25 at % hydrogen. Films with low tensile stress (~2×10^9 dynes/cm^2) can be prepared by plasma

deposition. Film resistivities range from 10^5 to 10^{21} Ω-cm, depending on silicon-to-nitrogen ratio, whereas dielectric strengths are between 1×10^6 and 6×10^6 V/cm.

4.4. Low-Dielectric-Constant Materials Deposition

As devices continue to scale down to the deep submicron region, they require multilevel interconnection architecture to minimize the time delay due to parasitic resistance (*R*) and capacitance (*C*). The gain in device speed at the gate level will be offset by the propagation delay at the metal interconnects because of the increased *RC* time constant, as shown in Fig. 4.7. For example, in devices with gate length of 250 nm or less, up to 50% of the time delay is due to the *RC* delay of long interconnection [8]. Therefore, the device interconnection network becomes a limiting factor in determining chip performance such as device speed, cross talk, and power consumption of ULSI circuits.

In order to reduce the *RC* time constant of ULSI circuits, interconnection materials with low resistivity and interlayer films with low capacitance are required. On the low capacitance topic ($C = \varepsilon_i A/d$, where ε_i, is the dielecric permittivity, A is the area, and d is the thickness of dielectric.

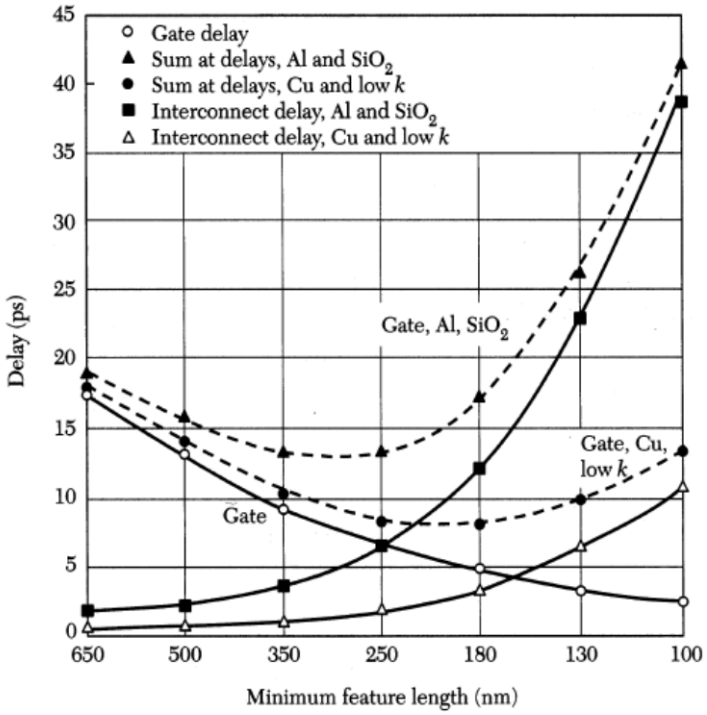


Fig. 4.7. Calculated gate and interconnect delay versus technology generation. The dielectric constant for the low-*k* material is 2.0. Both A1 and Cu interconnects are 0.8 μm thick and 43 μm long.

4.5. High-Dielectric-Constant Materials Deposition

High-*k* materials are required for ULSI circuits, especially for dynamic random access memory (DRAM). The storage capacitor in a DRAM has to maintain a certain value of capacitance for proper operation (e.g., 40 fF). For a given capacitance ($\varepsilon_i A/d$), usually a minimum d is selected to meet the conditions of the maximum allowed leakage current and the minimum required breakdown voltage. The area of the capacitor can be increased by using stacked or trench structures. However, for a planar structure, area A is reduced with increasing DRAM density. Therefore, the dielectric constant of the film must be increased.

Several high-*k* materials have been proposed, such as barium strontium titanate (BST) and lead zirconium titanate (PZT). They are shown in Table 4.2. In addition, there are titanates doped with one or more acceptors, such as alkaline earth metals, or doped with one or more donors, such as rare earth elements. The tantalum oxide (Ta_2O_5) has a dielectric constant in a range of 20-30. As a reference, the dielectric constant of Si_3N_4 is in a range of 6-7 and that for SiO_2 is 3.9. A Ta_2O_5 film can be deposited by a CVD process using gaseous $TaC1_5$ and O_5, as the starting materials.

Table 4.2. High-*k* materials

	Materials	Dielectric constant
Binary	Ta_2O_5	25
	TiO_2	40
	Y_2O_3	17
	Si_3N_4	7
Paraelectric perovskite	$SrTiO_3$ (STO)	140
	$(Ba_{1-x}Sr_x)TiO_3$ (BST)	300-500
	$Ba(Ti_{1-x}Zr_x)O_3$ (BZT)	300
	$(Pb_{1-x}La_x)(Zr_{1-y}Ti_y)O_3$ (PLZT)	800-1000
	$Pb(Mg_{1/3}Nb_{2/3})O_3$ (PMN)	1000-2000
Ferroelectric perovskite	$Pb(Zr_{0.47}Ti_{0.53})O_3$ (PZT)	>1000

4.6. Polysilicon Deposition

Using polysilicon as the gate electrode in MOS devices is a significant development in MOS technology. One important reason is that polysilicon surpasses aluminum for electrode reliability. Figure 4.8 shows the maximum time to breakdown for capacitors with both

polysilicon and aluminum electrodes [9]. The polysilicon is clearly superior, especially for thinner gate oxides. The inferior time to breakdown of aluminum electrode is due to the migration of aluminum atoms into the thin oxide under an electrical field. Polysilicon is also used as a diffusion source to create shallow junctions and to ensure ohmic contact to crystalline silicon. Additional uses include the manufacture of conductors and high value resistors.

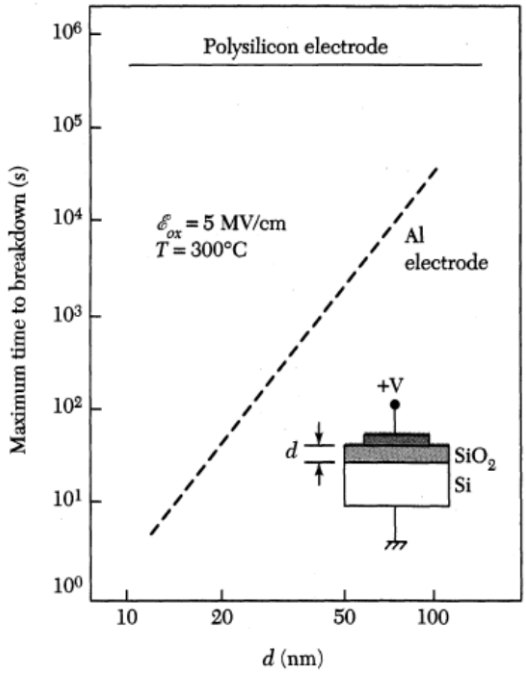


Fig. 4.8. Maximum time to breakdown versus oxide thickness for a polysilicon electrode and an aluminum electrode [9].

Over the many years that polysilicon has been used in commercial integrated circuits, the equipment used for its deposition has changed greatly, evolving from low-capacity, silicon epitaxial reactors operating at atmospheric pressure to high-volume, low-pressure CVD systems. In all commercial applications, however, chemical vapor deposition has been used to deposit polysilicon. In early prototype investigations, physical vapor deposition by evaporation was studied, but the step coverage provided by this technique was inadequate to cover the irregular surface topology of the integrated circuit; by contrast, CVD techniques provided material with excellent conformal step coverage, leading to the rapid acceptance of this more complex formation technology [10-13].

At the time that chemical vapor deposition was first used for polysilicon deposition, the most common CVD system was the horizontal, atmospheric-pressure epitaxial reactor widely used in the late 1960s and early 1970s. This type of system (illustrated in Fig. 4.9), allows operation over a wide temperature range, but its capacity is severely limited by the size of the

susceptor, on which the wafers are placed in a single layer. The low-pressure CVD (LPCVD) reactor developed to overcome the limited capacity of the horizontal system can form layers on 100-200 wafers simultaneously, although the range of conditions over which it can operate satisfactorily is severely limited. The basic elements of the LPCVD reactor are illustrated in Fig. 4.10.

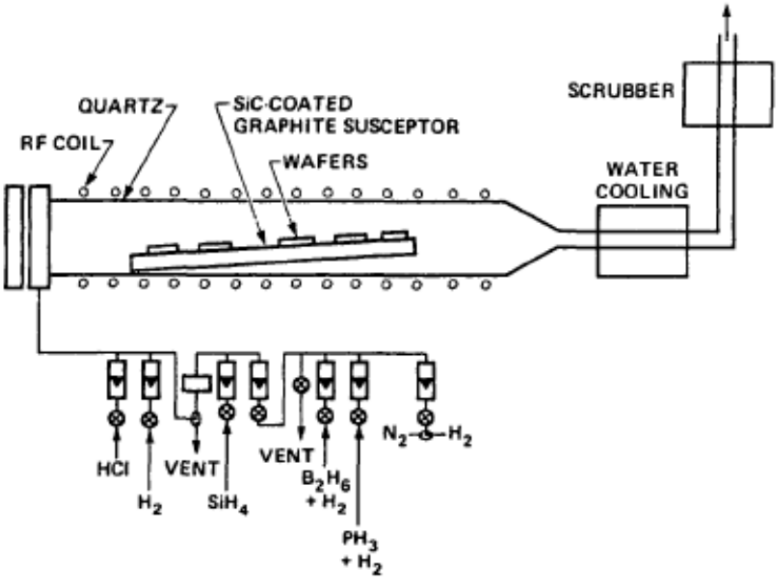


Fig. 4.9. The versatile, horizontal, cold-wall, atmospheric pressure reactor has the flexibility needed to develop many new CVD processes, but its limited wafer capacity makes it less desirable for routine manufacturing.

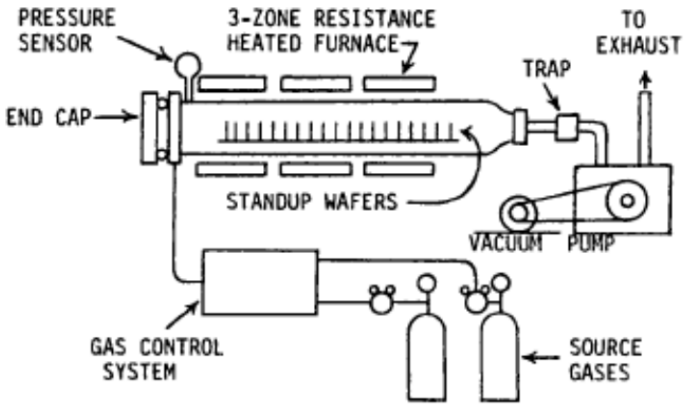


Fig. 4.10. The hot-wall, low-pressure reactor is used for routine deposition of polysilicon because of its high wafer capacity and simplicity.

4.6.1. Gas Dynamic

In an open-flow CVD reactor, the reactant gas is continuously forced through the reactor; the silicon-containing gas is, in many cases, mixed with a carrier gas. Hydrogen is usually used as the carrier gas when the deposition occurs at higher temperatures (≥ 800°C), where nitrogen may react with the silicon-containing gases; nitrogen can be used at lower temperatures, where hydrogen lowers the deposition rate because it is a reaction product.

Viscous forces exerted on the flowing gas by the susceptor and the walls of the deposition chamber slow the gas near these stationary surfaces, forming a *boundary layer*, which separates the *forced-convection* region from the wafer surfaces and the walls of the reaction chamber (Fig. 4.11) [14, 15]. The silicon-containing molecules diffuse from the forced convection region through the boundary layer to the wafer surface. The rate of diffusion R can be written [14]

$$R = D\frac{C_G}{\delta} \tag{4.10}$$

where C_G is the concentration of the silicon-containing species in the forced-convection region just above the boundary layer, D is the gas-phase diffusivity, and δ is the thickness of the boundary layer. Diffusion through the boundary layer is usually the most important of the gas-phase transport processes. Some reaction or decomposition of the silicon-containing gas can occur within the boundary layer *(homogeneous reaction),* but the final reaction should occur on the surface itself (by *heterogeneous reaction)* so that a dense silicon film is formed. Once the silicon-containing gas (or its partially reacted products) reach the surface, it is adsorbed, and further chemical reactions take place to reduce it to silicon and reaction products, which are in turn desorbed. The overall reaction rate can be written as

$$R = Ck = Ck_0 \exp(-\frac{E_a}{kT}) \tag{4.11}$$

where the reaction-rate coefficient k is characterized by an apparent activation energy E_a.

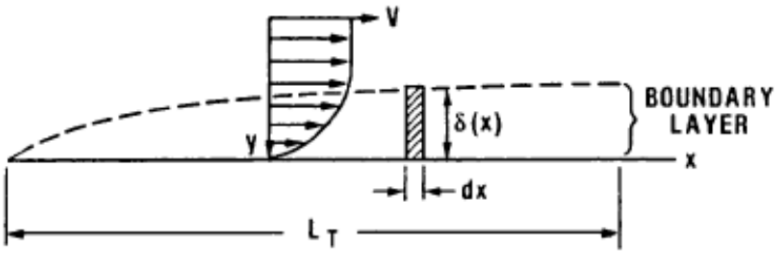


Fig. 4.11. A *boundary layer* separates the rapidly moving gas in the forced-convection region from the stationary surfaces in the deposition chamber.

Either diffusion through the boundary layer or reaction at the surface of the wafer may limit the overall deposition process. Gas-phase diffusion varies only slowly with temperature, increasing as $T^{1.5}$ or T^2, while the reaction rate increases rapidly with increasing temperature, varying as $exp(-E_a/kT)$. For silicon deposition, E_a is about 1.6 eV **(**38 kcal/mole). Therefore, at higher temperatures, the reaction proceeds rapidly, and diffusion of the gaseous silicon species through the boundary layer to the wafer surface limits the overall deposition process. In this *mass-transport-limited* regime, the deposition rate is only a weak function of temperature. **As** the temperature is reduced, the reaction rate decreases rapidly until it becomes the limiting step in the overall deposition process. In the *surface-reaction-limited* regime of operation, the deposition rate is a strong function of temperature (Fig. 4.12). Near 700°C the deposition rate changes by about 25% for a 10°C temperature variation, so that excellent temperature control is needed in this operating regime to achieve the film thickness uniformity required for controllable integrated-circuit fabrication. [Note that over a considerable range of temperatures (≥ 100°C), $C_G > C_S > 0$. In this range both mass transport and reaction influence the overall deposition process, and both must be considered.]

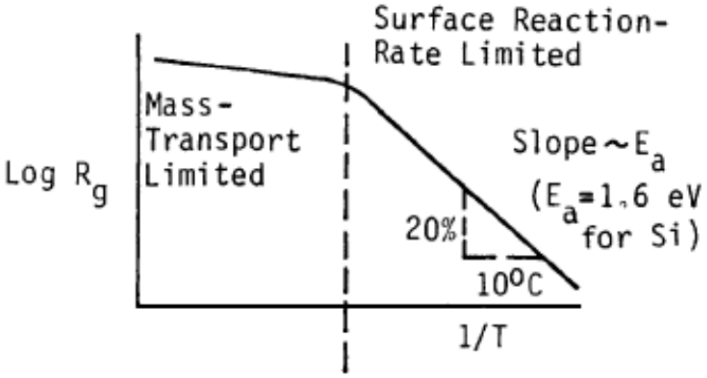


Fig. 4.12. The deposition rate is a rapidly varying function of temperature in the surface-reaction-rate-limited regime of operation (low temperatures), while it changes only slowly with temperature in the mass-transport-limited regime (higher temperatures).

Proper process design considers the geometry of the particular reactor to be used for the deposition; the relative ease of controlling the gas flow and the temperature determine the choice of operating regime. If the gas flow is well controlled, but the temperature is difficult to control, the process should operate in the temperature-insensitive, mass-transport-limited regime. If the temperature is better controlled than the gas flow, operation in the surface-reaction- limited regime is preferred. The importance of choosing the proper operating regime can be illustrated

by contrasting the deposition conditions in horizontal, atmospheric-pressure reactors with those in the low-pressure reactors now widely used.

Horizontal, Atmospheric-Pressure Reactor. In the horizontal, atmospheric-pressure, reactor, a small quantity of the silicon-containing gas is mixed with a large amount of carrier gas, and this gas mixture is forced through the reactor. Only the wafers and their supporting susceptor are heated; the walls of the chamber remain relatively cool and little deposit forms on them. The forced convection region is also relatively cool so that gases flowing in this region can travel long distances along the susceptor without reacting significantly. The majority of the temperature gradient occurs across the boundary layer. In this type of reactor, energy is often coupled into the susceptor by radio-frequency induction or by heating the opaque support plate with lamps without significantly heating the walls of the chamber. (The term "susceptor" is often used to describe the supporting plate on which the wafers sit even when heating methods other than rf-induction are used.) In either case, the temperature of the wafer can only be controlled within about 5 or 10°C, which would lead to unacceptable deposition rate and thickness variations if the reactor were operated in the reaction-limited regime. On the other hand, the gas flow, and especially diffusion through the well-defined boundary layer, is relatively well controlled, and operation of this reactor in the mass-transport-limited regime is preferred.

Although the horizontal reactor is very flexible and can operate over a wide temperature range because the forced convection flow region remains cool, its wafer capacity is severely limited by the size of quartz envelopes available and the large area needed when wafers are placed in a single layer on the susceptor surface.

Low-Pressure Reactor. To overcome the capacity limitations of the horizontal reactor, the high-capacity, low-pressure reactor was developed. In this reactor the wafers are placed in a resistance-heated furnace similar to an oxidation or diffusion furnace and are closely spaced, generally perpendicular to the axis of the tube. Figure 4.13 shows that, in this reactor, the gases flow first through the annular space between the chamber walls and the wafers. They then move between the closely spaced wafers to the wafer surfaces, where the reaction occurs. If mass transport limited the overall reaction, the deposition rate would be markedly higher near the readily accessible edges of the wafers. Therefore, in contrast to the horizontal reactor, this reactor must be operated in the reaction-limited regime. Fortunately, the temperature of the resistance-heated furnace used to heat this reactor can readily be controlled to a fraction of a degree so that the rapid variation of deposition rate with temperature in this operating regime does not degrade control of the deposited film thickness.

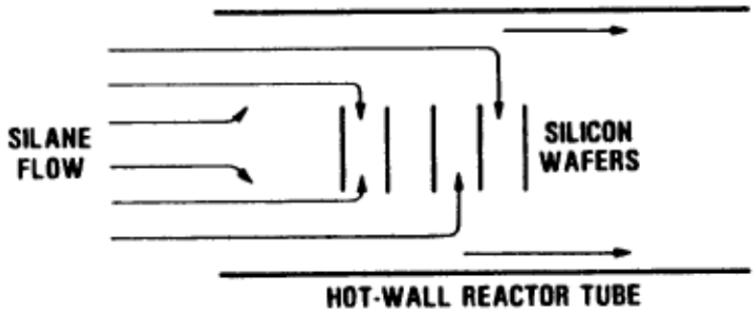


Fig. 4.13. The narrow space between wafers in the low pressure reactor makes gas diffusion to the centers of the wafers difficult.

Because of the long, narrow space through which the reactants must diffuse, obtaining reaction-limited operation requires considerable effort. Lowering the temperature alone is not adequate; the ease of mass transport must also be enhanced. As we have already seen, mass transport can be characterized by the ratio D/δ, where D is the gas-phase diffusivity and δ is the boundary-layer thickness or another characteristics dimension of the deposition system. The rapid variation of the diffusivity with pressure is the key to achieving surface-reaction-limited operation in this type of reactor. The diffusivity is inversely proportional to the total pressure; reducing the pressure by a factor of several thousand to a fraction of a torr (tens of pascals) increases the ease of gas-phase diffusion by a similar amount. Making the mass transport easier by operating at reduced pressures thus moves the deposition process into the reaction-rate-limited regime, as desired. To maximize the deposition rate, the partial pressure of the silicon-containing gas must be comparable to the total pressure, and little or no carrier gas is used in this reactor. Typical deposition conditions are shown in Table 4.3.

Table 4.3. Typical polysilicon deposition conditions.

	Atmosheric pressure reactors	Low pressure reactors	
Temperature (°C)	950	620	640
Silane partial pressure (torr)	0.3	0.2	0.2
Total pressure (torr)	760	0.2	1
Carrier Gas	H_2	None	N_2 or H_2
Deposition Rate (nm/min)	120	10	15
Wafer capacity	20	100	100
Throughput (wafers/hour) 40		100-150	100-150

One significant difference between the hot-wall LPCVD system and the cold wall reactor is the temperature of the gases in the forced-convection region. Because the gases are heated significantly in the forced-convection region of the hot-wall reactor, they may decompose or react in this region, leading to particles on the wafer surface or a porous film, as well as changing the deposition rate. Gas-phase decomposition is also promoted by the limited amount of carrier gas available to dilute the reactants and decrease the reaction probability. For example, a hydrogen carrier gas can suppress the thermal decomposition of silane because it is one of the products of the reaction

$$SiH_4(g) \rightarrow Si(s) + 2H_2(g) \qquad (4.12)$$

The presence of hot gases in the forced-convection region is even more deleterious when silicon compounds are to be deposited in the hot-wall reactor. One gas may tend to decompose before the desired reaction between different gases occurs, limiting the variety of gases that can be effectively used. For example silane decomposes more readily when it reacts with ammonia in the reaction

$$3SiH_4(g) + NH_3(g) \rightarrow Si_3N_4(s) + 12H_2(g) \qquad (4.13)$$

and the less easily decomposed, silicon-containing gas dichlorosilane (SiH_2C1_2) is generally used in place of silane, leading to other problems.

4.6.2. Wafer-to-Wafer Uniformity

In either the horizontal reactor or the conventional low-pressure reactor, the gases are inserted from one end of the reaction chamber, and flow along the wafer load. For efficient utilization of the silicon-containing gas, its partial pressure must decrease along the wafer load. Unless compensated, this gas depletion causes the deposition rate to decrease along the wafer load, with unacceptable variations in film thickness. In each reactor a means related to the parameters limiting the deposition must be found to achieve a uniform thickness regardless of the position of the wafer.

As it was already discussed, the horizontal reactor is usually operated so that the overall deposition process is controlled by diffusion through the boundary layer. From Eq. 4.10 the diffusion rate of the silicon-containing species is just CD/δ, where C is the concentration of the silicon-containing gas in the forced-convection region. Because this gas is consumed as silicon is deposited, its concentration varies with distance x along the deposition chamber. The ratio $C(x)D/S$ can only be kept constant if one of the other variables can be made a compensating function of position. In practice, δ can be readily made to decrease with position to compensate for the decrease in C.

The boundary layer is formed by the viscous forces exerted on the gas in the forced-convection region by the susceptor and walls of the chamber. If the gas can be forced to travel faster, it resists the viscous forces more readily, and the boundary layer becomes thinner. Because the amount of gas blowing in the forced-convection region remains almost constant along the length of the deposition chamber, its velocity can be increased by decreasing the effective cross section of the chamber. By tilting the susceptor so that the cross section through which the gas must flow decreases along the direction of gas flow, the velocity increases, and the boundary layer thickness decreases. Consequently, the ratio $DC(x)/\delta(x)$ is less sensitive to gas depletion along the length of the deposition chamber. In practice, tilting the susceptor by 1 or 2°is adequate to compensate for several tens of percent gas depletion.

Different mechanisms limit the deposition process in the low-pressure reactor. In this reactor, the deposition rate depends strongly on temperature; consequently, a slight increase in temperature along the length of the deposition chamber can compensate for moderate gas depletion, and a uniform deposition rate can again be obtained along the length of the reaction chamber.

Severe gas depletion cannot be overcome, however, and proper process design involves a trade-off between deposition rate, efficient gas utilization, and uniformity. Figure 4.14 shows the deposition rate along the reaction chamber at several different temperatures (with the temperature uniform along the length of the chamber at each temperature). At the low temperature of 525°C, little of the silicon-containing gas is consumed, and the deposition rate is uniform along the length of the deposition chamber even without a temperature gradient. However, the very low deposition rate leads to an unacceptably low reactor throughput. The average deposition rate increases by about 10 times between 525 and 625°C, but at 625°C the rate varies moderately along the length of the chamber. It is in this range that a temperature gradient can be used to improve uniformity. At still higher temperatures, the maximum deposition rate increases, as is desirable for higher wafer throughput, but severe gas depletion greatly decreases the deposition rate toward the downstream end of the deposition chamber. Because virtually all of the siliconcontaining gas is being depleted before reaching this region, uniformity cannot be achieved by using a temperature gradient. Proper process design requires operating at a lower average temperature, at which the gas depletion is only moderate and can be compensated by using a temperature gradient. (The uniformity can also be improved somewhat by using a higher gas velocity so that the gas travels farther along the deposition chamber before entering into the reaction; however, this decreases the efficiency of gas use [16].

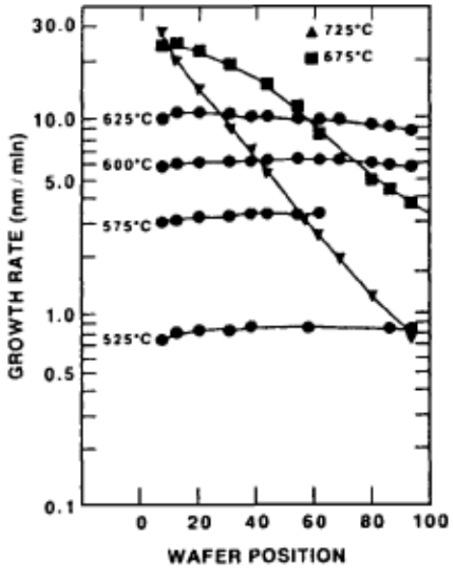


Fig. 4.14. In the hot-wall, low-pressure reactor, the deposition rate decreases along the length of the reactor when a significant fraction of the reactant gas is consumed. This gas depletion is more important at higher deposition temperatures.

Using a temperature gradient can have detrimental effects. Because the structure of polysilicon varies rapidly with deposition temperature, any device property which depends on the structure can vary with the position of the wafer in the reactor. The electrical conduction through oxide grown on polysilicon depends sensitively on the polysilicon structure; consequently, conduction through the oxide can vary with wafer position in the reactor during polysilicon deposition. Such variations cannot be tolerated in critical applications, such as in electrically erasable, programmable read-only memories (EEPROMs), which rely on controlled conduction through oxides grown on polysilicon. Therefore, for films which are to be used for critical applications, temperature gradients cannot be used, and modified LPCVD reactors must be employed. Two different approaches can be taken to avoid the need for a temperature gradient. One approach merely modifies the standard, tube-type LPCVD reactor by injecting the silicon-containing gas at several locations along the length of the reaction chamber, rather than only at one end. The gases are still exhausted at one end of the tubular reactor chamber, as in the conventional reactor. The second approach uses a markedly different reactor geometry, in which the wafers are placed in a constant-temperature, bell-jar-shaped furnace, as shown in Fig. 4.15. The gases are injected all along the wafer load; they flow through the narrow space between a single pair of wafers and then are immediately exhausted from the system. Both approaches provide good thickness uniformity without using a temperature gradient.

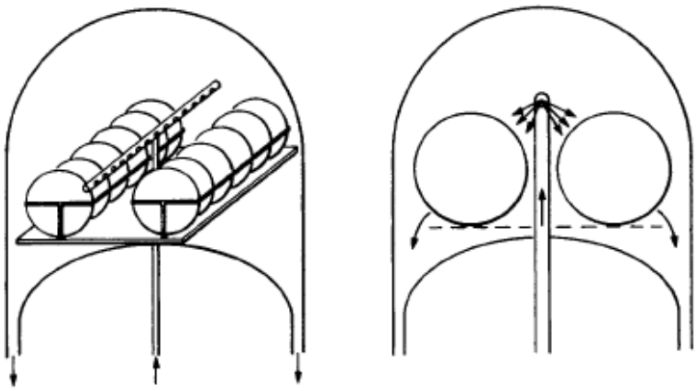


Fig. 4.15. In the *vertical-flow* reactor the gases pass between a single pair of wafers and are then removed from the deposition chamber, decreasing the amount of gas depletion.

4.6.3. Silicon Gas Sources

Several silicon-containing gas sources are available for the chemical vapor deposition of silicon. The most commonly used species contain silicon plus varying amounts of hydrogen and chlorine. Although several of these gases are used to deposit silicon on silicon, nucleation on an insulating oxide or nitride surface may be difficult with the chlorine-containing species, and polysilicon is usually deposited by the thermal decomposition or *pyrolysis* of silane (SiH_4). The overall reaction can be written

$$SiH_4(g) \text{ -+ } Si(s) + 2H_2(g) \qquad (4.14)$$

When polysilicon is deposited on surfaces on which nucleation occurs more readily, other gases, such as dichlorosilane (SiH_2Cl_2), trichlorosilane ($SiHCl_3$) or silicon tetrachloride ($SiCl_4$) can be used. All of the reactants in a CVD system enter into the reaction chamber as gases. This is most conveniently accomplished if the reactants are in the gas phase at room temperature so that they can be metered by conventional flowmeters and forced into the reaction chamber by the pressure difference between the gas source and the chamber. Silane is convenient in this respect because it is a gas at room temperature.

However, silane is highly explosive and must be used with great care; proper design of the storage and purging system is especially crucial. Although silane is *pyrophoric* and ignites spontaneously when exposed to oxygen, the reaction may not occur immediately if the silane and oxygen are not adequately mixed. A large quantity of silane may accumulate and then explode when contact between the silane and oxygen increases. To avoid silane accumulation if a leak occurs, any locations where stagnant gas can accumulate must be eliminated by forcing large quantities of air or nitrogen around the gas cylinder and other locations where leaks might occur. Extremely small leaks may not be a safety hazard, but the SiO_2 formed by the reaction of silane

and oxygen from air can cause particles which degrade film quality. It can also coat the walls of the plumbing and mass flow controllers, changing the calibration of the latter.

In an atmospheric-pressure system, the gases are pushed through the reaction chamber by the higher pressure of the incoming gases. In a low-pressure system, a pump pulls the gases through the chamber. Mechanical pumps, sometimes augmented by a “Roots blower’’ are generally adequate to obtain the pressures used in typical LPCVD reactors. Because of the corrosive nature of the gases, the pumps and pump oil must be carefully selected. Particles resulting from reaction of the gas species being pumped must be frequently filtered from the pump oil to avoid damaging the pump.

After the gases leave the reaction chamber, they are cooled and flow through a “scrubber” which removes toxic or environmentally damaging species before being discharged to the atmosphere. This scrubber may be dedicated to a particular reactor, or one scrubber may serve an entire fabrication area. In the latter case, possible reactions between gases from various sources must be considered.

4.6.4. Doping During Deposition

In many cases polysilicon is deposited undoped, and dopant atoms are subsequently added by ion implantation or from a gas-phase source. In some cases, however, the process flow can be simplified by adding the desired dopant to the polysilicon during deposition. High concentrations of phosphorus are frequently added during the deposition when highly conducting gate electrodes and interconnections are required. Phosphorus is added by introducing phosphine gas (PH_3) into the deposition chamber along with the silane. Because of the different reactivities of the two gases, however, the dopant concentration may vary along the wafer load. This variation is immediately visible as a variation in resistivity if the dopant concentration in the polysilicon grains is less than the solid solubility of the dopant in silicon. However, if the dopant concentration exceeds the solid solubility, excess dopant may be incorporated in the film without being electrically active. In this case, the resistivity appears uniform, but the excess dopant atoms can degrade the polysilicon. This degradation is especially important in the thinner polysilicon films are in advanced VLSI processes. The dopant gases phosphine (PH_3), arsine (AsH_3), and diborane (B_2H_6) are convenient gas-phase dopant sources. However, adding large quantities of these dopant gases may itself alter the deposition process. The *n*-type dopant gases phosphine and arsine can severely depress the desposition rate when introduced in large quantities. The deposition rate may change by as much as a factor of ten, and the change is more severe for arsine than for phosphine [17]**.** The effect is most visible in the surface-reaction-limited regime,

in which the LPCVD reactor operates. Unlike phosphine and arsine, the *p*-type dopant gas diborane *iizcreases* the deposition rate [18, 19].

Although the cause of these rate changes is not fully understood, it may be visualized in the following manner: Phosphine and arsine or their reaction products are strongly bound to surface adsorption sites, which are then unavailable to the incoming silicon-containing gas, decreasing the deposition rate. These adsorbed atoms on the surface can exert long-range repulsive forces so that the silicon-containing atoms are efficiently repelled by even moderate surface coverage. Phosphorus is known to efficiently repel other species from surfaces; in chemical kinetic studies, the walls of experimental chambers are often coated with phosphoruscontaining solutions to impede the heterogeneous recombination of hydrogen atoms and other species at the walls [20]. The efficiency of the repulsion reaction has been related to the electronegativity of the blocking species. The increase of deposition rate when boron is added is attributed to a catalytic effect of adsorbed boron atoms on the adsorption or decomposition of the silicon-containing species [21], thus providing an efficient parallel deposition path. The severe decrease of the silicon deposition rate as *n*-type dopant gas is added during the deposition can be avoided by using other silicon-containing gases, such as disilane (Si_2H_6) [22, 23]. Disilane decomposes into silane and silylene (SiH_2). The latter forms very strong bonds to silicon, and its adsorption is not readily blocked by phosphorus- or arsenic-containing species adsorbed on the depositing surface. While the deposition rate is reduced by a factor of ten for phosphorus concentrations in the $10^{20}cm^{-3}$ range when silane is used in the LPCVD reactor, the decrease is about twenty times less for disilane under the same conditions [22]. However, achieving uniform deposition across a wafer may be more difficult with disilane because of its more reactive nature - the very reason that it is not blocked by adsorbed dopant atoms.

The dopant gases are, of course, highly toxic. Because large quantities are needed to dope the films during deposition, proper handling is required. In particular, decomposition of phosphine or the other gases at the exhaust end of the deposition system must be considered in system design. Additional apparatus is sometimes added to complete the dopant-gas reactions as the gases leave the main deposition chamber.

4.6.5. Polysilicon Deposition Process

A low-pressure reactor (Fig. 4.2*a*) operated between 600°C and 650°C is used to deposit polysilicon by pyrolyzing silane according to the following reaction.

$$SiH4 \rightarrow^{600°C} Si + 2H_2 \qquad (4.15)$$

Of the two most common low-pressure processes one operates at a pressure of 25-130 Pa using 100% silane, whereas the other process involves a diluted mixture of 20%-30% silane in nitrogen at the same total pressure. Both processes can deposit polysilicon on hundreds of wafers per run with good uniformity (i.e., thickness within 5%).

Figure 4.16 shows the deposition rate at four deposition temperatures. At low silane partial pressure, the deposition rate is proportional to the silane pressure [3]. At higher silane concentrations, saturation of the deposition rate occurs. Deposition at reduced pressure is generally limited to temperatures between 600°C and 650°C. In this temperature range, the deposition rate varies as $exp(-E_a/kT)$, where the activation energy E_a is 1.7 eV, which is essentially independent of the total pressure in the reactor. At higher temperatures, gas-phase reactions that result in a rough, loosely adhering deposit become significant and silane depletion will occur, causing poor uniformity. At temperatures much lower than 600°C, the deposition rate is too slow to be practical.

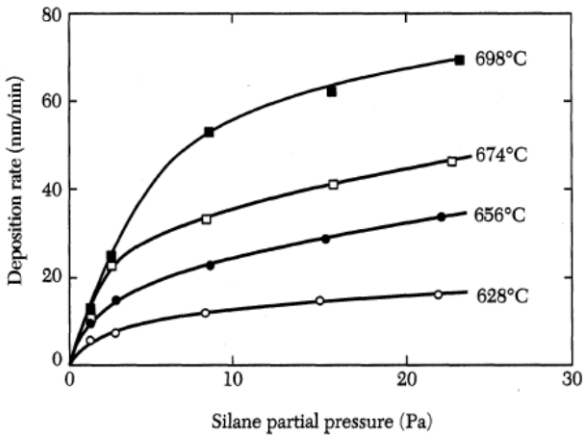


Fig. 4.16. Effect of silane concentration on the polysilicon deposition rate [3].

Process parameters that affect the polysilicon structure are deposition temperature, dopants, and the heat cycle applied following the deposition step. A columnar structure results when polysilicon is deposited at a temperature of 600°C - 650°C. This structure is comprised of polycrystalline grains ranging in size from 0.03 to 0.3 μm at a preferred orientation of (110). When phosphorus is diffused at 950°C, the structure changes to crystallite and grain size increases to a size between 0.5 and 1.0 μm. When temperature is increased to 1050°C during oxidation, the grains reach a final size of 13 μm. Although the initially deposited film appears amorphous when deposition occurs below 600°C, growth characteristics similar to the polycrystalline-grain columnar structure are observed after doping and heating.

Polysilicon can be doped by diffusion, ion implantation, or the addition of dopant gases during deposition, referred to as in situ doping. The implantation method is most commonly used because of its lower processing temperatures. Figure 4.17 shows the sheet resistance of single crystal silicon and of 500 nm polysilicon doped with phosphorus and antimony using ion implantation [24]. Implant dose, annealing temperature, and annealing time all influence the sheet resistance of implanted polysilicon. Carrier traps at the grain boundaries cause a very high resistance in the lightly implanted polysilicon. As Fig. 4.17 illustrates, resistance drops rapidly, approaching that of implanted single crystal silicon, as the carrier traps become saturated with dopants.

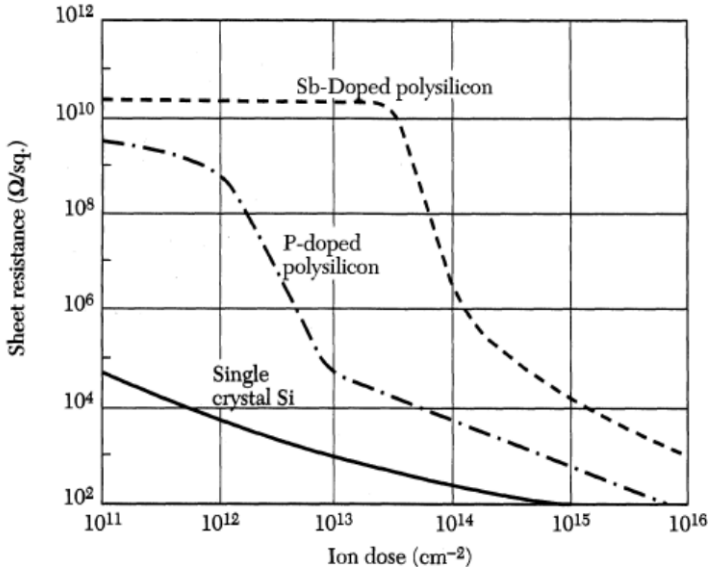


Fig. 4.17. Sheet resistance versus ion dose into 500 nm polysilicon at 30 keV [24].

4.7. Conclusions

Modern semiconductor device fabrication requires the use of thin films. Currently, there are four important types of films-thermal oxides, dielectric layers, polycrystalline silicon, and metal films. The major issues related to film formation are low-temperature processing, step coverage, selective deposition, uniformity, film quality, planarization, throughput, and large-wafer capacity.

Thermal oxidation offers the best quality for the Si-SiO_2 interface and has the lowest interface trap density. Therefore, it is used to form the gate oxide and the field oxide, LPCVD of dielectrics and polysilicon offer conformal step coverage. In contrast, PVD and atmospheric-pressure CVD generally result in noncomformal step coverage. To minimize the *RC* time delay due to parasitic resistance and capacitance low-dielectric-constant materials for interlayer films

are extensively used to meet the requirements of the multilevel interconnect structures of ULSI circuits. In addition, the high-dielectric-constant materials to improve the gate insulator performance and to increase the capacitance per unit area for DRAM have been considered.

Chapter 5. Silicon Oxidation

A.Evtukh

5.1. Introduction

Many different kinds of thin films are used to fabricate discrete devices and integrated circuits, including thermal oxides, dielectric layers, polycrystalline silicon, and metal films. An important oxide layer is the gate oxide, under which a conducting channel can be formed between the source and the drain. A related layer is the field oxide, which provides isolation from other devices. Both gate and field oxides generally are grown by a thermal oxidation process because only thermal oxidation can provide the highest-quality oxides having the lowest interface trap densities.

Semiconductors can be oxidized by some methods, including thermal oxidation and electrochemical anodization. Among these, thermal oxidation is the most important for silicon devices. It is a key process in modern silicon IC technology. The basic thermal oxidation apparatus (Fig. 5.1) consists of a resistance-heated furnace, a cylindrical fused-quartz tube containing the silicon wafers held vertically in a slotted quartz boat, and a source of either pure dry oxygen or pure water vapor. Oxidation temperature is generally in the range of 900–1200°C, and the typical gas flow rate is about 1 L/min. The oxidation system uses microprocessors to regulate the gas flow sequence, to control the automatic insertion and removal of silicon wafers, to ramp the temperature up (i.e., to increase the furnace temperature linearly) from a low temperature to the oxidation temperature, to maintain the oxidation temperature to within ±1°C, and to ramp the temperature down when oxidation is completed.

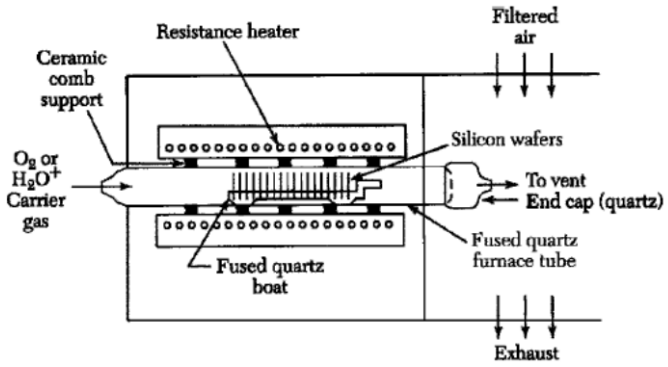


Fig. 5.1. Schematic of an oxidation furnace [1].

5.2. Thermal oxidation process

5.2.1. Growth Kinetics

The following chemical reactions describe the thermal oxidation of silicon in oxygen (“dry” oxidation) and water vapor (“wet” oxidation), respectively:

$$Si(solid) + O_2(gas) \rightarrow SiO_2(solid) \quad (5.1)$$

$$Si(solid) + 2H_2O(gas) \rightarrow SiO_2(solid) + 2H_2(gas) \quad (5.2)$$

The silicon–silicon dioxide interface moves into the silicon during the oxidation process. This creates a new interface region, with surface contamination on the original silicon ending up on the oxide surface. As a result of the densities and molecular weights of silicon and silicon dioxide, growing an oxide of thickness x consumes a layer of silicon $0.44x$ thick (Fig. 5.2).

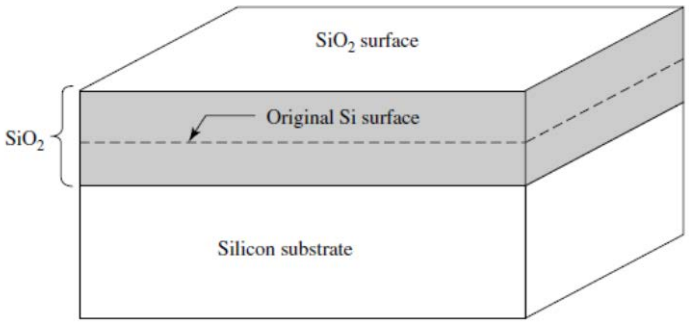


Fig. 5.2. Movement of silicon–silicon dioxide interface during oxide growth [1].

The kinetics of silicon oxidation can be described on the basis of the simple model illustrated in Fig. 5.3. A silicon slice contacts the oxidizing species (oxygen or water vapor), resulting in a surface concentration of C_0 molecules/cm^3 for these species. The magnitude of C_0 equals the equilibrium bulk concentration of the species at the oxidation temperature. The equilibrium concentration generally is proportional to the partial pressure of the oxidant adjacent to the oxide surface. At 1000°C and a pressure of 1 atm, the concentration C_0 is 5.2×10^{16} cm^{-3} for dry oxygen and 3×10^{19} cm^{-3} for water vapor.

The oxidizing species diffuses through the silicon dioxide layer, resulting in a concentration C_s at the surface of silicon. The flux F_1 can be written as

$$F_1 = D\frac{dC}{dx} \cong \frac{D(C_0 - C_s)}{x} \quad (5.3)$$

where D is the diffusion coefficient of the oxidizing species, and x is the thickness of the oxide layer already present.

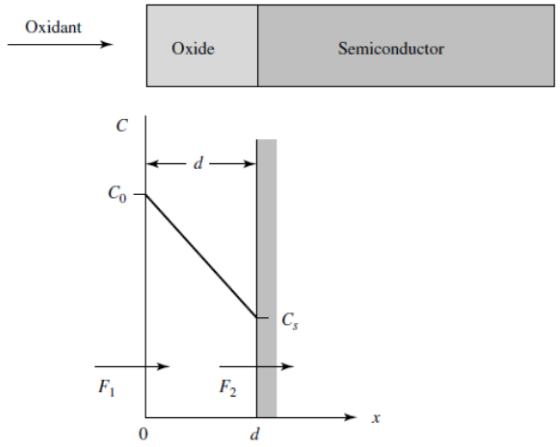


Fig. 5.3. Basic model for the thermal oxidation of silicon [1].

At the silicon surface, the oxidizing species reacts chemically with silicon. Assuming the rate of reaction to be proportional to the concentration of the species at the silicon surface, the flux F_2 is given by

$$F_2 = kC_s \tag{5.4}$$

where k is the surface reaction rate constant for oxidation. At the steady state, $F_1 = F_2 = F$. Combining Eqs. (5.3) and (5.4) gives

$$F = \frac{DC_0}{x + (D/k)} \tag{5.5}$$

The reaction of the oxidizing species with silicon forms silicon dioxide. Let C_1 be the number of molecules of the oxidizing species in a unit volume of the oxide. There are 2.2×10^{22} silicon dioxide molecules/cm^3 in the oxide, and one oxygen molecule (O_2) is added to each silicon dioxide molecule, whereas we add two water molecules (H_2O) to each SiO_2 molecule. Therefore, C_1 for oxidation in dry oxygen is 2.2×10^{22} cm^{-3}, and for oxidation in water vapor it is twice this number (4.4×10^{22} cm^{-3}). Thus, the growth rate of the oxide layer thickness is given by

$$\frac{dx}{dt} = \frac{F}{C_1} = \frac{DC_0 / C_1}{x + (D/k)} \tag{5.6}$$

This differential equation can be solved subject to the initial condition, $x(0) = d_0$, where d_0 is the initial oxide thickness; d_0 can also be regarded as the thickness of oxide layer grown in an earlier oxidation step. Solving Eq. (5.6) yields the general relationship for the oxidation of silicon:

$$x^2 + \frac{2D}{k}x = \frac{2DC_0}{C_1}(t + \tau) \tag{5.7}$$

where $\tau \equiv (d_0^2 + 2Dd_0/k)C_1/2DC_0$, which represents a time coordinate shift to account for the initial oxide layer d_0.

The oxide thickness after an oxidizing time t is given by

$$x = \frac{D}{k}\left[\sqrt{1 + \frac{2C_0k^2(t+\tau)}{DC_1}} - 1\right] \tag{5.8}$$

For small values of t, Eq. (5.8) reduces to

$$x \cong \frac{C_0k}{C_1}(t+\tau) \tag{5.9}$$

and for larger values of t, it reduces to

$$x \cong \sqrt{\frac{2DC_0}{C_1}(t+\tau)} \tag{5.10}$$

During the early stages of oxide growth, when surface reaction is the rate limiting factor, the oxide thickness varies linearly with time. As the oxide layer becomes thicker, the oxidant must diffuse through the oxide layer to react at the silicon–silicon dioxide interface and the reaction becomes diffusion-limited. The oxide growth then becomes proportional to the square root of the oxidizing time, which results in a parabolic growth rate.

Equation (5.7) is often written in a more compact form

$$x^2 + Ax = B(t+\tau) \tag{5.11}$$

where $A = 2D/k$, $B = 2DC_0/C_1$ and $B/A = kC_0/C_1$. Using this form, Eqs. (5.9) and (5.10) can be written as

$$x = \frac{B}{A}(t+\tau) \tag{5.12}$$

for the linear region and as

$$x^2 = B(t+\tau) \tag{5.13}$$

for the parabolic region. For this reason, the term B/A is referred to as the *linear rate constant* and B is the *parabolic rate constant*. Experimentally measured results agree with the predictions of this model over a wide range of oxidation conditions. For wet oxidation, the initial oxide thickness d_0 is very small, or $\tau \cong 0$. However, for dry oxidation, the extrapolated value of d_0 at $t = 0$ is about 25 nm. Thus, the use of Eq. (5.11) for dry oxidation on bare silicon requires a value for τ that can be generated using this initial thickness. Table 5.1 lists the values of the rate constants for wet oxidation of silicon, and Table 5.2 lists the values for dry oxidation.

Table 5.1. Rate constants for wet oxidation of silicon.

Temperature (°C)	A (μm)	B (μm^2/h)	τ (h)
1200	0.05	0.72	0
1100	0.11	0.51	0
1000	0.226	0.287	0
920	0.5	0.203	0

Table 5.2. Rate constants for dry oxidation of silicon.

Temperature (°C)	A (μm)	B (μm^2/h)	τ (h)
1200	0.04	0.045	0.027
1100	0.09	0.027	0.076
1000	0.165	0.0117	0.37
920	0.235	0.0049	1.4
800	0.37	0.0011	9.0
700	-	-	81.0

The temperature dependence of the linear rate constant B/A is shown in Fig. 5.4 for both dry and wet oxidation and for (111)- and (100)-oriented silicon wafers [1-6]. The linear rate constant varies as $exp(-E_a/k_BT)$, where the activation energy E_a is about 2 eV for both dry and wet oxidation. This closely agrees with the energy required to break silicon–silicon bonds, 1.83 eV/molecule. Under a given oxidation condition, the linear rate constant depends on crystal orientation. This is because the rate constant is related to the rate of incorporation of oxygen atoms into the silicon. The rate depends on the surface bond structure of silicon atoms, making it orientation-dependent. Because the density of available bonds on the (111) plane is higher than that on the (100) plane, the linear rate constant for (111) silicon is larger.

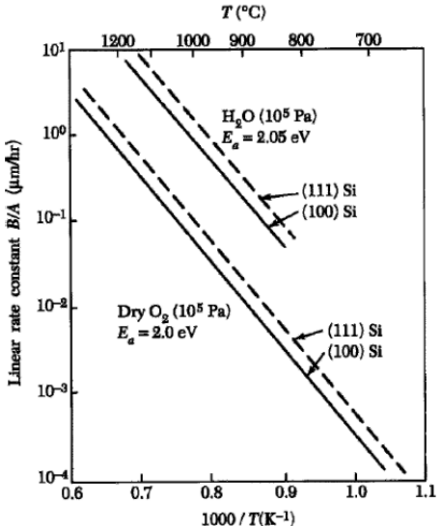


Fig. 5.4. Linear rate constant versus temperature [1].

Figure 5.5 shows the temperature dependence of the parabolic rate constant B, which can also be described by $exp(-E_a/kT)$. The activation energy E_a is 1.24 eV for dry oxidation. The comparable activation energy for oxygen diffusion in fused silica is 1.18 eV. The corresponding value for wet oxidation, 0.71 eV, compares favorably with the value of 0.79 eV for the activation energy of diffusion of water in fused silica. The parabolic rate constant is independent of crystal orientation. This independence is expected because it is a measure of the diffusion process of the oxidizing species through a random network layer of amorphous silica.

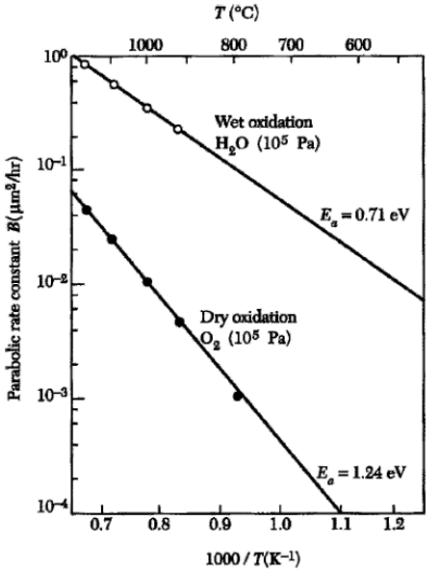


Fig. 5.5. Parabolic rate constant versus temperature [1].

Although oxides grown in dry oxygen have the best electrical properties, considerably more time is required to grow the same oxide thickness at a given temperature in dry oxygen than in water vapor. For relatively thin oxides such as the gate oxide in a MOSFET (typically $\leq$20 nm), dry

oxidation is used. However, for thicker oxides such as field oxides (≥20 nm) in MOS integrated circuits, and for bipolar devices, oxidation in water vapor (or steam) is used to provide both adequate isolation and passivation.

Figure 5.6 shows the experimental results of silicon dioxide thickness as a function of reaction time and temperature for two substrate orientation [2]. Under a given oxidation condition, the oxide thickness grown on a (111)-substrate is larger than that grown on a (100)-substrate because of the larger linear rate constant of the (111)-orientation. Note that for a give temperature and time, the oxide film obtained using wet oxidation is about 5-10 times thicker than that using dry oxidation.

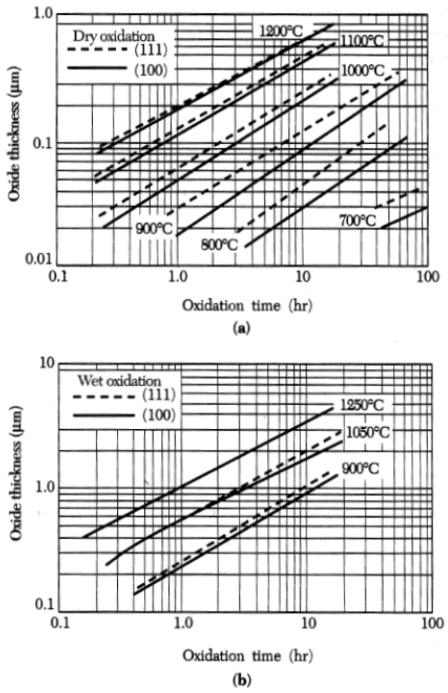


Fig. 5.6. Experimental results of silicon dioxide thickness as a function of reaction time and temperature for two substrate orientations. (*a*) Growth in dry oxygen. (*b*) Growth in steam.

5.2.2. Thin Oxide Growth

Relatively slow growth rates must be used to reproducibly grow thin oxide films of precise thickness. Approaches to achieve such slower growth rates include growth in dry O_2 at atmospheric pressure and lower temperatures (800–900°C); growth at pressures lower than atmospheric pressure; growth in a reduced partial pressures of O_2 by using a diluent inert gas, such as N_2, Ar, or He, together with the gas containing the oxidizing species; and the use of composite oxide films with the gate oxide films consisting of a layer of thermally grown SiO_2 and an overlayer of chemical vapor deposition (CVD) SiO_2. However, the mainstream approach for gate oxides 10–15 nm thick is to grow the oxide film at atmospheric pressure and lower temperatures (800–900°C).

With this approach, processing using modern *vertical* oxidation furnaces can grow reproducible, high-quality 10-nm oxides to within 0.1 nm across the wafer.

It was noted earlier that for dry oxidation, there is a rapid early growth that gives rise to an initial oxide thickness d_0 of about 20 nm. Therefore, the simple model given by Eq. (5.11) is not valid for dry oxidation with an oxide thickness ≤20 nm. For ultra-large-scale integration, the ability to grow thin (5–20 nm), uniform, high-quality reproducible gate oxides has become increasingly important.

In the early stage of growth in dry oxidation, there is a large compressive stress in the oxide layer that reduces the oxygen diffusion coefficient in the oxide. As the oxide becomes thicker, the stress will be reduced due to the viscous flow of silica and the diffusion coefficient will approach its stress-free value. Therefore, for thin oxides, the value of D/k may be sufficiently small that we can neglect the term Ax in Eq. (5.11) and obtain

$$x^2 + d_0^2 = Bt \tag{5.14}$$

where d_0 is equal to $\sqrt{2DC_0\tau/C_1}$, which is the initial oxide thickness when time is extrapolated to zero, and B is the parabolic rate constant defined previously.

5.3. Impurity redistribution during oxidation

Dopant impurities near the silicon surface will be redistributed during thermal oxidation The redistribution depends on several factors. When two solid phases are brought together, an impurity in one solid will redistribute between the two solids until it reaches equilibrium. The ratio of the equilibrium concentration of the impurity in the silicon to that in the silicon dioxide is called the segregation coefficient and is defined as

$$k = \frac{C_{Si}}{C_{SiO2}}, \tag{5.15}$$

where C_{Si} is the equilibrium concentration of impurity in silicon, C_{SiO2} is the equilibrium concentration of impurity in SiO_2.

A second factor that influences impurity distribution is that the inipurity may diffuse rapidly through the silicon dioxide and escape to the gaseous ambient. If the diffusivity of the impurity in silicon dioxide is large, this factor will be important. A third factor in the redistribution process is that the oxide is growing, and thus the boundary between the silicon and the oxide is advancing into the silicon as a function of time. The relative rate of this advance compared with the diffusion rate of the impuritv through the oxide is important in determining the extent of the redistribution. Note that even if the segregation coefficient of an impurity equals unity, some redistribution of the umpurity in the silicon will still take place.

As indicated in Fig. 5.2, the oxide layer will be about twice as thick as the silicon layer it replaced. Therefore, the same amount of impurity will now be distributed in a large volume, resulting in depletion of the impurity from the silicon.

Four possible redistribution processes are illustrated in Fig. 5.7 [3]. These processes can be classified into two groups. In one group, the oxide takes up the impurite (Fig. 5.7, *a* and *b* for $k<1$), and in the other the oxide rejects the impurity (Fig. 5.7, *c* and *d* for $k>1$). In each case, what happens depends on how rapidly the impurity can diffuse through the oxide. In group 1, the silicon surface is depleted of impurities; an example is boron, with *k* approximately equal to 0.3. Rapid diffusion of the impurity through the silicon dioxide increases the amount of depletion; an example is boron-doped silicon heated in a hydrogen ambient, because hydrogen in silicon dioxide enhances the diffusivity of boron. In group 2, *k* is greater than unify, so the oxide rejects the impurity. If diffusion of the impurity through the silicon dioxide is relatively slow, the impurity piles up near the silicon surface; an example is phosphorus, with *k* approximately equal to 10. When diffusion through the silicon dioxide is rapid, so much impurity may escape from the solid to the gaseous ambient that the overall effect will be a depletion of the impurity; an example is gallium, with *k* approximately equal to 20.

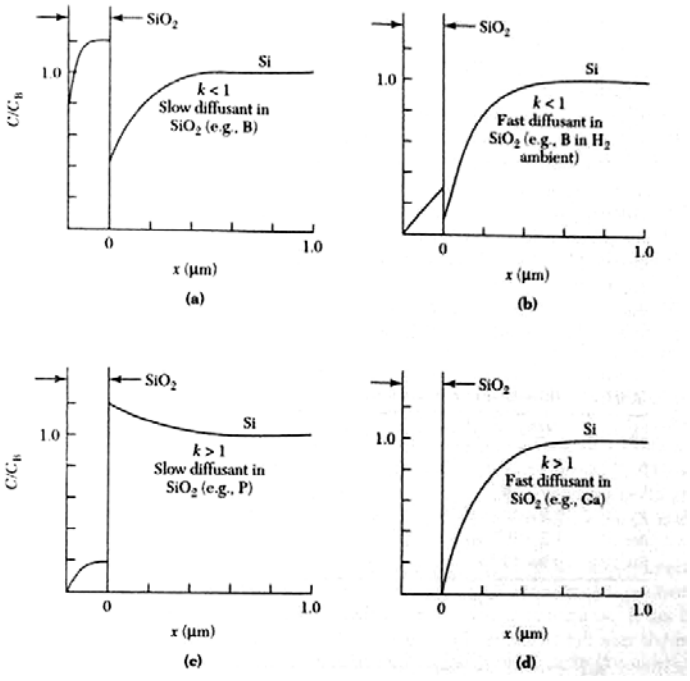


Fig. 5.7. Four different cases of impurity redistribution in silicon due to thermal oxidation.

The redistributed dopant impurities in silicon dioxide are seldom electrically active. However, redistribution in silicon has an important effect on processing and device performance. For example, nonuniform dopant distribution will modify the interpretation of

the measurements of interface trap properties, and the change of the surface concentration will modify the threshold voltage and device contact resistance.

5.4. Masking properties of silicon dioxide

A silicon dioxide layer can also provide a selective mask against the diffusion of dopants at elevated temperatures, a very useful property in IC fabrication. Predeposition of dopants, whether it be by ion implantation, chemical diffusion, or spin-on techniques, typically results in a dopant source at or near the surface of the oxide. During a subsequent high-temperature drive-in step, diffusion its oxide-masked regions must be slow enough with respect to diffusion in the silicon to prevent dopants from diffusing through the oxide mask to the silicon surface. The required thickness may be determined experimentally by measuring the oxide thickness necessary to prevent the inversion of a lightly doped silicon substrate of opposite conductivity at a particular temperature and time. Typically, oxides used for masking common impurities are 0.5 to 1.0 μm thick.

The values of diffusion constants for various dopants in various dopants in SiO_2 depend on the concentration, properties, and structure of the oxide. Table 5.3 lists diffusion constants for various common dopants, and Fig. 5.8 gives the oxide thickness required to mask boron and phosphorus as a function: of diffusion time and temperature. Note that SiO_2 is much more effective for masking boron than phosphorus. Nevertheless, the diffusivities of P, Sb, As, and B in SiO_2 are all orders of magnitude less than their corresponding values in silicon, so they are all compatible with oxide masking. This is not true, however, for Ga or Al. Silicon nitride is used as an alternative masking material for these elements.

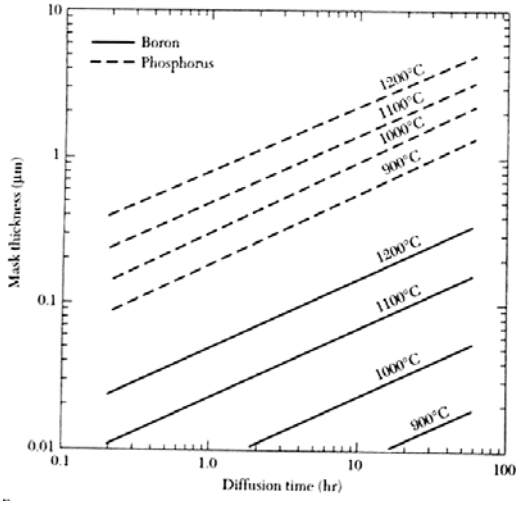


Fig. 5.8. Thickness of silicon dioxide needed to mask boron and phosphorus diffusions as a function of diffusion time and temperature.

Table 5.3. Diffusion constants in SiO_2

Dopants	Diffusion Constants at 1100 °C (cm^2/s)
B	3.4×10^{-17} to 2.0×10^{-14}
Ga	5.3×10^{-11}
P	2.9×10^{-16} to 2.0×10^{-13}
As	1.2×10^{-16} to 3.5×10^{-15}
Sb	9.9×10^{-17}

5.5. Silicon Oxide Quality

Oxides used for masking are usually grown by wet oxidation. A typical growth cycle consists of a dry–wet–dry sequence. Most of the growth in such a sequence occurs in the wet phase, since the SiO_2 growth rate is much higher when water is used as the oxidant. Dry oxidation, however, results in a higher quality oxide that is denser and has a higher breakdown voltage (5–10 MV/cm). It is for these reasons that the thin gate oxides in MOS devices are usually formed using dry oxidation.

It is well known that defects in noncrystalline SiO_2 films on silicon play an important role in determining the properties of $Si/Si0_2$ interface structures and, hence, of various silicon-based semiconductor devices.

MOS devices are also affected by charges in the oxide and traps at the SiO_2 –Si interface. The basic classification of these traps and charges, shown in Fig. 5.9, are interface-trapped charge, fixed-oxide charge, oxide-trapped charge, and mobile ionic charge.

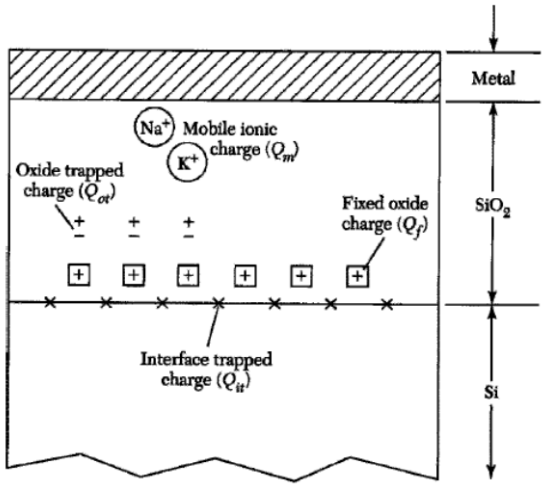


Fig. 5.9. Description of charges associated with thermal oxides [1].

Interface-trapped charges (Q_{it}) are due to the SiO_2 –Si interface properties and dependent on the chemical composition of this interface. The traps are located at the SiO_2 –Si interface with energy states in the silicon-forbidden bandgap. The interface trap density (i.e., number of interface traps per unit area and per eV) is orientation dependent. In silicon with a <100> crystal orientation, the interface trap density is about an order of magnitude smaller than that in the <111> orientation. Present-day MOS devices with thermally grown silicon dioxide on silicon have most of the interface trapped charges passivated by low-temperature (450°C) hydrogen annealing. The value of Q_{it} for <100> oriented silicon can be as low as 10^{10} cm^{-2}, which amounts to about one interface trapped charge per 10^5 surface atoms. For <111> oriented silicon, Q_{it} is about 10^{11} cm^{-2}.

The fixed charge (Q_f) is located within approximately 3 nm of the SiO_2 –Si interface. Generally, Q_f is positive and depends on oxidation and annealing conditions, as well as on the orientation of the silicon substrate. This charge is fixed and very difficult to charge or discharge. It has been suggested that when the oxidation is stopped, some ionic silicon is left near the interface. These ions, along with uncompleted silicon bonds (e.g., Si-Si or Si-O bonds) at the surface, may result in the positive interface charge. Q_f can be regarded as a charge sheet located at the SiO_2-Si interface. Typical fixed oxide charge densities for a carefully treated SiO_2-Si interface system are about 10^{10} cm^{-2} for <100> surface and about 5×10^{10} cm^{-2} for a <111> surface. Because of the lower values of Q_{it} and Q_f the <100> orientation is preferred for silicon MOSFETs.

Oxide-trapped charges (Q_{ot}) are associated with defects in the silicon dioxide. These charges can be created, for example, by X-ray radiation or high-energy electron bombardment. The traps are distributed inside the oxide layer. Most process-related Q_{ot} can be removed by low-temperature annealing.

Mobile ionic charges (Q_m), which result from contamination from sodium or other alkali ions, are mobile within the oxide under raised-temperatures (e.g., >100°C) and high-electric-field operations. Trace contamination by alkali metal ions may cause stability problems in semiconductor devices operated under high-bias and high-temperature conditions. Under these conditions mobile ionic charges can move back and forth through the oxide layer and cause threshold voltage shifts. Therefore, special attention must be paid to the elimination of mobile ions in device fabrication. For example, the effects of sodium contamination can be reduced by adding chlorine during oxidation. Chlorine immobilizes the sodium ions. A small amount (6% or less) of anhydrous HCl in the oxidizing gas can accomplish this, but the presence of chlorine during dry oxidation increases both the linear and parabolic rate constants, leading to a higher growth rate.

5.6. Silicon Oxide Structure

The basic structural unit of thermally grown silicon dioxide is a silicon atom surrounded tetrahedrally by four oxygen atoms, as illustrated1 in Fig. 5.10. The silicon-to-oxygen internuclear distance is 0.16 nm, and the oxygen-to-oxygen internuclear distance is 0.227 nm. These tetrahedra are joined together at their corners by oxygen bridges in a variety of ways to form the various phases or structures of silicon dioxide (also called silica). Silica has several crystalline structures (e.g., quartz) and an amorphous structure. When silicon is thermally oxidized, the silicon dioxide structure is amorphous. Typically amorphous silica has a density of 2.21 g/cm^3, compared with 2.65 g/cm^3 for quartz.

The basic difference between the crystalline and amorphous structures is that the former is a periodic structure, extending over many molecules, whereas the latter has no periodic structure at all. Figure 5.10, *b* is a two-dimensional schematic diagram of a quartz crystalline structure made up of rings with six silicon atoms. Figure 5.10, *c* is a two-dimensional schematic diagram of an amorphous structure for comparison. In the amorphous structure there is still a tendency to form characteristic rings with six silicon atoms. Note that the amorphous structure in Fig. 5.10, *c* is quite open because only 43% of the space is occupied by silicon dioxide molecules. The relatively open structure accounts for the lower density and allows a variety of impurities (such as sodium) to enter and diffuse readily through the silicon dioxide laver.

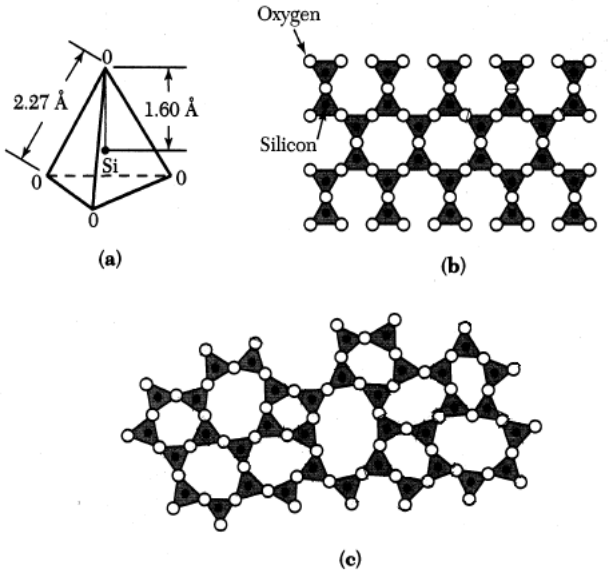


Fig. 5.10. (*a*) Basic structural unit of silicon dioxide. *(b)* Two dimensional representation of a quartz crystal lattice. *(c)* Two-dimensional representation of the amorphous structure of silicon dioxide.

The structural defects, particularly their electronic properties, are often described phenomenologically, i.e., without emphasizing their origin and relationship to the structure of noncrystalline SiO_2. Even when the immediate environment of some point defects was taken into account, for instance, in ESR studies, important features of the defect structure of the noncrystalline SiO_2 film (e.g., ordered regions termed channels and chemical interactions between defects) were usually not considered. A different approach was based on the localized properties of chemical bonds: It was suggested that a unique defect in noncrystalline SiO_2 film is the structural channel [7,8] and that network defects in vitreous SiO_2 are significantly different from point defects in crystals, particularly with respect to irradiation behavior. Numerous observations have been made which can be interpreted in a self-consistent manner within the framework of these ideas.

The structure of essentially all crystalline and non-crystalline polymorphs of SiO_2 is based on the tetrahedrical configuration of the oxygen atoms around the silicon: that is, the Si-O coordination is 4:2 as expressed by the $SiO_{4/2}$ formula. The exception is stishovite in which the silicon is surrounded by six oxygen atoms in an octahedral configuration. Stishovite can be written as $SiO_{6/3}$ indicating the 6:3 coordination. The properties of stishovite are drastically different from those of the 4:2 coordinated crystalline polymorphs and vitreous SiO_2; however, they illustrate in an extreme manner some trends exhibited by the polymorphs of 4:2 coordination.

Polymorphs based on $SiO_{4/2}$ tetrahedra.

One of the many remarkable features of SiO_2 is that there are nine polymorphs, including vitreous SiO_2, based on essentially identical $SiO_{4/2}$ tetrahedra; that is, the Si-O bond length varies only from 1.60 to 1.63 A° and the O-Si-O bond angle is ≈109° in all of them. The uniformity in the short-range order (SRO) is responsible for the very small difference (about 1%) in the standard free enthalpies of formation of the most stable and the least stable forms of silica , i.e., α-quartz and vitreous SiO_2 (the values are -197.2 kcal $mole^{-1}$ and -195.3 kcal $mole^{-1}$, respectively). Another manifestation of the same SRO is the fact that the optical properties of vitreous and crystalline SiO_2 are identical in the first approximation, whereas the behavior of most of the noncrystalline solids indicates some disorder (reduction in the average coordinate matrix as in amorphous Si or reduced second neighbour interaction as in amorphous Se).

In contrast, the crystal structure and density of SiO_2 polymorphs exhibit significant variation, for example, the densities of quartz and vitreous silica are 2.65 and 2.20 g cm^{-3} respectively (this corresponds to a change of 17%). Other properties also exhibit some variation, e.g., the bond overlap population changes by ≈3% from quartz to vitreous silica. The main cause of

the change in density is the variation in the packing arrangement (i.e. topology) of the SiO_2 tetrahedra. The Si-O-Si bond angle, which links the tetrahedra plays an important role m determining certain properties. This angle may exhibit a significant distribution even for some crystalline polymorphs, for instance, tridymite from 139.7° to 173.2° in the large unit cell comprising 320 $SiO_{4/2}$ units. The histogram of the Si-O-Si bond angle distribution is shown in Fig. 5.11. These observations indicate that the *variation among the $SiO_{4/2}$ polymorphs is essentially of conformational nature*, that is, significant changes in the long-range order (LRO) are associated with little change in the energy content. This fact has important implications for the structure and properties of noncrystalline (vitreous) SiO_2.

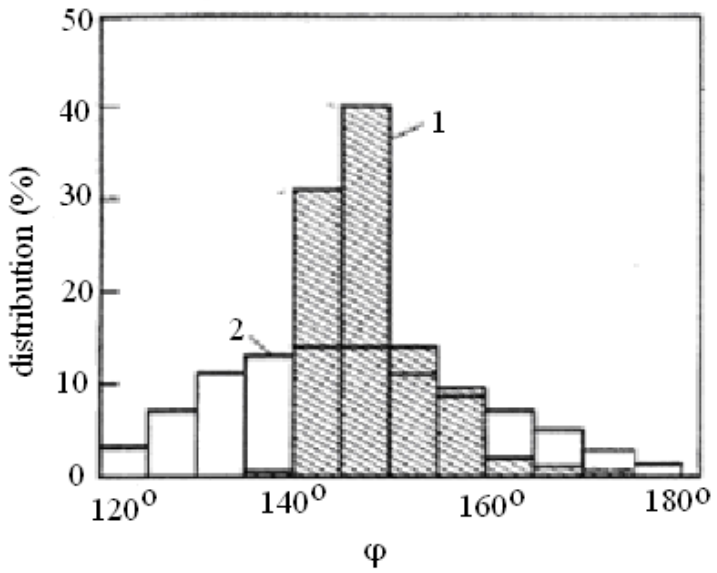


Fig. 5.11. Histogram of Si-O-Si bond angle distribution. (1) Tridymite (triclinic), (2) vitreous SiO_2.

Structure of noncrystalline SiO_2.

From the viewpoint of X-ray or electron diffraction analysis, noncrystalline SiO_2 films obtained by thermal or anodic oxidation of silicon are identical to fused vitreous silica. Therefore, the structure of fused silica will be considered first. The diffraction analysis of fused SiO_2 indicates that the shortest interatomic distances (i.e. the first Si-O, O-O, and Si-Si distances) correspond to those present in the crystalline phase, and that the Si-O-Si bond angles have a very wide distribution extending from 120° to 180°; the distribution curve peaks at 144°. According to a refined analysis, the mean value of the Si-O-Si bond angle is 152°. The histogram of the bond angle distribution in vitreous silica is shown in Fig. 5.11.

The results of the diffraction analysis are often interpreted in terms of random network model. A recent model based on -Si-O-Si-loops consisting of 4, 5, 6, 7, and 8 Si atoms resulted in 153° as the mean value of the Si-O-Si bond angle; therefore, it appears to be a realistic description

of the structure of vitreous SiO_2. It is clear from Fig. 5.11 that the range of the Si-O-Si bond angles is larger for noncrystalline SiO_2 than for tridymite, particularly in the region below 140°. The reason is that, in addition to the distorted 6-member loops in tridymite, vitreous silica contains 4- and 5-member loops. The Si-O-Si bond angle decreases as the loop size decreases, and puckering of the loops (i.e., deviation from nonplanaritv) also tends to decrease the bond angle. Despite the difference in the range of the Si-O-Si bond angle distribution, the essential structural feature of crystalline tridymite and vitreous silica is common: the Si-O-Si bond angles exhibit an unusually large variation.

A interpretation of the diffraction analysis of vitreous SiO_2 suggested that it consists of tridymite-like regions up to at least 20 A° in size, which are bonded together similarly to twinned crystals but in such a manner that isotropic properties would be present. It is important to realize that the size of these tridymite-like regions is smaller than that of the unit cell of the tridymite crystal (the *c*-axis is 82 A°); therefore, these regions cannot be considered as microcrystals. Rather, they represent a form of *structural ordering* in the sense that the bonding topologies of SiO_2 glass and tridymite show a distinct resemblance. In contrast, crystallographic ordering, which refers to the translational symmetry associated with the periodicity of the crystal is lacking in vitreous silica. From the mechanical and thermal properties of SiO_2 glass it has also been inferred that pre-ordered regions exist in silica, which exhibit some similarities to particular crystalline polymorphs. Similar conclusions were reached from the polarizability behavior of silica glass during pressure densification.

The flexibility of the SiO_2 structure, as manifested in the wide distribution of the Si-O-Si bond angles in noncrystalline SiO_2 and in crystalline tridymite, is the reason that despite the lack of LRO in noncrystalline SiO_2, the SRO is still very high and there is *no need to introduce defects* in significant density to obtain the noncrystalline structure. The model of vitreous SiO_2 based on 4- to 8-member loops does not contain broken Si-O bonds (i.e., nonbridging oxygens) except on the surface. The essentially identical SRO in vitreous and crystalline SiO_2 as well as the lack of defects in vitreous SiO_2 are the reasons that the configurational entropy of vitreous silica is very low, ≈ 0.9 cal K^{-1}. For these reasons, fused SiO_2 and thermally or anodically grown SiO_2 films are considered *vitreous* rather than amorphous.

Despite the structural similarity between fused silica and thermally or anodically grown SiO_2 films on silicon, there is a significant difference in their crystallization behavior. The crystallization product of fused silica is practically always cristobalite, whereas thermally grown SiO_2 films on silicon may crystallize to cristobalite, quartz, or tridymite. Anodic oxidation of silicon may even result in a thin single crystal quartz film beneath the noncrystalline oxide. The fact that

the unique crystallization behavior of SiO_2 films on silicon is very likely related to the presence of ordered regions (channels).

π-Bonding in SiO_2.

The structure and chemistry of siloxane (silicone) polymers, whose framework consists of Si-O-Si groups, is usually discussed in terms of the properties of the Si-O bond. Similar considerations have also been applied to the structural chemistry of SiO_2 poly morphs and silicates. The concept of $d\pi$-$p\pi$ bonding arising from the overlap between the 2*p* orbitals of the oxygen containing the lone pair electrons and the 3*d* orbitals of silicon plays an important role in these considerations. An important feature of the Si-O bond is that π-bonding increases and the ionicity of the Si-O bond decreases with the Si-O-Si bond angle. Also, π-bonding can be influenced by other bonds, e.g., OH and O^-Na^+. The concept of π-bonding was also employed to explain the bond polarizability and other properties of SiO_2 polymorphs.

In contrast, the optical spectra of SiO_2 can usually be adequately interpreted without considering the Si 3d orbitals because, based on siloxane chemistry, $d\pi$-$p\pi$ bonding does not have an appreciable effect on the bond energy; however, it has a great effect on structure and chemical behavior. This is demonstrated by the small (0.14%) difference in the total energy of H_4SiO_4 calculated with and without involving the Si 3d orbitals. In contrast, the difference in molecular orbital energies is much larger for those orbitals which are affected by including the Si 3d orbitals than for those which are not affected by these orbitals. This is particularly true for that orbital which corresponds to π-bonding, that is, the orbital energy is reduced by 48% when the Si 3d orbitals are considered. The total population of electrons in the molecular orbitals which involve the Si 3d orbitals is 1.43 (0.36 per Si-O bond). These considerations of the H_4SiO_4 molecule are very relevant to SiO_2 since the X-ray fluorescense spectrum of silica could be interpreted on this basis. The 92.3 and 94.5 eV peaks in the $L_{2,3}$ spectrum are products of transitions (to Si 2*p*) from orbitals winch contain significant 3d contributions. The X-ray spectra of silicate and SiO_2 crystals have been also interpreted in terms of Si 3*d* admixture into the oxygen lone pair bands. It was pointed out that, although this admixture, is relatively small, its effect can be decisive in the structure and chemistry of SiO_2.

This effect π-bonding arises from the fact that the Si-O bond is mixed ionic- covalent. The σ component of the covalent bond (overlap between Si $3sp^3$ hybrid and O 2*p* orbitals) is responsible for the almost invariant nature of the $SiO_{4/2}$ in SiO_2 polymorphs. According to molecular orbital calculations the overlap between oxygen orbitals, i.e., the bond overlap population, expressed as *n*(Si-O), increases with increasing Si-O-Si bond angle (see Fig. 5.12).

The Si-O bond overlap population increases even when the Si 3d orbitals are not included in the calculations. However, the confidence level of the correlation between Si-O bond length and *n*(Si-O) is better when the Si 3*d* orbitals are included, and with respect to *n*(Si-O), the Si-O_{nb} conation is separated from the Si-O-Si configuration. This separation reflects the large difference in the chemical behavior of bridging and non-bridging oxygens. This overlap population is higher than the value of ≈0.36 per bond for H_4SiO_4 because the hydrogen in the OH group acts as an electron acceptor relative to silicon and, hence, decreases the π–component of the Si-O bond. These *n*(Si-O) values represent trends rather than absolute numbers and, hence they should be only considered as means for grouping and classifying. Concomitant with the increase in π-bonding, the ionicity of the bond decreases with angle so that the overall bond strength remains approximatety unchanged. Consequently, many properties of SiO_2 polymorphs, such as optical and X-ray spectra, as well as the free enthalpy of formation, are practically structure invariant. In contrast, several important properties (e.g., bond polarizability, infrared absorption and etching), depend on the extent of π-bonding and, hence, display a trend as a function of the Si-O-Si bond angle as shown in Fig. 5.12.

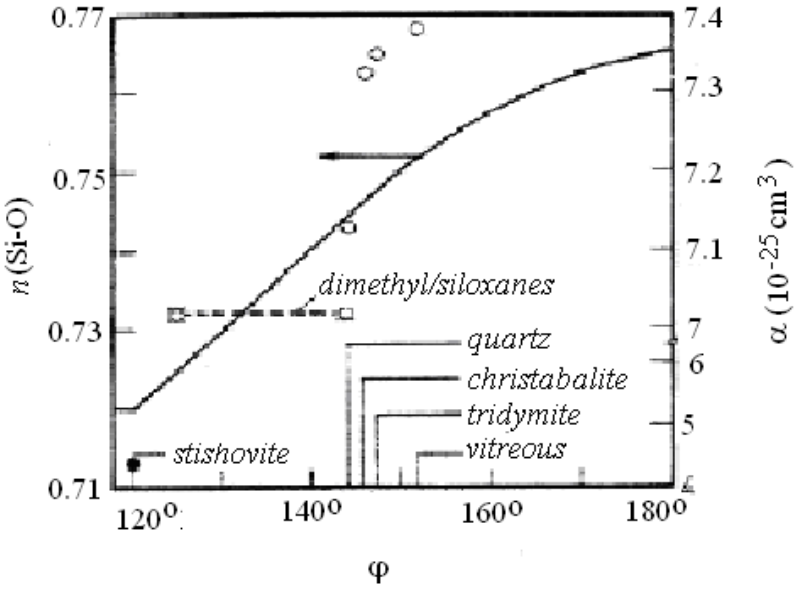


Fig. 5.12. Various properties of the Si-O bond as a function of the Si-O-Si bond angle. The solid line is the bond overlap population, *n*(Si-O) calculated from the Si(*spd*) basis. This curve does not hold for stishovite because the coordination of oxygen is three; i.e., it is not a bridging oxygen. The full and open symbols represent the Si-O bond polarizability, α, and refer to 6:3 and 4:2 Si-O coordinations, respectively. The bond angles representing tridymite and vitreous silica are the mean values of the corresponding angle distributions. Note the break in the α-scale.

Ionicity of the Si-O bond.

Figure 5.12 shows that the bond polarizability of stishovite, 4.32×10^{-25} cm^3 is much smaller than that of the other SiO_2 polymorphs and siloxanes ≈7.0 to ≈7.4×10^{-25} cm^3. Also, the Si-O stretching vibration frequency (not shown in Fig. 5.12) is much less for stishovite than for the other polymorphs, the respective values being 885 cm^{-1} and 1077 to 1106 cm^{-1}. Furthermore, the Si-O bond length is larger in stishovite (1 77 A°) than in $Si0_{4/2}$ polymorphs (1.59 to 1.63 A°). These observations indicate that Si-O bond is more ionic in stishovite than in the $SiO_{4/2}$ polymorphs. This conclusion is further strengthened by arguments based on the electronic dielectric properties The value of *f*, a coefficient associated with first and second neighbour delocalization, is 3 8 for stishovite; this value is characteristic of ionic crystals in which the anions are in close contact with each other as in rutile (TiO_2). In contrast, the *f* value for $SiO_{4/2}$ polymorphs are in the range of 5.0 to 5.1, which is typical of covalent solids The conclusion that the Si-O bond is more ionic in stishovite than m the $SiO_{4/2}$ polymorphs is contrary to the ionicity values 0.41 for stishovite and 0 53 to 0.65 for the $Si0_{4/2}$ polymorphs. The implications of the difference between the two interpretations is discussed below. Another great difference in the character of the Si-O bond in stishovite and $Si0_{4/2}$ polymorphs is the lack of π-bonding in the former: this is partially responsible for the fact that stishovite is soluble in H_2O but insoluble in HF, while the $SiO_{4/2}$ polymorphs behave in the opposite manner.

Since the SRO in the various $SiO_{4/2}$ polymorphs and siloxane polymers is essentially the same, the variation in their properties is not as great as the difference between $SiO_{4/2}$ polymorphs and stishovite. Nevertheless, this difference and the trend shown in Fig. 5.12, together with the increase in the ratio of the force constants of the Si-O-Si bending and Si-O stretching vibration with increasing Si-O-Si bond angle, demonstrate that *the ionicity of the Si-O bond decreases as the Si-O-Si angle increases.*

This conclusion is at variance with recent suggestions concerning the ionicity of the Si-O bond. The difference between the two interpretations, which is very pertinent to the chemical behavior of defects and other properties of SiO_2 films, is briefly discussed. One of the important conclusions is that the ionicity, F_i increases and the covalent character $(1 - F_i)$ of the Si-O bond decreases with increasing Si-O-Si bond angle, φ. This conclusion was reached on the basis of a relationship between the electronic dielectric constant, $\varepsilon(0)$, and F_i: $\varepsilon(0) = 1 + 1.26d^2(1 - F_i)$, where *d* is the Si-O bond length in angstroms. The fact that this relationship is incorrect in this case is clearly demonstrated by the wrong order of ionicity values for stishovite and $SiO_{4/2}$ polymorphs as discussed above. This gross discrepancy exists because the dielectric ionicity concept which underlies this relationship cannot be considered as a quantitative guide to the correlation between bond and dielectric properties;

the first-neighbour delocalization (covalency) and second-neighbour delocalization (anion-anion contact) as expressed by the structural parameter, *f*.

This point can be further illustrated by the internal inconsistency of using the above relationship and another closely related equation for the ionicity,

$$F_i = \frac{C^2}{E_h^2 + C^2} \tag{5.16}$$

where

$$E_h^2 + C^2 = E_g^2 \tag{5.17}$$

In this equation, E_g is the average bandgap, C and E_h are its ionic (coulombic) and covalent components, respectively. The values of F_i are 0.57 for quartz and 0.65 for vitreous $Si0_2$, representing a change of 14%. According to (5.16), the corresponding change in the value of $[C^2/(C^2 + E_h^2)]^{1/2}$ should be ≈4%. However, both the experimental and theoretical studies indicate that the optical bandgap of the $Si0_{4/2}$ polymorphs is essentially structure-invariant.

The F_l value of stishovite is ~ 30% less than that of quartz. Hence, the optical bandgap of stishovite should be significantly less than that of quartz. However, based on the similarity in the packing of oxygen atoms in rutile and stishovite, it can be easily shown that the optical bandgap of stishovite is essentially identical to that of quartz; the refractive index of stishovite calculated by this assumption (1.82) is practically identical to the experimental value (1.83). The 14-% variation in F_i, between quartz and vitreous SiO_2 is much larger than the corresponding change in the bond overlap population, ≈3% (Fig. 2) and is in the opposite direction. This discrepancy exists for the same reason as the discrepancy concerning the ionicity of stishovite discussed above.

The ionicity, F_i was represented as a monotonically increasing function of φ. Subsequently, this relationship was modified on the basis of a qualitative argument involving repulsion between adjacent Si atoms; the pertinent curves are shown in Fig. 5.13. However, the points in Fig. 5.13 clearly demonstrate that the F_i values simply exhibit a random scatter, particularly if the correct values for d and φ are used. Hence, the ionicity of the Si-O bond increases and the covalency decreases with the Si-O-Si bond angle is unjustified, even if the theoretical considerations were correct (however, they are not). Thus, not withstanding the rejection of the conceptual framework based on π-bonding as an "unclear point, the properties of SiO_2 polymorphs can be interpreted in a consistent manner on this basis. This was apparently recognized.

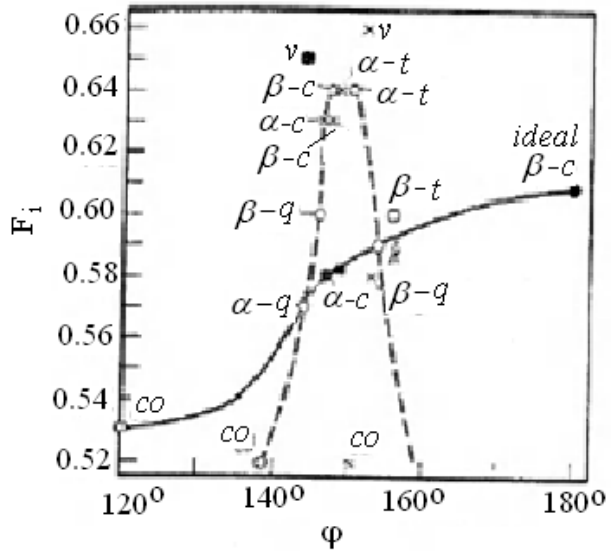


Fig. 5.13. Ionicity, Fi, of the Si-O bond as a function of the Si-O-Si bond angle φ. The abbreviations *v, t, c, q, k,* and *co* represent vitreous, tridymite, cristobalite, quartz, keatite, and coesite, respectively. Ideal β-cristobalite refers to a *hypothetical* structure with $\varphi = 180°$ The crosses represent Fi values calculated using up-to-data values for the Si-O bond length and plotted against up-to-date φ. Average *d* and φ values were used in these calculations for vitreous SiO_2, α and β tridymite, and coesite.

Si-O bond and defect structure of vitreous SiO_2.

With respect to the defect structure of vitreous SiO_2, an important effect is the decrease of the Si-O bond length with increasing Si-O bond overlap, i.e., with increasing Si-O bond angle as shown in Fig. 5.14. This means that the bond/structural flexibility of vitreous SiO_2 arising from the wide distribution of the Si-O-Si bond angles is further increased. Figure 5.14 also shows that for a given bond length the bond overlap is larger for the Si-O_{nb} (O_{nb}= non-bridging oxygen) than the Si-O-Si configuration (containing a bridging oxygen) As will be discussed below, this increased bond overlap in the Si-O_{nb} configuration has significant consequences with respect to network defects in vitreous SiO_2.

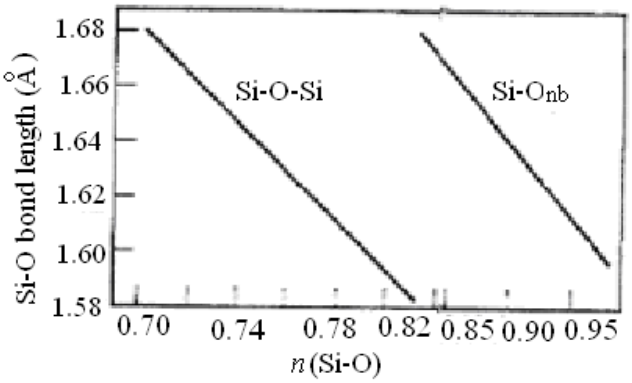


Fig. 5.14. Relationship between Si-O bond length and bond overlap population. Note the break in the abscissa.

Based on the behavior of bond polarizability, infrared spectra, etc., it has been suggested that vitreous SiO_2 has the highest π-bond order among SiO_2 polymorphs, excluding the unstable melanophlogite (see Fig. 5.12). This idea has been strengthened by a model of the α-β phase transformation of quartz crystal. As the temperature increases from 25 to 600 °C, the Si-O-Si bond angle increases from 144° to 153° and the Si-O bond length decreases from 1.607 to 1.59 A°; this has been interpreted as an increase in π-bonding with temperature. It is well known from crystallography that the high temperature polymorph is always characterized by increased symmetry relative to the low temperature polymorph. Vitreous SiO_2, which is isotropic, represents the highest symmetry among SiO_2 polymorphs and, hence, is expected to have the highest π-bond order.

Another feature of the α-β phase transformation is that increasing π-bond order tends to pull the oxygen atoms to the plane perpendicular to the c-axis; this explains the negative thermal expansion coefficient in the *c*-direction in β-quartz. The atomic arrangement along the *c*-axis is such that Si and O atoms form a helix along the *c*-axis where each loop consists of six Si atoms. This rearrangement of the oxygen atoms indicates that π-bonding increases preferentially along the axis of the structural channel as the temperature increases.

A similar phenomenon is that those five Si-O-Si bond angles in tridymite, which are larger than 160°, are between layers (consisting of Si-O-Si loops) rather than within a layer. The layers are connected in such a manner that structural channels are present along the c-axis. These observations demonstrate that structural channels in crystalline SiO_2 polymorphs represent increased π-bonding between the Si and O atoms forming these channels. Thus, the earlier suggestion that channel defects in vitreous SiO_2 are characterized by increased π-bonding appears to be substantiated.

5.7. Oxidation of Polycrystalline Silicon

In most integrated circuits, the polysilicon is electrically isolated from overlying conductors (either metal or additional layers of polysilicon) by silicon dioxide, which may be formed by thermal oxidation of the polysilicon or by chemical vapor deposition. In most integrated circuits these oxide layers must simply be highly insulating. However, specialized devices, such as the electrically erasable, programmable read-only memories (EEPROMs) used with increasing frequency in VLSI circuits require a thin oxide with well-controlled conductivity above the polysilicon.

Numerous studies have shown that the oxidation rate of polysilicon can differ substantially from that of single-crystal silicon and also that the electrical properties of the oxide grown on polysilicon are different from those of a similar thickness of oxide grown on single-crystal silicon. In this section, the differences in the oxidation rates of polysilicon and single-crystal silicon will be first examined and these differences will be related to the structure of the polysilicon. In some applications the differences in the oxidation rate can be used constructively, while in other applications, they complicate integrated-circuit fabrication.

5.7.1. Oxide Growth on Polysilicon

Undoped Films. Substantial differences can exist between the thickness of oxide grown on polysilicon and that of oxide simultaneously grown on single-crystal silicon. For lightly doped films one of the dominant factors leading to this difference is the presence of grains with different orientations in the polysilicon. To demonstrate the importance of differently oriented grains, in one study [13] thick layers of polysilicon were polished so that the resulting smooth surface contained regions with different crystal orientations. When oxidized, each differently oriented grain oxidizes at a rate characteristic of that particular orientation of crystalline silicon. Grains with a given orientation exhibit the same oxide color as seen on the similarly oriented, single-crystal control wafers. These differences are accentuated under surface-reaction-limited oxidation conditions and reduced when the oxidation is performed under diffusion-limited conditions. Although polishing the samples removed the faceted surface structure which causes the exposed crystal planes to differ from the grain orientation, the study did show that the macroscopically measured oxide thickness can be expected to be a suitable average of the oxide thicknesses grown on differently oriented grains.

A similar trend is seen in oxides grown on the fine grain polysilicon typically used in integrated circuits. The oxide thicknesses grown on thin polysilicon films deposited either in an atmospheric-pressure reactor at 960°C or in a low-pressure reactor at 625°C are between the oxide thicknesses grown on rapidly oxidizing (111)-oriented, single-crystal silicon and slowly oxidizing (100)-oriented silicon [6]. Because the oxide thickness grown depends on the dominant crystal orientations in the film being oxidized, it can differ significantly for polysilicon deposited under different conditions, necessitating process modification when the polysilicon deposition conditions change significantly. The crystalline texture of polysilicon films depends strongly on the deposition temperature and the oxide thickness should vary correspondingly.

Heavily Doped Films. Although oxidation of lightly doped polysilicon is similar to that of lightly doped single crystal silicon, marked differences are seen between the oxidation of heavily

doped polysilicon and similarly doped single-crystal silicon. Heavily doped single-crystal silicon oxidizes much more rapidly than does lightly doped single-crystal silicon because of the excess point defects present in the heavily doped material. **As** the influence of the added point defects begins to dominate the oxidation process, the differences between the oxidation rates of the different orientations of silicon decrease significantly. However, the oxidation-rate enhancement on polysilicon is usually much less than that on single-crystal silicon doped at the same time [9].

Figure 5.15 [9] shows the thickness of oxide grown on two different orientations of single-crystal silicon and on polysilicon deposited under two different deposition conditions. Varying amounts of phosphorus were added by gaseous diffusion from a $POC1_3$ source to sets of samples each containing all four types of material, and then all samples were simultaneously oxidized under surface-reaction-limited conditions. The resulting oxide thicknesses are shown as functions of the sheet resistance measured on the (100)-oriented single-crystal silicon contained in each set of samples. As the dopant concentration increases (decreasing sheet resistance), the oxide thickness grown on the polysilicon becomes a smaller fraction of that grown on single-crystal silicon, even though the amount of dopant added to the polysilicon is expected to be at least as great as that added to the single-crystal silicon.

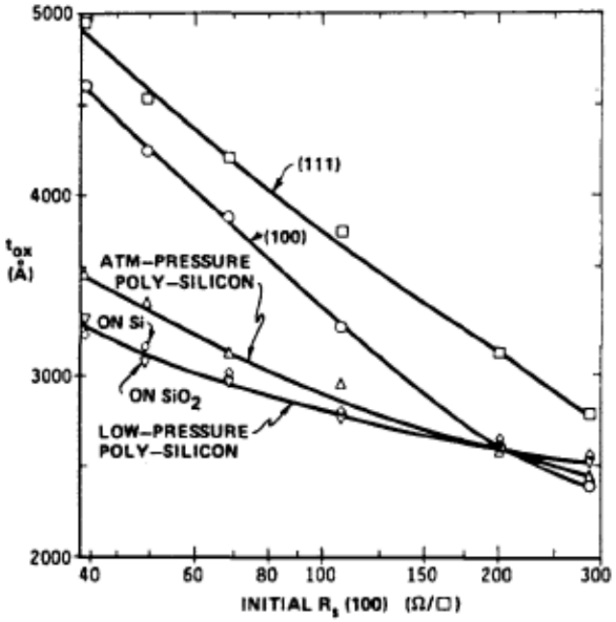


Fig. 5.15. Oxide thickness grown on phosphorus-doped single-crystal and polycrystalline silicon during a 150 min, 850°C, pyrogenic steam oxidation as a function of the sheet resistance R_S measured on the (100)-oriented, single-crystal silicon wafer in each set.

This apparently anomalous behavior is best understood by considering the interaction of diffusion and oxidation. The impurities diffuse much more rapidly in polysilicon than in single-

crystal silicon [10]. During doping and oxidation, therefore, the phosphorus added near the surface can diffuse toward the back of the silicon film more readily in polysilicon than in single-crystal silicon. The surface concentration is lower; and, consequently, a thinner oxide is grown. The differences between the oxide thicknesses grown on the different types of polysilicon depend on the ease of diffusion in each sample, which is governed by the detailed grain structure.

An alternate explanation suggests that the difference between the oxide thicknesses grown on polysilicon and on single-crystal silicon is dominated by the different electrical activity of the dopant in each type of material. Because the dopant is less active in polysilicon, the Fermi level is closer to the intrinsic Fermi level during oxidation; the charged point defects which enhance the oxidation are less numerous; and the oxide grown is thinner. The different behavior of the point defects in single-crystal silicon and in polysilicon may also affect the oxidation rate. Grain boundaries and other structural defects in polysilicon can act as recombination sites for point defects, increasing their concentration gradients near the surface. If the silicon interstitials injected by the oxidation process can diffuse away from the oxidizing surface more readily, the oxidation rate should increase, rather than decrease.

Differences in the dopant-diffusion rates between polysilicon and single-crystal silicon can also explain the *more rapid* oxidation of heavily doped polysilicon than single-crystal silicon sometimes seen. We can understand this behavior by considering the finite thickness of the polysilicon. Table 5.4 [9] shows that the oxide thickness grown on heavily doped polysilicon also depends on the thickness of the polysilicon film and is greater on thinner films than on thicker films. In thick films, the dopant can readily diffuse away from the polysilicon surface, reducing the surface dopant concentration and the oxidation rate, as discussed above. In thinner polysilicon films, however, the finite thickness confines the dopant atoms; as dopant atoms approach the back surface, the increasing concentration of dopant there decreases the concentration gradient driving the diffusion. The surface concentration remains high, and the oxide grown is, consequently, thicker.

Table 5.4. Oxide thicknesses grown on n^+ polysilicon films of different thicknesses during a 75 min, 850°C, pyrogenic steam oxidation.

Polysilicon thickness (μm)	Oxide thickness (nm)	
	LP	AP
0.5	393	378
1.0	280	301
1.5	276	310
n^+ (100) single crystal	515	

The increase in oxide thickness with increasing dopant concentration does not continue indefinitely, however. It saturates at a value approximately corresponding to the solid solubility of phosphorus in silicon at the oxidation temperature. As shown in Fig. 5.16 [11], when oxidation occurs at 750°C, the oxide growth rate remains constant for phosphorus chemical concentrations greater than about 10^{21} cm^{-3}, which is close to the solid solubility of phosphorus in silicon at this temperature, especially if segregation of the dopant at grain boundaries is considered. Thus, the oxide thickness depends on the phosphorus concentration from the concentration at which the Fermi level at the oxidation temperature departs from the intrinsic Fermi level up to the concentration corresponding to the solid solubility of the dopant in silicon.

The considerably greater oxide thickness grown on heavily doped polysilicon than on lightly doped, single-crystal silicon can be used advantageously in the fabrication of many integrated circuits. In a typical silicon-gate integrated circuit, the heavily doped, *n*-type polysilicon gate is adjacent to the lightly doped single-crystal silicon regions which subsequently form the source and drain of an MOS transistor. After definition of the polysilicon, however, these single-crystal regions are still lightly doped. Oxidation under suitable conditions produces a thick oxide on the polysilicon, while only a thin oxide is grown on the single-crystal silicon. The thick oxide on the polysilicon can serve as an implant mask during processing, or it can be used to reduce capacitance in the finished circuit. The differential oxidation rate is also useful in reducing critical mask alignment; the thin oxide on the single-crystal silicon can be removed while not exposing the heavily doped polysilicon, which is covered by a much thicker oxide.

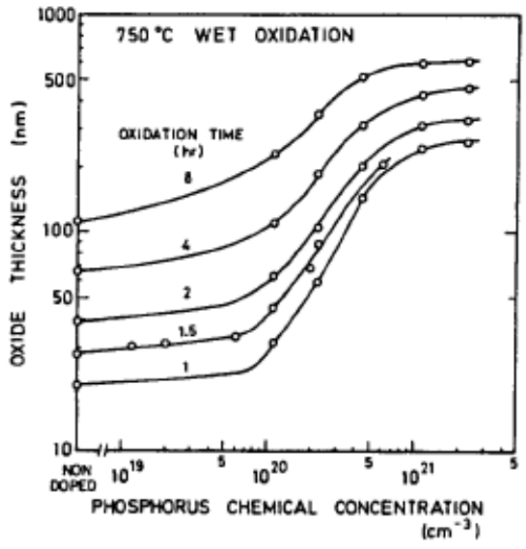


Fig. 5.16. Oxide thickness grown on heavily phosphorus doped polysilicon during a 750°C, wet-oxygen oxidation [11].

Grain-Boundary Oxidation. The disordered structure near the grain boundaries might be expected to oxidize more rapidly than the crystalline structure near the center of the grains. In the study of polished samples of undoped, large grain polysilicon, deposited at a high temperature, careful examination of the grain-boundary regions did not reveal any enhanced oxidation at the grain boundaries. Because grain boundaries occupy a larger fraction of fine-grain polysilicon, grain-boundary effects might be expected to play a more important role in the oxidation of fine-grain material, especially in films deposited at lower temperatures, which might be expected to be less ordered. If the grain boundaries oxidize rapidly, the oxide formed there should cause considerable compressive stress in the oxidized film. However, no tendency toward stress is caused by oxidation of *undoped* films [12]. In fact, a slight increase of tensile stress is found, suggesting that the heat treatment orders the structure near the grain boundaries. On the other hand, in phosphorus-doped films oxidation causes compressive stress [12]. In addition, examination of the local film thickness near the grain boundaries as oxidation proceeds shows that the entire thickness of the polysilicon film is consumed first at the grain boundaries while unoxidized silicon still remains near the centers of the grains [12]. Although differences in the thickness of the polysilicon film near the center of the grains and near the grain boundaries could also cause the polysilicon to be completely consumed by oxidation near the grain boundaries first, these observations suggest that the region near the grain boundaries oxidizes more rapidly than does the silicon away from the grain boundaries. Because the grain boundaries are the lowest portions of the polysilicon surface before oxidation and they oxidize most rapidly, the surface roughness of the polysilicon is expected to increase as oxidation proceeds.

Other studies have confirmed the more rapid oxidation near grain boundaries. High resolution, cross-section transmission electron microscopy shows that Si-P precipitates can form at grain boundaries when the phosphorus concentration exceeds its solid solubility at the oxidation temperature [13]. Because this phase oxidizes more rapidly than does silicon, the oxide near the grain boundaries can be considerably thicker than that over the centers of the grains. Consider a polysilicon film doped to solid solubility at an intermediate temperature. When it is oxidized at a higher temperature, the phosphorus concentration is below solid solubility at the oxidation temperature, and no enhanced oxidation is expected at the grain boundaries. At lower oxidation temperatures the phosphorus concentration is greater than solid solubility; excess phosphorus concentrates at the grain boundaries; and the grain-boundary regions oxidize more rapidly than do the centers of the grains. In addition to forming more rapidly, the oxide grown over grain-boundary precipitates appears to contain a high phosphorus concentration and, therefore, etches more rapidly than does SiO_2. As the oxide grown on polysilicon is etched, narrow grooves can be left in the

polysilicon surface. In the extreme case, the indentation can extend through the entire thickness of the polysilicon film so that the underlying oxide (e.g., the gate oxide) can be attacked.

5.7.2. Oxide-Thickness Evaluation

Although measuring the oxide thickness on single-crystal silicon nondestructively by optical techniques is straight forward using an ellipsometer or a spectrophotometer, measuring the oxide thickness grown on polysilicon is more complex because of the multilayer structure on which the oxide is grown. At the wavelengths typically used for oxide-thickness measurements (e.g. λ = 628 nm for ellipsometry or λ = 400-800 nm for spectrophotometry), polysilicon is transparent, and the reflected signal being analyzed is influenced by reflections from the underlying interfaces, as well as from the top and bottom interfaces of the oxide layer grown on the polysilicon. In theory, the reflectance of the total multilayer structure can be analyzed, and the oxide thickness of the polysilicon can be extracted. However, in most cases variations in the thicknesses of the underlying layers make the indicated thickness of the oxide grown on the polysilicon layer too uncertain for practical use.

The optical techniques can, however, be adapted to the ultraviolet wavelength range, in which polysilicon is opaque. In this case, the reflected signal is dominated by interference at the top and bottom of the oxide layer grown on the polysilicon. The wavelength range from 200 to 400 nm is suitable for this type of measurement. However, near the lower end of this wavelength range, the surface roughness of the polysilicon also affects the reflected signal, and at λ =280 and 370 nm, structural bands of crystalline silicon can influence the reflected signal. The interference of light reflected from the two surfaces of the top oxide is strong, however, and the technique is a useful, nondestructive, method of measuring the oxide thickness grown on polysilicon. Figure 5.17 shows the interference signal obtained over the wavelength range from 200 to 500 nm, with interference in the top oxide below about 400 nm and interference in the multilayer structure at higher wavelengths. Ellipsometry using ultraviolet light can also be employed to determine the oxide thickness, again taking advantage of the fact that silicon is opaque to ultraviolet light.

Of course, the oxide thickness can be determined destructively by etching a step and measuring the step height with a surface profilometer. Although this technique is the most straightforward, it can be more time consuming than the optical techniques, its resolution is limited by the surface roughness of the polysilicon, and it cannot readily be used on device wafers.

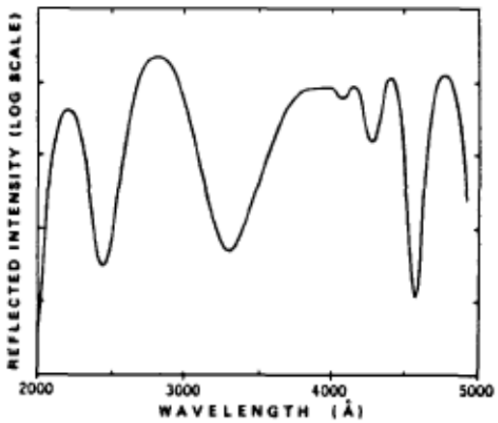


Fig. 5.17. The reflectance in the wavelength range from 200 to 400 nm arises from interference in the oxide above the polysilicon, while the signal at longer wavelengths is influenced by reflection at interfaces beneath the polysilicon as well.

5.8. Conclusions

Silicon dioxide is a high-quality insulator that can be thermally grown on silicon wafers. It can also serve as a barrier layer during impurity diffusion or implantation, and it is a key component of MOS devices and circuits. These factors have contributed significantly to silicon's current status as the dominant semiconductor material in use today.

This chapter described the mechanism of thermal oxidation of silicon and presented a kinetic model of oxide growth. This model accurately predicts oxide growth rate for a wide range of process conditions. The chapter also discussed dopant redistribution and the masking properties of oxides. Oxide characterization methods and oxide quality were discussed as well.

Chapter 6. Microlithography

V.Verbitsky

6.1 Properties of microlithography

Microlithography - technological process of surface creating of the mask or functional layer relief image of one topological layer of integrated circuits, in scale 1: 1 in a special lamina resist using photon, electron, X-ray or ion beams. Information about the picture layer topology can be stored in the form template or electronically in a computer.

In the technology of semiconductor integrated circuits this image layers are mostly temporary resist contact mask on the surface oxide or silicon nitride which are grown or deposited on semiconductor wafers. With operations of remove material image the topological layer in resist lamina is transferred to the oxide, nitride, or other heat-resistant film and get a mask that applies to the local processing plates (doping, removal, oxidation, etc.). Resist topological layer masks are also used for the direct creation of functional elements of topological layers of low-temperature circuit's technological operations: ion implantation, ion or ion-plasma etching, and others.

In the technology of hybrid integrated circuits (HIC) and microassemblies (MA) the high-temperature processes are not used because resist mask on the surface of the functional layer is usually final and it is used for shaping element via removal disadvantaged regions by chemical, ion, ion-plasma etching, etc. Modern technological methods of functional layers of microelectronic structures make it possible to create them in the direction perpendicular to the plane of the plate to within one monatomic layer. To form with such precision drawing element in a plane that is in the other two dimensions, much difficult.Powered by lithography provides that the first surface of the plate is applied a thin layer of resist sensitive to certain types of radiation. After that via exposure through group or local pattern exhibiting certain locations the latent image of topological layer is created in resist by chemical etching.

Improving lithography is the basis for reducing the size of transistors and increasing the degree of integration of chips. It used to be that the limit of optical lithography capabilities is 1 micron, but today manufacturers of integrated circuits have mastered the minimum topological size 0.13 microns, and advanced - 60 nm or less. Lithography progress in recent years has exceeded most predictions. During the research it was found that doubling the number of elements in integrated circuits par one year is mainly due to increased resolution lithography.

Now the size of the transistors in the plane of the plate is determined not so much performance as the existing level of productive capacity, mainly lithographic equipment with which creates a picture of transistors and conductors on the surface of the crystal. The main factors that determine the minimum size of topological planar circuits, are the wavelength of the beam used for lithography, precision patterns and electro-mechanical equipment, which possesses both, tread patterns on the surface of the plate.

Lithography characterizing the minimal line width compared to the wavelength of the beam, which is used for exposition. Reducing the size of the elements is made possible through lithography due to reduction of the wavelength of the beam, which is used for exposition, from 435 to 157 nm (excimer F-lasers). Today, the 90 th 60 th nanometer technology using photolithography with a wavelength of 193 nm. The technological process of topological rules for exhibiting 45 nm using an excimer F-lasers with a wavelength of 157 nm.

Creating lines with a width less than the wavelength of the radiation source is complicated by the diffraction of light. The dimensions smaller than the wavelength can be obtained by a variety of special techniques, such as off-axis exposure, masking of the phase shift and so on. To achieve high-resolution lithography is also necessary to create new photosensitive resists.

Non-optical lithography technique is also used for shaping. In particular, small items make it possible to create electron-beam lithography, as the wavelength of electrons is only about 0.01 nm. Electron beam lithography has long been used to make patterns and slow exposure. However, fabrication of complex circuits using electron-beam lithography requires much greater speed exposure. To achieve using thermal electron beams with a large area of overlap that pass through electronic templates diminutive and electron-optical lens.

Leading chip manufacturers for forming topological elements with sizes from 1 to 0.06 microns use the contact X-ray lithography. The wavelength of 1.1 nm separated from the spectrum of synchrotron radiation. The main problem is that no lenses and mirrors are for the beam with such small wavelength. Therefore, the use of templates patterned circuit in 1:1 scale. Making such patterns without distortion is extremely time-consuming and expensive process. Other problems is the need to dense placement template to the plate (10 nm or closer) and the presence of diffraction effects arising in connection with this.

In projection ultraviolet lithography the problem of scale pattern is solved using the radiation with a wavelength of 11- 13 nm. At this wavelength it is possible to use the focus system with quadruple reduction. However, this system requires concave mirror, which is made using 40 layers of special film with thickness of 2.3 nm. This is used for lithography process with minimal topological size 32 and 13.5 nm.

Now the intensive research of ion-beam lithography and lithography based on the emission of hot electrons are conducted. Ion-beam lithography allows carrying out the local doping impurities with resolution for planar structures up to 10 nm.

6.2. Types of lithography

Depending on the type of resist and type of radiation used for exposure, it is considered optical lithography (photolithography), electron beam and ion-beam or X-ray.

The *photolithography* is the most convenient in the technology of IC. It is based on the use of light-sensitive resist materials - *photoresist*, which can be *positive* and *negative.* Under the influence of light the negative photoresists are polymerized and become insoluble in the developer. After exposure through a photomask the resist film is chemical etched (developed). During film development the area of negative photoresist, protected by an opaque film on the photomask is removed. In a positive photoresist the radiation breaks the polymer chains, resulting in the irradiated area will be removed. Photoresists should have good light sensitivity, high optical resolution, stability to acid and high adhesion. Most photoresisters are two-component system, consisting of a polymer base (protector) and photosensitive component. To ensure the required viscosity on the plate while drawing a photoresist solvent is added.

The figure of topological layer set the masks which are the optical contrast image matrix of one topological layer of chips at a scale of 1:1. Photomask is a glass plate on one side of which the wear-resistant non-transparent film (Cr, CrO_3, Fe_2O) are formed images of topological layer. In the manufacture of chips the photolithography perform up to 25 times, which use a set of photomasks.

Process of making photomasks starts from machine design circuits in general topology and topological each layer separately. Figure topological layer made, for example, imagesetter equipment by-element on the photosensitive printing plate in the correct location of simple geometric shapes that are set image generator. Then, using the photoreduction camera its linear dimensions are reduced to the required. The working photomask are made using the photorepeaters. This usually matrix images of a topological layer is in 1:1 scale. Combining topological layers perform with special shapes that print on each photomask.

In *electron-beam lithography* (EBL) the electron-resists sensitive to electron beams is used. The EBL process is similar to conventional process of photolitography. In electron- and ion-beam exposure system the topology information is stored in the picture memory of electronic computer and immediately used it to control the beam.

EBL is used for making the template or pattern for the direct formation of electronresist the plate when the chips needed to create only one lithography. In this case, using direct scanning beam on a plate covered with resist. When manufacturing the chip must perform several lithographs, then use sets of patterns obtained by electron-beam lithography.

EBL may be a projection and scanning. For the production of integrated circuits two types of projection systems were developed: with preservation of picture size of topological layer and with reduce them. In systems of the first type the thin-film mask is applied to the surface of a flat photocathode, irradiated by light emission of electrons occurs from places not protected by the mask photocathode, causing electrons hit the electronresist. For electronoresist with a sensitivity of 0.1 Kl/m^2 at current density of 0.1 A/m^2 duration of exposure is about 1s. In EBL reduction projection type metallic mask is used depicting one topological layer chip that is irradiated parallel

beam of electrons. With the focus of reduced electron optical mask pattern is projected onto the plate. The wavelength of electrons with energies of 10 ... 20 keV is about 0.01 nm.

There is no template in scanning EBL, and exposure is occurs by shifting on the surface of the plate focused electron beam using given program.

In *X-ray lithography* the soft X-rays with a wavelength of 0.2 -1 nm is used. It is used for the manufacture of semiconductor chips. Template for X-ray lithography is a thin (about 5 μm) and transparent to X-rays the silicon membrane or other material which is applied to thin film transparent to X-ray beam pattern layer which is made from gold in 1:1 scale. Figure of template are created by electron-beam lithography. Plates are coated with the layer resist which is sensitive to X-rays (e.g, polymetilacrylate, PMMA). Resist exposures through template. Due to the small wavelength of X-rays a minimum line width is obtained and diffraction radiation almost does not limit the resolution. Resolution increases with increasing distance from the source of X-rays to the plate, but increases the duration of exposure.

The considered methods of forming the microelectronics are indirect, as the process of drawing film formation and properties of the layers forming elements (resistors, capacitors, and transistors) are separated in time.

The method *free the mask* is used for forming elements at film vacuum deposition . It is based on the screening of the substrate flux from substances deposited with a special stencil which is free mask that accurately reproduces with topological layer of IC. During the deposition film the formation is a direct picture element. *Free mask* is bimetallic thin screen with holes, the configuration and placement which meets the required configuration of topological layer.

6.3. Photolithography

Photolithography - the formation on the surface of the wafer or substrate using photosensitive material (photoresist) protective masks depicting elements of the topological layer. Following the transfer of the image topological layer on a heat-resistant mask or perform functional layer by chemical, plasma or plasma-chemical etching.

Widespread use of photolithography due to the following advantages: versatility of the method (applying for the establishment of intermediate masks and elements of IC); simplicity of the transition from one configuration to another picture; the default pattern masks; high reproducibility and resolution; high performance.

The basis of the photolithography process is photochemical reactions that take place in the photoresist under the influence of light radiation. As a result of photochemical reactions of changing the structure of photosensitive materials and their solubility in chemicals. The nature of the interaction of chemicals exhibited and unexposed land photoresist layer varies dramatically,

which makes it possible to carry out the following dimensional processing, which is formed as a result of flat relief mask.

6.3.1. Technology of photolithography

Photolithography is the main way of forming elements integrated circuits, and it is widely used in industrial processes producing bipolar and MOS integrated circuits as separate processes.

The process consists of several photolithography process operations (fig 6.1).

1. Applying photoresist. On the surface oxide layer of a semiconductor wafer spraying or centrifuging applied photoresist (Fig. 6.1 a). Thickness photoresist film can be from 500 to 1000 nm. Use photoresisters two types: positive and negative.

2. Drying photoresist layer. During the drying the solvent is removed from the photoresist. Alone resist becomes semisolid lamina.

3. Combination and exposure. Covered with photoresist plate is placed in unit combination, where it possesses both a photomask. Position wafer and photomask regulated so that the special tags for photomasks (signs of reference) to coincide with the corresponding plate. Masks made on a glass plate on one side of which a thin film of metal or emulsion creates a matrix of optically contrasting images of topological layer chips in 1:1 scale. Masks the picture plane is pressed against the plate and turn on the source of ultraviolet radiation (Fig. 6.1 b). High-energy radiation through the optically transparent region masks penetrates the surface of the photoresist in which photochemical reactions take place.

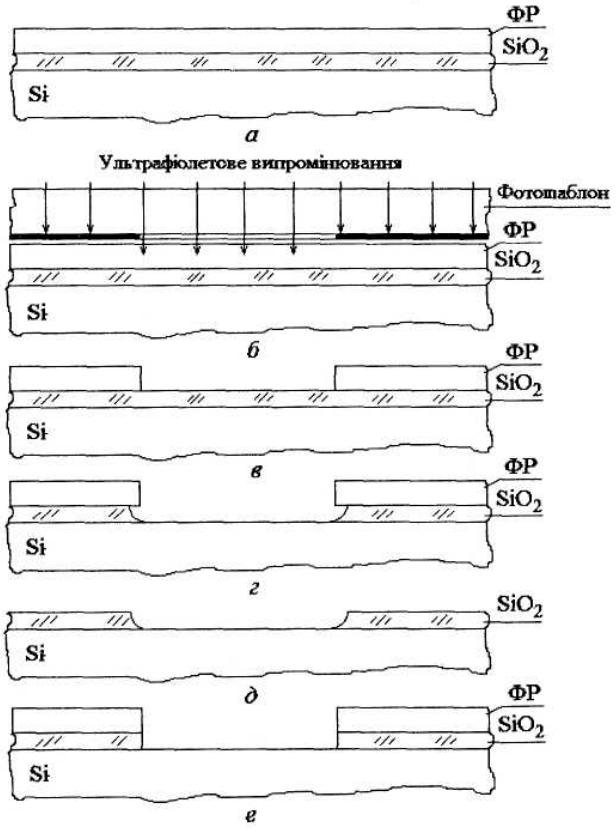


Fig 6.1. The sequence of the major technological operations photolithography

4. Developing. Positive photoresist irradiated through photomask and subsequent manifestation creates a direct image picture photomask. In areas of the photoresist exposed to ultraviolet the photochemical reactions occur degradation. Therefore, under the action of the developer exposed area of photoresist dissolve easily and are removed from the surface. Unexposed areas on the surface of the photoresist are almost insoluble (Fig. 6.1 c). The negative photoresist irradiation through photomask and next manifestation creates a reverse image on the photomask pattern. In areas of the photoresist exposed to ultraviolet light, there photochemical polymerization reaction and the photoresist become resistant to the action of the developer.

5. Fixing photoresist. In temperatures of 130 ... 150 ° C is dry photoresist, improving its adhesion to the oxide film and acid stability. After this, the surface of silicon dioxide is created photoresist mask pattern photomasks.

6. Transferring images photoresistive mask dioxide film silicon. Images can be transferred in many ways: by chemical etching, plasma and plasma-chemical removal, etc. Moving image by chemical etching plate is placed in a solution of acid and hydrogen fluorine perform isotropic etching process dioxide film in places not protected photoresist mask (Fig. 6.1, d). After etching the silicon dioxide layer is created in dioxide picture windows that replicate the pattern in a layer of photoresist and photomasks pattern. With the plate removed photoresist (Fig. 6.1 e). Duration digestion control with high precision: it shall be sufficient for the complete removal of silicon dioxide in the box and not too large to prevent significant underdigestion dioxide under photoresist mask. Lateral underdigestion leads to an increase in the size of the windows in the mask of silicon dioxide compared with the set in the design. Consider the process of transferring an image from photoresist mask on mask of silicon dioxide are called "wet" etching, which is based on an isotropic chemical etching process.

Today, modern methods of image transfer are used, which include cutting and plasma-chemical methods. The basis of the implementation of these methods is the process of anisotropic etching. Therefore, these methods are called dry etching methods. In the methods of dry etching with a high degree of anisotropy of the etching in the direction perpendicular to the surface of the plate is significantly shorter than the lateral (Fig. 6.1 f). Due to the anisotropy of the etching was obtained elements of integrated circuits with submicron dimensions.

Photolithography used to create the surface layers of masking nitride or silicon dioxide on the surface functional layer topological relief image of one layer of the chip. Relief image of a topological layer chip serves as the mask, which is transferred to other functional layers of camouflage or plate. If the relief of the topological layer of photoresist is transferred to camouflage layers of nitride or silicon dioxide, is formed through heat resistant mask on these then perform local diffusion, ion implantation, epitaxial single-crystal films accumulation, oxidation, plasma or plasma-chemical etching.

After performing photoresist mask formation in leading, resistive, dielectric and insulating layers and thin film hybrid circuits. Thus, the chip manufacturing topological image pattern layer masks transferred to photoresist mask, and the mask pattern is transferred to a photoresist functional layer. However, increasingly using more complex scheme of formation: button photoresist photomask is transferred to a mask, which is then transferred to a heat-resistant mask dioxide or silicon nitride, and through them perform elements forming circuits.

The task of photolithography is to provide the combination of topological layers and play in a two-dimensional pattern resist photomasks with accuracy within ± 10% of the minimum size of its elements, and 5% tolerance of the desired slope edges (Fig. 6.2). Layered combination of topological structures should be carried out with accuracy of better than ± 25% of the minimum size. The minimum width of the playback picture lines are such that the combination is a quadruple precision.

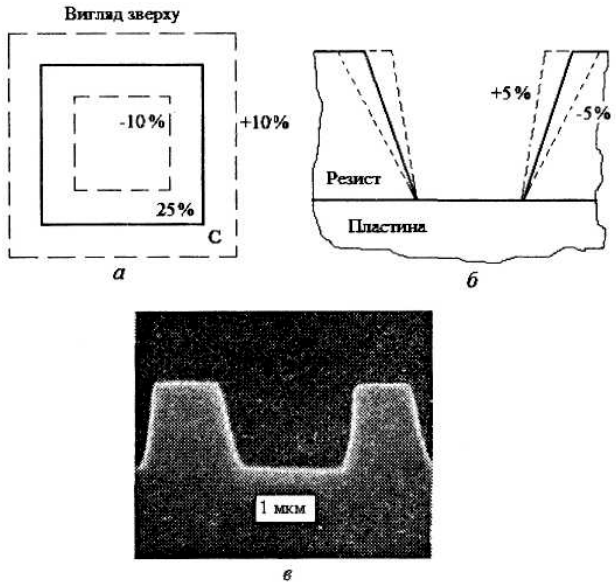


Fig 6.2. Profile in the resist for the next transfer of masking layer: a - reproduction size template; b reproduction tilting edge; c - profile in the resist

During the production of chips perform multiple overlapping and image transfer (up to twenty or more times) to resist in the technological layers. Assessment of the impact of the projection optics and technological systems combining the accuracy is defined as the sum of mean square errors of image transfer and combination.

To achieve a high level of yield and high performance circuit’s photolithographic equipment other than the correct reproduction of the topology of the chip and the precise combination of layers should be organized playable technical process.

In photolithography, there are problems associated with disabilities light beams (diffraction, interference, refraction and reflection); mechanical devices exhibit restrictions, thermal distortion of the wave plates and their relief, resist properties (contrast, the developing speed, etc.).

Working resist choosing a used light sensitivity to radiation-resolution acidstability. Devices exhibiting selected from highaperture optics (numerical aperture NA> 0,4), which allows the structure to form a sharp edge. However, in this case, the image field is reduced, and in one exposure can exhibit only one crystal or even part of it (Fig. 6.3). In addition, the depth of focus becomes close in meaning to the thickness of the photoresist and each field must again combination and focus before the next exposure.

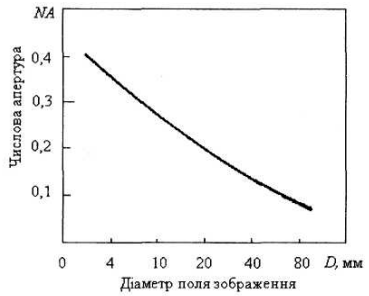


Fig 6.3. Dependence of the numerical aperture of the lens on the size of the image field

Radiation sources used in photolithography divided into point (lasers) and extended (mercury lamps). In this respect exposure photoresisters through photomask in projection systems can implement coherent and incoherent light stream. The emission spectrum of these sources is in three main spectral ranges: the far ultraviolet (UV) - from 100 to 300 nm; average UV from 300 to 360 nm; near-UV - from 360 to 450 nm.

Mercury lamps emit power in the near UV range at wavelengths 365 nm, 405 nm and 435 nm, the average UV range at wavelengths 313 and 334 nm in the far UV range -254 nm. Power output radiation of a mercury arc lamp in the far UV to the lowest. To create equal intensity in all bands need to develop new sources of energy. An example of such sources is F-Excimer laser with a wavelength of 157 nm. Lenses for far UV range (far UV - range) made of quartz and fluorite. However, their transparency and refractive indices are inadequate for today's requirements of optical materials and manufacturing lenses for far-UV remains problematic.

Go to the far UV range due to the fact that the minimum size of directly proportional to wavelength radiation. In addition, high energy photons (above 5 eV) make it possible to extend the set of applied photoresist, since it is enough to break any organic connection.

6.3.2. Photoresists

In optical lithography to create a protective mask on the surface of the plate using photoresists. They have a low light sensitivity and high resolution. Photoresist is applied to a thin film (0.1 ... 2.0 mm) on the surface of the plate, which creates an image layer. The film exhibits in the blue or ultraviolet light. Under the influence of light exposed areas of photoresist change their chemical properties, so that creates a latent image topological layer. During the manifestation

occurs selectively removing resist according to the resulting exposure. Figure photoresist on the wafer surface is used as a mask during the etching, sputtering or other technological operations which are used in microtechnology.

To perform these functions, photoresists must meet the following requirements:

- High sensitivity to light beams specified range wavelengths;
- High resolution, which determines the number of lines on one millimeter;
- Good adhesion to the surface of the wafer or substrate; photoresist layer must hold fast to the surface without peeling at drying, fixing, developing and etching;
- Acid stability - ability to withstand long layers photoresist impact dressers;
- Adaptability - a convenient and easy method of applying layers and subsequent their treatment, minimal toxicity and resist developer.

A prerequisite for successful application of photoresist is its sensitivity. Therefore, the technology of chips used only those polymers in which the ability to change their properties in response to light is the maximum and manageable. To increase the sensitivity of the photosensitive polymer injected impurities that reduce the activation energy of the reaction or curing photodestruction. As a result, regardless of the nature of the impurities and the mechanism of its action is possible tempo-driven flow of secondary chemical processes that lead to irreversible change properties of polymers.

Table 6.1. Typical resist composition

Фоторезист	Тип резисту		
	Позитивний електронно-, йонно-променевий, рентгенівський	Позитивний УФ	Негативний УФ
Склад:			
– резист	Поліметилметакрилат	Феноло-формальдегідна смола (новолак)	Циклічний каучук
– чутливий компонент	-	Нафтохінондіазид	Бісазид
– розчинник	Хлорбензол	два- етоксіетилацетат, діоксан, етилцелозольв	Ксилол
– проявник	Метилізобутилкетон	0,25 % КОН, водний розчин тринатрійфосфату	Ксилол
Операція:			
– сушіння	175 °С	75…90 °С	75…90 °С

Resists used in microlithography is film-forming polymers, low molecular weight impurities photosensitivity. To create photoresisters widely used synthetic polymers: polyvinyl alcohol, polyesters, polyamides, phenol-formaldehyde and epoxy resins (Table 6.1).

For photoresisters impurity radiation sensitive components ranging from 2 to 30% by weight of polymer base, depending on the molar absorption coefficient.

Properties resist, except sensitivity depend on the polymer. Since the absorption beam electron, ion and X-ray resist is determined by their mass absorption coefficient photosensitive component in this case is not required.

In response to thermal radiation photoresists are divided into positive and negative. The negative photoresist in exhibiting through photomask expose to light becomes resistant to solvents (in photoresist polymerization reaction occurs - cross-stitching), and no exposure tracts photoresist is removed. On the surface of the plate after the display of the image created by the reverse pattern masks. Positive photoresist in exhibiting through photomask expose to light becomes soluble in aqueous alkaline solutions or decreases its resistance to plasma etching, as when exposed to light it, the processes photodecay, photoregrouping, photoaccession or photosensitivity. Irradiation tracts removed and the surface of the plate creates a direct image masks.

Positive photoresist typically consists of three main components: resin, which provides manufacturing resist as a thin film photoactive compound (inhibitor) and the solvent, which allows applying resist to the plate in the liquid phase. In the dry film (thickness 0.1 ... 2.0 mm) photoactive compound prevents resist dissolution in manifestation aqueous-alkaline solution. The destruction of the photoresist side light creates compounds that do not prevent the dissolution of resist and increase the rate of its removal developer.

Today, about 20 manufacturers produce positive photoresisters DHN-resists consisting of hinondiazide ether and phenol-formaldehyde resin. Photolysis reaction scheme shown in Fig. 6.4. In addition, the released nitrogen foam polymer film, this facilitates the penetration of the developer. Solvents are alcohols, ketones, aromatic hydrocarbons, dioxin, xylene, etc. A common approach is a mixture of several solvents with different elasticity vapor improves film formation, reduces pores and mechanical stresses in the film.

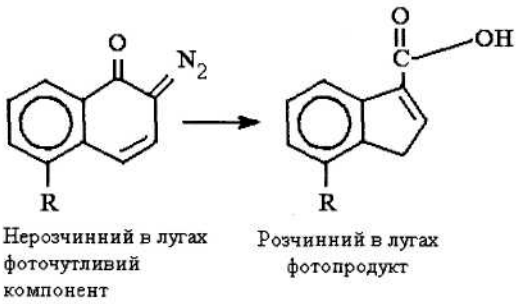


Fig 6.4. Scheme photolysis DHN-resist

The main parameters of the positive photoresisters shown in Table. 6.2.

DHN-positive photoresists have important advantages over negative:

- Insensitivity to oxygen;
- The highest stability in plasma, stability in plasma due to the presence of oxygen phenol components in the resin;

- Can be thermally stabilized up to 200 ° C;
- May manifestation;
- Use for two-layer resistive systems as a mask;
- Easy removal.

The disadvantages of DHN-positive resists are:

- A small tolerance for display of parameters for medium and low doses exposure;
- Average contrast (y <2), which leads to relatively large errors widths;
- Average sensitivity (750 J/m^2);
- Poor adhesion to the silicon surface.

Table 6.2 Main parameters of positive resist

Марка фоторезисту	Роздільна здатність, ліній/мм (товщина резисту 1 мкм)	Щільність дефектів, *n*/мм^2 (не більше)	Стійкість у проявнику, с
ФП-383	400	0,5	60
ФП-330	400	0,75	60
ФП-333	500	0,2	180
ФП-307	500	0,35	90

The negative photoresist after exposure light acquires the properties acidstability or insolubility of the developer. These properties are due to the increase in molecular weight during polymerization or cross-stitching; change of polarization in creating more or less polar functional groups; modify the degree of oxidation or ion ionization stable complex with charge transfer.

In common negative resists by light occurs disordered cross stitching and basic side chains. At each insertion requires at least one photochemical act to activate this process. Of all the types of polymerization ordered polarization is the most effective way of rapid conversion of monomers or monomer applications or high molecular weight crosslinked polymers.

Since the monomers are distributed in a solid film, which has a high viscosity, connection radicals can not move because the proliferation response is limited.

In modern modification of the classical free radical polymerization is used for negative resists surface after polymerization. The essence of this process lies in the fact that the surface of the polymer creates free radicals or cations (as a result of decomposition of onium salts exposure). In the latent image is applied monomer is polymerized on top of the image. At low doses (less than 50 J/m^2) you can get a dual-layer resist, to use siliconretention monomers.

Of all the ordered polymerization is the most effective way of rapid conversion of monomers or monomer applications or high molecular weight crosslinked polymers. However, to achieve real insolubility resists are mainly disordered intermolecular insertion.

Decreased solubility during exposure passes through the stages of agglomeration to gelation. In the gel fraction can be sol-fraction (soluble), which is due to polydispersity of the polymer, i.e., the presence of low molecular weight components. Polydispersity ratio serves as a molecular weight Mw average weight to average number M, which varies from unity (monodisperse distribution) to 15. Polydispersity of most free radical polymers is in the range of 1.5 to 3. Monodispersed polymer with the highest molecular weight Mw provide maximum sensitivity, resolution ability and yield in polymerization systems with stitching. But for the most part used in the manufacture of polydisperse polymers, which are much cheaper than monodisperse. When exhibiting resistive film at normal dose last 30% of its thickness at the substrate never fully exhibited. The duration of the exposure process is selected so as to ensure the polymerization of the film across the thickness (Fig. 6.5). Resist high adhesion to the substrate achieved only when over exposure.

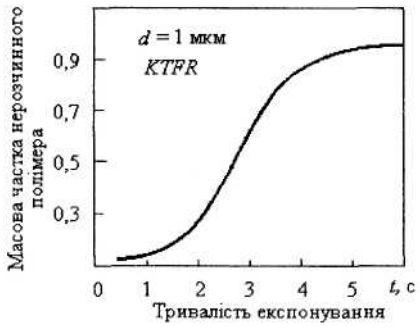


Fig 6.5. Dependence of polymerization negative photoresist on the duration of exposure

Resolution negative resist is limited by scattering of the incident and reflected exposure radiation and heterogeneity of their distribution as necessary to resist overexposure to ensure its insolubility and adhesion to the substrate. With increasing dose exposure increases the width of the lines. Small details picture may disappear during developing. Optimum sensitivity and uniform exposure can resist when the optical density D = 0,43. The optical contrast of the resist is inversely proportional to T and D is 2.3. Films with a high optical density (more than one) have less contrast and require significant over exposure. Resist with a lower optical density with greater contrast, but in need of a large dose of exposure. The low absorption and high resolution can be obtained using thin films of resist, but increases the density of defects in the film and reduced acid stability.

Since most of the resist is polydisperse (2 - 5), they have low contrast (y <2) (see Table 6.3). Resists with contrast close to the theoretical limits of 3.5, giving a cool edge mask profile compared with low contrast (at ~ 0.5 ... 1.5) negative resists.

Table 6.3 Contrast and doze of negative resists

Резист	Контраст γ	Доза E_g, Дж/м^2
KPR	0,9	700
KTFR	0,5	60
KOR	0,5	40

To print to resist the system of equal lines and spaces width of 1 m, you need to modulation transfer function values (TFV) lens on the appropriate frequency was at least 0.6. As contrast photoresisters less than 2, and the minimum critical modulation transfer function resist ranges 0.8 ... 1.0, even when set to TFV lens is equal to one, the minimum size is limited to linear magnitude 2.5 um film thickness in resist 0.5 microns. The resolution can be reduced by using thinner resist laminas. Photosensitivity negative resist depends on the quantum yield F sensitizer, which is modified by light (homoliz or regrouping) and then reacts with the polymer. Sensitizers can also transfer electronic excitation energy to activate the polymer.

The main advantages of negative resists:

- Wide tolerances on the parameters of the developing insensitivity to remanifestation;
- High adhesion and resistance to wet etching;
- Self-compensation of deviations of sizes of elements; restoration unexposed area to its original size due underdigestion at isotropic wet etching;
- A wide range of compositions; use of modern devices exhibits operating in DUF-range (200 ... 300 nm), allows in some cases, waive photosensitive components.

Disadvantages negative resists:

- Resolution limited thickness resists events underexposure, scattering of radiation in the forward and reverse directions;
- Resists desensitized in the presence of oxygen;
- Can not be used to create a metallization by an explosive lithography.

With the introduction of microtechnology in plasma and plasma-chemical etching using the negative resist is reduced from a positive DHN-resist.

Application of photoresist. In modern microtechnology lithographic process is fully automated. He performed in a clean room 0 - 1 class. Technological operations in a clean room place without people, and robotic and automated transport systems run by computers that are located outside the clean room.

Photoresists - Low-energy polymers have poor adhesion to hydrophobic surfaces coated with adsorbed water. Most of the water is removed at elevated temperatures. After removing the surface layer of water silane groups on oxidized silicon wafer covered with silicon active promoter

to improve adhesion DHN-resist and reduce surface tension to surface oxidation. Adhesion promoters applied by immersion, spray, centrifugation or sedimentation from the gas phase.

Putting photoresist layer on the surface of the substrate - one of the critical operations photolithographic process. Conditions and methods of this operation are provided by most of the above formulated requirements for photoresist.

Technological methods of deposition on the surface of the substrate photoresisters: dip, spin, spray (pulverization), vapor deposition, etc.

Given the need to automate the process of applying photoresist to the surface of the plates of large diameter (250 - 300 mm), and the use of rectangular plates of large size, the main methods of applying the resist under these conditions is sputtering (pulverization) and vapor deposition.

Spraying receive a wide range of thicknesses resist layers. In this case, the substrate can have a non-flat surface. Formation of films by sputtering occurs deposited on a substrate with discrete drops that are solvent atmosphere merge into a continuous layer. Compared with the method of centrifugation spray photoresist has the following advantages: the ability to precisely control the thickness of the film; uniformity in thickness (at a thickness of 0.1 ... 10 mm) within the whole area of the substrate; minimum internal stresses and defects; high adhesion to the substrate.

Method of forming monomolecular layers of polar molecules on inorganic substrates can be used to nanolithography with high resolution. Since backscatter electrons limit the resolution electron resist, film thickness monometer minimize this phenomenon. Due to the surface adsorption of molecules are placed in layers. Reactive groups are so close that the spontaneous polymerization of vinyl monomers and acetylene can produce highly sensitive resists (10 mkKl/m^2).

Drying is a process operation that completes the formation of a layer of photoresist. It is performed in two steps: low-temperature and high-temperature exposure. During the low-temperature aging for 10 - 15 minutes to gradually removes the solvent and the polymer macromolecules oriented stack. High temperature drying is accompanied by intense evaporation of the solvent and the polymer macromolecules transition to a stable state. The internal capacity of the system decreases, attaining zero. This process needs some relaxation and durability.

Photoresist lamina after application to a substrate and drying to be isotropic and homogeneous. She must have not only a constant thickness across the area in use, but also be chemically isotropic to her reaction to the exposure and manifestation was similar across the surface.

6.3.3 Exposure methods

Contact printing was the first method for forming image in photoresist layer. In this method, mask, which contain matrix of proper image in scale 1:1 after closely, without gap placing on plate, which covered by photoresist, expose with light (Fig. 6.6). After this, exposed and not exposed

areas appear on photoresist. In depends on photoresist type, during further developing exposed areas removes for positive resist and not exposed – for negative resist. Resolution capability is limited by diffraction phenomena between two neighbor lines. When coherent radiation used for exposure, intensity decay on border regions of windows become sharper, which leds to improve image contrast and increase definition. Resolution capability defines as number of high-contrast couples of lines per unit length. Width of lines equal to distance between them. Resolution capability determines minimal line width or minimal distance between two lines.

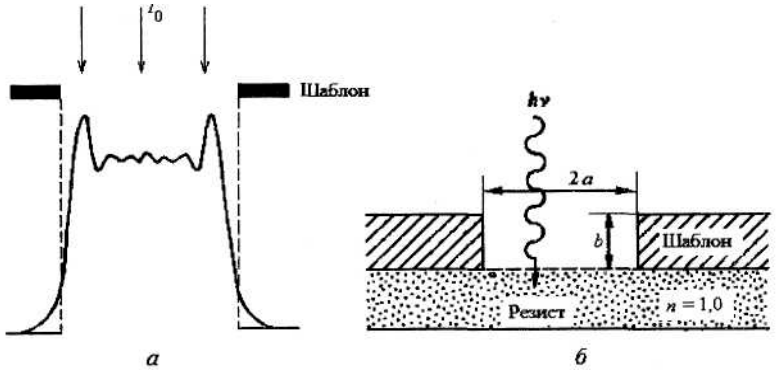


Fig. 6.6 Contact print: a – profile of intensity distribution, b – cutest of photomask and resist film

The main defect of contact printing is wearing-out of photomask with it multiple reusing. Close placing photomask on plate causes appearing of defects on photomask and resist surface. Storing of defects and photoresist particles attached to photomask leds to fast wearing-out of photomask. Maximal exposition number depends on template complexity and hardness of layer with photomask image.

Emulsion photomasks for exposuring large IC are usually used nearly ten times. Photomasks, which covered by chrome, oxides of iron or other metals, can be periodically cleaned, so they can be used more then 100 times. Resolution capability of contact printing improves by using UV- radiation (wavelength 200…260 nm) for exposuring of photoresist, which helps to reduce diffraction influence.

It is hard to realize conditions of contact printing in practice, as plates have thermic curvature and wave profile of surface, which create local gaps between photomask and photoresist, size of which can be near 10 mkm.

Contactless printing. Contactless printing method realizes by space removing of photomask on distance of few microns from covered with photoresist substrate, which significantly reduce number of defects, which appears because of contact with plate. Diffraction of exposure light reduce resolution capability and deteriorate image definition (Fig. 6.7). Diffraction cause photoresist exposure out of window range in photomask. Level of these negative effects depends on distance between phtomask and plate, which different on whole surface of plate. Angle of exposure

by diffracted radiation of photoresist protected plate surface depends on window size and lightbeam wavelength.

In case of small gaps, non flatness is the main source of distortions (especially in large diameter plates). The main phenomena for plates with large diameter is diffraction, so exposure possibilities in reproduction lines with minimal topological size are worse than in case of contact print.

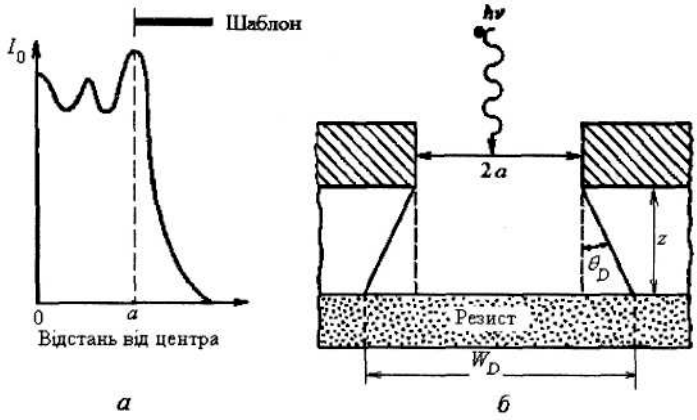


Fig. 6.7 Contactless print: a – profile of intensity distribution, b – cutest of photomask and resist film

Projection print. Projection print realizes by projecting photomask image on covered with photoresist plate by using high-definition lens system. In this method, photomask can be used many times, so using high grade mask is profitable. In projection system used lenses (Fig. 6.8) or mirrors which allows to project image of photmask (scale 10:1, 5:1 or 1:1) on photoresist.

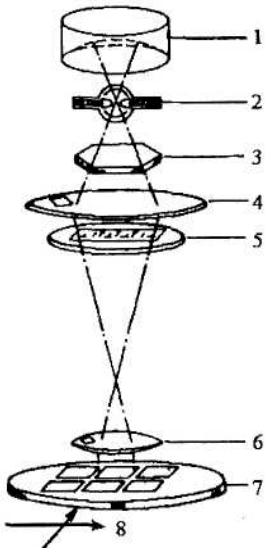


Fig. 6.8 Scheme of projection print device with decreasing based on refracting optics: 1 – mirror; 2 – mercury arc lamp; 3 – filter; 4 – condenser lens; 5 – photomask with topological layer image (4 – 10 XI); 6 – decreasing objective; 7 – plate; 8 – mechanical device of step shifting.

The main requests for step projection systems: high-precision alignment of nearby crystals for compensating plate and photomask distortions; precise laser coordinate table for moving

crystals after every alignment; efficiency – near 40 plates with diameter 300 mm per hour. Profile of lightbeam intensity distribution on photoresist for projection lithography showed on Fig. 6.9 a.

Projecting of two-dimensional topological layer image causes decreasing of slope. So special photoresist, in which by influence of sinusoidal-modulated beam intensity square mask will be formed for further image transference with etching or lift-off lithography, is needed.

Quality of optical image in projection print characterized by function of transference modulation (FTM), which shows dependence of light intensity modulation depth in image plane on grid space frequency, which used as image source (grid consists of direct lines sequence, width of which are equal to distance between them).

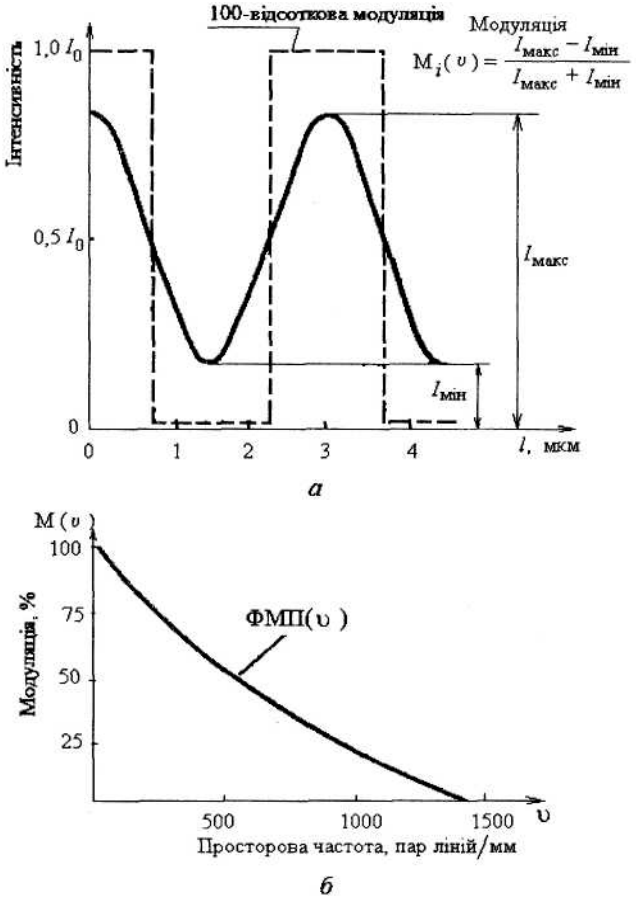


Fig. 6.9 Intensity modulation Mi in image plane (a) and function of transference modulation (b); space frequency – 333 couples of lines/mm; line width and distance between them – 1,5 mkm

FTM curve built for sinusoidal light intensity distribution in image plane (Fig. 6.9 b). Distribution characterized by grid space frequency V, which defines as number of line couples per millimeter and modulation coefficient

$$M_0 = \frac{I_{макс} - I_{мін}}{I_{макс} + I_{мін}}, \qquad (6.1)$$

where I_{max} and I_{min} – local intensity maximums and minimums.

Used to think, that minimal value of FTM must be 60% for reproducing minimal details by using positive photoresist. Image contrast

$$C = I_{макс} / I_{мін}. \quad (6.2)$$

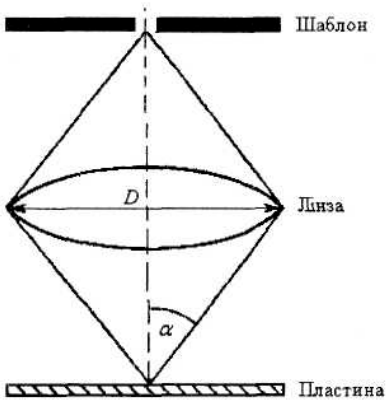
In scanning projection systems photmask image projects on some part of plate. During one exposure image prints on square near 100 mm^2 and then exposure continues after shifting image on the neighbor part of plate. Optical systems characteristics for projection print limited by diffraction phenomena. So projecting and creating of optical elements makes in way, when their parameters, which characterize image reproducing, defines by limited apertures, but not blooming characteristics.

The main parameters of lenses (Fig. 6.10) are: focus *f*, aperture diameter *D* and numerical aperture

$$NA = n \sin\alpha = D / 2f, \quad (6.3)$$

where *n* – refractive index in image space (usually near one); *2a* – maximal angle on cone vertex of beams, which fall on image point on optical axis of projection system. In standard projection systems, which make image transference, objective focus is function of aperture diameter. Effective number of projection system

$$F = 1 / 2NA = f / D. \quad (6.4)$$



а

Параметр	Позначення
Фокусна відстань	f
Діаметр апертури	D
Число F	$F = f/D$
Числова апертура	$NA = n \sin\alpha = D/2f$
Роздільна здатність	$\Delta x = 1{,}22\lambda F$
Глибина фокуса	$\pm\Delta a = \pm 2\lambda F^2$

б

Fig. 6.10 Scheme of projection lens (a) and its parameters (b)

The smallest line size, which can be achieved in resistive film by projective system with diffractive distortions

$$w = k\lambda / NA, \quad (6.5)$$

where *k* – coefficient, *k* <> 0,3 – for resist with image forming in upper surface layer; *k* <> 0,5 – for multilayer resists; *k* <> 0,75 – for single-layer resist; *k* <> 1,1 – for resists, located on reflecting surface. For practical smallest size takes threefold *w* value:

$$w = 1{,}83\lambda / NA. \quad (6.6)$$

By using far UV-radiation and objective with bigger numerical aperture value of minimal print size decreases. Focus depth *LI* also decrease with increasing of numerical aperture *NA*

$$\Delta l = \pm\lambda/2(NA)^2 = \pm 2\lambda F^2, \quad (6.7)$$

which causes appearing of length optical difference ±λ/4 in image plane. For keeping hidden image in focus it's necessary to make additional focusing for every exposure.

For objective with *NA = 0,35* and light exposuring wavelength 300 nm focus depth will be smaller than 1,5 mkm. In this case imparalelity of plates, topographical profile on surface and even thickness of resist will limit minimal size.

Focus depth of optical system must be less then ±10 mkm, which usually includes declination of plate surface from ideal plane. This limits lens aperture and resolution capability. So, it's almost impossible to create lens system, which simultaneously satisfies conditions of achieving high quality (limited by diffraction phenomena) image and evenly lighting of whole surface of plate with typical size 300 mm.

Optical image forming systems for projecting devices could be divided in depend on of light source type onto coherent and noncoherent. If photomask or grid illuminated from point source the image in projecting device will be coherent or almost coherent, as in image plane (of plate) light, diffracted by grid, is coherent by amplitude.

The majority of projection print systems have light source diameter much smaller then objective diameter. So, forming image is quasycoherent. For contact or contactless print with small gap coherent radiation are used. In case of noncoherent radiation image on plate surface will be blurred by penumbra.

6.4. Models exhibiting

Photochemical reactions in photoresist occurs when activation by photons. Activation is selective. Absorb quantum of light activates a specific molecule organic system without arousing others. Photochemical conversion process is divided into three stages:

- The absorption of photon, when the molecules acquire electron-excited state;
- Primary photochemical processes in which molecules are involved in electron-excited state;
- Second or "dark" reactions of various chemical compounds formed due to the primary process.

Falling on photoresist light creates in it the activated molecule. Activation molecules can be direct or sensitized. In the first case, the excited molecules react directly to light, and the second - the excited molecules do not react, and transfer their energy to other molecules. These molecules go

from a state of thermodynamic equilibrium with the environment, interact with the molecules of the photoresist, causing a chemical reaction.

The quantum yield of the photodecomposition reaction Φ photosensitive components of positive resist (naphthoquinonedoziad - DHN) indicates efficient use of photons. Define it as the ratio of the number active molecules to absorbed photons. At best, only one of five photons causes the necessary changes. This means it is necessary at dose 1000 J/m^2 ,10^{17} photons for photochemical conversion of 10^{16} molecules DHN. However, for the display of images in the resist sufficiently decomposed only about half DHN molecules. Quantum yield Φ positive DHN-resists are shown in table 3.4.

Exhibited positive photoresisters tracts occur at rate approximately an order of magnitude greater than the rate of unexposed development tracts after exhibiting by rays with a wavelength of 150…500 nm. Resist profiles required for different types of lithographic processes are created at various doses of rays and ratios of display rate.

Table 6.4 Quantum yield of positive resists

λ	365 нм	405 нм	435 нм	Денне світло
Φ	0,15	0,23	0,16	0,13…0,26

At high doses (more than 1 000 J/m2) shape of the profile depends on the absorbed radiation, reflected photons and quantum yield Φ of photochemical transformations in resists. Absorbed energy determines the dependence of the dissolution rate of resist from penetration depth:

$$\frac{d\vartheta_{E.\Pi}}{dz} = \left(\frac{d\vartheta_{E.\Pi}}{dE}\right)\left(\frac{dE}{dz}\right). \quad (6.8)$$

At medium doses contributed to the creation resist profile does as a factor that determines the solubility of the photoresist in the developing ($d\vartheta_{E.\Pi}/dE$) (6.8), and the factor that determines the absorption (dE/dz).

At low doses factor that determines the solubility in developing, is dominant profile is eroded and resist film becomes thinner. The angle of the edge photoresist mask ($d\vartheta_{E.\Pi}/dE$) in the design gap of minimum width W ((6.5), (6.5)) depends on the optical circuit device exhibiting and error minimum size of Aw -tilting edge.

$$\frac{dE}{dz} = \frac{2NA}{\lambda\left[1 - \Delta l(NA)^2/\lambda\right]^2}. \quad (6.9)$$

The effectiveness of the absorbed energy to resist depends on the contrast of the developing process γ. For positive photoresisters contrast is determined by the slope of the characteristic curve of film development (Figure 6.11), which is obtained as the ratio of the residual resist layer thickness after the initial manifestation of lg the absorbed dose.

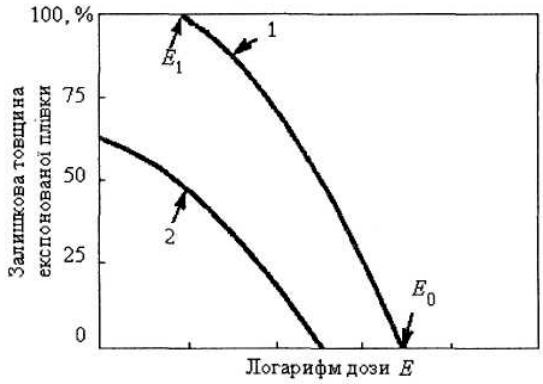


Fig. 6.11. Characteristics curves of positive resist: 1 – no reducing thickness of the photoresist; 2 – thinning unexposed resist tracts.

Contrast is calculated by the formula

$$\gamma = [lg(E_1 / E_0)]^{-1}, \quad (6.10)$$

Where E_1 – dose early development resist, E_0 – dose at which is not photoresist at substrate after development.

Dose E_1 is independent of the film thickness *d*, and E_0 increases with increasing thickness according to absorption law.

$$\gamma = (\beta + \alpha d)^{-1}, \quad (6.11)$$

Where β – constant; a – absorption coefficient resist. For a homogeneous expose resist at all depth of the films, absorption should be about 0,4, the maximum contrast of the resist – 3 – 4.

It is shown the value of contrast for positive DHN-resist at irradiation dose of 1000 J/m inTable. 6.5. The contrast was determined for silicon wafers coated with a layer of silicon dioxide thickness of 0,7 μm, grown in wet oxygen. The thickness of the resist film - 1 μm. At this dose imaging process in resist determines the film development.

Table 6.5 The value of contrast of positive resists

Діапазон експонування (довжина хвилі), нм	Резист	γ
340…440	AZ-111	1,3
	AZ-1370	2,0
	AZ-2400	2,0
	HR-204	1,43
300…330	AZ-2400	1,67
	Kodak 809	1,61
	AZ-1370	1,2
	HR-204	1,0
280…330	AZ-2400	0,96
	AZ-1370	1,2
	HR-204	0,85

The light that passed through the photoresist absorbed during exposure. For the mathematical description of the optical absorption light without reflection (the model is perfect) usually are model Dill-Shava, which is used to study the optical and mechanical parameters of the resist. Beer-Lambert Law describes this model as follows:

$$\frac{dI(z,t)}{dz} = -I(z,t)\sum_i a_i m_i , \quad (6.12)$$

Where I(z,t) – intensity of light at any time and anywhere in photoresist film; z – distance between the surface of the photoresist and some point in the film; a – molar absorption coefficient of the photoresist`s component I; m – molar concentration of component i.

Photoresist is usually regarded as a three-component system: inhibitor or photoinitiator (photosensitive component); rubber or protector (gum); solvent. It should be noted that the solvent evaporates mostly during drying inhibitor partially destroyed during the exposure, and the remainder converted to inhibitor photoreaction products. Inhibitor is a major component that absorbs light. Consider that the curve of absorption spectra and spectral sensitivity for a positive resist is equivalent. Therefore, equation (3.15) can be written as follows:

$$\frac{dI(z,t)}{dz} = -I(z,t)[a_1 m_1(z,t) + a_2 m_2(z,t) + a_3 m_3(z,t)], \quad (6.13)$$

Where I(z,t) – intensity in the photoresist at a depth z for the duration of exposure t, a_1, a_2 and a_3 – molar absorption coefficient of inhibitor, protector and products of photoreaction; m1*(z,t), m2(z,t)* and $m_3(z,t)$ – molar concentration of inhibitor, protector and products of photoreaction at depth z for the duraction of expose t.

Optical absorption of radiation is an important characteristic of photoresist as determines the rate of photochemical reactions in contrast. Therefore, optical properties of the resist are fully describes by refractive index and absorption coefficient a. Refractive index in complex form written as

$$\bar{n} = n - ik , \quad (6.14),$$

Where n is real part, n= 1.68 (λ= 404,7 nm)f or photoresists AZ1350J

$$k = \alpha\lambda / 4\pi . \quad (6.15)$$

To resist such AZ1350J at the wavelength of 404,7 nm extinction coefficient k=0.0187. The intensity of light that irradiates the surface of the photoresist on each micrometer of its thickness, weakened by the amount

$$exp(-\alpha x) = exp(-4\pi kx / \lambda) = exp(-0,58x) = 0,56 . \quad (6.16)$$

After the exposure concentration of the inhibitor significantly reduced. For most of the duration of exposure of resist can be considered transparent.

The film development of the positive photoresist. The film development – surface reaction removing the expose areas of the photoresist. If they remove the liquid chemical reagent, the rate of manifestation depends on the concentration of the inhibitor to the exposed surface of the resist chemical composition and dressers. Experimental dependence of the developing rate photoresist R from normalized concentration of inhibitor M (z, t) makes it possible to determine the ratio between the duration of exposure and the duration of film developing. According to experimental data dependence R (M) can be written as

$$R(M) \equiv exp\left(E_1 + E_2 M + E_3 M^2\right). \quad (6.17)$$

For photoresist AZ1350J thickness 583.4 nm after exposure radiation with an energy of 570 J/m at a wavelength 435.8 nm, the experimental values of the coefficient are as follows: E_1 =5.27, E_2 =8.19, E_3 = -12.5.

Exposure light of a certain wavelength causes a chemical modification of the photoresist, resulting in the loss of its protective properties, ie, the destruction of the inhibitor. The effect of chemical modifications localized in the vicinity of the point where the photon absorption occurred, and the level of exposure may vary widely in film resist. The film development is in the process of digestion, in which the selective removal of resist with a speed that depends on the amount of destroyed inhibitor. The relationship between exposure and film development process determined by the amount of inhibitor remaining unexposed.

6.5. Exposure of negative photoresist

A negative photoresist is a light-sensitive material with an effective threshold energy Em. If the photon energy is less than the threshold exposure, the \resist is removed during developing. When the photon energy E> Em resist becomes insoluble in developer and created an image of a topological layer acts as a protective mask. The value of the threshold energy depends on the type of photoresist thickness and so on.

Image size on photoresist lamina depends on the ratio between the effective threshold energy exhibition and distribution of energy in the diffraction of light at the edge photomask (Figure 6.12).

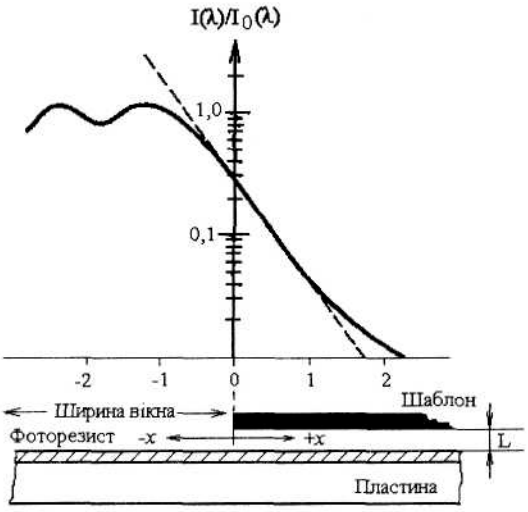


Fig 6.12. The diffraction of light at the edge photomasks

6.6. Advances photolithography

Minimum topological dimensions of the integrated circuits in microtechnology achieved through photolithography. Today, the industry has mastered the minimum topological dimensions of 130, 90 and 60 nm, exhibiting are performed mainly lasers with wavelengths of 248 and 193 nm. Introduces in the manufacture of ICs with topological size 45 and 32 nm, and 2010 before the chip manufacturers have released topological size of 13.5 nm. Back in 2006 Intel Corporation produced SRAM with minimal topological size of 45 nm with a capacity of over 153 Mb and an area of 119 mm. Memory cell has an area of 0.346 μm^2.

Chapter 7. Etching

V.Skryshevsky

7.1 Introduction

Single crystals in academic research and device technology are seldom used in as-grown form. They are usually cut, abraded and polished to prepare samples of desired dimensions and orientations. By physical and chemical means the quality of the surface so prepared is controlled and assessed and the distribution and density of defects regulated. Etching is applied to obtain desired mesas and grooves in semiconductor wafers and multilayers. Most of these operations essentially involve the removal of material by expending energy by mechanical, thermal, or chemical means. Material can be removed mechanically from the surface by abrasion or by bombarding it with energetic particles. Thermal means involve the heating of a crystal at elevated temperatures in vacuum or some gaseous atmosphere. Chemical means make use of chemical reagents which may be the undersaturated mother liquor of the dissolving crystalline substance or some other liquid or gaseous medium capable of reacting with the crystal to yield reaction products. Abrasion and grinding operations remove the surface, more or less uniformly, without revealing the crystalline microstructure, and hence this kind of process is called *mechanical polishing.*

Bombardment, and the thermal or chemical treatment of crystals invariably reveal defects and microstructure, and therefore these methods bear the general suffix *"etching"* (fig.7.1). In the case of chemical treatments it is also possible to remove the surface uniformly by changing the experimental parameters, e.g. composition of etching reactants, temperature of etching, etc. The process is then referred to as *chemical polishing.* A crystal can also be etched in a chemical reagent under the application of a potential; by changing the experimental conditions, such as current density, composition of the etching bath, etc., the crystal can be etched or polished. These processes are called *electrolytic etching* and *electrolytic polishing*, respectively. Electrolytic dissolution is applicable to materials (metals and semiconductors) which are good conductors of electricity. When a chemical reagent selectively reveals the surface microstructure, including defects, the process is referred to as *selective etching.* If the revelation of dislocations is of prime concern, the term *"dislocation etching"* is frequently used. However, in the case of semiconductors, etching and polishing are employed in the same sense; both simply mean the removal of material. In dissolution kinetics, on the other hand, dissolution and etching are taken as synonyms, although in general dissolution refers to the macroscopic removal of material from the crystal surface.

Generally, etching is consisted of 3 processes:

- Mass transport of reactants (through a boundary layer) to the surface to be etched.
- Reaction between reactants and the film to be etched at the surface.
- Mass transport of reaction products from the surface through the surface boundary layer.

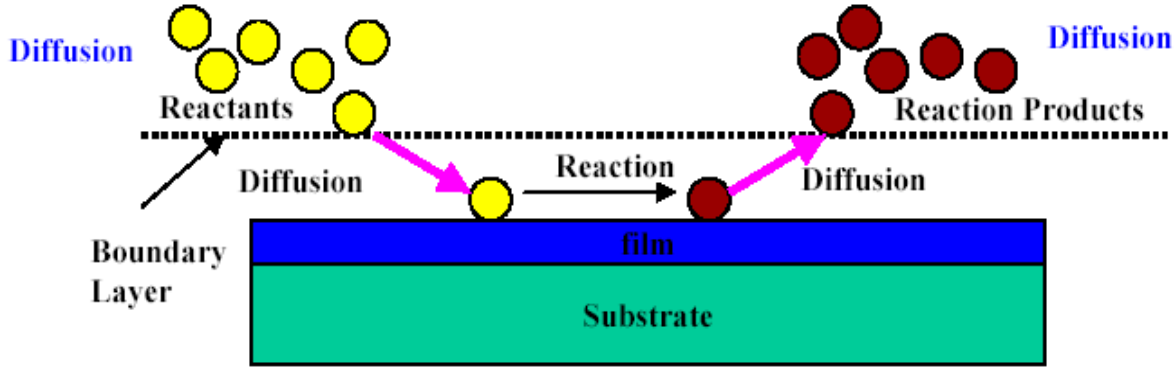


Fig.7.1 3 processes of etching.

7.2 Wet chemical etching.

Etching is done either in "*dry*" or "*wet*" methods. Wet etching is a material removal process that uses liquid chemicals or etchants to remove materials from a wafer. The specific patters are defined by masks on the wafer. Materials that are not protected by the masks are etched away by liquid chemicals. These masks are deposited and patterned on the wafers in a prior fabrication step using lithography. A wet etching process involves multiple chemical reactions that consume the original reactants and produce new reactants. The wet etch process can be described by three basic steps. (1) Diffusion of the liquid etchant to the structure that is to be removed. (2) The reaction between the liquid etchant and the material being etched away. A reduction-oxidation (redox) reaction usually occurs. This reaction entails the oxidation of the material then dissolving the oxidized material. (3) Diffusion of the byproducts in the reaction from the reacted surface. Wet etch is cheap and simple, but hard to control (not reproducible), not popular for *nanofabrication* for pattern transfer purpose.

When a material is attacked by a liquid or vapor etchant, it is removed *isotropically* (uniformly in all directions) or *anisotropic etching* (uniformity in vertical direction). The difference between isotropic etching and anisotropic etching is shown in Figure 7.2. Material removal rate for wet-etching is usually faster than the rates for many dry etching processes and can easily be changed by varying temperature or the concentration of active species. Liquid etchants etch crystalline materials at different rates depending upon which crystal face is exposed to the etchant.

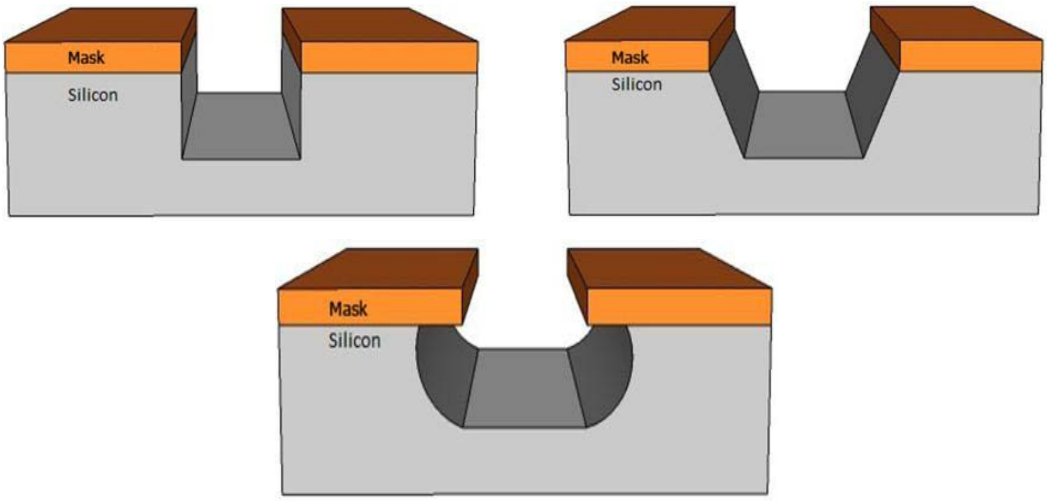


Figure 7.2. (a) Completely anisotropic (b) Partially anisotropic and (c) Isotropic etching of silicon

Advantages of wet etching are high selectivity, relatively inexpensive equipment, batch system with high throughput, etch rate can be very fast (many μm/min). Disadvantages: generally isotropic profile Fig. 7.3, high chemical usage, poor process control (not so reproducible), excessive particulate contamination.

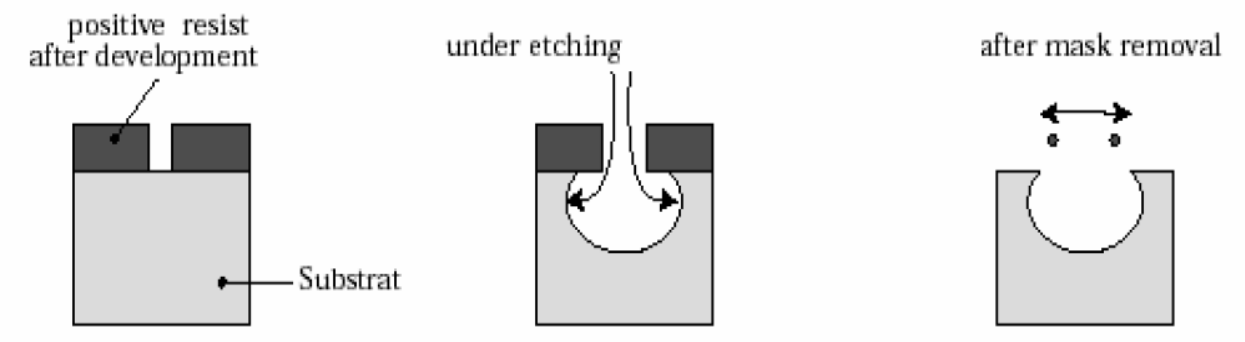


Fig.7.3 Wet etching through mask.

The etch rate can be controlled by any of the three serial processes: reactants transport to the surface (depends on chemical concentration and stirring…), reaction rate (depends on temperature), reaction products transport from the surface (depends on stirring…). Preference is to have reaction rate controlled process because etch rate can be increased by temperature and good control over reaction rate – temperature of a liquid is easy to control. Mass transport control will result in non-uniform etch rate: edge etches faster. Etchant is often stirred to minimize boundary layer and make etching more uniform.

7.3 Isotropic wet etching

7.3.1 Silicon dioxide

Wet etching of silicon dioxide ocurs in HF solution. SiO_2 and H_2O are polar molecules and attract each other, i.e. SiO_2 is hydrophilic material, whereas Si is non polar and hydrophobic. The etch process can be described by eq. (1)

$SiO_2 + 6HF \rightarrow H_2SiF_6 + 2H_2O$ (1)

- SiO_2 Etch is isotropic and easily controlled by dilution of HF in H_2O.
- Thermally grown oxide etches at
    - 120 nm/min in $6H_2O$:1HF (Fig.7.4).
    - ~1 μm/min in 49 wt% HF (i.e. undiluted as purchased HF).
- Faster etch rate for doped or deposited oxide.
- High etch selectivity (SiO_2/Si) > 100

In regular HF etches, HF is consumed and the etch rate drops. To overcome this problem the buffered HF (BHF) or buffered oxide etchant (BOE) are used that provides consistent etch rate HF buffered with NH_4F (40% concentration) to maintain HF (50% concentration), typically 6 (or 5) NH_4F : 1HF. The etch rate is constant due to reaction of:

$$NH_4F \rightarrow NH_3\uparrow + HF$$

The typical etch rate in standart BOE is 100 nm/min.

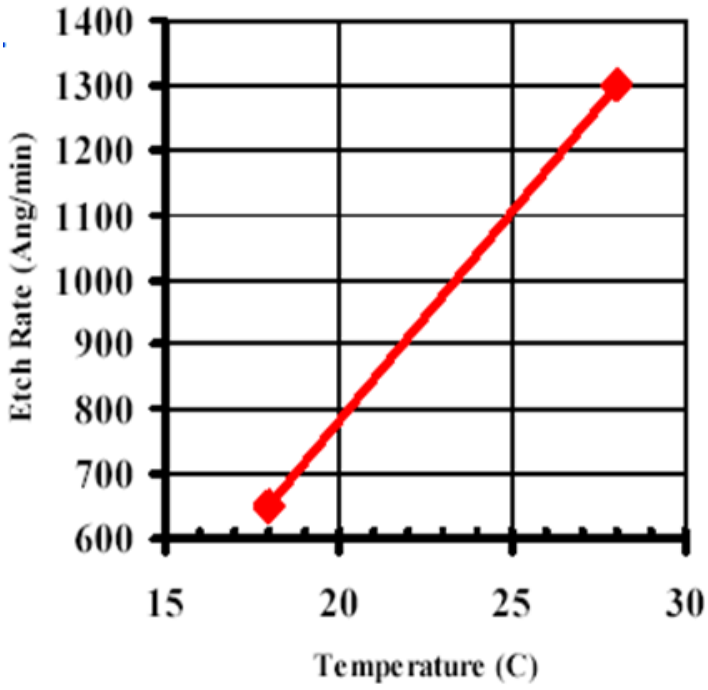


Fig.7.4 The rate of silicon dioxide etching versus temperature in HF solution.

7.3.2 Silicon isotropic etching

- Silicon is isotropic etched by nitric acid and hydrofluoric acid mixtures (HNO_3 may be replaced by other strong oxidants like H_2O_2)
- HNO_3 partially decomposes to NO_2, which oxidizes the surface of Si.

  $Si + 2NO_2 + 2H_2O \rightarrow SiO_2 + H_2 + 2HNO_2$

- The HF then dissolves the SiO_2. The overall reaction is

  $$Si + HNO_3 + 6HF \rightarrow H_2SiF_6 + HNO_2 + H_2O + H_2$$

- Excess nitric acid results in a lot of silicon dioxide formation and etch rate becomes limited by HF removal of oxide (polishing).
- CH_3COOH (acetic acid) or H_2O can be added as diluent, but etch differently.
- Acetic acid is preferred because it prevents HNO_3 dissociation.

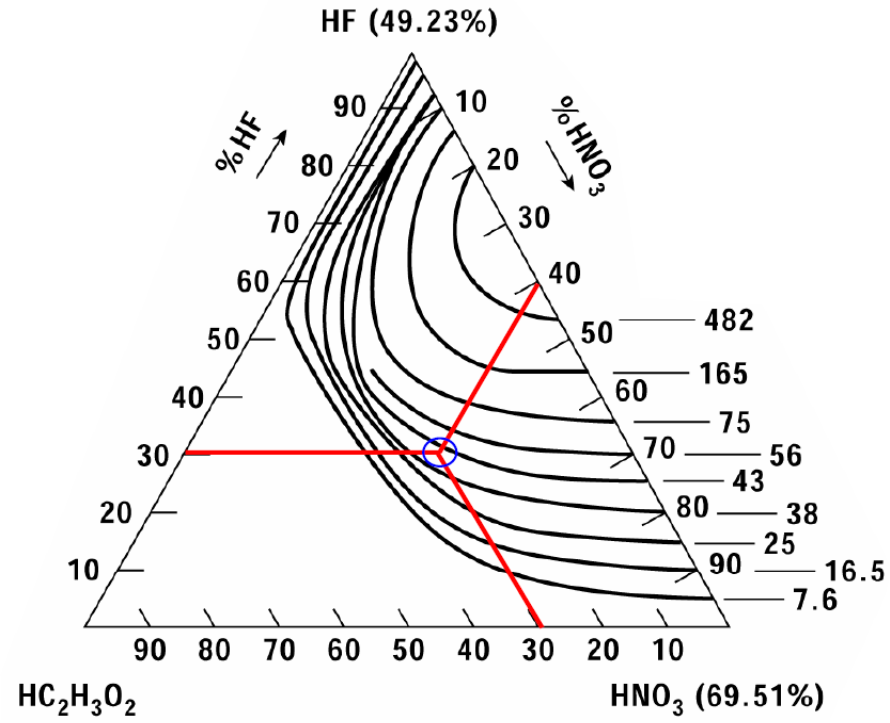


Fig.7.5 Regions of silicon etching versus reactant concentration.

Fig 7.5 shows the regions of silicon etching versus reactant concentration. Region 1 is high HF concentrations, reaction limited by HNO_3, follow constant HNO_3% lines. Rate limited by oxidation, etched wafer surface have some oxide. Region 2 is high HNO_3 concentrations, reaction limited by HF, follow constant HF % lines. Rate limited by reduction, etched wafer

surface have more oxide. Regions exist where the reduction reaction is so slow, the surface is very planar and ends up being “polished” after the etch.

For isotropic wet etching, a mixture of hydrofluoric acid, nitric acid, and acetic acid (HNA) is the most common etchant solvent for silicon. The concentrations of each etchant determines the etch rate. Silicon dioxide or silicon nitride is usually used as a masking material against HNA. As the reaction takes place, the material is removed laterally at a rate similar to the speed of etching downward. This lateral and downward etching process takes places even with isotropic dry etching which is described in the dry etch section. Wet chemical etching is generally isotropic even though a mask is present since the liquid etchant can penetrate underneath the mask (Figure 2b). If directionality is very important for high-resolution pattern transfer, wet chemical etching is normally not used.

7.3.3 Silicon nitride

Silicon nitride is etched very slowly by HF solution at room temperature. For 20:1 BOE at 20C eact rate of SiO_2 is 300 A/min, eacxt rate of S_3N_4 is 5-15 A/min. Silicon nitide etches in 49% HF at room temperature at about 500 A/min. Usually the phosphoric acid at 150-200C is usewd to etch silicon nitride (100 A/min) whereas etch SiO_2 is 10 A/min, selectivity etching of Si_3N_4 over SiO_2 is 10, selectivity etching of Si_3N_4 over Si is 30 (fig.7.6), eq:

$Si_3N_4+H_3PO_4 +H_2O \rightarrow NO\uparrow +NO_3^-+ H_2PO_4^- +H_2SiO_3$

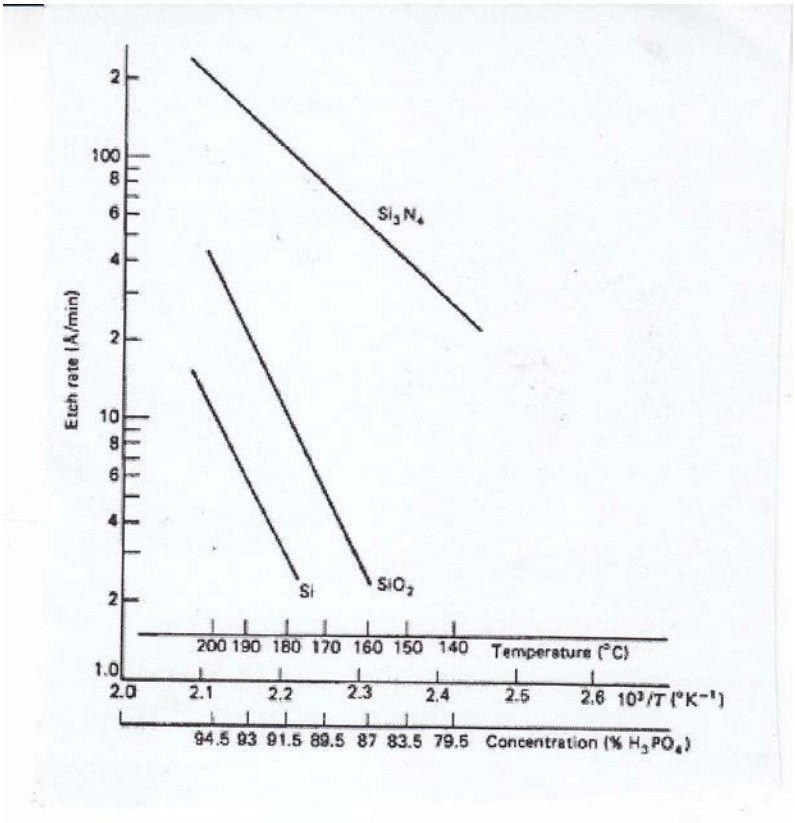


Fig.7.6 Phosphoric acid etch rate of silicon nitride, silicon oxide and silicon.

7.3.4 Aluminum and other materials

Aluminum etches in water, phosphoric, nitric and acetic acid mixtures:

$50H_3PO_4 : 20H_2O : 1HNO_3 : 1CH_3COOH$

The etching process includes the convertation of Al to $Al_2O_3$with nitric acid (evolves H_2), Dissolve Al_2O_3 in phosphoric acid. Since gas evolution leading to bubbles the local etch rate goes down where bubble is formed, leading to non-uniformity of etching. Al can also be etched in (diluted) acid or base, such as HCl, HNO_3, H_2SO_4, NaOH or KOH, but less controllable (etch the native oxide slowly and un-controllably, then once oxide all etched away, etch Al metal very fast).

The tables 7.1, 7.2 present different etching solution for various materials used at IC and MEMS fabrications

Table 7.1 Etching solution for various materials

Etchant	**Etches**	**Doesn't etch**
H_3PO_4(19), Hac(1), HNO_3(1), H_2O(2)	Al, SiN, M	SiO_2, Si, PR
HF, BOE (HF + NH_4F)	SiO_2, M	Si, SiN, Au
H_2SO_4(3), H_2O_2(1) pirahna	Organics, M	Si, SiO_2, SiN
I_2(I),KI(2),H_2O(10)	Au, M	Si, SiO_2, SiN, M, PR
NH_4OH(5), H_2O_2(1)	Polymers, Al	Si, SiO_2, SiN, M
HNO_3(64), NH_4F(3), H_2O(33)	Si, M	SiN, PR
HCl(3), HNO_3(1) (aqua regia)	Au, other M	Cr, Si, SiN, SiO_2

Here, M: metal; PR: photoresist; Hac: acetic acid

Table 7.2 Etching solution for various materials

Table 10–1 Common wet chemical etchants for various thin films used in IC fabrication

Material	Etchant	Comments
SiO_2	HF (49% in water) "straight HF"	Selective over Si (i.e., will etch Si very slowly in comparison). Etch rate depends on film density, doping.
	NH_4F:HF (6:1) "Buffered HF" or "BOE"	About $^1/_{20}$ th the etch rate of straight HF. Etch rate depends on film density, doping. Will not lift up photoresist like straight HF.
Si_3N_4	HF (49%)	Etch rate depends strongly on film density, O, H in film.
	H_3PO_4:H_2O (boiling @ 130–150°C)	Selective over SiO_2. Requires oxide mask.
Al	H_3PO_4:H_2O:HNO_3:CH_3COOH (16:2:1:1)	Selective over Si, SiO_2, and photoresist.
Polysilicon	HNO_3:H_2O:HF (+ CH_3COOH) (50:20:1)	Etch rate depends on etchant composition.
Single crystal Si	HNO_3:H_2O:HF (+ CH_3COOH) (50:20:1)	Etch rate depends on etchant composition.
	KOH:H_2O:IPA (23 wt. % KOH, 13 wt. % IPA)	Crystallographically selective; relative etch rates: (100): 100 (111): 1
Ti	NH_4OH:H_2O_2:H_2O (1:1:5)	Selective over $TiSi_2$.
TiN	NH_4OH:H_2O_2:H_2O (1:1:5)	Selective over $TiSi_2$.
$TiSi_2$	NH_4F:HF (6:1)	
Photoresist	H_2SO_4:H_2O_2 (125°C)	For wafers without metal.
	Organic strippers	For wafers with metal.

7.3 Anisotropic silicon etching

There is a large difference in the etch rate depending on the silicon crystalline plane. In materials such as silicon, this effect can allow for very high anisotropy. Some of the anisotropic wet etching agents for silicon are potassium hydroxide (KOH), ethylenediamine pyrocatechol (EDP), or tetramethylammonium hydroxide (TMAH). Etching a (100) silicon wafer would result in a pyramid shaped etch pit as shown in Figure 7.7a. The etched wall will be flat and angled. The angle to the surface of the wafer is 54.7o. Figure 7.7c-d depicts scanning electron micrographs of (110)-oriented two-dimensional silicon walls with micro and nanoscale dimensions generated based on KOH based wet etching.

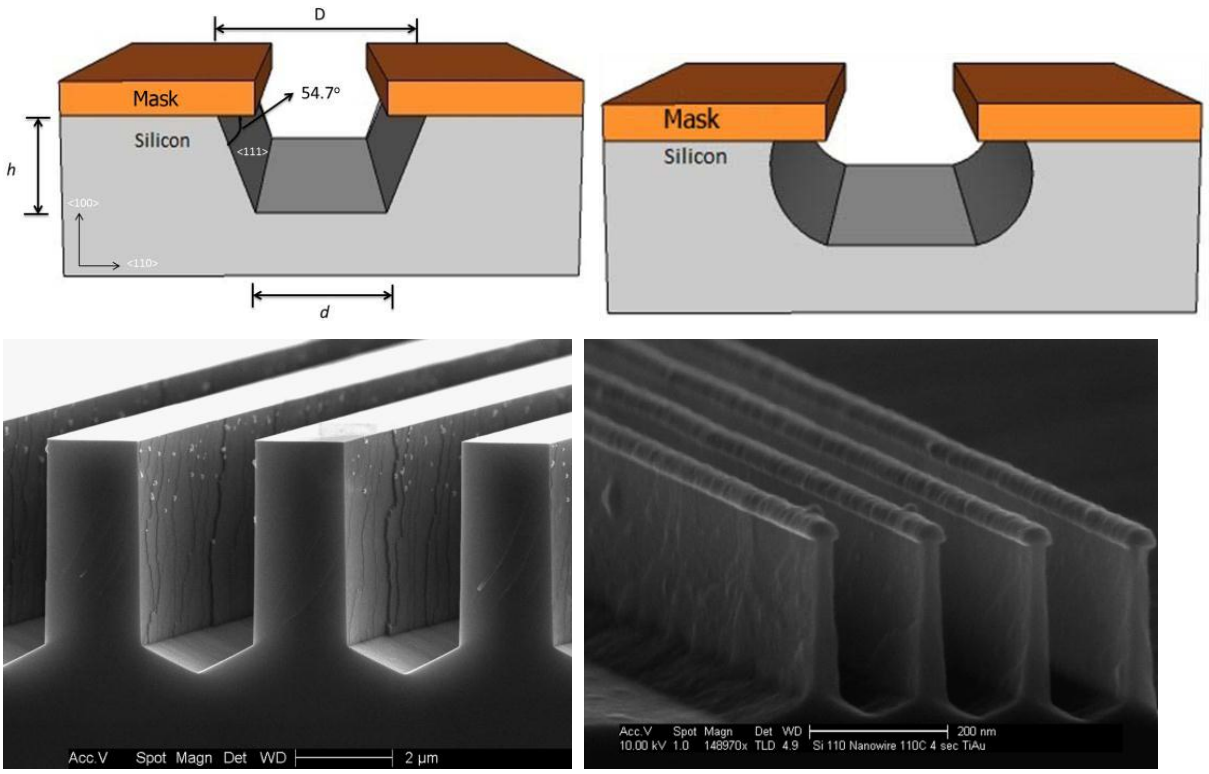


Figure 7. 7 Schematics of an etch profile in (a) an anisotropic and (b) an isotropic etch of a (100) oriented silicon surface. (c-d) KOH based wet etching of (110)-oriented Si surfaces with micro and nanoscale two-dimensional walls.

The relationship between mask dimensions, etch depth and the floor width is given in equation 1

$d = D - 2h/\tan 54.7^{o}$ [1]

Orientation selective etch of silicon occur in hydroxide solutions partly because of the closer packing of some orientations relative to other orientations . In fact:

Density of planes: <111> > <110>, <100>

Etch rate: R(111) << R(110), R(100)

<100> direction etches faster than <111> direction, with etch rate

R(100) = few 100 × R(111) , where reaction rate limited

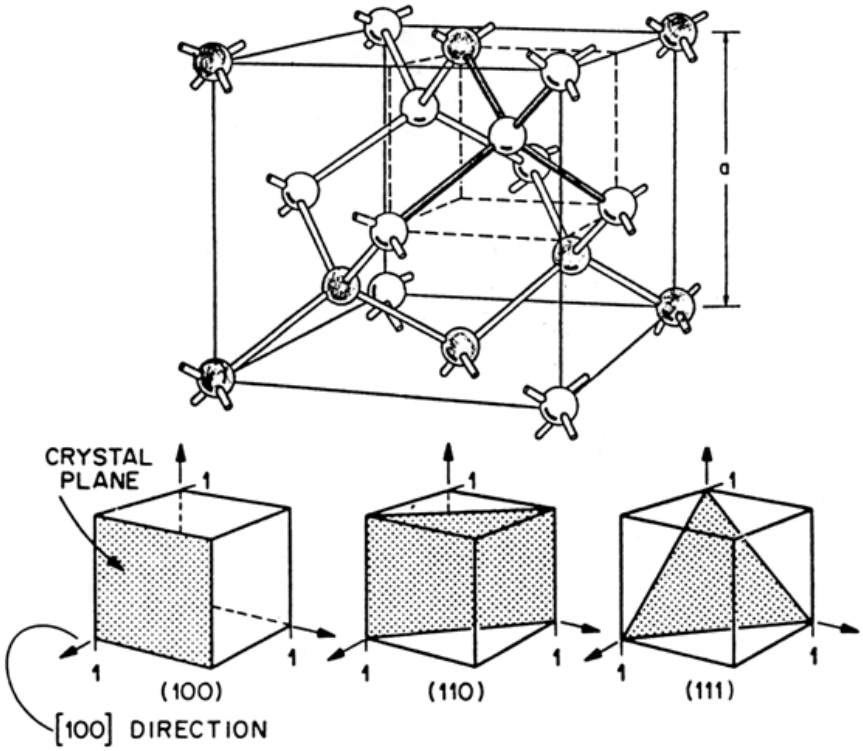


Fig.7. 8 Silicon crystal in [100], [110] and [111] direction.

The planes {100} and {110} have 2 bonds below surface and 2 dangling bonds that can react. {111} plane has three of its bonds below surface and only one dangling bond to react. It results in much slower etch rate.

The most popular anisotropic etch solutions:

1) KOH etch solutions:

250 g KOH: 200 g 2-propanol, 800 g H_2O at 80°C (fig.7.9) , the rate is 1000 nm/min of [100]

Etch stops at p++ layers, Selectivity: {111}:{110}:{100} ~ 1:600:40

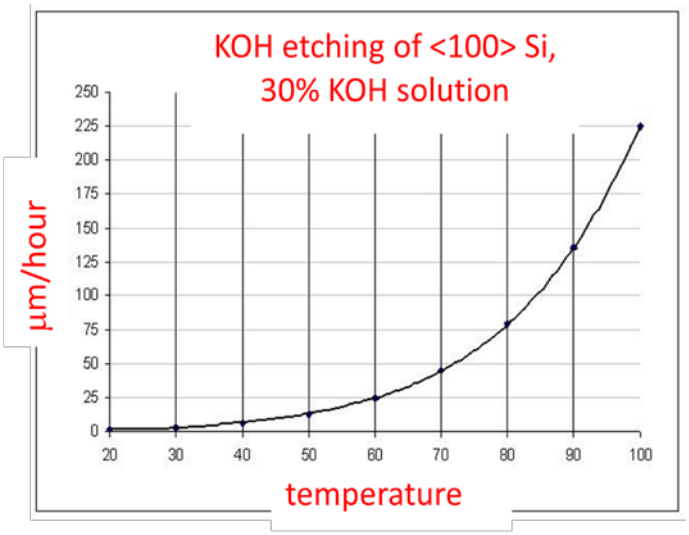


Fig.7.9 KOH etching of <100> Si versus temperature

2) Tetramethyl Ammonium Hydroxide (TMAH)

- Used widely as positive photoresist developer (since it contains no metal like K or Na, which are harmful for device.)
- Typical etching at 80-90^{o}C.
- Etching rate ~0.5-1.5 μm/min (10-40%w.t)
- Selectivity <100> : <111> ~ 10 - 35, much lower than KOH.
- Result in rough surface (H_2 bubble), KOH etch is smoother.
- Like KOH, attacks aluminum
- Like KOH, can use boron-stop-etching technique (etching rate decreases 40 times for $10^{20}/cm^3$ boron doping).
- Excellent selectivity of <100>Si : oxide/nitride (1: 5000-50000)

3) Ethylene Diamine Pyrochatechol (EDP) (table 7.3)

- Typical etching temperature 115^{o}C.
- Etching rate 1μm/min.
- Selectivity of <100>Si : oxide/nitride ~ 3000-7000.
- Doesn't attack metal (Au, Cr, Cu, Ta) but attacks Al.
- Selectivity <100> : <111> ~35; (100) etches faster than (110), ((110) etches faster for KOH).
- Excellent for boron stop technique, etching rate drops 50 times for $7x10^{19}/cm^3$ boron doping.
- Carcinogenic.
-

Table 7.3 Comparison of the etching properties of KOH and EDP

Materials	Etchants	Etch Rates
Silicon in <100>	KOH	0.25 –1.4 μm/min
Silicon in <100>	EDP	0.75 μm/min
Silicon dioxide	KOH	40 – 80 nm/hr
Silicon dioxide	EDP	12 nm/hr
Silicon nitride	KOH	5 nm/hr
Silicon nitride	EDP	6 nm/hr

Rate drops
Harder to etch

Some examples of MEMS preparation on Si substate using the anisotropic etching are presented in figs.7.10-7.12 .

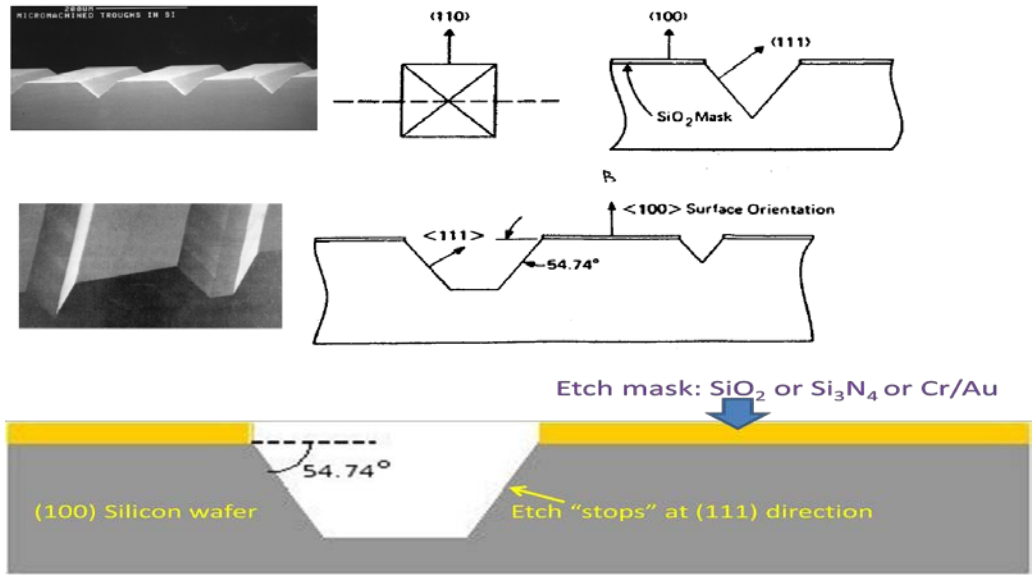


Fig.7.10. Effect of slow {111} etching with KOH: etching virtually stops at {111} plane.

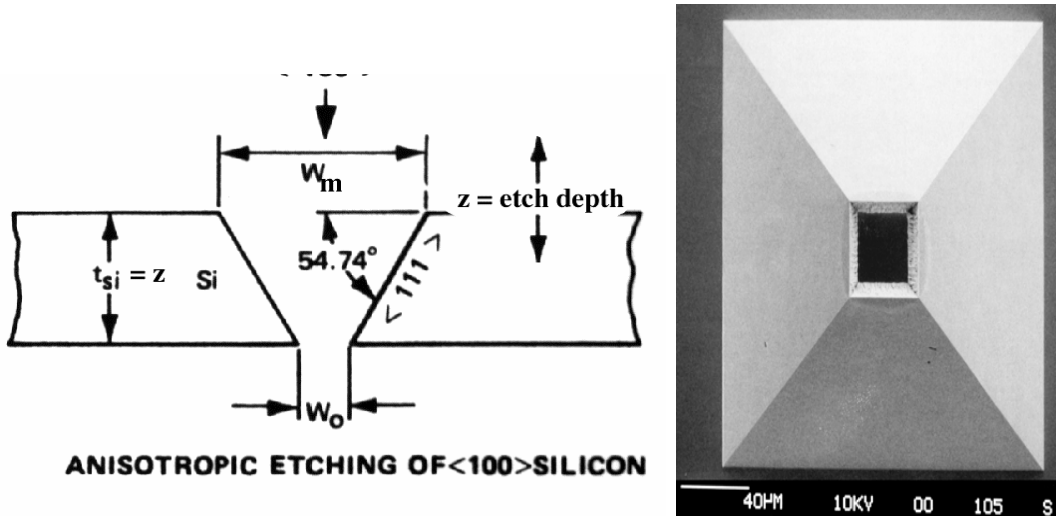


Fig.7.11 bubble-jet printer nozzle base on anisotripic etching of <100> silicon

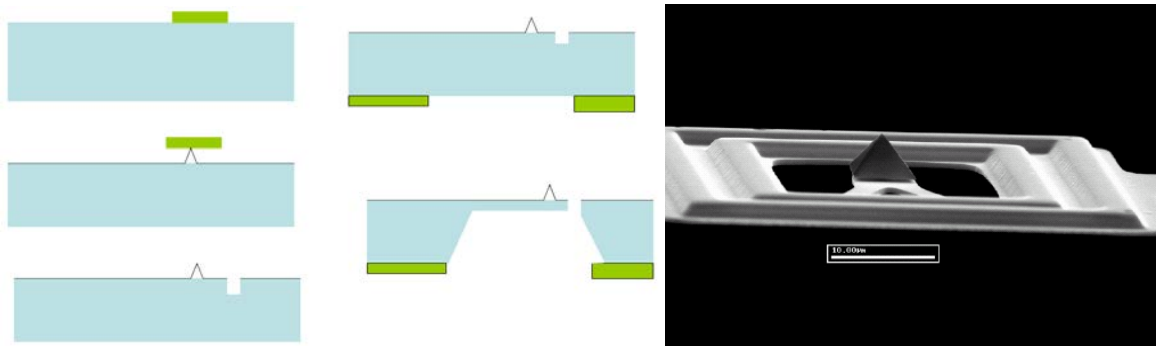


Fig.7.12 Atomic force tip preparation

The bottom of pits in Si substrate are 1) flat ({110} plane) if KOH is used since {100} etches slower than {110} and 2) V-shaped ({100} plane) if EDP is used since {110} etches slower than {100} (fig.7.13).

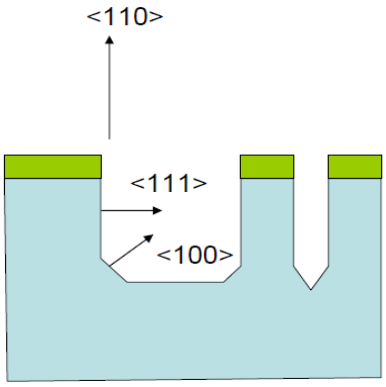


Fig.7.13. Formations of flat and V shaped pits in silicon

7.4 Etch stop process

In wet etching process, etching depth is hard to control, so need etch stop layer. Besides oxide and nitride, etching may be stopped by the following two methods, both related to doping of the silicon substrates.

- Controlled by doping: doped Si dissolved slower than pure Si (fig.7.14).
- Controlled by electrochemical etch stop (fig.7.15).

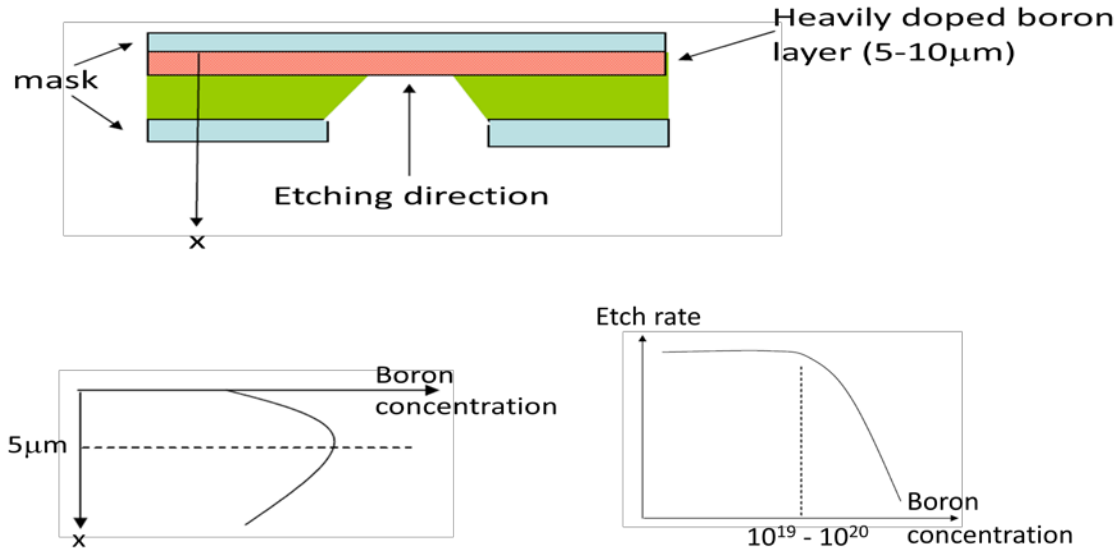


Fig.7.14 Etch stop by boron doping

When silicon is biased with a sufficiently large anodic potential relative to the etchant, it get oxidized due to electrochemical passivation, which then prevents etching. For passivation to occur, current flow is required. So if current flow can be prevented, there will be no oxide growth and etching can proceed. Current flow can be prevented by adding a reverse-biased diode structure (fig.7.15).

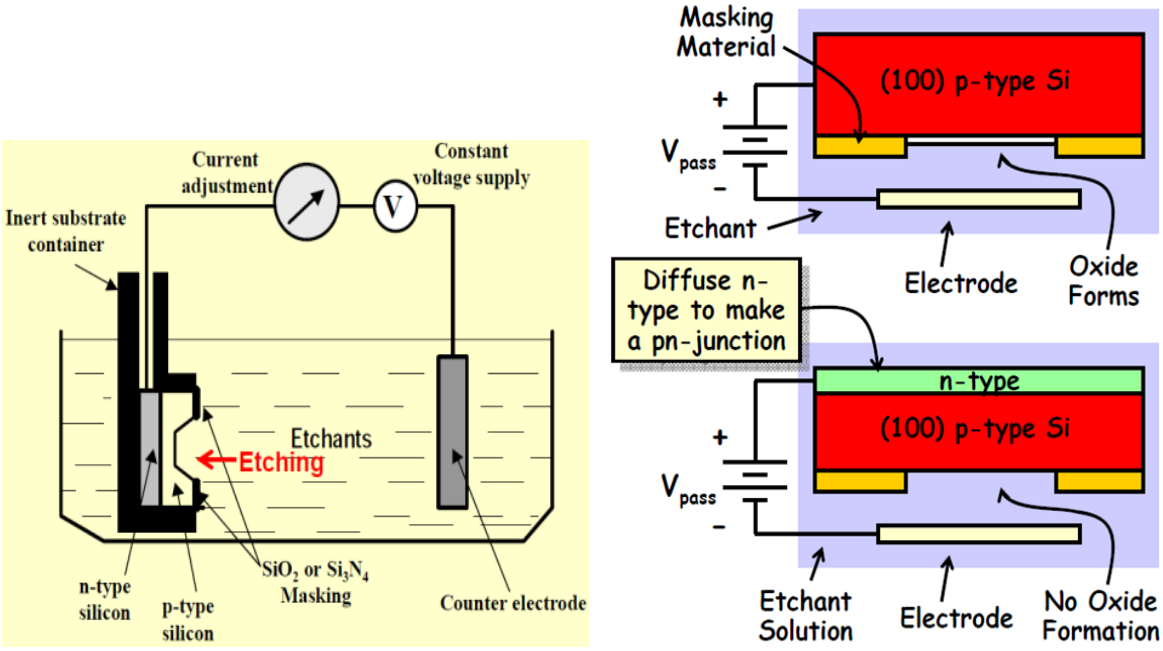


Fig.7.15 Electrochemical etch stop

7.5 Physical dry etching

In *dry etching*, plasmas or etchant gasses remove the substrate material. The reaction that takes place can be done utilizing high kinetic energy of particle beams, chemical reaction or a combination of both. Dry plasma etch works for many dielectric materials and some metals (Al, Ti, Cr, Ta, W…).For other metals, ion milling (Ar^+) can be used, but with low etching selectivity. As a result, for metals that cannot be dry-etched, it is better to pattern them using lift off.

Physical dry etching requires high energy kinetic energy (ion, electron, or photon) beams to etch off the substrate atoms. When the high energy particles knock out the atoms from the substrate surface, the material evaporates after leaving the substrate. There is no chemical reaction taking place and therefore only the material that is unmasked will be removed. The physical reaction taking place is illustrated in Figure 7.15.

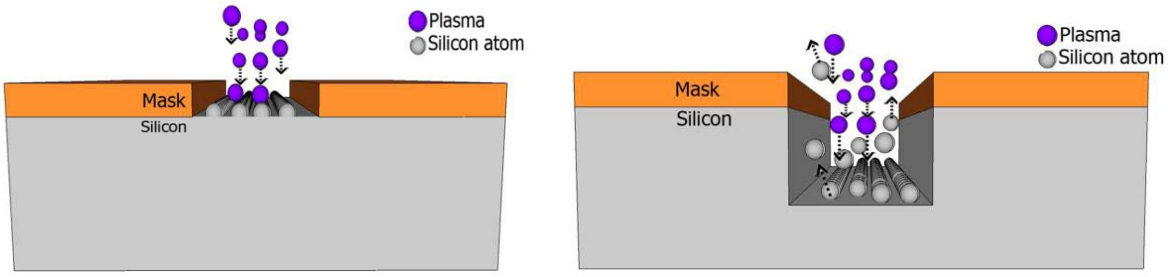


Figure 7.15. The plasma hits the silicon wafer with high energy to knock-off the Si atoms on the surface. (a) The plasma atoms hitting the surface. (b) The silicon atoms being evaporated off from the surface

Chemical dry etching (also called vapor phase etching) does not use liquid chemicals or etchants. This process involves a chemical reaction between etchant gases to attack the silicon surface. The chemical dry etching process is usually isotropic and exhibits high selectively. Anisotropic dry etching has the ability to etch with finer resolution and higher aspect ratio than isotropic etching. Due to the directional nature of dry etching, undercutting can be avoided. Figure 16 shows a rendition of the reaction that takes place in chemical dry etching. Some of the ions that are used in chemical dry etching is tetrafluoromethane (CH4), sulfur hexafluoride (SF6), nitrogen trifluoride (NF3), chlorine gas (Cl2), or fluorine (F2).

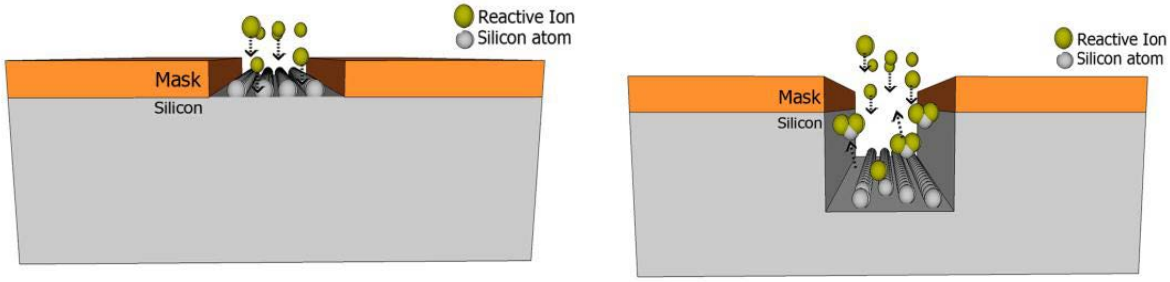


Figure 7.16. Process of a reactive ion interacting with the silicon surface. (a) The interaction between the reactive ion and the silicon atom. (b) A bond between the reactive ion and the silicon atom then chemically remove the silicon atoms from the surface.

Reactive ion etching (RIE) uses both physical and chemical mechanisms to achieve high levels of resolution. The process is one of the most diverse and most widely used processes in industry and research. Since the process combines both physical and chemical interactions, the process is much faster. The high energy collision from the ionization helps to dissociate the etchant molecules into more reactive species.

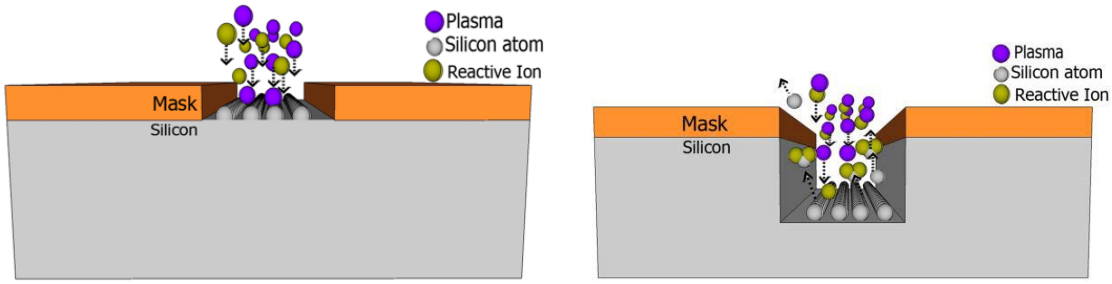


Figure 7.17**.** The RIE process. This process involves both physical and chemical reactions to etch off the silicon.

In the RIE-process, cations are produced from reactive gases which are accelerated with high energy to the substrate and chemically react with the silicon. The typical RIE gasses for Si are CF_4, SF_6 and BCl_2 + Cl_2. As seen in Figure 7.17, both physical and chemical reaction is taking place. Figure 7.18 depicts some micro/nano structures with high aspect ration etched using RIE process.

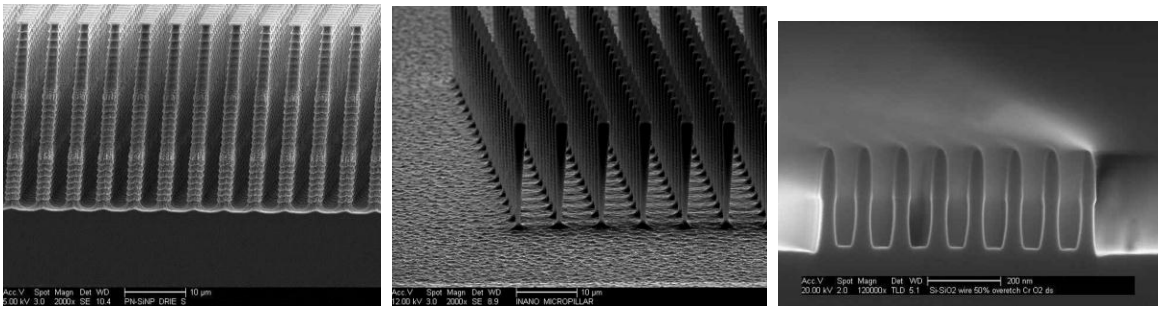


Figure 7.18**.** (a-b) Silicon micro-pillars fabricated using deep reactive ion etching (DRIE). These pillars were made using a Bosch process. The Bosch process is a pulsed-multiplexed etching technique which alternates between two modes to achieve extremely long and vertical micron-scaled nanowires (c) SiO_2 and Si nanowalls etched via RIE process.

Chapter 8. Diffusion

V.Ilchenko

8.1. Basic diffusion process

Diffusion is the process whereby a particle moves from regions of higher concentrations to regions of lower concentrations. The process could be visualized by thinking of a drop of black ink dropped into a glass of clean water. Initially, the ink stays in a localized area, appearing as a dark region in the clean water. Gradually, some of the ink moves away from the region of high concentration, and instead of there being a dark region and a clean region, there is a graduation of colors. As time passes, the ink spreads out until it is possible to see through it. Finally, after a very long time, a steady state is reached and the ink is uniformly distributed in the water. The movement of the ink from the region of high concentration (ink drop) to the region of low concentration (the rest of the glass of water) is an illustration of the process of diffusion.

Doping is a method to control the electrical properties in semiconductors. Doping is achieved by replacing the constituting atoms of the semiconductor with atoms which contain fewer or more electrons. Through doping, the crystal composition is thus slightly altered so that it contains either a higher concentration of electrons or holes, which makes the semiconductor *n*-type or *p*-type, respectively.

The doping of semiconductors can be performed during the bulk crystal or epitaxial film growth by introducing the dopant along with the precursor chemicals. This way, the entire crystal or film is uniformly doped with the same concentration of dopants. Another method consists of carrying out the doping after the film deposition by performing the diffusion of dopants. In this chapter, we will focus on the diffusion of dopants and we will illustrate our discussion with the doping of silicon. The diffusion of dopants in compound semiconductor epitaxial films generally follows a similar model. However, the effects of diffusion doping in compound semiconductor heterostructures are subtler and have been discussed in detail in many specialized papers. In the doping of silicon by diffusion, the silicon wafer is placed in an atmosphere containing the impurity or dopant to incorporate. Because the silicon does not initially contain the dopant in its lattice, we are in the presence of two regions with different concentrations of impurities. At high temperatures (900–1200 °C), the impurity atoms can move into the crystal and diffusion can therefore occur, as schematically illustrated in Fig. 8.1. The wafers are loaded vertically into a quartz boat and put into a furnace similar to the furnace used

for oxidation. There are three types of sources to be used for the dopant atoms: solid, liquid, and gas, as shown in Fig. 8.2.

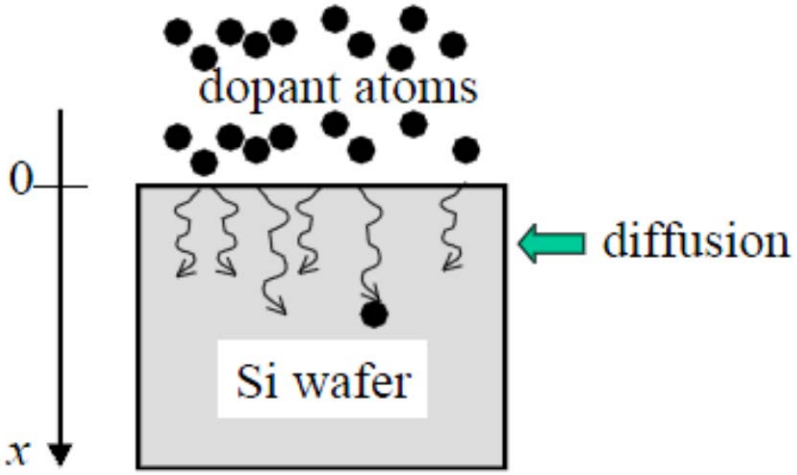


Fig. 8.1. Diffusion of dopants in a silicon wafer. The wafer is placed in an atmosphere containing the dopant. The gradient of dopant concentration between the atmosphere and the silicon crystal leads to their diffusion into the silicon.

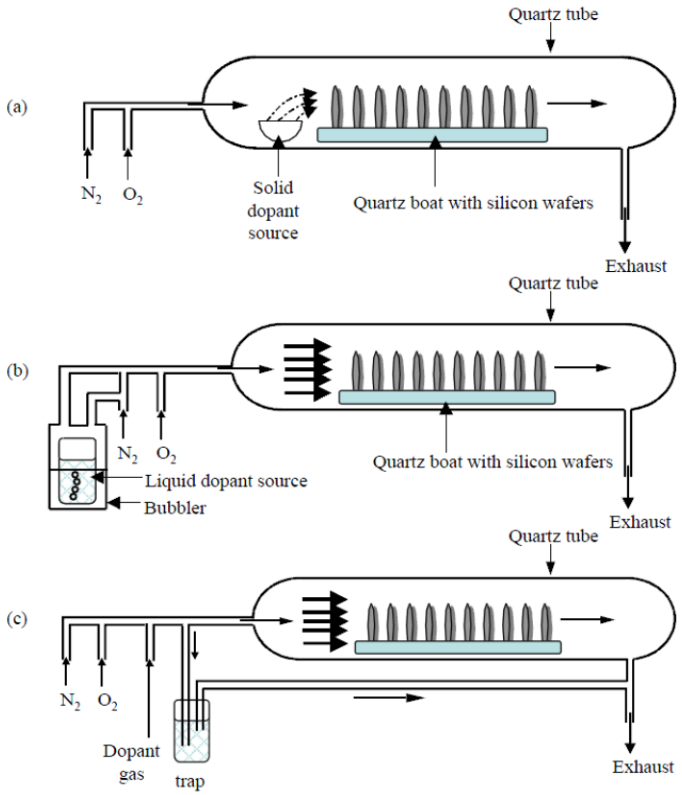


Fig. 8.2. Diffusion furnaces. (a) solid source diffusion with the source in a platinum source boat, (b) liquid source diffusion with the carrier gas passing through the bath, and (c) gas source diffusion with gaseous impurity sources.

There exist several types of diffusion mechanisms. An impurity can diffuse into an interstitial site in the lattice and can move from there to another interstitial site, as shown in Fig. 8.3(a). We then talk about interstitial diffusion Sometimes a silicon atom can be knocked into an interstitial site, leaving a vacancy in the lattice where a diffusing dopant atom can fit, as shown in Fig. 8.3(b). A third possible mechanism consists of a dopant directly diffusing into a lattice vacancy (Fig. 8.3 (c)). We then talk about substitutional diffusion. It is only in the cases that an impurity occupies a vacated lattice site that *n*-type or *p*-type doping can occur. The presence of such vacancies in the lattice can be due to defects or to heat which increases atomic vibrations thus giving enough energy to the silicon atoms to move out of their equilibrium positions into interstitial sites.

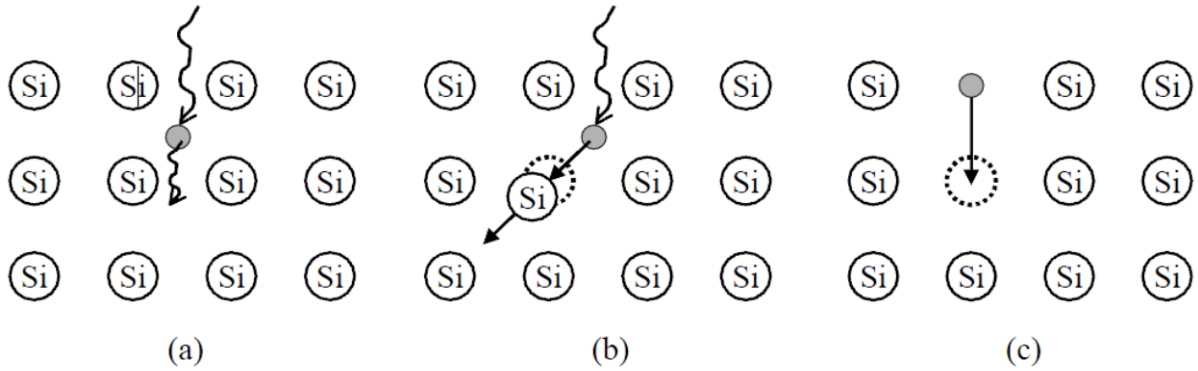


Fig. 8.3. Three possible diffusion mechanisms in a silicon wafer: (a) an impurity moves from one interstitial site to another, (b) a silicon atom is knocked into an interstitial site, thus leaving a vacancy which can be occupied by a diffusing impurity, (c) an impurity diffuses directly into a vacancy.

There are many different types of impurities that can be used for diffusion, the most common being boron, phosphorus, arsenic, and antimony. Table 8.1 lists the reactions for the materials for the three different types of diffusion sources. The rate at which the diffusion of impurities takes place depends on how fast they are moving through the lattice. This phenomenon is quantitatively characterized by the diffusion coefficient of the impurity in silicon. Table 8.2 lists diffusion coefficient values for common impurities in silicon.

Table 8.1. Diffusion reactions for common impurity types.

Impurity	Type	Reaction
Boron	Solid	$2(CH_3O_3)B + 9O_2 \rightarrow B_2O_3 + 6CO_2 + 9H_2O$
	Liquid	$4BBr_3 + 3O_2 \rightarrow 2B_2O_3 + 6Br_2$
	Gas	$2B_2O_3 + 3Si \rightarrow 4B + 3SiO_2$
Phosphorus	Solid	$2P_2O_5 + 5Si \rightarrow 4P + 5SiO_2$
	Liquid	$4POCl_3 + 3O_2 \rightarrow 2P_2O_5 + 6Cl_2$
	Gas	$2PH_3 + 4O_2 \rightarrow P_2O_5 + 3H_2O$
Arsenic	Solid	$2As_2O_3 + 3Si \rightarrow 3SiO_2 + 4As$
Antimony	Solid	$2Sb_2O_3 + 3Si \rightarrow 3SiO_2 + 4Sb$

Table 8.1. Diffusion coefficient and activation energy values for common impurities in silicon.

Element	D_0 ($cm^2 \cdot s^{-1}$)	E_A (eV)
B	10.50	3.69
Al	8.00	3.47
Ga	3.60	3.51
In	16.5	3.9
P	10.5	3.69
As	0.32	3.56
Sb	5.6	3.95

8.1.1. Diffusion Equation

We can then model the diffusion process by combining Fick's first and second law of diffusion:

$$J = -D\left(dN / dx\right) \tag{8.2}$$

$$\frac{\partial N}{\partial t} = D\left(\partial^2 N / \partial x^2\right)$$

The technology of diffusion in semiconductor processing consists of introducing a controlled amount of chosen impurities into selected regions of the semiconductor crystal. To prevent the diffusion of dopants in undesired areas, it is common to use a dielectric mask such as SiO2 to selectively block the diffusion as show in Fig. 8.4. Fig. 8.5 shows a plot of the minimum mask thickness needed for a given diffusion time for boron and phosphorus diffusion.

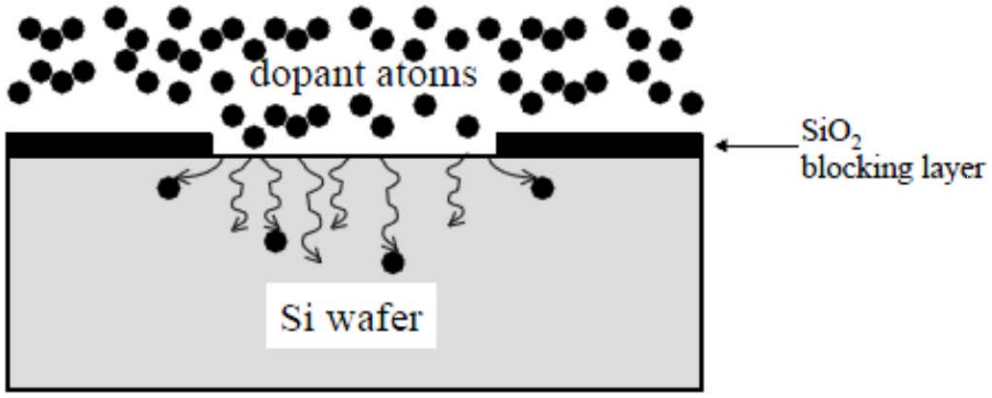


Fig 8.4. Schematic illustration of the selective diffusion in a silicon wafer. The SiO_2 layer acts as a blocking layer for the diffusion of dopant atoms. Some dopant atoms can diffuse laterally under the blocking layer to some extent.

There are two major techniques for conducting diffusion, depending on the state of the dopant on the surface of the wafer: (1) constatnt-source diffusion, also called predeposition or thermal predeposition, in which the concentration of the desired impurity at the surface of the semiconductor is kept constant; and (2) limited-source diffusion, or drive-in, in which a fixed total quantity of impurity is diffused and redistributed into the semiconductor to obtain the final profile.

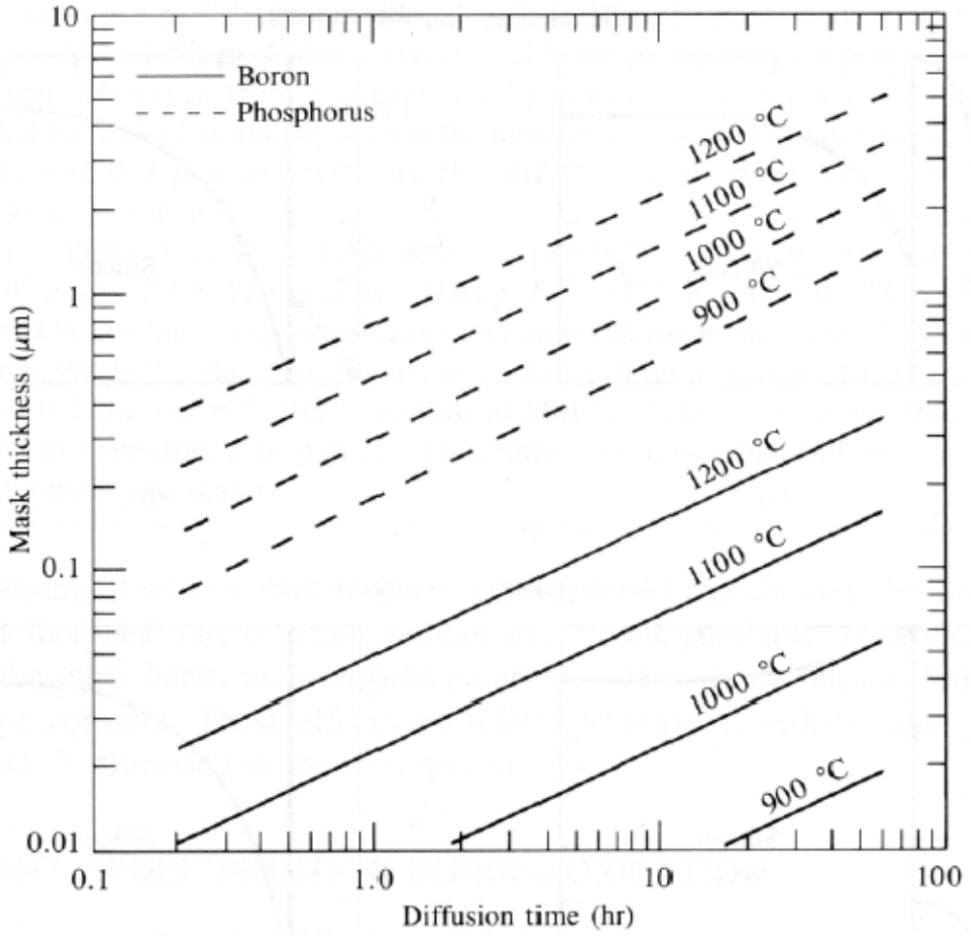


Fig. 8.5. Minimum SiO_2 mask thickness needed for successful diffusion of boron and phosphorus in silicon for a given temperature and time.

8.1.1.1.Constant-source diffusion: predeposition

During predeposition, the silicon wafer is heated to a specific temperature, and an excess of the desired dopant is maintained above the wafer. The dopants diffuse into the crystal until their concentration near the surface is in equilibrium with the concentration in the surrounding ambient above it. At a given temperature, the maximum concentration that can be diffused into a solid is called the solid solubility. Having more dopants available outside the solid than can enter the solid guarantees that solid solubility will be maintained during the predeposition. For example, the solid solubility of phosphorus in silicon at 1000°C is 9×10^{20} atoms/cm^3, while for boron in silicon at the same temperature it is only 2×10^{20} atoms/cm^3. These values only depend on the temperature, for a given dopant in a given semiconductor. As a result, the substrate temperature also determines the concentration of the dopant at the surface of the crystal wafer during diffusion. Under predeposition conditions, let us denote N_0 the dopant concentration in

the wafer near the surface. N_0 would be equal to the solid solubility of the dopant at the predeposition temperature if the excess dopant in the ambient above the wafer is sufficient. The concentration of dopant in the crystal at a depth x below the surface and after a diffusion time t can be known and is equal to:

$$N(x,t) = N_0 erfc\left(x / 2\sqrt{Dt}\right), \tag{8.3}$$

obtained for the boundary conditions

$$N(0,t) = N_0 = const, \tag{8.4}$$

where D is the diffusion coefficient of the dopant at the predeposition temperature and ***erfc*** refers to the complementary error function. The complementary error function is found by complementing the integral of the normalized Gaussian function, and is shown in Fig. 8.6:

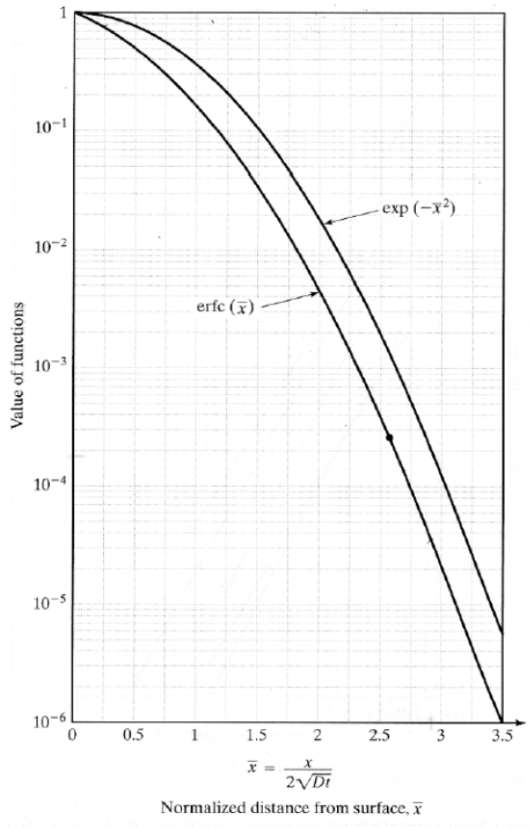


Fig.8.6. Complementary error function, used in the calculation of the dopant concentration.

$$erfc(\bar{x}) = 1 - erf(\bar{x}) = \frac{2}{\sqrt{\pi}} \int_{\bar{x}}^{\infty} e^{-t^2} dt \tag{8.5}$$

The shape of the dopant concentration function is shown in Fig. 8.7 for several values of the product *Dt*. We see that, as the diffusion coefficient increases or, equivalently, as the diffusion time increases, the dopant reaches deeper into the crystal. The surface concentration remains the same at *N0*. The concentration *NB* represents the background carrier concentration and refers to the concentration of majority carriers in the semiconductor before diffusion. The value of x for which N(x,t) is equal to N_B is conventionally termed the junction depth.

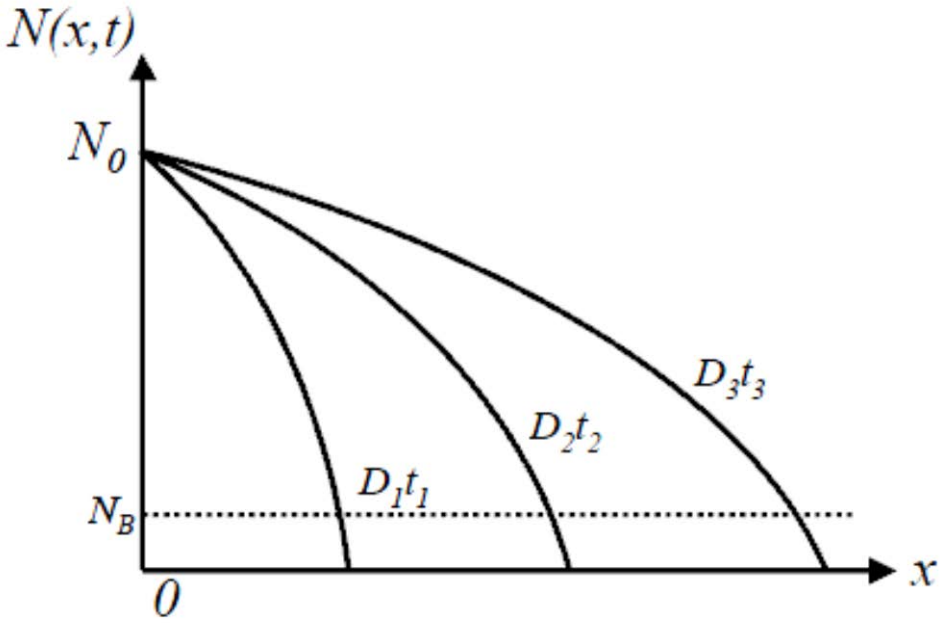


Fig. 8.7. Semi-logarithmic graph of the complementary error function, representing the dopant concentration in the crystal during predeposition, where the surface concentration is kept constant, for several values of Dt: $D_3t_3 > D_2t_2 > D_1t_1$. As the diffusion coefficient and/or the diffusion time is increased, the dopant reaches deeper into the crystal.

The total amount of impurities Q introduced per unit area, also called the dose, after a diffusion duration t in a predeposition process is found by integrating the function in Eq. (8.3) for values of x>0, which leads to the following expression:

$$Q(t) = N_0 \sqrt{\frac{4Dt}{\pi}}.$$

8.1.1.2. Limited-source diffusion: drive-in.

Unlike predeposition, the drive-in diffusion process is carried out with (8.6) amount of impurity. This method allows us to better control the resulting doping profile and depth, which are important parameters in the fabrication of semiconductor devices. During drive-in, the parameters which can be controlled include the duration of diffusion, the temperature, and the ambient gases. The dopant concentration profile of the drive-in has the shape of a Gaussian function, as shown in Fig. 8.8. In this type of diffusion, the dose remains constant causing the surface concentration to decrease. This relationship explains the shape of the curve, which can be expressed by solving Eq. (8.7) and using the boundary condition that the impurity concentration

$$N(x,t) = \frac{Q}{\sqrt{\pi}\sqrt{Dt}} \exp\left(-\frac{x^2}{4Dt}\right) \tag{8.7}$$

at the surface is equal to the dose:

which is expressed in units of atoms per unit volume. D is the diffusion coefficient of the impurity at the drive-in temperature and t is the drive-in time. The drive-in can be performed after a predeposition step, in a high temperature diffusion furnace once the excess dopant remaining on the surface of the wafer from the predeposition step has been removed. In this case, Q is the total dose introduced into the crystal during a predeposition step.

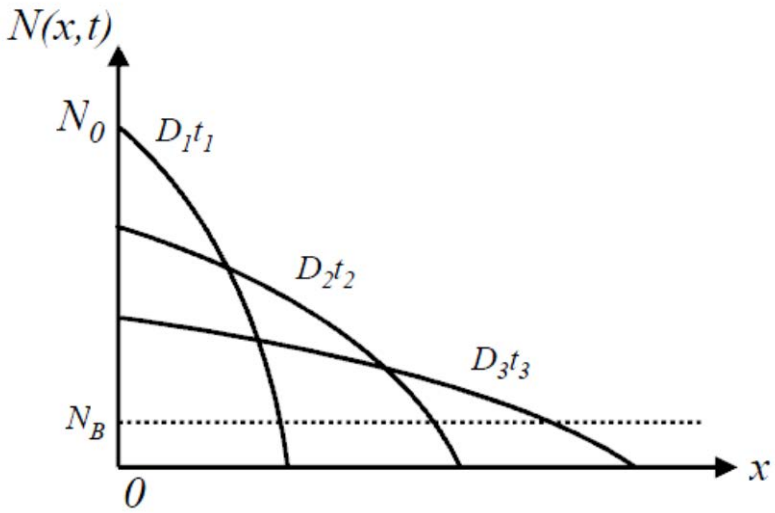


Fig. 8.8. Semi-logarithmic graph of the dopant concentration in the crystal during drive-in for several values of Dt: $D_3t_3 > D_2t_2 > D_1t_1$. As the diffusion coefficient and/or the diffusion time is increased, the dopant reaches deeper into the crystal. At the same time, the concentration at the surface is reduced because the drive-in is a limited-source diffusion process.

8.1.2. Diffusion profiles

When diffusing *p*-type impurity dopants in an originally *n*-type doped semiconductor, a *p*-*n* junction can be formed, as shown in Fig. 2.15. In fact, the purpose of most diffusion processes is to form a *p*-*n* junction by changing a region of an *n*-type semiconductor into a *p*-type or vice versa. Let us consider the example of an *n*-type doped silicon wafer which exhibits a background concentration *NB*, and a *p*-type diffusing impurity with surface concentration *N0*. Where the diffusing impurity profile concentration intersects the background concentration *NB*, a metallurgical junction depth, x_j, is formed as shown in Fig. 8.9.

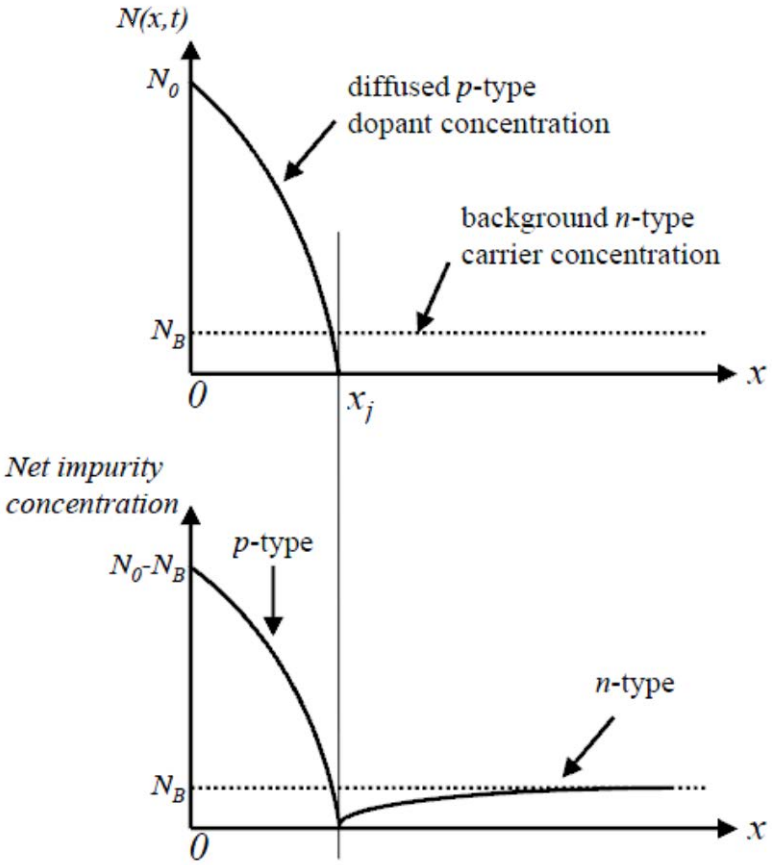


Fig. 8.9. Semi-logarithmic graphs illustrating the formation of a p-n junction through diffusion. A p-type dopant is diffused into an n-type semiconductor which has a background concentration of NB. The p-type dopant concentration profile after diffusion is shown in the top graph. The p-n junction will occur where the p-type dopant concentration is equal to the n-type background concentration as shown on the bottom graph.

At the metallurgical junction depth, the background concentration is equal to the surface concentration, so the net impurity concentration is zero. In the predeposition process with a

complementary diffusion profile, the junction depth is found by solving Eq. (8.3) and using the boundary condition that $N(x_j,t) = N_B$:

$$x_j = \left(2\sqrt{Dt}\right) erfc^{-1}\left(\frac{N_B}{N_0}\right), \tag{8.8}$$

where ***erfc*$^{-1}$** refers to the reciprocal function of the complementary error function. In the drive-in process with a Gaussian diffusion profile, the junction depth is found by solving Eq. (2.34) and using the boundary condition that $N(x_j,t) = N_B$:

$$x_j = 2\sqrt{Dt\ Ln\left(\frac{N_0}{N_B}\right)}. \tag{8.9}$$

By successively diffusing two impurities of different types into an originally doped wafer, more complex structures can be achieved, such as for example an *n*-*p*-*n* transistor structure as illustrated in Fig. 8.10. The starting wafer would be an *n*-type (with a background concentration *NC* in this example), the first diffusion process would introduce *p*-type dopants (N_B in this example) and the second diffusion would introduce *n*-type impurities (N_E) such that $N_E >> N_B >> N_C$.

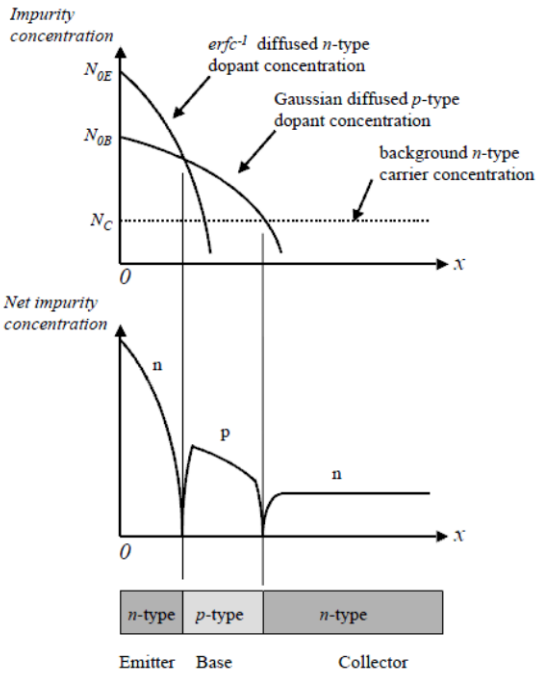


Fig. 8.10. Semi-logarithmic graphs illustrating the formation of an n-p-n transistor through diffusion. p-type dopants are first diffused into an n-type semiconductor to form the first junction. n-type dopants are subsequently diffused to form the second junction.

8.2. Extrinsic diffusion

In this section, we will classify and characterize point defects in crystalline solids as far as this is relevant for diffusion processes in semiconductors. Let us discuss the various types of point defects. silicon. We distinguish between *instrinsic* and *extrinsic* point defects. Contrary to extrinsic point defects, intrinsic point defects do not involve foreign atoms. The most simple and basic intrinsic point defects in crystals are vacancies (V) and self-interstitials (I). Vacancies are unoccupied lattice sites, whereas self-interstitials are additional atoms (of the same species the crystal consists of) squeezed-in between the atoms on lattice sites.

Extrinsic point defects are foreign atoms which may either be incorporated into *interstitial* sites between the lattice atoms, such as Li or Fe in silicon or on *substitutional* lattice sites such as the Group III elements (e.g. B, AI, Ga) or the Group V elements (P, As, Sb) commonly used as acceptors or donors in silicon devices. The isoelectronic Group IV elements such as C, Ge, or Sn are substitutional dissolved in a neutral charge state. Some elements. A, such as Au or Pt in silicon, Cu in germanium, or Zn in GaAs may exist in both *interstitial* (A_i) and *substitutional* sites (A_s) which gives rise to peculiar diffusion mechanisms discussed in later sections. In compound semiconductors, which usually consist of two sublattices, the number and configurations of point defects increases accordingly. In the *interstitialcy* (or *indirect interstitial)* mechanism a self-interstitial replaces a substitutional atom which then in return replaces a neighboring lattice atom (Fig. 8.11, 8.12)

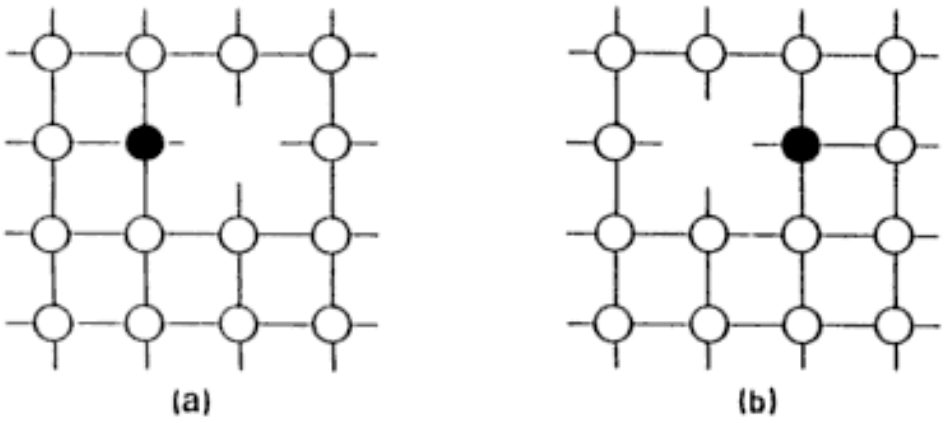


Fig. 8.11. Vacancy mechanism (schematic).

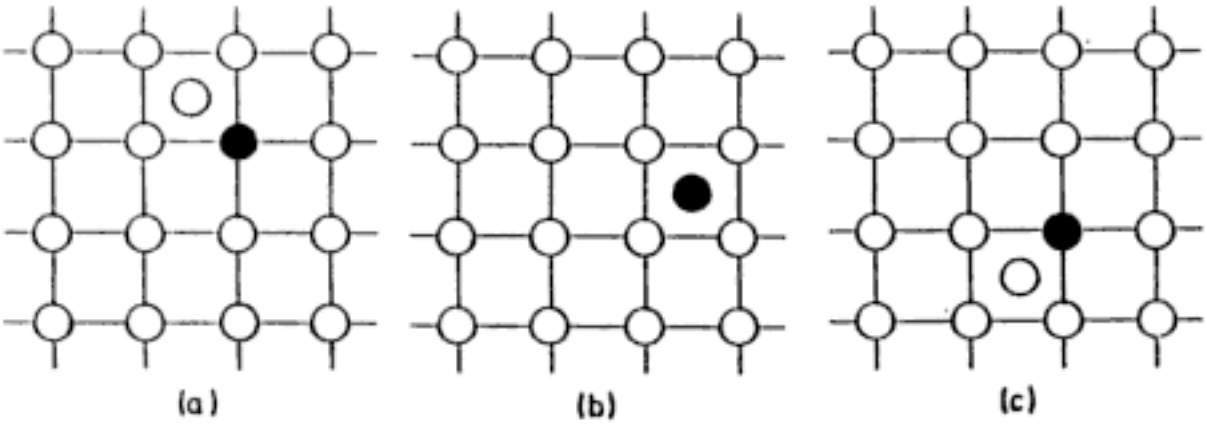


Fig. 8.11. Interstitialy mechanism (schematic).

In the *interstitialcy* (or *indirect interstitial)* mechanism a self- interstitial replaces a substitutional atom which then in return replaces a neighboring lattice atom. Attractive or repulsive elastic or coulombic forces between intrinsic point defects and substitutionally dissolved foreign atoms may lead to higher or lower diffusivitics of these atoms than realized for the lattice atoms themselves. The higher diffusivities of all Group III and Group V dopants in silicon as compared to silicon self-diffusion (diffusion of Si in Si), is attributed to fast moving dopant-point defect complexes. At sufficiently high temperatures, a dynamical equilibrium exists between unbound intrinsic point defects and those tight-up in various charge states, the diffusivities of which depend on the specificcharge state.

For a given Fermi-level the diffusivity D^s of a substitutionally dissolved element migrating via the vacancy mechanism will be proportional to the concentration C_v of vacancies, which may serve as *"diffusion vehicles"*

Analogously

$$D^s \propto C_V \tag{8.10}$$

holds for an interstitialcy mechanism. Under thermal equilibrium conditions intrinsic point

$$D^s \propto C_I \tag{8.11}$$

defects will be present in their equilibrium concentrations $C_V = C_V^{eq}$ and $C_I = C_I^{eq}$. Deviations from these values, leading to *enhanced* or *retarded* diffusion processes, may be introduced by a number of non-equilibrium phenomena. Examples arc: (i) particle irradiation with a sufficiently high energy to create vacancy-self-interstitial pairs ("Frenkel pairs") such as ion implantation or proton irradiation, (ii) surface treatments such as oxidation or nitridation of

silicon surfaces, (iii) formation of precipitates associated with large volume changes, such as oxygen precipitation in oxygen-rich silicon, or (iv) cooling of crystals during crystal growth.

Some elements A may be dissolved on interstitial (A_i) and also on substitutional sites (A_s). Let us consider cases in which the solubility on substitutional sites (C_s^{eq}) is much higher than that on interstitial sites (C_i^{eq}) and in which the difusivity D_i or A_i is much higher than that of A_s by a normal vacancy mechanism. Under these circumstances the movement of A atoms may be accomplished by fast diffusion as A_i interstitials and their subsequent change-over to substitutional position. The interchange between A_i and A_s requires intrinsic point defects. Two atomistic realizations of this interchange have been suggested. The *Frank-Turnbull, dissociative,* or *Longini* mechanism,

$$A_i + V \leftrightarrow A_s, \tag{8.12}$$

involves vacancies, whereas the ***kick-out*** mechanism,

$$A_i \leftrightarrow A_s + I, \tag{8.13}$$

involves self-interstitials. Both mechanisms are schematically shown in Fig. 8.12. If charged species are involved in (8.12) or (8.13), the appropriate number of electrons or holes has to be added.

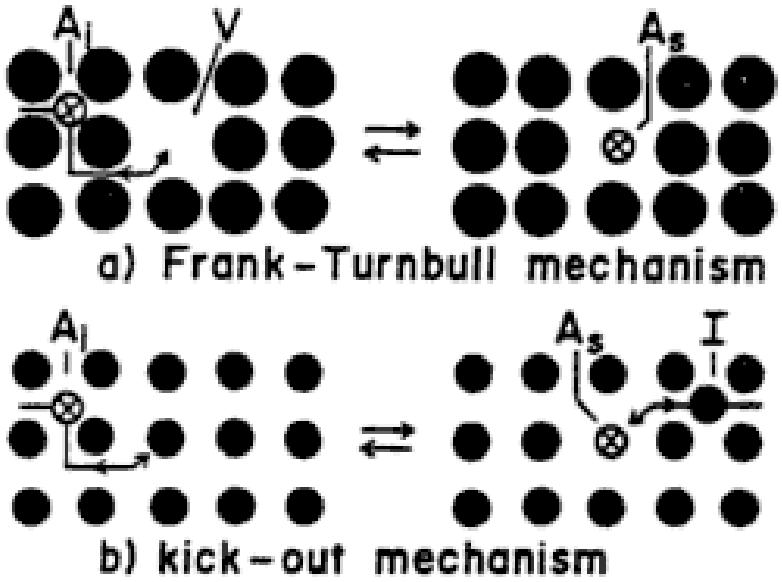


Fig. 8.12. Frank-Turnbull and kick-oul mechanism.

In understanding the differences between the various diffusion mechanisms, it is helpful to compare the kick-out mechanism to the interstitialcy mechanism. In the interstitialcy mechanism in its pure form (Fig. 8.11) the interstitial position is just a transition state for the

diffusing atom during moving from one substitutional site to a neighboring one. In the case of a highly mobile complex between the substitutional atom and a self-interstitial, this complex will also move many lattice constants before decaying and releasing a self-interstitial and the atom to a substitutional position. This diffusion mode is closely related to the kick-out mechanism. The only difference is that in the kick-out mechanism the moving entity is truly interstitial and can therefore be expected to move faster than a complex. For the mathematical description of the diffusion process and many physical effects this does not make an essential difference. For example, a self-interstitial supersaturation ($C_i > C_i^{eq}$) will enhance a kick-out diffusion process by shifting the balance in (8.13) more to the mobile $A_;$ configuration as well as the diffusion of a substitutionally dissolved atom migrating via the interstitialy mechanism according to Eq. (8.11). A comparison of the vacancy mechanism with the Frank-Turnbull mechanism shows significant differences even if highly mobile complexes are considered.

8.3. Lateral diffusion

There are two serious problems which can arise during dopant diffusion through a mask: (i) penetration of the dopant under the edge of the mask by lateral diffusion; and (ii) diffusion through the mask itself. Both of these effects can lead to the distortion of the device structures. Lateral diffusion will always occur because diffusion is a roughly isotropic process in semiconductor materials. Clearly a metallurgical p-n junction will be formed at approximately the same distance laterally under the mask as it is deep in the semiconductor wafer. This effect limits the packing density of devices which can be achieved, and in particular diffusion doping cannot be used to fabricate devices with very narrow gates. If the diffusion mask is made of silicon nitride, then the semiconductor material under the mask can be under considerable strain and this strain can result in a considerable enhancement in dopant diffusion rates and even worse lateral penetration (Lin *et al* 1979).

Analog bipolar circuits employ both pnp and npn transistors, together with junction diodes, resistors and sometimes small-value capacitors. Precise fabrication and layout details depend very largely upon the individual IC manufacturers, and although not generally disclosed in detail, are based upon the fundamentals introduced. There are, however, two distinct methods of achieving isolation between the devices fabricated on a wafer. The first, and oldest, method is the junction- isolated process, which relies upon the presence of a reverse-biased p-n junction barrier between adjacent devices to achieve isolation. The second and more recent method is silicon dioxide isolation, which provides an insulating barrier between devices. Figure 8.13 illustrates the cross-sectional arrangement of a bipolar transistor with junction isolation. This process is sometimes termed the Standard Buried Collector (SBC) process, since it requires a

buried area of n' silicon to be first made below each transistor on the p-type substrate, which will form a low resistance subcollector contact area for each transistor. The seven process steps are as follows: appropriate masks are used at each step to define each window, the silicon dioxide protecting layer at the surface being reformed between steps when necessary:

- Using mask No. 1. open window in ihe silicon dioxide (SiO_2) protecting layer on the lightly doped p-type substrate, and diffuse in a n^+ buried subcollector.
- Remove the remaining SiO_2 and grow a lightly doped n-type silicon epitaxial layer over the whole wafer.
- Regrow the SiO_2 and open a new window using mask No. 2 to diffuse in a deep p^+ trench ("moal") around the perimeter of the device so as to form an isolated island.
- Regrow the SiO_2 and open a new window using mask No. 3 to diffuse in the p-type base area, leaving a thin n-type collector width in the epi. layer between base and buried subcollector.
- Regrow the $SiO_;$ and open a new window using mask No. 4 to diffuse in the n^+ transistor emitter area and also the collector contact.
- Regrow the SiO- and open a new window using mask No. 5 for contacts.
- Cover with interconnection metal, and then etch away using mask No.5 so as to form the final required interconnection pattern.

With the substrate of this structure connected to the most negative voltage of the circuit, the pn junction around each device formed by the p-type moat will be reverse-biased, and therefore will provide device isolation. The n^+ buried layer also prevents the formation of parasitic vertical pnp transistors which may other-wise occur between base, collector and substrate regions.

While still ocasionally used, this SBC process has drawbacks in that the active area of the transistor is only the area below the emitter, which is within the base area, and the base area is itself within the collector area. Also due to lateral diffusion, the minimum width of the p^+ isolation moal will be about twice the depth of the epitaxial layer, so that in total the useful active area of the transistor is often less than 5% of the total device area. Figure 8.13 illustrates this shortcoming, the active area under the emitter being only 2.67% of the total device area.

Minimum feature size	5 μm
Worst-case alignment tolerance between levels	2 μm
Epitaxial-layer thickness	10 μm
Collector-base junction depth	5 μm
Emitter-base junction depth	3 μm
Minimum emitter-to-collector spacing at surface	5 μm
Minimum base-to-isolation spacing at surface	5 μm
Minimum collector contact n^+ diffusion to isolation spacing	5 μm
Minimum collector contact n^+ diffusion to base spacing	5 μm
Buried-layer diffusion (both up and down)	2 μm
Buried layer to isolation spacing	5 μm
Lateral diffusion = vertical diffusion	

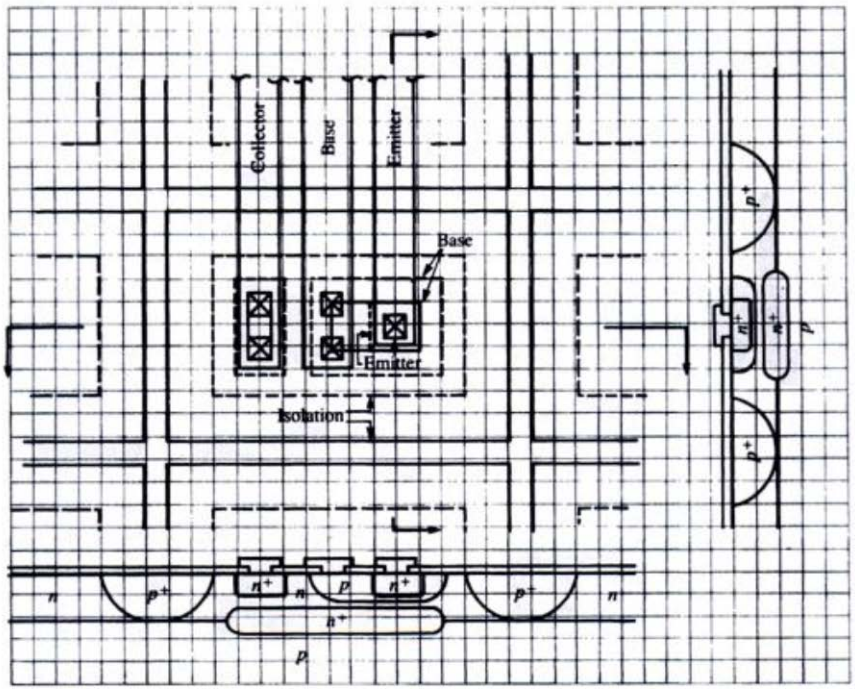


Fig. 8.13. An example of minimum area bipolar transistor layout on a 5 (im grid with a given feature size of 5 цт. The solid lines in the plan view represent the edges of the step-and-repeat mask patterns, with the dotted lines representing the final lateral spread of emitter, base and collector regions assuming equal vertical and lateral diffusion distances. Note that the isolation area occupies over 60% of the total transistor area, illustrating the inefficiency of isolation in this early fabrication process.

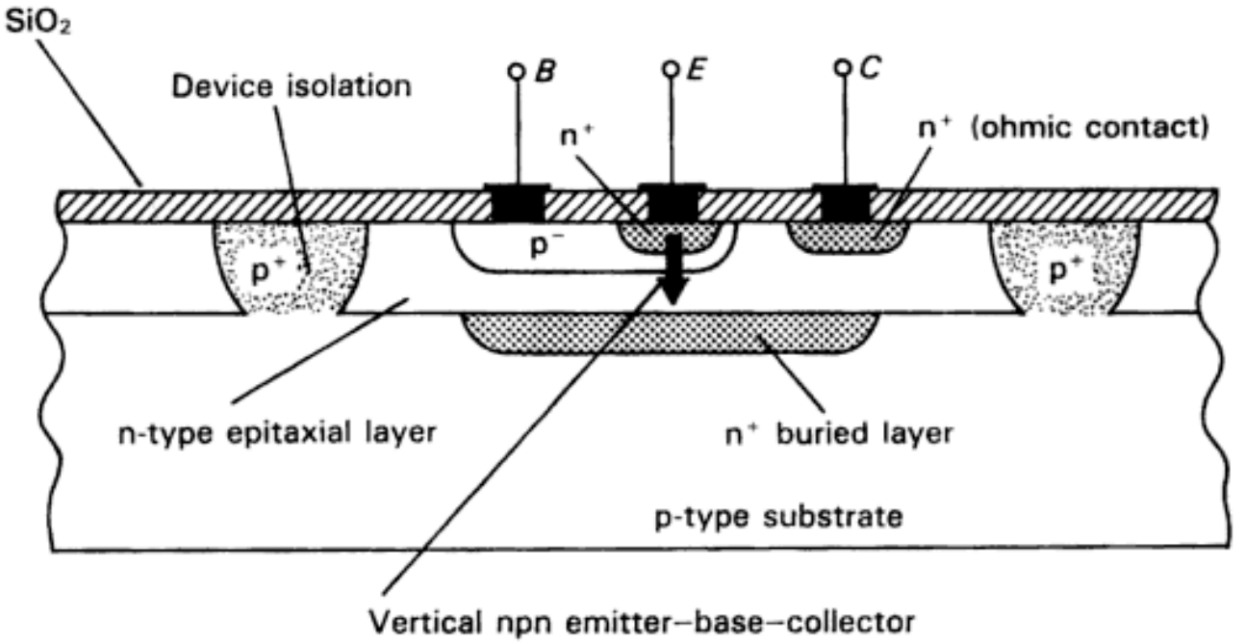


Fig. 8.14. The cross-sectional fabrication details of an SBC bipolar npn junction transistorwith diode isolation (not to scale).

The use of silicon dioxide isolation instead of pn junction isolation is an example of MOS fabrication techniques being applied to bipolar technology. This allows a considerable decrease in transistor area compared with junction isolation, and hence an increase in device density and operating speed.

The cross-section of a typical oxide-isolated npn transistor is shown in Figure 8.14. The first fabrication steps of making the buried n^+ region and the growth of the epitaxial layer remain as previously described. The next three steps, however, to create the SiO_2 isolation boundaries are as follows:

- A thin SiO_2 layer followed by a thick protecting Si_3N_4 (silicon nitride) layer is grown over the whole surface, the latter then being lightly oxidized. (If Si_3N_4 is formed directly on the epitaxial surface it causes surface damage due to differing thermal expansions, and hence ihe need for the thin interleaving SiO_2 layer.)
- Windows where oxide-isolation boundaries are required are cut through the Si_3N_4 and the SiO_2, followed by an etch which is allowed to dissolve away about half ihe thickness of the exposed epitaxial layer.
- A boron implant to form a p" region is then made on ihe surface of all these etched boundaries, followed by a long high-temperature cycle to grow a thick SiO_2 layer in these boundaries.

Since SiO_2 occupies about twice the volume of the silicon from which it is produced, the effect of the last step is to cause the SiO_2 to grow deeper into and fill the cuts in the epitaxial layer. The SiO_2 growth finally reaches the silicon substrate, with the p^+ implant being driven ahead of it.

The subsequent processing stages window and implant the n^+ and p^+ regions into the epitaxial layer after removal of the thick Si_3N_4 layer. However, unlike the SBC process where the emitter region is within the base and collector regions, the base, emitter and collector regions of the oxide-isolated junction transistor are side by side. Also in the detailed fabrication steps, the emitter, base and contact regions are self-aligned, that is, they are positioned by the windows made in the SiO; rather than relying upon the accuracy and positioning of separate masks of the areas required—this is a carry-over from CMOS processing and will be mentioned again later. The total effect of this method of bipolar fabrication is to make the transistor size considerably smaller than the SBC process, giving better performance and much higher packing density.

Different manufacturers have different variations of this oxide-isolated process. Polysilicidc, a low resistance form of polysilicon, may be used instead of metal for the interconnections into the base, emitter and collector regions, another technique which has been copied from MOS technology.

Yet another way that has been pursued to provide device isolation in bipolar technology is ”deep-trench” isolation. The deep trenches, which reach down to the substrate, are cut by a reactive-ion etching process, which produces a sharp etch with minimum side-spreading. The surfaces of these trenches are then oxidized to produce the isolating SiO_2, and the remaining volume is then filled with polysilicon. This isolation process is done after a SiN_4 protective layer has been applied and windowed, as in the oxide-isolated process described above.

The above illustrations each show npn transistors. The alternative polarity pnp devices are available, but inherently do not have the same maximum performance as npn since the majority carriers in a pnp device are holes rather than electrons. In the SBC process the pnp transistors are frequently lateral devices, with the pnp action lying along the plane of the device, rather than vertically as in Figure 8.14; in oxide-isolated and deep-trench fabrication the n" buried layer forms part of the base rather than the collector. The remaining devices required for analog circuit designs, namely diodes, resistors and possibly capacitors, are all available in silicon technology. Diodes are frequently and conveniently formed by using one of the p-n junctions of a npn transistor, either the collector-to-base junction with the base possibly shorted to the emitter, or the base-to-emitter junction with possibly the collector shorted to the base. In general, therefore, the silicon area occupied by a diode is the same as for a transistor.

Chapter 9. Ion implantation

V.Skryshevsky

9.1 Introduction

Ion implantation is the most common method of semiconductors doping. Ion implantation consists of introducing charged atoms (*projectiles*) into a material (the t*arget),* by communicating to them sufficient energy so that they enter beyond the surface area. The projectiles' energy is thus clearly different from those of techniques aiming for a surface process or deposition (plasma, adsorption). Moreover, the implantation is distinguished from these other methods by the purity of the beam, selected at the isotopic level.

Nowadays, there are thousands of ion implantation machines functioning in this industry. Most of them are devoted to the production of integrated circuits based on silicon. The method, which in the 1970s definitively led to ion implantation supplanting diffusion in the manufacturing technology of microelectronics devices, is the adjustment of the threshold voltage of MOS transistor. The implantation quickly imposes itself into all the other doping stages, and even in processes of insulation or purification of the silicon: formation of drain and source areas and the deep n- or p-wells of CMOS circuits, the realization of burred collectors, emitter and base doping in bipolar technology, impurity gettering, realization of buried insulating silica layers by implantation of an important quantity of oxygen, etc. Nowadays, for example, the fabrication of the circuits of the most recent CMOS technology requires more than ten stages of ion implantation.

The main *advantage* of implantation is the precise control of the number and of the penetration depth of the ions. The dose indicates the number of ions implanted per unit of surface of the target (often given in atoms/cm^2) and the energy of ions (generally in keV) is the parameter controlling the spatial distribution of the atoms.

The major *disadvantag*e of the technique is the inevitable damaging of the target during the slowing down of the ions. The collisions of the projectiles with the atoms of the target move the latter from their initial equilibrium position. If the atom thus displaced acquires enough energy, it can cause other displacements, and be at the origin of a cascade of collisions. The damage created by the implantation must be repaired (we speak of healing), before being able to profit from the effects expected from implantation, i.e. full electrical activity of the dopants. This healing intervenes during a thermal treatment (annealing), following the implantation.

All these mentioned aspects – implanters, ion path, damaging, annealing, dopant activation, application to the development of devices – will be developed in this chapter.

9.2 Set up and work of ion implanters

The implanters are classified into two categories: the *medium current implanters* (a few mA at maximum) and *high current implanters* (from a few mA up to several tens of mA). The range of accessible energy is the same in the two categories, typically from 10 to 200 keV. The difference between the two groups lies in their process capacity and in their respective flexibility in use. A medium current implanter processes one wafer at a time. The ions are distributed there uniformly with the help of a system of electrostatic scanning. The sample holder can often be oriented, which is useful for oblique implantations. The limitation in current is related to the difficulty in evacuating, as calories, a sufficient quantity of the energy deposited by the beam. On a high current implanter, typically several tens of targets are installed on a carousel revolving permanently in front of the beam. This target displacement constitutes one of the scanning directions, the other being of an electrostatic or mechanical nature.

Figure 9.1 schematically illustrates the general disposition of a medium current implanter An ion implanter consists of a source of ions into which the elements to be ionized are introduced in the form of gas or vapor. With a polarized electrode, we can extract ions from the source. A magnetic analysis system separates the ions according to their mass and an acceleration column communicates to them the desired speed. Lastly, a set of polarized plates uniformly distributes the ions on the target. The loading of the wafers and their positioning in front of the beam and then their unloading are executed by a robotized system. On a medium current implanter, with the master controller of the target we can often tilt the wafers up to 60 degrees in relation to the direction of the beam.

An ion source essentially consists of a closed vessel, called an arcing chamber, in which an electrical discharge is maintained in a gas or in a vapor containing the desired element. In most high current sources, the discharge is initiated and maintained by electrons emitted by a filament heated by the transit of an intense current and by a magnetic field, perpendicular to the direction of electron emission, which considerably increases their course between the cathode and anode. Thus, the probability of ionizing the vapor atoms increases by a significant factor. The source with a hot cathode is very frequently used in industrial implanters . In this type of source, a voltage of about 50-100 V is applied between the cathode and anode and the intensity of the magnetic field is about 100 G. In the sources with cold cathode, the discharge is maintained with the help of higher voltages and magnetic fields, respectively 1 to 2 kV and a few

kG. Most elements are introduced into the arcing chamber in the form of a pure gas or as components of gaseous molecular compounds. Other elements, liquid or solid, with a vapor pressure of about 10-3 Torr in a controllable temperature range (300-700°C), are set in a furnace close to the arcing chamber. Sometimes, solids, which cannot be vaporized in the elementary form, exist in the form of compounds, which can often be more easily sublimed. Lastly, low vapor pressure materials can be vaporized using a chemical reaction.

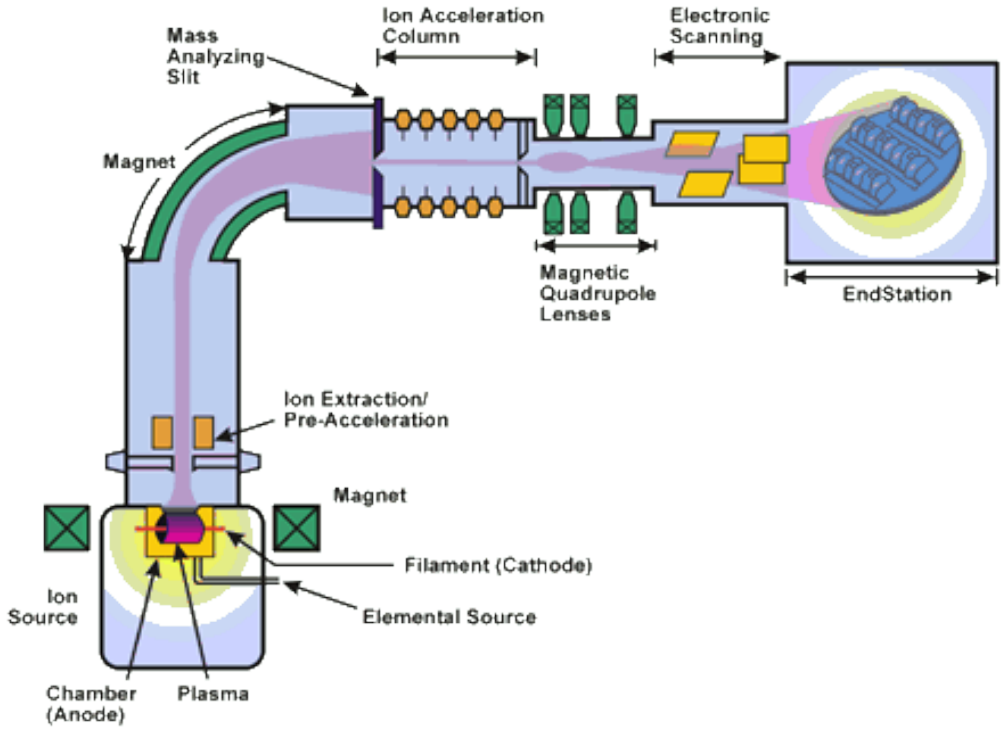


Fig.9.1 Variant of schematic representation of a ion implanter

A large amount of ions, monoatomic or molecular, with different masses or charge states, is extracted from the source. For example, when boron trifluoride is injected into the arcing chamber, the ions extracted from the source are: $10B^{++}$, $11B^{++}$, $10B^{+}$, $11B^{+}$, F^{+}, $10BF^{+}$, $11BF^{+}$, $10BF_2^{+}$, $11BF_2^{+}$, $10BF_3^{+}$, $11BF_3^{+}$. For each ion group, their intensities are in the isotopic abundance, namely 1 to 5, in the case of boron 10 and 11. Only one type of ion generally interests the user, for example $11B^{+}$. It is thus necessary to proceed to a selection according to the mass of these ions. An analysis magnet separates the various beams. Indeed, the path of the ions, accelerated by a potential V (in volts), follows, in a magnetic field B (in Gauss), a circle trajectory of radius R (in centimeters) so that $RB \approx 144\ (M_1 V / n)^{1/2}$, where M_1 is the mass of the ion (in atomic units) and n its charge state. The common magnets installed in the implanters have average resolutions of about $M_1/\Delta M_1 \approx 100$. This means that they are able to differentiate the ion of mass 100 from the ion of mass 101.

Most installations have a magnet deflecting the ion trajectory at 90 degrees. Under these conditions, the magnets also have a significant focusing role: the image of the extraction slot is located at the same distance from the output pole of the magnet, as the source from the input pole. It is there that a shutter is often set (a Faraday cage or a definition slot) allowing us to control the beam or to stop the undesirable ions. Beyond this point, the beam is again divergent. It is taken back a little further in quadripolar electromagnetic lenses, which refocus it before its entrance into the scanning plates.

An important aspect of the design of these chambers relies in measuring the current, and thus the dose. Generally, on medium current implanters, the target is directly connected to an ammeter and the current is integrated there, in order to determine the dose ϕ, according to the relation:

$$\varphi(ions/cm^2) = \frac{1}{A}\int_0^{t_{impl}} \frac{I}{ne}dt \qquad (9.1)$$

where A (in cm^2) is the implanted surface, usually defined by a beam limiter at the entrance of the implantation chamber. *I* (in ampere) is the beam current, and e is the elementary charge (in coulomb), t_{impl} corresponds to the entire duration of the implantation. A certain number of phenomena possibly come to distort this measurement: (1) the ions, on their course to the target, can capture an electron and thus be found in a neutral state; (2) the impact of the ions tears off electrons from the target which, if they do not go back there, give place to a parasitic surplus current; (3) the back diffusion and the surface sputtering, by ejecting atoms, decrease the number of ions actually introduced into the target.

For the strongest doses, the implantation duration starts to play a role and the output will then strongly depend on the current. However, as mentioned previously, an ill-considered increase in the current density leads to significant heating of the target. Figure 9.2 reports, for an implanter with electrostatic scanning, the theoretical value of the heating of a silicon wafer, as a function of the time, for various power densities (or irradiances, in W/m^2).Such temperatures are prohibitive: photoresists, for example, can tolerate only moderated temperatures, of about 100°C. Beyond this level, they inflate, crackle or decompose. There is thus a compromise to be found between high output and reasonable heating. With a purely electrostatic scanning, the beam remains almost permanently on the same wafer that quickly heats up.

Ion implantation in the semiconductor industry often implies insulating structures. A significant charge effect can be induced there by ions, causing a progressive deterioration of the dielectric qualities of the oxides, going, possibly, up to a disastrous breakdown. In addition, this charge effect can also disturb the behavior of the ions up to the target, producing a divergence of

the beam. These problems are all the more severe, as the currents become more important. It is then necessary to maintain the target neutrality. This is generally carried out with the help of a “shower” of secondary electrons (therefore, of low energy) placed near the target, like, for example, the one in Figure 9.3.

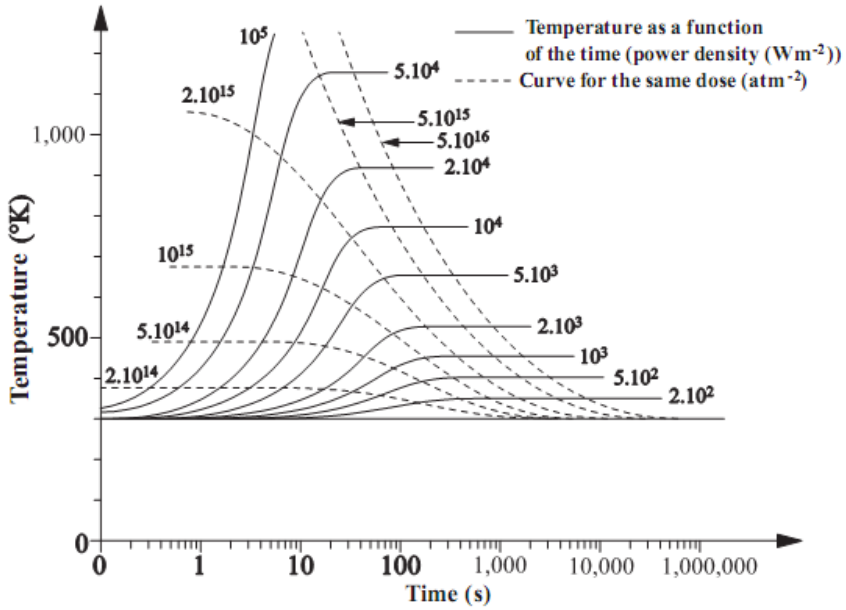


Figure 9.2. Calculated values (full line) of the silicon temperature as a function of time and of the power density (irradiance). The curves in dotted lines correspond to the iso-dose contours reached under these conditions.

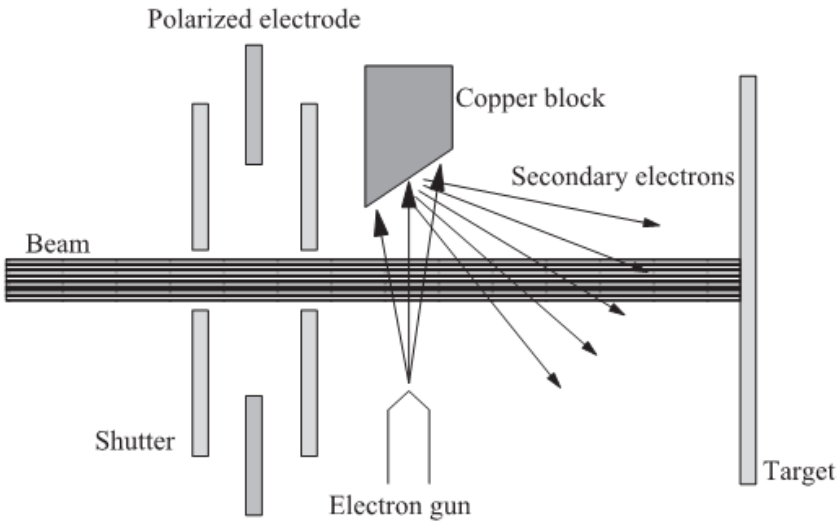


Figure 9.3. Example of an “electronic shower” device intended to neutralize the charge induced in the target by positive ions

9.3 Ion range

9.3.1. Binary collision and stopping power

The slowing down of the ions in the material is characterized by the *stopping power*, noted (dE/dx), corresponding to the quantity of energy dE lost per elemental distance dx. It is a function of the energy E. The interaction between the projectile and a target atom is described by assuming two distinct processes: the *collision* between the two nucleui and the *interactions* with the electrons (Fig.9.4). The first corresponds to the Coulomb repulsion, causing an important *deviation of the trajectory*. The second is related to *excitations and ionizations*, leading to significant *energy loss* for the ion, but which do not alter its direction.

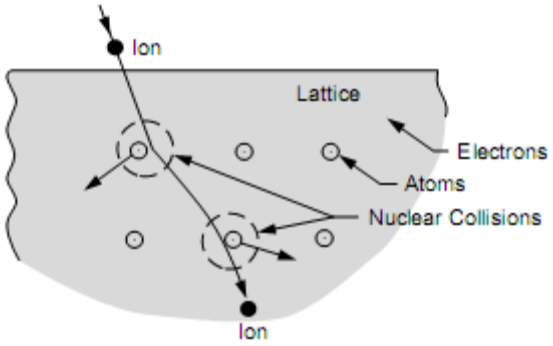


Fig.9.4. An ion incident on a crystal lattice is deflected in nuclear collisions with the lattice atoms and also loses energy in collisions with electrons .

The stopping power can thus be separated into two components: nuclear $(dE/dx)_n$ and lectronics $(dE/dx)_e$:

$$\frac{dE}{dx} = (\frac{dE}{dx})_n + (\frac{dE}{dx})_e = N(S_n + S_e) \qquad \textbf{(9.2)}$$

The stopping cross-sections for each type of interaction, S_n and S_e, are physical quantities independent of the density of the slowing environment, N is the atomic density. In the context of classical mechanics, the elastic binary collision with an atom of the target initially immobile, following the laws of conservation of energy and of momentum, we can write the energy transfer *T* as:

$$T = \frac{4M_1M_2E}{(M_1+M_2)^2}\sin^2\frac{\theta_c}{2} \quad , \qquad (9.3)$$

where M_1 and M_2 are respectively masses of the ion and of the target atom. E is the energy and θ_c the scattering angle in the system of the center of mass (see Figure 9.5). The scattering angle is related to the interaction potential V(r) by the relation:

$$\theta_c = \pi - 2p \int_{r_{min}}^{\infty} \frac{dr}{r^2 \sqrt{1 - \frac{V(r)}{E_c} - \frac{p^2}{r^2}}} \tag{9.4}$$

where p is the impact parameter (see Figure 9.5) and E_C is the energy in the system of the center of mass ($Ec=EM_2/(M_1+M_2)$), r_{min} is the collision radius between the two partners.

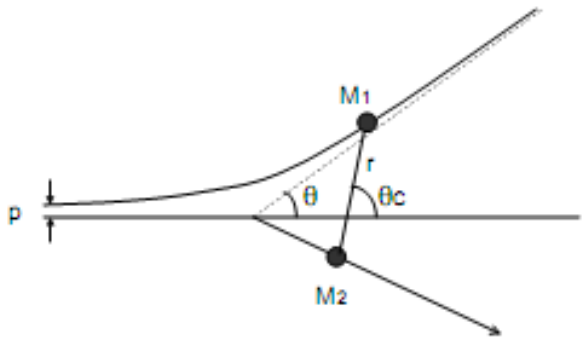


Figure 9.5. Diagram of the collision between the projectile (M_1) and the target atom (M_2)

The potential used at implantation energies is of the Coulomb type, moderated by the *shielding* effect of the electrons. It is written:

$$V(r) = \frac{Z_1 Z_2 e^2}{4\pi\varepsilon_0 r} \varphi\left(\frac{r}{a}\right) \tag{9.5}$$

where Z_1 and Z_2 are the atomic numbers of the projectile and of the target, and ϕ□ is the shielding function, tending to 1 when r decreases, *a* is the shield radius. The cross-section (i.e. the probability that a collision, leading to the deflection θ_C, occurs) is:

$$d\sigma(\theta_c) = 2\pi p dp = -2\pi p\left(\frac{dp}{d\theta_c}\right) d\theta_c \tag{9.6}$$

And finally, the stopping cross-section is:

$$S_n = \int_{T_{min}}^{T_{max}} T(\theta_c) d\sigma(\theta_c) \tag{9.7}$$

Lindhard, Scharf and Schiott (LSS) introduce *reduced variables*, ε and ρ □respectively proportional to the energy and to the ion range *R*:

$$\varepsilon = \left[\frac{4\pi\varepsilon_0 a}{Z_1 Z_2 e^2}\frac{M_2}{M_1+M_2}\right]E \qquad \rho = N\pi a^2 \frac{4M_1M_2}{(M_1+M_2)^2}R \qquad (9.8)$$

The nuclear stopping power dominates at very low speed. At higher energies, the electronic stopping power takes increasing importance and, always in the domain of implantation, its value is proportional to the speed:

$$S_n = \frac{\ln(1+1.21\varepsilon)}{2(\varepsilon+0.0065\varepsilon^{0.154}+0.242\varepsilon^{1/2})} \qquad S_e = k\varepsilon^{1/2} \qquad (9.9)$$

k is a constant without dimension, which depends on M_1, M_2, Z_1, Z_2. Thus, in opposition to the case of the nuclear stopping power, the electronic deceleration cannot be represented in the form of a universal curve. Nevertheless, for common projectiles, k varies between 0.1 and 0.2 (from the heaviest to the lightest) and, in the domain of interest, the stopping powers depend on the energy, as indicated in Figure 9.6. In this figure, the electronic stopping power is represented for k = 0.15. The nuclear stopping power is at maximum at ε≈□0.35, which corresponds to 3 keV for boron, 15 keV for phosphorus or 70 keV for arsenic in silicon. The electronic stopping power becomes dominant at ε>□2 for the lightest ions, and at ε>□4, for the heaviest, which, always in silicon, corresponds to 20 keV for boron, 140 keV for phosphorus or 800 keV for arsenic.

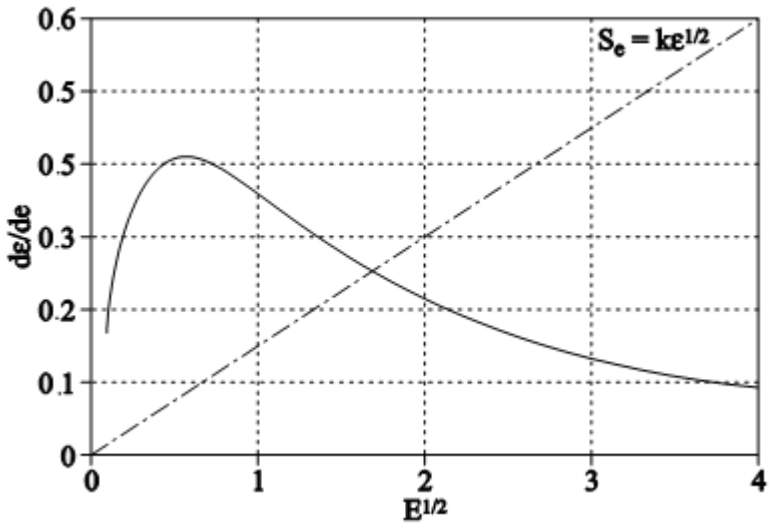


Figure 9.6. Stopping power cross-sections as a function of the square root of energy in LSS units.

9.3.2.Profile of the implanted ions

The total distance covered by an ion along its trajectory is called the *range* (noted R), but what interests the user is the distance covered in the normal direction to the target, which we call the *projected range*. The slowing down process being essentially statistical, the relevant parameters are the average projected range (noted R_P) and the scattering of the values around R_P, characterized by a longitudinal standard deviation ΔR_P and a transverse standard deviation $\Delta R\perp$ (see Figure 9.7). To calculate these magnitudes, the concepts related to the binary collision and to the stopping powers are applied to the statistical problem of a series of collisions. For that, there are two solutions: the resolution of a transport equation and the Monte Carlo approach.

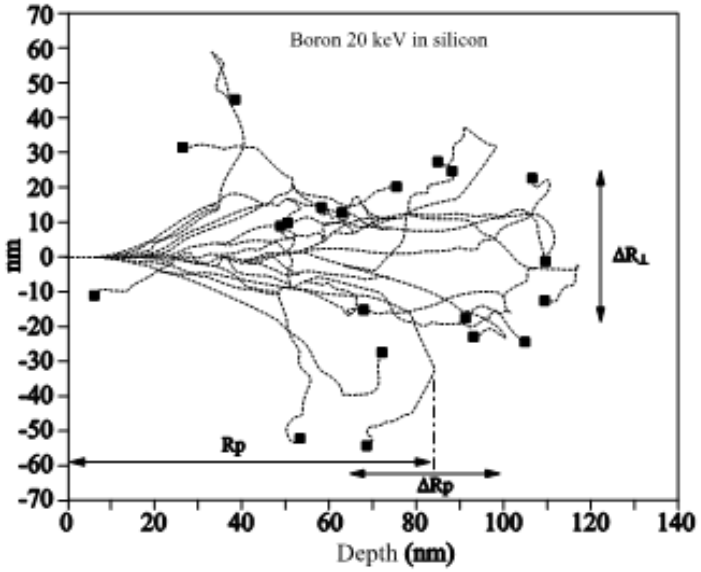


Figure 9. 7. Range of about 20 boron ions of 20 keV in silicon (the trajectories are represented by a dotted line and the stopped boron by a black square) and definitions of the average projected range and of the longitudinal and transverse scatterings.

The two methods give similar results, namely Rp and ΔRp. As a first approximation, we can then describe the profile of the implanted atoms as a Gaussian function:

$$C(x) = \frac{\varphi}{\sqrt{2\pi}\Delta R_p} \exp\left[-\frac{(x-R_p)^2}{2\Delta R_p^2}\right] \tag{9.10}$$

Figure 9. 8 shows the calculated from eq.10 Gauss profile of B in Si for diferent energy of *projectiles.*The depth increases with increasing of energy.

However, at comparing to the experimentally measured distributions, it is clear that, except when $M_1 = M_2$, the Gaussian approximation is not satisfactory. Two dimensionless parameters, γ and β, respectively describing the profile asymmetry (*skewness*) and the sharp or squashed character of the maximum concentration peak (*kurtosis*) can be defined as:

$$\gamma = \frac{\int_{-\infty}^{+\infty} (x - R_p)^3 f(x,E)dx}{(\Delta R_p)^3} \quad , \quad \beta = \frac{\int_{-\infty}^{+\infty} (x - R_p)^4 f(x,E)dx}{(\Delta R_p)^4} \qquad (9.11)$$

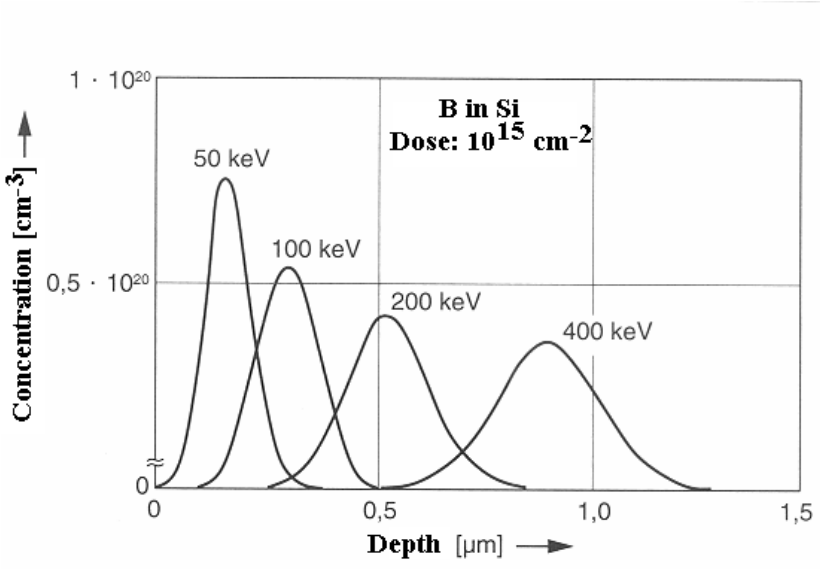


Fig. 9.8 Calculated Gauss profile for B implantation at different energy

If $\gamma < 0$, the peak is beyond R_P and the concentration is more important towards the surface than towards the depth. This is the case for light projectiles in a heavy target, for example boron in silicon (see Figure 9.9). The situation is reversed for $\gamma > 0$. We find this shape for a heavy projectile in a light target, such as arsenic in silicon (see Figure 9.10). $\gamma = 0$ and $\beta = 3$ correspond to a Gaussian distribution.

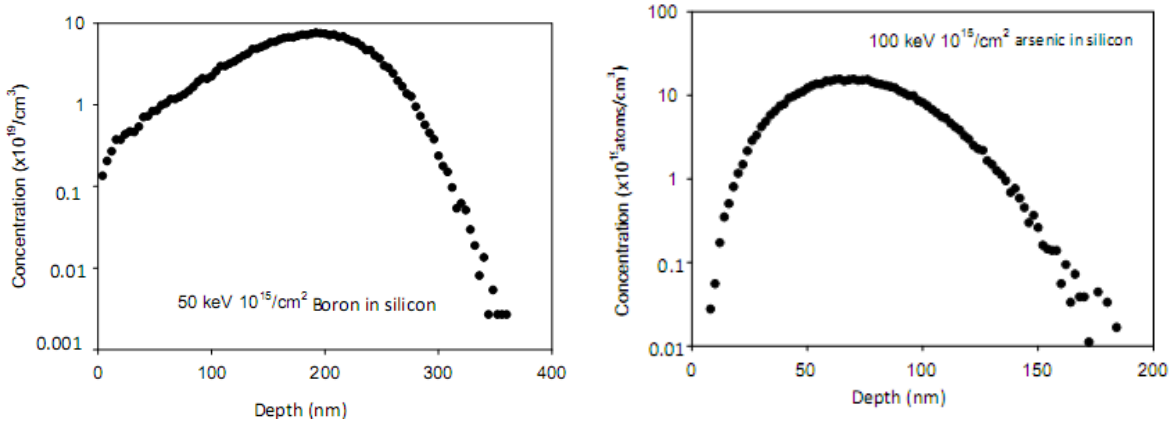


Figure 9.9. Simulation of the profile of boron atoms (50 keV 10^{15}/cm^2) implanted in silicon, **Figure 9.10** Simulation of the profile of arsenic atoms (100 keV 10^{15}/cm^2) implanted in silicon.

9.4. Backscattering, surface sputter and channeling

A certain number of phenomena can alter the shape of the concentration profile of implanted atoms (Fig.9.11). They appear under some specific conditions and are generally not taken into account in simulators. *Backscattering* corresponds to the events of almost head-on collisions during which an energy close to $T_{max}=\gamma E$ is transferred to the *recoil* atom, which can lead, if $M_1<M_2$, to a reflection of the projectile on the target surface. The backscattering coefficient, i.e. the number of reflected ions divided by the total number of projectiles can go up to 30% for low energy boron ions into silicon. This phenomenon must thus be taken into account to estimate the dose actually received by the target. *Sputtering* corresponds to the ejection of surface atoms of the target under the action of elastic energy transfers. It is characterized by a sputtering coefficient S, corresponding to the number of atoms ejected per incident ion. It is a function of M_1, M_2 and E. S reaches its strongest value at the energy corresponding to the maximum of nuclear stopping power. In silicon, it is then of the order of 5 for an antimony beam, of 3 for arsenic, of 1.5 for phosphorus and lower than 1 for boron.

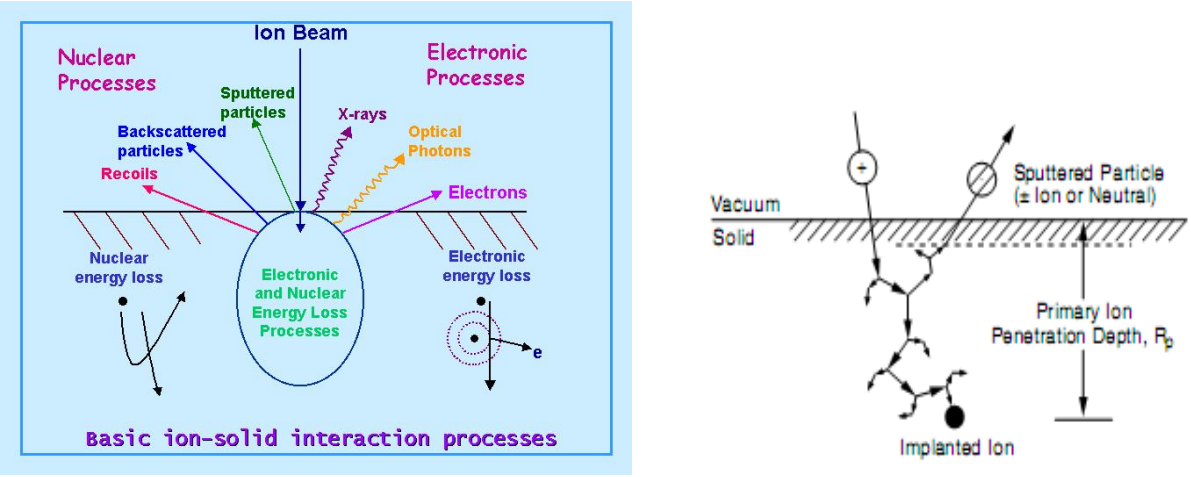


Fig.9.11 Basics of ion-solid interaction process and sputtering process.

By assuming that the sputtering rate remains constant during implantation, and that the profile is Gaussian, for high-dose implantation, the distribution will obey:

$$C(x)=\frac{N}{2S}\,erfc\,\frac{x-R_p}{\sqrt{2}\Delta R_p} \qquad (9.12)$$

For high implantation doses, the profiles will evolve as shown in Figure 9.12. The surface sputtering is thus the only phenomenon actually limiting the quantity of atoms that we can introduce into a material by implantation

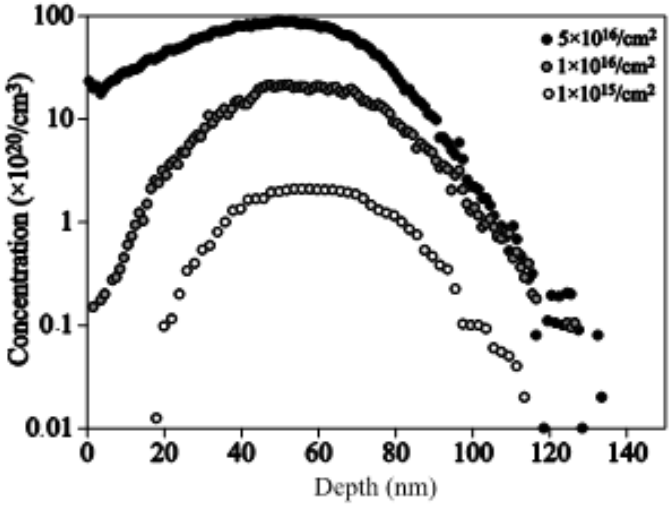


Figure 9.12. S*imulation of the profile of antimony ions of 100 keV implanted in silicon at various doses*

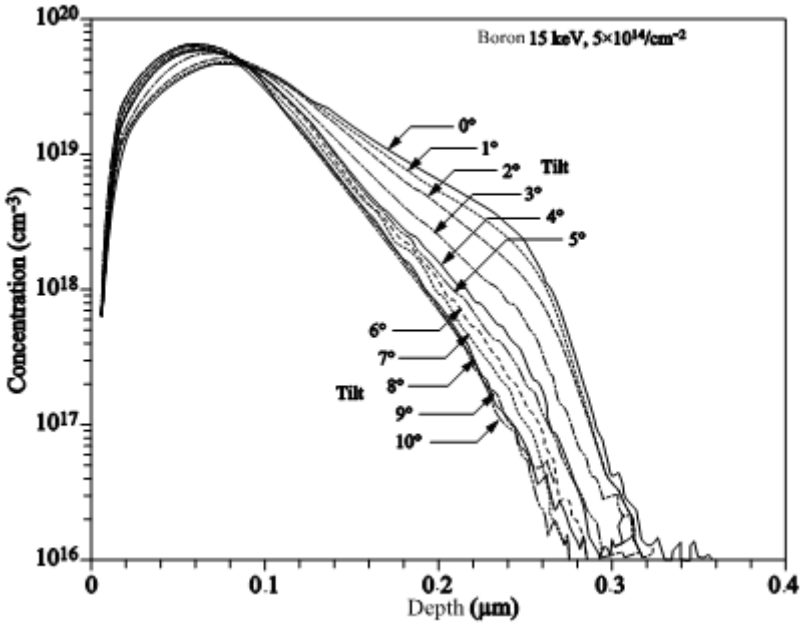


Figure 13. *SIMS profiles of the boron implanted at 15 keV in silicon according to the incidence angle compared to the direction <100> of the crystal*

The order of magnitude is of a few degrees for usual implantation energies in silicon. It is higher in axial channeling than in planar channeling. The latter constitutes a major limitation to the realization of surface junctions by boron implantation. The tails in distributions can obviously be reduced, thanks to an accurate control of the orientaion of the wafers as a function

of the beam direction. But, even with these precautions, the angular divergence of the ions compared to their initial direction, after a certain number of collisions, quickly becomes very important. This means that a part of them will always end up adopting a trajectory focused along the axis or secondary planes of the crystal (see Figure 9.14). This phenomenon explains the persistence of a tail in the distributions of Figure 9.13, even for incidence directions far from any channeling axis. The only way to completely remove this effect consists of destroying the crystalline order by preamorphizing the target.

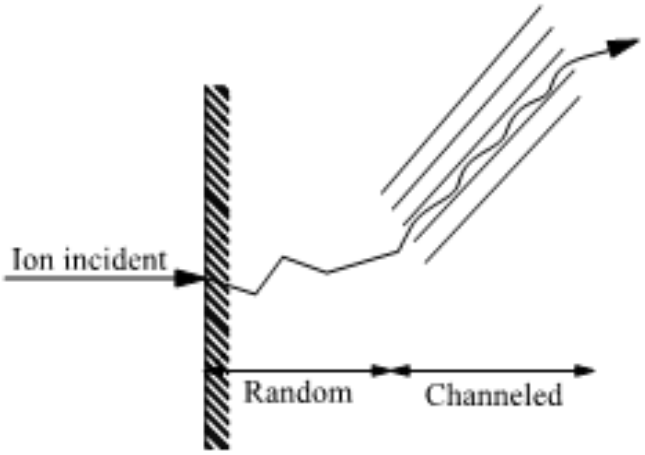


Figure 14. Schematic representation of the ion channeling at the end of the course

9.5.Implantation through a mask

In the fabrication technology of electronic circuits, the implantations are almost always carried out through a mask containing openings, corresponding to zones to dope, and protected zones, where ions must be completely stopped. Resists, oxide layers (SiO_2) or silicon nitride (Si_3N_4) are used to stop the projectiles. Because of their nature and their thickness d, the efficiency of these films is characterized by the transmission coefficient τ, i.e. the number of ions managing to cross, normalized to the total number of incident ions.

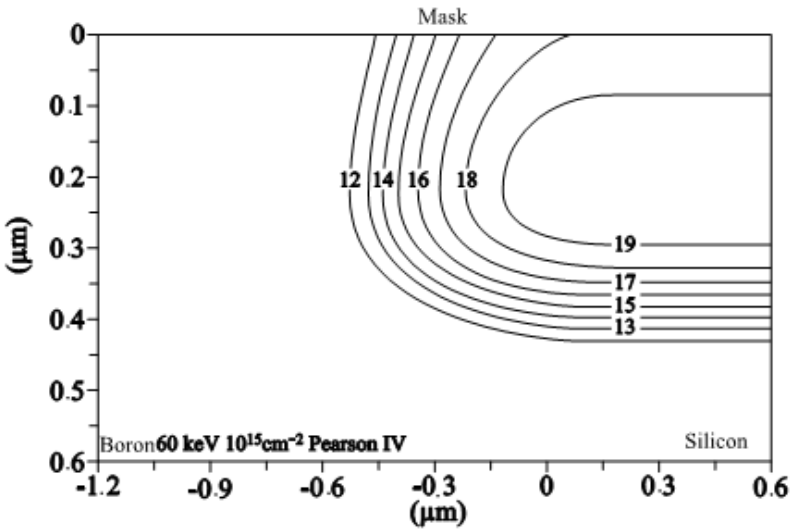


Figure 9.15. *2D* profile calculated in the case of a vertical mask edge

Figure 9.15 shows the contours of equal concentration (from 10^{19} to $10^{12}cm^{-3}$) for boron ions of 60 keV, implanted at the dose of 10^{15} cm^{-2} through a mask with 1µm width. We can note that the doping can extend a few hundred nanometers and risk to interfer with the neighboring device, if the latter is set too close. The horizontal spread is thus a serious limitation to the integration of a growing number of devices per unit of silicon surface.

9.6. Creation and healing of the defects

9.6.1.Primary collision and cascade

When an ion of energy E comes to strike an atom of the target, the transferred energy is T (see eq.9.3). If T is higher than a *threshold energy* E_d, called the *displacement* energy (of several eV), the recoil atom thus released will be able in its turn to cover a certain distance in the target, losing its energy through nuclear and electronic collisions. Repeating this, a series of atomic displacements is created, called a cascade (Fig.9.16).

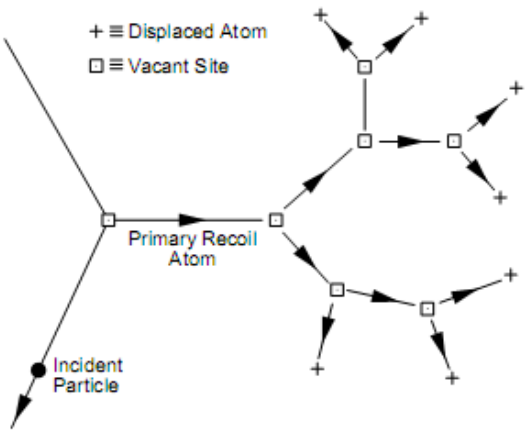


Fig. 9.16 Schematic of the formation of collision cascade by a primary knock-on atom

This lasts until the remaining energy in each individual cascade, and in particular the one of the primary ion, is lower than E_d. The entire duration of the cascade is very short: less than one picosecond. In addition, transfers under the threshold often represent a significant part of the energy transfered to the slowing-down environment. They sometimes constitute the majority of events, resulting in an important creation of phonons, i.e. thermal oscillations, and thus finally in a local temperature rise. This thermalization proceeds on a time scale definitely longer than the cascade duration, of about one nanosecond. In a very simple model, Kinchin and Pease estimate that at each collision, the average transferred energy is divided by two and thus, at the *n*th collision, the average transferred energy is $T/2^n$. Finally, this model leads to a number of displaced atoms (per unit of target surface), proportional to the dose.

However, in practice, we notice (see Figure 9.17) that the number of created defects (here in silicon at low-temperature) is often much higher than predicted by the Kinchin-Pease (KP) model. The order of magnitude is a few hundred atoms moved by incident ion for a light projectile, a few thousand for a medium mass and several tens of thousands for a heavy projectile. At higher energy, the curves again take the linear slope of the KP relation, indicating that the creation of damage at high-energy obeys this law. At low energy, the important creation of defects is due to an effect related to the high density of transferred energy, leading to the overlay of the cascades initiated by the recoil atoms. This is equivalent to a collapse of the displacement energy. This phenomenon is usually called a *displacement spike*

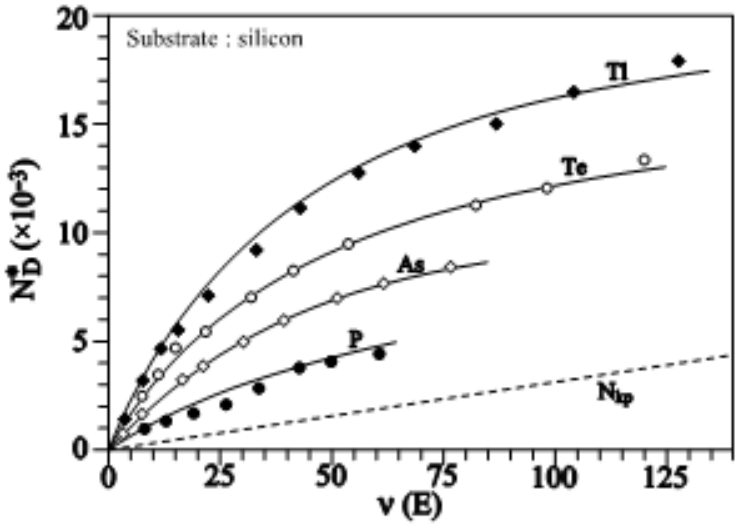


Figure 9.17. Total number of silicon atoms moved by incident ion, according to the total quantity of energy lost in elastic collisions. The dashed line corresponds to the standard Kinchin and Pease model

This phenomenon of cascade overlay is obviously even more important during the implantation of polyatomic ions, such as BF_2^+ or decaborane, commonly used in microelectronic technology. Figure 9.17 shows how the number of displaced silicon atoms gradually deviates from the linear behavior, depending on the number of atoms contained in the molecule. Figure 9.18 shows the profile of energy deposition in elastic collisions, for 100 keV arsenic in silicon, compared to the distribution of the implanted ions. These profiles do not coincide: a maximum of energy is deposited before the ions stopping. The defects are created according to the profile of energy deposition.

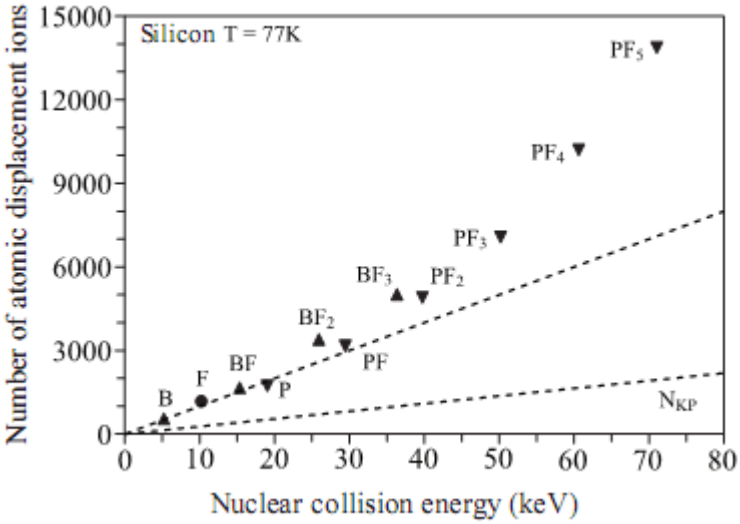


Figure 9.17. Number of silicon atoms moved at 77 K (measured by RBS-C in situ) by the implantation of various polyatomic projectiles, according to the energy lost in elastic collisions. This number grows quicker than linearly (dashed line) and it is much more important than the Kinchin and Pease standard modelpredicting it.

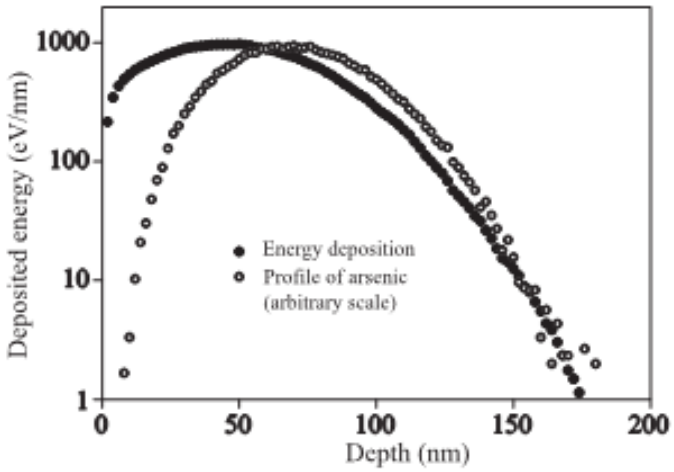


Figure 9.18. Comparison between the profile of the deposited energy in elastic collisions (black dots and left scale) and distribution of the arsenic implanted (white dots and arbitrary scale) at 100 keV in silicon.

9.6.2 Point defects

The concept of point defect is thus relatively easy to define, especially in the microelectronics materials, which are often perfect single crystals. Elementary point defects are thus the vacancies, i.e. *vacant sites* in the network, and the *interstitials*, which are supernumerary atoms located between the atoms of the network. After implantation, the interstitial species are the implanted atoms, but also the atoms moved from the target, called self-interstitials. Several configurations are possible for these interstitials, all of them allowing the atom to be inserted in the network. Moreover, around the vacancy, as well as around the interstitial, a local rearrangement is necessary. In silicon the defects' formation energy, is about 1 eV for a vacancy and 3 to 5 eV for an interstitial. An energy transfer higher than the displacement threshold produces a *Frenkel pair* (vacancy + interstitial). The energy formation of Frenkel pair in silicon at approximately 15 eV since it is necessary take ino account the electrostatic attraction which cause their recombination.

It is quite obvious that, beyond this little simplistic classification in vacancies and interstitials, the defects also differ by their configuration and by the connections made in their environment, including possible hybridizations of the target atoms or of impurities. For example, an interstitial often takes a configuration of shared site with an atom of the material, in energetically favorable crystallographic directions such as <100> or <110>. This is particularly true in ionic-covalent crystals, where there is an important lattice–electron interaction, which often produces important relaxation effects around a defect.

We said that the first stage, the cascade, occurs in approximately 10^{-13} seconds, and the second one, thermalization, in 10^{-9} s. The rearrangement described above is the third stage, bringing the system towards a new equilibrium. It occurs on a much longer period of time. In the second stage and especially in the third stage, events such as diffusion, precipitation or chemical reactions can occur. We then understand that other defects can be formed. They are more complex and considered as point defects, but are already defect combinations: for example the double or triple vacancy, etc., the vacancy-impurity pairs and, in covalent compounds, the antisite defects.

9.6.3. Accumulation of damages, amorphization

During implantation, elementary defects (interstitial and vacancies) are generated proportionally to the projectiles' flux. They can recombine and/or destroy themselves on "sinks" (grain boundary, surfaces, interfaces, dislocations, etc.) or diffuse, proportionally to their local gradient. Without writing too complex equations, two extreme situations can arise:

– the generation rate (i.e. the projectile flux or mass) is low and/or the diffusion is important. The creation of Frenkel pairs is permanently compensated by recombinations and annihilations. We thus reach a stationary state, with a low concentration of vacancies and isolated interstitials. This is the case during a high temperature implantation or with light ions;

– the ion flux is important and/or the diffusion is low, so that the generation rate of the Frenkel pairs is more important than the recombinations or the other forms of elimination of point defects. These pairs will thus accumulate proportionally with the in-depth profile of the energy deposition, leading first to the formation of small defects aggregates, and then finally of real amorphous zones.

All these behaviors are evidenced on the experimental curve in Figure 19, describing silicon damaging by high-energy silicon ions. Using an amorphization model based on similar mechanisms, it can determin, for various projectiles, the critical dose for silicon amorphization as a function of the implantation temperature (Figure 9.19,b).

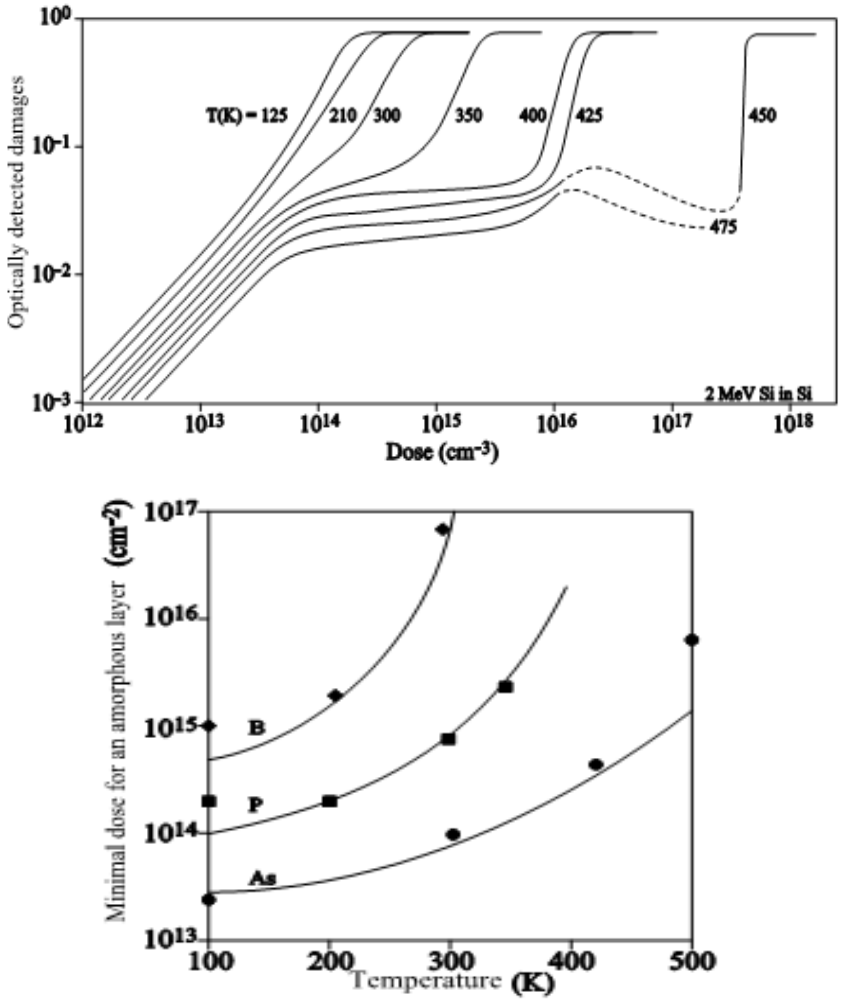


Figure 9.19. a) Silicon fraction damaged by Si ions of 2 MeV as a function of the implantation dose and temperature, b) Necessary critical dose to form an amorphous layer in silicon at various temperatures and using various projectiles

At ordinary temperature, in silicon, amorphization occurs, when critical deposition is about 10 to 12 eV/atom. For example, boron, at the doses usually used in microelectronics, does not reach this threshold and will thus not amorphize silicon at ambient temperature. However, even at ordinary temperature, point defects (vacancies and interstitials) are very mobile. Behaviors other than the simple accumulation in the form of amorphous agglomerates will thus appear: first of all the already quoted recombination, equivalent to a self-healing of the damages; the combination, also already mentioned, of interstitials and vacancies, between them or with impurities; the clustering of the vacancies and interstitials in a more ordered form, which leads, especially when the temperature is raised, to the formation of extended defects (for example dislocation loops). A simple calculation shows, for example, that above a certain number of atoms, the insertion of a loop in the silicon lattice is less expensive in energy than the insertion of an amorphous sphere. Obviously, the transition from one to the other also requires a certain energy, which can be provided, for example, by a rise in temperature during or after the implantation. Table gives the value of this temperature in some common semiconductors.

Material	Critical temperature	Material	Critical temperature
Si	120	SiC	350
GaAs	140	InSb	-40
GaP	220		

Table Critical temperature (in °C) beyond which it becomes impossible to amorphize some common semiconductors

9.6.4 Damage healing and dopant activation

The aim of the implantation method in semiconductors is to set the dopants in substitutional position, where they are electrically active. The first condition is thus to restore the initial crystalline order. This recrystallization occurs during a thermal process (annealing). The conditions of this annealing will ideally be selected, so that the dopants can be positioned at the same time in substitutional sites. The first condition is related to the morphology of the damaged area, according to whether it is amorphized or not. In the first case, it is the recrystallization front (the amorphous-crystal interface), which, by passing through the implanted area, will cause dopant activation. In the second case, there is no epitaxy front. The dopan itself will have to diffuse to the vacant sites. Figure 9.20 illustrates these two situations with the active phosphorus fraction in silicon, depending on the implantation temperature.

At ambient temperature, i.e. below the critical temperature reported in Table, an almost total phosphorus activation is reached with the annealing at 600°C. Above this value (implantations at 200 or 600°C in this example), the full dopant activation requires a much higher annealing temperature. Moreover, for annealing between 300 and 600°C, the activity increases and then decreases again. This effect of reverse annealing is well-known in the case of boron implantations which, at traditional microelectronics doses, do not amorphize silicon at room temperature.

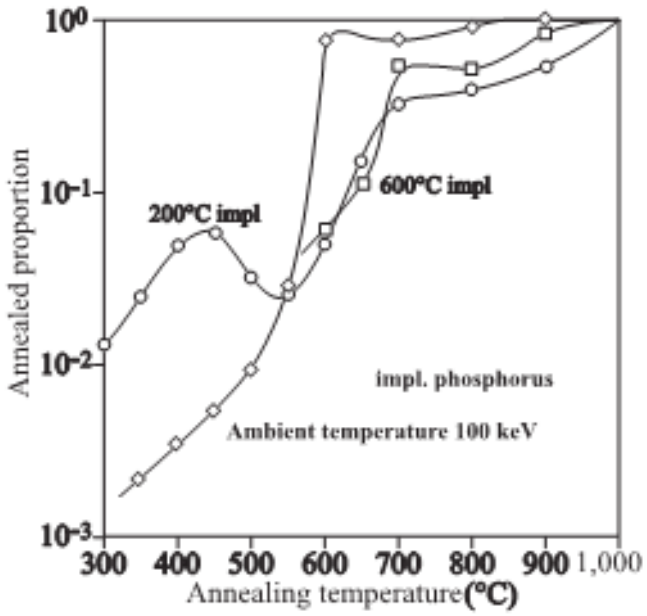


Figure 9.20. Variations of the carriers concentration with the annealing temperature in silicon implanted with phosphorus at various temperatures

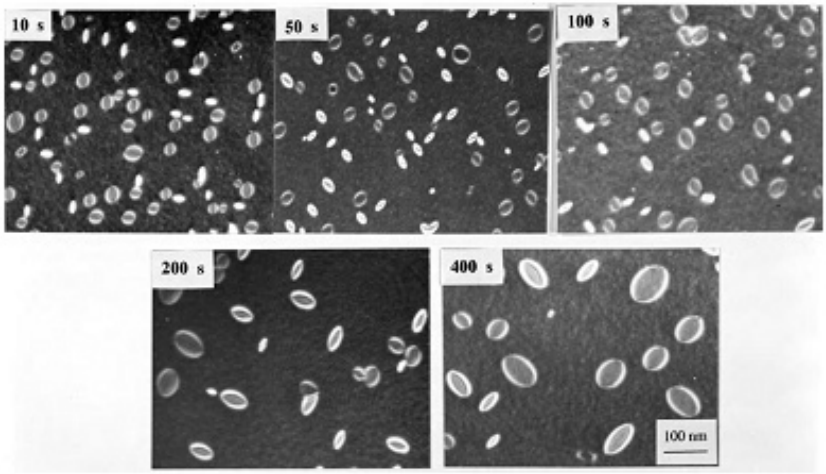


Figure 9.21. Images of transmission electron microscopy (plane view) of the "end of range" defects formed in amorphized silicon (Ge 150 keV, $2\times10^{15}/cm^2$) and annealed at 1,000°C for increasing times (10, 50,100,200 and 400 s)

Some residual defects remain after the solid phase epitaxy of an amorphous layer. They result from the presence, beyond the original amorphous-crystal interface, of an excess of recoil self-interstitials, compared to the local concentration of vacancies. In silicon, during annealing, a significant part of them will condense in {311} defects in the temperature range 600-750°C or in dislocation loops in the range 800-1 100°C. These defects are called EOR (*end of range defects)*. They are set in the (111) planes. The series of photomicrographs in Figure 9.21 and 9.22 show the morphology and development in size and population of these defects during an annealing at 1,000°C.

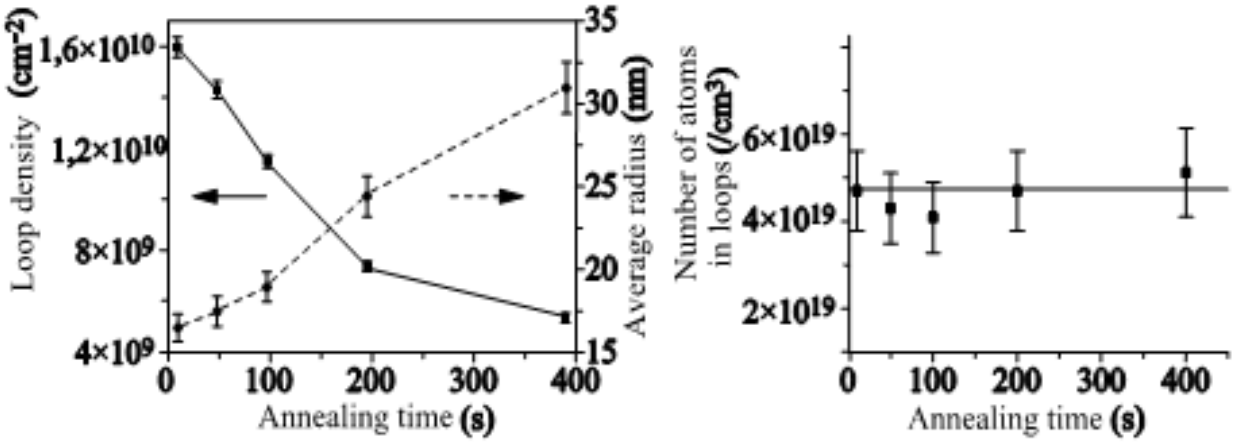


Figure 9.22. Left: development during the annealing time (at 1,000°C) of the density and of the average radius of EOR defects observed in Figure 2.21. Right: development of the total number of atoms (per cm^2) taken in the loops functions of the annealing time at 1,000°C

These defects can be traps for impurities or for charge carriers, altering the electric behavior of the devices. Moreover, excess interstitials can also, transitorily, take part in the diffusion of the dopants located in this zone. Generally, since the dopants diffuse in semiconductors via mechanisms assisted by vacancies and/or selfinterstitials, any local and/or temporal supersaturation (or undersaturation) of one or the others, created by implantation, will cause an “abnormal” diffusion. This aspect of implantation is sometimes used to adjust a doping profile, in particular to slow down diffusion.

Regarding dopant activation, we need to recall that the quantity of impurities that can be set in substitutional sites is limited. Beyond this limit, they precipitate in the form of immobile (in terms of diffusion) and inactive (from an electric point of view) clusters. However, for some dopants, such as arsenic or phosphorus in silicon, there is an “absolute” solubility limit and another one concerning their electrical activity. Table gives the maximum solubility value (at high temperature) of the main silicon dopants. Except at high temperature, where it saturates and sometimes decreases, solubility is the result of a thermal activated process. It is thus much

lower at low-temperatures and the dopants can thus precipitate during annealing at these temperatures. The forced recrystallization around these precipitates also leads to the formation of dislocations. However, the fast motion of an solid phase epitaxy front through the dopants profile can result in largely exceeding the solubility values and thus in incorporating, in a metastable state, a large amount of impurities

Dopant	Solubility (cm^{-3})	Dopant	Solubility (cm^{-3})
B	5×10^{20}	As	1.5×10^{21} (5×10^{20})
P	10^{21} (4×10^{20})	Sb	7×10^{19}
Ga	4.5×10^{19}	In	8×10^{17}

Table Maximum value of the solubility of various dopants in silicon. The value between brackets is the solubility of the electrically active atoms

9.7. Applications of ion implantation in traditional technologies CMOS

The implantation is used in most doping operations of MOS and bipolar technologies. The common operations are schematized in Figure 9.23, in the case of CMOS technology:

– doping of the p- and n-wells;

– adjustment of the threshold voltage by a light over doping of the channel under the gate oxide;

– doping of the source and drain extensions (LDD);

– pockets or “anti-punch through” halos;

– source and drain doping;

– gate polycrystalline silicon doping.

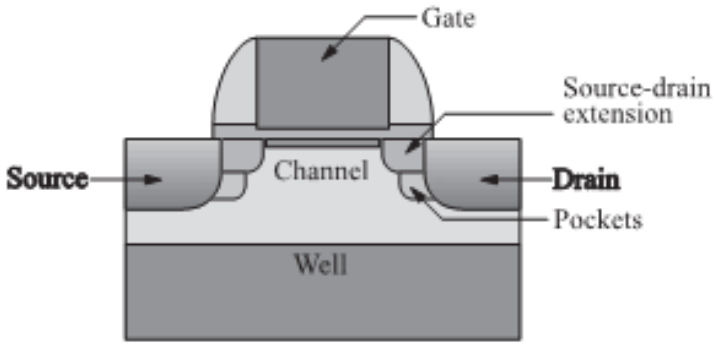


Figure 9.23. Areas implanted in traditional CMOS technology

The performances of the submicron transistors depend on the impurity profile close to the drains. A gradual profile, called LDD (lightly doped drain) can reduce the electric field in this zone and thus avoid the injection of “hot” electrons, which could deteriorate the N-transistor. The effect also exists, but in lower proportions, for the holes in the P-transistor. This doping of the channel-drain transition zone is carried out by a self-aligned implantation, possibly at an oblique incidence (up to 60 degrees) of phosphorus or arsenic (for n-type dopant) and boron (for p-doping). Doses are about several 10^{13} ions/cm^2 (see Figure 9.24). In this same zone, when the channel is very short, the transistor behavior can be alter by the “punch through” effect, if, under functioning (under polarization), the depletion zone of the drain-channel junction reaches the source-channel junction. This punch through effect is avoided by a phosphorus or boron implantation, possibly again in oblique incidence, with doses of about several 1012 ions/cm2, right under the active channel, in the area adjacent to the drain and source. Accurate control of the zone to be doped requires slightly higher energies: about 100 keV for boron and 150 keV for phosphorus (see Figure 9.24).

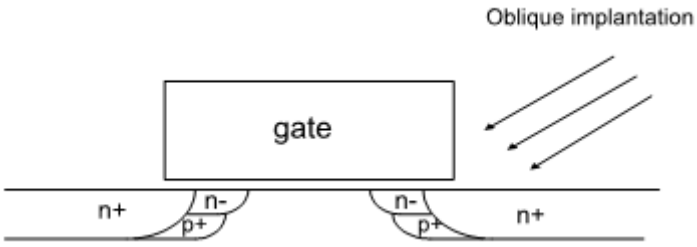


Figure 9.24. Schematic representation of the realization of LDD and anti-“punch through” doping

Figure 9.25 summarizes all these operations in the form of a dose-energy graph.

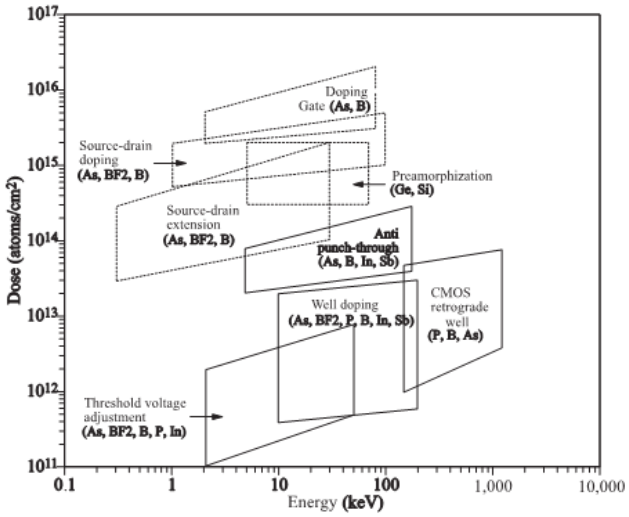


Figure 9.25. Domains of energies and doses corresponding to the various operations of doping or amorphization of silicon in CMOS technology

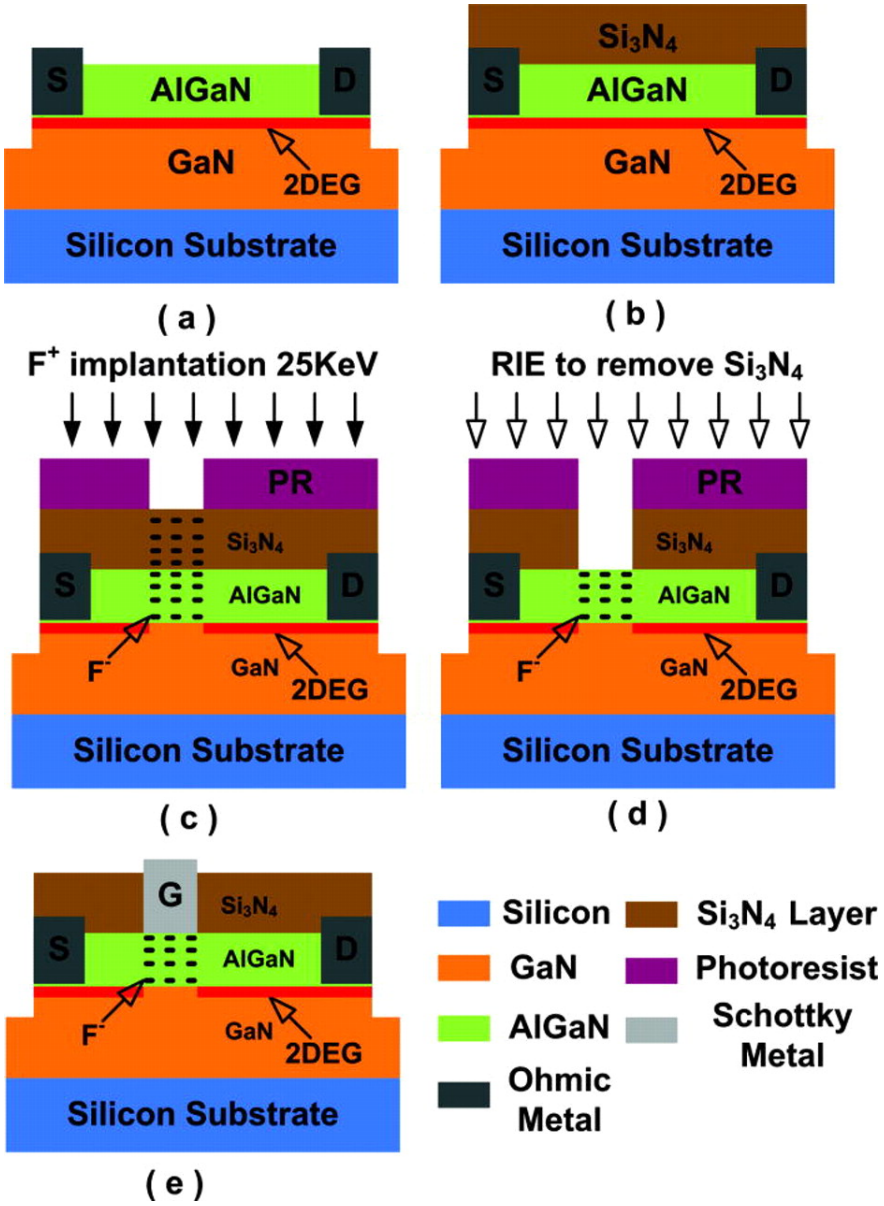


Fig.9.26 Application of ion implantation in heterostuctures.

New methods are proposed with using of ion impalantation, which are less conventional than simple doping, such as proximity gettering or the realization of semiconductor structures on an insulator (SOI). In this field, a new branch of the silicon industry has been recently developed, thanks to hydrogen implantation and to the Smart-Cut method. Finally, the arrival in force of nanotechnologies seems to open up new research tracks for ion implantation The example of quatum well forming with ion implantation is shown in Fig.9.26.

Chapter 10 Method of thin film deposition. Metallization

V.Skryshevsky, V.Verbitsky

10.1. Introduction

Thin films are needed to make metal wires and to insulate those wires, to make capacitors, resistors, inductors, membranes, channels, nozzles, mirrors, beams and plates, and to protect those structures against mechanical and chemical damage. Thin films have roles as permanent parts of finished devices, but they are also used intermittently during wafer processing, as protective films, as sacrificial layers, and as etch and diffusion masks. A great many solid materials are available as thin films: aluminum, gold, copper, tungsten and nickel are routinely used in microfabrication. Oxides of silicon, aluminum, hafnium and tantalum are used, as are nitrides of silicon and titanium. Diamond-like and Teflon-like films offer special properties, as do various alloys like PtMn, TiW, SiGe and CoFe. This chapter deals with the most common deposition processes for thin films, with the basic characteristics which make thin films different from the bulk, as well as some important applications.

10.2 Physical Vapor Deposition

The general idea of physical vapor deposition (PVD) is material ejection from a solid target material, transported in a vacuum to the substrate surface where film deposition takes place. Atoms can be ejected from the target by various means: resistive heating, electron beam heating, ion bombardment or laser beam bombardment (known as laser ablation). All aluminum films in microfabrication are deposited by PVD, and PVD is used for copper, refractory metals and for metal alloys and compounds like TiW, WN, TiN, $MoSi_2$, ZnO and AlN.

10.2.1 Evaporation

Evaporation of elemental metals is fairly straightforward: hot metals have high vapor pressures and in a high vacuum the evaporated atoms will be transported to the substrate. Typical deposition rates in evaporation are 0.1–1 nm/s, which is very slow. Evaporation systems are either high vacuum (HV) or ultrahigh vacuum (UHV) systems, with the best UHV deposition systems with 10^{-11} torr base pressures, and 10^{-12} torr oxygen partial pressures. In (ultra)high vacuum the atoms do not experience collisions, and therefore they take a line-of-sight route from source to substrate. The mean free path (MFP) is the measure of collisionless transport, and below about 10^{-4} torr the MFP is larger than the size of a typical deposition chamber. Low deposition temperature combined with line-of-sight transport means that evaporated films will

not coat sidewalls of holes and ridges well, even though film quality on planar surfaces is good. There are very few parameters in evaporation that can be used to tailor film properties. Atoms arrive at thermal speeds, which results in basically room temperature deposition. There is no bombardment in addition to the thermalized atoms themselves, which bring very little energy to the surface. Substrate heating can be done to improve film quality. This works because impurities are desorbed and adsorbed atoms can diffuse and find energetically favorable lattice sites. Low-melting-point metals, such as gold and aluminum, can easily be evaporated, but refractory metals require more sophisticated heating methods. Localized heating by an electron beam (Figure 10.1) can vaporize even tungsten (m.p. 3650 K), but deposition rates are, however, very low. Additionally, X-rays will be generated, which can damage sensitive devices.

Evaporation of alloys and compounds is tricky: the component with higher vapor pressure will evaporate more readily, and it can happen that the minority atoms in the starting material end up as the majority atoms in the thin film. Most compounds decompose when heated, therefore evaporation of compounds is limited to a few special cases, like silicon dioxide. It is possible that the molten metal reacts with the crucible because temperatures are very high, even though this is being minimized by the use of refractory materials for crucibles: namely, Mo, Ta, W, graphite, BN, SiO_2 and ZrO_2. Some crucible material can be incorporated into film also in the case of electron beam misalignment: if a misaligned e-beam hits the crucible, crucible material will be evaporated and incorporated in the deposited film.

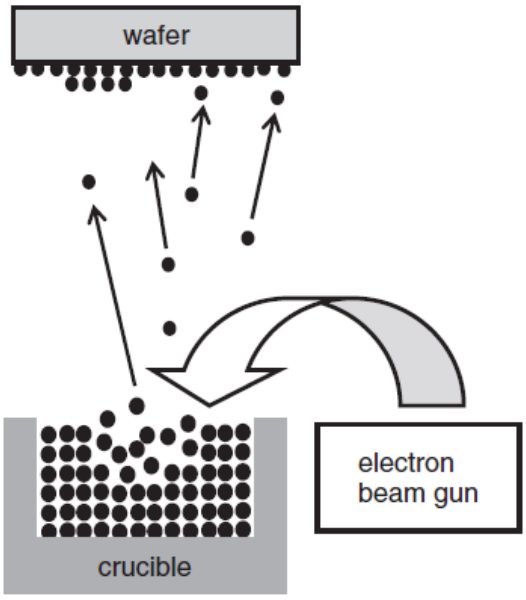


Figure 10.1 Electron beam evaporation: heated metal vaporizes and the evaporated atoms are transported in high vacuum to the substrate wafer

10.2.2 Sputtering

Sputtering (Figure 10.2) is the most important PVD method. Argon ions (Ar^+) from a glow discharge plasma hit the negatively biased target and eject typically one target atom. The ejected target atoms will be transported to the substrate wafers in a vacuum. These atoms are energetic and hit the substrate with considerable energy, which has both beneficial and detrimental effects on the growing film. Typical sputtering rates are 1–10 nm/s, significantly higher than in evaporation. Sputtering of nonconductive films necessitates use of RF fields to prevent charging of the target.

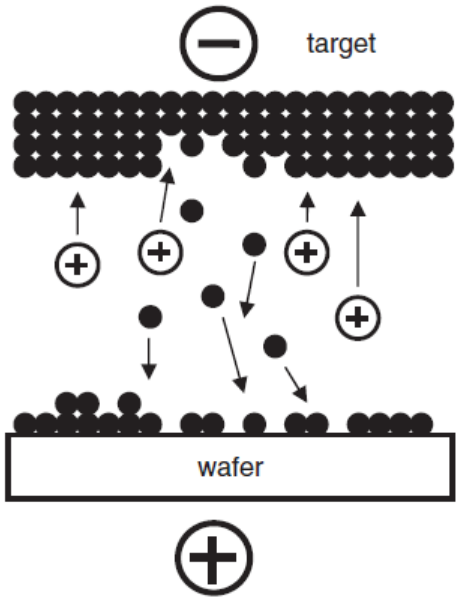


Figure 10.2 Sputtering: argon ions knock atoms out of a target, and the ejected atoms travel in a vacuum and deposit on the wafer

Because sputtering pressures are quite high, 1–10 mtorr (cf. evaporation 10^{-6} torr), sputtered atoms will experience many collisions before reaching the substrate. In a process called thermalization, the high-energy sputtered particles (5 eV corresponds to about 60 000 K!) collide with argon gas (T = 300 K) and cool down. Thermalization occurs also in other species present in the plasma, namely the reflected neutrals (some argon ions are neutralized upon target collision). These neutrals provide energy to the substrate. Thermalization reduces the energy of particles reaching the substrate and it reduces the flux of particles to the substrate. Lower flux means lower deposition rate.

In contrast to evaporation, the energy flux to the substrate wafer can be substantial. This has both beneficial and detrimental effects: loosely bound atoms (both film forming atoms as well as unwanted impurities) will be knocked out, improving adhesion and making the film denser. But energies that are too high can cause damage to the film, the substrate and underlying

structures (thinoxide breakdown because of high voltages). There will always be some argon trapped in the film, but its effect can usually be neglected because argon is a noble gas and therefore non-reactive. Incorporation of residual oxygen or nitrogen is much more pronounced because they are reactive and form oxides and nitrides. Sputtering yield is the number of target atoms ejected per incident ion. Sputtering yields of metals range from about 0.5 (for carbon, silicon and refractory metals Ti, Nb, Ta, W) to 1–2 for aluminum and copper, to 4 for silver at 1000 eV argon ion energy. Refractory metals have low sputtering yields, which is the fundamental reason for lower deposition rates. In practice, there is another reason which further lowers the deposition rate: refractory metals tend to have higher resistivity and thus lower thermal conductivity, which means that high sputtering powers cannot be applied to refractory sputtering targets. For heavy metals like tungsten and tantalum, sputtering yields are higher with xenon and krypton: these heavy gases transfer energy more efficiently to similar mass target atoms. However, argon is almost exclusively used. If oxygen is added to the sputtering atmosphere intentionally (usually together with argon), oxide films will result. The method is called reactive sputtering. In similar vein, nitrogen additions lead to nitrides. This is the way for example that Ta_2O_5 and TiN are made by sputtering.

10.3 Chemical Vapor Deposition

In chemical vapor deposition (CVD) the source materials are brought into the reactor in the gas phase, they are activated in the plasma, diffuse to the wafer surface, and react there to deposit film. Byproducts are desorbed and pumped away as shown in Figure 10.3. Deposition rates are temperature dependent according to the Arrhenius equation , but they are on the order of 0.1–10 nm/s. Common CVD processes include

$$SiH_4\,(g) \_\Rightarrow \; Si\,(s) + 2\,H_2\,(g) \sim 600\,^{\circ}C \qquad (10.1)$$

$$SiCl_4\,(g) + 2\,H_2\,(g) \Rightarrow \; Si\,(s) + 4\,HCl\,(g) \sim 1200\,^{\circ}C \qquad (10.2)$$

$$SiCl_4\,(g) + 2\,H_2\,(g) + O_2\,(g) \Rightarrow \; SiO_2\,(s) + 4\,HCl\,(g) \sim 900\,^{\circ}C \qquad (10.3)$$

$$3\,SiH_2Cl_2\,(g) + 4\,NH_3\,(g) \Rightarrow \; Si_3N_4\,(s) + 6\,H_2\,(g) + 6\,HCl\,(g) \sim 800\,^{\circ}C \qquad (10.4)$$

CVD processes depend on both chemical reactions and flow dynamics. There are two main cases: high flow rate supplies enough reactants and film deposition is limited by slow surface chemical reactions (termed “surface reaction limited”); or fast surface reaction consumes source gas rapidly and the deposition rate is limited by gas supply. This is termed “mass transport limited” or “diffusion limited”. Silicon deposition (Equation 10.1 or 10.2) on a single crystalline silicon wafer can result in a single crystalline thin film. This is termed epitaxy and it is an important special case of thin-film deposition. The chapter 3 is devoted to epitaxial deposition. Most deposition processes lead to amorphous or polycrystalline films.

Silicon dioxide can be deposited by many reactions, for example:

SiH_4 *(g)* +2 N_2O *(g)*→ SiO_2 *(s)* +2 H_2 *(g)* +2 N_2 *(g)* (10.5)

SiH_4 *(g)* + O_2 *(g)* → SiO_2 *(s)* +2 H_2 *(g)* (10.6)

$Si(OC_2H_5)_4$ −→ SiO_2 *(s)* + gaseous byproducts (10.7)

The simple reaction in Equation 10.6 is, however, problematic. Silane and oxygen can already react in the gas phase, which means that solid oxide particles are formed in the gas stream. These will then float around the reactor and sporadically deposit on the wafers. In the nitrous oxide process, oxide is formed by a surface reaction, therefore particle contamination is reduced (but in both cases oxide is formed on the reactor walls, and these films will be a source of flakes and particles if the reactor is not cleaned regularly).

The names for CVD oxides are unfortunately many. LTO, for low-temperature oxide, refers to oxide deposited by the reaction in Equation 10.6. The low deposition temperature of 425 ◦C is desirable in many cases. HTO obviously stands for high-temperature oxide (Equation 10.3), but the difference is deeper: different source gases are used, and the resulting film quality is much better at high temperatures. TEOS (Equation 10.7) is the name of the precursor molecule tetraethoxysilane $Si(OC_2H_5)_4$, but it is used as the name for the resulting oxide too (deposited at 700 ◦C, resulting in high-quality oxide). Sometimes the name USG is used: it stands for undoped silica glass. However, there are no metals in USG, so it is not glass in the traditional sense The addition of $POCl_3$ gas to the source gas flow leads to phosphorus-doped oxide deposition. The resulting film is called PSG, for phosphorus-doped silica glass.

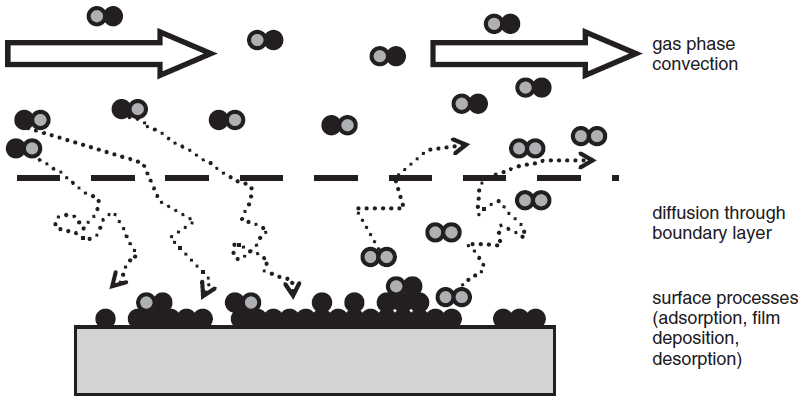


Figure 10.3 CVD: source gas molecules adsorb and react on surface to form a film, and the reaction products are desorbed, diffused and pumped away

A few percent of phosphorus (5 atomic % maximum) modifies the oxide in many ways. Phosphorous getters sodium ions which are detrimental to MOS transistors, and therefore PSG is used as a passivation layer in integrated circuits. In MEMS PSG is used as a sacrificial layer

because its etch rate in hydrofluoric acid is much faster than that of undoped CVD oxide. Phosphorus also lowers the glass transition temperature of PSG, making it possible to flow PSG at about 1000 ◦C. If both boron and phosphorus are added, we get BPSG. This oxide film flows at about 950 ◦C, resulting in smoothly sloping walls. CVD tungsten is deposited in two steps. The silane reduction step (Equation 10.8) deposits a thin nucleation layer over every surface in the system and high-rate blanket deposition with hydrogen reduction (Equation 10.9) is used to achieve the desired total thickness:

$WF_6\ (g) + SiH_4\ (g) \_\Rightarrow W\ (s) + 2\ HF + H_2\ (g) + SiF_4\ (g)$ (5.8)

$WF_6\ (g) + 3\ H_2\ (g) \_\Rightarrow W\ (s) + 6\ HF\ (g)$ (5.9)

This process is able to fill holes and trenches and is very important in multilevel metallization.

10.3.1 PECVD: Plasma-Enhanced CVD

Because high temperatures cannot be used in many cases, for example when oxide needs to be deposited on aluminum (m.p. 650 ◦C), one has to find new solutions. New source gas chemistries which enable lower deposition temperatures is one way to go. Another solution is to enhance source gas decomposition and reactions by plasmas. This results in deposition rates similar to thermal CVD, 0.1–10 nm/s, at much lower temperatures, typically around 300 ◦C, enabling deposition on most metals, for instance. Unfortunately lower deposition temperature results in less dense films. A simple parallel-plate diode reactor for PECVD is shown in Figure 10.4.Wafers are placed on a heated bottom electrode, the source gases are introduced from the top, and pumped away around the bottom electrode. The operating frequency is often 400 kHz, which is slow enough for ions to follow the field, which means that heavy ion bombardment is present. At 13.56MHz only the electrons can follow the field, and the ion bombardment effect is reduced. In thermal CVD, pressure, temperature, flow rate and flow rate ratio are the main variables. In PECVD there is additionally the RF power that can be varied.

In advanced PECVD reactors, RF power can be applied to both electrodes, and the two power sources can supply different frequencies, duty cycles and power levels. The ratio of 13.56MHz power to kilohertz power is important for film stress tailoring. PECVD shares many beneficial features of both thermal CVD and sputtering. Whereas thermal oxide or LPCVD nitride is stoichiometric SiO_2 and Si_3N_4, with ratios 1:2 and 3:4 of atoms, many other (PE)CVD films are non-stoichiometric: for example, plasma nitride is best described as SiN_x ($x \approx 0.8$). Amorphous silicon, a-Si, or more specifically designated as a-Si:H, is made by PECVD, the overall reaction being the same as that of LPCVD silicon (Equation 10.1). Hydrogen is incorporated into deposited films up to 30 at.%. Hydrogen release during annealing has to be considered: it has both beneficial and detrimental effects. PECVD can be used to deposit mixed

oxides, nitridesand carbides, as well as doped oxides just like thermal CVD. A mixture of silane, nitrous oxide and ammonia will result in oxynitride, SiO_xN_y , with varying ratios of nitrogen and oxygen, covering the whole range of compositions (and material properties) between oxide and nitride.

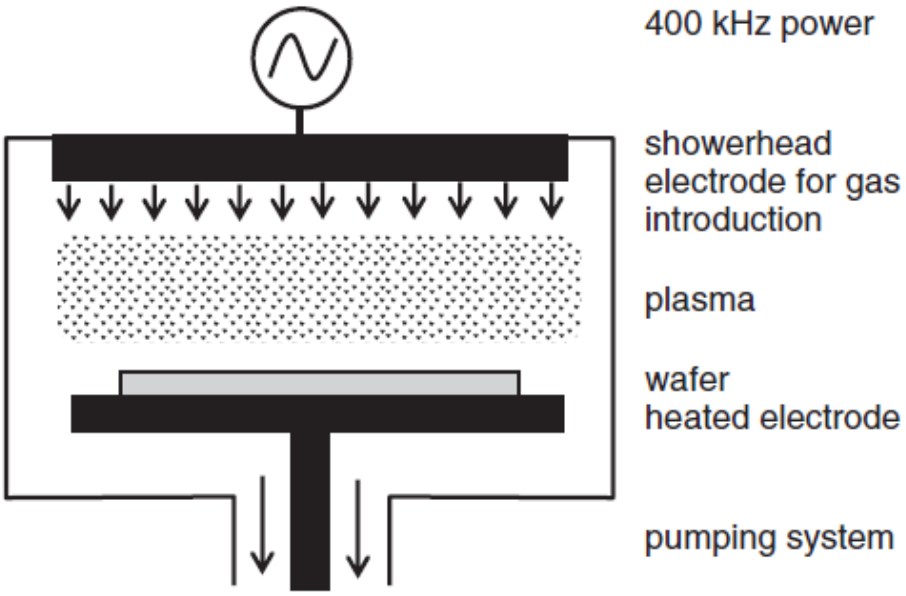


Figure 10.4 Schematic PECVD system

Silicon carbide is deposited via the reaction

$SiH_4\ (g) + CH_4\ (g) \longrightarrow SiC\ (s) + 4\ H_2\ (g)$ (10.10)

Carbon is deposited by the reaction (resembling silicon deposition, Equation 10.1)

$CH_4\ (g) \longrightarrow C\ (s) + 2\ H_2\ (g)$ (10.11)

Depending on the exact process conditions, many allotropes of carbon can be made. Non-conducting hydrogenated carbon films resemble diamond in some, but not all, respects, and they are known as diamond-like carbon, DLC. Films with less hydrogen have sp3 bonds similar to diamond, and they are referred to as ta-C, tetrahedral amorphous carbon. If intense plasma or a hot filament is used, highly reactive atomic hydrogen is produced. In this case it is possible to grow polycrystalline diamond films. Under different CVD conditions carbon nanotubes (CNTs) are made. The important factor for CNT deposition is the presence of metallic catalyst particles, for example iron or nickel.

10.4 ALD: Atomic Layer Deposition

In ALD, film is deposited one atomic layer at a time, offering ultimate thickness control. ALD works in pulsed mode: chemical bonds are formed between precursor gas molecules and

the surface atoms. Once all possible reaction sites are occupied, no more reactions can take place (Figure 10.5). A purging nitrogen pulse then removes all unreacted precursor molecules. A pulse of second precursor is then introduced. It reacts with the first reacted layer, the surface saturates similarly, and unreacted precursor gases are purged away. Repetition of successive reactant and purge pulses leads to film deposition in a layer-by-layer fashion. The ability of ALD to coat over steps is excellent because all surfaces are coated alike. Figure 10.6 shows ALD alumina and titanium multilayer films deposited over steps. This ability to coat steep topographical features is increasingly in demand as both ICs and MEMS are made more 3D.

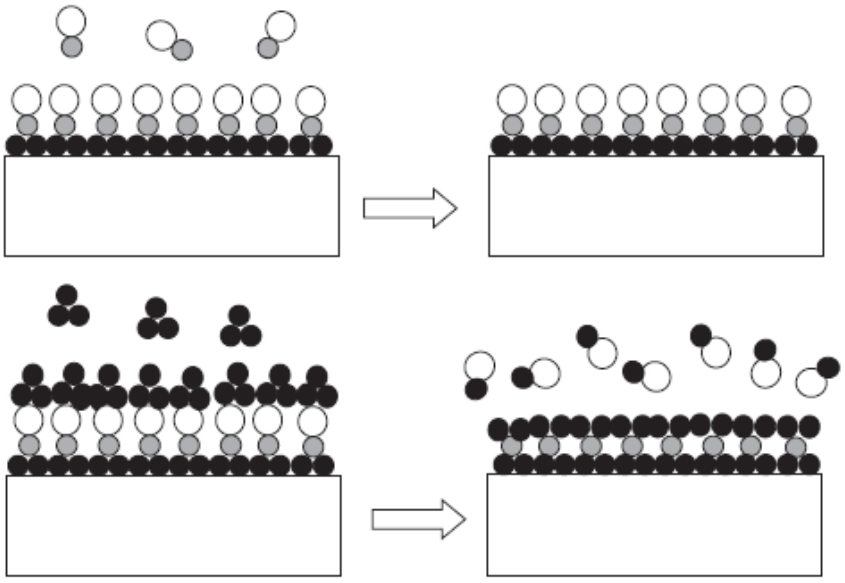


Figure 10.5 ALD: first pulse of precursors saturate wafer surface, and extra precursors are purged away by a nitrogen pulse; second precursor gases react with first layer, and reaction products are purged away

As an example of the ALD process, hafnium dioxide deposition is discussed. HfO_2 is a material with a high dielectric constant and is being used as the gate oxide in advanced CMOS. Hafnium chloride reacts with surface hydroxyl groups to form Hf–O bonds. The second precursor is water, and the oxygen in water reacts with the hafnium to form Hf–O bonds again, with hydrogen chloride formed as a byproduct. The overall reaction for hafnium dioxide deposition is given by Equation 5.12. The notation (ad) emphasizes that the reactions take place between adsorbed molecules on the surface, not in the gas phase:

$$HfCl_4\,(ad) + 2\,H_2O\,(ad) \Rightarrow HfO_2\,(s) + 4\,HCl\,(g) \qquad (10.12)$$

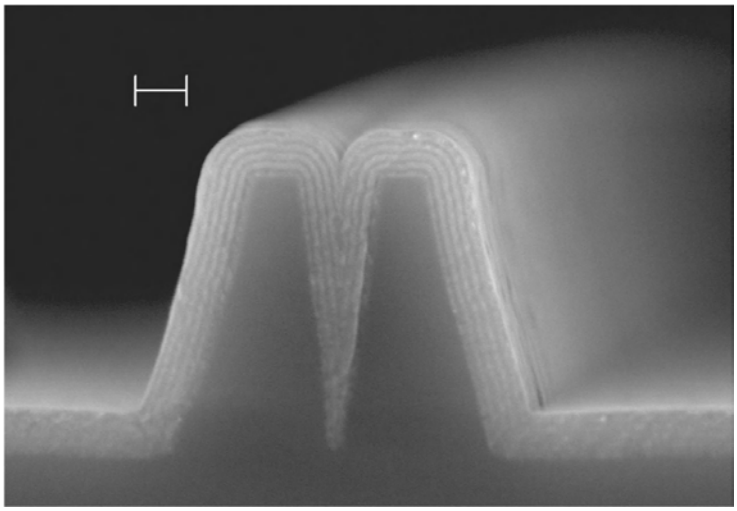


Figure 10.6 Atomic layer deposited aluminum oxide and titanium oxide thin films over silicon waveguide ridges

ALD is free of one of the main mechanisms of irreproducibility in CVD: homogeneous gas phase reactions. Because only one gas is introduced at a time, there cannot be gas phase reactions between precursors. The layer thickness is given by the number of pulses times the monolayer thickness. In theory one monolayer per pulse is deposited, but in many cases sub-monolayer growth is seen. One explanation is steric hindrance: large precursor molecules take up space, so it is simply impossible for another precursor molecule to come close enough, and some surface atoms will not react with precursor molecules. This is depicted in Figure 10.7. It can also be noted that not all surface sites are reactive enough for the ALD reaction to take place Both monolayer and sub-monolayer deposition are selflimiting. Practical growth rates range are around 1A° /cycle (0.1 nm/cycle): for Al_2O_3 deposition they are 1.1A° /cycle and for TiN, 0.2A° /cycle. When thickness/cycle numbers are translated into deposition rates, one has to take into account the flushing cycles between the pulses. Overall rates of a few nanometers per minute are typical for ALD. This is slow: for example, the LPCVD rate of polysilicon is typically 10 nm/min. But there are many applications where films of a few nanometers are needed, for example CMOS gate oxides and diffusion barriers in copper metallization.

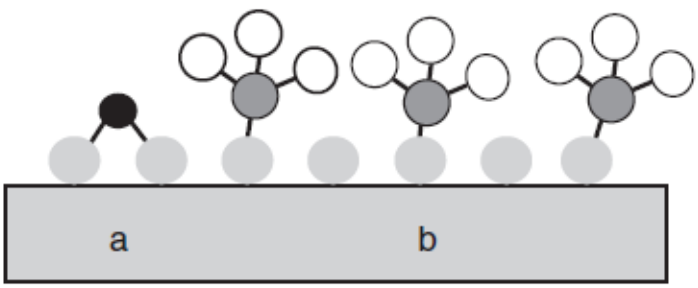


Figure 10.7 Sub-monolayer deposition in ALD: (a) nonreactive surface site; (b) steric hindrance by a large precursor molecule prevents another precursor molecule from approaching the reactive site

10.5 Electrochemical Deposition (ECD)

10.5.1 Electroplating/galvanic deposition

Electroplating takes place on a wafer that is connected as a cathode in metal ion-containing electrolyte solution. The counter-electrode is either passive, like platinum, or made of the metal to be deposited (Figure 10.8). Electroplating can be very simple: copper is deposited on the cathode according to the reduction reaction, Equation 10.13, while at the anode copper is dissolved into the electrolyte solution:

At cathode $Cu^{2+} + 2\ e^{-} \rightarrow Cu(s)$

electrolyte solution: $CuSO_4$

At anode $Cu \rightarrow Cu^{2+} + 2\ e^{-}$ (10.13)

Gold is plated in a two-step process (Equation 10.14) with the second, charge transfer reaction, as the rate limiting step:

$$Au(CN)_2^{-} \Leftrightarrows AuCN + CN^{-}$$
$$AuCN + e^{-} \Rightarrow Au\ (s) + CN^{-} \quad (10.14)$$

Electroplating rates vary a lot but are generally in the range of 0.1–10 μm/min. Deposited mass is calculated from

$$\text{mass} = \alpha ItMnF \qquad (10.15)$$

where I is the current, t the time, M the molar mass, n the species charge state, α the deposition efficiency and F the Faraday constant, 96 500 coulombs.

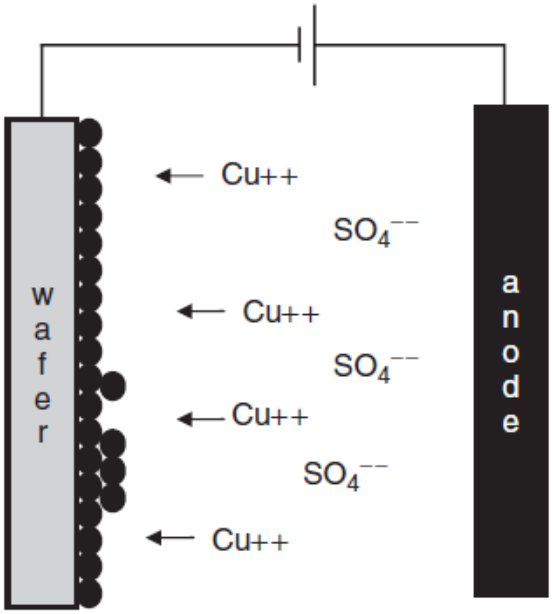


Figure 10.8 Electroplating: $CuSO_4$ electrolyte ionizes to produce Cu^{++} and SO_4^{2+} ions, copper film deposits at the Cathode

Noble metals can be deposited at 100% efficiency (α = 1.00). In less noble metal deposition hydrogen evolution makes α smaller, and for some non-metals like phosphorus co-deposition with cobalt (Co:P 12%, a soft magnetic material), α can be as low as 0.2. Other typical electroplated metals include nickel and iron–nickel (81% Ni, 19% Fe, Permalloy). Many metals have no plating processes available: aluminum, titanium, tungsten, tantalum and niobium cannot be plated. Three transport processes are active during ECD: diffusion at the electrodes due to local depletion of the reactant via deposition; migration in the electrolyte; and convective transport in the plating bath. The last is connected to electrochemical cell design, and it is affected by factors such as stirring, heating, recirculation and hydrogen evolution.Macroscopic current distribution is determined by the plating bath electrode arrangement and wafer and bath conductivity. Electrical contact to the wafer also needs careful consideration. Microscopic (local) current distribution depends on pattern density and pattern shapes. The third scale in ECD is the feature scale: potential gradients inside structures are important, especially when deep and narrow grooves are filled.

In practice the plating solutions are complex mixtures of electrolytes, salts (for conductivity control), modifiers for film uniformity and morphology improvement, as well as surfactants. Accelerators (brighteners) are additives that modify the number of growth sites. Suppressors are additives for surface diffusion control. Taken together, these additives increase the number of nucleation sites and keep the size of each nucleation site small, which drives smooth growth. Pulsed plating can also be used in balancing nucleation and grain growth: high overpotential and low surface diffusion favor nucleation, and the opposite conditions favor grain growth. Many plating solutions are proprietary. Plating baths are rather aggressive solutions, and photoresist leaching into the plating bath or adhesion loss are real concerns for reproducible plating.

10.5.2 Plating on structured wafer

Electroplating onto a photoresist pattern easily produces elaborate microstructures, like the gears shown in Figure 10.9. The process is described in Figure 10.10. A conductive seed layer is sputtered on the wafer. This seed layer, also known as the plating base or field metal, can be very thin, tens of nanometers. Photoresist is exposed and developed, and metal plating then follows. Photoresist is then removed and the seed metal is etched away, resulting in metallic microstructures.

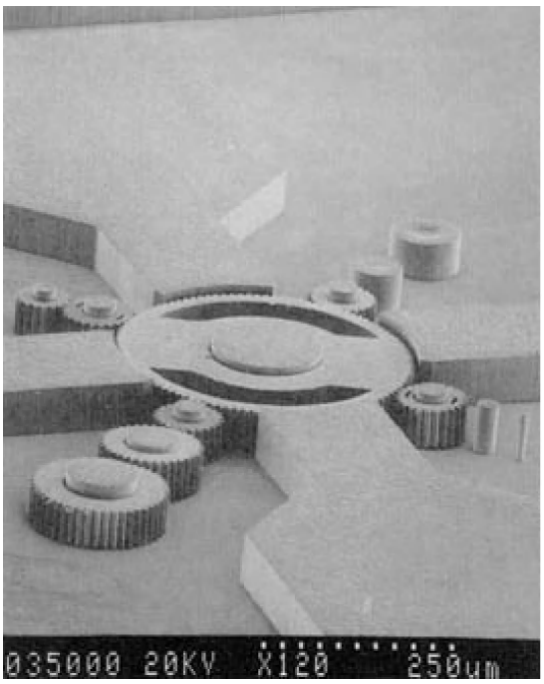


Figure 10.9 Nickel gear structures (50 μm high) made by electroplating.

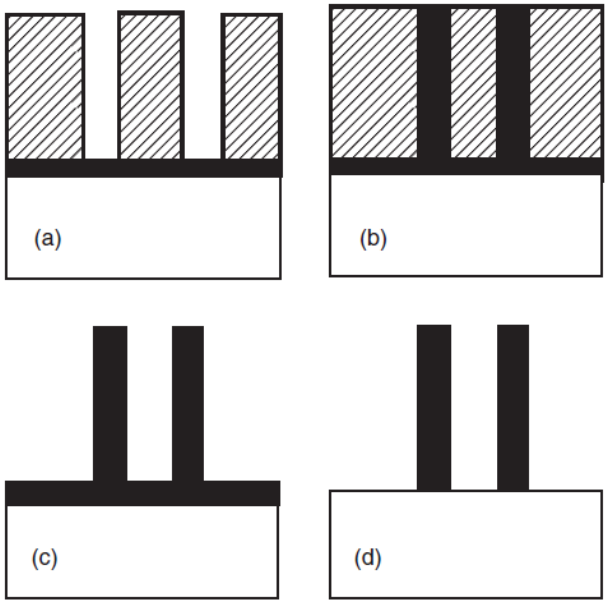


Figure 10.10 Resist masked plating (LIGA, for Lithography and Galvanic plating): (a) seed layer deposition and lithography; (b) plating; (c) resist stripping; (d) seed layer removal

The seed layer needs to be removed after plating, otherwise it would electrically short all metallized structures. Often the deposited metal itself can act as an etch mask for seed layer removal because the seed layer is always very thin compared to the plated metal; in many cases the seed layer thickness is less than the plating thickness variation. Thickness uniformity of plated metals is about 5–10%, so that 50 nm seed layer thickness is less than thickness fluctuations of plated metal 1 μm thick. Electroplating is suitable for extremely small structures, too: modern IC metallization is done by electroplating copper into trenches narrower than 100

nm wide and 200 nm high. Electroplating can fill trenches 500 μm deep and 5 μm wide (aspect ratio 100:1). Usually plating is allowed to proceed till resist top surface level but not above. It is, however, possible to overplate, and to form mushroom-shaped structures (Figure 10.11). After resist stripping, such a mushroom can be annealed (reflown) to form a ball-like bump. Bumps of Sn–Pb and In are used for flip-chip packaging. Alternatively, plating can be continued until metal fronts touch. Removal of resist underneath results in freestanding metal bridges, or in fluidic channels, depending on design details. The applications can be in RF circuits as air bridges or as cooling channels for high-power electronics.

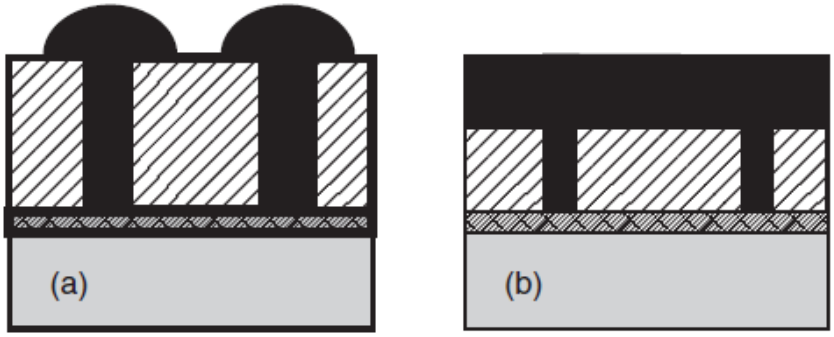


Figure 10.11 (a) Overplating; (b) backplating

10.5.3 Electroless deposition

Electroless deposition depends on a reduction reaction in an aqueous solution which contains metal salts and a reducing agent. Metal deposition takes place as a result of metal ion reduction. The surface needs to be suitable for electroless deposition and this is achieved by exposing it to a catalyst, such as $PdCl_2$. This reducing agent starts the reduction reaction which then continues locally. Selective deposition is thus possible. Gold, nickel and copper are the usual metals to be deposited by the electroless method. The major advantage of electroless deposition compared to electroplating is elimination of the need to make electrical contacts to the wafer. Copper electroless deposition chemistries traditionally use sodium hydroxide in the plating bath, but sodium is a contaminant in transistors. Alternative pH adjustment can be done with TMAH (tetramethyl ammonium hydroxide). Copper sulfate $CuSO_4$ in formaldehyde (HCHO) and EDTA (ethylene diamine tetraacetic acid) complexing agent are the basic constituents of the bath. Surfactants (polyethylene glycol) and stabilizers (2,2 -dipyridyl) can be added. The reaction is described by

$$CuEDTA2^{-} + 2\,HCHO + 4\,OH^{-} \Rightarrow Cu + H2 + 2\,H_2O + 2\,HCOO^{-} + EDTA^{4-} \quad (10.16)$$

The deposition rate is on the order of 100 nm/min. The electroless deposition set-up is extremely simple. Selectivity, however, is difficult to maintain. Hydrogen evolution and

incorporation into the film are a problem because hydrogen is mobile; carbon incorporation is another problem.Gold can be deposited from KOH, KCN, KBH4, $KAu(CN)_2$ mixture at rates exceeding 5 μm/min, even though much lower rates are usually used. Temperatures for electrochemical deposition processes range from room temperature to 100 ◦C.

10.6 Application of Metallic Thin Films in MEMS and IC

Metallic thin films have various application in microfabricated devices.

• **Conductors:** Resistivity is the main consideration: aluminum and copper are main choices for most applications, and gold is often used in RF devices, like inductor coils, to minimize resistive losses. Doped silicon and polycrystalline silicon can be used as conductors, but their resistivity is very high compared with metals.

• **Contacts to semiconductors:** Ohmic (metal-like) and Schottky (diode-like) contacts are possible. Aluminum, itself p-type dopant in silicon, makes good ohmic contact to p-type silicon. Platinum silicide is one candidate for silicon Schottky contacts

• **Capacitor electrodes:** Capacitor electrodes need not be highly conductive. The most important capacitor electrode, the MOSFET gate, is chosen to be polycrystalline silicon because its interface with silicon dioxide is stable, and its lithography and etching properties are good.

• **Plug fills:** When vertical holes need to be filled with a conducting material, CVD tungsten and electrodeposition of copper are employed. Because distances are short, it is rather step coverage than resistivity which determines the choices.

• **Resistors:** Doped semiconductors, metals, metal compounds and alloys can be used as resistors. Heating resistors can be made of almost any material, but precision resistors are difficult to make.

• **Adhesion layers:** Noble metals like gold and platinum do not adhere well to substrates, and therefore thin (10–20 nm thick) “glue” layers of titanium or chromium are needed.

• **Barriers:** Barriers are needed to prevent unwanted reactions between thin films or diffusion of unwanted atoms. Amorphous metal alloys and compounds like tungsten nitride W:N, titanium–tungsten TiW, TiN and TaN are used as barriers between metals and silicon.

• **Mechanical materials:** Aluminum, nickel and TiAl alloys are materials for micromechanical free-standing beams and cantilevers, in e.g. micromirrors and resonators. Films like TiN can be used as mechanical stiffening layers to prevent mechanical changes in the underlying, softer films, like aluminum.

Optical materials: In image sensors metals act as light shields, and chromium is used in photomasks to block light. TiN is often deposited on top of aluminum to reduce reflectivity,

because lithography is difficult of highly reflecting surface. Transparent conductors like indium doped tin oxide (ITO; $In_xSn_yO_2$) are needed in displays and light emitting devices.

• **Magnetic materials:** Nickel and nickel alloys, Ni:Fe, are used for magnetic structures in microactuators. Cores of microtransformers are also made of these materials, which are usually deposited by electroplating.

• **Catalysts and chemically active layers:** Chemical sensors, microreactors and fuel cells use films like palladium and platinum as catalysts.

• **Electron emitters:** Vacuum microemitter tips are often made of molybdenum because of its high melting point and low work function.

• **Infrared emitters and other IR components:** Heated wires emit infrared, and porous metallic films, like aluminum black, act as IR absorbers. Metallic meshes actas IR filters and aluminum is used as an IR mirror.

• **Sacrificial layers:** Many devices require free-standing structures. These must be fabricated on solid films, which will subsequently be etched away. Copper is often used as a sacrificial material under nickel or gold.

• **Protective coatings:** Sometimes the role of the topmost layer is simply to protect the underlying layers from the ambient: from etching agents or environmental stressors. Nickel and chromium are used as masks for etching.

• **X-ray components:** Masks for X-ray lithography require high atomic mass materials which effectively block X-rays. Tungsten, gold and lead are prime candidates. X-ray mirrors are made by alternating heavy (tungsten, molybdenum) and light materials (carbon or silicon) with layer thicknesses in the nanometer range.

• **Bonding layers:** Gold and tin layers are used in eutectic bonding. Tight, hermetic bonds can be obtained at fairly low temperatures when eutectic alloys are formed.

• **Bonding pads:** Wires have to be attached to chips, and this is best achieved with soft metals like aluminum and gold, while hard metals like tungsten or chromium are unsuitable for wire bonding, and also difficult to probe by probe needles.

Deposition process greatly influences the choice of metals. Not all materials are amenable to all deposition methods, and the resulting film properties (resistivity, phase, texture, adhesion, stress, surface morphology) are closely connected with the details of the deposition process, and may well be idiosyncratic with the equipment. Reproducing results that have been obtained with another piece of equipment can be a nightmare.

10.6.1 Properties of metallic thin films

Low resistivity is required of metals in thin-film form. Thin-film resistivity is usually much higher than bulk resistivity. Aluminum, copper and gold thin-film resistivities are close to bulk values; for most others thin films, resistivities are factor of two higher. Important microfabrication metals are listed in Table 5.2. Resistivities are strongly deposition process dependent, and the tabulated values should be used as guidelines with every deposition process being characterized individually. It should also be borne in mind that thermal conductivity similarly depends on the details of deposition process and film thickness.

Table 5.2 Properties of metals

Metal	Resistivity bulk (ohm-cm)	Resistivity thin film (ohm-cm)	CTE (ppm)	Thermal conductivity (W/cm-K)	Melting point (°C)
Al	3	3–4	23	2.4	650
Cu	1.7	2–4	16	4	1083
Mo	5.6	10–20	5	1.4	2610
W	5.6	10–100[a]	4.5	1.7	3387
Ta	12	20–200[a]	6.5	0.6	3000
Ti	48	100–200	8.6	0.2	1660
Co	6.2	10–20	12.5	0.7	1500
Ni	6.8	10–20	13	0.9	1455
Cr	13	20–40	6	0.7	1875
Pt	10	20–100	9	0.7	1769
Au	1.7	2–3	14	3	1064

[a] Tungsten and tantalum can exist in two different phases which have different resistivities; a minor change in sputtering conditions can result in either phase.

Alloys and compounds TiW, TiN*x* and TaN*x* have resistivities that are even more strongly deposition process dependent than simple metals, and the exact composition will also have a profound effect. Resistivities of TiW, TiN*x* and TaN*x* are usually in the range of100–500 μohm-cm. Young's moduli are the same order of magnitude for all metals, from 100 GPa for soft metals to 600 GPa for refractory metals. Many metal properties are related to the melting point. High melting point equals high bond strength, and a stable atomic arrangement in solids. This translates to a high current density tolerance.

10.7 Multilevel metallization

Thin-film deposition is seldom the last process step. The films will be modified intentionally or unintentionally in subsequent process steps. For example, all elevated temperature steps will modify thin films. So far we have been dealing mostly with single layer films. But processes and structures can be made much more functional and reliable by adopting multilayer films.

In IC metallization multiple layers of metal are used for various reasons: titanium improves adhesion, TiN acts as a barrier between materials and prevents reactions, CVD-W is used because it can fill contact holes, etc. Dielectrics are similarly used in double layer structures: passivation is provided by PSG/nitride: phosphorus-doped oxide is a good barrier for sodium ion diffusion, and nitride is an excellent mechanical scratch-protective coating. These are depicted schematically in Figure 10.12 and in the SEM micrograph of Figure 10.13.

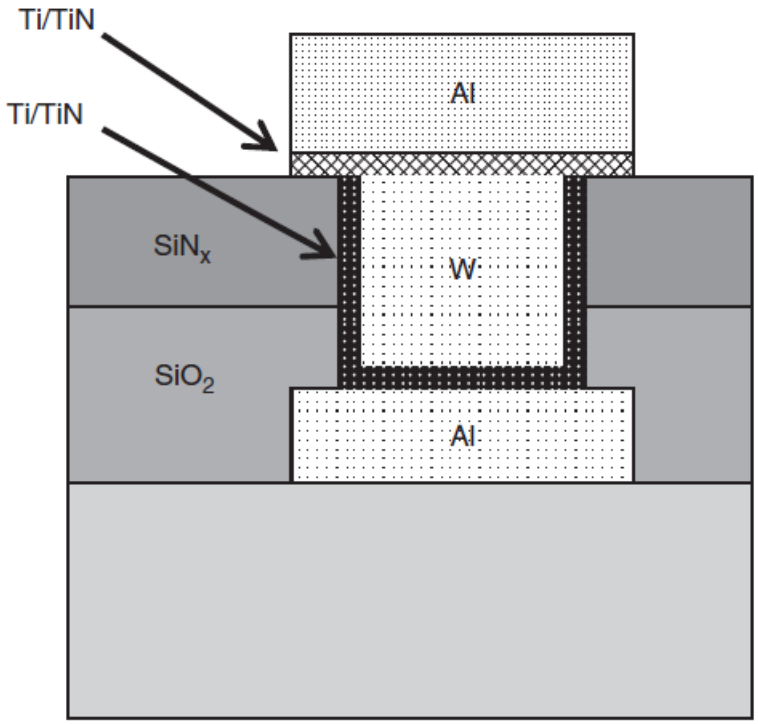


Figure 10.12 Cross-section of multilevel metallization: double layer dielectric (SiO2/SiN*x*), triple layer plug fill metallization (Ti/TiN/W) and triple layer top metallization (Ti/TiN/Al)

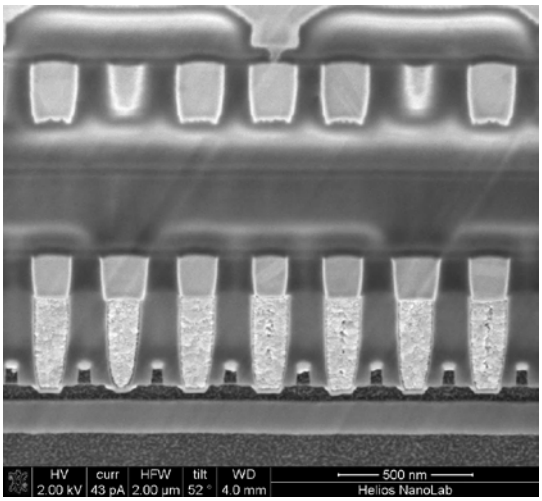


Figure 10.13 Contact plug filled by Ti/TiN/CVD-W.

10.8 Polymer Films

Polymer films can be deposited by a number of methods:

• spin coating

• dip coating

• sputtering

• evaporation

• CVD and PECVD

• self-limiting vapor phase and liquid phase reactions.

10.8.1 Spin coating

Spin coating is a very widely used method for resist spinning and increasingly for other materials as well, for example spin-on-glasses (SOGs) and polymers (known together as spin-on-dielectrics, SODs) are usually spin coated. Briefly the material is dissolved in a suitable solvent, dispensed on a wafer and spun at high speed (e.g., 1000–5000 rpm), see Figure 10.14.

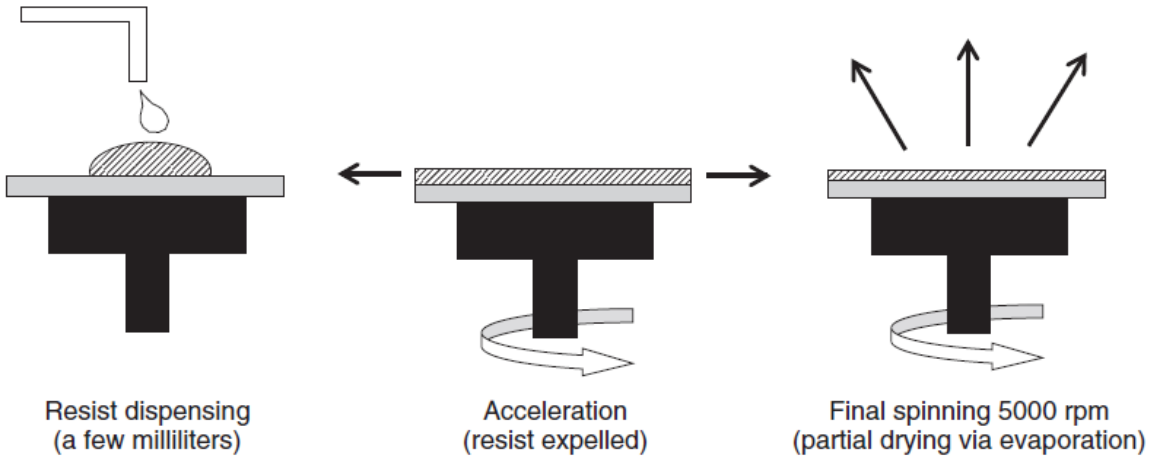


Figure 10.14 Spin coating process

Polymeric films can replace inorganic films, especially when thick films are needed. Thicknesses up to 1000 μm can be made by spin coating; inorganic films made either by CVD or by PVD cannot usually be thicker than a few micrometers. Spin-coated films fill cavities and recesses because they are liquids during spin coating. This is advantageous for filling gaps and smoothing, but if a uniform thickness over the topography is desired, spinning is not ideal. Room temperature spinning is always accompanied by baking (in the range 100–250 ◦C).

10.8.2 Self-limiting methods

After a fashion resembling ALD, monolayer thick polymer films are made by covalently bonding the molecule to a surface. Self-assembled monolayers (SAMs) are madethis way. The molecules have a reactive group at one end and a non-reactive group at the other end. In Figure

10.15 the reactive group is the trichloro group ($SiCl_3$) and the polymer chain consists of a chain of 18 carbon atoms (octadecane, or C-18) with methyl group (CH_3) at the other end. Silicon reacts with hydroxyl groups on the silicon surface, forming strong Si–O bonds, and HCl is released. When all hydroxyl groups on the surface have reacted, no more reactions can take place: there are no more reactive sites available once the surface is covered by one molecular layer.

Another SAM is shown by Equation 10.17, where fluorinated SAM reacts to form a hydrophobic (Teflon-like) layer on the surface:

$$F_3C–(CF_2)_n–Si–OH + HO–Si\ (surf) \rightarrow F_2C–(CF_2)_n–Si–O–Si\ (surf) + H_2O \quad (10.17)$$

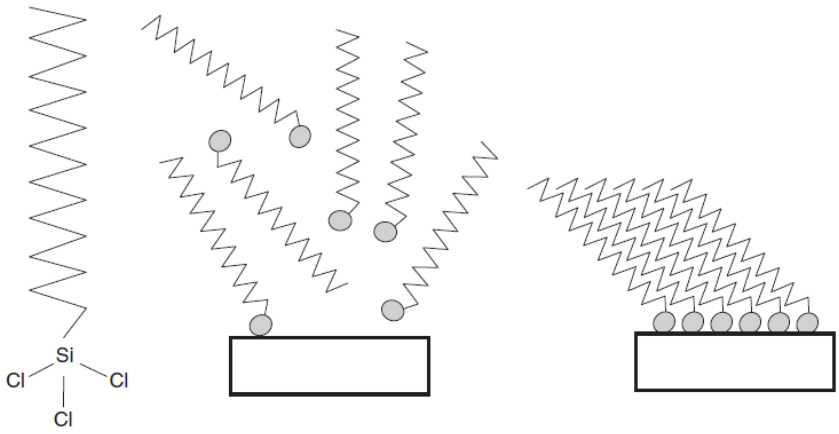


Figure 10.15 Self-assembled monolayer of octadecane trichlorosilane (OTS) on a silicon surface

10.8.3 Properties of polymers

Polymer films can offer exceptional properties, like softness, which may be required for low-pressure sensors or sensitive cantilever sensors. Many sensors use polymers as active parts of the devices: for example, capacitive humidity sensors work on the principle that capacitance changes when the polymeric capacitor dielectric absorbs water. Thin-film polymer is paramount for device operation so that humidity rapidly penetrates the whole film. Polymers are also used as structural materials in microsystems. Those structures can be thin or thick, up to millimeters. Widely used polymer materials in microfabrication include thermally stable aromatic polymers (BCB), epoxies (SU-8) and polyimides (PI). All of these are available as photoresists, too, acting in the negative mode. Non-photoactive polyimides are also widely used. Thermoplast polymers PMMA, PC and COC are used in embossing/imprinting applications. Parylene is used as a structural material, protective coating and thermal insulator. Various fluoropolymers are used to make hydrophobic surfaces. Perfluorinated films like Teflon have other uses, too, because of

their exceptional properties, like low water absorption, low friction, extreme chemical tolerance and very good electrical properties at high frequencies.

Polymers are inferior to inorganic films in terms of mechanical strength. Tensile strengths of polymers are in the range of 100–400 MPa, and Young's moduli on the order of 1–10GPa, compared to 50–500 GPa for inorganic solids and elemental metals. Stresses in polymers are inherently low (<100 MPa) whereas stress minimization in oxides and nitrides is quite a challenge. In addition to normal process-related variation, polymer properties vary from manufacturer to manufacturer, and the listed properties are indicative of some typical values only.

Polymers have thermal limitations: maximum usable temperatures are often in the range of 100–200 ◦C, but some exceptional polymers tolerate 400 ◦C. The coefficients of thermal expansion are in the range of 30–100 ppm/◦Cvs. 1–20 ppm/◦C for elemental metal films and simple inorganic compounds, which is a considerable mismatch. Many of the thin-film deposition methods described in this chapter can be applied to polymer thin films. Evaporation is used to deposit small organic molecules like pentacene ($C_{14}H_{22}$) which will deposit as a conductive thin film and can be used as a channel material in organic electronics (Figure 10.15).

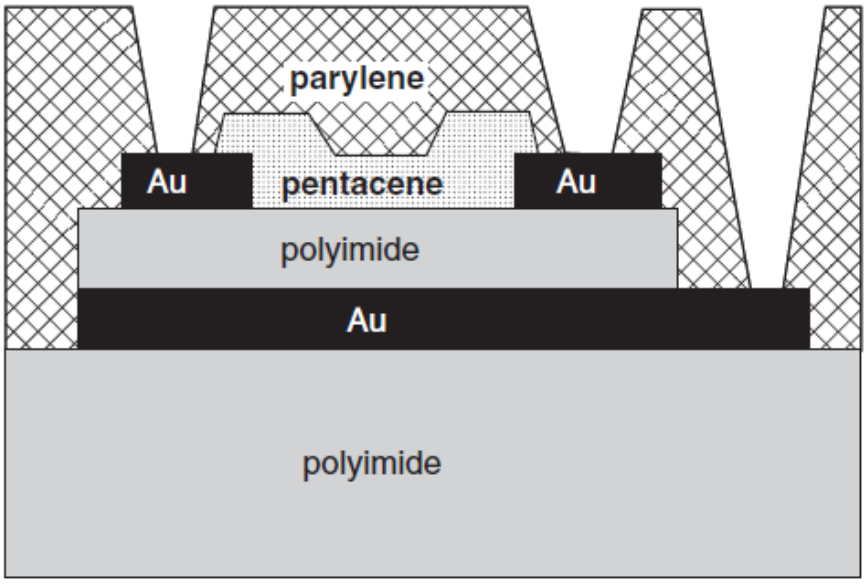


Figure 10.15 Bottom gate TFT on polyimide substrate.Spin-coated polyimide as gate dielectric, pentacene as active channel material and parylene as passivation. Metallization of Cr/Au

The same process conditions and process performance apply for the evaporation as of organics, as for any other material, for example 5 $\times 10^{-7}$ mbar pressure and 0.1 nm/s deposition rate. Polymers cannot be evaporated: they will decompose rather than evaporate.

Parylene is deposited by thermal CVD. Layer thicknesses are similar to other CVD processes in the thin end, but because of polymer softness, stress build-up is less, and layers tens of micrometers thick can be made. These were the traditional parylene applications in microtechnology: thick protecting layers for finished devices. Today, the conformal deposition of thin films, low deposition temperature, basically room temperature, enable novel applications. Teflon, an insulator, must be sputtered in a RF system (this applies to inorganic insulators as well). Deposition rate tends to be low (e.g., 0.05 nm/s), but high-density, pinhole-free film can then be obtained. Polyimides, polypropylene and polyethylene have also been sputter deposited.

Teflon-like perfluoropolymer films $(CF_2)_n$ can also be plasma deposited. Fluorine-containing source gases CHF_3 or C_4F_8 (which are readily available because they are used to plasma etch silicon dioxide) are broken down in plasma, and fluoropolymer is deposited at wafer, at room temperature. FTIR and XPS analyses will reveal the C–F bonds, and also C–H bonds which indicate incomplete fluorination.

Chapter 11. Silicon-on-Insulator technology

V.Skryshevsky

11.1 Introduction

Reducing power consumption has become the most important issue in silicon IC technology. Bulk-Si devices are now running into a number of fundamental physical limits. Among the problems are that the carrier mobility is decreasing due to impurity scattering, the gate tunneling current is increasing as the gate insulator becomes thinner, and the p-n junction leakage is increasing as the junction becomes shallower. These trends make conventional scaling less and less feasible. As a result, the operating voltage tends to be set higher than what a scaled-down device was expected to need to achieve the desired speed performance. This is certainly the case for microprocessors (MPU), for which speed is the top priority. The heat generated by today's MPUs is close to exceeding the level that can be handled by conventional cooling schemes.

Silicon-on-insulator (SOI) technology features a low capacitance, which enables high-speed operation. That is, the supply voltage can be lowered to cut power consumption while adequate speed is provided. The advantages of SOI technology also include good radiation hardness, the ability to withstand high temperatures, and the ability to handle high voltages. This technology also enables the fabrication of micro-electro-mechanical systems (MEMS) for control systems. Furthermore, it allows flexible device design; for example, the properties of the substrate can be set independently of those of the device layer because an insulator separates devices from the substrate.

To design and fabricate high performance integrated devices, one fruitful solution is to isolate electrically the semiconductor layer, in which the active elements are made, from the substrate. This can be accomplished:1) either by depositing the "active layer" on an isolating substrate, 2) or introducing a dielectric layer between the active layer and the substrate. A lot of effort has been devoted in the past twenty years to produce such structures for which two acronyms are currently being used: 1) SOI: which can be understood as "Silicon On Insulator" or "Semiconductor On Insulator" and which usually refers to (SOI) devices; and BOX: which stands for "Buried OXide" and which usually refers to the (BOX) structure. The MOSFET structures of bulk-Si and SOI substrates are presented in Fig.11.1.

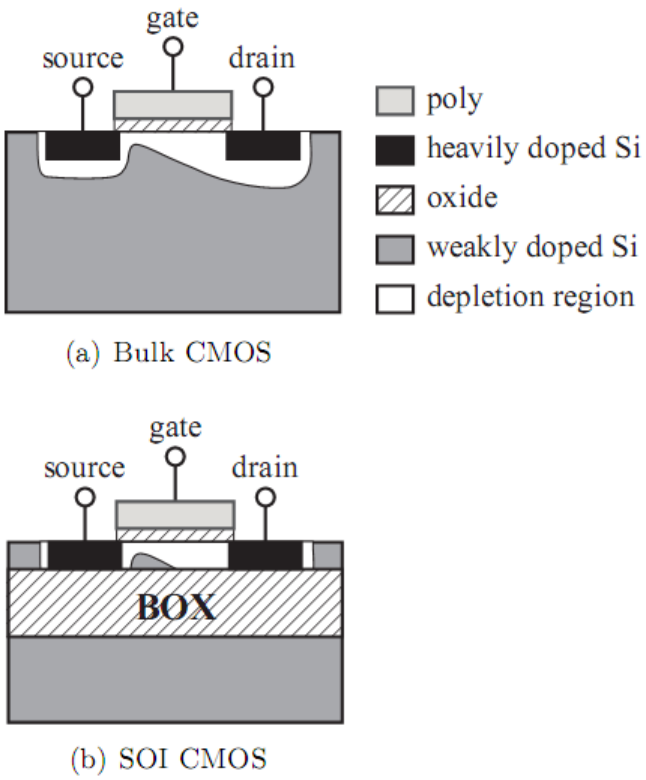


Figure 11.1 The MOSFET structures of bulk-Si and SOI substrates. The key feature of the SOI structure is the layer of silicon dioxide just below the surface.

11.2 Properties and advantages of SOI devices

At the beginning, the motivation of the development of the SOI technology is from the radiation hard properties of the SOI devices. Owing to the excellent isolation provided by the buried oxide, immunity of the SOI devices against high-energy particle illumination is excellent.

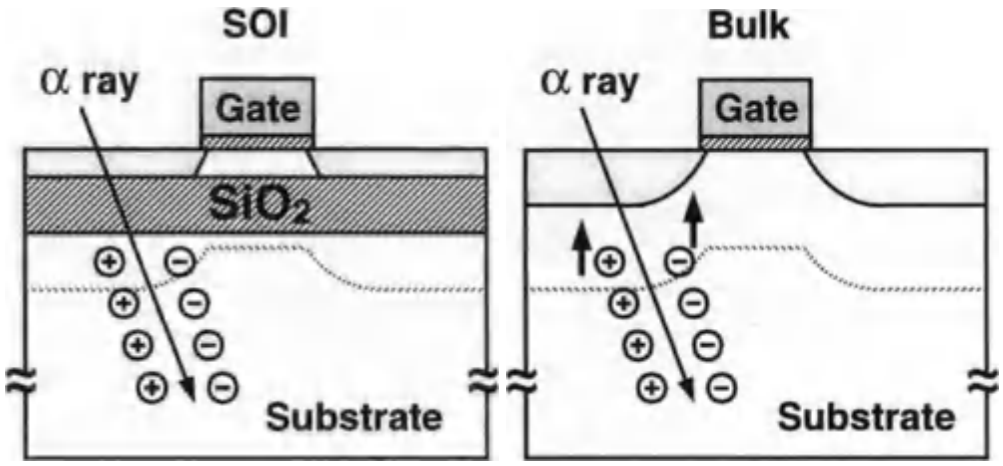


Fig. 11.2 Radiation effect on bulk and SOl MOS devices.

Fig. 11.2 shows the radiation effects on bulk and SOI MOS devices. When illuminated by a ray of the α–particles, in the silicon region where the high-energy particles pass, electron-hole pairs are generated. As a result, a large amount of current is produced. For the SOI MOS device, the active thin-film region is totally isolated from the substrate. Therefore, the α-particle induced current has little influence in the device performance. In contrast, for the bulk MOS devIce, this α-particle induced current may flow to the active region. As a result, the operation of the bulk MOS device is seriously affected. Therefore, compared to the conventional bulk MOS devices, the SOl MOS devices are more suitable for operation in the environment with high energy radiation.

Owing to the buried oxide layer, the parasitic capacitances of SOI MOS devices are smaller than those of bulk ones. In SOI devices, the capacitance between the drain (source) and the substrate is negligibly small because of the dielectric constant of SiO2, which is lower than that of Si, and the thickness of the BOX (Fig.11.3a). This helps improve thes witching speed of CMOS devices. Fig. 3b shows the comparison of the parasitic capacitances between bulk and SOI CMOS devices and the multiplication time of a 16bit x 16bit multiplier circuit built on a 0.6μm CMOS gate array in bulk and SOI technologies. As shown in the figure, except the gate-related delay, other delays of the multiplier circuit due to the junction capacitances and the wiring capacitances of the SOI CMOS have been improved substantially as compared to the bulk CMOS. As a result, the speed of the multiplier circuit using the SOl CMOS technology has been enhanced.

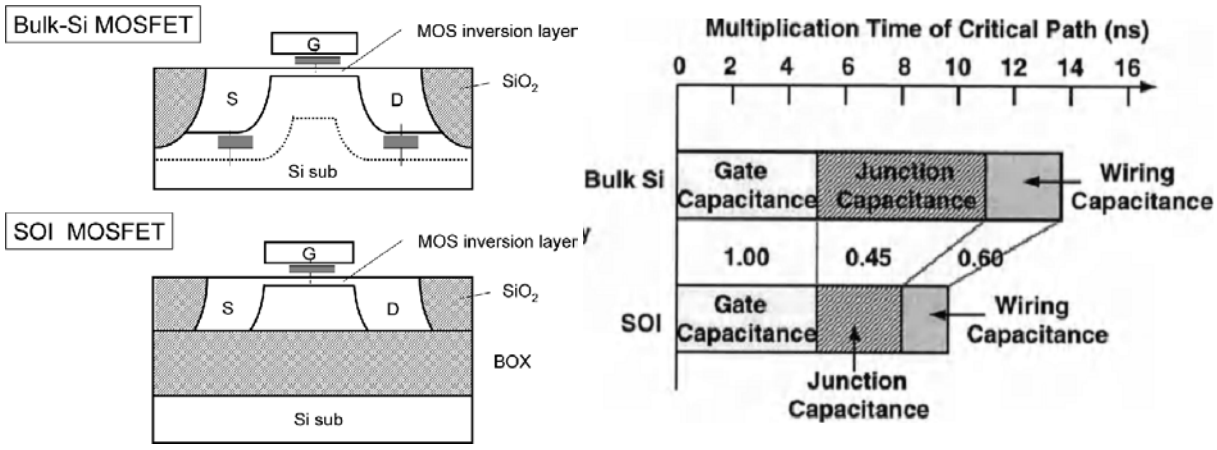


Figure 11.3. a) Capacitances of bulk-Si and SOI MOSFETs, b) Comparison of the parasitic capacitances between bulk and SOI CMOS devices and the multiplication time of a 16bit x 16bit multiplier circuit built on a 0.6μm CMOS gate array in bulk and SOl technologies.

Fig. 11.4 shows the power versus delay characteristics of the CMOS inverter circuit using SOl and bulk CMOS devices with a channel length of 0.15 μm. From this figure, owing to

smaller parasitic capacitances in the SOl MOS devices, the power-delay product of the SOl inverter circuit is much smaller as compared to the bulk one. From this figure, SOI CMOS technology has high-speed and low-power properties. For SOI CMOS devices, buried oxide has been used for isolation. For bulk CMOS devices, between device and substrate, depletion regions of the reverse-biased pn-junction have been used for isolation. Therefore, device density of the bulk CMOS technology cannot be high. For a deep submicron down-scaled bulk CMOS technology based on n-well or p-well, the increased substrate doping density may degrade the device performance. In addition, the parasitic npn and pnp bipolar junction transistor (BJT) in the well and the substrate may be accidentally turned on to cause latch-up. In contrast, in the SOI technology, owing to the buried oxide isolation, no latch-up exists. In addition, device isolation is much simpler for the SOl CMOS technology. Therefore, from processing point of view, SOI CMOS technology has a higher device density and an easier device isolation structure.

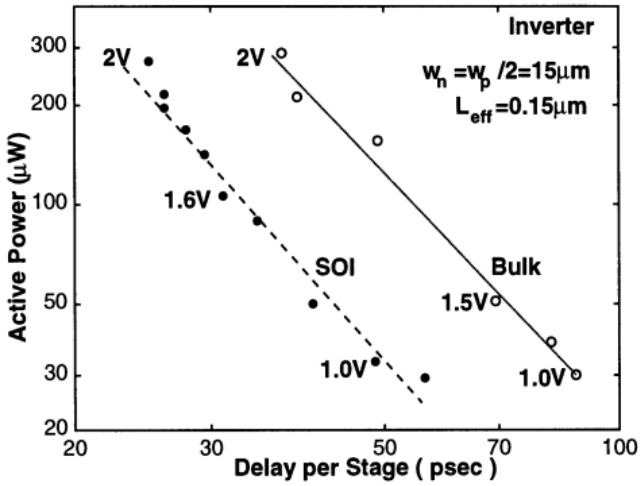


Figure 11.4 Power versus delay of the inverter circuit using bulk and SOI CMOS devices.

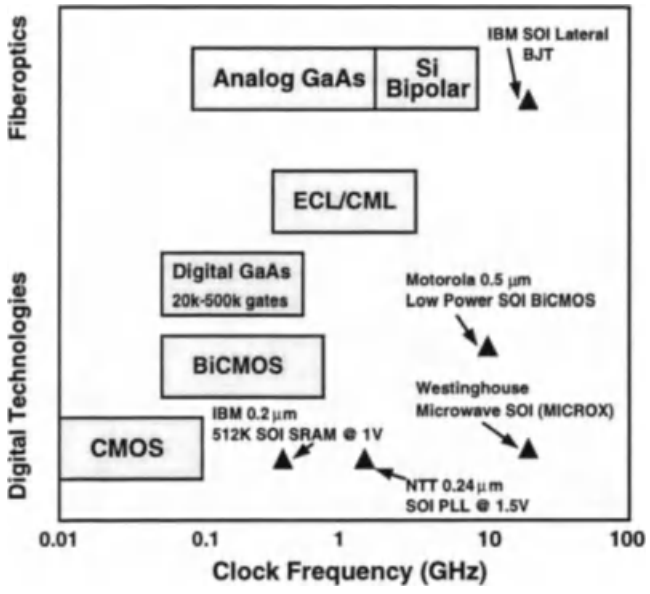


Fig.11.5 Technologies for high-speed digital circuits.

Fig. 11.5 shows the technologies for high-speed digital circuits. As shown in the figure, in addition to VLSI applications, SOI technology has also been used to realize communication circuits, microwave devices, BiCMOS devices, and even fiber optics applications for its superior performance.

With silicon-on-insulator technology, MOSFETs are formed in a thin top Si layer (SOI layer) separated from the Si substrate by an insulating film. This structural feature provides SOI devices with several advantages for high-speed, low-power operation. SOI devices are divided into two types, depending on the operating mode: fully-depleted (FD) and partially-depleted (PD) (Fig.11.6). The difference mainly derives from the thickness of the SOI layer, with the layer generally being thinner for FD-SOI devices. In PD-SOI devices, there is an undepleted neutral region at the bottom of the SOI layer; but in FD-SOI devices, the entire body region is fully depleted. This provides FD-SOI devices with additional advantages, such as a low threshold voltage, a small leakage current, and smaller floating-body effects. Because of these features, FD-SOI devices exhibit better performance at low supply voltages and lowering the supply voltage is one of the most effective ways to reduce power consumption.

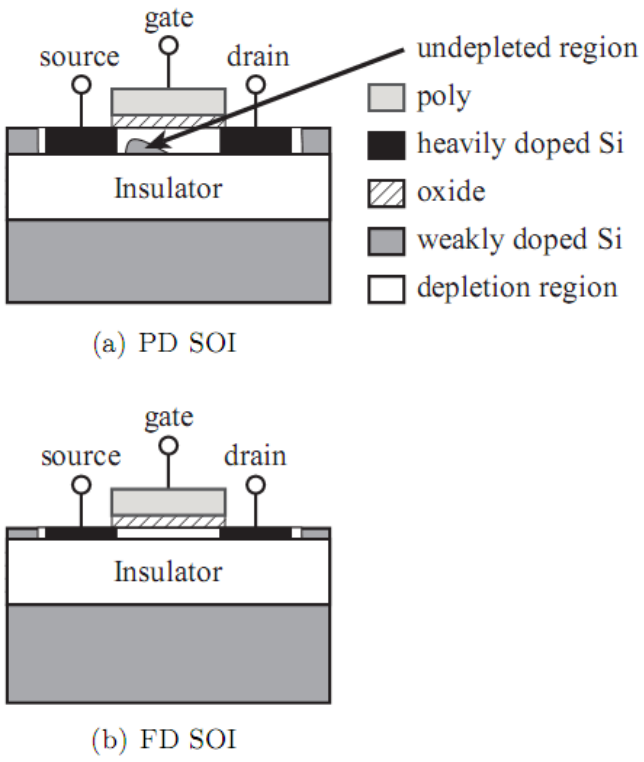


Fig.11.6 Cross section of a partially depleted and a fully depleted SOI MOSFET device.

11.2.1 The summary of Advantages of SOI technology:

1. Negligible drain-to-substrate capacitance
2. Small body effects, fast stacked gates
3. No latch-up
4. Simple device isolation, smaller area
5. Excellent radiation hardness
6. Small junction leakage current
7. Reduced short-channel effects

Table 1 Brief History of development of SOI technology.

1978 Demonstration of CMOS by SIMOX (NTT, K.Izumi)
1981-90 3-dimensional IC project in Japan
1985 High-temperature annealing in SIMOX
1986 Fully-depleted SOI MOSFETs (HP, Toshiba)
1990 Lowering the operating voltage of LSIs (<2V)
1992 PACE (Hughes)
1994 ELTRAN (Canon)
1994 Low-dose SIMOX, ITOX technology (NTT)
1995 UNIBOND (LETI)
1997-2003 Gate array (Mitsubishi), PowerPC (IBM), MPC
(Motorola), Watch controller (Oki), Opteron (AMD),
CELL (Sony, IBM, Toshiba)

Table 2 presents the most important SOl materials and the techniques involved in their fabrication.

Table 2-1. SOI Materials and associated fabrication techniques [1]

Generic Process	**Fabrication Technique (acronym)**	**Process Variations**
Silicon heteroepitaxy	Silicon-on-Sapphire (SOS)	SPEAR DSPE UTSi
	Graphoepitaxy	
	Silicon on Cubic Zirconia (SOZ)	
	Silicon on Spinel	
	Silicon on CaF_2	
Thick polysilicon deposition	Dielectric Isolation (DI)	
Polysilicon melting and recrystallization	Laser Recrystallization	Selective Annealing
	Electron Beam Recrystallization	
	Zone-Melting Recrystallization (ZMR)	LEGO
Silicon homoepitaxy	Epitaxial Lateral Overgrowth (ELO)	CLSEG PACE
	Lateral Solid-Phase Epitaxy (LSPE)	MILC
Formation of porous silicon	Full Isolation by Porous Silicon (FIPOS)	
Ion beam synthesis of a buried insulator	Separation by Implanted Oxygen (SIMOX)	ITOX SIMOX MLD
	Separation by Implanted Nitrogen (SIMNI)	
	Separation by Implanted O+N (SIMON)	
	Synthesis of silicon carbide (SiCOI)	
Wafer bonding	Wafer Bonding and Etch Back (BESOI)	PACE
Layer transfer	H^+-Induced Splitting (Smart-Cut®)	UNIBOND® NanoCleave®
	Porous Silicon Splitting	Eltran®
SiGe epitaxy	Strained silicon-on-insulator (SSOI)	
	Strained SiGe-on-insulator (SiGeOI)	
Diamond layer formation	Silicon on Diamond (SOD)	
Preferential etching	Silicon on Nothing (SON)	

11.3 Heteroepitaxial techniques

Heteroepitaxial SOI films are obtained by epitaxially growing a silicon layer on a single-crystal insulator. Reasonably good epitaxial growth is possible on insulating materials when the lattice parameters are sufficiently close to those of single-crystal silicon. Substrates can either be single-crystal bulk material, such as Al_20_3 (sapphire) or thin insulating films grown on a silicon substrate (epitaxial CaF_2). Heteroepitaxial growth of a silicon film can never produce a defect-free material by itself if the lattice parameters of the insulator do not perfectly match those of silicon. Table 3 illustrates the differing materials properties . Furthermore, the silicon film will never be stress-free if the thermal expansion coefficients of the silicon and the insulating substrate are not equal.

Table 2-2. Important parameters of materials used in heteroepitaxial SOI materials fabrication. [2]

Material	Crystal Structure	Dielectric Constant	Lattice Parameter (nm)	Thermal Expansion Coefficient (K^{-1})
Si	Diamond	11.7	0.5430	3.8×10^{-6}
Sapphire (0 1 –1 2)	Rhombohedral	9.3	0.4759	9.2×10^{-6}
Cubic Zirconia	Cubic	20	0.5206	11.4×10^{-6}
Spinel	Cubic	8.4	0.808	8.1×10^{-6}
CaF_2	Cubic	6.8	0.5464	26.5×10^{-6}

Heteroepitaxial silicon-on-insulator films are grown using silane or dichlorosilane at temperatures around 1000°C. All the insulating substrates have thermal expansion coefficients that are 2 to 3 times higher than that of silicon. Therefore, thermal mismatch is the single most important factor determining the physical and electrical properties of heteroepitaxial silicon films grown on bulk insulators. Indeed, the silicon films have a thickness which is typically 1000 times smaller than that of the insulating substrate. While the films are basically stress-free at growth temperature, the important thermal coefficient mismatch results in a compressive stress in the silicon film which reaches $=-7xI0^9$ dynes/cm^2 at the surface of a 0.5 μm silicon-on sapphire (SOS) film. An even higher value is reached at the Si-sapphire interface. Such stresses may equal or exceed the yield stress of silicon, resulting in relaxation in the silicon film via generation of crystallographic defects such as microtwins, stacking faults and dislocations.

11.3.1 Silicon-on-Sapphire (SOS)

Silicon-on-Sapphire (SOS) material was first introduced 1964. It is obtained by epitaxial growth of silicon on a (1 1 -1 2)-oriented crystalline alumina (α-AI_20_3 also called sapphire) wafer. The sapphire crystals are produced using Czochralski growth. The sapphire boule is sliced into wafers that are then subjected to mechanical and chemical polishing. The sapphire wafers

receive a final hydrogen etching at 1150°C in an epitaxial reactor, and a silicon film is deposited using the pyrolysis of silane at temperatures between 900 and 1000°C. The lattice constant of silicon and sapphire are 0.543 and 0.475 nm, respectively, The thermal expansion for silicon and sapphire are 3.8×10^{-6} and 9.2×10^{-6} K^{-1}, respectively. Due to the lattice mismatch between sapphire and silicon the defect density in the silicon film is quite high, especially in very thin films. As the film thickness increases, however, the defect density appears to decrease as a simple power law function of the distance from the Si-Sapphire interface. The main defects present in the as-grown SOS films are stacking faults and (micro)twins. Typical defect densities near the Si-Sapphire interface reach values as high as 10^6 planar faults/em and 10^9 line defects/cm^2. These account for low values of resistivity, mobility, and lifetime near the interface. Because the epitaxial silicon is deposited at high temperature and because the thermal expansion coefficients of silicon and sapphire are different the silicon film is under compressive stress at room temperature. Several techniques have been developed to reduce both the defect density and the stress in the SOS films . The Solid-Phase Epitaxy and Regrowth (SPEAR) and the Double Solid-Phase Epitaxy (DSPE) techniques are other more successful methods for improving the crystal quality of SOS films. These techniques employ the following steps. First, silicon implantation is used to amorphize the silicon film, with the exception of a thin superficial layer, where the original defect density is lowest. Then a thermal annealing step is used to induce solid-phase regrowth of the amorphized silicon, the top silicon layer acting as a seed. A second silicon implant is then used to amorphize the top of the silicon layer, which is subsequently recrystallized in a solid-phase regrowth step using the bottom of the film as a seed. In the SPEAR process, an additional epitaxy step is performed after solid-phase regrowth. Using such techniques, substantial improvement of the defect density is obtained. Noise in MOS devices is reduced, and the minority carrier lifetime is increased by two to three orders of magnitude

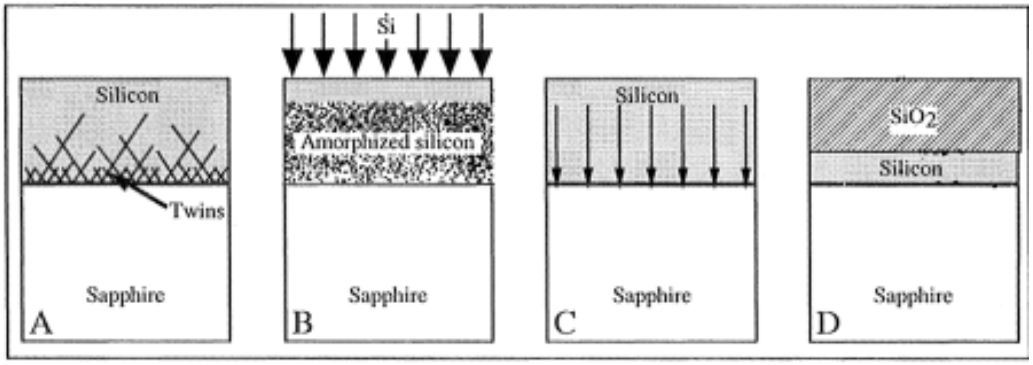


Fig. 11.7 UTSi SOS process. A: Growth of a relatively thick epitaxial silicon film; B: Amorphization using silicon ion implantation; C: Solid-phase regrowth downward from the defect-free surface; D: Thinning of the silicon film by thermal oxidation.

The most recent technique used to produce high-quality SOS is the UTSi (Ultra-Thin Silicon) process: a relatively thick film of silicon is grown on sapphire and, as in the SPEAR process, silicon ion implantation is used to amorphize the silicon film below the most superficial layer, which is relatively defect -free (Fig.11.7). Low-temperature annealing is then used to regrow the defect-free silicon downward from the surface through a solidphase epitaxy mechanism. The silicon film is then thinned to the desired thickness (100 nm) by thermal oxidation and oxide strip. This process delivers relatively defect-free and stress free SOS material in which devices
with a high effective mobility can be fabricated.

11.3.2 Silicon-on-Zirconia (SOZ)

Yttrium-stabilized cubic zirconia [$(Y_2O_3)_m \cdot (ZrO_2)_{1-m}$] can also be used as an alternative dielectric substrate for silicon epitaxy. Indeed, zirconia is an oxygen conductor at high temperature. This means that , while being an excellent insulator at room temperature ($p > 10^{13}$ Ω.cm), cubic zirconia is permeable to oxygen at high temperature. This unique property has been used to grow an SiO_2 layer at the silicon-zirconia interface *(i.e.* to oxidize the most defective part of the silicon film) by the transport of oxygen through a 500 μm-thick zirconia substrate. The growth of a 160 nm thick film at the interface necessitates only 100 min at 925°C in pyrogenic steam.

11.3.3 Silicon-on-Spinel

Spinel [$(MgO)_m \cdot (AI2O3)_{1-m}$] can be used as a bulk insulator material or can be grown epitaxially on a silicon substrate at a temperature between 900 and 1000°C. Stress-free 0.6 μm silicon-on-spinel films have been grown, but the properties of MOSFETs made in this material are inferior to those of devices made in SOS films due to higher defect density.

11.3.4Silicon on Calcium Fluoride

Like spinel, calcium fluoride (CaF_2) can be grown epitaxially on silicon. Fluoride mixtures can also be formed and their lattice parameters can be matched to those of most semiconductors. For example, $(CaF_2)_{0.55} \cdot (CdF_2)_{0.45}$ has the same lattice parameters as Si at Troom, and $(CaF_2)_{0.42} \cdot (SrF_2)_{0.58}$ is matched to Ge. Unfortunately, the thermal expansion coefficient of these fluorides is quite different from that of Si, and lattice match cannot be maintained over any appreciable temperature range. Silicon can, in turn, be grown on CaF_2 using MBE or e-gun evaporation at T= 800°C. As in the case of Si films on epitaxial spinel, $Si/CaF_2/Si$ films are essentially stress-free, which can be readily understood by noticing that the mechanical support is a silicon wafer, which, of course, has the same thermal expansion

coefficient as the top silicon film. MOSFETs have been made in $Si/CaF_2/Si$ material and exhibit surface electron and hole mobility of 570 and 240 $cm^2/V.s$, respectively.

11.4 Polysilicon melting and recrystallization

MOS transistors can be fabricated in a layer of polysilicon deposited on an oxidized silicon wafer, but the presence of grain boundaries brings about low surface mobility values (=10 $cm^2/V.s$) and high threshold voltages (several volts). Grain boundaries contain silicon dangling bonds giving rise to a high density of interface states (several 10^{12} $cm^{-2}V^{-1}$) which must be filled with channel carriers before threshold voltage is reached. Above threshold, once the traps are filled, the grain boundaries generate potential barriers which have to be overcome by the channel carriers flowing from source to drain . This gives rise to the low values of mobility observed in polysilicon devices. Mobility can be improved and more practical threshold voltage values can be reached if the silicon dangling bonds in the grain boundaries are passivated. This can be performed by exposing the wafers to a hydrogen plasma, during which the fast diffusing atomic hydrogen can penetrate the grain boundaries and passivate the dangling bonds. This treatment can improve the drive current of polysilicon devices by a factor of 10, due to both an increase of mobility and a reduction of threshold voltage. Hydrogen passivation also significantly improves the leakage current of the devices. High-performance IC applications, however, require much better device properties, and grain boundaries must be eliminated from the silicon film . This is the goal of the poly silicon melting and recrystallization techniques. In these techniques a polysilicon film is deposited on an oxidized silicon wafer and melted using a focused laser beam, an electron beam, a heated carbon strip or a halogen lamp. Quenching conditions are then controlled such that the polysilicon film is converted into relatively large silicon crystals.

11.4.1 Laser recrystallization

The laser recrystallization of polycrystalline silicon have been carried out with pulsed lasers (ruby laser) and continuous-wave (cw) lasers such as CO_2. Silicon is transparent at the 10.6 μm wavelength produced by CO_2 lasers. Therefore, silicon films cannot be directly heated by CO_2 lasers; SiO_2, on the other hand, absorbs the 10.6 μm wavelength with a penetration depth of ~ 10 μm. Polysilicon films deposited on SiO_2 (quartz or an oxidized silicon wafer) or covered by an SiO_2 cap can thus be melted through "indirect" heating produced by CO_2 laser irradiation (>100W). CW YAG:Nd lasers can output high power beams (300 W) at a wavelength of 1.06 μm. Silicon is transparent at this wavelength, but if the wafer is preheated to a temperature of 1200-1300°C, free carriers are generated in the silicon and the 1.06 μm wavelength can be

absorbed. CW Ar lasers, on the other hand , emit two main spectral lines at 488.0 and 514.5 nm (blue and green), and can reach an output power of 25 W when operated in the multiline mode. These wavelengths are well absorbed by silicon. In addition to this, the reflectivity of silicon increases abruptly once melting is reached. This effect is very convenient since it acts as negative feedback on the power absorption and prevents the silicon from overheating above melting point.

The laser beam is focused on the sample by means of an achromatic lens into a circular or, more often, an elliptical spot. Scanning of the beam is achieved through the motion of galvanometer-driven mirrors. The size of the molten zone and the texture of the recrystallized silicon depend on parameters such as laser power, laser intensity profile, substrate preheating temperature (the wafer is held on a heated vacuum chuck), and scanning speed. Typical recrystallization conditions of a 500 nm-thick LPCVD poly silicon film deposited on a 1 μm thermal oxide grown on a silicon wafer are : spot size of 50-150 μm (defined as the laser spot diameter at 1/e intensity, TEM_{00} mode), power of 10-15 W, scanning speed of 5-50 cm/sec, and substrate heating at 300-600°C.

Silicon films recrystallized on an amorphous SiO_2 substrate have a random crystal orientation. X-ray diffraction studies of polysilicon recrystallized with a Gaussian laser profile indicate the presence of crystallites having (111) , (220), (311) , (400), (331) , (110) , and (100) orientations. This is clearly unacceptable for device fabrication, since different crystal orientations will result in different gate oxide growth rates. In addition, in between the crystallites, grain boundaries exist which act to reduce carrier mobility. Ideally, one wants a uniform (100) orientation for all crystallites. From there comes the idea of opening a window (seeding area) in the insulator to allow contact between the silicon substrate and the polysilicon layer. Upon melting and recrystallization, lateral epitaxy can take place and the recrystallized silicon will have a uniform (100) orientation, as shown in Fig.11.8.

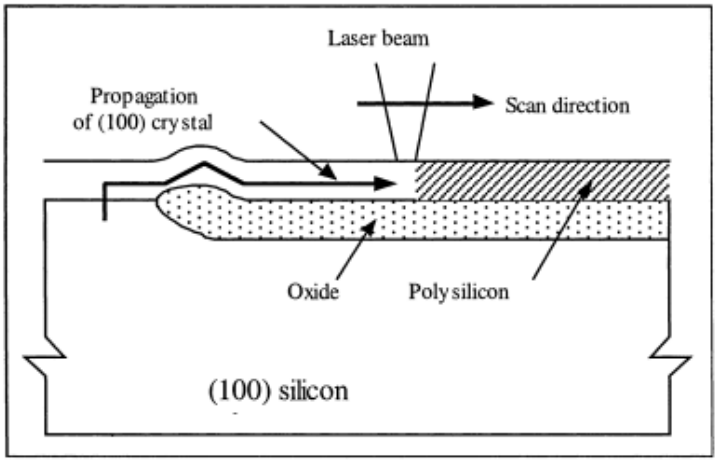


Fig. 11.8. Principle of the lateral seeding process.

In order to obtain not only a single large crystal, but a large single-crystal area, the laser beam must be raster-scanned on the wafer with some overlap between the scans. Unfortunately, small random crystallites arise at the edges of the large crystals, which precludes the formation of large single crystal areas, and grain boundaries are formed between the single-crystal stripes. The location of these grain boundaries depends on the scanning parameters and the stability of the beam. In other words, from a macroscopic point of view, the location of the boundaries is quasi-random, and the yield of large circuits made in this material will be zero. A solution to this problem is to use stripes of an anti reflecting (AR) material (SiO_2 and/or Si_3N_4) to obtain the photolithographically-controlled shaping of the molten zone and produce a succession of concave solidification fronts at the trailing edge of the molten zone. This technique is called "selective annealing" because more energy is selectively deposited on the silicon covered by AR material. It permits the growth of large adjacent crystals with straight grain boundaries, the location of which is controlled by a lithography step (Figure 11.9). The technique can be used with a laser scan parallel, slanted or perpendicular to the anti reflection stripes (AR stripes). Using this technique, chip-wide (several mm x several mm) defect-free, (100)-oriented single-crystal areas have been produced.

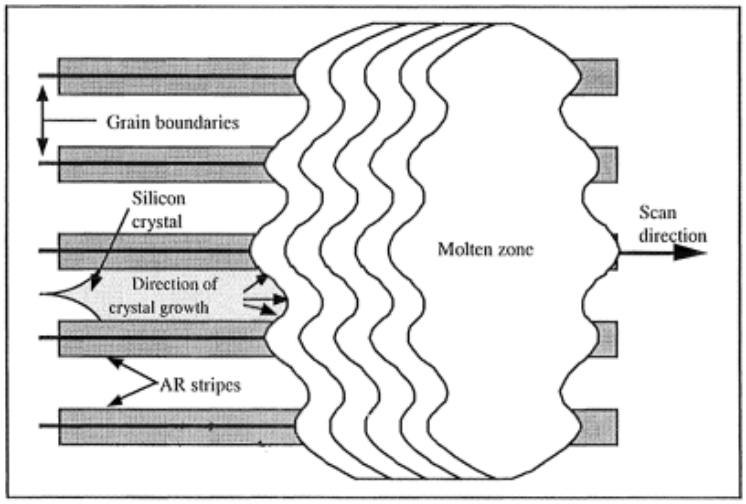


Figure 11.9 Recrystallization using anti reflection stripes (selective annealing).

11.4.2 E-beam recrystallization

The recrystallization of a polysilicon film on an insulator using anelectron beam (e-beam) is in many respects very similar to the recrystallization using a continuous-wave (cw) laser. Similar seeding techniques are used, and an (SiO_2 and/or Si_3N_4) encapsulation layer is used to prevent the melted silicon from de-wetting. The use of an e-beam for recrystallizing Sal layers has some potential advantages over laser recrystallization since the scanning of the beam can be controlled

by electrostatic deflection, which is far more flexible than the galvanometric deflection of mirrors . Indeed, the oscillation frequency of a laser beam scanned using galvanometer-driven mirrors is limited to a few hundred hertz, while e-beam scan frequencies of 50 MHz have been utilized. The absorption of the energy deposited by the electron beam is almost the same in most materials, such that the energy absorption in a sample is quite independent of crystalline state and optical reflectivity of the different materials composing it. This improves the uniformity of the recrystallization of silicon deposited over an uneven substrate, but precludes the use of a patterned anti reflection coating. Structures with tungsten stripes have, however, been proposed to achieve differential absorption.

11.4.3 Zone-melting recrystallization

One of the main limitations of laser recrystallization is the small molten zone produced by the focused beam, which results in a long processing time needed to recrystallize a whole wafer. Recrystallization of a polysilicon film on an insulator can also be carried out using incoherent light (visible or near IR) sources. In this case, a narrow (a few millimeters) but long molten zone can be created on the wafer. A molten zone length of the size of an entire wafer diameter can readily be obtained. As a result, full recrystallization of a wafer can be carried out in a single pass. Such a recrystallization technique is generally referred to as Zone-Melting Recrystallization (ZMR) because of the analogy between this technique and the float-zone refining process used to produce silicon ingots. The first method which successfully achieved recrystallization of large-area samples makes use of a heated graphite strip which is scanned across the sample to be recrystallized. The set-up is called "graphite strip heater" (Figure 11.10). A heated graphite susceptor is used to raise the temperature of the entire sample up to within a few hundred degrees below the melting temperature of silicon. Additional heating is locally produced at the surface of the wafer using a heated graphite strip located a few millimeters above the sample and scanned across it heater.

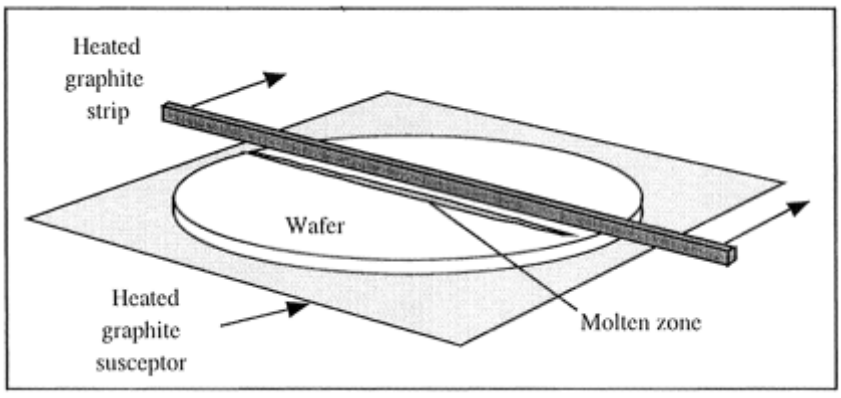


Figure 11.10: Zone-Melting Recrystallization of an SOl wafer using a grap hite strip

A typical sample is made of a silicon wafer on which a 1 μm-thick oxide is grown and a 0.5 μm -thick layer of LPCVD amorphous or polycrystalline silicon is then deposited. The whole structure is capped with a 2 μm –thick layer of deposited SiO_2 covered by a thin Si_3N_4 layer. The capping layer helps minimize mass transport and protects the molten silicon from contaminants (such as carbon from the strip heater). Recrystallization is carried out in a vacuum or an inert gas ambient in order to keep the graphite elements from burning. Both the graphite susceptor and the graphite strip can be replaced by lamps to achieve ZMR of SOI wafers. A lamp recrystallization system is composed of a bank of halogen lamps which is used to heat the wafer from the back to a high temperature (1100°C or above), and a top halogen or mercury lamp whose light is focused on the sample by means of an elliptical reflector (Figure 11.11*).* An unpolished quartz plate may be inserted between the lamp bank and the wafer in order to diffuse the light and homogenize the energy deposition at the back of the wafer. As in the case of strip heater recrystallization, a narrow, wafer-long molten zone is created and scanned across the wafer with a speed on the order of 0.1-1 mm/sec. ZMR can also be carried out using an elongated laser spot; a linear molten zone can be created using a high-power (300 W) continuous-wave YAG:Nd laser (l.06 μm).

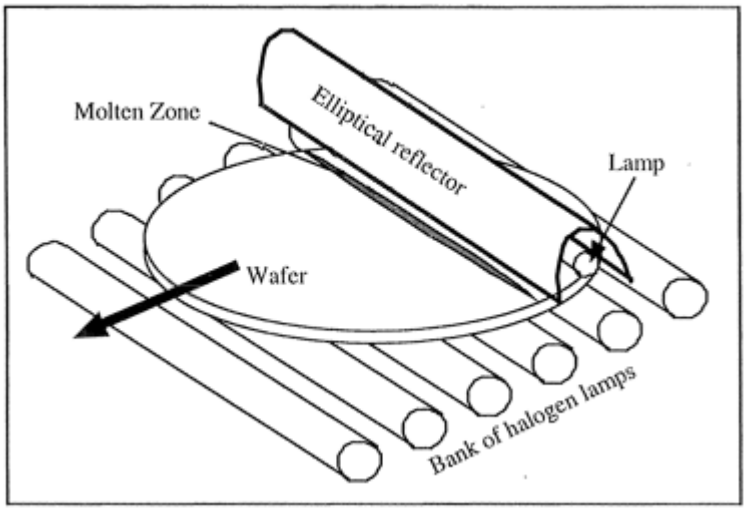


Figure 11.11: ZMR rccrystallization of an SOl wafer using lamps.

11.5. Homoepitaxial techniques

Silicon-on-insulator can be produced by homoepitaxial growth of silicon on silicon, provided that the crystal growth can extend laterally on an insulator (usually, SiO_2). This can be achieved either using a classical epitaxy reactor or by lateral solid-phase crystallization of a deposited amorphous silicon layer.

11.5.1 Epitaxial lateral overgrowth

The Epitaxial Lateral Overgrowth technique (ELO) consists of the epitaxial growth of silicon from seeding windows over SiO_2 islands or devices capped with an insulator. It can be performed in an atmospheric or in a reduced - pressure epitaxial reactor. The principle of ELO is illustrated in Figure 11.12. Typical sample preparation for ELO involves patterning windows in an oxide layer grown on a (100) silicon wafer. The edges of the windows are oriented along the <010> direction. After cleaning, the wafer is loaded into an epitaxial reactor and submitted to a high-temperature hydrogen bake to remove the native oxide from the seeding windows. Epitaxial growth is performed using *e.g.* an $SiH_2Cl_2 + H_2 + HCl$ gas mixture. However, nucleation of small silicon crystals with random orientation occurs on the oxide. These crystallites can be removed by an *insitu* HCl etch step. Once the small nuclei are removed, a new epitaxial growth step is performed, followed by an etch step, and so on, until the oxide is covered by epitaxial silicon. The epitaxial growth proceeds from the seeding windows both vertically and laterally, and the silicon crystal is limited by <100> and <101> facets (Figure 11.12,A). When two growth fronts, seeded from opposite sides of the oxide, join together, a continuous siliconon-insulator film is formed, which contains a low-angle subgrain boundary where the two growth fronts meet.

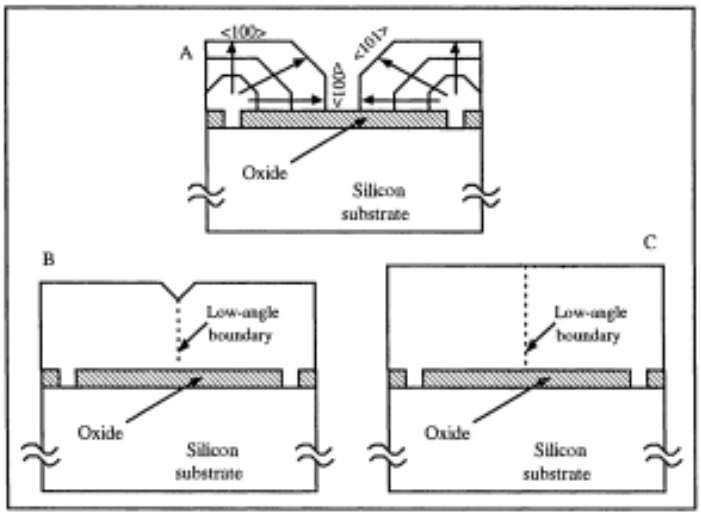


Figure 11. 12 Epitaxial Lateral Overgrowth (ELO): growth from seeding windows (A), coalescence of adjacent crystals (B), self-planarization of the surface (C).

Because of the presence of <101> facets on the growing crystals, a groove is observed over the center of the SOI area (Figure 11.12,B). This groove, however, eventually disappears if additional epitaxial growth is performed (Figure 11.12,C). Three-dimensional stacked CMOS inverters have been realized by lateral overgrowth of silicon over MOS devices. The major disadvantage of the ELO technique is the nearly 1:1 lateral-to-vertical growth ratio, which means that a 10 μm -thick film must be grown to cover 20 μm -wide oxide patterns (10 μm from each

side). Furthermore, 10 additional micrometers must be grown in order to get a planar surface. Thinner SOI films can, however, be obtained by polishing the wafers after the growth of a thick ELO film. The ELO technique has been used to fabricate three-dimensional and double-gate devices.

A variation of the ELO technique, called "tunnel epitaxy", "confined lateral selective epitaxy" (CLSEG) or "pattern-constrained epitaxy" (PACE). In this technique, a "tunnel" of SiO_2 is created, which forces the epitaxial silicon to propagate laterally (Figure 11.13). With this method, a 7: I lateral-to-vertical growth ratio has been obtained.

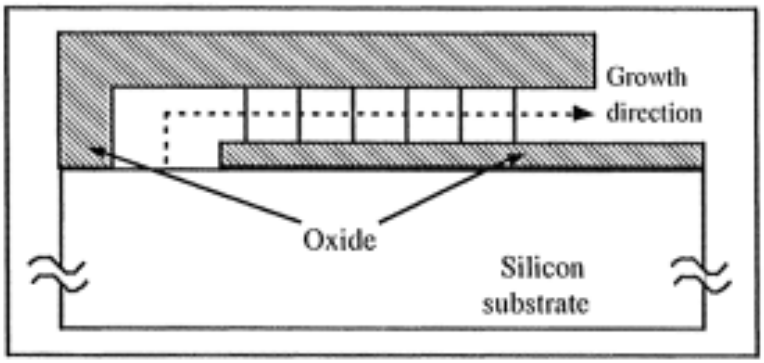


Figure 11.13 Principle of tunnel epitaxy .

11.5.2 Lateral solid-phase epitaxy

Lateral Solid-Phase Epitaxy (LSPE) is based on the lateral epitaxial growth of crystalline silicon through the controlled crystallization of amorphous silicon (a-Si). A seed is needed to provide the crystalline information necessary for the growth. The thin amorphous silicon film can either be deposited or obtained by amorphizing a polysilicon film by means of a silicon ion implantation step. LSPE is performed at relatively low temperature (575-600°C) in order to obtain regrowth while minimizing random nucleation in the amorphous silicon film. The LSPE technique has been used to fabricate double-gate SOI MOSFETs.

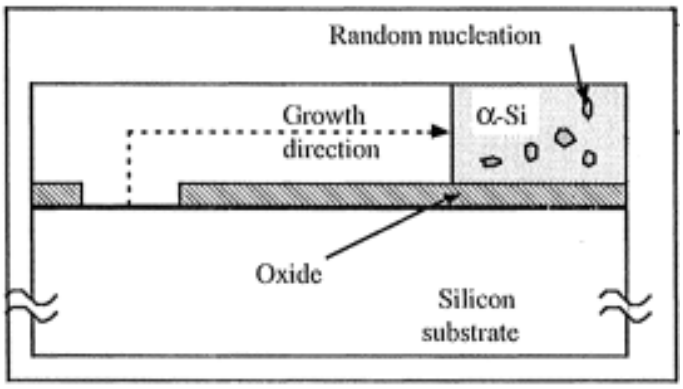


Figure 11.14: Principle of lateral solid-phase epitaxy.

The lateral epitaxy rate is on the order of 0.1 nm/s in undoped a-Si and 0.7 nm/s in heavily ($3x10^{20}cm^{-3}$) phosphorous-doped a-Si. The distance over which lateral epitaxy can be performed over an oxide layer is in the order of 8 μm for undoped material, and 40 μm if a heavy phosphorus doping is used, further lateral extension of LSPE being limited by random nucleation in the a -Si film (Figure 11.14).

Solid-phase growth of crystals from amorphous silicon can also be stimulated by the presence or contact of a metal such as palladium, aluminum or nickel. This technique, called "metal-induced lateral crystallization" (MILC) has been used to fabricate thin-film transistors (TFTs), gate-all-around transistors, and three-dimensional CMOS SOI integrated circuits. In a typical MILC process used for making TFTs, amorphous silicon is deposited on SiO_2 and capped with a low-temperature deposited oxide layer. After gate formation, contact holes are opened in the source/drain areas and nickel is deposited. Upon annealing at a temperature close to 400°C $NiSi_2$ forms in the contact holes. Annealing at 500°C is then performed during which MILC occurs, as part of the original $NiSi_2$ moves through the amorphous silicon, leaving behind a trail of long, needle-shaped silicon crystals. The passage of $NiSi_2$, through the silicon leaves behind approximately 0.02 atomic percent of nickel in the crystallized silicon. MILC recrystallization of over 40 μm has been demonstrated. The typical growth rate ranges between 0.25 and 1 μm /hour.

11.6 FIPOS

The process of Full Isolation by Porous Oxidized Silicon (FIPOS) was invented in 1981. It relies on the conversion of a layer of silicon into porous silicon and on the subsequent oxidation of this porous layer. The oxidation rate of porous silicon being orders of magnitude higher than that of monolithic silicon, a full porous silicon buried layer can be oxidized while barely growing a thin oxide on silicon islands on top of it. The original FIPOS process is described in Figure 11.15. It relies on the fact that p-type silicon can readily be converted into porous silicon by electrochemical dissolution of p-type silicon in HF (the sample is immersed into an HF solution and a potential drop is applied between the sample and a platinum electrode dipped into the electrolyte) . The conversion rate of n-type silicon is much lower. Porous silicon formation proceeds as follows: at first, an Si_3N_4 film is patterned over a p-type silicon wafer, and boron is implanted to control the density of the porous silicon surface layer. The N^- material is formed by conversion of the P^- silicon into N^- silicon by proton (H^+ ion) implantation. The p-type silicon is then converted into porous silicon by anodization in a hydrogen fluoride solution. Optimal conversion yields a 56% porosity (porosity is controlled by the HF concentration in the

electrolyte, the applied potential, and the current density at the sample surface during anodization). In this way, the volume of the buried oxide formed by oxidation of the porous layer is equal to that of the porous silicon, and stress in the films can be minimized.

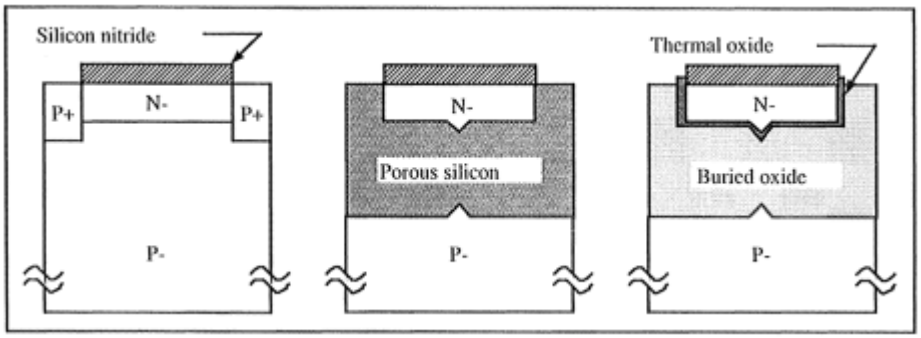


Figure 11.15*:* The original FlPOS process. From left to right: formation of N^- islands and P^+ current paths, formation of porous silicon, and oxidation of the porous silicon.

Since the surface area of silicon exposed to the ambient is extremely high, and porous silicon oxidizes very rapidly. This allows the grow of a thick buried oxide (oxidation depth is comparable to the width of the N^- silicon islands) while growing only a thin oxide at the edges of the N^- silicon islands in which the devices will be made. The thermally grown oxide provides the silicon islands with a high-quality bottom interface.

The original FIPOS technique produces high-quality SOI islands, and has been used to produce devices with good electrical characteristics but it has several limitations. Indeed, the formation of a thick oxidized porous silicon layer is needed to isolate even small islands. Such a thick oxide can induce wafer warpage, especially if the islands are unevenly distributed across the wafer. A second limitation is the formation of a little cusp of unanodized silicon at the bottom-center of the silicon islands, where the two (left and right) anodization fronts meet during porous silicon formation (Figure 11.15). This limits the use of such a technique for thin-film SOI applications, where thickness uniformity of the silicon islands is of crucial importance. Several variations of the process have been proposed, such as the formation of a buried P^+ layer (over a P^- substrate) located below N^- silicon islands, the whole structure being produced *e.g.* by epitaxy. This technique solves the problem of having to produce a very thick buried layer to isolate wide islands, since it permits an island width-to-porous silicon layer thickness ratio larger than 50. A different approach for the fabrication of FIPOS structures is based on the preferential anodization of the N^+ layer of an $N^-/N^+/N^-$ structure (Figure 11.16). With this method, the thickness of both the silicon islands and the porous silicon layer are uniform and easily controlled by the N^+ doping profile *(e.g.* obtained by antimony implant on a lightly-doped, n-type wafer, and subsequent epitaxy of an N^- superficial layer). Another advantage of the $N^-/N^+/N^-$ approach is the

automatic endpoint on the island isolation. As soon as all of the N^+ layer is converted into porous silicon the anodization current drops, and anodization stops due to a change in anodization potential threshold between the N^+ layer and the N^- silicon in the substrate and the islands. Such an automatic end of reaction control is not available in the $N^-/P^+/P^-$ approach where anodization of the P^- substrate occurs as soon as the P^+-layer has been converted into porous silicon . After oxidation of the porous silicon layer, a dense buried oxide layer is obtained.

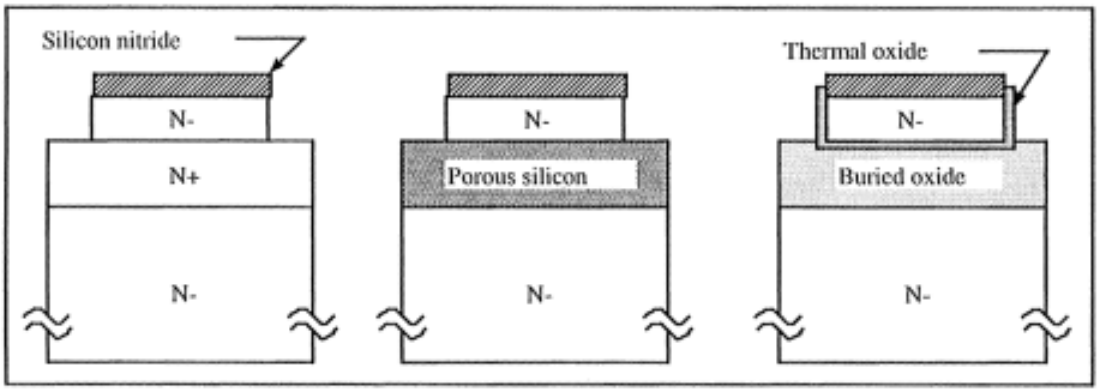


Figure 11.16. FIPOS formation ($N^-/N^+/N^-$ technique). From left to right: $N^-/N^+/N$- structure, formation of porous silicon, and oxidation of the porous silicon.

It is worth noticing that porous silicon is a single-crystal material, in spite of the fact that it contains many voids. A blanket porous silicon layer can, therefore, be created on a wafer, and epitaxial single-crystal silicon can be grown on it using MBE (molecular-beam epitaxy) or low-temperature PECVD (plasma-enhanced chemical vapor deposition). Other semiconductors such as GaAs can also be grown on porous silicon. The growth of epitaxial silicon on porous silicon is a key processing step of the Eltran® process which will be described below.

11.7 Separation by implanted oxygen (SIMOX)

SOI material can be obtained by implanting ions of oxygen or nitrogen into silicon and annealing the structure to form a buried insulator layer, commonly referred to as the BOX (buried oxide). Separation by implanted oxygen (SIMOX) is the most successful material based on this technique The SIMOX technique was invented by K. Izumi, of NTT in 1978. In this technique a high dose of oxygen ions is implanted in a silicon wafer followed by high-temperature annealing step to form a buried oxide layer (Figure 11.17). Ion implantation is traditionally used in the semiconductor industry to introduce dopant atoms, and doses higher than a few 10^{15} cm^{-2} are rarely employed. In the SIMOX technique implanted oxygen atoms are used to synthesize a new material, namely silicon dioxide. As a result a very high dose of oxygen

ions (typically 1.8×10^{18} cm^{-2} at 200 keV in the "standard" SIMOX process) must be implanted to form the buried oxide (BOX) layer.

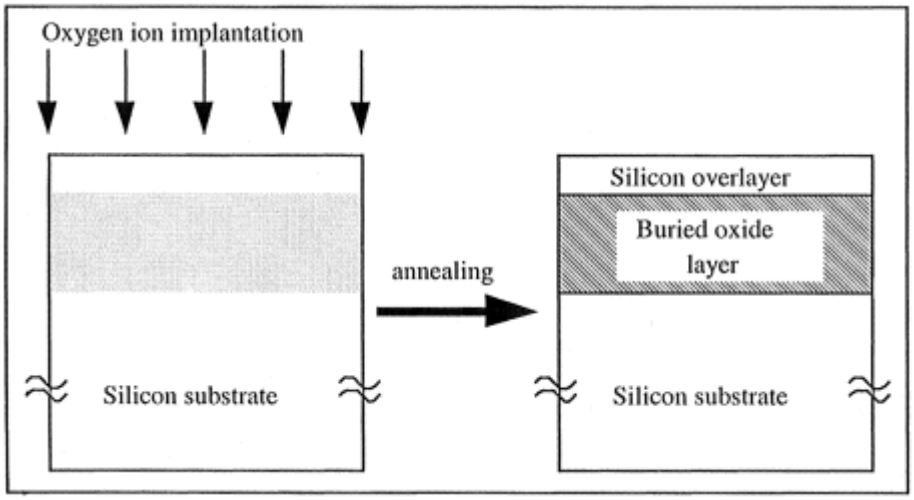


Figure 11.17: The principle of SIMOX: a high dose of oxygen is implanted into silicon, followed by an annealing step. The result is a buried layer of silicon dioxide below a thin, single-crystal silicon overlayer.

11.7.1 Standard SIMOX

Stoichiometric SiO_2 contains 4.4×10^{22} oxygen atoms/cm^3. Therefore, the implantation of 4.4×10^{17} atoms/cm^{-2} should be sufficient to produce a 100 nm-thick buried oxide layer. Unfortunately, due to the statistical nature of ion implantation, the oxygen profile in silicon does not have a box shape, but rather a skewed Gaussian profile. The implanted atoms spread over more than 100 nm, such that SiO_2 stoichiometry is not reached (Figure 11.18). If the wafer is annealed after implanting an oxygen dose that is too low, oxide precipitates form at a depth equal to the depth of maximum oxygen concentration, but no continuous layer of SiO_2 is produced. The most commonly used dose is 1.8×10^{18} cm^{-2}, which produces a 400 nm-thick buried oxide layer upon annealing . Figure 19 illustrates the evolution of the profile of oxygen atoms implanted into silicon with an energy of 200 keV. At low doses, a Gaussian oxygen profile is obtained. When the dose reaches 1.4×10^{18} cm^{-2} ("critical dose"), stoichiometric SiO_2 is formed (66 at.% of oxygen for 33 at.% of silicon), and further implantation does not increase the peak oxygen concentration, but rather broadens the overall profile *(i.e.* the buried oxide layer becomes thicker).

The temperature at which the implantation is performed is also an important parameter that influences the quality of the silicon overlayer. Indeed, the oxygen implantation step does amorphize the silicon which is located above the projected range. If the temperature of the silicon wafer during implantation is too low, the silicon overlayer becomes completely amorphized, and forms polycrystalline silicon upon subsequent annealing, an undesirable effect.

When the implantation is carried out at higher temperatures (above 500°C) the amorphization damage anneals out during the implantation process ("self annealing"), and the single-crystal nature of the top silicon layer is maintained. The silicon overlayer, however, is highly defective and the ion implantation step must be followed by a high temperature anneal step to improve the quality of both the BOX and the silicon layer. Since every single implanted oxygen atom must traverse the top silicon layer (the future silicon-on-insulator layer) a large number of defects is created. To reduce defects density the maintaining of the wafer at a temperature where most defects would self anneal during implantation, and performing a subsequent thermal treatment at high temperature (1350°C) in an appropriate ambient (argon + 2% oxygen) are used. This allows for the stabilization and densification of the BOX, as well as for the removal of oxide precipitates and other defects in the top silicon layer.

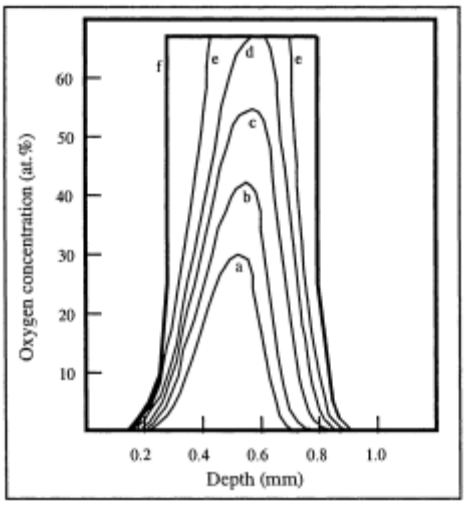


Figure 11.18: Evolution of the oxygen concentration profile with the implanted dose for an implantation energy of 200 keV: a) $4x10^{17}$, b) $6x10^{17}$, c) 10^{18} , d) $1.2x10^{18}$, e) $1.8x10^{18}$, and f) $1.4x10^{18}$ cm^{-2}.

During annealing, both dissolution of the oxide precipitates and precipitation of the dissolved oxide take place in the oxygen-rich silicon layers. In order to minimize the total surface energy of the $Si0_2$ precipitates, thus creating a more stable system, small precipitates dissolve into silicon, and large precipitates grow from the dissolved oxygen. At any given temperature (and for a given concentration of oxygen in the silicon), there exists a critical precipitate radius, below which a precipitate will disappear, and above which it will be stable. The use of a nitrogen ambient during annealing can induce the formation of silicon oxynitride around the oxide precipitates and inhibit their dissolution into the silicon matrix. Therefore, the use of an inert gas, such as argon, is preferred to nitrogen for the high-temperature annealing step used in the SIMOX formation process. Figure 11.19 shows TEM cross sections of SIMOX

samples. The small silicon inclusions at the bottom of the BOX in Figure 20 are characteristic of the standard SIMOX material.

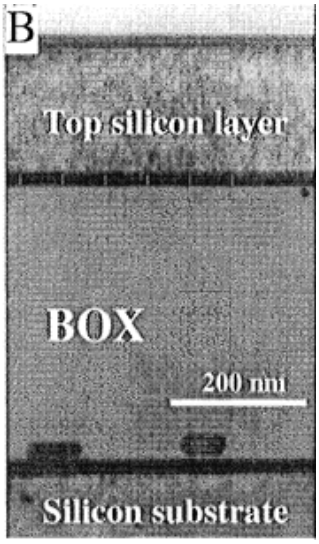


Figure 11.19: TEM cross section of SIMOX

11.7.2 Low-dose SIMOX

In 1990 Nakashima and Izumi proposed reducing the implanted oxygen dose to drastically reduce the dislocation density in the silicon overlayer film. They found that the dislocation density drops significantly as the dose is reduced below 1.4×10^{18} cm^{-2} with an implantation energy of 180 keV (Figure 11.20). Beside the reduction of defect density in the silicon layer there are other motivations for reducing the oxygen dose used to produce SIMOX material. At first, the total-dose radiation hardness of thin buried oxides is expected to be better than that of thicker ones. Secondly, direct fabrication of thin buried oxides is attractive for thin-film device applications. Finally, the production cost of a SIMOX wafer is proportional to the implanted dose. A potential additional benefit from this technique is the reduction of contamination of the wafers by impurities (carbon, heavy metals), which is proportional to the implanted oxygen dose. Low-dose SIMOX is obtained by implanting O^+ ions within a narrow dose window of approximately 4×10^{17} $atom.cm^{-2}$, called the "Izumi window". Under these conditions implantation followed by a 6 hour anneal at 1320°C forms a continuous BOX having a thickness of 80-nm.

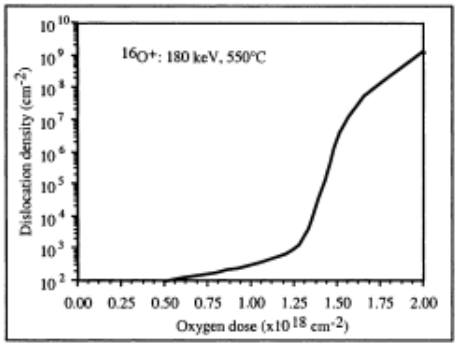


Figure 11.20 Evolution of dislocation density in the silicon overlayer with implanted oxygen dose

Figure 11.21 schematically represents the structure of the buried oxide versus dose around the process window for an implant energy of 120 keV. At a doses of $3x10^{17}$ cm^{-2} isolated oxide precipitates are formed. For a dose of $5x10^{17}$ cm^{-2} silicon precipitates form in the BOX. Only doses within this region produce a continuous, precipitate-free BOX.

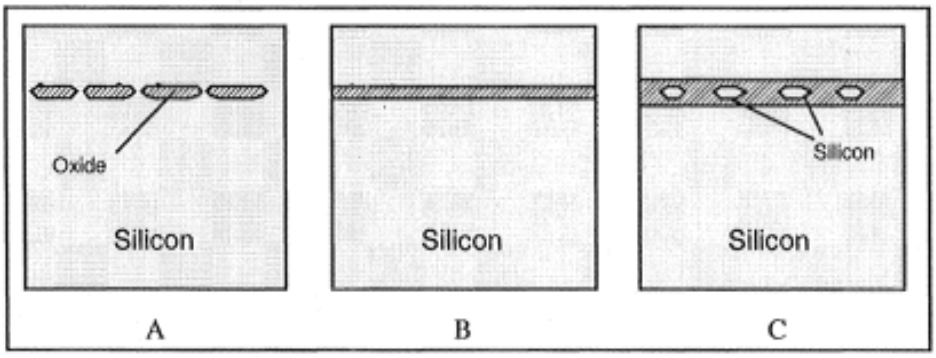


Figure 11.21 Evolution of the buried oxide structure for a dose of A: $3x10^{17}$ cm^{-2}; B: $4x10^{17}$ cm^{-2}; C: $5x10^{17}$ cm^{-2} at energy of 120 keV

11.7. 3 Internal thermal oxidation (ITOX)

It is possible to increase the thickness of the BOX produced by low-dose oxygen implantation. Indeed, high-temperature (1350°C) oxidation of a lowdose SIMOX wafer causes an increase of the buried oxide thickness (Figure 11.22). This phenomenon is called high-temperature internal oxidation (ITOX). As long as the thermal oxide grown on the silicon overlayer is thinner than 500 nm, there exists a linear relationship between the thickness of this oxide layer (t_{ox}) and the thickness increase of the buried oxide. The internal oxide grows at the expense of the bottom of the silicon overlayer. High-temperature internal oxidation has been shown to significantly improve the roughness of interface between the silicon overlayer and the BOX and to densify the oxide itself.

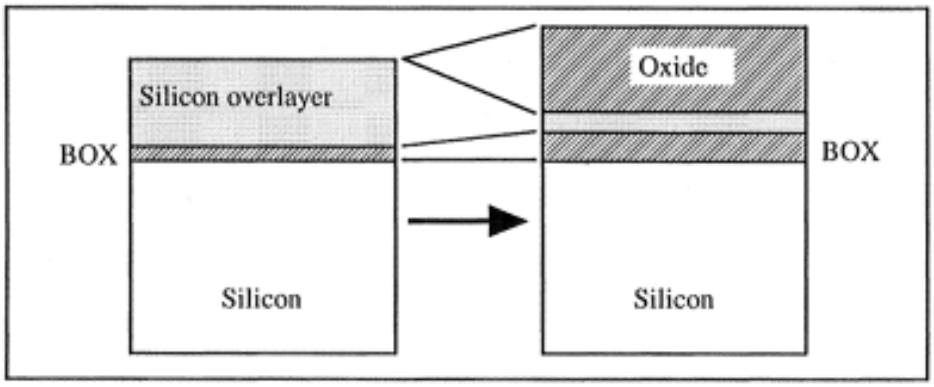


Figure 11.22: Principle of internal thermal oxidation (ITOX).

11.7. 4 Modified low-dose (MLD) SIMOX

In low-dose SIMOX the formation of a continuous buried oxide layer is possible if the peak oxygen concentration is located in highly defective silicon. This observation is the basis for the modified low-dose (MLD) SIMOX process. The MLD process overcomes the problem of oxide continuity encountered during low-dose SIMOX processing. To promote the formation of an ultrathin buried oxide during post-implantation annealing, the implantation process is modified to produce a microstructure that promotes coalescence of the oxygen into a continuous layer. This is accomplished by performing a two-step implant.

11.7. 5 Related techniques

It is possible to form a buried oxide layer without actually implanting oxygen. As we have seen earlier the simultaneous presence of defects and oxygen in silicon can result in the formation of a continuous buried oxide layer. In 2001, A. Ogura reported the formation of a continuous BOX obtained by implantation of light ions (H^+ or He^+) and subsequent annealing in an oxygen-containing ambient. The implantation of hydrogen or helium ions creates a defective layer near the projected range of the ions. Few defects, however, are created in the rest of the silicon, including in the future silicon overlayer, since H^+ and He^+ are light ions. Reported implant conditions are H^+, $5x10^{16}$ cm^{-2}, 45 keV, and He^+, $1\text{-}5x10^{17}$ cm^{-2}, 45 keV, all implanted at room temperature. Upon annealing in an argon/oxygen ambient at temperatures ranging from 1200 to 1350°C, oxygen diffuses through the top of the silicon layer and an internal oxidation process takes place which forms a buried oxide layer where the defects were presents (Figure 11.23). This technique makes it possible to produce a SIMOX-like SOI structure without the need for oxygen implantation and with less damage to the silicon overlayer.

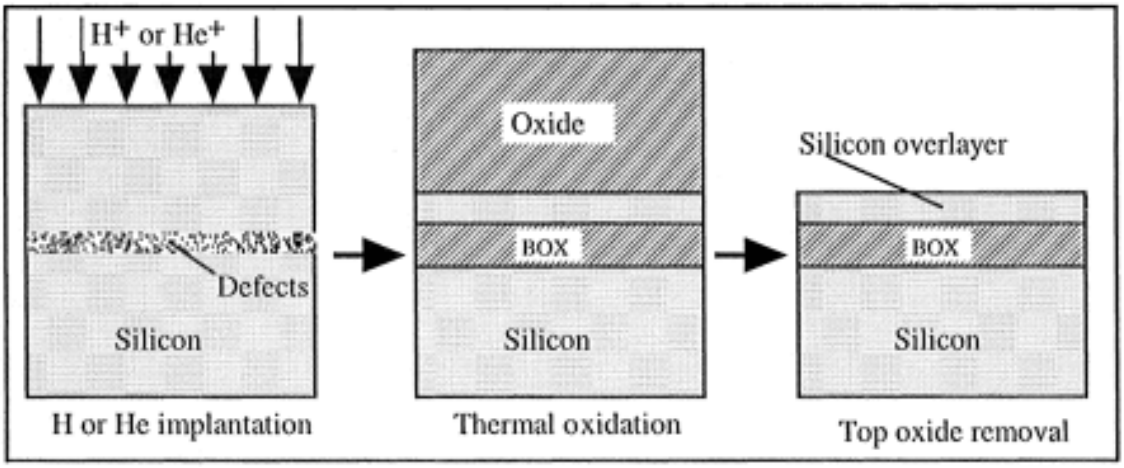


Figure 11.23 Buried oxide formation by light ion implantation and annealing in oxygen atmosphere

The table presents the summary of material quality of SIMOX SOI

Table 2-3. SIMOX material properties (in 2003) [1, 136]

Parameter	**Standard SIMOX**	**SIMOX MLD**
Wafer diameter	up to 200 mm	up to 300 mm
Silicon film thickness	210 nm	20 to 145 nm
Silicon film thickness uniformity	±2.5 %	±2 nm
Buried oxide (BOX) thickness	375 nm	135 or 145 nm
Buried oxide (BOX) thickness uniformity	±10 nm	±5 nm
Surface roughness (RMS)	0.7 nm	<0.15 nm
Dislocation density	< 1000 cm^{-2}	< 1000 cm^{-2}
HF defect density	< 0.5 cm^{-2}	< 0.1 cm^{-2}
BOX pipe (pinhole) density	< 0.1 cm^{-2}	< 0.1 cm^{-2}
Metallic contamination	< $5x10^{10}$ cm^{-2}	< $3x10^{10}$ cm^{-2}
BOX dielectric breakdown	>5 MV/cm	>7 MV/cm

11.7. 6 Separation by implanted nitrogen (SIMNI)

Just as buried oxide can be synthesized by oxygen ion implantation, a buried silicon nitride layer (Si_3N_4) can be obtained by implanting nitrogen into silicon. The critical nitrogen dose for forming a buried nitride layer is $1.lx10^{18}$ cm^{-2} at 200 keV , but lower doses can be employed if the implant energy is lower. The obtained buried nitride and silicon overlayer thickness are 190 nm and 215 nm, respectively, in the case of a $7.5x10^{17}$ N^+ ions cm^{-2} implantation at 160 keV followed by a 1200°C anneal. The most significant difference between oxygen and nitrogen inmplantation is that the peak of the nitrogen distribution does not saturate once stoichiometry is attained. This is due to the low diffusion coefficient of nitrogen in Si_3N_4

(10^{-28} cm^2 s^{-1} at a temperature of 500°C compared with 10^{-17} cm^2 s^{-1} for O_2 in SiO_2). The result of this low diffusion coefficient is that unreacted, "free" nitrogen is found in the buried layer if supercritical doses are implanted, and, as a result, nitrogen bubbles form. Nitride buried layers are polycrystalline, as opposed to buried oxide layers which are amorphous, and the grain boundaries between the Si_3N_4 crystallites can cause leakage currents between devices made in the silicon overlayer and the underlying silicon substrate. Furthermore, Si-Si_3N_4 interfaces are known to have a higher density of surface states than Si-SiO_2 interfaces. MOS device fabrication has, however, been demonstrated in SIMNI material.

11.7.7 Separation by implanted oxygen and nitrogen (SIMON)

Implantation of both oxygen and nitrogen ions into silicon is also used. These were attempting to combine the advantages of both SIMNI (formation of a buried layer by implantation of a relatively low dose and low defect generation) with those of SIMOX (formation of an amorphous rather than polycrystalline buried layer with good Si-dielectric interfaces). Buried oxynitride layers may also present better radiation hardness performances than pure SIMOX material. Buried oxynitride layers have been formed by implantation of different doses of both nitrogen and oxygen into silicon. Different implant schemes have been proposed (oxygen can be implanted before nitrogen, or vice-versa). The kinetics of synthesis of oxynitrides by ion implantation is more complicated than that of pure SIMOX or pure SIMNI materials. In some instances, gas (nitrogen) bubbles can be formed within the buried layer. It is, however, possible to synthesize buried oxynitride layers that are stable and remain amorphous after annealing at 1200°C. The resistivity of the buried oxynitride layers can reach up to 10^{15} Ω.cm, which is comparable to that of oxynitride layers formed by other means.

11.8 WAFER BONDING AND ETCH BACK (BESOI)

The expression "wafer bonding" refers to the phenomenon whereby mirror-polished, flat and clean wafers of almost any material, when brought into contact, are locally attracted to each other by Van der Waals forces and adhere or "bond" to each other. Bonds formed at room temperature are usually relatively weak. Therefore, for many applications the room temperature bonded wafers must undergo a heat treatment to strengthen the bonds across the interface. After wafer bonding one of the wafers is subsequently polished or etched down to a thickness suitable for SOI applications. The other wafer serves as a mechanical substrate, and is called handle wafer (Figure 11.24).

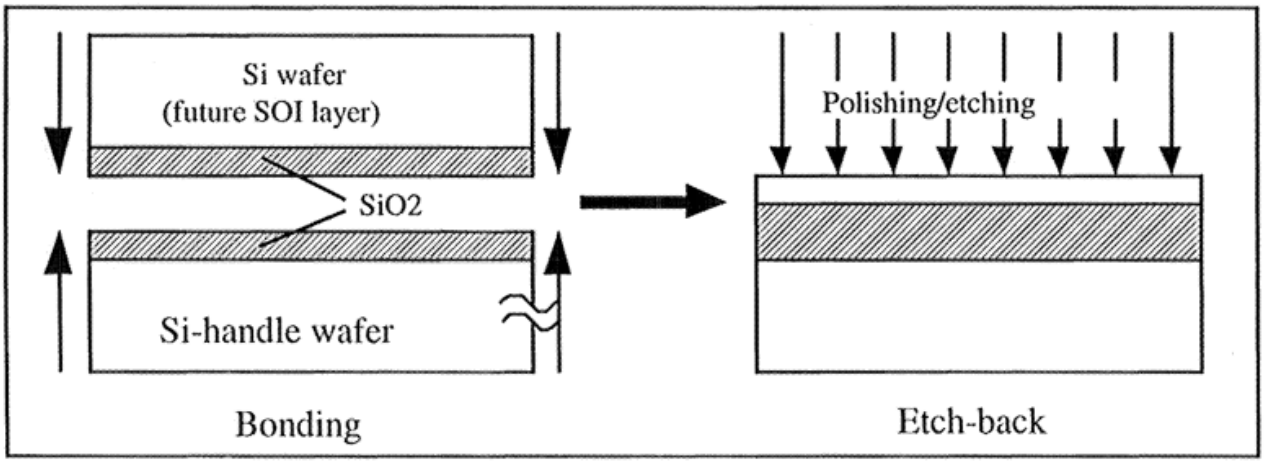


Figure 11.24 Bonding of two oxidized silicon wafers (left) , and polishing/etching back of one of the wafers.

11.8 .1 Hydrophilic wafer bonding

When two flat, hydrophilic surfaces such as oxidized silicon wafers are placed against one another, bonding naturally occurs, even at room temperature. The contacting force is caused by the attraction of hydroxyl groups OH^- adsorbed on the two surfaces. This attraction propagates from a first site of contact across the whole wafer in the form of a "bonding wave" with a speed of several cm/s. The presence of particles between the wafers creates unbonded areas called "voids". Voids with a diameter of several millimeters are readily created by particles 1 μm or less in size. Because of the usually organic nature of the particulates the extrinsic voids cannot be eliminated by high-temperature annealing. The only way of obtaining void-free bonding is, therefore, to clean the wafers with ultra-pure chemicals and water and to carry out all bonding operations in an ultra-clean environment.

The presence of voids is easily revealed by infrared, ultrasonic imaging, or X-ray tomography. The presence of voids is, of course, undesirable, since they can lead to delamination of devices located in imperfectly bonded areas during device processing. Wafer bonding must be carried out in a very clean environment to avoid the presence of particulates between the bonded wafers. A micro-clean room bonding apparatus has been developed,which involves bonding the wafers in a particle-free enclosure. Deionized filtered water is first flushed between the two closely-spaced wafers. The wafers are then dried by spinning under an infrared lamp and finally brought in contact.

Directly after room temperature bonding the adhesion between the two wafers is determined by Van der Waals interactions or hydrogen bridge bonds and one or two orders of magnitude lower than typical for covalent bonding. The typical bonding energy due to the Van der Waals is on the order of 50-100 mJ/m^2. For most practical applications a higher bond energy

is required which may be accomplished by an appropriate heating step which frequently, for commercial SOI production, is performed at temperatures as high as 1100^{o}C. When silica surfaces are hydrated, water molecules cluster on the oxidized wafer surface. When two such surfaces are brought into contact, hydrogen bonding occurs via the adsorbed water, as shown in Figure 26,A. At temperatures above 200°C the adsorbed water separates from the SiOH group and forms a tetramer water cluster (Figure 11.25,B). For temperatures greater than 700°C the water clusters decompose and diffuse away leaving Si-O-Si bonds (Figure 11.256,C).

Bond strength for room temperature contact bonded wafers varies between 60-85 mJ/m^2, which is consistent with the surface energy of silica bonded through hydrogen bonding. In addition, the surface energy increases at the transition temperature of 300°C. This is the temperature where hydrogen bonds begin to convert to Si-O-Si bonds. Surface energy is constant for bonded wafers annealed in the region of 600°C to 1100°C for anneal times between 10 seconds to 6 hours and reaches values in excess of 1000 mJ/m^2. The bonding process does not involve mass transport in that temperature range. Rather the bond strength is limited in that regime by the amount of contacted area of the bonded wafers, which is a function of how well the wafers can elastically deform at a specific temperature. The kinetics of the deformation is so fast that the bond strength appears to be a function of temperature only. For temperatures greater than 1100°C, the bond energy does increase with time of anneal, but this is due to the viscous flow of the oxides at these high temperatures. It is worth noting that high bonding energies (> 1000 mJ/m^2) can be obtained by activating the surface of the wafers using a plasma *(e.g :.* oxygen plasma) prior to bonding and then annealing the wafer pair at low temperature *(e.g.:* 400°C) (Fig.11.26). To be complete, one should mention that the bonding of silicon wafers to sodium-containing glass substrates can be facilitated by the application of an electric field during the annealing step . This process is called "anodic bonding"

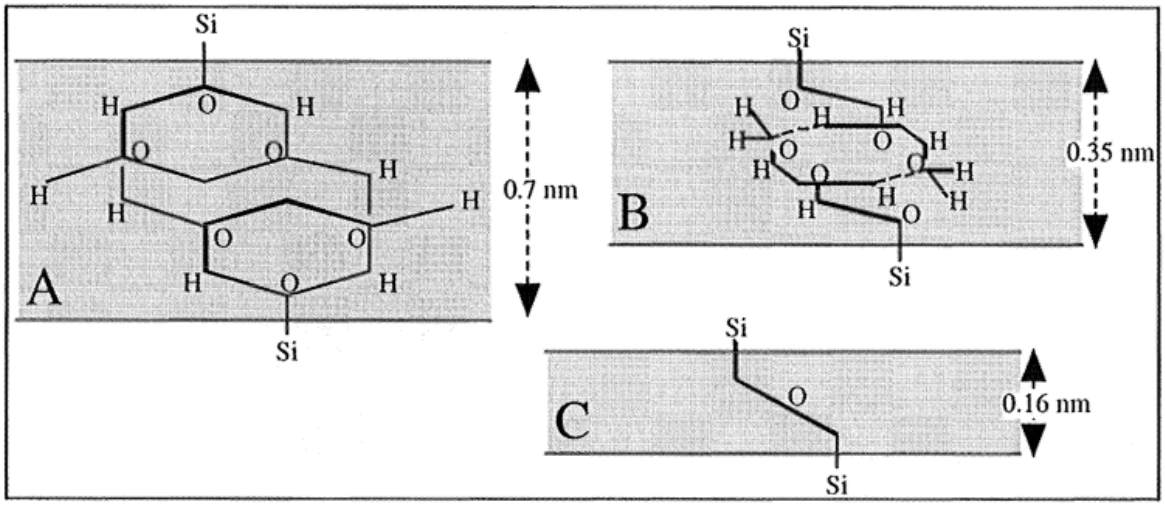


Figure 11.25: Model for silicon wafer bonding at different temperatures. A: Room temperature, $SiOH:(OH_2)_2:(OH_2)_2:HOSi$; B: T = 200°C, $SiOH:HOSi + (H_2O)_4$, C: T> 700°C, $SiOSi + H_2O$

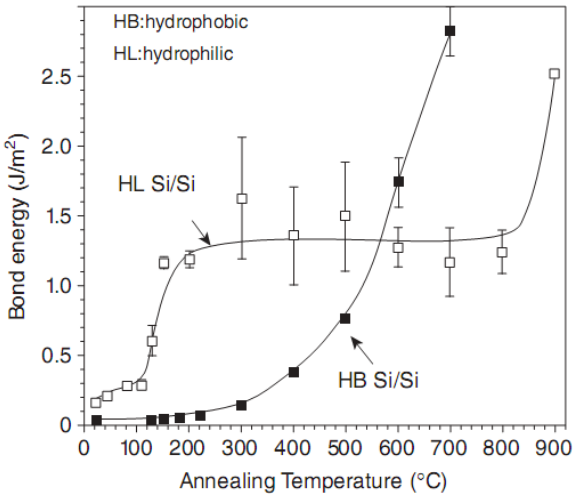


Fig.11.26 Dependences of bond energy on annealing temperature for hydrophobic and hydrophilic Si surfaces

11.8.2 Etch back

After bonding of the wafers has been carried out, the top wafer has to be thinned down from a thickness of 600 μm to a few micrometers or less in order to be useful for SOI device applications. Two basic thinning approaches can be used: grinding followed by chemicomechanical polishing, and grinding followed by selective etch-back. The grinding operation is a rather crude but rapid step that is used to remove all but the last several micrometers of the (top) bonded wafer. The thinning method using chemico-mechanical polishing is cheap, but its use is, so far, limited to the fabrication of rather thick SOI films because of the absence of an etch stop. Much more accurate are the techniques using, after initial grinding, a chemical etch-back procedure with etch stop(s). The etch stop is usually obtained by creating doping concentration gradients at the surface *(i.e.* right next to the oxide layer used for bonding) of the top wafer. For instance, in the double etch-stop technique, a lightly doped wafer is used, and a P^{++} layer is created at its surface by ion implantation. Then, a lowdoped epitaxial layer is grown onto it. This epitaxial layer will be the SOI layer at the end of the process. After bonding and grinding, two chemical etch steps are used. First, a potassium hydroxide solution is used to selectively etch the substrate and to stop on the P^{++} layer. Then a 1:3:8 $HF:HNO_3:CH_3COOH$ etch is used to remove the P^{++} layer. The combined selectivity of the two etching solutions is better than 10,000:1. The final thickness uniformity of the SOI layer depends on the uniformity of the silicon thickness grown epitaxially, as well as on the uniformity of the P^{++} layer formation, but thickness standard deviations better than 12 nm can be obtained. Other etch stop techniques can be used as well (SiGe layer, carbon or nitrogen-doped silicon layer, etc.). Precision polishing of the top silicon film can also be achieved using a computer-controlled

scanning plasma electrode. This technique, called plasma-assisted chemical etching (PACE) is used to planarize the silicon film and can produce 100 nm-thick SOI material with an thickness variation better than 2 nm. A more recent polishing technique, called the magnetorheological finishing (MRF) technique can reach a silicon film thickness uniformity of 0.8 nm. It makes use of a polishing fluidic slurry whose viscosity can be locally controlled by the application of a magnetic field. The rate of material removal during MRF polishing is controlled by a computer to achieve high-accuracy results.

11.9 LAYER TRANSFER TECHNIQUES.

11.9.1 Smart-Cut®

In the layer transfer techniques a thin superficial layer is peeled off from a silicon wafer and transferred to an oxidized silicon handle wafer. Splitting the superficial layer from a wafer is achieved either by gas implantation and anneal and/or by applying pressure along a weakened crystal plane underneath the surface of a wafer.

The Smart-Cut® process combines ion implantation technology and wafer bonding to transfer a thin surface layer from a wafer onto another wafer or an insulating substrate. It consists of the three steps: 1) Implantation of gas ions (usually hydrogen), 2) bonding to a stiffener, and 3) thermal annealing. SOI wafers fabricated using the Smart- Cut® process are called UNIBOND® wafers (Figure 11.27).

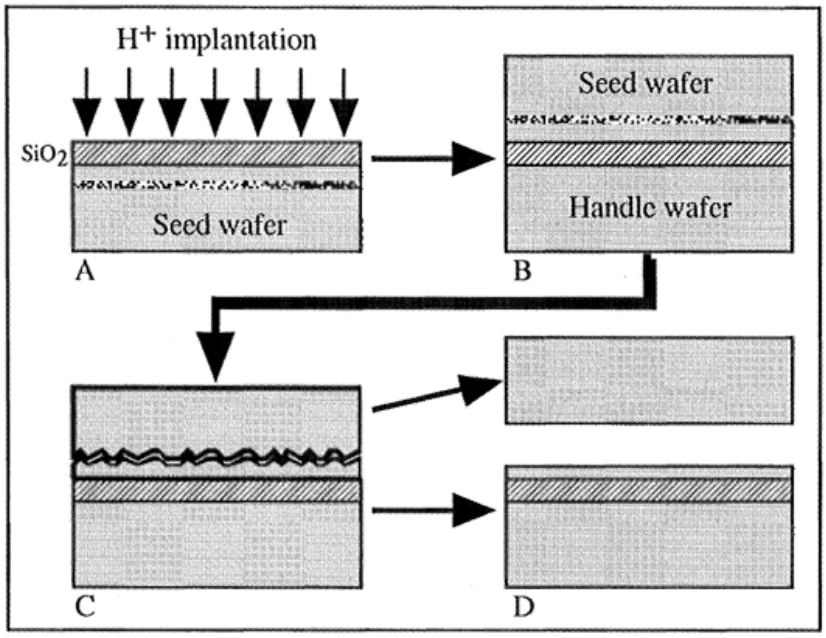


Figure 11.27: The Smart-Cuts process. A: Hydrogen implantation; B: Wafer bonding; C: Splitting of wafer A; D: Polishing of both wafers. Wafer A is recycled as a future handle wafer.

1) Ion implantation of hydrogen ions into an oxidized silicon wafer (called the "seed wafer"). The implanted dose is on the order of $5x10^{16}$ cm^{-2}. At this stage hydrogen-decorated defects (microbubbles) are formed at a depth equal to the implantation range (Rp). The wafer is preferably capped with thermally grown SiO_2 prior to implantation to protect the silicon top surface during implantation. This oxide layer will be used for hydrophilic bonding to a separate oxidized wafer.
2) Hydrophilic bonding of the seed wafer to another oxidized silicon wafer, called the to "handle wafer", is performed.
3) A two-phase heat treatment of the bonded wafers is then carried out.

During the first phase, which takes places at a temperature around 500°C, a crystalline rearrangement and coalescence of the hydrogen-decorated defects into larger structures occurs in the hydrogen-implanted region of the seed wafer. Hydrogen pressure builds up in the growing cavities and eventually the seed wafer splits in two parts: a thin layer of monocrystalline silicon which remains bonded onto the handle wafer, and the remainder of the seed wafer which can be recycled for later use. The basic mechanism of the wafer splitting upon hydrogen implantation and thermal treatment is similar to surface flaking and blistering of materials exposed to helium or proton bombardment. During annealing, the average size of the microcavities increases. This size increase takes place along a (100) direction *(i.e.* parallel to the wafer surface) and an interaction between cavities is observed, which eventually results into the propagation of a crack across the whole wafer. This crack is quite parallel to the bonding wafer. The second heat treatment takes place at a higher temperature (1100°C) and is aimed at strengthening the bond between the handle wafer and the SOI film. Finally, chemo-mechanical polishing is performed on the SOI film to give it the desired mirror-like surface. Indeed, this layer exhibits some microroughness after the splitting of wafer wafer A, and a final touch-polish step is necessary. This polishing step reduces the surface roughness to less than 0.15 nm and consumes a few hundred angstroms of the SOI film. A great advantage is that the seed wafer can be recycled, resulting in just *N+1* silicon wafers processed to produce *N* SOl substrates.

Implantation of gas ions into materials is long known to lead to the formation of blisters at the material surface. It is also possible to form blisters by implanting a moderate dose $<10^{17}cm^{-2}$ of gas ions and then performing a thermal annealing step to promote the coalescence of small gas-containing defects into larger ones. For the fabrication of SOl material it is suitable to use hydrogen rather than helium or any other rare gas because less energy is deposited in the material above the projected range of the ions. As a result fewer defects are created in the silicon surface layer (the future SOI film) and most defects are situated at a depth near the projected

range. In the particular case of hydrogen (proton) implantation into silicon the defects created by implantation, or hydrogen-related cavities (HCRs) consist of a mix of hydrogen-decorated vacancies, vacancy clusters, and platelets. Platelets are flat, disk-shaped microcavities containing hydrogen. Their thickness is approximately 2 lattice parameters (l nm), their diameter is approximately 10 nm, and they are mainly oriented along (100) planes along the (100) surface of the wafer (Figure 11.28). Typical implantation dose ranges between $lx10^{16}$ and $7x10^{16}$ H^+/cm^2. Upon annealing, typically above 500°C, the hydrogen atoms in the HRCs sever their bonds with silicon atoms and diffuse in the silicon. The hydrogen atoms aggregate in the larger defects, forcing them to grow in size, through a mechanism called "Ostwald ripening"; the hydrogen atoms lost by the small cavities are captured by larger ones, such that larger cavities grow at the expense of the smaller ones.

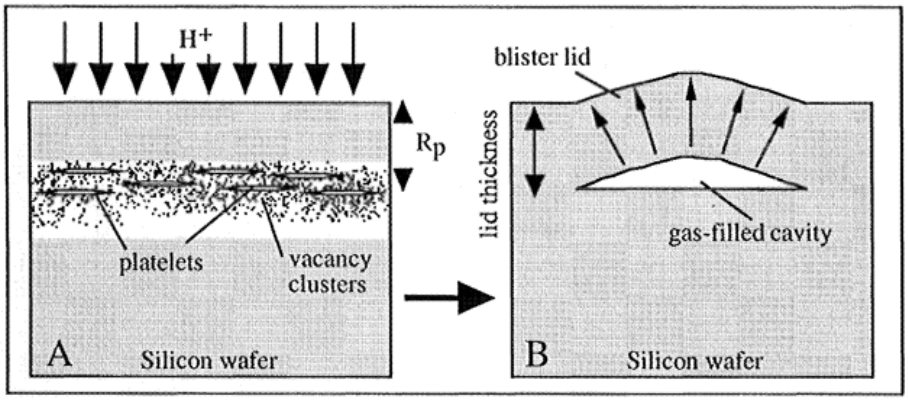


Figure 11.28 A: Hydrogen implant ation ad formation of defects (vacancies, vacancy clusters and platelets; B: Formation of a blister upon annealing. Rp is the projected range of the implanted ions.

The Smart-Cut® process is based on harnessing the destructive forces produced during blister formation to produce a thin silicon film. This is accomplished by attaching a stiffener to the implanted silicon wafer. Usually the stiffener is an oxidized silicon wafer called the "handle wafer" (Figure 11.29), but glass, quartz and other materials can also be used. The implanted wafer is usually called the "seed wafer" because it is the wafer from which the thin SOI film will come from.

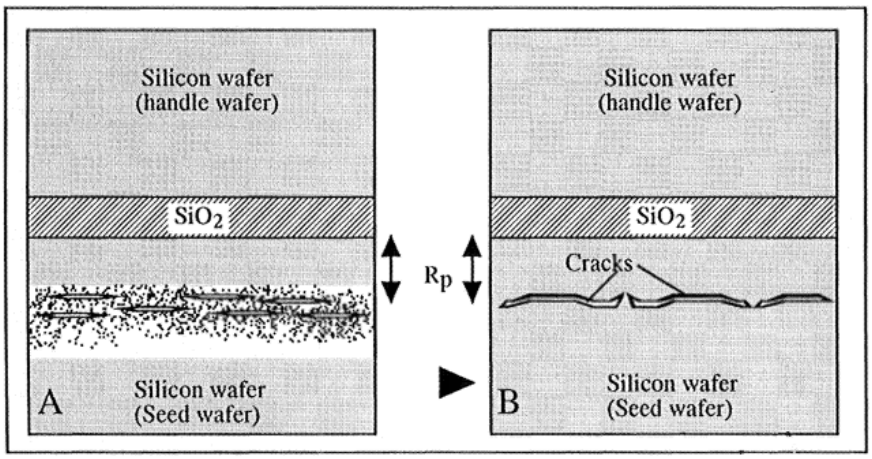
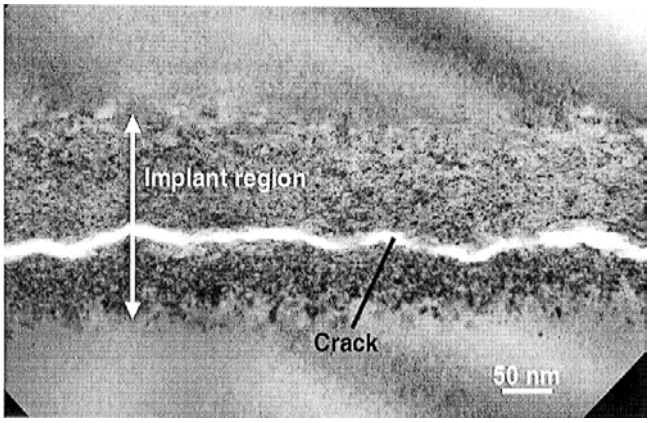


Figure 11.29 Formation of cracks ncar the projected range of a silicon wafer implanted with hydrogen. A: Bonding of the handle wafer (stiffener) to the seed wafer; B: Formation of a crack network ncar the projected range upon annealing .



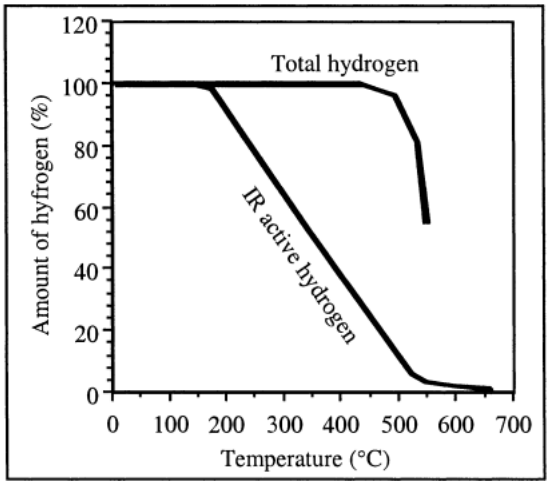


Figure 11.30. TEM micrograph of a crack

Figure 11.31 Evolution of hydrogen in Si-H bonds and total amount of hydrogen vs annealing temperature . Annealing time is 30 min.

An annealing step is then performed to both strengthen the bond between the two wafers and to make it possible for the smaller HCR to coalesce into larger, hydrogen-filled structures called "cracks" or "microcracks". These cracks correspond to the gas-filled cavities forming blisters in the absence of a stiffener. The role of the stiffener is to prevent mechanical deformation of the thin silicon layer between the defect region and the SrO, layer. In addition, the presence of the stiffener provides a restoring force that opposes the vertical lift that leads to blistering and the vertical force is transformed into lateral crack propagation. As a result a network of horizontal cracks filled with pressurized hydrogen is created, which creates a "perforated line"-like separation (Figures 11.30,11.31). The annealing step is responsible for removing hydrogen atoms from the vacancy complexes. The hydrogen diffuses to the growing microcracks.

Splitting of the seed wafer takes place naturally at the end of the annealing process, when the silicon is sufficiently weakened near the projected range depth by the network of microcracks and the buildup of pressure in them. The duration of the splitting itself is extremely short (probably less than one millisecond) and generates a small audible noise. It is, however, possible to induce splitting of the seed wafer by other means, once the silicon has been sufficiently weakened by hydrogen implantation and some annealing. For instance , splitting can be obtained by dipping the wafer pair in liquid nitrogen or applying a mechanical force to the edge of the seed wafer.

The table presentss the **Soitec Roadmap** for SOI wafers preparation. Fig.11.32 shows AFM picture of 200 mm XUT product thickness map 580 A of Si on 1500 A BOX. It can be seen nice homogeneity of SOI thickness along 200 mm wafer.

Time line	2000	2001	2002	2003	2004
Process generation	PD	FD	UT1	UT2	XUT
Silicon layer thickness (Å)	1000	500–700	200–700	200–300	200
Si Unif 6σ (Å)	150	100	70	30	10
Si Unif 6σ; all wrs all sites (%)	± 7%	± 7%	± 6%	± 5%	± 2.5%
AFM Roughness (RMS) 1×1 µm	1.0 Å	1.0 Å	1.5 Å	2 Å	2 Å
AFM Roughness (RMS) 10×10 µm	1.5 Å	1.5 Å	3 Å	4 Å	5 Å
Box Layer Thickness (nm)	150–200	100–150	100–150	80–150	80–150
Box Layer TTV	± 4%	± 4%	±3%	± 3%	± 3%

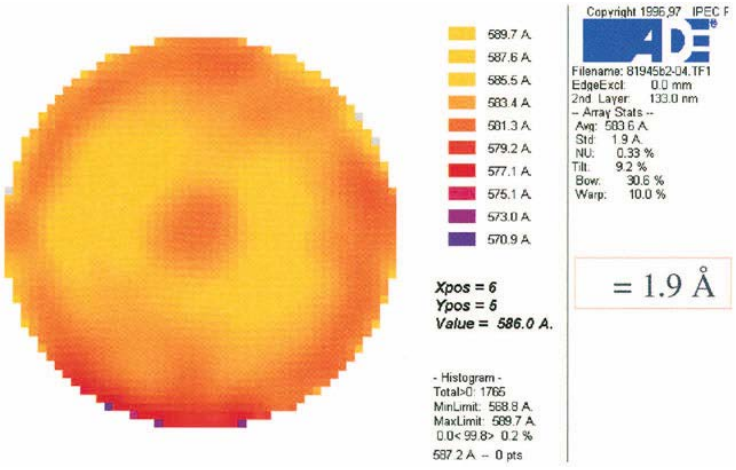


Fig. 11.32. 200 mm XUT product thickness map. 580 A of Si on 1500 A BOX

11.9.2 Eltran®

Epitaxial Layer TRANsfer ((Eltran®), registered trademark, 1994, Canon (Japan) technique combines the formation of a porous silicon layer, epitaxy, and wafer bonding to produce SOI wafers. The original Eltran®process was published in 1994. It comprises the following steps (Figure 11.33) : the formation of a blanket porous silicon layer on a silicon wafer (seed wafer) followed an hydrogen bake step to seal the
surface pores , and the growth of a single-crystal epitaxial silicon film. After thermal growth of an oxide the seed wafer is bonded to an oxidized wafer called the "handle wafer" .

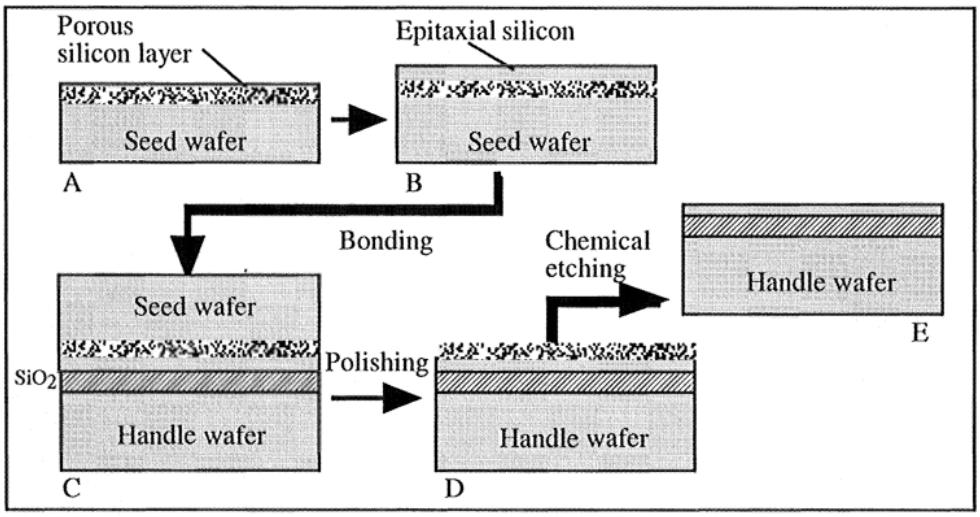


Figure 11.33: The original Eltran®process. A: Formation of a porous silicon layer; B: Growth of epitaxial silicon ; C: Bonding to a handle wafer; D: Polishing of the silicon wafer; E: Porous silicon etching .

The bulk of the handle wafer is then removed by grinding and polishing until the porous silicon layer is reached. The porous silicon is removed by etching in an $HF:H_2O_2$ solution and a final H_2 annealing is applied to smooth the surface. One drawback of the original Eltran®process is that it requires two silicon wafers to produce a single SOI wafer.

Similar to the Smart-Cut process, the second-generation Eltran® process uses $N+1$ silicon wafers to produce N SOI wafers. The process is based on the formation of a double porous silicon layer, or more exactly a layer with two different porosities. The application of a high-pressure (20-60 MPa) water jet "unzips" the seed wafer from the handle wafer along the stress region between the two porous silicon layers, thereby producing an SOI substrate and a recyclable seed wafer (Figure 11.34).

In the Eltran® process a low current density is first used to form a low-porosity layer at the surface of the seed wafer. A low-porosity surface layer is suitable for the subsequent growth

of a high-quality epitaxial layer. The current density is then increased to produce a higher-porosity layer bebeath the low-porosity layer. The difference in porosity generates mechanical stress between the two layers. A hydrogen bake step is applied to seal the pores at the surface, and a silicon layer (the future SOI layer) is grown by epitaxy. Hydrophilic bonding is used to attach the seed wafer to the handle wafer. Then a jet of pressurized water with a diameter of 0.1 mm is directed at the edge of the wafer assembly. The water jet acts as a liquid wedge that splits the porous silicon at the region of maximum stress, *i.e.* where the two porosities meet. Because a liquid rather than a solid wedge is used, the splitting easily propagates across the entire wafer assembly, and the porous silicon layer opens like a zipper. Once splitting has been achieved the porous silicon layers can be removed from both the seed and the SOI wafers by etching in an $HF:H_2O_2$ solution. A final H_2 annealing is applied to smooth the surface.

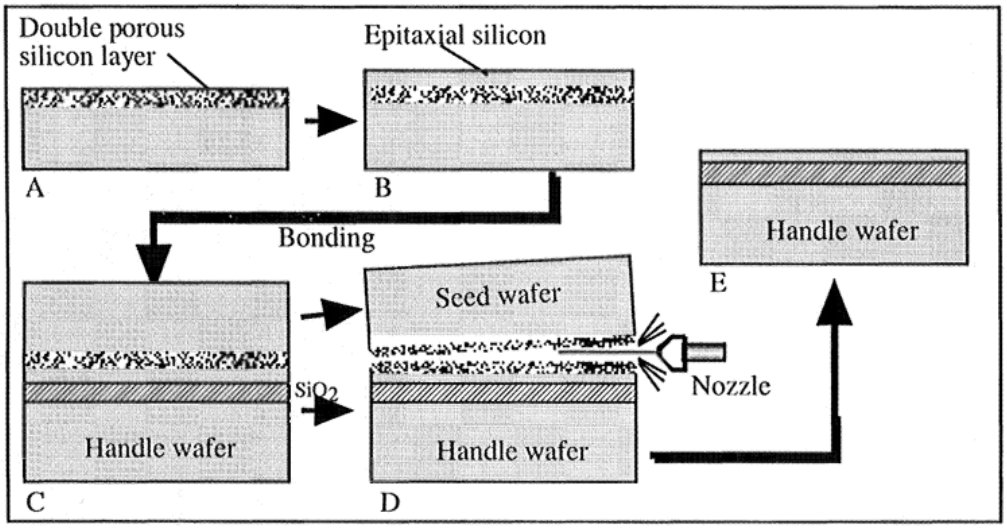


Figure 11.34: Second-generation Eltran®process. A: Formation of a porous silicon layer; B: Growth of epitaxial silicon; C: Bonding to a handle wafer; D: Porous silicon layer splitting using a water jet; E: Etching and H_2 annealing.

Fig.11.35 presents the photo of 300 mm-diameter ELTRAN® Wafer, fig.11.36 – the history of ELTRAN® development.

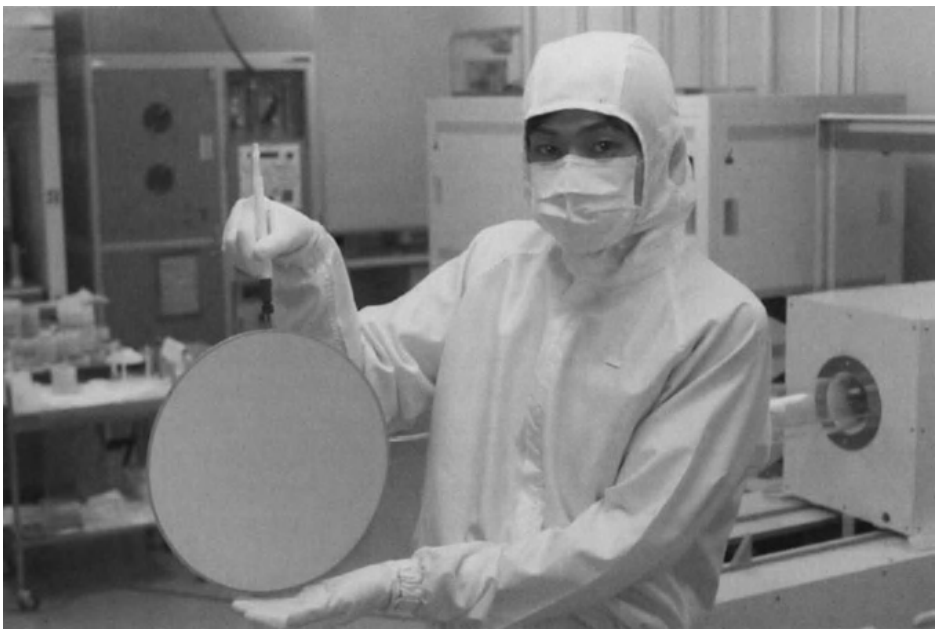


Fig. 11.35 300 mm-diameter ELTRAN® Wafer

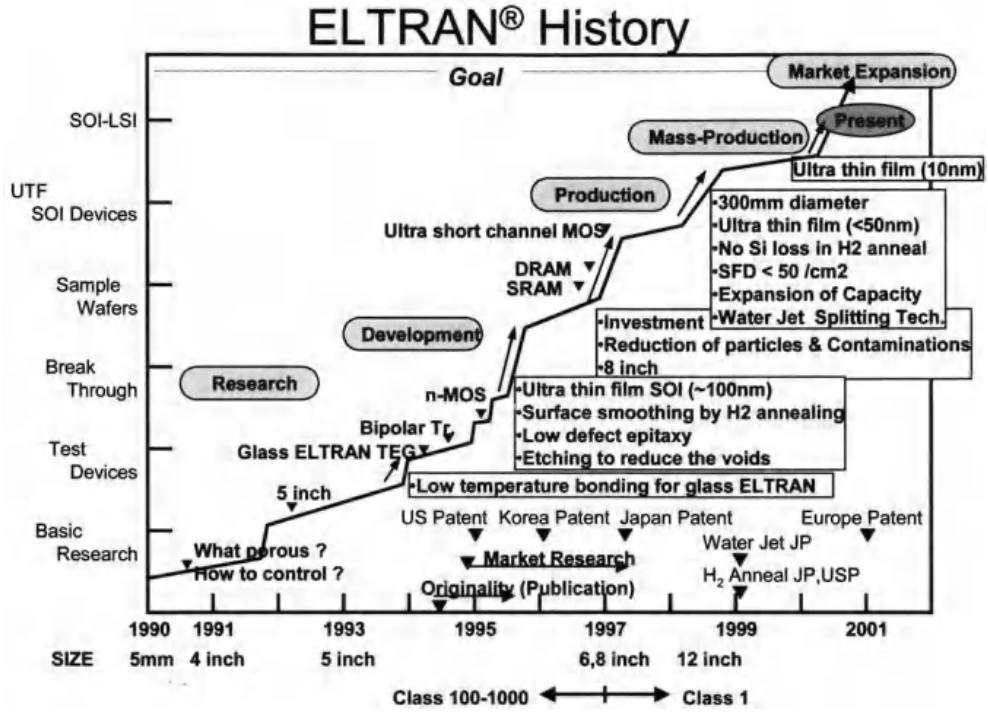


Fig.11.36 The history of ELTRAN®

11.10 STRAINED SILICON ON INSULATOR (SSOI)

Silicon films with in-plane tensile stress (strain) have a higher electron mobility than relaxed (unstressed) films. Conversely, silicon and SiGe films with in-plane compressive strain have a higher hole mobility. Several techniques have been developed to produced strained silicon-on-insulator (SSOI) material. A first technique employs oxygen implantation and resembles the SIMOX process (Figure 11.37). At first a graded $Si_{1-x}Ge_x$ layer is grown on a silicon wafer. The germanium content, *x*, is increased from 0 to 0.1 during growth. Then a

relaxed layer of $Si_{0,9}Ge_{0,1}$ is grown. Next, oxygen ions (180 keV, $4x10^{17}$ cm^{-2}) are implanted into the relaxed SiGe layer and high-temperature annealing (1350°C for 6 h) is carried out to grow a buried Si02 layer inside the top SiGe layer. A strained layer is silicon can directly be grown on the relaxed SiGe layer, or a combination of strained SiGe ($Si_{0,82}Ge_{0,12}$) and strained Si films can be grown. The strain in the top SiGe layer is compressive, which increases hole mobility, while the strain in the silicon layer is tensile, which provides enhances electron mobility. SOI MOSFETs have been fabricated using this process, and electron and hole mobility 60% and 30 % higher than in regular SOI devices has been measured, respectively.

Strained silicon-on-insulator can also be produced using the Smart-Cut.

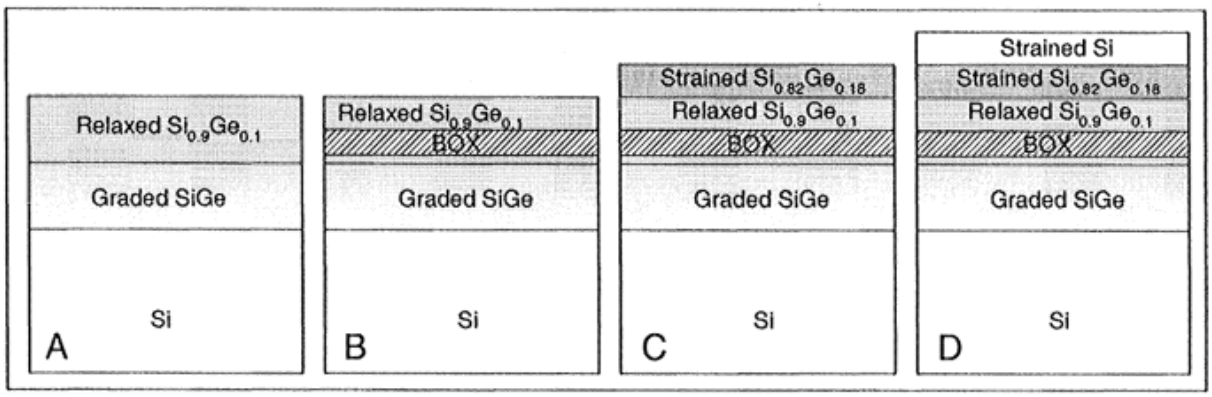


Figure 11.37: Formation of SSOI using a SIMOX-like process. A: Growth of a relaxed SiGe layer; B: Oxygen implantation and annealing to form a buried oxide (BOX) layer; C: Growth of a SiGe layer richer in Ge than the first layer; D: Growth of the strained silicon layer process.

Chapter 12. Integrated Devices

(A. Evtukh)

12.1. Introduction

Microwave, photonic, and power applications generally employ discrete devices. For example, an IMPATT diode is used as a microwave generator, an injection laser as an optical source, and a thyristor as a high-power switch. However, most electronic systems are built on the integrated circuit (IC), which is an ensemble of both active (e.g., transistor) and passive devices (e.g., resistor, capacitor, and inductor) formed on and within a single-crystal semiconductor substrate and interconnected by a metallization pattern [1-4]. ICs have enormous advantages over discrete devices connected by wire bondings. The advantages includes (*a*) reduction of the interconnection parasitics, because an IC with multilevel metallization can substantially reduce the overall wiring length, (*b*) full utilization of semiconductor wafer's "real estate," because devices can be closely packed within an IC chip, and (*c*) drastic reduction in processing cost, because wire bonding is a time-consuming and error-prone operation.

In this chapter the basic processes described in previous chapters to fabricate active and passive components in an IC are we combined. Because the key element of an IC is the transistor, specific processing sequences are developed to optimize its performance. Three major IC technologies associated with the three transistor families: the bipolar transistor, the MOSFET, and the MESFET are considered.

Specifically, the following topics are covered:

- The design and fabrication of IC resistor, capacitor, and inductor.
- The processing sequence for standard bipolar transistor and advanced bipolar devices.
- The processing sequence for MOSFET with special emphasis on CMOS and memory devices.
- The processing sequence for high-performance MESFET and monolithic microwave IC.
- The major challenges for future microelectronics, including ultrashallow junction, ultrathin oxide, new interconnection materials, low power dissipation, and isolation.

Figure 12.1 illustrates the interrelationship between the major process steps used for IC fabrication. Polished wafers with a specific resistivity and orientation are used as the starting material. The film formation steps include thermally grown oxide films, deposited polysilicon, dielectric, and metal films. Film formation is often followed by lithography or impurity doping. Lithography is generally followed by etching, which in turn is often followed by another impurity doping or film formation. The final IC is made by sequentially transferring the patterns

from each mask, level by level, onto the surface of the semiconductor wafer. After processing, each wafer contains hundreds of identical rectangular chips (or dice), typically between 1 and 20 mm on each side, as shown in Fig. 12.2a. The chips are separated by sawing or laser cutting; Figure 12.2b shows a separated chip. Schematic top views of a single MOSFET and a single bipolar transistor are shown in Fig. 12.2c to give some perspective of the relative size of a component in an IC chip. Prior to chip separation, each chip is electrically tested. Defective chips are usually marked with a dab of black ink. Good chips are selected and packaged to provide an appropriate thermal, electrical, and interconnection environment for electronic application [5].

IC chips may contain from a few components (transistors, diodes, resistors, capacitors, etc.) to as many as a billion or more. Since the invention of the monolithic IC in 1959, the number of components on a state-of-the-art IC chip has grown exponentially.

We usually refer to the complexity of an IC as small-scale integration (SSI) for up to 100 components per chip, medium-scale integration (MSI) for up to 1000 components per chip, large-scale integration (LSI) for up to 100,000 components per chip, very-large scale integrated (VLSI) for up to $l0^7$ components per chip, and ultra large-scale integration (ULSI) for larger numbers of components per chip. Examples of two ULSI chips: a 32-bit microprocessor chip, which contains over 42 million components, and a 1 Gbit dynamic random access memory (DRAM) chip, which contains over 2 billion components.

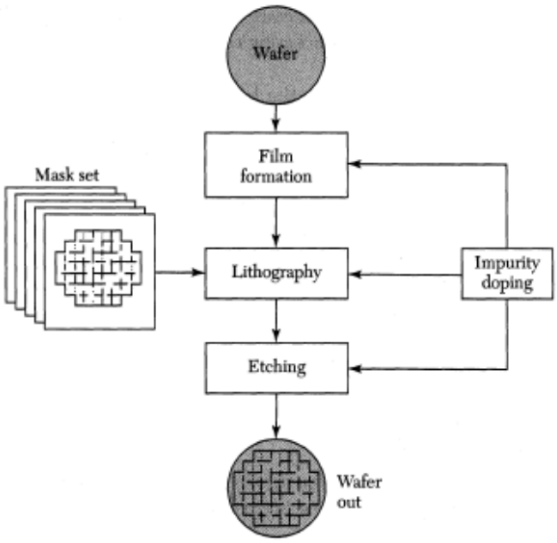


Fig. 12.1. Schematic flow diagram of integrated-circuit fabrication.

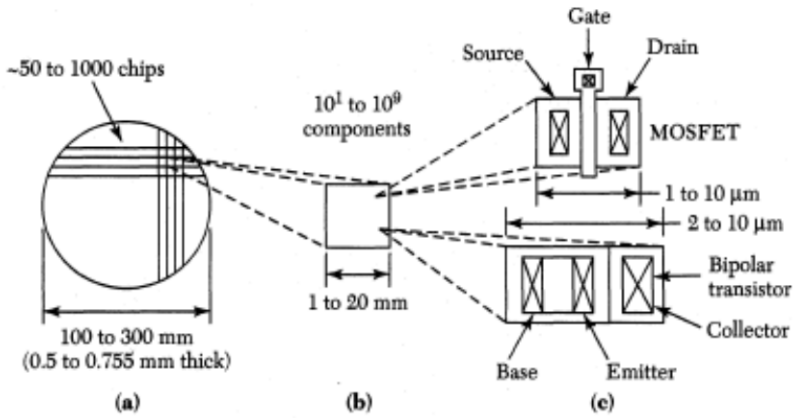


Fig. 12.2. Size comparison of a wafer to individual components. (a) Semiconductor wafer. (b) Chip, (c) MOSFET and bipolar transistor.

12.2. Passive Components

12.2.1. The Integrated-Circuit Resistor

To form an IC resistor, it is possible to deposit a resistive layer on a silicon substrate, then pattern the layer by lithography and etching. It is also to define a window in a silicon dioxide layer grown thermally on a silicon substrate and then implant (or diffuse) impurities of the opposite conductivity type into the wafer. Figure 12.3 shows the top and cross-sectional views of two resistors formed by the latter approach: one has a meander shape and the other has a bar shape.

Consider the bar-shaped resistor first. The differential conductance dG of a thin layer of the p-type material that is of thickness dx parallel to the surface and at a depth x (as shown by the B-B cross section) is

$$dG = q\mu_p p(x)\frac{W}{L}dx, \tag{12.1}$$

where W is the width of the bar, L is the length of the bar (we neglect the end contact areas for the time being), μ_p is mobility of hole, and $p(x)$ is the doping concentration. The total conductance of the entire implanted region of the bar is given by

$$G = \int_0^{x_j} dG = q\frac{W}{L}\int_0^{x_j} \mu_p p(x)dx, \tag{12.2}$$

where x_j is the junction depth. If the value of μ_p, which is a function of the hole concentration, and the distribution of $p(x)$ are known, the total conductance can be evaluated from Eq. 12.2. We can write

$$G \equiv g\frac{W}{L}, \tag{12.3}$$

where $g \equiv q \int_0^{x_j} \mu_p p(x)dx$ is the conductance of a square resistor pattern, that is, $G = g$ when $L = W$.

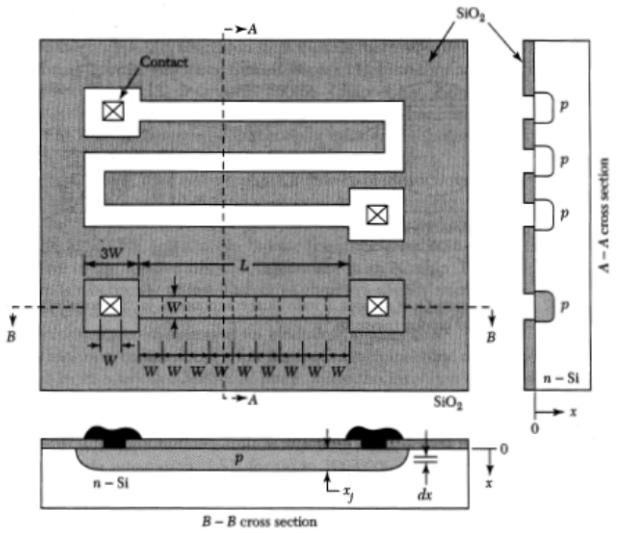


Fig. 12.3. Integrated-circuit resistor. All narrow lines in the large square area have the same width *W*, and all contacts are the same size.

The resistance is therefore given by

$$R \equiv \frac{1}{G} = \frac{L}{W}\left(\frac{1}{g}\right), \tag{12.4}$$

where *1/g* usually is defined by the symbol $R_\square$, and is called the sheet resistance. The sheet resistance has units of ohms but is conventionally specified in units of ohms per square (Ω/□).

Many resistors in an integrated circuit are fabricated simultaneously by defining different geometric patterns in the mask such as those shown in Fig. 12.3. Since the same processing cycle is used for all these resistors, it is convenient to separate the resistance into two parts: the sheet resistance $R_\square$, determined by the implantation (or diffusion) process; and the ratio *L/W*, determined by the pattern dimensions. Once the value of $R_\square$, is known, the resistance is given by the ratio *L/W*, or the number of squares (each square has an area of $W \times W$) in the resistor pattern. The end contact areas will introduce additional resistance to the IC resistors. For the type shown in Fig. 12.3, each end contact corresponds to approximately 0.65 square. For the meander-shape resistor, the electric-field lines at the bends are not spaced uniformly across the width of the resistor but are crowded toward the inside corner. A square at the bend does not contribute exactly 1 square, but rather 0.65 square.

12.2.2. The Integrated-Circuit Capacitor

There are basically two types of capacitors used in integrated circuits: MOS capacitors and *p-n* junctions. The MOS (metal-oxide-semiconductor) capacitor can be fabricated by using a heavily doped region (such as an emitter region) as one plate, the top metal electrode as the other plate, and the intervening oxide layer as the dielectric. The top and cross-sectional views of a MOS capacitor are shown in Fig. 12.4a. To form a MOS capacitor, a thick oxide layer is thermally grown on a silicon substrate. Next, a window is lithographically defined and then etched in the oxide. Diffusion or ion implantation is used to form a p^+-region in the window area, whereas the surrounding thick oxide serves as a mask. A thin oxide layer is then thermally grown in the window area, followed by a metallization step. The capacitance per unit area is given by

$$C = \frac{\varepsilon_{ox}}{d}, \text{ F/cm}^2, \tag{12.5}$$

where ε_{ox} is the dielectric permittivity of silicon dioxide (the dielectric constant $\varepsilon_{ox}/\varepsilon_0$ is 3.9) and d is the thin-oxide thickness. To increase the capacitance further, insulators with higher dielectric constants are being studied, such as Si_3N_4, and Ta_20_5, with dielectric constants of 7 and 25, respectively. The MOS capacitance is essentially independent of the applied voltage, because the lower plate of the capacitor is made of heavily doped material. This also reduces the series resistance associated with it.

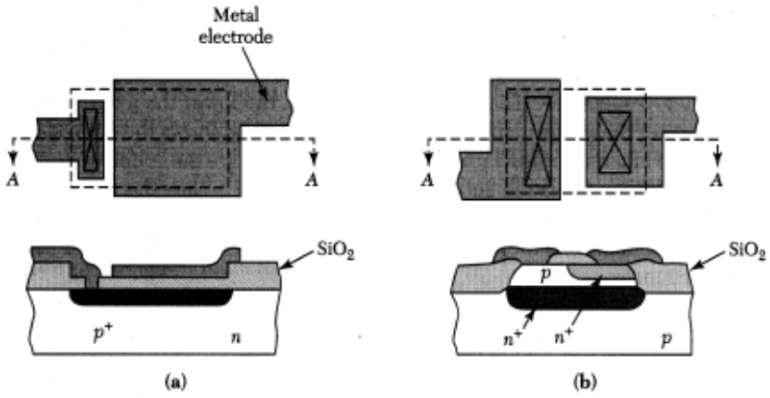


Fig. 12.4. (a) Integrated MOS capacitor. (b) Integrated *p-n* junction capacitor.

A *p-n* junction is sometimes used as a capacitor in an integrated circuit. The top and cross sectional views of an n^+-p junction capacitor are shown in Fig. 12.4b. The detailed fabrication process is considered in Section 12.3, because this structure forms part of a bipolar transistor. As a capacitor, the device is usually reverse biased, that is, the *p*-region is reverse-biased with respect to the n^+-region. The capacitance is not a constant but varies as $(V_R + V_{bi})^{-1/2}$, where V_R is the applied reverse voltage and V_{bi} is the built-in potential. The series resistance is considerably

higher than that of a MOS capacitor because the *p*-region has higher resistivity than does the p^+-region.

12.2.3 The Integrated-Circuit Inductor

IC inductors have been widely used in III-V based monolithic microwave integrated circuits (MMIC) [6]. With the increased speed of silicon devices and advancement in multilevel interconnection technology, IC inductors have started to receive more and more attentions in silicon-based radio frequency (rf) and high-frequency applications. Many kinds of inductors can be fabricated using IC processes. The most popular method is the thin-film spiral inductor. Figure 12.5a and b shows the top-view and the cross section of a silicon-based, two-level-metal spiral inductor. To form a spiral inductor, a thick oxide is thermally grown or deposited on a silicon substrate. The first metal is then deposited and defined as one end of the inductor. Next, another dielectric is deposited onto the metal 1. A via hole is defined lithographically and etched in the oxide. Metal 2 is deposited and the via hole is filled. The spiral patterned can be defined and etched on the metal 2 as the second end of the inductor.

To evaluate the inductor, an important figure of merit is the quality factor, Q. The Q is defined as $Q = L\omega /R$, where L, R, and ω are the inductance, resistance, and frequency, respectively. The higher Q values, the lower the loss from resistance, hence the better the performance of the circuits. Figure 12.5c shows the equivalent circuit model. R_1 is the inherent resistivity of the metal, C_{pl} and C_{p2} are the coupling capacitances between the metal lines and the substrate, and R_{sub1} and R_{sub2} are the resistances of the silicon substrate associated with the metal lines, respectively. The Q increases linearly with frequency initially and then drops at higher frequencies because of parasitic resistances and capacitances.

There are some approaches to improve the Q value. The first is to use low-dielectric-constant materials (<3.9) to reduce the C_p. The other is to use a thick film metal or low-resistivity metals (e.g., Cu, Au to replace Al) to reduce the R_1. The third approach uses an insulating substrate (e.g., silicon-on-sapphire, silicon-on-glass, or quartz) to reduce $R_{sub\ loss}$.

To obtain the exact value of a thin-film inductor, complicated simulation tool, such as computer aided design, must be employed for both circuit simulation and inductor optimization. The model for thin-film inductor must take into account the resistance of the metal, the capacitance of the oxide, line-to-line capacitance, the resistance of the substrate, the capacitance to the substrate, and the inductance and mutual inductance of the metal lines. Hence, it is more difficult to calculate the integrated inductance compared with the integrated capacitors or resistors. However, a simple equation to estimate the square planar spiral inductor is given as [3]

$$L = \mu_0 n^2 r \approx 1.2\times 10^{-6} n^2 r, \qquad (12.6)$$

where μ_0 is the permeability in vacuum ($4\pi\times10^{-7}$ H / m), L is in henries, n is the number of turns, and r is the radius of the spiral in meters.

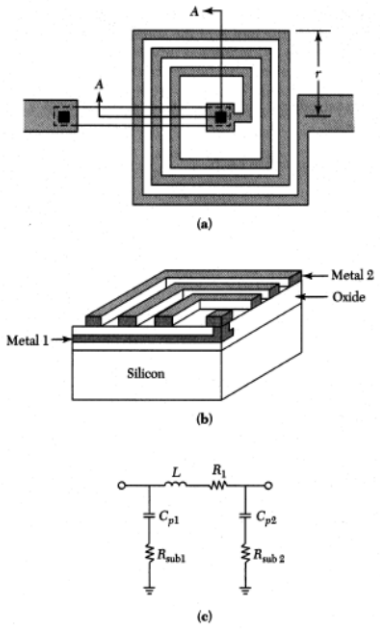


Fig. 12.5. (a) Schematic view of a spiral inductor on a silicon substrate. (b) Perspective view along A-A'. (c) An equivalent circuit model for an integrated inductor.

12.3. Bipolar Technology

For IC applications, especially for VLSI and ULSI, the size of bipolar transistors must be reduced to meet the high-density requirement. Figure 12.6 illustrates the reduction in the size of the bipolar transistor in recent years [7]. The main differences in a bipolar transistor in an IC compared with a discrete transistor are that all electrode contacts are located on the top surface of the IC wafer, and each transistor must be electrically isolated to prevent interactions between devices. Prior to 1970, both the lateral and vertical isolations were provided by *p*-*n* junctions (Fig. 12.6a) and the lateral *p*-isolation region was always reverse biased with respect to the *n*-type collector. In 1971, thermal oxide was used for lateral isolation, resulting in a substantial reduction in device size (Fig. 12.6b), because the base and collector contacts abut the isolation region. In the mid-1970s, the emitter extended to the walls of the oxide, resulting in an additional reduction in area (Fig. 12.6c). At the present time, all the lateral and vertical dimensions have been scaled down and emitter stripe widths have dimensions in the submicron region (Fig. 12.6d).

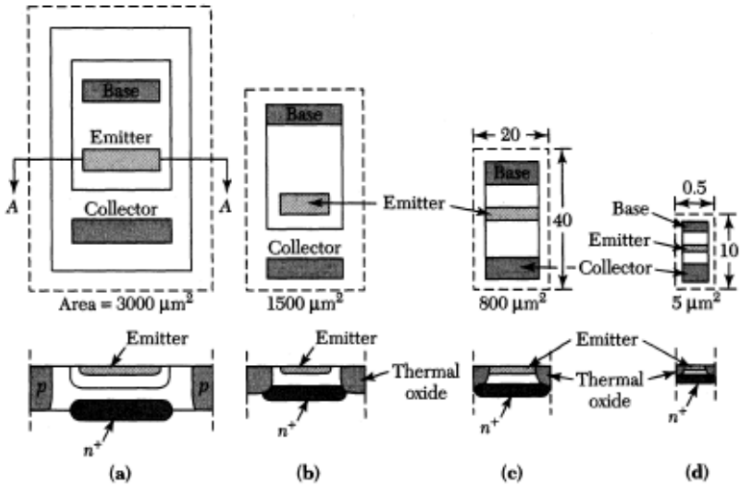


Fig. 12.6. Reduction of the horizontal and vertical dimensions of a bipolar transistor. (a) Junction isolation. (b) Oxide isolation. (c and d) Scaled oxide isolation [7].

12.3.1 The basic fabrication process

The majority of bipolar transistor used in ICs are of the *n*-*p*-*n* type because the higher mobility of minority carriers (electrons) in the base region results in higher-speed performance than can be obtained with *p*-*n*-*p* types. Figure 12.7 shows a perspective view of an *n*-*p*-*n* bipolar transistor, in which lateral isolation is provided by oxide walls and vertical isolation is provided by the n^+-*p* junction. The lateral oxide isolation approach reduces not only the device size but also the parasitic capacitance because of the smaller dielectric constant of silicon dioxide (3.9, compared with 11.9 for silicon). Let's consider the major process steps that are used to fabricate the device shown in Fig. 12.7.

For an *n*-*p*-*n* bipolar transistor, the starting material is a *p*-type lightly doped (~10^{15} cm^3), <111> or <100>-oriented, polished silicon wafer. Because the junctions are formed inside the semiconductor, the choice of crystal orientation is not as critical as for MOS devices. The first step is to form a buried layer. The main purpose of this layer is to minimize the series resistance of the collector. A thick oxide (0.5-1 μm) is thermally grown on the wafer, and a window is then opened in the oxide. A precisely controlled amount of low-energy arsenic ions (~30 keV, ~10^{15} cm^{-2}) is implanted into the window region to serve as a predeposit (Fig. 12.8a). Next, a high temperature (~1100°C) drive-in step forms the n^+-buried layer, which has a typical sheet resistance of 20 Ω/□.

The second step is to deposit an *n*-type epitaxial layer. The oxide is removed and the wafer is placed in an epitaxial reactor for epitaxial growth. The thickness and the doping concentration of the epitaxial layer are determined by the ultimate use of the device. Analog circuits (with their higher voltages for amplification) require thicker layer (~10 μm) and lower

dopings (~5 × 10^{15} cm^{-3}), whereas digital circuits (with their lower voltages for switching) require thinner layers (~3 μm) and higher dopings **(**~2 × 10^{16} cm^{-3}). Figure 12.8b shows a cross-sectional view of the device after the epitaxial process. Note that there is some outdiffusion from the buried layer into the epitaxial layer. To minimize the outdiffusion, a low-temperature epitaxial process should be employed, and low-diffusivity impurities should be used in the buried layer (e.g., As).

The third step is to form the lateral oxide isolation region. A thin-oxide pad (~50 nm) is thermally grown on the epitaxial layer, followed by a silicon-nitride deposition (~100 nm). If nitride is deposited directly onto the silicon without the thin-oxide pad, the nitride may cause damages to the silicon surface during the subsequent high-temperature steps. Next, the nitride-oxide layers and about half of the epitaxial layer are etched using a photoresist as mask (Fig. 12.8c and 12.8d). Boron ions are then implanted into the exposed silicon areas (Fig. 12.8d).

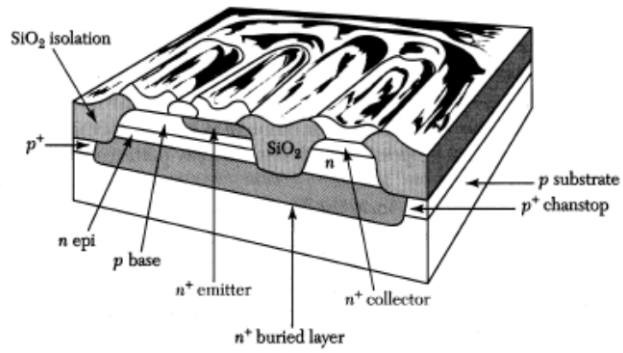


Fig. 12.7. Perspective view of an oxide-isolated bipolar transistor.

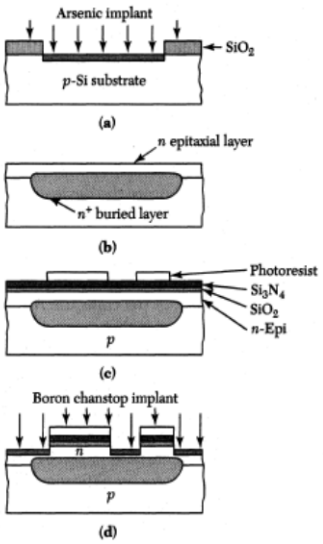


Fig. 12.8. Cross-sectional views of bipolar transistor fabrication. (a) Buried-layer implantation. (b) Epitaxial layer. (c) Photoresist mask. (d) Chanstop implant.

The photoresist is removed and the wafer is placed in an oxidation furnace. Since the nitride layer has a very low oxidation rate, thick oxides will be grown only in the areas not protected by the nitride layer. The isolation oxide is usually grown to a thickness such that the top of the oxide becomes coplanar with the original silicon surface to minimize the surface topography. This oxide isolation process is called local oxidation of silicon (LOCOS). Figure 12.9a shows the cross section of the isolation oxide after the removal of the nitride layer. Because of segregation effects, most of the implanted boron ions are pushed underneath the isolation oxide to form a p^+-layer. This is called p^+ channel stop (or chanstop), because the high concentration of *p*-type semiconductor will prevent surface inversion and eliminate possible high-conductivity paths (or channels) among neighboring buried layers.

The fourth step is to form the base region. A photoresist is used as a mask to protect the right half of the device; then, boron ions **(**~10^{12} cm^{-2}) are implanted to form the base regions, as shown in Fig. 12.9b. Another lithographic process removes all the thin-pad oxide except a small area near the center of the base region (Fig. 12.9c).

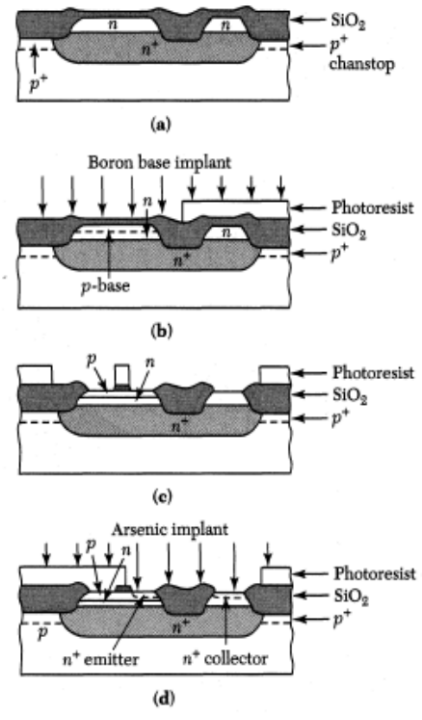


Fig. 12.9. Cross-section views of bipolar transistor fabrication. (a) Oxide isolation. (b) Base implant. (c) Removal of thin oxide. (d) Emitter and collector implant.

The fifth step is to form the emitter region. As shown in Fig. 12.9d, the base contact area is protected by a photoresist mask; then, a low-energy, high-arsenic-dose **(**~10^{16} **c**m^{-2}) implantation forms the n^+-emitter and the n^+-collector contact regions. The photoresist is

removed; and a final metallization step forms the contacts to the base, emitter, and collector as shown in Fig. 12.7.

In this basic bipolar process, there are six film formation operations, six lithographic operations, four ion implantations, and four etching operations. Each operation must be precisely controlled and monitored. Failure of any one of the operations generally will render the wafer useless.

The doping profiles of the completed transistor along a coordinate perpendicular to the surface and passing through the emitter, base, and collector are shown in Fig. 12.10. The emitter profile is abrupt because of the concentration-dependent diffusivity of arsenic. The base doping profile beneath the emitter can be approximated by a Gaussian distribution for a limited-source diffusion. The collector doping is given by the epitaxial doping level (~2×10^{16} cm^{-3}) for a representative switching transistor; however, at larger depths, the collector doping concentration increases because of outdiffusion from the buried layer.

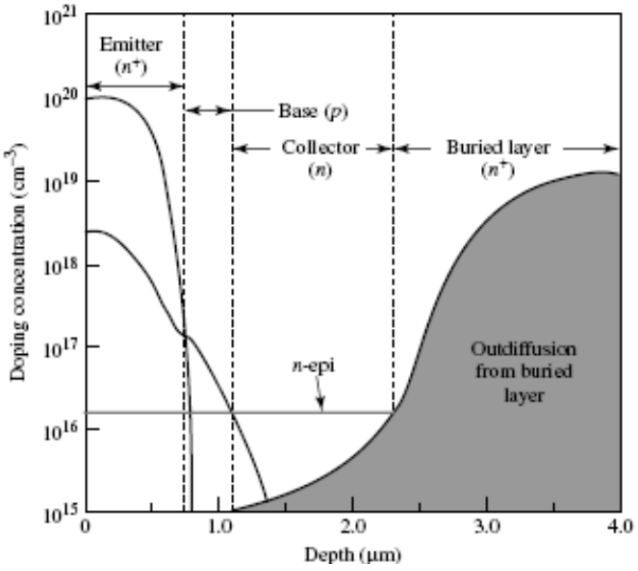


Fig. 12.10. *n-p-n* transistor doping profiles.

12.3.2. Dielectric isolation

In the isolation scheme described previously for the bipolar transistor, the device is isolated from other devices by the oxide layer around its periphery and is isolated from its common substrate by a n^+-p junction (buried layer). In high-voltage applications, a different approach, called dielectric isolation, is used to form insulating tubs to isolate a number of pockets of single-crystal semiconductors. In this approach the device is isolated from both its common substrate and its surrounding neighbors by a dielectric layer.

A process sequence for the dielectric isolation is shown in Fig. 12.11. An oxide layer is formed inside a <100>-oriented *n*-type silicon substrate using high-energy oxygen ion implantation (Fig. 12.11a). Next, the wafer undergoes a high-temperature annealing process so that the implanted oxygen will react with silicon to form the oxide layer. The damage resulting from implantation is also annealed out in this process (Fig. 12.11b). After this, we can obtain an *n*-silicon layer that is fully isolated on an oxide [namely, silicon-on-insulator, (SOI)]. This process is called SIMOX (separation by implanted oxygen). Since the top silicon is so thin, the isolation region is easily formed by the LOCOS process illustrated in Fig. 12.8c or by etching a trench (Fig. 12.11c) and refilling it with oxide (Fig. 12.11d). The other processes are almost the same as those from Fig. 12.8c through Fig. 12.9 to form the *p*-type base, n^+-emitter, and collector.

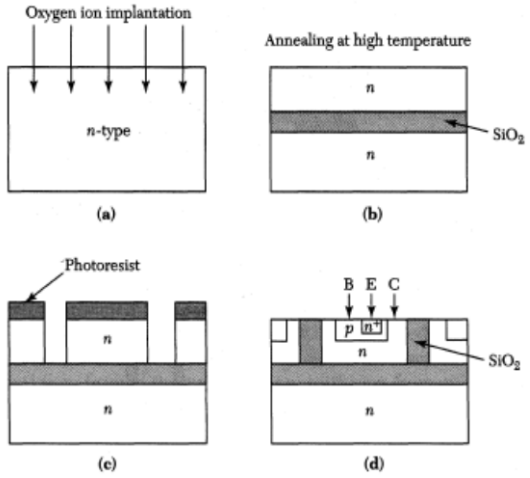


Fig. 12.11. Process sequence for dielectric isolation bipolar device using silicon-on-insulator for high-voltage application. (a) Oxygen ion implantation. (b) Annealing at high temperature to form the isolation dielectric. (c) Trench isolation formed by a dry-etching process. (d) Base, emitter, and collector formation.

The main advantage of this technique is its high breakdown voltage between the emitter and the collector, which can be in excess of several hundreds volts. This technique is also compatible with modern CMOS integration. This CMOS-compatible process is very useful for mixed high-voltage and high-density IC.

12.3.3. Self-Aligned Double-Polysilicon Bipolar Structure

The process shown in Fig. 12.9c needs another lithographic process to define an oxide region to separate the base and emitter contact regions. This gives rise to a large inactive device area within the isolated boundary, which increases not only the parasitic capacitances but also

the resistance that degrades the transistor performance. The most effective way to reduce these effects is by using the self-aligned structure.

The most widely used self-aligned structure is the double-polysilicon structure with the advanced isolation provided by a trench refilled with polysilicon [8], shown in Fig. 12.12. Figure 12.13 shows the detail sequence of the steps for the self-aligned double-polysilicon (*n*-*p*-*n*) bipolar structures [9]. The transistor is built on an *n*-type epitaxial layer. A trench of 5.0 μm in depth is etched by reactive ion etching through the n^+-subcollector region into the *p*-substrate region. A thin layer of thermal oxide is then grown and serves as the screen oxide for the channel stop implant of boron at the bottom of the trench. The trench is then filled with undoped polysilicon and capped by a thick planar field oxide.

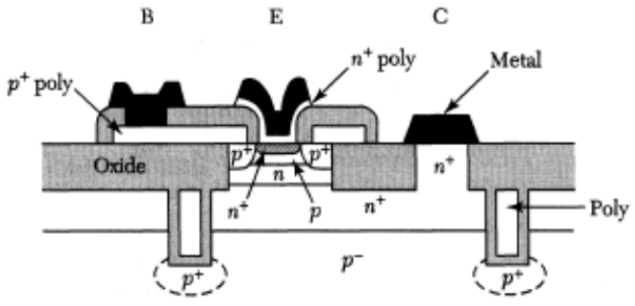


Fig. 12.12. Cross-section of a self-aligned, double-polysilicon bipolar transistor with advanced trench inolation [8].

The first polysilicon layer is deposited and heavily doped with boron. The p^+-polysilicon (called poly 1) will be used as a solid-phase diffusion source to form the extrinsic base region and the base electrode. This layer is covered with a chemical-vapor deposition (CVD) oxide and nitride (Fig. 12.13a). The emitter mask is used to pattern the emitter-area regions, and a dry-etch process is used to produce an opening in the CVD oxide and poly 1 (Fig. 12.13b). A thermal oxide is then grown over the etched structure, and a relatively thick oxide (approximately 0.1-0.4 μm) is grown on the vertical sidewalls of the heavily doped poly. The thickness of this oxide determines the spacing between the edges of the base and emitter contacts. The extrinsic p^+ base regions are also formed during the thermal-oxide growth step as a result of the outdiffusion of boron from the poly 1 into the substrate (Fig. 12.13c). Because boron diffuses laterally as well as vertically, the extrinsic base region will be able to make contact with the intrinsic base region that is formed next, under the emitter contact.

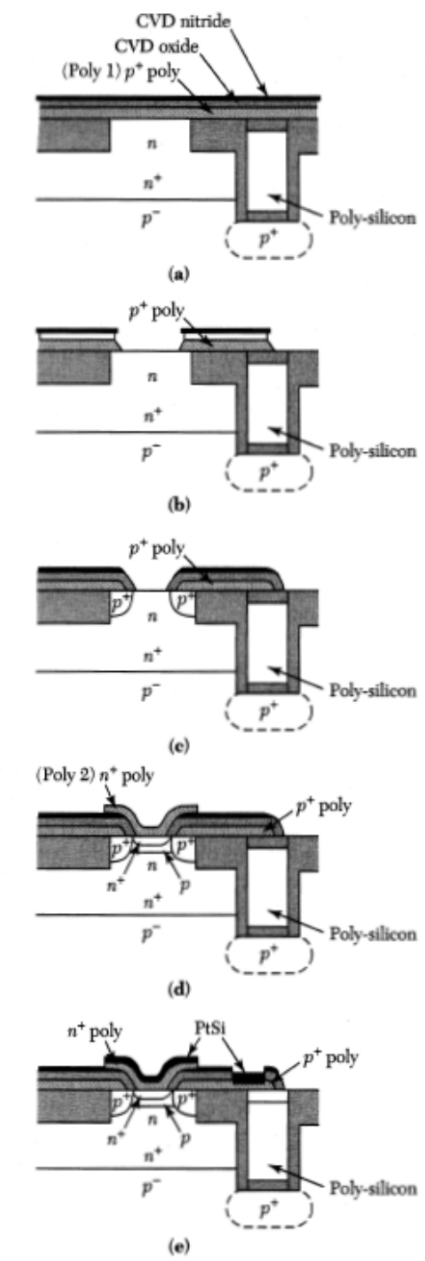


Fig. 12.13. Process sequence for fabricating double-polysilicon self-aligned *n-p-n* transistors [9].

Following the oxide-grown step, the intrinsic base region is formed using ion implantation of boron (Fig. 12.13d). This serves to self-align the intrinsic and extrinsic base regions. After the contact is cleaned to remove any oxide layer, the second polysilicon layer is deposited and implanted with As or P. The n^+-polysilicon (called poly 2**)** is used as a solid phase diffusion source to form the emitter region and the emitter electrode. A shallow emitter region is then formed through dopant outdiffusion from poly 2. A rapid thermal anneal for the base and emitter outdiffusion steps facilitates the formation of shallow emitter-base and collector-base junctions. Finally, Pt film is deposited and sintered to form PtSi over the n^+-polysilicon emitter and the p^+-polysilicon base contact (Fig. 12.13e).

This self-aligned structure allows the fabrication of emitter regions smaller than the minimum lithographic dimension. When the sidewall-spacer oxide is grown, it fills the contact hole to some degree because the thermal oxide occupies a larger volume than the original volume of polysilicon. Thus, an opening 0.8 μm wide will shrink to about 0.4 μm if sidewall oxide a 0.2 μm thick is grown on each side.

12.4. MOSFET technology

At present, the MOSFET is the dominant device used in ULSI circuits because it can be scaled to smaller dimensions than other types of devices. The dominant technology for MOSFET is the CMOS (complementary MOSFET) technology, in which both *n*-channel and *p*-channel MOSFETs (called NMOS and PMOS, respectively) are provided on the same chip. CMOS technology is particular attractive for ULSI circuits because it has the lowest power consumption of all IC technology. Figure 12.14 shows the reduction in the size of the MOSFET in recent years. In the early 1970s, the gate length was 7.5 μm and the corresponding device area was about 6000 μm^2. As the device is scaled down, there is a drastic reduction in the device area. For a MOSFET with a gate length of 0.5 μm, the device area shrinks to less than 1% of the early MOSFET. We expect that device miniaturization will continue. The gate length became less than 0.10 μm in the early twenty-first century.

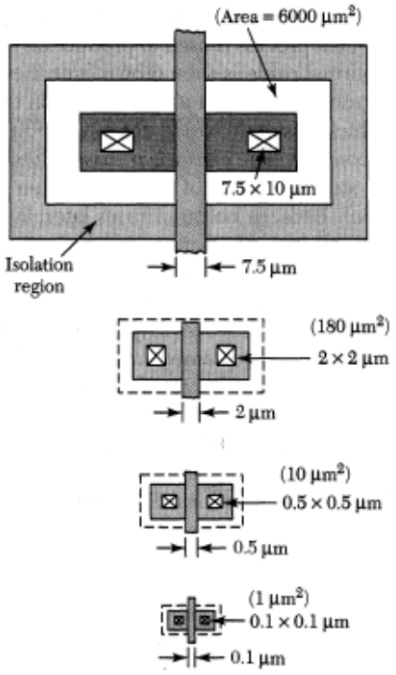


Fig. 12.14. Reduction in the area of the MOSFET as the gate length (minimum feature length) is reduced.

12.4.1 The Basic Fabrication Process

Figure 12.15 shows a perspective view of an *n*-channel MOSFET prior to its final metallization [10]. The top layer is a phosphorus-doped silicon dioxide (P-glass) that is used as an insulator between the polysilicon gate and the gate metallization and also as a gettering layer for mobile ions. Compare Fig. 12.15 with Fig. 12.7 for the bipolar transistor and note that a MOSFET is considerably simpler in its basic structure. Although both devices use lateral oxide isolation, there is no need for vertical isolation in the MOSFET, whereas a buried-layer n^+-p junction is required in the bipolar transistor. The doping profile in a MOSFET is not as complicated as that in a bipolar transistor and the control of the dopant distribution is also less critical. Let's consider the major process steps that are used to fabricate the device shown in Fig. 12.15.

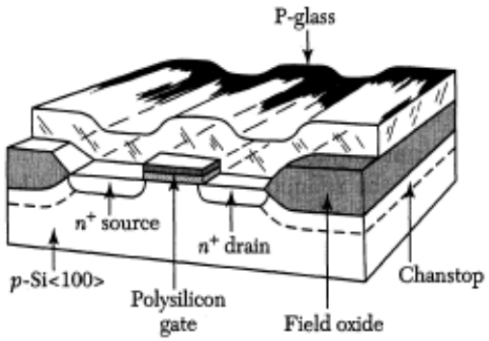


Fig. 12.15. Perspective view of an *n*-channel MOSFET [10].

To process an *n*-channel MOSFET (NMOS), the starting material is a *p*-type, lightly doped (~10^{15} cm^{-3}), <100>-oriented, polished silicon wafer. The <100>-orientation is preferred over <111> because it has an interface-trap density that is about one-tenth that of <111>.

The first step is to form the oxide isolation region using LOCOS technology. The process sequence for this step is similar to that for the bipolar transistor. A thin-pad oxide (~35 nm) is thermally grown, followed by a silicon nitride (~150 nm) deposition (Fig. 12.16a) [10]. The active device area is defined by a photoresist mask and a boron chanstop layer is then implanted through the composite nitride-oxide layer (Fig. 12.16b). The nitride layer not covered by the photoresist mask is subsequently removed by etching. After stripping the photoresist, the wafer is placed in an oxidation furnace to grow an oxide (called the field oxide), where the nitride layer is removed, and to drive in the boron implant. The thickness of the field oxide is typically 0.5-1 μm.

The second step is to grow the gate oxide and to adjust the threshold voltage. The composite nitride-oxide layer over the active device area is removed, and a thin-gate oxide layer (less than 10 nm) is grown. For an enhancement-mode *n*-channel device, boron ions are implanted in the channel region, as shown in Fig. 12.16c, to increase the threshold voltage to a predetermined value (e.g., + 0.5V). For a depletion-mode *n*-channel device, arsenic ions are implanted in the channel region to decrease the threshold voltage (e.g., - 0.5V).

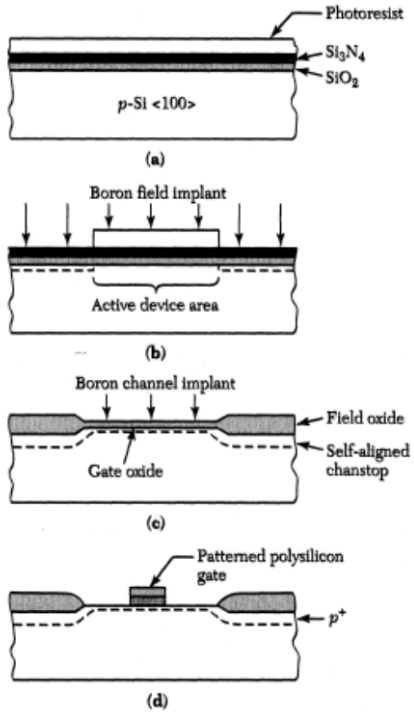


Fig. 12.16. Cross-sectional view of NMOS fabrication sequence [10]. (*a*) Formation of SiO_2, Si_3N_4, and photoresist layer, *(b)* Boron implant, *(c)* Field oxide, *(d)* Gate.

The third step is to form the gate. A polysilicon is deposited and is heavily doped by diffusion or implantation of phosphorus to a typical sheet resistance of 20-30 Ω/□. This resistance is adequate for MOSFETs with gate lengths larger than 3 μm. For smaller devices, polycide, a composite layer of metal silicide and polysilicon such as W-polycide, can be used as the gate materials to reduce the sheet resistance to about 1 Ω/□.

The fourth step is to form the source and drain. After the gate is patterned (Fig. 12.16d), it serves as a mask for the arsenic implantation (~30 keV, ~5×10^{15} cm^{-2}) to form the source and drain (Fig. 12.17a), which are self-aligned with respect to the gate [10]. At this stage, the only overlapping of the gate is due to lateral straggling of the implanted ions (for 30 keV As, $\sigma_\perp$ is only 5 nm). If low-temperature processes are used for subsequent steps to minimize lateral

diffusion, the parasitic gate-drain and gate-source coupling capacitances can be much smaller than the gate-channel capacitance.

The last step is the metallization. A phosphorus-doped oxide (P-glass) is deposited over the entire wafer and is flowed by heating the wafer to give a smooth surface topography (Fig. 12.17b). Contact windows are defined and etched in the P-glass. A metal layer, such as aluminum, is then deposited and patterned. A cross-section view of the completed MOSFET is shown in Fig. 12.17c, and the corresponding top view is shown in Fig. 12.17d. The gate contact is usually made outside the active device area to avoid possible damage to the thin-gate oxide.

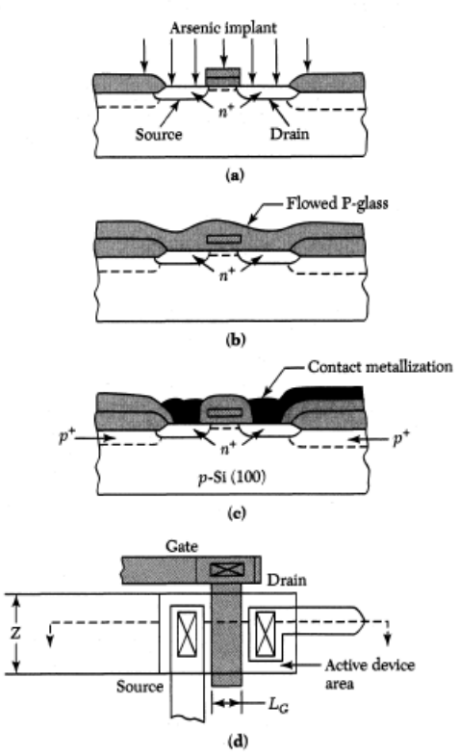


Fig. 12.17. NMOS fabrication sequence [10]. (*a*) Source and drain. (*b*) P-glass deposition. (*c*) Cross section of the MOSFET. (*d*) Top view of the MOSFET.

12.4.2. Memory Devices

Memories are devices that can store digital information (or data) in terms of bits (binary digits). Various memory chips have been designed and fabricated using NMOS technology. For most large memories, the random access memory (RAM) organization is preferred. In a RAM, memory cells are organized in a matrix structure and data can be accessed (i.e., stored, retrieved, or erased) in random order, independent of their physical locations. A static random access memory (SRAM) can retain stored data indefinitely as long as the power supply is on. The SRAM is basically a flip-flop circuit that can store one bit of information. A SRAM cell has four enhancement-mode MOSFETs and two depletion-mode MOSFETs. The depletion-mode

MOSFETs can be replaced by resistors formed in undoped polysilicon to minimize power consumption [8].

To reduce the cell area and power consumption, the dynamic random access memory (DRAM) has been developed. Figure 12.18a shows the circuit diagram of the one-transistor DRAM cell in which the transistor serves as a switch and one bit of information can be stored in the storage capacitor. The voltage level on the capacitor determines the state of the cell. For example, +1.5 V may be defined as logic 1 and 0 V defined as logic 0. The stored charge will be removed typically in a few milliseconds mainly because of the leakage current of the capacitors; thus, dynamic memories require periodic "refreshing" of the stored charge.

Figure 12.18b shows the layout of a DRAM cell, and Fig. 12.18c shows the corresponding cross section through AA'**.** The storage capacitor uses the channel region as one plate, the polysilicon gate as the other plate, and the gate oxide as the dielectric. The row line is a metal track to minimize the delay due to parasitic resistance *(R)* and parasitic capacitance (*C*), the *RC* delay. The column line is formed by n^+-diffusion. The internal drain region of the MOSFET serves as a conductive link between the inversion layers under the storage gate and the transfer gate. The drain region can be eliminated by using the double-level polysilicon approach shown in Fig. 12.18d. The second polysilicon electrode is separated from the first polysilicon capacitor plate by an oxide layer that is thermally grown on the first-level polysilicon before the second electrode has been defined. The charge from the column line can therefore be transmitted directly to the area under the storage gate by the continuity of inversion layers under the transfer and storage gates.

To meet the requirements of high-density DRAM, the DRAM structure has been extended to the third dimension with stacked or trench capacitors. Figure 12.19a shows a simple trench cell structure [12]. The advantage of the trench type is that the capacitance of the cell could be increased by increasing the depth of the trench without increasing the surface area of silicon occupied by the cell. The main difficulties of making trench type cells are the etching of the deep trench, which needs a rounded bottom corner and the growth of a uniform thin dielectric film on trench walls. Figure 12.19b shows a stacked cell structure. The storage capacitance increases as a result of stacking the storage capacitor on top of the access transistor. The dielectric is formed using the thermal oxidation or CVD nitride methods between the two-polysilicon plates. Hence, the stacked cell process is easier than the trench type process.

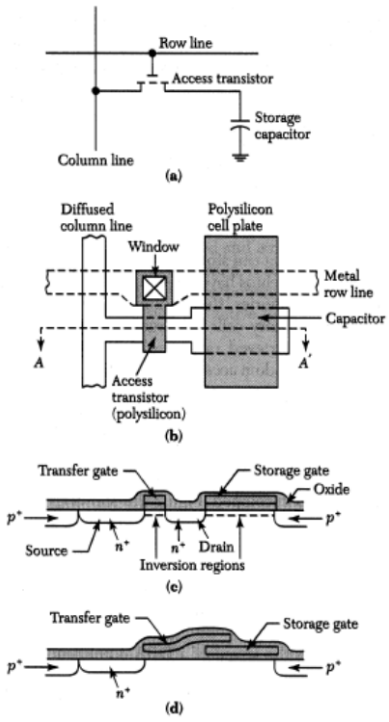


Fig. 12.18. Single-transistor dynamic random access memory (DRAM) cell with a storage capacitor [11]. (*a*) Circuit diagram. (*b*) Cell layout. *(c)* Cross section through A-A', *(d)* Double-level polysilicon.

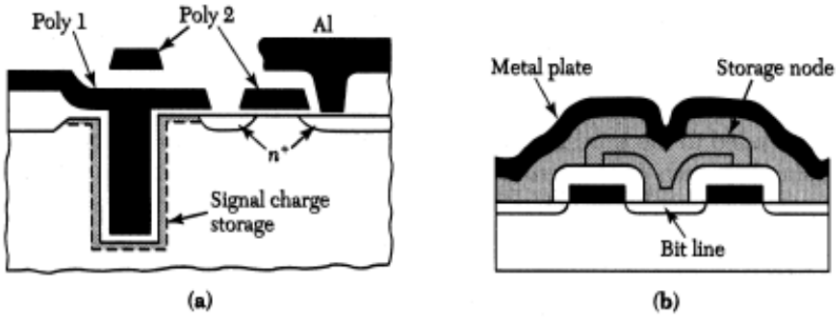


Fig. 12.19. (a) DRAM with a trench cell structure [12]. (b) DRAM with a single-layer stacked-capacitor cell.

Figure 12.20 shows a 1 Gb DRAM chip. This memory chip uses 0.18 μm design rules. Trench capacitors and its peripheral circuits are in CMOS. The memory chip has an area of 380 mm^2 (14.3 mm x 27.3 mm) that contains over 2 billion components and operates at 2.5 V. This 1 Gb DRAM is mounted in an 88-pin ceramic package, which can provide adequate heat dissipation.

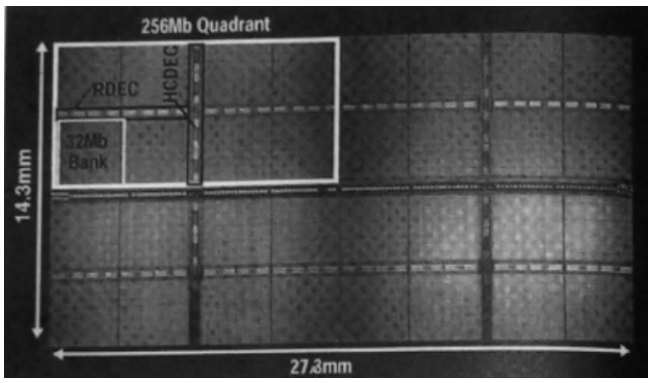


Fig. 12.20. A 1 Gb DRAM that contains over 2 billion components.

Both SRAM and DRAM are volatile memories, that is, they lose their stored data when power is switched off. Nonvolatile memories, on the other hand, can retain their data. Figure 12.21a shows a floating-gate nonvolatile memory, which is basically a conventional MOSFET that has a modified gate electrode. The composite gate has a regular (control gate) and a floating gate which is surrounded by insulators. When a large positive voltage is applied to the control gate, charge will be injected from the channel region through the gate oxide into the floating gate. When the applied voltage is removed, the injected charge can be stored in the floating gate for a long time. To remove this charge, a large negative voltage must be applied to the control gate, so that the charge will be injected back into the channel region.

Another version of the nonvolatile memory is the metal-nitride-oxide-semiconductor (MNOS) type shown inn Fig. 12.21b. When a positive gate voltage is applied, electrons can tunnel through the thin oxide layer (~2 nm) and be captured by the traps at the oxidenitride interface, and thus become stored charges there. The equivalent circuit for both types of nonvolatile memories can be represented by two capacitors in series for the gate structure, as illustrated in Fig. 12.21c. The charge stored in the capacitor C_1 causes a shift in the threshold voltage, and the device remains at the higher threshold voltage-state (logic 1). For a well-designed memory device, the charge retention time can be over 100 years. To erase the memory (e.g., the store charge) and return the device to a lower threshold voltage state (logic 0), a gate voltage or other means (such as ultraviolet light) can be used.

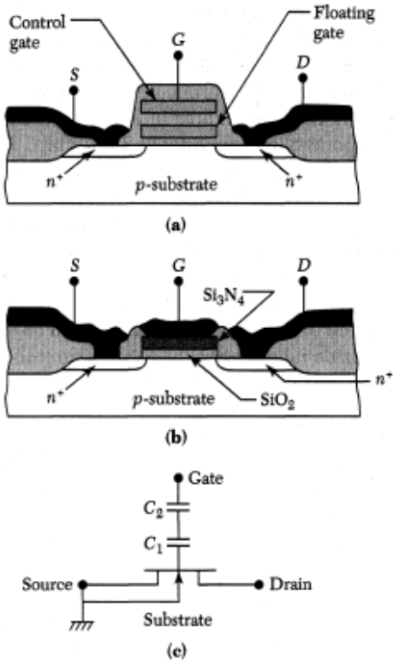


Fig. 12.21. Nonvolatile memory devices. (*a*) Floating-gate, nonvolatile memory, (*b*) MNOS nonvolatile memory. (*c*) Equivalent circuit of either type of nonvolatile memory.

The nonvolatile semiconductor memory (NVSM) has been extensively used in portable electronics systems, such as cellular phones and the digital cameras. Another interesting application is the chip card, also called IC card. The top photo in Fig. 12.22 shows an IC card. The diagram at the bottom of Fig. 12.22 illustrates the nonvolatile memory device that stores the data that can be read and written through the bus to a central processing unit (CPU). In contrast to the limited volume (1 kbytes) inside a conventional magnetic tape card, the size of the nonvolatile memory can be increased to 16 kbytes, 64 kbytes, or even larger depending on the applications (e.g., you can store personal photos or finger prints). Through the IC card read/write machines, the data can be used in numerous applications, such as telecommunications (card telephone, mobile radio), payment transactions (electronic purse, credit card), pay television, transport (electronic ticket, public transport), health care (patient-data card), and access control. The IC card will play a central role in the global information and service society of the future [13].

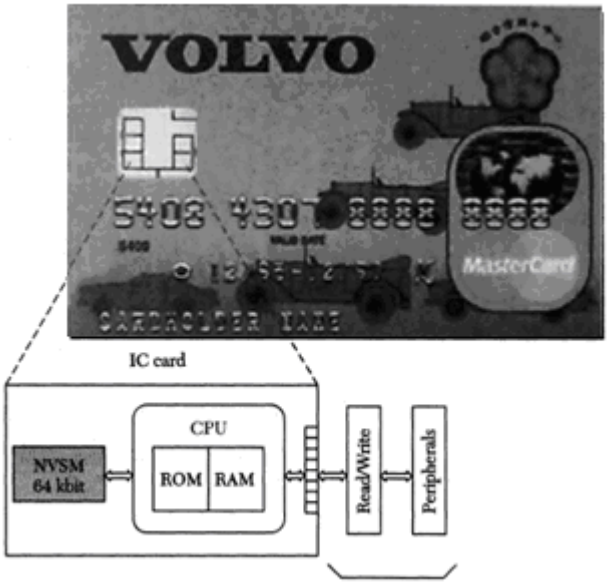


Fig. 12.22. An integrated-circuit (IC) card. The data stored in the NVSM can be accessed through the bus of the central processing unit (CPU). There are several metal pads connecting to the read/write machine.

12.4.3 CMOS Technology

Figure 12.23a shows a CMOS inverter. The gate of the upper PMOS device is connected to the gate of the lower NMOS device. Both devices are enhancement-mode MOSFETs with the threshold voltage V_{Tp} less than zero for the PMOS device and V_{Tn} greater than zero for the NMOS device (typically the threshold voltage is about 1/4 V_{DD}). When the input voltage V_i is at ground or at small positive values, the PMOS device is turned on (the gate-to-ground potential of PMOS is $-V_{DD}$, which is more negative than V_{Tp}), and the NMOS device is off. Hence, the output voltage V_o is very close to V_{DD} (logic 1). When the input is at V_{DD}, the PMOS (with $V_{GS} = 0$) is turned off, and the NMOS is turned on ($V_i = V_{DD} > V_{Tn}$). Therefore, the output voltage V_o equals zero (logic 0). The CMOS inverter has a unique feature: in either logic state, one device in the series path from V_{DD} to ground is nonconductive. The current that flows in either steady state is a small leakage current, and only when both devices are on during switching does a significant current flow through the CMOS inverter. Thus, the average power dissipation is small, on the order of nanowatts. As the number of components per chip increases, the power dissipation becomes a major limiting factor. The low power consumption is the most attractive feature of the CMOS circuit.

Figure 12.23b shows a layout of the CMOS inverter, and Fig. 12.23c shows the device cross section along the A-A' line. In the processing, a *p*-tub (also called a *p*-well) is first implanted and subsequently driven into the *n*-substrate. The *p*-type dopant concentration must be high enough to overcompensate the background doping of the *n*-substrate. The subsequent

processes for the *n*-channel MOSFET in the *p*-tub are identical to those described previously. For the *p*-channel MOSFET, $^{11}B^+$ or $^{49}(BF_2)^+$ ions are implanted into the *n*-substrate to form the source and drain regions. A channel implant of $^{75}As^+$ ions may be used to adjust the threshold voltage and a n^+-chanstop is formed underneath the field oxide around the *p*-channel device. Because of the *p*-tub and the additional steps needed to make the *p*-channel MOSFET, the number of steps to make a CMOS circuit is essentially double that to make an NMOS circuit. Thus, we have a trade-off between the complexity of processing and a reduction in power consumption.

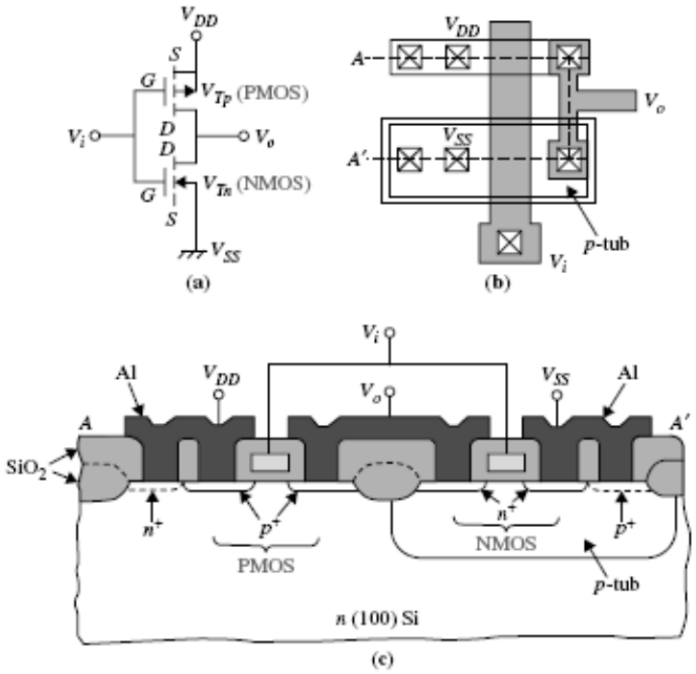


Fig. 12.23. Complementary MOS (CMOS) inverter. (*a*) Circuit diagram. (*b*) Circuit layout. (*c*) Cross section along dotted A-A' line of (*b*).

Instead of the *p*-tub described above, an alternate approach is to use an *n*-tub formed in *p*-type substrate, as shown in Fig. 12.24a. In this case, the *n*-type dopant concentration must be high enough to overcompensate for the background doping of the *p*-substrate (i.e., $N_D > N_A$). In both the *p*-tub and the *n*-tub approach, the channel mobility will be degraded because mobility is determined by the total dopant concentration ($N_A + N_D$). A recent approach using two separated tubs implanted into a lightly doped substrate is shown in Fig. 12.24b. This structure is called a twin tub [4]. Because no overcompensation is needed in either of the twin tubs, higher channel mobility can be obtained. All CMOS circuits have the potential for a trouble some problem called latchup that is associated with parasitic bipolar transistors. An effective processing technique to eliminate latchup problem is to use the deep-trench isolation, as shown [14] in Fig.

12.24c. In this technique, a trench with a depth deeper than the well is formed in the silicon by anisotropic reactive sputter etching. An oxide layer is thermally grown on the bottom and walls of the trench, which is then refilled by deposited polysilicon or silicon dioxide. This technique can eliminate latchup because the *n*-channel and *p*-channel devices are physically isolated by the refilled trench. The detailed steps for trench isolation and some related CMOS processes are now considered.

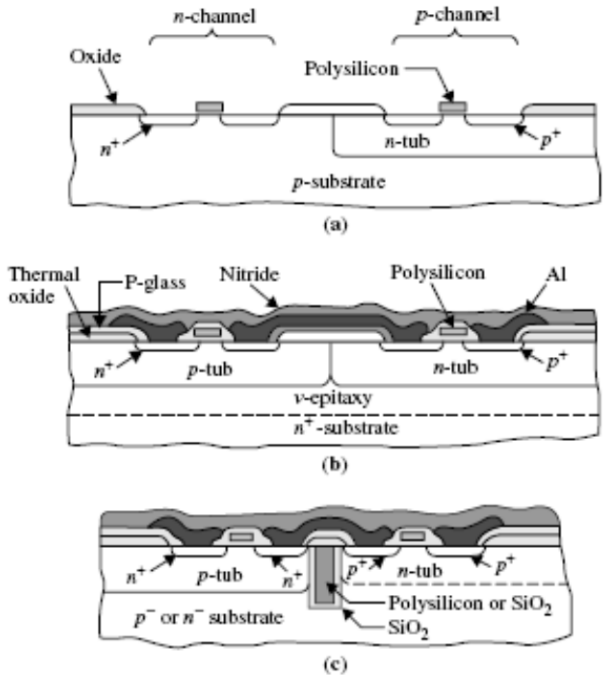


Fig. 12.24. Various CMOS structures. (*a*) *n*-tub. (*b*) Twin tub [4]. (*c*) Refilled trench [14].

Well-Formation Technology

The well of a CMOS can be a single well, a twin well, or a retrograde well. The twin well process exhibits some disadvantages, e.g., it needs high temperature processing (above 1050°C) and a long diffusion time (longer than 8 hours) to achieve the required depth of 2-3 μm. In this process, the doping concentration is highest at the surface and decreases monotonically with depth. To reduce the process temperature and time, high-energy implantation is used, i.e., implanting the ion to the desired depth instead of diffusion from the surface. Since the depth is determined by the implantion energy, we can design the well depth with different implantation energy. The profile of the well in this case can have a peak at a certain depth in the silicon substrate. This is called a retrograde well. Figure 12.25 shows a comparison of the impurity profiles in the retrograde well and the conventional thermal diffused well [15]. The energy for the *n*- and *p*-type retrograde wells is around 700 keV and 400 keV, respectively. As mentioned

above, the advantage of the high energy implantation is that it can form the well under low-temperature and short-time conditions; hence, it can reduce the lateral diffusion and increase the device density. The retrograde well can offer some additional advantages over the conventional well: (*a*) because of high doping near the bottom, the well resistivity is lower than that of the conventional well and the latchup problem can be minimized, (*b*) the chanstop can be formed at the same time as the retrograde well implantation, reducing processing steps and time, (*c*) higher well doping in the bottom can reduce the chance of punchthrough from the drain to the source.

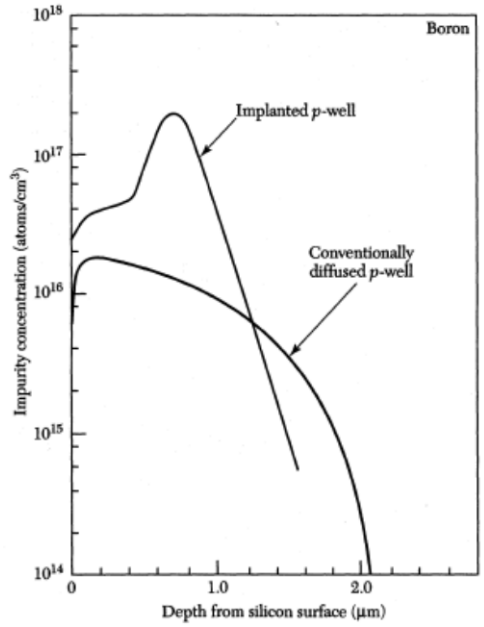


Fig. 12.25. Retrograded *p*-well implanted impurity concentration profile. Also shown is a conventionally diffused well [15].

Advanced Isolation Technology

The conventional isolation process (Section 12.4.1) has some disadvantages that make it unsuitable for deep-submicron (0.25 μm and smaller) fabrications. The high-temperature oxidation of silicon and long oxidation time result in the encroachment of the chanstop implantation (usually boron for *n*-MOSFET) to the active region and cause V_T shift. The area of the active region is reduced because of the lateral oxidation. In addition, the field oxide thickness in submicron-isolation spacings is significantly less than the thickness of field oxide grown in wider spacings. The trench isolation technology can avoid these problems and has become the mainstream technology for isolation. Figure 12.26 shows the process sequence for forming a deep (larger than 3 μm) but narrow (less than 2 μm) trench-isolation structure. There are four steps: patterning the area, trench etching and oxide growth, refilling with dielectric materials

such as oxide or undoped polysilicon, and planarization. This deep trench isolation can be used in both advanced CMOS and bipolar devices and for the trench-type DRAM. Since the isolation material is deposited by CVD, it does not need a long-time or a high-temperature process, and it eliminates the lateral oxidation and boron encroachment problems.

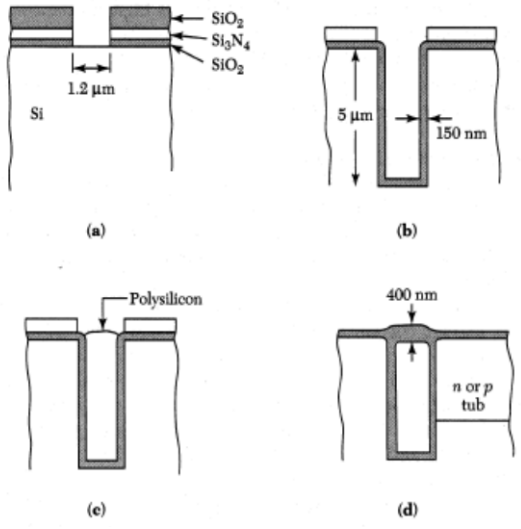


Fig. 12.26. Process sequence for forming a deep, narrow-trench, isolation structure. (*a*) Trench mask patterning. (*b)* Trench etching and oxide growth. (*c*) Polysilicon deposition to fill the trench. (*d)* planarization.

Another example is the shallow trench (depth is less than 1 μm) isolation for CMOS, shown in Fig. 12.27. After patterning (Fig. 12.27a), the trench area is etched (Fig. 12.27b) and then re-filled with oxide (Fig. 12.27c). Before refilling, a chanstop implantation can be performed. Since the oxide has over filled the trench, the oxide on the nitride should be removed. Chemical-mechanical polishiig (CMP) is used to remove the oxide on the nitride and to get a flat surface (Fig. 12.27d). Due to its high resistance to polishing, the nitride acts as a stop-layer for the CMP process. After the polishing, the nitride layer and the oxide layer can be removed by H_3PO_4 and HF, respectively. This initial planarization step at the beginning is helpful for the subsequent polysilicon patterning and planarizations of the multilevel interconnection processes.

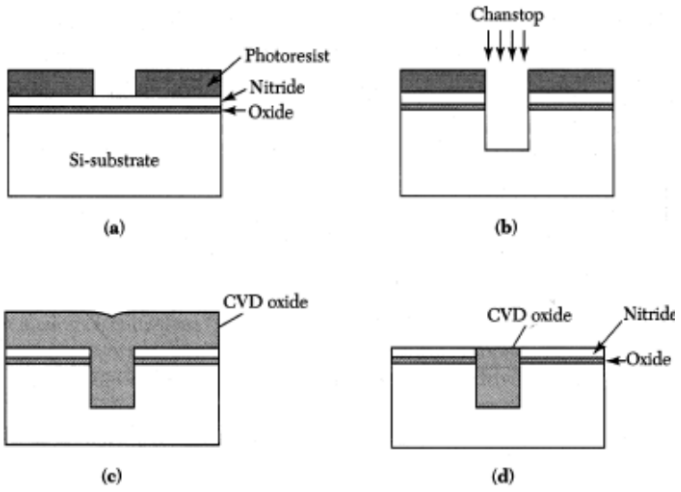


Fig. 12.27. A shallow trench isolation for CMOS. (*a*) Patterning with photoresist on nitride/oxide films. (*b*) Dry etching and chanstop implantation. (*c*) Chemical-vapor deposition (CVD) oxide to refill. (*d)* Planar surface after chemical-mechanical polishing (CMP).

Gate-Engineering Technology

If we use n^+-polysilicon for both PMOS and NMOS gates, the threshold voltage for PMOS ($V_{Tp} \cong$ -0.5 to -1.0 V) has to be adjusted by boron implantation. This makes the channel of the PMOS a buried type, shown in Fig. 12.28a. The buried-type PMOS suffers serious short-channel effects as the device size shrinks to 0.25 μm and less. The most noticeable phenomena for short-channel effects are the V_T roll-off, drain-induced barrier lowering (DIBL), and the large leakage current at the off state so that even with the gate voltage at zero, leakage current flows through source and drain. To alleviate this problem, one can change n^+-polysilicon to p^+-polysilicon for PMOS. Due to the work function difference (there is a 1.0 eV difference from n^+- to p^+-polysilicon), one can obtain a surface *p*-type channel device without the boron V_T adjustment implantation. Hence, as the technology shrinks to 0.25 μm and less, dual-gate structures are required, i.e., p^+-polysilicon gate for PMOS, and n^+-polysilicon for NMOS (Fig. 12.28b). A comparison of V_T for the surface channel and the buried channel is shown in Fig. 12.29. We note that the V_T of surface channel rolls off slowly in the deep-submicron regime compared with the buried-channel device. This makes the surface-channel device with the p^+-polysilicon suitable for deep submicron device operation.

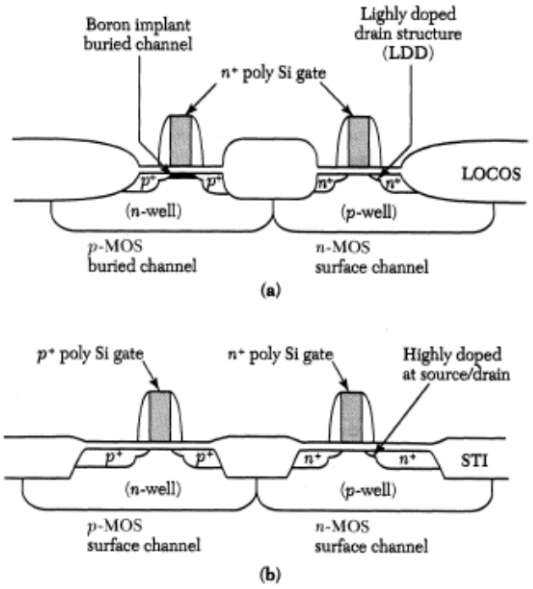


Fig. 12.28. (*a*) A conventional long-channel CMOS structure with a single-polysilicon gate *(n^+). (b)* Advanced CMOS structures with dual-polysilicon gates.

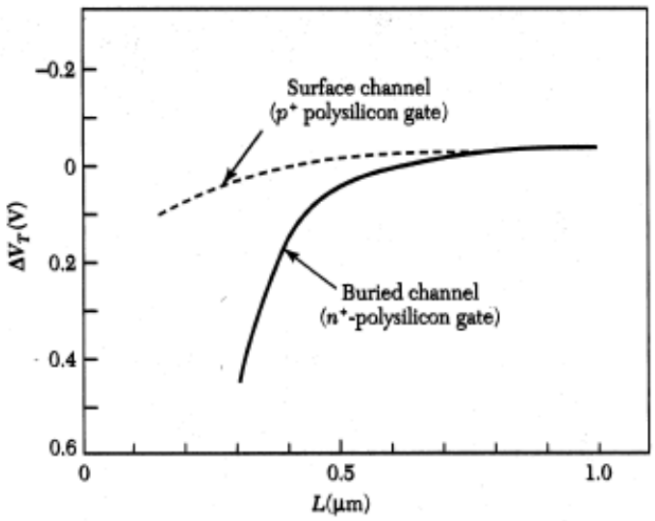


Fig. 12.29. The V_T roll-off for a buried type channel and for a surface type channel. The V_T drops very quickly as the channel length becomes less than 0.5 μm.

To form the p^+-polysilicon gate, ion implantation of BF_2^+ is commonly used. However, boron penetrates easily from the polysilicon through the oxide into the silicon substrate at high temperatures, resulting in a V_T shift. This penetration is enhanced in the presence of a F-atom. There are methods to reduce this effect: use of rapid thermal annealing to reduce the time at high temperatures and, consequently, the diffusion of boron; use of nitrided oxide to suppress the boron penetration, since boron can easily combine with nitrogen and becomes less mobile; and

the making of a multilayer of polysilicon to trap the boron atoms at the interface of the two layers.

Figure 12.30 shows a microprocessor chip (Pentium 4) that has an area of about 200 mm^2 and contains 42 million components. This ULSI chip is fabricated using 0.18 μm CMOS technology with a six-level aluminum metallization.

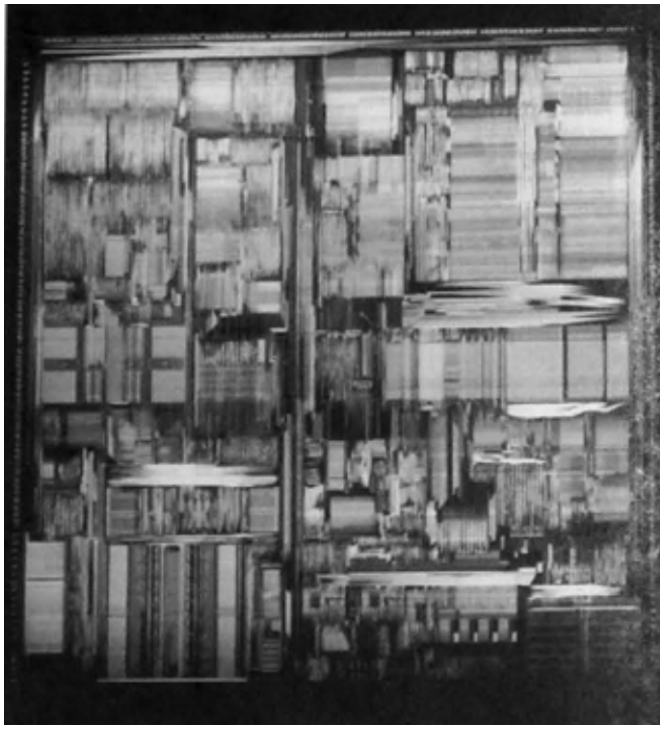


Fig. 12.30. Micrograph of a 32-bit microprocessor chip, Pentium 4.

12.4.4 BiCMOS Technology

BiCMOS is a technology that combines both CMOS and bipolar device structures in a single IC. The reason to combine these two different technologies is to create an IC chip that has the advantages of both CMOS and bipolar devices. As we know that CMOS exhibits advantages in power dissipation, noise margin, and packing density, whereas bipolar shows advantages in switching speed, current drive capability, and analog capability. As a result, for a given design rule, BiCMOS can have a higher speed than CMOS, better performance in analog circuits than CMOS, a lower power-dissipation than bipolar, and a higher component density than bipolar, Figure 12.31 shows the comparison of a BiCMOS and a CMOS logic gates. For a CMOS inverter, the current to drive (or to charging) the next loading, C_L, is the drain current I_{DS}, For a BiCMOS inverter, the current is $h_{fe}I_{DS}$, where h_{fe} is the current-gain of the bipolar transistor and I_{DS} is the base current of the bipolar transistor and is equal to the drain current of M_2 in the CMOS. Since h_{fe} is much larger than 1, the speed can be substantially enhanced.

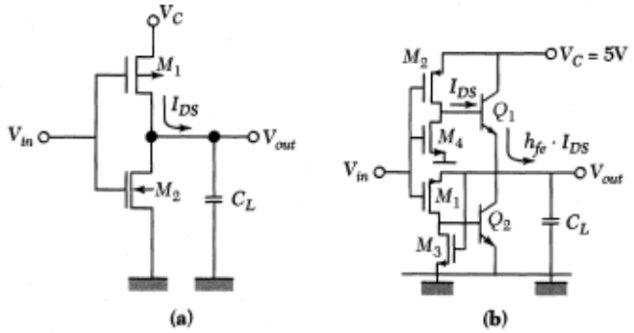


Fig. 12.31. (*a*) CMOS logic gate. (*b*) Bipolar CMOS (BiCMOS) logic gate.

BiCMOS has been widely used in many applications. In the early days, it was used in SRAM. At the present time, BiCMOS technology has been successfully developed for transceiver, amplifier, and oscillator applications in wireless-communication equipment. Most of the BiCMOS processes are based on the CMOS process, with some modifications, such as adding masks for bipolar transistor fabrication. The following example is for a high-performance BiCMOS process based on the twin-well CMOS process, shown [16] in Fig. 12.32.

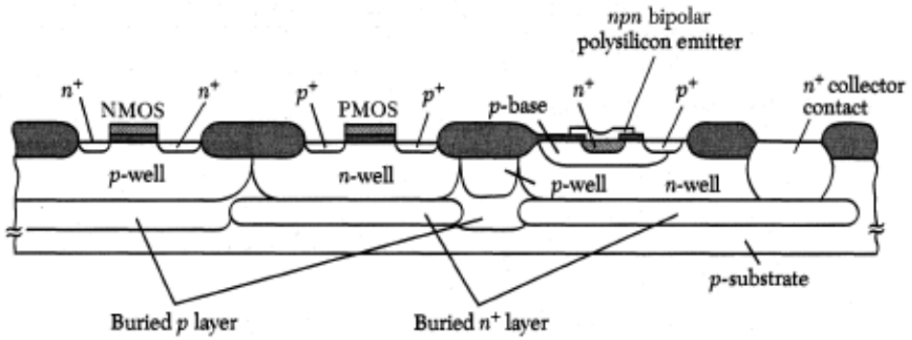


Fig. 12.32. Optimized BiCMOS device structure. Key features include self-aligned *p* and n^+ buried layers for improved packing density, separately optimized *n*- and *p*-well (twin-well CMOS) formed in an epitaxial layer with intrinsic background doping, and a polysilicon emitter for improved bipolar performance [16].

The initial material is a *p*-type silicon substrate, and then an n^+-buried layer is formed to reduce the collector's resistance. The buried *p*-layer is formed through ion implantation to increase the doping level to prevent punchthrough. **A** lightly doped *n*-epi layer is grown on the wafer and a twin-well process for the CMOS is performed. To achieve high performance of the bipolar transistor, four additional masks are needed. They are the buried n^+-mask, the collector

deep-n^+-mask, the base p-mask, and the poly-emitter mask. In other processing steps, the p^+-region for base contact can be formed with the p^+-implant in the source/drain implantation of the PMOS, and the n^+-emitter can be formed with the source/drain implantation of the NMOS. The additional masks and longer processing time compared with a standard CMOS are the main drawbacks of BiCMOS. The additional cost should be justified by the enhanced performances of BiCMOS.

12.5. MESFET Technology

Recent advances in gallium arsenide processing techniques in conjunction with new fabrication and circuit approaches have made possible the development of "silicon-like" gallium arsenide IC technology. There are three inherent advantages of gallium arsenide compared with silicon: *(i)* higher electron mobility, which results in lower series resistance for a given device geometry; *(ii)* higher drift velocity at a given electric field, which improves device speed; *(iii)* and the ability to be made semiinsulating, which can provide a latticematched dielectric-insulated substrate. However, gallium arsenide also has three disadvantages: *(i)* a very short minority-carrier lifetime; *(ii)* lack of a stable, passivating native oxide; and *(iii)* crystal defects that are many orders of magnitude higher than in silicon. The short minority-carrier lifetime and the lack of high-quality insulating films have prevented the development of bipolar devices and delayed MOS technology using gallium arsenide. Thus, the emphasis of gallium arsenide IC technology is in the MESFET area, in which our main concerns are the majority carriers transport and the metal-semiconductor contact.

A typical fabrication sequence [17] for a high-performance MESFET is shown in Fig. 12.33. A layer of GaAs is epitaxially grown on a semiinsulating GaAs substrate, followed by an n^+-contact layer (Fig. 12.33a). A mesa etch step is performed for isolation (Fig. 12.33b), and a metal layer is evaporated for the source and drain ohmic contacts (Fig. 12.33c). A channel recess etch is followed by a gate recess etch and gate evaporation (Fig. 12.33d and e). After a liftoff process that removes the photoresist, shown in Fig. 12.33e, the MESFET is completed (Fig. 12.33f).

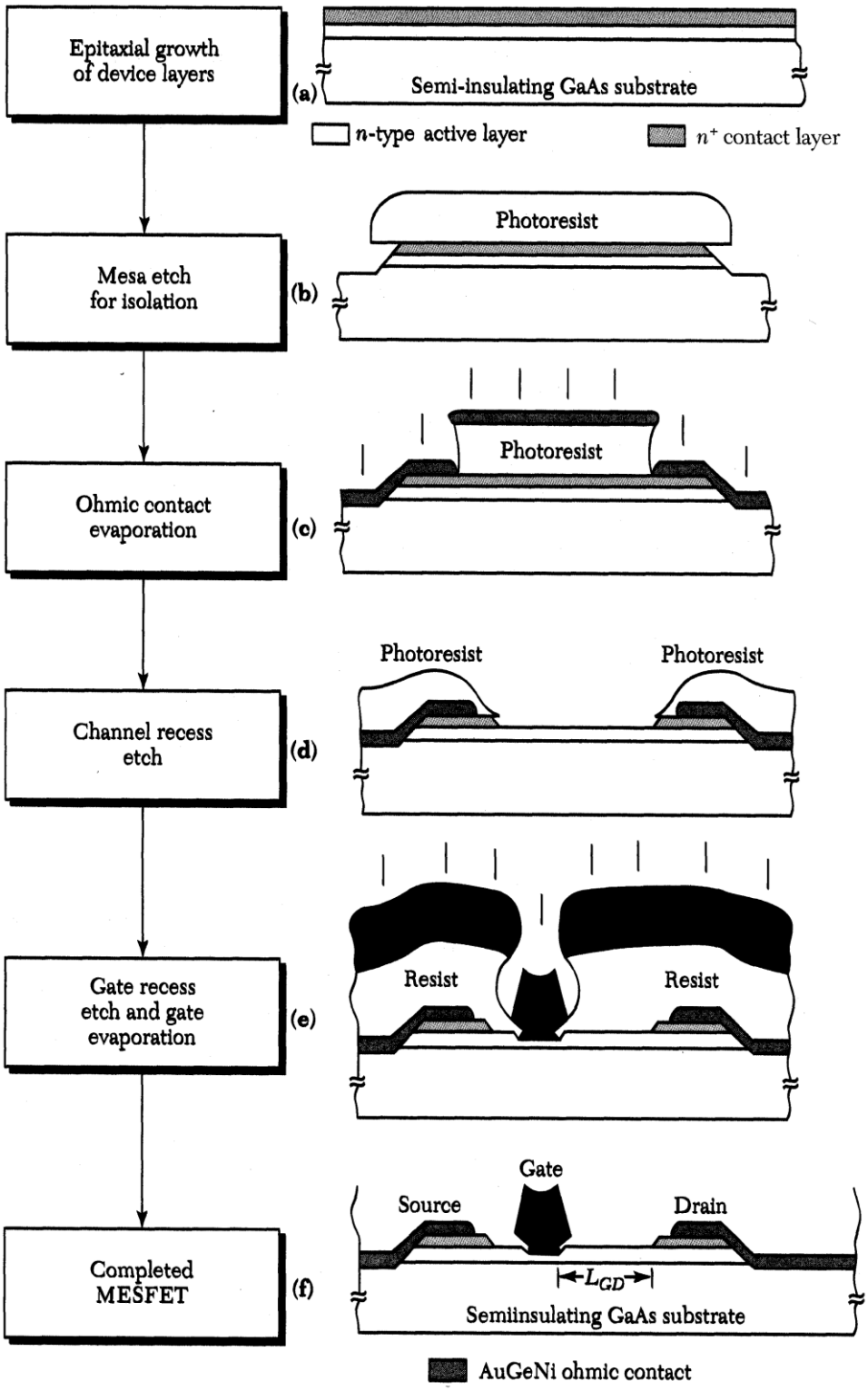


Fig. 12.33. Fabrication sequence of a GaAs MESFET [17].

The n^+-contact layer reduces the source and drain ohmic contact resistances. Note that the gate is offset toward the source to minimize the source resistance. The epitaxial layer is thick enough to minimize the effect of surface depletion on the source and drain resistance. The gate electrode has maximal cross-sectional area with a minimal foot print, which provides low gate

resistance and minimal gate length. In addition, the length L_{GD} is designed to be greater than the depletion width at gate-drain breakdown.

A representative fabrication sequence for a MESFET integrated circuit is shown [18] in Fig. 12.34. In this process, n^+-source and drain regions are self-aligned to the gate of each MESFET. A relatively light channel implant is used for the enhancement-mode switching device and a heavier implant is used for the depletion-mode load device. A gate recess is usually not used for such digital IC fabrication because the uniformity of each depth has been difficult to control, leading to an unacceptable variation of the threshold voltage. This process sequence can also be used for a monolithic microwave integrated circuit (MMIC). Note that the gallium arsenide MESFET processing technology is similar to the silicon-based MOSFET processing technology.

Gallium arsenide ICs with complexities up to the large-scale integration level (~10,000 components per chip) have been fabricated. Because of the higher drift velocity (~20% higher than silicon), gallium arsenide ICs will have a 20% higher speed than silicon ICs that use the same design rules. However, substantial improvements in crystal quality and processing technology are needed before gallium arsenide can seriously challenge the preeminent position of silicon in ULSI applications.

12.6. Challenges for Microelectronics

Since the beginning of the integrated-circuit era in 1959, the minimum device dimension, also called the minimum feature length, has been reduced at an annual rate of about 13% (i.e., a reduction of 30% every 3 years). According to the prediction by the International Technology Roadmap for Semicondutors [19], the minimum feature length will shrink from 130 nm (0.13 μm) in the year 2002 to 35 nm (0.035 μm) around 2014, as shown in Table 12.1. Also shown in Table 12.1 is the DRAM size. The DRAM has increased its memory cell capacity four times every 3 years and 64 Gbit DRAM became available in year 2011 using 50 nm design rules. The table also shows that the wafer size will increase to 450 mm (18 in. diameter) in 2014. In addition to the feature size reduction, challenges come from the device level, material level, and system level, discussed in the following subsections.

Table 12.1. The Technology Generation [19] from 1997 to 2014

Year of the first product shipment	1997	1999	2002	2005	2008	2011	2014
Feature size (nm)	250	180	130	100	70	50	35
DRAMa size (bit)	256M	1G	-	8G	-	64G	-

Wafer size (mm)	200	300	300	300	300	300	450
Gate oxide (nm)	3-4	1.9-2.5	1.3-1.7	0.9-1.1	<1.0	-	-
Junction depth (nm)	50-100	42-70	25-43	20-33	15-30	-	-

DRAM, dynamic random access memory.

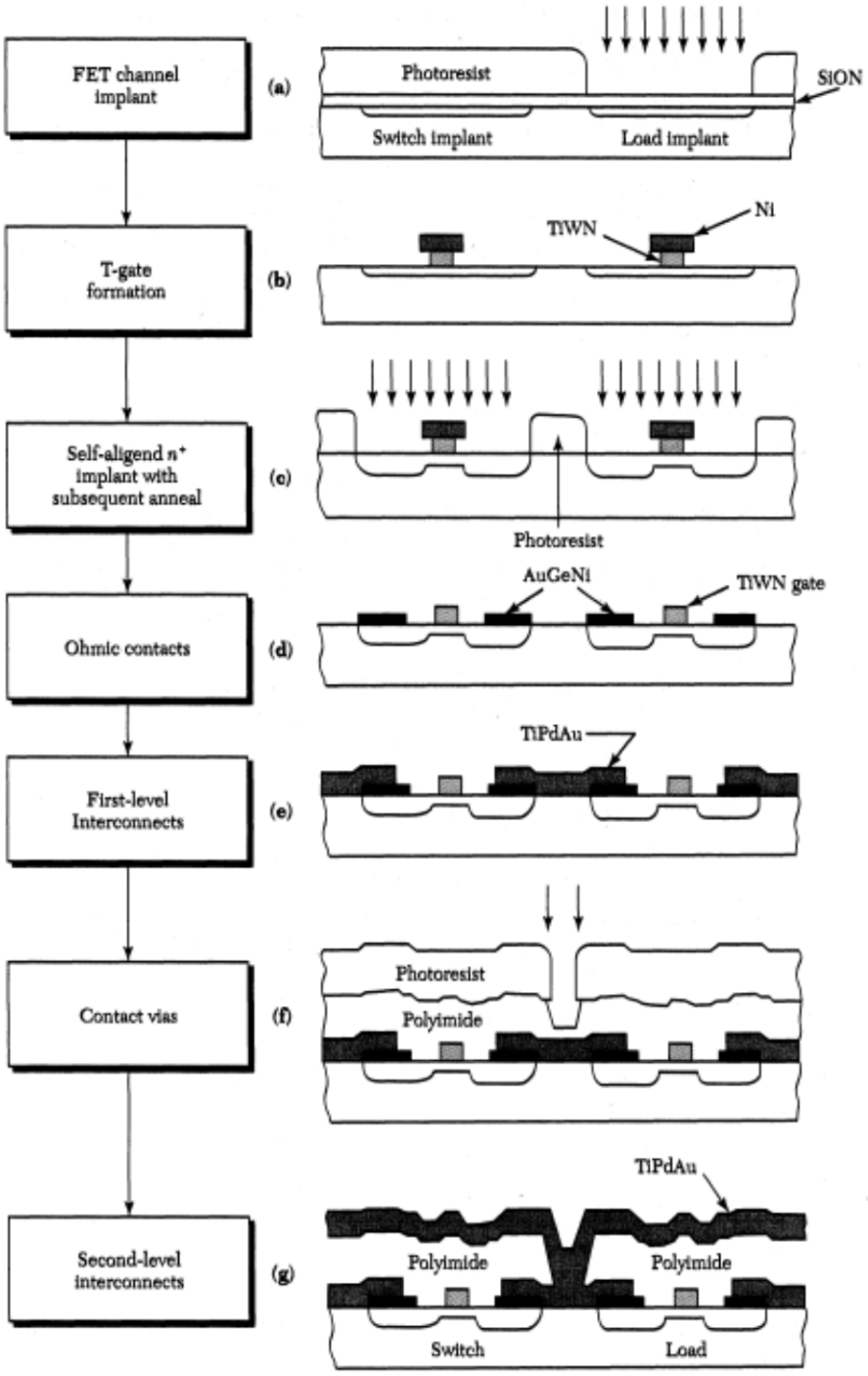


Fig. 12.34. Fabrication process for MESFET direct-coupled FET logic (DCFL) with active loads. Note that the n^+-source and drain regions are self-aligned to the gate [18].

12.6.1 Challenges for Integration

Figure 12.35 shows the trends of power supply voltage V_{DD}, threshold voltage V_T, and gate oxide thickness d versus channel length for CMOS logic technology [20]. From the figure, one can find that the gate oxide thickness will soon approach the tunneling-current limit of 2 nm. V_{DD} scaling will slow down because of nonscalable V_T (i.e., to a minimum V_T of about 0.3 V due to subthreshold leakage and circuit noise immunity). Some challenges of the 180 nm technology and beyond are shown [21] in Fig. 12.36. The most stringent requirements are as follows.

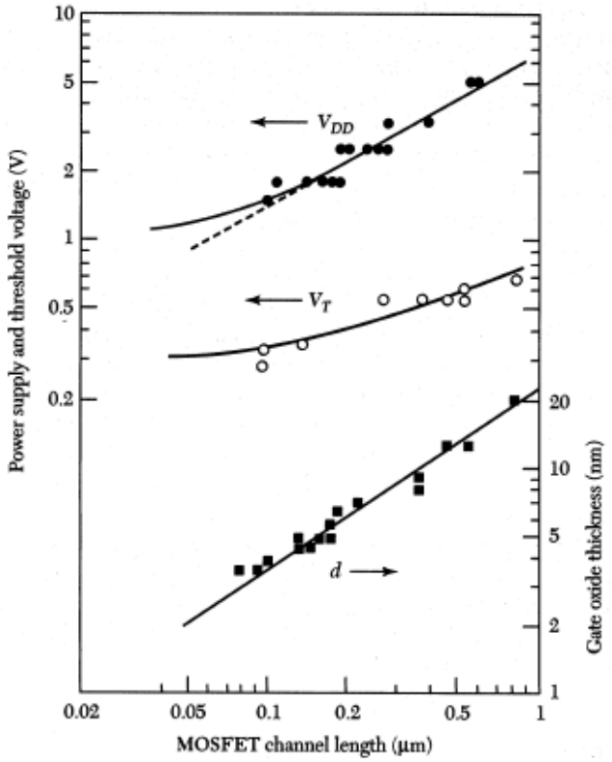


Fig. 12.35. Trends of power supply voltage V_{DD}, threshold voltage V_T, and gate oxide thickness d versus channel length for CMOS logic technologies. Points are collected from data published over recent years [20].

Ultrashallow Junction Formation

The short-channel effect happens as the channel length is reduced. This problem becomes critical as the device dimension is scaled down to 100 nm. To achieve an ultrashallow junction with low sheet resistance, low-energy (less than 1 keV) implantation technology with high dosage must be employed to reduce the shortchannel effect. Table 12.1 shows the required junction depth versus the technology generation. The requirements of the junction for 100 nm are depths around 20-33 nm with a doping concentration of $1 \times 10^{20}/\text{cm}^3$.

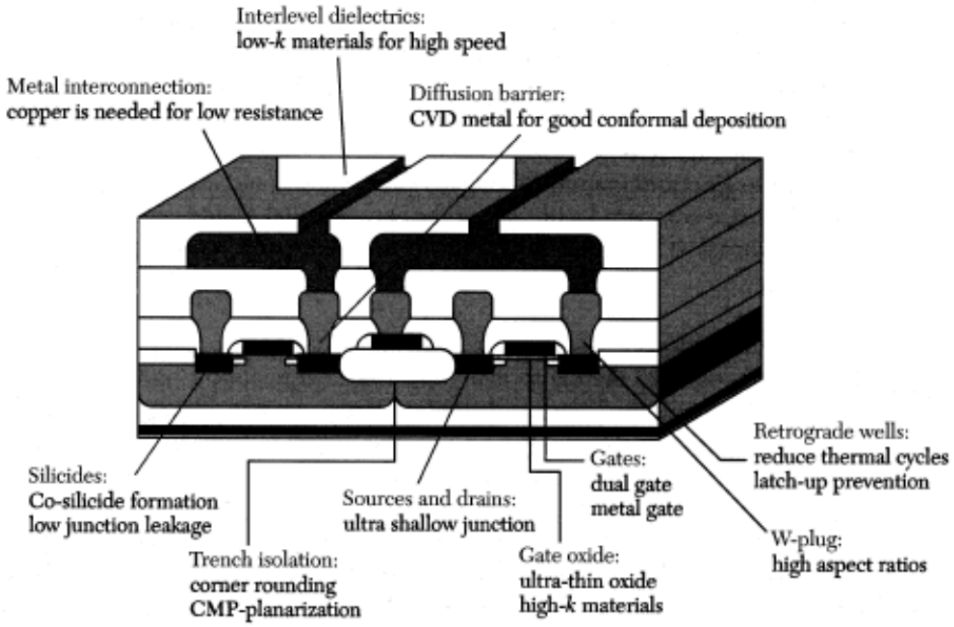


Fig. 12.36. Challenges for 180 nm and smaller MOSFET [21].

Ultrathin Oxide

As the gate length shrinks below 130 nm, the oxide equivalent thickness of gate dielectric must be reduced around 2 nm to maintain the performance. However, if only SiO_2 (with a dielectric constant of 3.9) is used, the leakage through the gate becomes very high because of direct tunneling. For this reason, thicker high-*k* dielectric materials that have lower leakage current are suggested to replace oxide. Candidates for the short term are silicon nitride (with a dielectric constant of 7), Ta_2O_5 (25), and TiO_2 (60-100).

Silicide Formation

Silicide-related technology has become an integral part of submicron devices for reducing the parasitic resistance to improve device and circuit performance. The conventional Ti-silicide process has been widely use in 350-250 nm technology. However, the sheet resistance of a $TiSi_2$ line increases with decreasing line width, which limits the use of $TiSi_2$ in 100 nm CMOS applications and beyond. $CoSi_2$ or NiSi processes will replace $TiSi_2$ in the technology beyond 100 nm.

New Materials for Interconnection

To achieve high-speed operation, the RC time delay of the interconnection must be reduced [22]. It is obvious that the gate delay decreases as the channel length decreases, meanwhile the delay resulting from interconnect increases significantly as the size decreases. This causes the total delay time to increase as the dimension of the device size scales down to

250 nm. Consequently, both high-conductivity metals, such as Cu, and low-dielectric constant (low-*k*) insulators, such as organic (polyimide) or inorganic (F-doped oxide) materials offer major performance gains. Cu exhibits superior performance because of its high conductivity (1.7 μΩ-cm compared with 2.7 μΩ-cm of Al) and is 10-100 times more resistant to electromigration. The delay using the Cu and low-*k* material shows a significant decrease compared with that of the conventional A1 and oxide. Hence, Cu with the low-*k* material is essential in multilevel interconnection for future deep-submicron technology.

Power Limitations

The power required merely to charge and discharge circuit nodes in an IC is proportional to the number of gates and the frequency at which they are switched (clock frequency). The power can be expressed as $P \cong 1/2CV^2nf$, when C is the capacitance per device, V is the applied voltage, n is the number of devices per chip, and f is the clock frequency The temperature rise caused by this power dissipation in an IC package is limited by the thermal conductivity of the package material, unless auxiliary liquid or gas cooling is used. The maximum allowable temperature rise is limited by the bandgap of the semiconductor (~100°C for Si with a bandgap of 1.1 eV). For such a temperature rise, the maximum power dissipation of a typical high-performance package is about 10 W. As a result, we must limit either the maximum clock rate or the number of gates on a chip. As an example, in an IC containing 100 nm MOS devices with $C = 5 \times 10^{-2}$ fF, running at a 20 GHz clock rate, the maximum number of gates we can have is about 10^7 if we assume a 10% duty cycle. This is a design constraint fixed by basic material parameters.

SOI Integration

SOI wafer can be used as the isolation. Recently SOI technology has received more attention. The advantages of the SOI integration become significant as the minimum feature length approaches 100 nm. From the process point of view, SOI does not need the complex well structure and isolation processes. In addition, shallow junctions are directly obtained through the SOI film thickness. There is no risk of nonuniform interdiffusion of silicon and Al in the contact regions because of oxide isolation at the bottom of the junction. Hence, the contact barrier is not necessary. From the device point of view, the modern bulk silicon device needs high doping at the drain and substrate to eliminate short-channel effects and punch-through. This high doping results in high capacitance when the junction is reversed bias. On the other hand, in SOI, the maximum capacitance between the junction and substrate is the capacitance of the buried insulator whose dielectric constant is three times smaller than that of silicon (3.9 versus 11.9),

Based on the ring oscillator performance, the 130 nm SOI CMOS technology can achieve 25% faster speed or require 50% less power compared to a similar bulk technology [23]. SRAM, DRAM, CPU, and rf CMOS have all been successfully fabricated using SOI technology. Therefore SOI is a key candidate for the future system-on-a-chip technology.

12.6.2. System-on-a-Chip

The increased component density and improved fabrication technology have helped the realization of the system-on-a-chip (SOC), that is, an IC chip that contains a complete electronic system. The designers can build all the circuitry needed for a complete electronic system, such as a camera, radio, television, or personal computer (PC), on a single chip. Figure 12.37 shows the SOC application in the PC's mother-board. Components (11 chips in this case) once found on boards are becoming virtual components on the chip at the right [24].

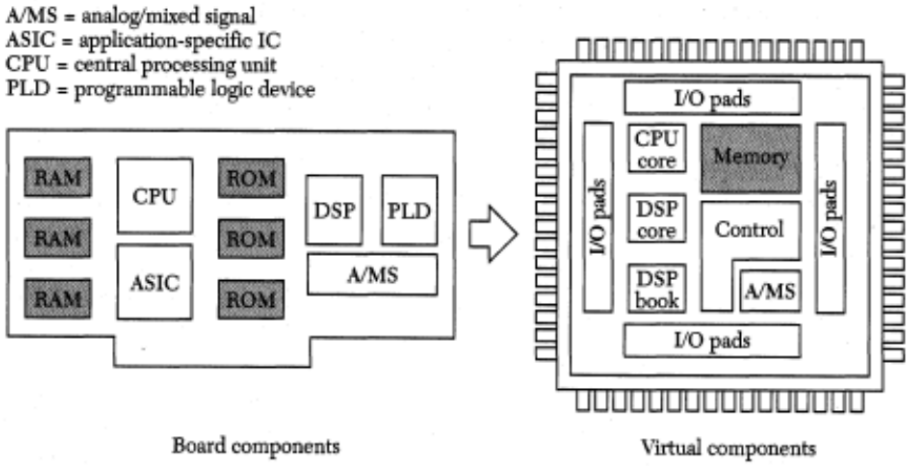


Fig. 12.37. System-on-a-chip of a conventional personal computer mother-board [24].

There are two obstacles in the realization of the SOC. The first is the huge complexity of the design. Since the component board is presently designed by different companies and different design tools, it is difficult to integrate them into one chip. The other is the difficulty of fabrication. In general, the fabricating processes of the DRAM are significantly different from those of logic IC (e.g., CPU). Speed is the first priority for the logic, whereas leakage of the stored charge is the priority for memory. Therefore, multilevel interconnection using five to six levels of metals is essential for logic IC to improve the speed. However, DRAM needs only two to three levels. In addition, to increase the speed, a silicide process must be used to reduce the series resistance, and ultrathin gate oxide is needed to increase the drive current. These requirements are not critical for the memory.

To achieve the SOC goal, an embedded DRAM technology is introduced, i.e., to merge logic and DRAM into a single chip with compatible processes. Figure 12.38 shows the schematic cross section of the embedded DRAM, including the DRAM cells and the logic CMOS devices [25]. Some processing steps are modified as a compromise. The trench-type capacitor, instead of the stacked type, is used so that there is no height difference in the DRAM cell structure. In addition, multiple gate oxide thicknesses exist on the same wafer to accommodate multiple supply voltages and/or combine memory and logic circuits on one chip.

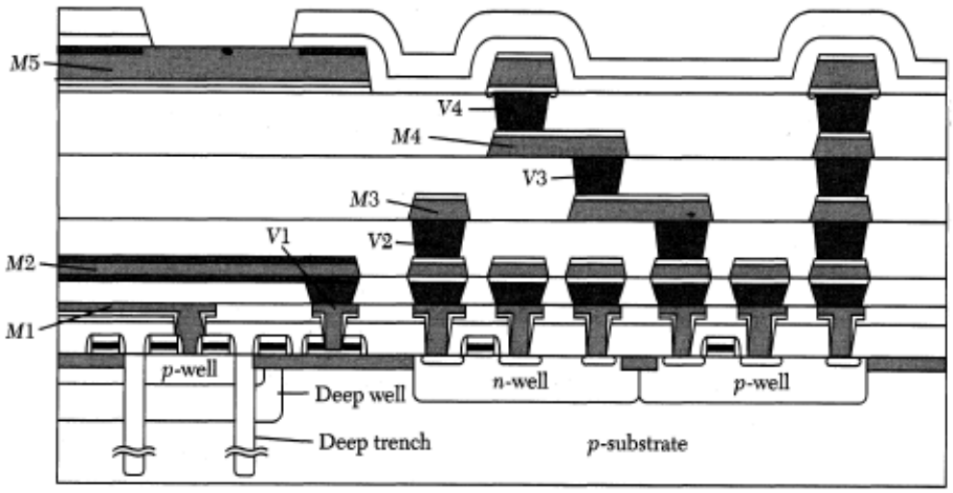


Fig. 12.38. Schematic cross section of the embedded DRAM including DRAM cells and logic MOSFETs. There is no height difference in the trench capacitor cell because of the DRAM cell structure. *M1* to *M5* are metal interconnections, and *V1* to *V4* are via holes [25].

12.7. Summary

In this chapter we considered processing technologies for passive components, active devices, and IC. Three major IC technologies based on the bipolar transistor, the MOSFET, and the MESFET were discussed in detail. It appears that the MOSFET will be the dominant technology at least until 2014 because of its superior performance compared with the bipolar transistor. For 100 nm CMOS technology, a good candidate is the combination of an SOI-substrate with interconnections using Cu and low-*k* materials.

Because the rapid reduction in feature length, the technology will soon reach its practical limit as the channel length is reduced to about 20 nm. What will be the device beyond the CMOS is the question being asked by research scientists. Major candidates include many innovative devices based on quantum mechanical effects. This is because when the lateral dimension is reduced to below 100 nm, depending on the materials and the temperature of operation, electronic structures will exhibit nonclassical behaviors. The operation of such devices will be on

the scale of single-electron transport. This approach has been demonstrated by the single-electron memory cell. The realization of such systems with trillions of components will be a major challenge beyond CMOS [26].

Chapter 13. Basic MEMS and NEMS technologies. Micromashining

V.Skryshevsky

13.1 Introduction

The microelectromechanical and nanoelectromechanical systems (MEMS and NEMS), are the batch-fabricated integrated micro (nano) scale system (motion, electromagnetic, radiating energy and optical microdevices/microstructures – driving/sensing circuitry – controlling/processing ICs) that:

1. Converts physical actions, events, and parameters to electrical, mechanical, and optical signals and vice versa;
2. Performs actuation, sensing, and other functions;
3. Comprises control (intelligence, decision-making, evolutionary learning, adaptation, self-organization, etc.), diagnostics, signal processing, and data acquisition features, and microscale features of electromechanical, electronic, optical, and biological components (structures, devices, and subsystems), architectures, and operating principles are basics of the MEMS operation, design, analysis, and fabrication. Figure 13.1 illustrates the functional block-diagram of MEMS.

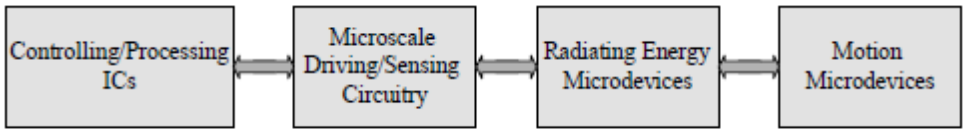


Fig.13.1 The functional block-diagram of MEMS

Micromachining is the basic technology for fabrication of MEMS and NEMS. Today micromachining is used for fabricating micro-channels and micro-grooves in micro-fluidics applications, micro-filters, drug delivery systems, micro-needles, and micro-probes in biotechnology applications. Micromachined components are crucial for practical advancement in MEMS, Microelectronics (semiconductor devices and integrated circuit technology) and Nanotechnology. The main reason of micromachining development is the *miniaturization of devices*, both electrical and mechanical systems. Engineering criteria, such as cost, performance, and system integration provide incentives to miniaturize. Miniaturized components allow complete systems to be created in a single package. One important benefit of system integration is avoiding the assembly of discrete components in the final device, thus *lowering cost.* Additionally, batch fabrication methods can drastically reduce the cost of a system's components.

For example, while manufacturers may spend quite a lot of money processing a wafer into microelectronics systems, that wafer can contain billions of transistors, making the individual transistors extremely inexpensive.

Miniature devices can have *increased reliability.* The mass decreases faster than structural strength. Thus miniature components can withstand higher drops and larger vibrations. System integration also increases reliability. The interconnections between discrete components can also fail, so monolithically manufactured systems typically exhibit greater reliability.

Miniature devices can have *faster response times.* Smaller devices obviously have less inertia, less thermal mass, less capacitance, etc. For many types of devices, these parameters play important roles in the device's response time. System integration, by reducing the size of interconnections, both electrical and mechanical, reduces the amount of parasitic loads on the system. Thus system integration can also improve the response times. Finally, smaller devices can simply *fit more places.*

For the last decades, micromachining has been closely associated with microelectronics fabrication processes. The use of traditional semiconductor manufacturing techniques, such as photolithography, thin-film deposition, and etching, provide a novel approach to the shaping and processing of materials to produce functional mechanical devices. Early micromachining research and development was closely tied to existing techniques provided by the already established microelectronics industry.

However, micromachining is increasingly moving to fabrication techniques developed specifically for micromachining. These include new methods of depositing and etching films, expanding the range of film thicknesses possible. Modern micromachining is also making use of materials not found in the microelectronics industry.

Several main micromachining processes exist. They all have advantages and disadvantages. The three main technologies are *bulk micromachining, LIGA, and surface micromachining.*

Bulk micromachining refers to the process of selectively etching away a portion of the substrate to form free standing structures bound by a cavity.

LIGA uses thick layers of photoresist to serve as a mould for electroplated parts.

Surface-micromachining builds structures out of multiple thin-films, each of which are patterned using photolithography. A release step is used to remove some of the thin films.

In the bulk machining, the materials with the dimensions of more than in the range of micrometer or above centimetre scale are being used for the machining process. This process can be applicable to produce 3D MEMS structures. This also uses anisotropic etching of single crystal silicon. For example, silicon cantilever beam for atomic force microscope (AFM).

Surface micromachining is another new technique/process for producing MEMS structures. This uses etching techniques to pattern microscale structures from polycrystalline (poly) silicon, or metal alloys. Example: accelerometers, pressure sensors, micro gears and transmission, and micro mirrors etc.

The surface micromachining has also advantageous because it is so closely related to manufacturing processes used in microelectronics. Surface micromachining benefits from lots of cross-over expertise in thin film deposition and patterning. It is also capable of creating devices with smaller features than either bulk micromachining or LIGA. One disadvantage of surface micromachining is that all devices are made from thin films. Perpendicular to the wafer surface, conventional surface micromachined devices can be too small for the desired application. It is difficult to build thicker structures, and several important technologies have been created to fill this need. Often, it is difficult to build structures large enough. In these cases, bulk micromachining or LIGA may be more suitable.

Micromachining has evolved greatly in the past few decades, to include various techniques, broadly classified into mask based and tool based, as depicted in Fig.13.2. The size evolution of MEMS devices with appropriate technology methods is presented in Fig.13.3. It can classified machining in three divisions: normal, precision, and ultraprecision. The term micromachining is now associated with the qualities of precision and ultraprecision.

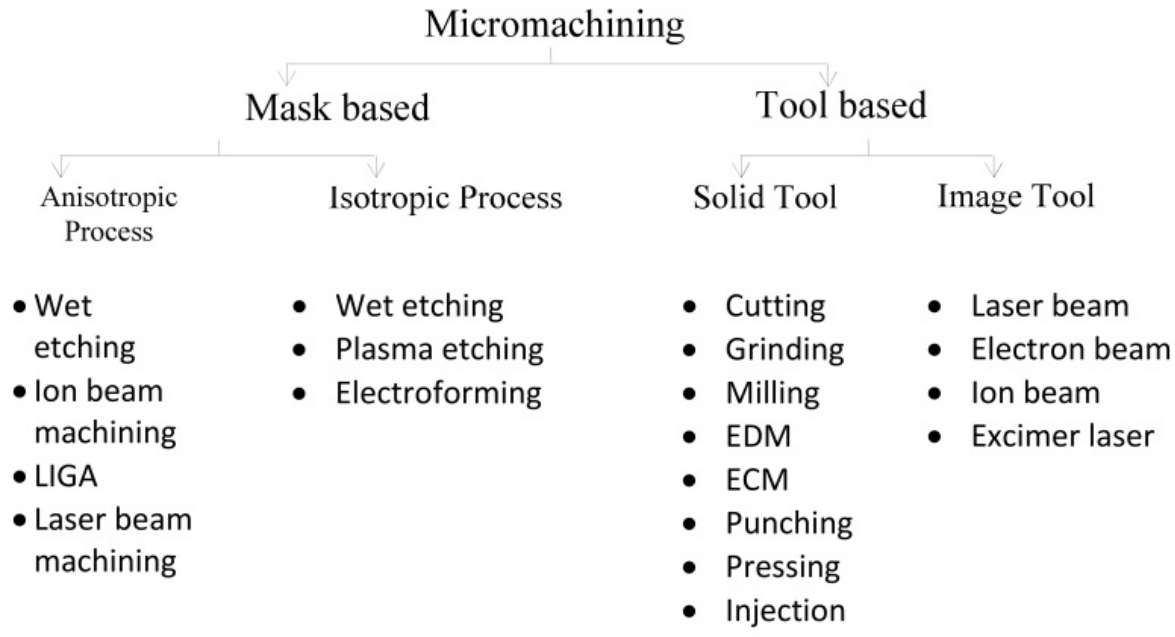


Fig.13.2 The method of micromashining.

As can be seen from fig.2 besides mentioned basic methods the other methods are developed too. Lithography or etching methods are not capable of making true 3D structures

e.g. free form surfaces and also limited in range of materials. Mechanical machining is capable of making free form surfaces in wide range of materials. There are two approaches used to machine micron and sub-micron scale features.

1. Design ultra precision (nanometer positioning resolution) machine tools and cutting tools. For this, ultra precision diamond turning machines can be used.

2. Design miniature but precise machine tools. Example: Micro-lathe, micro-mill, micro-EDM, etc.

Mechanical micromachining process:

-Can produce extremely smooth, precise, high resolution true 3D structures

-Expensive, non-parallel, but handles much larger substrates

-□Precision cutting on lathes produces miniature screws, etc with 12 μm accuracy

-□Relative tolerances are typically 1/10 to 1/1000 of feature

-□Absolute tolerances are typically similar to those for conventional precision machining (Micrometer to sub-micrometer).

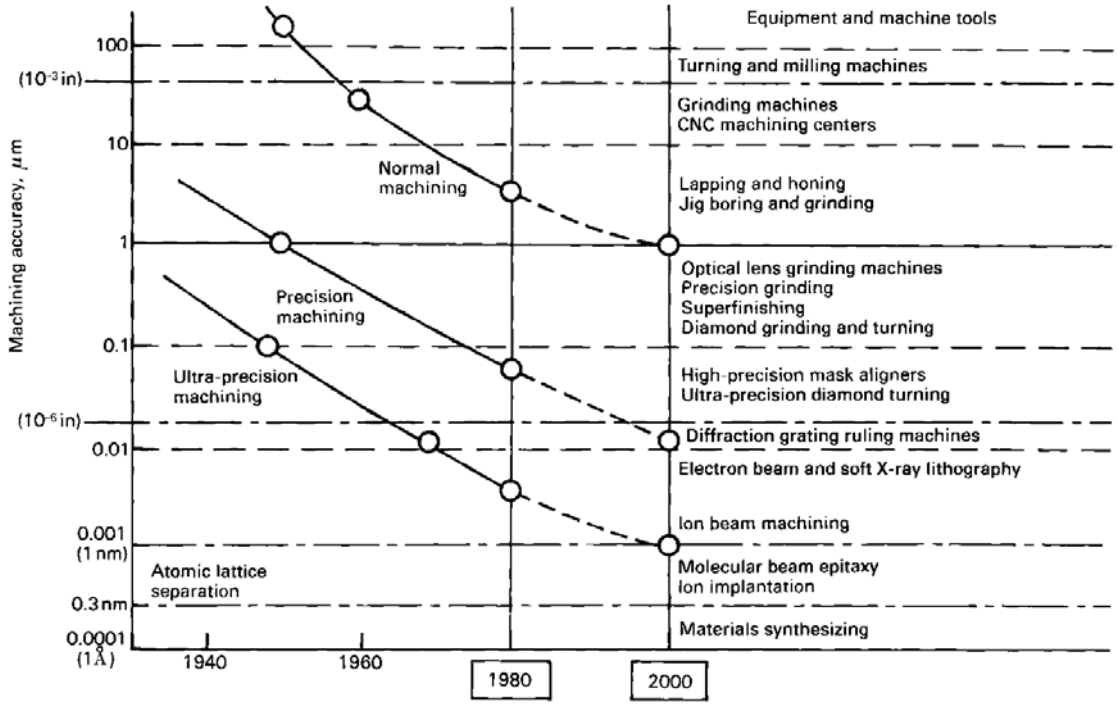


Figure 13.3 MEMs technology evolutions.

13.2 Bulk Micromachining

Bulk micromachining was developed more than thirty years ago to fabricate three dimensional microstructures. Bulk micromachining of silicon uses wet and dry etching

techniques in conjunction with etch masks and etch stop layers and wafer-to-wafer bonding, to develop microstructures from silicon substrates.

13.2.1 Isotropic and Anisotropic Etching

Wet etching is the process of removing material by immersing the wafer in a liquid bath of the chemical etchant. Wet etchants are categorized as *isotropic* etchants (attack the material being etched at the same rate in all directions) and *anisotropic* etchants (attack the material or silicon wafer at different rates in different directions, and therefore, shapes/geometry can be precisely controlled). In other words, the *isotropic* etching has a uniform etch rate at all orientations, while for *anisotropic* etching, the etch rate depends on crystal orientation. 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 silicon, silicon oxide, silicon nitride, polysilicon, gold, aluminum, and other commonly used materials. 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. The hydrofluoric acid etches the silicon oxide faster than the silicon. *Isotropic* etching in liquid reagents is the most widely used process for removal of damaged surfaces, creating structures in single-crystal slices, and patterning single crystal or polycrystalline semiconductor films. For *isotropic* etching of silicon, the most commonly used etchants are mixtures of hydrofluoric (HF) and nitric (HNO_3) acids in water or acetic acid (CH_3COOH). In this co-called HNA etchant system, after the hole injection and OH- attachment to the silicon to form $Si(OH)_2$, hydrogen is released to form SiO_2. Hydrofluoric acid is used to dissolve SiO_2 to form water soluble H_2SiF_6. The reaction is

$$Si + HNO_3 + 6HF \rightarrow H_2SiF_6 + H_2NO_2 + H_2O + H_2.$$

Water can be used as a diluent for this etchant. However, acetic acid CH_3COOH is preferred because it controls the dissociation of the nitric acid and preserves the oxidizing power of HNO_3 for a wide range of dilution (i.e., it acts as a buffer). Thus, the oxidizing power of the etchant remains almost constant. Tables 13.1 and 13.2 show the typical wet and dry etchants and etch rate.

Table 13.1. Wet etchant

Material	Etchant and Etch Rate
Polysilicon	6 ml HF, 100 ml HNO_3, 40 ml H_2O, 8000 Å/min, smooth edges 1 ml HF, 26 ml HNO_3, 33 ml CH_3COOH, 1500 Å/min
Phosphorous-doped silicon dioxide (PSG)	Buffered hydrofluoric acid (BHF) 28 ml HF, 170 ml H_2O, and 113 g NH_4F, 5000 Å/min 1 ml BHF and 7 ml H_2O, 800 Å/min
Silicon nitride (Si_3N_4)	Hydrofluoric acid (HF) 140 Å/min CVD at 1100°C 750 Å/min CVD at 900°C 1000 Å/min, CVD at 800°C
Silicon dioxide (SiO_2)	Buffered hydrofluoric acid (BHF) 28 ml HF, 170 ml H_2O, and 113 g NH_4F, 1000-2500 Å/min 1 ml BHF and 7 ml H_2O, 700-900 Å/min
Aluminum (Al)	4 ml H_3PO_4, 1 ml HNO_3, 4 ml CH_3COOH, 1 ml H_2O, 350 Å/min 16-19 ml H_3PO_4, 1 ml HNO_3, 0-4 ml H_2O, 1500-2400 Å/min
Gold (Au)	3 ml HCl, 1 ml HNO_3, 25-50 µm/min. 4 g KI, 1 g I2, 40 ml H_2O, 0.5-1 µm/min
Chromium (Cr)	1 ml HCl, 1 ml glycerine, 800 Å/min, (need depassivation) 1 ml HCl, 9 ml saturated $CeSO_4$ solution, 800 Å/min (need depassivation) 1 ml (1 g NaOH in 2 ml H_2O), 3 ml (1 g $K_3Fe(CN)_6$ in 3 ml H_2O), 250-100 Å/min (photoresist mask)
Tungsten (W)	34 g KH_2PO_4, 13.4 g KOH, 33 g $K_3Fe(CN)_6$, and H_2O to make 1 liter, 1600 Å/min (photoresist mask)

Table 13.2 . Dry etchant

Material	Etchant (Gas) and Etch Rate
Silicon dioxide (SiO_2) Phosphorous-doped silicon dioxide (PSG)	$CF_4 + H_2$, C_2F_6, C_3F_8, or CHF_3, 500-800 Å/min
Silicon (single-crystal and polycrystalline)	$SF_6 + Cl_2$, 1000-5000 Å/min CF_4, CF_4O_2, CF_3Cl, SF_6Cl, Cl_2+H_2, $C_2ClF_5O_2$, SF_6O_2, SiF_4O_2, NF_3, $C_2Cl_3F_5$, or CCl_4He
Silicon nitride (Si_3N_4)	CF_4O_2, CF_4+H_2, C_2F_6, or C_3F_8, SF_6He
Polysilicon	Cl_2, 500-900 Å/min
Aluminum (Al)	BCl_3, CCl_4, $SiCl_4$, BCl_3Cl_2, CCl_4Cl_2, or $SiCl_4Cl_2$
Gold (Au)	$C_2Cl_2F_4$ or Cl_2
Tungsten (W)	CF_4, CF_4O_2, C_2F_6, or SF_6
Al, Al-Si, Al-Cu	$BCl_3 + Cl_2$, 500 Å/min

The examples of Bulk Micromachining with *Isotropic* etching are presented in Fig.13.4

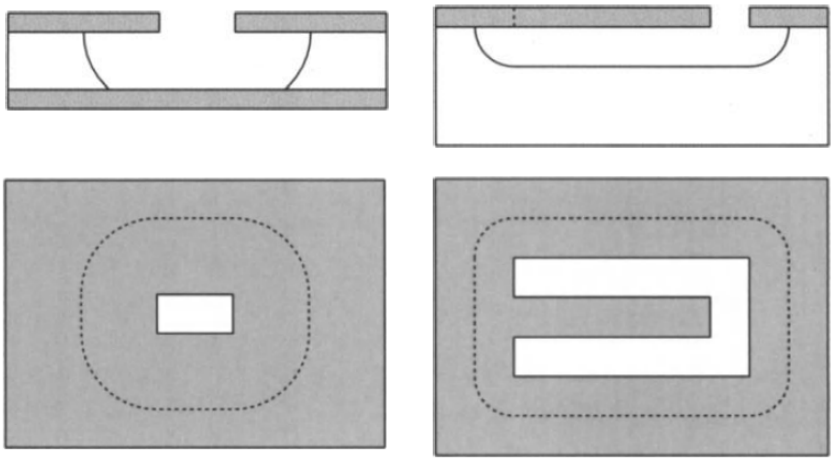


Figure 13.4. The design for a membrane and cantilever micromashined with an isotropic etchant.

There are three important anisotropic etchants for silicon: ethylene diamine, pyrocatechol, and water (EDP); KOH and water; HF, HNO_3, and acetic acid (HNA). The most popular *anisotropic* etchant is potassium hydroxide (KOH) because it is the safest one to use. There are other anisotropic etchants, and etchant chemistries can become quite complex. That is, they etch the different crystal orientations with different etch rates. Anisotropic etchants etch the (100) and (110) silicon crystal planes faster than the (111) crystal planes. For example, the etch rates are 500:1 for (100) versus (111) orientations, respectively.

Silicon dioxide, silicon nitride, and metallic thin films (chromium and gold) provide good etch masks for typical silicon anisotropic etchants. These films are used to mask areas of silicon that must be protected from etching and to define the initial geometry of the regions to be etched. Using an anisotropic etchant for bulk micromachining allows crystallographic planes to be used as *etch stops* to control the shape of the etched regions. To make proper use of the crystallographic planes, the crystal orientation has to be fixed before etching begins. Fortunately, silicon wafers come in three common orientations: (100) oriented silicon, (110) oriented silicon, and (111) oriented silicon. For all three orientations, the listed crystallographic direction is perpendicular to the substrate surface. In (100) oriented silicon, anisotropic etching typically leads to V –shaped grooves and pyramids, since these shapes are bounded by the (111) planes. The (111) planes are at an angle of 54.7° to the surface. This leads to profiles as shown in fig. 13.5.

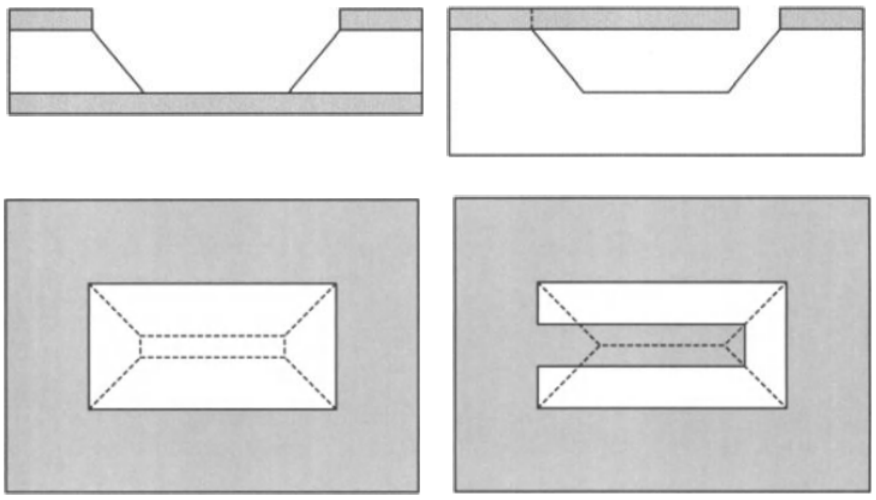


Figure 13.5. The design for a membrane and cantilever micromashined with an anisotropic etchant. Substrate has (100) orientation.

In (110) oriented silicon, the (111) planes lie perpendicular to the wafer surface. This allows the etching of square structures into the substrate. However, the anisotropic etchant will not undercut a thin-film on the top surface, so it cannot be used to release cantilevers, unless the structure is etched from the back side of the wafer.

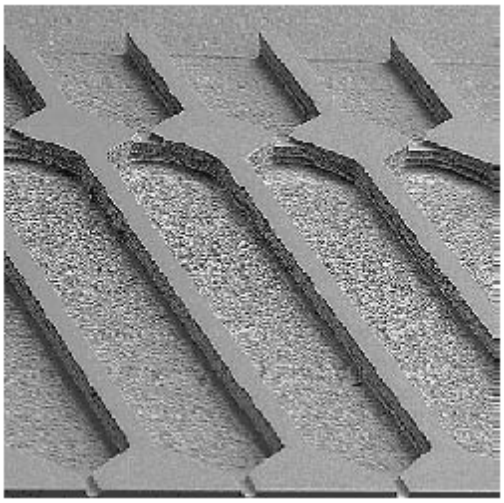


Figure 13.6 Reactive ion etching of silicon

The widely implemented dry etching process in micromachining applications is *reactive ion etching*. In this process, ions are accelerated towards the material to be etched, and the etching reaction is enhanced in the direction of travel the ions. Reactive ion etching is an *anisotropic* etching process. Deep trenches and pits (up to a few tens of microns) of the specified shape with vertical walls can be etched in a variety of commonly used materials, e.g., silicon,

polysilicon, silicon oxide, and silicon nitride. Compared with the *anisotropic* wet etching, dry etching is not limited by the crystal planes in the silicon. Figure 13.6 illustrates the *anisotropic* etched 400 μm deep and 20 μm width grooves (in 110-silicon), and three-dimensional silicon structure are made using reactive ion etching.

In addition to etching, bulk micromachining often makes use of wafer bonding. Wafer bonding refers to a process where two wafers are permanently joined. By etching multiple wafers individually, and then bonding them together, more complicated devices can be fabricated. The two most common methods of wafer bonding are anodic bonding and silicon fusion bonding.

13.2.2 Etch Stops

In addition to crystallographie direction, material type is another method of controlling etching (fig.13.7). The etch rate is heavily dependent on material. This is key to surface micromachining, as etchants selective to the sacrificial material must be found. When used to control the extent of etching in bulk micromachining, material changes are often referred to as *etch stops*. Two techniques have been widely used in conjunction with silicon *anisotropic* etching to guarantee the etch-stop. Heavily-boron-doped silicon (so-called p^+ *etch-stop*) is effective in stopping the etch. The *pn-junction* technique can be used to stop etching when one side of a reverse-biased junction-diode is etched away.

(a) p-type silicon (epi or implanted) over n-type silicon

(b) silicon dioxide over silicon

Figure.13.7. Bulk micromachined membrane using etch stops.

Silicon-on-insulator (SOI) Wafers: SOI wafers are made up of at least two different materials. Typically, the manufacturer starts with a substrate and epitaxially grows a few layers on one side of the wafer. This leads to substrates that comprise mostly, for example, silicon, with a few thin-films on top. However, the quality and dimensional control of the thin films is very high. This means that structures, etched using the thickness of the thin films as a control can be also be fabricated with high dimensional control.

Doping: In addition to pure material changes, such as the difference between silicon and silicon dioxide, some etchants are sensitive to changes in doping. A good example of this is Boron doped silicon, which at high concentrations ($> 5 \times l0^{18} cm^{-3}$) stops several etchants.

Electrochemical: Applying a voltage across a pn-junction will prevent some etchants from etching the n-type material. The above techniques can be used to create etch stops parallel to the wafer surface. This makes them useful manufacturing membranes, as illustrated in figure 7.

13.3 Surface Micromachining

13.3.1. The base of Surface Micromachining

Surface micromachining technology allows one to fabricate the structure as layers of thin films. Surface micromachining with single crystal silicon, polysilicon, silicon nitride, silicon oxide, and silicon dioxide (as structural and sacrificial materials which are deposited and etched), as well as metals and alloys, is widely used to fabricate thin micromechanical structures and devices on the surface of a silicon wafer. This technology guarantees the fabrication of three-dimensional microdevices with high accuracy, and the surface micromachining can be called a thin film technology. Each thin film is usually limited to thickness up to 5 μm which leads to fabrication of high performance planar-type microscale structures and device. This affordable low-cost high-yield technology is integrated with electromechanical microstructures – ICs fabrication processes guaranteeing the needed microstructures-IC fabrication compatibility.

Surface micromachining is based on the application of *sacrificial (temporary) layers* that are used to maintain subsequent layers and are removed to reveal (release) fabricated microstructures. This technology was first demonstrated for ICs, and applied to fabricate motion microstructures in the 1980s. On the surface of a silicon wafer, thin layers of structural and sacrificial materials are deposited and patterned. Then, the sacrificial material is removed, and microelectromechanical structure or device is fabricated. Figure 13.8 illustrates a typical process sequence of the surface micromachining fabrication technology.

Usually, the sacrificial layer is made using silicon dioxide (SiO_2), phosphorous-doped silicon dioxide (PSG), or silicon nitride (Si_3N_4). The structural layers are then typically formed with polysilicon, metals, and alloys. The sacrificial layer is removed. In particular, after fabrication of the surface microstructures and microdevices (micromachines), the silicon wafer can be wet bulk etched to form cavities below the surface components, which allows a wider range of desired motion for the device. The wet etching can be done using: Hydrofluoric acid (HF), Buffered hydrofluoric acid (BHF), Potassium hydroxide (KOH), Ethylene-diamene-pyrocatecol (EDP), Tetramethylammonium hydroxide (TMAH), Sodium hydroxide (NaOH).

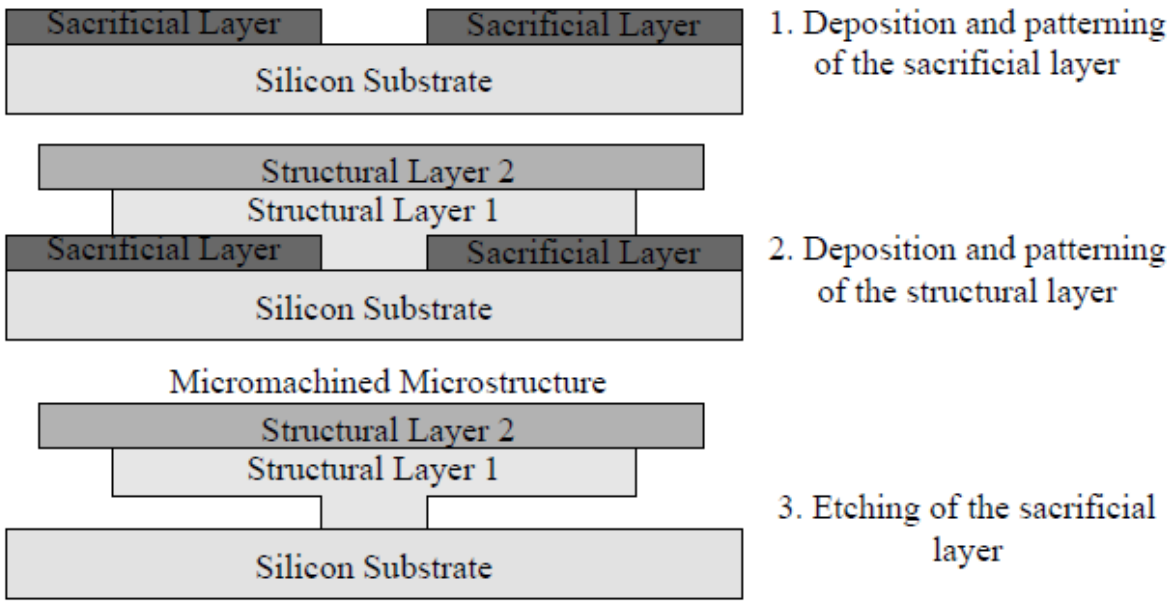


Figure 13.8 Surface micromachining.

A simple surface micromachined cantilever beam is shown in Figure 13.9. A sacrificial layer of silicon oxide is deposited on the surface of the silicon wafer. Two layers of polysilicon and ferromagnetic alloy are then deposited and patterned using dry etching. The wafer is wet etched to remove the silicon oxide layer under the beam releasing the beam which is attached to the wafer by the anchor.

13.3.2 Micromachining fabrication of the polysilicon thin film membrane

The design and fabrication of motion microstructures start with the microstructure synthesis, identification of the microstructure functionality, specifications, and performance. The basic steps of surface micromachining we can consider on the example of the thin film membranes preparation taking into account that membranes are widely applied in MEMS. Let us develop the fabrication flow to fabricate the thin membrane. The polysilicon membrane can be fabricated by oxidizing a silicon substrate, patterning the silicon dioxide, deposition and patterning of polysilicon over the silicon dioxide, and removal of the silicon dioxide. To attain the actuation features, the NiFe thin film alloy (magnetic material) should be then deposited. The major fabrications steps and processes for thin membrane are given in Table 13.3.

Table.13.3 Process steps of thin film membrane

Process Steps	Description
Step 1.	Silicon dioxide is grown thermally on a silicon substrate. For example,

Silicon dioxide grow	growth can be performed in a water vapor ambient at 1000^0C for one hour. The silicon surfaces will be covered by 0.5 to 1 μm of silicon dioxide (thermal oxide thickness is limited to a few microns due to the diffusion of water vapor through silicon dioxide). Silicon dioxide can be deposited without modifying the surface of the substrate, but this process is slow to minimize the thin film stress. Silicon nitride may also be deposited, and its thickness is limited to 4-5 μm.
Step 2. Photoresist	A photoresist (photo sensitive material) is applied to the surface of the silicon dioxide. This can be done by spin coating the photoresist suspended in a solvent. The result after spinning and driving-off the solvent is a photoresist with thickness from 0.2 to 2 μm. The photoresist is then soft baked to drive off the solvents inside.
Step 3. Photolithography Exposure, and Development	The photoresist is exposed to ultraviolet light patterned by a photolithography mask (photomask). This photo-mask blocks the light and defines the pattern to guarantee the desired surface topography. Photomasks are usually made using fused silica, and optical transparency at the exposure wavelength, flatness, and thermal expansion coefficient must be met. On one surface of the glass (or quartz), an opaque layer is patterned (usually hundreds of Å thick chromium layer). A photomask is generated based upon the desired form of the polysilicon membrane. The surface topography is specified by the mask. The photoresist is developed next. The exposed areas are removed in the developer. In a positive photoresist, the light will decrease the molecular weight of the photoresist, and the developer is selectively remove (etch) the lower molecular weight material.
Step 4. Etch silicon dioxide	The silicon dioxide is etched. The remaining photoresist will be used as a *hard mask* which protects sections of the silicon dioxide. The photoresist is removed by the wet etching (hydrofluoric acid, sulfuric acid, and hydrogen peroxide) or dry etching (oxygen plasma). The result is a silicon dioxide thin film on the silicon substrate.
Step 5. Deposit polysilicon	Polysilicon thin film is deposited over the silicon dioxide. For example, polysilicon can be deposited in the LPCVD system at 6000C in a silane (SiH_4) ambient. The typical deposition rate is 65-80 Å/min to minimize the internal stress and prevent bending and buckling (polysilicon thin film must be stress free or have a tensile internal stress). The thickness

	of the thin film is up to 4 μm.
Step 6. Photoresist	Photoresist is applied to the polysilicon thin film, and the planarization must be done. The patterned silicon dioxide thin film changes the topology of the substrate surface. It is difficult to apply a uniform coat of photoresist over a surface with different heights. This results in photoresist film which different thickness, nonuniformity, and corners and edges of the patterns may not be covered. For 1 μm (or less) height, this problem is not significant; but for thicker films and multiple layers, replanarization is required.
Step 7. Photolithography Exposure, and Development	A photomask containing the desired topography (form) of the polysilicon membrane is aligned to the silicon dioxide membrane. Alignment accuracy (tolerances) can be done within the nanometer range, and the accuracy depends upon the size features of the microstructure.
Step 8. Etch polysilicon	The polysilicon thin film is etched with the photoresist protecting the desired polysilicon membrane form. It is difficult to find a wet etch for polysilicon which does not attack photoresist. Therefore, dry etching through plasma etching can be applied. Selectivity of the plasma between polysilicon and silicon dioxide is not concerned because the silicon dioxide will be removed later. Therefore, the polysilicon can be overetched by etching it longer than needed. This results in higher yield.
Step 9. Photoresist remove	The photoresist protecting the polysilicon membrane is removed.
Step 10. Deposit NiFe	NiFe thin film is deposited.
Step 11. Remove silicon dioxide: release the thin film membrane	The silicon dioxide is removed by wet etching (hydrofluoric or buffered hydrofluoric acids) because plasma etching cannot easily remove the silicon dioxide in the confined space under the polysilicon thin film. Hydrofluoric acid does not attack pure silicon. Hence, the polysilicon membrane and silicon substrate will not be etched. After the silicon dioxide is removed, the polysilicon membrane is formed (released). This membrane can bend down and stick to the surface of the substrate during drying after the wet etch. To avoid this, a rough polysilicon, which does not stick, can be used. Other solution is to fabricate the polysilicon membrane with the internal stress attaining that the

	polysilicon membrane it bended (curved) up during drying. Both solutions lead to the specific mechanical properties of the polysilicon membrane which might not be optimal from the operating requirements standpoints. Therefore, the alternatives are sought, and, in general, it possible to fabricate the polysilicon membrane with no stress.

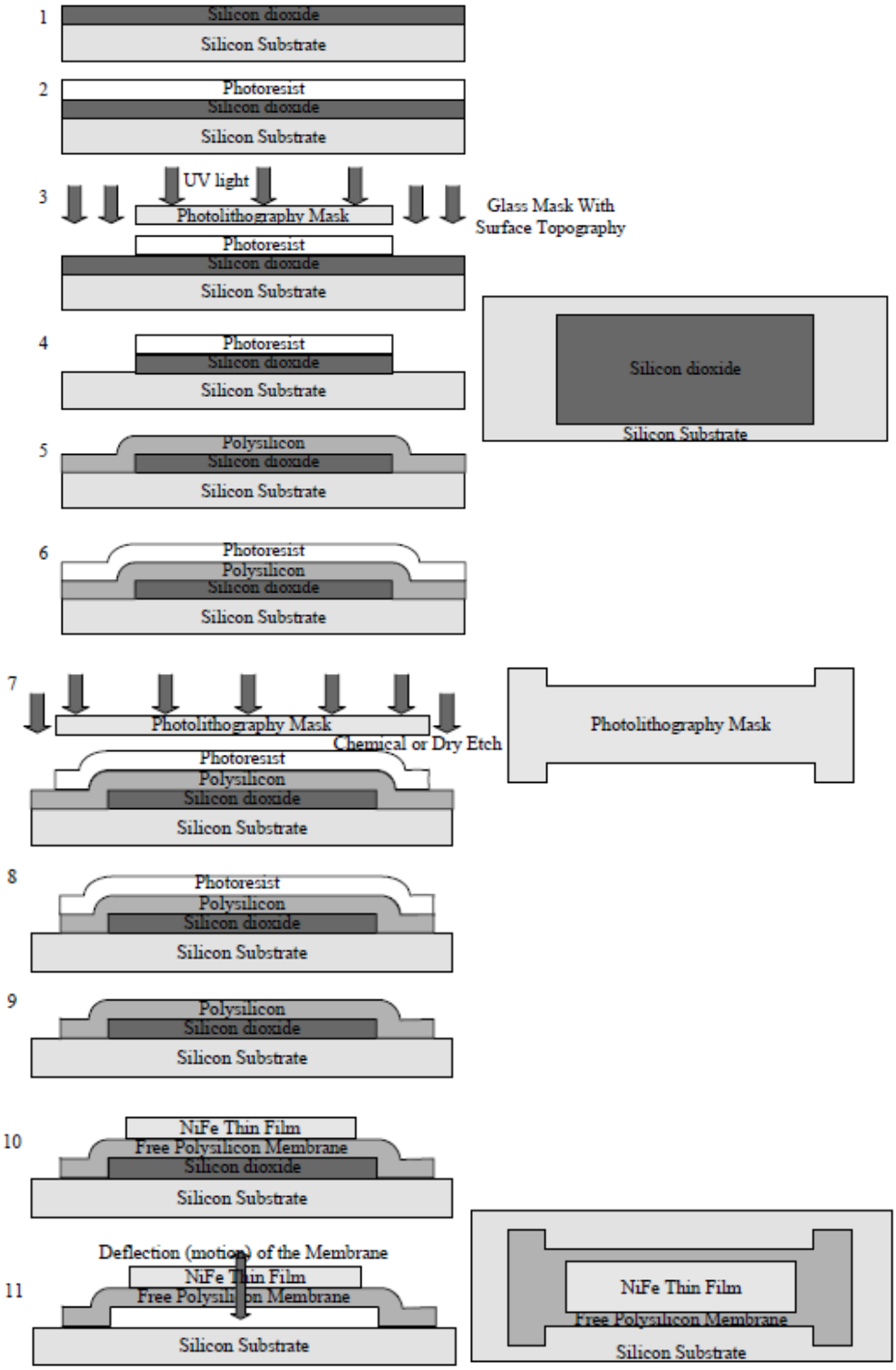


Figure 13.9 Micromachining fabrication of the polysilicon thin film membrane.

It is evident that the conventional CMOS processes and materials were used to develop the fabrication flow (steps) in order to fabricate thin film membrane. Therefore, CMOS fabrication facilities can be converted to fabricate microstructures, microdevices, and MEMS. Figure 9 illustrates the application of the surface micromachining technology to fabricate the polysilicon thin film membrane on the silicon substrate according presented in table.

In electromagnetic microstructures and microdevices, metals, alloys, ferromagnetic materials, magnets, and wires (windings) must be deposited. Using electron-beam lithography (a photographic process that uses an electron microscope to project an image of the required structures onto a silicon substrate coated with a photosensitive resist layer), the process of fabrication of micromagnets on the silicon substrate is illustrated in Figure 13.10. Development removes the photoresist which has been exposed to the electron beam. A ferromagnetic metal or alloy is then deposited, followed by lift-off of the unwanted material. This process allows one to make microstructures within nanometer dimension.

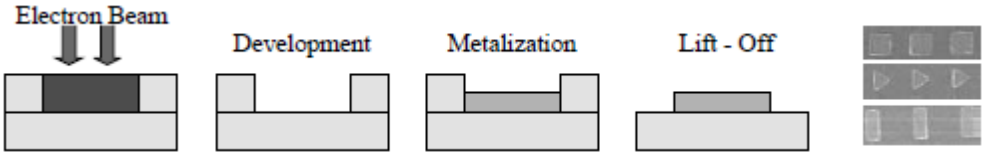


Figure 13.10. Electronic beam lithography in micromagnet fabrications

13.3.3 Sacrificial and Structural Materials

The materials used in a particular surface-micromachining process can be divided into two groups: structural materials and sacrificial materials (Table 13.4). The structural materials are relatively untouched by the etching during the release process and so remain in the final device. The sacrificial materials are removed during the release process and so are not present in the final device. The sacrificial materials are used to define spaces, or voids, in the structures. It is important to note that the role a material plays in a particular surface micromachining process depends strongly on that process. For example, surface micromachining based on copper and nickel is possible. Etchants can be found that will selectively remove either the copper or nickel. Copper and nickel can therefore swap roles of being the structural and sacrificial materials. Since, in surface micromachining, materials are deposited as thin films, it is common to refer to the different films, or layers, as being structural or mechanical.

Table 13.4. Materials in some surfacemicromachining processes.

Process	Structural	Sacrificial
MEMGen (a)	Nickel	Copper
MEMGen (b)	Copper	Nickel
PolyMUMPs™	Polycrystalline silicon and gold	Silicon dioxide
MetalMUMPs™	Polycrystalline silicon, copper, nickel and gold	Silicon dioxide
SUMMiT V™	Polycrystalline silicon	Silicon dioxide

Thus the various layers assume different roles.

• The *structural materials* should have good mechanical properties. For most devices, the structural material should also have good electrical properties.

• The *sacrificial materials* should be stable during deposition and throughout the fabrication process. However, it should etch quickly during the release step.

Ideally, all materials used in a surface micromachining process are easily pattemed. Handling the different materials will be much easier if they can be deposited and pattemed using standard microelectronics processes. Finally, the etchant should have a high selectivity for the sacrificial materials during release.

Table lists some surface micromachining processes and the materials they use. Historically, surface micromachining processes have depended on materials readily available from microelectronics, and so processes based on silicon/silicon dioxide, silicon/silicon nitride, and aluminium/silicon nitride are common.

For example, to fabricate microscale gears (microgear train), a sacrificial silicon dioxide is deposited on the wafer and patterned. Then, a structural layer of polysilicon is deposited and patterned. This polysilicon layer becomes the structural microgears element. Other layers are then deposited and patterned making the rest of the microstructure (microscale gears). Etching in the hydrofluoric or buffered hydrofluoric acids removes the sacrificial layers releasing the microgear. As more layers are added, more mechanical structures can be created (figures 13.11). More complicated devices, such as pin-joints and moving platform, require additional layers. For example, both PolyMUMPs™ and SUMMiT™ are surface micromachining processes that use polycrystalline silicon and silicon dioxide. However, the PolyMUMPs ™ process is a three layer process, while SUMMiT™ has up to five structural layers. SUMMiT™ is therefore capable of creating several structures not available in PolyMUMPs™. Additional layers require some type of planarization. For example, the wafer is planarized before depositing the fourth structural layer in SUMMiT™(fig.13.12).

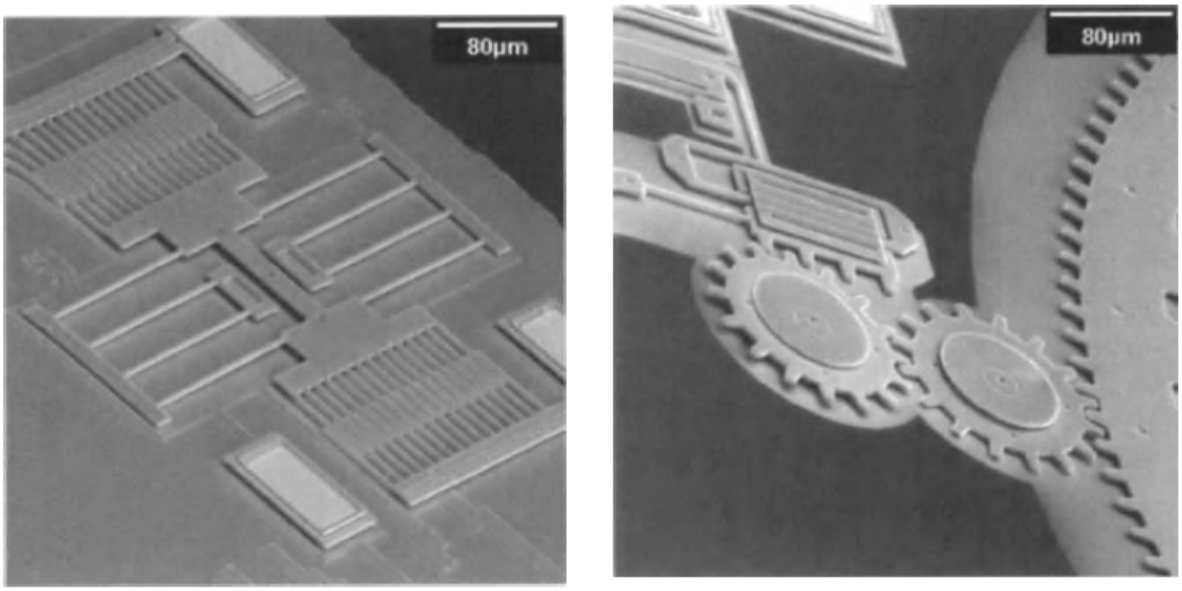


Figure 13.11. *a)*SEM of electrostatic comb drive resonator, which can be built in two layers processes. b) SEM of a gear train, which can be built in three layer processes. Fabricated using PolyMUMPs™.

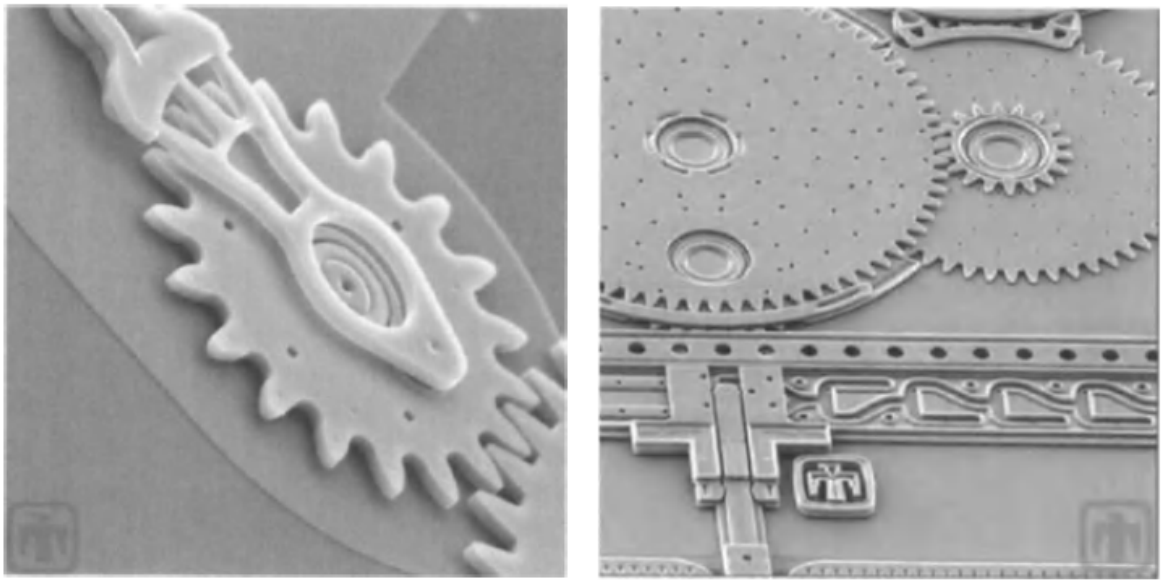


Figure 13.12. a) SEM of a gear containing a pin-joint, which can be built in four layers processes, b) SEM of multi level gears, which can be built in five layers processes. Fabricated using SUMMiT™. www.mems.sandia.gov.

Because of their similarities, it is possible to develop a manufacturing process which includes both microelectronics and micromechanics processes. This allows extremely tight integration of the sensors, actuators, and electronics. Unfortunately, a monolithic process that includes both micromechanics and microelectronics requires many processing steps, and so these combined processes have relatively low yields compared to separate microelectronics or

micromachining processes. There exists a range of options for systems that combine micromechanics and microelectronics. These options vary from completely separate packages, to multiple chips in the same package. Increasing integration typically means increased performance, since parasitics and noise associated with the interconnects is reduced. However, increased integration also typically means increased cost due to increased manufacturing complexity.

13.4 LIGA

The LIGA process, which denotes *Lithography–Galvanoforming–Molding* (in German Lithografie–Galvanik–Abformung), is capable of producing three dimensional microstructures of a few centimeters high with the aspect ratio (*depth versus lateral dimension)* of more than 100.
The detailed LIGA process description is discussed below:

- Deep X-ray lithography and mask technology - Deep X-ray (0.01 – 1nm wavelength) lithography can produce high aspect ratios (1 mm high and a lateral resolution of 0.2 μm).

- X-rays break chemical bonds in the resist; exposed resist is dissolved using wetetching process.

- Electroforming

- The spaces generated by the removal of the irradiated plastic material are filled with metal (e.g. Ni) using electro-deposition process.

- Precision grinding with diamond slurry-based metal plate used to remove substrate layer/metal layer.

- Resist Removal

- PMMA resist exposed to X-ray and removed by exposure to oxygen plasma or through wet-etching.

- Plastic Molding

- Metal mold from LIGA used for injection molding of MEMS.

LIGA Process Capability:

- High aspect ratio structures: 10-50 μm with Max. height of 1-500 μm
- Surface roughness < 50 nm
- High accuracy < 1μm

The LIGA technology is based on the x-*ray lithography* which guarantees shorter wavelength (from few to ten Å which lead to negligible diffraction effects) and larger depth of focus compared with optical lithography. The ability to fabricate microstructures and microdevices in the centimeter range is particularly important in the actuators applications since the specifications are imposed on the rated force and torque developed by the microdevices. Due

to the limited force and torque densities, the designer faces the need to increase the actuator dimensions.

Desides deep x-ray lithography the LIGA and LIGA-like processes are based on and *electroplating* of metal and alloy structures, allowing one to achieve structural heights in the centimeter range. This type of processing expands the material base significantly and allows the fabrication of new high-performance electromechanical microtransducers. High performance actuators with maximized active volumes and minimized surface areas have been designed and fabricated using LIGA and LIGA-like technologies.

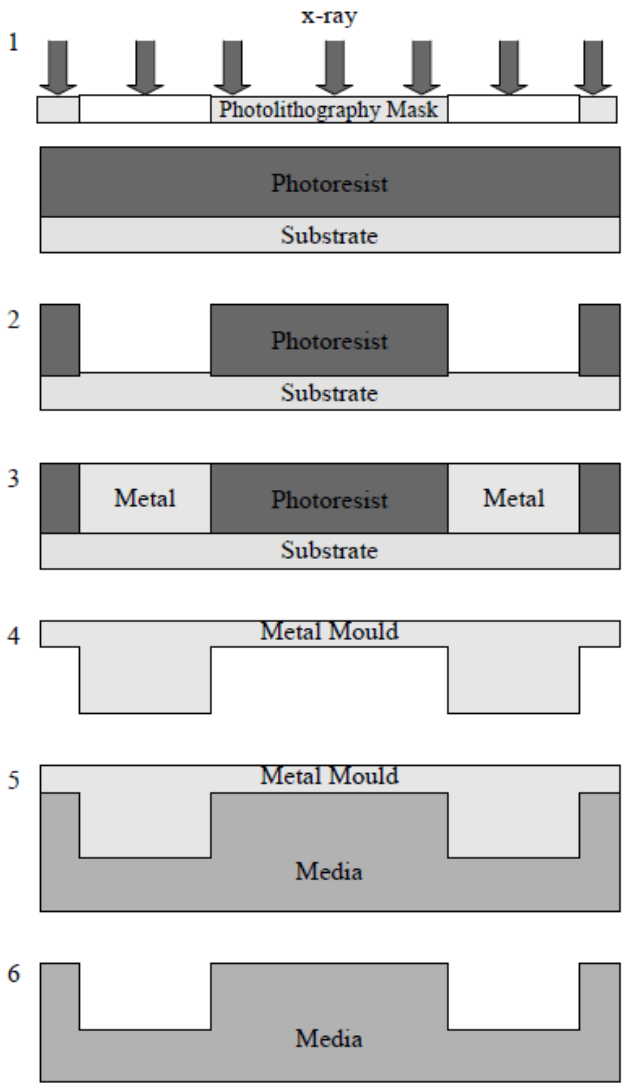


Figure 13.13 LIGA fabrication technology.

In these processes, a substrate with a plating base is covered with a thick photoresist (thickness can be in the centimeter range). The photoresist is cured and exposed by x-rays from a synchrotron source (x-ray lithography). The photoresist strain, which is due adhesion, causes well known difficulties. This problem is solved by combining surface micromachining,

patterning the sacrificial layers under the plating base, and optimizing the processes. The achievable structural height of LIGA or LIGA-like fabricated structures is defined mainly by the photoresist processing. The photoresist procedures have been optimized and used to produce low strain photoresist layers with thickness from 50 μm to the centimeters range. Large area exposures of photoresist with thickness up to 10 cm have been achieved with x-ray masks. After electroplating, replanarization can be made through precision polishing.

Figure 13 illustrates the basic sequential processes (steps) in LIGA technology. Here, the 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 (step 1). This photoresist is then developed (step 2). The pattern formed is electroplated with metal (step 3). The metal structures produced can be the final product, however, it is common to produce a metal mould (step 4). This mould can then be filled with a suitable media (e.g., metal, alloy, polymer, etc.) as shown in step 5. The final structure is released (step 6).

The described LIGA technology (frequently referred to as the *high aspect ratio* technique) allows one to fabricate microstructures with small lateral dimensions compared with thickness. Thick and narrow microstructures guarantee high ruggedness in the direction perpendicular to the substrate and compliance in the lateral directions. For actuators, high aspect-ratio technology offers the possibility to fabricate high torque and force density microtransducers. As was emphasized, high-intensity, lowdivergence, and hard x-rays are used as the exposure source for the lithography. Due to short exposure wavelength, the desired features size is achieved. These x-rays are usually produced by a synchrotron radiation source. Polymethyl-methacrylate (PMMA) and polylactides are used as the x-ray resists because PMMA (PlexiglasTM or LuciteTM) and polylactides photoresists have high sensitivity to x-rays, thermal stability, desired absorption, as well as high resolution and resistance to chemical, ion, and plasma etching. Polyglycidyl-methacryl-atecoethylacrylate (PGMA) is used as the negative x-ray resist. The exposure wavelength varies depending upon the x-ray radiation source used. For example, the 0.2 nm x-ray wavelength allows one to transfer the pattern from the high-contrast x-ray mask into the photoresist layer with a few centimeters thickness so that the photoresist relief may be fabricated with an extremely high depth to width ratio.

A critical part of the high-aspect-ratio processes is plating to form the metallic electromechanical microstructures in the mold. Using plating, metal is deposited from ions in a solution following the shape of the plating mold. This is the additive process, and the thickness of the plated metal can be large since the plating rate can be high. A variety of metals (Al, Au, Cu, Fe, Ni, and W) and alloys (NiCo, NiFe, and NiSi) can be deposited or codeposited. It is

important that roughness (smoothness) of the reflective metal surfaces with the desired shape can be achieved even for optical applications.

Despite the wide range of micromachining technologies, they share close links with the microelectronics industry. Many of the processes used in micromachining were pioneered for microelectronics. This allows micromachining fabrication to draw on established expertise and equipment. For the future, it is hoped that the close links between microelectronics and micromachining will lead to manufacturing technologies capable of creating both mechanical and electrical elements. Such a process would be ideal for sensors, actuators, and other combined typesof systems. System integration, where increasingly complicated systems are built at fixed costs, is the ultimate goal.

13.5 Microelectrodischarge Machining

13.5.1 General description

Microelectrodischarge machining (also known as *micro-EDM*, *μ-EDM*, and *electrodischarge micromachining*) has been developed in the past 30 years from the nonconventional manufacturing technique of electrodischarge machining (EDM) commonly known as spark erosion. In essence, the principle of EDM is simple:

1. Two electrodes are separated from each other by a dielectric fluid.
2. A voltage difference is applied between the two electrodes, a cathode (negatively charged) and an anode (positively charged).
3. If the two electrodes are moved close enough together and the voltage is high enough, the dielectric fluid will break down and conduct an electrical current, causing an electrical discharge (a spark) between them.
4. The sparks will produce an extremely high temperature (of the order of 10,000 K) at localized spots on the electrodes such that the electrode materials (especially the anode) will be vaporized, leaving craters behind on the surfaces as evidence of material removal.
5. By careful choice of the dielectric, voltage generator, and electrodes this material removal process can be used as a manufacturing tool in the following ways.

Current micro EDM technology used for manufacturing micro-features can be categorized into four different types:

a) *Die-sinking micro-EDM*, where an electrode with micro-features is employed to produce its mirror image in the workpiece.

b) *Micro-ED drilling*, where micro-electrodes (of diameters down to 5–10 μm) are used to ‘drill’ micro-holes in the workpiece.

c) *Micro-ED milling*, where micro-electrodes (of diameters down to 5–10 μm) are employed to produce complex 3D cavities by adopting a movement strategy similar to that in conventional milling.

d) Micro-wire EDM, where a wire of diameter down to 20 μm is used to cut through a conductive workpiece.

13.5.2 EDM Die-Sinking

By fabricating a cathode (tool) in a conductive material such as copper or graphite and feeding the tool under servocontrol towards a planar (workpiece) anode immersed in kerosene, the shape of the tool can be impressed into the workpiece. Thus a (male) tool can be used to form a (female) workpiece of inverse shape, of great use in the production of a mold which can be used to replicate many ‘‘male’’ parts economically. The concept is illustrated in Figure 13.14.

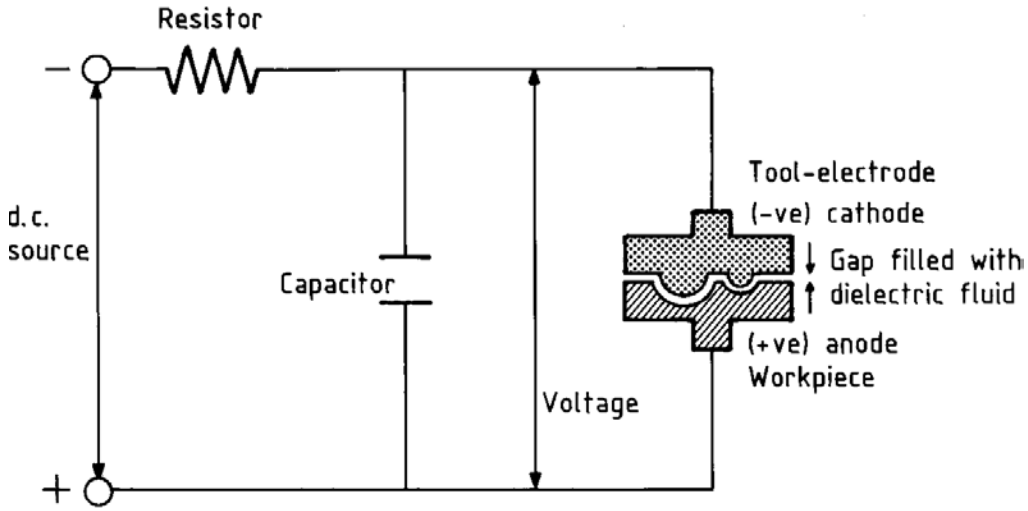


Figure 13.14 EDM by die-sinking connected to a relaxation (RC) circuit

13.5.3 Wire EDM (WEDM)

By using a continuous wire as the guiding cathode, and keeping it flooded with a dielectric such as deionized water, shapes may be machined into conductive materials, such as flat sheets and plates as illustrated in Figure 13.15. Typically wires comprise copper, brass, or tungsten with diameters ranging from 50 to 250 μm. The technique may therefore be described as a type of noncontact jigsaw for production of both simple and complex shapes.

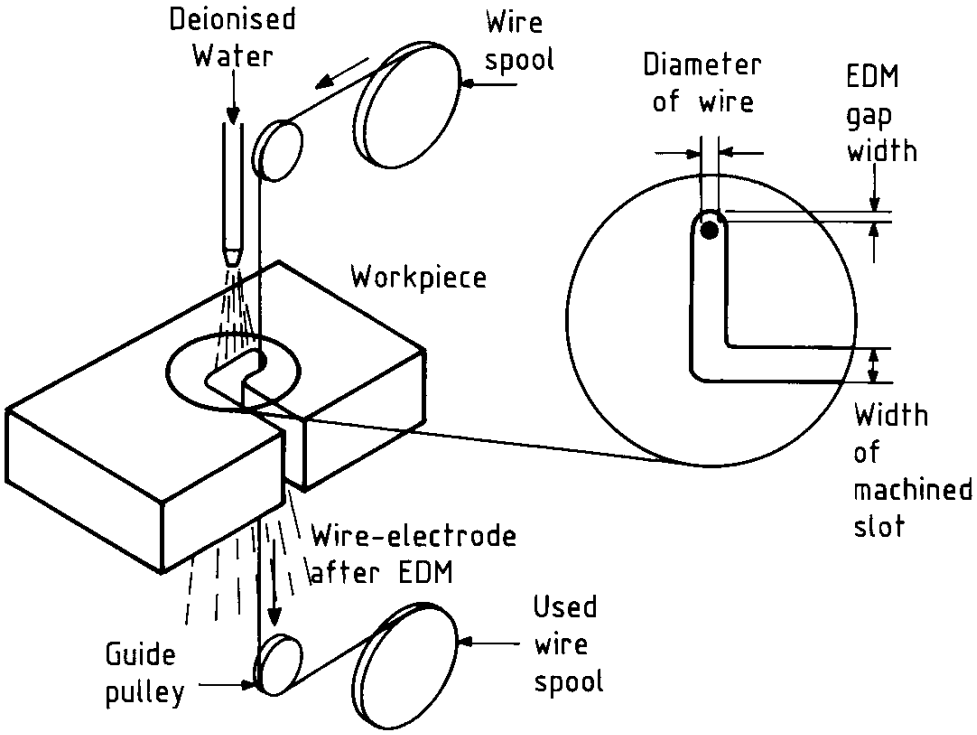


Figure 13.15 Wire EDM.

13.5.4 Electrodischarge Grinding (EDG)

This technique has been given its title due to its similarity in some respects to the grinding process. However, EDG is still a noncontact machining process and usually employs a spinning flat cathode and a dielectric fluid to form a smooth surface on a workpiece. Thus, as can be seen in Figure 13.16, a method of producing cylinders with parallel sides, stepped features, or tapers was developed to provide the required dimensions and tolerances. This work was also commercialized by Matsushita in 1988 as it allowed fabrication of rods with diameters less than 10 μm to be made routinely and thus complement the production of 10-μm diameter holes by μ-EDM die-sinking.

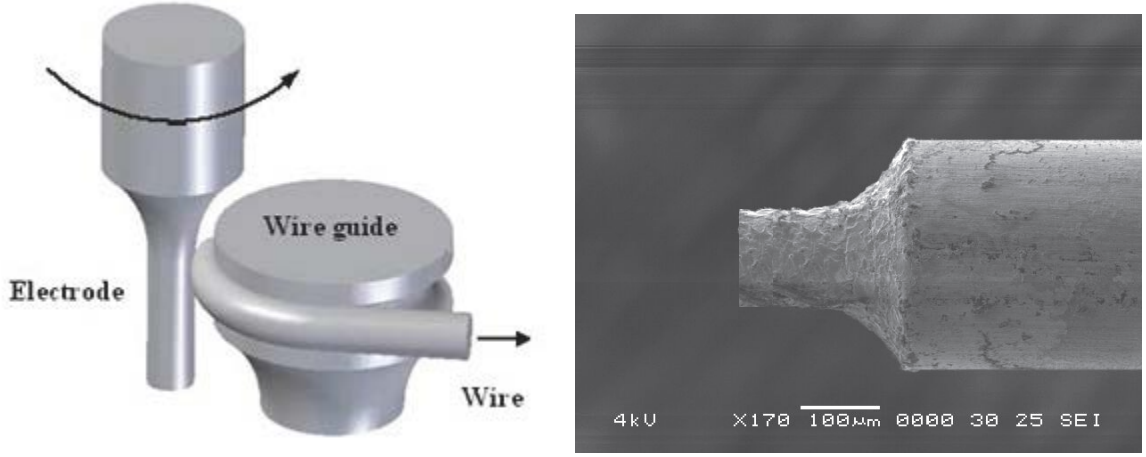


Figure 13.16 Principle of WEDG and SEM of rod after WEDG teatment.

The rate of machining is dependent on the discharge energy, which for the attainment of fine surface finish is kept low (<10^7 J per pulse) in μ-EDM. Table 13.5. shows rates of machining for sulphur-killed (SK) steels, stainless steels, and silicon.

Table 13.5. Rates of machining

μ-EDM	Feature	Feature size (μm)	Dielectric	Material	Thickness (μm)	Machining time (s)	Metal removal rate (mm^3/min)
Die-sinking	Circular hole	ϕ = 100	D. I. water	SK steel	1000	120–260	$1.8–3.9 \times 10^{-3}$
Die-sinking	Circular hole	ϕ = 55	D. I. water	SK steel	500	21–25	$2.9–3.4 \times 10^{-3}$
Die-sinking	Circular hole	ϕ = 100	D. I. water	AISI 304 stainless steel	1000	600	0.79×10^{-3}
Die-sinking	Circular hole	ϕ = 38	Kerosene	17-7PH stainless steel	80	182	0.03×10^{-3}
Die-sinking	Circular hole	ϕ = 85		AISI 304 stainless steel	1000	40	8.5×10^{-3}
μ-WEDM	Slot	Width = 150	D. I. water	50 mΩ.cm silicon	350	Wire cutting at 0.175mm/s	550×10^{-3}

13.5.5 Application of μ-EDM

The applications of μ-EDM are many and various. They are described below according to the method used for fabrication.

1) Microelectrodes for μ-EDM Die-Sinking. Typically a WEDG machine is used to make cylindrical electrodes with diameters down to 5 μm. If these electrodes are to be used for μ-EDM die-sinking precision holes, then the μ- electrodes need to be straight and have a high length-to-diameter (*L*/*D*) ratio so that many holes can be fabricated from the same electrode before it wears away.

2) Microshafts and Pins. These parts are especially important in the assembly of miniature Devices.

3) Micropipes. A novel process has been developed for the manufacture of micropipes, combining WEDG and electroforming, as shown in Figure 13.17 a.

4) Inkjet Nozzles

5) Micro mold cavity (fig.13.17b)

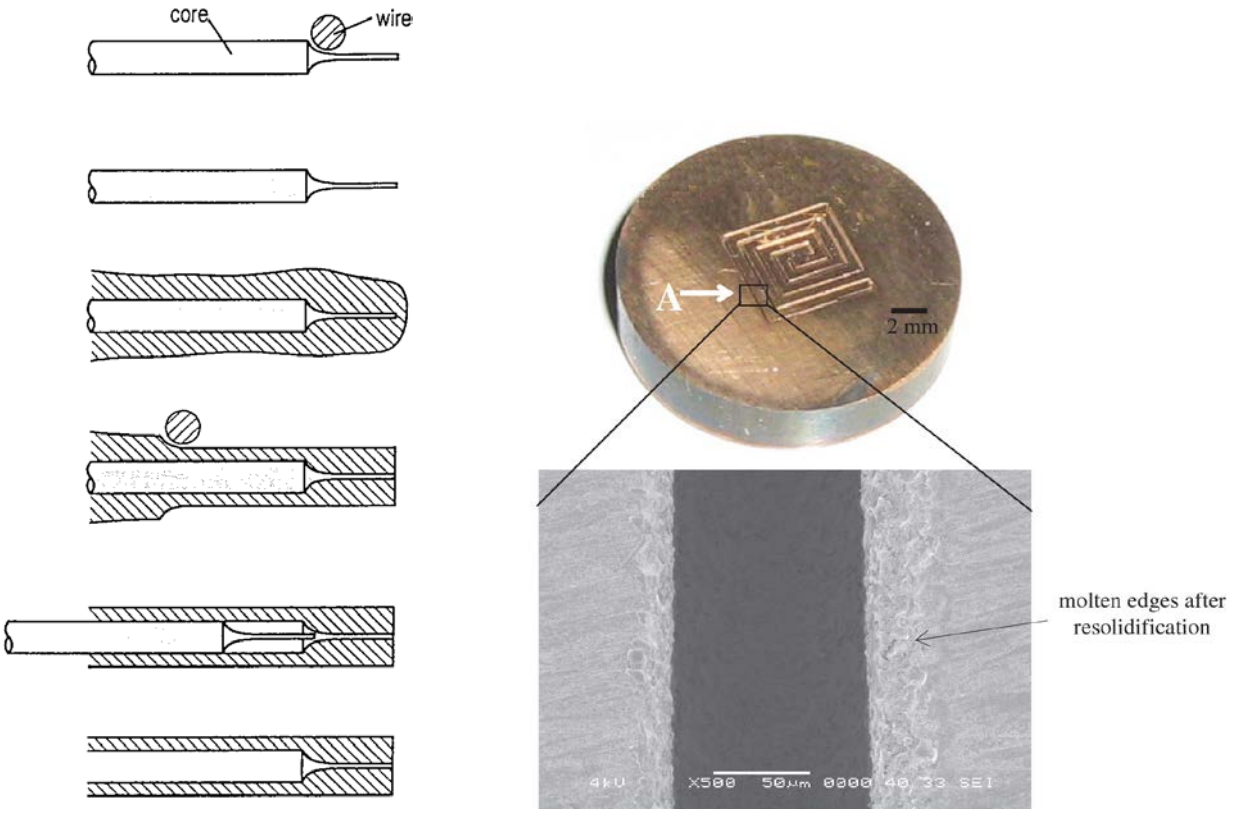


Figure 17 a) Micropipe fabrication: (a) Co re preparation by WEDG; (b) parting film formation; (c) deposition (electroforming); (d) forming outside by WEDG; (e) parting; (f) finished nozzle
b) Fabrication of micro swiss-roll combustor mold cavity by micro ED milling: fabricated micro swiss-roll combustor mold cavity and SEM micrographs of window A.

13.6 Nanofabrication by Focused-ion-beam technique

The *focused ion beam (FIB)* based nanofabrication method can be followed for the fabricating the nanoscale devices on materials based on metal and non-metallic elements, particularly the layered structure materials like graphite, Bi-2212 and YBCO which are recently attracted the world scientific community due to their interesting electrical and electronic properties. FIB can be used as a *direct milling* method to make microstructures without involving complicated masks and pattern transfer processes. FIB machining has advantages of high feature resolution, and imposes no limitations on fabrication materials and geometry. Focused ion beams operate in the range of 10-200 keV. As the ions penetrate the material, they loose their energy and remove substrate atoms. FIB has proven to be an essential tool for highly localized implantation doping, mixing, micromachining, controlled damage as well as ion-induced deposition. The FIB technique allowing us to engrave materials at very low dimensions is a complement of usual lithographic techniques such as optical lithography. The main difference is that FIB allows direct patterning and therefore does not require an intermediate sensitive media

or process (resist, metal deposited film, etching process). FIB allows 3D patterning of target materials using a finely focused pencil of ions having speeds of several hundreds of km s−1 at impact. Concerning the nature of the ions most existing metals can be used in FIB technology as pure elements or in the form of alloys, although gallium (Ga+ ions) is preferred in most cases.

The FIB milling involves two processes: 1) Sputtering, ions with high energy displace and remove atoms of substrate material, and the ions lose their energy as they go into the substrate; 2) Re-deposition, the displaced substrate atoms, that have gained energy from ions through energy transfer, go through similar process as ions, sputtering other atoms, taking their vacancy, or flying out.

13.6.1 Nanoscale stack fabrication by focused-ion-beam

Using an FIB, perfect stacks can be fabricated more easily along the c-axis in thin films and single-crystal whiskers. FIB 3D etching has been recognized as a well-known method for fabricating high-precision ultra-small devices, in which etching is a direct milling process that does not involve the use of any masking and process chemicals and that demonstrates a submicrometer resolution. Thus, these our proposal is focused on the fabrication of a nanoscale stack from the layered structured materials like thin graphite flake and BSCCO, via FIB 3D etching. The detailed schematic of fabrication process is shown in Fig. 13.18.

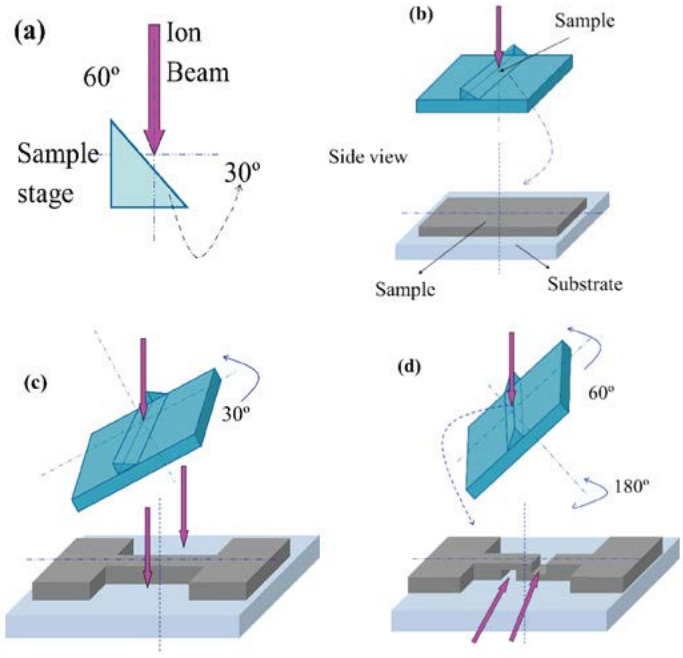


Fig. 13.18. FIB 3-D fabrication process (a) Scheme of the inclined plane has an angle of 60° with ion beam (where we mount sample). (b) The initial orientation of sample and sample stage. (c) Sample stage titled by 30° anticlockwise with respect to ion beam and milling along abplane. (d) The sample stage rotated by an angle of 180° and also tilted by 60° anticlockwise with respect to ion beam and milled along the c-axis.

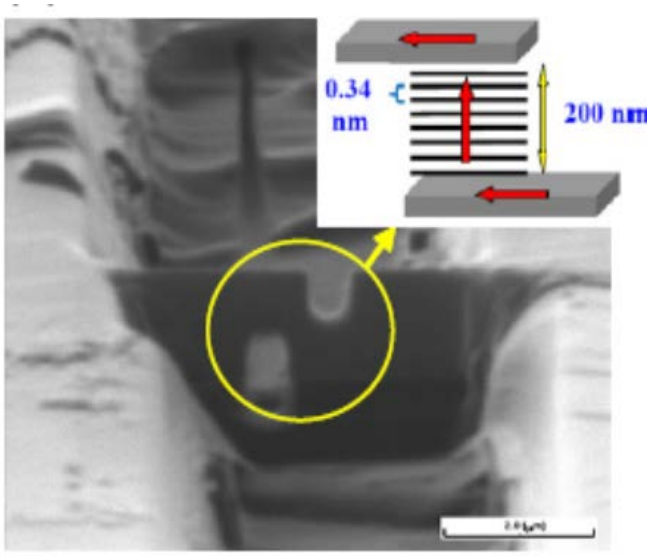


Fig. 13.19. FIB image of the nanoscale stack fabricated on a thin graphite flake along the c-axis height of 200 nm (image scale bar is 2 μm). Inset shows the schematic diagram of stack arrangement along the c-axis. The vertical red arrow indicates the current flow direction through the nanoscale stack.

The 3D etching technique is followed by tilting the substrate stage up to 90° automatically for etching thin graphite flake. We have freedom to tilt the substrate stage up to 60° and rotate up to 360°. To achieve our goal, we used sample stage that itself inclined by 60° with respect to the direction of the ion beam (fig 13.18a). The lateral dimensions of the sample were 0.5×0.5 μm^2. The in-plane area was defined by tilting the sample stage by 30° anticlockwise with respect to the ion beam and milling along the *ab*-plane. The in-plane etching process is shown in Fig. 18(a)–(c). The out of plane or the c-axis plane was fabricated by rotating the sample stage by an angle of 180°, then tilting by 60° anticlockwise with respect to the ion beam, and milling along the c-axis direction. The schematic diagram of the fabrication process for the side-plane is shown in Fig. 13.18(d). The dimensions of the side-plane was W=0.5 μm, L=0.5 μm, and H=200 nm. The c-axis height length (H) of the stack was set as 200 nm. An FIB image of fabricated stack is shown in Fig. 19 in which the schematic of stack arrangement (graphene layers with interlayer distance 0.34 nm) was also shown in the inset (top right) in Fig. 13.19. The vertical red arrow indicates the current flow direction through the stack.

13.6.2 FIB nano fabrication on superconducting devices

Superconductivity is a phenomenon when the resistance of the material becomes zero and it expels all the magnetic field below a certain temperature usually at very low temperature. The quantum application of superconductivity was introduced in 1962. B. D. Josephson

discovered a tunnel junction consists of two strips of superconductors separated by an insulator where the insulator is so thin that electrons can tunnel through it known as Josephson junction. The schematic of different types of Josephson junctions are shown below in Fig.13.20. S stands for superconductor, S' for a superconductor above *Tc*, N for normal metal, Se for semiconductor, and I for an insulator.

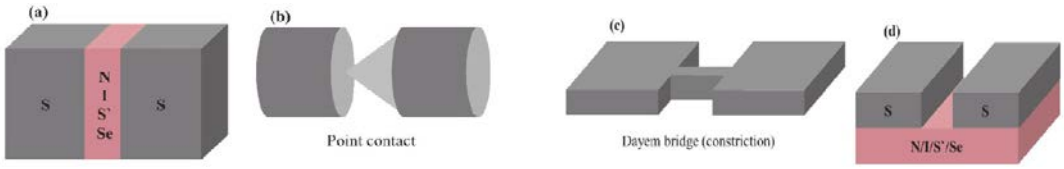


Fig. 13.20. The schematics of different types of superconducting devices.

Now Bi-family material is a one of the famous emerging material for electron tunneling devices, such as *intrinsic Josephson junctions* (IJJ) in layered high-*Tc* superconductors. Considering Bi-family as a layered structure material, there are three compounds in the Bi family high-temperature superconductors, differing in the type of planar CuO_2 layers; single-layered $Bi_2Sr_2CuO_{6+\delta}$ (Bi-2201) single crystal, double-layered $Bi_2Sr_2CaCu_2O_{8+\delta}$ (Bi-2212) single crystal, and triple-layered $Bi_2Sr_2Ca_2Cu_3O_{10+\delta}$ (Bi-2223) single crystal. The spacing of consecutive copper-oxide double planes in the most anisotropic cuprate superconductors is greater than the coherence length in the out-of-plane *c*-direction. When a current flows along the *c*-direction in such a material, it therefore flows through a series array of "intrinsic" Josephson junctions (IJJs). These junctions and junction arrays are showing promise for a wide variety of applications, including as voltage standards and sub-mm-wave oscillators, electric-field sensors and quantum current standards. FIB image of a submicron stack fabricated on Bi-2212 single crystal whiskers with in-plane area of 0.4 μm × 0.4 μm and schematic of the IJJs configuration are shown in Fig. 13.21, in which FIB fabrication procedures followed as described before.

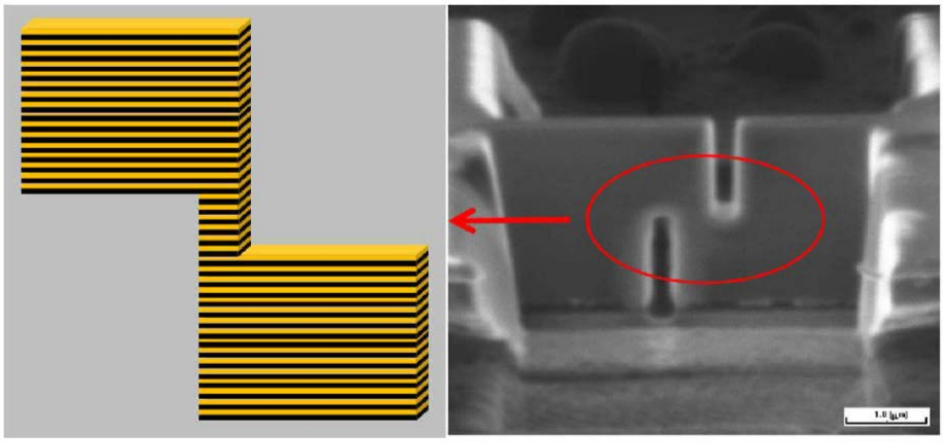


Fig. 13.21. FIB image of a submicron stack fabricated on Bi-2212 single crystal whisker. The red color circular part shown in stack contains many IJJs.

13.7 Laser micromashining

13.7 .1 General remarks

Laser micromachining is based on the interaction of laser light with solid matter. As a result of a complex process, small amounts of material can be removed from the surface of the solid. Two different phenomena may be identified: *pyrolithic (thermal)* and *photolithic* processes. In both cases short to ultrashort laser pulses are applied in order to remove small amounts of material in a controlled way. Pyrolithic processes are based on a rapid thermal cycle, heating, melting, and (partly) evaporation of the heated volume. In the case of photolithic processes the photon energy is sufficient for direct breaking of the chemical bonds in a wide variety of materials. It is applied mostly on polymers by use of ultraviolet lasers in wavelengths of 157 to 351 nm. Because the photon energy is converted directly in breaking chemical bonds there is almost no thermal interaction with the product itself. The reaction products escape as gas or small particles. A presentation of laser processes is given in Table 13.6. *Laser micromachining* includes a wide range of processes where material is removed accurately but the term is also used to describe processes such as microjoining and microadjustment by laser beam.

Table 13.6 Laser Micromachining Processes

Ablation	Wire stripping	Soldering
Cutting	Marking	Trimming
Drilling	Welding	Hardening
Decoration	Structuring	Texturing

Schematic diagram of the laser micromachining setup is drawn in Figure 13.22a. A pair of aluminium mirrors is used to adjust the direction of the collimated laser beam to be coaxial with the UV objective lens. For *laser ablation* the high-power laser pulses are used to evaporate matter from a target surface. In this process, a supersonic jet of particles (plume) is ejected normal to the target surface which condenses on substrate opposite to target. The ablation process takes place in a vacuum chamber- either in vacuum or in the presence of some background gas. The graphical process scheme is given in Fig.13.22b.

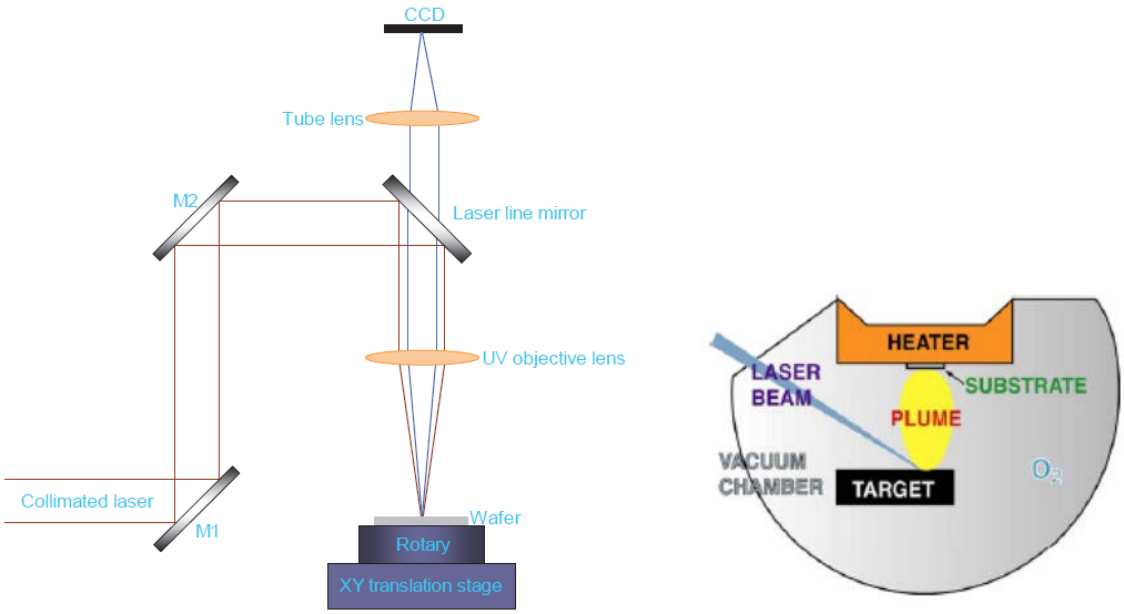


Fig. 13.22. Schematic of laser micromachining setup (a) and laser ablation (b)

Lasers used for micromachining are characterized by short pulse lengths from the millisecond range for applications like microwelding to the pico- and even femtosecond area for ablation of metals. Wavelengths vary from $\lambda = 10.6$ μm for the CO_2 laser to 157 nm for a fluorine excimer laser. The beam is further characterized by the (half) divergence angle θ and the radius *w* of the beam waist. For the ideal beam the following equation applies: $\Theta \cdot w = \lambda/\pi$. This quantity is invariant which means that with ‘‘ideal’’ optics this relation is valid over the whole beam trajectory. A significant expression for the beam quality is the M^2 number, which is the ratio between the θ*w* product for the real beam, compared to an ‘‘ideal’’ Gaussian beam. A minimum spot diameter:

$$\delta = \frac{4}{\pi} \cdot M^2 \lambda \frac{f}{D}.$$

Where D is lens diameter, f is focal length. For micromachining a small spot is usually required. This can be obtained with a high beam quality M^2, with a short wavelength and a short focal length lens *f*. Some examples are given in Table 13.7.

Table 13.7 Examples of Beam Properties

Laser	Wavelength λ (μm)	Power P (W)	w·θ (mm·mrad)	Beam quality M^2 (−)	Spot diameter with f/4 lens δ (μm)
HeNe	0.63	0.002	0.2	0.98	3
Nd:YAG[a]	1.06	100	6	10	50
Nd:YAG[b]		1000	25	80	500
Q-Switched		10	2	3	15
Nd:YAG	1.06	100	6	10	50
		1000	25	80	500
CO_2	10.6	1000	10	1.5	80
Copper vapor	0.51	20	0.5	3	8
Ti:Sapphire	0.78	1			
Excimer	0.193–0.351	100	20	200	—

13.7.2 Principles of laser material removal

The mechanism of laser beam interaction and material removal is shown in Figure 13.23a. Laser energy is focused on the material surface and partly absorbed. The absorptivity depends

on the material, the surface structure, the power density, and the wavelength. With a CO_2 laser about 20% is absorbed with laser micromachining while with shorter wavelengths (Nd:YAG and excimer lasers) 40 to 80% is absorbed. The remaining part is reflected. Absorption occurs in a very thin surface layer, where the optical energy is converted into heat. The optical penetration depth (fig.13.23b) is defined as the depth for which the power density is reduced to 1/*e* of the initial density. For steel this depth is on the order of 15 nm for CO_2 radiation or 5 nm for Nd:YAG radiation. The absorbed energy diffuses into the bulk material by conduction. For short pulses the heat flow is approximately one-dimensional. The temperature at the center of the spot follows with an absorbed power density. With, for example, 10^9 W/cm^2 on steel the melting point is reached in 300 ns. However, the time to melt is reduced by a factor of 100 to only 3 ns if the power density is increased tenfold. The high vaporization rate (vapor speeds have been reported in the range of 3 to 10 km/s) causes a shock wave and a high vapor pressure at the liquid surface considerably increases the boiling temperature. Finally the material is removed as a vapor by the expulsion of melt, as result of the high pressure and by an explosivelike boiling of the superheated liquid after the end of the laser pulse. In metals a rim of resolidified material caused by laser micromachining is clearly evident (Figs. 13.23c).

In plastics, however, the process is quite different; here the material is removed by breaking the chemical bonds of the macromolecules, and is dispersed as gas or small particles. No melt is found (Fig. 13.24). High-power laser pulses are used to evaporate matter from a target surface. In this process, a supersonic jet of particles (plume) is ejected normal to the target surface which condenses on substrate opposite to target. The ablation process takes place in a

vacuum chamber-either in vacuum or in the presence of some background gas. The graphical process scheme is given below in Fig.13.22b.

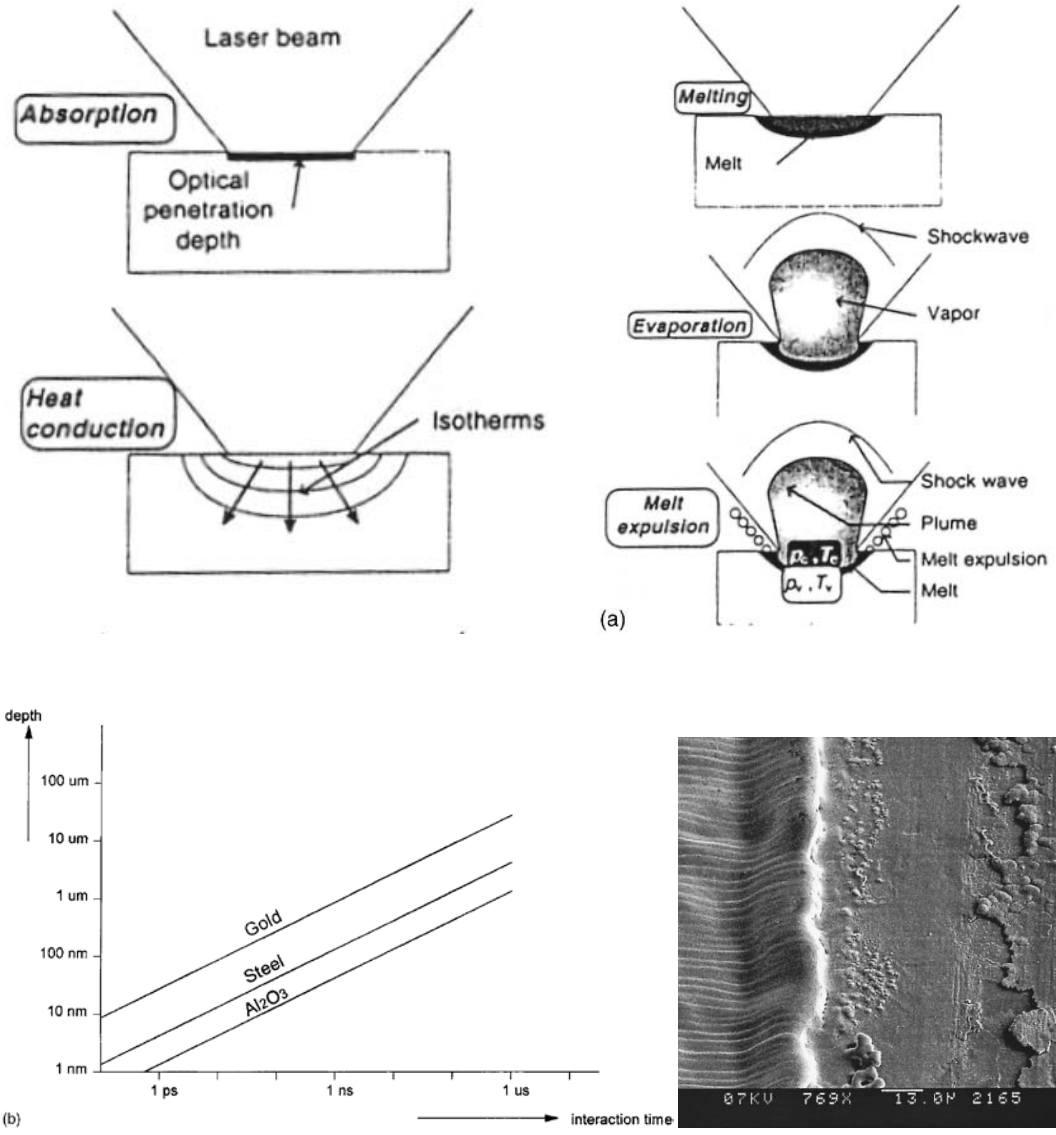


Figure 13.23 Penetration depth of heat for short laser pulses: (a) laser beam interaction and material removal; (b) micron and submicron penetration levels, c) Rim of resolidified material after micromachining of stainless steel. Also some redeposition is evident (right): excimer laser, KrF 248 nm, 20 ns, 12 J/cm^2, 90 Hz. The ablation depth is 15 μm; rim height is 10 μm.

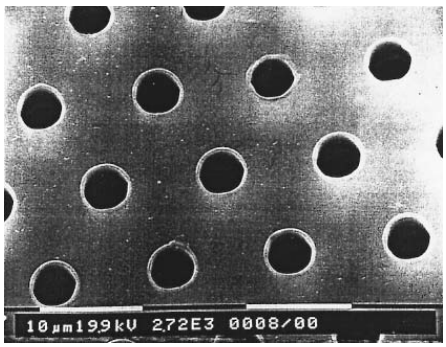


Figure 13.24 Fine hole drilling in polyamide: diameter 5 µm, depth 25 µm. No significant liquid phase could be observed.

The results for micromachining of steel are given in Table 13.8. From the table ablation still occurs at the end of the pulse owing to the presence of a relatively thick superheated layer, which continues to evaporate while the surface is already cooling down but still remaining above the normal boiling temperature. From the model it was found that with a given (total) fluence the ablation depth has an optimum (maximum) for a given pulse length. For 2.5 J/cm^2 this optimum is obtained with 15 ns pulses, for 5 J/cm^2 with 50 ns, and for 10 J/cm^2 with 150 ns pulses. The ablation depth given in Table 13.8 is based on material removal only by evaporation.

Table 13.8 Numerical Results for Laser Micromachining of Steel

Fluence	1 J/cm^2			2.5 J/cm^2			5 J/cm^2		
Time (ns)	5	10	20	5	10	20	5	10	20
Surface temperature (K)	2000	2500	3200	4000	4300	4300	4300	4300	4300
Ablation velocity (m/s)	0	0	0.8	5	12	12	12	12	12
Melt layer (µm)	0.03	0.15	0.3	0.2	0.3	0.65	0.35	0.8	1.4
Dielectric layer (µm)	0	0	0	0	0.1	0.4	0.3	0.6	1.3
Ablation depth at end of pulse (µm)		0			0.18			0.24	
Total ablation depth (µm)		0			0.22			0.40	

The mechanism of material removal for plastics is based on photochemical reactions with photons. The typical bonding energy for many macromolecules is in the 3 to 15 eV range, which corresponds approximately with the photon energy in the ultraviolet. The process consists of three steps: the UV photons are absorbed in the top layer typically of 0.2-µm thickness, the long chain molecules in this layer are broken into parts, and finally they are removed from the processing area in the form of vapor and small particles. The photon energy obtained from different lasers is listed in Table 13.9. Only photons with higher energies can release the chemical bonds. Machining of Teflon, for instance, requires at least the photon energy as produced by the fluorine laser The threshold fluence for a wide variety of plastics is about 120

mJ/cm^2. At low fluence the walls become tapered from about 2° at 500 mJ/cm^2 to 20° at 150 mJ/cm^2. Useful fluences are given in Table 13.10.

Table 13.9 Photon Energies from Different Laser Sources and Required Dissociation Energies for Several Chemical Bonds

Laser	Wavelength λ (nm)	Photon energy E (eV)	Chemical bond	Bond energy E (eV)
CO_2	10600	0.12		
Nd:YAG	1064	1.16		
XeF	351	3.53	Si–Si, Cl–Cl	1.8–3
XeCl	308	4.03	C–N, C–C	3–3.5
Nd:YAG 4th harm.	266	4.65		
KrF	248	5.00	C–H, O–H	4.5–4.9
KrCl	222	5.50		
ArF	193	6.42		
F_2	157	7.43	C=C	7

Table 13.10 Ablation Parameters for Drilling 100-μm Holes

Type of material	Wavelength (nm)	Fluence (J/cm^2)	Ablation rate per pulse (μm)
Polycarbonate	248	4	0.4
Polyester	248	4	0.8
Polyethylene	248	3.7	1.0
	193	6	0.4
Silicone rubber	308	10	1.5
	248	30	0.2
Kapton foil	308	10	1.2
Plexiglass	193	1	0.3
Hostaform	248	2.8	0.6

13.7.3 Typical examples of Laser micromashining Applications

Laser Microdrilling

Laser drilling of small holes is a widespread application. Holes are drilled in hard materials such as metals, ceramics, or diamond, in softer materials for microelectronics or medical purposes, and also in plastics to perforate foils or for ventilation. Two different techniques are used: direct focusing of the beam to the desired (small) diameter of the hole and alternatively by imaging a mask. Example of applications are drilling holes in wire drawing Dies is presented in Fig. 13.25.

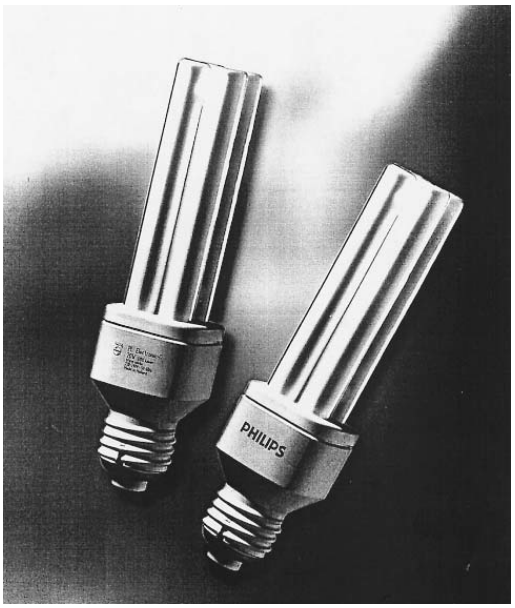


Figure 13.25 Diamond wire drawing die ‘‘drilled’’ by a *Q*-switched Nd:YAG laser. Wire opening 50 μm.

Figure 13.26 Laser surface modification: laser power 9 W, pulse length 120 ns.

Laser Microwelding

Although the technique differs from micromachining a brief discussion of laser microwelding is appropriate as it is probably the most common application. It is employed in the massproduction of electronic products such as television sets, quartz lamps, and other consumer products.

Laser Microadjustment

The basis of laser adjustment is the generation of thermal mechanical stresses in metal structures. An example is the fine adjustment of an audio head enabling precision adjustment in a short time Laser Surface Structuring. Laser surface structuring and texturing are applied to obtain special surface effects on molds and dies, for instance, to obtain cosmetically attractive surfaces on plastic (consumer) products. *Q*-switched Nd:YAG and excimer lasers are used. Products can also directly be textured, marked, or colored b lasers. An example is given in Figure 13.26. A controlled carbonization causes the required grey color. The text is written with a speed of 1000 mm/s.

Chip shaping of light-emitting diodes to improve light extraction

Light extraction from GaN-based light-emitting diodes is seriously suppressed by total internal reflections within the semiconductor layers. With a high refractive index around 2.4, light extraction from the top surface is limited within a 23° emission cone as depicted in Figure 13.27 (c). One effective method to enhance light extraction is employing tilting sidewalls such as

those in a truncated pyramid (TP) LED, where the conventionally confined light rays are extracted from the top surface via sidewall reflections that redirect the light ray into the top surface emission cone. Accordingly, top surface emission of a laser fabricated TP LED shown in Figure 13.27 (b) is particularly stronger compared to the conventional rectangular chip in (a). The overall light extraction can be enhanced by 85%. The improvement is attributed to the additional indirect light extraction from top surface via sidewall reflections.

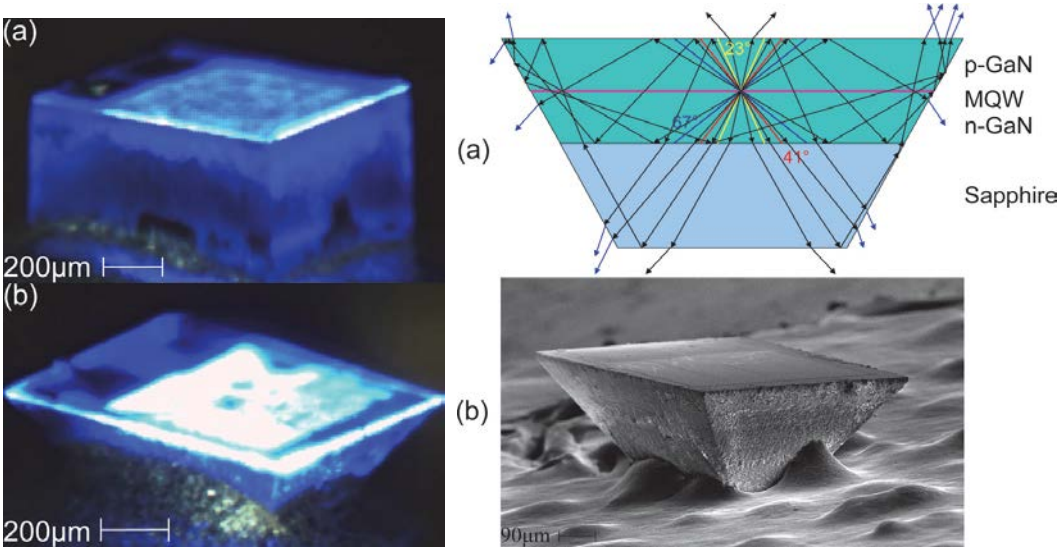


Fig. 13.27. Optical micrographs of (a) conventional cuboid LED and that of (b)truncated pyramid LED with tilting sidewall,(c)shematic diagram of ehanced top surface light extraction via sidewall reflections. (d) SEM image of the truncated pyramid LED chip shaped by laser micromachining.

Additional indirect light extraction also exists in a triangular LED, making it unique among polygonal LEDs. However, the mechanism is slightly different with that of a TP LED. In a triangular LED chip, enhanced light extraction is due to indirect light extraction from the sidewall via reflections on neighbouring sidewalls, while in the case of a rectangular chip or other polygons, the indirect extraction is trivial. Actual chip geometry from triangle to heptagon are fabricated with the laser micromachining system and shown in Figure 13.28.

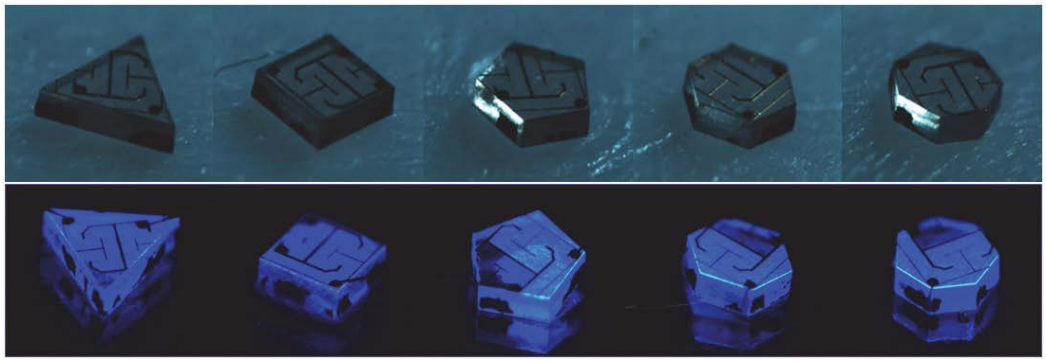


Fig. 13.28. Optical Micrograph Polygonal LEDs as fabricated by laser beam (upper row) and biased at 2.5 V (lower row).

Device isolation on GaN-on-sapphire wafer via laser micro-patterning

GaN is the major material for the fabrication of state-of-the-art blue light-emitting diodes. It is conventionally grown on sapphire substrates by metalorganic chemical vapour deposition (MOCVD), since sapphire is stable and can withstand the high temperature during the growth process. Although there are many issues involved with sapphire, such as lattice mismatch with GaN and poor heat conductivity, sapphire is still prevalent in the fabrication of low-power blue and white LEDs. In addition, being an electrical insulator, sapphire does not interfere with the current conduction in GaN. By selectively removing certain area of GaN, the GaN layer can be separated into multiple electrically isolated small-area LEDs. These LEDs can be connected together by metal interconnects at a later stage, allowing a variety of integrated optoelectronic circuits to be developed.

13.8 Femtosecond Laser Processing

13.8.1 Peculiarities of Femtosecond Laser irradiation

The rapid development of femtosecond lasers over the past few decades has opened up new doors for materials processing since such lasers have many advantages over conventional pulsed lasers (i.e., nanosecond lasers): 1) Femtosecond lasers have significantly lower ablation thresholds than nanosecond lasers, 2) due to nonlinear absorption (i.e. multiphoton absorption at extremely high peak intensities) the femtosecond lasers can induce strong absorption even in materials that are transparent to the laser wavelength such as NaCl and polytetrafluoroethylene (PTFE), 3) femtosecond laser processing reduces heat diffusion to surrounding regions of the processed area, resulting in highquality microfabrication of soft materials such as biological tissues and hard or brittle materials such as semiconductors and insulators., 4) This suppression of heat diffusion to the surroundings also improves the spatial resolution of nanoscale processing, 5) femtosecond laser irradiation at intensities near the ablation threshold produces nanoripple structures on various materials with periodicities much shorter than the irradiation wavelength.

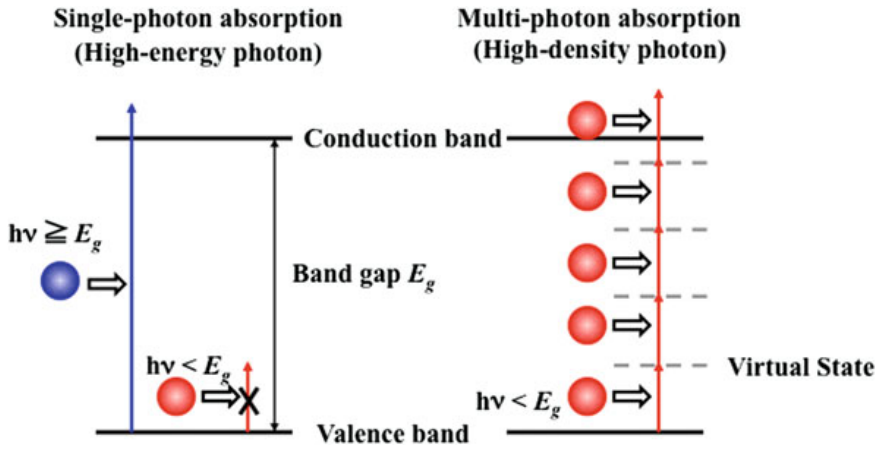


Fig. 13.29 Electron excitation in materials by single and multiphoton absorption

Multiphoton absorption permits not only surface modification but also three dimensional (3D) internal microfabrication of transparent materials such as glass and polymers. The ability of femtosecond lasers to directly form microstructures inside glass makes them suitable for fabricating microfluidic and optofluidic devices. Figure 13.29 depicts single and multiphoton absorption based on electron excitation. Conventional absorption is linear single-photon absorption. When light whose photon energy exceeds the band gap of a material is incident on the material, it is absorbed and a single photon excites an electron from the valence band to the conduction band. On the other hand, light whose photon energy is smaller than the band gap cannot excite electrons, so that no absorption occurs in the stationary state. However, when an extremely high density of photons is incident on the material, an electron can be excited by multiple photons, even when the band gap exceeds the photon energy. This process is known as multiphoton absorption. The extremely high density of photons required to induce multiphoton absorption can be easily obtained by femtosecond lasers due to their ultrashort pulse widths.

Besides, at high laser intensities and low frequencies, electron excitation is induced by tunneling ionization rather than multiphoton absorption. In tunneling ionization, the potential barrier formed by the valence and conduction band structures is first drastically deformed by the intense electric field of a femtosecond laser and then the barrier length is reduced. When this occurs, an electron can tunnel through the barrier and eventually easily escape from the molecule to be excited from the valence band to the conduction band.

Femtosecond lasers can suppress heat diffusion to the surrounding of the processed area, which is advantageous since it gives higher spatial resolutions (fig.13.30). When a 10-ns laser pulse irradiates Cu with a spot size equal to the laser wavelength (typically several hundred nanometers to 1 μm), the processed region becomes larger than the spot size due to the thermal diffusion length of 1.5 μm. In contrast, since there is almost negligible thermal diffusion in femtosecond laser irradiation, the processed region is unlikely to extend beyond the spot size.

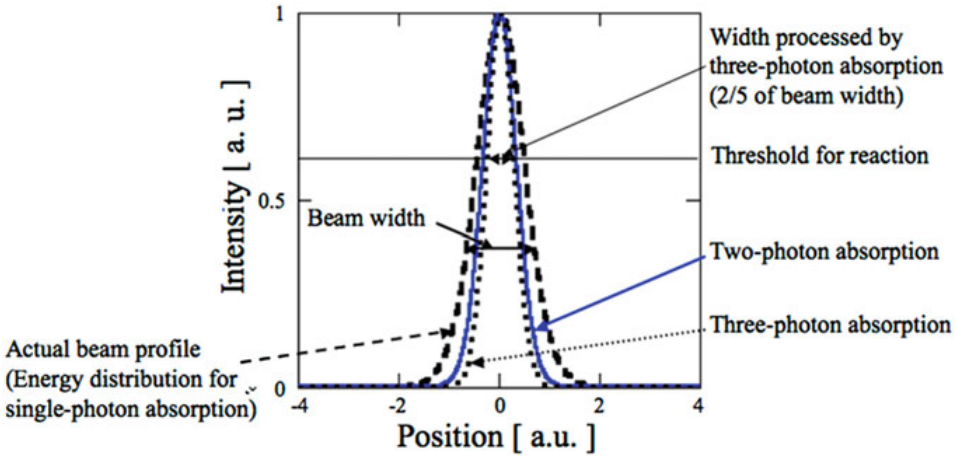


Fig. 13.30 Actual beam profile (thick dashed line) and spatial distributions of laser energy absorbed by transparent materials by two (solid line) and three (thin dashed line) photon absorption. The solid horizontal line indicates the reaction threshold.

Interaction of Femtosecond Laser Radiation with Glass Femtosecond laser irradiation induces electron excitation and relaxation processes in glass, as depicted in Fig. 13.31. Electrons are initially excited from the valence band to the conduction band by multiphoton absorption or by tunneling ionization. For relatively low laser intensities of the incident femtosecond laser beam, the generated free electrons contribute to photochemical reactions (e.g., photoreduction of ions doped in glass resulting in precipitation of atoms). At higher intensities, excited electrons can sequentially absorb several photons in the same laser pulse so that they are excited to higher energy states for which free carrier absorption is efficient. This sequential excitation is termed electron heating and it results in non-thermal bond breaking.

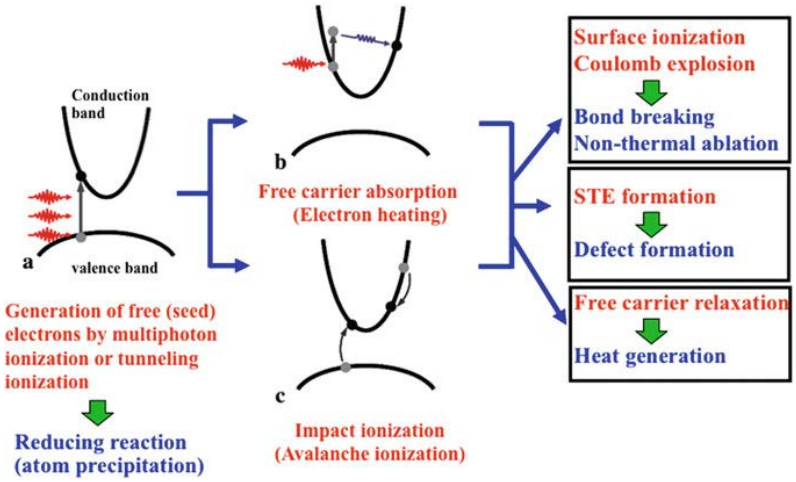


Fig. 13.31 Electron excitation and relaxation processes in glass induced by femtosecond laser irradiation. Only multiphoton absorption is shown for the initial excitation of free electrons

13.8.2 Femtosecond-Laser-Assisted Wet Chemical Etching

Microfluidic devices have undergone rapid development in the past two decades, enabling systems for chemical and biological analysis to be miniaturized. A wide variety of microfluidic devices have been fabricated for controlling and manipulating tiny volumes of liquids with high precision and ease of operation. The most popular microfluidic fabrication technology is soft lithography using poly (dimethylsiloxane) (PDMS) substrates. Although soft lithography is rapid and cost effective, it cannot be used to directly form three dimensional (3D) microfluidic structures such as buried microchannels and microchambers without stacking and bonding. This difficulty can be overcome using the advanced femtosecond laser 3D microprocessing techniques.

Figure 13.32 schematically illustrates the process used to fabricate 3D microfluidic structures in Foturan glass: formation of a latent image by femtosecond laser direct writing (Fig. 13.32a), transformation of the latent image into an etchable phase bythermal treatment (Fig. 13.32b), and removal of the modified material by wet chemical etching in a 5–10 % aqueous solution of HF acid in an ultrasonic bath (Fig. 13.32c). The ultrasonic bath is critical because it can significantly enhance the etch rate bysimply increasing the mass transfer of the chemical etchant in the thin channel.

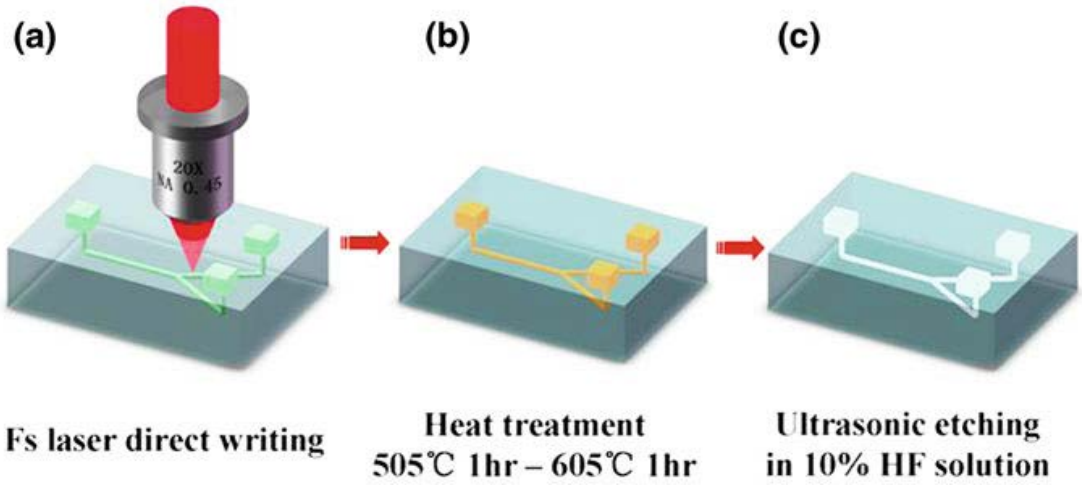


Fig. 13.32 Schematic depiction of the processes for fabricating 3D microfluidic structures in Foturan glass.

Figure 13.33a and b respectively show micrographs of the glass substrate surface and microchannels embedded in glass of a typical microfluidic mixer fabricated in Foturan glass. The two microreservoirs on the right-hand side serve as solution inlets while the horizontal hole on the left-hand side serves as the solution outlet. To perform mixing, two different liquids are introduced into the micromixer.

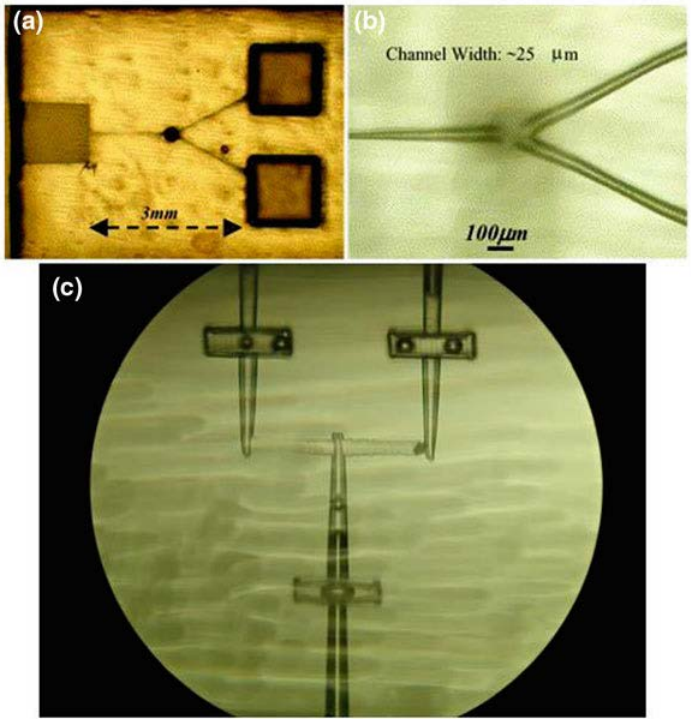


Fig. 13.33 Optical micrographs showing (a) top view of a micromixer, and (b) close-up view of microchannels embedded in glass. C) Optical micrograph of a complex 3D microfluidic structure with microchannels and chambers arranged in a multilayer configuration

A new strategy has recently been developed for fabricating microchannels with nearly unlimited lengths and arbitrary geometries by femtosecond laser direct writing in mesoporous glass immersed in water followed by post-annealing, realizing long square-wave-shaped microchannels and large-volume microfluidic chambers. Figure 13.34a and b respectively show a schematic view of the experimental setup and a flow diagram of the fabrication process. The main fabrication process involves two steps: (1) direct formation of hollow microchannels in the porous glass substrate immersed in water by femtosecond laser ablation and (2) post-annealing at 1150 ^{o}C to consolidate the porous glass substrate. As a demonstration, a square-wave-like microchannel with a total length of 14 mm and a diameter of 64 μm has been fabricated at a depth of *250 μm below the glass surface (see Fig. 13.34c). The channel diameter has an excellent homogeneity. Using this technique, a passive microfluidic mixer consisting of 3D mixing units has been constructed. The 3D micromixer has a higher mixing efficiency than a 1D micromixer.

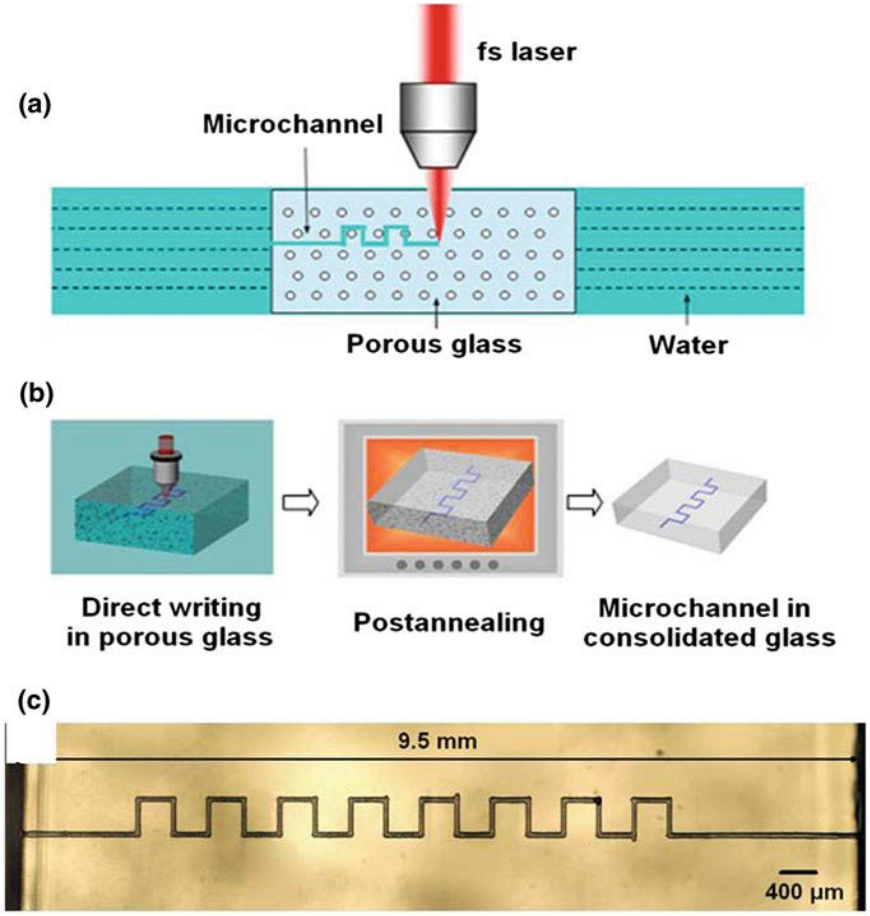


Fig. 13.34. a Schematic diagram of experimental setup, and (b) flow diagram of fabrication process. C) Optical micrograph of a square-wave-like microchannel fabricated in glass

Chapter 14. Nanotechnologies.

V.Ilchenko

14.1. What is nanotechnologies?

Nanoscience is the study of objects that are very small, between 1 and 100 nanometers, about 80,000 times smaller than the width of human hair. Nanotechnology has been applied to numerous industries such as electronics, medicine, food, energy, and even in clothing. A fundamental operation in nanoscience is the separation and concentration of nano-sized objects.

"Nanotechnology" has become a word around which a remarkable range of science and , engineering is being organized. Why has it emerged into the limelight, while other areas of technology that might have as much potential ("intelligent machines", the biological/ computational interface, sustainable development, and others) have not? The answer to this question is complicated, with components of economic necessity, scientific opportunity, and public engagement. Some of the more technical aspirations of nanotechnology are these:

14.1.1. Characterization of the nanostructures.

A key aspiration of nanotechnology is to demonstrate the proposition that as things become small, they become different. "Different" often translates into "interesting", and sometimes into "useful and valuable". There are a number of demonstrations of the emergence of differences in nanostructures. Although buckyballs (among the first of the synthetic nanostructures to catch the interest of both the technical community and the public) have so far fallen in the category of "interesting but not especially useful", buckytubes (or carbon nanotubes) have genuinely remarkable properties: especially their high electrical conductivity and unique mechanical strength. The properties of surfactant-stabilized colloids are the basis for many bioanalytical systems. Fluorescent CdSe quantum dots, unlike fluorescent organic dyes, photobleach only slowly (or not at all), and show interesting (if annoying) optical phenomena such as "blinking". SAMs provide an unequaled ability to tailor the properties of surfaces.

Quantum behavior becomes increasingly prominent as structures become smaller. Many of the behaviors of atoms and molecules are, of course, only explicable on the basis of quantum mechanics. The properties of objects and structures larger than a few microns are usually classical. In the intermediate region-the region of nanometer-scale structures- quantum and classical behaviors mix. This mixture offers the promise of new phenomena and/or new technologies. The fluorescent behavior of semiconductor quantum dots can only be explained quantum mechanically; as can the tunneling currents that characterize scanning tunneling microscopes, and electron emission from the tips of buckytubes. response of electrical resistance

to magnetic field in GMR materials is already useful in magnetic information storage**,** and the behavior of spin-polarized electrons in magnetic semiconductors forms one foundation for the emerging field of spintronics. The ability to make structures in the region where quantum behavior emerges, or where classical and quantum behaviors merge in new ways, is one with enormous opportunity for discovery. And because quantum behavior is fundamentally counterintuitive, there is the optimistic expectation that nanostructures and nanostructured materials will found fundamentally new technologies.

The argument for the development of nanotechnology for use in the microelectronics industry is clear. Information technology (IT) has been the technology that has most changed society in the last 50 years. Its development has not yet subsided, although the progression of dimensions to ever-smaller sizes -described by Moore's Law- must inevitably come to an end when these dimensions reach the size of molecules and individual atoms. In between the current structures (-100 nm) and the minimum size limit (-1 nm), developments of immense economic importance are inevitable. Beyond evolutionary developments in silicon-based technology, there are a host of possibilities in new, but not necessarily fundamentally different, technologies for IT. Will there be important technologies built around organic semiconductors? Will it be possible and practical to take advantage of the high mobility of electrons in semiconducting buckytubes to make new electronic devices? Will some combination of bottom-up synthesis of monodisperse, magnetic colloids, surfactant-assisted crystallization, and materials fabrication generate practical, ultradense magnetic information storage media? Is there a way to use self-assembly of small circuit elements as the basis for a new strategy for fabricating microprocessors, mass storage devices, or displays? It is too early to judge the practical importance of these areas, although they are showing technical feasibility in demonstrations in research laboratories. In the longer term loom the potentially revolutionary technologies, to which nanostructures and nanomaterials may make a contribution: quantum computing, computing using cellular automata, photonic computing, semiconductor/biological hybrid computing, molecular electronics and others. The history of predicting revolutions is poor, and it is likely that any revolution will emerge from an unexpected direction.

Since, however, nanostructures constrain electrons and photons in new ways, and since nanostructures have been difficult to fabricate, and hence are still relatively unexplored, the possibility that a revolution in IT will appear, unexpectedly, from the exploration of some area of nanoscience is higher than it might be in more familiar areas. The relation between manufacturing and nanotechnology is less explored than that in many apparently higher-technology areas. Manufacturing is a field that permeates technology: any successful technology

must be transferred from its developer to its users, and most technologies (even software is ultimately housed in hardware) require manufacturing something. There are complex but important relations between nanoscale features of manufacturing systems, the cost of manufacturing processes, and the performance of manufactured objects. Would the performance of ball-bearings in a heavy-duty transmission improve if there were no defects larger than a few nanometers? What would robotic assembly systems be like if every part were identical to within a few nanometers? It is not possible, at present, to answer these and related questions, since nanoscience is just beginning to generate the tools and metrologies necessary to explore them. Nanoscience does, however, have the potential to make important contributions to future manufacturing systems. At the foundation, underlying the technologies, is the fundamental science of phenomena at the nanoscale. Nanoscience, in its broad sense, is a new area. We do not know what will develop, but we do know that in order for it to develop, it must have materials, procedures, and tools. Developing new ways of manipulating matter at the smallest scales-scales that bridge between atoms and molecules (chemistry) and mesoscopic matter, (materials science)-is a centrally important part of fundamental scientific inquiry.

Nanotechnology promises to deliver faster and smarter devices where the dimension of an individual component can approach the size of atoms and molecules. Achieving the goals of nanotechnology requires nanofabrication techniques with an ability to generate building blocks with at least one dimension less than 100 nm. Obviously, the description can be generalized for the case of 'confinement' in more than one dimension. One subdivides:

- confinement in one dimension: quantum films
- confinement in two dimension: quantum wires
- confinement in three dimension: quantum dots.

In quantum dots both electrons and holes can have only discrete energy values. Therefore, these systems are often denoted as ***artificial atoms***.

Compared to natural atoms the energy spectrum can be designed by changing the structure. The following figure 14.1. summarizes the different confinement regimes.

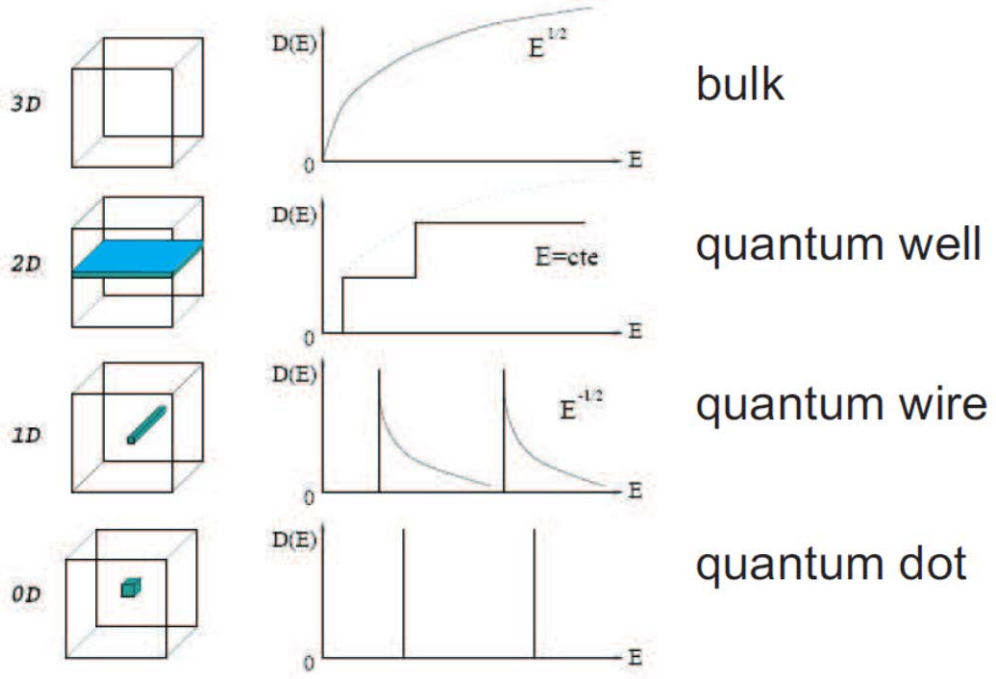


Fig. 14.1. Different confinement regimes and density of states $D(E)$ for electrons in bulk, quantum wells, quantum wires, and quantum dots.

14.2. Fabrication methods.

For the fabrication principles in nanotechnology, one distinguishes between socalled top-down and bottom-up approaches. Top-down approaches seek to create nanoscale devices by using larger, externally-controlled tools to direct their assembly. Here, often traditional workshop or microfabrication methods are used to cut, mill, and shape materials into the desired shape and order. Micropatterning techniques, such as photolithography and inkjet printing belong to this category. Bottom-up approaches, in contrast, seek to have smaller (usually molecular) components built up into more complex assemblies. These techniques use the chemical properties of single molecules to cause single-molecule components to (a) self-organize or self-assemble into some useful conformation, or (b) rely on positional assembly.

These approaches utilize the concepts of molecular self-assembly and/or molecular recognition.

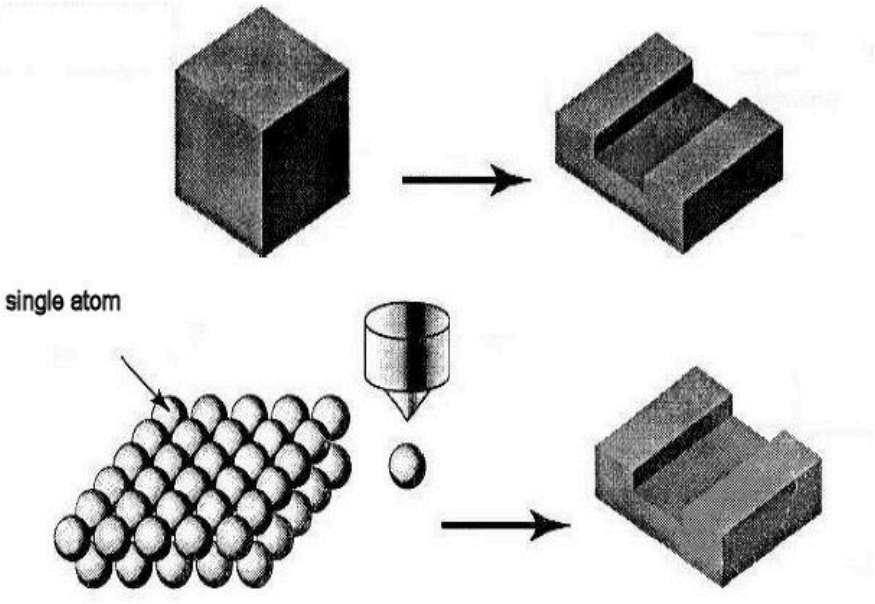


Fig. 14.2. Top-down vs. bottom-up technologies.

14.2.1 Top-down process.

Photolithography has been the method of choice for patterning the ever-decreasing feature sizes required in the semiconductor industry. However, the cost of such equipment is rapidly approaching $50 million as it is advanced to the nanofabrication regime, in part due to the momentum in industry driven by Moore's law. Hence, there are growing efforts in developing alternative techniques for highthroughput and low-cost nanomanufacturing. In practice, nanofabrication embraces two steps—the generation and replication of the patterned nanostructures. Conventional techniques use light or electrons to generate patterns through processes that begin with designing a computer-aided design (CAD) file and end with transferring patterns from the design to an array of small features on the surface of a substrate. Unconventional techniques, those reminiscent of macroscopic molding, embossing, printing, and skiving technologies, provide the ultimate, lowcost solutions to replicate nanostructures with potential applications in diversified research areas that are weakly connected to engineering. This paragraph of the chapter provides a brief overview of various techniques in nanofabrication with a focus on the **top–down techniques**. Our goal is to illustrate the different operational and mechanistic principles and shed light on their practices, limitations, and potential applications.

Electron Beam Lithography (e-beam lithography or EBL) refers to a lithographic process that uses a focused beam of electrons to form the circuit patterns needed for material deposition on (or removal from) the wafer, in contrast with optical lithography which uses light. Electron lithography orders higher patterning resolution than optical lithography because of the shorter wavelength possessed by the 10-50 keV electrons that it employs. Given the availability of technology that allows a small-diameter focused beam of electrons to be scanned over a surface, an EBL system doesn't need masks anymore to perform its task. An EBL system simply 'draws' the pattern over the resist wafer using the electron beam as its drawing pen. Thus, EBL systems produce the resist pattern in a 'serial' manner, making it slow compared to optical systems. The following figure 14.3 shows the major components of an e-beam system:

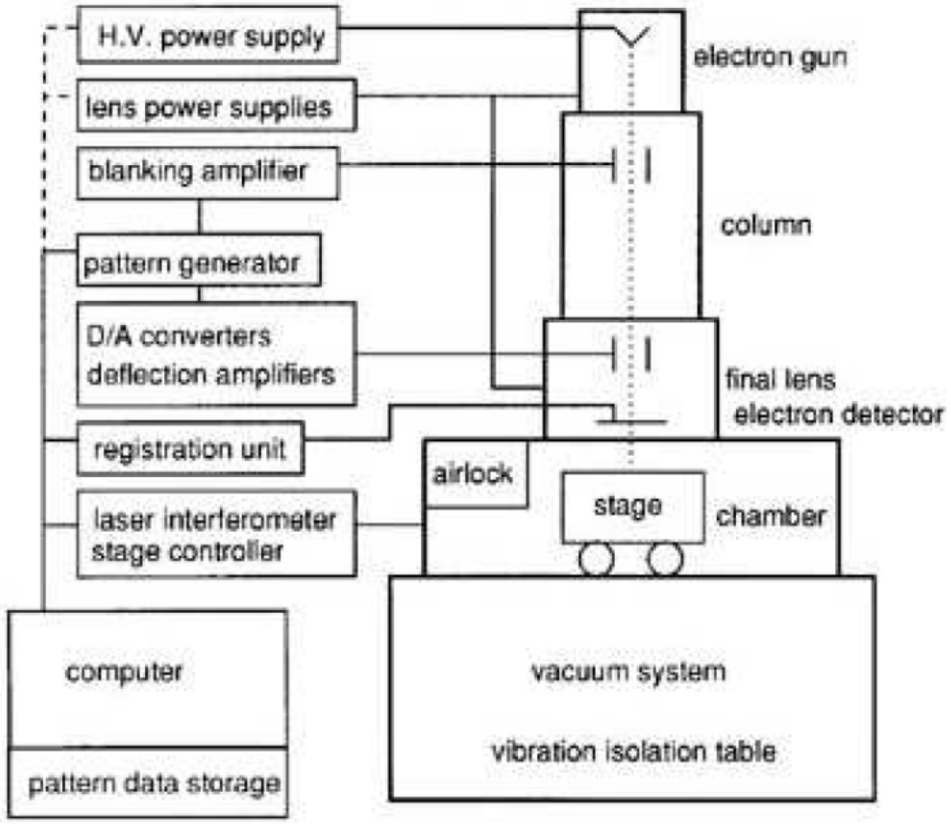


Fig. 14.3. Block diagram showing the major components of a typical electron beam lithography system.

Mainly, the developing of practical EBL drawing system has been started since 1960s, and fine pattern formation has also been studied together with the system development. Regarding EBL-drawn pattern size, at first, micron and submicron-sized pattern has been drawn on mask blank and directly on the device. Today, the pattern size miniaturizes to nanometer-size of less than 20

nm in research. Especially, we have to be focused the EBL into the possibility to form fine dot and fine pitched dot 2-dimensional arrays for patterned media and quantum devices.

A typical EBL system consists of the following parts:

- an *electron gun or electron source* that supplies the electrons
- an *electron optics* that 'shapes' and focuses the electron beam
- a *mechanical scanning stage* that positions the wafer under the electron beam
- a *wafer handling system* that automatically feeds wafers to the system and
- unloads them after processing
- a *computer system* that controls the equipment.

Just like optical lithography, electron lithography also uses positive and negative resists, which in this case are referred to as electron beam resists (or e-beam resists). E-beam resists are e-beam-sensitive materials that are used to cover the wafer according to the defined pattern. A common resist is polymethyl methacrylate (PMMA).The resolution of optical lithography is theoretically limited by di®raction (wave-length of the electrons on the order of 0.2-0.5 angstroms). However, main constraints are due to the following effects:

- electron scattering

  During electron beam lithography, scattering occurs as the electron beam interacts with the resist and substrate atoms (see figure 14.4). Scattering broadens the diameter of the incident electron beam and gives the resist unintended extra doses of electron exposure. Both effects result in wider images than what can be ideally produced from the e-beam diameter. In fact, closely-spaced adjacent lines can 'add' electron exposure to each other, a phenomenon known as p*roximity effect*.

- resist swelling

  Resist swelling occurs as the developer penetrates the resist material. The resulting increase in volume can distort the pattern, to the point that some adjacent lines that are not supposed to touch become in contact with each other.

- aberrations

  Compared to optical systems electron beam optics su®ers from enhanced aberrations. As in optics this increases the size of the smallest available spot an thus the resolution.

The last figure 14.4 of this section provide an example of complex integrated circuits written by electron beam lithography.

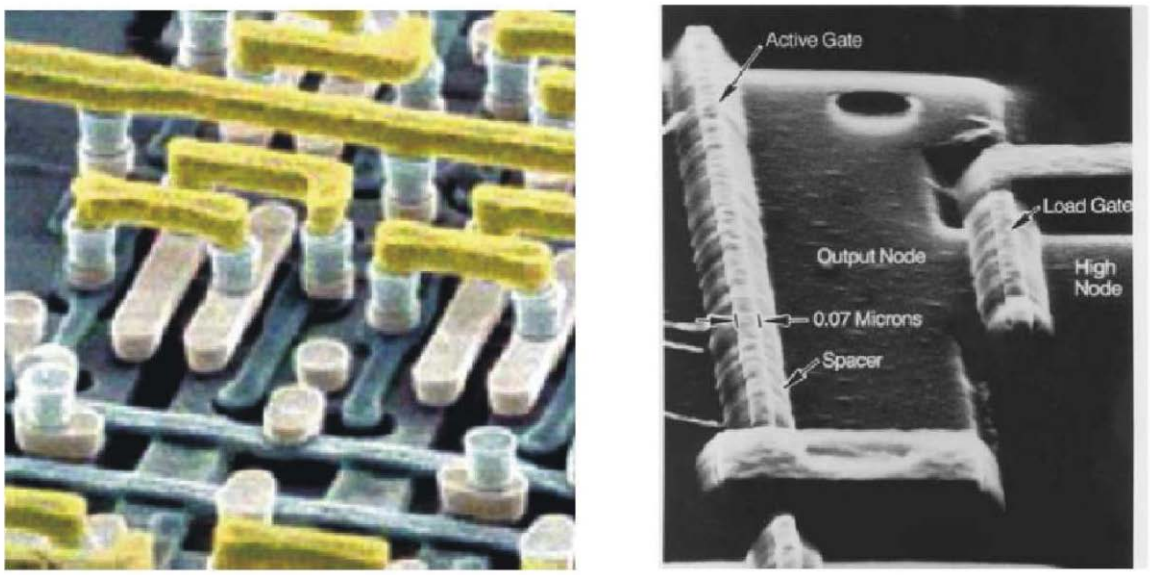


Fig 14.4. Examples of two complex structures written by e-beam lithography.

Recently, EBL has been applied to mask and reticle pattern draw, for fabricating semiconductor devices, and nanometer-sized pattern direct writing for developing of new concept nano-device.

14.2.2 Bottom-up process.

Yesterday we wrote about air bridges in nanotechnology fabrication. Today we show a practical example. Traditionally, electronic devices have been fabricated by top-down fabrication methods. Conducting polymers, for instance, have been synthesized as micro- and nanoscale fibers, tubes and wires for more than 10 years now. More recently, nanowires have been integrated into electronic circuits, making possible the development of devices such as polymer nanowire chemical sensors with superior performance. What most of these fabrication techniques have in common is that they are template-based (e.g. lithography or DNA templates) or depend on specialized fiber forming techniques such as electrospinning. However, as electronic components become smaller and smaller it is increasingly more difficult to use existing methods of fabrication. New methods must be developed. A group of researchers in Australia have demonstrated a technique for growing ordered polymer nanowires within a pre-patterned electronic circuit such that electrical contacts to the nanowires are made *in situ* during the growth procedure, avoiding the time-consuming and challenging task of manipulating nanowires into position and making electrical contacts post-synthesis. *Bottom-up fabricating method* can be usually based on epitaxial growth.

Epitaxial techniques are today the mostly utilized techniques to create solid-state nano-scaled devices. The term *epitaxy* comes from a Greek root (*epi* “above” and *taxis* “in ordered manner”) and refers to the method of depositing a mono-crystalline Film on a mono-crystalline substrate. The deposited film is denoted as epitaxial film or epitaxial layer.

Epitaxial films may be grown from gaseous or liquid precursors. Because the substrate acts as a seed crystal, the deposited film takes on a lattice structure and orientation identical to those of the substrate. This is different from other thin film deposition methods which deposit polycrystalline or amorphous films, even on single-crystal substrates. If a film is deposited on a substrate of the same composition, the process is called homo-epitaxy; otherwise it is called hetero-epitaxy. The latter, is often applied to growing crystalline films of materials of which single crystals cannot be obtained and to fabricating integrated crystalline layers of different materials, such as quantum dots.

Here we discuss two major epitaxial techniques, Molecular Beam Epitaxy (MBE) and Metal-Organic Chemical Vapour Deposition (MOCVD).

Molecular Beam Epitaxy (MBE) was invented in the late 1960s at Bell Telephone Laboratories by J. R. Arthur and Alfred Y. Cho. The growth process takes place in an ultra-high vacuum environment ($< 10^{-9}$ mbar), such that the mean free path of the particles is larger than the geometrical size of the chamber. The term "beam" means that evaporated atoms do not interact with each other or with vacuum chamber gases until they reach the wafer. Figure 14.5 shows a typical MBE chamber. Ultra-pure elements such as gallium and arsenic are heated in separate effusion cells until they begin to slowly sublimate from the solid or evaporate from the liquid phase. The effusion cell temperature is typically used to control the flux of the atomic beam. For some molecules/atoms (e.g. nitrogen), gas cells are used as well. The composition of the different materials is controlled, using mechanical shutters in front of the output ports of the cells. The gaseous elements then condense on the substrate, where they may react with each other. In order to obtain a high mobility of the adatoms, the substrate is heated to high temperatures (typ. ≈300° C for II-VI materials, such as ZnSe, and ≈600 °C for III-V materials, such as GaAs). In order to maintain the crystal structure, the substrates have to be mono-crystalline wafers with carefully cleaned surfaces.

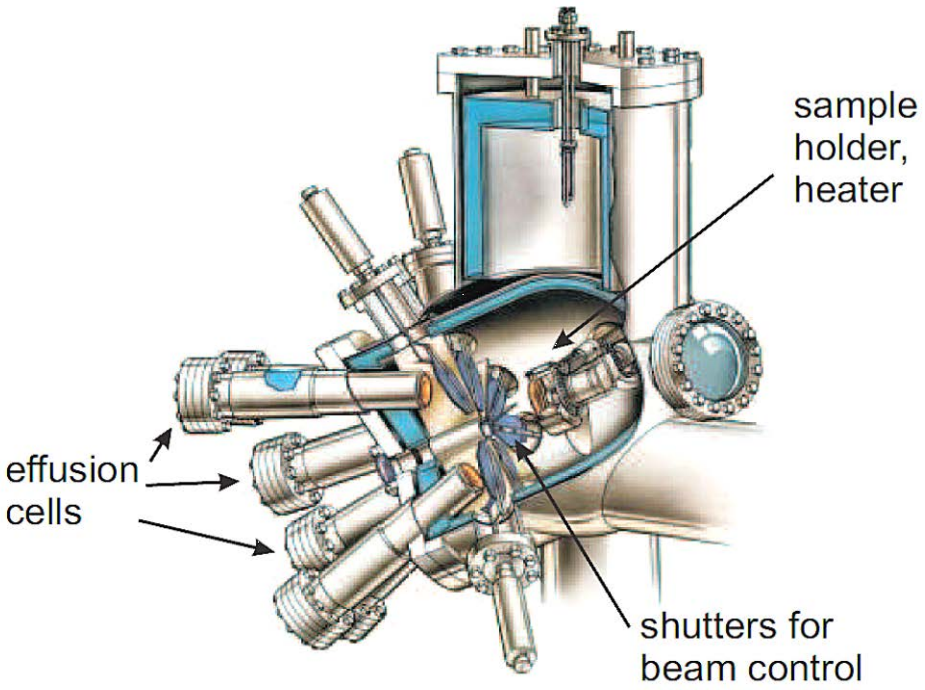


Fig. 14.5 Schematics of an MBE chamber.

Figure 14.6 shows the different processes which the adatoms undergo during growth. Nucleation of atoms can take place on mono-atomic steps, on defects, or directly on the surface. The nucleation process is in competition with sublimation of atoms from the surface, depending on the substrate temperature and the molecular beam flux. During operation, RHEED (Reflection High Energy Electron Diffraction) is often used for monitoring the growth of the crystal layers. This technique uses an electron beam with energies of some ten keV. The electrons are accelerated in an electron gun and are focussed with a shallow angle on the sample surface. Incident electrons diffract from atoms at the surface of the sample, and a small fraction of the diffracted electrons interfere constructively at specific angles and form regular patterns on the detector. The diffraction pattern is monitored on a fluorescing screen.

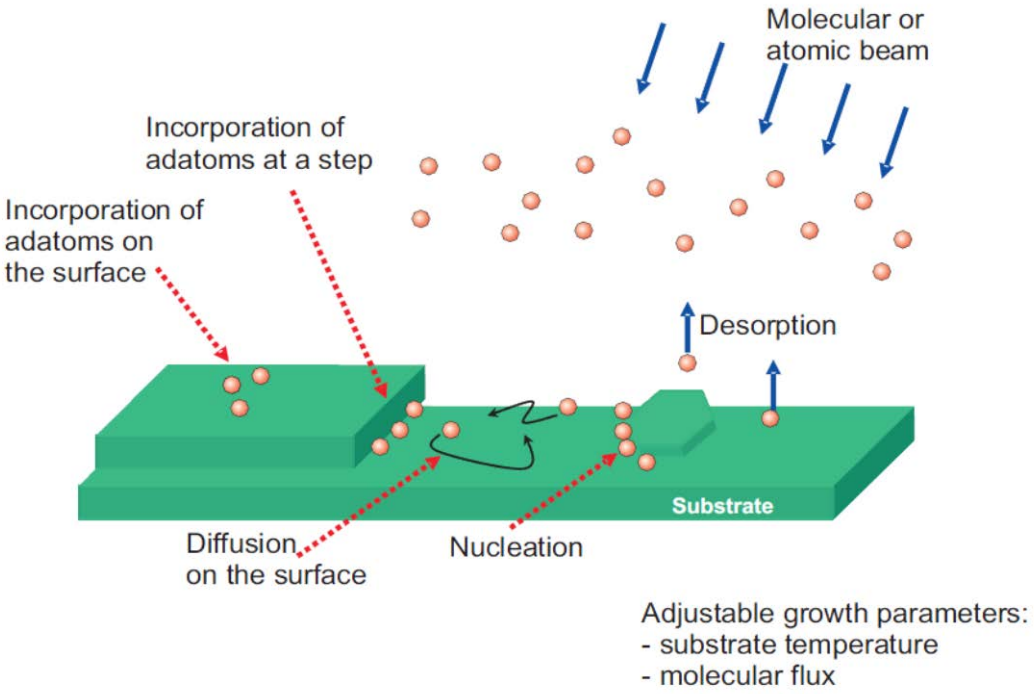


Fig. 14.6. Different processes during the MBE growth.

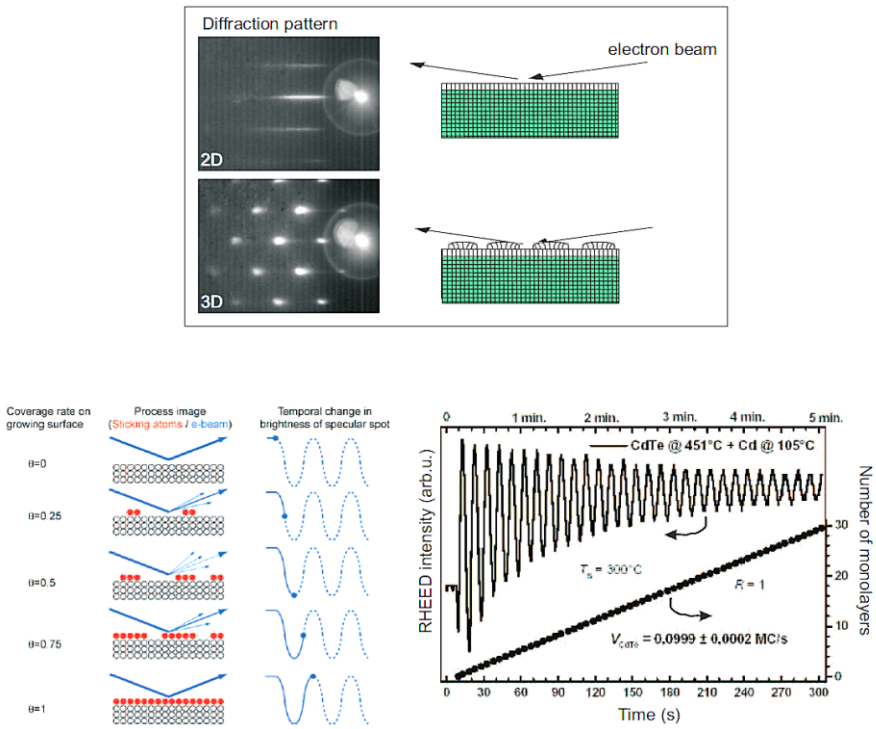


Fig. 14.7. Top: RHEED patterns for a °at/2D surface and a structured/3D surface. Bottom left: principle of RHEED oscillations. Bottom right: experimental RHEED oscillations and number of monolayers for CdTe growth.

The diffraction pattern provides various information about the surface structure, as the penetration depth of the electrons is only few monolayers. Figure 14.7 (top) compares the pattern of a flat surface (stripes) with that of a roughened surface, e.g. due to quantum dot formation (spots). The number of stripes/spots indicates the crystal structure of the surface (so-called surface reconstruction), while the distance between neighbouring stripes is inverse proportional to the lattice constant at the surface. Finally, the intensity variation of the stripes allows a direct counting of the number of deposited monolayers (figure 14.7, bottom). Here one makes use of the fact, that at a semi-finished monolayer, increased scattering of the electrons appears and consequently the intensity of the directly reflected e-beam is reduced. Monitoring these RHEED oscillations allows to track the material deposition with sub-monolayer accuracy.

Metal-Organic Chemical Vapour Deposition (MOCVD), also known as Metal-Organic Vapour Phase Epitaxy (MOVPE), is a chemical vapour deposition method of epitaxial growth of materials, especially compound semiconductors from the surface reaction of organic compounds or metalorganics and metal hydrides containing the required chemical elements. In contrast to MBE, the growth of crystals is by chemical reaction instead of physical deposition. This takes place not in a vacuum, but from the gas phase at moderate pressures (2 to 100 kPa).

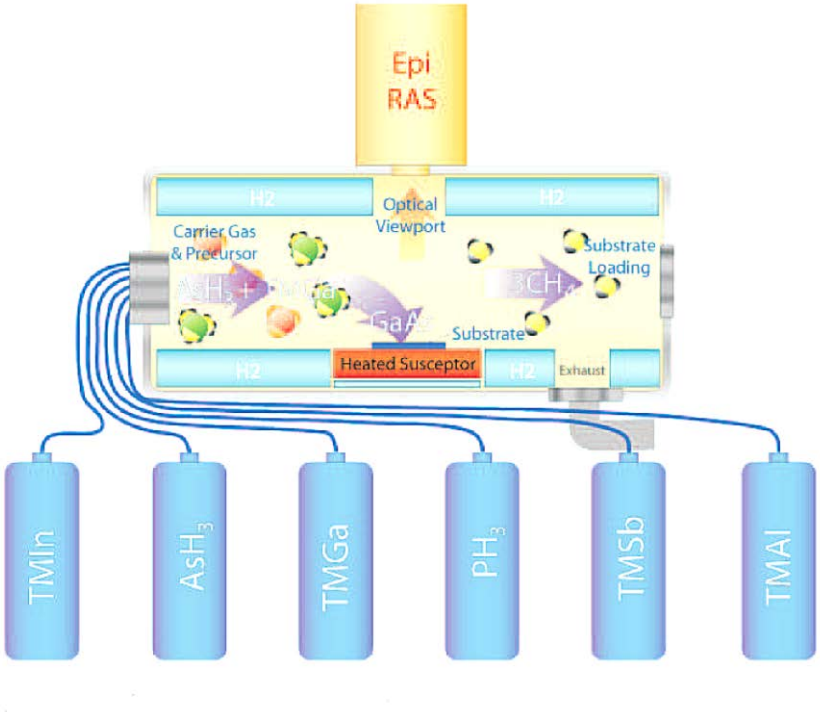


Fig. 14.8. Schematics of MOCVD growth.

The principle of MOCVD on the example of a GaAs film is the following: gaseous compounds of gallium or arsenic are needed as so-called precursors. For As, the arsenic hydride (AsH3) and

for Ga a metal-organic compound such as trimethyl gallium (TMGa) is mainly used, respectively. They are fed into the reactor with the aid of a carrier gas (hydrogen or nitrogen). The reactor contains the substrate (GaAs wafer) which is heated. The temperatures range from about 500 to 1500 °C depending on the material system to be produced. For pure GaAs the chemical bounds of the compounds have to be broken. This already occurs partially in the gas phase due to the heat emitted from the substrate or by collisions with the molecules of the carrier gas. The fragments move to the substrate surface, together with undamaged arsine and TMGa, where they settle and migrate over the wafer. Due to the high temperature and the reactions accelerated by the substrate, additional bonds are split up so that ultimately pure gallium or arsenic can be deposited. In this way, a new GaAs layer grows on the wafer monolayer by monolayer. The remainder of the starting molecules, the methyl groups of TMGa and the hydrogen of arsine, partially combine forming methane. Together with molecules which have not reacted they detach from the surface and are °ushed out of the reactor by the carrier gas stream.

The advantage of MOCVD over MBE is the much increased growth speed up to 1 nm/s, which makes this technique ideal for mass production. Due to the absence of high-vacuum components, MOCVD is comparably inexpensive and easy to maintain. The main expenses are the high-purity precursors and (compared to MBE) the low material efficiency. The main disadvantage of MOCVD compared to MBE is, that MOCVD utilizes elemental compounds. So a comparably large amount of impurities (such as hydrogen, nitrogen, or oxygen) are implanted into the crystal.

A heterostructures is the interface that occurs between two layers or regions of different crystalline semiconductors. These semiconducting materials have in general unequal band gaps. Heterostructures allow *band-gap engineering* which is used to make and optimize the electronic energy bands in many solid state device applications, such as semiconductor lasers, solar cells and transistors to name a few. In heterostructures of very small spatial dimensions, the motion of charge carriers is restricted. They are forced into a quantum confinement regime, leading to formation of a set of discrete energy levels. In this way arbitrary potentials can created for electrons and holes in a heterostructure.

While the bandgap influences the optical and electronic properties, for the growth of mono-crystalline heterostructures, the lattice mismatch $(a1 - a2)/a2$ between the substrate material (lattice constant $a1$) and the overgrown material ($a2$) becomes an important parameter. The growth of non-matched layers (e.g. $a1 = 0.605nm$ for InAs, $a2 = 0.565nm$ for GaAs, $\Delta a/a = 7\%$) introduces a strain, whose energy accumulates with increasing layer thickness. When this energy

exceeds certain critical levels, plastic relaxation occurs, such as the formation of misfit dislocations.

Figure 14.9 shows an example of a defect-free GaAs-AlAs heterostructure. The combination GaAs-Al(Ga)As is of high interest in semiconductor optics, as the small lattice mismatch (0.28%) allows to grow large layers, while they have a comparably large difference in their refractive index (AlAs: n = 3:0; GaAs: n = 3:6 at λ = 900 nm). This makes them ideal candidates for the formation of large monolithic Bragg reflectors.

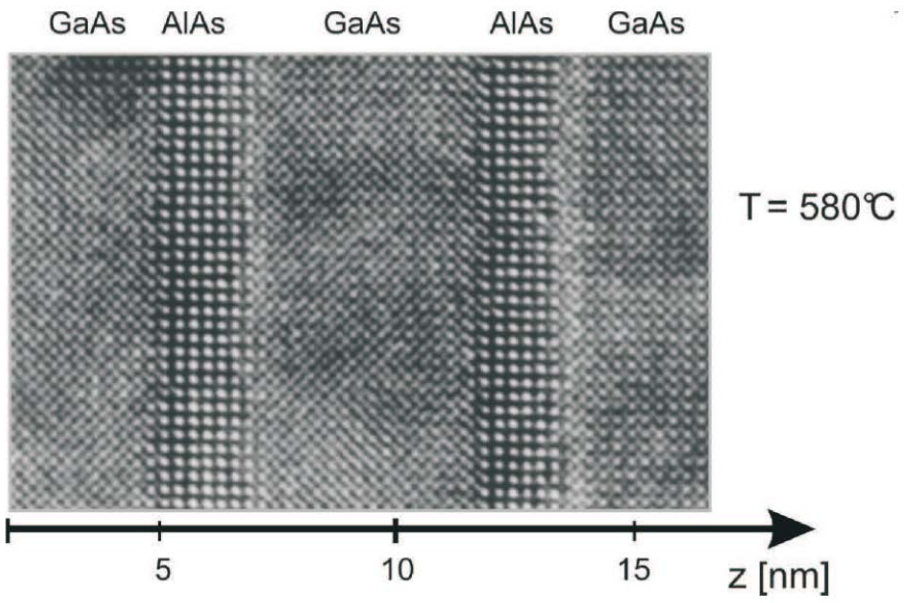


Fig. 14.9. GaAs-AlAs super-lattice grown by MBE.

Quantum dots (QDs) are particularly significant for optical applications due to their high quantum yield and discrete energy level structure. There are several approaches to use quantum dots as light-emitting diodes and as laser active material. In electronic applications they have been proven to operate like a single-electron transistor and show the Coulomb blockade effect. Quantum dots have also been suggested as implementations of qubits for quantum information processing, and as sources for single photon states.

In contrast to the growth of 2-dimensional layers (quantum wells) the formation of quantum dots is less straightforward, as it requires the formation of small islands with sizes on the order of a few 10 nm. Cleaved-edge overgrowth and natural quantum dots In the past, QDs have been processed by starting from higher dimensional semiconductor heterostructures, like etching pillars in quantum well systems or forming intersections of quantum wells or quantum wires. Also the growth of nano-structures on patterned substrates, such as grooves and pyramids led to

successful quantum dot formation. So-called natural quantum dots are formed by width fluctuations mainly of quantum wells. In this environment the *excitons*, i.e., correlated electron-hole pairs, are trapped in broader regions of the quantum well, where the confinement energy is lowered, so that a potential minimum is formed.

For epitaxially grown QDs, the most common technique exploits self-assembly of localized islands. Self-assembled QDs are formed when growing a semiconductor layer on top of a substrate material of smaller lattice constant. Above the critical thickness and under certain conditions, the strain relaxes by forming small islands, where at the surface the QD lattice constant relaxes to its bulk value (see Figure 14.10).

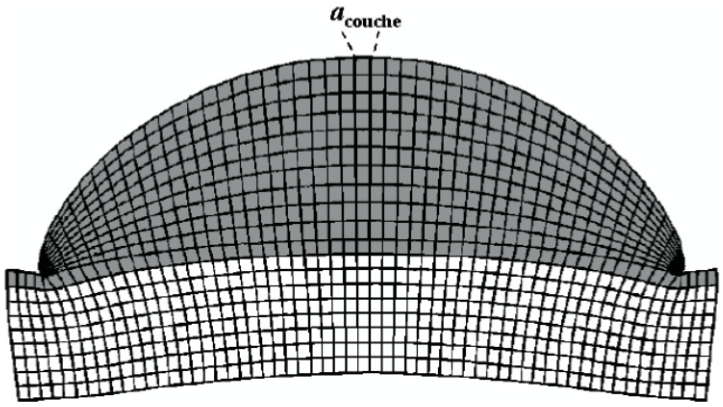


Fig. 14.10. Under certain conditions, strain in lattice-mismatched heterostructures relaxes to form islands, where at the surface the lattice constant of the added material relaxes to its bulk value. This growth mode is called *Stranski-Krastanov growth*. Generally, a thin layer, which is known as the *wetting layer*, will remain, completely covering the substrate. The wetting layer forms a quantum well, which usually shows photoluminescence below the quantum dot emission wavelength.

QDs are finally capped with (usually) the substrate semiconductor material to obtain a high quantum yield, i.e., avoiding non-radiative recombination via surface states. As the emission properties of QDs do not only depend on the material, but also on size and shape, they can be used as optical emitters covering large spectral ranges from UV to IR. The InAs/GaAs system is by far the most studied of all QD systems, emitting in the infrared regime. InGaN quantum dots imbedded in GaN have the potential to cover the complete visible range. Finally II-VI systems such as CdSe/ZnSe or CdTe/ZnTe QDs luminescence in the 500-600 nm regime.

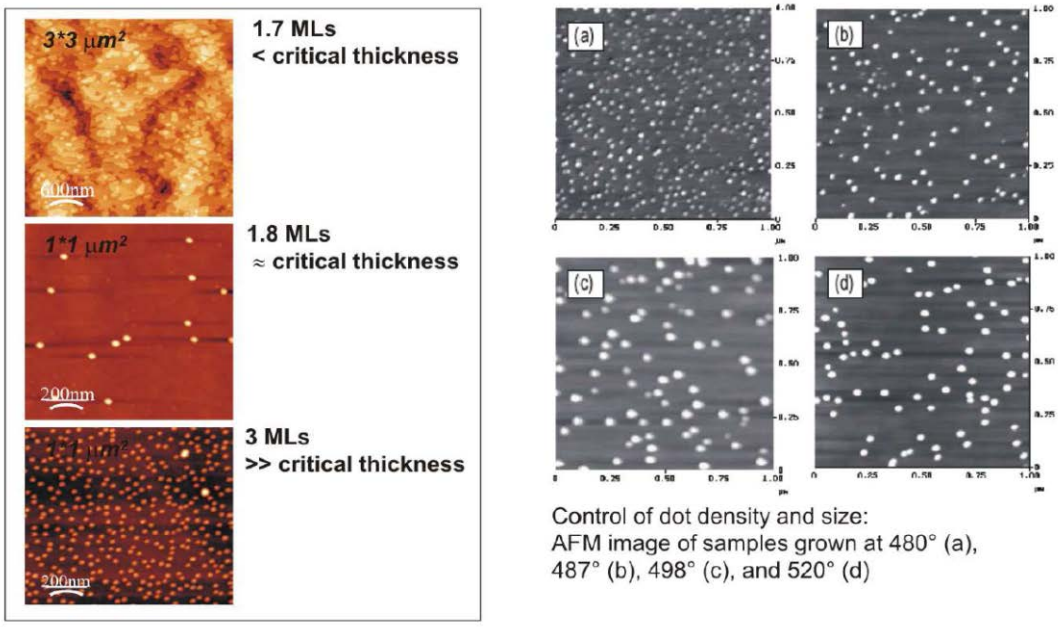


Fig. 14.11. Left: Formation of quantum dots on a GaAs substrate while In deposition; Right: AFM images of quantum dots grown under different growth conditions.

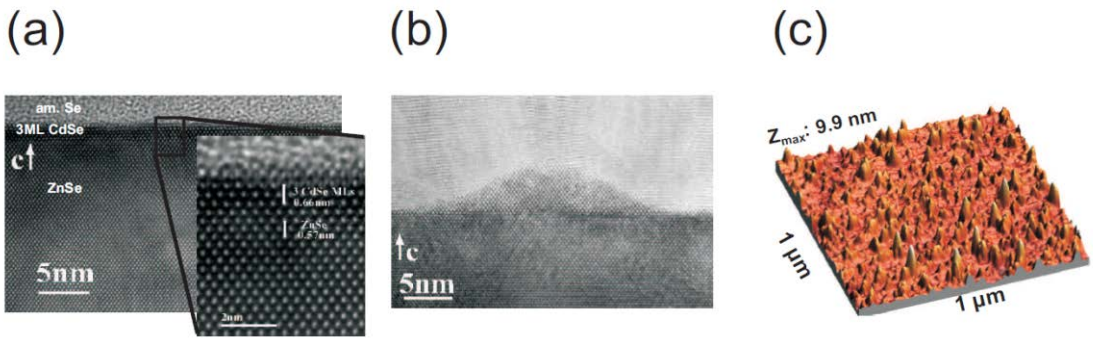


Fig. 14.12. (a) TEM image of a CdSe layer on ZnSe below the critical thickness (3 ML). (b) Above the critical thickness it relaxes by forming a QD. (c) AFM image of CdSe QDs distributed on a ZnSe surface.

The Figures 14.11 and 14.12 show quantum dots in two different material systems (III-IV and II-VI materials). As pointed out quantum dots are a unique system as they are artificial light emitting objects with a discrete energy level structure. The following Figure 14.13 shows a photoluminescence image of a sample containing approximately a dozen InP quantum dots, which are excited by a green laser. The spectrum of one of these dots is also displayed and shows the characteristic spectral lines.

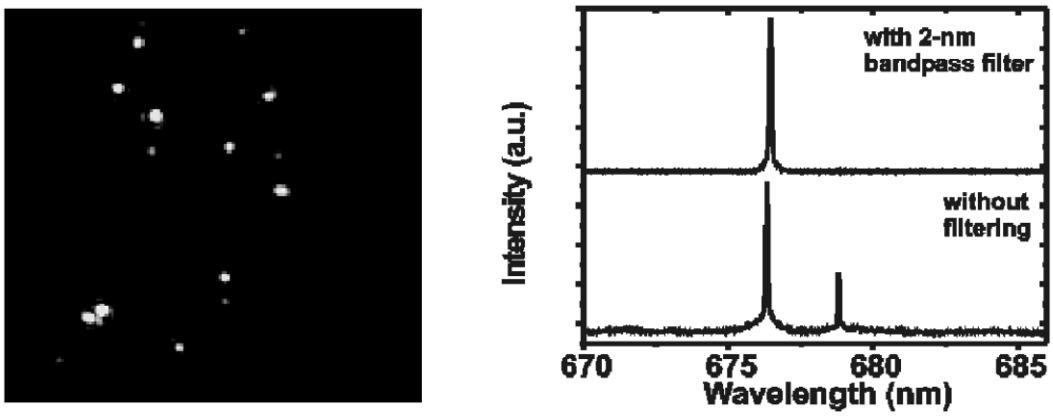


Fig. 14.13. Left: Photoluminescence image of a sample containing several quantum dots (area approximately $10^1 m$ x $10^1 m$; Right: Discrete spectral lines from a single quantum dot.

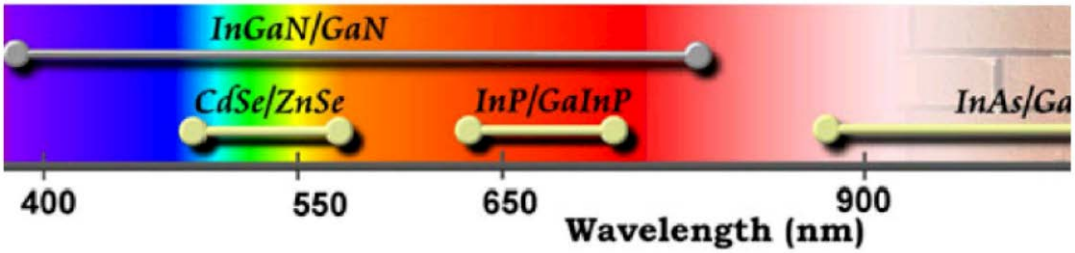


Fig. 14.14. Schematic representation of the wavelength ranges accessible with different Stranski-Krastanow QD material systems.

Almost a continuous spectral regime from the UV to the near-infrared is accessible with quantum dot emitters grown by Stranski-Krastanow self-assembly as illustrated in Figure 14.14.

14.3. Ordering of nanosystems.

Among all recently proposed techniques, self-assembly possesses the advantages of allowing patterning the smallest possible size, along with robustness in patterning large areas. However, technical barriers remain for its large-scale implementation. It is instructive to review the progress of the past few decades, and most common manufacturing techniques, in predicting the

future of the technology. Photolithography has dominated as a manufacturing approach in the microelectronics industry since its introduction with the first integrated circuit.

Current photolithography has been shown to produce features as small as 30 nm on chips, but smaller features require different types of processes. Proposed methods have included decreasing the imaging wavelength, application of extreme ultraviolet (EUV) light, or X-rays. All of these require significant capital investments. Scanning beam lithography, a relatively slower approach to manufacture versus photolithography, is also widely used in chip manufacture, employing either electron or ion beams. Both types are intrinsically serial processes and as such are often used to produce photomasks for projection lithography rather than for actual device fabrication.

Fabrication times depend upon the pattern density and feature size; arrays of sub 20 nm features over an area of 1 cm require 24 h. The slow rate of fabrication, high available precision, and high cost of usage and maintenance restrict scanning beam lithography techniques to small areas or low densities of features, primarily in research applications.

Overall, the method can be useful for transistor fabrication and repair, and the ability to write with different ions is potentially useful in tuning the properties of electronic nanostructures. Unconventional nanofabrication methods that overcome some of the limitations of photolithography and scanning beam lithography, i.e., high capital and operational costs, and/or low resolution, have been developed recently.

Soft lithography provides an inexpensive approach to reproduce patterns created by other lithographic means, wherein numerous molds and replicas can be made from the same master. This technique usually uses elastomeric polydimethylsiloxane (PDMS) stamps with patterned relief on the surface to produce features. The stamps can be prepared by casting polymers against masters patterned with conventional lithographic techniques. Replication of block copolymer templates has led to the fabrication of 20-nm-wide and 27-nm-deep holes. Recently, periodic vertical patterns with peak-to-trough dimension of 1.5 nm have been replicated with PDMS. Distortion or deformation of polymer nanostructures, optimization of conditions for pattern transfer and replication of nanoscale features, and registration of nanoscale features in soft materials for patterning multilayered structures presently limit application of soft lithography.

Unconventional techniques offer new routes to nanofabrication and present insightful new directions to manufacturing of nanostructures at low costs. Here we will focus our discussion on a set of techniques based on embossing, molding, printing, and skiving. These techniques have the ability to generate patterns on a variety of substrates, ranging from rigid, planar substrates to soft, flexible, and curved surfaces. They also allow for rapid prototyping of features in a variety of materials, and are not limited to photoresist polymers used in the conventional techniques. Here, we aim to introduce the fundamentals of these selected techniques, illustrate the interdisciplinary nature of nanofabrication, and elucidate the possibilities to allow chemists, biologists, and other professions outside the engineering community to work on nanofabrication.

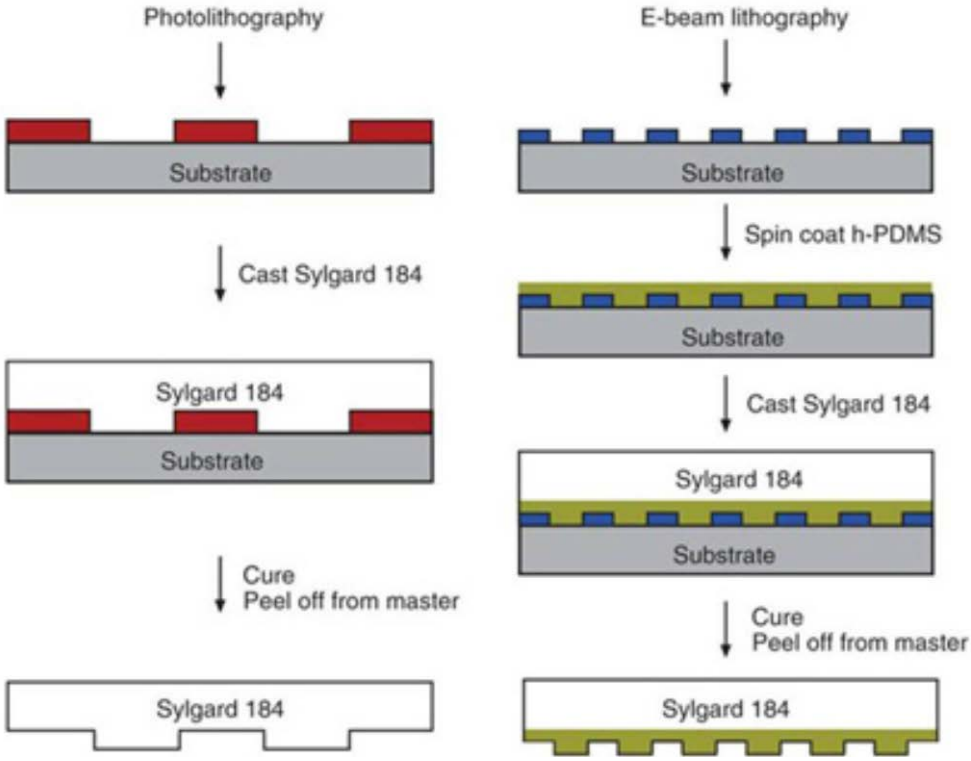


Fig. 14.15. Schematic illustration of the procedures for fabricating Corning Sylgard 184 PDMS and *h*-PDMS.

Embossing and molding are two techniques that use hard or soft molds to transfer a patterned topography into a layer of polymer on a substrate. For example, embossing with hard molds has been used to mass-produce surface-relief structures on compact disks and diffraction gratings in a commercial setting. In fact, hard molds with superb mechanical properties can replicate nanoscale features with a resolution of 20 nm. Hard molds are typically fabricated by conventional techniques, followed by transferring the resist patterns into silicon or quartz.

Unfortunately, it is expensive to fabricate hard molds with regard to lithographic steps involved and the life span of these molds endured in a replication process is rather limited. For example, any contamination of thermoplastic polymer on the hard mold could comprise the resolution and uniformity of a replication process. In contrast, soft molds—often referred as elastomeric stamps based on poly(dimethylsiloxane) (PDMS)—are much easier and less expensive to fabricate. In a typical process, a PDMS mold is fabricated by casting a liquid prepolymer against a master whose surface has been patterned with complementary structures by photolithography or EBL. The PDMS stamp is often fabricated using Dow Corning Sylgard 184 to accommodate fabrication at feature sizes larger than 500 nm. Composite stamps consisting of two layers—a flexible layer (Sylgard 184) supported by a hard layer (*h*-PDMS)—can extend the feature sizes down to the 50–100 nm regime. Fig. 14.15 outlines a procedure for fabricating Sylgard 184 and *h*-/184 PDMS composite stamps from masters.

Nano-imprint lithography (NIL) refers to the pressureinduced transfer of a topographic pattern from a hard mold (silicon) into a thermoplastic polymer film heated above its glass-transition temperature. In NIL, the mold is brought into conformal contact with the polymer, followed by softening the polymer to allow a viscous liquid flow into the mold. Upon cooling down to room temperature, the polymer solidifies to generate a replicated pattern on the substrate. NIL offers a potential solution to manufacturing of nanostructures, but it has a number of pitfalls. First, the molds often suffer significantly reduced lifetimes due to the wear experienced under the conditions of heating and cooling cycles. Second, air bubbles are often trapped in the embossed film at an elevated temperature. As a result, the film often contains defects in the replicated structures. Finally, the process itself is not capable of addressing registration issues, and thus, it is not suitable for fabrication of complex devices. NIL has only been used to fabricate simple devices (single or a few layers) for electrical, optical, photonic, and biological applications. Step-and-flash imprint lithography (SFIL) is a technique derived from NIL that uses a hard mold in transparent quartz to imprint features at room temperature and low applied pressure. In SFIL, a low-viscosity photocurable prepolymer is first dispensed onto the surface of a substrate. When the coated substrate is brought into contact with the mold, the prepolymer spreads across the surface and fills into the relief structures of the mold. After UV-induced photo-polymerization (i.e., curing), the mold is removed, leaving behind replicated structures on the substrate. SFIL has resolution close to sub-10 nm, which is ultimately limited by the size of the features that can be fabricated on the quartz mold. SFIL is useful in device fabrication because transparent quartz allows one to view the alignment marks and thus achieve multilayer fabrication steps with precise registration.

Replica molding (REM) provides an efficient method for replicating patterned relief structures on a PDMS stamp with thermally curable epoxy or UV-curable polyurethane faithfully with shrinkage of less than 3% on curing. Fig. 14.16. shows a diagram and a typical replicated structure in polymer. The fidelity of REM is largely determined by van der Waals interactions, wetting, and kinetic factors such as filling the mold. It was reported that REM had the ability to transfer ~10 nm features on a chrome master into PU using Sylgard 184 PDMS. When *h*-PDMS is used, the minimum feature size of REM can go down to <5 nm. REM can also produce numerous PDMS stamps from the original high-resolution, high-cost master, extending the capability of nanofabrication to nanostructures in a variety of materials. The simplicity and low cost of REM also support its potential use in the manufacturing of nanometersized structures.

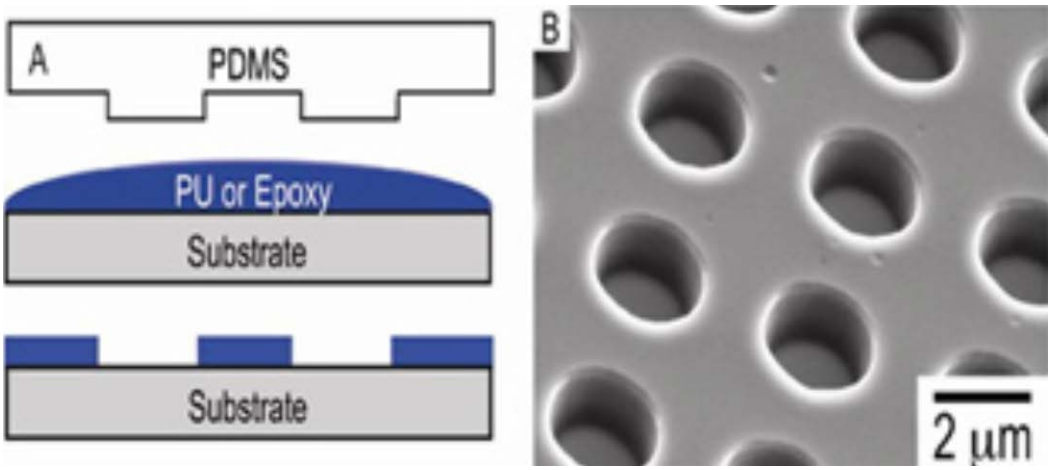


Fig. 14.16. (A) Schematic illustration of REM and (B) an SEM image of an array of holes replicated in polyurethane.

Printing materials onto a surface provides an effective route to transfer molecules and form well-defined nanometerscale features. Microcontact printing (μCP) represents a classic approach to transfer “ink” from a patterned PDMS stamp to a substrate. In μCP of self-assembled monolayers (SAMs), a solution of molecules in ethanol is inked onto the surface of a PDMS stamp, which is dried and brought into conformal contact with a thin layer of gold or silver evaporated on a substrate. The ink molecules transfer from the stamp and form SAMs in a pattern defined by the topography of the stamp. These patterned SAMs can serve as resist or template for selective etching or deposition, respectively. Additionally, μCP allows patterning with other materials that include biomolecules, colloidal particles, and polymers. The printing resolution is determined by the surface diffusion of printed molecules and the distortion of the features within the stamp during printing. Submicrometer printing resolution is achievable

through the use of a normal PDMS stamp. Fig. 14.17 shows examples of structures fabricated using μCP. Recently, Perl et al. published a review article and discussed a number of "extended" μCP methods with reproducible resolution approaching the sub-100 nm regime.

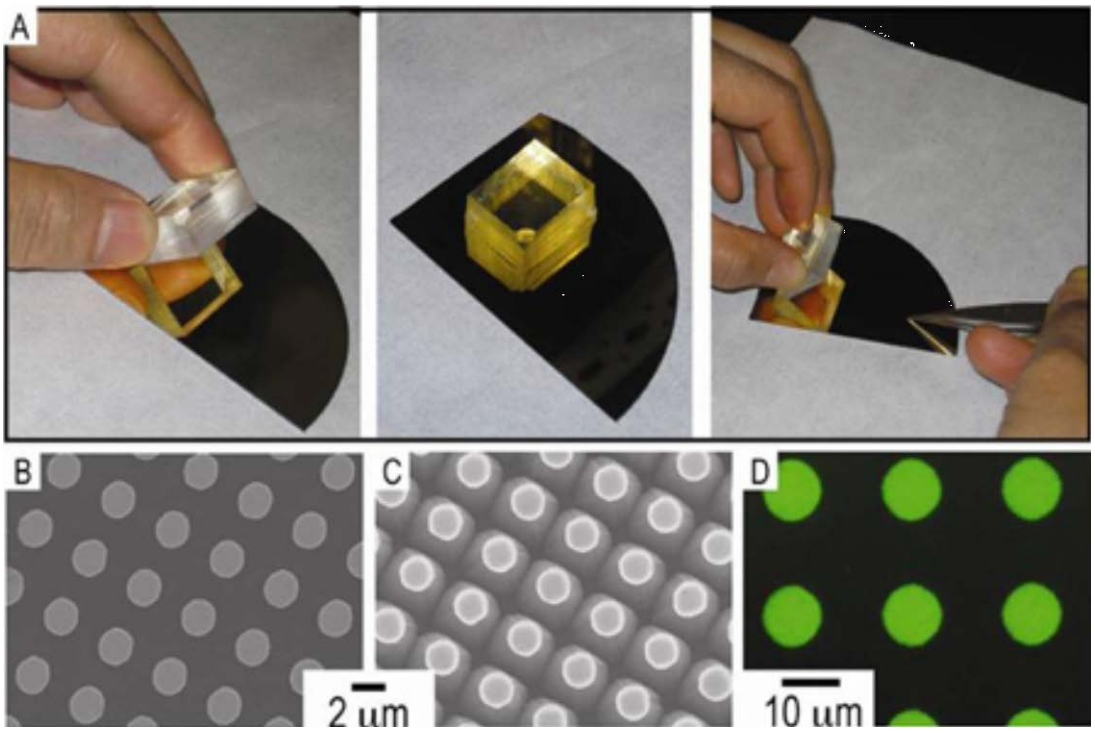


Fig. 14.17. (A) Photographs of three major steps involved in μCP. (B) An SEM image of silver disks fabricated by μCP of SAM followed by selective etching of silver film. (C) An SEM image of structures fabricated by using silver (B) as resist layer for underneath Si(100) etching. (D) Fluorescence optical micrograph of an array of IgG dots fabricated by μCP.

As an extension to μCP, nanotransfer printing (nTP) uses a PDMS stamp coated with a solid layer of ink such as gold to form threedimensional structures with feature sizes between tens of nanometers and tens of microns over areas of several square millimeters. One approach to facilitate efficient transfer is to promote surface chemical bonding between the inked stamp and substrate during contact printing. Alternatively, the process can also rely on the noncovalent surface forces to guild the transfer from the low-energy surface of a PDMS stamp to a variety of substrates. nTP is useful for the fabrication of electronic components, nanoelectromechanical systems, nanofluidic networks, and photonic and plasmonic structures.

Dip Pen Nanolithography (DPN), an alternative approach to printing materials, uses an atomic force microscope (AFM) tip as the “pen,” a solid substrate as the “paper,” and molecules as the ink to directly print patterns through molecular diffusion. Fig. 14.18 shows a schematic of DPN, where the molecular material is first coated and dried onto an AFM tip. The subsequent transfer of molecules from the tip to the surface occurs through a water meniscus that forms spontaneously under the ambient condition. The resolution of DPN is determined by the diffusion rate of molecules and the size of the water meniscus that bridges the tip and the substrate.

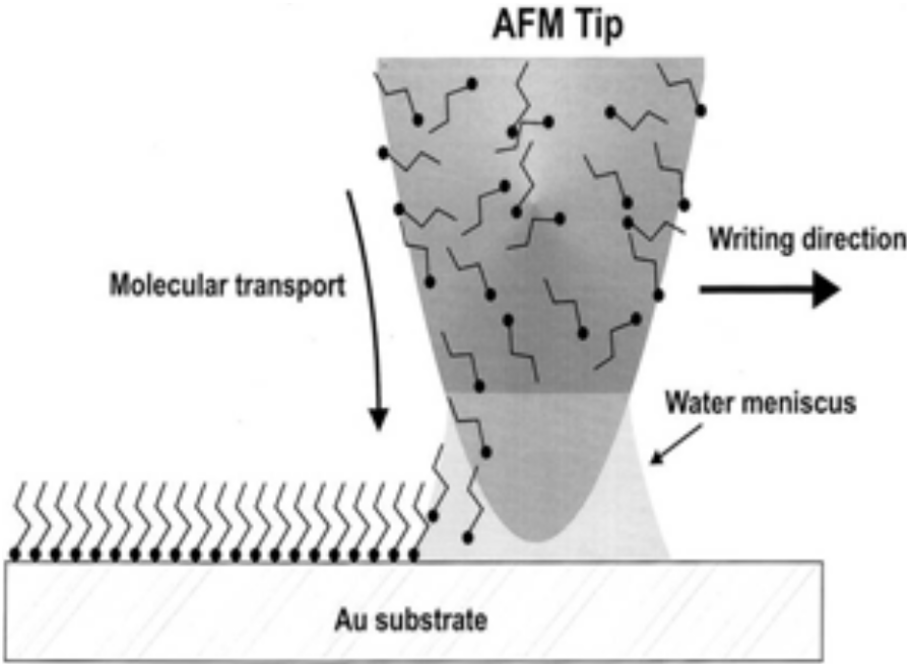


Fig. 14.18. Schematic illustration of DPN.

As a result, DPN allows one to maneuver the sizes of features through a control of humidity (meniscus size) and the tip–substrate contact time (diffusion rate of molecules). In practice, features as small as 50 nm and as large as 1000 nm can be generated reproducibly.

Additionally, DPN has been extended to print other materials that include polymers, proteins, peptides, DNA, and nanoparticles. Recently, polymer pen lithography (PPL) was introduced to combine tributes of μCP and DPN.[34] Specifically, PPL replaced the AFM tip with an array of PDMS tips to print materials. Similar to μCP, PPL has a resolution that is determined by the force that is applied to the PDMS tips and the printing time that keeps the tips in contact with the substrate. Unlike a serial process such as DPN, PPL offers a massive parallel printing capacity

that includes millions of PDMS tips in printing over areas on the order of many square centimeters.

Scanning probe lithography (SPL) is another recent approach for nanofabrication; it uses a conductive scanning probe tip to pattern a thin layer of electron sensitive material and has demonstrated significant potential as an alternative to conventional scanning beam lithography. SPL techniques include scanning tunneling microscopy (STM), atomic force microscopy (AFM), and near-field scanning optical microscopy (NSOM). This technique has shown precise positioning of atoms with an STM tip. Dip-pen nanolithography (DPN) using AFM has produced features as small as 15 nm. Scanning near-field photolithography using NSOM has generated molecular features of 20 nm. The commercial availability of AFM and STM instrumentation and probes makes SPL a convenient approach for nanoscale patterning.

However, the inherent serial nature of SPL using a single tip results in undesirably slow writing. Further, SPL lacks the ability to pattern a high density of features over large areas and is also limited to a small set of materials, e.g., thin films of semiconductors, polymers, and some organic molecules. Parallel approaches in SPL are being developed to overcome the serial limitations of standard SPL technologies. However, it is difficult to fabricate an array of functioning probes with high yields and to pattern complex designs by simultaneously addressing each probe. Edge lithography uses the edge of a topographic feature in the fabrication as well as the developmental stage of nanoscale features. These methods, in which the edges of one pattern become the features of a second pattern, can produce parallel arrays with scale less than 100 nm. There are different forms of edge lithography. One type of approach transfers the edges of a patterned thin film into another material. A second type converts films that are thin in the vertical direction into structures that are thin in the lateral direction. Currently, edge lithographic technique is restricted to making certain types of noncrossing line structures that can be achieved in one step. It is also necessary to increase the density of the patterned features over large areas and to establish the electrical connection among individual features.

One of the biggest challenges in the fabrication of self-assembled semiconductor nanostructures is their controlled lateral ordering on a flat substrate surface. Such an ordering would allow us to address precisely one or just a few nanostructures in postepitaxially processed devices. To a certain degree ordering of self-assembled ~self-ordered nanostructures can be achieved by choosing appropriate growth conditions and material systems. Sometimes, highly indexed or tilted substrate surfaces are used for the growth of selfordered nanostructures.

However, entropy prevents perfect island arrangements and the ordering of self- assembled nanostructures is only of short range. A more promising approach seems to be a combination of

defined prepatterning and self-assembly. Ordered arrays of InGaAs islands were grown on a GaAs(311)*B* substrate by selective epitaxy on a SiN prepatterned surface. Without doubt the Si (001) substrate constitutes very important template in semiconductor technology. Recently, it was shown that self-assembled Ge islands align themselves along the edges of Si mesas, which were grown into prepatterned SiO_2 windows. By changing the SiO_2 pattern different arrangements of Ge islands have been produced. For most device relevant processes or large integration of devices, it would be desireable to have these ordered islands on a planar surface. In this letter we present long-range ordered lines of self-assembled Ge islands on a flat Si (001) surface. The ordering is initiated by the prepatterning of the Si substrate.

The prepatterning can be produced with electron beam lithography and reactive ion etching. The process results usually used in 10 nm deep and about 100 nm wide trenches with a periodicity of 250 nm. The orientation of the trenches is 60° off the [110] direction. After deoxidation the substrate is overgrown with solid source molecular beam epitaxy. First a 19 nm thick Si buffer layer is grown while ramping the growth temperature *Tg* from 400 to 700 °C. The Si buffer layer is followed by a 15-period 9 nm Si/2.5 nm $Si_{0.75}Ge_{0.25}$ superlattice at T_g=700 °C. The sample was finished off with 6 ML Ge islands at T_g=700 °C. The temperature 700 °C since at this specific growth temperature the mean distance of Ge islands agrees well with the period of the processed trenches has been chosen. A schematic illustration of the layer stack and its evolution is given in Figure 14.19. Once the ordered thickness modulation of the SiGe layers is induced by the prepatterned substrate it can propagate through the whole superlattice via its strain fields. This effect is well known in multiple layers of self-assembled Ge/Si islands, where Ge islands can reproduce themselves in a vertical direction via their strain fields.

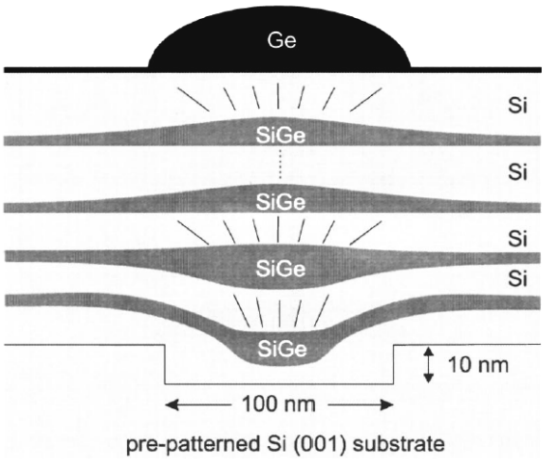


Fig. 14.19. Schematic illustration of the transformation from a surface modulation over a strain field modulation to the final island nucleation on a flat surface.

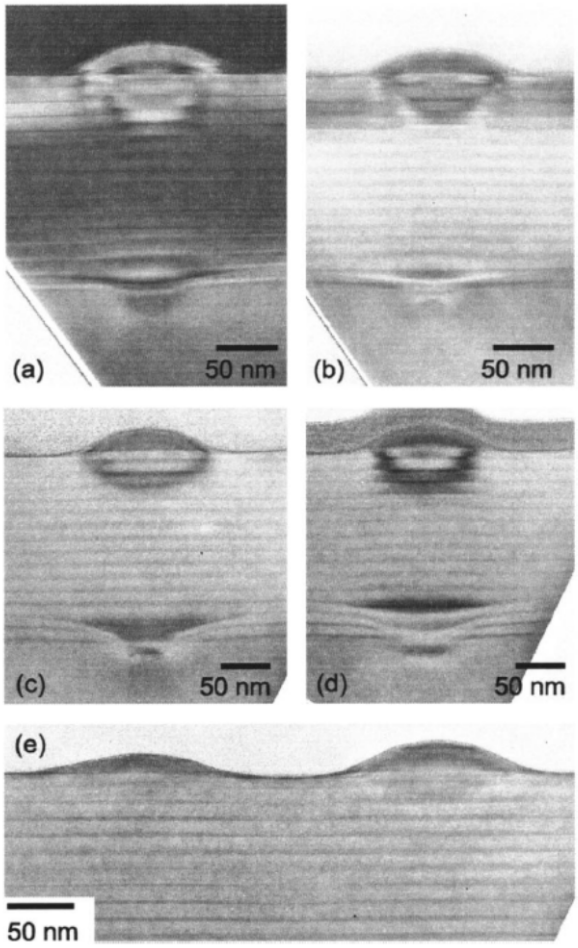


Fig. 14.20. Cross-sectional TEM images from different locations of the Ge islands on Si/SiGe superlattice on a prepatterned substrate: (a) {002} dark field image; (b)–(d) {004} bright-field images; (e) many-beam image at {110} pole position, illustrating the vertical propagation of thickness fluctuation within the superlattice.

In this case the ‘‘dome-like’’ islands16 in the final 6 ML Ge layer sit exactly above the thicker SiGe regions of the superlattice and hence above the initially processed grooves in the Si substrate. It is noteworthy that planarization of the trenches has already been accomplished after the fourth period of the superlattice. For another samples it would therefore be sufficient to reduce the number of periods from 15 to about five. On the other hand, many periods produce a thick buffer, which ensures high crystal quality. Figure 14.20 demonstrates that the vertical reproduction of thickness undulations or islands via their strain fields is perfectly suited to convey a predefined periodicity through a buffer layer (the superlattice) to an epitaxial surface. Unlike to the initial substrate surface — which might be damaged by the etching process — the epitaxial surface is free from any crystal defects and can be used for ordered growth of an active region without any quality restrictions.

The result of the growth procedure is given in Fig. 4.21. Self-assembled Ge dots have formed along perfectly periodic lines (60° off the [110] direction) on the Si surface. Figure 14.21 shows

AFM images (Figs. 4a and 4b) together with their autocorrelations (Fig. 4c and 4d). The 1.731.7 mkm^2 scan in Fig. 4a demonstrates that the dots not only align in the predefined direction but tend to arrange themselves into hexagonal arrays. The ordering into arrays is a self-assembling process, though, and hence is not strictly repeated over the whole surface. The autocorrelation verifies that the periodicity of the hexagon persists over 4–5 periods only. The 10X10 mkm^2 AFM scan in Fig. 4c and its autocorrelation in Fig. 4d demonstrate the long-range ordering in one direction of the self-assembled dots.

The periodic arrangement of the dots is perfect and is maintained over all of the prepatterned area. A slight dot size inhomogeneity in real space is noted, though. It is necessary to point out that these islands grow on a flat surface and can easily be overgrown with a flat Si layer.

It should be mentioned the flat surface to be essential for post-epitaxial processing of devices, especially for those which are part of large-scale integration. As it is clear from written above it is possible to produce not only ordered lines but also ordered arrays of self-assembled Ge nanostructures on Si(001) surfaces. In this case the structure should be grown on a prepatterned array of shallow mesas.

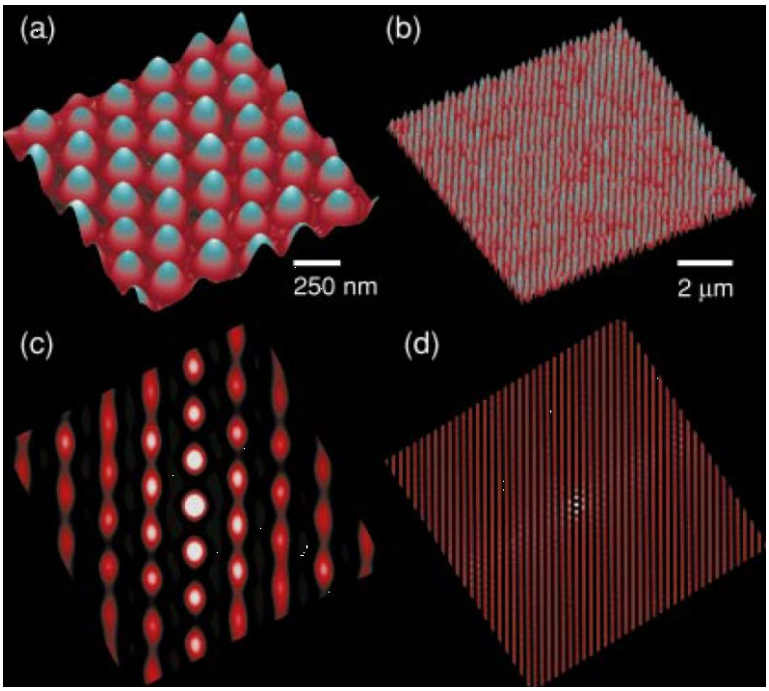


Fig. 14.21. Long-range ordering of self-assembled Ge dots on a planar Si surface. The images show: (a) 1.7 31.7 mm2 and (b) 10310 mm2 AFM scans of 6 ML Ge deposited on a 15-period 2.5 nm $Si_{0.75}Ge_{0.25}$/9 nm Si superlattice. The superlattice was grown on a prepatterned substrate. Autocorrelations in (c) and (d) demonstrate the perfect periodicity in one direction. All dots order along lines 60° off the [110] direction.

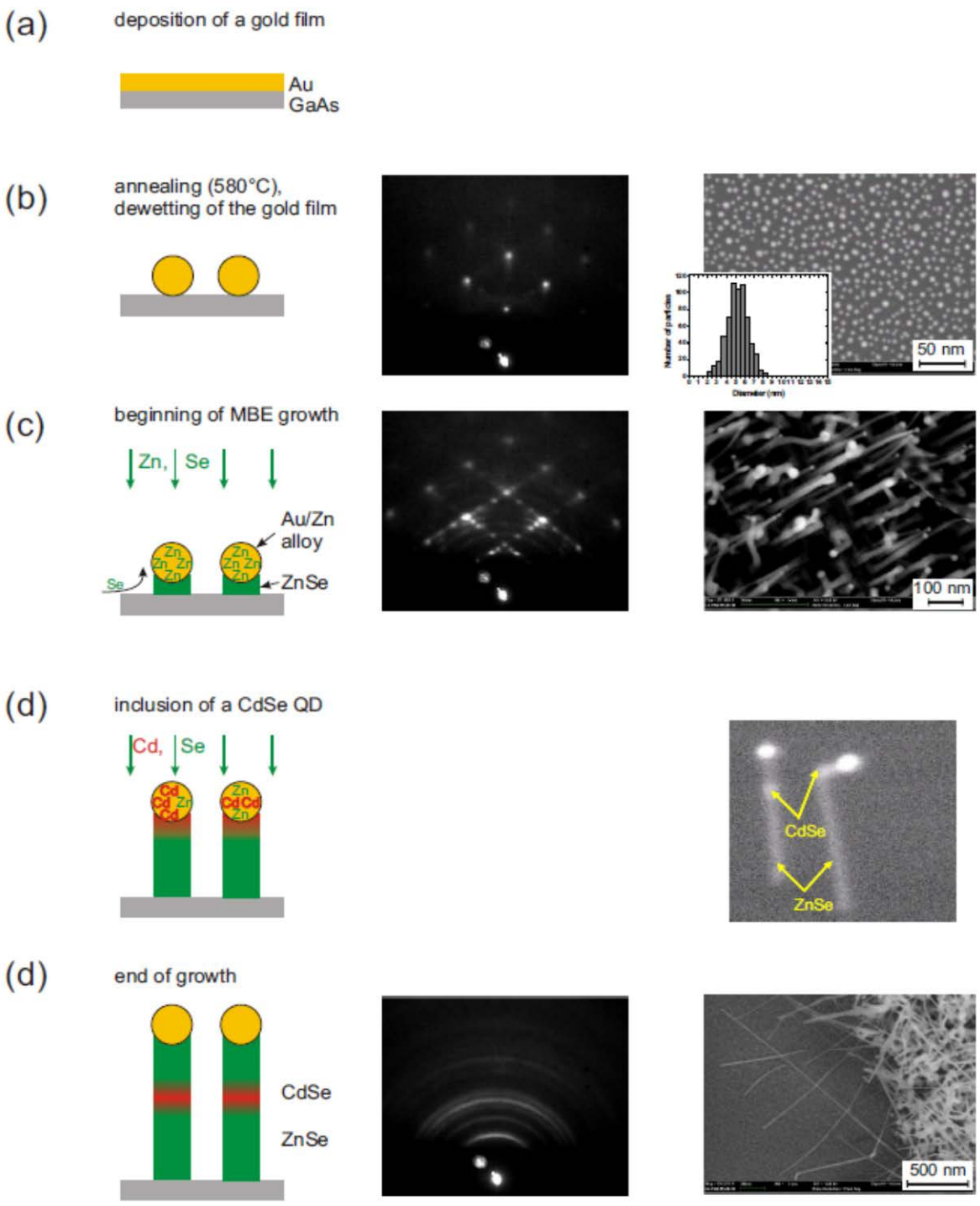


Fig. 14.22. Principle of the VLS growth on the example of ZnSe MBE growth at 450±C. Left column: schematics of the growth.

In the past few years, the growth of nanowires (NWs) have found an increasing attention. These systems are semiconductors with ultra-high aspect ratios (length up to micrometers, and

diameters even below 10 nm are possible). For inclusion of QD heterostructures they have the special advantage that strain relaxation can elastically happen on the narrow sidewalls, so that there are no restrictions to the sizes any more, which makes self-assembly obsolete. As a further consequence, there is no wetting layer, that otherwise can introduce non-radiative escape channels for the charge carriers out of the QD.

For biological/medical analysis NW sample are interesting as sensors due to the huge surface-to-volume ratio. One of the most frequently employed technique for epitaxial NW growth is the Vapour-Liquid-Solid (VLS) growth method. This process was originally developed by Wagner & Ellis in the 1960s to produce micrometer sized whiskers. Starting from 1990s, this technique was employed by many researches to form nanowires and nanorods from a rich variety of materials. In the VLS method, one starts with nanometer-sized metal particles, that are deposited on the surface. During the growth the substrate is heated above the melting point of the metal nanoparticles to a temperature at which it forms an eutectic phase with one of the epitaxial semiconductor reactants. The continued feeding of the semiconductor atoms into the liquid droplet supersaturates the eutectic. This alloy acts as a reservoir of reactants, which favours the growth at the solid-liquid interface and thus forms a one-dimensional nanowire with the alloy droplet remaining on the top. The size of the metal particle also affects the diameter of the nanowire and its growth speed (see Fig. 14.22)

14.4. Method for templating the growth of nanomaterials.

The fabrication of structures (circuits) on a wafer requires a process by which specific patterns of various materials can be deposited on or removed from the wafer's surface. The process of defining these patterns on the wafer is known as *lithography*. Lithography uses photoresist materials to cover areas on the wafer that will not be subjected to material deposition or removal. The Electron Beam Lithography (e-beam lithography or EBL) commonly used at this case. At the same time much more challenging are self-organized methods.

Self-organized nanopore arrays of valve metal oxides can be formed non-lithographically by the electrochemical anodization of valve metals such as Al, Ti, Ta, Hf, Zr, etc. The anodization process is simple and economical and the resulting structures are mechanically robust and chemically resistant even at elevated temperatures. Therefore, anodically formed nanoporous valve metal oxidesare excellent architectures for templating and pattern transfer and a wide variety of functional nanostructures have been formed using nanoporous alumina and titania.

The anodic formation of porous alumina has been known since 1956 but has been extended to the other valve metals only in the last decade. We shall restrict our discussion to anodic aluminum oxide (AAO) and nanotubular ТИХ In AAO, the thickness of the nanoporous film, the

size of the nanopores and their spacing arc the morphological parameters of interest and these can be controlled by tuning the anodizalion potential, the duration of the anodization and by choosing the appropriate electrolyte to perform the anodization. In $Ti0_2$ nanotube (TNT) arrays, the tubular architecture results in an additional morphological parameter, namely the wall-thickness.

Both AAO and TNT's have been fabricated on a variety of different substrates such as glass, Si wafers, flexible polymeric substrates and even curved surfaces such as metallic pipes. Also, both AAO and TNT's can be transformed into free-standing membranes several hundreds of mkm in thickness by detaching the nanoporous film from the underlying substrate. It has been difficult to form high quality nanoporous Al_2O_3 on the technologically important transparent conducting oxide (TCO) coated glass substrates due to issues of adhesion, though a relatively recent technique that uses a very thin Ti adhesion promoter appears to be promising. On the other hand, the formation of TNT's on TCO coated glass substrates, while challenging, has been successfully achieved and utilized for devices. Another important difference between AAO and TNT's lies in their differing conductivities; Al_2O_3 is an insulator whereas crystalline TiO_2 is an n-type semiconductor.

Nanoporous alumina is perfectly ordered when the pores form a hexagonal honey-comb structure consisting of close-packed nanochannels of high-aspeel ratio (shown in Fig. 14.23). The largest sizes of domains with defect-free ordering is restricted to a few mkm^2. To demonstrate scalability and increase throughput for templating applications, nanoporous alumina with much larger defect-free domains is demanded. The size of self-ordered domains have been increased by prepatterning the aluminum (Al) surface prior to the anodization process and using the patterns as the nucleation sites to guide the growth of the nanochannels.

In 1997, Masuda pioneered the use of hard-stamping to form the necessary nucleation sites on the surface of polished aluminum. In Masuda's process, conventional electron beam lithography was used to pattern a hexagonally arranged array of convexes on a master mold.

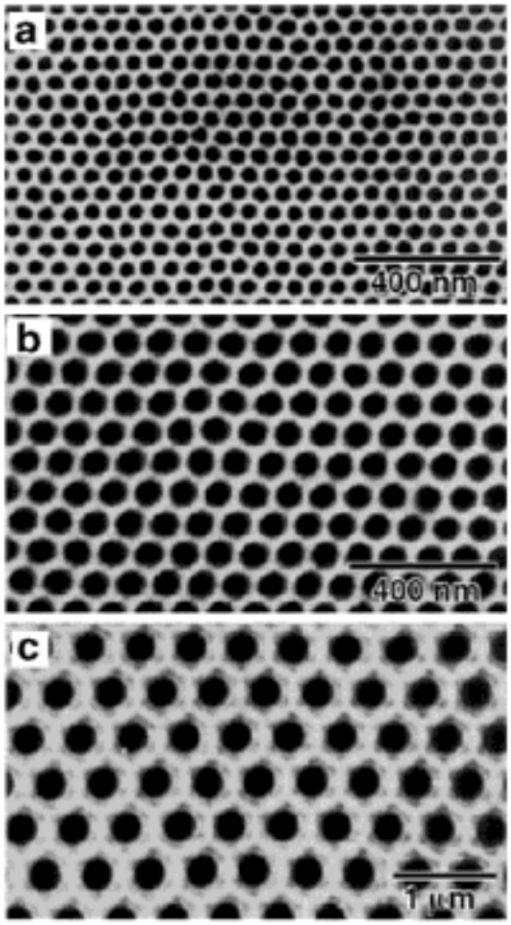


Fig. 14.23. SEM micrographs of naturally occurring long-range ordered anodic porous alumina formed in three types of acid electrolytes: (a) sulfuric acid, (b) oxalic acid and (c) phosphoric acid.

The mold was made from a mechanically hard material such as SiC. The master mold was then pressed onto the aluminum surface using an oil press at room temperature to generate the required array of concaves on the surface of aluminum. The soft-imprinting technique introduced by the Gao group uses Ar^+ plasma etching through a free standing nanoporous alumina membrane etch mask to create ordered nanoindcntalions on the Al surface. The AAO etch mask was itself formed by a regular two-step anodization. Using this technique, highly ordered porous anodic alumina templates were fabricated on different substrates (such as Si, glass slides, and flexible polyimide films) over large areas (>1.5 cm^2). Another interesting technique employing self-assembly is based on the spontaneous organization of monodisperse polystyrene nanoparticles into a 2D array. In this process, shallow concaves were formed on Al by replicating the ordered structure of the 2D array of polystyrene particles on the submicrometer scale and initiating hole development during anodization. Other approaches in the service of the same objective include direct focused ion beam (FIB) lithography, interference lithography,

holographic lithography, colloidal lithography, and block copolymer self-assembly. A more recent technique namely, step and flash imprint lithography (SFIL) was used to demonstrate near-perfect ordered AAO with square and hexagonal lattice configuration on silicon substrate over 4 in. wafer areas (See Figure 14.24).

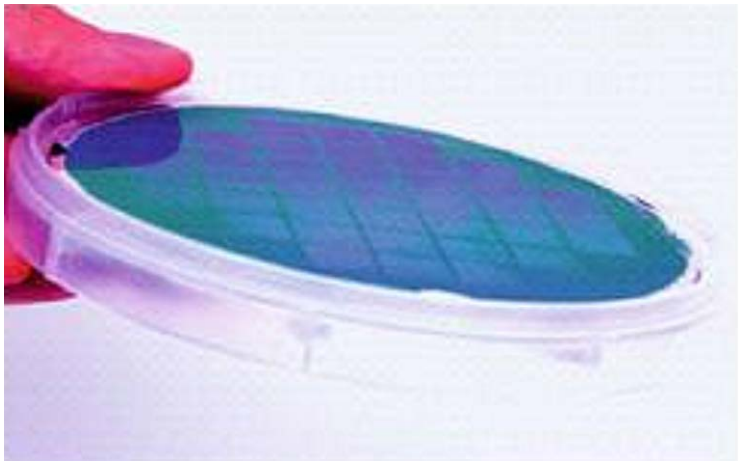


Fig. 14.24. Photograph of a near-perfect AAO template on a 4 in. silicon wafer. The 10 mm x 10 mm square areas with bright light diffraction indicate the anodized sample that wasprcpattcrned using SFIL.

In 2003, Masuda el al. introduced a new method based on pretexturing the aluminum to engineer differentiation of the self-organized AAO pores into two types with a controllable period. To appreciate how this is accomplished consider the following process: A SiC master mold is patterned using conventional electron-beam lithography to create an ideally ordered arrangement of convex dots with a period of 200 nm but with the important difference that every sixth site had a defect (no dot). When this master mold is stamped onto a polished Al foil, the pattern on the mold is replicated on the Al surface as an ordered array of concave dimples but with a dimple missing at every sixth site. The imprinted Al foil is anodized al 80 V in 0.05 M oxalic acid at 16 ^{o}C to form the nanoporous architecture. When the aluminum is subsequently selectively removed to reveal the barrier layer it is found that the barrier layer is thicker at the imprinted sites than at the non-imprinted sites. Likewise, the pores formed at the non-imprinted sites are smaller in diameter than the pores at the imprinted sites. The difference in thickness between the barrier layer at the imprinted and non-imprinted sites is exploited to selectively open the bottoms of the smaller non-imprinted pores.

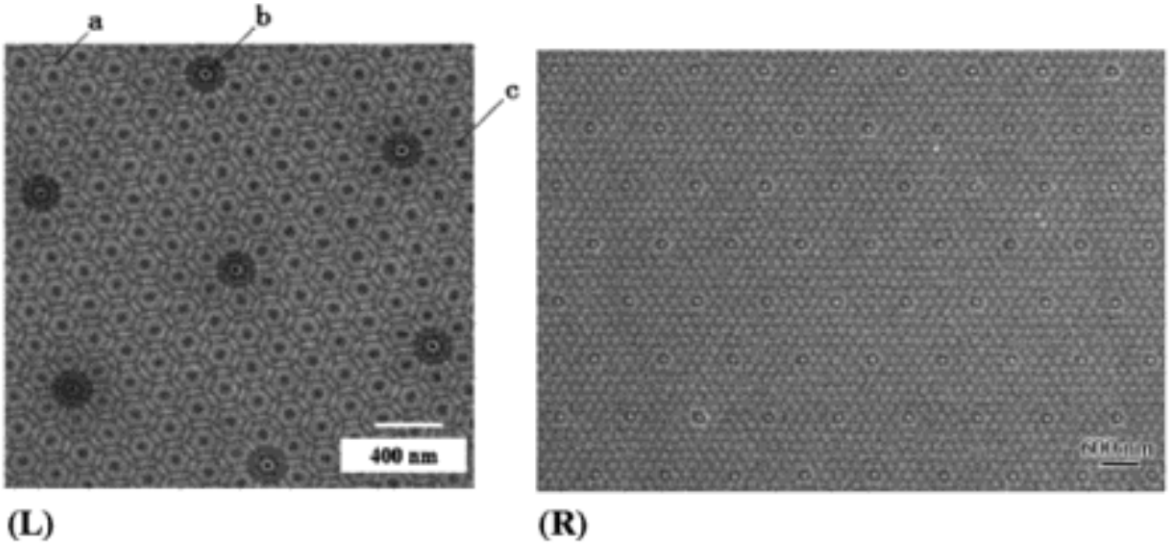


Fig. 14.25. SEM image, laken from the barrier layer side, of an alumina substrate with selectively opened holes (a) porous alumina (b) selectively opened hole and (c) barrier layer (R).

Figure 14.25 shows the SEM image taken from the barrier layer side of such an AAO membrane with selectively opened holes. It is clear from this image that pores of slightly smaller diameter, at every sixth site (corresponding to the non-imprinted sites) are selectively opened. In the next step a contact layer is deposited on the top-side of the membrane to provide electrical contact during the subsequent electrodeposition of gold into the membrane. Since the pore bottoms have been selectively opened only at the non-imprinted sites, Au is selectively deposited into these pores as seen at the site labeled ‘b’.

Clearly, the selective opening of the pores is contingent upon performing ihe wet etch of the barrier layer for just enough time for the thinner barrier layer at the bottom of the non-imprinted pores to be completely removed thus opening the pore but not so long as lo also cause the remaining pores to open. In this particular report, a dipping time of 77 min in phosphoric acid was found to be optimal for selective opening of the pores. The entire process sequence is illustrated in the schematic diagram of Fig. 14.26. The same pore differentiation process can be further continued lo open all the remaining pores after the deposition of metal (say gold) into the selectively opened pores. In the subsequent step, a second material (say nickel) is deposited into the remaining pores (now open) thus resulting in a binary mosaic-like nanocomposite. It also follows that the pore differentiation process can be extended to vary the period of the differentiated pores. Also, the two materials used to form the mosaic composite need not be restricted to metals such as Ni and Au but could be any two materials that can be grown in the pores by electrochemical or electro phoretic deposition.

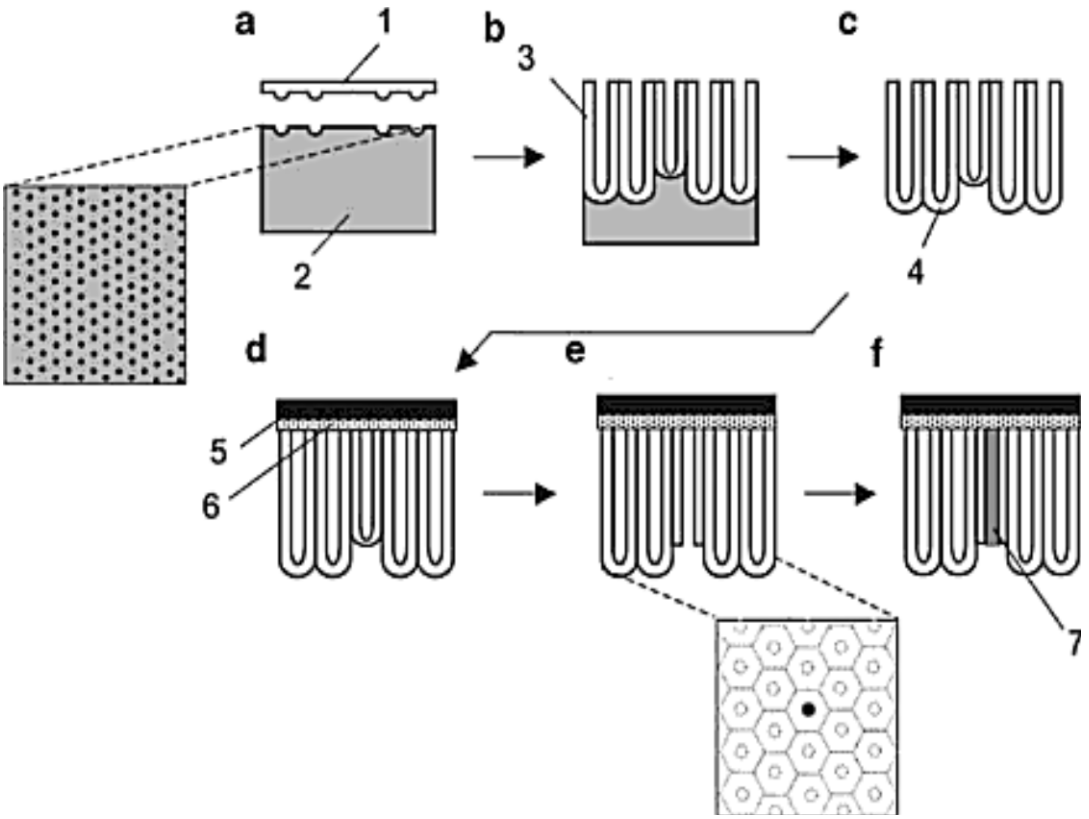


Fig. 14.26. Schematic of the preparation process for a pore-differentiated Au-nanodisk array: (a) imprinting the aluminum with a SiC mold; (b) anodization of the Al; (c) removal of the Al; (d) formation of an electrode on the top-surface of the AAO membrane; (e) selective etching of the barrier layer at shallow sites to form selectively opened holes; (f) deposition of Au. (ICP) etching through the template. The purpose of the nano-patteming was to reduce the high dislocation density that results when GaN is grown in Si substrates due to large mismatch in lattice constant and thermal conductivity between Si and GaN. In another papers antireflective sub-wavelength structured surfaces (for photovoltaic application) were fabricated on single-crystalline Si waters patterned using an AAO etch-mask. The AAO pattern was transferred to the Si substrate by inductively coupled plasma.

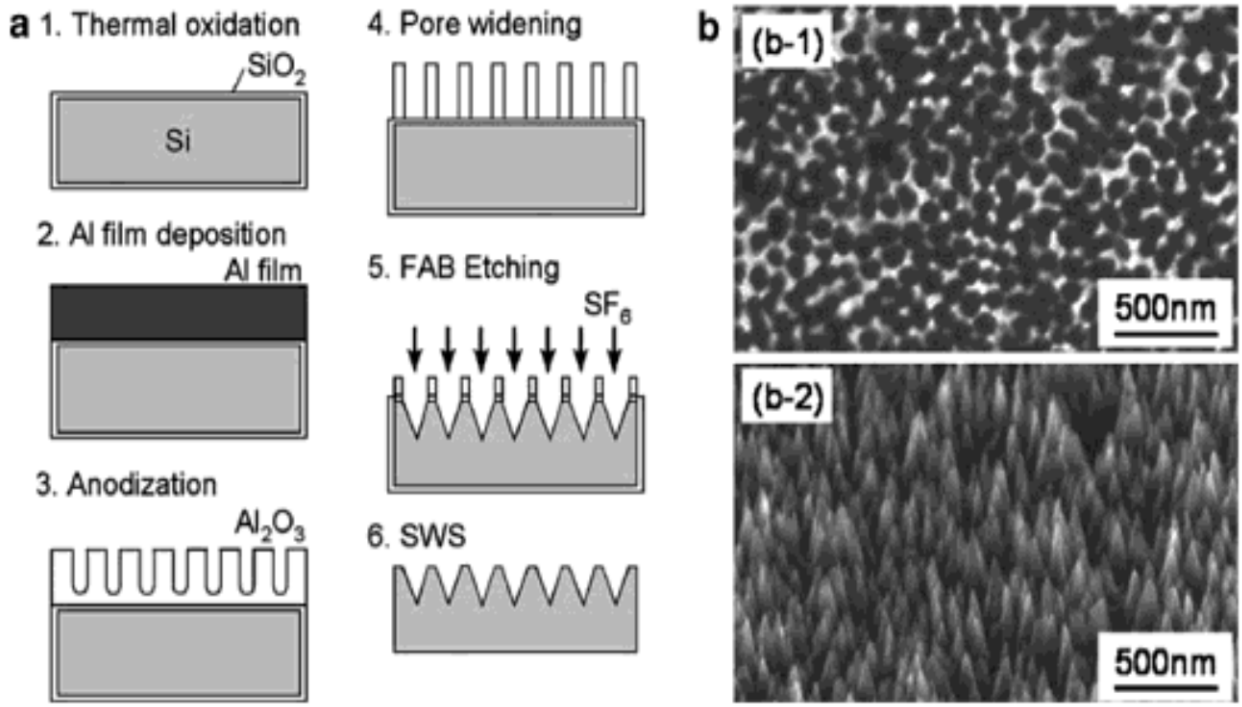


Fig. 14.27. Sub-wavelengh structured (SWS) silicon wafers patterned using AAO: (a) fabrication sequence used to prepare SWS Si; (b) SEM images of the Si surface after fast atom beam etching through the AAO mask for 50 min showing (b-1) top-view and (b-2) cross-section.

Figure 14.27a shows a schematic of the fabrication process used to achieve the sub-wavelength structured surfaces while Fig. 14.27b shows SEM images of the resulting SWS Si. The starting step, namely the thermal oxidation of Si, was performed in a standard quartz furnace to form a thin protective oxide barrier for anodic oxidation. The thickness of the deposited Al film was 500 nm. The anodization was performed in 0.3 M oxalic acid at 40 V to obtain a periodicity of ~ 100 nm and was followed by pore-widening in 5 wt% H_3PO_4 to remove the barrier layer.

Fast atom beam etching (FAB) with SF_6 gas performed under an acceleration voltage was used to generate the SWS surface relief grating. Figure 14.28 demonstrates that the reflectivity of the resulting SWS Si is less than 1% across the entire wavelength range from 300 to 1,000 nm.

Also shown in Fig. 14.28 is the high reflectance of a polished Si wafer and the reflectance of a conventionally alkali-textured Si wafer.

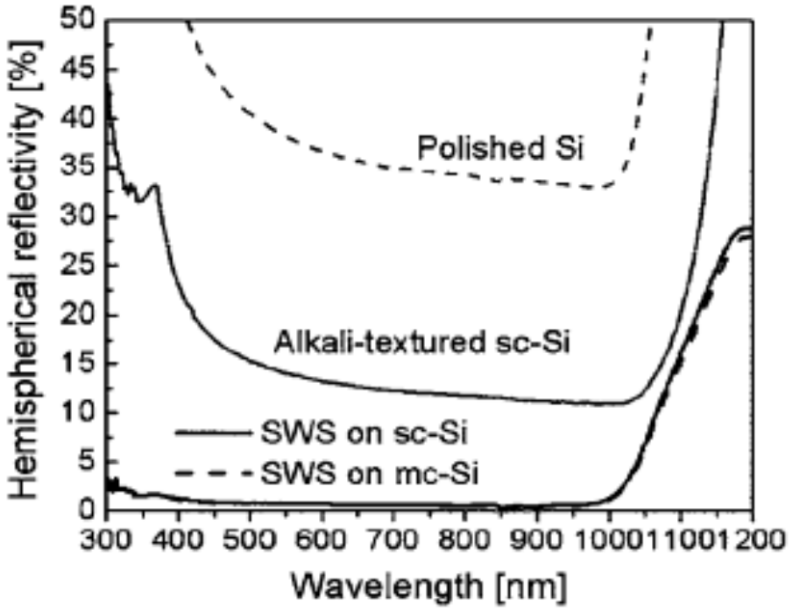


Fig. 14.28. Measured reflectivity spectra of AAO-patterned SWS Si samples fabricated by the process depicted in Fig. 14.19a.

The lateral ordering of semiconductor nanocrystals in 2D arrays is a subjeel of both fundamental and applied research in nanophotonics and nanoelectronics. Si nanocrystals embedded wilhin thin dielectric layers form the building blocks of certain next generation non-volatile memory devices. Quantum dot (QD) lasers constructed from arrays of InP and InAs nano-islands on GaAs and other substrates are becoming increasingly important for applications such as high density optical data storage systems. The formation of lateral 2D arrays of QD's is also important in fundamental experiments designed to investigate the collective behavior of large assemblies of coupled nano-elemenls. In all these applications, high regularity in the arrangement of the QD arrays and control of their spacing is required lo obtain precise control over the device properties. Furthermore, critical feature sizes smaller than 100 nm need to be fabricated over a large area. These requirements are beyond the reach of conventional nanofabrication techniques. The use of electrochemically formed self-organized nanoporous alumina stencils offers an attractive route towards the formation of 2D QD arrays.

In some papers, CdS nanodot arrays were formed on Si substrates by thermal evaporation of CdS through an ultrathin AAO membrane.

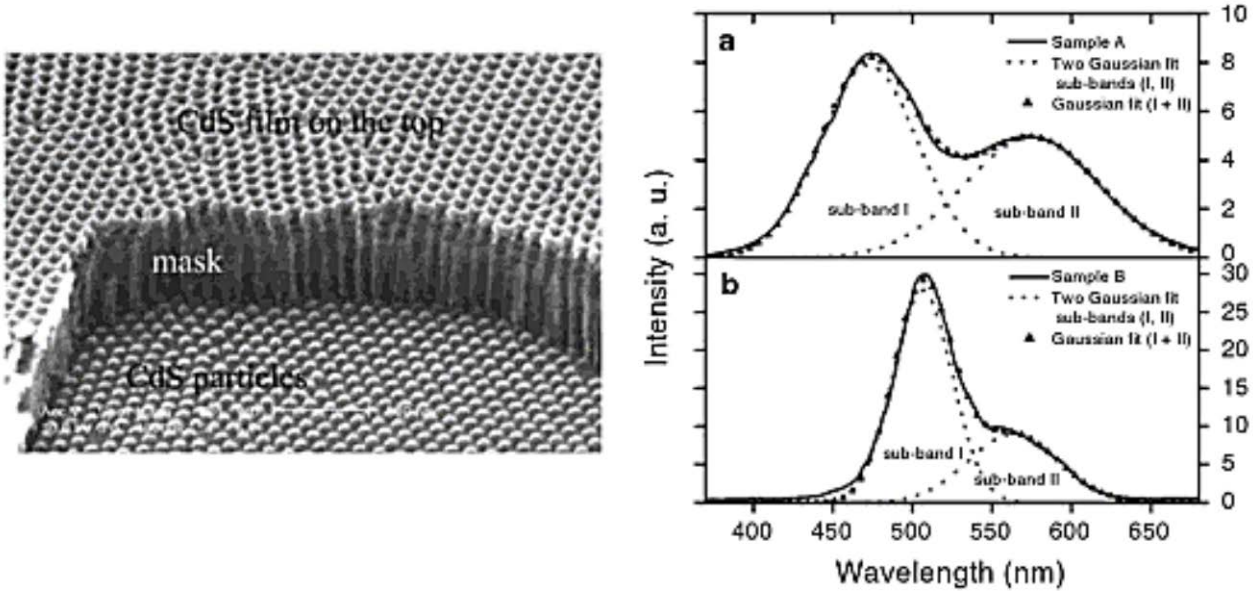


Fig. 14.29. Left, CMS nanodol arrays with an average diameter and spacing of 80 nm and 105 nm respectively, formed on Si substrates using an AAO mask; and right, photoluminescence spectra of CdS nanodot arrays of (a) 10 nm height and (b) 50 nm height and their two Gaussian fit subbands. The excitation wavelength is 350 nm. The peak positions of the subbands I and II are located at about 473 and 575 nm in sad and 506 and 563 nm in sbd, respectively.

The resulting CdS nanodots (shown in Fig. 14.29) were polycrystalline with a (002) preferred orientation and possessed a mono-dispersed size distribution.The photoluminescence (PL) characteristics of CdS NP’s of two different sizes 10 nm in (a) and 50 nm in (b), are shown in Fig. 14.29R and each show two features, namely band-edge emission (sub-band I) and surface-defect emission (sub-band II), Band-edge emission is ascribed to the radiative recombination of excitons in the NP's and therefore the corresponding peak energy of such band-edge emission is usually slightly lower than the band-gap energy of the CdS NP's.The 10 nm nanodols had apanicle crystallite size smaller than 4 times the Bohr exciton radius and therefore exhibited the effeels of quantum confinement, which manifested itself in the strongly blue-shifted broad peak of sub-band I. On the other hand, the 50 nm nanodols, which were outside the excitonic confinement regime, exhibited a narrow size-independent band-edge emission at 506 nm.

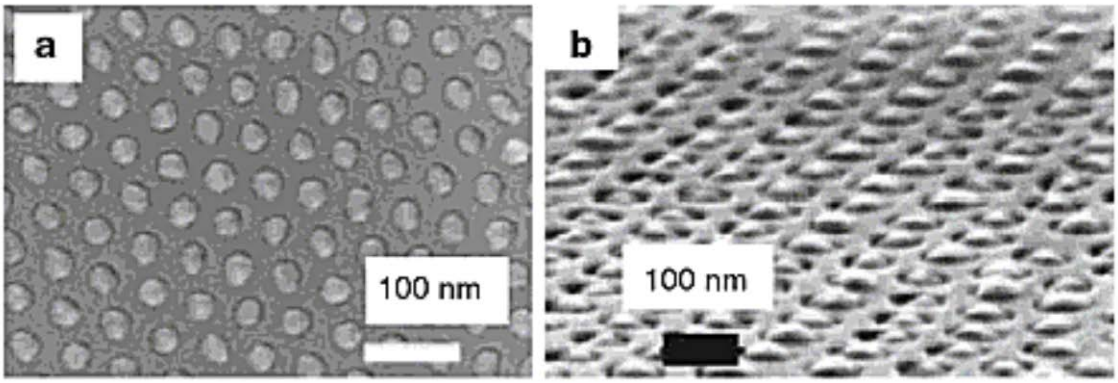


Fig. 14.30. (a) SEM top-view image and (b) SEM oblique view image of a hexagonally ordered InAs quantum dot array on a GaAs substrate formed using an AAO membrane as a stencil.

By reactive ion etching (RIE) using BCl_3 through an AAO membrane etch-mask with a periodicity of 110 nm and a diameter of 55 nm, transferred the nanopore array pattern onto a GaAs substrate. Subsequently, they grew a highly ordered InAs nanodol arrays by molecular-beam epitaxy on non-lithographically nanopattemed GaAs. The resulting dots (see Fig. 14.30) were organized in a dense hexagonal lateral superlattice.

For plasmonic device applications, it is desirable lo create periodic arrays of size-controlled noble metal panicles (Ag and Au) on arbitrary substrates. Here too, AAO templates have been used. Some researchers demonstrated an 8% increase in short circuit density in an optically thin GaAs solar cell decorated with densely packed high-aspect ratio Ag nanoparticles (NP's) fabricated by masked deposition through an AAO template. Enhanced light absorption occurred in this device due to the longer optical path of incident light in the absorber layer, which was itself due to the strong scattering by interacting surface plasmons formed in the Ag NP's.

14.5. Nanocrystalline semiconductors.

Semiconductor nanocrystals (NC) or colloidal QDs are small semiconductor crystallites which are made by organometallic chemical methods and are composed of a semiconductor core capped with a layer of organic molecules (Murray et al. 1993). The organic capping prevents uncontrolled growth and agglomeration of the nanoparticles. It also allows NCs to be chemically manipulated as if they were large molecules, with solubility and chemical reactivity determined by the identity of the organic molecules, Figure 14.31. The capping also provides passivation of NCs; that is, it terminates dangling bonds that remain on the semiconductors surface. The unterminated dangling bonds can affect the NCs emission efficiency because they lead to a loss mechanism wherein electrons are rapidly trapped at the surface before they have a chance to emit a photon. Using colloidal chemical syntheses, one can prepare NC with nearly atomic precision;

their diameters range from nanometers to tens of nanometers and size dispersions as narrow as 5%. Because of the quantum-size effect, this ability to tune the NC size translates into a means of controlling various NC properties, such as emission and absorption wavelengths.

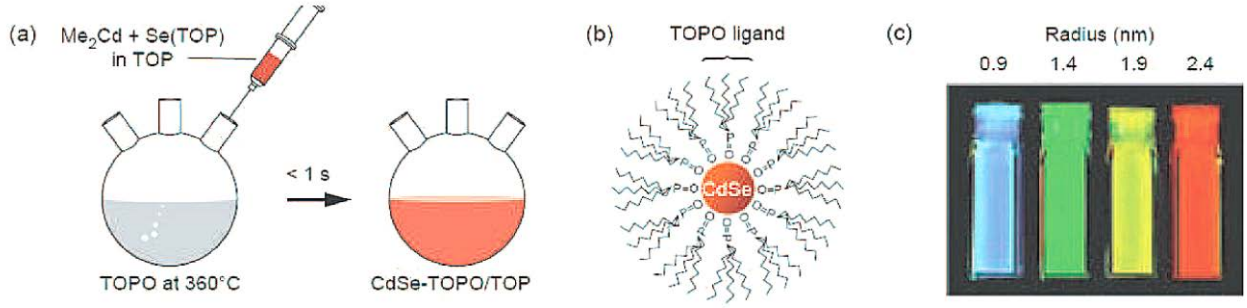


Fig. 14.31. (a) An organometallic method is used for the fabrication of highly monodisperse CdSe NCs. Nucleation and subsequent growth of NCs occurs after a quick injection of metal and chalcogenide precursors into the hot, strongly coordinating solventa mixture of trioctylphosphine (TOP) and trioctylphosphine oxide (TOPO) in the case shown. After a fixed period, removing the heat source stops the reaction. As a result, NCs of a particular size form. (b) The colloidal NCs obtained by the method illustrated in (a) consist of an inorganic CdSe core capped with a layer of TOPO/TOP molecules. (c) Solutions of CdSe NCs of different radii, under ultraviolet illumination, emit different colours because of the quantum size effect. A 2.4-nm-radius dot has an energy gap of about 2 eV and emits in the orange, whereas a dot of radius 0.9 nm has a gap of about 2.7 eV and emits a blue colour.

Colloidal NCs with variable surface chemistry are ideal building blocks for creation of different superstructures, like composite polymer/nanocrystal films, composite core-shell microspheres, 1D, 2D and 3D NC arrays. An example of a crystal of nanocrystals is shown in Figure 14.32. Advanced optical spectroscopy studies on NCs and their superstructures address energy transfer, charge separation, and single particle luminescence.

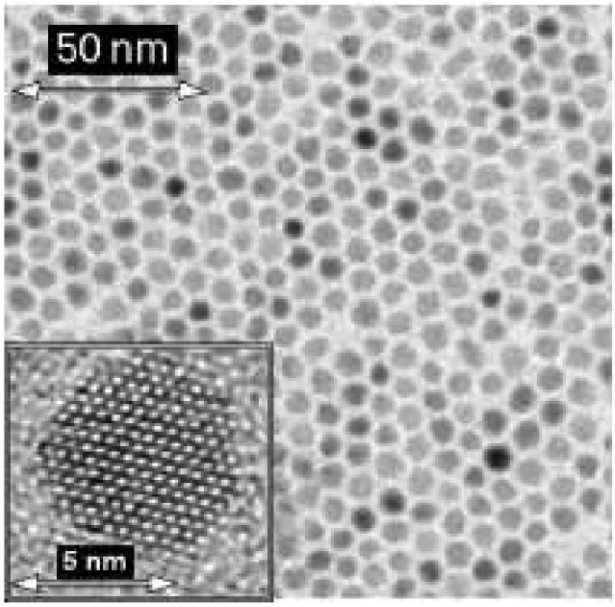


Fig. 14.32. TEM image of an individual nanocrystal (inset) and an array of nanocrystals.

Highly luminescent semiconductor NCs are interesting for different applications, ranging from solar cells to biological fluorescent labels.

A real breakthrough in Si nanoclusters growing was reported in 2002 by using a superlattice of SiO and $Si0_2$ that is nowadays a standard approach for fabricating ordered, layer arranged, size-controlled Si nanocrystals. This method is based on the preparation of amorphous $SiO_x/Si0_2$ superlattices and thermal annealing for phase separation and crystallization. The preparation of $SiO_x/Si0_2$ superlattices ($1 < x < 2$) is a simple, elegant, and efficient method for the synthesis of size-controlled Si nanocrystals and could be used for a number of applications. Nonstoichiometric oxides (SiO_x) are not stable at high temperatures and decompose by a phase

$$SiO_x \rightarrow \frac{x}{2} SiO_2 + \left(1 - \frac{x}{2}\right) Si.$$

separation into the two stable components (Si and $Si0_2$):

Depending on temperature, amorphous Si clusters or Si nanocrystals will be observed. Figure 8.1 schematically demonstrates the desired control of size, separation, and density for applications that can be achieved using the SiO_x/SiO_2

superlattice approach in the following way: the size of the nanocrystals is controlled by the layer thickness of the SiO_x layers (for more details, please see below). The stoichiometry of the SiO_x influences the number of Si nanocrystals within the layers and their average density. An additional density control can be achieved by thicker barrier $Si0_2$ layers between adjacent SiO_x

layers. The preparation can be done by evaporation of SiO powder under high vacuum. Adding oxygen during the growth can be used for change in stoichiometry. Please note that already an oxygen partial pressure of 10^{-4} mbar is enough to completely oxidize the growing films into $Si0_2$. The base pressure in our evaporation chamber was 1 x 10^{-6} mbar for SiO layers. We observed that the particle size of the SiO powder and the chosen evaporation boat tan also influence the deposition process. SiO pellets of 99.99% purity arc used, which we grind to a powder. Tantalum boats are used for evaporation. A boat temperature exceeding 1000 C (as used here for the evaporation of the SiO powder under oxygen partial pressure) results in an oxidation of the boat material Only the vapor pressure of Ta_2O_5, is low enough to prevent a coevaporation of the metal oxide covering the boat. We observed metal contamination in the grown films that drastically deteriorate, especially the electrical properties if tungsten or molybdenum boats are used.

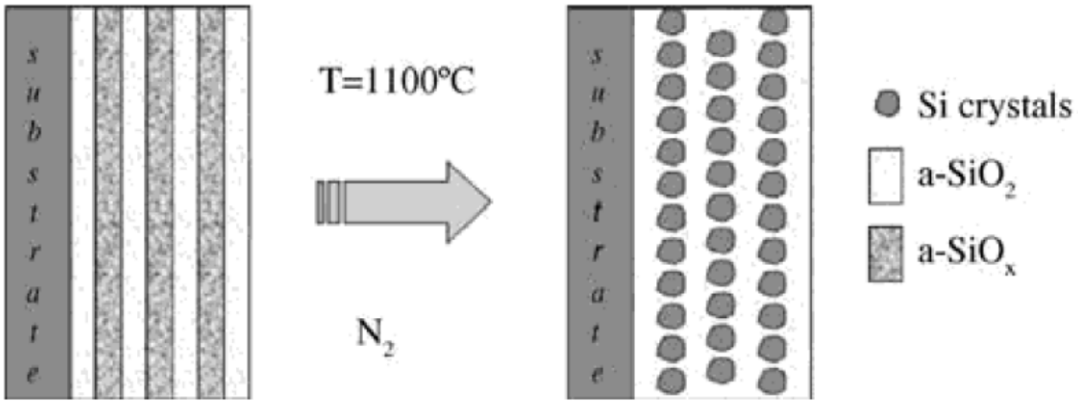


Fig. 14.33. Model of process for preparing layer arranged Si nanocrystals.

A detailed investigation of the phase separation of SiO_x resulting in Si nanocrystals after 1100 °C annealing under nitrogen atmosphere was reported by L.X. Yi. The annealing under nitrogen atmosphere gives the better luminescence intensity than when Ar is used. It was difficult to prove nitrogen in the films after the annealing process, thus only traces (of below 1%) might be included by diffusion during annealing. The evaporation process used for SiO/SiO_2 superlattices has a number of advantages: it is simple and the resulting layers do not contain any hydrogen and a very low level of nitrogen that is important for the crystallization process.

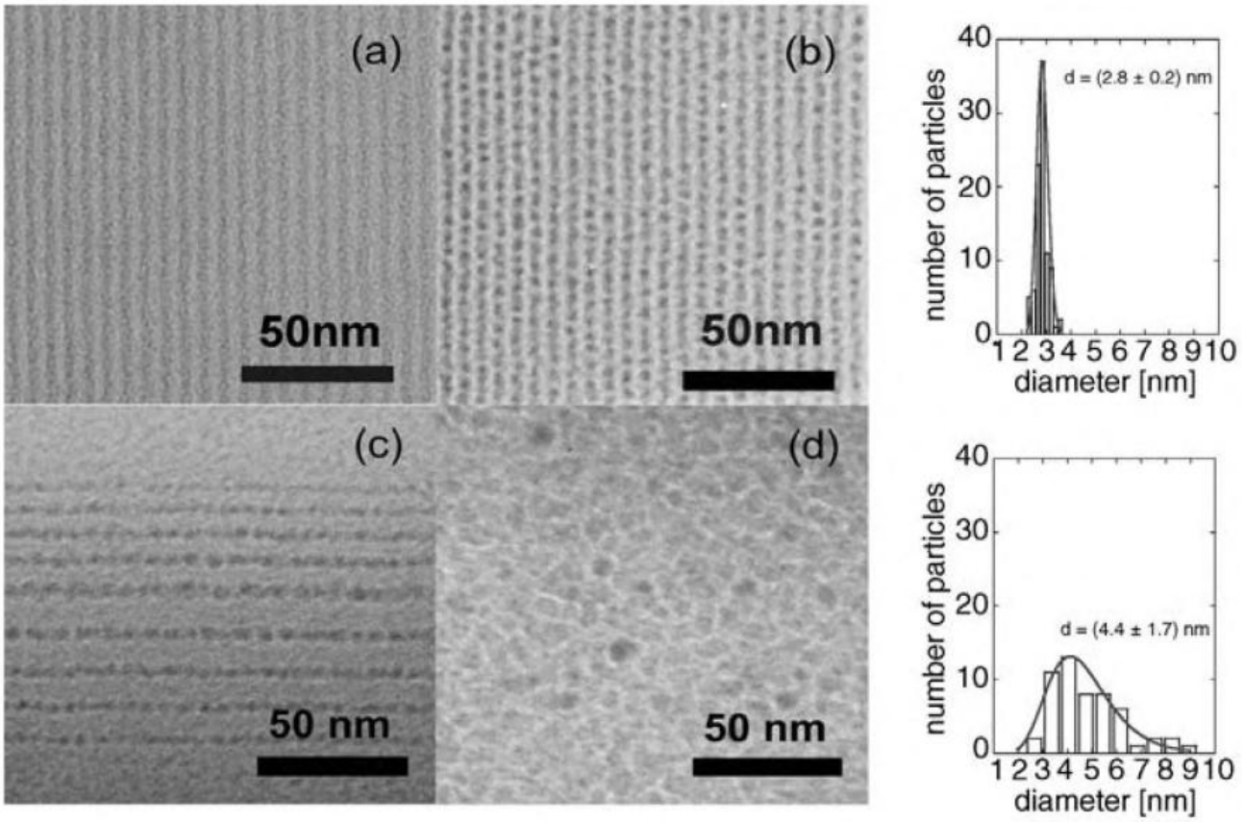


Figure 14.34. Cross-sectional bright field TEM images of (a) an as-prepared amorphous $SiO/Si0_2$ superlattice, (b) the same film after 1100 °C annealing under nitrogen, (c) the of changing the thickness of the SiO layer and the correspondent change in Si NCs size, and (d) a bulk SiO film after annealing for comparison.

Hydrogen stays in the films up to 450-500 °C and might hinder the phase separation that already starts at a lower temperature. The influence of nitrogen is not completely clear as yet, but traces supplied by diffusion during annealing seem to be of advantage. However, SiO_x films grown by CVD with N_20 as oxygen source normally contain up to 12% nitrogen, which stays in the films even at higher temperatures. Furthermore, CVD- grown SiO_x contains high concentrations of hydrogen that can be diffused out by annealing around 450-500 °C but it might hinder the phase separation that starts at lower temperatures. Such films should be called a nonstoichiometric oxynitride and not a SiO_x film as often done in literature. Nitrogen obviously influences and even hinders the clustering of silicon in such oxynitride films, even from the beginning. We will only discuss films prepared by evaporation, that is, prepared without nitrogen.

In Figure 14.34, we compare cross-sectional bright field TEM images of (a) an as- prepared amorphous SiO/SiO_2 superlattice and (b) the same film after 1100 °C annealing under nitrogen.

In (c), we demonstrate the effect of changing the thickness of the SiO layer. Figure 14.34d presents a bulk SiO film after annealing for comparison. The Si nanoctyslals are observed as the darker contrast Size control can be clearly seen here. The thick bulk film contains rather big crystals randomly distributed within the $Si0_2$ matrix whereas the nanocrystals are confined in the former SiO layers with uniform sizes controlled by the former SiO layer thickness. The size control works well for thickness of 2-6 nm. Figure 14.34b and d also gives the respective size distribution estimated from dark field images of samples (b) and (d). If the layer thickness is below 2 nm, then the resulting crystals are still in the range of around 2nm but less in numbers. The reason for that is based on crystallization theory resulting in a critical crystallization diameter of the nanocrystals.

Crystals below the critical size are not stable; they stay as amorphous clusters and contribute to the growth of the bigger ones by Ostwald ripening. If the SiO layer is thicker than 7 nm, then more than one starting nucleus will be established over the layer thickness and the size control is more and more lost. In bulk SiO_x films, there is a random nucleation within the matrix and an Ostwald ripening of the nuclei by diffusion and rearrangement of silicon and oxygen atoms.

The phase separation can be monitored by IR spectroscopy. In the range of 700-1500 cm^{-1}, various silicon-oxygen-related absorption bands can be seen. The band around 810 cm^{-1} can be assigned to Si—O—Si bond bending motion in SiO_2. With higher annealing, a new band at 880 cm^{-1} appears and increases in intensity up to 400-500 °C. For annealing temperatures above 500 °C, this absorption band loses intensity and vanishes at 800 °C.

The main factor of self-assembled processes is the tendency of atomic system to the configuration with minimal of potential energy. Most important process from the other processes taking place in the solid state is spontaneous crystallization. Crystalline state is more stable than amorphous state. Thus amorphous state always predispose to crystallization. It is necessary to note that physical and chemical mechanisms of this process are always depend on properties of surroundings oneself and some external conditions. It is absolutely naturally that the main parameter at this case is the temperature.

The formation of crystalline nucleus is always going to decrease the Gibbs energy of the system $\Delta g = g_{am} - g_{cr}$, where g_{am} - Gibbs energy for amorphous state, g_{cr} - Gibbs energy for crystalline state.

The surface energy of growing nucleus is going to increase the energy of the system. So, the changing of the energy of the system at this case can be written for the unit of the space by means of expression as

$$\Delta G = 4\pi r^2 \sigma^* - \frac{4}{3}\pi r^3 \Delta g,$$

where r – radius of the particle, σ^* - surface energy per unit of the square.

The changing of the energy with the growing of the radius of the nucleus is shown on the Figure 14.35.

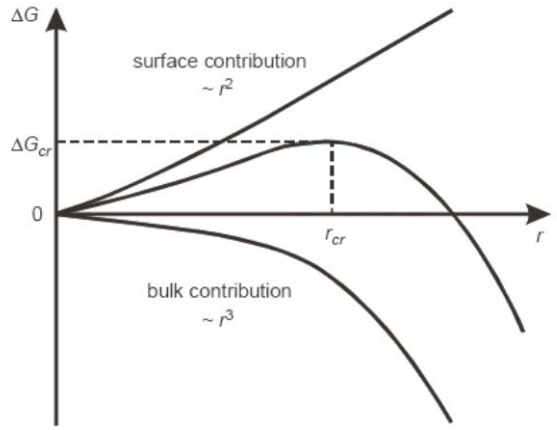


Fig. 14.35. The changing of the Gibbs energy from the radius of nucleus.

$$r_{cr} = \frac{2\sigma^*}{\Delta g}.$$

The changing of the energy has the maximum for the nucleus with critical radius

The creation of nucleus with radius less than r_{cr} can be happened at positive increasing of the energy of the system. Some dynamically equilibrium concentration of these clusters can be exist at this conditions. But system is unstable at this case. Another nucleus with the radius more than r_{cr} has the more preferable energy conditions for growth.

The previously mentioned models of dust formation by condensation are not solely but partly based on equilibrium conditions. Deviations from the equilibrium are hard to describe and hence this is a suitable approach, which is also justified by the observation of the equilibrium products. Reality is likely to be more exciting than equilibrium. The times scales that the dust particles spend in their respective cosmic environments are long and surrounding particles (electrons protons, high energy cosmic rays) and fields provide the energy to potentially change the internal structure of the solids. The small size of the particles may give a further boost to this.

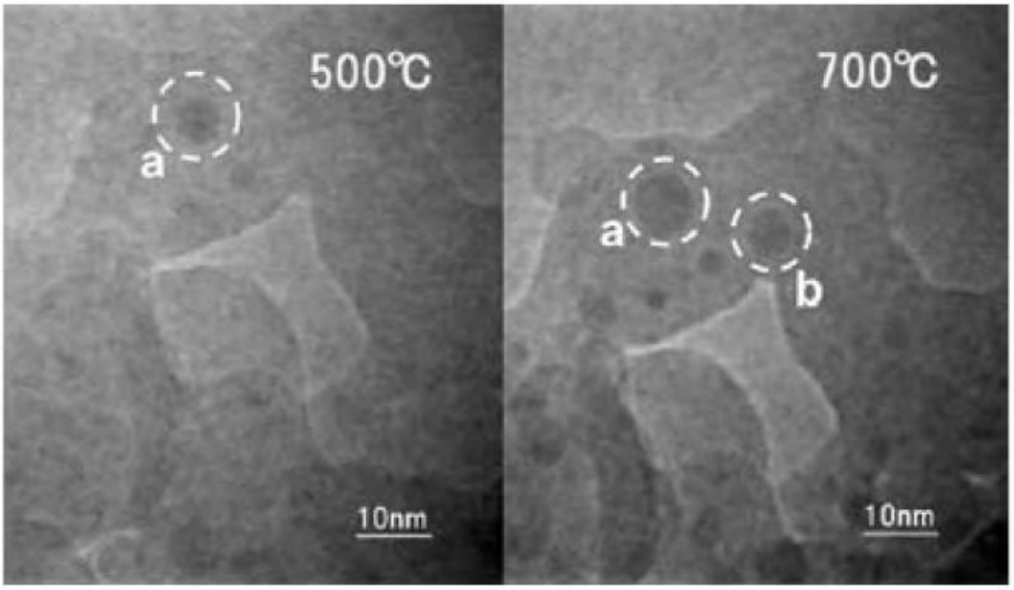


Figure 14.36. Transmission electron microscope image of a mixture of SiO, after heating to 500 and 700 °C Growth of Si nanocrystallites is seen in the black regions marked with white circles.

It is worthwhile to consider laboratory measurements regarding these issues. It was experimentally studied the formation of dust particles in a cooling gas flow to simulate the conditions in stellar outflows: the sizes of the forming condensates are of the order of 20-50 nm and their surfaces are highly reactive. The initial condensation temperature is of the order of 1500 K and the gas subsequently cools down. The detected forming condensates are not those expected for thermodynamic equilibrium conditions. The initial stoichiometric composition being that of Mg and Fe silicates. It was shown that Mg oxides and Fe oxides (such as Fe_2O_3 and Fe_30_4) condensed individually and that subsequently Mg silicate and Fe silicate form separately from the silicon oxide particles. Observations of the atomic ratio using an energy dispersion X-ray analysis system equipped with transmission electron microscope (TEM) showed that the forming SiO_x particles had x close to 1. It is suggested that the Mg and Fe silicates form by aggregation of the condensates and their subsequent processing, so that the minerals that finally form may be close to those that form under equilibrium conditions. Laboratory experiments showed, for instance, the formation of crystalline forsterite grains by coalescence and growth of Mg and SiO smoke particles.

The formation of silicon nanoparticles may occur as a result of internal alteration of other silicon-containing materials. It is shown that since SiO is one of the most abundant and most strongly bonded molecules, it forms and condenses into oxygen-rich stellar outflows. They suggested that following the nucleation of SiO particles, the annealing and the separation into a Si core and a SiO_2 mantle would form the silicon nanoparticles and that the SiO_2 would facilitate

the "passivation" of the Si crystal that is required for quantum confinement effects to occur for the Si nanoparticles. The formation of silicon nanoparticles was also observed in laboratory experiments where evaporated SiO powder recondensed to Si and SiO_2 crystallites. Moreover, experimental studies of SiO_x particles support the hypothesis that Si nanocrystallites form by annealing within a bulk sample. It was reported the detection of Si nanocrystallites of about 10 nm diameters after heating a mixture of SiO_x particles with x=1 (see Figure 14.36). These are detected in transmission electron microscope images by an enhanced contrast at the location of the Si nanocrystallites.

Typical "stacking faults" of the Si nanocrystallites are observed at high-resolution images and the electron diffraction images show the characteristic rings of the silicon cube structure (see Figure 14.37). The Si nanocrysiallites that are formed from SiO_x ($x = 1$) particles are still observed in the samples after heating to about 700 °C when subsequently cooled to room temperature again. Those Si nanocrysiallites formed in $Si0_2$ particle samples survive heating to about 900 °C The Si nanocrysiallites have noinfluenceon the measured IR spectra and it is still open whether photoluminescence can be observed. Luminescence caused by electron irradiation (cathodoluminescence) was observed at 400nm though. The Si nanoparticles in this latter sample were produced by coevaporation of Ag and SiO.

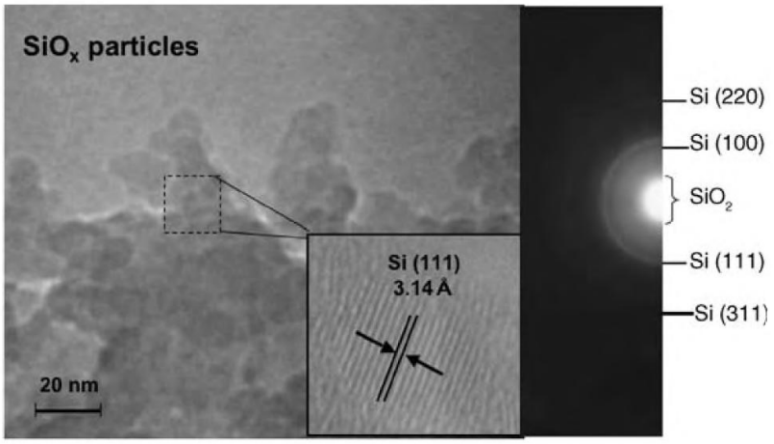


Figure 14.37. SiO_x particles at room temperature after heating to 900 °C. The left-hand side is a transmission electron microscope image and the right-hand side shows the electron diffraction pattern. Dark contrasts with approximately 10 nm diameter indicate the formation of Si crystallites. The high-resolution part of the photograph is taken in approximately the region that is indicated with the dashed square, it shows a structure forming along the (111) planes of cubic Si crystallites. The electron diffraction pattern shows the presence of silicon crystals. The given numbers denote the lattice planes for each crystal and the silicon patterns are for cubic silicon. The innerblurred diffraction rings arc for β-cristobalite, a high-temperature phase of SiO_2.

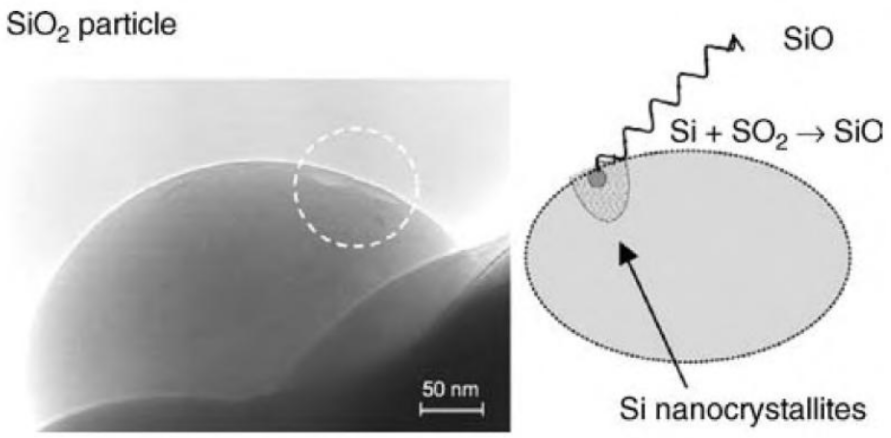


Fig. 14.38. Transmission electron of Si crystallites. A possible mechanism of microscope image (on the left) of SiO_2 particle partial evaporation of the $Si0_2$ particle is at room temperature after annealing at 900 C sketched on the right.

During the formation of Si particles, a significant amount of oxygen can dissolve into the Si particles while the structure of the silicon structure remains. Formation of Si crystallites is also observed during partial sublimation of SiO_2 particles: the SiO_2 on the surface dissolves into SiO and Si, and while the SiO sublimates, the Si crystallites remain in the surface layer of the particle. An image of a particle after heating and a sketch of the suggested process are shown in Figure 14.38. All these findings indicate that Si nanocrystallites most likely do form in cosmic environments and that they typically form within a larger bulk material. Moreover, their formation within the bulk material leaves the element abundances in the solid phase constant, so that it cannot be traced by gas-phase observations.

Chapter 15. Methods for control.

V.Ilchenko

15.1. General classification of characterization methods.

The relatively short period which has elapsed since the invention of the bipolar transistor has seen a dramatic and unprecedented industrial development based on the electrical and optical properties of semiconducting materials. Alongside this industrial development, and closely related to it, has also grown up a basic research activity concerned with gaining a deeper and wider understanding of semiconductors in general and of device materials in particular.

First and foremost of these materials is silicon which provides the basis for the important business in integrated circuits.

At the same time a number of other materials have also reached an advanced state of development where their properties are in demand to provide specific functions not available from silicon.

Examples of materials in this category are GaAs, AlAs, GaP and associated ternary alloys for making light-emitting diodes and injection lasers, GaAs and InP for microwave devices and CdHgTe whose band gap can be tailored for application in the field of infra-red photodetection.

The electrical and optical properties of semiconductors are determined by:

(i) the **chemical composition** of the pure perfect crystal—this determines intrinsic properties such as the fundamental energy gap and the effective masses of the carriers;

(ii) **lattice defects** such as vacancies and interstitials and complexes thereof which introduce electron states within the band gap of the material—such defects occur as a consequence of the way in which the crystal is grown, its thermal history, and as a by-product of doping by ion implantation, though they may also be introduced intentionally by irradiation with energetic particles;

(iii) **chemical impurities** which introduce electron states within the band gap; these states may be near one of the band edges, and have the effect of doping the crystal, or may be deep within the band gap—impurities are usually introduced intentionally during or after crystal growth, but may also be incorporated unintentionally as contamination during any stage of the growth or processing of the material;

(iv) the **dimensions** of the structure—when the dimensions become similar to or less than the de Broglie wavelength of the charge carriers (10-100 Å**)** the electron energy levels of the structures are changed by quantum size effects.

In setting up a characterization activity the first question to ask is that of motivation: why do we want to characterize the material, what is the material to be used for, what properties are important in this application?

Motivations tend to fall into two groups—measurement of material properties with reference to a specification, and experiments intended to give a greater understanding of the physics of the material. The specification may be set by the requirements for a device or by the requirements for some other experiment or investigation; it may refer to material as produced by the growth process or to material which has been modified after growth by other processes such as diffusion, ion implantation and thermal treatment. Measurement against a specification usually implies a feedback of information for adjustment of the process and in most cases the speed of the measurement is of paramount importance. This task of process control may be undertaken on a phenomenological basis, relying on experience and statistical correlations where understanding of fundamental relationships between cause and effect is lacking.

The chemical and physical characterization techniques in this chapter include optical microscopy techniques, electron microscopy, field ion microscopy, scanning probe techniques, X-ray, and spectroscopy techniques methods, some most well-known electrical methods as well as a few others. These characterization techniques are generally more specialized and require more complicated and more expensive equipment than those of the previous chapters. Some methods are used a great deal by a few specialists or are offered as services. For example, secondary ion mass spectrometry is one of the most common characterization methods. Others find use less frequently, but they give important information, often not obtainable by other techniques. Because of the specialized nature of each of the methods in this chapter, only a brief description is given of the principles, the instrumentation, and the most important areas of application. The specialist using any of the methods is already familiar with the details; the nonspecialist is usually not interested in the details, but may be interested in an overview, in the detection limits, the required sample size, and so on.

All analytical techniques are based on similar principles. A primary electron, ion, or photon beam on the sample causes backscattering of the incident particles-waves or the emission of secondary particles-waves. The mass, energy, or wavelength of the emitted entities is characteristic of the target element or compound from which it originated. The distribution of the unknown can be mapped in the *x*-*y* plane and frequently also in depth.

Each of the techniques has particular strengths and weaknesses, and frequently more than one method must be utilized for unambiguous identification. Differences between the various techniques include sensitivity, elemental or molecular information, spatial resolution in *x*, *y*, and z directions, destructiveness, matrix effects, speed, imaging capability, and cost.

The characterization of semiconductor materials and devices frequently requires a measurement of an impurity spatially in the ***x*** and v as well as in the z-dimension. Typical ***x-y*** resolution capabilities are shown in Fig. 15.1. Electron beam and probe methods are suitable for diameters less than 1 μm. Electron beams can be focused to diameters as small as 0.2 nm. Ion beams cover the 1 to 100 μm range and X-rays typically have diameters of 100 /xm and above. There is a dichotomy in the characterization of materials at small dimensions: high sensitivity and small volume sampling are mutually exclusive. Generally decreasing beam diameter results in poorer sensitivity. High sensitivity requires large excitation beam diameters.

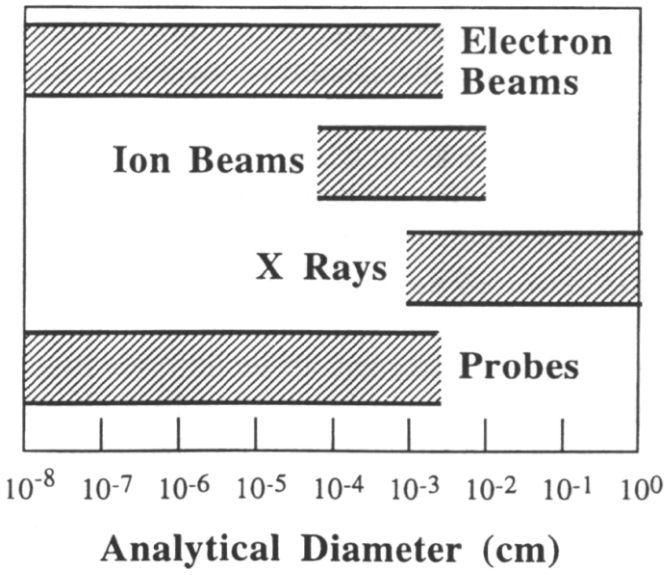


Fig. 15.1 Diameter capabilities of electron-beam, ion-beam, X-ray, and probe characterization techniques.

The term *spectroscopy* is used for characterization techniques that are primarily qualitative in their ability to determine densities even though they may be very quantitative for identifying impurities; *spectrometry* is used for quantitative methods.

15.2. Microscopy techniques.

First of all we discuss those optical characterization techniques most commonly used in the semiconductor industry. Optical measurements are attractive because they are almost always noncontacting with minimal sample preparation—a major advantage when contact formation is detrimental. The instrumentation for many optical techniques is commercially available and has become easier to use and is often automated. The measurements can have very high sensitivity. Optical measurements use the portion of the electromagnetic spectrum from the ultraviolet to the far infrared. Parameters are wavelength (A), energy $(E$ or $h\nu)$, and wavenumber (WN). The most common units are: wavelength in nanometer (1 nm = 10^{-9} m = 10^{-7}cm – 10^{-3} μm), angstrom (1 Å= 10^{-10} m = 1(T^8 cm = 10^{-4} μm) or micrometer (1 μm = 10^{-6} m =10^{-4} cm); energy in electron

volt (1 eV = 1.6 x 10^{-19} J); and wavenumber in inverse wavelength (1 WN = 1/λ). The relationship between energy and wavelength is

The main optical techniques are summarized in Fig. 15.2. Light is either reflected,

$$E = h\nu = \frac{hc}{\lambda} = \frac{1.2397 \times 10^3}{\lambda(nm)} = \frac{1.2397}{\lambda(\mu m)}(eV). \tag{15.1}$$

absorbed, emitted, or transmitted. Most of the techniques in that figure are discussed here; some have been discussed in earlier chapters (e.g., photoconductance) and some are not discussed at all (e.g., ultraviolet photoelectron spectroscopy). For completeness, we also discuss several nonoptical methods for film thickness and line-width determination in this chapter.

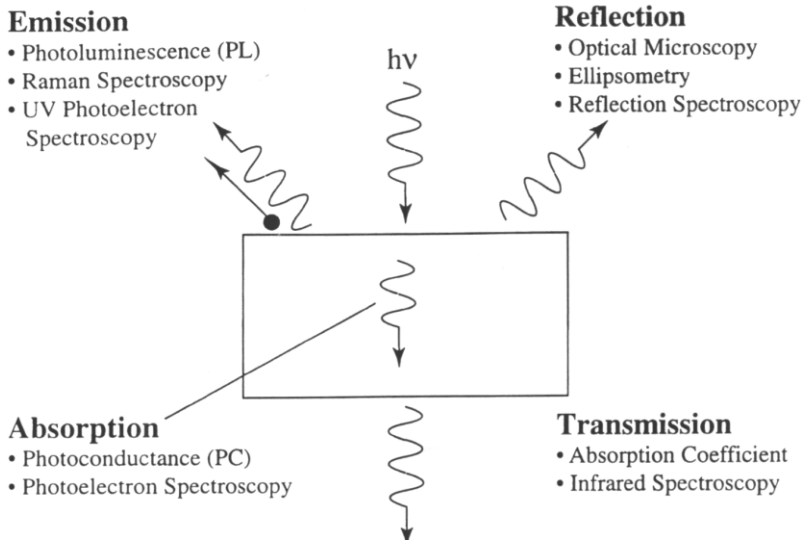


Fig. 15.2 Optical characterization techniques.

The compound optical microscope is one of the most versatile and useful instruments in a semiconductor laboratory. Many of the features of integrated circuits and other semiconductor devices are sufficiently gross to be seen through such a microscope. However, optical microscopy becomes useless as feature sizes shrink to the submicron regime. Typically optical microscopy remains useful for feature sizes above about 0.6-0.8 μ*m*. For smaller sizes, electron beam microscopes become useful. The basic optical microscope can be enhanced by adding phase and differential interference contrast as well as polarizing filters. Optical microcopy is not only used to view the features of integrated circuits; it is also useful for analyzing particles found on such circuits.

To identify and analyze particles requires a skilled and practiced microscopist. The technique is most useful for particles larger than one micron and the analysis depends on matching the unknown with data on known particles. Particle atlases are available to aid in identification. The essential elements of a compound optical microscope are illustrated in Fig. 15.3. Its optical elements, the *objective* and the *ocular* or *eyepiece,* are shown as simple lenses;

in modern microscopes they consist of six or more highly corrected compound lenses. Object O is placed just beyond the first focal point f_{obj} of the objective lens that forms a real and enlarged image I. This image lies just within the first focal point f_{oc} of the ocular, forming avirtual image of I and Γ. A virtual image is an image that does not actually exist and cannot be observed on a screen, for example. The position of Γ may lie anywhere between near and far points of the eye. The objective merely forms an enlarged real image, which is examined by the eye looking through the ocular. The overall magnification *M* is a product of the lateral magnification of the objective and the angular magnification of the ocular. The simplest microscope is the monocular microscope, with only one eyepiece. The binocular instrument has two eyepieces to make viewing of the sample more convenient. When one objective is used with a binocular microscope, the observed image is generally not stereoscopic.

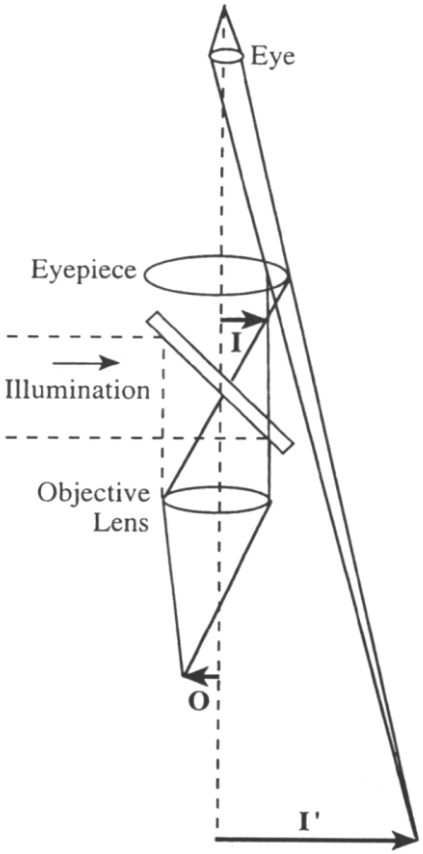


Fig. 15.3 Simplified representation of a compound microscopes optical paths.

Light can be thought of as waves as well as particles. To explain some experimental results, it is easier to use the wave concept, while for others the particle concept is more useful. Waves interfere with one another, placing certain limits on the performance of microscopes. Airy first computed the diffracted image and showed in 1834 that, for diffraction at a circular aperture

$$Sin(\alpha) = \frac{1.22\lambda}{d} \qquad (15.2)$$

of diameter d, the angular position of the first minimum (measured from the center) is given by where λ is the wavelength of light in free space (See Fig. 15.4(a)).

The central spot containing most of the light is called the Airy or diffraction disc. You can do your own experiment by looking at a bright point source at a distance of several meters, for example a microscope lamp, through a small pinhole in a cardboard sheet. The same kind of pattern is formed when a point object is imaged by a microscope. There is no lower size limit of an object that can be detected in isolation, given adequate illumination.

Generally one is not interested in detecting a point object, but a two- or three-dimensional object. Two point objects, a distance s apart, produce overlapping images, as shown in Fig. 15.4(b).

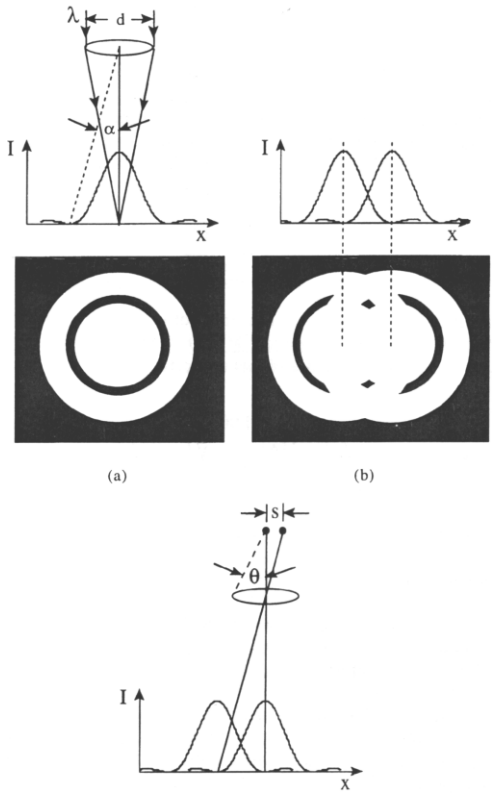


Fig. 15.4 (a) Diffraction at the aperture of a lens showing the Airy disc, (b) the Raleigh criterion for resolution, (c) the resolution limit of an optical microscope. *I* represents the intensity.

If they are too close, it is impossible to resolve them. As they are separated, it becomes possible to tell that there are two objects. The definition of this spacing is somewhat arbitrary. Raleigh suggested that objects can be distinguished when the central maximum of one coincides with the first minimum of the other. The intensity between the two peaks then decreases to 80% of the peak height, as shown in Fig. 15.4(c). The equation

(15.3)

$$s = \frac{0.61\lambda}{n\ Sin(\theta)} = \frac{0.61\lambda}{NA}$$

gives the **resolution** (the minimum distance between points or parts of an object) that satisfies Raleigh's criterion. In the equation, *n* is the refractive index of the medium separating the object from the objective and θ is the half angle subtended by the lens at the object. The **numerical aperture** (NA), usually engraved on the objective mount, is a number that expresses the resolving power of the lens and the brightness of the image it forms.

It is possible to purchase long working distance objectives that allow greater clearance between the objective lens and the sample. While normal objectives may have working distances of 3 mm (20 **X** magnification, 0.46 NA) or 0.5 mm (50 **X** , 0.8 NA), long working distance objectives have 11 mm (20 **X** , 0.4 NA) or 8 mm (50 **X** , 0.55 NA). However, they generally have lower NA and require stronger illumination.

According to Eq. (15.3) three variables may be adjusted to reduce *s* or increase the resolution. The wavelength may be reduced. Blue light has higher resolution than red light. One frequently uses a green filter with its transmission peak at the wavelength for which the objective is chromatically corrected and the eye is most sensitive. Green light also causes the least eye fatigue. The resolution may be improved by increasing the angle θ toward the theoretical maximum of 90°. NA ≈ 0.95 is the upper practical limit. Beyond this, further gain in resolution is achieved by use of immersion objectives in which a fluid with higher index of refraction than air is placed between the sample and the front lens of the objective. With air as the immersion medium, the numerical aperture is sometimes referred to as "dry" NA. Immersion fluids can be water (n = 1.33), glycerin (n = 1.44), oil (n = 1.5 — 1.6), cargille (n = 1.52), or monobromonaphthalene (n = 1.66). Oil is rarely used for semiconductors, but distilled water is sometimes used. Practical limits of NA $\approx$ 1.3-1.4 for oil-immersion optics limit the resolution to $s \approx 0.25$ μm for green light with $\lambda \approx 0.5$ μm.

Magnification *M* is related to the resolving power of the microscope objective and the eye. However, the image must be magnified sufficiently for detailed to be visible to the eye. The **resolving power** is the ability to reveal detail in an object by means of the eye, microscope, camera, or photograph. An approximate relationship for the magnification is

$$M = \frac{\max NA\ of\ microscope}{\min NA\ of\ eye} \approx \frac{1.4}{0.002} = 700 \qquad (15.4)$$

Magnification is sometimes expressed as the ratio of the resolution limits

(15.5)

$$M = \frac{\lim of\ resolution\ (eye)}{\lim of\ resolution\ (microscope)} \approx \frac{200\ \mu m}{0.61\ \lambda / NA} = \frac{200\ \mu m}{0.25\ \mu m} NA = 800$$

where the eye resolution is related to the distance between the rod and cone receptors on the retina of the eye. The maximum magnification of a microscope when the image is viewed by the eye is around 750 **X** . Magnification above this is **empty magnification***;* it gives no additional information. It is also useful when the light detector is not the eye, but photographic film; then higher magnification than that implied by the equations above is possible. The eye fatigues easily if used at its limits of resolving power so it is desirable to supply more magnification than the minimum required for convenience. A reasonable rule is to make the magnification about 750 NA, but one should always use the lowest magnification that permits comfortable viewing. Excessive magnification produces images of lower brilliance and poorer definition, with the result that the amount of object detail that can be seen is frequently reduced.

Contrast—the ability to distinguish between parts of an object—depends on many factors. Dirty eyepieces or objectives degrade image quality. Glare will reduce contrast, especially if the sample is highly reflecting. It is most serious when viewing samples with little contrast and can be controlled to some extent by controlling the field diaphragm, the opening that controls the area of the lighted region. The diaphragm should never be open more than just enough to illuminate the complete field of the microscope. For critical cases it may be reduced to illuminate only a small portion of the normal field.

15.3. Electron microscopy.

Electron beam techniques are presented in Fig. 15.5. Incident electrons are absorbed, emitted, reflected, or transmitted and can, in turn, cause light or X-ray emission. An electron beam of energy E_t incident on the sample surface causes emission of electrons from the surface over a wide range of energies, as illustrated in Fig. 15.6, where the number of electrons emitted by a sample or the electron yield $N(E)$ is plotted against the electron energy E. Three groups of electrons can be distinguished. $N(E)$ shows a maximum for low energy or secondary electrons. The interaction of an electron beam with a solid can lead to the ejection of loosely bound electrons from the conduction band. These are the secondary electrons with energies below about 50 eV with a maximum $N(E)$ at 2 to 3 eV. Auger electrons are emitted in an intermediate energy range. Backscattered electrons that have undergone large-angle elastic collisions leave the sample with essentially the same energy as the incident electrons.

Electrons can be focused, deflected, and accelerated by appropriate potentials; they can be efficiently detected and counted, their energy and angular distribution can be measured, and they do not contaminate the sample or the vacuum system. However, because they are charged, they can cause sample charging that may distort the measurement.

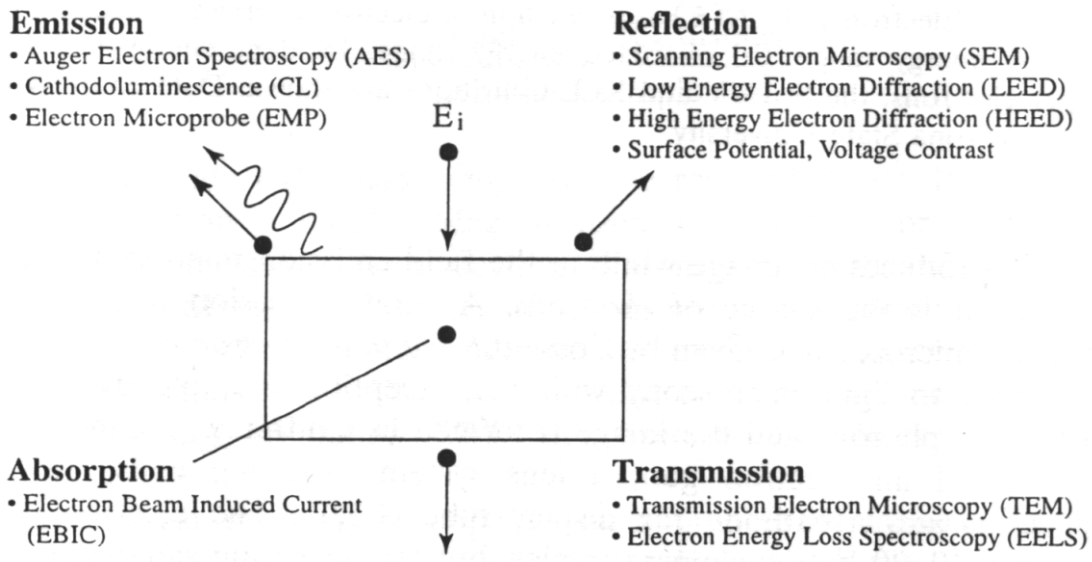
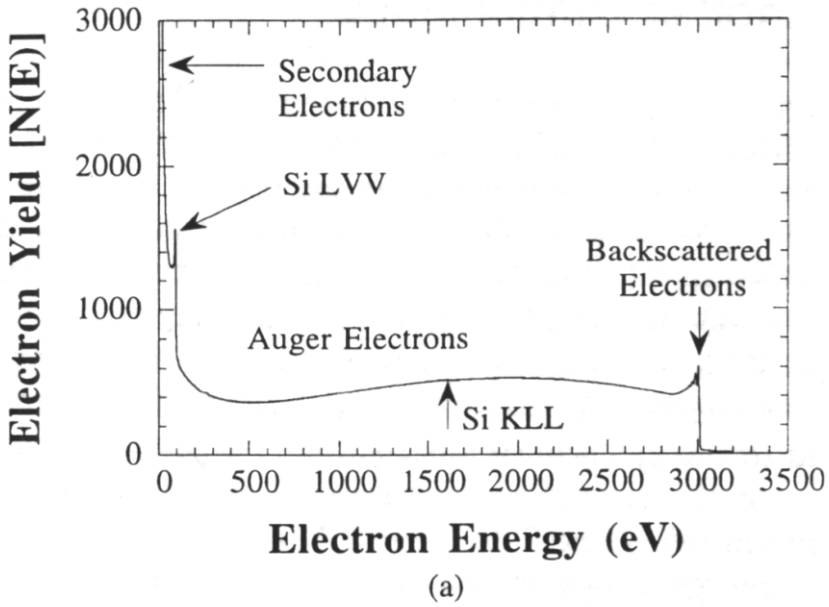


Fig. 15.5 Electron beam characterization techniques.

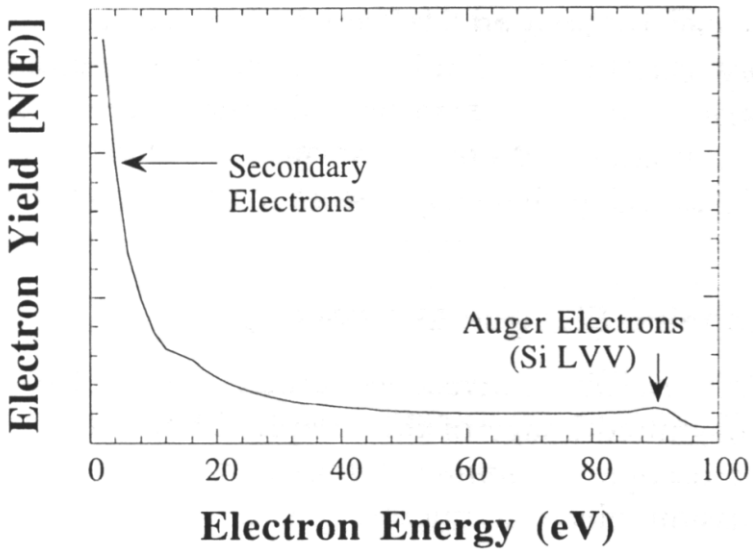
An electron microscope utilizes an electron beam (e-beam) to produce a magnified image of the sample. There are three principal types of electron microscopes: scanning, transmission, and emission. In the scanning and transmission electron microscope, an electron beam incident on the sample produces an image while in the field-emission microscope the specimen itself is the source of electrons. A good discussion of the history of electron microscopy is given by Cosslett. Scanning electron microscopy (SEM) is similar to light microscopy with the exception that electrons are used instead of photons and the image is formed in a different manner. An SEM consists of an electron gun, a lens system, scanning coils, an electron collector, and a cathode ray display tube (CRT). The electron energy is typically 10-30 keV for most samples, but for insulating samples the energy can be as low as several hundred eV. The use of electrons has two main advantages over optical microscopes: much larger magnifications are possible since electron wavelengths are much smaller than photon wavelengths and the depth of field is much higher.

De Broglie proposed in 1923 that particles can also behave as waves. The electron wavelength λ_e depends on the electron velocity υ or the accelerating voltage V as

$$\lambda_e = \frac{h}{m\upsilon} = \frac{h}{\sqrt{2qmV}} = \frac{1.22}{\sqrt{V}}(nm) \qquad (15.6)$$



(a)



(b)

Fig. 15.6 Electron yield $N(E)$ as a function of electron energy for silicon, (a) Entire electron energy range, (b) restricted energy range. Incident energy is *3* keV. For Auger electrons, the Si LW and KLL transitions are shown.

The wavelength is 0.012 nm for V = 10,000 V—a wavelength significantly below the 400 to 700 nm wavelength range of visible light—making the resolution of an SEM much better than that of an optical microscope.

The image in an SEM is produced by scanning the sample with a focused electron beam and detecting the secondary and/or backscattered electrons. We will not concern ourselves with the details of forming a focused electron beam because this is discussed in appropriate books and papers. Electrons and photons are emitted at each beam location and subsequently detected.

Secondary electrons form the conventional SEM image, backscattered electrons can also form an image, X-rays are used in the electron microprobe, emitted light is known as cathodoluminescence, and absorbed electrons are measured as electron-beam induced current. All of these signals can be detected and amplified to control the brightness of a CRT scanned in synchronism with the sample beam scan in the SEM. A one-to-one correspondence is thus established between each point on the display and each point on the sample. Magnification M results from the mapping process according to the ratio of the dimension scanned on the CRT to

$$M = \frac{Length\ of\ CRT\ display}{Length\ of\ sample\ scan} \qquad (15.7)$$

the dimension of the scanned sample

For a 10-cm wide CRT displaying a sample scanned over a 100-μm length, the magnification is 1000 X . Magnifications of 100,000 X or slightly higher are possible in SEMs. It is obvious that high magnifications are easily achieved with SEMs, but low magnifications are more difficult. For a magnification of 10 X the scanned length on the sample is one centimeter, only 10 X smaller than the CRT scan. An SEM typically has one large viewing CRT and a high-resolution CRT with typically 2500 lines resolution forphotography.

The contrast in an SEM depends on a number of factors. For a flat, uniform sample the image shows no contrast. If, however, the sample consists of materials with different atomic numbers, a contrast is observed if the signal is obtained from backscattered electrons, because the backscattering coefficient increases with the atomic number Z. The secondary electron emission coefficient, however, is not a strong function of Z and atomic number variations give no appreciable contrast. Contrast is also influenced by surface conditions and by local electric fields. But the main contrast-enhancing feature is the sample topography. Secondary electrons are emitted from the top 10 nm or so of the sample surface. When the sample surface is tilted from normal beam incidence, the electron beam path lying within this 10 nm is increased by the factor $1/Cos\,\theta$ where θ is the angle from normal incidence (θ = 0° for normal incidence). The interaction of the incident beam with the sample increases with path length and the

$$C = Tan(\theta)\ d\theta \qquad (15.8)$$

secondary electron emission coefficient increases. The contrast C depends on the angle as

For θ = 45°, a change in angle of $d\theta$ =1° produces a contrast of 1.75% while at 60° the contrast increases to 3% for $d\theta$ =1°.

The sample stage is an important component in SEMs. It must allow precise movement in tilt and rotation to enable the sample to be viewed at the appropriate angle. The angle effect is

responsible for the striking three-dimensional nature of SEM images, but the striking pictures come about also due to the signal collection. Secondary electrons are attracted and collected by the detector even if they leave the sample in a direction away from the detector. This does not happen in optical microscopes, where light reflected away from the detector (the eye) is not observed. An SEM forms its picture in an entirely different manner than an optical microscope, where light reflected from a sample passes through a lens and is formed into an image. In an SEM no true image exists. The secondary electrons that make up the conventional SEM image are collected, and their density is amplified and displayed on a CRT. Image formation is produced by mapping, which transforms information from specimen space to CRT space.

A schematic representation of an SEM is shown in Fig. 15.7. Electrons emitted from an electron gun pass through a series of lenses to be focused and scanned across the sample.

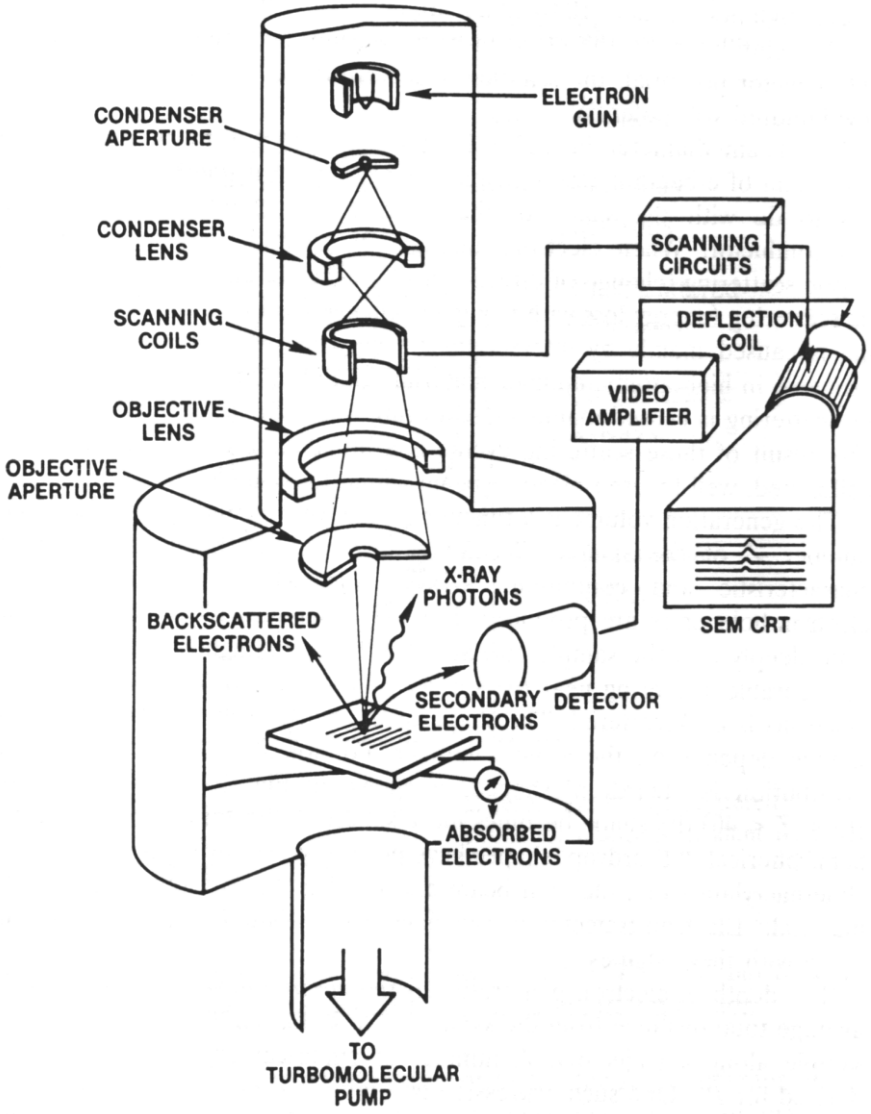


Fig 15.7 Schematic of a scanning electron microscope.

The electron beam should be bright with small energy spread. The most common electron gun is a tungsten "hairpin" filament emitting electrons thermionically with an energy spread of around 2 eV. Tungsten sources have been largely replaced by lanthanum hexaboride (LaB_6) sources with higher brightness, lower energy spread (~ 1 eV) and longer life and by field-emission guns with an energy spread of about 0.2 to 0.3 eV. Field-emission guns are about 100 X brighter than LaB_6 sources and 1000 X brighter than tungsten sources, and they have longer lifetimes. The incident or primary electron beam causes secondary electrons to be emitted from the sample and these are ultimately accelerated to 10 to 12 kV.

They are most commonly detected with an Everhart-Thornley (ET) detector. The basic component of this detector is a scintillation material that emits light when struck by energetic electrons accelerated from the sample to the detector. The light from the scintillator is channeled through a lightpipe to a photomultiplier, where the light incident on a photocathode produces electrons that are multiplied, creating the very high gains necessary to drive the CRT.

High potentials of 10 to 12 kV are necessary for efficient light emission by the scintillator. So the electron beam will not be influenced by the high ET detector potential, the scintillator is surrounded by a Faraday cage at a few hundred volt potential.

The beam diameter in SEMs is in the range of 1 to 10 nm. Yet the resolution of e-beam measurements is not always that good. Why is that? It has to do with the shape of the electron-hole cloud generated in the semiconductor. When electrons impinge on a solid, they lose energy by elastic scattering (change of direction with negligible energy loss) and inelastic scattering (energy loss with negligible change in direction). Elastic scattering is caused mainly by interactions of electrons with nuclei and is more probable in high atomic number materials and at low beam energies. Inelastic scattering is caused mainly by scattering from valence and core electrons. The result of these scattering events is a broadening of the original nearly collimated, well-focused electron beam within the sample.

The generation volume is a function of the e-beam energy and the atomic number Z of the sample. Secondary electrons, backscattered electrons, characteristic and continuum X-rays, Auger electrons, photons, and electron-hole pairs are produced. For low-Z samples most electrons penetrate deeply into the sample and are absorbed. For high-Z samples there is considerable scattering near the surface and a large fraction of the incident electrons is backscattered. The shape of the electron distribution within the sample depends on the atomic number. For low-Z material ($Z < 15$) the distribution is "teardrop"-shaped, as shown in Fig. 15.8. As Z increases ($15 < Z < 40$) the shape becomes more spherical and for $Z > 40$ it becomes hemispherical. "Teardrop" shapes have been observed by exposing polymethylmethacrylate to an

electron beam and etching the exposed portion of the material. Electron trajectories, calculated with Monte Carlo techniques, also agree with these shapes.

The depth of electron penetration is the electron range R_e, defined as the average total distance from the sample surface that an electron travels in the sample along a trajectory. A number of empirical expressions have been derived for R_e. One such expression is

$$R_e = \frac{4.28\times10^{-6}\ E^{1.75}}{\rho} \tag{15.9}$$

where ρ is the sample density (g/cm^3) and E the electron energy (keV).

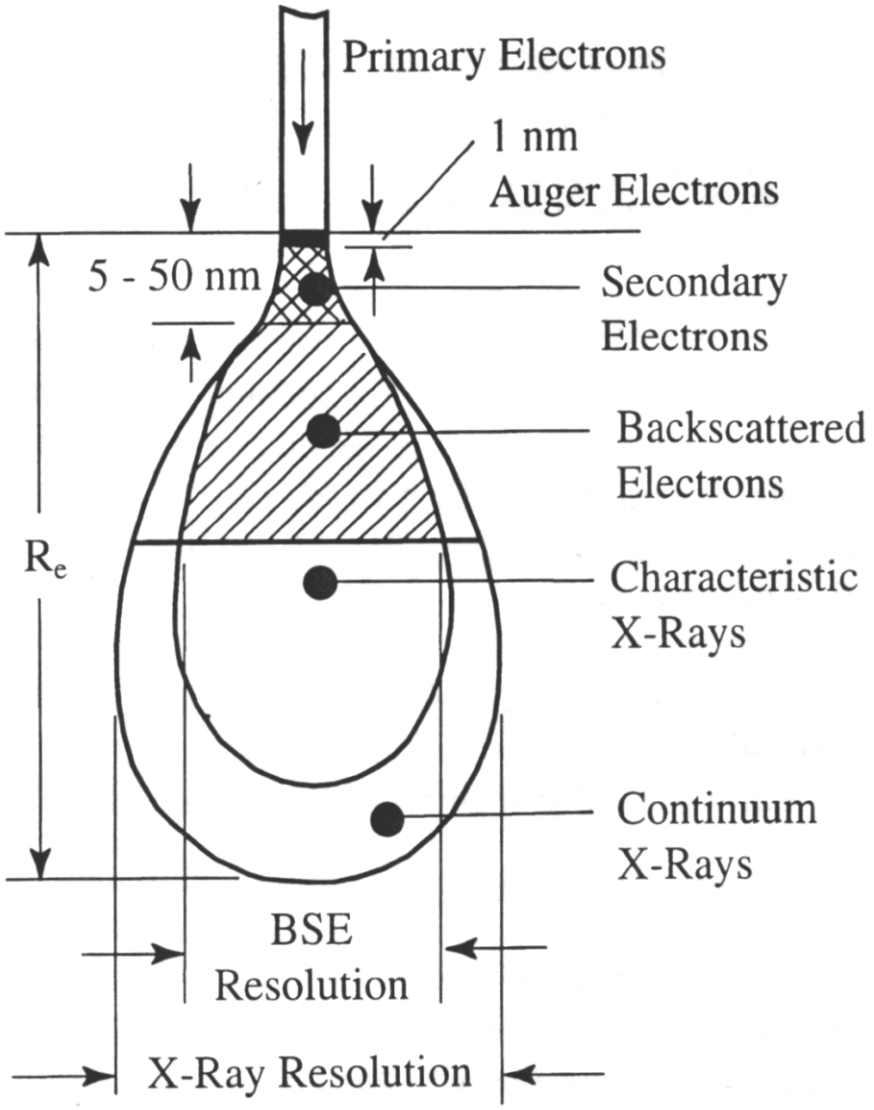


Fig. 15.8 Summary of the range and spatial resolution of backscattered electrons, secondary electrons, X-rays, and Auger electrons for electrons incident on a solid.

The most common use of SEMs for semiconductor applications, when used as a microscope, is to view the surface of the device, frequently during failure analysis and for cross-sectional analysis to determine device dimensions, for example MOSFET channel length, junction depth, and so on. SEM resolution is high (as good as ~ 1 nm), because secondary electrons emerge from the top few nm of the sample. In the past, the SEM was mainly a research tool, but more recently it has moved to the wafer processing production line for on-line inspection and line-width measurement. When inspecting integrated circuits, it is important to reduce or eliminate surface charging when electrons landing on insulators cannot discharge to ground potential. Surface charging is eliminated by coating the surface with a thin conductive layer (Au, AuPd, Pt, PtPd, and Ag provide an oxide-free surface) or by reducing the beam energy until the number of primary electrons is roughly equal to the number of secondary and backscattered electrons. The energy for this balance is around 1 keV, which is also sufficiently low to minimize electron beam damage to devices. The reduced signal-to-noise ratio of low-energy beams is optimized by using high beam brightness and digital frame storage for signal enhancement.

The most commonly used electron beam technique for failure analysis is voltage contrast. It makes use of the secondary yield influenced by local electric fields. We illustrate this effect in Fig. 15.9 with three conductors or lines at various potentials. In Fig. 15.9(a) all three lines are at ground potential and a certain number of electrons are collected by the detector. Fewer electrons are collected at the detector from a line at a 5 V potential in Fig. 15.9(b) than from a line at ground potential. Similarly, a -5 V line gives a still higher signal. The reason, of course, is that electrons emitted from a line at a positive potential experience not only the attractive potential of the detector, but also the attractive potential of the emitting line. This allows line voltages to be determined. Using stroboscopic techniques, one can measure the transient behavior of an IC, i.e., observe the circuit switch from one state to another. This is the principle of voltage contrast.

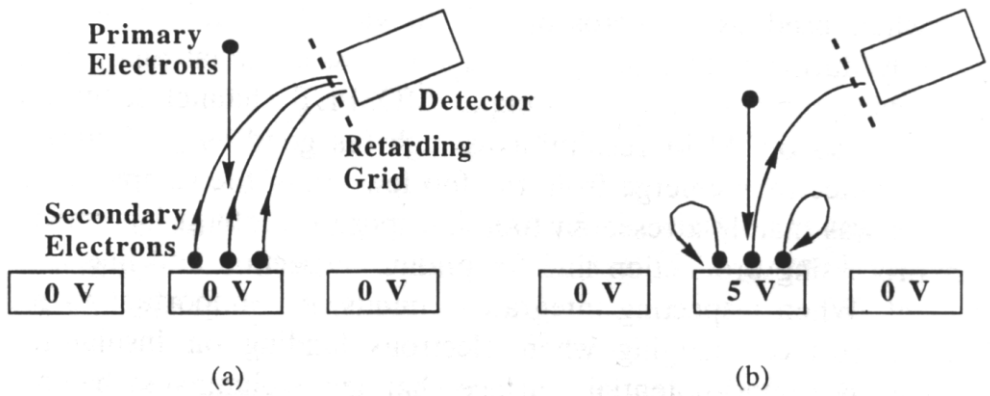


Fig. 15.9 Voltage contrast showing the effects of (a) ground potential and (b) positive potential on electron detection.

Electron beams have a number of advantages over mechanical probes for IC failure analysis. These are that the beam is small, allowing narrow lines to be contacted, there is no capacitive loading of the circuit (important during transient analysis), it has high resolution, voltages can be measured to the millivolt range and voltage waveforms into the subnanosecond range. Voltage contrast measurements are illustrated in Fig. 15.10. The e-beam voltage x-y image in Fig. 15.10(a) shows the state of the various IC lines. Light corresponds to high voltage and dark to low voltage. If, for example, one of the lines had an open circuit, as might happen from electromigration, this wouldclearly show in such an image, but would be very difficult to detect by other means.

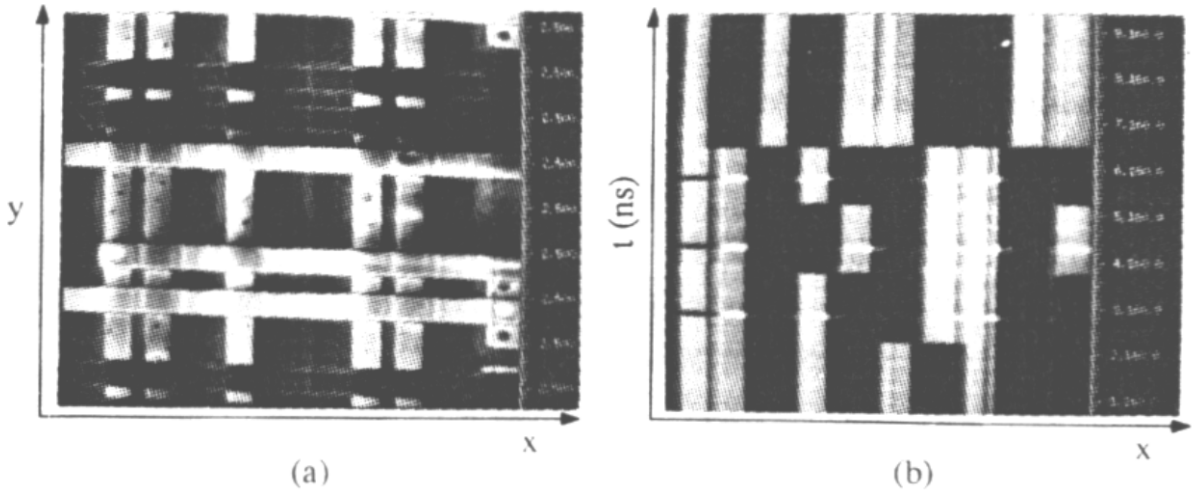


Fig. 15.10 Voltage contrast images in the (a) *x*-*y* and (b) *x*-time configuration.

In Fig. 15.10(b) we show the time dependent behavior. Such an image is obtained by setting the beam at a particular y location and scanning the beam in x and time. The transition of a line from high to low or low to high is clearly shown. In this mode, one can compare the circuit behavior to see if the various portions of an IC switch as they should. If, for example, the line resistance increases due to electromigration, the RC switching time may be affected. This voltage contrast measurement will display it.

Transmission electron microscopy (TEM) was originally used for highly magnified sample images. Later, analytical capabilities such as electron energy loss detectors and light and X-ray detectors were added to the instrument and the technique is now also known as analytical transmission electron microscopy (AEM). The "M" in TEM and AEM stands for either "microscopy" or"microscope." Transmission electron microscopes are, in principle, similar to optical microscopes; both contain a series of lenses to magnify the sample. The main strength of TEM lies in its extremely high resolution, approaching 0.15 nm. The reason for this high resolution can be found in the resolution equation Eq. (15.3), s=0.61 λ/NA. In optical

microscopy, the numerical aperture NA≈1 and λ≈500 nm, giving s≈300 nm. In electron microscopy, NA is approximately 0.01 due to larger electron lens imperfections, but the wavelength is much shorter. According to Eq. (15.6), $\lambda_e \approx 0.004$ nm for V=100,000 V, giving a resolution of s≈0.25 nm and magnifications of several hundred thousand—much better than optical microscopy. The actual resolution expression is more complicated and this simple calculation should only be taken as a coarse estimate. A shortcoming of TEM is its limited depth resolution.

A schematic of a transmission electron microscope is shown in Fig. 15.11. Electrons from an electron gun are accelerated to high voltages—typically 100 to 400 kV—and focused on the sample by the condenser lenses. The sample is placed on a small copper grid a few mm in diameter. The static beam has a diameter of a few microns. The sample must be sufficiently thin (a few tens to a few hundred nm) to be transparent to electrons. This circumvents the resolution problem of Fig. 15.8 because the beam does not have a chance to spread when the sample is so thin.

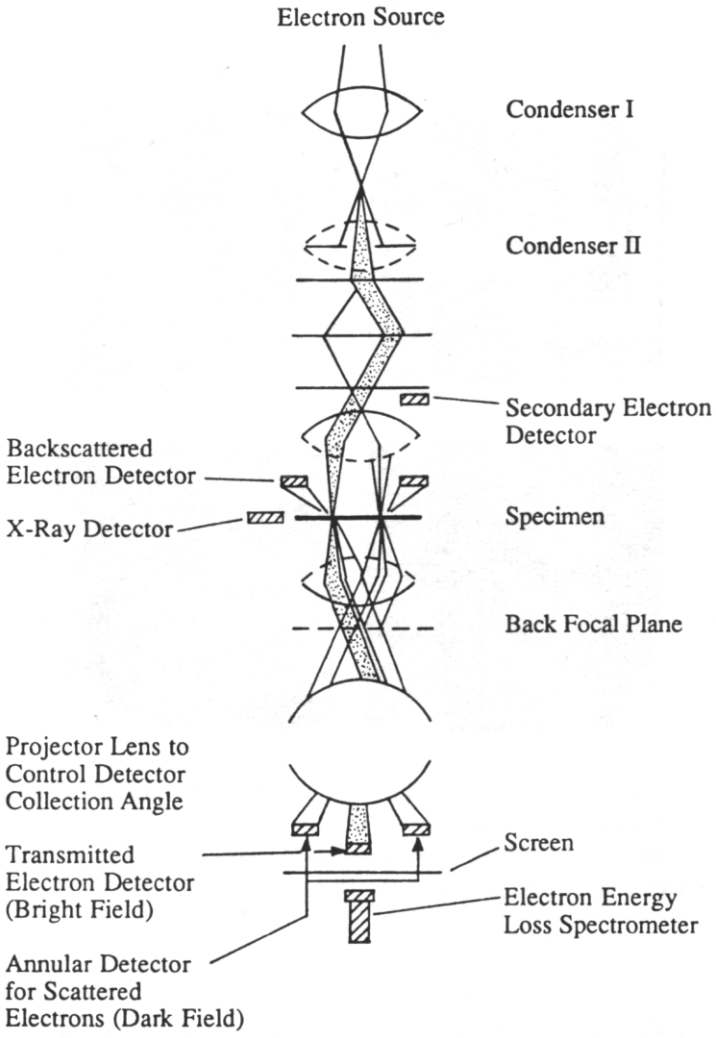


Fig. 15.11 Schematic of a transmission electron microscope.

The transmitted and forward scattered electrons form a diffraction pattern in the back focal plane and a magnified image in the image plane. With additional lenses, either the image or the diffraction pattern is projected onto a fluorescent screen for viewing or for electronic or photographic recording. The ability to form a diffraction pattern allows structural information to be obtained.

The three primary imaging modes are bright-field, dark-field, and high-resolution microscopy. Image contrast does not depend very much on absorption, as it does in optical transmission microscopy, but rather on scattering and diffraction of electrons in the sample. Images formed using only the transmitted electrons are bright-field images and images formed using a specific diffracted beam are dark-field images. Few electrons are absorbed in the sample. Absorbed electrons lead to sample heating to change the sample during the measurement.

Consider an amorphous sample consisting of atoms ***A*** with inclusions of atoms ***B,*** where $Z_B > Z_A$ (Z = atomic number). Electrons experience very little scattering from atoms ***A,*** but are more strongly scattered by atoms ***B.*** The more strongly scattered electrons are not transmitted by the image forming lenses and do not reach the fluorescent screen but the weakly scattered electrons do. Hence, the heavier elements do not appear on the screen. In other words, the image brightness is determined by the intensity of those electrons transmitted through the sample that pass through the image forming lenses. For crystalline specimen, the wave nature of electrons must be considered and Bragg diffraction of electrons by the sample crystal planes occurs. Electrons "make it" to the screen if they are not deflected by Bragg diffraction. Contrast comes about by mass contrast, thickness contrast, diffraction contrast, and phase contrast.

A stationary, parallel, coherent electron beam passes through the sample in TEM forming a magnified image in the image plane which is then simply projected onto a fluorescent screen. In *scanning transmission electron microscopy* (STEM) a fine beam (diameter ~ 0.5 nm) is scanned across the sample in a raster fashion. The objective lens recombines the transmitted electrons from all points scanned by the probe beam to a fixed region in the back focal plane where they are detected by an electron detector. The detector output modulates the brightness of a CRT, much as secondary electrons do in an SEM. The primary electrons in an STEM also produce secondary electrons, backscattered electrons, X-rays, and light (cathodoluminescence) above the sample much as in SEMs. Below the sample, inelastically scattered transmitted electrons can be analyzed for electron energy loss, making the instrument truly an analytical electron microscope. X-ray analysis has become an important aspect of transmission electron microscopy at magnifications much higher than possible for EMP in an SEM. However, the volume for X-ray generation is much smaller, giving much weaker X-ray intensity than in SEMs. The integration time for each picture element in STEM is limited since the data are collected

serially. This makes for poorer images than obtained in TEM, where all picture elements are integrated simultaneously in an imaging system.

Pictures by TEM have taken on an important role in semiconductor integrated circuit development. In particular, cross-sectional images through semiconductor devices have brought out many aspects of process-induced features not obtainable with other techniques. AEM success is very much dependent on sample preparation. Semiconductor devices usually contain semiconductors, insulators, and metals and it is difficult to use chemical etches to thin all layers simultaneously. Additionally, the devices are small and it is very difficult to locate particular features of interest.

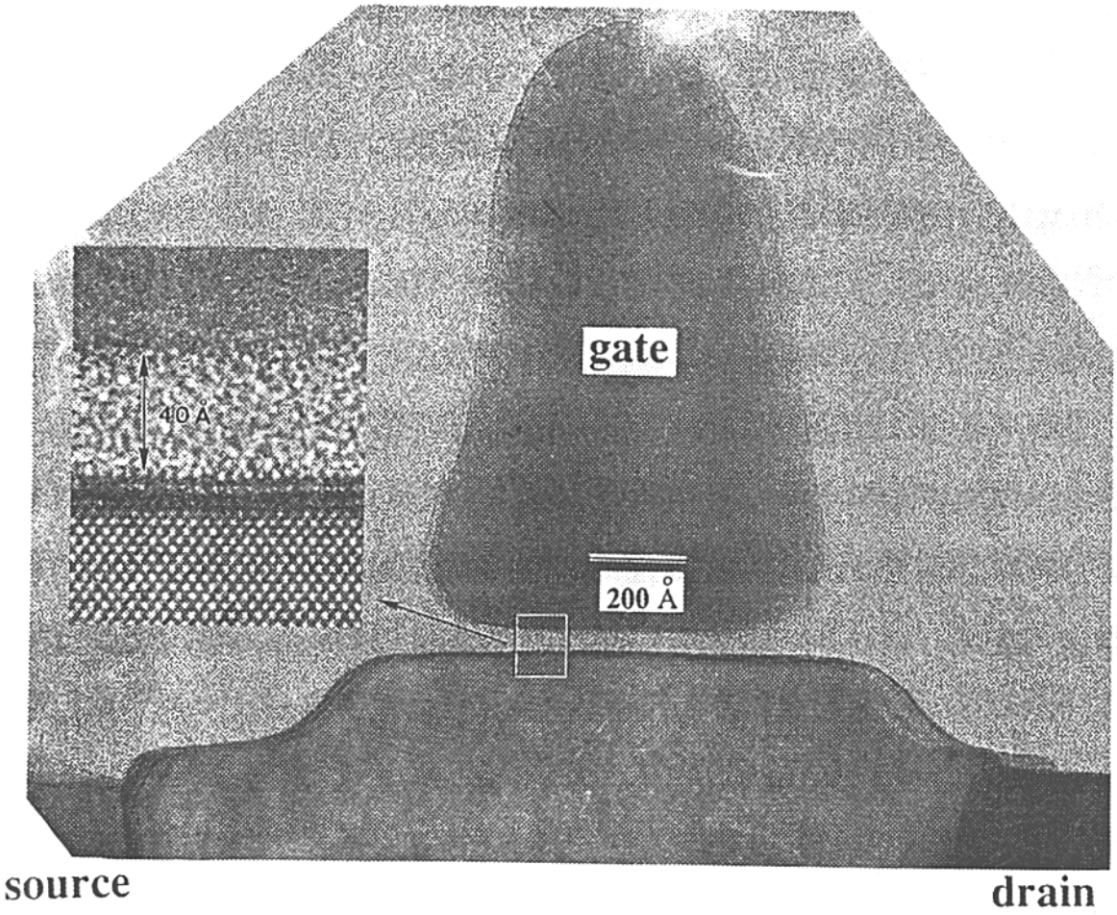


Fig. 15.12 TEM micrograph of a Si MOSFET.

High resolution TEM (HREM) has come of age during the past decade or so. It gives structural information on the atomic size level, is known as lattice imaging, and has become very important for interface analysis. For example, oxide-semiconductor, metal-semiconductor, and semiconductor-semiconductor interfaces have benefited a great deal from HREM images. In lattice imaging a number of different diffracted beams are combined to give an interference image. In Fig. 15.12 we show a cross section of a MOSFET with a 0.1 μm channel length and a 4 nm oxide thickness.

15.4. Field ion microscopy.

Ion beam characterization techniques are illustrated in Fig. 15.13. Incident ions are absorbed, emitted, scattered, or reflected and can, in turn, cause light, electron or X-ray emission. Aside from characterization, ion beams are also used for processing as in ion implantation. People usually discuss most widespread two main ion beam material characterization methods: secondary ion mass spectrometry (SIMS) and Rutherford backscattering spectrometry (RBS).

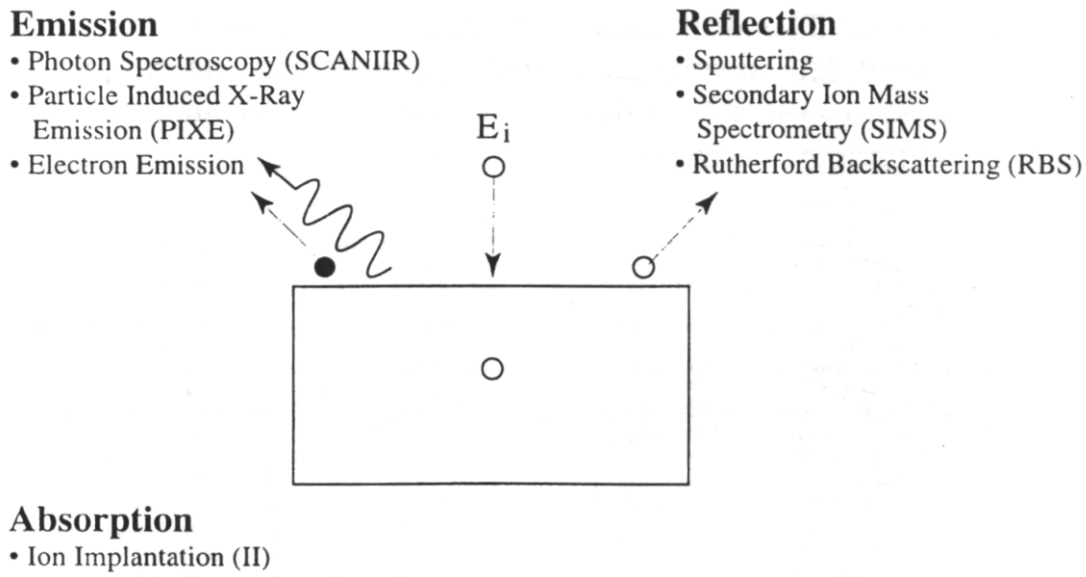


Fig. 15.13 Ion beam characterization.

Secondary ion mass spectrometry (SIMS), also known as ion microprobe and ion microscope, is one of the most powerful and versatile analytical techniques for semiconductor characterization. It was developed inde pendently by Castaing and Slodzian at the University of Paris and by Herzog and collaborators at the GCA Corp. in the United States in the early 1960s, but did not become practical until Benninghoven showed that it was possible to maintain the surface integrity for periods well in excess of the analysis time. Benninghoven did much to further the evolution and advances of SIMS. The technique is element specific and is capable of detecting all elements as well as isotopes and molecular species. Of all the beam techniques it is the most sensitive, with detection limits for some elements in the 10^{14} to 10^{15} cm^{-3} range because there is very little background interfere ence signal. Lateral resolution is typically 100 μ*m* but can be as small as 0.5 μ*m* with depth resolution being 5 to 10 nm.

The basis of SIMS, shown in Fig. 15.14, is the destructive removal of material from the sample by sputtering and the analysis of that material by a mass analyzer. A primary ion beam impinges on the sample and atoms from the sample are sputtered or ejected from the sample.

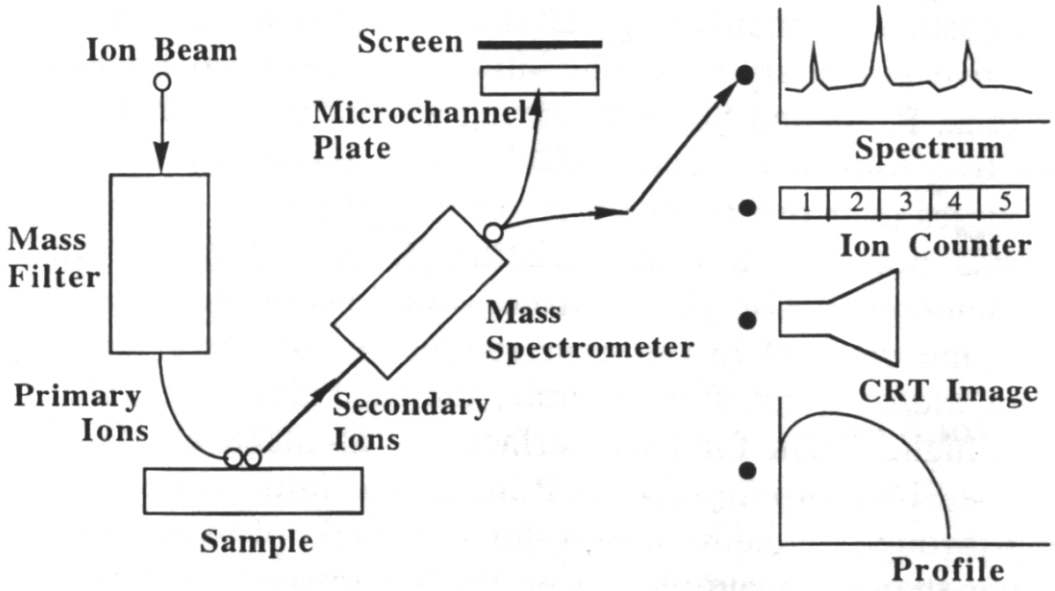


Fig. 15.14 SIMS schematic.

Most of the ejected atoms are neutral and cannot be detected by conventional SIMS. But a small fraction is ejected as positive or negative ions. This fraction was estimated as about 1% of the total in 1910, an estimate that is still considered reasonable. The mass/charge ratio of the ions is analyzed, detected as a mass spectrum, as a count, or displayed on a fluorescent screen or on a CRT. The detection of the mass/charge ratio can present a problem, since various complex molecules form during the sputtering process between the sputtered ions and light elements like H, C, O, and N typically found in SIMS vacuum systems. The mass spectrometer only recognizes the total mass/charge ratio and can mistake one ion for another.

Sputtering is a process in which incident ions lose their energy mainly by momentum transfer as they come to rest within the solid. In the process they displace atoms within the sample. The incident projectiles need not be ions; neutral beam bombardment causes sputtering also, but ions are used in SIMS. Sputtering takes place when atoms near the surface receive sufficient energy from the incident ion to be ejected from the sample. The escape depth of the sputtered atoms is generally a few monolayers for primary energies of 10 to 20 keV typically used in SIMS. The primary ion loses its energy in the process and comes to rest tens of nm below the sample surface. Ion bombardment leads not only to sputtering, but also to ion implantation and lattice damage. The sputtering yield is the average number of atoms sputtered per incident primary ion; it depends on the sample or target material, its crystallographic orientation, and the nature, energy, and incidence angle of the primary ions. Selective or preferential sputtering can occur in multicomponent or polycrystalline targets when the components have different sputtering yields. The component with lowest yield becomes enriched at the surface while that with the highest yield becomes depleted. However, once an equilibrium

situation is reached, the sputtered material leaving the surface has the same composition as the bulk material and preferential sputtering is not a problem in SIMS analysis.

The yield for SIMS measurements with Cs^+, , O^-, and Ar^+ ions of 1 to 20 keV energy ranges from 1 to 20. What is important, however, this is not the total yield, but the yield of ionized ejected atoms or the *secondary ion yield,* because only ions can be detected. The secondary ion yield is significantly lower than the total yield, but can be influenced by the type of primary ion. Electronegative oxygen (O_2^+) is a secondary ion yield, enhancing species for electropositive elements (e.g., B and Al in Si) which produce predominantly positive secondary ions. The situation is reversed for electronegative elements (e.g., P, As and Sb in Si) having greater yields when sputtered with electropositive ions like cesium (Cs^+). The secondary ion yield for the elements varies over five to six orders of magnitude.

SIMS has not only a wide variation in secondary ion yield between different elements, it also shows strong variations in the secondary ion yield from the same element in different samples or matrices. The latter is the well-known *matrix effect.* For example, the secondary ion yield for oxidized surfaces is higher than for bare surfaces by as much as 1000. A striking example is a SIMS profile of B or P implanted into oxidized Si obtained by sputtering through an oxidized Si wafer. The yield of Si in SiO_2 is about 100 times higher than the yield of Si from the Si substrate. A plot of yield versus sputtering time shows a sharp drop when the sample is sputtered through the $Si0_2$-Si interface.

SIMS can give three types of results. For low incident ion beam current or low sputtering rate ($\sim$0.1 nm per hour), a complete mass spectrum can be recorded for surface analysis of the outer 0.5 nm or so. This mode of operation is known as *static* SIMS. In *dynamic* SIMS, the intensity of one peak for one particular mass is recorded as a function of time as the sample is sputtered at a higher sputter rate ($\sim$10 μm per hour), yielding a depth profile. It is also possible to display the intensity of one peak as a two-dimensional image. The various output signals are illustrated in Fig. 15.14.

Quantitative depth profiling is unquestionably the major strength of SIMS, with one selected mass plotted as secondary ion yield versus sputtering time. Such a plot must be converted to density versus depth. The conversion of signal intensity to density can, in principle, be calculated knowing the primary ion beam current, the sputter yield, the ionization efficiency, the atomic fraction of the ion to be analyzed, and an instrumental factor. Some of these factors are generally poorly known and a successful technique for routine quantitative SIMS analysis has not yet emerged. The usual approach is one of using standards with composition and matrices identical or similar to the unknown. Ion implanted standards are very convenient and also very accurate.

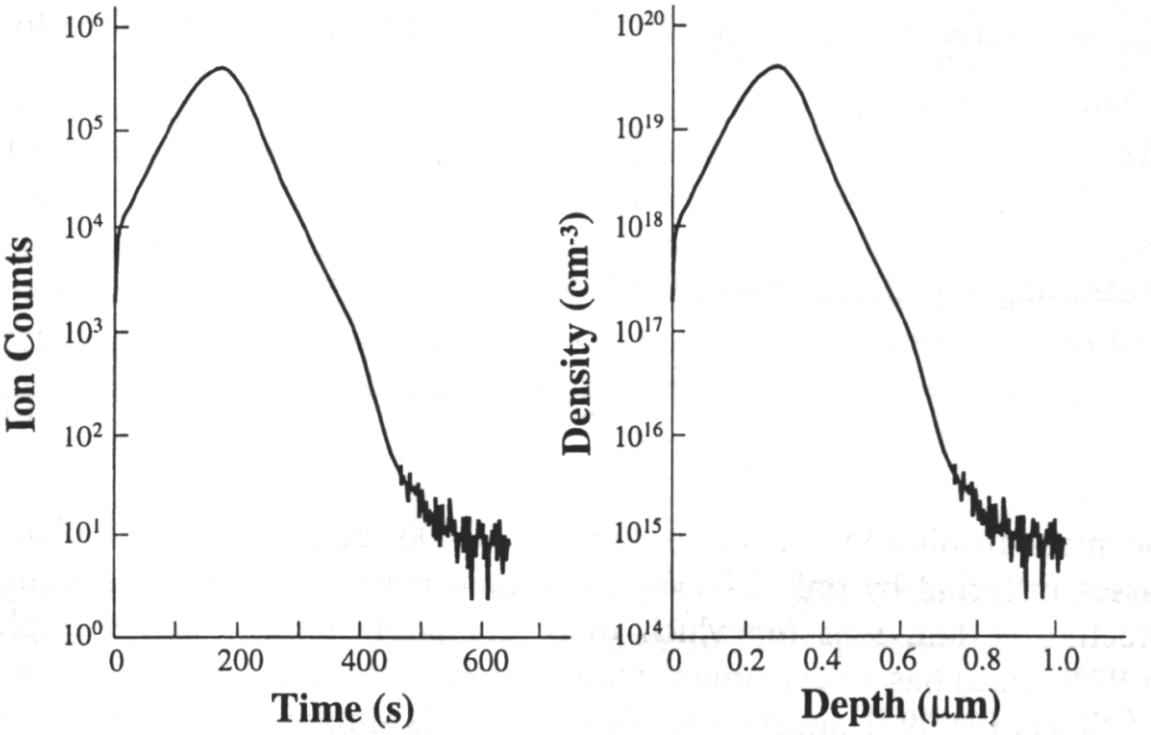


Fig 15.15 (a) Raw $^{11}B^{+}$ secondary ion signal versus sputtering time, (b) boron profile for a boron implant into a silicon substrate.

The implant dose of an ion-implanted standard can be controlled to an accuracy of 5% or better. When such a standard is measured, one calibrates the SIMS system by integrating the secondary ion yield signal over the entire profile. Calibrated standards are, therefore, very important for accurate SIMS measurements. The time-to-depth conversion is usually made by measuring the sputter crater depth after the analysis is completed. An example of the conversion of yield or intensity versus time to density versus depth profile is given in Fig. 15.15, showing both the raw SIMS plot and the dopant density profile.

There are two instrumentation approaches to SIMS: the *ion microprobe* and the *ion microscope.* The ion microprobe is an ion analog of the electron microprobe. The primary ion beam is focused to a fine spot and rastered over the sample surface. The secondary ions are mass analyzed and the mass spectrometer output signal is displayed on a CRT in synchronism with the primary beam to produce a map of secondary ion intensity across the surface. The spatial resolution is determined by the spot size of the primary ion beam and resolutions less than 1 μ*m* are possible. The mass spectrometer consists of electrostatic and magnetic sector analyzers in tandem. In the electrostatic analyzer, the ions travel between two parallel plates separated a distance *d* with a radius of curvature r_V. A potential *V* between the two plates permits only those ions with the properenergy *E* to be transmitted without striking either plate, where *E* is

$$E = \frac{qVr_V}{2d} \tag{15.10}$$

In the magnetic sector spectrometer, a magnetic field B curves the ion of mass m, charge q, and energy E into a path of radius r_B according to

$$\frac{m}{q} = \frac{qB^2 r_B^2}{2E} \tag{15.11}$$

Substituting Eq. (15.10) into (15.11) gives

$$\frac{m}{q} = \frac{B^2 r_B^2 d}{Vr_V} \tag{15.12}$$

The mass resolution can be as high as 40,000, equivalent to resolving two masses differing by only 0.003%. Such high mass resolution is required for detecting certain ions for which there are interferences. For example, ^{31}P (31.9738 amu) has a very similar mass/charge ratio to a $^{30}Si^{1}H$ (31.9816 amu) or $^{29}Si^{1}H_2$ (31.9921 amu) molecule and ^{54}Fe is similar to the $^{28}Si_2$ dimer.

The ion microscope is a direct imaging system, analogous to an optical microscope or a TEM. The primary ion flood beam illuminates the sample, and secondary ions are imultaneously collected over the entire imaged area with a resolution on the order of 1 μm. The spatial distribution of the secondary image is preserved through the system using an electrostatic and magnetic sector analyzer in tandem, amplified by a microchannel plate, and displayed on a fluorescent screen. A small aperture may be inserted to select an area for analysis with a resolution of 1 μm or better. Ion imaging is also done by raster scanning the ion beam across the sample and measuring and displaying the secondary ion intensity as a function of the lateral position of the small spot scanning ion beam. The lateral resolution of this imaging method is dependent on the beam size, which can be as small as 50 nm.

SIMS has found its greatest utility in semiconductor characterization. In particular, it has found widespread application in dopant profiling. For a more detailed discussion and comparison with spreading resistance measurements. SIMS measurements are well suited for semiconductor applications, because matrix effects are of minor consequence and ion yields can be assumed to be linearly proportional to densities up to 1%. Furthermore, the substrate sputters very uniformly, at least for Si. An example profile in Fig. 15.16, shows that arsenic, boron, and oxygen can be determined in a single measurement. This sample was formed by diffusing As and

B from a poly-Si layer deposited on the Si substrate. The plot shows the location of the junction ($N_{As}=N_B$) and the location of the poly- Si substrate interface (oxygen peak).

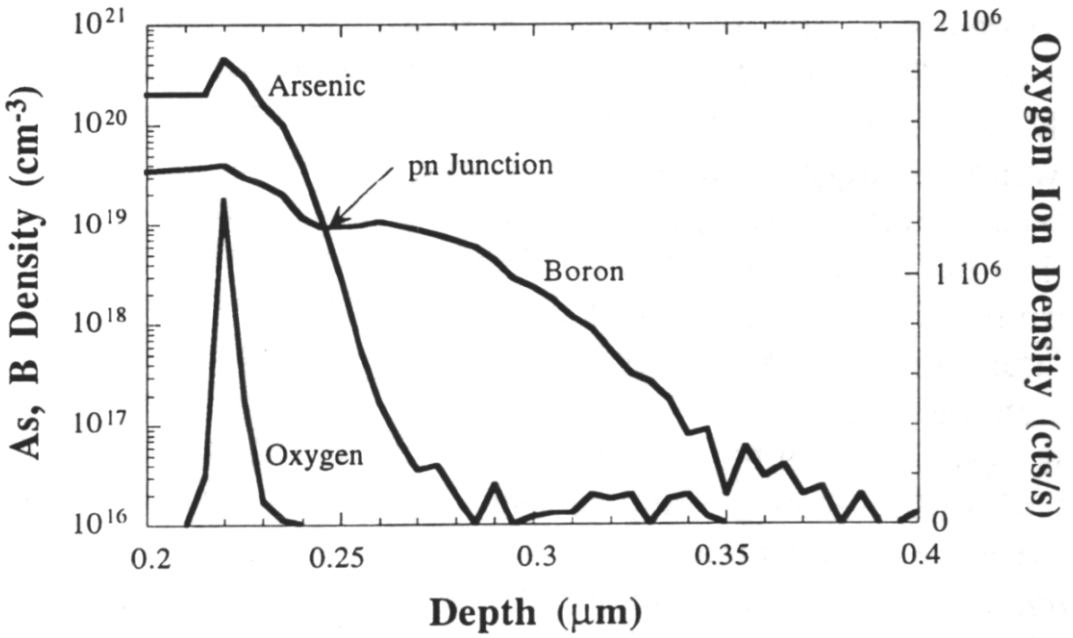


Fig 15.16 SIMS depth profile of a shallow Si *p-n* junction. Both ***A*** and ***B*** were measured with 3 keV Cs ions at 60° incidence.

Factors that need to be considered in data analysis are crater wall effects, ion knock-on, atomic mixing, diffusion, preferential sputtering, and surface roughening. Some of these are instrumental and can be alleviated to some extent, but others are intrinsic to the sputtering process. For SIMS, the most important type of atomic mixing is "cascade mixing," resulting from primary ions striking sample atoms and displacing them from their lattice positions. This causes a homogenization of all atoms within the depth affected by the collision cascade. Dopant atoms originally present at a given depth in the sample will distribute throughout this "mixing depth" as sputtering proceeds and the dopant profile will give a deeper distribution than the true distribution. It is hence important that the primary ion penetration depth be kept to a minimum for shallow dopant profiling. These deeper junctions are often observed when junctions measured by SIMS are compared to junctions measured by spreading resistance. A high vacuum is very important for SIMS. The arrival rate of gaseous species from the vacuum chamber should be less than that of the primary ion beam; otherwise it is vacuum contamination that is measured, not the sample. This is particularly important for low mass species like hydrogen.

Rutherford backscattering spectrometry (RBS) is based on backscattering of ions or projectiles incident on a sample. The technique is also known as *high-energy ion (back)-scattering spectrometry* (HEIS). It is quantitative without recourse to calibrated standards. Experiments by Rutherford and his students in the early 1900s proved the existence of nuclei and scattering from these nuclei. The field of ion interactions in solids was very intensively researched and

developed following the discovery of fission and nuclear weapons development. But it was not until the late 1950s that nuclear backscattering was put to practical use to detect a variety of elements. Further developments in the 1960s led to identification of minerals and determination of properties of thin films as well as thick samples. More recently the concepts have been adopted and implemented for routine materials characterization.

RBS is based on bombarding a sample with energetic ions—typically He ions of 1 to 3 MeV energy—and measuring the energy of the backscattered He ions. It allows determination of the *masses* of the elements in a sample, their *depth distribution* over distances from 10 nm to a few microns from the surface, their areal density, and the *crystalline structure* in a nondestructive manner. The depth resolution is on the order of 10 nm. The use of ion backscattering as a quantitative materials analysis tool depends on an accurate knowledge of the nuclear and the atomic scattering processes. Fortunately, these are generally very well known.

The method is illustrated in Fig. 15.17. Ions of mass M_1 atomic number Z_1, energy E_0, and velocity υ_0 are incident on a solid sample or target composed of atoms of mass M_2 and atomic number Z_2. Most of the incident ions come to rest within the solid, losing their energy through interactions with valence electrons. A small fraction—around 10^{-6} of the number of incident ions—undergoes elastic collisions and is backscattered from the sample at various angles. The incident ions lose energy traversing the sample until they experience a scattering event and then lose energy again as they travel back to the surface, leaving the sample with reduced energy.

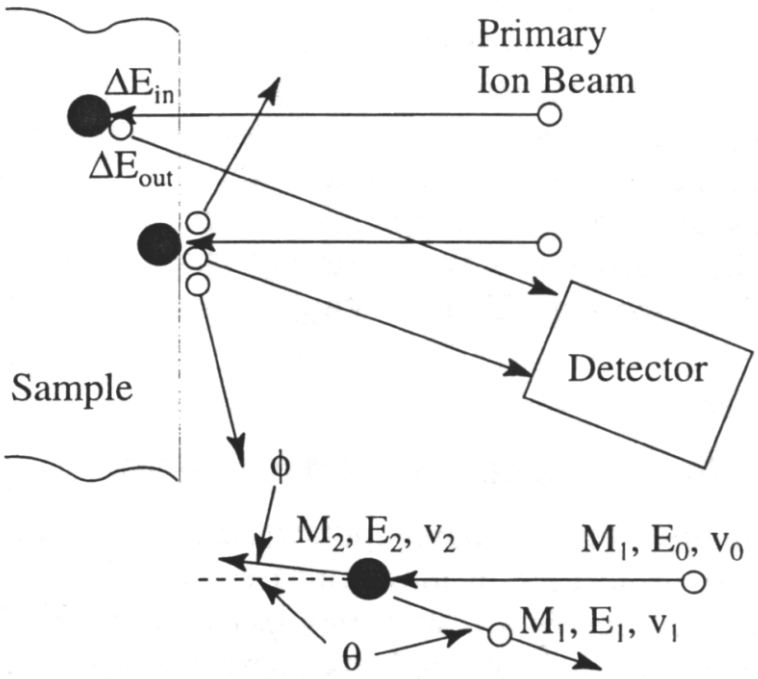


Fig. 15.17 Rutherford backscattering schematic.

After scattering, atoms M_2 have energy E_2 and velocity υ_2 and the scattered ions M_1

$$E_0 = \frac{M_1\upsilon_0^2}{2} = E_1 + E_2 = \frac{M_1\upsilon_1^2}{2} + \frac{M_2\upsilon_2^2}{2} \tag{15.13}$$

have energy E_x and velocity υ_1. Conservation of energy gives

Conservation of momentum in the directions parallel and perpendicular to the incidence direction gives

Eliminating ϕ and υ_2 and taking the ratio $E_1 / E_0 = \left(M_1\upsilon_1^2 / 2\right) / \left(M_1\upsilon_0^2 / 2\right)$, gives the *kinematic factor K:*

$$M_1\upsilon_0 = M_1\upsilon_1 Cos(\theta) + M_2\upsilon_2 Cos(\phi) \qquad 0 = M_1\upsilon_1 Sin(\theta) - M_2\upsilon_2 Sin(\phi) \tag{15.14}$$

$$K = \frac{E_1}{E_0} = \frac{\left[\sqrt{1-\left(RSin\theta\right)^2} + RCos\theta\right]^2}{\left(1+R\right)^2} \approx 1 - \frac{2R(1-Cos\theta)}{\left(1+R\right)^2} \tag{15.15}$$

where $R = M_1 / R_2$ and θ is the scattering angle. The approximation in Eq. (15.15) holds for R << 1 and θ close to 180°. Equation (15.15) is the key RBS equation. The kinematic factor is a measure of the primary ion energy loss. The scattering angle should be as large as possible. It is obviously impossible for the ions to be scattered by 180° (ions scattered back into the source), but angles of 100° to 170° are commonly used. The unknown mass M_2 is calculated from the measured energy E_1 through the kinematic factor. We will illustrate the use of RBS with the two examples in Fig. 15.18. In Fig. 15.18(a) we show a silicon substrate with a very thin film (approximately one monolayer coverage) of nitrogen, silver, and gold. The atomic weight and calculated R, K, and E_x are for θ = 170° and incident helium ions (M_1 = 4) with E_0 = 2.5 MeV. Helium ions have energies of 0.78, 1.41, 2.16 and 2.31 MeV after scattering from the N, Si, Ag, and Au atoms at the sample surface. Since N, Ag, and Au is only at the surface in this example, RBS signals from these elements have a narrow spectral distribution con firmed by experimental data. The yield is not to scale on this figure.

Figure 15.18(a) brings out two important properties of RBS plots: the RBS yield increases with element atomic number and the RBS signal of elements lighter than the substrate or matrix rides on the matrix background while elements heavier than the matrix are displayed by themselves. This makes the nitrogen signal more difficult to detect because it rides on the Si

signal. The Si background count represents the "noise," and the signal-to-noise ratio is degraded compared to heavy elements on a light matrix.

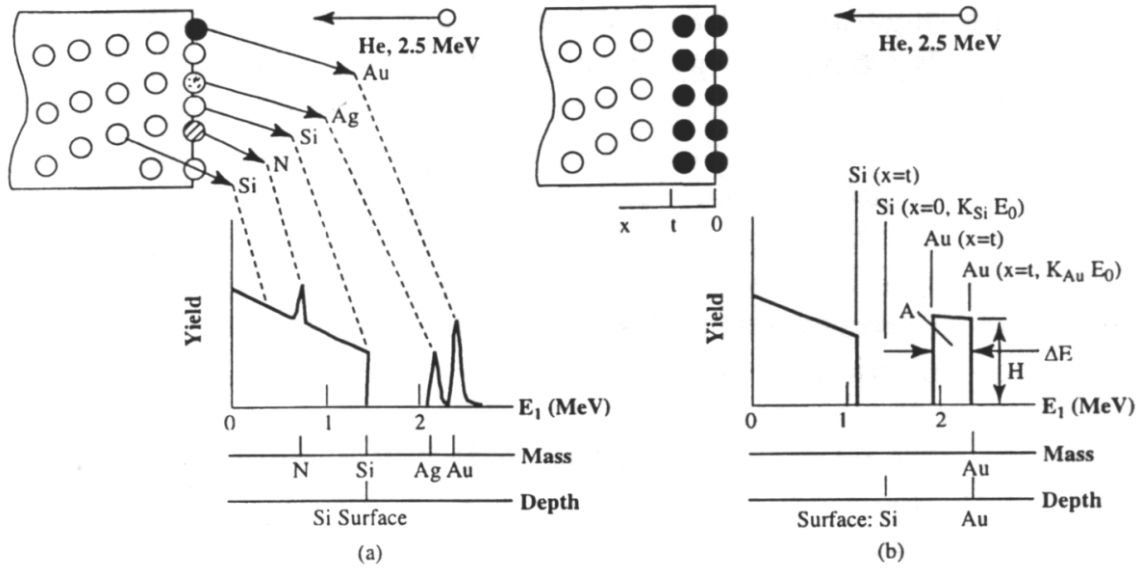


Fig. 15.18 (a) RBS calculated spectrum for N, Ag, and Au on Si, (b) schematic spectrum for a Au film on Si. *"A"is* the area under the curve.

RBS plots become slightly more complicated for layers of finite thicknesses. In Fig. 15.18(b) we consider a gold film of thickness *d on a* silicon substrate. The He ions are backscattered from surface gold atoms with $E_{1,\mathrm{Au}}$ = 2.31 MeV as in Fig. 15.18(a). However, those ions backscattered from deeper within the Au film emerge with lower energies, due to additional losses within the film. These losses come from Coulombic interactions between helium ions and electrons with He ions. Consider a scattering event from those Au atoms at the Si-Au interface at $x = d$. The He ion loses energy ΔE_{in} traveling through the Au film before the scattering event at the back gold surface. Upon scattering, it loses additional energy $(E_0 - \Delta E_{in})$ $(1 - K_{Au})$. To reach the detector it must traverse the film a second time, losing energy ΔE_{out} The total energy loss is the sum of these three losses. The energy of He ions scattered from

$$E_1(d) = (E_0 - \Delta E_{in})K_{Au} - \Delta E_{out} \qquad (15.15)$$

the sample at depth *d* is

The energy losses are slightly energy dependent and are listed in tables of stopping powers. The energy difference of the ions backscattered from the surface and from the interface ΔE can be related to the film thickness *d* by

$$\Delta E = \Delta E_{in}K_{Au} + \Delta E_{out} = [S_0]d \qquad (15.16)$$

where $[S_0]$ is often referred to as the backscattering energy loss factor; it is in units of eV/Å and is tabulated for pure-element samples, e.g., $[S_0]$ = 133.6 eV/Å for gold films with a 2 MeV beam energy.

An RBS system consists of an evacuated chamber that contains the He ion generator, the accelerator, the sample, and the detector. Negative He ions are generated in the ion accelerator at close to ground potential. In a tandem accelerator, these ions are accelerated to 1 MeV, traversing a gas-filled tube or "stripper canal," where either two or three electrons are stripped from the He" to form He^+ or He^{2+}, respectively. These ions with energies of around 1 MeV are accelerated a second time to ground potential at which point the He^+ ions have 2 MeV and the He^{2+} ions have 3 MeV energy. A magnet separates the two high-energy species.

In the sample chamber, the He ions are incident on the sample and the backscattered ions are detected by a Si surface barrier detector that operates much like the X-ray EDS detector. The energetic ions generate many electron-hole pairs in the detector, resulting in output voltage pulses from the detector. The pulse height, proportional to the incident energy, is detected by a pulse height or multichannel analyzer that stores pulses of a given magnitude in a given voltage bin or channel. The spectrum is displayed as yield or counts versus channel number with channel number proportional to the incident energy. The energy resolution of Si detectors, set by statistical fluctuations, is around 10 to 20 keV for typical RBS energies. The sample is mounted on a goniometer for precise sample-beam alignment or channeling measurements. Typical RBS runs take 15-30 minutes.

Typical semiconductor applications include measurements of thickness, thickness uniformity, stoichiometry, nature, amount, and distribution of impurities in thin films, such as silicides and Si- and Cu-doped Al. The technique is also very useful to investigate the crystallinity of a sample to determine if it is crystalline or amorphous. Backscattering is strongly affected by the alignment of atoms in a single crystal sample with the incident He ion beam. If the atoms are well aligned with the beam, those He ions falling between atoms in the channels penetrate deeply into the sample and have a low probability of being backscattered. Those He ions that encounter sample atoms "head-on" are, of course, scattered. The yield from a well-aligned single crystal sample can be two orders of magnitude less than that from a randomly aligned sample. This effect is referred to as *channeling* and has been extensively used to study ion implantation damage in semiconductors with the yield decreasing as the single crystal nature of an implanted sample is restored by annealing.

RBS is particularly suited for heavy elements on light substrates. Contacts to semiconductors generally fall into this category. Consequently, RBS has been used extensively in the study of such contacts. For example, Fig. 15.19 shows RBS spectra for platinum and

platinum silicide on silicon. Initially a Pt film is deposited on a Si substrate. The RBS spectrum clearly shows the Pt film. The Si signal is consistent with E_v taking into account the loss into and out of the Pt film. As the film is heated, PtSi forms. Note the formation from the Pt-Si interface indicated by the Pt yield decrease for that part of the film near the Si substrate.

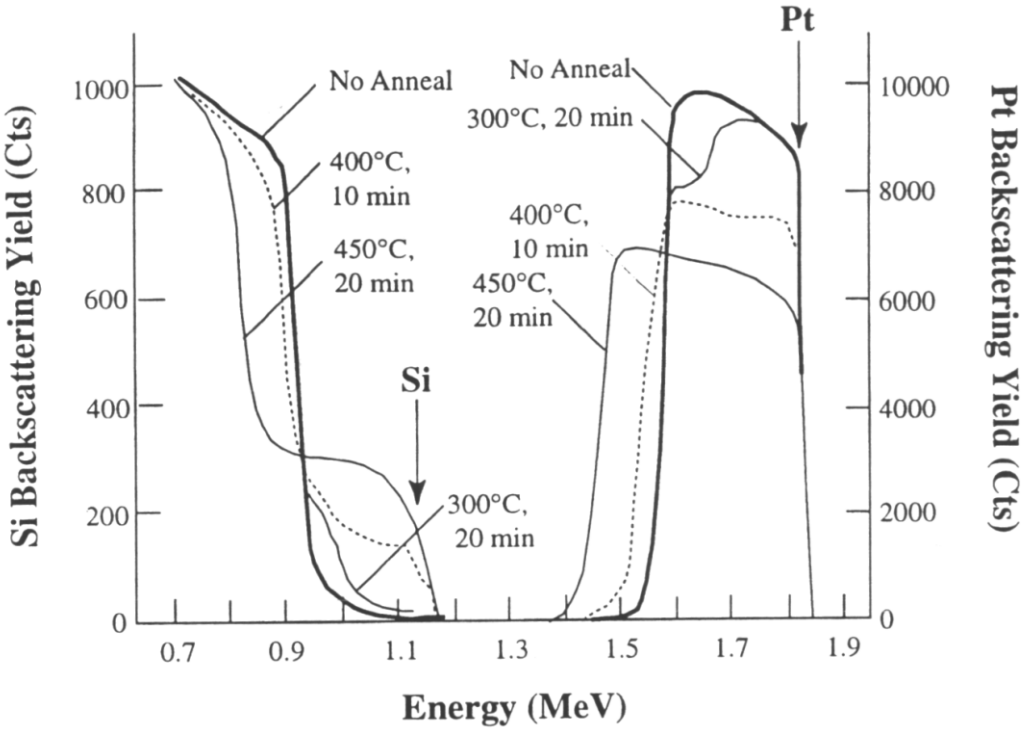


Fig. 15.19 RBS spectra for a 2000 A Pt film on Si before and after heat treatment. Platinum silicide is formed first at the interface and then throughout the film. E_0=2 MeV.

At the same time, the Si signal moves to higher energies, indicative of Si moving into the Pt film. When stoichiometry is attained, the Pt signal is uniform, but reduced and the Si signal has risen. It would have been very difficult to obtain these data with other techniques nondestructively.

RBS can provide both atomic composition and depth scales to accuracies of 5% or better. The detection limit lies in the 10^{17} to 10^{20} cm"^{3} range, but depends on the element and on energy. The sensitivity to light elements, e.g., oxygen, carbon, and nitrogen, in the presence of heavier elements is poor, because the differential scattering cross section is low for such elements. However, the cross section can be enhanced by using ion beams for which the elastic scattering is resonant. For example, the resonance at 3.08 MeV for oxygen enhances the cross section 25 times compared to its corresponding Rutherford cross section. Typical RBS depth resolutions are 10 to 20 nm for film thicknesses of less than or equal to 200 nm. The penetration depth of 2 MeV He ions is about 10 μm in silicon and 3 μm in gold. Beam diameters are

commonly around 1 to 2 mm, but microbeam backscattering with beam diameters as small as 1 μm is possible. Lateral nonuniformities over the area of the analyzing beam cannot be resolved.

A particular difficulty is the ambiguity of RBS spectra, because the horizontal axis is simultaneously a depth and a mass scale. A light mass at the surface of a sample generates a signal that may be indistinguishable from that of a heavier mass located within the sample. Through the use of tabulated constants, experimental techniques such as beam tilting, detector angle changes, and incident energy variations as well as good analytical reasoning, sample analysis is usually successful, but additional information may have to be provided to resolve ambiguities. Computer programs are extensively used in spectrum analysis. As with some other physical and chemical characterization techniques, the more is known about the sample before the analysis, the less ambiguous are the results.

15.5. Diffraction techniques.

Low energy electron diffraction (LEED), first demonstrated in 1927 by Davisson and Germer, is one of the oldest surface characterization techniques for investigating the crystallography of sample surfaces. It provides structural, not elemental information, and is illustrated in Fig. 15.20(a). A low-energy (10 to 1000 eV), narrow-energy spread electron beam incident on the sample penetrates only the first few atomic layers. Electrons are diffracted by the periodic atomic arrangement of the atoms. The elastically scattered, diffracted electrons emerge from the surface in directions satisfying interference conditions from the crystal periodicity and strike a fluorescent screen, forming a distinct array of diffraction spots due to the orientation of the crystal lattice of the sample. The diffraction pattern is viewed through a window behind the screen. The pattern can also be photographed or viewed with a TV camera. A series of grids filter the scattered electrons.

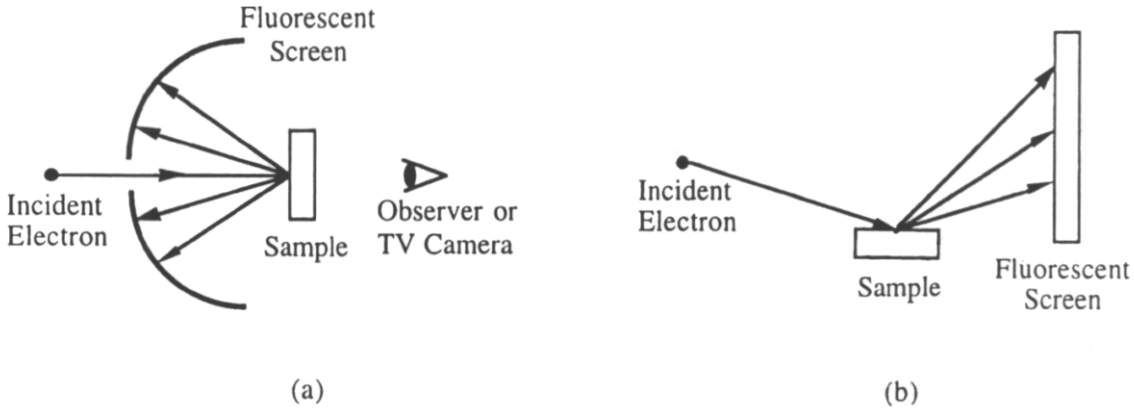


Fig. 15.20 (a) LEED diffractometer, (b) RHEED diffractometer.

LEED provides information on the atomic arrangement and is sensitive to crystallographic defects. It is typically used to determine surface atomic structure, surface structural disorder, surface morphology, and surface changes with time. The diffraction conditions can be most easily studied using a reciprocal lattice and an Ewald sphere. Sample preparation is important. To study the properties of the surface, it is important for the *surface to be clean, for contaminated surfaces generally* do not give diffract tion patterns. Consequently, LEED measurements are generally made in an ultra-high vacuum (UHV) of less than 10^{-10} torr. A monolayer of contamination takes about one second to form at a pressure of 10^{-6} torr, but takes about one day at 10^{-10} torr. Even a fraction of a monolayer is sufficient to prevent accurate surface crystallography measurements. Samples should be cleaved in vacuum, if at all possible, to expose the appropriate surfaces that have not been subjected to ambient contamination.

Electron diffraction by high energy electrons is known as *reflection high-energy electron diffraction* (RHEED). As shown in Fig. 15.20(b), 1 to 100 keV electrons are incident on the sample, but because such energetic electrons penetrate deeply, they are made to strike the sample at a shallow, glancing angle of typically less than 5°. Forward scattered electrons are utilized as there is little backscattering. RHEED gives information on surface crystal structure, surface orientation, and surface roughness. Molecular beam epitaxial growth (MBE) has done much to foster the use of RHEED by allowing continuous monitoring of the growth of epitaxial films. The experimental arrangement of Fig. 15.20(b) leaves the front of the sample clear for growth beams. Additionally, since the electron beam strikes the sample at a glancing angle, it is a more critical characterization method, since it picks out surface irregularities more effectively than LEED.

Another well-known methods based on diffraction processes used X-ray. X-Ray topography (XRT) or X-ray diffraction is a nondestructive technique for determining structural crystal defects. It requires little sample preparation and gives structural information over entire semiconductor wafers but it does not identify impurities as most of the other techniques in this chapter do. The XRT image is not magnified because no lenses are used. It is, therefore, not a high-resolution technique, but does give microscopic information through photographic enlargement of the topograph.

Consider a perfect crystal arranged to diffract monochromatic X-rays of wavelength A from lattice planes spaced d. The X-rays are incident on the sample at an angle a, as shown in Fig. 15.21(a). The primary beam is absorbed by or transmitted through the sample; only the diffracted beam is recorded on the film. The diffracted beam emerges at twice the Bragg angle θ_B defined by

$$\theta_B = Sin^{-1}\left(\frac{\lambda}{2d}\right) \qquad (15.17)$$

The diffracted X-rays are detected on a high-resolution, fine-grained photo graphic plate or film held as close as possible to the sample without intercepting the incident beam. The plate should be held perpendicular to the secondary X-rays for highest resolution. If the lattice spacing or lattice plane orientation vary locally due to structural defects, Eq. (15.17) no longer applies simultaneously to the perfect and the distorted regions. Consequently there is a difference in X-ray intensity from the two regions.

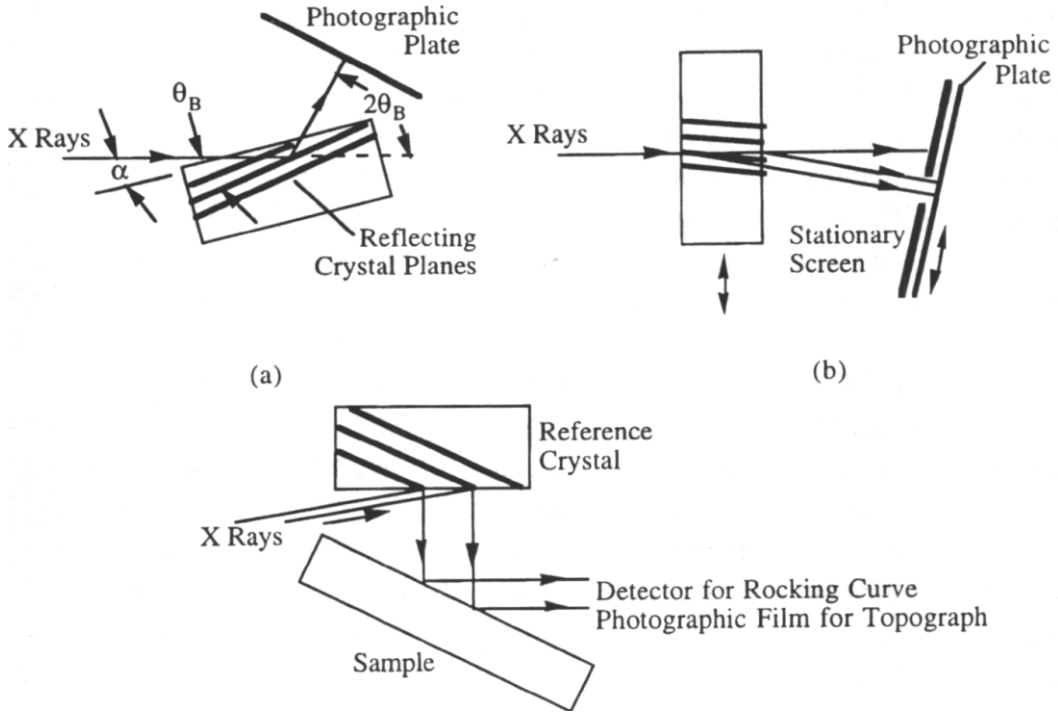


Fig. 15.21 (a) Berg-Barrett reflection topography, (b) Lang transmission topography,(c) double-crystal topography with a rocking curve.

For example, the diffracted beam from dislocations is more intense than from an area without defects caused by the mitigation of extinction and by Bragg defocusing. Dislocations produce a more heavily exposed image on the film. The image is formed as a result of diffraction from an anomaly such as strain in the crystal but does not image the defect directly. Strain ***S*** is

$$S = \frac{d_{unstrained} - d_{strained}}{d_{unstrained}} \qquad (15.18)$$

the amount of elastic deformation defined by

By determining d in unstrained and strained regions, using Eq. (15.18), one can determine S.

The reflection method illustrated in Fig. 15.21(a), known as the Berg-Barrett method, is based on the original work of Berg, modified by Barrett and further refined by Newkirk. It is the simplest X-ray topography method. There are neither lenses nor moving parts except for the sample alignment goniometer. Reflection XRT probes a thin sample region near the surface,

since the shallow incident angle a confines X-ray penetration to the near-surface region. This method is used to determine dislocations, for example, and is useful for dislocation densities up to about 10^6 cm^{-2}. The resolution is about 10^{-4} cm, and areas as large as 200 mm diameter wafers can be examined with the Berg-Barrett method.

Transmission XRT, illustrated in Fig. 15.21(b), introduced by Lang, is by far the most popular XRT technique. Monochromatic X-rays pass through a narrow slit and strike the sample aligned to an appropriate Bragg angle. The tall and narrow primary beam is transmitted through the sample and strikes a lead screen. The diffracted beam falls on the photographic plate through a slit in the screen. X-rays are absorbed in a solid. However, absorption is considerably reduced when the X-rays are aligned for diffraction along certain crystal planes. A topography is generated by scanning the sample and the film in synchronism holding the screen stationary. Scanning combined with oscillation is effective when extreme sample warpage prevents large area imaging. While the crystal is scanned, both crystal and film are also oscillating simultaneously around the normal to the plane containing the incident and reflected beam. Entire large-area wafers can be imaged. Large-diameter wafers become warped during processing, making it necessary to adjust the specimen continuously during topography measurements to ensure that it stays on the chosen Bragg angle.

To "photograph" defects, one usually chooses a weakly diffracting plane. A uniform sample gives a featureless image. Structural defects cause stronger X-ray diffraction, thus providing film contrast or topographic features. The Lang technique has also been adapted to reflection topography. Scanning provides for considerably more flexibility than is possible with the Berg-Barrett technique. For semiconductors, the Lang method is used primarily to study defects introduced during crystal growth or during wafer processing. Transmission topographs provide information on defects through the entire sample; reflection topographs provide information of 10 to 30 μm depth from the surface. X-ray topographs of silicon wafers are shown in Fig. 15.22.

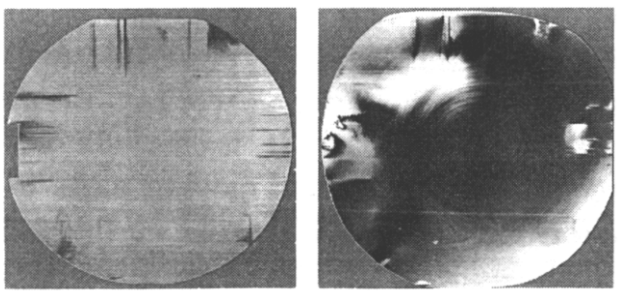


Fig. 15.22 (a) X-ray topographs of a 7 μm, (100)-oriented epitaxial silicon wafer using the Lang and double crystal topography methods, (b) crystal defects by the Lang transmission method.

In section topography the sample and film are stationary and a narrow "section" of the sample—the cross section—is imaged. The stationary sample is illuminated by a narrow X-ray beam and the sample cross section is imaged on the film. The method is like that in Fig. 15.21(b), except both sample and photographic plate are stationary. Section topography has proven to be very valuable for defect depth information. For example, it is common in integrated fabrication to precipitate oxygen in silicon wafers. Section topography is a convenient method to obtain a nondestructive cross-sectional picture through the wafer, clearly showing the precipitated regions. An example of such a section topograph is shown in Fig. 15.23.

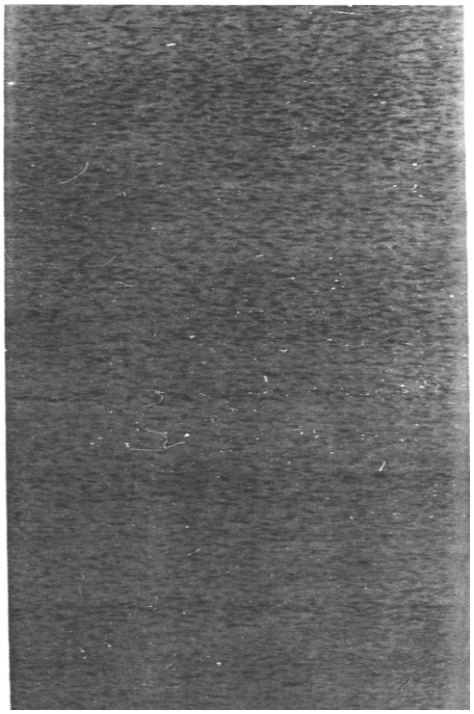


Fig 15.23 X-ray section topograph of an n-type (100) Si wafer. Wafer is 675 μm thick.

Double-crystal diffraction provides higher accuracy because the beam is more highly collimated than is possible with single crystal topography. Thetechnique, shown in Fig. 15.21(c), consists of two successive Bragg reflections from reference and sample crystals. Reflection from the first, carefully selected "perfect" crystal produces a monochromatic and highly parallel beam to probe the sample. The double crystal technique is used not only for topography, but also for rocking curve determination. To record a rocking curve, the sample is slowly rotated or "rocked" about an axis normal to the diffraction plane and the scattered intensity is recorded as a function of the angle as shown in Fig. 15.21(c).

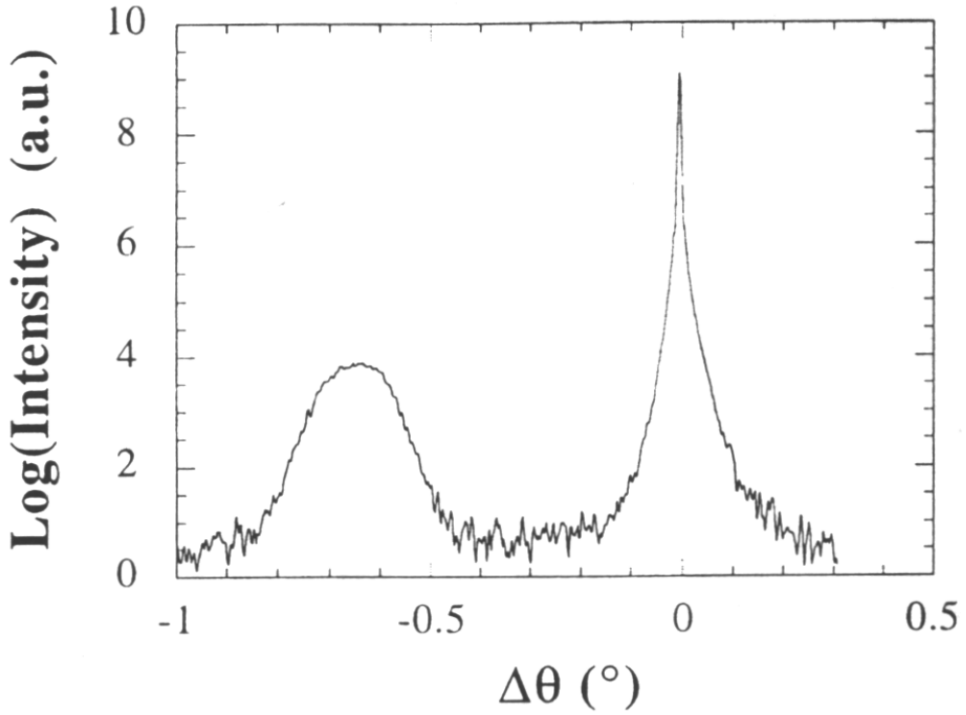


Fig. 15.24 Rocking curve of a heteroepitaxial $Si_{080}Ge_{020}$ film (150 nm) on (100) Si. The film is diffracting at a smaller angle than the substrate. From Bragg's law, this implies that the film has a large d-spacing and thus has a larger lattice parameter than the substrate.

Such a rocking curve is shown in Fig. 15.24. The rocking curve width is a measure of crystal perfection. The narrower the curve, the more perfect is the material. For epitaxial layers it provides data on lattice mismatch, layer thickness, layer and substrate perfection, and wafer curvature. Double crystal diffraction has been extended to four-crystal diffraction where four crystals are used to collimate the X-ray beam further.

15.6. Spectroscopy techniques.

Most well-known and widespread spectroscopy method is photoluminescence (PL) also known as *fluorometry*. This method can be also contributed as the one of the oldest. Photoluminescence provides a nondestructive technique for the determination of certain impurities in semiconductors. It is particularly suited for the detection of shallow-level impurities, but can be applied to certain deep-level impurities, provided their recombination is radiative. Photoluminescence is also used in other applications. For example, ultraviolet light in fluorescent tubes, generated by an electric discharge, is absorbed by a phosphor inside the tube and visible light is emitted by photoluminescence. We discuss PL only briefly by giving the main concepts and a few examples. *Identification* of impurities is easy with PL, but measurement of the impurity *density* is more difficult. PL can provide simultaneous information on many types of impurities in a sample. However, only those impurities that produce radiative

recombination processes can be detected. Fortunately, many impurities fall within this category. A typical PL set-up is illustrated in Fig. 15.25. The sample is placed in a cryostat and cooled to temperatures near liquid helium. Low temperature measurements are necessary to obtain the fullest spectroscopic information by minimizing thermally activated nonradiative recombination processes and thermal line broadening. The thermal distribution of carriers excited into a band contributes a width of approximately $kT/2$ to an emission line originating from that band. This makes it necessary to cool the sample to reduce the width. The thermal energy $kT/2$ is only 1.8 meV at T = 4.2 K. For many measurements this is sufficiently low, but occasionally it is necessary to reduce this broadening further by reducing the sample temperature below 4.2 K. It is possible, however, to make room temperature PL measurements, albeit with decreased sensitivity. Using a scanned photon beam or moving the sample allows PL maps to be generated.

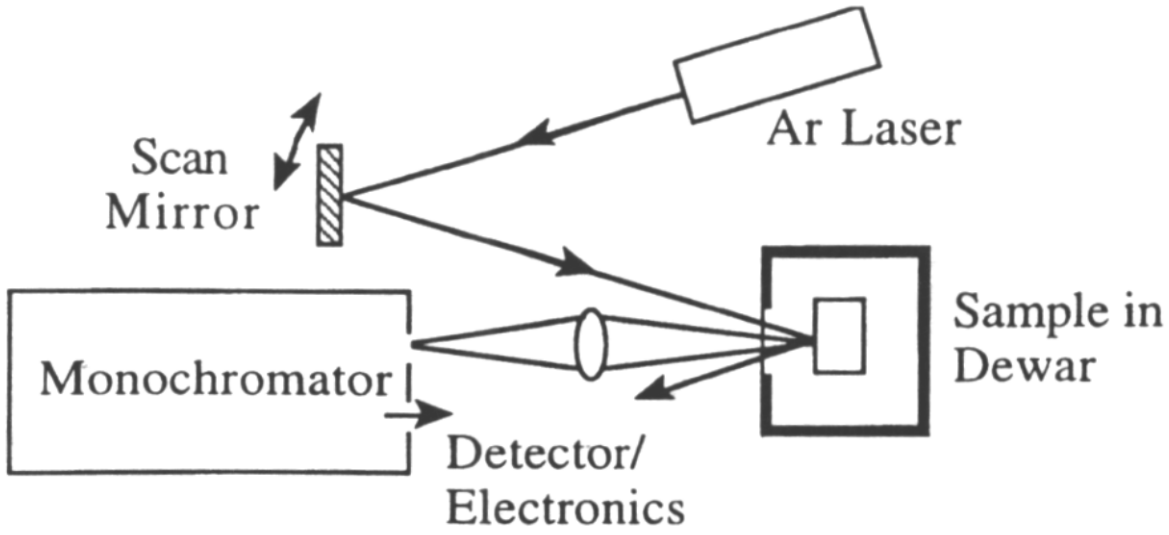


Fig 15.25 Schematic photoluminescence arrangement.

The sample is excited with an optical source, typically a laser with energy $h\nu > E_g$, generating electron-hole pairs (ehps) which recombine by one of several mechanisms. Photons are emitted for *radiative* recombination. Photons are not emitted for *nonradiative* recombination in the bulk or at the surface. For good PL output, the majority of the recombination processes should be radiative. Some of the photons may be reabsorbed in the sample,provided the photons are directed at the surface within the critical angle. Some of the photons directed to the back surface may be reflected and emitted. Accounting for these processes allows for PL intensity Φ_{PL} to be written as

$$\Phi_{PL} = \frac{\Phi(1-R)Cos(\theta)}{\pi n(n+1)^2}\frac{L}{\tau_{rad}s_r} \tag{15.19}$$

where Φ_{PL} is the incident photon flux density, R the reflectivity, θ the emission angle, n the index of refraction, L the minority carrier diffusion length, τ_{rad} the radiative lifetime, and s_r the surface recombination velocity. Equation (15.19) shows the PL signal to be very sensitive to surface recombination.

The photon energy depends on the recombination process, illustrated in Fig. 15.26, where five of the most commonly observed PL transitions are shown. Band-to-band recombination [Fig 15.26(a)] dominates at room temperature but is rarely observed at low temperatures in materials with small effective masses due to the large electron orbital radii.

Excitonic recombination is commonly observed, but what are excitons? When a photon generates an ehp, Coulombic attraction can lead to the formation of an excited state in which an electron and a hole remain bound to each other in a hydrogen-like state. This excited state is referred to as a free *exciton* (FE). Its energy, shown in Fig. 15.26(b), is slightly less than the band-gap energy required to create a *separated* ehp. An exciton can move through the crystal, but because it is a *bound* ehp, both electron and hole move together and neither photoconductivity nor current results.

A free hole can combine with a neutral donor [Fig. 15.26(c)] to form a positively charged excitonic ion or *bound exciton* (BE). The electron bound to the donor travels in a wide orbit about the donor. Similarly electrons combining with neutral acceptors also form boundexcitons.

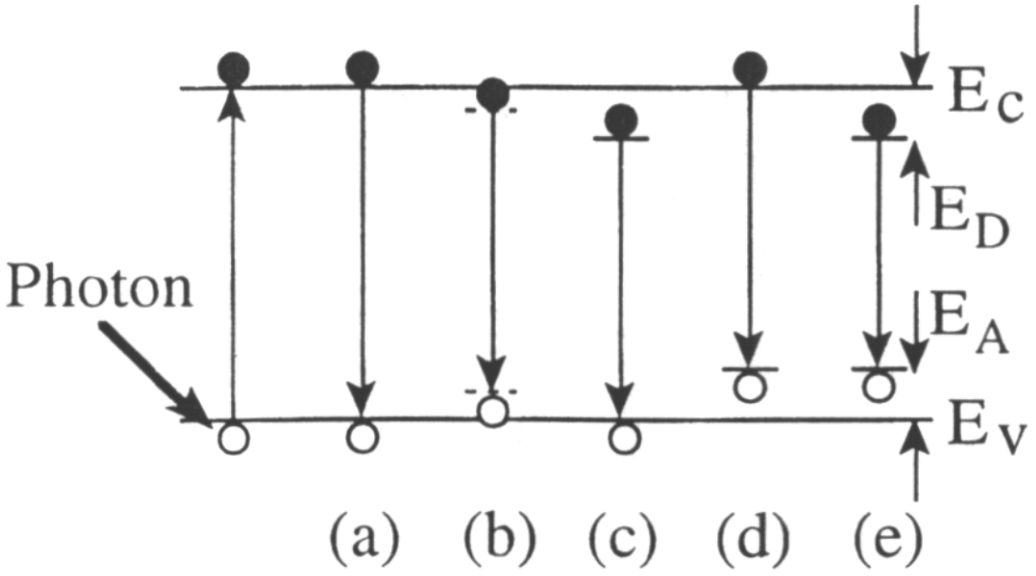


Fig. 15.26 Radiative transitions observed with photoluminescence.

If the material is sufficiently pure, free excitons form and recombine by emitting photons. The photon energy in direct band-gap semiconductors of band gap energy E_g is

$$h\nu = E_g - E_x \tag{15.20}$$

where E_x is the excitonic binding energy. In indirect band-gap semiconductors, momentum conservation requires the emission of a phonon, giving

$$h\nu = E_g - E_x - E_p \tag{15.21}$$

with E_p *is the phonon energy. Bound exiton recombination dominates over* free exciton recombination for less pure material. A free electron can also recombine with a hole on a neutral acceptor (see Fig. 15.26(d)), and similarly a free hole can recombine with an electron on a neutral donor.

Lastly, an electron on a neutral donor can recombine with a hole on a neutral acceptor, the well-known donor-acceptor (D-A) recombination, illustrated in Fig. 15.26(e). The emission line has an energy modified by the Coulombic interaction between donors and acceptors

$$h\nu = E_g - (E_A + E_D) + \frac{q^2}{K_s \varepsilon_0 r} \tag{15.22}$$

where r is the distance between donor and acceptor. The photon energy in Eq. (15.22) can be higher than the band gap for low $(E_A + E_D)$. Such photons are generally reabsorbed in the sample. The full widths at half maximum (FWHM) for bound exciton transitions are typically < $kT/2$ and resemble slightly broadened delta functions. This distinguishes them from donor-valence band transitions which are usually a few kT wide. Energies for these two transitions are frequently similar and the line widths are used to determine the transition type. A PL spectrum showing BE-D, BE-A, and D-A recombination in InP is shown in Fig. 15.27.

The optics in a PL apparatus are designed to ensure maximum light collection. The PL-emitted light from the sample can be analyzed by a grating monochromator and detected by a photodetector. Replacing the monochromator with a Michelson interferometer leads to enhanced sensitivity and reduced measurement time. PL radiation from shallow-level impurities in Si and GaAs can be detected with a photomultiplier tube with an S-1 photocathode able to detect wavelengths from about 0.4 to 1.1 μm. Lower-energy light from deeper levels requires a PbS (1-3 μm) or doped germanium detector.

The volume analyzed in PL measurements is determined by the absorption depth of the exciting laser light and the diffusion length of the minority carriers. Usually the absorption depth is on the order of microns or so. It is possible, however, to confine the absorbed light to a very thin layer near the surface by using ultraviolet light.

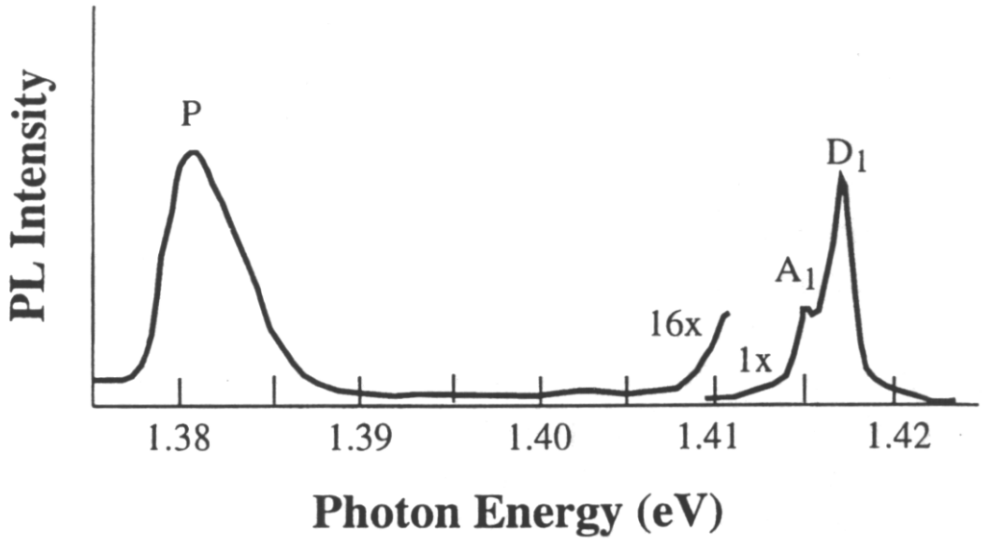


Fig. 15.27 Low temperature PL spectrum of InP. D_1 and A_1 are broadened donor and acceptor bound exciton lines, and P is a D-A line due to carbon acceptors.

This is useful in such materials as siliconon-insulator, in which the active Si layer is only about 0.1 μm thick. It is generally difficult to correlate the intensity of a given PL spectral line with the density of the impurity giving rise to that line.

This is due to nonradiative bulk and surface recombination that can vary significantly from sample to sample and from location to location on a given sample. A novel approach to this problem is due to Tajima. For Si samples of different resistivity, he found spectra with both intrinsic and extrinsic peaks as shown in Fig. 15.28. Higher resistivity samples showed higher intrinsic peaks. The ratio $X_{TO}(BE)/I_{TO}(FE)$ is proportional to the doping density, where $X_{TO}(BE)$ is the transverse optical phonon PL intensity peak of the bound exciton for element X (boron or phosphorus) and $I_{TO}(FE)$ is the transverse optical phonon intrinsic PL intensity peak of the free exciton.

Calibration curves of photoluminescence intensity ratio versus impurity density for Si are shown in Fig. 15.26. Good agreement is found between the resistivity measured electrically and the resistivity calculated from the carrier density measured by photoluminescence. Very pure float-zone Si was used and varying amounts of phosphorus were introduced using neutron transmutation doping to generate calibration curves for the PL data.[101] It is estimated that for samples with areas of 0.3 cm^2 and 300 μm thickness, the detection limits for P, B. Al and As in

Si are around 5 X 10^{10}, 10^{11}, 2 X 10^{11}, and 5 X 10^{11} cm^{-3}, respectively. Various impurities in Si have been catalogued. The interpretation has also been applied to InP, where the donor density as well as the compensation ratio was determined.

The ionization energies of donors in GaAs are typically around 6 meV and the energy difference between the various donor impurities is too small to be observable by conventional PL.

However, acceptors with their wider spread of ionization energies can be detected by using the transitions: free electron to neutral acceptor (Fig. 15.26(d)) and electron on a neutral donor to hole on a neutral acceptor (Fig. 15.26(e)). Acceptors in GaAs determined with PL have also been catalogued.

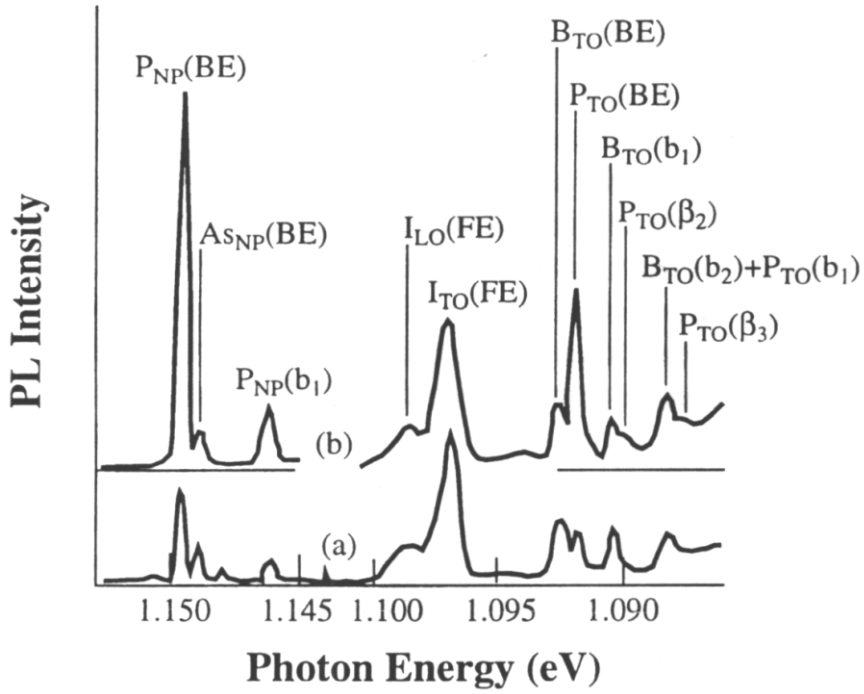


Fig. 15.28 Photoluminescence spectra for Si at $\boldsymbol{T}$ = 4.2 K. (a) Starting material, (b) after neutron transmutation doping. Base lines for measuring the peak heights are shown by the horizontal lines. Symbols: I = intrinsic, TO = transverse optical phonon, LO = longitudinal optical phonon, BE = bound exciton, FE = free exciton. The sample contains residual arsenic. Components labeled b_n and β_n are due to recombination of multiple bound excitons.

Complications arise when the energy difference between the ground states of two or more acceptors is identical to the difference between their band-acceptor and donor-acceptor pair transitions. When this occurs, transitions can often be differentiated through variable temperature measurements or through variable excitation power measuremwnt that cause a shift of the donor-acceptor pair transition to higher energies. Donors in GaAs can be detected by *magneto-*

photoluminescence measurements. The magnetic field splits some of the spectral lines into several components due to magnetic field splitting of the bound exciton initial states. Quantitative correlation of PL data to impurity densities can, of course, also be made by Hall measurements.

Raman spectroscopy is a vibrational spectroscopic technique that can detect both organic and inorganic species and measure the crystallinity of solids. It is free from charging effects that can influence electron and ion beam techniques. We mention it here because it is finding increased use in semiconductor characterization. For example, it is sensitive to strain, allowing it to be used to detect stress in a semiconductor material or device. Since the light beam can be focused to a small diameter, one can measure stress in localized regions.

When light is scattered from the surface of a sample, the scattered light is found to contain mainly wavelengths that were incident on the sample (Raleigh scattering) but also at different wavelengths at very low intensities (few parts per million or less) that represent an interaction of the incident light with the material. The interaction of the incident light with optical phonons is called Raman scattering while the interaction with acoustic phonons results in Brillouin scattering. Optical phonons have higher energies than acoustic phonons giving larger photon energy shifts, illustrated in Fig. 15.29. Hence Raman scattering is easier to detect than Brillouin scattering. However, even for Raman scattering, the energy shift is small. For example, the optical phonon energy in Si is about 0.067 eV, while the exciting photon energy is several eV (Ar laser light with λ=488 nm has an energy of $h\nu$=2.54 eV). Since the intensity of Raman scattered light is very weak (about 1 in 10^8 parts), Raman spectroscopy is only practical when an intense monochromatic light source like a laser is used.

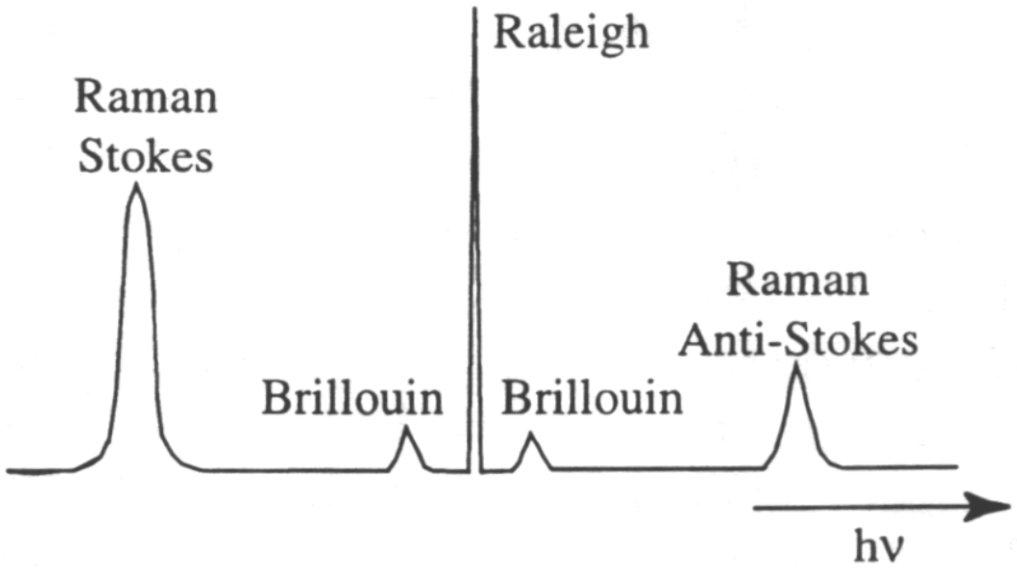


Fig. 15.29 Energy distribution of scattered light.

Raman spectroscopy is based on the Raman effect, first reported by Raman in 1928. If the incident photon imparts part of its energy to the lattice in the form of a phonon (phonon emission) it emerges as a lower-energy photon. This down-converted frequency shift is known as Stokes-shifted scattering. Anti-Stokes-shifted scattering results when the photon absorbs a phonon and emerges with higher energy. The anti-Stokes mode is much weaker than the Stokes mode and it is Stokes-mode scattering that is usually monitored.

During Raman spectroscopy measurements a laser beam, referred to as the pump, is incident on the sample. The weak scattered light or signal is passed through a double monochromator to reject the Raleigh scattered light and the Raman-shifted wavelengths are detected by a photodetector. In the Raman microprobe, a laser illuminates the sample through a commercial microscope. Laser power is usually held below 5 mW to reduce sample heating and specimen decomposition.[112] In order to separate the signal from the pump it is necessary that the pump be a bright, monochromatic source. Detection is made difficult by the weak signal against an intense background of scattered pump radiation. The signal-to-noise ratio is enhanced if the Raman radiation can be observed at right angles to the pump beam. A major limitation in Raman spectroscopy is the interference caused by fluorescence, either of impurities or the sample itself. The fluorescent background problem is eliminated by combining Raman spectroscopy with FTIR, dramatically demonstrated with the spectra in Fig. 15.30.

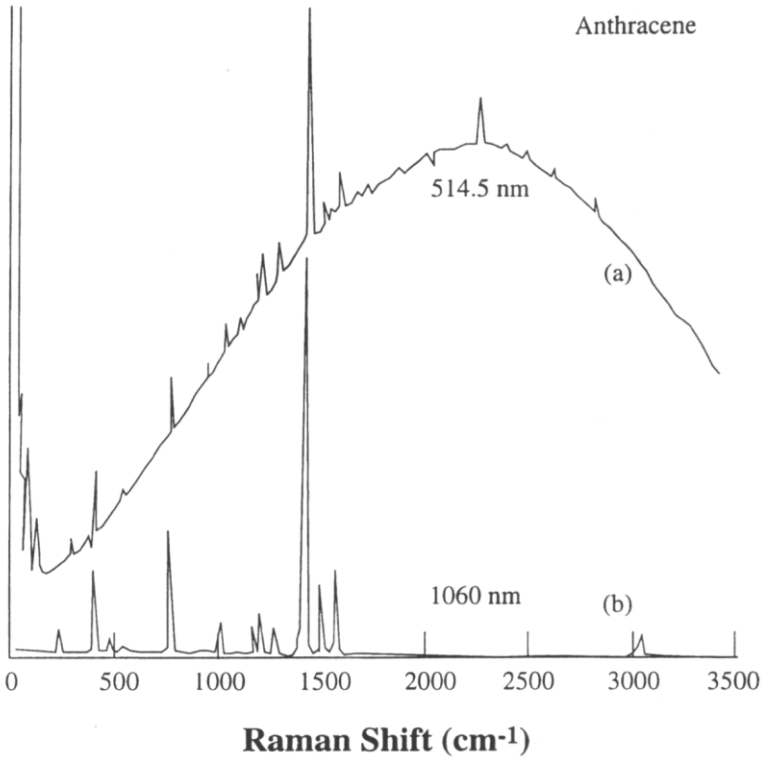


Fig. 15.30 (a) Conventional Raman spectrum and (b) FTIR Raman spectrum of anthracene.

By using lasers with varying wavelengths and hence different absorption depths, it is possible to profile the sample to some depth. The technique is nondestructive and requires no contacts to the sample. Most semiconductors can be characterized by Raman spectroscopy. The wavelengths of the scattered light are analyzed and matched to known wavelengths for identification.

Various properties of the sample can be characterized. Its composition can be determined. Raman spectroscopy is also sensitive to crystal structure. For example, different crystal orientations give slightly different Raman shifts, e.g., scattering by transverse optical (TO) phonons is forbidden in(100)-oriented GaAs. However, damage and structural imperfections induce scattering by the forbidden TO phonons, allowing implant damage to be monitored, for example. The Stokes line shifts, broadens and becomes asymmetric for microcrystalline Si with grain sizes below 100Å. The lines become very broad for amorphous semiconductors, allowing a distinction to be made between single crystal, polycrystalline, and amorphous materials

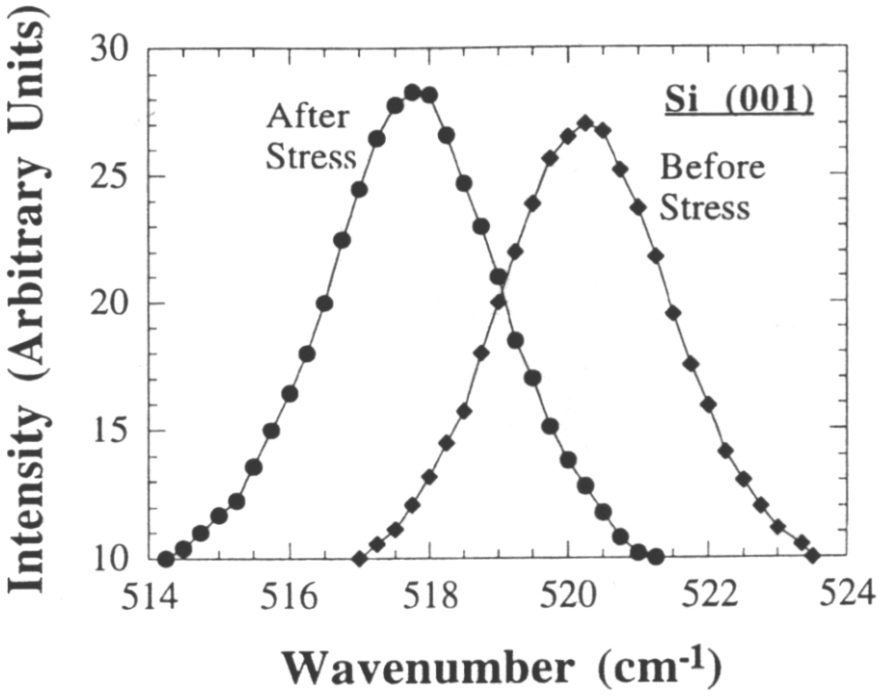


Fig. 15.31 Raman spectra for (001) Si unstressed and under biaxial stress.

The frequency is also shifted by stress and strain in thin film. An example is shown in Fig. 15.31, where the Raman spectra are shown for stressed and unstressed (001) Si. Note the Si signal at $1/\lambda \sim 520$ cm^{-1}, corresponding to an energy of 0.067 eV. A summary of semiconductor applications including structural defects, ion damage, laser annealing, alloy fluctuations, interfaces, and heterojunctions is given by Pollack and Tsu. The Raman microprobe is able to

identify organic contaminants that appear as particles as small as 2 μm or as films as thin as 1 μm. The technique is most successful for organic materials because organic spectral data bases exist. For example, silicone films, teflon, cellulose, and other contaminants have been detected. Raman spectroscopy is very effective, when coupled with other characterization techniques, for problem solving in semiconductor processing.

15.7. Scanning probe techniques.

Scanning probe microscopy (SPM) refers to all techniques using a mechanism to scan a sharp tip across a sample surface at very small distances to obtain two- or three-dimensional images of the surface at nanometer or less resolution both laterally and vertically. In the extreme, one can btain lateral resolution on the order of 0.1 nm and vertical resolution of 0.01 nm. The original application of SPM was the scanning tunneling microscope (STM), invented in the early 1980s[116] based on the earlier topografiner. Still today it is the only technique for imaging at atomic resolution other than TEM. A myriad of SPM instruments has been developed over the past decade, and one can sense current, voltage, resistance, force, temperature, magnetic field, work function, and so on with these instruments at high resolution as outlined in Table 15.1.

Table 15.1

AFM	Atomic force microscopy
BEEM	Ballistic electron emission microscopy
CFM	Chemical force microscopy
IFM	Interfacial force microscopy
MFM	Magnetic force microscopy
MRFM	Magnetic resonance force microscopy
MSMS	Micromagnetic scanning microprobe system
Nano-Field	Nanometer electric field gradient
Nano-NMR	Nanometer nuclear magnetic resonance
Nano-SRP	Nanometer spreading resistance profiling
NSOM	Near-field optical microscopy
SCM	Scanning capacitance microscopy
SCPM	Scanning chemical potential microscopy
SEcM	Scanning electrochemical microscopy
SICM	Scanning ion-conductance microscopy
SKPM	Scanning Kelvin probe microscopy
SThM	Scanning thermal microscopy
STOS	Scanning tunneling optical spectroscopy
STM	Scanning tunneling microscopy

We describe only scanning tunneling and atomic force microscopy here. For a more detailed description of these and other probe techniques, the reader is referred to the extensive literature that has been published over the past 25 years.

The schematic in Fig. 15.32 shows the major features of a scanning tunneling microscope (STM).

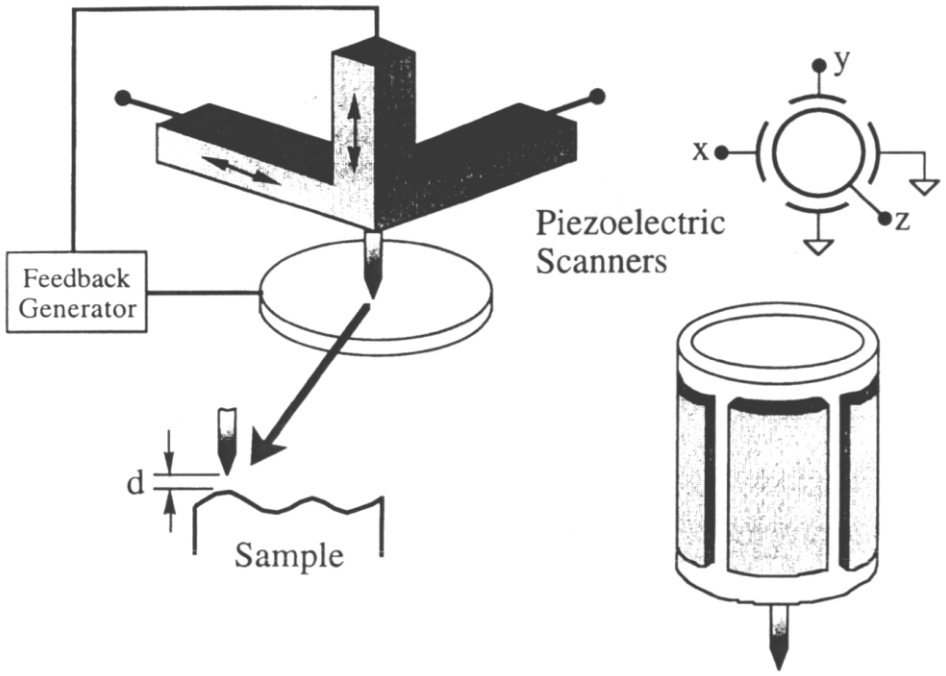


Fig. 15.32 Schematic illustration of a scanning tunneling microscope.

It consists of a metallic probe tip scanned across the sample at distances of about 1 nm, with a bias voltage between the tip and the sample that is less than the work function of the tip or the sample. The probe is usually made from tungsten or Pt-Ir. Piezoelectric elements provide the scanning mechanism. A piezoelectric material is one that changes dimension upon application of a voltage. By applying voltages to x, y, and z-oriented piezoelectric elements, the tip or the sample can be scanned in all three directions. Early implementations used the three-arm arrangement in Fig. 15.32. However, this arrangement is subject to low resonance frequencies and was later changed to the tubular implementation. Since the probe tip is very close to the sample surface, a tunnel current of typically 1 nA flows between the two across the gap. Clearly, both probe and sample must be conducting for this technique to work. For high resolution images it is very important that the tip be extremely sharp. In fact it is believed that a single atom at the probe tip primarily determines the device operation. The current is given by

$$J = \frac{C_1 V}{d} \exp\left(-2d \sqrt{\frac{8\pi^2 m \Phi_B}{h^2}} \right) = \frac{C_1 V}{d} \exp\left(-1.025 d \sqrt{\Phi_B} \right) \qquad (15.23)$$

for d in Å and Φ_B in eV, where C_1 is a constant, V the voltage, d the gap spacing between tip and sample, and Φ_B an effective work function defined by $\Phi_B = \left(\Phi_{B1} + \Phi_{B2}\right)/2$ with Φ_{B1} and Φ_{B2} the work functions of the tip and sample, respectively. For $\Phi_B \approx 4$ eV, a typical work function, a gap spacing change from 10 Å to 11 Å, for example, changes the current density by about a factor of eight. Hence, small changes in gap spacing produce large current changes. This suggests application in surface flatness characterization.

There are two modes of operation. In the first, the gap spacing is held constant, as the probe is scanned in the x and y dimensions, through a feedback circuit holding the current constant. The voltage on the piezoelectric transducer is then proportional to the vertical displacement giving a contour plot. This is the dominant mode. In the second mode, the probe is scanned across the sample with the gap, and hence the current, varying. The current is now used to determine the wafer flatness. Equation (15.22) is somewhat simplified. The tunnel current is actually a measure of the overlap of the electronic wave functions of probe and sample in the gap separating the two and the probe actually images surface wave functions rather than just atomic positions. However, the current is largely determined by the gap spacing or sample topography. Holding the probe above a given location of the sample and varying the probe voltage gives the tunneling spectroscopy current, allowing the band gap and the density of states to be probed. By using the STM in its spectroscopic mode, the instrument probes the electronic states of a surface located within a few electron volts on either side of the Fermi energy. The sensitivity of STM to electronic structure can lead to undesirable artifacts. For example, a region of lower conductivity appears asa dip in the image.

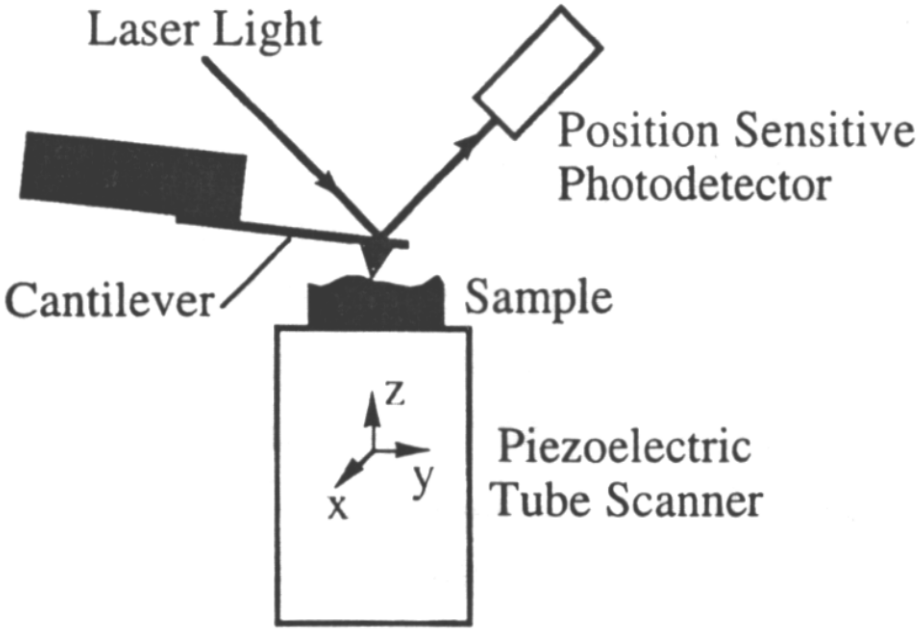


Fig. 15.33 Schematic illustration of anatomic force microscope.

Atomic force microscopy (APM) operates by measuring the forces between a probe and the sample. These forces depend on the nature of the sample, the distance between the probe and the sample, the probe geometry, and sample surface contamination. In contrast to scanning tunneling microscopy, which requires electrically conducting samples, AFM is suitable for conducting as well as insulating samples. The AFM principle is illustrated in Fig. 15.33. The instrument consists of a cantilever with a sharp tip mounted on its end. The cantilever is typically formed from silicon, silicon oxide, or silicon nitride and is about 100-200 μm long and 0.5 to 5 μm thick. Silicon nitride cantilevers and tips are formed by depositing Si_3N_4 on a Si surface containing a pyramidal etch pit. The vertical sensitivity depends on the cantilever length. For topographic imaging, the tip is brought into continuous or intermittent contact with the sample and scanned across the sample surface. Depending on the design, piezoelectric scanners translate either the sample under the cantilever or the cantilever over the sample. The motion of the cantilever is sensed by a segmented, position sensitive photodetector. Holding the signal constant, equivalent to constant cantilever deflection, by varying the sample height through a feedback arrangement, gives the sample height variation.

An AFM can operate in several modes. In the contact mode, the sample topography is measured by scanning the tip, which contacts the surface, across the sample. The contact mode is frequently used in AFM measurements. However, samples are typically covered with a thin layer of water or other contaminants. When the probe touches the surface, it is pulled toward the sample by capillary action. This force, coupled with possible electrostatic forces, creates a

substantial frictional force as the tip is scanned across the surface, leading to possible sample damage. In the noncontact mode, the instrument senses van der Waal attractive forces between surface and probe tip held above the sample surface. This mode has lower resolution than the contact mode. In a third mode, the cantilever is excited close to its resonant frequency by an external signal applied to a piezoelectric ceramic to which the cantilever is attached. Resonance frequencies depend on the cantilever mechanical properties and vary typically from 15 kHz to 500 kHz.

As the tip approaches the sample surface, the attractive force increases, leading to a resonance frequency decrease. When the tip is brought closer to the surface, there is a relatively abrupt change to a strong repulsive force as the tip touches the surface. The repulsive force acts for a short time, and the tip strikes briefly against the sample before bouncing off and continues to vibrate through the spring restoring forces. In the low amplitude mode (swing amplitude less than 5 nm), often referred to as the true contactless mode, the swing amplitude is so small that the probe is embedded in the attractive force field between the probe and the sample surface. The local gradient of the attractive forces shifts the resonance frequency of the cantilever, leading to a change in frequency, phase, and amplitude of the oscillation. It is these shifts that are used for sensing the tip-sample distance variation. This leads to very small disturbance of the sample by the probe but also to a limited resolution. At high amplitudes (swing amplitude typically 5-50 nm), the probe moves in and out of the force fields. Most commonly, one uses the amplitude damping that results from moving the probe close to the surface, in the feedback loop. This mode is the tapping mode. By keeping the amplitude constant, one also keeps the tip-sample distance constant. The probe exerts negligible frictional force on the sample, since the contact is intermittent, and surface damage is minimal. Examples of AFM images are shown in Fig. 15.34 .

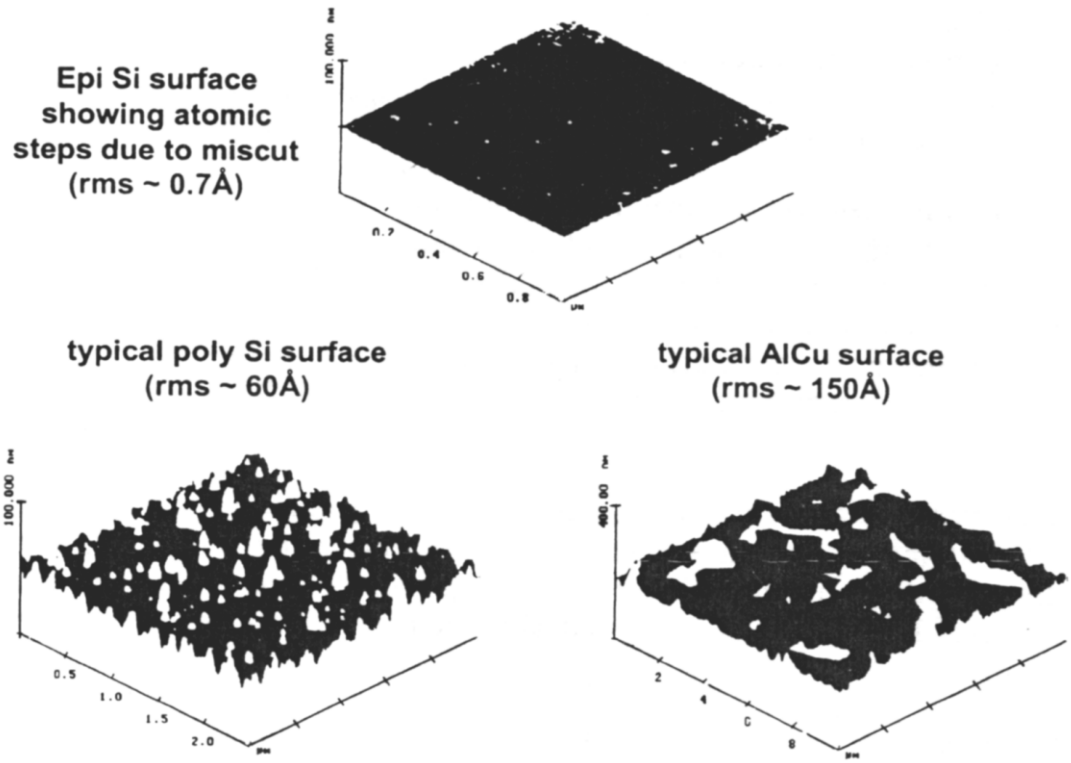


Fig. 15.34 AFM images. Epitaxial silicon surface: 1μm x 1 μm area, vertical scale: 100 nm/div, polycrystalline silicon surface: 2.5 /xm X 2.5 дт area, vertical scale:100 nm/div, AlCu surface: 10 μm x 10 μm area, vertical scale: 400 nm/div.

Tapping mode imaging works well for soft, adhesive, or fragile samples, allowing high resolution topographic imaging of sample surfaces that are easily damaged, loosely held to their substrate, or otherwise difficult to image by other AFM techniques. Specifically, tapping mode overcomes problems associated with friction, adhesion, electrostatic forces, and other difficulties that can plague conventional AFM scanning methods. When the probe passes over a bump in the surface, the cantilever has less room to oscillate and the oscillation amplitude decreases. Conversely, when the probe passes over a depression, the cantilever has more room to oscillate and the amplitude increases and approaches the maximum free air amplitude. The oscillation amplitude of the tip is measured by the detector and a feedback loop adjusts the probe-sample separation to maintain a constant amplitude and force on the sample.

Imaging in a fluid medium tends to dampen the cantilever's normal resonant frequency. In this case, the entire fluid cell can be made to oscillate to drive the cantilever into oscillation. For an appropriate frequency (usually in the range of 5 kHz to 40 kHz), the amplitude of the cantilever decreases when the probe begins to tap the sample, similar to tapping mode operation in air. The probe shape plays an important role in AFM imaging. For example, a large radius

probe is more strongly attracted than a small, high-aspect ratio probe. The probe shape, of course, also plays an important role in replicating the geometry of the sample.

15.8. Surface analysis and depth profiling.

The discussions in this chapter are applicable to all oxide-semiconductor systems. However, the examples are generally directed at the $Si0_2$-Si system, since that is the most important one. There are four general types of charges associated with the $Si0_2$-Si system shown in Fig. 15.35. They are *fixed oxide charge, mobile oxide charge, oxide trapped charge,* and *interface trapped charge.* This nomenclature was standardized in 1978. The abbreviations of the various charges are given below. In each case, Q is the net effective charge per unit area at the $Si0_2$-Si interface (C/cm^2), N is the net effective number of charges per unit area at the $Si0_2$-Si interface (number/cm^2), and D_{it} is given in units of number/$cm^2 \times$eV, $N = |Q/q|$ where Q can be positive or negative, but N is always positive.

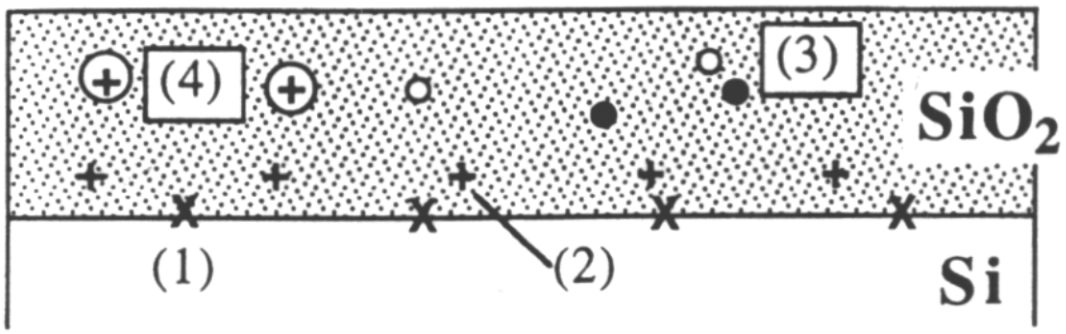


Fig. 15.35 Charges and their location for thermally oxidized silicon.

The charges are described in the following.

1. *Interface Trapped Charge (Q_{it}, N_{it}, D_{it}).* These are positive or negative charges, due to structural defects, oxidation-induced defects, metal impureties, or other defects caused by radiation or similar bond-breaking processes (e.g., hot electrons). The interface trapped charge is located at the Si-$Si0_2$ interface. Unlike fixed charge or trapped charge, interface trapped charge is in electrical communication with the underlying silicon. Interface traps can be charged or discharged, depending on the surface potential. Most of the interface trapped charge can be neutralized by low-temperature (450°C) hydrogen or forming gas (hydrogen-nitrogen mixture) anneals. This charge type has been called surface states, fast states, interface states, and so on. In the past it has been designated by N_{ss}, N_{st} and other symbols.

2. *Fixed Oxide Charge (Q_f, N_f).* This is a positive charge, due primarily to structural defects (ionized silicon) in the oxide layer less than 2 nm from the Si-$Si0_2$ interface. The density of this charge, whose origin is related to the oxidation process, depends on the oxidation ambient and temperature, cooling conditions, and on silicon orientation. Since the fixed oxide charge cannot be determined unambiguously in the presence of moderate

densities of interface trapped charge, it is only measured after a low-temperature (~ 450°C) hydrogen or forming gas anneal, which minimizes interface trapped charge. The fixed oxide charge is not in electrical communication with the underlying silicon. It has been found that Q_f depends on the final oxidation temperature. The higher the oxidation temperature, the lower is Q_f. However, if it is not permissible to oxidize at high temperatures, it is possible to lower Q_f by annealing the oxidized wafer in a nitrogen or argon ambient after oxidation. This has resulted in the well-known "Deal triangle" in Fig. 15.36, which clearly shows the relationship between Q_f and oxidation and annealing. It shows that the processes are reversible. An oxidized sample may be prepared at any temperature and then subjected to dry oxygen at any other temperature, with the resulting value of Q_f being associated with the final temperature. Any Q_f value resulting from a previous oxidation can be reduced to a constant value. Fixed charge was often designated as Q_{ss} in the past .

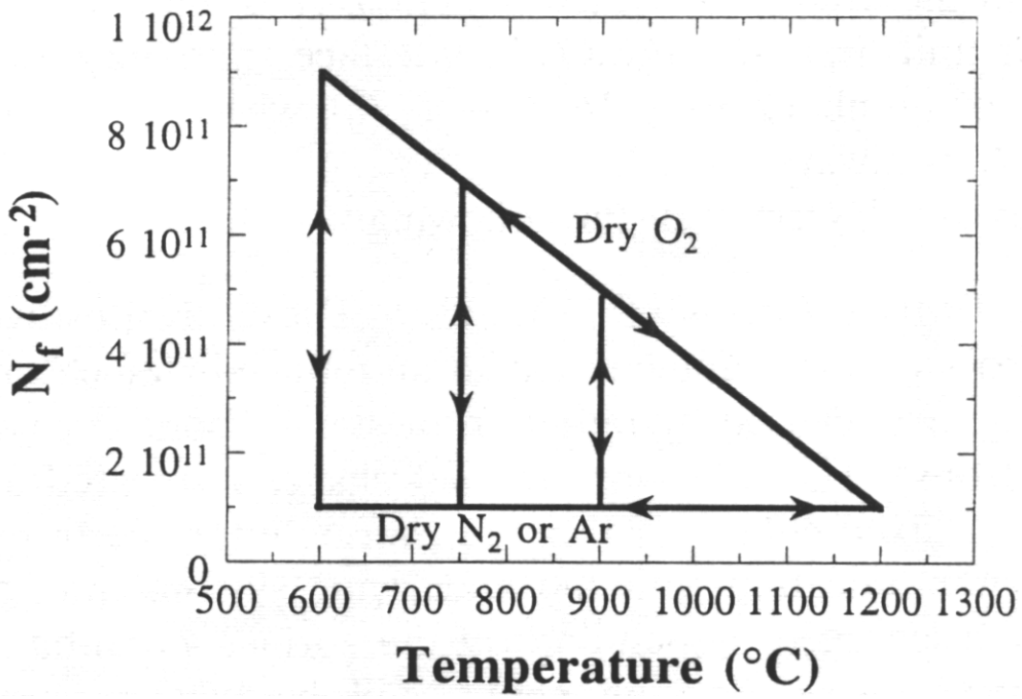


Fig. 15.36 "Deal triangle" showing the reversibility of heat treatment effects on Q_f.

3. Oxide Trapped Charge (Q_{ot}, N_{ot}). This charge may be positive or negative due to holes or electrons trapped in the bulk of the oxide. Trapping may result from ionizing radiation, avalanche injection, Fowler-Nordheim tunneling, or other mechanisms. Unlike fixed charge, oxide trapped charge is sometimes annealed by low-temperature (< 500°C) treatments, although neutral traps may remain.

4. Mobile Oxide Charge (Q_m, N_m). This is caused primarily by ionic impurities such as Na^+, Li^+, K^+, and possibly H^+. Negative ions and heavy metals may contribute to this charge even though they are typically not mobile below 500°C.

The various charges can be measured by many methods, the most popular being the capacitance-voltage (C-V) measurement of a metal-oxide-semiconductor capacitor (MOS-C). Before discussing measurement methods, we derive the capacitance relationships and describe the C-V curves. The energy band diagram of an MOS capacitor on a p-type substrate is shown in Fig. 15.37. The intrinsic energy level E_i is taken as the zero reference potential. The surface

$$C = \frac{dQ}{dV} \tag{15.24}$$

potential ϕ_s is measured from this reference level. The capacitance is defined as

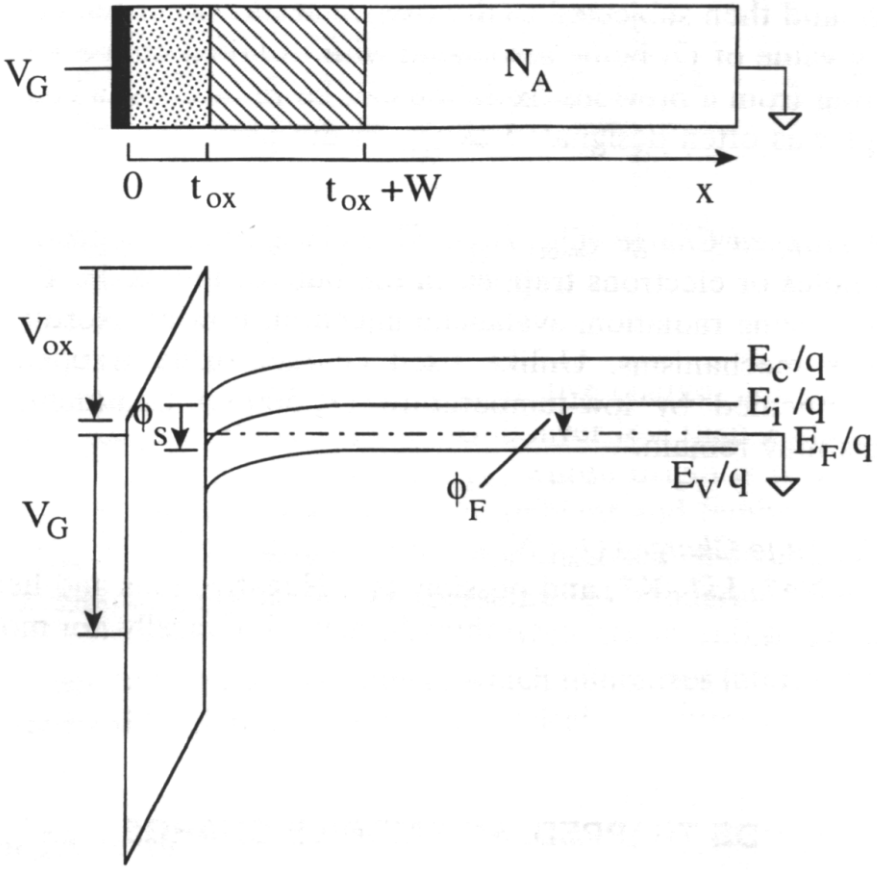


Fig. 15.37 Cross section and potential band diagram of an MOS capacitor.

It is the change of charge due to a change of voltage. During a capacitance measurement, a small-signal ac voltage is applied to the device. The resulting charge variation gives rise to the

$$C = -\frac{dQ_s + dQ_{it}}{dV_{ox} + d\phi_s}$$

capacitance. Looking at an MOS-C from the gate, we find $C = dQ_G/dV_G$, where Q_G and V_G are the gate charge and the gate voltage, respectively. Since the total charge in the device must be zero, $Q_G = -(Q_s + Q_{it})$, assuming no oxide charge. The gate voltage is partially dropped across the oxide and partially across the semiconductor. This gives $V_G = V_{FB} + V_{ox} + \phi_s$, where V_{FB} is the flatband voltage, V_{ox} the oxide voltage, and ϕ_s the semiconductor voltage or surface potential, allowing Eq. (15.24) to be rewritten as

The semiconductor charge Q_s, in general, consists of hole charge Q_p, space-charge region bulk charge Q_b, and electron charge Q_n. With $Q_s = Q_p + Q_b + Q_n$. Eq. 15.24 becomes

(15.25)

$$C = -\frac{1}{\dfrac{dV_{ox}}{dQ_s + dQ_{it}} + \dfrac{d\phi_s}{Q_p + Q_b + Q_n + dQ_{it}}}$$

Utilizing the general capacitance definition of Eq. (15.24), Eq. (15.26) becomes

The positive accumulation charge Q_p dominates for negative oxide and negative gate

(15.26)

$$C = \frac{1}{\dfrac{1}{C_{ox}} + \dfrac{1}{C_p + C_b + C_n + C_{it}}} = \frac{C_{ox}(C_p + C_b + C_n + C_{it})}{C_{ox} + C_p + C_b + C_n + C_{it}} \quad (15.27)$$

voltages for p-substrate devices. For positive V_G, the semiconductor charges are negative. The minus sign in Eq. (15.26) cancels in either case.

Equation (15.27) is represented by the equivalent circuit in Fig. 15.38(a). For negative gate voltages, the surface is heavily accumulated and Q_p dominates. C_p is very high, approaching a short circuit. Hence, the four capacitances are shorted as shown in Fig. 15.38(b). For small positive gate voltages, the surface is depleted and the space-charge region charge, Q_b =--qN_AW, dominates. Trapped interface charge capacitance also contributes. The total capacitance is the combination of C_{ox} in series with C_b in parallel with C_{it}. In weak inversion C_n begins to appear. Figure 15.38(c) shows the equivalent circuit for weak inversion. For strong inversion, C_n dominates because Q_n is very high.

If Q_n is able to follow the applied ac voltage, the *low-frequency* equivalent circuit (Fig. 15.38(d)) becomes the oxide capacitance again. This gives the *low-frequency* C-V curve. When the inversion charge is unable to follow the ac voltage, the circuit in Fig. 15.38(e) applies in inversion, with $C_b = K_s \varepsilon_0 / W_{inv}$ giving the *high-frequency* C-V curve. W_{inv} is the minimum or final space-charge region width.

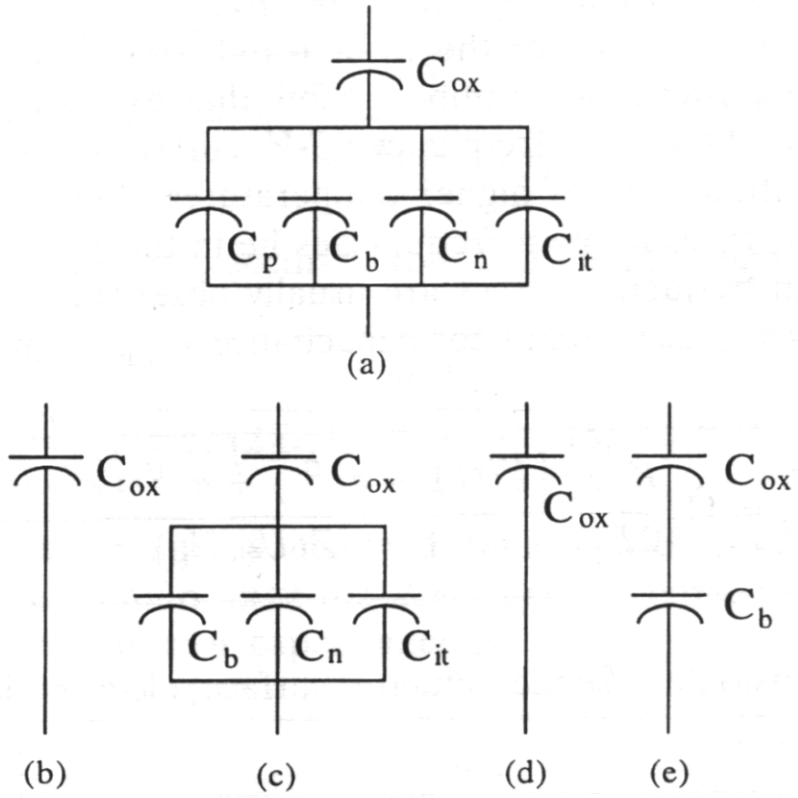


Fig. 15.38 Capacitances of an MOS capacitor for various bias conditions as discussed in the text.

The inversion capacitance dominates only if the inversion charge is able to follow the frequency of the applied ac voltage, also called the ac probe frequency. With the MOS-C biased in inversion, the ac voltage drives the device periodically above and below the dc bias point. During the phase when the device is driven to a slightly higher bias voltage, electron-hole pairs (ehp) are thermally generated in the space-charge region (scr).

Ideal low-frequency C-V curves (lf), consisting of $C_{s\ lf}$ in series with C_{ox}, are shown in Fig. 15.39 for zero Q_{it} and V_{FB}. The normalized flatband capacitance is 0.7, as shown by the dots. Ideal high-frequency (hf) and deep-depletion (dd) curves are also shown in Fig. 6.5. They coincide with the low-frequency curve in accumulation and depletion but deviate in inversion, because the inversion charge is unable to follow the applied ac voltage for the hf case and does not exist for the dd case. Which of these three curves is obtained during a C-V measurement? That depends on the measurement conditions. Let us consider an MOS-C on a p-substrate with

the dc gate voltage swept from negative to positive voltages. Superimposed on the dc voltage is a small-amplitude ac voltage of typically 10-15 mV amplitude. The ac voltage is necessary to measure the capacitance, while the dc voltage determines the bias condition. All three curves are identical in accumulation and depletion. The curves deviate from one another when the device enters inversion. If the dc voltage is swept sufficiently slowly to allow the inversion charge to form and if the ac voltage is of a sufficiently low frequency for the inversion charge to be able to respond to the ac probe frequency, then the low-frequency curve is obtained. If the dc voltage is swept sufficiently slowly to allow the inversion charge to form but the ac probe frequency is too high for the inversion charge to be able to respond, then the high-frequency curve is obtained. The deep-depletion curve obtains for either high or low frequency if the dc sweep rate is too high and no inversion charge can form during the sweep.

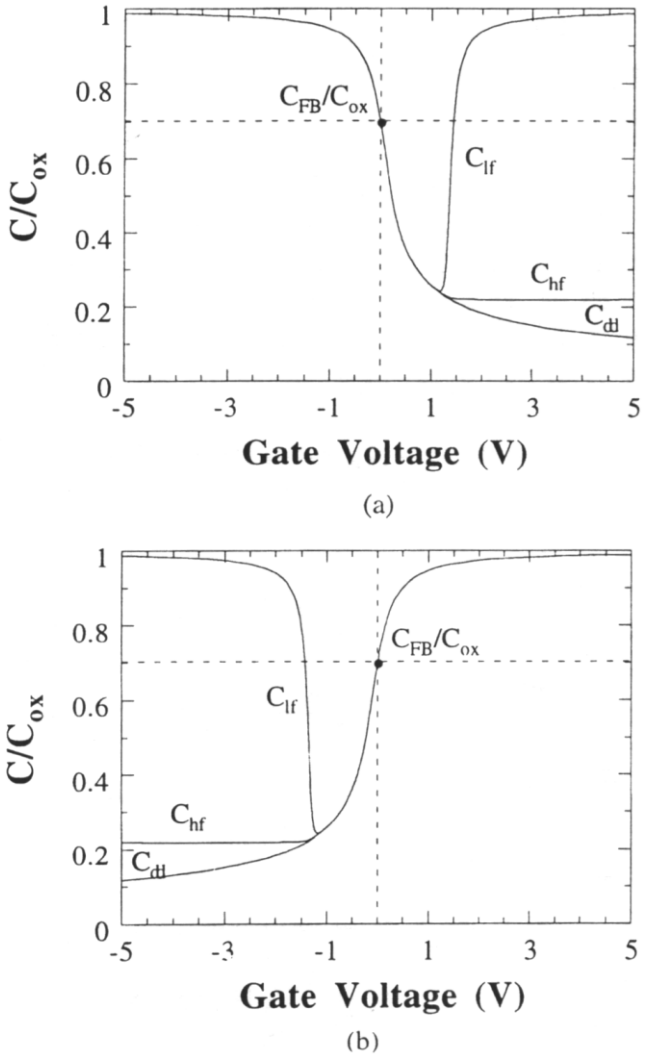


Fig. 15.39 Low-frequency (lf), high-frequency (hf), and deep-depletion (dd) normalized SiO_2-Si capacitance-voltage curves of an MOS-C; (a) p-substrate $N_A = 10^{17}$ cm^{-3}, (b) n-substrate $N_D = 10^{17}$ cm^{-3}, $t_{ox} = 10$ nm, T = 300 K.

High-frequency C-V curves are typically measured at 100 kHz-1 MHz, but sometimes lower frequencies are used. Occasionally one uses higher frequencies. The basic capacitance measuring circuit, shown in Fig. 15.40, consists of the device to be measured and an output resistor R. The MOS device to be measured has a gate-induced space-charge region (scr), represented by the parallel G-C circuit, with G being the conductance of the scr and C its capacitance.

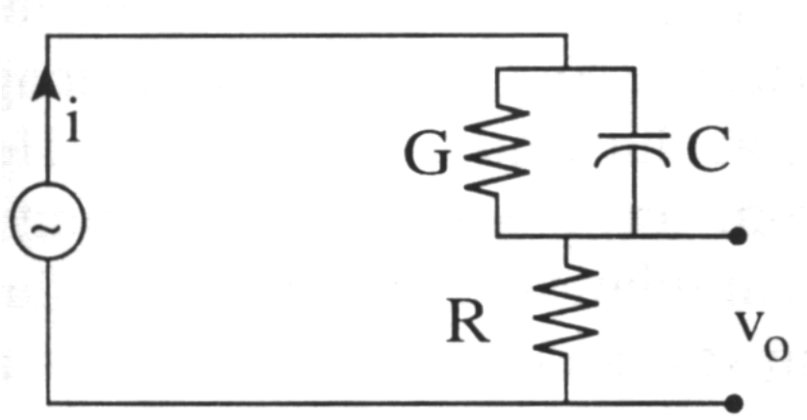


Fig. 15.40 Simplified capacitance measuring circuit.

An ac current i flows through the device and the resistor, giving the output voltage as

For R such that RG<<1 and $(\omega RC)^2 << RG$, Eq. (15.28) reduced to

$$\upsilon_0 = iR = \frac{R}{Z}\upsilon_i = \frac{R}{R+(G+j\omega C)^{-1}}\upsilon_i = \frac{RG(1+RG)+(\omega RC)^2+j\omega RC}{(1+RG)^2+(\omega RC)^2}\upsilon_i \quad (15.28)$$

$$\upsilon_0 = (RG + j\omega RC)\upsilon_i \quad (15.29)$$

The output voltage has two components: the in-phase RG and the out-of-phase $j\omega RC$, with $\upsilon_0 = RG\upsilon_i$ for the 0° phase and $\omega RC\upsilon_i$ for the 90° phase components. Using a phase sensitive detector, one can determine the conductance G or the capacitance C, knowing R and $\omega = 2\pi f$.

The low-frequency capacitance is usually not obtained by measuring the capacitance, but rather by measuring a current or a charge, because capacitance measurements are difficult to do at low frequencies. The quasistatic or linear ramp voltage method overcomes this problem. The current is measured by applying a slowly varying voltage ramp as shown in Fig. 15.41(a).

The op-amp circuit with a resistive feedback connected to the MOS-C gate is an ammeter. When a changing voltage is applied to a capacitor, the resulting displacement current is given by

$$I = \frac{dQ_G}{dt} = \frac{dQ_G}{dV_G}\frac{dV_G}{dt} = C\frac{dV_G}{dt} \qquad (15.30)$$

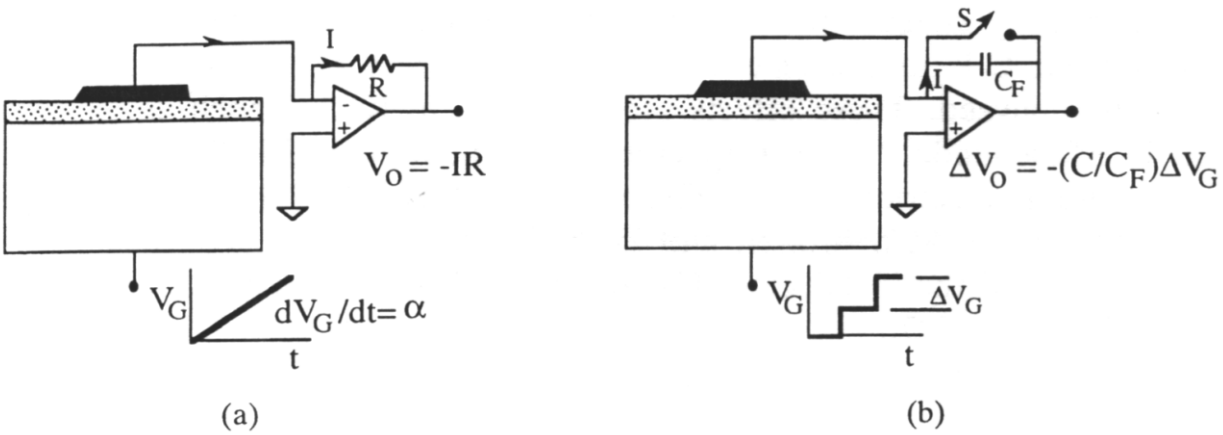


Fig. 15.40 Block diagram of circuit to measure the current and charge of MOS capacitor.

If dV_G/dt is constant as for a linear voltage ramp, then / is proportional to C. If, furthermore, dV_G/dt is sufficiently low, then the low-frequency C-V curve is obtained.

The nature of interface trapped charge is not completely understood, but Q_{it} can be controlled to the extent that very low values are achieved in today's devices. A number of measurement techniques have been developed over the years; we will describe the main ones in this chapter.

The low-frequency or quasistatic method is the most common interface trapped charge measurement method. It uses standard semiconductor laboratory components and the evaluation theory is simple. It provides information only on the interface trapped charge density, but not on the capture cross section of the traps. In this chapter we use the terms "interface trapped charge" and "interface traps" interchangeably. The effect of interface traps on both hf and If C-V curves is shown in Fig. 15.41. For the calculated hf curves in Fig. 15.41(a), we assume an interface trap density distribution uniform in energy through the band gap with D_{it} acceptor-like in the upper half and donor-like in the lower half of the band gap. Further, we assume that the interface traps cannot respond to the probe frequency. The distortion in the "$D_{it} \neq 0$" hf curve is due to voltage stretch-out with interface trap charge adding to the gate voltage. For $\phi_s = \phi_F$ the upper half band gap donor-type and lower half band gap acceptor-type interface traps cancel one another, leading to the coincidence of the ideal and distorted C-V curves. The lf C-V curves distort due to the additional capacitance when the interface traps can respond to the low probe frequencies and because of voltage stretch-out as shown in Fig. 15.41(b). If interface traps have a peaked distribution, then C_{lf} also has a peaked shape.

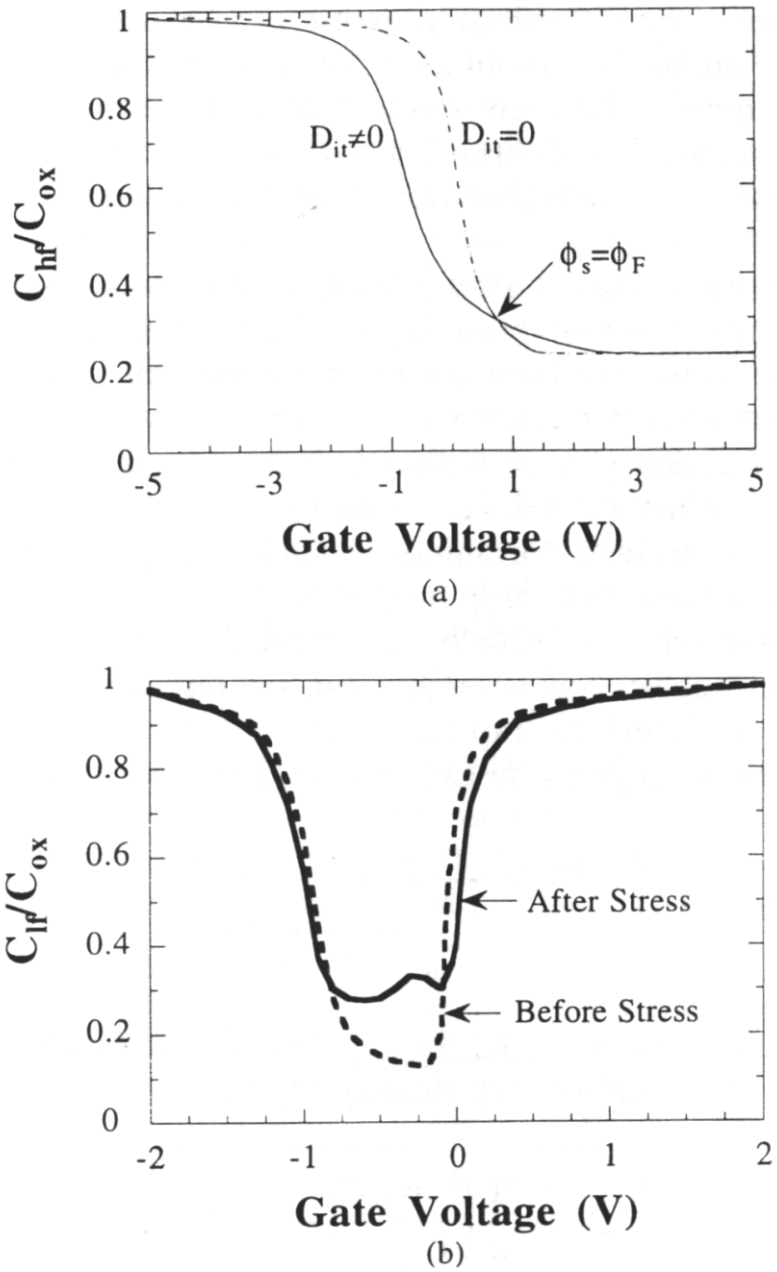


Fig. 15.41 Theoretical ideal ($D_{it} = 0$) and $D_{it} \neq 0$ (a) hf and (b) lf C-V curves.

The room-temperature, high-frequency capacitance method developed by Terman was one of the first methods for determining the interface trap density. The method relies on a hf C-V measurement at a frequency sufficiently high that interface traps are assumed not to respond. They should, therefore, not contribute any capacitance.

How can one measure interface traps if they do not respond to the applied ac signal? Although interface traps do not respond to the ac probe frequency, they do respond to the slowly varying dc gate voltage and cause the hf C-V curve to stretch out along the gate voltage axis as interface trap occupancy changes with gate bias.

In other words, for an MOS-C in depletion or inversion additional charge placed on the gate by a gate voltage induces additional semiconductor charge $Q_G = -(Q_b + Q_n + Q_{it})$.

With

$$V_G = V_{FB} + \phi_s + V_{ox} = V_{FB} + \phi_s + \frac{Q_G}{C_{ox}} \tag{15.30}$$

it is obvious that for a given surface potential ϕ_s, V_G varies when interface traps are present. This is the cause of the C-V "stretch-out" shown in Fig. 14.41. The stretch-out produces a nonparallel shift of the C-V curve. Interface traps distributed uniformly through the semiconductor band gap produce a fairly smoothly varying but distorted C-V curve. Interface traps with distinct structure, for example, peaked distributions, produce more abrupt distortions in the C-V curve.

The relevant equivalent circuit of the hf MOS-C is that in Fig. 15.38(c) with $C_{it} = 0$, that is, $C_{hf} = C_{0x}\, C_S/(C_{0x} + C_s)$ where $C_s = C_b + C_n$. C_{hf} is the same as that of a device without interface traps provided C_s is the same. The variation of C_s with surface potential is known for an ideal device. Knowing ϕ_s for a given C_{hf} in a device without Q_{it} allows us to construct a ϕ_s versus V_G curve of the actual capacitor. This is done as follows: From the ideal MOS-C C-V curve, find ϕ_5 for a given C_{hf}. Then find V_G on the experimental curve for the same C_{hf}, giving one point of a ϕ_5 versus V_G curve. Repeat for other points until a satisfactory ϕ_s-V_G curve is constructed. It is this ϕ_s-V_G curve that contains the relevant interface trap information.

The experimental ϕ_s versus V_G curve is a stretched-out version of the theoretical curve and the interface trap density is determined from this curve by

$$D_{it} = \frac{C_{ox}}{q}\left(\frac{dV_G}{d\phi_s} - 1\right) - \frac{C_s}{q} = \frac{C_{ox}}{q}\frac{d\Delta V_G}{d\phi_s} \tag{15.31}$$

where $\Delta V_G = V_G - V_G(ideal)$ is the voltage shift of the experimental from the ideal curve, with V_G being the experimental gate voltage. The method is not widely used, but is generally considered to be useful for measuring interface trap densities of 10^{10} cm^{-2} eV^{-1} and above. The Terman method has been widely critiqued. Its limitations were originally pointed out to be due to inaccurate capacitance measurements and insufficiently high frequencies. A later theoretical study concluded that D_{it} in the 10^9 cm^{-2} eV^{-1} range can be determined provided the capacitance is measured to a precision of 0.001 to 0.002 pF.

To compare experimental curves with theoretical curves, one needs to know the doping density exactly. Any dopant pile up or out-diffusion introduces errors. Surface potential fluctuations can cause fictitious interface trap peaks near the band edges. The assumption that

interface traps do not follow the ac probe frequency may not be satisfied for surface potentials near flatband and towards accumulation unless exceptionally high frequencies are used. Lastly, graphical differentiation of the ϕ_s versus V_G curve can cause errors. Large discrepancies were found for D_{it} determined by the Terman technique compared with deep level transient spectroscopy (DLTS), hence the Terman method is thought to be of questionable accuracy.

15.9. Summary of techniques for property measurements.

Electron Microscopy and Transmission Electron Microscopy. Its major strength lies in its unprecedented atomic resolution imaging ability. To achieve this, one must prepare extremely thin samples, therefore making sample preparation its major weakness. This has been somewhat alleviated by focused ion beam sample preparation.

Electron Microprobe. Its major strength is the availability of EMP on many scanning electron microscopes and the relatively simple way of obtaining quantitative elemental information. The energy resolution of energy dispersive spectroscopy is modest, but usually sufficient. For higher energy resolution one needs to use wavelength dispersive spectroscopy, which is more difficult to use. A weakness is the modest spatial resolution of the technique and the damage the electron beam can inflict on semiconductor samples.

Secondary Ion Mass Spectrometry. Its major strength lies in its sensitivity (better than most beam techniques) and the ability to detect all impurities. Furthermore, it is one of the most commonly used beam techniques for dopant profiling and, with the recent time-of-flight SIMS, for organic contaminants. Its weaknesses include matrix effects, molecular interferences, the destructive nature of the measurement, and the need for calibrated samples.

X-Ray diffraction and fluorescence. Its major strength is the ability for rapid, contactless survey of elements. Its weakness is the modest resolution due to the difficulty of focusing X-rays and the presence of matrix effects. The sensitivity to surface contamination is greatly extended by total reflection X-ray fluorescence. *X*-Ray Photoelectron Spectroscopy. Its major strength is the ability to characterize the elemental and molecular nature of thin layers. Its weakness is the modest resolution due to the difficulty of focusing X-rays, the high vacuum requirement, and its low sensitivity.

Rutherford Backscattering. Its major strength lies in its contactless and absolute measurements without recourse to calibrated standards. Its major weakness is the specialized nature of the instrumentation and the difficulty of measuring light elements on a heavy element substrate.

Scanning Probe Microscopy. Its major strength is the ability to image samples at very high resolution both laterally and vertically. No other technique competes here. Its weakness is the

fragile nature of the probe tips and the interpretational difficulties, because one does not always know what is being imaged.

Interface Trapped Charge. For MOS capacitors the choice for the most practical methods lies between the high-frequency capacitance, conductance and the quasistatic methods. These are the two most widely used techniques. The strength high-frequency of the capacitance and conductance method lies in its sensitivity and its ability to give the majority carrier capture cross sections. Its major weakness is the limited surface potential range over which D_{it} is obtained and the amount of work required to extract D_{it}, although simplified methods have been proposed. The main strengths of the quasistatic method (both the *I-V* and the *Q-V*) are the relative ease of measurement and the large surface potential range over which D_{it} is obtained. A weakness for the *I-V* version is the current measurement requirement. The currents are usually low because the sweep rates must be low to ensure quasi equilibrium. The recent *Q-V* version alleviates some of these problems and commercial instruments are available.

Chapter 16. Process monitoring

A.Shkavro, A.Evtukh

16.1. Process flow and key measurement points

The basic unit processes used in fabricating an integrated circuit, as well as the process flows for several major IC technologies were considered in detail. In order for these processes to repeatably produce reliable, high-quality devices and circuits, each unit process must be strictly controlled. Many diagnostic tools are used to maintain systematic control. Such control requires that the key output variables for each process step (i.e., those that are correlated with product functionality and performance) be carefully monitored.

Process monitoring enables operators and engineers to detect problems early on to minimize their impact. The economic benefit of effective monitoring systems increases with the complexity of the manufacturing process. Manufacturing line monitors consist of extremely sophisticated metrology equipment that characterizes the state of features on the semiconductor wafers themselves.

When we monitor a physical system, we observe that system's behavior. On the basis of these observations, we take appropriate actions to influence that behavior in order to guide the system to some desirable state. Semiconductor manufacturing systems consist of a series of sequential process steps in which layers of materials are deposited on substrates, doped with impurities, and patterned using photolithography to produce sophisticated integrated circuits and devices. As an example of such a system, Figure 16.1 depicts a typical CMOS process flow. Inserted into this flow diagram in various places are symbols denoting key measurement points. Clearly, CMOS technology involves many unit processes with high complexity and tight tolerances. This necessitates frequent and thorough inline process monitoring to assure high-quality final products.

The measurements required may characterize physical parameters, such as film thickness, uniformity, and feature dimensions; or electrical parameters, such as resistance and capacitance. These measurements may be performed directly on product wafers, either directly or using test structures, or alternatively, on nonfunctional monitor wafers (or "dummy" wafers). In addition to these, some measurements are actually performed "in situ," or *during* a fabrication step. When a process sequence is complete, the product wafer is diced, packaged, and subjected to final electrical and reliability testing.

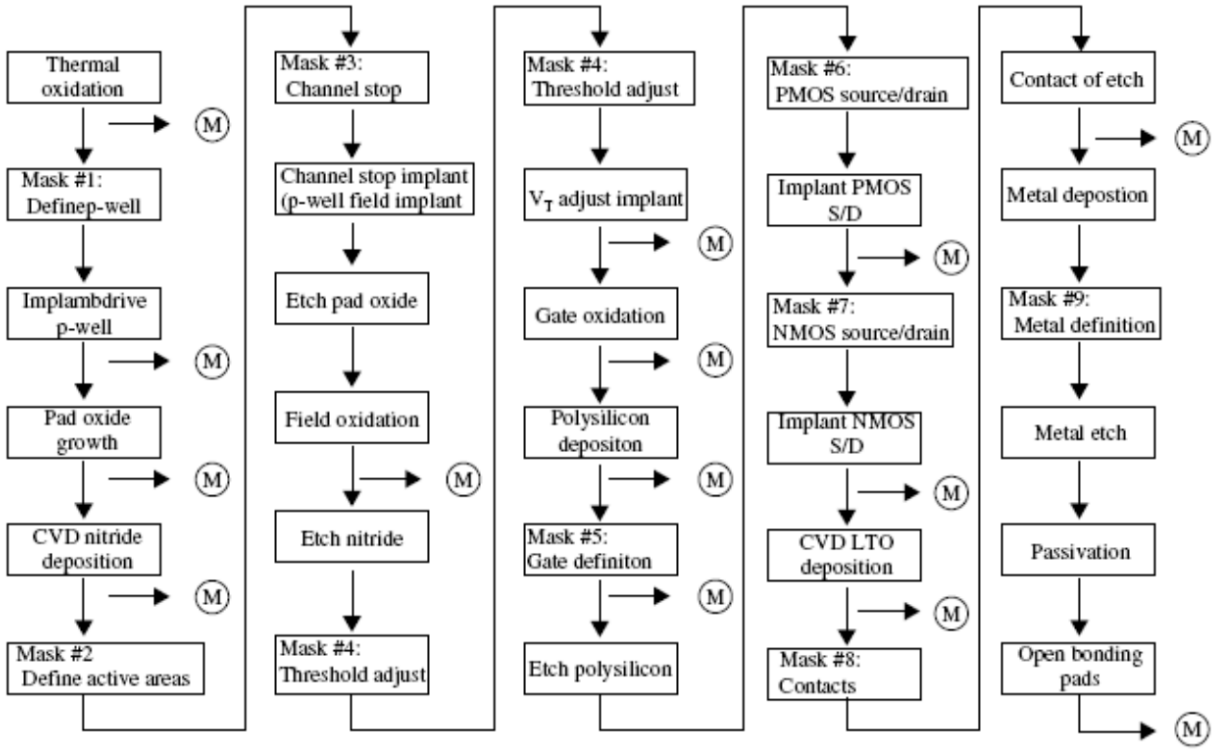


Fig. 16.1. CMOS process flow showing key measurement points (denoted by ‘‘M’’) [1].

16.2. WAFER STATE MEASUREMENTS

There is no substitute for regular inspection of products during manufacturing to ensure high quality. Inspections can reveal contamination, structural flaws, or other problems. Such investigations must not be limited to visual inspections, however, since not all processes have a visible effect on electronic products. Thin-film deposition and ion implantation are two important examples of this. In addition, with ever-increasing levels of integration, features on wafers become smaller and more difficult to inspect. As a result, visual inspection must be supplemented by sophisticated physical and electrical measurements of various characteristics that describe the state of a wafer.

Wafer state characterization includes the measurement of the physical parameters related to each manufacturing process step. Examples include

Lithography: (*i*) linewidth, (*ii*) overlay, (*iii*) print bias, (*iv*) resist profiles.

Etch: (*i*) etch rate, (*ii*) selectivity, (*iii*) uniformity, (*iv*) anisotropy, (*v*) etch bias.

Deposition or Epitaxial Growth: (*i*) sheet resistance, (*ii*) film thickness, (*iii*) surface concentration, (*iv*) dielectric constant, (*v*) refractive index.

Diffusion or Implantation: (*i*) sheet resistance, (*ii*) junction depth, (*iii*) surface concentration.

The total collection of such measurements relate to the physical characteristics of product wafers, and these physical characteristics can be correlated with the electrical performance of devices and circuits.

16.2.1. Blanket Thin Film

Let's begin the discussion of wafer state measurements with those measurements that are performed on blanket thin films. The term "blanket" is used to differentiate wafers that have been uniformly coated by a thin film from those in which the film has been patterned using photolithography and etching.

16.2.1.1. Interferometry

Perhaps the simplest method for determining the thickness of an oxide is to compare the color of the wafer with a reference color chart, such as the one in Table 16.1. When an oxide-coated wafer is illuminated with white light perpendicular to the surface, the light penetrates the oxide and is reflected by the underlying silicon wafer. Constructive interference leads to enhancement of a certain wavelength of reflected light, and the color of the wafer corresponds to that wavelength. For example, a wafer with a 500-nm silicon dioxide layer will appear blue green.

Optical metrology provides fast and precise measurements of film thickness and optical constants. In semiconductor manufacturing, *interferometry* (sometimes called *reflectometry*) is a widely used optical method for measuring such parameters. Single- or multiple wavelength interferometers are commonly used for both in situ and postprocess measurements of film thickness. In this method, a light source, usually a laser is focused on a semiconductor wafer while a detector measures the reflected light intensity. The wafer consists of a parallel stack of partially transparent thin films. The reflected light intensity varies as a function of time depending on the thickness of the top layer due to constructive and destructive interference caused by multiple reflections. To illustrate the basic concept, consider Fig. 16.2, which shows a film of uniform thickness d and index of refraction n, with the eye of the observer focused on point a. The film is illuminated by broad source of monochromatic light S. There is a point P on the source such that two rays (represented by the single and double arrows) can leave P and enter the eye after traveling through point a. These two rays follow different paths, one reflected from the upper surface of the film and the other from the lower surface. Whether point a appears bright or dark depends on the nature of the interference (i.e., constructive or destructive) between the two waves that diverge from a. The two factors that impact the nature of the interference are differences in optical path length and phase changes on reflection.

Table 16.1. Color chart for thermally grown SiO_2 films observed perpendicularly under daylight fluorescent lighting.

Film tickness (μm)	Color and comments	Film Thickness (μm)	Color and comments
0.05	Tan	0.60	Carnation pink
0.07	Brown	0.63	Violet red
0.10	**Dark violet to red violet**	0.68	'Bluish' (not blue but borderline
0.12	Royal blue		between violet and blue green;
0.15	Light blue to metallic blue		appears more like a mixture
0.17	Mettallic to very light yellow green		between violet red and blue green and looks grayish)
0.20	Light gold or yellow, slightly metallic	0.72	Blue green to green (quite broad)
		0.77	"Yellowish"
0.22	Gold with slight yellow orange	0.80	Orange (rather broad for orange)
		0.82	Salmon
0.25	Orange to melon	0.85	Dull, light red violet
0.27	Red violet	0.86	Violet
0,30	Blue to violet blue	0.87	Blue violet
0.31	Blue	0.89	Blue
0.32	Blue to blue green	0.92	Blue green
0.34	Light green	0.95	Dull yellow green
0.35	Green to yellow green	0.97	Yellow to "yellowish"
0.36	Yellow green	0.99	Orange
0.37	Green yellow	1.00	Carnation pink
0 39	Yellow	1.02	Violet red
0.41	Light orange	1.05	Red violet
0.42	Carnation pink	1.06	Violet
0.44	Violet red	1.07	Blue violet
0.46	Red violet	1.10	Green
0.47	Violet	1.11	Yellow green
0.48	Blue violet	1.12	Green
0.49	Blue	1.18	Violet
0.50	Blue green	1.19	Bed violet
0.52	Green (broad)	1.21	Violet red
0.54	Yellow green	1.24	Carnation pink to salmon
0.56	Green yellow	1.25	Orange
0.57	Yellow to "yellowish" (not	1.28	"Yellowish"
	yellow but is in the position	1.32	Sky blue to green blue
	where yellow is to be	1.40	Orange
	expected; at times appears to	1.45	Violet
	be light creamy gray or	1.46	Blue violet
	metallic)	1.50	Blue
0.58	Light orange or yellow to pink	1.54	Dull yellow green

For the two rays to combine to give maximum intensity, we must have

$$2dn\cos\theta = (m + 0.5)\lambda \qquad (16.1)$$

where m = 0, 1, 2, . . . and θ is the angle of the refracted beam relative to the surface normal. The term 0.5λ accounts for the phase change that occurs on reflection since a phase change of 180° is equivalent to half a wavelength. The condition for minimum intensity is

$$2dn\cos\theta = m\lambda \qquad (16.2)$$

Equations (16.1) and (16.2) hold when the index of refraction of the film is either greater or less than the indices of the media on *each* side of the film.

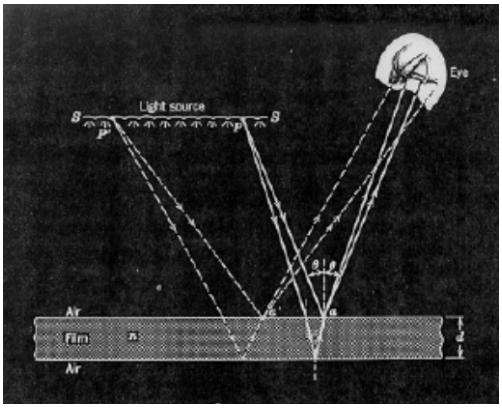


Fig. 16.2. Interference by reflection from a thin film [2].

Therefore, if the index of refraction is known, the thickness of the film may be computed by simply counting peaks or valleys in the reflected waveform. Interferometry becomes more complex when applied to stacks of several thin films. The overall goal, however, is still to obtain film thickness information from the time-varying reflected intensity signal. The reflected light intensity is given by [3]

$$I_r(d,\lambda) = I_0(\lambda)r(d,\lambda,\phi_1,\phi_2,...,\phi_N) \qquad (16.3)$$

where I_0 is the incident light intensity, r is the reflection coefficient, d is the thickness of the top layer, and ϕ_i are physical constants (i.e., thicknesses and refractive indices) associated with the lower films in the film stack. The reflected intensity is monitored using a detector consisting of a lightsensitive transducer, such as a photodiode, in conjunction with an optical filter or diffraction grating to select the wavelength(s) of interest. The output of the detector corresponding to a particular wavelength is of the form

$$y_\lambda(kT) = \alpha(\lambda,kT)A(\lambda,kT)I_0(\lambda,kT)r(d(kT),\lambda)e_\lambda(kT) \qquad (16.4)$$

where T denotes the sampling period, k is an integer, α represents losses in the optical system, A is the gain of the detector, and e_λ is measurement noise. The physical parameters ϕ_i are considered to be fixed and known in this formulation and are not shown. For multiple-wavelength (or

spectroscopic) measurements, this expression is repeated for each wavelength used. For *p* wavelengths, in matrix form, this is written as

$$\vec{y}(kT) = diag(\vec{h}(kT)\vec{r}(d(kT)) + e(kT) \tag{16.5}$$

where *diag(**x**)* represents a matrix with the elements of the vector ***x*** along the diagonal and

$$\vec{y}(kT) = [y_{\lambda 1}(kT)...y_{\lambda p}(kT)] \tag{16.6}$$

$$\vec{h}(kT) = [\alpha(\lambda_1, kT)A(\lambda_1, kT)I_0(\lambda_1, kT)...\alpha(\lambda_p, kT)A(\lambda_p, kT)I_0(\lambda_p, kT)] \tag{16.7}$$

$$\vec{r}(d(kT) = [r(d(kT), \lambda_1.....r(d(kT), \lambda_p]^T \tag{16.8}$$

$$\vec{e}(kT) = [e_{\lambda 1}(kT)....e_{\lambda p}(kT)]^T \tag{16.9}$$

where the superscript *T* represents the transpose operation.

To obtain film thickness or the rate of change of thickness (i.e., etch or deposition rate), the detector output is processed in one of two ways: (1) extrema counting or (2) least-squares fitting. Extrema counting takes advantage of the fact that the reflected light intensity varies approximately periodically with both the wavelength of the incident light and the thickness of the top film. The distance between peaks and valleys is a known function of the top film thickness. Thus, if many wavelengths are available, thickness can be determined by counting the peaks in a plot of reflectance versus wavelength. If only a single wavelength is available, the movement of peaks and valleys over time during in situ measurements indicates that a specific amount of material has been etched or deposited. This provides the average etch or deposition rate between successive minima and maxima.

To use the least-squares approach, at each timepoint, the following nonlinear optimization problem is posed:

$$\min_d[(\vec{y}(kT) - diag(\vec{h})\vec{r}(d))^T(\vec{y}(kT) - diag(\vec{h})\vec{r}(d))] \tag{16.10}$$

The film thickness is then that for which the minimum is achieved. Etch rate or deposition rate is then calculated from the resulting thickness versus time curve.

The final variation of interferometry is one that is particularly applicable to thickness monitoring during plasma etching. During etching, the emission from the plasma itself may be used as the light source. As this light is reflected from the etched film and underlying film surfaces while the thickness of the etched film decreases, the optical path difference between light rays varies and the changing constructive and destructive interference results in periodic signals in the same manner as previously described. If a chargecoupled device (CCD) camera is placed in such a way that it can view these signals (see Fig. 16.3), each pixel of the CCD camera then acts as an individual interferometer monitoring a different part of the wafer. This arrangement is called *full-wafer interferometry* [4].

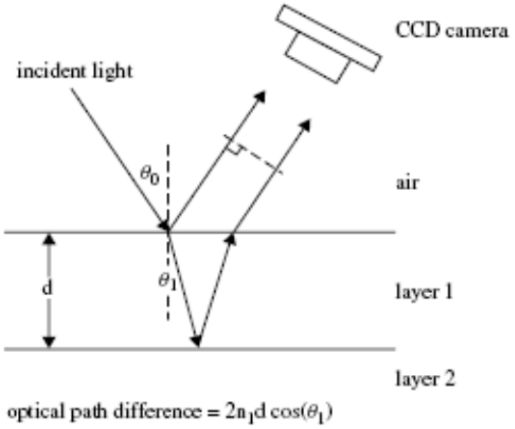


Fig. 16.3. Schematic of full-wafer interferometry [4].

Clearly color chart comparisons are subjective and are therefore not the most accurate mechanism for determining oxide thickness. A more accurate measurement can be obtained using techniques such as profilometry or ellipsometry.

16.2.1.2. Ellipsometry

Ellipsometry is another widely used measurement technique that is based on the polarization changes that occur when light is reflected from or transmitted through a medium. Changes in polarization are a function of the optical properties of the material (i.e., its complex refractive indices), its thickness, and the wavelength and angle of incidence of the light beam relative to the surface normal (Fig. 16.4). These differences in polarization are measured by an ellipsometer, and the oxide thickness can then be calculated.

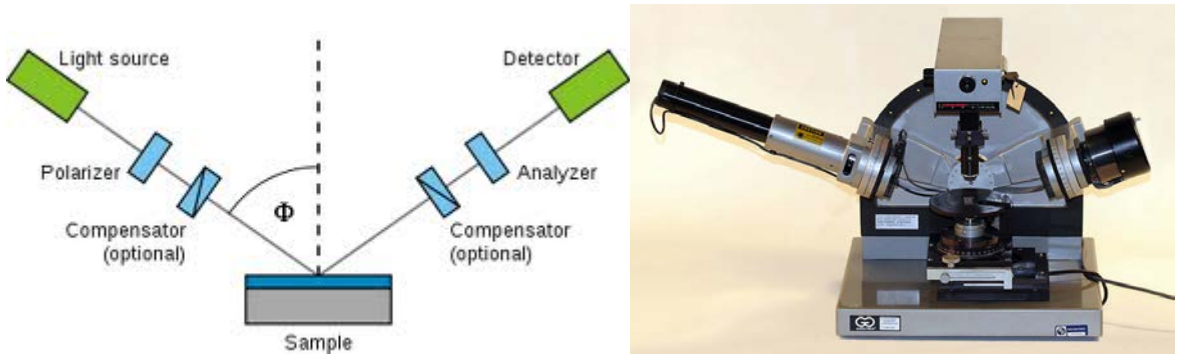


Fig. 16.4. Schematic image and photograph of the ellipsometer.

When multiple light beams of varying wavelength are used, the technique is referred to as *spectroscopic ellipsometry* (SE). SE, which can be used to make in situ or postprocess measurements, is a fundamentally more accurate technique than interferometry for obtaining film thickness and optical dielectric function information. In general, SE measurements are performed at

an off-normal angle with respect to the sample. In this configuration, the measurement is sensitive to the polarization state of both the incident and reflected waves. Figure 16.5 shows an unpolarized beam of light falling on a dielectric surface. In this case, the dielectric is glass. The electric field vector for each wave train in the beam can be resolved into two components—one perpendicular to the plane of incidence (i.e., the plane of the figure) and another parallel to this plane. The perpendicular component, represented by the dots, is the σ component (or "*s* component"). The parallel component, represented by the arrows, is the π component (or "*p* component"). On average, for completely unpolarized incident light, these two components are of equal amplitude. However, if the incident beam is polarized (as is the case in ellipsometry), this is no longer true. In the most common configuration, linearly polarized light is incident on the surface, and the elliptical polarization status of the reflected light is analyzed. Measured ellipsometry data are usually written in the form of the ratio (ρ) of the *total reflection coefficients* for *s* and p polarization (R^s and R^p, respectively). In other words

$$\rho = R^p / R^s = \tan(\psi)e^{i\Delta} \tag{16.11}$$

where $\tan(\psi)$ is the ratio of the magnitude of the *p*-polarized light to the *s*-polarized reflected light and Δ is the difference in phase shifts on reflection for the *p* and *s* polarizations, respectively.

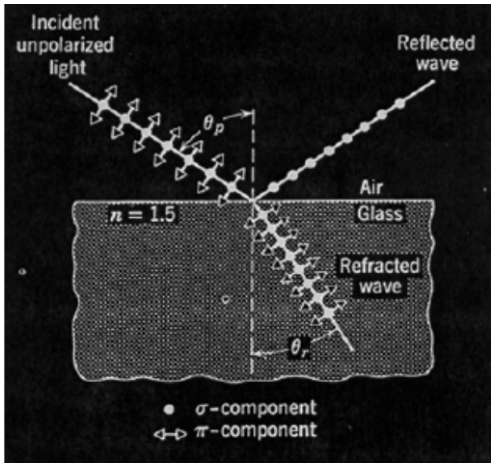


Fig. 16.5. Illustration of components of polarization [2].

Another set of expressions called the *Fresnel equations* relate [Eq. (16.11)] to the bulk complex dielectric function (ε). The dielectric function represents the degree to which the material may be polarized by an applied external electric field, and as a complex number, it is expressed as

$$\varepsilon = \varepsilon_1 + j\varepsilon_2 \tag{16.12}$$

where ε_1 and ε_2 are the real and imaginary parts, respectively. For heterogeneous samples consisting of multiple layers, the dielectric function determined by ellipsometry is an average over the region penetrated by the incident light called the *effective dielectric function*, $\langle\varepsilon\rangle$. If the sample structure is

not too complicated, <ε> can be simulated by appropriate models (such as the "ambient–film–substrate" model). In this case, film and substrate properties can be separated, and film properties (i.e., thickness or dielectric function) can be determined as follows. Because there are a maximum of two independent optical parameters (ψ and Δ) measured at each wavelength, the maximum number of unknowns that can be determined from a single spectral measurement is 2*w*, where *w* is the number of wavelengths scanned. Thus far, we have discussed the index of refraction as if it were a single parameter. However, in general, the *complex index of refraction* (*N*) consists of a real part (*n*) and an imaginary part (*k*), or

$$N = n - jk \tag{16.13}$$

where *k* is the *extinction coefficient*, which is a measure of how rapidly the intensity decreases as light passes through a material. The dielectric function is related to the complex index of refraction by the relationship

$$\varepsilon = N^2 \tag{16.14}$$

Therefore, we can obtain values for *n* and *k* in terms of ε_1 and ε_2 using

$$n = \sqrt{\frac{1}{2}[\varepsilon_1^2 + \varepsilon_2^2)^{1/2} + \varepsilon_1]} \tag{16.15}$$

$$k = \sqrt{\frac{1}{2}[\varepsilon_1^2 + \varepsilon_2^2)^{1/2} - \varepsilon_1]} \tag{16.16}$$

As mentioned above, the complex index of refraction is related to the total reflection coefficients by the Fresnel equations, which are given by [5]

$$R^p = \frac{r_{12}^p + r_{23}^p \exp(-j2\beta)}{1 + r_{12}^p r_{23}^p \exp(-j2\beta)} \tag{16.17}$$

$$R^s = \frac{r_{12}^s + r_{23}^s \exp(-j2\beta)}{1 + r_{12}^s r_{23}^s \exp(-j2\beta)} \tag{16.18}$$

where the Fresnel reflection coefficients at the individual interfaces are of the form

$$r_{12}^p = \frac{N_2 \cos\phi_1 - N_1 \cos\phi_2}{N_2 \cos\phi_1 + N_1 \cos\phi_2} \tag{16.19}$$

$$r_{12}^p = \frac{N_1 \cos\phi_1 - N_2 \cos\phi_2}{N_1 \cos\phi_1 + N_2 \cos\phi_2} \tag{16.20}$$

and

$$\beta = 2\pi\left(\frac{d}{\lambda}\right) N_2 \cos\phi_2 \tag{16.21}$$

All subscripts and angles mentioned in Eqs. (16.17)–(16.21) are described in Fig. 16.6.

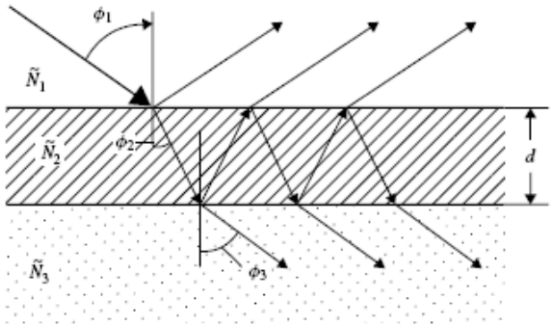


Fig. 16.6. Reflections and transmissions in ambient (1), film (2), and substrate (3) [1].

Thus, materials with finite light absorption have two unknowns (ε_1 and ε_2, or equivalently, n and k), at each wavelength and one additional unknown in the film thickness. Thus, the total number of unknowns is $2w + 1$. Because this number of unknowns is one too many to be determined from spectroscopic ellipsometry data, it is necessary to employ a dispersion model. Such a model describes the functional dependence of n and k on λ based on P fitting parameters. Therefore, the total number of unknowns becomes $P + 1$. As long as $2w > P + 1$, film thickness and the optical constants may be determined simultaneously by numerically iterating the $P + 1$ fitting parameters to fit spectra [3]. For example, for a thin film on a substrate, the usual objective is to determine thickness d for a known substrate and film dielectric function. To do so, the value of d is found that minimizes the function

$$\sum_{\lambda} \left| \langle \varepsilon \rangle - \langle \varepsilon \rangle_{calc} \right|^2 \tag{16.22}$$

(or similar functions using ρ, or ψ and Δ) [6]. Here, the first term represents measured values, and the second term represents theoretically calculated values. This expression can be minimized using well-known procedures such as Newton's method or the Levenberg–Marquardt algorithm [7].

16.2.1.3. Quartz Crystal Monitor

The deposition of metals such as aluminum is often accomplished using the evaporation technique. The deposition rate during evaporation operations is commonly measured using a device known as a *quartz crystal monitor*. This device is a vibrating crystal sensor that is allowed to oscillate at its resonant frequency as the frequency is monitored. This resonant frequency then shifts as a result of mass loading as additional mass from the evaporated metal is deposited on top of the crystal. When enough material has been added, the resonant frequency shifts by several percent. By feeding the frequency measurements to the mechanical shutters of the evaporation system, the thickness of the deposited layer, as well as its time rate of change, can be readily monitored. The sensing elements needed to detect such shifts are quite inexpensive and easy to replace. This method is effective for a wide range of deposition rates.

16.2.1.4. Four-Point Probe

The *four-point probe* is an instrument used to measure the resistivity and sheet resistance of diffused layers. As depicted schematically in Fig. 16.7, this technique requires a fixed current to be injected into the wafer surface through two outer probes. The resulting voltage is measured between two inner probes. If the probes have a uniform spacing (s, in cm), and the sample is infinite, then the resistivity in Ω-·cm is given by [8]

$$\rho = 2\pi sV / I \tag{16.23}$$

for $t >> s$ and

$$\rho = (\pi t / \ln 2)V / I \tag{16.24}$$

for $s >> t$. For shallow layers such as this, Eq. (16.24) means that the sheet resistance (R_s) is then given by

$$R_s = \rho / t = (\pi / \ln 2)V / I = 4.53V / I \tag{16.25}$$

Although the approximations used in Eqs. (16.24) and (16.25) are valid for shallow diffused layers in silicon, different correction factors must be used for sheet resistance measurements on bulk wafers.

It should be noted that monitor wafers used for sheet resistance measurements can also be used to determine junction depth (x_j). After the wafers are diffused or implanted with dopants, the thickness of the diffused region is defined as the junction depth. This parameter may be determined from sheet resistance measurements by replacing t with x_j in Eq. (16.25).

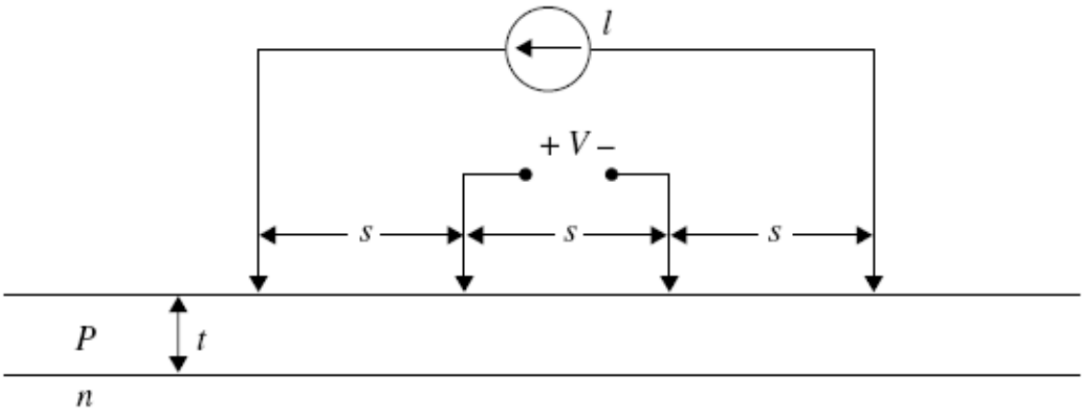


Fig. 16.7. Schematic of four-point probe measurement [8]. In this example, the sheet resistance of a *p*-type epitaxial layer of thickness t on an *n*-type substrate is measured.

16.2.2. Patterned Thin Film

Let’s continue the discussion of wafer state measurements with those measurements that are performed primarily on wafers that have previously been patterned using photolithography and etching to form specific structures or devices.

16.2.2.1. Profilometry

Profilomentry is a very common method of film thickness measurement. In this technique, a step feature in the grown or deposited film is first created, either by masking during deposition or by etching afterward. The profilometer then drags a fine stylus across the film surface (see Fig. 16.8). When the stylus encounters a step, a signal variation indicates the step height. This information is then displayed on a chart recorder or CRT screen. Films of thickness of less than 100 nm to greater than 5 µm can be measured with this instrument. The measurement of thin films is difficult because of vibration, surface roughness, and the precision required in leveling the instrument. Some more recently developed surface profilometers use atomic force microscopy.

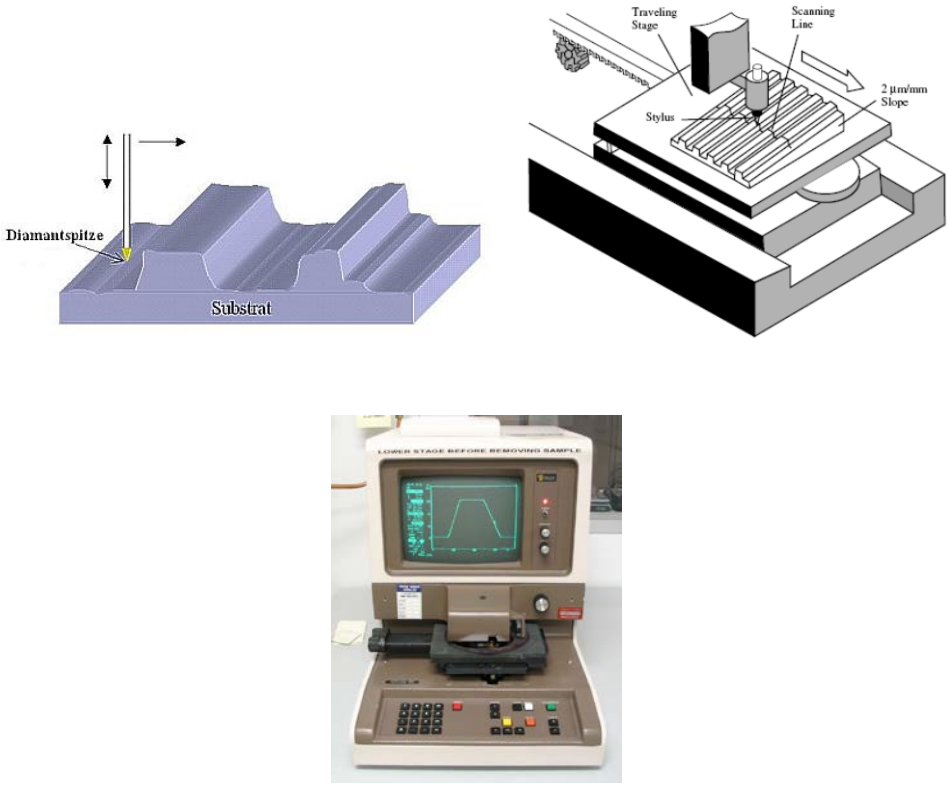


Fig. 16.8. Schematic drawing and photograph of the surface profilometer.

16.2.2.2. Atomic Force Microscopy

Atomic force microscopy (AFM) is a method for measuring surface properties and/or profiles with atomic-scale topographical definition. In this technique, a sharp tip built at the end of a soft cantilever arm is vibrated perpendicular to the surface at close to the resonant frequency of the cantilever–tip mass as the probe tip traverses laterally across the feature to be characterized. The tip is in atomically close proximity to the surface, so a van der Waals electrostatic force is created between them. This force, which has a strong dependence on the gap between the tip and surface topography, modifies the resonant frequency of the system. The changes in resonance are monitored by an interferometric detection technique that provides a corresponding displacement signal, resulting in a direct measure of the atomic-scale surface topography. A schematic of an AFM system is shown in Fig. 16.9. Figure 16.10 shows a typical AFM scan of a surface structure. One disadvantage of this technique compared to conventional methods is its low throughput.

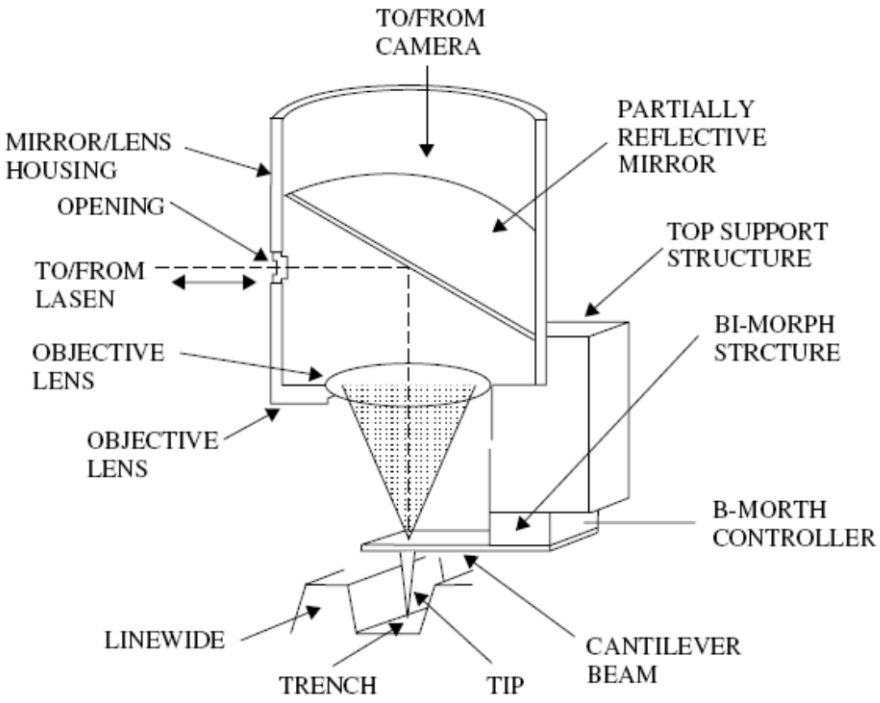


Fig.16.9. Schematic of atomic force microscopy system [9].

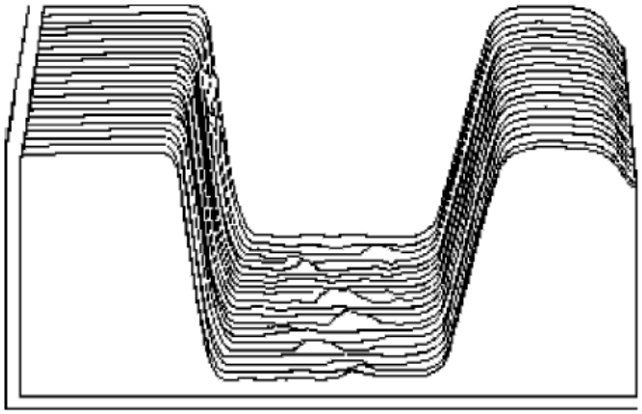


Fig. 16.10. Typical AFM image of a surface feature (in this case, a trench) [9].

16.2.2.3. Scanning Electron Microscopy

Scanning electron microscopy (SEM) is a key technique for assessing minimum feature size in semiconductor manufacturing. The minimum feature size is often expressed in terms of the critical dimension (CD) or minimum *linewidth* that can be resolved by the photolithography system. The decrease in linewidths toward the scale of fractions of a micrometer has rendered conventional optical microscopes nearly obsolete. However, linewidth measurements based on SEM can overcome the limitations of optical techniques for submicrometer geometry features.

The fine imaging capability of the SEM is due to the fact that the wavelength of electrons is four orders of magnitude less than that of optical systems. At such small wavelengths, diffraction effects are usually negligible and spatial resolution is excellent. Features as small as 100 nm can be readily resolved [13]. The electron beam may be based on thermionic or field emission sources. A schematic of a typical field emission SEM is shown in Fig. 16.11.

As shown in this figure, the electron gun consists of a tip, first anode, and second anode. A voltage is established between the tip and first anode to facilitate field emission from the tip. An accelerating voltage is then applied between the tip and the second anode to accelerate the electrons. The electron beam emitted from the tip passes through the aperture provided at the center of the first anode, is accelerated, and passes through the center aperture of the second anode to the condenser lens. Electron beams are collected by the condenser lens and aggregated into a small spot on the objective lens. Figure 16.12 shows a typical digital photo output of an SEM. The CD of the feature is usually determined by an arbitrary edge criterion. While lateral resolution offers a tremendous benefit, it must be pointed out that SEM still suffers from several disadvantages, including high cost, low throughput (only ~30 wafers per hour), and the destructive nature of the measurement (i.e., wafers must be cleaved to expose the feature to be imaged).

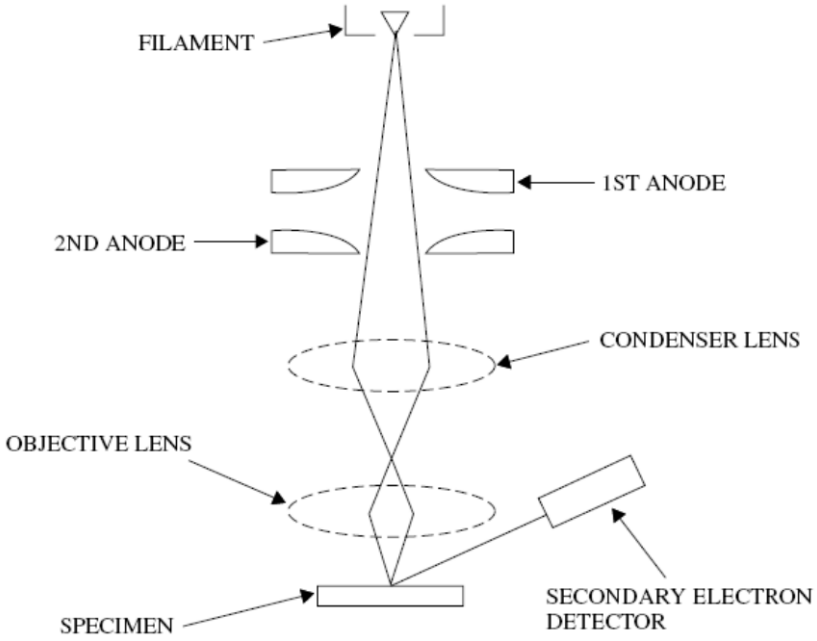


Fig. 16.11. Schematic of field emission SEM optics [9].

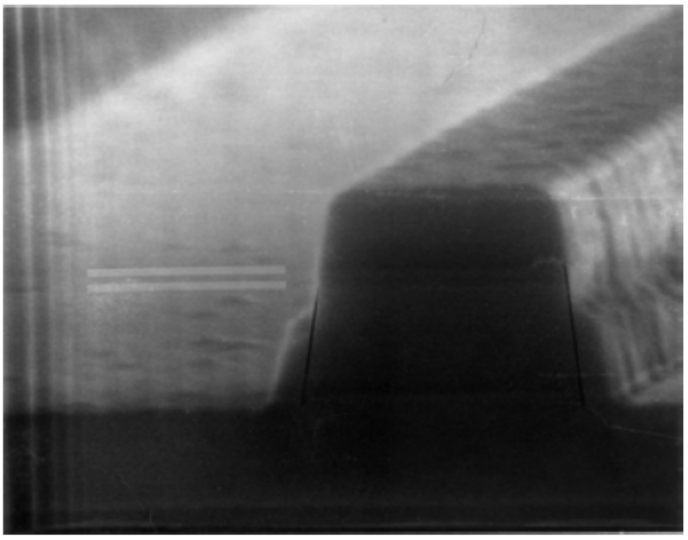


Fig. 16.12. Sample SEM output (the parallel lines are calibration marks).

16.2.2.4. Scatterometry

Scatterometry is another optical measurement technique. It is used for patterned features based on an analysis of the light diffracted (or *scattered*) from a periodic structure such as a grating of photoresist lines. Figure 16.13 shows a schematic of an angle-resolved scatterometer, which measures the intensity of the light diffracted as a function of incident angle and polarization. Scatterometry is used to characterize surface roughness, defects, particle density on the surface, film thickness, or the CD of the periodic structure.

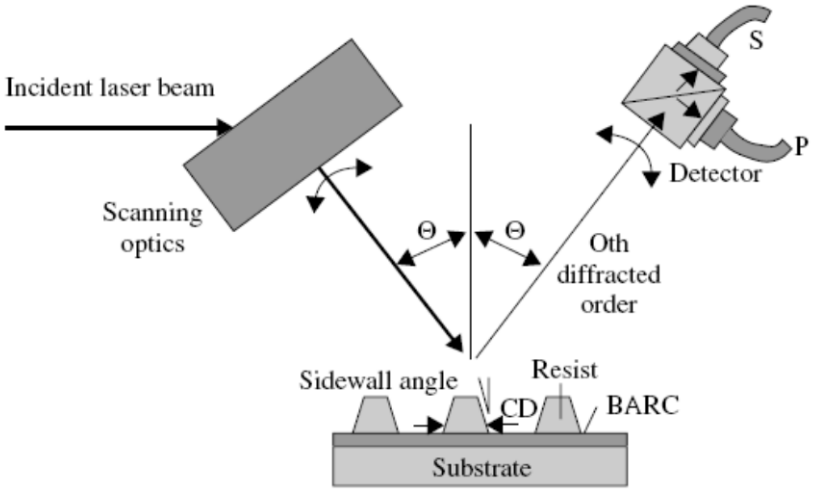


Fig. 16.13. Schematic of a 2θ angle-resolved scatterometer [10].

The most common type of angle-resolved scatterometer is called a "2θ" scatterometer due to the two angles (incident and measurement) associated with the method. An incident laser is focused on a sample and scanned through some range of incident angles (θ_i). The light is scattered by the periodic patterns into distinct diffraction orders at angular locations specified by the grating equation

$$\sin\theta_i + \sin\theta_n = n\lambda / d \tag{16.26}$$

where θ_i is taken to be negative, θ_n is the angular location of the nth diffraction order, λ is the wavelength of the incident light, and d is the spatial period (or pitch) of the periodic structure. Because of the complex interaction between the incident light and the periodic features, the fraction of power diffracted into each order is a function of the dimensions of the structure and thus may be used to characterize them. Capturing diffracted light "signatures" (such as those depicted in Fig. 16.14) is just the first phase of scatterometry. In the subsequent analytical phase, a diffraction model is used to interpret the experimental signatures in terms of key parameters such as CD or film thickness. Doing so requires a library of theoretical signatures for comparison to the measured data. The generation of such a library is accomplished by first specifying nominal film stack dimensions and the expected variation of each parameter to be measured. A computerized diffraction model is then used to produce the library of scatter signatures that encompasses all combinations of these parameters for subsequent analysis.

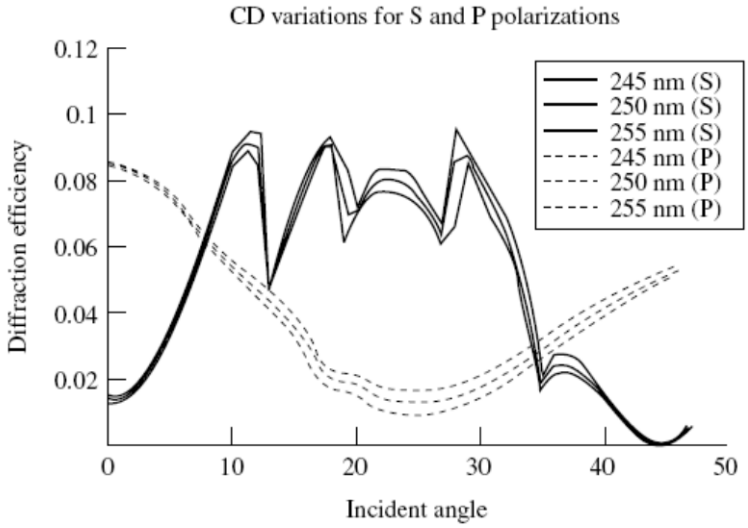


Fig. 16.14. Sample scatterometry signatures for 5-nm photoresist CD variations [10].

16.2.2.5. Electrical Linewidth Measurement

Another CD characterization technique depends on direct-current electrical measurement. The most common test structure for this measurement is shown in Fig. 16.15. In this configuration, two structures are combined to perform resistance measurements. The upper portion, a four-terminal Van der Pauw structure, is used to measure sheet resistance. This structure is designed to account for doping or film thickness variations. The lower structure is a four-terminal crossbridge linear resistor pattern used to determine the average linewidth (W). The length of the line segment (L) between pads 4 and 6 is known. When a known current is applied through pads 3 and 5, the resulting voltage is measured at pads 4 and 6. The average linewidth may then be calculated as the product f the measured sheet resistance and length, divided by the measured resistance (V_{46}/I_{35}). The key advantages of electrical linewidth measurement are resolutions on the order of 1 nm and short cycle time. The main disadvantages are the requirement that the film be conductive and the need for physical contact with the wafer.

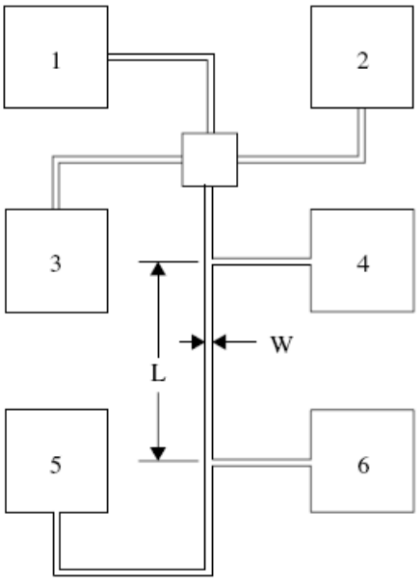


Fig. 16.15. Electrical linewidth measurement test structure [9].

16.2.3. Particle/Defect Inspection

Contamination is a major concern in semiconductor manufacturing, and billions of dollars are spent annually by manufacturers in order to reduce it. Contamination often takes the form of particles that can appear on the surface of wafers and cause defects in devices or circuits. The fraction of the product that is sensitive to particles depends in part on the particle size. A general rule of thumb is that particles as small as one-tenth the size of a structure can cause the structure to fail. With the industry currently immersed in manufacturing devices with submicrometer features, even nanometer-scale particles are of great concern. Inspection and characterization of particles are therefore critical.

16.2.3.1. Cleanroom Air Monitoring

One method of controlling particulate contamination is performing manufacturing operations in a *cleanroom* environment, such as the one schematically depicted in Fig. 16.16. Air enters the cleanroom through high-efficiency particulate air (HEPA) or ultra-high-efficiency particulate air (ULPA) filters. The air is forced o flow laminarly (as opposed to turbulently) so that lateral dispersion of contaminants generated in the room is minimized. Cleanrooms are categorized by their “class,” which quantifies the number of particles of a given size per cubic foot of air. Various aspects of cleanroom performance affect product quality, and as feature sizes continue to decrease, cleanroom specifications are likewise becoming progressively tighter.

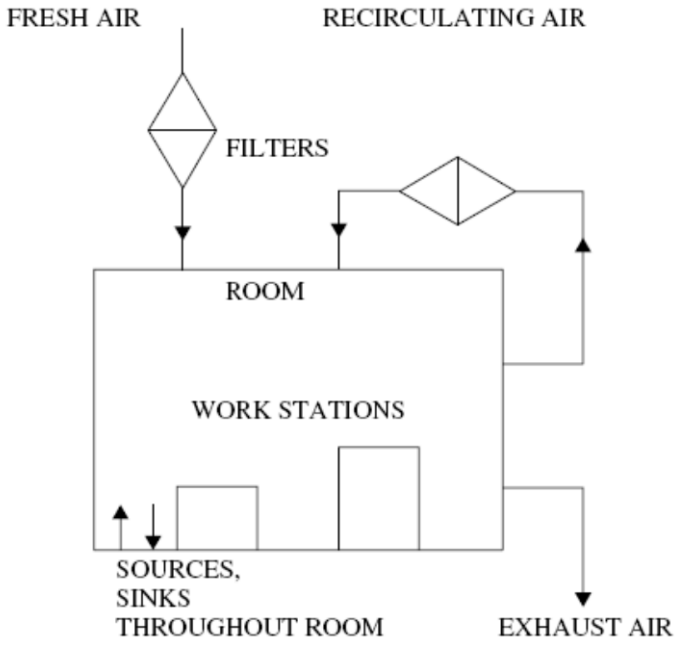


Fig. 16.16. Cleanroom schematic [9].

Despite the use of cleanrooms, semiconductor fabrication processes, as well as manufacturing personnel themselves, still generate materials that can contaminate products. Such contamination may originate from process gases and vapors, process liquids, processes that break up bulk material (such as sputtering), deposition processes, metallic impurities, wafer handling, or tool wear, to name just a few. The usual methods for quantitatively determining cleanroom air quality involve sampling via optical particle counters and sampling onto “witness plates” that are later read by surface particle counters.

In the latter approach, a preinspected clean silicon wafer is placed in a location to be monitored. After a fixed time period, the plate is removed and reinspected. The particles per unit area added to the plate are counted. Surface particle counters can inspect an entire plate within minutes with nearly complete detection of particles of sizes a low as fractions of micrometers.

For gases, liquids, and many types of surfaces, optical particle counters are used. Using these devices, particles are illuminated as they pass through a focused laser beam (see Fig. 16.17). The light scattered from the particles is then measured and correlated with the number of particles present. The amount of light scattered into the sensing element will depend on the light (intensity, wavelength, polarization), the characteristics of the particles (size, shape, orientation, refractive index), and the measurement geometry (position and solid angle subtended by the optics with respect to the beam and the particle). In addition to cleanroom monitoring, this technique is also used for *in situ particle monitoring* (ISPM) inside of processing equipment that produces particles, such as ion implantation or sputtering equipment.

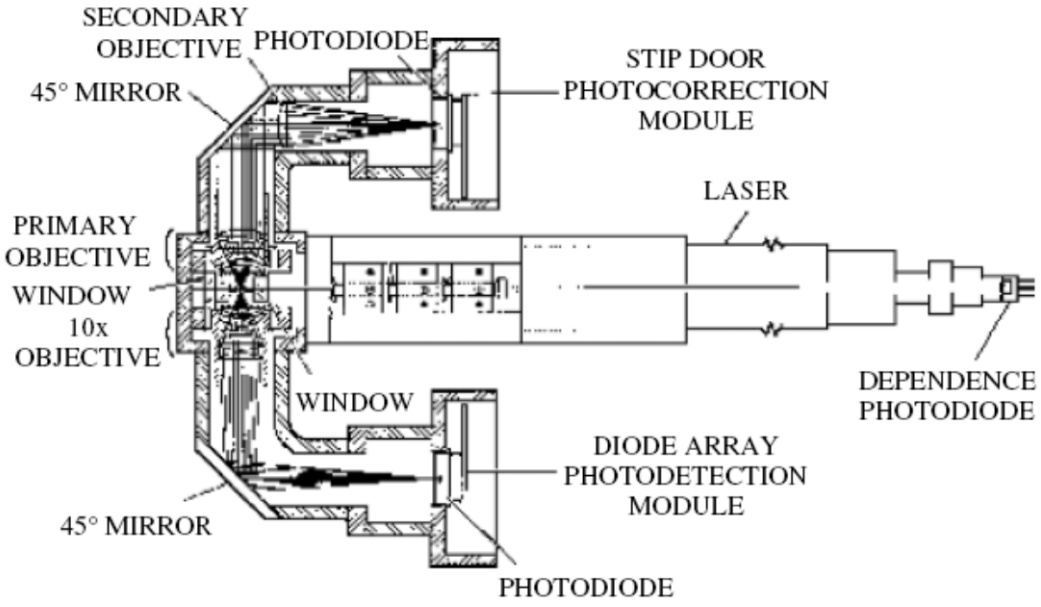


Fig. 16.17. Optical particle counter [9].

16.2.3.2. Product Monitoring

In addition to monitoring contamination in the ambient environment, it is perhaps more crucial to monitor particles that actually wind up on the wafer surface, since these are the particles that can cause circuit defects. Experience has shown that most processing-related defects tend to occur in a few layers of the complete process [11]. For CMOS processes, for example, defects in the gate oxide and interconnect layers represent the vast majority of all defects.

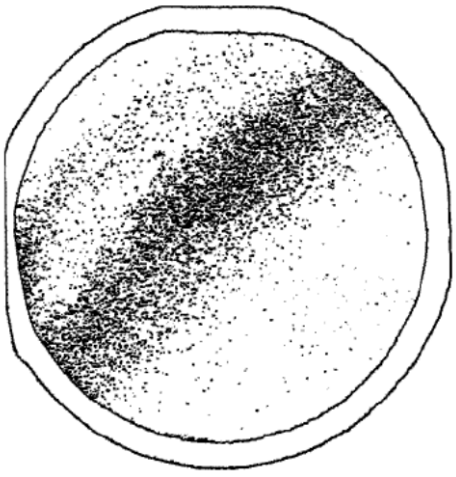


Fig. 16.18. Sample surfscan [11].

To control the formation of such defects, special *inline monitoring* techniques are required. These techniques involve inspection of product wafers at various stages in the process. Two common approaches for local defects are “surfscan” and image evaluation. The surfscan technique uses scattered laser light and analyzes reflections to count the particles on the wafer surface (see Fig.

16.18). Surfscan is usually applied to unpatterned wafers. Image evaluation techniques, on the other hand, make use of automated inspection equipment to check the occurrence of local defects on patterned wafers at several critical points in the manufacturing process.

Generic particle counts are useful, but limited. In order to assess the impact of the presence of defects caused by particles, specially designed test structures are used. These structures, also known as *process control monitors* (PCMs), include single transistors, single lines of conducting material, MOS capacitors, via chains, and interconnect monitors. Product wafers typically contain several PCMs distributed across the surface, either in die sites or in the scribe lines between die (see Fig. 16.19).

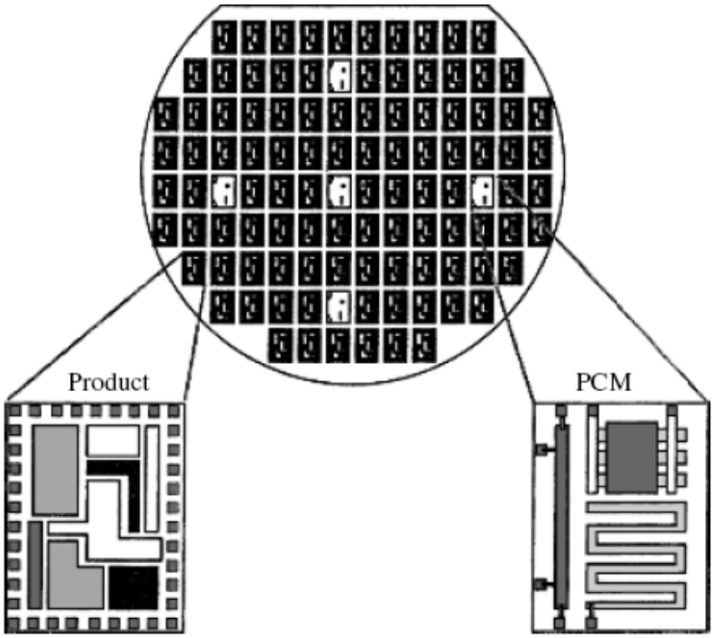


Fig. 16.19. Configuration of products and PCMs on a typical wafer [11].

Process quality can be checked at various stages of manufacturing through inline measurements on PCM structures. Three typical interconnect test structures are shown in Fig. 16.20. Using such test structures, measurements are performed to assess the presence of defects, which can be inferred by the presence of short circuits or open circuits using simple resistance measurements. For example, the meander structure facilitates the detection of open circuits through increased end-to-end resistance of the meander. The double-comb structure can likewise be used to detect shorts (short circuits), since any extra conducting material bridging the two combs will reduce the resistance between combs significantly. The comb–meander–comb structure combines the capabilities of the other two structures and permits the detection of both shorts and opens. Various combinations of widths of lines and spaces in these test structures allow the collection of statistics on defects of various sizes.

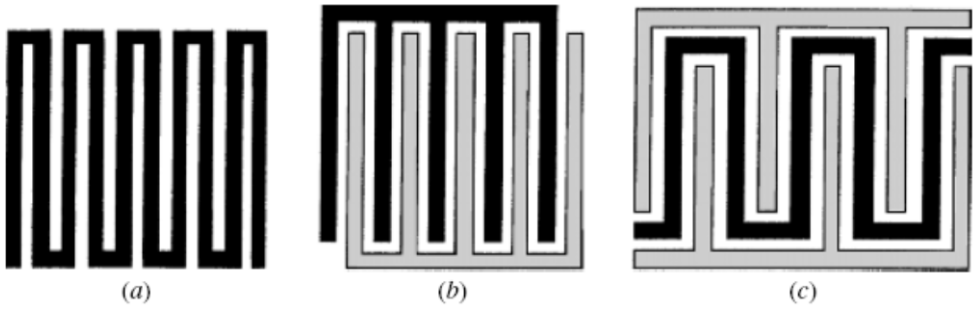


Fig. 16.20. Basic test structures for interconnect layers: (*a*) meander structure; (*b*) double-comb structure; (*c*) comb–meander–comb structure [11].

16.2.4. Electrical Testing

In the preceding section, the concept of test structures for process monitoring was introduced. Although this introduction was presented in the context of particle and defect monitoring, it should not be construed that this is the only use of test structures. In fact, electrical measurements performed on test structures are a major mechanism for assessing *yield* and other indicators of product performance as well. Such measurements are performed on an inline basis and also at the conclusion of the fabrication process. In addition, electrical testing of the final product is crucial to ensure quality.

16.2.4.1. Test Structures

Figures 16.20–16.26 are examples of electrical test structures used for process monitoring. However, these by no means represent a comprehensive set, as dozens of possible structures exist for monitoring hundreds of process variables.

Figure 16.21 shows a high-density bipolar transistor chain used to monitor the leakage current between transistor terminals (emitter–base leakage, emitter–collector leakage, etc.). The emitters and bases are wired into parallel chains. Collectors are contacted via the substrate, which eliminates metal short interference in the emitter–collector leakage test. The collectors are also wired to the second level of metal to test collector isolation leakage. Transistor chains can also be used to monitor base–base shorts, as shown in Fig. 16.22. In this example, shorts due to polysilicon bridging can be detected by forming the polysilicon bases on field oxide to eliminate the possibility of shorts through the substrate. It is also important to monitor transistor contacts for open circuits. Series-type chains similar to the meander structure in Fig. 16.20 can be used for this purpose by connecting the contacts for the various transistor terminals. Figure 3.23 shows an example of a collector contact chain. Note that although the structures depicted in Figs 16.21–16.23 were designed for bipolar circuits, analogous structures can be fabricated to evaluate MOS circuits by wiring up chains connecting their source, drain, and gate terminals in similar chains.

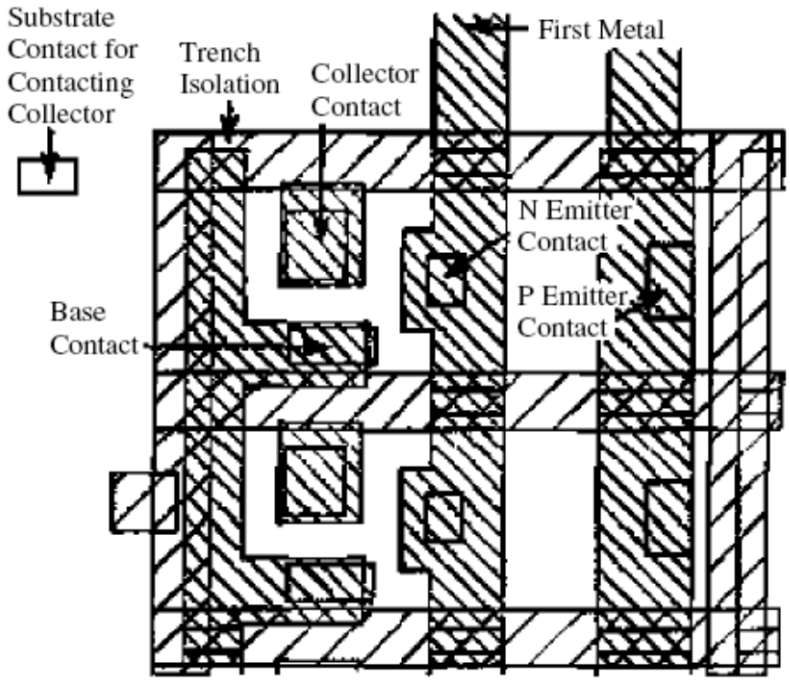


Fig. 16.21. Bipolar transistor chain [9].

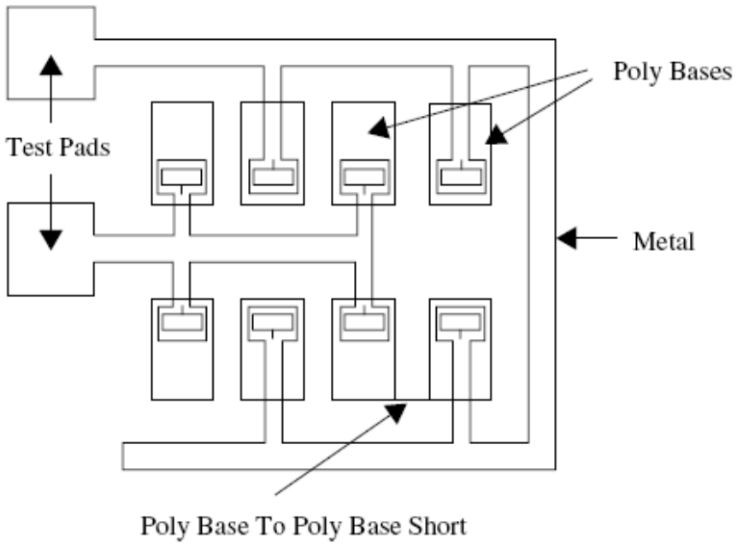


Fig. 16.22. Polysilicon base to polysilicon base short chain [9].

Figure 16.24 is a typical example of a via chain structure used to test connectivity between metal layers. This chain also includes a first-level metal stripe running parallel to the chain as a mask misalignment monitor. An adjacent metal stripe runs on every level, but never along the full length of the chain. They instead appear at certain sections, alternating with each other.

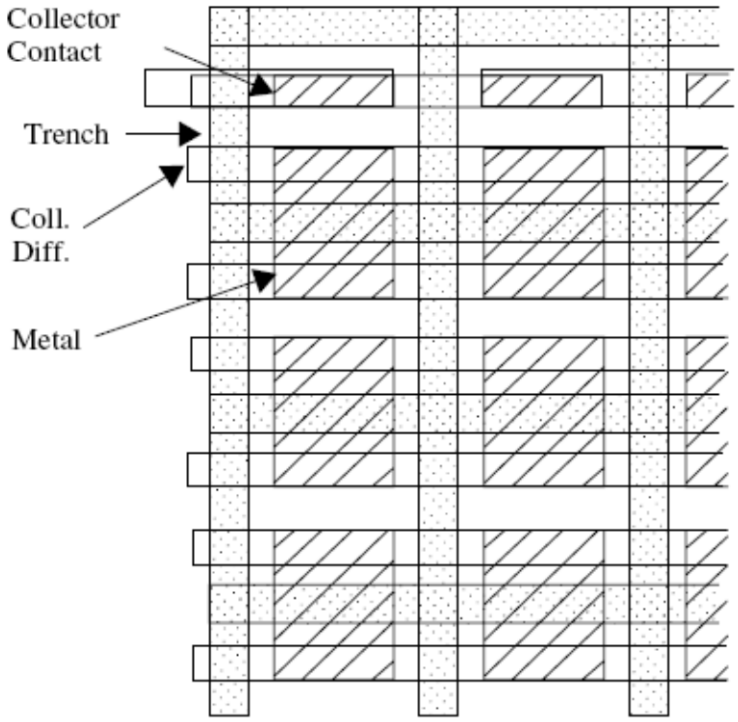


Fig. 16.23. Collector contact chain [9].

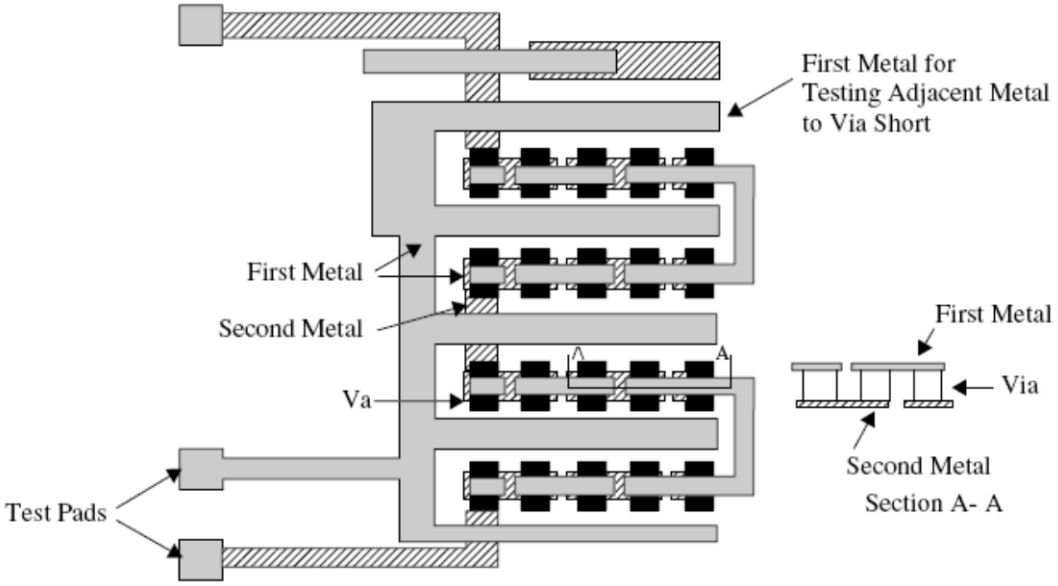


Fig. 16.24. Via chain [9].

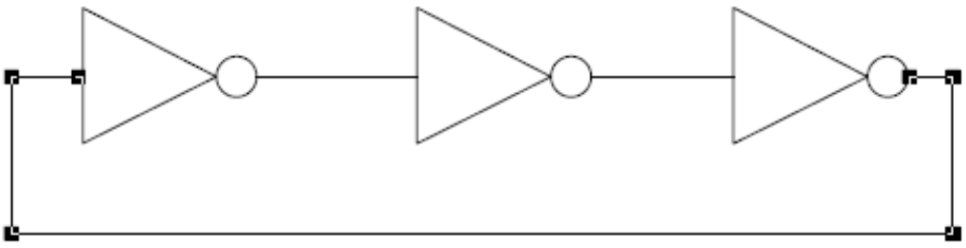


Fig. 16.25. Ring oscillator.

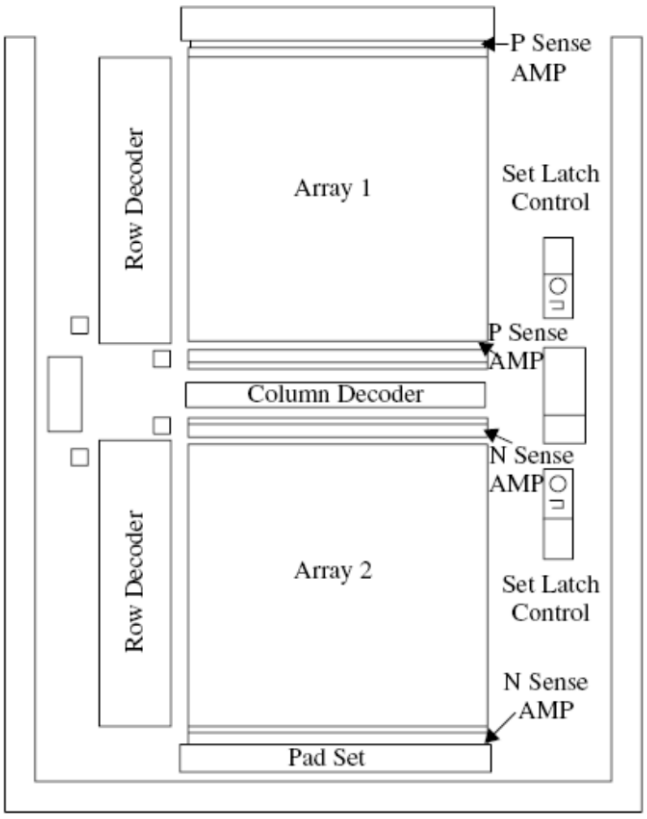


Fig. 16.26. Array diagnostic monitor [9].

In addition to defect monitoring, test structures are also used to assess functional characteristics of the semiconductor devices and their dependence on processing conditions. These can be individual devices or simple subcircuits. A common example of such a structure is a *ring oscillator*, which is used to measure speed and capacitive loading effects. A ring oscillator is essentially a chain of inverters (see Fig. 16.25). It is formed by connecting an odd number of inverters in a loop. In general, a ring with N inverters will oscillate with a period of $2N\tau_p$ and a frequency of $1/2N\tau_p$, where τ_p is the propagation delay through a single inverter. Inverter chains can also be used to monitor transistor current gain or voltage drops across transistors [13]. An example of a more elaborate functional test structure is the array diagnostic monitor (ADM) shown in block diagram form in Fig. 16.26. The ADM, which is used to assess CMOS DRAM circuits, has DC and AC diagnostic capabilities. It is essentially a simplified, yet fully functional duplicate version of a memory array. ADM testing allows for rapid process feedback and ultimately translates into accelerated process improvement.

16.2.4.2. C-V Measurements

Capacity - voltage measurement methods are extremely varied. One category includes measurements at quasi-static equilibrium conditions. In this case in addition to the bias voltage the

changing of experiment parameters such as temperature or sample frequency test signal can be applied. Such information as the concentration of ionized impurities or impurity concentration profile in the bulk semiconductor can be obtained from the measurements. In addition the method allows to determine the characteristics of the interface, such as barrier height and density of surface states. The second category includes the measurement of transient processes in non-equilibrium conditions caused by changing in the state of the charge over time through the carrier generation and recombination processes. In this type of experiments the stationary state is exitated by voltage bias pulse or optically. Such measurements can be used to study the energy states at the interface or in the bandgap of semiconductors. One of the most important methods in this category is deep levels relaxation spectroscopy (DLTS).

The specific parameters and characteristics as well as methods of their determination from C-V measurements depend on the type and characteristics of the model barrier structures under investigation.

Schottky barrier.

Contact with metal-semiconductor with Schottky barrier can be used as a discrete device, part of the more complex device, or as a test structure. The general physical model of the metal-semiconductor Schottky barrier is shown in Fig. 16.27. It takes into account the intermediate layer at the metal-semiconductor boundary and surface electronic states (Fig. 16.27a), as well as a model of tight contact (Fig. 16.27b).

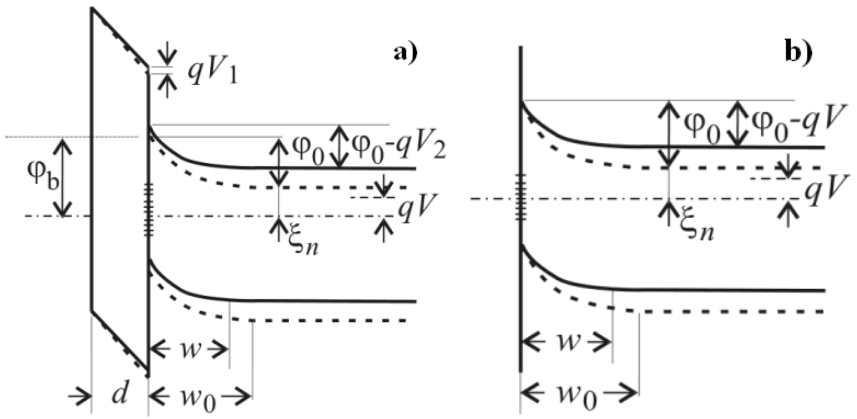


Fig. 16.27. General physical model of the metal-semiconductor Schottky barrier with (a) and without (b) intermediate layer.

The main parameters of this model is the height of the potential barrier (φ_b), the thickness of the space charge region (SCR) of semiconductor (w), as well as the distribution of the applied to the

contact voltage (V) between the transition layer (V_1) and SCR (V_2), which is determined by the parameters of all regions.

It can be shown that in the absence of the intermediate layer (tight contact) the capacity of the metal-semiconductor Schottky barrier with equal

$$C(V) = \frac{dQ}{dV} = \frac{\varepsilon_0 \varepsilon_s S}{w(V)} \tag{16.27}$$

where S is the contact area. In the case of uniform distribution of the impurities w is described by

$$w(V) = \sqrt{\frac{2\varepsilon_0 \varepsilon_s (\varphi_0 - qV)}{q^2 n_0}} \tag{16.28}$$

and

$$C(V) = S\sqrt{\frac{\varepsilon_0 \varepsilon_s q^2 n_0}{2(\varphi_0 - qV)}}. \tag{16.29}$$

As it follows from equation (16.29) the dependence

$$(\frac{S}{C})^2(V) = \frac{2}{\varepsilon_0 \varepsilon_s q n_0}(\frac{\varphi_0}{q} - V). \tag{16.30}$$

is the straight line (Fig. 16.28). The slope of the line is determined by the concentration of impurities n_0, and cutoff voltage on the voltage axis φ_0/q.

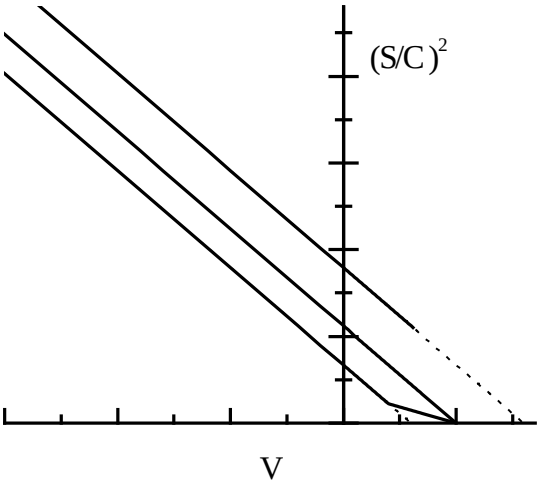


Fig. 16.28. $(S/C)^2 - f(V)$ dependencies of tight metal-semiconductor Schottky barrier with the uniform distribution of impurities.

This is one of the main methods of the Schottky barrier height determination. However, it should be noted that it is applicable only in the case of tight contact with the uniform distribution of impurities $N_d=n_0$. The ideality of current-voltage characteristic points out on possibility to neglect the intermediate layer.

In the case of non-uniform distribution of impurities the expression for the capacitance of Schottky diode is also described by (16.27), but the dependence $(S/C)^2 = f(V)$ is not straight line, and therefore φ_0/q can not be determined. Tangent to the curve at each point is determined by the concentration of ionized impurities in the semiconductor at $x=w$. This means that it is possible to determine the coordinate dependence of impurities concentration in the parametric form from experimental C-V curve: you can get depending on the experimental dependence of the concentration of impurities from the coordinates:

$$N_d(x) = \frac{2}{q\varepsilon_0\varepsilon_s S^2}\frac{dV}{d(1/C^2)}. \qquad (16.31)$$

This method is widely used as the non-destructive method for investigation of doping profile in surface region of semiconductor.

Case stepped distribution of impurities. In practice, there may be cases where the concentration of impurities in the thin layer of surface region is different from volume one. This layer may be formed, for example, due to impurity segregation during oxidation of the semiconductor surface, resulting in solid phase interaction at the metal-semiconductor interface during annealing, or may be created specifically to produce certain specified electrical characteristics of the contact. To simplify the analysis it is usually supposed that the impurity concentration in thin layer of surface region is also uniformly distributed, i.e. the stepped distribution of impurities in semiconductors is considered.

$$\rho(x) = qn_1 = qKn_0 \quad \text{in the range} \quad 0 \le x < l. \text{при} \qquad (16.32)$$

$$\rho(x) = qn_0 \quad \text{in the range} \quad l < x < w \qquad (16.33)$$

It can be shown that in the absence of the intermediate layer (tight contact) the capacity of the metal-semiconductor Schottky barrier is equal

$$C(V) = \frac{\varepsilon_0\varepsilon_s S}{w(V)}, \qquad (16.34)$$

where S is the contact area, and w is described by

$$w(V) = \sqrt{\frac{2\varepsilon_0\varepsilon_s(\varphi_0 - qV)}{q^2 n_0}} \qquad (16.35)$$

in the case of a uniform, or

$$w = \{w_p^2 + l^2(1-K)]^{1/2} \qquad (16.36)$$

in the case of the stepped distribution of impurities. Then it is possible to show

$$(S/C)^2 = \frac{2}{\varepsilon_0\varepsilon_s qn_0}\left(\frac{\varphi_0^*}{q} - V\right) \qquad (16.37)$$

All dependencies of $(S/C)^2$ on V at negative and small forward voltage is the straifht lines that cross the V axis ($(S/C)^2$ =0) at the points $V=\varphi^*_0/q$. The slope of the lines is the same and is determined by the concentration of impurities in semiconductors. In the case of uniform distribution of impurities in semiconductors $\varphi^*_0 = \varphi_0$.

The indication that the concentration of impurities in the surface layer of the semiconductor is different from the bulk can be 'not ideality' of current-voltage characteristic.

In case of contact with an intermediate layer the applied voltage (V) is divided between the intermediate layer (V_1) and the space charge region (V_2), i.e.

$$V = V_1 + V_2 . \tag{16.38}$$

Accordingly, the high-frequency capacitance of the contact can be represented as a series connection of the intermediate layer capacity (C_1) and the space charge region capacity (C_2):

$$C = \frac{C_1 \times C_2}{C_1 + C_2} \tag{16.39}$$

where

$$C_1 = \frac{\varepsilon_0 \varepsilon_1 S}{d}, \quad C_2 = \frac{\varepsilon_0 \varepsilon_2 S}{w}. \tag{16.40}$$

The expression for the thickness of the SCR in this case has the same type as in the case of tight contact if instead of V substitute V_2. The relationship (16.39) can be written in the form

$$C = \frac{S\sqrt{q\varepsilon_0\varepsilon_2 n_0 / 2}}{\sqrt{(\frac{\sqrt{q}a_n}{2} + \frac{\sqrt{\varphi_0}}{q})^2 - (V - V_1^i)}} \tag{16.41}$$

where

$$a_n = \frac{d}{\varepsilon_0 \varepsilon_1}(2n_0\varepsilon_0\varepsilon_2)^{1/2} \tag{16.42}$$

$$V_1^i = q\frac{d}{\varepsilon_0 \varepsilon_1}(m_i - m_i^0) \tag{16.43}$$

V_1^i is the potential associated with the recharge of surface states, m_i and m_{i0} are the electron concentration on the surface states (per unit area) in apposition with the bias voltage and at V=0, respectively.

As can it be seen from (16.41)

$$(\frac{S}{C})^2 = \frac{2}{q\varepsilon_0\varepsilon_2 n_0}\ [(\frac{\sqrt{q}a_n}{2} + \sqrt{\frac{\varphi_0}{q}})^2 - (V - V_1^i)]. \tag{16.44}$$

Since, in general case $V_1^i = f(V)$ the dependence $(S/C)^2$ can not be straight line even in the case of the uniform distribution of donors.

In the absence of recharge of surface states at changing the bias voltage (V_1^i=0) the dependence $(S/C)^2=f(V)$ is the straight line. Its slope is determined by the impurity concentration in the semiconductor, and the cut-off on the voltage axis at $(S/C)^2$=0 is equal to

$$(\frac{\sqrt{q}a_n}{2}+\sqrt{\frac{\varphi_0}{q}})^2.$$

At d=0 the first term in brackets is also zero, since a_n=0 and the cutoff voltage is φ_0/q as it should be. Thus the dependence $(S/C)^2=f(V)$ can be straight line not only in the case of tight contact with the uniform distribution of impurities, accordingly the cutoff voltage of the line does not necessarily give value φ_0/q. It is obviously that for the contact with the intermediate layer the barrier height determined from the capacity-voltage characteristics will be more than φ_b-ξ_n.

P-n junction.

Capacity of *p-n*-junction can be represented in the form of parallel connection of the barrier and diffusion capacities. Barrier capacity of *p-n*-junction capacitance as the capacity of Schottky barrier is due to overcharging the space charge region and is described by (16.27). The thickness of the SCR w depends on the distribution of impurities in *p-n*-junction and the applied voltage. The real distribution of impurities in *p-n*-junction is usually extrapolated by linear (smooth junction) or stepped (sharp junction) ones.

In the smooth *p-n*-junction the concentration of impurities in the SCR varies linearly:

$$w=[\frac{12\varepsilon_0\varepsilon_s}{qb}(\frac{\varphi_0}{q}-V)]^{1/2} \qquad (16.45)$$

where $b=dN/dx$ is the gradient of the impurities concentration in *p-n*-junction.

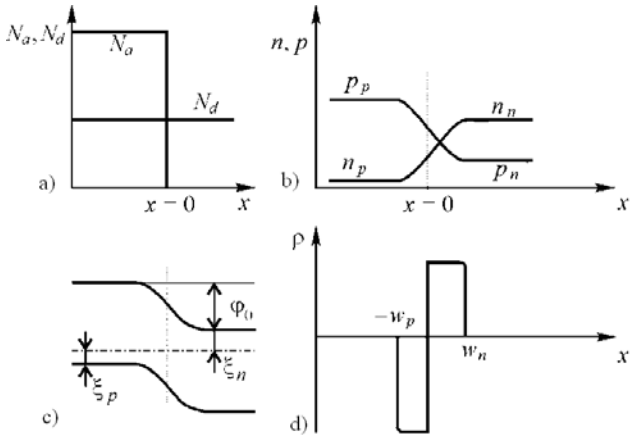


Fig. 16.29. The impurities distribution of in sharp *p-n*-junction (a), distribution of charge carriers (b), energy band diagram of p-n-junction (c), distribution of space charge in the depletion approximation of SCR (d).

In the case of sharp *p*-*n*-junction with the uniform distribution of impurities in *p*- and *n*-regions (Fig. 16.29) thickness of *p*-*n*-junction

$$w = [\frac{\varepsilon_0 \varepsilon_s (N_d + N_a)}{qN_a N_d}(\frac{\varphi_0}{q} - V)]^{1/2} \qquad (16.46)$$

Thus, the barrier capacitance of *p*-*n*-junction is proportional to $(\varphi_0 / q - V)^{-1/m}$, where *m* is equal to 2 or 3 for sharp and smooth *p*-*n*-junction respectively.

The diffusion capacitance is associated with the injection of minority carriers. Its value is proportional to the current, and therefore exponentially increases with the growth of forward voltage. At the forward bias the diffusion capacity increases with the voltage much faster than the barrier one and is dominant. At reverse voltage there is no injection and the capacity of *p*-*n*-junction is the barrier one. Namely this barrier capacity, i.e. the capacity of *p*-*n*-junction at reverse bias is used to control the parameters of the structures.

As can it be seen from Eqs. (16.27), (16.45) and (16.46), the $(S/C)^2=f(V)$ dependences for sharp, and smooth *p*-*n*-junctions are straight lines that when extrapolating cross voltage axis at $V=\varphi_0/q$. The slope of the line in the case of the smooth *p*-*n*-junction is determined by the concentration gradient of impurities *b* and in the case of the sharp *p*-*n*-junction by the conversion ratio $N_dN_a/(N_d+N_a)$.

In the case of the sharp asymmetrical *p*-*n*-junction when the concentration of impurities in one region is much higher than in other, for example $N_a>>N_d$, the thickness of the SCR *p*-*n*-junction

$$w = [\frac{\varepsilon_0 \varepsilon_s}{qN_d}(\frac{\varphi_0}{q} - V)]^{1/2} \qquad (16.47)$$

is determined by the concentration of impurities in low doped region. Since the thickness of the SCR *w* is the sum of the thicknesses of the SCR of *p*-and *n*-regions

$$w = w_p + w_n \qquad (16.48)$$

The condition of neutrality is expressed as

$$N_a w_p = N_d w_n \qquad (16.49)$$

The values of w_n and w_p can be obtained in the form

$$w_p = \frac{N_d}{N_a + N_d} w, \qquad w_n = \frac{N_a}{N_a + N_d} w. \qquad (16.50)$$

As it can be seen from (16.50), in the case of a sharp asymmetrical *p*-*n*-junction the thickness of the SCR of the low doped region is much more higher then heavily doped one and practically equal to the full SCR thickness. This fact is true and for the asymmetrical *p*-*n*-junction case with non-uniform distribution of impurities. Then, from the capacity-voltage characteristics of

asymmetric *p*-*n*-junction, for example sharp p^+-*n*-junction it is possible to get the distribution of impurities in low doped region according to (16.31).

Engineering of C-V measurements.

In general, the equivalent circuit of a Schottky barrier or the *p*-*n*-junction is shown in Fig. 16.30.

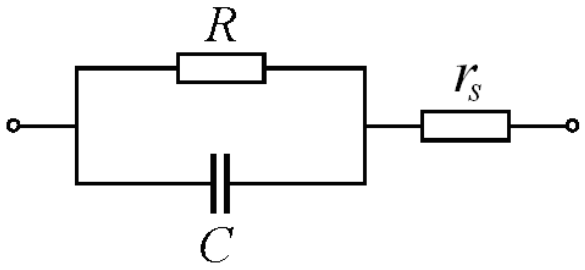


Fig. 16.30. The equivalent circuit of the diode

Here *R* is the differential resistance, *C* is the capacitance, r_s is the series resistance of quasi-neutral region of the semiconductor. In case of *p*-*n*-junction *C* consists of parallel-connected diffusion and barrier capacitances. The barrier capacity is dominant at reverse voltage. To measure the capacity of two-terminal unit the digital *R, L, C* meters are commonly used. The frequency and amplitude of the testing signal is chosen within wide range. The devices can contain block of bias voltage, which is controlled by external signals and interfaces for coupling to computer.

Such devices present the investigated two-terminal unit as series or parallel connection of active R_M and reactive $(iwC_M)^{-1}$ resistivity. This index "*M*" means "measured." It is possible to calculate the *R* and *C* parameters of equivalent circuit with taking into account of series resistance rs in case of parallel equivalent circuit:

$$R = R_M \frac{[(1 - r_s / R_M)^2 + w^2 C_M^2 r_s^2]}{1 - r_s / R_M - w^2 C_M^2 r_s^2} \qquad (16.51)$$

$$C = \frac{C_M}{1 - r_s / R_M - w^2 C_M^2 r_s^2} \qquad (16.52)$$

If the inequalities $r_s << R_M$ and $w^2 C_M^2 r_s^2$ are performed then $R \approx R_M$ and $C \approx C_M$. The first inequality is usually performed at reverse voltage, and the second one at sufficiently low frequency of AC signal.

16.2.4.3. Final Test

Functional testing at the completion of manufacturing is the final arbiter of process quality and yield. The purpose of final testing is to ensure that all products perform to the specifications for which they were designed. For integrated circuits, the test process depends a great deal on whether the chip tested is a logic or memory device. In either case, automated test equipment (ATE) is used to apply a measurement stimulus to the chip and record the results. The major functions of the ATE are input pattern generation, pattern application, and output response detection. A block diagram for a basic ATE is shown in Fig. 16.31.

For logic devices, during each functional test cycle, input vectors are sent through the chip by the ATE in a timed sequence. Output responses are read and compared to expected results. This sequence is repeated for each input pattern. It is often necessary to perform such tests at various supply voltages and operating temperatures to ensure device operation at all potential regimes. The number and sequence of failures in the output signature are indicative of manufacturing process faults.

The test process for memory products is very similar to that used for logic. However, one important variation is the availability of the redundancy technique. For dynamic RAM circuits, a widely used approach is to add a few extra word and/or bit lines that can replace faulty lines in the main array. Replacement of these faulty lines is accomplished by fusing them to redirect a bad word or bit address to a redundant line. Testing the redundant lines requires two passes. During the first pass, the addresses of errors are recorded and stored. As long as the number of faults is less than the number of extra lines, the chip is repairable. Although redundancy adds considerable cost and complexity to testing, the yield benefit achieved more than compensates for this.

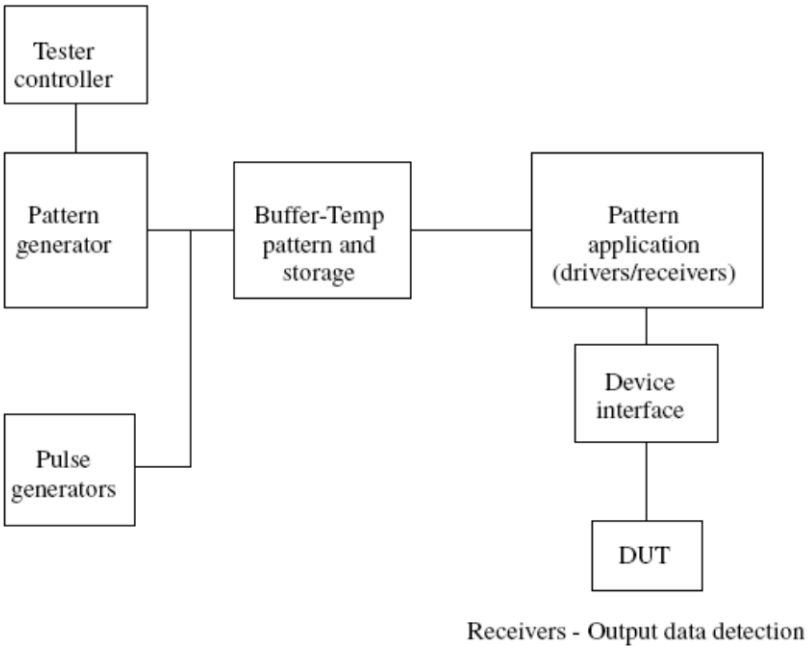


Fig. 16.31. Block diagram of basic test system (DUT = ‘‘device under test’’) [9].

Test results may be expressed in a variety of ways. A couple of examples are shown in Figs. 16.32 and 16.33. Figure 16.32 shows a plot of a two-dimensional plot called a “shmoo” plot for a hypothetical bipolar product. In a shmoo plot, the outlined shaded region is where the device is intended to operate, while the blank area outside represents the failure region. Another typical test output is the cell map shown in Fig. 16.33. Cell maps are very useful in identifying and isolating device failures, particularly in memory arrays. In addition, the patterns generated in the cell map may be compiled, catalogued, and later compared to a library of existing defect types, thereby aiding in the diagnosis of faults.

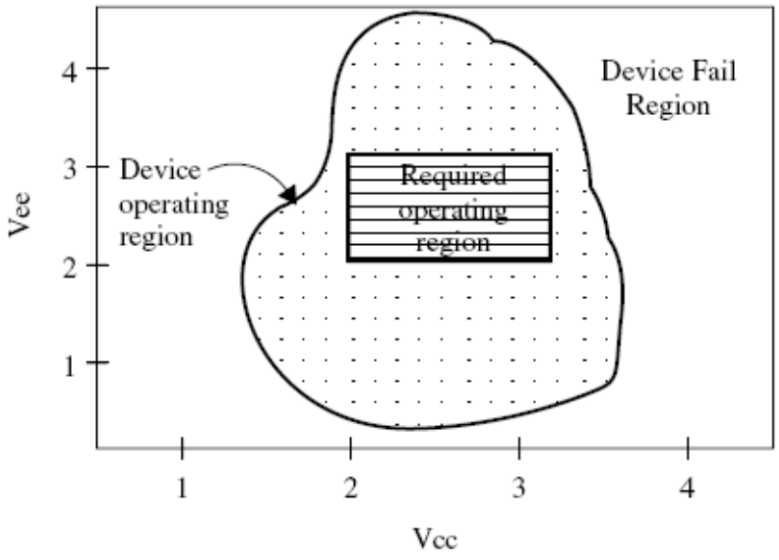


Fig. 16.32. Example of two-dimensional voltage *shmoo* plot for hypothetical bipolar chip [9].

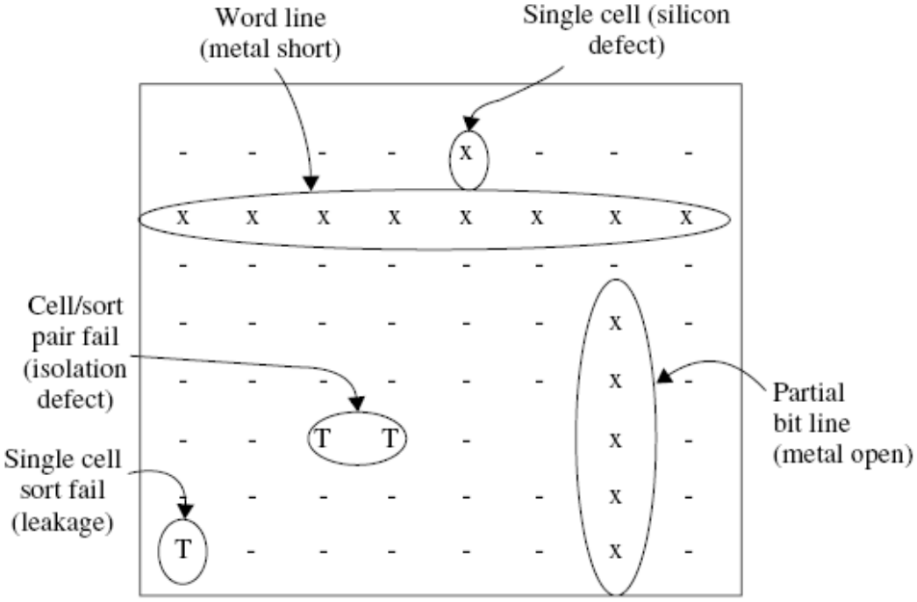


Fig. 16.33. Cell map showing examples of failure patterns and defect types [9].

16.3. TECHNIQUES FOR CHARACTERIZATION AND FAILURE ANALYSIS OF INTEGRATED CIRCUITS

The prototype development of an integrated circuit (IC) comprises the design, processing and testing of an IC such as that shown schematically in Fig. 16.1(*a*). Prototype development time should be kept as short as possible. This implies that the time needed for each step shown in Fig. 1(a) should be carefully considered so that it can be further reduced. As a rule prototype development requires several loops (redesigns) before an error-free and optimally functioning device is obtained, complying with the process technology. To minimize the number of redesigns it is important to increase the amount of information obtained in the testing phase. One possibility is by repairing fatal errors within the IC. Figure 16.34(*a*) shows how this in-circuit repair fits in with the prototype development cycle [12].

Figure 16.34(*b*) shows the testing phase more closely. In external testing the IC is connected to a tester or a measurement set-up which supplies electrical control signals and measures the electrical output signals of the IC. This way of testing normally precedes other testing activities, because it can give a lot of information in a short time. In-circuit testing comprises measurements or excitation at locations within the IC. With in-circuit measurements the response of the IC is measured in its interior while it is driven externally. With in-circuit excitation a local disturbance is generated within the IC and the response of the IC is measured externally. In-circuit testing gives detailed and additional information on the (local) operation characteristics of an IC. This information can be of prime importance to the characterization or failure analysis of an IC. Here characterization means the measurement of a parameter for varying conditions like supply voltage, temperature, etc. Failure analysis is an activity where a malfunction of the IC is tracked down.

It is the purpose of this part to give an overview of in-circuit testing techniques and internal repair. The various techniques are shown in Fig. 16.35. Liquid crystal (LC) is used to measure local temperature differences and light emission locates spots that emit (faint) light. Electron beam testing and electro-optic (EO) sampling are used to measure local electrical signals. Electron beam testing can also be used to display the voltage distribution within the IC.

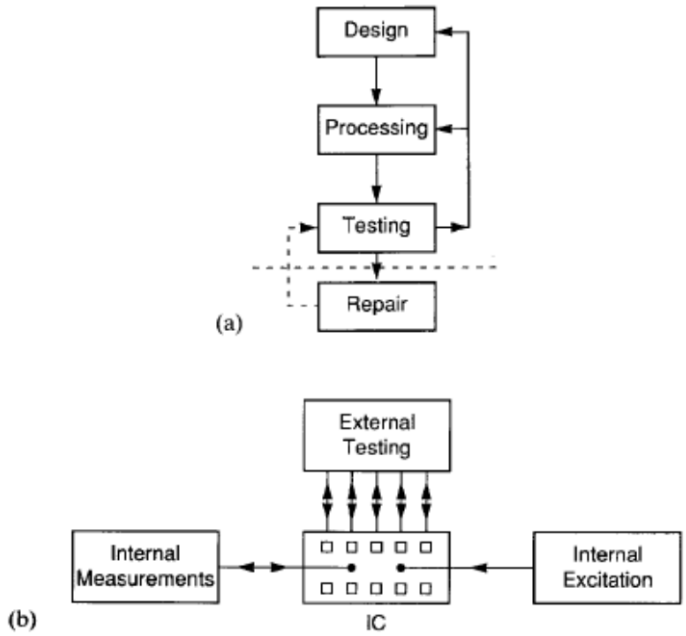


Fig. 16.34. (*a*) Prototype development of ICs [13]. (*b*) External and internal testing.

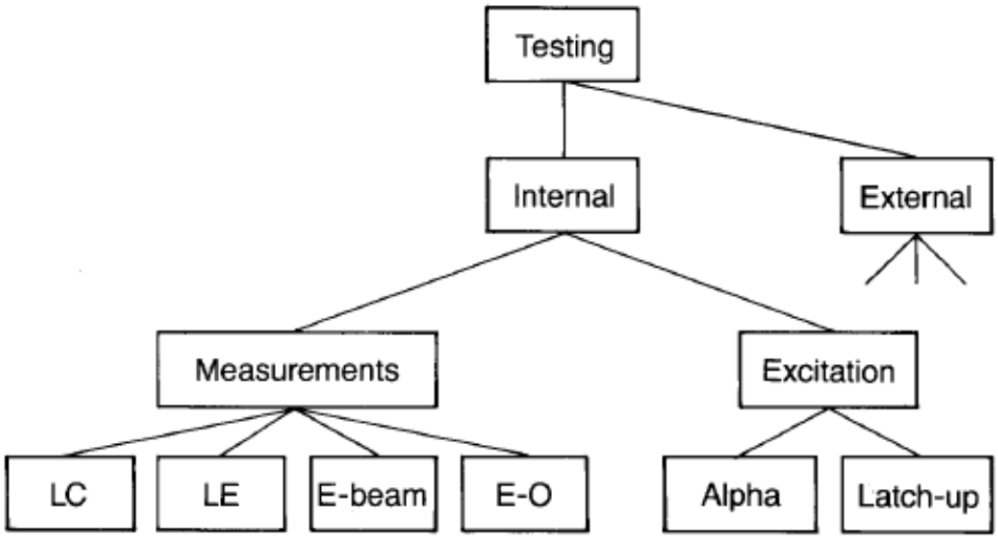


Fig. 16.35. Internal testing methods. LC: liquid crystal; LE: light emission; E-beam: electron beam testing; EO: electro-optic sampling; alpha: errors induced by alpha particles; latch-up: laser-induced latch-up [13].

16.3.1. In-circuit Measurements

In-circuit measurements probe the local value of a physical parameter, which is influenced by the local functioning or malfunctioning of the IC. The way each measurement technique works is described in this section. Examples illustrate the application of in-circuit measurements.

16.3.1.1. Liquid crystal techniques

Failures in ICs often lead to an increase in current, which causes a small rise in temperature. Liquid crystals (LCs) are used to visualize the temperature distribution within the IC [14-20]. Nematic LCs (liquid crystals) with a first-order phase transition at the clearing point T_c are most appropriate for this purpose. Below T_c the LC material is in the nematic (ordered) state, showing optical anisotropy, while above T_c it is isotropic. The difference in optical properties is used to determine whether or not the local temperature is above T_c.

In the experimental set-up, shown schematically in Fig. 16.36, a microscope is used to examine the IC. The latter, in the form of a wafer or a mounted device, is fixed on a heating stage. The whole assembly can be moved by a computer-controlled *x-y* stage. Electrical connections arc provided to drive the IC. In the microscope LC material in the isotropic state appears black, whereas ordered LC appears colorful. This is caused by the effect of the LC on the polarization of light passing through it. The linearly polarized incident beam remains linearly polarized after reflection from the IC if the LC is isotropic; the polarization becomes elliptical if the LC is anisotropic. Here we assume that the polarization of the incident beam is not parallel to the optical axis, i.e. along the preferred direction within the LC. The analyser transmits the polarization component perpendicular to the polarization of the incident beam. This perpendicular component is only present when the reflected light has an elliptical polarization. The image of the IC covered with LC therefore shows black regions corresponding to the sites where the LC is in the isotropic slate. To visualize small local temperature differences the assembly is heated to a bias temperature T_0, close to but below T_c. When a local power dissipation results in a temperature increase dT, with

$$dT > T_c - T_0 \tag{16.53}$$

a black spot appears at that location.

Factors affecting the temperature sensitivity of the set-up are the temperature stability (T_c-T_0), local thermal capacitances and thermal conductivities, the distance between heat source and surface, and the type of LC material. In our case T_c-T_0 can be kept within 0.1 °C over long periods (1 h) and 0.01 °C for short periods (1 s). The LC most frequently used is ROCE 1510 from Hoffmann LaRoche [19], which has a T_c of 48.0 °C and a very narrow phase transition (10^{-5}°C). At room temperature this material is in the solid phase and it melts at 44.5 °C. A few drops of a saturated acetone solution of the LC are put on the IC (a wafer or a packaged IC) [17]. At 20 °C the acetone evaporates in about one minute, leaving behind a thin layer of solid LC material. Then the IC is assembled and heated so that it can be examined through the microscope. When working on wafers electrical contacts are made with a probe card. By moving and positioning the wafer with respect to the probe card, each IC (or die) present on the wafer can be examined.

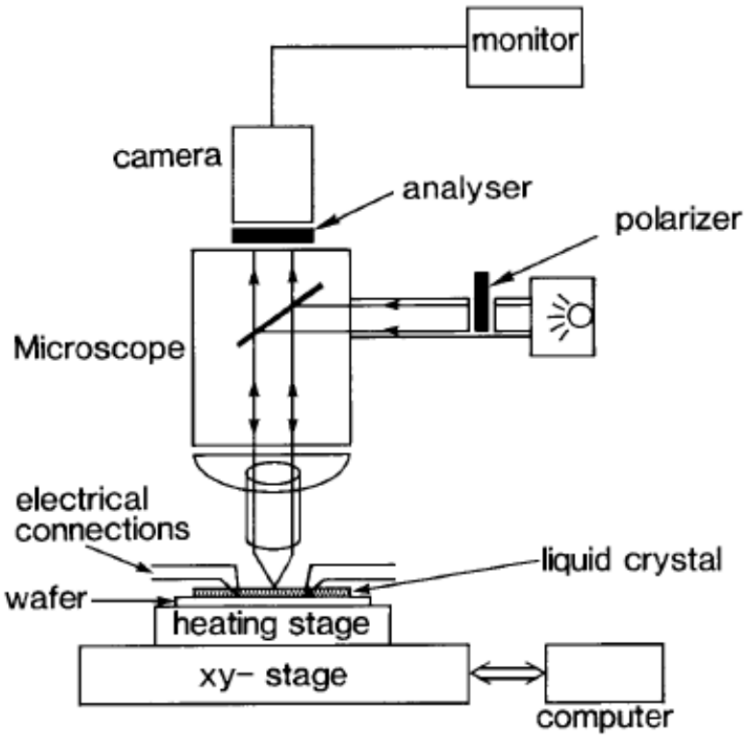


Fig. 16.36. Experimental set-up for determining locations of heat dissipation within ICs using LC [13].

The LC method has already been successfully applied to a large number of ICs. Typical failures found in this way are: failures caused by electrostatic discharge (ESD), interlayer shorts, oxide shorts, latch-up and design failures. An example is shown in Fig. 16.37. The upper micrograph shows a small part of a VLSI circuit containing a black spot. When the supply voltage is decreased, the size of the spot also decreases. This allows an accurate determination of the location of power dissipation. Having found this location, deprocessing (layer-by-layer removal) is performed to establish the cause of power dissipation. The deprocessed IC is shown in Fig. 16.37(*b*). In this SEM micrograph a stripe of material can be seen within the black ellipse. This stripe connects the wide metal track and the dot in the lower right-hand corner, causing a short-circuit.

Within an IC, LCs display not only thermal effects but also electric fields [21]. The electric field exerts a force on the LC molecules and changes the anisotropy direction of the LC. This affects the polarization of the reflected beam, which makes it observable through the microscope. In this case there is no need to have T_0 close to T_c and it is this feature that permits a distinction to be made between thermal and electrical effects. The response time of the LC to changes in the electric field lies in the millisecond range, but even when the period of the *ac* signal is shorter than the response lime of the LC the average direction of anisotropy is changed. For ROCE 1510 electric

fields can be visualized best at 1 kHz. This allows tracking of electrical signals within the IC, a feature that can be helpful for failure analysis.

Improvements of this technique are possible with respect to temperature resolution and spatial resolution. The temperature resolution is determined by the temperature regulation and this can be improved to give a higher sensitivity of this method. Thermal isolation from the environment will then be required. The spatial resolution is determined by the optics and the heat diffusion in the top layer of the IC and the LC. The optical part can be improved by using UV light, but this imposes severe limitations on the class of suitable LC materials, which must be chemically stable and transparent for UV illumination. Improvement of the optics is useful for visualizing smaller geometries, but for most hot spots heat diffusion rather than optics will determine the spatial resolution.

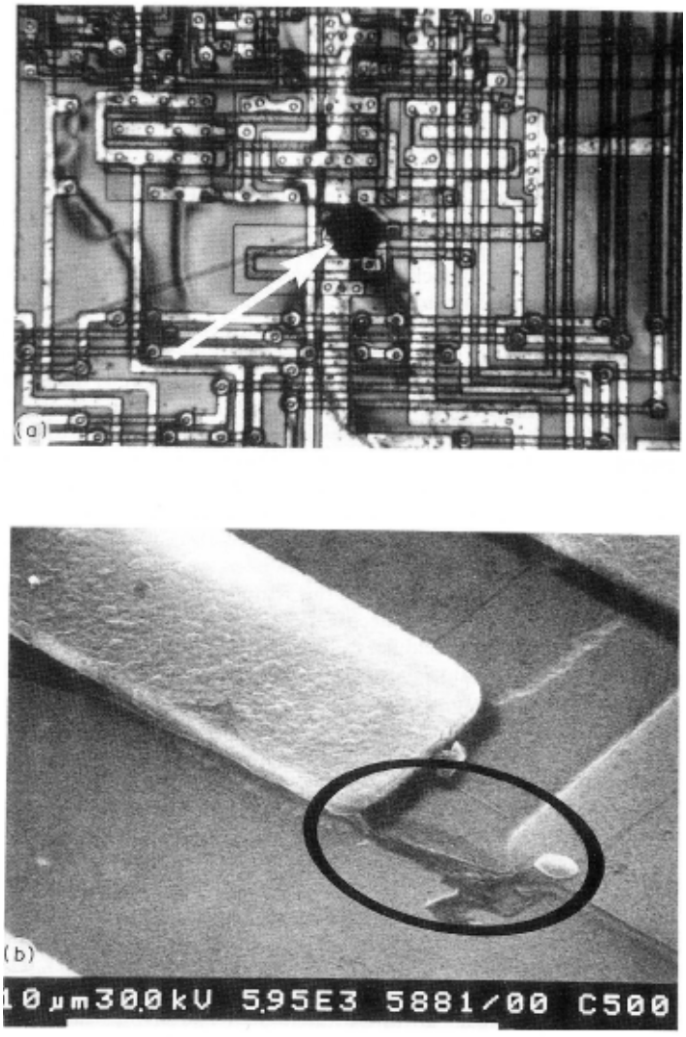


Fig. 16.37. Application of the LC method: (*a*) optical micrograph of an IC containing a hot spot; (*b*) SEM micrograph of the deprocessed IC showing the cause of heat dissipation [13].

Voltage contrast can be improved by choosing different LC materials, where switching speed and orientation properties could be optimized. LC layers on top of ICs can be oriented by

covering them with specially prepared glass plates. However, this requires a long preparation time. *dc* voltage contrast will remain difficult, if not impossible, because of charging of the insulating layer present on most ICs [16].

16.3.1.2. Light emission

When an IC emits light two important types of information can be obtained: (*a*) the location of light emission, (*b*) the nature of the emitted light. The location tells where something is happening whereas the nature (spectrum, time dependence) of the emitted light contains information about the physical mechanism causing light emission.

Light emission can occur when energetic or trapped charge carriers decay to a lower energetic state. The excess energy is obtained from the electric fields applied on or built into the IC. This happens, for example, when electrons are accelerated in the channel of a transistor (hot electrons), when electrons (or holes) traverse a potential barrier or when breakdown occurs (electrons acquire enough energy to generate electron-hole pairs, resulting in an avalanche of charge carriers). These situations are encountered in a functional IC. For non-functional ICs it has been found that, where gate oxide breaks down, light emission persists when a bias is applied over the oxide. In addition, the typical light emission due to hot electrons or punch-through can be modified when anomalous electrical behavior of transistors occurs. These effects make like emission a valuable technique. Until recently, however, it was of limited value because of the low intensity of the emitted light. This situation has changed dramatically with the introduction of high-gain image intensifiers, which allow the detection of very faint light emitted by an IC [22].

The experimental set-up we use consists of an image intensifier replacing the camera on the arrangement shown in Fig. 16.36. The output of the image intensifier, with a gain variable from 70000 to 700000, is coupled to a CCD camera and an image processing unit. The image processing unit incorporates a frame buffer, an arithmetic logic unit for performing calculations and an analogue/digital interface to perform input-output operation. This unit is driven by dedicated software, which allows averaging, integration and overlaying of images.

An example of a typical measurement is shown in Fig. 16.38. Here we first recorded an image of the IC with a weak illumination. The illumination was then turned off and a light emission picture of the same area was recorded. This was subsequently overlaid with the image of the IC. In Fig. 16.38 the location of light emission lies within the drawn circle. After deprocessing of this area the light emission was found to have resulted from damaged gated oxide; the damage occurred at the interface of gate oxide and field oxide (thick oxide layer). The excess current caused by this error depends strongly on the supply voltage. For this particular example a current of 2 μA could be visualized. The spatial resolution was about 1 μm. In general, the relation between light emission

and current depends on the excitation mechanism of the photons. Since there are several mechanisms that can play a role, no simple relationship exists between electric current and the intensity of emitted light.

It is important to note that this method is not only suited to localize failures but can also be used to study functional devices and test structures. Here one can use the spectrum and time dependence to get information on the physical processes which on the one hand emit light and on the other are important for the electrical behaviour of ICs. This would be particularly useful if a unique relation exists between the emitted spectrum and the mechanism giving rise to light emission. One complication in measuring spectra is the low intensity of the light (particularly in the visible). Here photocathodes with a good quantum yield over a broad spectral range would improve the situation. The sensitivity of commercially available equipment (Hamamatsu) is already good enough to permit photon counting and here only the background count could be minimized. In addition, the image intensifier can be gated, which opens up the possibility for 'stroboscopic* pictures. Also, the time dependence of light generated by a chip in pulsed operation can be measured in this way. Information thus obtained could be valuable for determining characteristic times of charge carriers.

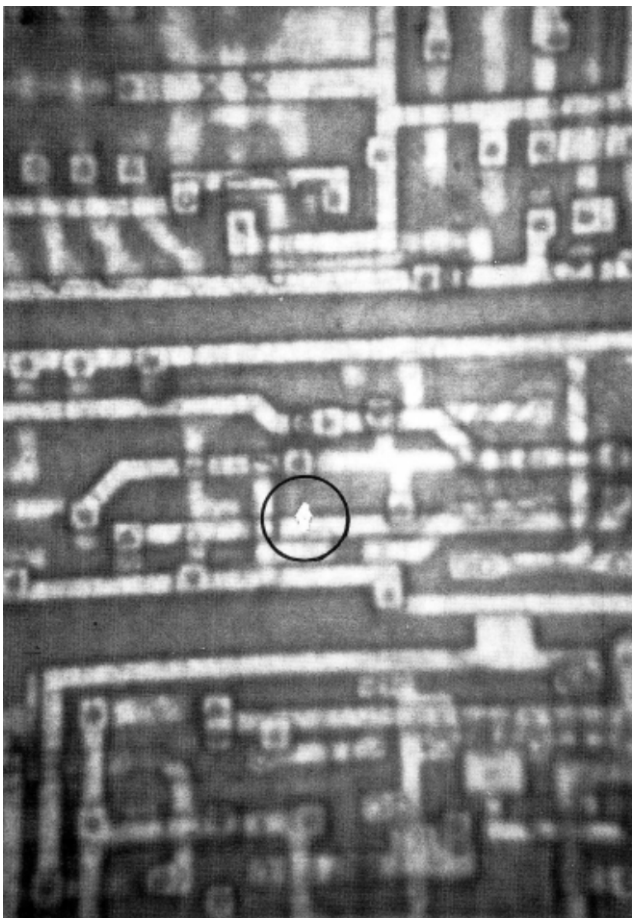


Fig. 16.38. Optical micrograph of a 64k SRAM showing the location of light emission [13].

These developments might be hampered to some extent by the fact that metal lines screen the light from underlying 'sources'. Of course, the light on the screened source may escape as a result of multiple reflections, but this will limit the spatial resolution.

16.3.1.3. Eleclrooptic sampling

Electrooptic (EO) sampling is a technique to measure the time dependence of electric signals in ICs [23-26]. This is achieved by bringing an electrooptic (EO) crystal so close to the surface of an IC that the emanating electric fields from the IC enter the EO crystal. The interaction of the electric fields with the crystal causes a change in optical properties which influences the polarization of an incident laser beam. A schematic arrangement for measurements in ICs is shown in Fig.16.39. Here the electric field changes the extraordinary refractive index (n_e), causing the field-dependent phase difference between light polarized parallel to the c axis and light polarized perpendicular to the c axis. When the polarization of the incoming light is at 45° to the c axis 'both polarizations are excited' and the elliptical polarization of the reflected laser beam becomes a measure of the phase difference. As will be explained later, this phase difference is linearly proportional to voltage. The time dependence of the electric field is sampled by using a pulsed laser, where the laser pulses are synchronized with the (repetitive) electrical signal. By varying the delay between laser pulses and the electrical signal the latter can be recorded. The laser pulse width determines the time resolution, provided the pulse width is longer than the response time of the EO material.

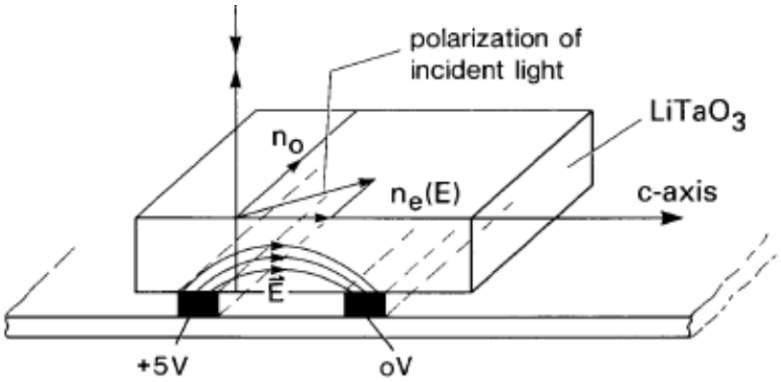


Fig. 16.39. A possible configuration for measuring signals employing the electrooptic effect. The dark regions in the cross-section represent metal lines at a different potential [13].

The experimental set-up is shown in Fig. 16.40. Here laser pulses of 50 ps are generated by a semiconductor laser driven by an electronic unit. The laser beam enters a microscope (partly shown in Fig.16.40) where it is focused on the backplane of a bismuth silicate (BSO) crystal. The [100]

direction of the BSO crystal is parallel to the propagation direction of the laser. With the microscope the circuit and laser spot can be seen simultaneously and with a ×50 objective the minimum laser spot is around 1 μm. The BSO crystal has a conical shape and is attached to the objective lens. In this way the crystal can be placed anywhere within the IC and adjustments on the 45° mirror allow fine positioning of the laser spot within the field of view. The reflected laser beam enters a detection chain consisting of a ¼λ plate, a Wollaston prism and a dual detector. With a proper choice of optic axes the difference signal of the two detectors is proportional to the electric field induced phase delay in the BSO. To facilitate the detection of the small difference signal, pulse generator PG2 generates a low-frequency (10 kHz) modulation of the electrical signal. By using the lock-in amplifier tuned at this frequency the signal-to-noise ratio is improved and drift is reduced. Pulse generator PG1 synchronizes the laser pulses and electrical signals, the latter being generated by a programmable data generator (HP 8080 A). The HP 8080 A is also used to vary the delay between laser pulses and the electrical signal.

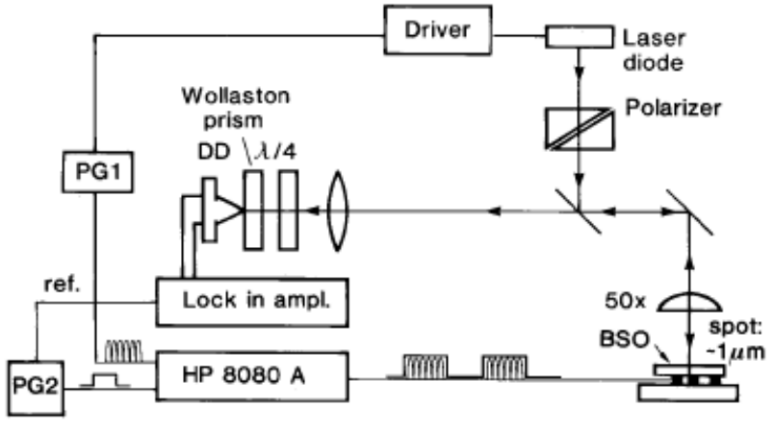


Fig. 16.40. Experimental set-up used to perform electrooptic sampling in ICs. DD, dual detector; PG, pulse generator; BSO, bismuth silicate; λ/4, λ/4 plate [13].

A measurement of 10 ns wide pulses with a repetition rate of 20 MHz is shown in Fig. 16.41, where we plot the output of the lock-in amplifier versus the delay. The pulses are applied Co a test structure with a 5 μm wide transmission line. The rise time of these pulses is around 1 ns. Here we used a computer connected to the HP 8080 A to increment the delay with 100 ps steps. The total measurement time for Fig. 16.41 was 3 s. It is seen from the figure that accurate timing information is easily obtained, but the amplitude (voltage) of the signal is not readily obtained. The amplitude of the generated pulse is 5 V, and with this set-up signals with amplitude of 50 mV could be measured. Factors determining the amplitude of the signal measured by the lock-in amplifier will now be discussed.

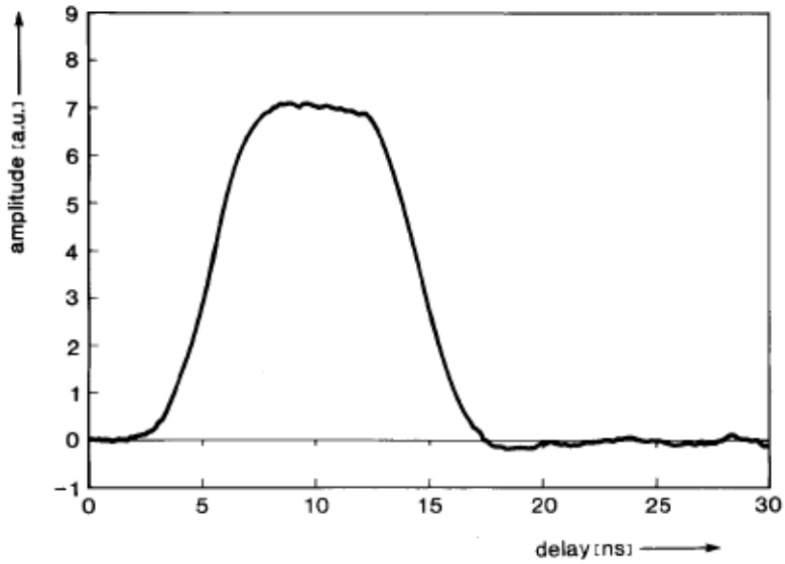


Fig. 16.41. Signal measured on a test structure using EO sampling.

The crystal symmetry of the EO material and the orientation with respect to the laser beam and electric field determines which components of the electric field are being measured. For $LiTaO_3$ shown in Fig. 16.39 one obtains the transverse geometry [27], where the main contribution is given by the component of the electric field parallel to the *c* axis (E_y); however, the perpendicular component (E_z) also gives a contribution which cannot be neglected:

$$\Delta n_e = -0.5 n_e^3 r_{33} E_y \tag{16.54}$$

$$\Delta n_0 = 0.5 n_0^3 (r_{22} E_z - r_{13} E_y) \tag{16.55}$$

Here n_0 and n_e are the ordinary and extraordinary refractive indices without applied electric field and Δn_e, Δn_a are the changes caused by the electric field; r_{33}, r_{22} and r_{13} are components of the EO tensor r which determines the relation between changes in dielectric constant and applied electric field. The non-zero components of this tensor are determined by the symmetry of the crystal. Because the laser beam crosses the crystal twice the phase difference present in the reflected laser beam is given by

$$\Delta\Gamma = \frac{4\pi}{\lambda} \int_0^d (\Delta n_0 - \Delta n_e) dz \tag{16.56}$$

and by using equations (16.54) and (16.55) one obtains

$$\Delta\Gamma = \frac{2\pi}{\lambda} \left\{ n_0^3 r_{22} \int_0^d E_z dz + (n_e^3 r_{33} - n_0^3 r_{13}) \int_0^d E_y dz \right\} \tag{16.57}$$

Here λ is the wavelength of the laser and *d* is the thickness of the crystal. The integral of E_y depends on *y* and is maximum at the mid-point of the two tracks. The integral of E_z equals the potential drop over crystal, which is at a maximum on the 5 V track, provided the potential at the other crystal side

is 0 V. From equation (16.57) it is apparent that calibration of the amplitude of the EO effect is difficult for the transverse geometry. Also potential variations on nearby tracks disturb the measurement (through the integral of E_y).

The longitudinal configuration [27] is more promising. A representant of this configuration is BSO. Here only the E_z gives a contribution:

$$\Delta n_x = 0.5 n_0^3 r_{41} E_z \tag{16.58}$$

$$\Delta n_y = -0.5 n_0^3 r_{41} E_z \tag{16.59}$$

where r_{41} is the EO tensor component of BSO and x and y are at 45 to the [010] and [001] directions of the BSO. The phase difference is now given by

$$\Delta\Gamma = (4\pi n_0^3 \frac{r_{41}}{\lambda})V \tag{16.60}$$

where V is again the potential drop over the crystal. If the potential on the front side of the crystal is 0 V one directly measures the potential on the track where the laser beam is focused. However, the strength of the electric field falls substantially over a distance comparable with the spaces between the metal lines, which means that the EO material must be brought close to the metal track. The gap that can be allowed without loss of signal amplitude is the subject of further study. Here important parameters are the dielectric constants (at frequencies of the signal to be measured) of the EO crystal and gap material. Different ways to define the potential on the front of the crystal are also being investigated. These studies are necessary for obtaining a calibration procedure for measurements in silicon ICs. For GaAs ICs the situation is simplified because it is in itself an EO material with the same symmetry as BSO. Therefore longitudinal sampling can be readily applied with zero gap.

The time resolution of 50 ps obtainable with semiconductor lasers [28-30] seems sufficient for measurements in state-of-the-art silicon ICs. Since the response time of many EO materials lies below 1 ps, the time resolution can be easily improved by using shorter laser pulses [23]. Then EO sampling can be used for high-frequency characterization of single devices and interconnections.

16.3.2. In-circuit Excitation

In-circuit excitation is carried out to measure the response of the IC to a local disturbance. Depending on the type and magnitude of the excitation the response can vary from a small change in parameters to an induced fault state. Excitation can be carried out with radiation or particles. In the following we will describe how alpha particles (helium nuclei) and light can give rise to fault states. To qualify for an IC the occurrence of these faults has to comply with stringent limits.

16.3.2.1. Alpha particle induced errors

In silicon an alpha particle with energy of 4 MeV loses its energy within a distance of 30 μm and the energy loss mechanism is due to the creation of electron-hole (e-h) pairs. The creation of each e-h pair requires 3.6 eV and some 1 million e-h pairs are therefore generated along a path of such an alpha particle. This generation process takes about 1 ps. After their generation the e-h pairs diffuse and/or are accelerated by the electric fields within the IC. Finally, ihe non-recombincd electrons and holes reach the substrate or a circuit node. For the circuit under study (SRAMs made in an rt-well CMOS process) the electrons will be collected on a circuit node within a typical time of 1 ns. The resulting peak current is in the milliampere range and can cause a change of the memory content [31]. When the affected memory cell is rewritten this error disappears. These alpha-induced errors are therefore referred to as soft errors and the rate at which they occur is referred to as the soft error rate (SHR). Note that for DRAMs alpha particles can more easily lead to errors than for SRAMs.

Alpha particles emitted by natural abundant radioactive impurities have energy of between 2 and 9 MeV and the number of alpha particles reaching an IC is determined by the radioactive impurity content of the materials surrounding the IC. Due to the penetration depth of alpha particles only a layer of some 30 μm facing the IC will contribute. The materials used in the IC production also contribute in so far as they contain radioactive isotopes. The package filler HC 10-2 TX 003509 emits one alpha particle into an IC of one square centimetre every hundred hours [12, 32], Assuming that one out of every thousand alpha particles creates an error, this means that an error occurs once in every 10^5 hours. Though this error rate may seem low, it still creates a serious threat to the reliability of memory devices. The requirements as to the reliability are expressed in FITS (failure in ten to the ninth device hours). For the SER the requirement is generally below 1000 FIT and the example given above would exceed this by a factor of ten. However obscure as the occurrence of alpha particles may seem, therefore, the alpha particle sensitivity of present-day memory devices has to be known.

To determine the SER the measurements are accelerated by exposing the chip surface to a larger alpha flux. In our laboratory we use Am24T sources with an intensity varying from 500 Bq to 100 kBq. If all the emitted alpha particles enter the chip, this corresponds to an accelerating factor of between 2×10^8 and 4×10^{11} as compared with the example given earlier. The number of errors per incident alpha particle obtained when the IC is exposed to a high flux of alpha particles is called the accelerated SER (ASER). The ASER can be used to determine the influence of chip parameters such as supply voltage and temperature on the alpha sensitivity. The SER can be calculated from the ASER when the alpha emission from materials surrounding the chip and materials used in the chip production is known. The SER can also be determined when a large number of devices is

tested over a long period of time (system test). ASER measurements can be validated by comparing calculations based on them with a system test.

As an example of the influence of chip parameters on the SER we consider a 64k SRAM made in a 1 μm CMOS process. Figure 16.42 shows the influence of supply voltage on the SER. The number of alpha particles incident on the chip in its package (no acceleration) was not accurately known. This caused a large and unknown systematic error in the derivation of the absolute value of the SER, but the relative values are much more accurate. It can be seen that the SER increases exponentially over four orders of magnitude when the supply voltage decreases from 5 to I V. The qualitative explanation is that at a lower supply voltage the stored charge is also lower, which makes it easier for the 'fixed amount' of injected charge to cause an error. These measurements were done for two different patterns written into the SRAM. Both the dependence on supply voltage and the dependence on pattern agree with circuit simulations [12]. The pattern dependence finds its origin in the non-symmetrical layout of the memory cell. The strong dependence on supply voltage is important when one is considering memories with low-voltage battery back-up or future generations of memories operating at lower supply voltages.

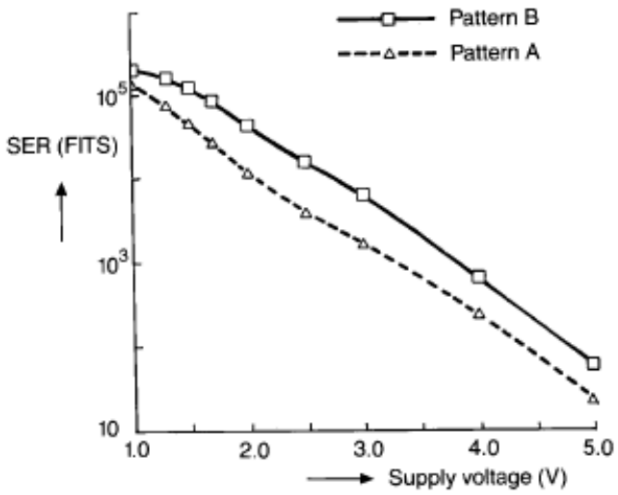


Fig. 16.42. Soft error rate induced by alpha particles as a function of supply voltage for a 64k SRAM. Measurements were done for two different patterns (A and B) resident in the memory [13].

16.3.2.2. Laser excitation of latch-up

Latch-up in a CMOS IC occurs when a thyristor between the supply voltage and ground opens. Such an unwanted, parasitic thyristor exists when PMOS and NMOS transistors are in close proximity [12]. This situation is depicted in Fig.16.43. The PMOS transistor is made in an n-well, whereas the NMOS transistor is made in the *p*-substrate. The *n*-well is connected to the supply voltage V_{dd} and the back of the substrate is connected to ground. In this configuration there are two parasitic bipolar transistors, one being a *pnp* connected to the source S_p of the PMOS transistor, and

the other an npn connected to the source S_n of the NMOS transistor. These two parasitic transistors form a thyristor as shown in Fig. 16.43. When V_{dd} is increased to a certain value, V_{high}, this thyristor may open and V_{dd} is then connected to ground through substrate resistances R_1 and R_2. Because R_1 and R_2 are fairly low a large current starts to flow, which can destroy the chip. Even when the supply voltage is lowered afterwards the thyristor remains in the conducting state until a value V_{hold} is reached. Special measures [33] are therefore taken to prevent this happening. Nonetheless, latch-up sometimes occurs in a chip and then it is important to know which are the sensitive places.

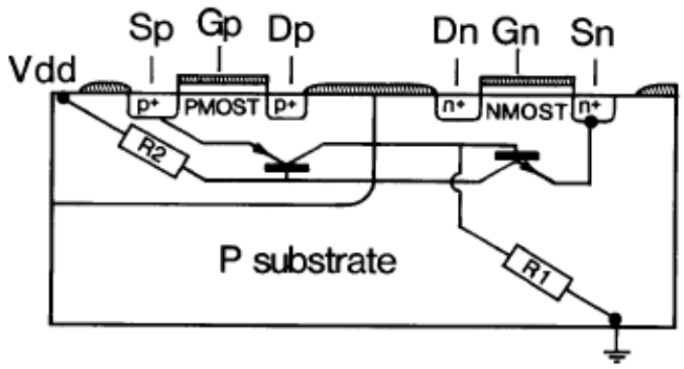


Fig. 16.43. The p and n transistor in a CMOS process showing the presence of parasitic bipolar transistors forming a thyristor. Latch-up occurs when this thyristor opens [13].

Here we will describe an experimental set-up for detecting sites in an IC that are prone to latch-up. This set-up was originally developed in [34] and it is used here with some modifications. The basic principle is the local creation of electron-hole (e-h) pairs with a focused laser beam. These e-h pairs are separated by the local electric fields in the IC and consequently give rise to a small current. The supply voltage is set close to V_{high}. When the e-h pairs are created in a region sensitive to latch-up the extra current will trigger latch-up, resulting in a large supply current.

The experimental set-up is given in Fig. 16.44. A helium-neon laser using a maximum power of 5 mW is focused with a microscope objective to give a spot of 1 μm with an objective with NA = 0.6 (at ×50 magnification). Before entering the microscope the laser beam crosses an electrooptic modulator and a beam expander to fill the entrance pupil of the objective. The modulator is used to switch the laser on and off and to adjust the power level. The chip is mounted on a computer-controlled *x-y* table. To examine the chip the table is first scanned along *y* in a line-by-line sequence. After each line scan the laser and supply voltage are switched off in order to reset any latch-up condition that might have occurred. If latch-up occurs a large increase of supply current is detected and the corresponding *x* value is recorded. Then this procedure is repeated for scans along the *x* direction in order to determine the *y* value of 'latch-up sites'. After the

measurement the table is positioned at the recorded (*x*, *y*) values where an image is taken from the IC with the laser on. In this way the laser spot marks the latch-up sensitive location.

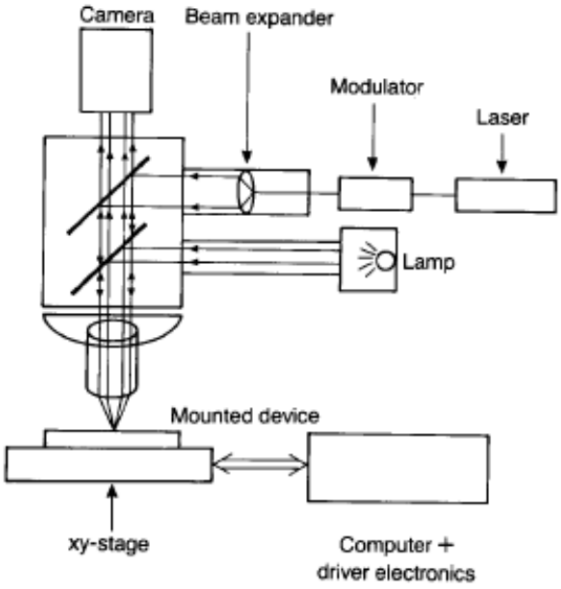


Fig. 16.44. Experimenial set-up for detecting locations within the IC which can cause latch-up. The laser is used to excite electron-hole pairs [13].

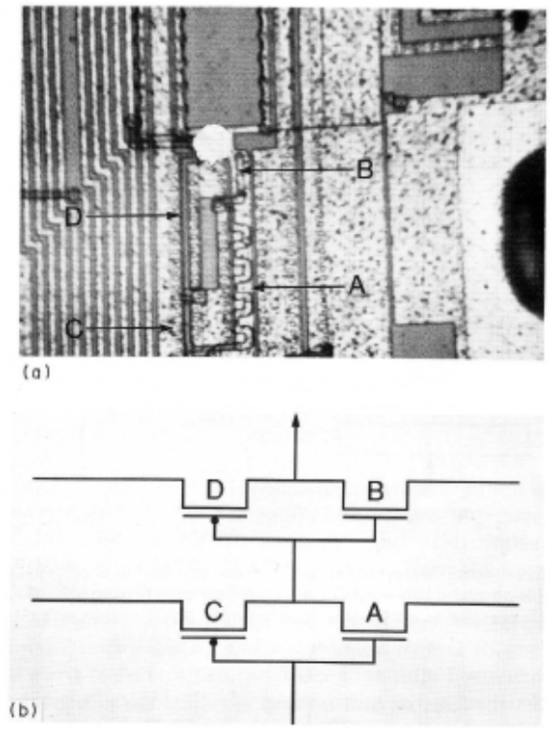


Fig. 16.45. (*a*) Oplical micrograph of an IC showing the location which is most sensitive to latch-up: A, B, C and D point towards transistor gates, (*b*) The electric circuit giving the interconnection of A, B, C and D and the type of transistor (dot indicates p transistor) [13].

An example of a chip examined using this technique is given in Fig.16.45. Here the white region is the image of the laser spot and therefore marks the location where latch-up was induced during the measurement. A, B, C and D indicate the gates of the four transistors present in this picture. The type (*p*-type with dot) and electrical connection of these transistors is indicated in Fig. 16.45(*b*). As can be seen, it resembles the situation shown in Fig. 16.43 (adjacent *p* and *n* transistors).

16.3.3. Repair Techniques

The testing and analysis of prototype ICs is often hampered by fatal errors in either the design or the processing of these ICs. The ability to repair these fatal errors on the spot leads to fewer redesigns. This implies a shorter time to market, a reduction of development cost and more efficient use of processing equipment. A laser-based facility has been built by which the two basic repair operations—the breaking and making of connections—can be implemented. The breaking of connections is performed by a MEL 31 laser system (Florod) incorporating a pulsed xenon laser emitting 1 μs pulses with a maximum energy of 0.7 mJ in the wavelength range of 0.48 to 0.54 μm. The laser beam is incident on a rectangular diaphragm which is projected onto the IC through a microscope. The diaphragm is also projected with visible light visualizing the size and location of the area to be removed by the laser. The IC (either a mounted device or a wafer) is placed on a dedicated (*x, y, z*) stage with a resolution of 0.5 μm and a maximum scan of 200 mm in both the *x* and *y* directions. The *x* and *y* motions are computer controlled, which enables automatic die-stepping when using wafers.

When the laser beam is projected within the IC the local temperature can rise above the melting or boiling point of the constituent IC materials, causing them to evaporate. This temperature increase depends on various parameters: the power density and the pulse time of the laser, the absorption coefficient and the thermal conductivity of chip materials. The situation is further complicated by the temperature dependence of these material parameters. These factors have led to an 'experimental' approach in breaking connections.

In a typical IC we can find the following sequence of layers: a top layer of Si_3N_4 or SiO_2 (scratch protection): a metallization layer (second metal); an insulation layer; a metallization layer (first metal); an insulting layer; a polysilicon layer; the active region (where the transistors are located). The following operations can be performed:

(*a*) openings in scratch protection;

(*b*) removal of first metal and overlying scratch protection;

(*c*) removal of second metal and overlying insulators.

For making new connections laser-induced liquid-phase metal deposition was developed [12, 32]. In the process metal ion complexes, which form when ammonia (NH_3) is added to a metal salt solution, produce the supply of metal. When this solution is applied on top of the IC, local deposition of the metal can be initiated by local heating with a focused Ar^+ laser beam. Within the laser spot the temperature can be increased to a level where a chemical reaction starts with the net effect that the metal ion complexes decompose and metal atoms are deposited. The local increase in temperature also gives rise to local boiling and a stationary jet flow, which result in a very efficient replenishment of reactants. A very high deposition rate is therefore achieved. Up to now we have mostly used palladium salt solutions from which palladium tracks are deposited, but the use of other metals is equally possible [35].

One application taken from reference [12] is shown in Fig. 16.46. Here the input protections of the IC (the meander-like structures) did not function, which prevented further testing of the device. After bypassing the input protections with palladium tracks (black), the devices could be tested and revealed some design errors! Without this repair these errors would not have shown up until the next redesign.

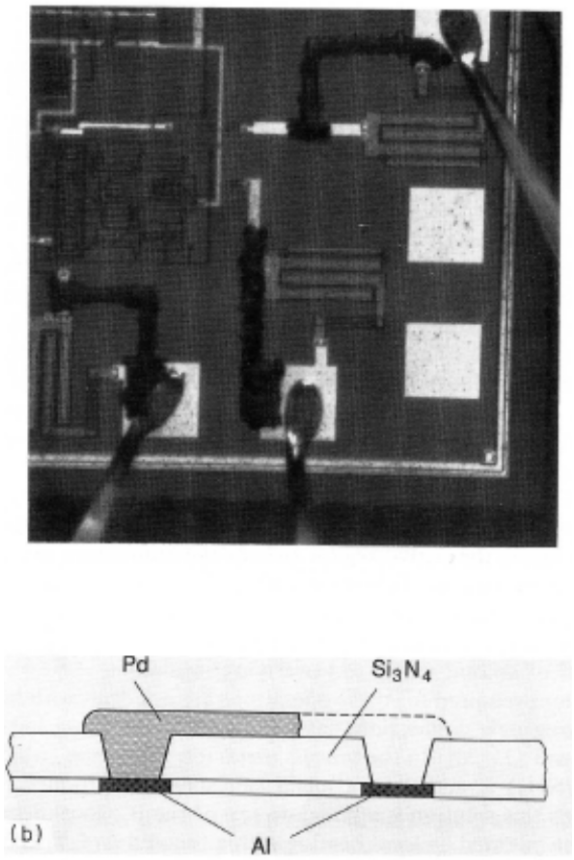


Fig. 16.46. (*a*) Repaired IC showing ihe deposited palladium tracks, (*b*) Configuration showing how lines of small geometry can be connected [13].

To connect metal lines of small dimension we follow a procedure that can be outlined on the basis of Fig. 16.46(*b*), showing two metal lines covered by scratch protection. Two holes are made in the scratch protection with the pulsed laser and then a palladium track is grown using the method explained above. The minimum geometry of lines that can be connected in this way is determined by the minimum size of the holes and not by the minimum width of the palladium tracks.

16.3.4. Comparison and Outlook

In this section the perspective is broadened in two ways:

(*a*) by comparing the diagnostic techniques described so far with other existing diagnostic techniques;

(*b*) by indicating possible developments leading to tomorrow's diagnostic methods.

Only those diagnostic methods required for obtaining information on electrically active devices will be considered. The same division of topics as applied before will be used: in-circuit measurements, in-circuit excitation and in-circuit repair.

16.3.4.1. In-circuit measurements

(a) Local temperature detection

Far-infrared microscopy and detection of local boiling can also be used for the localization of local heat sources in ICs. If we compare temperature resolution, spatial resolution and image acquisition time, LC outperforms the other methods. Only for special applications like high temperature or large structures can far infrared be a better alternative. Although the trend in ICs is towards smaller geometries, LC can still be used to determine the approximate site of an error and this will continue to be important information.

(b) Light emission

Light emission is a unique tool to determine the location of a fault and it will remain valuable for future generations of ICs. It is our experience that light emission is often complementary to LC. New developments including light emission will be directed to the use of information contained in its spectrum and its time dependence. Not only could this help to identify failures, but it can also be used to give an understanding of the physics behind electrical phenomena in functional devices.

(c) Signal measurements

Today electron beam testing is the most commonly used technique for waveform measurements. It should be noted here that even when other methods for waveform measurements have better

specifications than electron beam testing, the required degree of integration, which has been realized for electron beam testing, can be obtained only after a lot of work.

Related methods to perform internal signal measurement are the photoelectron SEM (PSEM) [36], photoemission from ICs [37] and photoconductive sampling which relies on very fast switches driven by picosecond laser pulses [38]. Comparing the various signal measurement techniques, it is clear that in order to reach picosecond time resolution laser-based systems are most adequate. When these systems must be interfaced with testers several difficulties with respect to triggering have to be solved. For triggering, photoconductive switches could be necessary. These difficulties are avoided when a laser-based test system [39] is used. Conventional electron beam systems employing a faster beam blank driver could also be used. An advantage of EO sampling over electron beam testing is that the former can be done in air, while the latter has to operate in vacuum. Also some specific problems of electron beam testing (beam displacement caused by electric signals) are absent. The advantage of electron beam testing is its superior spatial resolution, which will be important in future generations of ICs. Signal measurements on future ICs will therefore probably show the merging of a laser-based approach, to achieve good time resolution, and electron beam testing, giving good spatial resolution. At present the PSEM comes close to this situation.

16.3.4.2. In-circuit excitation

OBIC and EBIC also employ in-circuit excitation. In both cases the changes in supply current caused by either an optical beam (OBIC) or an electron beam (EBIC) are detected. They can be useful for the detection of faults in special applications (pn junctions) and larger geometries, where the exciting beams can easily reach the sensitive layers. Widespread applicability of these techniques is not to be expected. The main reason is the inability to reach active material in dense circuits. Optical beams will not penetrate overlying metal layers and once electron beams arc energetic enough they also will cause damage to the circuit. These same considerations apply to laser-induced latch up. With optical beams one can resort to allowing longer wavelength light to enter from the substrate side, giving decreased spatial resolution. Alpha particle sensitivity will remain an issue for the future generations of memories. This statement is based on the fact that both the node capacitance and node voltage will be lower for smaller geometries. More simulation work will thus be required, but basically the measurements can be carried out in the same way.

16.3.4.3. In-circuit repair

Repair based on laser systems has the disadvantage of a fairly limited spatial resolution. The most appropriate technique for increasing resolution is the use of focused ion beams (FIB). Culling

with an FIB has already proved to give a spatial resolution of 0.1 μm [40]. A further advantage of this method is its greater controllability and selectivity. This allows holes to be made in one layer without damaging neighboring layers, which opens up the possibility of making connections to underlying layers. However, the deposition of conductive material still presents some problems. At present, deposition is still mostly performed by the decomposition of (toxic) gases, which leads to complicated hardware. The development of alternative deposition methods seems important.

Features of Techniques for Characterization of Failure Analysis of ICs

- *Liquid crystal* is a method to quickly localize hot spots. Spatial resolution is 1 jtm and minimum detectable power dissipation is between 0.1 and 10 $\mu W/\mu m^2$.

- *Light emission* quickly localizes points that emit faint light. Spatial resolution is between 0.5 and 2 μm. Currents of 1 to 10 μA through defective oxide can be easily localized.

- *EO sampling* can be used to measure signals within ICs. Spatial resolution is 1 um and time resolution is 50 ps, but can be brought below 1 ps. Employing BSO crystals, the signal-to-noise ratio allows 50 mV signals to be measured, although absolute amplitude calibration is lacking for the moment.

- *Alpha particle* sensitivity of memories is measured by accelerated testing. An SER of 1 FIT can easily be detected

- *Latch-up* sites are quickly located by laser-induced latch-up. Spatial resolution is around 1 μm.

- *In-circuit repair* comprises laser cutting and laser-induced growing of metal (palladium) tracks

Chapter 17. Processes to produce integrated circuits

V.Verbitsky

17.1. The technological features of production of integrated circuits

Technology of production of integrated circuits (IC) and micro assemblies (MA) is a set of technological operations (TO) or technological processes (TP) of transformation of properties and shapes, test operations and tests that will be carried out with the incoming materials, semi-finished or some electronic elements to creation of IC and MA as completed products that have specified electrical parameters for acceptable economic and social indicators. TP also includes special hardware to perform the International Standard (IS).

The process of production of (IC) consists of a certain number of TO and passes that perform in the given order. If all TO are performed, products will be finished. In the process of manufacturing integrated circuits distinguish two cycles:

- manufacture of crystals, boards;
- the division of bases for crystals and planes; compilation and installation of the IC.

Technological operations of the first cycle, for example, for semiconductor integrated circuits aimed at creating the required number of on-chip transistors, resistors, capacitors and creating of connections between them to implement the necessary function in IC. The operations of the second cycle are aimed at creating of completed IC design, GIC or MA.

The integral group method is a specific feature of the IC fabrication.The hundreds of billions of elements are produced during one TO at the same time using this method. For example, on a plate with a diameter of 300 mm placed about 400 crystals IC dynamic random access storage devices with area of 160 mm^2, each of which has 64 millions of transistors and 64 millions of capacitors. Process of manufacturing capacitors is preceding to process of transistors. Therefore, on a single chip at the same time64 millions of transistors or 64 millions of capacitors are creating at the same time. On the one plate about $2.56 \cdot 10^{10}$ elements are producing simultaneously. Since IC manufactured on batches, each with 20 to 40 plates, the total number of transistors generated at the same time will be from $5.12 \cdot 10^{11}$ to $1.024 \cdot 10^{12}$. These reach proximity parameters of transistors, increase the percentage of suitable components and reduce their cost and the cost of integrated circuits. Group methods combine all the processes and manufacturing operations that are applied to semiconductor wafers or substrates in general, that is the manufacture of crystals, boards and in some cases - in the preparation.

Using typical TP is an important feature of manufacturing technology. A typical process – is a set of TP or TO that is performed in a specific sequence in a particular technological equipment to produce IC with defined structure and design by using group method. Using typical processes makes it possible to create many types of integrated circuits on a single structural and technological base that provides them with the same level of quality and reliability.

Silicon is the primary semiconductor material in integrated microelectronics, and for the next twenty years, it will retain its dominant position as a material for integrated circuits. Silicon has advantages not found in other semiconductor materials, namely that on silicon surface can be grown dense homogeneous film of silicon dioxide that is stable within a wide range of temperatures.

Oxide film used to make the under gate dielectric, selective mask holding local technological operations, a protective layer for the p-n transitions and the whole crystal. These properties of silicon emphasize its process ability.

Elements of semiconductor integrated circuits (transistors, resistors, capacitors, wires, etc.) created in monocrystalline wafer (chip) based on the same type of transistor structure - bipolar or MIS (metal - insulator - semiconductor). Therefore, the type of transistor and method of forming transistor structures in the crystal are assigned to the classification processes of manufacturing semiconductor integrated circuits. There are processes of manufacture: bipolar and metal - insulator - semiconductor (MIS) integrated circuits.

Passive components of hybrid circuits on the surface of the dielectric base or dielectric film create in the form of multilayer film structures of a given shape from materials with different physical properties. The active components (transistors, diodes, integrated circuits) attached to passive components by using assembly operations. As components you can also use resistors, capacitors, transformers, inductive components and so on. To create passive elements of circuits the thin (2 mm) and thick (20 to 40 micron) films are used. Processes are called thin-film and thick-film respectively.

The morphogenesis of all elements of integrated circuits performs on one side (surface) of semiconductor wafers or dielectric base. Outputs from all regions of the created elements are also placed on the same surface of the plates. Thus, all elements of integrated circuits are placed in the plan at the plate. Exploring the surface of the chip, each element of the IC, all regions of each element and contacts to regions can be identify. That determined the name of the technology "planar" (Fig.17.1).

All elements of semiconductor integrated circuits (transistors, diodes, resistors, capacitors, etc.) are created in a common semiconductor-base that has a degree of electrical conductivity. For the normal functioning of each of the elements and circuits in general is

necessary to ensure reliable isolation between elements that would exclude or minimize parasitic drove mutual influence and interaction of individual elements of the chip. Therefore, elements of semiconductor IC are provided in isolated areas, called pockets or chambers.

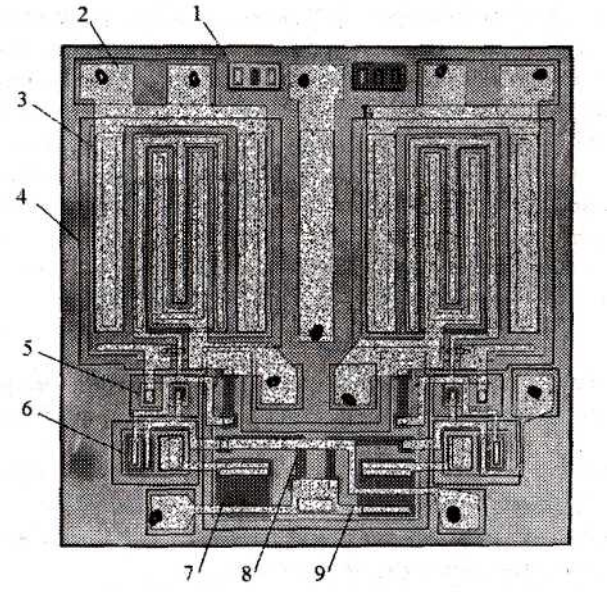


Fig.17.1. Crystal of semiconductor integrated circuits manufactured by the planar-epitaxial technology:

1 - semiconductor base;2 - contact plane; 3 - a powerful bipolar transistor; 4 - insulating region; 5 - diode; 6 - bipolar transistor of average power; 7 - low-resistance resistor; 8 - a high-resistor; 9 - conductor

There are several methods of isolation pockets, which are using in technology of semiconductor integrated circuits. The most important of which are *isolation by backward displaced p-n junction; dielectric regions isolation and mixed insulation*.

Since the elements of film and hybrid circuits create on a dielectric base or dielectric film, they are well isolated from each other and there is no need to use special insulating elements of design and manufacturing operations.

17.2. Technological processes of manufacturing a bipolar IC

The main structural element of semiconductor integrated circuits on bipolar transistors (BT) is a transistor structure type n^+-*p*-*n* with uniform distribution of impurities in the collector region.

Planar and planar-epitaxial technologies are most commonly used to produce bipolar IC. The collector areas of BT region arecreated in epitaxial layer of silicon *p*-type conductivity that increasing on the *p*-type substrate. Base and emitter region areproduced by diffusion or ion

implantation alloying impurities in the epitaxial layer. Emitter area sare created by donor impurities of the maximum possible concentration. Such impurity is preferably phosphorus.

Many technologies are used to produce the bipolar IC. These technologies differ by method of forming a transistor structure and method of isolation of elements.

17.3. Processes to produce bipolar circuits with isolation by back-biased *p-n* junction

Depending on the method of forming the insulating regions bipolar integrated circuits are created for the following technologies:

- ✓ standard planar-epitaxial technology with grooves *n*-layer and isolation by backward displaced *p-n* junction;
- ✓ planar-epitaxial technology with grooves *n*-layer and the collector insulating diffusion (KID-technology);
- ✓ planar-epitaxial technology with basic insulating diffusion (BID technology);
- ✓ planar-epitaxial technology with grooves n^+-layer and insulation by double diffusion, and others.

Consider a simplified flow diagrams and features of designs of transistor structures created by them in some of the above processes. The transistor is the most difficult part of the chip, the creation of which will indicate of the entire production of integrated circuits.

17.3.1 Standard planar- epitaxial technology with grooves n^+-layer and isolation by backward displaced *p-n* junction

This technology was one of the first technologies of bipolar integrated chips. The first operational amplifiers, switches, current multipliers, modulators and other integrated chips (Fig.1.2) were created by this technology.

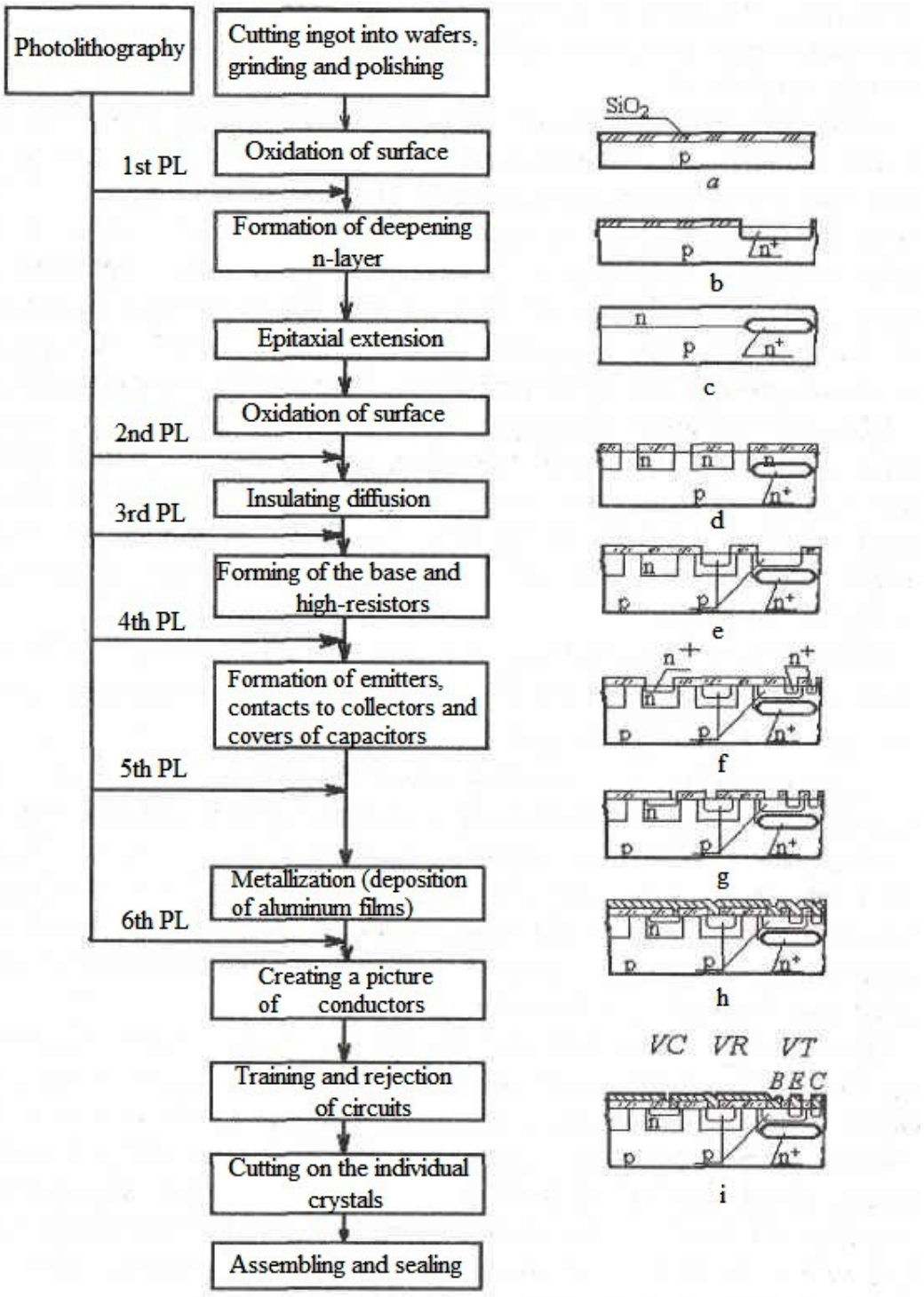


Fig.17.2.The sequence of basic technological operations of manufacturing bipolar integrated chips for planar-epitaxial technology with deepenn^+-layer and insulation backward shifted *p-n* transition

The technology is well proven, easily absorbed in the production, ensures high and stable electrical parameters of chips, has been developed from the first chips with topological dimensions of tens of micrometers to a minimum topological size of 10 microns, and then - to 5

or even 3 microns topological rules in which ensured good reproducibility of the electrical parameters of chips.

Integrated chips on this technology often create high-quality semiconductor epitaxial films deposited on semiconductor wafers of monocrystalline silicon stamps EKDB-10/0.1, obtained by the Czochralski method. In the letter designation E defines the scope of using silicon - epitaxial building films, K - Silicon, D - hole conductivity, B - doped with boron. The figure 10 indicates that the resistivity of the plate is 10 Om/cm, and the mean free path of carriers - 0.1 mm, or 100 μm. Size of the plate microirregularities *p*-type corresponds to the 14th grade.

At manufacturing of bipolar semiconductor IMC can use two-or three-layer plate with the following structure: silicon-based *p*-type, the surface of which was increased in high-epitaxial layer of silicon n^+-type resistivity (2÷5)-10^{-2}Ohm·m, and on its surface coated with low-epitaxial layer of *n*-type resistivity (0÷0.5)·10^2Ohm·m.

Consider a sequence of basic technological operations of manufacturing a bipolar semiconductor integrated circuits (see Fig.17.2) consisting of UT transistor, resistor URi capacitor UC (Fig.17.3a).

If used as a starting material the plate that has already prepared the surface to required purity level, the first oxidation process operation will be the working surface of the plate (Fig.17.2a). Film of silicon dioxide in the process takes two important functions: protecting semiconductor surface contamination from impurities and provides an opportunity for local diffusion, ion implantation or local epitaxy when elements of chips are creating.

The next process operation will be the first photolithography, by which to create the necessary ground plate windows in the oxide at the local diffusion. The diffusion of donor impurities is performed in semiconductor wafer of *p*-type conductivity, creating deep en n^+-region (Fig.17.2b). When creating depressions n^+-region as alloying material using antimony or arsenic which have a smaller diffusion coefficient compared to phosphorus. Therefore created p-type region will not significantly change their geometrical dimensions on the next thermal process operations.

After this technological operation part of the plate material, which created n^+-region and on which in the following TO will be generated collector region of the transistor, becomes low resistance. This makes it possible to significantly reduce the resistance created by the collector region of the transistor.

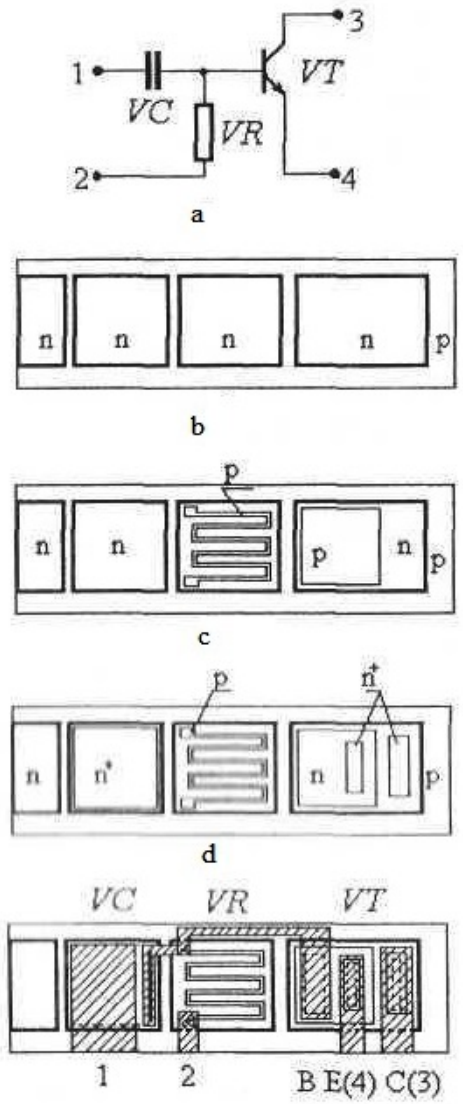


Fig.17.3.The topology of the basic elements of semiconductor chips (capacitor, resistor, and transistor)

Subsequent TO is the removal of the oxide layer on the surface of the plate and epitaxial accumulation of *n*-typesilicon on the working surface of the plate (Fig.17.2c). Options of depth layers and epitaxial process for the 10-micrometer topological rules are given in Table.17.1. For processes with smaller topological regions norms settings will change in proportion to the accepted ratio.

The surface of epitaxial filmis oxidized and performed a second photolithography by which in certain places of the plate creates a window in oxide film for diffusion separator. Separator insulating diffusion (Fig.17.2g, 17.3b) is performed. Acceptor impurity diffusion is hold in open windows. Oxide, which partially remained on the plate surface, serves as a mask through which atoms of alloying agents almost don't diffuse as diffusion coefficient of diffusant in dioxide of silica is smaller than the diffusion coefficient of diffusant in silicon.

Table 17.1

Options depth n^+-layers and epitaxial films

Options	n-layer(arsenic)	n-layer (antimony)
Specific surface resistance of deepening n-layer, Ohm/kV	11	20-75
The thickness of the deepeningn-layer, mcm	7±2	5
The thickness of the epitaxial n-layer, mcm	7,5÷9,5	10÷15
Specific volume resistance of epitaxial n-layer, Ohm·m	$(0,1÷0,4)10^{-2}$	$(1,5÷2,5)\text{-}10^{-2}$

After this operation, the islands of *p*-type silicon are detected surrounded on all sides by *p*-type silicon. Therefore they are electrically isolated from each other by *p*-*n* junction. As the insulating separator diffusion is performed in two stages, dioxide silicon layer is grown on the *p*-type regions in the second stage. This layer serves as the reflective boundaries for alloying impurity and simultaneously protects the surface of diffusion of unwanted elements.

A third photolithography is performed. In silicon dioxide creates a window for the base diffusion. If with the base area create high-resistor, in the relevant areas of the plate open windows under resistive strip of the high-resistors. Diffusion of acceptor impurity (boron) is used to creating a database field and high-resistive strip resistors (Fig.17.2d). As in the previous TO, diffusion is performed in two stages, and layer of silicon dioxide is increased on the surface that will perform these functions of camouflage on following TO. Depending on the value of resistor configurations it may be different. In this case, the resistor is made in the form of square wave (Fig.17.3c).

A fourth photolithography is used. In silicon dioxide creates a window for emitter diffusion and formation of the bottom plate of the capacitor (Fig.17.2g). The concentration of alloying impurities in the emitter region should be much larger than the original epitaxial layer. The conductivity of silicon emitter region is very high, which makes it possible to make it highly effective. In addition, it also provides an opportunity to make a low-bottom plate of the capacitor. In Fig.1.3g shows the top view of the wafer after the emitter diffusion. The surface of plate is oxidized again.

A fifth photolithography is performed. In the oxide layer opens windows for contacts to the regions of transistor, resistor, and capacitor and so on. (Fig.17.2h). The aluminum film is applied on the surface ofplate (Fig.17.2d). This TO is commonly called metallization.

A sixth photolithography (Fig.17.2i) is performed on the metallization. On the surface of the crystal the system of interconnect conductors is formed including the top plate of the capacitor. In Fig.1.3d shows the part of working surface of the crystal after forming layer of conductors.

Ready plates are subjected to testing and rejection, and then they are separated into individual crystals. Suitable crystals are mounted in the trunk. Set the crystal solves two major problems: protects the crystal from external influences and provides an electrical connection to the crystal through the pins. Many types of buildings and methods of compilation and installation of the crystals are developed.

Depending on the technology of the described above major manufacturing operations in a particular technological process can be repeated several times. Even in the considered simplified technological process of bipolar integrated circuits (see Fig.17.2) workflow photolithography was repeated for six times. As photolithography is widely used in industrial processes of producing bipolar and MDS-integrated chips as a separate process, let supplement the technological process of technological operations of photolithography.

Photolithography is intended for creation of the nitride film or silicon dioxide relief image of one topological layer chip.

Relief image of a topological layer of chip serves as the mask on the surface of plate through which the local diffusion, localion implantation, local application of epitaxial single-crystal films, local oxidation, the local ion plasma or chemical etching and others are performed in the following TO.

The technologies process of photolithography consists of several TO:

1. Applying of photoresist to the surface of the oxide film of semiconductor wafer (Fig.17.4a). Photoresist – is a light-sensitive material that is applied to a surface, such as spraying or centrifuging. Thickness of photoresist film can be from 200 to 1000 nm. Use two types of photoresists: positive and negative.
2. Drying of photoresist film, in which the solvent is removed from the photoresist. Resist becomes semisolid film.
3. Combination and exposure. Covered with photoresist plate is placed in a device for combining, where she alignment with photomasks. Locationofthe plate and photomask is regulated so that the special tags for photomask (reference signs) will coincide with respective tags on the plate. Masks are made on a glass plate on one side of which in thin

film of metal or an emulsion create matrix of optically contrasting images of topological layer of chips at a scale of 1:1. The picture plane of the masks is pressed to the plate and turn on the source of ultraviolet radiation (Fig.17.4b). High-energy radiation through the optically transparent region of photomasks penetrates to the surface of the photoresist in which photochemical reactions take place.

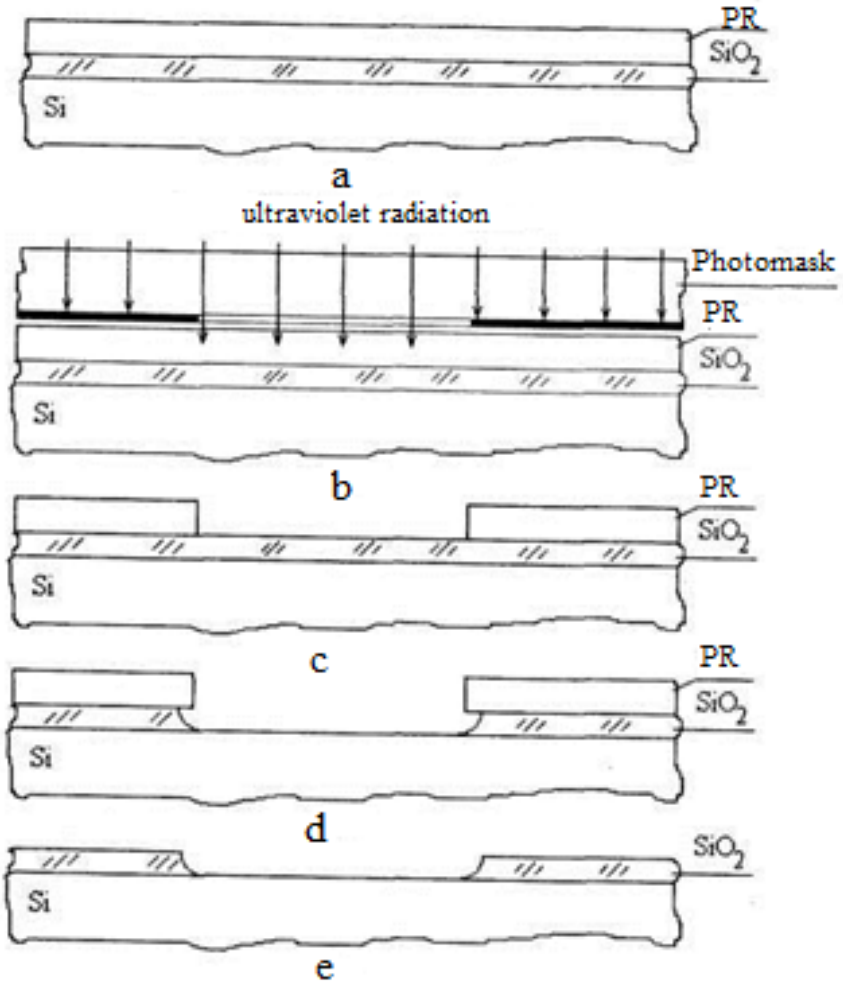


Fig.17.4. The sequence of the major technological operations of photolithography

4. The manifestation. Positive photoresist during irradiation through a photomask and during following manifestation creates a direct image of figure of masks. In areas of the photoresist where it exposed by ultraviolet, photochemical reactions of degradation is occurring. Therefore, under the action of the developer exposed regions of photoresist is easily dissolved and removed from the surface. Unexposed regions on the surface of the photoresist are almost insoluble (Fig.17.4c).
5. Negative photoresist during irradiation through a photomask and subsequent manifestation creates a reverse image on the photomask. In areas on the photoresist

exposed by ultraviolet light, photochemical polymerization reactions take place, and the photoresist becomes resistant to the action of the developer.

6. Stiffen of photoresist. In temperatures around 150 ° C photoresist is dried, resulting in are improving its adhesion to the oxide film and acid stability. After this TO a photoresist mask of image is created on the surface of silica.
7. Transferring of images of photo resistive mask on film of dioxide silicon. Images can be transferred in many ways: by chemical etching, plasma and plasma-chemical removal, etc. For the considered technology transfer process is performed by chemical digestion.

The plates are placed into solution of hydrofluoric acid and performed isotropic etching process by dioxide film in places, that unprotected by photoresist mask. After etching of the silicon dioxide windows picture is formed in dioxide layer, which replicates the pattern in the photoresist layer, and respectively photomask pattern (Fig.17.4e). Duration of etching is controlled with high precision: it shall be sufficient for the complete removal of the silica in the box and not large enough to prevent significant etching of dioxide under photoresist mask. Lateral etching causes increasing of the windows size in the mask of silicon dioxide in compare with the set in the design. After etching the surface of the plate is cleared of photoresist.

Considered process of transferring images from photoresist mask to silicon dioxide mask are called "wet" etching, based on a process of isotropic chemical etching.

Today, modern methods of image transfer, which include cutting and plasma-chemical methods, are used. The implementation of these methods is based on anisotropic etching process. Therefore, these methods are called "dry" etching methods. In these methods with a high degree of etching anisotropy in the direction perpendicular to the surface of the plate is significantly faster than in lateral. Due to the etching anisotropy was obtained integrated circuits with submicron size elements.

Considered a simplified scheme of the integrated circuits production process for bipolar plenary-epitaxial technology with deepen *n*-type layer and isolation by back-biased *p*-*n* transition contained only the basic manufacturing operations.

The typical production process of an integrated circuits series by the considered technology consists of 131 technology crystal manufacturing operations and 39 assembly technical operations. All technological operations are conventionally divided into basic, measuring, control and preparative. In Table.17.2 is shown the number of manufacturing operations for these groups to crystal manufacturing process.

Table 17.2

TP composition for manufacturing bipolar integrated circuits of crystal chips with planar-epitaxial technology by TO

Number ofTO	Basic		Measuring		Test		Preparative	
	Amount	%	Amount	%	Amount	%	Amount	%
131	71	54	20	15	34	26	6	5

IC manufacturing process groups of operations, depending on the characteristics of the physical-chemical transformations of the original material properties and equipment used to perform in different areas. The list of possible production areas and distribution of TO of the crystal shown in Table. 17.3.

Table 17.3

TO distribution of manufacturing crystal

Area	TO		Main		Measurement		Control	
	Amount	%	Amount	%	Amount	%	Amount	%
Chemical	24	18	14	58	-	-	10	42
Thermal	42	32	11	26	19	45	12	29
Photolithograph	56	43	44	79	-	-	12	21
Spraying	2	2	1	50	1	50	-	-
Epitaxy	1	1	1	100	-		-	-
Preparation	6	5	6	100	-	-	-	-

Considered IC production process characterized by repeating of similar TO. So with 131 crystal manufacturing TO 24 perform chemical processing surface plates, 42 - diffusion, thermal oxidation and other, 56 - photolithographic process operations and only two TO (building epitaxial single-crystal films on the surface of the plate and the plate surface metallization) performed one time.

The structure and topology of the bipolar transistor manufactured by the epitaxial planar-technology with deepen n^+-layer and isolation by back-biased *p*-*n* transition shown in Fig.17.5, where 1 - the basis with *p*-type conductivity; 2 – isolating p^+-type region; 3 - collector region; 4 – deepen *n*-layer; 5 - base area with *p*-type conductivity; 6 – emitter *n*-type area; 7 - *p*-type region, created with the emitter and which is intended to produce ohmic contacts to high-resistivity collector region 3.

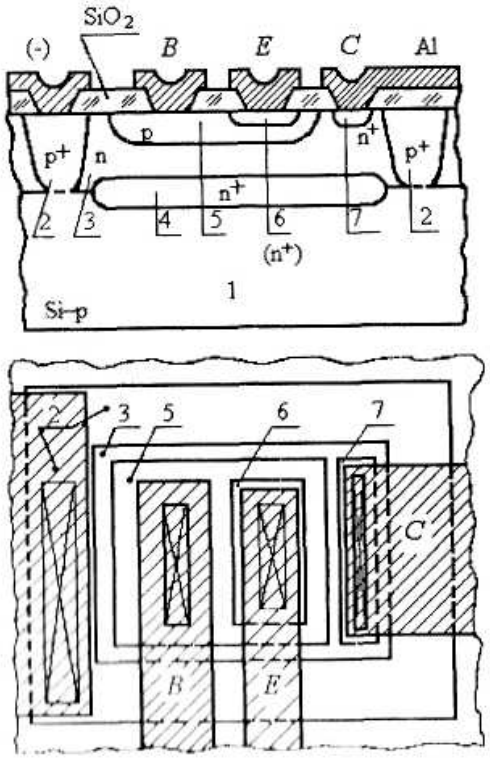


Fig.17.5. The structure and topology of the bipolar transistor, manufactured by planar-epitaxial technology with deepen n^+-layer and isolation with back-biased *p-n* transition

Lateral surfaces of area 2 and the top surface of area 4 are borders of the collector region 3. Base 1 and connected with it isolating region 2 are connected to a negative potential of supply source that backwardly shifts isolative *p-n* transition between regions 2 and 3.

As the reverse current of isolating transition is small, satisfactory transistor isolation from base is provided. The area surrounded on all sides by isolating *p-n* transitions, called pockets or isolated regions. They contain not just BT, but other elements of the IC (see Fig.17.2).

Low-resistance deepen layer 4 shunts a high-resistance epitaxial *n*-type layer placed over it, reducing the resistance of the collector region 3. This gives possibility to improve frequency properties of the transistor and reduce saturation voltage and thus reduce the voltage of low-level digital integrated circuits in which the transistor is turned on by the scheme "common emitter" and operate in saturation mode.

Considered technology used for IC small and medium degree of integration. The main advantage of this isolation method and bipolar transistor construction is a simple technology, but the isolation by *p-n* transition is imperfect: isolative *p*-type regions occupy a large area of the crystal; isolation transmission forms a barrier crossing capacity, which increases the switch delay of digital ICs and reduces the maximum frequency of analog IC; with increasing temperature and ionizing radiation the reverse current increases, worsening isolation. The structure of the

transistor, isolated by p-n transition, besides the basic n^+-p-n transistor contains parasite p-n-p transistor.

By the considered technology different types of bipolar integrated circuits are produced: transistor-transistor logic, Schottky transistor-transistor logic, emitter-dual logic, operational amplifiers, modulators, switches, etc.

17.3.2 The planar-epitaxial technology with deepen n^+-layer and collector insulating diffusion (KID-technology).

This technology uses an effective way to improve the density of IC elements and also the degree of integration of chips. Its essence is that the insulation of elements of circuits is performed by the collector insulating diffusion in place of separation diffusion of p^+-type, as was done in previous technologies.

The simplified schemeof technological process, that reflects the features of KID technology, is shown in Fig.17.6. As in the previous technology, the basis is silicon p-type conductivity, the surface of which oxide and in the right place open windows to create deepen n^+-regions (Fig.17.6a).

To create deepen regions n^+-type (Fig.17.6b) conduct a local diffusion of donor impurities. From surface of the plate remove silica; perform chemical cleaning and building up thin (about 2 mcm) single-crystal silicon epitaxial layer of p-type (Fig.17.6c). Oxide film is applied on the surface of epitaxial film, where windows areopenedby photolithographyto create insulating regions of transistor (Fig.17.6d).

Local diffusion of donor impurities creates insulating region of n^{++}-type in the entire depth of the epitaxial layer to contacts with deepen layer of n^+-type. Created diffusion region of n^{++}-type surround deepen n^+-layer and locate with him in physical and electrical contact. As a result, local epitaxial area of$p$-type (Fig. 17.6d), which are isolating from the base by collector layer of n^+-type, is formed, where a base and the emitter area of transistor structures will be create.Areas of n^+-type perform insulation of transistor structures and create a deep high-doped in contact region to the collector, which reduce the consistent resistance of the collector.

A layer of silicon dioxide is removed from the plate and the thin base p^+-diffusion is carried over the entire surface of plate (Fig.17.6f). As the basic diffusion is performed without dioxide mask, the process of masking layerapplication and photolithography are excluded. Therefore the manufacturing process is simplified.

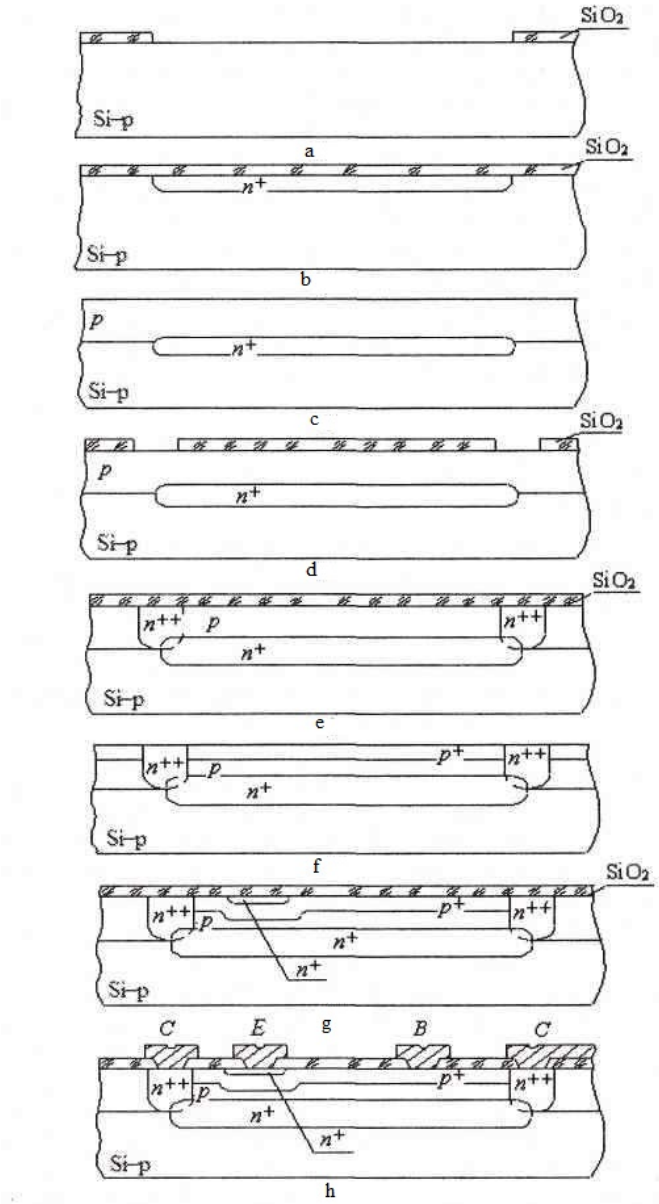


Fig . 17 .6. The scheme of technological process of manufacturingbipolar ICs by KID-technology

Some important problems in the design of the transistor are solved due to the process of applying a thin base p^+-layer:

- reduced transition resistance between the uniformly doped epitaxial layer of *p*-type and the metal contacts to the field base that prevents the formation of a contact rectifier;
- surface resistance of base region formed by diffusive p^+-layer and uniformly doped epitaxial *p*-type layer, becomes necessary to create high-value resistors.

The surface of plate ise xposed to oxide on the following TO. Windows in a protective SiO_2 film under the field emitter are opened by photolithography.Shallow diffusion or ionic

implantation of donor impurities in the base region of p-type is conducted to create the emitter area of the transistor (Fig.17.6g).

The superficial p^+-layer is pushed into deep of the epitaxial *p*-type layer during such TO, reducing the thickness of region and resistance of the active region of base under the emitter. As a result, the frequency of operation of the transistor increases and thus-speed of switching in digital integrated circuits and a frequency limit of circuit's functioning in a linear mode also increases.

Windows under the contacts to the transistor regions are opened in silica dioxide. The surface metallization are performed and system of conductors (Fig.1.6h) is created by photolithography. The passivation layer of silicon dioxide is applied on the surface of plate and windows over external contact planes are opened. As a result, transistor structures, which are isolated by back-biased *p*-*n* transitions between diffusion regions n^{++}-type and epitaxial *p*-type layer, are created. Although actual insulating diffusion was not carried out.

The thickness of the base of the bipolar transistor, created by KID-technology, depends on the thickness of the epitaxial layer, thickness backward n^+-diffusion of deepen layer and depth of the emitter diffusion from the surface. In such circumstances, the spread of thickness of the base will be much larger than for BT, isolated dividing diffusion. Therefore, great variation of the coefficients of the gain of transistor by current takes place in this technology and coordination characteristics is worse. A thin epitaxial layer of *p*-type base limits break down voltage the collector - base. Also, because the *n*-collector region directly contact with base area of *p*-type, the breakdown voltage collector-emitter is reduced. The area of the collector of the transistor is heavily doped, so to increase the speed of transistors no need to conduct an additional diffusion of gold to reduce the lifetime of minority charge carriers, as was done for the previous technology. Bipolar transistors are made by KID-technology with size 5x5 mcm. The structure and topology of the bipolar transistor manufactured by KID-technology, shown in Fig.17.7, where 1 - base of p-type conductivity, 2 - insulating region of *n*++-type, 3 - base area, 4 - n^+-deepen layer (collector), 5 - n^+-type emitter region.

KID technology is simpler than the standard (has less manufacturing operations), provides a higher percentage of suitable IC and has 1.5-2 times greater density of accommodation of elements and performance. Due to these advantages KID technology is widely used for both linear and digital IC.

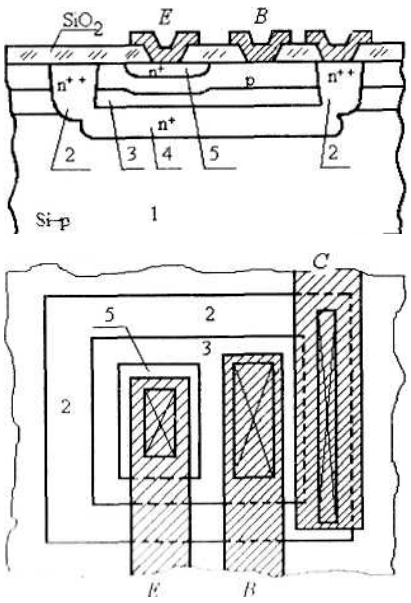


Fig.17.7. The structure and topology of bipolar transistor manufactured by KID-technology

17.3.3. Planar-epitaxial technology with base insulated diffusion (BID technology)

Technology with a base insulated diffusion is simpler than KID technology, but to ensure normal functioning of ICs produced by this technology, it is necessary to have an additional source for reverse bias insulating areas. By the way the size of transistor produced with BID technology (9 microns) is greater than produced with KID-technology.

The simplified scheme of technological process that reflects the peculiarities of BID technology is shown in Fig.17.8. As in KID technology its base is silicon of *p*-type conductivity, the surface of which is increased with monocrystalline silicon epitaxial layer of *p*-type conductivity (Fig.1.8a). The surface of epitaxial film is oxidized and windows are opened with the first photolithography to create a base circular and ring insulating areas around the base (Fig.1.8b).

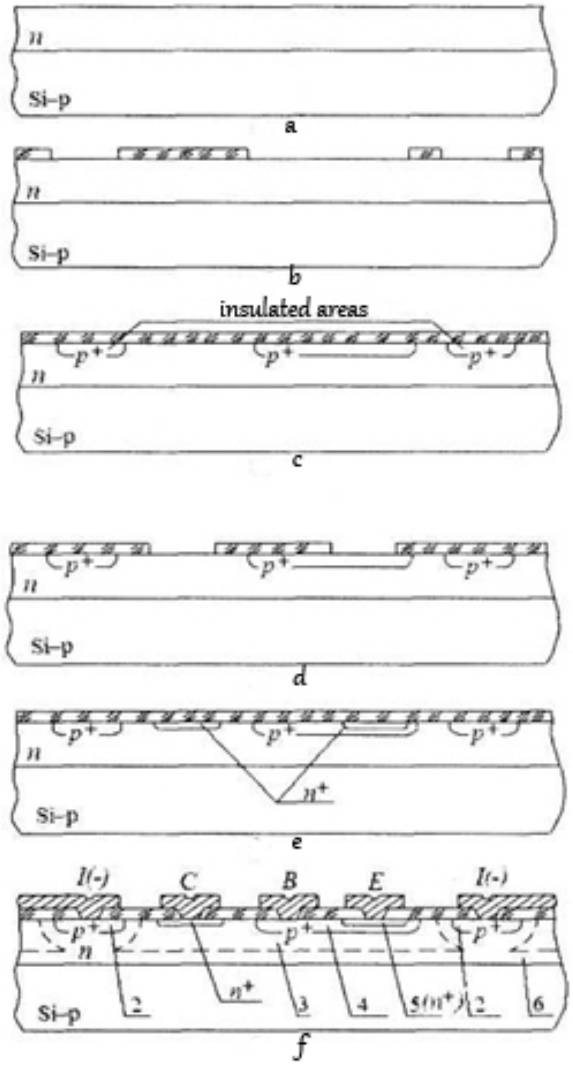


Fig.17.8. Scheme of the technological process of manufacturing bipolar ICs for BID technology.

Simultaneously acceptor impurities diffusion is performed in the base area and the insulating region. Insulating areas p^+-type does not penetrate the entire depth of the epitaxial layer (Fig.17.8c). During second photolithography windows are opened for emitter and collector contacts to the area is occurring (Fig.17.8d).

Donor impurities diffusion is performed in open windows, creating an emitter field and contacts to the collector area (Fig.17.8e). During the following technological operations windows are opened for contacts to the transistor areas and insulating areas, then performing metallization and creating system interconnect conductors (Fig.17.8.f).

The structure and topology of the bipolar transistor manufactured by BID technology is shown in fig. 1.9, where 1 - base *p*-type of conductivity, 2 - insulating area of p^+-type, 3 - collector area, 4 - area of base, 5 - emitter area of n^+-type.

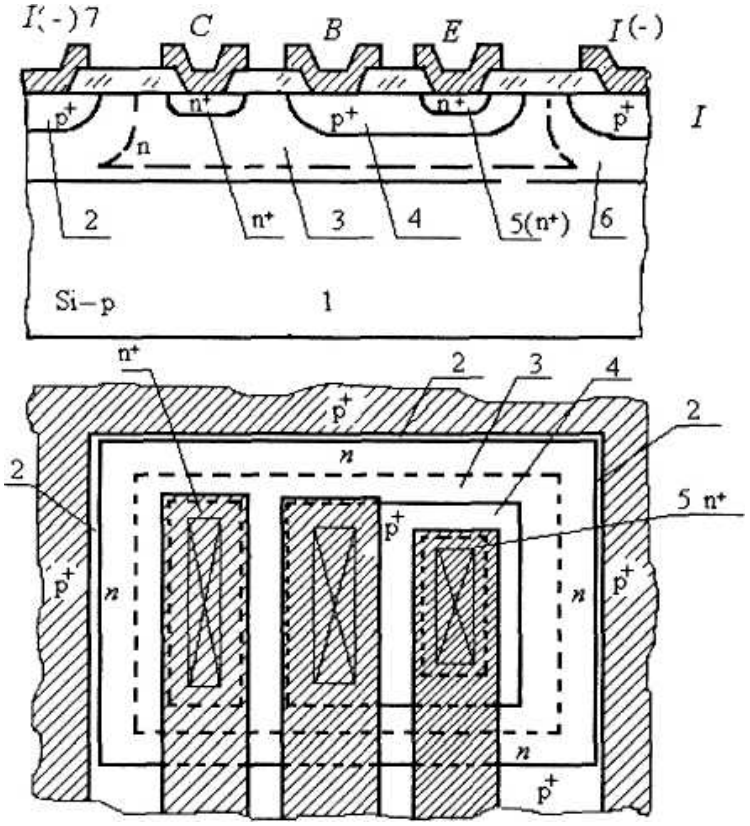


Fig.17.9.The structure and topology of the bipolar transistor manufactured by BID technology

Isolation of transistors performed functionally. To ensure reliable isolation of transistors insulating region 2 is connected to a negative pole, and the epitaxial layer of n-type to positive pole of an additional source of bias. The region of space charge (SCR) of insulating p-n-transition expands to docking with SCR p-n- transition, basis – epitaxial layer of n-type 6. After that they form new insulated area of collector 3.

17.3.4. Planar-epitaxial technology with buried p-layer and insulation with double diffusion

In considered planar epitaxial technology with buried n-layer and insulated inversely biased *p*-*n*-junction the depth of separation of p^+-diffusion must be large enough, another way divided area won't be able to pass through the epitaxial layer of *p*-type and get foundation of *p*-type (Fig.17.2e).

With great depth of diffusion also significant lateral diffusion under the edge of the window is occurred. So dividing p^+-area is wide enough and will cover a large area on the chip. If the thickness of the epitaxial film is 10 microns, the total width of p^+-area of the plate's

surface equals 30 ... 50 microns. As a result, size of transistors increases and degree of integration is reduced. Insulation of elements of IC with double diffusion significantly reduces the area of delimited plate and crystal.

The scheme of technological process of manufacturing bipolar circuits insulated with double diffusion is shown in Fig.17.10. The basis is silicon of *p*-type conductivity with oxidized surface. First photolithography is carried out in the separating areas locations and it opens the windows in SiO_2 to create deep p^+-regions (Fig.17.10a). To create areas of depth p^+-type diffusion a local acceptor impurity is performed (Fig.17.10b).

During the second photolithography in the required places it opens windows in SiO_2 for creating deep n^+-areas (Fig.17.10c). Through SiO_2 mask a local diffusion of donor impurities is performed to create deep regions n^+-type (Fig.17.10d).

Silica is removed from the surface of the plate, then chemical cleaning and building up of thin (about 2 microns) monocrystalline epitaxial silicon layer n-type is performed (Fig.17.10e). During the accumulation of epitaxial layer from the areas of p^+-type and n-type the diffusion of acceptor and donor impurities is occurred in stackable monocrystalline layer. Areas will grow both in vertical and horizontal directions.

Silica film is applied on the surface of epitaxial film. During the third photolithography, windows are opened to create insulating regions of the transistor. Insulating *p*-type regions are created with local diffusion of acceptor impurities. Diffusion is performed nearly half the thickness of the epitaxial layer (Fig.17.10f).

In these high-temperature technology operations of diffusion and oxidation counter diffusion of acceptor impurities in the epitaxial layer of *n*-type regions of deep p^+-type regions and upper insulating *p*-type regions occurs until their closure.

During fourth photolithography opening windows occur to create contact regions to deep *p*-layer. Donor impurity diffusion is conducted near half of thickness of the epitaxial layer of *n*-type. Created by diffusion n^+-type regions surround *n*-type regions as a ring (Fig.17.10g).

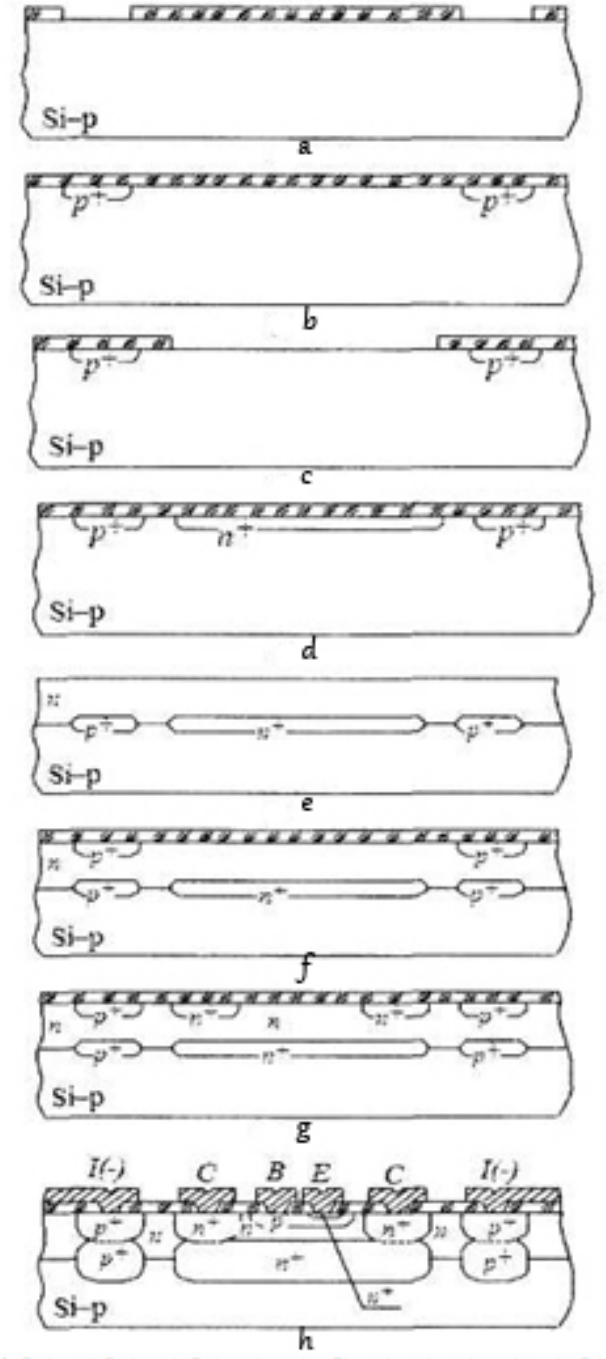


Fig.17.10.The scheme of technological process of manufacturing bipolarcircuits with isolation element method of double diffusion

During the following technological operations bases and emitters regions are created with diffusion or ion implantation.

Layer of metallization is put on surface and interconnect conductors formation is performed. During performing high-temperature technology operations of diffusion and oxidation counter diffusion in epitaxial layer of *n*-type regions of deep n^+-type regions and upper insulating n^+-type regions occurs until their closure (Fig.17.10.).

When all technical operations are performed, buried p^+-type areas will close up with surface isolated regions *n*-type, creating reliable circular insulating areas around the transistor. Contact area of n^+-type collector will close up with buried n^+-region. As a result, local isolated from all sides and from the base *n*-type epitaxial area(collector) will formed, which provide base and emitter areas of transistor structures. The n^+-type areas isolate the transistor structure and create high-alloy contact areas to the collector, which reduces the collector series resistance and improve the electrical parameters of the transistor.

In this technology the square of insulating areas is significantly decreased, but it increased the number of technological operations, including very crucial operations, so that complicated manufacturing process.

The structure and topology of the bipolar transistor manufactured by the epitaxial planar technology with buried n^+-layer and with insulation of double diffusion method are shown in Fig.17.11, where 1 - base of p-type conductivity, 2 - circular buried insulated p^+-type area, 3 - surface circular insulated p^+-type area, 4 – buried n^+-type layer, 5 - circular contact area of the collector n^+-type to n^+-buried layer, 6 – emitter n^+-type area, 7 – base n^+-type area.

The relative complexity of the technological processes is characterized with manufacturing operations of photolithography and basic manufacturing operations. Comparative data for the considered technological processes is shown in Table 17.4.

The main advantages of considered technological manufacturing processes of bipolar circuits with isolated elements of inversely biased *p-n* junctions are: relatively simple technological process, the absent of crucial technological operations, a high percentage yield of fit chips. The disadvantages include: large size of the bipolar transistor, a limited degree of integration, the long duration of the process cycle and insufficient insulation of elements of chips, manifested through the reverse saturation current of p-n-transitions. Besides, consumption current from the power supply increases with large squares of insulated p-n junctions.

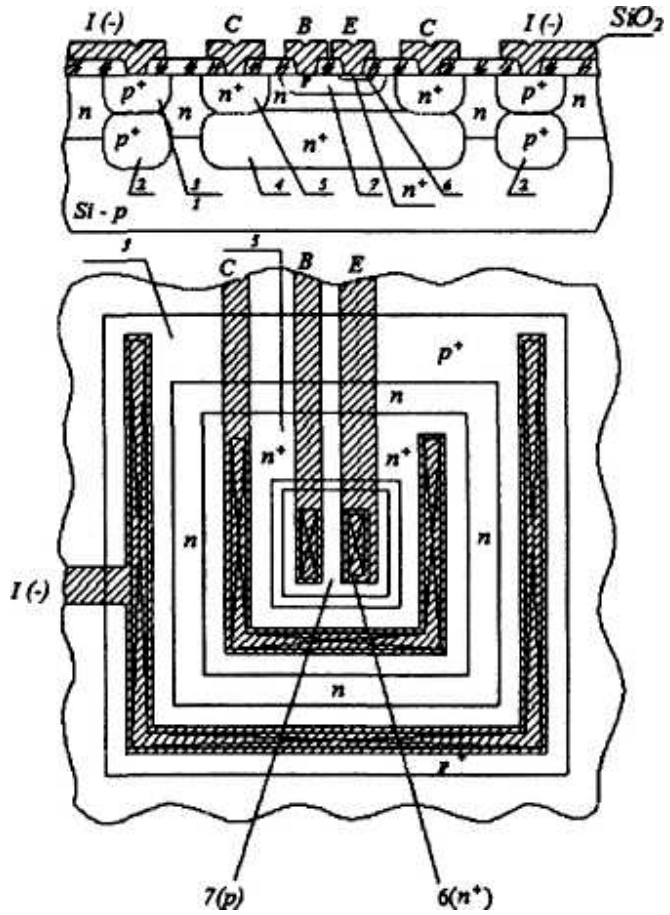


Fig.17.11.The structure and topology of the bipolar transistor manufactured by the epitaxial planar technology with insulation with double diffusion.

Table.17.4

Comparative characteristics for considered manufacturing technologies of bipolar circuits

Technology	Number of photolithography operations	Number of basic operations
Epitaxial planar technology with inversely biased insulated p-n transition	6	21
Epitaxial planar technology with double diffusion isolation	8	24
Epitaxial planar technology with isolation of collector isolated diffusion	5	20
Epitaxial planar technology with isolation of base isolated diffusion	4	15
Epitaxial planar technology with three photomask formation	3	11

17.4. Technological manufacturing processes of bipolar circuits with dielectric insulation elements

Insulation elements with dielectric provide better electrical parameters of chips.

According to the manufacturing technology of bipolar IC with dielectric isolation it provides for the establishment of crystals in which every element is fully isolated with dielectric layer. Depending on the material used for insulation and methods of possible technological implementations those technological processes are developed:

- microplanar epitaxial (planar) technology with dielectric isolation and poly-silicon application (EPIC-technology);
- microplanar epitaxial (planar) technology with glass isolation, signals or ceramics;
- microplanar epitaxial (planar) technology with insulated V-shaped grooves created by anisotropic etching of silicon (VlP-technology);
- microplanar epitaxial (planar) technology “silicon on sapphire”(SOS)
- microplanar epitaxial (planar) technology “silicon on dielectric”(SOD)

17.4.1. Microplanar epitaxial (planar) technology with dielectric isolation and poly-silicon application (EPIC-technology)

Circuit elements in this technology are isolated with dielectric film (silica) by using polycrystalline silicon as a material bearing structure. The scheme of technological manufacturing process of integrated circuits for EPIC-technology is shown in Fig. 17.12.

The starting point is a silicon wafer of n-type conductivity, which is applied to the surface of single-crystal epitaxial layer of n-type conductivity (Fig.17.12a). The surface is precipitated with epitaxial film silicon nitride (SiN_4) and then first photolithography is performed. Windows in silicon nitride film are opened for the future insulating region (Fig.17.12b)

Microstructures are created on the surface of the plate. Chemical or plasma-chemical etching of silicon to a depth of 15 m is performed through the open windows in SiN_4 (Fig.17.12c). Silicon nitride film is removed from surface of the plate, it is performed a chemical cleaning of surfaces and thermally grown or precipitated silica film with thickness of about 2 microns (Fig.17.12d).

High-resistance poly-silicon layer with thickness about 200 microns is applied to the surface of the oxide film (Fig.17.12e). The opposite surface of the plate is polished to the surface of the poly-silicon in isolated areas (Fig.17.12f). After making the mentioned technology

operations in poly-silicon base isolated with dielectric monocrystalline areas n^+-n-type are created. Isolated layer of silica replaced by silicon nitride or double layer of SiO_2 - $Si_3N_4$to improve the insulation elements.

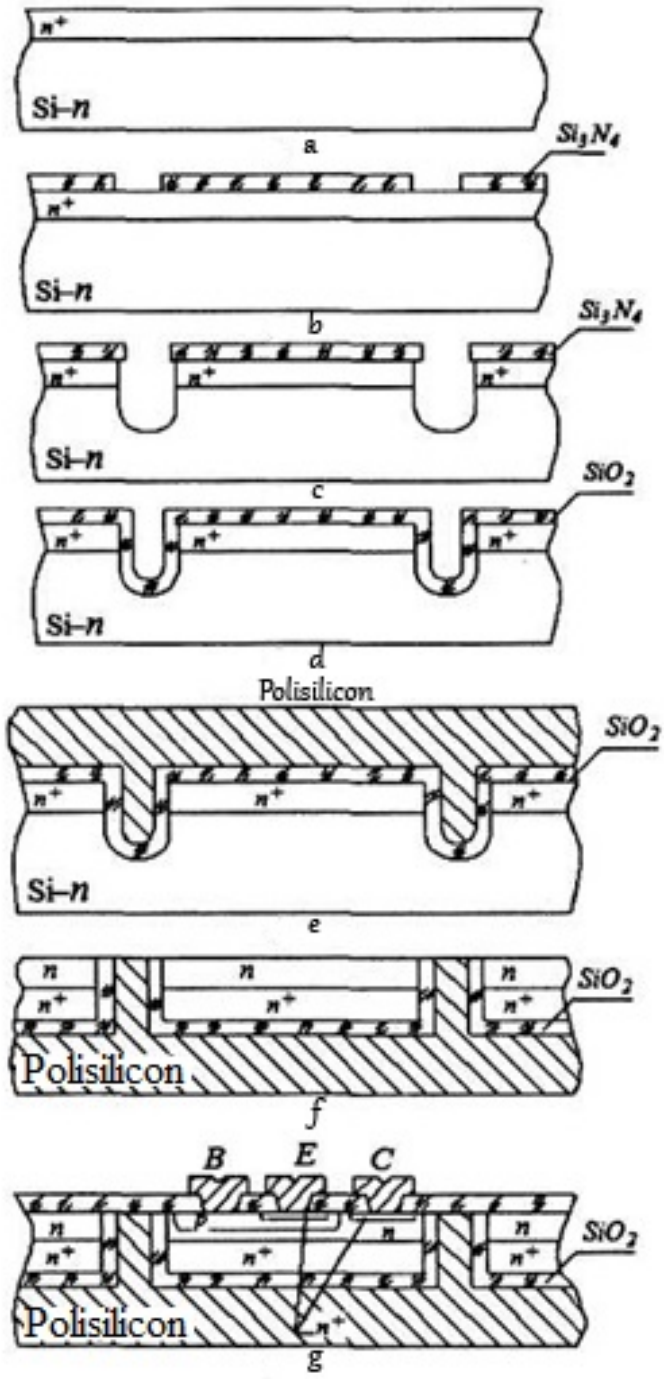


Fig.17.12. The scheme of technological manufacturing process of bipolar ICs for EPIC-technology.

Base areas are created with diffusion or ion implantation in isolated areas of n^+-n-type during the following technological operations. Then emitter areas are created and metallization is performed and interconnecting conductors systems are formed (Fig.17.12.g). The considered technological process is characterized as micro-epitaxial planar. The topology of the transistors

isn’t different from the topology of planar-epitaxial transistor with buried n^+-layer (see Fig.17.5, which insulated p^+-areas are replaced with dielectric areas of silica and are filled with poly-silicon. This construction hasn’t conductors that serve the voltage on the isolated areas. Along with the considered process it is developed some modifications, which enable to create complementary bipolar integrated circuit (Fig.17.13). Silicon of *p*-type conductivity is used as a basis for manufacturing circuits. Silicon surface is oxidized, photolithography is performed, and windows in SiO_2 areopened for deep diffusion (Fig.17.13a).

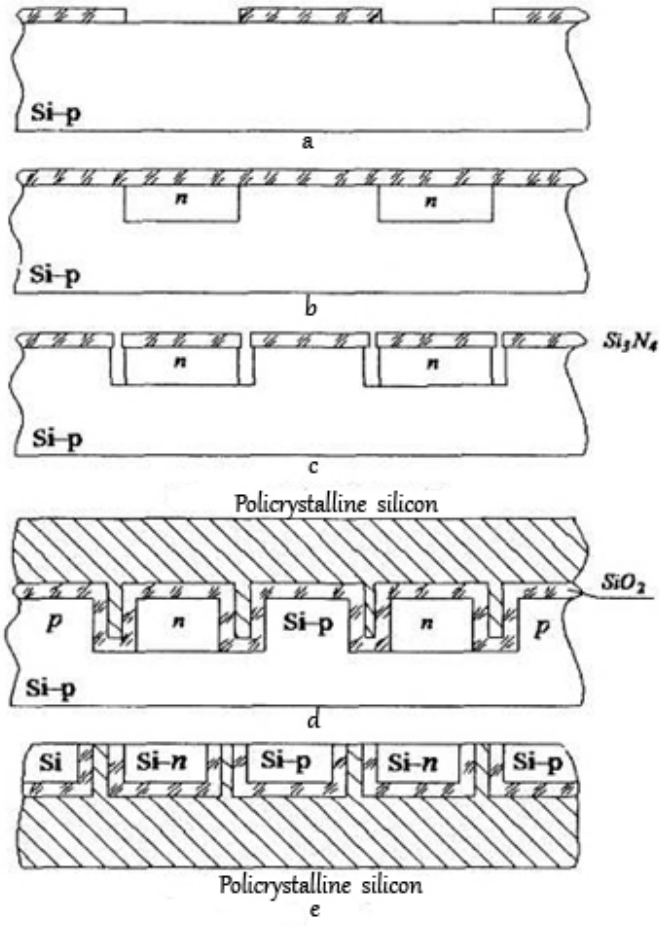


Fig.17.13. Scheme of technological manufacturing process of complementary bipolar ICs by EPIC-technology.

Donor impurity diffusion is performed to a depth of 15 microns (Fig.17.13b). Silica is removed and silicon nitride is precipitated on the surface of the plate. Photolithography is performed, windows are opened for separation areas and deep removal of silicon from disadvantaged areas is performed (Fig.17.13c). Silicon nitride is removed from the surface of the plate and oxidizing is performed. From the opposite side of the plate it is performs polishing to monocrystalline areas of *n*-and *p*-type conductivity, which make *p*-*n*-*p*-and *n*-*p*-*n*-transistors (Fig.17.13d).

Integrated circuits are made by EPIC-technology, have good electrical parameters that are resistant to radiation, but are expensive. ERIS technology is quite complex and time consuming. Disadvantages are the low-scale integration chips produced by this technology.

17.4.2. Microplanar epitaxial (planar) technology with glass isolation, signals or ceramics.

To reduce the disadvantages of the previous technology developed technology for manufacturing bipolar circuits, in which polycrystalline silicon replaced with glass, sital or ceramic. This type of processes is generally called "silicon on insulator". Scheme technological process of manufacturing microchips insulated elements with glass is shown in Fig.17.14.

Integrated circuits are created in a silicon wafer of *p*-type conductivity, which is applied to the surface of a two-layer epitaxial structure: layer of n^+-type and layer of *n*-type (Fig.17.14a). In epitaxial layer of *n*-type by planar-epitaxial technology transistor structures are created (Fig.17.14b). Silicon nitride layer is applied to the plate with the created elements, photolithography is performed and then windows are opened in sites of future isolated areas. Through the windows deep silicon etching is performed. Microstructures occur on the surface of the plate, each of which there is a transistor structure (Fig.17.14c).

Silicon nitride is removed from the surface of the plate. Plate from microstructure is pasted to a supporting plate (Fig.17.14d) and on the opposite side it is polished to the creating microstructure (Fig.17.14e). On the design side that is an opposite to the supporting plate, put insulating dielectric, thickness of about 200 microns (Fig.17.14f).

Then the supporting plate is removed, silica is applied to the surface of transistor structures windows are opened for contacts to areas of transistors metallization is performed and conductors are formed (Fig.17.14g).

Topology of the transistor structures by this technology is similar to the topology planar-transistor structures discussed above.

17.4.3. Microplanar epitaxial (planar) technology with insulated V-shaped grooves created by anisotropic etching of silicon (VIP-technology)

In the above technological manufacturing processes of bipolar ICs size of insulated areas around transistors depends on the chosen etching method of monocrystalline silicon. If isotropic chemical etching methods are used, the size of separating insulating areas will be large, simultaneously the density constrain elements on the chip will decrease.

In order to reduce the square of insulating areas formed during etching of monocrystalline silicon, it is developed methods and technologies of vertical anisotropic etching.

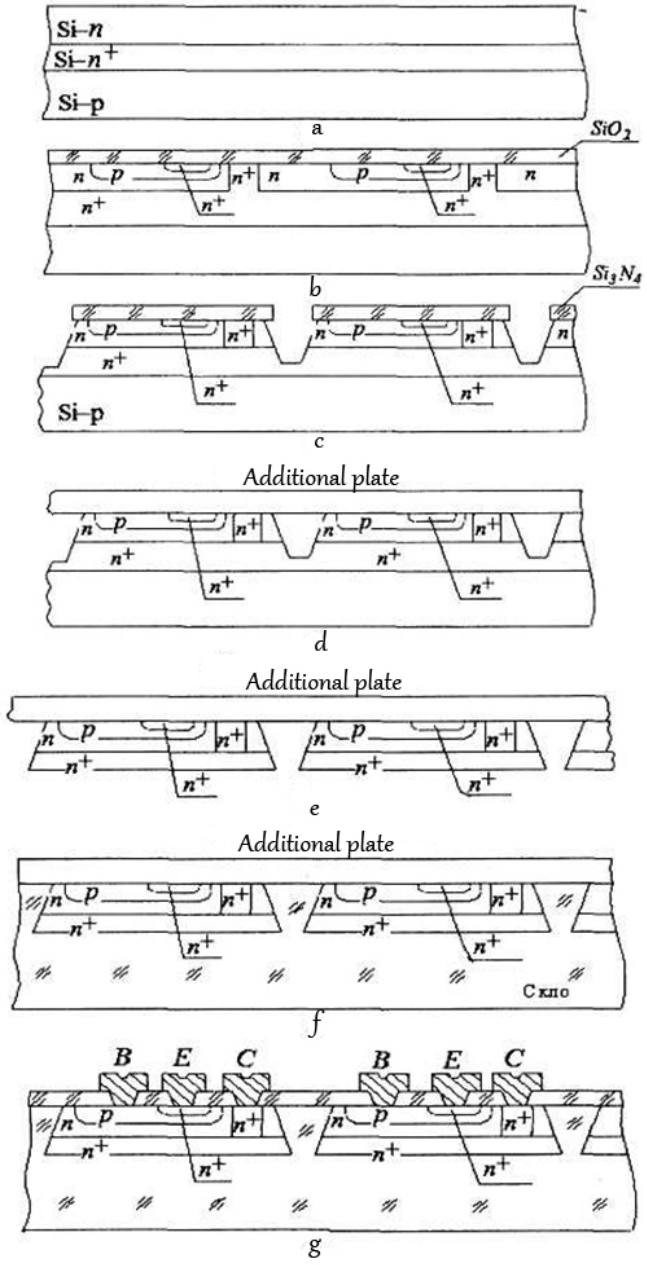


Fig.17.14. The scheme of technological manufacturing process of bipolar circuits with glass insulated elements

The method is based on the fact that the etching of silicon with crystallographic plane orientation (100) is faster than the crystallographic orientation (111). Therefore, the separation areas become V-shaped form and have a square less than the same for ERIS technology. The scheme of the technological process is shown in Fig.17.15.

Monocrystalline epitaxial layer of n^+-type conductivity silicon nitride film is applied to the silicon wafer *n*-type conductivity. Photolithography is performed; windows are opened for

insulating areas (Fig.17.15a). Anisotropic etching of silicon is carried out through a window in the silicon nitride. During the etching of silicon V-shaped groove is formed sidewalls which have crystallographic orientation (111) and it is placed at an angle of 54.74 ° to the surface of the plate that has the orientation (100). The depth of the groove L depends on the width of window: (Fig.17.15b):

$$L = \frac{W}{V - 2}$$

Silicon nitride is removed from the surface of plate, chemical cleaning plates is performed and thermal oxidation is conducted. The surface of the plate and the sidewalls of the groove are coated with silica. Poly-silicon layer wit thickness of about 200 microns precipitated in oxide film (Fig.17.15c). The opposite side of the plate is polished to the vertices of V-shaped grooves (Fig.17.15d). Monocrystalline n-n+-type areas are formed on the surface of the plate, isolated on polycrystalline silicon substrate and are isolated apart thermally extension silica. Polycrystalline silicon provides the mechanical strength of the integrated circuits. Polycrystalline silicon is well coordinated by the temperature coefficient of linear expansion of monocrystalline silicon; therefore it can endure high-temperature technological operations of the next manufacturing cycle.

Critical technological operations for all considered manufacturing processes of bipolar circuits with dielectric insulation is a polishing silicon wafers to the created transistor structures (to opening the tops of V-shaped grooves). If to stop polishing before, the n-areas will remain electrically interconnected. If to continue polishing longer than it is necessary, the n-areas will become too thin or even disappear completely.

Epitaxial layer of n^+-type that was on the plate's surface of *n*-type, is now at the bottom of the insulated areas of *n*-type and serves as the deepening layer that is designed for reducing the series resistance of the collector n^+-*p*-*n* transistor. The following technological operations of the manufacturing of integrated circuits is performed the same sequence as the operations in manufacturing chips with insulated elements inversely biased *p*-*n*-junction (Fig. 17.15d).

Manufacturing circuits with dielectric insulation is more expensive than circuits with isolated *p*-*n* junctions. But dielectric insulation has more advantages in the production of high-voltage radiation-resistant integrated circuits.

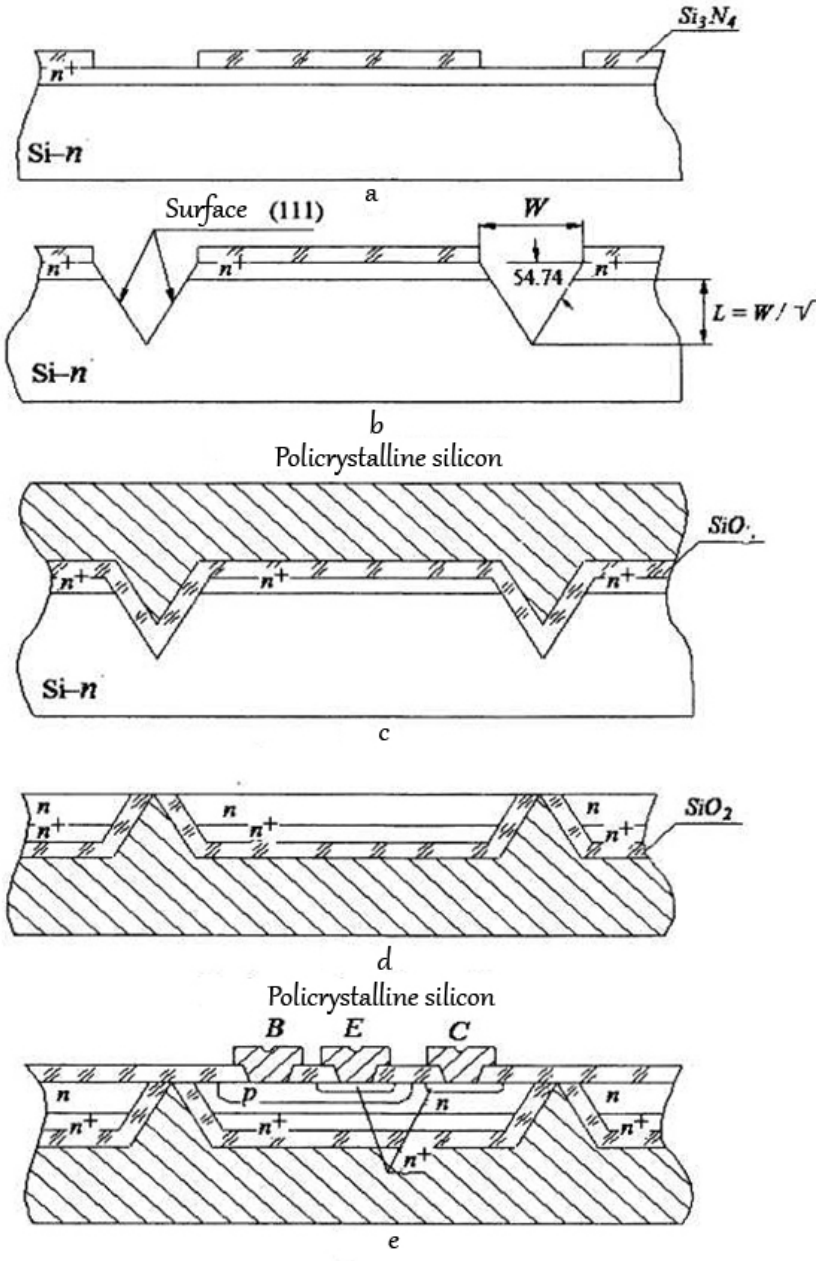


Fig.17.15. The scheme of technological manufacturing process of bipolar circuits with dielectric insulation using anisotropic V-grooves etching

Dielectric resistance silica is about 600 V/micron. Therefore, the layer with two oxidized thickness of 0.5 microns has breakdown voltage between isolated p-area and basis about 300 V. in circuits with isolated p-n-junctions breakdown voltage between the p-region and basis don't exceed 50 V. Therefore, using of dielectric insulation allows creating a high-voltage transistors and diodes.

Ionizing radiation (X-rays or gamma rays) creates a large quantity of excess free electrons and holes in silicon. X-rays have energies above 100 eV, and gamma rays –has energy more than 100 keV. Generating electron-hole pairs in silicon requires energy 1.1 eV. Therefore, it is clear

that such radiation can generate a large number of free electrons and holes. These excess carriers significantly increase the reverse current of *p-n* junctions. In circuits with isolated *p-n*-junctions significant leaps of current occur over the influence of ionizing radiation pulse. Chips with dielectric insulation are more resistant. It provides by dielectric layer between the *n*-type areas and the polycrystalline substrate.

Integrated circuits insulated with glass have an even better insulating capabilities and higher electrical parameters.

17.4.4. Microplanar epitaxial (planar) technology “silicon on sapphire” (SOS)

Considering requirements of modern circuits this technology belongs to the perspective, but the high cost of chip production limits its use. Fig.17.16 shows the structure of the bipolar transistor manufactured by this technology.

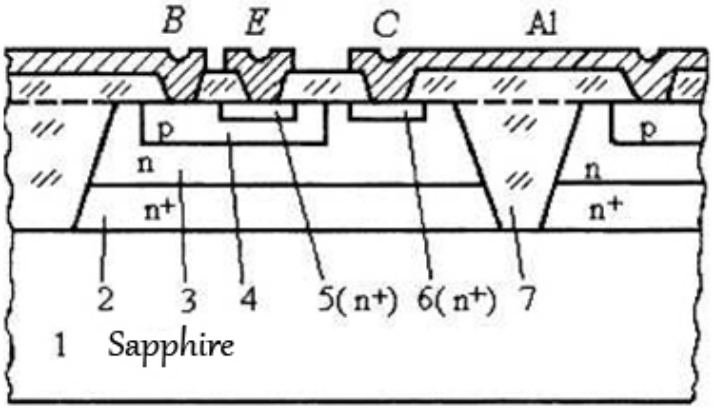


Fig.17.16.The structure of the bipolar transistor, made by "silicon-on-sapphire" technology (SOS)

On sapphire basis 1 epitaxial monocrystalline layers of n^+-type conductivity 2 and n-type conductivity 3 are grown. The thickness of the layers does not exceed 1-2 microns. Monocrystalline islands of the double-layer structure (microstructures) isolated between each other with two oxidized isolated areas 7 create local oxidation of epitaxial films all over the depth. There are so formed areas in each of islands by planar technology: base - 4, 5 - emitter, collector contact to the area 6 and switching conductors.

Significant mechanical tensions arising at the interface of silicon - sapphire through their crystal lattice mismatch and dissimilar coefficients of linear expansion of the silicon layer that has thickness of about 1 micron. It leads to structural defects. As a result lifetime of minority

carriers in silicon at sapphire basics significantly reduces. This phenomenon may be one of the factors limiting the use of such technology for the production of bipolar circuits, but it is widely used for the production of metal - insulator - semiconductor (MOS) integrated circuits. This technology is especially useful for the production of complementary MOS (KMDN) chips, which will be considered in the next section. Using dielectric foundations reduces parasitic capacitance and consequently it increases the speed of the chips.

17.4.5. Microplanar epitaxial (planar) technology “silicon on dielectric”(SOD)

Silicon-on-insulator technology makes it possible to apply a thin layer of monocrystalline silicon on surface of twice oxidized film. On the surface of the first monocrystalline layer put the second, and in a two-layer structure it is created a bipolar structure. Some typical production operations of this technology is shown in Fig.17.17.

High-quality film of silica is thermally increased on the surface of the silicon wafer. In silica windows open photolithography, that passes across the plate or across confined areas surrounding the islands of SiO_2 (Fig.17.17a). Chemical vapor deposition of silicon is performed from gas phase. On the surface of monocrystalline silicon film is built up in the grooves and on the surface of silicon dioxide polycrystalline film is built up (Fig.17.17b). In the film directed recrystallization is performed using a scanning laser or electron beam or resistive tape heater. Those fields of deposited film in contact with silicon-base act as crystallization centers from which recrystallization of polycrystalline film begins. Along with the movement of the heating zone monocrystallization of silicon films is distributed by crystallization centers to areas of the film that are above film of silica. As a result of recrystallization on the surface of the plate a thin layer of monocrystalline silicon is created, which in the following technological operations is applied to monocrystalline layers of required type and level of conductivity (Fig. 17.17c). Technological operations of deep oxidation of the entire thickness of two-layer epitaxial structures to silica layer are used to isolate elements (bipolar transistors, resistors, etc.). Areas of base, emitter and contact to the collector and system of conductors are formed in planar technology microstructure. (Fig.17.17)

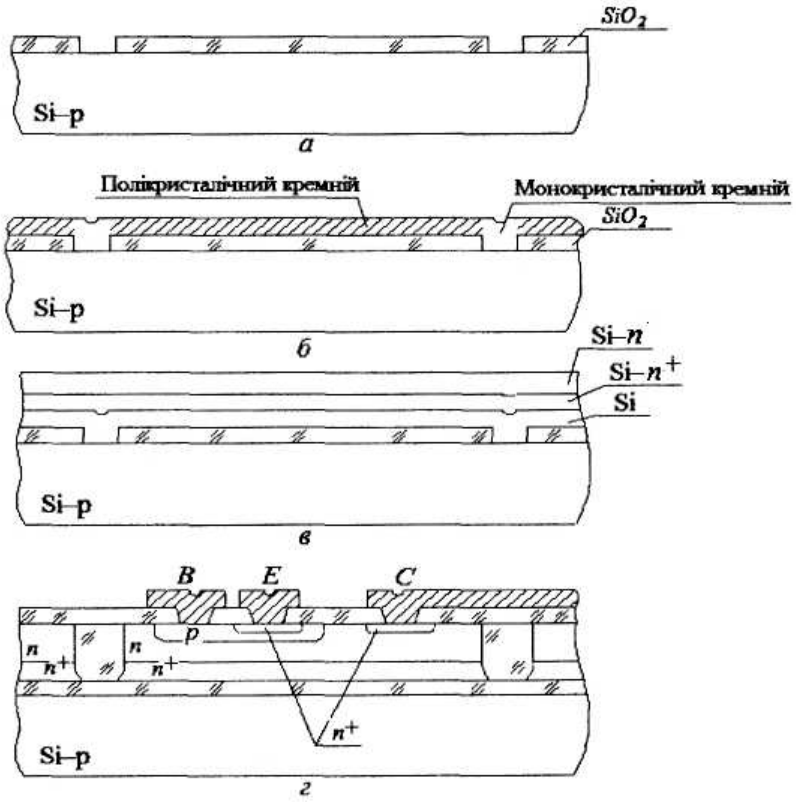


Fig.17.17. The scheme of technological process of manufacturing bipolar circuits on the "silicon on insulator".

A variety of technologies "silicon on insulator" is a horizontal building up epitaxial technology. As for SOS technology on the surface of monocrystalline silicon wafer, silica layer is thermally increased where windows are opened which they are shaped grooves across the plate or closed contours around areas of silica. On the surface of the plate silicon is precipitated from the gas phase, followed by silicon etching gas. On the surface of monocrystalline silicon films are created in the grooves and on the surface of silicon dioxide polycrystalline films are created. Mainly poly-silicon is removed at gas etching. After one cycle gas etching completing, areas of monocrystalline silicon grown on the free surface of SiO_2 leaves on the surface of the plate. Such deposition – etching cycles is repeated several times. In subsequent cycles monocrystalline silicon islands start to grow in a horizontal direction, covering the oxidized surface. Since polycrystalline silicon in each cycle is removed, and then later high quality monocrystalline film is created. Areas of monocrystalline films that are located above silica film are epitaxial hetero-structures, but such parameters as mobility and lifetime of carriers are the same parameters of homo epitaxial layers. So this technology of creation of high-quality monocrystalline layers is used to make bipolar and field-effect IC. Ion implantation of oxygen to the desired depth is used to create structures "silicon on insulator" in modern technological processes of production of

integrated circuits. After thermal processing in depth of semiconductor wafer a continuous silica layer is created that isolates the monocrystalline silicon layer on the surface of monocrystalline base. The surface layer of monocrystalline silicon is divided with deep oxidation into separate pockets, in which creates circuits element are created.

Designs of bipolar transistors with dielectric isolation are differed from structures of insulated inversely biased p-n-junction so that the transistors are placed in microstructures isolated from all sides and from the bottom dielectric layer. The quality of this insulation is much higher. Currents penetrations through the dielectric are much less than through the *p*-*n*-junction at its reverse bias. Since the dielectric constant of silicon dioxide is four times smaller than silicon, the specific capacitance of the dielectric insulation is also smaller.

Technology of manufacturing of bipolar IC with full dielectric isolation is used for the manufacture of integrated circuits for small and medium degrees of integration, which have the special requirements of radiation resistance and electrical insulation at high frequencies, or to create a high degree of integration of ICs with good insulation elements.

The considered technological processes of production of bipolar circuits with dielectric insulation elements are perspective for analog integrated circuits (differential and operational amplifiers) and microwave integrated circuits in which current of penetration should be small. They are used them in the manufacture of integrated circuits for special systems with increased radiation resistance.

Chapter 18. Reliability of IC and microsystem devices

A.Shkavro

18.1 Introduction

Any electronic item is characterized by its quality and reliability. Product quality, particularly of an electronic item, is a total scope of the properties that distinguish one from another, and define its fitness for the use by appointment. Reliability is one of the quality indicators that characterizes the permanence and stability of other indicators.

It is possible the appearance of some percentage of defective items during the production, i.e. those that cannot be used by appointment. Percentage of serviceable ones among the total quantity is called *the serviceable items output*. It depends on a number of factors, particularly on the item complexity and level of technique. The serviceable product output influences directly their prime cost and hence the expedience.

In general, applying IC and increasing degree of integration justified the expectations as to raising the reliability of the systems based on them. However, increase of the integration degree influences their versatility and yield.

The optimal size of IC with respect to the quantity of the functions performed by the system is determined by several interrelated factors: the division of the system (subsystem) on the circuits, the cost of pressurization and assembling, general reliability of the full system.

Cost of the scheme, based on the large integral circuits with relatively low percentage of the serviceable items output, could be significantly smaller than the total cost of the pressurization of huge amount of the simpler crystals with higher yield percentage, and theirs setting up on the board. Wherein, quality and reliability of the set-up are improved uniquely on account of reducing of the shells, connections, signals delay, and mutual interactions in the board conductors. Choice of the optimum division of the scheme on the separate IC is a complex task, to solve which it is necessary to have the possibility to forecast the yield of serviceable crystals and cost and reliability of IC depending on size, forecast general reliability of the scheme.

18.2 Factors that influence on the yield of serviceable IC

During manufacturing the IC, even one board contains few chips that could be defective. Besides, defects could appear during further technique stages – setting up the crystals into the shells, pins connection, pressurization. In practice, the amount of serviceable schemes could vary

from near 100% to one or few schemes per board. Commonly, the causes of the yield reduction could be divided into 3 classes: technological factors, design factors, and occasional point defects.

18.2.1. Technological factors

At the figure 18.1 there is shown the photograph of the board with crystals (chips) IC that was tested. Defective crystals are marked with paint dots. As could be seen, there are areas with very high content percentage of serviceable crystals and areas with relatively low percentage of those crystals. The appearance of defective chips is determined by the series of factors, particularly, technological. These factors include: deviation of the thickness of the oxide layer and multi-crystalline silicon, deviation of the resistance of the implanted layers, error of a measurement of the element size during lithography, formation of the scheme's topography, and errorswhile combining photo-mask with scheme topology which was formed during previous stages.Many of these factors are interrelated. For example, in the area where the multi-crystalline silicon layer thickness is less than average value, etch depth is too high if etching time is chosen based on the thickness of multi-crystalline silicon layer that is more than average value.

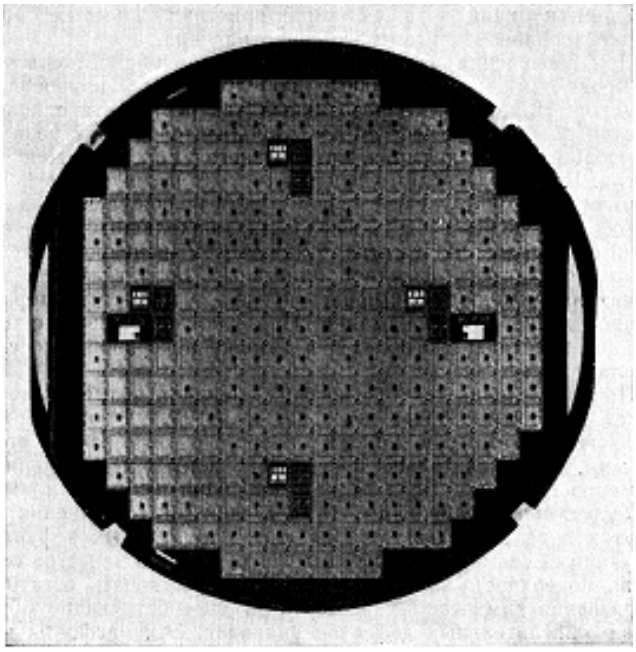


Fig. 18.1.The photography of the IC board where in the defective crystals are marked with points. (Crystals that contain small number of elements are the trial modules.

In the areas where multi-crystalline silicon layer thickness is less, the gates of MOS-devices have the less size. It leads to the too small length of the conducting channel of MOS-transistor resulting that transistors are not disabled when corresponding voltage is applied to the

gate. Thus, the scheme functioning could be violated, or output current could be enhanced overly.

The deviation of the doping level and implanted levels could lead to the change of the resistance of the contact to implanted layers; and deviation of dielectric layer thickness – to the change of the size of contact window. Both these factors could be the cause to functioning failure at the lanes that are characterized by the value of the contact resistance.

There are small but critical changes in the board size during different processing operations. For example, SiO_2layer created by the board oxidation has the volume that is twice more than those of Si used for receiving of the oxide layer. After oxidation the board consists of inner silicon layer having stretching stress, and oxide layers on the both surfaces having internal compression stress. The diameter of the oxygenated board is more than diameter of the primary board. If the level of interior stress is more than breaking point then the deformation occurs. While removing the oxide from one side, the board bends at the side of the surface with the oxide layer.

The board size can change more than of $2 \cdot 10^{-3}\%$ while the technological operations take place. So, the board with diameter 125 mm could change it on 2.5 μm that is more than permissible deviation. If such deviations are not countervailed, the errors of combining would cause the appearance of areas that are occupied by idle schemes. Besides that, under poor conditions of clearance, on the surface of the board could be present residual contamination by chemical reagents that contribute to the formation of oxide defect of the shell. These defects could result in excessive output current and further scheme failure.

Due to improvement of the technological procedures, the influence of many of these factors could be reduced or removed, however it is possible the appearance of new reasons to scheme failure.

18.2.2. Factors of scheme projection (design factors)

Some board areas could have low yield of appropriate devices not only due to the excess of established deviations of device parameters associated with the manufacturing technique, but due to the fact that during scheme design there are not taken into account possible deviations of the device parameters and correlation between those.

For instance, during MOS-scheme design, the most important parameter is threshold voltage $V_T$and conductive channel length$L$ of MOS-transistors. Deviations of substrate doping level, ion implantation dose, gate oxide thickness result in the changes of threshold voltage. Deviation of gate length and depth of source-drain transition result in the changes of the conductive channel length. Usually, threshold voltage and conductive channel length are not

interrelated. However, as a rule, the scheme speed is increased while decreasing these parameters. The scheme performance are often simulated for two conditions: high speed (low values of $V_T$and L) and low speed (high values of $V_T$and$L$). Besides that, the scheme performance have to be simulated at low value of $V_T$and high of$L$, and high value of $V_T$and low of$L$. While scheme design, other scheme parameters deviation have to be considered, for instance, the resistance of areas implanted, capacitance between conductive layer and substrate, contact resistance, and leakage currents.

Two schemes with the same nominal size and element composition, received using the same technique, could differ considerably by yield of appropriate items. Low yield of appropriate devices in this case is explained by the fact that during scheme design, its sensitivity to changes of device parameters is not taken into account.

18.2.3.The Point Defects

As a rule, the yield of appropriate schemes differs from 100% even in case if all deviations of technological parameters are within acceptable limits. Commonly, the cause of yield reducing existence of point defects on board. There, point defects are defined as an imperfect board area which size is less compared with the crystal size. As an example, let’s have a look on the crystal 2000 × 2000that consists of elements which size is compared with 2 mcm. The dust particle with diameter of 200 mcm could cause the break of metallic conductor. Similarly, the dust particle with diameter of 200 mcm could cause detachment from crystal of large area of metallic layer. Both these particles are considered as a cause for point defects appearance. At the same time, the board area that contains several crystals without metallization is not considered as the point defect.

The appearance of point defects could be associated with heterogeneity of original silicon board. To identify such dependency, one could analyze the distribution of defected crystals through the board; and thus receive more exact information, one also could analyze the correlation of specific causes for crystal defectiveness with theirs disposition through the board.

Despite defects caused by output board defectiveness, there are many other types of technological defects. One of the most common reasons is dust and other particles from the environment. They could reach the board surface during its movement through the technological region, or could be implemented during its sedimentation. Solid particles could be present in resist solutions and could sediment during the lithography process. Moreover, it is possible the adhesion to the board surface of silicon particles that split off the board while manipulating during technological operations. Isolated oxide defects of the shell that cause the increase of

output current and the scheme failure, also could be considered as the reason of point defects, as well as isolated hills on epitaxial layers or breakdown of dielectric films.

Point defects could also be present on lithographic patters as well as on silicon boards. Dust and other particles that reached pattern blanks while manufacturing, could be the cause of formation of permanent defects on the pattern. Particles that reached the pattern during its use, cause gradual increase of the density of point defects that are present on the board. These defects could be removed during periodical depuration of the pattern surface.

18.2.4. Ways to increase the yield of appropriate crystals

The main measures that assist to increase the yield of appropriate crystals are measures aimed to reducing and removing the influence of factors listed: technique improvement, design of scheme topology that is resistant to random deviations of material parameters and geometry, reducing the density of point defects. Particularly, there is effective to use input control of the board quality, materials, equipment, premises cleanliness, and inter-operational control. This control could be performed by the assessment of the scheme quality during all stages of its manufacturing, and be implemented with Scanning Electron Microscope. If any change in the defects density is detected during specific operation, then there should be taken the appropriate measures for the correction of the parameters of technological process. The defects density control could be performed with the use of special operations. Boards could be etched for removing all films, and silicon substrate is handled for detection of shells' defects which density is controlled later. Topology could be formed on the other boards using the special patterns that allow to carry out the electrical measurements for detecting the defects of dielectric breakdown. It is necessary to introduce new methods to control these defects, while new types of defects are detected.

In some cases reservation could be applied for the increase of the appropriate IC yield. In particular, the reservation is effective if applied to the schemes with regular structure, for example to the memory circuits. On the crystals of these schemes a fragments could be connected to substitute the defected ones. Sometimes, crystals with structured fragments cannot operate in full scope due to failure of separate fragments, however they are appropriate for the use in more simple schemes, or together with another scheme that contains defective fragments that are pressurized and marked as a certain device. During some period of technology development, the manufacturing and using the micro-schemes with limited functions is economically proved. For example, memory schemes, different variants of simplified processor. It is clear, that schemes, which are able to serve given functions, could differ by the value of some parameters, for example, maximum working frequency, admissible voltage or current,

output current, noise factor, etc. Apparently, it is not advisable to set maximum requirements for the parameters so culling a significant amount of devices as those that do not meet the requirements; similarly, it is not advisable to set the values of technical parameters at a low level so deliberately restricting the system characteristics. Usually, based on special tests one divide devices into groups that differ by the values of specific parameters and thus by the price.

18.3. Characteristics of IC reliability

18.3.1. General concepts and terms of reliability

As were mentioned in the previous section, any device is characterized by its quality and reliability. Device quality is defined as full range of properties that differ the device from others, and determine its suitability for the use by appointment. Reliability is one of the quality parameters that characterizes the permanence and stability of other parameters. If the device reliability does not meet the certain requirements then it cannot be considered as a qualitative.

Reliability is the device attribute to perform given functions keeping in time the values of operational parameters that meet the requirements for specific conditions of regime and terms of use, maintenance, storage, and hauling.

The fact of the reliability characteristics of objects in specific work environment and storage, for example, in high humidity or in aggressive environment are commonly indicated in technical documentation, and in the case of high radiation exposure, in general, additionally new term is used – resistance to radiation damage.

Integral circuits and separate semiconductor devices are not used apartly, they always have to work in some system. The scope of objects that act together solving some problems is called the system.

One divides systems on renewable and non-renewable. Renewable systemsare repaired after the failure and continue working. Non-renewable systems are not repaired due to technical impossibility or economical unreasonableness. There are systems that are served, and there are systems that are not served. The latest could be self-renewable, for example, by automatic backup.

It is clear, that the system reliability depends on the reliability of components. The reliability of renewable system, and non-renewable one is characterized by somewhat different indicators.

In general, the reliability is assessed by following parameters of the device: workability, durability, infallibility, maintainability, safety.

Workability – the item state, when it is able to perform given operations with parameters that are established by requirements of technical documentation.

Durability – the item property to keep the workability up to boundary state for a long time, with required breaks for the preventing service. The boundary state is defined by impossibility of further device exploitation that is caused by either reducing its effectiveness, or safety conditions that are indicated in technical documentation.

Infallibility – the item property to keep workability during some period without interruption.

Maintainability – the item ability for the prevention, detection, and removing the failures by preventing maintenance and repairing.

Safety – the item property to keep its performance indicators during and after the term of storage and transportation that is established by requirements of technical documentation.

18.3.2. Quantitative indicators of reliability

The definition given above is qualitative indicator. It is possible to speak of high or low reliability, but this characteristic will be subjective.The introduction of quantitative indicators of reliability initiated the scientific methods of the reliability research. The reliability theory is based upon the probability theory. The failure is considered as the transition from the working state to the off-, and is a random event.

While assessing the quality of non-renewable devices including semiconductor items and integral circuits, there are used the following main indicators:

$P(t)$ – the probability to work without failure, function of reliability;

$Q(t)$ – the probability of failure;

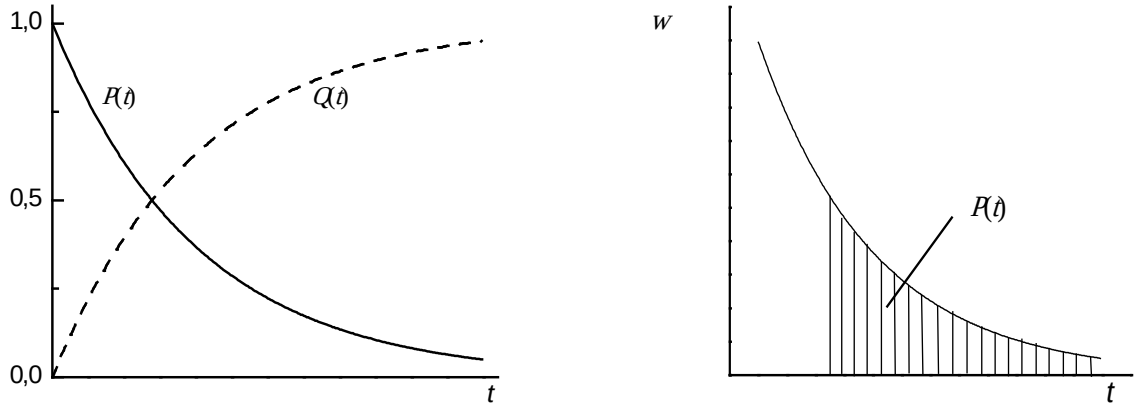
$\lambda(t)$ – the intensity of failure.

The probability to work without failure during time $tP(t)$is defined as the probability that uptime ξwill be more than t

$$P(t) = P(\xi > t) \tag{18.1}$$

The probability of failure $Q(t)$ is defined as the probability that up time will not be more than t. It is apparently that

$$Q(t) = P(\xi \leq t) = 1 - P(t) \tag{18.2}$$



A B

Fig. 18.2. The probability to work without failure and the probability of the failure (a), the density of failure distribution (b)

As for any random variable, it is possible to define the function of probability density for$Q(t)$

$$w(t) = \frac{dQ(t)}{dt} = -\frac{dP(t)}{dt} \tag{18.3}$$

Taking into account (18.3), the probability that the failure would not occur until time *t* could be written as $P(t) = \int_t^\infty w(t)dt$, that is equal to the square of the shaded area under the curve at the Fig. 18.2b. The density of the failure distribution$w(t)$is used in the reliability theory as an attribute for speed of changing the reliability of object in time, i.e. the frequency of failures.

The intensity of failures $\lambda(t)$ is a probabilistic parameter:

$$\lambda(t) = \frac{w(t)}{P(t)}, \tag{18.4}$$

that shows, what percentage of working elements will refuse in the moment *t* per time unit. Taking into account (18.2) and (18.3) one could obtain:

$$\lambda(t) = -\frac{1}{P(t)}\frac{dP}{dt} \tag{18.5}$$

$$\ln P(t) = -\int_0^t \lambda(t)dt \tag{18.6}$$

$$P(t) = exp\left\{-\int_0^t \lambda(t)dt\right\} \tag{18.7}$$

Term(18.7) is the mathematical expression of the main law of reliability. It contains the functional connection between two main reliability indicators.

In practice, often the uptime distribution in a wide time period is approaching the exponential law, and the failure intensity does not depend on time, namely

$$P(t) = e^{-\lambda t}, \tag{18.8}$$

and $\lambda = const$.

Besides described main indicators of reliability, there are used others, particularly, meantime of working without failures:

$$t_{mid} = \int_0^\infty tw(t)dt = \int_0^\infty P(t)dt \tag{18.9}$$

Substitute (18.4) in (18.9), then:

$$t_{mid} = \int_0^\infty \frac{w(t)}{\lambda(t)} dt \tag{18.10}$$

If$\lambda = const$, as it often takes place in practice,

$$t_{mid} = \frac{1}{\lambda}\int_0^\infty w(t)\, dt = \frac{1}{\lambda},$$

Namely

$$t_{mid} = \frac{1}{\lambda} \tag{18.11}$$

In general case, the experimental dependence$\lambda(t)$looks as it is shown on fig. 18.3. Due to characteristic form, sometimes it is called “the curve of reliability”, or “the reliability bucket”. Three characteristic areas could be marked here.

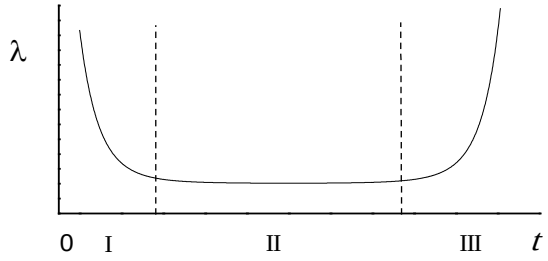


Fig. 18.3. Time dependence of the failure intensity

I – area of early failures that is characterized by high failure intensity. It is caused by the failures due to the presence of hidden defects.

II – working area that is characterized by the constant (or almost constant) failure intensity.

III – area of depreciation that is characterized by the sharp increase of failure intensity.

The curve of reliability (λ - parameter) has quite a general nature. Such kind of dependence characterizes not only the artificial products, but any object, in particular, the living organisms. That is why, the area I is sometimes called a period of infant mortality, even for the artificial devices. It is necessary to mention that for the commercial products, the area I could not be observed. It is connected with the fact that devices are tested before the production, and according to the results of testing, unreliable items are culled, i.e. they are removed from

consignment. Conversely, $\lambda$is the parameters obtained for the items that have not previously tested, and in the area I could have a peak.

Area III could be not observed for some devices, particularly for the microelectronic items, at least during the periods that are reasonable from the point of the use.

18.3.3. Laws of distribution of random variables

The failure, namely the device transition from working state to the inoperable, is random event. Respectively, time to failure of the device is the random parameter. Apparently, that values of all parameters obtained experimentally, will be the random ones. As were already mentioned, the mathematical tools for the assessment of reliability are based on the theory of probability.

The main concepts of the theory of probability.

The probability of independent event$v$, for example, that the item of the consignment of N items would be defective, is equal

$$Q = \frac{D}{N} \ , \tag{18.12}$$

if in consignment there are *D* items defective. So, the probability of occasional item to be defect less

$$P = \frac{N-D}{N} = 1 - Q. \tag{18.13}$$

When the event probability is equal 1, for example, event$v$with condition $D = N$, the event is reliable. When D = 0, the event v is not possible. In practice, there are often occur the event probabilities either near 1 (almost, reliable event) or near 0 (event is almost not possible).

If from the whole consignment of *N* items (from general set) one in the way of random selection forms picks of *n* items *d* of which are defective, then the ratio

$$q_n^d = \frac{d}{n} \tag{18.14}$$

is the statistical probability of defective item by the definition. Accordingly, the statistical probability of item's being defectless equals

$$p_n^d = \frac{n-d}{n} = 1 - q_n^d. \tag{18.15}$$

Apparently, that unlike the probabilities *P* and *Q*, statistical probabilities$p_n^d$and$q_n^d$are random parameters that are approaching $P$and$Q$at$n \to N$.

The random parameters could be either discrete or continuous. There are often used hypergeometric, or binomial, or Poison distribution for the probability distribution of appearance of discrete values. In the case of continuous random values, there are commonly used exponential, normal (Gauss), lognormal law, Weibull law etc.

One use expectation and dispersion for the characterization of random value. Take the random parameter$X$ that is equal to discrete valuesx_1, x_2, … $x_n$with relevant probabilitiesp_1, p_2, … p_n, then the expectation$M[X]$is:

$$M[X] = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i}, \tag{18.16}$$

ortaking into account that$\sum_{i=1}^{n} p_i = 1$, one obtains

$$M[X] = \sum_{i=1}^{n} x_i p_i. \tag{18.17}$$

The dispersion of discrete value of random parameter$X$is marked as$\sigma^2$and is defined by formula

$$\sigma^2[X] = \frac{\sum_{i=1}^{n} \{M[X] - x_i\}^2}{\sum_{i=1}^{n} m_i}, \tag{18.18}$$

where m_i is the quantity of repetitions of value x_i from all set of random variable$X$. Apparently, that in the case of the absence of repetitions in the denominator of formula (18.18) will be just n.

In the case of continuous random value (18.16) looks like

$$M[X] = \int_0^{\infty} x w(x)\, dx, \tag{18.19}$$

where,$w(x) = dF(x)/dx$–the probability density, a $F(x)$ – integral function of distribution of random parameter$X$.

Let us consider some of distributions.

Hypergeometric law of distribution characterizes that in *n* observations a certain event will occur *d* times, if from possible *N* observations their quantity is *D*

$$P_n^d = \frac{C_D^d \cdot C_{N-D}^{n-d}}{C_N^n} \tag{18.20}$$

where

$$C_D^d = \frac{D!}{d!(D-d)!} \tag{18.21}$$

-the number of combinations of *D* to *d*.

In this case, the expectation and dispersion are:

$$M[d] = nQ, \qquad \sigma^2[d] = nQP\left(1 - \frac{n-1}{N-1}\right), \tag{18.21}$$

where$Q = D/N$ i $P = 1 - Q$.

The problems where items from consignment are chosen accidentally and the items chosen are not returned to the consignment.

At$n \ll N$dispersion goes to $n\,QP$. While this, hypergeometric law is reduced to the binomial.

Binomial distribution law characterizes the possibility of the random event v to occur in n independent observations. If in one independent observation the probability of event$v$ is equal p, then the probability of this event to occur *d* times in *n* independent observations is:

$$P_n^d = C_n^d p^n (1-p)^{n-d} \tag{18.22}$$

and $C_n^d$is the number of combinations of *d* in *n*.

Checking the devices for the presence of some attribute (for example, serviceability) by consequent random selections with return is an evident scheme that leads to binomial distribution. The contradistinction to the mentioned above is hypergeometric distribution, where *n* is not the sample size, but *n* is the number of independent observations that are held in constant conditions.

The expectation and dispersion of binomial distribution, respectively are:

$$M[v] = nQ, \qquad \sigma^2[v] = nQP. \tag{18.23}$$

when$D < 0.1N$, thebinomialdistributioncoincides with the Poison distribution.

The Poison distribution

In practice, there are often used small sample with size $n < 0.1N$, and $Q < 0.1$ thus random values are distributed according to the Poison distribution.

$$q = \frac{a^d}{d!} e^{-a}, \tag{18.24}$$

where $d = 0, 1, 2, \ldots$ i $a = nQ$

In this case

$$M[d] = a, \tag{18.25}$$

$$\sigma^2[d] = a. \tag{18.26}$$

The Poison law describes the probability distribution of the appearance various quantity of defective items *d* depending on the value $a = nQ$ (parameter of distribution) in relatively small sample.

As were mentioned above, there are both the discrete and continuous random values. For example, time to failure of the device is continuous random value.

Exponential distribution law is quite often used in practice. The random value *t* is distributed according to exponential law if the distribution density

$$f(t) = \lambda e^{-\lambda t}, \tag{18.27}$$

If *t*is time to failure of the device, then λ is the failure intensity.

There is often used the integral distribution function *F(t)* to characterize the continuous random values:

$$F(t) = \int_0^t f(t)\, dt \tag{18.28}$$

If$f(t)$ is defined by (18.27), then the integral distribution functions is

$$F(t) = 1 - e^{-\lambda t} \tag{18.29}$$

The expectation and dispersion for the exponential law are

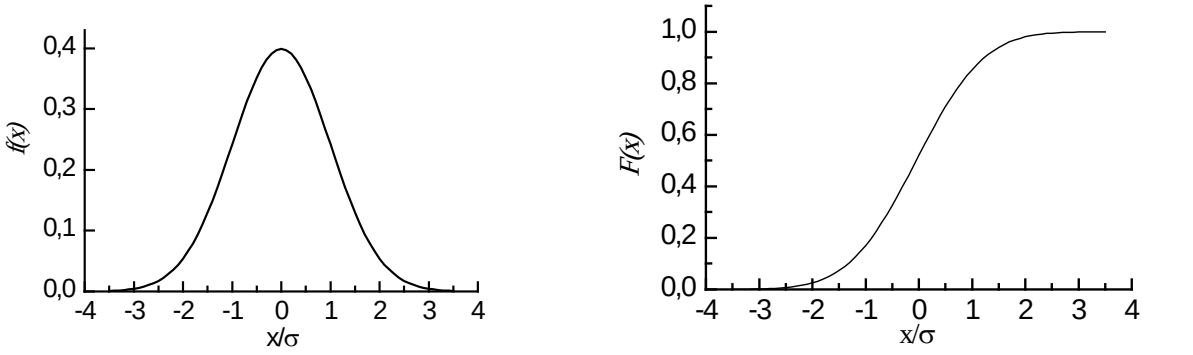
$$M[t] = \frac{1}{\lambda} \tag{18.30}$$

$$\sigma^2[t] = \frac{1}{\lambda^2} \tag{18.31}$$

Normal distribution law (Gauss law)

Normal probability distribution law of continuous random value that can take both positive and negative values, ranging from $-\infty$ to ∞.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-M[x])^2}{2\sigma^2}} \tag{18.32}$$



a b

Рис.18.4.The probability density (a), and integral function (b) of normal distribution of continuous random value

Curve (18.32) is symmetrical about$x = M[x]$ (fig.18.4).

In the case of the normal law, integral function of distribution is

$$F(x) = \int_{-\infty}^{x} f(x)\,dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-M[x])^2}{2\sigma^2}}\,dx \tag{18.33}$$

$F(x)$–probability that random value $X$is less than some value$x$.

Weibull distribution

The Weibull distribution characterizes the distribution of continuous random value at $t \geq 0$. In the Weibull law, the failure frequency is

$$f(t) = \frac{\delta}{t_0} t^{\delta-1} e^{-\frac{t^\delta}{t_0}} \tag{18.34}$$

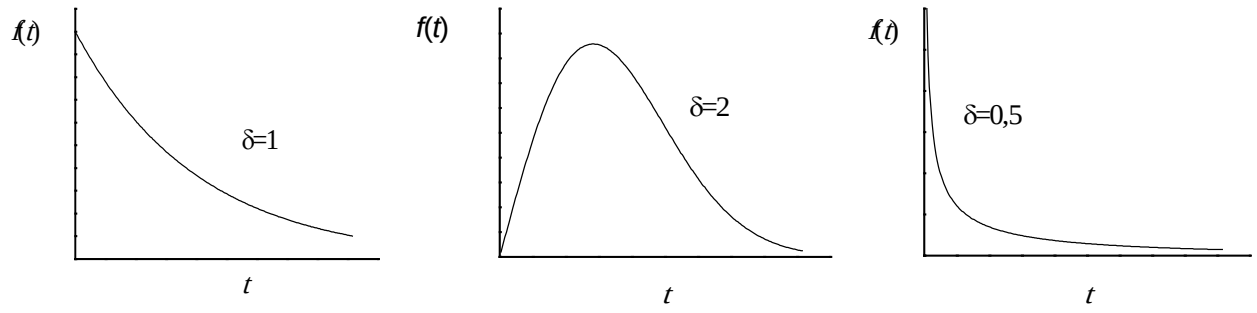


Рис.18.5. Density of the Weibull distribution at$\lambda = 1$

where t_0 and δ are constant values of the Weibull law. δ is parameter of shape. The curve shape of the density of probability $f(t)$ depends on it (fig.18.5). From (18.34) one can see that parameter t_0 is equal $\frac{1}{\lambda}$, where λ is the failure intensity at $\delta = 1$.

18.3.4. Methods to forecast and assess the reliability.

There are two main approaches in the science of reliability: the static and physical. In the first one, the device failure is considered as a random value without the reason analysis. Only after all devices of the consignment are fault, one could obtain the valid data of the failure distribution. Practical value of this approach is under the question, since these parameters for another consignment could differ. That's why, one tests some consignment for the reliability and according to this data, the device parameters of reliability are derived. In fact, there are built the histograms of failure distribution in time, and distribution parameters are derived according to the results of this observation. The obtained data extends to the whole consignment. As was mentioned above, the accuracy of such assessment depends on the sample size that is approaching the actual data if the sample size tends to the quantity of the consignment. Commonly, the observation time is reduced however, to meet the requirements of exact assessment of statistical probability, time of observation have to be long enough allowing the major part of consignment to failure. Apparently, that in case of low level of failure intensity, it is necessary to increase the device-hours (i.e. the quantity of the devices that are tested times the period of testing), in order to receive the characteristics of reliability. While high cost of testing, the device cost, and their testing might be inappropriate, or it is impossible to obtain the correct data during the period when they are actual.

Physical approach is based on the obvious statement that failures do not exist without cause. That is why this approach could be called as the physics of failures. The main tasks of this approach are the physical grounding of item insecurity, the recommendations for increasing the reliability, particularly, terms of use, working out non-destructive ways of control.

Since failure is consequence of some physico-chemical processes, its presence and speed are conditioned by inner and external factors that are called the factors of unreliability. Physico-statistical approaches are especially prospective in determination and forecasting the reliability. In particular, it is possible the following technique. The value and time changes of all parameters of so called training sample are researched. The most instructive methods, regarding determination and forecasting, are found out. Then the dynamics of these instructive parameters for other consignment is determined and researched during some period, and by the mathematical processing the extrapolation of parameter changes and the assessment of failure intensity in future are held.

The method of forced testing is used for reducing the research time. In this case the research is held in the conditions of increased temperature, electrical, or others load; and results are extrapolated for the case of normal (working) conditions. The main problem of forced test isproof of, so called, auto-modelity condition, that means mechanism causing faults is not changed during forced test.

18.4. Failure types and insecurity factors.

Violation of device workability is defined as the failure. While this, the device don’t meet the requirements of the quality conditions. Criterions of failure are established by normative and technical documentation.

18.4.1. Failure types

One divide failures by various attributes:

a. By degree of influence workability–complete and incomplete;

b. By physical character of direct manifestation–catastrophic (sudden) and parametrical (gradual);

c. By connection with others failures – dependent and independent;

d. By uptime–stable (non-convertible), temporary (convertible), shimmering.

Let’s look at the division on sudden and gradual.

Sudden failures is the consequence of hidden defects of material or construction that are manifesting in the process of exploitation, appearing due to abrupt changes, and having final character.

Gradual failures are conditioned by degradation of physico-chemical material properties under the influence of exploitation factors and natural senescence; are appearing due to drift of working parameters and their going beyond accepted values; as a rule, they have non-convertible character.

Relaxation failures – are resulted by gradual accumulation of changes in device state and appearance of abrupt transition to the nonworking state.

18.4.2. Insecurity factors and causes to failure

Depending on in what life stage the failures could be removed, they are divided on constructive industrial and performance. Respectively, one could select constructive, industrial, and performance factors of insecurity.

Regarding integral circuits, constructive factors are, in particular, chosen inappropriately materials, non-optimality of scheme decision, not the best chosen electrical and thermal properties, inefficiency of security, and other errors in constructive decision. Such disadvantages result in the device overvoltage and overheating during their working, and as a consequence in an acceleration of degradation processes.

The appearance of various defects is possible during production and technical stage, for example, lithography and diffusion defects, unsatisfactory adhesion and metallization quality, too large resistance of ohmic contacts, and other disadvantages conditioned by non-optimality of technology.

Above all things, among performance factors are its use in inappropriate electrical and thermal regimes, in climatic, or radioactive, or unforeseen overload.

According to mentioned above item sorting, integral circuits are elements of a system. Even the most complex and completed ones in terms of performance of given functions, are still components of electronic device. In turn, integral circuit consists of many elements, for instance, transistor, resistor etc. That is why, an integral circuit is also the system that consists of other elements. Integral circuit is the system that could not be restored, that is why it could not be repaired by replacement of out-of-use element. Even if it is possible, it is inexpedient economically, for example in the case of hybrid integral circuits. As were expected, the reliability of integral circuit without constructive disadvantages goes forward reliability of the consisting elements. In particular, it is caused by the unity of the technology, high level of automation of technical processes, and use of inter-operational control.

In terms of reliability, integral circuit contains not only radio-technical elements but also metallization, shell, etc. Degradation and failure of these elements cause the failure of whole IC at all. In turn the reliability of separate elements of integral circuit is defined as the stability characteristics of its elements, in particular as stability of parameters of separate area. Certainly, external actions influence the presence and speed of degradation processes.

The research demonstrates that the failure of items of single consignment could be caused by different reason. The failure distribution by type depends on the level of technology and

previous experimentation. The failure distribution of discrete transistors before and after culling process and also after improving the technique is shown on fig. 18.6. It is clear that some types of failure are removed effectively while culling. Improvement of the technique also leads to enhancing the reliability due to significant reducing some failure types, however portion of other types become more noticeable.

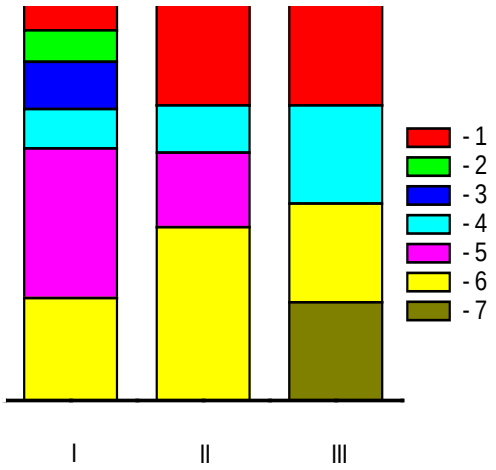


Fig. 18.6. Failure type distribution for silicon planar transistors, before (I) and after (II) culling; moreover, after improving the technique (III):1 – various failures; 2 –failures of metallization; 3 – change of electrical parameters; 4 – failures caused by deficiency of pressurization; 5 – failures caused by deficiency of wired pins and welding;6 – failures caused by surface effects; 7 – failures caused by deficiency of photolithography.

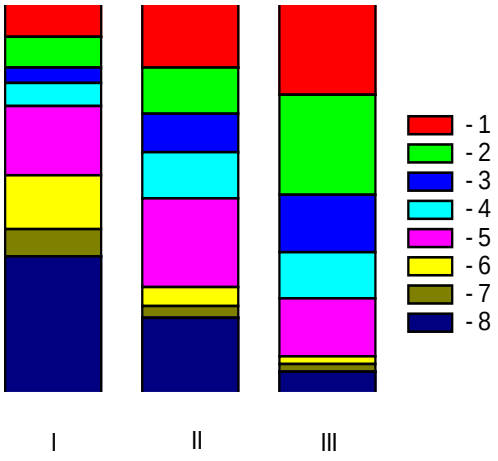


Fig. 18.7. Failure type distribution for IC of different degree of integration (IC of low (I), medium (II), and large (III) degree of integration):
1 –failures of metallization; 2 – diffusion errors;3 – unrelated particles; 4 – other failures; 5 – deficiency of crystal and oxide; 6 – deficiency of pressurization and welding; 7 – deficiency of crystal-holder; 8 – incorrect use.

The change in the distribution by failure cause and types while increasing the degree of integration of scheme is shown at the fig. 18.7. It is clear that while increasing the degree of integration, the influence of failures caused by metallization deficiency, diffusion errors, and unrelated particles also increases that is related to reducing the size of elements.

18.4.3. Types of culling tests

The experimental conditions and list of under-controlling parameters are regulated by corresponding technical documentation. These conditions regulate:

- General electric, mechanic and climatic requirements to the item construction, and special requirements to their production;
- Ways to conduct electric, mechanic and climatic tests for reliability;
- Norms and values of parameters at normal and boundary temperature of the environment;
- Regime to conduct various types of test etc.

Depending on aim, all tests could be divided into two main groups: researching that are conducted in order to study the device, and controlling that are conducted to control the item quality.

Among researching tests the most interesting are:

- Boundary tests that are held in order to determine the dependence between maximum permissible and exploit parameters.
- Comparing quality characteristics of two or more devices that is conducted in identical conditions;
- Accelerated tests which methods and conditions result in obtaining the necessary amount of information during more short term than at the normal conditions.

The next, more frequently occurring group of tests is controlling tests including first of all, such ones:

- Acceptor-donor that are conducted for each consignment. Commonly, the 10%-sample is tested. The correspondence to the set parameters is tested, then rough defects are found;
- Periodical testing is conducted regularly after definite period, and when the beginning of emission and if manufacturing was stopped temporarily but then resumed. These tests always are selective. They are conducted in order to check stability of the technology;
- Typical tests are conducted in the case of changing the construction, technology, and materials used;
- Testing for durability and safety is conducted to proof correspondence between set value of minimum working time to failure and preservation term;
- Qualification tests are conducted in order to assess production readiness to release the device and correspondence to its regulatory-technical documentation;
- Resource tests are conducted during the period of research and design work, and while the construction or technology is modernized;
- Boundary tests are devoted to determining the maximum exploit characteristic.

If testing is not researching (but controlling) then quantitative characteristics are regulated. For example, for hybrid integral circuits such working regimes are set:

- Centrifugation with acceleration 20000*g*;

- Testing the mechanical strength of wired pins to substrate by the way of breakaway from junction;
- Mechanical shock ($1500g$, 1mcs);
- Thermo-cycling ($-55 \div +125$℃);
- Thermo-stabilization (obsolescence under the temperature of 150 *K*, and others);
- Thermal shock ($-55 \div +125$℃);
- Climatic tests (98%ofdampness, high and low temperatures, surrounding containing salt;
- Electro-thermo-training (electrical resistance is rated, temperature is 125℃, testing period – 168 hours.);
- Turning over temperature range while electrical regime;
- 100% testing of shell tightness by the helium method

According to the character load that influence device, testing are divided:

- Mechanic (vibrations, accelerations, shock);
- Climatic (dampness, thermo-cycling);
- Electric (electrical regime under normal and elevated temperature);
- Radioactive;
- Stability while static electricity (as an artificial source of static electricity, it is used the discharge of the capacitor having the capacity of $C = 1\ nF$, charged to the voltage of 1 кВ through the resistor of$R = 1$mOhm.

The lists and testing regimes that are used by various companies might differ in some extent for different items.

18.5. Experimental methods to analyze the items' quality, defectiveness, and malfunctions.

Device malfunction or failure (fault) areconsidered to be due to changes in the structure. One carries out the failure analysis in order to localize these changes. To be more exact, the analysis of devices and integral circuits is held in order to find out the causes and ways of failures during storage, test, and exploit.

To do this, there are used both the well-known in practice methods: *radioscopy, X-ray analysis, optical spectrum analysis, metallography, masspectroscopy, gas chromatography, IR fault detection, IR radiometry, IR absorption spectroscopy, chemical and electro-chemical analysis, electronography, electron microscopy, micro-photography, binocular microscope, electrical measurements,* and relatively novel methods and techniques that are used in microelectronics including *scanning electron and ion microscopy including specular, electron*

and ion microprobes, high-resolution radiography, stereo-optical methods, method of liquid crystals, method of electrical measurement of IV-curves noise, escarpment, nonlinearity, methods of non-destructive testing the heterogeneous dielectrics by measurement of dielectric absorption etc. The order to analyze IC physically, there are used: external studying with the microscope; electrical test for functioning and measuring the electric parameters; physical analysis of the scheme in shell (radioscopy, IR fault detection, X-ray radiography of shell, test by leak-finder for leakproofness, analysis of residual atmosphere by gas chromatography method, etc.): cutting of shell; visual inspection (optical and scanning electron microscopy); removing coatings by etching and recrudescent visual inspection; inner microscopy inspection and phisico-chemical analysis (SEM, electron and ion microprobes, electronography, X-ray, chemical, metallurgical, and other types of analyses); insulation of element that faulted andexperiment repetition.

Depending on IC type and failure that occur, the analysis order could be changed. Let's consider the main stages of failure physical analysis. If non-functionality were determined by electrical test, and external review did not show serious errors, then the further research direction could be indicated by shell leak proofness test. Shell leak proffness is assessed by testing for rough leaks (10^{-6}cm^3/s, freon) and weak leaks (10^{-7}-10^{-8}cm^3/d, helium masspectrometer). The composition of residual atmosphere is researched by gas chromatography method in the case of shell leakage. It appears the natural necessity to find out the causes of shoddy shell, and also to assess the effectiveness of compound applied and possibility of moisture, polluting particles penetration and corrosion appearance. These questions are solved after shell cut, review, coatings removing and visual inspection using binocular or electron microscope.

Electron microscopy is the most famous method to inspect quality and determine large class of IC failures. Modern electron microscopes have huge magnification, high resolution, large depth of view. Currently using scanning tunneling microscopy (SEM) is particularly effective to study failures of IC element of complex topology (for example, high-frequency transistor having complex geometry, thin-film capacitor, multilayer metallization, etc). 3D scanning tunneling microscope allows to obtain an embossment of the surface. Information of the object state could be obtained using TEM while electron beam scanning to research surface, and is based on such types of interaction between electron beam and surface material: cathodo-luminescence, X-rays, Auger effect, high energy reflected electron scattering, secondary electron emission, appearance of induction currents or potential contrasts, etc.

While using the stream of secondary ions, one obtains huge magnification with high resolution for the depth. Method of electron microscopy with potential contrast is implemented if there is used effect of modulation of the stream of secondary ions with low energy of local

electric field. Analysis of failures of bipolar non-passive schemes is convenient to analyze in such way. Passivation reduces the potential contrasts. Then, use of currents directing by electron beam in crystals or film elements of IC allows, while working with REM, to study not only surface events but also obtain information about processes taking place in sub-surface layer and in the volume of thin films. SEM allows to detect specific faults that result in breakdown of p-n – junction or oxide, inversion of the surface, poor contacts between metallization and semiconductor, area with abnormally elevated conductivity, and also to carry out the microanalysis of the topology of the p-n – junction.

The most important stage of analysis of scheme devoid of protective coatings is, first of all, examination of the surface state. Here, the most important criterion is presence of contamination. Its source could be various: chemicals using in manufacturing, deionized water, air environment, industrial clothing, etc. Also, the criterion of surface state is presence of corrosion products. There isa variety of ways to analyze physico-chemically the contamination and corrosion products: SEM, roentgen-fluorescence analysis, emission and IR spectroscopy, micro-probing, etc.

While studying the out-of-work IC by use of ionic micro-probe, the destroying of sample occurs. Primary stream of accelerated argon or oxygen ions examines the surface, resulting in erosion of sub-surface layer to the depth of approximately 250 μm. While this knocking out, the positive and negative ions get into analyzer (masspectrograph) where are distributed in respect to the mass-charge ratio. This ratio and also the intensity of secondary emission of these ions are commonly used for identification and assessment of the percentage-content of different atoms and compounds in the observed contaminations, impurities, products of corrosion. The responsiveness of ionic micro-probe is more than those of electron microscope by few orders of magnitude (10^{19}atoms/cm^{3}and 10^9-10^{15}atoms/cm^{2}respectively).

The analysis of events that occur while breaking fast electrons at the IC surface, and also in the depth of 2-5 monomolecular layers from the surface, could be conducted by the Auger effect spectrograph. Using the intensity of energy spectrum lines, one could not only identify heavy atoms of foreign substances, but also obtain information about chemical bonds between these atoms. For example, using Auger-spectrometer, one managed to obtain the information of contact area junctions, silicon and oxygen traces produced by thermo-compression and ultrasound welding, remains of SiO_2coating, and large amount of nickel, cobalt, ferrum, oxigenand less amount of carbon, nitrogen, chlorine.

These contamination, inclusions, impurities worsened the quality of junctions of received by as thermo-compression and ultrasound welding. So, the method is suitable to analyze the quality of IC junctions.

While researching the causes of IC failures, it appears the necessity to obtain the conception of the surface state not only by its contamination level,but also by quality of its micro-relief taking into account the presence of roughness, protrusions, depths and other geometrical heterogeneities. It is necessary to find out the dependence of the surface micro-relief on various factors during time of exploiting (or testing); to find out what heterogeneities are acquired even after IC manufacturing, and how they influence its workability. Such information is especially for establishing the causes of oxide failures in semiconductor IC, thin-film capacitors in the hybrid IC.

The surface micro-relief is studied by the method of measurement of the surface profiles. Needle profilometer allows to assess roughness. Serving as indicator of inequalities, the profilometer probe as diamond needle provides back-translational motion at the surface (of the substrate, crystal, film) that is transformed into calibrated electrical signals registering by recorder. More detailed analysis of the surface quality is studied by the replica method. The thin film reproducing relief is deposited on the researching surface, then it is removed and studied in scanning electron microscope. Moreover, methods of electron diffraction and optical multipath interferometry are widely used.

Scanning and atom-force microscopy

The oxide film defects in semiconductor IC as well as capacitor dielectrics in thin-film microcircuits are detected with micro-probing. The area of defect dislocation is detected and, if needed its isolation from other circuit, the jumpers of respective metallization lines are cut (by scratching, etching, laser beam). The electrically isolated defect area is studied by SEM or method of liquid crystals. In the last case, while applying (through the micro-probes) voltage to defective area of microscheme. Liquid crystals allow to detect the abnormality in distribution of electrical and thermal fields. One managed to detect the micro-holes of diameter 0.5 mcm in the oxide layer by the method of liquid crystals.

Other ways to detect oxide defects are: optical contrast microscopy, optical scanning, laser scanning (external photo-emission), radioactive tracer method, neutron-activation analysis, and mentioned above methods of profile measurement, interferometry, electronic diffraction, etc. Physical analysis proves that defects in the oxide layer of IC are conditioned by insufficient control while performing technological operations, for example, errors when combining the masks, coating with photo-resist, etching operations, etc.

In some cases it is advisable to make the oblique cut of microscheme defect area. Using the oblique cuts there are detected defects of diffusion areas and insulating pockets and inclusions in the depth of the crystal of semiconductor scheme. The researching failures of bipolar and MOS IC caused by the appearance of inverse layers at the surface because of contamination, presence

of the surface charges and leaks, is better to conduct on special testing structures. Measurement of potential diagrams is conducted by SEM. Inverse layer are detected while scanning the surface with sharply-focused light beam (spot diameter of 2-50 mcm). Also there are researched current-voltage characteristics of separate areas of p-n – junctions.

An important part in failure analysis is studying the distribution of thermal fields in the IC by radiometric methods. There are used micro-radiometric methods to obtain the thermal profile of transistoror IC. The infrared radiometric microscopes are able to measure the temperature difference in 4 C while scattering power is 100-300 mW in the microscheme.

There are often used common methods in order to study the reliability of IC physically that proved themselves while failure testing. For instance, the shell tightness is examined by thermo-shock. The method of thermo-cycling is used in order to examine the film's adhesion. After these tests, the appearing rigidity revealed by SEM is an evidence of insufficient adhesion. There are tests for the presence of O.C. or S.C. by the voltage impulses of$V \leq 0.5$ V, in order to assess the strength of junction between wired pins and metallization and shell traverses. For this goal, IC are also tested by constantly directed vibration of constant or variable frequency. Finally, mechanical strength of contacts is measured by analysis of tension and direct measuring of destructive force. The final control of junctions and contacts is held after cutting the microcircuit and reviewing in optical and scanning electron microscope. Metalographic analysis of contacts allows to reveal the presence of intermetallic phases, and to clarify its composition and structure.

Physical failure analysis by mentioned above methods opens the possibility to forecast physically the reliability and stability of IC. There are also special methods that allow to obtain an idea of IC state, and to forecast the reliability grounding scientifically, by measuring the indirect parameters. The non-destructive tests take especial place, they provide information of physical state of object not changing this state, despite of a number methods mentioned above related with the destruction of the researching sample. It is clear, that methods of qualitative physico-chemical and those of non-destructive testing should not be contradistinguished. It is necessary to mention that the non-destructive methodshave paramount meaning that is acquired by the tasks of quantitative reliability prediction comparing with qualitative failure studying. From the other hand, the grounded choice of indirect qualificative parameters allow one to cull the IC during donor-accepting tests.

The question of the choice of indirect parameters that carry reliable information about the IC reliabilityis discussed currently. Apparently, such parameters should be more informative than common electrical. For instance, it is hardly to obtain full idea of the stability of technological process and forecast the behavior of consignment items during the exploit, while measuring the values of thin-film resistors and capacitors directly after their manufacturing. And

such indirect parameter as a coefficient of the third harmonic, measuring after sinusoidal signal at the resistive or capacitive circuit output, is more informative and suitable for culling thin-film elements and forecasting their reliability.

Method of forecasting the reliability and stability by the level of low-frequency noise is mentioned to be prospective. There are examples where this method was really informative. For example, the presence of defective conducting films while high current ($j = 10^7$ A/cm^2) is detected due to the increase of current noise level in few orders of magnitude. However, this level of LF noise could be used rather for culling IC; for quantitative forecasting it is necessary to find out the appropriate correlation. The same difficulties exist while choosing the value of the slop of IC elements' current-voltage characteristic as the main informative parameter.

Methods of non-destructive testing received substantial development, based on using the properties of inhomogeneous dielectrics. The basis of the method is fenomenological model that despite of its formality, allows to recognize the clear connection between the properties of inhomogeneous dielectrics, results of non-destructive testing, and characteristics of reliability. Also there are widely used researching the polarization processes that are slowly established in inhomogeneous dielectrics, i.e. absorption processes in the range of low and infra-low frequencies (lower than 10^{-4}Hz). As main parameters there are used coefficients of absorption and non-expotentiality, recovering voltage, and also frequency dependence of$\tan\delta$andε are determined. Inhomogeneous areas are potential reason for dielectric fault, if there is a possiblity to assess state and distribution of external electric field and formation of space charge in such dielectric, then it is equivalent to assessing the reliability. It is possible to receive quite a full idea of dielectric state based on physical models of absorption processes using the results of non-destructive testing, and this allows to judge reducing its electrical strength, and thus to determine the correlative connection between inhomogeneity of internal field and reliability.

Methods of studying the dielectrics absorption could be used for researching the failures of thin-film capacitors and semiconductor IC (capacity of p-n – junction). The disadvantage of this methods is the complexity of direct measurement of absorption and desorption currents. The measurement of indirect parameters: coefficient of non-expotentiality, potential of recovering, etc. – are less work-consuming, but mathematical processing of the results is also complex from the point of view of providing the correlation between results of non-destructive tests and indexes of reliability.

18.5.1 Testing structures

Modern IC that are able to perform different functions, also are distinguished by the complexity of technology. That is why the direct physico-chemical analysis of these schemes is

quite a complex. For the simplification of researching method and providing more access to each element, IC could be represented as a set of testing structures and analyzed by physico-chemical analysis and tested for reliability as a part of testing structures. The set of testing structures (TS) is the formal equivalent of real IC and is created being based on mathematical model of respective IC. Each TS, in turn, is a set of structural physical layers of the simplest topology without taking into accountthe definite configuration of elements that respects to topological drafts. For example, there are created such structural layers for MOS: initial substrate of n-type; “pocket” of p-type; doped areas of sources and drains of n- and p-type, metallic strips on thin and thick dielectric layer. Recruitment and combining of structured layers in each TS, and also recruitment and combining of TS for given IC are depending on the definite research aim and are determined by the level of development and specific features of the technology.

TS is characterized by the number of interacting layers that are situated at the same vertical plane, the relative position at horizontal plane, and the form of layers. Usually, each TS contains 2-5 structured layers. Depending on the research aim, all TS could be divided on four types that are assigned to researching: physical processes and electro-physical parameters of IC – physical TS; electrical values of elements (transistors, diods, resistors, conducting strips, etc.) – functional TS;working the different types of circuits – schemo-technical TS; physical principles of failures and parasitic effects appearing – reliability TS. Depending on the research aim, there are: different choice of the type and optimal TS combining, determining the necessary set of structured layers and theirs combining, minimization of the number of the structures. For instance, the TS of reliability that are relative to some semiconductor IC, is manufactured in the way to reproduce real conditions, for example: the number of defects, speed of defects’ development and respecting failure time, electrical and thermal regimes. TS should provide the possibility to research the main directions of the insecurity of semiconductor IC: oxide, surface, doped regions, pins, shell. For example, MOS IC are characterized by the following faults: the breakdown of thin oxide layer under the gate as the result of defect development and large electrical field, burnout and break of metallic strips due to electro-diffusion, appearance of output currents due to the formation of surface channels between the diffusion buses that are under different potentials, and space outflows of p-n – junction, etc.

Special TS with definite morphological variables should be created for researching each type of these faults. For instance, for analysis of the oxide break mechanism, the TS should represent not only physical properties, dielectric thickness and electric field inside, but also the conditions of its growth, specific of adjacent doped regions: their geometrical characteristics etc.

TS of reliability are restricted not only by physical analysis of faults. TS allow the possibility to determine quantitative characteristics of reliability by the results of accelerated

testing. Knowledge of physics of failures allows to select the main circle of destructive mechanisms, to consider their possible combinationsin various conditions and thus to reduce purely random factor of the faults during such testing. Element being a part of TS are tested for reliability, thus there are optimal conditions for defect detection and acceleration of development of IC failures. The result of determinative approach to the testing could be described as the possibility of reducing the amount of trialsfor the number of testing samples and for the necessary time required, compared with usual statistics experiment.

TS of reliability are also aimed to working out the measures and structures that increase the reliability. IC should be used to assess effectiveness of various constructive, technological, exploitation-conservation measures that prevent the appearance of faults and parasitic effects. Required quantity of “reliable TS” is reduced while constant technological process and optimal topology of IC.

18.6. Causes and mechanisms of failures of discrete devices and IC

18.6.1. Failures of interconnections in ICs

The most often faults of crystal connection to external pins are observed as breaks of burnout of wired conductors and breaks of contact connections of aurum or aluminum conductors to contact area. This faults are determined by the processes of electro-diffusion, formation of inter-metallic and Kirkendall effect.

Electro-diffusion in wired conductors.

It is determined experimentally that, while constant current of large density, break (burnout) of conductor appears. For example, for *Al* conductor, burnout is observed due to current density of$j > 10^6$A/cm^2. This effect is determined by electro-diffusion or mass-moving due to “electron wind”.

In continuous metallic conductor, there are two forces that act on thermally excited interstitial atom: electrostatic force

$$\vec{F} = q_i\vec{\mathcal{E}}, \tag{18.32}$$

where q_i– ion charge, $\vec{\mathcal{E}}$ - electric field intensity, and force of electron wind $\overrightarrow{F_e}\sim\overrightarrow{\mathcal{E}}$ that acts due to interchange of momentum between electron and ion while collision. Coefficient of proportionality $q_{ei} = \vec{F}/\vec{\mathcal{E}}$ depends on concentration of conducting electrons, their free path, and cross-section of ion scattering. It could be considered as a charge conditioned by interaction between electrons and ions. Then

$$\vec{F} = (q_i - q_{ei})\vec{\mathcal{E}} = q_i^*\vec{\mathcal{E}}, \tag{18.35}$$

where q_i^* - effective charge of ion. If $q_i^* < 0$, then the force acting on ion is directed according to electron movement. If large current density, then ions also move towards electron movement under this force (force of "electron wind"). Electron speed in this case could be written as

$$v_i = \frac{\mathcal{D}_i}{kT} q_i^*\vec{\mathcal{E}} \tag{18.36}$$

where $\mathcal{D}_i$–is coefficient of ion diffusion.

If cross-section of the conductor is inhomogeneous along the length, then while large current density, the most mass-movement would be in the area of maximum current density, so in the area of minimum cross-section. The ion movement would cause the following reducing the conductor cross-section in the region where this cross-section is minimum. Finally, this process will lead to the break of the conductor. In the moment of break, an electric arc appears between regions of the conductor that causes melting of the ends of the conductor in the site of break

Considering temperature dependence of concentration of thermally activated ions and coefficient of their diffusion, one could show that reciprocal average median time to failure of the conductor due to elecro-mass movement is

$$t_{50}^{-1} \sim j e^{-E_a/kT} \tag{18.37}$$

For massive aluminum and even for thin conductor, energy of activation of the electro-diffusion process is $E_a \cong 0.5 \div 0.7$ eB.

Formation of intermetallic.

As was mentioned above, in schemes of low and medium-scale integration there are ways to connect the scheme to external pins using thin aurum or aluminum conductors that are wired to the contact areas by thermo or ultra-sound compression.

The main reason of faults $Au - Al$contacts –breakdown or resistance increase (with following breakdown) are:

- Formation of intermetallic phases;

- Formation of emptiness due to unequal inter-diffusion in the contact$Au - Al$ (Kirkendall effect).

Earlier, failures of contact $Au - Al$were associated with formation of compound$AuAl_2$ that have purple color and is called "purple plague". However, research demonstrated that composition of intermetallic structures is complex and depends on conditions of their formation (concentration ofAu, AlandSiin the region of reaction, temperature$T$, time$t$, etc.). Today, there known at least 5 compounds and all of these have hardness more those ofAuorAl. That is why, intermetallic itself could not be the cause of reducing the mechanical stability of the contacts.

However, reaction result in changing in amount of substance, formation of pores that could evolve in opening due to different coefficient of thermal expansion, etc.

Kirkendall effect.

The difference in metal inter-diffusion coefficients results in formation of emptiness in the metal that is more effective diffusor. In the systemof $Au - Al$, the formation of microscopic emptiness is the result of vacancy accumulation, and is observed on the side of aurum that diffuses in aluminum with higher speed than in reverse.

Since in the system $Au - Al$ there is buffer region that consists of intermetallic phases, so the process of diffusion has complex character, that is why the vacancy accumulation is possible from the side of aluminum. In order to prevent the reactions in the boundary of $Au - Al$there are used the barrier layers (Cr, Ti, Moetc), but this do not solve the issue fully, since these metals also interact with aluminum.

The alternative to system $Au - Al$ could be the system $Al - Al$, that means use aluminum conductors for the connection of the micro-scheme to external pins. An absolute advantage of use of aluminum conductors is the absence of intermetallic. However, *Au* wire yields *Al* one by strength and rigidity. In order to increase strength, Al wire is doped by $1\%Si$. However, after heating, recrystallization results in reducing the strength of conductor.

18.6.2. Failures of IC metallization

The metallization is considered as a system of film metallic conductors disposed along the crystal surface that connect separate elements of micro-scheme. There are series of requirements for metallization, the main of them are:

- High electro-conductivity;
- High adhesion to silicon and silicon dioxide;
- Providing fine (low-resistance) ohmic contact to the$n$, as well as to p – semiconductor;
- Easiness of formation of topology (picture) of wiring;
- Possibility to connect external conductors.

Single material that meet these requirements is aluminum. Besides that, aluminum forms Shottky barrier at the contact to low-doped n-silicon that allows simple realization of some scheme decisions. However, aluminum metallization have series of disadvantages. The use of other metals to reduce some disadvantages requires formation of additional metallization layers including, as a rule, the aluminum layer. So that, aluminum is the single metal that allows to create single-layer metallization of silicon integral circuits.

The most frequently, metallization failures occur due to the currents of high density, high temperatures, and also during prolonged exposure to temperature and electrical loads.

Main failures:

Break as a result of metal electro-diffusion at locations of high current density;

Break as a result of chemical or elecrochemical corrosion of aluminum herewith shoddy protective layer and contamination of the crystal surface of IC substrate;

Short circuit through the holes in oxide, as the result of formation of bridges between conductors that are formed while electro-corrosion.

Electro diffusion in aluminum films could result in two types of faults:

- Break of semiconductor at the lowest cross-section;

- Formation of short circuit between adjacent conductors, or between conductor and semiconductor, that is conditioned by metal accumulation in the form of “mustache” in the region that is situated near the area of the lowest cross-section directed toward movement of electrons (direction of electron wind).

Despite of thin aluminum conductor, for aluminum films (metallization tracks) the average median time to failure caused by electro-mas movement is proportional to the square of the current density$t_m^{-1} \sim j^2$. As shown experimentally, the speed of mas movement is more in films with lower crystallite size. It is conditioned by accelerated diffusion at the grain boundary.

The most likely place to conductor break due to electro-diffusion is defect, mechanical fault of conductor that results in local reducing its cross-section, and also in the region of rung at the edge of contact window inSiO_2.

Corrosion and electro-corrosion.

As a consequence of previous technological operations, the real ways is a contamination at the crystal surface, particularly chloride ions. There is always the penetration of wet inside the shell due to laxity at the place of external pins, particularly in the case of plastic shell. This process result in formation of active compounds (substances), for example HCl. The corrosion appears due to this substances, and also the violation of integrity of aluminum conductor. If there is the potential difference between adjacent conductors, then there is the ion movement, and as a result the conductor break due to electro-corrosion, or short-circuit between bridges that are created while ions electro-migration.

Solid-phase interaction at the boundary $Al - Si$.

This interaction appears rather while micro-scheme manufacturing, and could be the cause of defect. As is known, one of the operations in technology of integral circuits is annealing of the structures after formation of metallization. It is conducted for the increase of adhesion of aluminum film to silicon and SiO_2. It is considered to be that while the temperature $200 \div$

450℃, in the system $Al - SiO_2$ there are only diffusion processes, and while higher temperature silicon recovering from SiO_2 takes place by the reaction

$$4Al + 3SiO_2 = 2Al_2O_3 + 3Si \qquad (18.38)$$

With the formation of multi-crystalline silicon and$\gamma - Al_2O_3$. This interaction is explained by not only adhesion increase but with resistance reducing of ohmic contacts. It is considered to be, that due to there action the layer of “natural” oxide is destroyed that always forms at the silicon surface in contact window. But solid-phase interaction in contact window is not restricted only by described reaction. After the formation of direct contact of Al to Si in the contact window there are observed the silicon dissolution in aluminum film and its diffusion along aluminum track for a long distance. The solubility of silicon in aluminum is near 1% at the temperature 500℃. Aluminum, in turn, fills the volume from which silicon was diffusing. The destroying of “natural” oxide in contact window occurs unevenly and, as a consequence, silicon dissolving in aluminum is uneven also. Schematically it is shown at the fig.18.8. As could be seen, the most dissolving and respectively the most penetrating of aluminum in silicon is observed at theperiphery of contact.

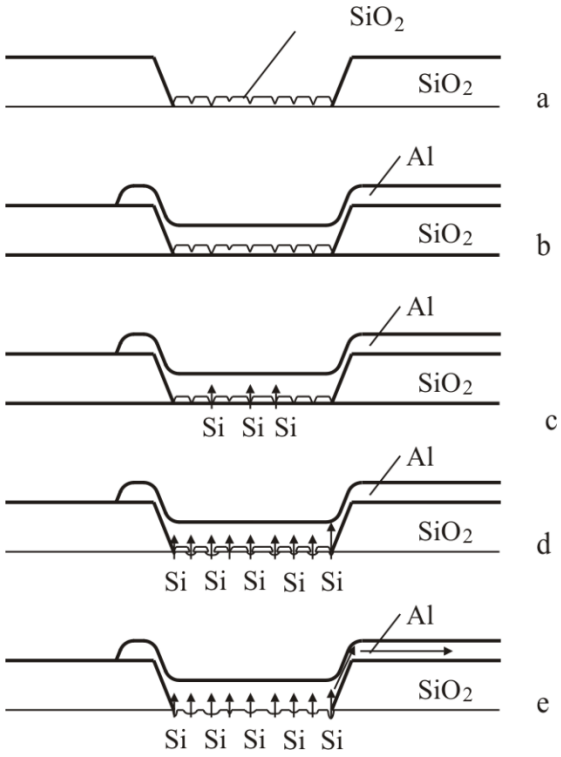


Fig. 18.8. Formation ofc ontact $Al - Si$: a – «window» SiO_2on the silicon with layer of “natural” oxide; b –structure after depositin galuminum film and photolithography; c – local destroying of natural oxide and beginning of silicon substrate dissolving in aluminum; d – formation of hollows on silicon surface and their filling by aluminum; e – final destroying of natural oxide and formation of embossed border $Al - Si$.

It could be caused by three reasons. Firstly, due to exothermal character of reaction (18.38), the temperature at contact periphery where the SiO_2has unlimited quantity, would be maximum, that is why the solubility and coefficient of silicon to aluminum diffusion would be maximum. Secondly, due to difference between coefficients of thermal expansion ofAl, SiandSiO_2at the contact periphery there are maximum mechanical stress that increases the solubility of silicon in aluminum. Thirdly, silicon that dissolves in aluminum at the contact periphery has the possibility to diffuse along the track. As a result of penetration of aluminum pile at the contact periphery it could “whip through” and short-circuit the shallow p-n = junction fig. 18.9.

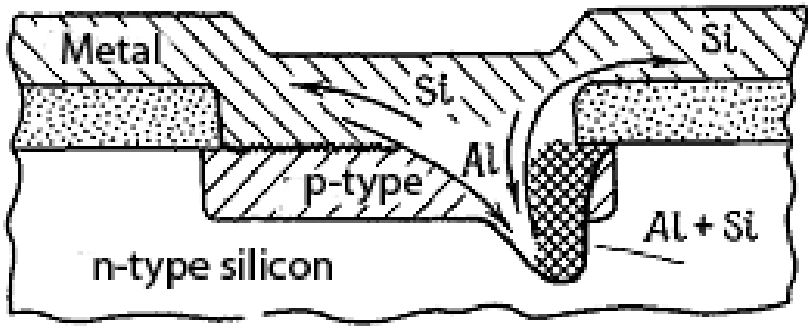


Fig. 18.9. Whip through shallow $p-n$ – junction while annealing.

One of the ways to prevent silicon dissolving in aluminum film is to deposit thin layer of amorphous silicon atop the metallization before depositing the aluminum metallization, or addition approximately 1% of silicon in aluminum film during its depositing. In the first two cases while annealing, silicon from additional amorphous film dissolves firstly, and if after that it is not enough to saturate aluminum then the monocrystalline silicon in contact window would be dissolving. Thus, if the thickness of additional silicon film is not enough then it do not perform fully its function. If the thickness is too large and it would not dissolve fully then this will result in impairment of ohmic contact in the window, or in impairment of connection between aurum conductor and contact area. The formation of aluminum film containing some silicon concentration could be implemented either by use of aluminum doped by silicon as the initial material for metallization, or by depositing metallization simultaneously from two sources – aluminum and silicon.

It is necessary to mention that in any case, silicon dissolved in aluminum while cooling after annealing is re-crystallized partially at silicon surface, creating thin layer doped by aluminum that influence the contact characteristics.

18.6.3. Degradation and failure of contact metal-semiconductor in silicon IC with single-layer metallization

Causes and reasons of instability of discrete and integral Shottky diodes.

The contact metal-semiconductor with Shottky barrier is one of the oldest semiconductor devices. Traditionally, it is used in SHF electronics (detectors, faucets, varactor, etc.). Nowadays, Shottky barrier is widely used as in integral circuits (IC) to increase its speed and in force diode in secondary (impulse) power sources.

In each concrete case, the work of contact metal-semiconductor device or IC is determined by some of the following electro-physical characteristics:

1. Forward voltage drop V_{F} at definite value of forward current;
2. Coefficient of ideality $n \equiv \frac{q}{kT}\frac{dV}{d\ln I}$ (or the slope of CVC in semi-logarithmic scale $\alpha = d\ln I/dV$);
3. Value of reverse current I_{R}at definite voltageV_{R} (or breaking voltage at definite value of reverse current);
4. Value of the contact capacity $C$and its voltage dependence,

Listed characteristics of contact metal-semiconductor are determined, in general, by the parameters of metal, semiconductor, interlayer between metal and semiconductor, and also electronic states at the boundary semiconductor-interlayer. Let's remark that as a metal could be as pure metals and different alloys, solid solutions and compounds of metal and semiconductor. Any change in parameters of listed areas occurring while storage, working, or due to some external actions could result change of elecrophysical characteristics of contact metal-semiconductor.

To describe the correlation between elecrophysical characteristics of contact metal-semiconductor and Shottky barrier with physical parameters of volume of semiconductor, metal, interlayer and surface states, there is used physical model presented on the fig. 18.10a; the main parameters of this model are height of potential barrierφ_b, semiconductor space charge region thickness (SCR) – w, and also distribution of applied to contact voltage$V$between interlayer (V_1) and SCR (V_2) that is determined by the parameters of all listed areas of contact metal-semiconductor.

Physical parameters of contact $Al-nSi$ depend significantly on the technology of their manufacturing. For instance, height of barrier of $Al-nSi$ junction obtained by spraying of Al onchemically etched surface ofnSithat was not annealed, approximately equals $\varphi_b \approx 0.45$ eB.

After annealing the height increases to the value of $\varphi_b \approx 0.7$ eB corresponding to obtained by Aldeposition on chieftain surface ofnSi in vacuum. Such an increase of barrier height during contact annealing could be explained by destroying natural oxide that appears on silicon after chemical etching. However as was observed, the height of barrier of$Al - nSi$contact depends on the annealing temperature (fig.18.10) and cooling speed after it.

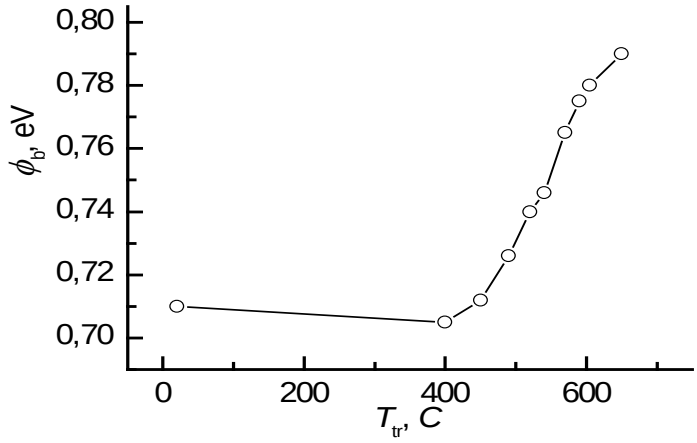


Fig. 18.10 Dependence of barrier height in contact $Al - nSi$ at 300 ℃ on the annealing temperature.

As could be seen from figure, beginning from 400℃, while annealing temperature increases, the barrier height is growing to the value much more bigger than$\varphi_b \approx 0.7$ eB. For structures that were annealed at 500℃ in case of natural contact cooling, after annealing the barrier height increases to nearly $\varphi_b \approx 0.72$ eB, while rapid cooling (structure looks like cool metallic table), the barrier height approximately equals$\varphi_b \approx 0.68$ eB and while gradual cooling increases to$\varphi_b \approx 0.74$ eB.

Such dependence of barrier height of the contact$Al - nSi$on the regime of heat treatment, could be explained by solid-phase interaction at the boundary aluminum-silicon. During annealing sub-surface silicon layer is dissolving in the aluminum film. Its concentration comes to 1% at the annealing temperature 500℃.While gradual cooling, solid solution$Al - Si$becomes supersaturated and part silicon dissolving in aluminum re-crystalizes at the silicon surface. Taking into account that at300 ℃silicon solubility reduces nearly by an order of magnitude, and diffusion coefficient remains considerable value, so the most of solute silicon have time to re-crystallize. If the thickness of Al film is nearly 1 μm then the thickness of re-crystallized layer could be up to100Å. Since while temperature increase also solubility of aluminum in silicon increases (to $10^{18} \div 20^{19}$ cm^{-3}), so re-crystallized silicon layer turns out to be doped by acceptors (Al – is acceptor in silicon with the depth of $E_a - E_v = 0.057$ eV). So, while gradual cooling,

there is observed the formation of the contact of $Al - p^+ - nS$. Band diagram of such contact is shown on the fig. 16.27 (Chapter 16) , it means that the presence of p^+ - layer could result in increase of barrier height from φ_b toφ_b'.The faster cooling – the less thickness of p^+ - layer. While hardening, namely fast cooling, silicon does not have time to re-crystallization, and remains in aluminum. Silicon addition before, after, or during depositing the aluminum film really reduces or even removes the process of dissolution of sub-contact layer of silicon substrate in metallization, but should not influence the process of silicon re-crystallization after annealing. So in this case, the formation of contact $Al - p^+ - nS$ should be expected after annealing. If the silicon concentration in aluminum exceeds an optimal value (that respects solubility at the annealing temperature) then p^+ - layer starts to be formed while annealing, and is characterized by high aluminum concentration.

The increase in thickness of either p^+ - layer l or acceptor concentration result in considerable growth of barrier height and ideality coefficient.

Since the formation ofp^+ - layer is held while annealing temperature is $450 \div 550$ ℃ then one could expect, at least for the structures cooled quickly after annealing, the instability of electrophysical characteristics during working or storage at high temperatures.

Thermal aging of contacts $Al - nSi$. At fig. 18.11.there is shown a change in $I - V$characteristics of contacts $Al - nSi$ in semi-logarithmical scale after endurance at 300℃ during different times.

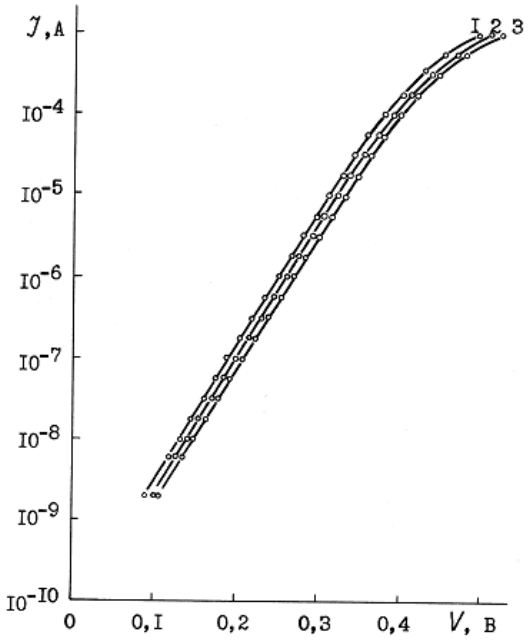


Fig. 18.11. $I - V$ characteristics of contacts $Al - nSi$ before – curve 1, and after endurance at temperature 300℃ during different times: 2 - $t = 3$ hours., 3 - $t =$ 16 hours.

One could see that thermal aging result in a slight decrease of current and slope of characteristic in whole voltage range.

By these characteristics there are identified dependences of barrier height and ideality coefficient, that are presented at fig. 18.12.

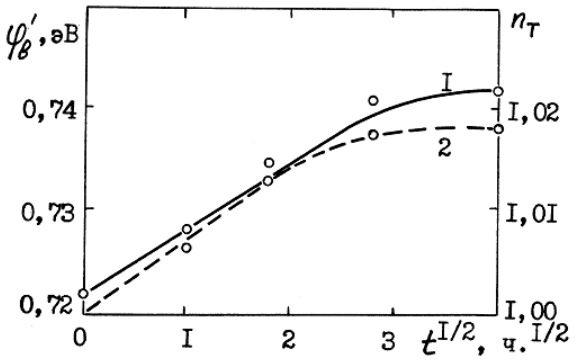


Fig. 18.12. Dependence of barrier height $\varphi_b'$and ideality coefficient n of $Al-nSi$ contacts on aging time at 300℃

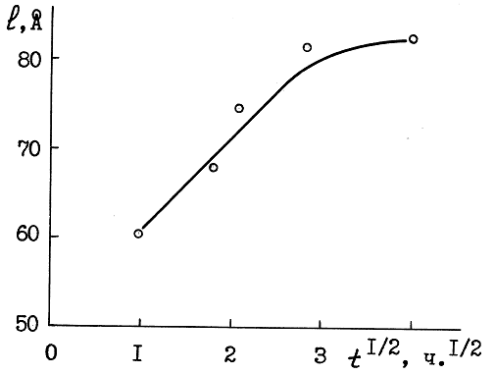


Fig. 18.13Dependence of p^+ - layer thickness l in contact $Al-nSi$ on aging time at the temperature of 300℃

These dependences could be explained by growth of p^+ - layer thickness from $l \approx 60$Å to $l \approx 80$Å with acceptor concentration $N_a = (7 \div 8)10^{17}$см$^{-3}$, due to re-crystallization of silicon that remained in aluminum after annealing (fig. 18.13).

As can be seen, if aging period is approximately to $8 \div 9$ hours then $l \sim t^2$. It means that re-crystallization process is restricted (controlled) by diffusion.

18.6.4. Processes of degradation in IC with multilayer metallization

Solid-phase interaction at the boundary $Al-Si$ considerably influences the problem of manufacturing the Shottki barriers with reproducing characteristics either in various working cycles or within single silicon board, and also result in the characteristics instability, could be the cause of device defects and failures. The large-scale research aimed to find out alternative materials for IC metallization had no result. Aluminum remains single material for creating single-layer metallization.

In order to increase quality and reliability of metal-semiconductor contacts (either rectifying or ohmic) there is used multi-layered metallization. In order to form Shottky barriers in such layers there are used metal silicide with metallic properties, for example, Pd_2Si, or $PtSi$, and wiring performed based on multi-layered systems, for example, $MoAu$, $MoPt$, $Ti-Pt-Au$, or $Ti-Pt$.

Ifwiringismadewithaluminumthatcontactdirectlytosilicide$PtSi$, orPd_2Si then the height of Shottki barrier is dependent on annealing the respective structures$Al-PtSi-nSi$, or$Al-Pd_2Si-nSi$. The depositing of additional layer ofTiWbetween aluminum and silicide stabilize

considerably the contact parameters, and in contacts without aluminum aging is not observed at all. So, degradation of characteristics of contacts $Al-PtSi-nSi,$ and $Al-Pd_2Si-nSi$ is connected with the aluminum. Thus, the degradation of a contact of $Al-PtSi-nSi$ takes place due to interaction between Al and $PtSi$ with the formation of a compound Al_2Pt (Fig.18.14). When $PtSi$ will react completely then a contact of $Al_2Pt-nSi$ forms. The process of degradation of the contact $Al-PtSi-nSi$ is controlled by Aldiffusion to the reaction area. Activation energy of the process is$E_a = 1.8$eV. Since the formation of the contact Al_2Pt-nStakes place non-simultaneously at all area of the contact window then during some time the contact is a parallel connection of two contacts $PtSi-nSi$ and Al_2Pt-nSi with different barrier height - $\varphi_b \approx 0.85$ eVand $\varphi_b \approx 0.70$eV respectively.

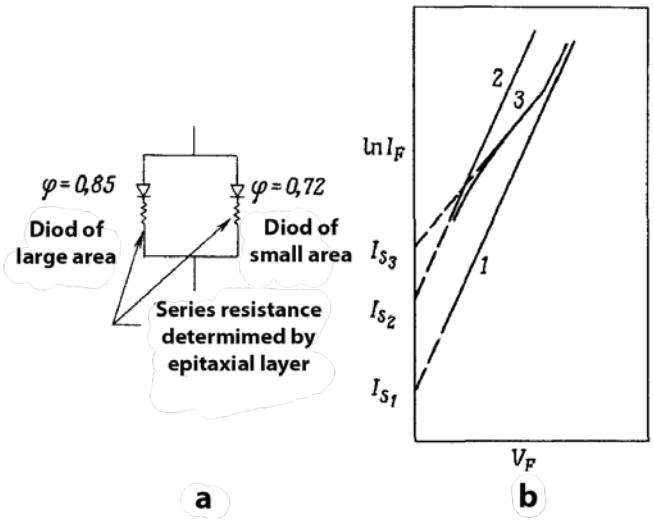


Fig. 18.14. Model of contact $Al-PtSi-nSi$–taking into account the region of area of $Al_2Pt-nSi$– (a), and change in CVC of the contact of $Al-PtSi-nSi$ when annealed – (b).

While Al_2Pt has no contact to silicon, curren-voltage characteristics shows $\varphi_b \approx 0.85$ (curve 1 at fig.18.14b). When the local contact between $Al_2Pt-nSi$ and$\varphi_b \approx 0.70$ forms then much more bigger current density will not result in considerable increase of current through the contact. However, since the series, relatively to the local contact Al_2Pt-nS, base resistance is more considerably than this of other circuit then the voltage drop makes itself felt at smaller bias voltage. So that, the considerable contribution to the whole current through the contact $Al_2Pt-nSi$ would be only at a small voltage (curve 2 at fig.18.14b). While increase of contact area of $Al_2Pt-nSi$, there is increase not only of its contribution to the total current, butthe voltage which restricts considerably this current due to voltage drop across the base resistance. Finally, the contact $Al_2Pt-nSi$ is formed for a whole contact area and current-voltage characteristics show the barrier height of $\varphi_b \approx 0.70$ eV(curve 3 at fig.18.14b).

For the structures $Al - Ti - nSi$ and $Al - V - nSi$, the same instability of characteristics is observed that is determined by formation of compounds $TiAl_3$ and VAl_3, respectively. These processes are also controlled by diffusion of aluminum, and are characterized by activation energy $E_a = 1.8$ eVand $E_a = 1.7$ eB respectively.

Institute of High Technologies

Taras Shevchenko National University of Kyiv

Laboratory work

Determination of surface resistivity of semiconductor wafers and technological layers using four-probe method

Department
Nanophysics of Condensed Matter

Determination of surface resistivity of semiconductor wafers and technological layers using four-probe method

Objective: To determine the surface resistivity distribution of silicon wafers and technological silicon layers on Sapphire by four-probe method.

1. Basic theory of the four-probe method for surface resistance investigation

One of the main electrical parameters of semiconductor materials is their resistivity ρ (Ohm · cm). The resistance of the material is determined by the concentration of charge carriers n and their mobility μ

$$\rho = \frac{1}{qn\mu} \qquad (1)$$

where q is charge of the electron.``

In the study of thin layers the value of resistance per unit area or surface resistance is used (Ohm/□ - Ohms per square). Surface resistance Rs of homogeneous layer is defined through ρ

$$R_S = \frac{\rho}{t} \qquad (2)$$

where t – is thickness of the layer.

The most common method of measurement in microelectronics technology is the four-probe method. Consider its relation to a semi-infinite sample. On its surface are positioned four thin metal probes S equidistant from each other, Figure 1. All probes are collinear.

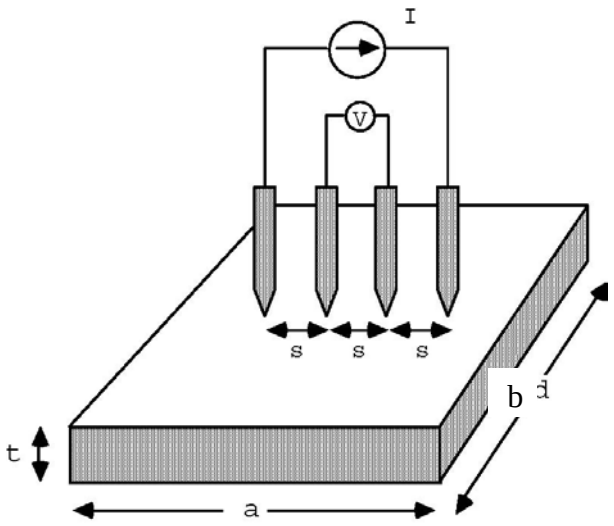


Fig. 1

Through the external probes 1 and 4 passes DC current I, and between probes 2 and 3 voltmeter measures potential difference U. In this case,

$$\rho = 2\pi S \frac{U}{I} \qquad (3)$$

In fact, in many cases it is necessary to determine the value of ρ in thin wafer of semiconductor material and technological layers with a thickness much smaller than the distance between the probes S (t «S). Then

$$\rho = 4,532t \frac{U}{I} \qquad (4)$$

For thicker layers, where t / S> 0,4 ratio (4) is specified by using the correction function Ft, depending on the thickness t

$$\rho = 4,532tF_t \frac{U}{I} \qquad (5)$$

The Ft values are listed in the table 1

Table 1

t/s	Ft
0.4	**0.9995**
0.5	**0.9974**
0.5555	**0.9948**
0.6250	**0.9898**
0.7143	**0.9798**
0.8333	**0.9600**
1.0	**0.9214**

If the measurements are carried out near the edge of the plates or layers, the boundary effects must be taken into account, i.e. the influence of the ratio D/S, where D - diameter in the case of the sample as a circle or b/S and a/b in the case of a rectangular sample. In this case, the relation (4) is specified by using the correction function F_D, depending on the D/S,, b/S and a/b.

Then for resistivity have

$$\rho = 4,532tF_D \frac{U}{I} \qquad (6)$$

The F_D values are listed in the table 2

Table 2

D/S, b/S	Circle	Rectangle			
		a/b=1	a/b=2	a/b=3	a/b ≥ 4
1.0				0.9988	0.9994
1.25				1.2467	1.2228
1.5			1.4788	1.4893	1.4893
1.75			1.7196	1.7238	1.7238
2.0			1.9454	1.9475	1.9475
2.5			2.3532	2.3541	2.3541
3.0	2.2662	2.4575	2.7000	2.7005	2.7005
4.0	2.9289	3.1137	3.2246	3.2248	3.2248
5.0	3.3625	3.5098	3.5749	3.5750	3.5750
7.5	3.9273	4.0095	4.0361	4.0362	4.0362
10.0	4.1716	4.2209	4.2357	4.2357	4.2357
15.0	4.3646	4.3882	4.3947	4.3947	4.3947
20.0	4.4364	4.4516	4.4553	4.4553	4.4553
40.0	4.5076	4.5120	4.5129	4.5129	4.5129
∞	4.5324	4.5324	4.5324	4.5324	4.5324

2. Equipment and samples for research

Determination of surface resistance is performed using automated equipment RS 30 "Prometrix" (USA). Accuracy of Rs is 1% .

Measurement range is within Rs $1 \cdot 10^{-3}$ - $1 \cdot 10^{4}$ Ohm·cm. Minimum size of the sample to determine the parameter at one point (in the center) should be at least 10 x 10 mm, the maximum size of the analysis of the parameter area distribution must not exceed 200 mm in diameter. The maximum number of points in the measurements on the plate is 625.

The samples are n- type conductivity silicon wafers, p-type conductivity silicon wafer and silicon-on-sapphire. All samples should be pre-determined by the thickness of silicon.

3. Work plan

1. Switch on the RS 30 equipment according to the instructions on its use.
2. Load the automation program on the computer.
3. Place the plate under study on the experimental table.
4. In the total menu select the type of computer research "Scan by diameter"
5. Menu "Scan by diameter" (Fig. 2): set the main parameters determining the resistivity of the plate:

- Number of points in the analysis,
- The diameter of the plate,
- Diameter of the scanning area,
- Direction of the scan,

- Select mode of the measurements (automatic or manual)
6. Start the measurements.
During scanning along diameter of the plate the Rs values are displayed on the screen.
7. After completion of measurement: calculation of the resistivity, statistical analysis of the results and building the ρ distribution along diameter of the sample.

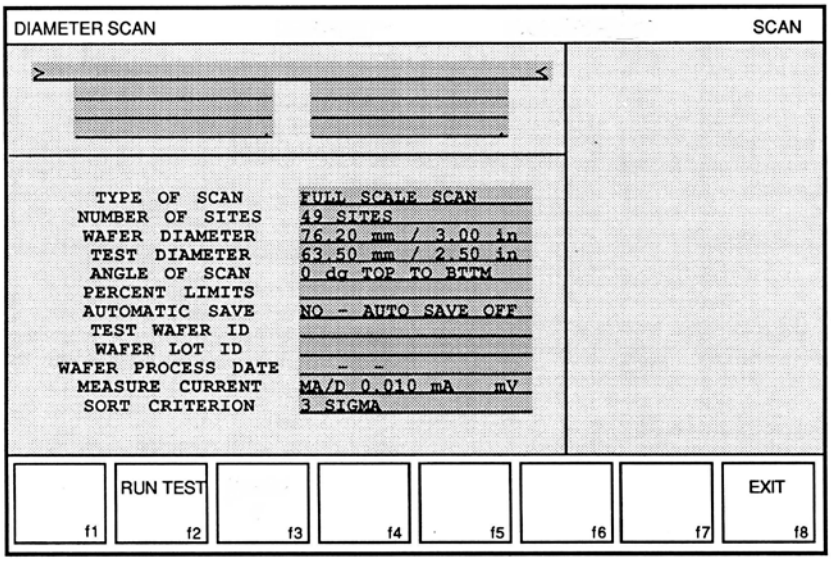


Fig. 2

4. Control questions

1. Which physical parameters of the material determine its resistance?
2. Essence of the four-probe method for determining the resistance of semiconductor materials
3. What physical sense is in correction functions when determining resistance of semiconductor materials

5. Literature

1. L.P Pavlov. Measurement methods for semiconductor materials, "Higher School", Moscow, 1987
2. S. Zi. Physics of semiconductor devices, v.1, "Mir", Moscow, 1984
3. Suhano T., T. Ykoma, E. Takeysy. Introduction to microelectronics. "Mir", Moscow, 1988

Institute of High Technologies

Taras Shevchenko National University of Kyiv

Laboratory work

Investigation of MIS structures electrical parameters in microelectronics technology by high-frequency C-V measurements

Department
Nanophysics of Condensed Matter

Investigation of MIS structures electrical parameters in microelectronics technology by high-frequency C-V measurements

Objective: Identify the main electrical parameters of MIS structures by high-frequency C-V measurements.

1. Theory basics of high-frequency C-V characteristics of MIS structures

Structure of metal-insulator-semiconductor (MIS) is one of the major structural and technological elements of semiconductor devices and integrated circuits. A typical MIS structure consists of a semiconductor substrate, a thin dielectric layer and deposited metal or other material that conducts electricity (e.g., polysilicon), Figure 1.

When applying voltage to the metallic electrode, electric field penetrates the thin insulator into the semiconductor, which creates a space charge region (SCR). This phenomenon is called the field effect. Depending on the sign and magnitude of the voltage on the electrode (gate) V_g the four energy states can be implemented in the surface region of the semiconductor: enrichment of the major carriers, depletion, weak and strong inversion. Gate voltage is the sum of the voltage drop across the insulator V_0 and semiconductor ψ_S.

$$V_g = V_0 + \psi_S \qquad (1)$$

Surface potential ψ_S - this is one of the main parameters of the field effect in semiconductors. The solution of Poisson's equation gives the dependence of the charge in the SCR Q_{SC} from ψ_S

$$Q_{sc} = \varepsilon_s \varepsilon_0 E_s = \pm \frac{\sqrt{2}\varepsilon_s \varepsilon_0 kT}{qL_d} \cdot F(\psi_s, \varphi_0) \qquad (2)$$

where L_d is Debye screening length.

$$L_d = \sqrt{\frac{kT}{q} \cdot \frac{\varepsilon_s \varepsilon_0}{qN_d}}$$

Function $F(\psi_S, \varphi_0)$ for nondegenerate semiconductor of p-type has the form

$$F(\psi_s, \varphi_0) = \sqrt{\left(e^{-\beta\psi_s} + \beta\psi_s - 1\right) + e^{-2\beta\varphi_0}\left(e^{\beta\psi_s} - \beta\psi_s - 1\right)} \qquad (3)$$

$\beta = q / kT$, q - electron charge, k - Boltzmann constant, T - temperature,
N_d - dopant concentration,
φ_0 - the position of the Fermi level in the semiconductor quasi-neutral,
ε_0 - absolute dielectric permittivity
ε_S - dielectric constant of the semiconductor

Change in surface potential causes a charge change in the semiconductor QSC. Then the capacity of the SCR has the form

$$C_{sc} = \frac{\partial Q_{sc}}{\partial \psi_s} = \frac{\varepsilon_s \varepsilon_0}{\sqrt{2} L_d} \cdot \frac{\sqrt{(1 - e^{-\beta\psi_s}) + e^{-2\beta\varphi_0}(e^{\beta\psi_s} - 1)}}{F(\psi_s, \phi_0)}$$

At enrichment ($\Psi_S<0$): $C_{sc} = C_p = \frac{\varepsilon_s \varepsilon_0}{L_d} e^{-\beta\psi_s}$

At depletion ($\varphi_0>\Psi_S>0$) or weak inversion ($2\varphi_0>\Psi_S>\varphi_0$)

$$C_{sc} = C_B = \sqrt{\frac{\varepsilon_s \varepsilon_0 q N_B}{2(\varphi_0 - \frac{kT}{q})}} = \frac{\varepsilon_s \varepsilon_0}{W},$$

where W – SCR width.

At strong inversion ($\Psi_S>2\varphi_0$) $C_{sc} = C_n = \frac{\varepsilon_s \varepsilon_0}{L_d} e^{\frac{\beta(\psi_s - \varphi_0)}{2}}$

At «flat zones» ($\Psi_S=0$) $C_{sc} = C_{FB} = \sqrt{\frac{\varepsilon_s \varepsilon_0 q N_B}{\frac{kT}{q}}} = \frac{\varepsilon_s \varepsilon_0}{L_d}$

The capacity of the whole MIS structure can be represented as a chain of series-connected capacitance of the dielectric (oxide) C_{OX} and capacitance of the semiconductor SCR C_{SC}. In the presence of surface states their capacitance C_{ss} is connected in parallel with capacitance SCR, Figure 2.
Then the capacity of the MIS structure C is

$$C = C_{ox}\left(1 - \frac{C_{ox}}{C_{ox} + C_{sc} + C_{ss}}\right)$$

The gate voltage is

$$V_g = V_{FB} + \psi_s + \frac{qN_{ss}\psi_s}{C_{ox}} + \frac{Q_{sc}}{C_{ox}}$$

Voltage of "flat zones» V_{FB} in the absence of surface states:

$$V_{FB} = \Delta\varphi_{ms} - \frac{Q_{ox}}{C_{ox}}$$

where $\Delta\varphi_{ms}$ - contact potential difference between the substrate and the gate
Q_{OX} – fixed charge in the dielectric.

In the study of MIS structures by high-frequency C-V measurements performed on alternating test signal with a period much shorter than the lifetime of minority carriers and surface states overcharging time. Then the actual capacity of the semiconductor is only the capacitance of SCR.

Applying the gate voltage Vg, generates a change in surface potential and thus the capacity of the MIS structure, getting the high-frequency C-V, fig. 3.

On the basis of high-frequency C-V the following important parameters are found:

1. Conductivity type semiconductor.
2. Thickness of the dielectric.

$$d_{ox} = \frac{\varepsilon_{ox}\varepsilon_0 S}{C_{max}}$$

where S – the gate surface, $C_{max}= C_{OX}$

3. The level of doping in the semiconductor N_B using minimum capacity C_{min}

$$C_{min} = \frac{\varepsilon_s\varepsilon_0}{d_{ox} + \frac{\varepsilon_s}{\varepsilon_{ox}} W_{max}}$$

$$N_B = \frac{4\varphi_0}{q\varepsilon_s\varepsilon_0} \cdot \frac{C_{min}}{\left(1 - \frac{C_{min}}{C_{max}}\right)}$$

$$\phi_0 = \frac{kT}{q}\ln\frac{N_B}{n_i}$$

where n_i concentration of intrinsic charge carriers in the semiconductor. Since φ_0 is also affected by N_B, the doping level obtained by iteration.

4. Fixed charge in the dielectric Q_{OX} (on experimentally determined V_{FB}).
5. Threshold voltage V_T.

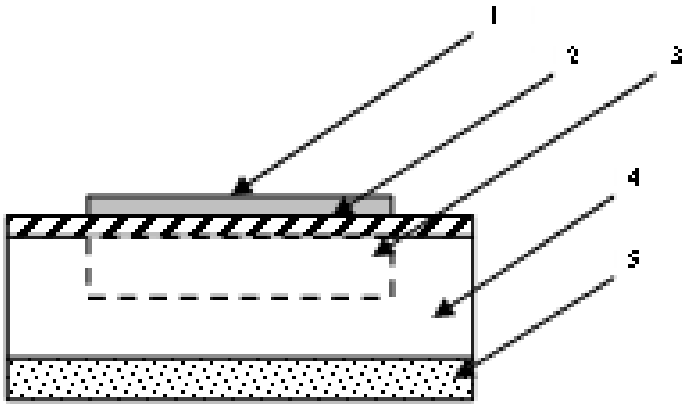


Fig.1 MIS structure: 1- electrode, 2- dielectric, 3- space charge region in semiconductor, 4- quazi neutral region in semiconductor, 5-Ohmic contact

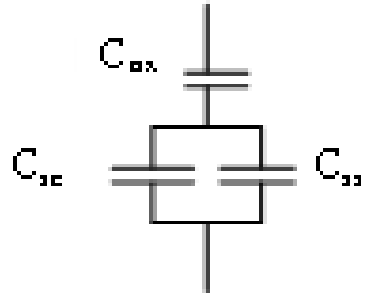


Fig.2 Equivalent schema of MIS structure

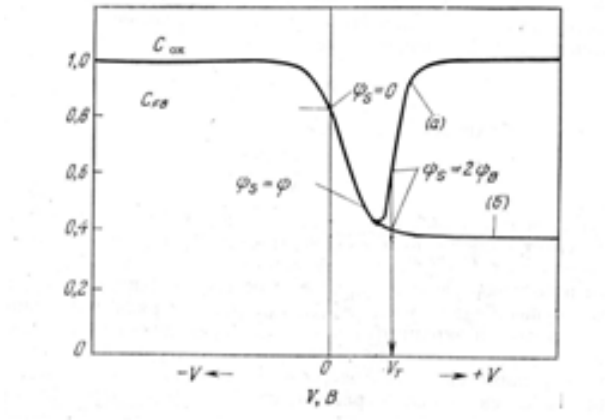


Fig.3 C-V curve of MIS structure at low (a) and high (b) frequency

Fig. 4. Equipment HP 4061A

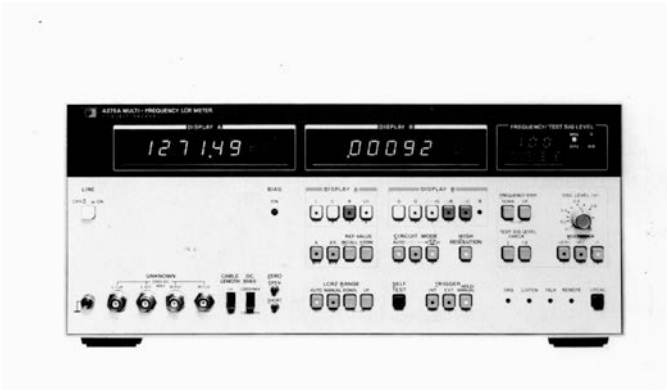


Fig.5 High-frequency C-V measurement device HP4275A

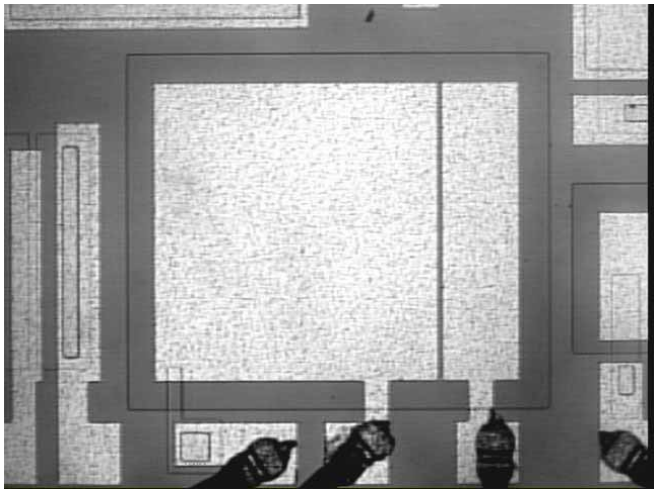


Fig 6. Test structure CT45T

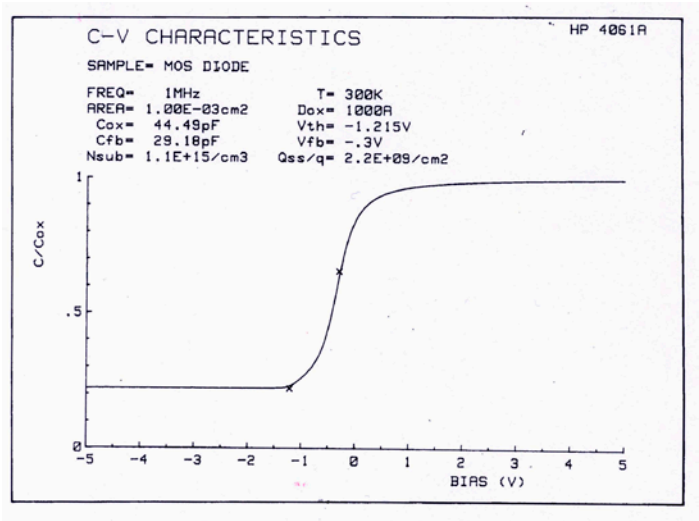


Fig.7. High-frequency C-V of the test structure

2. Equipment and samples for research

Research of high-frequency C-V is performed using automated installation NR4061A (company "Hewlett-Packard", USA) to determine the electrical parameters of semiconductor devices, Fig. 4, equipped with a precision capacitance meter NR4275A ,Figure 5. Capacitance sensitivity is 10^{-15} F, measurement accuracy depends on the capacitance and signal frequency is

in the range 0.1-5 %. Typical measurements are made at a frequency of 1MHz and signal amplitude of 25 mV, respectively. The maximum number of points in the analysis of high-frequency C-V is 200.

The contact to the sample (test MOS structures) is made on an analytical probe station MM 7000 (firm "Micromanipulator", USA).

Samples for research are ST45T test patterns produced by MIS technology for p-channel integrated circuits on n-type silicon, Figure 6. Investigated MIS capacitors formed in accordance with the growing technological regimes during shutter oxide MIS transistors. The thickness of the under-gate oxide is 1200-1400 Å.

Typical high-frequency C-V is received at the facility NR4061A shown in Figure 7. On the basis of the characteristics the values V_{FB}, Q_{OX}, V_T, N_B, d_{ox} are calculated.

3. Work plan.

1. Switch on the equipment NR4061A according to the instructions for its operation and load the operating system.
2. Put the plate with test structures on a table ST45T sample analytical probe station MM 7000.
3. Connect analytical probe station with the equipment NR4061A.
4. Upload work program for measuring high-frequency C-V "HFALLAlPRC"
5. On interactive mode:
   - Set the measurement mode (frequency, voltage interval, then test the voltage level of the signal).
   - Enter the area of the sample.
   - Calibrate the equipment.
6. Perform high-frequency C-V measurements.
7. Plot high-frequency C-V and calculate the parameters of the MIS structure: V_{FB}, Q_{OX}, V_T, N_B, d_{ox}

4. Control questions

1. Which physical phenomena on the surface of the semiconductor MIS structures occur in the process of changing the external voltage from accumulation to inversion?
2. Give an example of a typical high-frequency C-V of MIS structure?
3. How does the change in the thickness of the dielectric influence on the form of normalized high-frequency C-V of MIS structure?
4. How does the change in the concentration of dopant in the semiconductor influence on the form of normalized high-frequency C-V of MIS structure?
5. How is the value of the fixed charge in the dielectric of MIS structure calculated?

5. Literature

1. S. Zi. Physics of semiconductor devices, v.1, "Mir", Moscow, 1984
2. R.Maler, T.Keymyns. Elements of integrated circuits. "Mir", Moscow, 1988

Institute of High Technologies

Taras Shevchenko National University of Kyiv

Laboratory work

Research of dopant concentration profiles at surface of semiconductor in microelectronics technology

Department
Nanophysics of Condensed Matter

Research of dopant concentration profiles at surface of semiconductor in microelectronics technology

Objective: Determine the dopant profile in the surface layer of semiconductor by high-frequency C-V measurements for the Schottky diode.

1. Basic theory of the high-frequency C-V measurements for the Schottky diode

The metal-semiconductor junctions (Schottky diodes) are the basis of many semiconductor devices (photodetectors, sensors, capacitors with variable capacity) and an important element in the design of integrated circuits (e.g., FETs of microwave electronics). In addition, the Schottky diode is simply a technological framework for the study of important electrical properties of the semiconductor, including dopant concentration profile at the surface. Schottky diodes are widely used in the control of semiconductor parameters in microelectronics technology. In some cases, for the rapid analysis of dopant concentration, the mercury contacts (probes) are used instead of metal electrode, so the Schottky diodes are created on chemically clean surface of silicon or gallium arsenide.

The main feature of Schottky diodes is the formation of depletion region for the major carriers (space charge region) on the semiconductor surface under the metal electrode, which is characterized by integrated potential φ_i. When applying a reverse voltage V_g the space charge region (SCR) is distributed in the depth of the semiconductor. The value of the space charge per unit area is of the form

$$Q_S = \left[2q\varepsilon_0\varepsilon_s N_d \left(\phi_i - V_g \right) \right]^{1/2} \qquad (1)$$

where $\beta = q / kT$, q - electron charge, k - Boltzmann constant, T - temperature,
N_d - dopant concentration,
φ_0 - the position of the Fermi level in the semiconductor quasi-neutral,
ε_0 - absolute dielectric permittivity
ε_S - dielectric constant of the semiconductor

Corresponding SCR capacity is:

$$C_S = \frac{dQ_S}{dV_S} = \left[q\varepsilon_0\varepsilon_s N_d / 2\left(\phi_i - V_g \right) \right]^{1/2} = \frac{\varepsilon_0\varepsilon_s}{x_d} \qquad (2)$$

where x_d is SCR thickness.

In the study of Schottky diode using high-frequency C-V measurements the capacitance measurements performed on alternating test signal with a period much shorter than the lifetime of minority carriers and the surface states overcharging time. Then the actual capacity of the semiconductor is the only SCR capacity.

Applying the reverse voltage to the gate V_g, the SCR extension in depth of the semiconductor occurs as well as the associated decrease in capacitance of structure obtained by high-frequency C-V measurements

From eq. (2):

$$\phi_i - V_g = q\varepsilon_0\varepsilon_s N_d / 2C^2 \quad (3)$$

Thus, the graph of $1/C^2$ dependence from V_g is a straight line. Knowing its slope N_d can be determined. The point of its intersection with the abscissa gives the value φ_i.

Typical high-frequency C-V of Schottky diode on silicon is shown in Figure 2. Example of $1/C^2$ dependence from Vg is shown in Figure 3.

Measurement of high-frequency small-signal (differential) capacitance also allow to determine the concentration of dopant profile in the case where the value of N_d varies with distance from the surface of the semiconductor.

Then

$$N_d(x) = \frac{2}{q\varepsilon_0\varepsilon_s A^2}\left(\frac{d\frac{1}{C^2}}{dV}\right)^{-1} \quad (4)$$

where the SCR thickness is determined by the value of high-frequency C-V of Schottky diode.

$$W = \frac{\varepsilon_0\varepsilon_s A}{C_S} \quad (5)$$

Note that when calculating the profile of $N_d(x)$ it would remain not defined at a distance from the surface x_0, the starting SCR thickness, or the built-in potential φ_i. It is obvious that with increasing N_d value of x_0 decreases.

2. Equipment and samples for research

Research of high-frequency C-V is performed using automated installation NR4061A (company "Hewlett-Packard", USA) to determine the electrical parameters of semiconductor devices, Fig. 4, equipped with a precision capacitance meter NR4275A ,Figure 5. Capacitance sensitivity is 10^{-15} F, measurement accuracy depends on the capacitance and signal frequency is in the range 0.1-5 %. Typical measurements are made at a frequency of 1MHz and signal amplitude of 25 mV, respectively. The maximum number of points in the analysis of high-frequency C-V is 200.

The contact to the sample (test MOS structures) is made on an analytical probe station MM 7000 (firm "Micromanipulator", USA).

Samples for research are Schottky diodes in test structures ST45T produced by MIS technology for p-channel integrated circuits on n-type silicon. We study the profile $N_d(x)$ in silicon Schottky diodes formed in areas where previously MOS transistors were grown and then under-gate oxide was etched. The thickness of the under-gate oxide is 1200-1400 Å.

Example of dopant profile in silicon obtained by NR4061A equipment is shown in Figure 6.

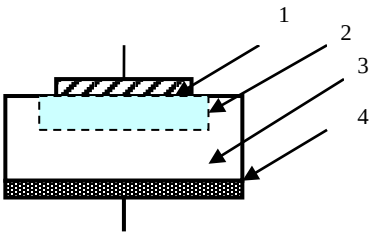


Fig. 1. Schottky diode. 1 - metal electrode, 2 - space charge region in the semiconductor, 3 - semiconductor, 3 - ohmic contact.

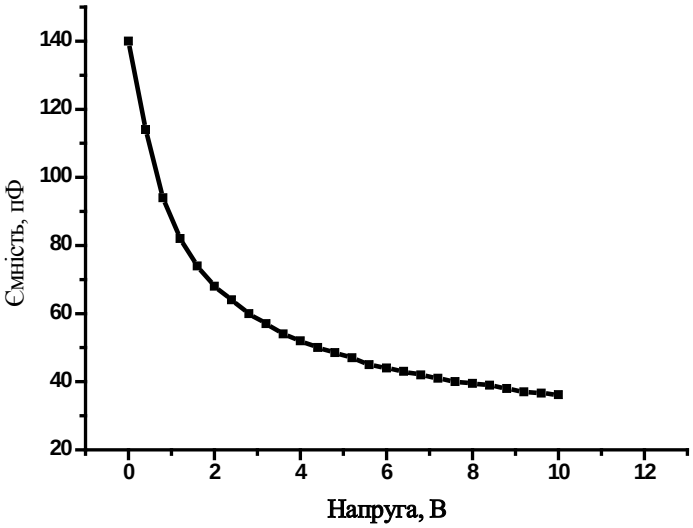


Fig. 2. Typical high-frequency C-V of the Schottky diode

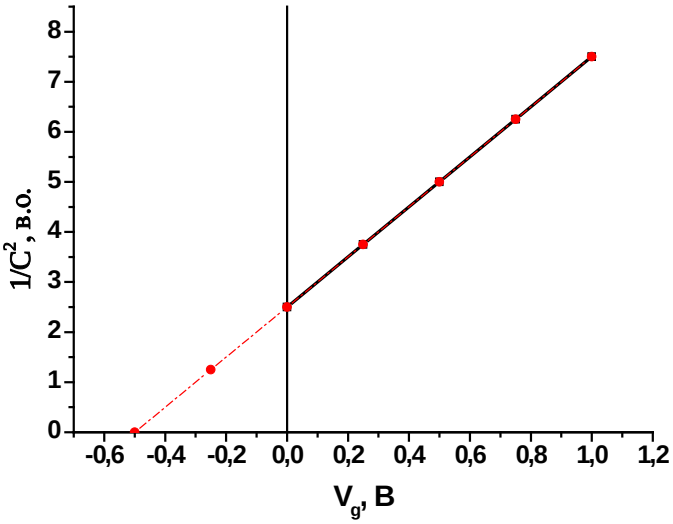


Рис. 3. $1/C^2$ dependence from voltage for the Schottky diode

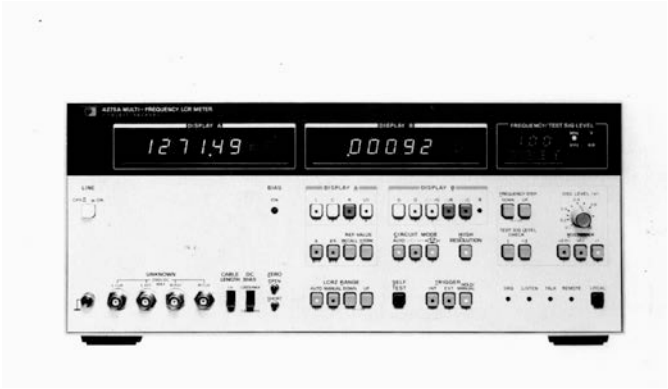
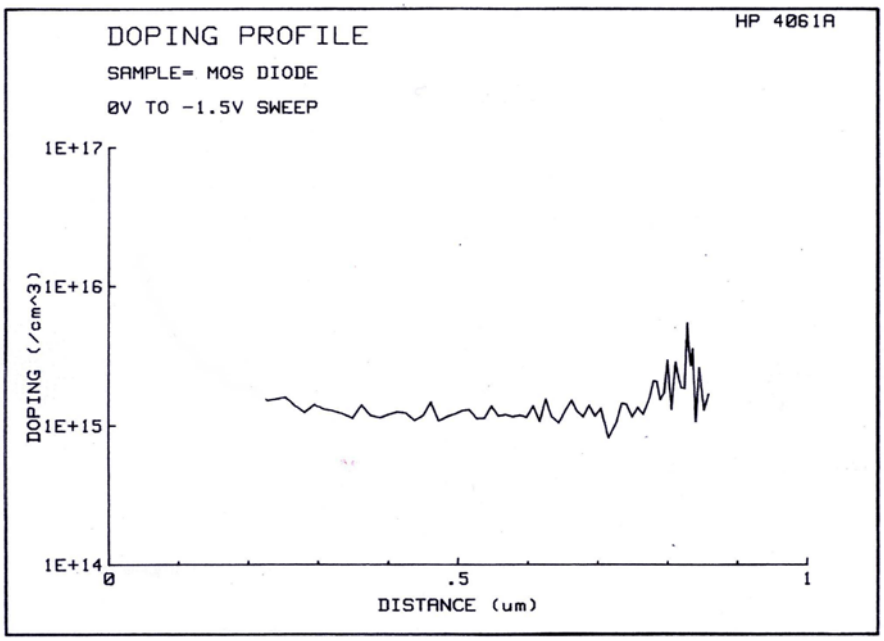


Fig. 4. Equipment HP 4061A Fig.5 High-frequency C-V measurement device HP4275A



.

Fig. 6. Example of dopant profile in silicon

3. Work plan.

1. Switch on the equipment NR4061A according to the instructions for its operation and load the operating system.

2. Put the plate with test structures on a table ST45T sample analytical probe station MM 7000.

3. Connect analytical probe station with the equipment NR4061A.

4. Upload work program for measuring high-frequency C-V "HFALLAlPRC"

5. On interactive mode:

- Set the measurement mode (frequency, voltage interval, then test the voltage level of the signal).
- Enter the area of the sample.
- Calibrate the equipment.

6. Perform high-frequency C-V measurements.

7. Plot high-frequency C-V and calculate dopant profile for the sample

4. Control questions

1. Physical content of high-frequency C-V measurements for determining the dopant profile in the surface layer of semiconductor for the Schottky diode?
2. Formulas determining dopant concentration profile in high-frequency C-V of the Schottky diode?
3. Typical view of high-frequency C-V of the Schottky diode?
4. How does the level of dopant concentration influence on the form of high-frequency C-V of the Schottky diode?
5. What is the reason for the inability to measure the concentration of dopant directly near the interface of a metal-semiconductor in the Schottky diode?

5. Literature

1. S. Zi. Physics of semiconductor devices, v.1, "Mir", Moscow, 1984
2. R.Maler, T.Keymyns. Elements of integrated circuits. "Mir", Moscow, 1988

Institute of High Technologies

Taras Shevchenko National University of Kyiv

Laboratory work

Investigation of electrical characteristics and parameters of the MIS transistors in microelectronics technology

Department
Nanophysics of Condensed Matter

Investigation of electrical characteristics and parameters of the MIS transistors in microelectronics technology

Objective: To measure the basic electrical characteristics of MOS transistors and determine their basic parameters

1. Basic theory of the MIS transistors

The transistor structure of the metal - insulator - semiconductor (MIS) is one of the most widespread microelectronic devices. Simple design MIS transistor (MIST) and high-density integrated circuit elements that are based on it, identified great practical importance of this device.

Simplified design of MIST is shown in Figure 1. MOS transistor is four-terminal device that consists of a semiconductor substrate, where two other heavy doped regions of other type conductivity (p-n junctions): source and drain are formed (by diffusion or ion implantation). The metal electrode is separated from the substrate dielectric layer, called the gate.

When an electric voltage on the gate is absent and the reverse voltage exists on the drain, the current between the source and the drain does not exist (in fact it is a small reverse current of p-n junction flow). If the applied gate voltage high enough, a thin inversion layer or channel is formed on the surface, that connects the drain and source region. In this case, between the source and drain regions exists an electric current. The magnitude of this current is modulated by the gate voltage.

The basic structural and technological parameters of MIST is channel length is L - the distance between the borders of p-n junctions of source and drain, width Z, the thickness of the insulator layer d_{ox}, deep p-n junctions and substrate doping level N_d.

Consider the basic expressions for the current in channel of MIS transistor on a substrate of n-type conductivity depending on the gate voltage V_G and the drain V_D.

At low voltages V_D ($V_D << V_G - V_T$, where V_T - threshold voltage at which the channel begins to leak current) mode is implemented so-called "smooth" channel. In this mode, the electric field gradient in the direction of the current in the channel is much smaller than the gradient of the electric field in the direction perpendicular to the current. In this case we have

$$I_D = \frac{Z}{L}\mu_p C_{ox}\left(V_G - V_T\right)V_D - {V_D}^2/2 \qquad (1)$$

$$V_T = 2\psi_B + \frac{\sqrt{2\varepsilon_S\varepsilon_0 qN_D}\,2\psi_B}{C_{OX}} \qquad (2)$$

$$\psi_B = \frac{kT}{q}\ln\frac{N_D}{n_i} \qquad (3)$$

$$C_{ox} = \frac{\varepsilon_0\varepsilon_{ox}}{d_{ox}} \qquad (4)$$

where q - electron charge, μ_p - mobility of minority carriers (holes) in the channel, C_{ox} - specific capacitance of the dielectric, ε_0 - absolute permittivity, ε_S - relative permittivity of the semiconductor, ε_{ox} - relative permittivity of the dielectric under the gate, ψ_S - potential in the semiconductor, k - Boltzmann constant, T - absolute temperature, n_i - concentration of intrinsic charge carriers in the semiconductor, d_{ox} - the thickness of the dielectric.

Equation (1) describes the output current- voltage characteristics (I-V) of MIS transistor in a steep area ($V_D < V_G - V_T$) when the inversion layer exists along the entire length of the channel.

As V_D growth in the channel near the drain, the electric field contributes to the increase of space charge in the semiconductor and reduces minority carrier charge. As a result of the inversion layer near the drain disappears, the current in the channel reaches saturation. In the saturation current I-V output can be displayed around the horizontal line. Typical output current-voltage characteristics of the MIS transistor is shown in Figure 2. In fact, the saturation current continues to grow as a result of different mechanisms of action, including effect of modulation of the channel.

The equation for the saturation current, which is valid for a fixed value of V_D and provided $V_D > V_G - V_T$, called transfer characteristic and has the form

$$I_D = \frac{Z}{2L}\mu_p C_{ox}\left(V_G - V_T\right)^2 \qquad (5)$$

It reflects the quadratic dependence of current on gate voltage. From this formula, depending on the square root of the saturation current of the gate voltage, it is determined practically important value of the threshold voltage V_T. This value can also be determined from the current-voltage characteristics of transfer by direct measurement of the gate voltage at which the drain current reaches a certain value, such as 10 mA. It is this method used in the control of MIS transistor parameters. Typical output current-voltage characteristics of the MIS transistor is shown in Fig.2b .

Enhancement in MIS transistor small signal mode is characterized by its slope

$$g_m = \frac{\partial I_D}{\partial V_G} = \frac{Z}{L}\mu_p C_{ox} V_D \qquad (6)$$

which increases linearly with increasing voltage at the drain.

From the output I-V in the "smooth" channel in equation (1) the effective mobility of charge carriers in the channel μ_p can be obtained.

2. Equipment and samples for research

Research of MIST I-V is performed on the automated equipment NR4145A (company "Hewlett-Packard", USA) to determine the electrical characteristics of semiconductor devices, Figure 3. The minimum step of programmable voltage is 10^{-4} V. The maximum number of points in the analysis is 500. Maximum sensitivity of current is 10^{-12} A.

The contact to the sample (test MOS structures) is made on an analytical probe station MM 7000 (firm "Micromanipulator", USA).

Samples for research are MOS transistors in test structures ST45T of semiconductor manufacturing technology created by MIS p-channel integrated circuits on n-type silicon. MIS transistors are investigated with different geometrical parameters (length and width of the channel). The thickness of the under-gate oxide is 1200-1400 Å.

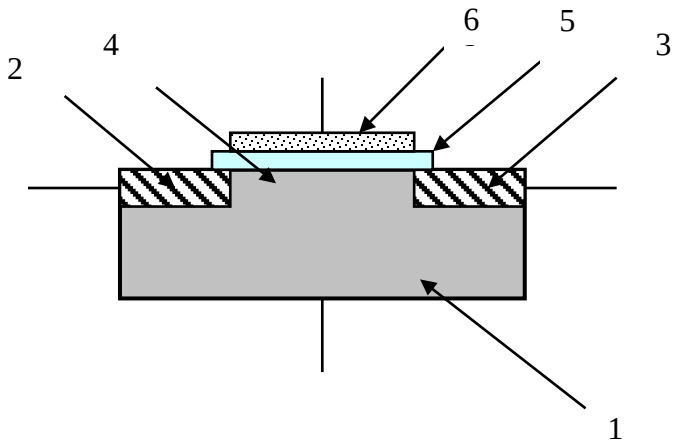
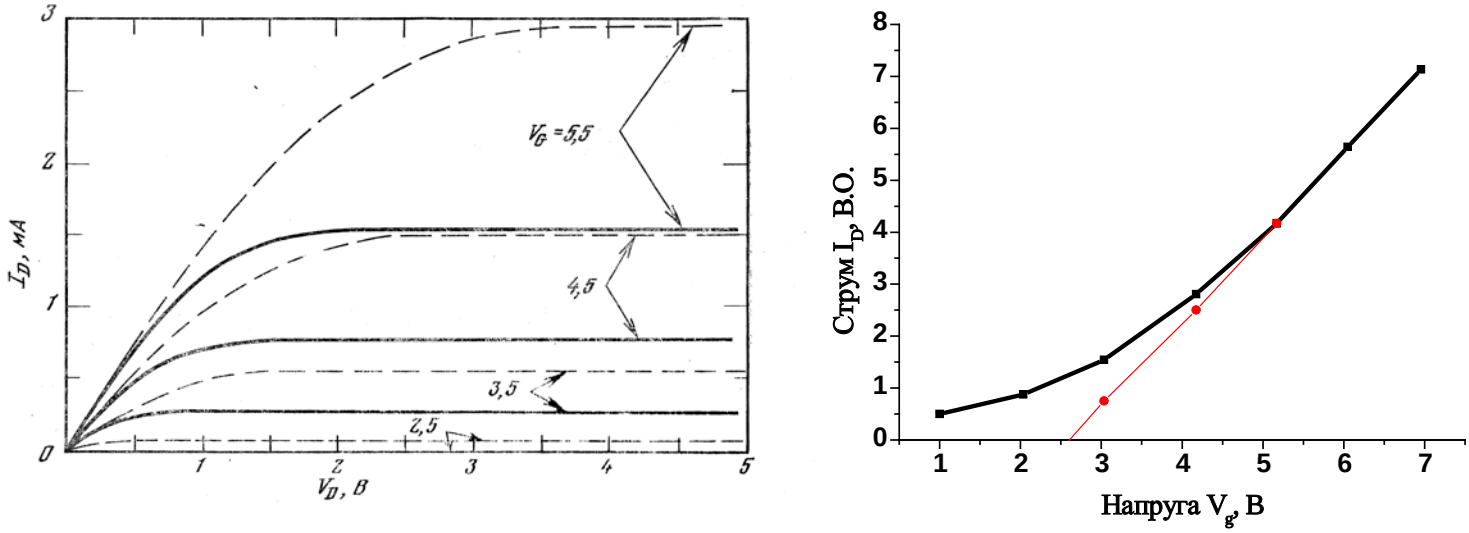


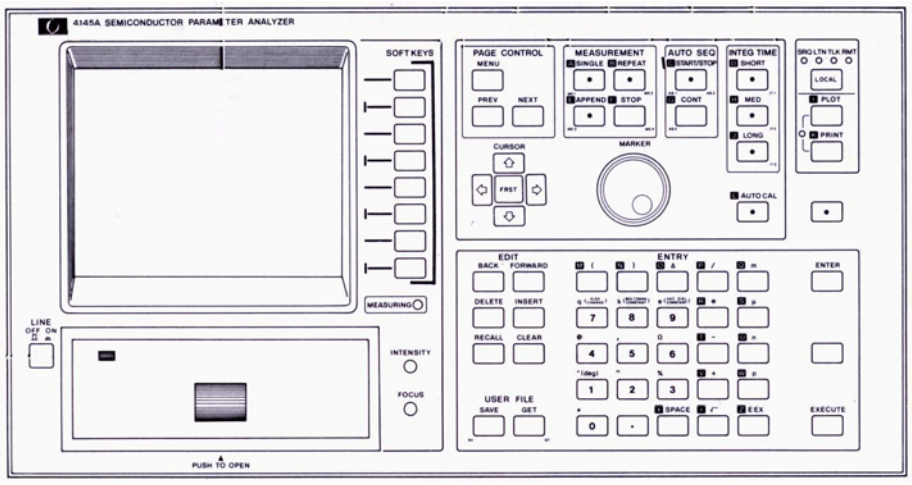
Fig. 1. Schematic view of MIS transistor: 1 – semiconductor, 2 - source, 3 - drain, 4 - channel region, 5 - insulator, 6 – control electrode (gate).



a) b)

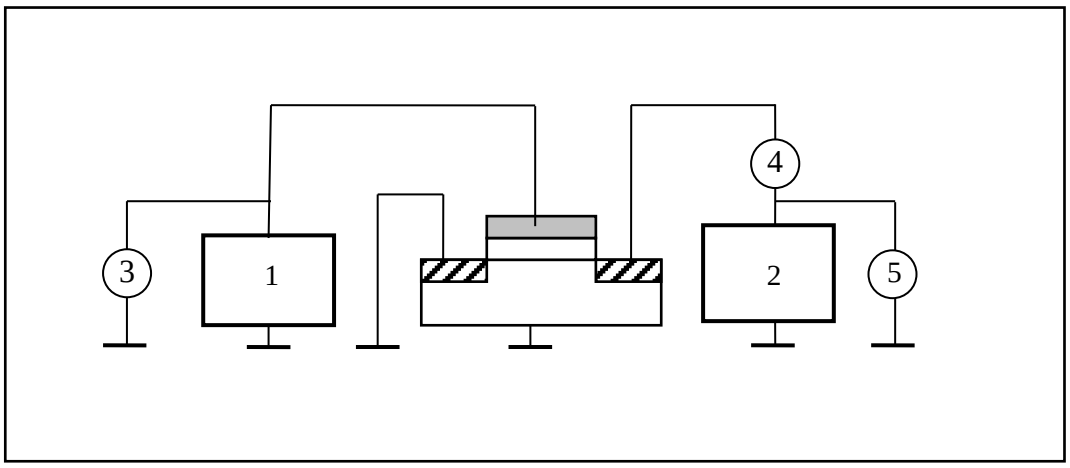
Fig. 2. Typical output current-voltage characteristics of two MIS transistors with different constructive and technological parameters (solid and dashed lines) (a). Typical transfer current-voltage characteristics of the MIS transistor (b)

a)

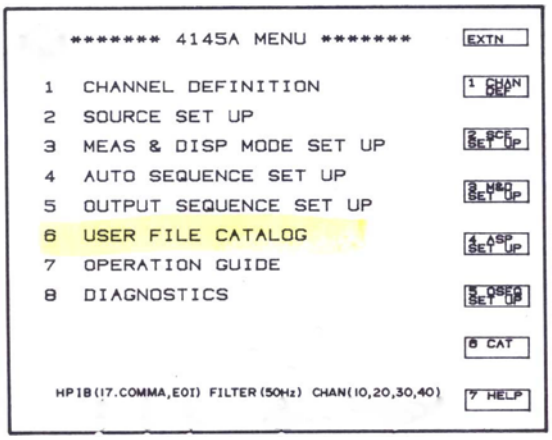


b)

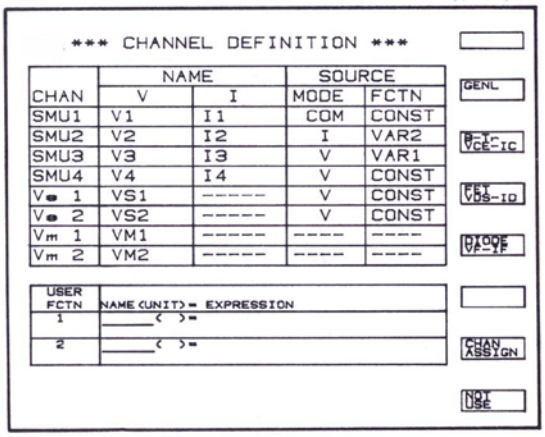
Fig. 3. General view of the automated equipment HP 4145 (a) for measuring MIS transistor current-voltage characteristics and its control panel (b)



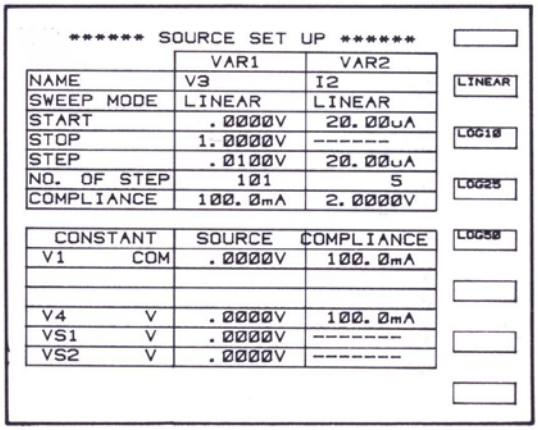
Rice. 4. Flow chart of measuring MIS transistor current-voltage characteristics: 1 - gate source voltage, 2 - drain voltage source, 3, 5 – voltage meters, 4 – ampere meter.



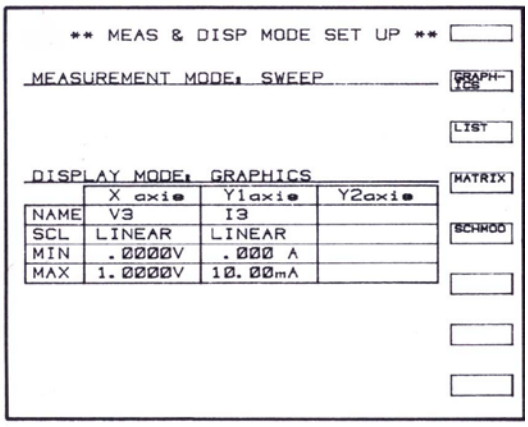
Menu № 1
System control and diagnostics



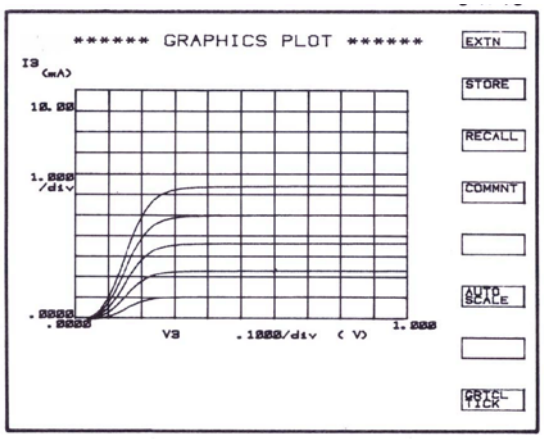
Menu № 2
Tracing sources and channels of measurements



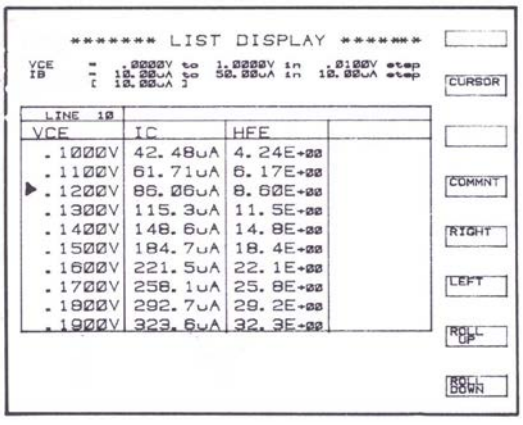
Menu № 3
Measurement modes



Menu № 4
Display parameters



Menu № 5
Building plots



Menu № 6
Table results ob measurements

Fig. 5. Sequence of main menu pages of equipment NR4145A for measuring I-V for MIS transistors when selecting modes and study characteristics

3.Work plan.

1. Gather electrical circuit of transfer and measurement of initial I-V of MIST according to Fig. 4.
2. Connect analytical probe station (measuring probes in micromanipulators) to the equipment NR4145A.
3. Put the plate with test structures ST45T on the sample table of analytical probe station MM 7000.
4. Switch on NR4145A according to the instructions on its operation and load an operating system.
5. Interactively, consistently passing different menus on the monitor, Fig. 5, set the power source voltage and registration currents: set status measurements (voltage range of V_D and V_G, voltage intervals, number of points and measurement speed, ranges of measuring currents I_D). Set the coordinate properties of final graphs.
6. Perform measurements of output and transfer I-Vs.
7. Plot the transfer and output I-Vs, and calculate values of threshold voltage V_T, slope of characteristics and mobility of charge carriers in channel in the so-called "smooth channel" approximation.

4. Control questions

1. Principle of operation of MIS transistor.
2. What is the mode of "smooth" channel?
3. Initial form of transfer and output I-Vs of MIS transistors.
4. How to determine the threshold voltage of MIS transistor?
5. How to determine the effective mobility of charge carriers in the channel of MIS transistor?

5. Literature

1. S. Zi. Physics of semiconductor devices, v.1, "Mir", Moscow, 1984
2. R. Krouford. Circuit applications of MOS transistors, "Mir", Moscow, 1970
3. R. Maler, T. Keymyns. Elements of integrated circuits. "Mir", Moscow, 1988

Institute of High Technologies

Taras Shevchenko National University of Kyiv

Laboratory work

Study of efficient lasing lifetime of minority carriers in a semiconductor MIS structures in microelectronics technology by dynamic nonequilibrium I-V

Department
Nanophysics of Condensed Matter

Study of efficient lasing lifetime of minority carriers in a semiconductor MIS structures in microelectronics technology by dynamic nonequilibrium I-V

Objective: To determine the effective lasing lifetime of minority carriers in a semiconductor of insulator-semiconductor (IS) and metal-insulator-semiconductor (MIS) structures by method of dynamic nonequilibrium current-voltage characteristics (DN I-V)

1. Basic theory of dynamic nonequilibrium current-voltage characteristics (DN I-V) of MIS structures

Effective lasing lifetime of minority carriers in a semiconductor MIS structures τ_{g0} is important electrophysical parameter characterizing the structural and impurity excellence of surface layer of the semiconductor in microelectronics structures, primarily silicon. Value τ_{g0} determine the reverse current p-n junctions, the electrical parameters of the charge-coupled devices (CCD), elements of dynamic memory devices in integrated circuits and other important characteristics of microelectronic products. Value τ_{g0} displays creation speed of electron-hole pairs in the semiconductor in a state of complete depletion of carriers, including the rate of formation of the inversion layer in MIS structure which transforms into non-equilibrium state after applying the inverse voltage. Conversion of MIS structure in equilibrium state takes time and depends on the dynamics of various generation processes of minority carriers in a semiconductor. Therefore, one of the most common ways to determine τ_{g0} is to analyze the speed of formation of the inversion charge in MIS structures.

In general, in the absence of minority carriers diffusion from the quasi-neutral semiconductor (at room temperature) establishing equilibrium inversion layer in MIS structure is the result of such mechanisms, Figure 1:

- The bulk generation in the space charge region (SCR) in the semiconductor;
- Generation at the electrode surface;
- Surface generation at the periphery of the electrode in thin layer, where the SCR is not covered by the electrode and reaches the surface of the semiconductor;
- Avalanche ionization in SCR;
- An abnormally high rate of generation due to the presence of some electrically active defects in the interfaces of IS.

The main source of charge generation in MIS structure is the bulk generation in SCR. Then in general τ_{g0} can be represented by the expression for bulk lasing lifetime of minority carriers:

$$\tau_{g0} = \tau_{p0} \exp\left[\left(E_T - E_i\right)/kT\right] + \tau_{n0} \exp\left[-\left(E_T - E_i\right)/kT\right] \qquad (1)$$

$$\tau_{p0} = 1/\sigma_p v_{th} N_t$$

$$\tau_{n0} = 1/\sigma_n v_{th} N_t$$

where N_t - concentration of generation active centers in semiconductor, σ_n, σ_p - capture cross section of electrons and holes by generation active center, v_{th} - thermal velocity of electrons, E_T, E_i - energy position of deep generation active center and the middle of the band gap of the semiconductor, k - Boltzmann constant, T - temperature.

One of the most effective methods for determining τ_{g0} is based on the use of dynamic nonequilibrium current-voltage characteristics (DN I-V) of MIS structures [1,2].

The DN I-V method is application of linearly increasing voltage (LIV) of inverse polarity to the MIS structure, that creates nonequilibrium conditions of the inversion layer formation, recording and analysis of capacitive charging current of the structure.

Analyze the main features DN I-V. Known that the application of an alternating voltage V (t) to a capacitor with capacitance C (V) causes capacitor current i_C, which has the form

$$i_C(V) = C(V)\frac{\partial V}{\partial t} \qquad (2)$$

In the case when V (t) = k • t, where k is constant growth rate of the signal, we have

$$i_C(V) = C(V)k \qquad (3)$$

Thus, the dependence of capacitive current on the applied voltage is directly shown by match the nonlinear capacitance of the MIS structure at a given rate of change of LIV.

When measuring DN I-V during application inverse LIV to the MIS structure the space charge region (SCR) on the surface of the semiconductor under electrode is gradually increasing. Nonequilibrium capacitive current in MIS structure decreases. But due to the growth of the SCR width the rate of generation of minority carriers in a semiconductor increases. There comes a time when the capacitive current reaches its minimum value i_{min} and begins to grow. In the usual case where the generation of minority carriers in the bulk has been determined by SCR is due to the classical mechanism of Shockley-Read-Hall, capacitive current generation gradually increases and reaches an equilibrium level of charging of the dielectric capacitance of the MIS structure i_0. During the generation of charge carriers in the MIS structure DN I-V crosses the value $i_{FB} = k \cdot C_{FB}$, where C_{FB} - capacity of "flat" zones, depending on the thickness of the insulator d_{ox} and concentration of dopant in the semiconductor, Figure 2. In the interval of inverse voltage ΔV_g, between levels i_{min} and i_{FB} DN I-V has a linear form. Approximation of DN I-V to the linear dependence makes possible to greatly simplify the expression of determining τ_{g0}, in particular to avoid the calculation of the surface potential in MIS structure. The result is

$$\tau_{g0} = \frac{A}{K}\sqrt{\Delta V_g\left(1 - \frac{i_{min} + i_{FB}}{2i_0}\right)} \qquad (4)$$

where

$$A = \left(\frac{2qn_i^2\varepsilon_0\varepsilon_S}{N_d C_{FB}^2}\right)^{1/2}$$

q - electron charge, ε_0 - absolute permittivity, ε_S - relative permittivity of the semiconductor, n_i - concentration of intrinsic charge carriers in the semiconductor.

Thus, registering DN I-V in a fixed K and finding specific points i_{min}, i_{FB}, interval ΔV_g, with known dielectric thickness and dopant concentration N_d, allows to define effective generation time of minority carriers. The value C_{FB} is obtained from nomograms of N_d and d_{ox} dependences and listed in different sources, such as [1], or calculated separately.

Selection of LIV speed depends on τ_{g0} and it is implemented on the demands of creating the conditions for a sufficiently substantial depletion of the semiconductor surface carriers. At low values of K there are quasi-static current-voltage characteristics, while large K gives a full description of nonequilibrium depletion of semiconductor surface.

When $\tau_{g0} > 10^{-6}$ s the generators of LIV use $K = 10^{-1}$ - 10 V/s and DN I-V is measured by the electrometer and recorders, Figure 3, while $\tau_{g0} < 10^{-6}$ s the oscillographic method is used for detecting characteristics.

Note that comparing with high-frequency capacitance-voltage characteristics (HF C-V) DN I-V have certain advantages, including possibility of direct calculation of the surface potential, which is important in non-uniform doping profiles in semiconductor MIS structure, and increased sensitivity to the processes of charge generation. Also, DN I-V normalization at the level of dielectric charging current i_0 is:

$$\frac{I(V)}{I_0} = \frac{C(V)}{C_0} + \frac{I_g(V)}{K\left[C_0 + C_d(V)\right]} \qquad (5)$$

where I(V /I_0 - normalized DN I-V, C/C(V) - normalized HF C-V, I_g(V) - charge generation current, C_0 - dielectric capacitance, C_d(V) - SCR capacitance of the semiconductor. In Figure 4 is the comparison of DN I-V and HF C-V in the presence of electrically active defects with high speed of charge generation in MIS structures. It can be seen, when you turn on the generation in defect the modulation of DN I-V is significantly higher than for HF C-V.

2. Equipment and samples for research

DN I-V measurement is conducted at the equipment, which block diagram is shown in Figure 4.

LIV is produced by voltage generator of equipment MDC for measurement HF C-V.

Capacitive currents of DN I-V are registered by B7-30 electrometer and displayed on a graphics device H-307. Contact to the MIS structures under research is performed on analytical probe station MP1100.

Samples for research are MIS transistors in test structures ST45T of semiconductor manufacturing technology created by MIS p-channel integrated circuits on n-type silicon. Dopant concentration $N_d = 1 \cdot 10^{15}$ cm $^{-3}$. Investigated MIS capacitors are formed in accordance with the technological regimes during under-gate oxide growing for MIS transistors. The thickness of the under-gate oxide is 1200-1400 Å.

3. Work plan.

1. Switch on the equipment for DN I-V measurements according to the instructions on their use.

2. Put the plate with test structures on a table ST45T sample analytical probe station MP 1100

3. Connect analytical probe station with the equipment for measuring DN I-V.

4. Set the speed of LIV K = 1 V/s. Measure DN I-V. If the speed is not sufficient to create nonequilibrium conditions for generation of minority carriers in the MIS structure, it is necessary to increase the speed to 2.5-5 V/s. Conversely, if the speed is great and creates conditions close to full depletion of the semiconductor surface by nonequilibrium charge carriers in the MIS structure, it must be reduced to 0.1-0.5 V/s. In general, this is the optimal speed LIV, which provides a surface potential level of 5-15 V at the applied voltage of 10-30 V.

5. Process DN I-V graphics. On the basis of known data of capacity "flat areas" C_{FB} dependence on the thickness of the dielectric and the level of dopant in silicon [3], find their values and further, determine the appropriate value of C_{FB}, or i_{FB} on the DN I-V of MIS structure. Mark the level i_{min} on DN I-V and determine the voltage interval ΔV_g, corresponding to the achievement of the current generation value i_{FB}. Calculate the value τ_{g0} by formula (4).

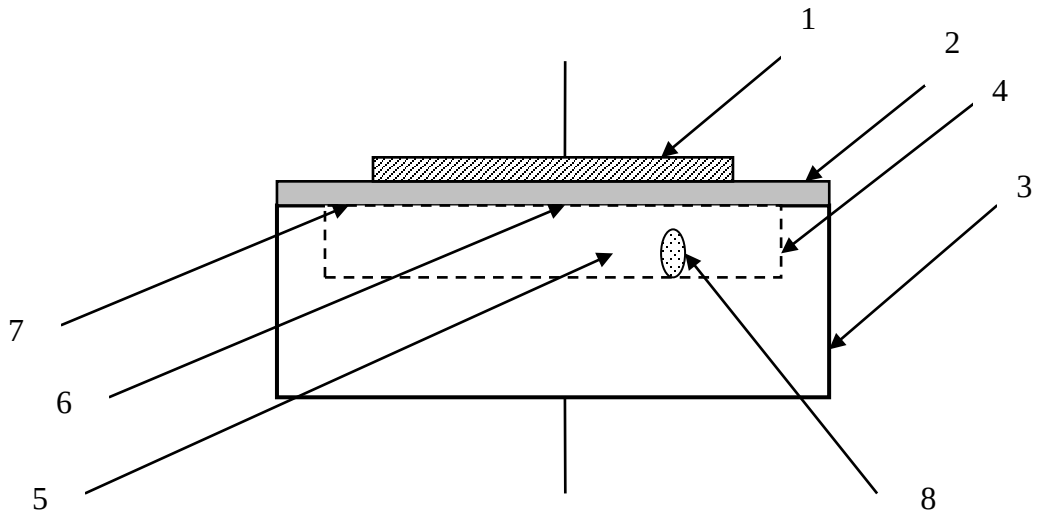


Fig. 1. Mechanisms of minority carriers generation in MIS structure:
1 – metal, 2 - insulator, 3 - semiconductor, 4 – SCR of semiconductor, 5 - generation in the SCR, 6 - surface generation under the electrode, 7 - generation at the periphery of the electrode, 8 - defect of abnormally high generation speed

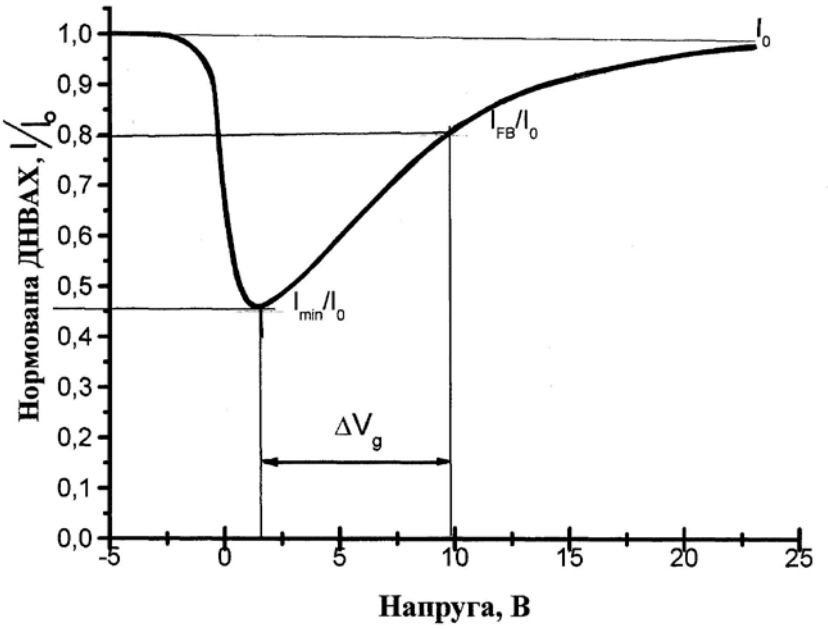


Fig. 2. Normalized DN I-V of MIS structure

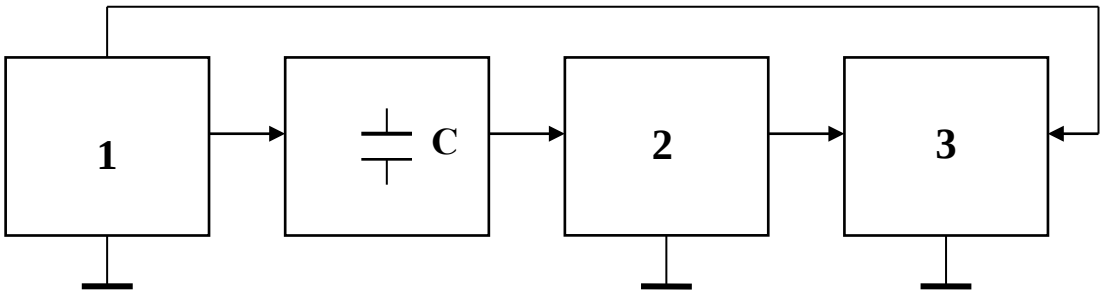


Fig. 3. Block diagram of equipment for measurement DN I-V of MIS structures: 1 - linearly increasing voltage generator, 2 - electrometer, 3 - the graphics device, C - MIS structure.

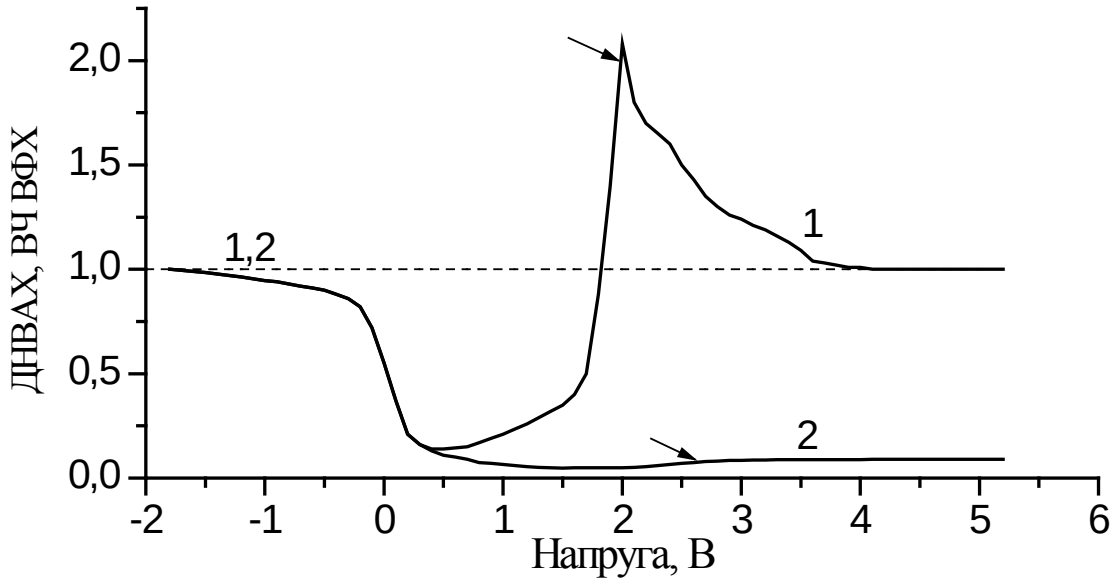


Fig. 4. Comparison of DN I-V (1) and HF C-V (2) of MIS structure with electrically active defects with high speed charge generation (process of generation is indicated by the arrow)

4. Control questions

1. Physical meaning of the effective lasing lifetime of minority carriers in a semiconductor
2. Basic mechanisms of the minority carriers generation in MIS structures
3. DN I-V method
4. Advantages of DN I-V comparing with HF C-V of MIS structures
5. How does the growth rate of inverse linearly increasing voltage influence on the DN I-V form?

5. Literature

1. V. Zakharov, V.M. Popov. Definition of bulk generation lifetime by dynamic nonequilibrium current-voltage characteristics of MIS structures.Microelectronics, v.5, 1976, № 2, p. 164-169.
2. Popov V.M. Research of profiles of bulk generation lifetime distribution for minority carriers in the MIS structures. Optoelectronics and semiconductor technics. 1991, Vol. 20, pp.12-15.
3. S. Zi. Physics of semiconductor devices, v.1, "Mir", Moscow, 1984

References to Chapter 1.

1. MICROELECTRONICS TO NANOELECTRONICS, Edited by ANUPAMA B. KAUL, CRC Press Taylor & Francis Group, Boca Raton, 2013

2.J.M. Martínez-Duart, R.J. Martín-Palma, F. Agulló-Rueda, NANOTECHNOLOGY FOR MICROELECTRONICS AND OPTOELECTRONICS, Elsevier, AMSTERDAM, 2006

3. Regina Luttge (Auth.) Microfabrication for Industrial Applications Elsevier Inc. 2011

4. Franssila, Sami. Introduction to microfabrication / Sami Franssila. – 2nd ed./ 2010, John Wiley & Sons, Ltd

5. N. P. Mahalik, Micromanufacturing and Nanotechnology, Springer, Berlin, 2006

6. O.Geschke, H.Klank, P.Telleman, Microsystem Enginnering of Lab on a chip Devices, Wiley-VCH Verlag, Weinheim, 2004.

7. Microfabrication and Nanomanufacturing, ed Mark J. Jackson, CRC Press, Taylor & Francis Group, Boca Raton, 2006.

8. VLSI MICRO- and NANOPHOTONICS, Science, Technology, and Applications, Eds. El-Hang Lee, Louay A. Eldada, Manijeh Razeghi • Chennupati Jagadish, CRC Press, Taylor & Francis Group, Boca Raton, 2011.

9. VLSI TECHNOLOGY, ed Wai-Kai Chen, CRC Press, Boca Raton, 2003.

References to Chapter 2.

1. S. M. Sze, Semiconductor devices: Physics and technology, John Wiley & Sons, Inc., 2002.

2. G.R. May, S. M. Sze, Fundamentals of semiconductor fabrication, John Wiley & Sons, Inc., 2004.

3. C. W. Pearce, "Crystal Growth and Wafer Preparation" and "Epitaxy," in S. M. Sze, Ed., VLSI Technology, McGraw-Hill, New York, 1983.

4. T. Abe, "Silicon Crystals for Giga-Bit Scale Integration," in T. S. Moss, Ed., Handbook on Semiconductors, Vol. 3, Elsevier Science B. V., Amsterdam/New York, 1994.

5. W. R. Runyan, Silicon Semiconductor Technology, McGraw-Hill, New York, 1965.

6. W. G. Pfann, Zone Melting, 2nd Ed., Wiey, New York, 1966

7. H.M. Liaw, Crystal growth of silicon, in *Handbook of semiconductor silicon technology,* eds. W.C. O'Mara, R.B. Herring, L.P. Hunt, P.94, Noyes Publ., New Jersey, USA (2009).

8. Liaw, H.M., U.S. Patent 4,394,532; assigned to Motorola Inc. (1983).

9. Lane, R.L. and Kachare, A.H., Multiple Czochralski Growth of Silicon Crystals from a Single Crucible, J. Crystal Growth 50:437-444 (1980).

10. Fiegl, G., Recent Advances and Future Directions in Cz-Silicon Crystal Growth Technology, Solid State Technology 26(8); 121 - 13 1 (1983).

11. Hopkins, R.H., Seidensticker, R.G., Davis, J.R., Rai-Choudhury, P., Blais, P.D. and McCormick, J.R., Crystal Growth Considerations in the Use of Solar Grade Silicon, J. Crystal Growth 42:493-498 (1977).

12. Bonora, A.C., Silicon Crystal Growth and Processing Technology: A Review, in: Silicon Processing, ASTM STP 804 (D.C. Gupta, ed.) pp. 5-15, Am. SOC. for Testings and Materials (1983).

13. Allgaier, R.S., Interpretation of Transport Measurements in Electrically Conducting Liquids, Phys. Rev. 185227-244 (1969).

14. Ohwa, M., Higuchi, T., Toji, E., Watanabe, M., Homma, K. and Takasu, S., Growth of Large Diameter Silicon Single Crystal under Horizontal or Vertical Magnetic Field, in: *Semiconductor Silicon* (H.R. Huff, T. Abe, and B. Kolbesen, eds.) pp. 117-128, The Electrochem. Society, Pennington, NJ (1986).

15. Barroclough, K.C., Series, R.W., Bae, G.J., and Kemp, D.S., Axial Magnetic Czochralski Silicon Growth, in: *Semiconductor Silicon* (H.R. Huff, T. Abe, and B. Kolbesen, eds.) pp. 129-141, The Electrochem. Society, Pennington, NJ (1986).

16. Moody, J.W., Oxygen in Czochralski Crystals and Melts-A Review, in: *Semiconductor Silicon*, (H.R. Huff, T. Abe, and B. Kolbesen, eds.) pp. 100-116, The Electrochem. Society. Pennington, N.J. (1986).

17. Suzuki, T., Isawa, N., Hoshi, K., Kato, Y. and Okubo, Y., MCz Silicon Crystals Grown at High Pulling Rates, in: *Semiconductor Silicon* (H.R. Huff, T. Abe, and B. Kolbesen, eds.) pp. 142- 152, The Electrochem. Society. Pennington, NJ (1986).

18. Kuroda, E., Matsubara, S., Saitoh, T., Czochralski Growth of Square Silicon Single Crystals, Jap. J. Appl. Phys. 19:L361 -L364 (1980).

19. Liaw, H.M., Growth of Single Crystal Silicon Square Ingots, The Eelectrochem. Soc. Meeting, Extended Abstract, 80- 12306-807 (1980).

20. Bennett, A.I. and Logini, R.L., Dendritic Growth of Germanium Crystals, Phy. Rev. 11653-61 (1959).

21. LaBelle, H.E., Jr. and Mlavsky, A.I., Growth of Controlled Profile Crystals from the Melts: part I- Sapphire Filaments, Mater. Res. Bull. 6:571-580 (1971).

22. LaBelle, H.E., Jr., Growth of Controlled Profile Crystals from the Melts: part II-Edge-Defined, Film-Fed Growth (EFG), Mater. Res. Bull. 6:581-590 (1971).

23. Ciszek, T.F., Edge-Defined, Film-feed Growth (EFG) of Silicon Ribbon, Mat. Res. Bull. 7:731-737 (1972).

24. Seidensticker, R.G., Dendritic Web Silicon for Solar Cell Application, J. Crystal Growth 39:17-22 (1977).

25. Faust, J.W., Jr. and John, H.F., Germanium Dendrite Studies, I. Studies of Twin Structures and the Seeding Mechanism, J. EIectrochem. Soc. 108:855-860 (1961).
26. O'Hara, S. and Bennett, A.I., Web Growth of Semiconductors, J. Appl. Phys. 35:686-693 (1964).
27. Hamilton, D.R. and Seidensticker, R.G., Propagation Mechanism of Germanium Dendrites, J. Appl. Phys. 31:1165- 1168 (1960).
28. Tucker, T.N., and Schwuttke, G.H., Growth of Dislocation- Free Silicon Web Crystals, Appl. Phys. Lett. 9:219-221 (1966).
29. Ravi, K.V., The Growth of EFG Silicon Ribbons, J. Crystal Growth 39: 1 - 16 (1977).
30. E. W. Hass and M. S. Schnoller, "Phosphorus Doping of Silicon by Means of Neutron Irradiation," IEEE Trans. Electron Devices, **ED-23**, 803 (1976).
31. M. Hansen, Constitution of Binary Alloys, McGraw-Hill, New York, 1958.
32. S. K. Ghandhi, VLSl Fabrlcation Principles, Wiley, New York, 1983.
33. J. R. Arthur, "Vapor Pressures and Phase Equilibria in the GaAs System," J. Phys. Chem. Solids, **28**, 2257 (1967).
34. B. El-Kareh, Fundamentals of Semiconductor Processing Technology, Kluwer Academic, Boston, 1995.
35. C. A. Wert and R. M. Thomson, Physics of Solids, McGraw-Hill, New York, 1964.
36. (*a*) F. A. Trumbore, "Solid Solubilities of Impurity Elements in Germanium and Silicon," Bell Syst. Tech. J., **39**, 205 (1960); (*b*) R. Hull, Propetties of Cystaline Silicon, INSPEC, London, 1999.
37. Y. Matsushita, "Trend of Silicon Substrate Technologies for 0.25 μm Devices," Proc. VLSl Technol. - Workshop, Honolulu, (1996).
38. The International Technology Roadmap for Semiconductors, Semiconductor Industry Association, San Jose, CA, 1999.

References to Chapter 3.

1. A.A. Evtukh, Epitaxial Structures, in *Inorganic Materials. Materials and Technologies,* G.G. Gnyesin, V.V. Skorokhod, Eds., 2008, V.2, P.823-831.
2. R.B. Herring. Silicon Epitaxy, in Handbook of Semiconductor silicon technology.Eds. W.C. O'Mara, R.B. Herring, L.P. Hunt, Noyes publ., New Jersey, USA, 2009.
3. P.K. Vasudev, Silicon-on-Sapphire Heteroepitaxy. Chapter 4 in Epitaxial Silicon Technology, (B.J. Baliga, ed.), Academic PressJnc., New York (1986).
4. S.M. Sze. Semiconductor devices. Physics and technology. 2nd Ed. John Wiley & Sons, Inc., 2002.

5. A. S. Grove, Physics and Technology of Semiconductor Devices, Wiley, New York, 1967.

6. R. Reif, T. I . Kamins, and K. C. Saraswat, "A Model for Dopant Incorporation into Growing Silicon Epitaxial Films," J. Electrochem. Soc., 126,644,653 (1979).

7. H.H. Lee, Silicon Growth at Low Temperatures: SiH_4-HCl-H_2 System, J. Cryst. Growth 69 (1984) 82-90.

8. J. Bloem, and L.J. Giling, Epitaxial Growth of Silicon by Chemical Vapor Deposition, Ch. 3 in VLSI Electronics Microstructure Science, V. 12 Silicon Materials (N.G. Einspruch, ed.), pp. 89-139, Academic Press, New York (1985)

9. J. Bloem, and W.A.P. Classen, Rate-determining Reactions and Surface Species in CVD Silicon. J. Cryst. Growth 49 (1980) 435-444.

10. D. Foster, A. Learn, and T.I. Kamins, Deposition Properties of Silicon Films Formed from Silane in a Vertical Reactor. Vac. Science and Technology 4 (1986) 1182- 1186.

11. J.C. Brice, The Growth of Crystals from Liquids, North-Holland Pub., Amsterdam (1973).

12. F.C. Eversteijn, Gas-Phase Decomposition of Silane in a Horizontal Epitaxial Reactor. Philips Res. Rep. 26 (1971) 134-144.

13. F.C. Eversteijn, Chemical Reaction Engineering in the Semiconductor Industry. Philips Res. Rep. 29 (1974) 45-66.

14. J. Bloem, High Chemical Vapor Deposition Rates of Epitaxial Silicon Layers. J. Cryst. Growth 18 (1973) 70-76.

15. H.M. Liaw, D. Weston, B. Reuss, M. Birritella, and J. Rose, in Proc. 9th Int. Conf. on Chemical Vapor Deposition (McD. Robinson, et al., eds.), pp. 463-475, Electrochem. SOC., Pennington, NJ (1984).

16. W.A.P. Classen, and J. Bloem, The Nucleation and Growth of Silicon via Chemical Vapor Deposition, in Semiconductor Silicoii 1981, (H.R. Huff, R.J. Kriegler, and T. Takeishi, eds.), pp. 365-376, Electrochem. SOC., Pennington, NJ (1981)

17. V. K. Jain, and S. K. Sharma, Solid State Electron. 13 (1970) 1145.

18. J. J. Heish, Liquid phase epitaxy, in Handbook on Semiconductor, vol. 3, pp. 415-497, ed. by S. P. Keller, North-Holland (1980).

19. M. B. Panish, and A. Y. Cho, IEEE Spectrum (April 1983) 18.

20. H. M. Manasevit, J. Cryst. Growth 55 (1981) 1.

21. L. P. Chen, Ph.D. dissertation, National Cheng Kung University, Taiwan (1987).

22. E. Johnson, R. Tsui, D. Convey, N. Meller, and J. Curless, J. Cryst. Growth 69 (1984): 497.

23. P.D. Dapkus, J. Cryst. Growth 68 (1984): 345.

24. G. B. Stringfellow, in Semiconductor and Semimetals, vol. 22, p. 209, ed. by W. Tsang Academic Press (1985).

25. R. D Dupuis, "Metalorganic Chemical Vapor Deposition of III-V Semiconductors", Science, 226 (1984) 623.
26. G. B. Stringfellow, Organometallic Vapor-Phase Epitaxy, Academic Press (1989).
27. R. Bhat, P. O'Connor, H. Temkin, R. Dingle, and V. G. Keramidas, Inst. Phys. Conf. Ser. 63 (1982) 101.
28. C. P. Kuo, R. M. Cohen, and G. B. Stringfellow, J. Cryst. Growth 64 (1983) 461.
29. T. F. Kuech, and R. Potemski, Appl Phys. Lett. 411 (1985) 821.
30. C. Y. Chang, Y. K. Su, M. K. Lee, L. G. Chen, M. P. Houng, J. Cryst. Growth 55 (1981) 24.
31. R. H. Moss, and J. S. Evans, J. Cryst. Growth 55 (1981): 129.
32. M. J. Cherng, R. M. Cohen, and G. B. Stringfellow, J. Electron. Mater. 13 (1984) 799.
33. C. B. Cooper, M. J. Ludowise, V. Aebi, and R. L. Moon, J. Electron. Mat. 9 (1980) 299.
34. R. B. Clough, and J. J. Tietjen, Trans. Metall Soc. AIME 245 (1969) 583.
35. G. B. Stringfellow, J. Electron. Mat. 17, 4 (1988) 327-335.
36. CRC Handbook of Laboratory Safety, ed. by N. Y. Steere, Chemical Rubber Co., Cleveland, OH (1967).
37. N, I. Sax, in Dangerous Properties of Industrial Materials, Yan Nostrand Reinhold (1979).
38. S. P. DenBaars, B. Y. Maa, P. D. Dapkus, A. D. Danner, and H. C. Lee, J. Cryst. Growth 77 (1986) 188.
39. C. A. Larsen, and G. B. Stringfellow, J. Cryst. Growth 75 (1986) 247.
40. G. B. Stringfellow, in Semiconductors and Semimetals, vol. 22, p. 209, ed. by W. Tsang, Academic Press (1985).
41. S. J. Bass, J. Cryst. Growth 31 (1975) 172.
42. H. M. Manasevit, Appl Phys. Lett. 12 (1968) 156.
43. D. H. Reep, and S. K. Ghandhi, J. Electrochem. Soc. 130 (1983) 675.
44. J. P. Duchemin, M. Bonnet, F. Koelsch, and Huggi, D. J. Cryst. Growth 45 (1978) 181.
45. S. K. Ghandhi, VLSI Fabrication Principles, Wiley (1983).
46. S. M. Sze, Physics of Semiconductor Devices, Wiley-Interscience (1981).
47. G. B. Stringfellow, J. Cryst. Growth 68 (1984) 111.
48. M. Koppitz, O. Vestavik, W. Plestschen, A. Mircea, M. Heyen, and W. Richter, J. Cryst. Growth 68 (1984) 136.
49. Y. Kusumoto, T. Hayashi, and S. Komiya, Jap. J. Appl Phys. 24 (1985) 620.
50. J. J. Coleman, and P. D. Dapkus, in Gallium Arsenide Technology, ed. by D. K. Ferry, Howard W. Sams & Co. (1985).
51. P. D. Dapkus, H. M. Manesevit, K. L. Hess, T. S. Low, and Stillman, G. E. J. Crystal Growth 55 (1981) 10.

52. M. Oishi, and K. Kuroiwa, Jap. J. Appl Phys. 21 (1982) 203.
53. R. H. Moss, and J. S. Evans, J. Cryst. Growth 55 (1981): 129.
54. I. A. Frolor, J. Beldyrevskii, B. L. Druz, and E. B. Sokolpv, Inorg. Mater. 13 (1977) 632.
55. J. P. Duchemin, J. P. Hirtz, M. Razeghi, M. Bonnet, and S. D. Hersee, J. Cryst. Growth 55 (1981) 64.
56. J. Yoshino, T. Iwamoto, and H. Kukimoto, J. Cryst. Growth 55 (1981) 74.
57. J. Yoshino, T. Iwamoto, and H. Kukimoto, Jap. J. Appl. Phys. 20 (1981) L290.
58. J. P. Duchemin, M. Bonnet, G. Beuchet, and F. Koelsch, Gallium Arsenide and Related Compounds, Institute of Physics Conference, series no. 45, pp. 10-18 (1978).
59. M. R. Leys, and H. Veerivliet, J. Cryst. Growth 55 (1981): 145.
60. K. W. Benz, H. Renz, J. Widlein, and M. H. Pilkuhn, J. Electron. Mater. 10 (1981) 185.
61. C. B. Cooper, M. J. Ludowise, V. Aebi, and R. L. Moon, J. Electron. Mater. 9 (1980) 299.
62. D. Hung, in Chemical Vapor Deposition, ch. 6, p. 245, ed. by M. L. Hitchman and J. E. Jensen, Academic Press, (1992).
63. A. Brauers, J. Cryst. Growth, 107 (1991) 281.
64. P. D. Dapkus, H. M. Manasevit, K. L. Hess, T. S. Low, and G. E. Stillman, J. Crystal Growth 55 (1981) 10.
65. T. Nakanishi, T. Udagawa, A. Tanaka, K. Kamei, J. Crystal Growth 38 (1977) 23.
66. P. Rai-Chudhury, J. Electrochem. Soc. 116 (1969) 1745.
67. Y. Seki, K. Tanno, K. lida, and E. Ichiki, J. Electrochem. Soc. 122 (1975) 1108.
68. T. F Kuech, and R. Potemski, Appl. Phys. Lett. All (1985) 821.
69. N. Kobayashi, and T. Makimoto, Jap. J. Appl. Phys. 24 (1985) L824.
70. M. Yoshida, H. Watanabe, and F. Uesugi, J. Electrochem. Soc. 132 (1985) 677.
71. N. Putz, H. Heinecke, M. Heyen, P. Balk, M. Weyers, and H. Liith, J. Cryst. Growth 74 (1986) 292.
72. T. F. Kuech, Materials Science Reports 2, p. 3 (1987).
73. A. Forster, and H. Liith, J. Vac. Sci Tech. B, 7(4) (1988) 720-724.
74. G. Wilkinson, G. A. Stone, and E. W. Abel, ed., Comprehensive Organomatallic Chemistry, Pergamon Press (1982).
75. G. B. Stringfellow, and G. Horn, Appl Phys. Lett. 34 (1979) 794.
76. G. W. Hooft, C Van Opdorp, H. Veenvliet, and A. T. Vink, J. Cryst. Growth 55 (1981) 173.
77. D. Kisker, J. N. Miller, and G. B. Stringfellow, Appl Phys. Lett. 40 (1982) 614.
78. E. E. Wagner, G. Hom, and G. B. stringfellow, J. Electron. Mater. 10 (1981) 239.
79. P. D. Dapkus, H. M. Manasevit, K. Hess, T. S. Low, and G. E. Stillman, J. Cryst. Growth 55 (1981) 10.

80. J. Nishizawa, J. Metals 25 (1961) 149.
81. Y. Mori, and J. Watanabe, Appl Phys. 52 (1981) 2792.
82. H. Morkos, *Handbook of Nitride Semiconductors and Devices*, Vol. 1: *Materials Properties, Physics and Growth*, WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2008.
83. H.M. Manasevit, F.M. Erdman, and W.I. Simpson, (1971) Journal of the Electrochemical Society, 118, 1864.
84. H. Amano, N. Sawaki, I. Akasaki, and Y. Toyoda, (1986) Applied Physics Letters, 48, 353.
85. C. Yuan, T. Salagaj, A. Gurary, P. Zavadski, C.S. Chen, W. Kroll, R.A. Stall, Y. Li, M. Schurman, C.Y. Hwang, W.E. Mayo, Y. Lu, S.J. Pearton, S. Krishnankutty, and R.M. Kolbas, (1995) Journal of the Electrochemical Society, 142, L163.
86. S. Nakamura, T. Mukai, and M.Senoh, (1994) Journal of Applied Physics, 76, 8189.
87. S. Nakamura, M. Senoh, S. Magahama, N. Iwasha, T. Yamada, T. Matsushita, H. Kioku, and Y. Sugimoto, (1996) Japanese Journal of Applied Physics, L74, 1998.
88. M.A. Khan, J.N. Kuznia, A.R. Bhattarai, and D.T. Olson, (1993) Applied Physics Letters, 62, 1786.
89. F. Binet, J.Y. Duboz, E. Rosencer, F. Scholz, and V. Harle, (1996) Applied Physics Letters, 69, 1002.
90. S. Nakamura, Y. Harada, and M. Seno, (1991) Applied Physics Letters, 58, 2021.
91. J.L. Dupuie, and E. Gulari, (1991) Applied Physics Letters, 59, 549.
92. D.K. Wickenden, J.A. Miragliotta, W.A. Bryden, and T.J. Kistnmacher, (1994) Journal of Applied Physiology, 75, 7585.
93. J.C. Knights, and R.A. Lujan (1978) Journal of Applied Physiology, 49, 1291.
94. A.E. Wickenden, D.K. Wickenden, and T.J. Kistenmacher, (1994) Journal of Applied Physiology, 75, 5367.
95. M. Matloubian, and M. Gershenzon, (1985) Journal of Electronic Materials, 40, 633–644.
96. A. Wakahara, and A. Yoshida, (1989) Applied Physics Letters, 54, 709.
97. T.Y. Sheng, Z.Q. Yu, and G.J. Collins, (1988) Applied Physics Letters, 52, 576.
98. D.K. Gaskill, N. Bottka, and M.C. Lin, (1986) Applied Physics Letters, 48, 1449.
99. M. Mizuta, S. Fujieda, T. Jitsukawa, and Y. Matsumoto, (1987) Proceedings of the International Symposium on GaAs and Related Compounds, 1986, Las Vegas, NV, Adam Hilger, Bristol.
100. S. Fujieda, M. Mizuta, and Y. Matsumoto, (1987) Japanese Journal of Applied Physics, 26, 2067.
101. S. Miyoshi, K. Onabe, N. Ohkouchi, H. Yaguchi, and R. Ito, (1992) Journal of Crystal Growth, 124, 439.

102. M. Ishii, T. Minami, T. Miyata, S. Karaki, and S. Takata, (1996) Institute of Physics Conference Series, 142, 899.
103. H.M. Manasevit, F.M. Erdmann, and W.I. Simpson, (1971) Journal of the Electrochemical Society, 118, 1864.
104. F.A. Pizzarello, and J.E. Coker, (1975) Journal of Electronic Materials, 4, 25.
105. G.D.O.Clock, and M.T. Duffy, (1973) Applied Physics Letters, 23, 55.
106. J.K. Liu, K.M. Lakin, and K.L. Wang, (1975) Journal of Applied Physiology, 46, 3703.
107. M. Morita, N. Uesugi, S. Isogai, K. Tsubouchi, and N. Mikoshiba, (1981) Japanese Journal ofApplied Physics, 20, 17.
108. M.A. Khan, R.A. Skogman, R.G. Schulze, and M. Gershenzon, (1983) Applied Physics Letters, 42, 430.
109. M. Hashimoto, H. Amano, N. Sawaki, and I. Akasaki, (1984) Journal of Crystal Growth, 68, 163.
110. M. Matloubian, and M. Gershenzon, (1985) Journal of Electronic Materials, 14, 633.
111. T. Sasaki, and T. Matsuoka, (1988) Journal of Applied Physiology, 64, 4531.
112. T. Nagatomo, T. Kuboyama, H. Minamino, and O. Omoto, (1989) Japanese Journal of Applied Physics, 28, L1334.
113. M.A. Khan, J.N. Kuznia, J.M. Van Hove, D.T. Olsen, S. Krishnankutty, and R.M. Kolbas, (1991) Applied Physics Letters, 58, 526.
114. S. Nakamura, Y. Harada, and M. Seno, (1991) Applied Physics Letters, 58, 2021.
115. D.K. Wickenden, J.A. Miragliotta, W.A. Bryden, and T.J. Kistnmachr, (1994) Journal of Applied Physiology, 75, 7585.
116. J.C. Knights, and R.A. Lujan, (1978) Journal of Applied Physiology, 49, 1291; A.E. Wickenden, D.K. Wickenden, and T.J. Kistenmacher, (1994) Journal of Applied Physiology, 75, 5367.
117. M. Shiloh, and J. Gutman, (1973) Materials Research Bulletin, 8, 711.
118. R. Lappa, G. Glowacki, And Galkowski, S. (1976) Thin Solid Films, 32, 73.
119. T.Y.Sheng, Z.Q. Yu, and G.J. Collins, (1988) Applied Physics Letters, 52, 576.
120. A. Wakahara, and A. Yoshida, (1989) Applied Physics Letters, 54, 709.
121. E.N. Eremin, L.I. Nekrasov, E.A. Rubtsova, V.M. Belova, V.L. Ivanter, L.N. Zakharov, and L.N. Petukhov, (1982) Russian Journal of Physical Chemistry, 56, 788.
122. M. Matloubian, and M. Gershenzon, (1985) Journal of Electronic Materials, 40, 633–644.
123. J.L. Dupuie, and E. Gulari, (1991)Applied Physics Letters, 59, 549.
124. H.H. Liu, D.C. Bertolet, and J.W. Rogers, (1994) Surface Science, 320, 145.
125. A.C. Jones, C.R. Whitehouse, and J.S. Roberts, (1995) Chemical Vapor Deposition, 1, 65.

126. A.C. Jones, J. Auld, S.A. Rushworth, D.J. Houlton, and G.W. Critchlow, (1994) Journal of Materials Chemistry, 4, 1591,

127. A.C. Jones, J. Auld, S.A. Rushworth, E.W. Williams, P.W. Haycock, C.C. Tang, and G.W. Critchlow, (1994) Advanced Materials, 6, 6.

128. F. Takeda, T. Mori, and T. Takahashi, (1981) Japanese Journal of Applied Physiology, 20, L169.

129. H.C. Lee, K.Y. Lee, Y.J. Yong, J.Y. Lee, and H. Kim, (1995) Thin Solid Films, 271, 50.

130. M. Yoshida, H. Watanabe, and F. Uesugi, (1985) Journal of the Electrochemical Society, 132, 677.

131. O. Ambacher, (1998) Journal of Physics D: Applied Physics, 31, 2653.

132. A. Y. Cho, and J. R. Arthur, Progress in Solid-State Chemistry, ed. by G. Somorjai and J, McCaldin, vol. 10, p. 157, Pergamon (1975).

133. A. Y. Cho, J. Vac. Sci. Tech. 16 (1979) 275.

134. K. Ploog, Crystal Growth, Properties, and Applications, ed. by H. C. Freyhardt, vol. 3, p. 73, Springer-Verlag (1980).

135. K. Ploog, Am. Rev. Mater. Sci. 11 (1981) 171.

136. L. L. Chang, and R. Ludeke, Epitaxial Growth, p. 37, ed. by J. W. Mathews, Academic Press (1975).

137. C. T. Foxon, and B. A. Joyce, Current Topics in Materials Science, vol. 7, ed. By E. Kaldis, North-Holland (1980).

138. R. F. C. Farrow, Crystal Growth and Materials, vol. 1, p. 237, ed. by E. Kaldis and H. J. Schul, North-Holland (1977).

139. J. C. Beam, Growth of Doped Silicon by Molecular Beam Epitaxy, p. 177, ed. by F. F. Y. Wang, North-Holland (1981).

140. M. A. Herman and H. Sitter, Molecular beam Epitaxy, Springer-Verlag, Berlin, 1996.

141. W. G. Herrenden-Harker, and R. H. Williams, Epitaxial Growth ofGaAs: MBE and MOCVD, p. 57, ed. by H. Thomas, D. V. Morgan, B. Thomas, I E. Aubrey, and G. B. Morgan, IEE UWIST (1985).

142. A. Y. Cho, Surface Sci. 17 (1969) 494.

143. A. Y. Cho, J. Appl. Phys. 42 (1971) 2074.

144. J. R. Arthur, Surface Sci. 43 (1974) 449.

145. A. Roth, Vacuum Technology, North-Holland, Amsterdam, 1976.

146. A. Y. Cho, M. B. Panish, and I. Hayashi, Proc. 3rd Int. Symp. on GaAs, p. 18, London (1970).

147. A. Y. Cho, J. Appl. Phys. 41 (1970) 2780.

148. A. Y. Cho, H. C. Casey, Jr., and P. W. Foy, Appl. Phys. Lett. 3§ (1977) 397.
149. A. Y. Cho, J. Vac. Sci. Tech. 16 (1979) 275.
150. A. Y. Cho, Appl Phys. Lett. 19 (1971) 467.
151. H. C. Casey, A. Y. Cho, and P. A. Barnes, IEEE J. Quantum Elect. QE-11 (1975) 467.
152. W. T. Tsang, Appl Phys. Lett. 34 (1979) 473.
153. W. T. Tsang, C. Weisbuch, and R. C. Miller, Appl Phys. Lett. 35 (1979) 673.
154. W. T. Tsang, and R. A. Logan, IEEE Quantum Elect. QE-15 (1979) 451.
155. D. M. Collins, 1982 MBE Workshop, Urbana, IL, Oct. 21-22 (1982).
156. S. Yamakoshi, O. Wada, T. Fujii, S. Hiyamizu, and T. Sakurai, IEEE IEDM, San Francisco (1982).
157. H. Morkoc, T. J. Drummond, and M. Omori, IEEE Trans. Elect. Dev. ED-29 (1982) 222.
158. M. Feng, Y. K. Eu, I. J. D'Haenens, and M. Braunstein, Appl Phys. Lett. 41 (1982) 633.
159. P. O'Connor, T. P. Pearsall, K. Y. Cheng, A. Y. Cho, J. C. M. Hwang, and K. Alavi, IEEE Elect. Dev. Lett. EDL-3 (1982) 64.
160. K. H. G. Duh, P. C. Chao, P. M. Smith, L. F. Lester, B. R. Lee, and J. C. M. Hwang, 44th Annual Device Research Conf., June 23-25, Amherst, MA (1986).
161. J. C. M. Hwang, D. G. Flahive, and S. H. Wemple, IEEE Elect. Dev. Lett. EDL-3 (1982) 320.
162. C. Y. Chen, A. Y. Cho, K. Y. Cheng, T. P. Pearsall, and P. O'Connor, IEEE Elect. Dev. Lett. EDL-3 (1982) 152.
163. A. Y. Cho, and K. Y. Cheng, Appl Phys. Lett. 38 (1981) 360.
164. K. Y. Cheng, A. Y. Cho, and W. R. Wagner, Appl Phys. Lett. 39 (1981) 607.
165. L. P. Erickson, G. L. Carpenter, D. D. Siebel, P. W. Palmberg, P. Pearah, W. Kopp. and H. Morkoc, J. Vac. Sci. Tech. B 3 (1985) 536-537.
166. A. Y. Cho, in Molecular Beam Epitaxy and Heterostructures, p. 191, ed. by L. L. Chang and K. Ploog, Kluwer, Academic Press (1985).
167. L. L. Chang, A. Segmulle, and L. Esaki, Appl Phys. Lett. 28 (1976) 39.
168. J. H. Neave, P. Blood, and B. A. Joyce, Appl Phys. Lett. 36 (1980) 311.
169. J. R. Arthur, J. Appl Phys. 39 (1968) 4032.
170. C. T. Foxon, J. W. Boudry, and B. A. Joyce, Surf. Sci. 44 (1974) 69.
171. B. A. Joyce, Kinetic and surface aspects of MBE, in Molecular Beam Epitaxy and Heterostructures^ ed. by L. L. Chang and K. Ploog, Martinus Nijhoff (1985).
172. J. R. Arthur, Structure and Chemistry of Solid Surface, ed. by G. A. Somoijai 46-1 Wiley (1969).
173. C. T. Foxon, and B. A. Joyce, Surf. Sci. 64 (1977) 293.

174. H. Kunzel, J. Knecht, H. Jung, K. Wiinstel, and K. Ploog, Appl Phys. A28. p 167 (1982).
175. K. Y. Cheng, A. Y. Cho, W. R. Wagner, and W. A. Bonner, J. Appl Phys. 52 (1981) 1015.
176. A. Y. Cho, J. Appl Phys. 42 (1971) 2074.
177. A. Y. Cho, J. Appl Phys. 41 (1970) 2780.
178. J. R. Arthur, Surf Sci., 43 (1974) 449.
179. J. H. Neave, and B. A. Joyce, J. Cryst. Growth 44 (1978) 387.
180. A. Y. Cho, J. Appl Phys. 47 (1976) 2841.
181. K. Ploog, and A. Fischer, Appl Phys. 13 (1977) 111.
182. H. C. Casey, A. Y. Cho, and P. A. Barnes, IEEE J. Quantum Elect. QE-11 (1975) 467.
183. R. Fischer, J. Klem, T. J. Drummond, R. E. Thorn, W. Kopp, H. Morkoc, and A. Y. Cho, J. Appl Phys. 54, (1983) 2508.
184. V.A. Shchukin, N.N. Ledentsov, and D. Bimberg, (2004) Epitaxy of Nanostructures, Springer, Berlin.
185. N.N. Ledentsov, (1999) Growth Processes and Surface Phase Equilibria in Molecular Beam Epitaxy, Springer, Berlin.
186. M.A. Herman, and H. Sitter, (1989) Molecular Beam Epitaxy: Fundamental and Current Status, Springer, Berlin.
187. D.D. Koleske, A.E. Wiekenden, R.L. Henry, W.J. DeSisto, and R.J. Gorman, (1998) Journal of Applied Physics, 84 (4), 1998–2010.
188. S. Yoshida, S. Misawa, and A. Itoh, (1975) Applied Physics Letters, 26, 461.
189. S. Winsztal, B. Wauk, H. Majewska-Minor, and T. Niemyski, (1976) Thin Solid Films, 32, 251.
190. K.R. Elliott, and R.W. Grant, (1984) Rockwell Project Final Report MRDC41116.2FR.
191. H.U. Baier, and W. Monch, (1990) Journal of Applied Physiology, 68, 586.
192. S. Yoshida, S. Misawa, and S. Gonda, (1983) Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures, 1, 250.
193. K. Oura, V.G. Lifshits, A.A. Saranin, A.V. Zotov, and M. Katayama, (2003) Surface Science, Springer, Berlin.
194. C. Adelmann, J. Brault, G. Mula, B. Daudin, L. Lymperakis, and J. Neugebauer, (2003) Physical Review B: Condensed Matter, 67, 165419.
195. S. Guha, N.A. Bojarczuk, and D.W. Kisker, (1996) Applied Physics Letters, 69, 2879.
196. O. Brandt, Y.J. Sun, L. Daweritz, and K.H. Ploog, (2004) Physical Review B: Condensed Matter, 69, 165326.
197. Q. Xue, O.K. Xue, R.Z. Bakhtizin, Y. Hasegawa, I.S.T. Tsong, T. Sakurai, and T. Ohno, (1999) Physical Review B: Condensed Matter, 59, 12604.

198. G. Koblmuller, R. Averbeck, H. Tiechert, and P. Pongratz, (2004) Physical Review B: Condensed Matter, 69, 035325.
199. C. Adelmann, J. Brault, D. Jalabert, P. Gentile, H. Mariette, G. Mula, and B. Daudin, (2002) Journal of Applied Physics, 91, 9638.
200. L.X. Zheng, M.H. Xie, and S.Y. Tong, (2000) Physical Review B: Condensed Matter, 61, 4890.
201. T. Zywietz, J. Neugebauer, and M. Scheffler, (1998) Applied Physics Letters, 73, 487.
202. J.M. Myoung, O. Gluschenkov, K. Kim, and S. Kim, (1999) Journal of Vacuum Science & Technology A: Vacuum Surfaces and Films, 17, 3019.
203. J. Neugebauer, T.K. Zywietz, M. Scheffler, J.E. Northrup, H. Chen, and R.M. Feenstra, (2003) Review Letters, 90, 056101.
204. J.M. Myoung, O. Gluschenkov, K. Kim, and S. Kim, (1999) Journal of Vacuum Science & Technology A: Vacuum Surfaces and Films, 17, 3019.
205. E. Iliopoulos, and T.D. Moustakas, (2002) Applied Physics Letters, 81, 295.
206. W.A. Harrison, (1980) Electronic Structure and the Properties of Solid, Dover, New York.
207. J.C. Phillips, (1973) Bonds and Bands in Semiconductors, Academic Press, New York.
208. Z. Bachrach, and B. S. Krusor, J. Vac. Sci. Tech. 18 (1981) 754.
209. M. Bafleur, A. Munoz-Yague, and A. Rocher, J. Cryst. Growth 59 (1982) 531.
210. M. Ohring, The Materials Science of Thin Film, Academic, New York, 1992.
211. J. C. Bean, "The drowth of Novel Silicon Materials," Physics Today, 39, 10, 36 (1986).
212. Y. G. Chai, and R. Chow, Appl Phys. Lett. 38 (1981) 796.
213. C. E. C. Wood, L. Rathbaum, H. Ohrio, D. Desimone, J. Cryst. Growth 51 (1981) 299.
214. M. Bafleur, A. Munoz-Yague, and A. Rocher, J. Cryst. Growth 59 (1982) 531.
215. H. Morkoc, R. Stamberg, and E. Krikorinar, Jap. J. Appl. Phys. 21, (1982) L234.
216. Y. Suzuki, M. Seiki, Y. Horikoshi, and H. Okamoto, Jap. J. Appl Phys. 23 (1984) 1641.
217. Y. H. Wang, W. C. Liu, S. A. Liao, K. Y. Cheng, and C. Y. Chang, Jap. J. Appl. Phys. 24 (1985) 514.
218. M. Bafleur, and A. Munoz-Yague, Thin Solid Films 101 (1983) 299.
219. T. Ito, M. Shinohara, and Y. Imamura, Jap. J. Appl Phys. 23 (1984) L524.
220. R. Dingle, M. D. FeUer, and C. W. Tu, in VLSI Electronics, vol. 11, ed. by N. G. Einspruch and W. Wisseman, Academic Press (1985).
221. Y. H. Wang, W. C. Liu, C. Y. Chang, and S. A. Laio, J. Vac. Sci. Tech. B 4, 1 (1986) 30.
222. G. D. Petit, J. A. Woodall, S. L. Wright, P. D. Kirchner and J. L. Freeout, J. Vac. Sci. Tech. B 2 (1984) 241.

223. P. D. Kirchner, J. M. Woodall, J. F. Freeouf, and G. D. Pettit, Appl Phys. Lett. 38 (1981) 427.
224. G. M. Metze, A. R. Calawa, and J. G. Mavroides, J. Vac. Sci. Tech. B1 (1983) 166.
225. J. K. Abrokwah, N. C. Cirillo, Jr., M. J. Helix, and M. Longerbone, J. Vac. Sci. Tech. B 2 (1984) 252.

References to Chapter 4

1. S.M. Sze, *Semiconductor devices: Physics and technology*, John Wiley & Sons, Inc., New York, 2002.
2. G.S. May, S.M. Sze, *Fundamentals of semiconductor fabrication*, John Wiley & Sons, Inc., New York, 2004.
3. A.C. Adams, "Dielectric and Polysilicon Film Deposition" in S.M. Sze, Ed., *VLSI Technology*, Mc-Graw-Hill, New York, 1983.
4. K. Eujino, et.al., Doped silicon oxide deposition by atmospheric pressure and low temperature chemical vapor deposition using tetraethoxysilane and ozone, J. Electrochem. Soc. **138**, 3019 (1991).
5. A.C. Adams, C.D. Capio, Planarization of phosphorus-doped silicon dioxide, J. Electrochem. Soc. **127**, 2222 (1980).
6. T. Yamamoto et.al., An advanced 2.5 nm oxidized nitride gate dielectric for highly reliable 0.25 μm MOSFETs, Symp. on VLSI Technol. Dig. of Tech. Pap., 1997, p.45.
7. K. Kumar, et.al., Optimization of some 3 nm gate dielectricgrown by rapid thermal oxidation in nitric oxide ambient, Appl. Phys. Lett., **70**, 384 (1997).
8. T. Homma, Low dielectric constant materials and method for interlayer dielectric films in ultralarge-scale integrated circuit multilevel interconnection, Mat. Sci. Eng., **23**, 243 (1998).
9. H.N. Yu, et.al., 1 m MOSFET VLSI technology. Part I – An overview, IEEE Trans.cElectron. Dev., ED-26, 318 (1979).
10. T.I. Kamins, Preparation and properties of polycrystalline silicon films, in *Handbook of semiconductor silicon technology*, eds. W.C. O'Mara, R.B. Herring, L.P. Hunt, P.640, Noyes Publ., New Jersey, USA (2009).
11. T.I. Kamins, *Polycrystalline Silicon for Integrated -Circuit Applications,* Kluwer Academic Publishing, Norwell MA (1988).
12. L.L. Kazmerski, *Polycrystalline and Amorphous Thin Films and Devices,* Academic Press, New York (1980).
13. H.C. deGraaff, in *Polycrystalline Semiconductors*, ed. G. Harbeke, p. 170, Springer-Verlag, Berlin, New York (1985).

14. F.C. Eversteyn, P.J.W. Severin, C.H.J. van den Brekel, and H.L. Peek, A stagnant layer model for the epitaxial growth of silicon from silane in a horizontal reactor, J. Electrochem. Soc. 117, 925-931 (1970).

15. A.S. Grove, *Physics and Technology o f Semiconductor Devices,* p. 18, Wiley, New York (1967).

16. R.S. Rosler, Low pressure CVD production processes for poly, nitride, and oxide, Solid State Technology, pp. 63-70 (1977).

17. C. Eversteyn, and B.H. Put. Influence of AsH_3, PH_3, and B_2H_6 on the growth rate and resistivity of polycrystalline silicon films deposited from a SiH_4-H_2 mixture, J. Electrochem. Soc. 120, 106-109 (1973).

18. L.H. Hall, and K.M. Koliwad, Low temperature chemical vapor deposition of boron doped silicon films, J. Electrochem. Soc. 120, 1438-1440 (1973).

19. S. Nakayama, I. Kawashima, and J. Murota, Boron doping effect on silicon film deposition in the Si_2H_6-B_2-H_6-He gas system, J. Electrochem. Soc. 133, 1721-1724 (1986).

20. D.W. Goodman, and R.R. Rye, in *Tungsten and Other Refractory Metals for VLSZ Applications,* Ed. R.S. Blewer, Materials Research Society, Pittsburgh, PA (1986).

21. Y. Yasuda, , and T. Moriya, Marked effects of borondoping on the growth and properties of polycrystalline silicon films, *Semiconductor Silicon 1973,* pp. 271 -284, Electrochemical Society, Pennington, NJ (1973).

22. D.B. Meakin, and W. Ahmed, LPCVD of *in-situ* phosphorus doped polysilicon from PH_3/Si_2H_6 mixtures, Fall 1986 *Electrochemical Society Meeting,* abstract 267, pp.398-399 (1986).

23. W. Ahmed, and D.B. Meakin, Phosphorus-doped silicon films prepared by low pressure chemical vapour deposition of disilane and phosphine, Thin Solid Films 148, L63-L65 (1 987).

24. J.M. Andrews, Electrical conduction in implanted polycrystalline silicon, J. Electron. Mater., 8, 227 (1979).

References to Chapter 5.

1. B.E. Deal and A.S. Grove, General relationship for the thermal oxidation of silicon, J. Appl. Phys. **36**, 3770 (1965).

2. J.D. Meindl, et. al. Silicon epitaxy and oxidation, in Process and device modeling for integrated circuits design, Eds. F. Van de Wiele, W.L. Engl, and P.O. Jespers, Noorhoff, Leyden, 1977.

3. A.S. Grove, Physics and technology of semiconductor devices, Wiley, New York, 1967.

4. G.S. May, S.M. Sze, Fundamentals of semiconductor fabrication, John Wiley & Sons, Inc., New York, 2004.
5. S.M. Sze, Semiconductor devices: Physics and technology, John Wiley & Sons, Inc., New York, 2002.
6. G.S. May, C.J. Spanos, Fundamentals of semiconductor manufacturing and process control, John Wiley & Sons, Inc., New York, 2006.
7. A.G. Revesz, J. Non-cryatalline solids **4**, 347 (1970).
8. A.G. Revesz, Phys. Stat. Sol (a) **57**, 235 (1980).
9. Kamins, T.I., Oxidation of phosphorus-doped low pressure and atmospheric pressure CVD polycrystalline- silicon films, J. Electrochem. Soc. **126**, 838-844 (1979).
10. T.I. Kamins, Preparation and properties of polycrystalline silicon films, in Handbook of semiconductor silicon technology, eds. W.C. O'Mara, R.B. Herring, L.P. Hunt, P.640, Noyes Publ., New Jersey, USA (2009).
11. Sunami, H., Thermal oxidation of phosphorus-doped polycrystalline silicon in wet oxygen, J. Electrochem. Soc. **125,** 892-897 (1978).
12. Irene, E.A., Tierney, E. and Dong, D.W., Silicon oxidation studies: Morphological aspects of the oxidation of polycrystalline silicon, J. Electrochem. Soc. **127**, 705-7 13 (1980).
13 Bravman, J.C. and Sinclair, R., Transmission electron microscopy studies of the polycrystalline silicon-SiO_2 interface, Thin Solid Films, **104**, 153-161 (1983).

References to Chapter 7.

[1] V. J. Logeeswaran, et al., "Harvesting and Transferring Vertical Pillar Arrays of Single-Crystal Semiconductor Devices to Arbitrary Substrates," Electron Devices, IEEE Transactions on, vol. 57, pp. 1856-1864, 2010.
[2] R. C. Jaeger, Introduction To Microelectronic Fabrication, 2 ed. Auburn, Upper Saddle River - Prentice Hall, 2002.
[3] M. J. Madou, Fundamentals of Microfabrication: The Science of Miniaturization, 2 ed.: CRC Press, 2002.
[4] Chen, et al., Effect of process parameters on the surface morphology and mechanical performance of silicon structures after deep reactive ion etching (DRIE) vol. 11. New York, NY, ETATS-UNIS: Institute of Electrical and Electronics Engineers, 2002.

References to Chapter 9.

1. Silicon Technologies. *Ion Implantation and Thermal Treatment.* Ed. Annie Baudrant, ISTE Ltd and John Wiley & Sons, Inc., London, Hoboken, 2011.

2. M. Nastasi, J.W. Mayer, Ion Implantation and Synthesis of Materials, Springer-Verlag Berlin, Heidelberg, 2006.

3. Ion Implantation, Ed. Mark Goorsky, Published by InTech, Rijeka, 2012.

4. ION IMPLANTATION, Science and Technology, Second Edition, Ed. J. F. Ziegler. ACADEMIC PRESS INC. , Boston, 1988.

5. Handbook of Plasma Immersion Ion Implantation and Deposition, ed. Andre Anders, John Wiley & Sons Inc. Hoboken, 2000.

References to Chapter 11.

1. Takayasu Sakurai, Akira Matsuzawa, Takakuni Douseki , Fully-depleted soi CMOS circuits and technology for ultralow-power applications, Springer, Dordrecht, The Netherlands, 2006.

2. Jean-Pierre Colinge, SILICON-ON-INSULATOR TECHNOLOGY: MATERIALS TO VLSI, Springer Science+Business Media, LLC, 2004

3. INSTABILITIES IN SILICON DEVICES, New Insulators, Devices and Radiation Effects Eds.Gerard BARBOTTIN and Andre VAPAILLE, ELSEVIER, AMSTERDAM, 1999.

4. Handbook of Wafer Bonding, eds*Peter Ramm, James Jian-Qiang Lu, and Maaike M.V. Taklo,* 2012 Wiley-VCH Verlag & Co. KGaA, Weinheim, Germany

5. Wafer Bonding. Applications and Technology, M. Alexe U. Gosele (Eds.), Springer-Verlag Berlin Heidelberg, 2004.

6. CMOS VLSI ENGINEERING, Silicon-on-Insulator (SOI). Eds. JAMESB.KUO and KER-WEI SU, 1998 Springer Science+Business Media Dordrech

7. SIMOX, eds.M.J.Anc, Institution of Engineering and Technology, London, UK, 2004

References to Chapter 12.

1. G.S. May, S.M. Sze, Fundamentals of semiconductor fabrication, John Wiley & Sons, Inc., New York, 2004.

2. S.M. Sze, Semiconductor devices: Physics and technology, John Wiley & Sons, Inc., New York, 2002.

3. G.S. May, C.J. Spanos, Fundamentals of semiconductor manufacturing and process control, John Wiley & Sons, Inc., New York, 2006.

4. C. Y. Liu and W. **Y.** Lee, "Process Integration," in C. Y. Chang and S. M. Sze, Eds., ULSI Technology, McGraw-Hill, New York, 1996.

5. T. Tachikawa, "Assembly and Packaging," in C. Y. Chang and S. M. Sze, Eds., ULSI Technology, McGraw-Hill, New York, 1996.

6. T. H. Lee, The Design of CMOS Radio-Frequency Integrated Circuits, Cambridge Univ. Press, Cambridge, U.K., 1998, Ch. 2.

7. D. Rise, "Isoplanar-S Scales Down for New Heights in Performance," Electronics, **53**, 137 (1979).

8. T. C. Chen, et. al., "A submicrometer High-Performance Bipolar Technology," IEEE Electron. Device Lett., **10**(8), 364, (1989).

9. G. P. Li et. al., "An Advanced High-Performance Trench-Isolated Self-Aligned Bipolar Technology," IEEE Trans. Electron Devices, **34**(10), 2246 (1987).

10. W. E. Beasle, J. C. C. Tsai, and R. D. Plummer, Eds., Quick Reference Manual for Semiconductor Engineering, Wiley, New York, 1985.

11. R. W. Hunt, "Memory Design and Technology," in M. J. Howes and D. V. Morgan, Eds., Large Scale Integration, Wiley, New York, 1981.

12. A. K. Sharma, Semiconductor Memodes-Technology, Testing, and Reliability, IEEE, New York, 1997.

13. U. Hamann, "Chip Cards-The Application Revolution," IEEE Tech. Dig. Int. Electron Devices Meet., p. 15, 1997.

14. R. D. Rung, H. Momose, and Y. Nagakubo, "Deep Trench Isolation CMOS Devices," IEEE Tech. Dig. Int. Electron. Devices Meet., p. 237, 1982.

15. D. M. Bron, M. Ghezzo, and J. M. Primbley, "Trends in Advanced CMOS Process Technology," Proc. IEEE, p. 1646, (1986).

16. H. Higuchi, et al., "Performance and Structure of Scaled-Down Bipolar Devices Merge with CMOSFETs," IEEE Tech. Dtg. Int. Electron. Devices Meet., 694, 1984.

17. M. A. Hollis and R. A. Murphy, "Homogeneous Field-Effect Transistors," in S. M. Sze, Ed., High-Speed Semiconductor Devices, Wiley, New York, 1990.

18. H. P. Singh, et al., "GaAs Low Power Integrated Circuits for a High Speed Digital Signal Processor," IEEE Trans. Electron. Devices, **36**,240 (1989).

19. International Technology Roadmap for Semiconductor (ITRS), Semiconductor Ind. Assoc., San Jose, 1999.

20. Y. Taur and E. J. Nowak, "CMOS Devices below 0.1 pm: How High Will Performance Go?" IEEE Tech. Dig. Int. Electron Devices Meet., 215, 1997.

21. L. Peters, "Is the 0.18 μm Node Just a Roadside Attraction," Semicond. Int., **22**, 46 (1999).

22. M. T. Bohr, "Interconnect Scaling-The Red Limiter to High Performance ULSI," IEEE Tech. Dig. Int. Electron Devices Meet., p. 241, 1995.

23. E. Leobandung, et al.,"Scalability of SO1 Technology into 0.13 μm 1.2 V CMOS Generation," IEEE Int. Electron Devices Meet., p. 403, 1998.
24. B. Martin, "Electronic Design Automation," IEEE Specr., **36**, 61 (1999).
25. H. Ishiuchi, et al., " Embedded DRAM Technologies," IEEE Tech. Dig. Int. Electron Devices Meet., p. 33, 1997.
26. S. Luryi, J. Xu, and A. Zaslavsky, Eds, Future Trends in Microelectronics, Wiley, New York, 1999.

References to Chapter 13.

1. Micromachining Techniques for Fabrication of Micro and Nano Structures, ed. Mojtaba Kahrizi, InTech, 2012
2. Micromachining of Engineering materials, *Ed. Joseph McGeough,* Marcel Dekker Inc New York, 2002
3. Nano and micromachining, Edited by J. Paulo Davim, Mark J. Jackson, ISTE Ltd and John Wiley & Sons, Inc,. London, Hoboken, 2009
4. Robert W. Johnstone & M. Parameswaran , An Introduction to surface-Micromachining, Kluwer Academic Publishers, Boston, 2004
5. Lyshevski, Sergey Edward, MEMS and NEMS : systems, devices, and structures, CRC Press LLC, N.W., 2002
6. Ronald D. Schaeffer , FUNDAMENTALS OF LASER MICROMACHINING, CRC Press Taylor & Francis Group, 2012
7. Rolf Wuthrich, MICROMACHINING USING ELECTROCHEMICAL DISCHARGE PHENOMENON, Elsevier Inc, Oxford, 2009
8. Femtosecond Laser Micromachining, Photonic and Microfluidic Devices in Transparent Materials, eds. Roberto Osellame, Giulio Cerullo, Roberta Ramponi, Springer, Heidelberg, 2012
9. Mark J. Jackson, Micromachining with Nanostructured Cutting Tools, Springer, London, 2013
10. Koji Sugioka, Ya Cheng, Femtosecond Laser 3D Micromachining for Microfluidic and Optofluidic Applications, Springer, London, 2014
11. Handbook of Silicon Based MEMS Materials and Technologies, eds.Veikko Lindroos, Markku Tilli , Ari Lehto and Teruaki Motooka, Oxford , Elsevier, 2010

References to Chapter 16.

1. G.S. May, C.J. Spanos, Fundamentals of semiconductor manufacturing and process control, John Wiley & Sons, Inc., New York, 2006.

2. D. Halliday and R. Resnick, Physics, NY: Wiley, New York, 1978.

3. F. Yang, W. McGahan, C. Mohler, and L. Booms, "Using Optical Metrology to Monitor Low-K Dielectric Thin Films," Micro **31–38** (May 2000).

4. K. Wong, D. Boning, H. Sawin, S. Butler, and E. Sachs, "Endpoint Prediction for Polysilicon Plasma Etch via Optical Emission Interferometry," J. Vac. Sci. Technol. A. **15**(3), (May/June 1997).

5. H. Tompkins and W. McGahan, Spectroscopic Ellipsometry and Reflectometry,Wiley, New York, 1999.

6. J. McGilp, D. Weaire, and C. Patterson (eds), Epioptics, Springer-Verlag, New York, 1995.

7. W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, Numerical Recipes in C, Cambridge Univ. Press, Cambridge, MA, 1988.

8. R. Jaeger, Introduction to Microelectronic Fabrication, Addison-Wesley, Reading, MA, 1993.

9. A. Landzberg, Microelectronics Manufacturing Diagnostics Handbook, Van Nostrand Reinhold, New York, 1993.

10. C. Raymond, "Angle-Resolved Scatterometry for Semiconductor Manufacturing," Microlithogry. World (winter 2000).

11. J. Pineda de Gyvez and D. Pradhan, Integrated Circuit Manufacturability, IEEE Press, Piscataway, NJ, 1999.

12. Kees de Kort, Techniques for Characterization and Failure Analysis of Integrated Circuits, in Analysiis of the Microelectronic Materials and Devices, Eds.: M. Grasserbauer, and H.W. WernerJohn Wiley & Sons, New York, pp. 895-919, 1991.

13. C. G. C de Kort, Philips J. Pes., 44, 295 (1989).

14. G. II. Heilmeier, L. A. Zanoni and L. A. Barton, IEEE Trans. Electron. Dev., 17, 22 (1970).

15. J. L. Lanning, Appl. Phys. Lett., 21, 173 (1972).

16. D. J. Channin, IEEE Trans. Electron. Dev., 1974, 650 (1974).

17. E. M. Fleuren, Proceedings 21st International Reliability Physics Symposium (Phoenix, Ariz.), p. 148, IEEE, New York (1983).

18. D. Burgess and P. Tan, Proceedings of 22nd International Reliability Physics Symposium (Las Vegas. Nev.), p. 119, IEEE, New York (1984).

19. R. Weyl, B. Lischke, R. Kappelmeyer and F. Beck, Mitteilungen aus den Forschungs Laboratorien und dem Werk fur Baulemente, p. 1, Siemens AG, Munchen (1985).

20. F. Beck. Elektronik, 13. 82 (1986).

21. P. Hendriks, K. de Kort, R. E. Horstman. J-P. Andre, C. T. Foxon and J. Wolter. Semicond. Sci. Techno!., 3. 521 (1988).
22. N. Khurana and C-L. Chiang, Proceedings of 25th International Reliability Physics Symposium (San Diego, Calif.), p. 72, IEEE, New York (1987).
23. J. A. Valdmanis, Electron. Lett., 23, 1308 (1987).
24. R. K. Jain, Test and Measurement World, 4, 40 (1984).
25. B. H. Kolner and D. M. Bloom, IEEE J. Quantum Electron., 22. 79 (1986).
26. J. A. Valdmanis and G. Mourou, IEEE J. Quantum Electron., 22, 69 (1986).
27. G. Mourou, in Characterization of Very High Speed Semiconductor Devices and Integrated Circuits (Ed. Ravi Jain), in Proc. Soc. Photo-opt. Instrum. Enging., 795. 300 (1987).
28. K. Kelterer, E. H. Boucher and D. Bimberg, Appl. Phys. Lett., 50, (21), 1471 (1987).
29. J. A. Valdmanis and S. S. Pei, Proceedings of Conference on Picosecond Electronics and Optoelectronics (Incline Village. USA, 14-16 January 1987). pp. 4—6, Springer, New York (1987).
30. J. Nees and G. Mourou, Electron. Lett., 22, (17), 918 (1986).
31. J. A. Valdmanis and G. Mourou, Laser FocusjElectro-optics, p. 96, March 1986.
32. T. C. May and M. H. Woods. IEEE Trans. Electron. Dev., 26, 2 (1979).
33. R. R. Troutman (Ed.), Latchup in CMOS Technology, Kluwer, Dordrecht (1986).
34. R. J. G. Goossens and J. H. A. van der Wielcn, Philips J. Res., 44, 241 (1989).
35. A. M. T. P. van der Putten, J. W. M. Jacobs, J. M. G. Rikken and C. Q. C. de Kort, in Laser Assisted Processing (Eds. L. D. Laude and G. Rauscher), in Proc. Soc. Photo-opt. Instrum. Engng.. 1022, 71 (1989).
36. P. May, J.-M. Halbout and G. Chiu, in Characterization of Very High Speed Semiconductor Devices and Integrated Circuits (Ed. Ravi Jain), in Proc. Soc. Photo-opt. Engng., 795, 201 (1987).
37. A. Blacha, R. Clauberg, H. Seitz. W. Wolz and H. Beha. in Characterization of Very High Speed Semiconductor Devices and Integrated Circuits (Ed. Ravi Jain), in Proc. Soc. Photo-opt, Engng., 795, 286 (1987).
38. J.-M. Halboui, P. G. May and M. B. Ketchcn, in Characterization of Very High Speed Semiconductor Devices and Integrated Circuits (Ed. Ravi Jain), in Proc. Soc. Photo-opt. Engng., 795, 247 (1987).
39. Francois J. Henley and Hee-June Choi, Proceedings IEEE International Test Conference (Washington, D.C.), p. 700, Computer Society Press of IEEE, Washington, DC. (1988).
40. R. Clampitt, R. Watkins and J. Whitaker, Microetectron. Engng., 6, 605 (1987).