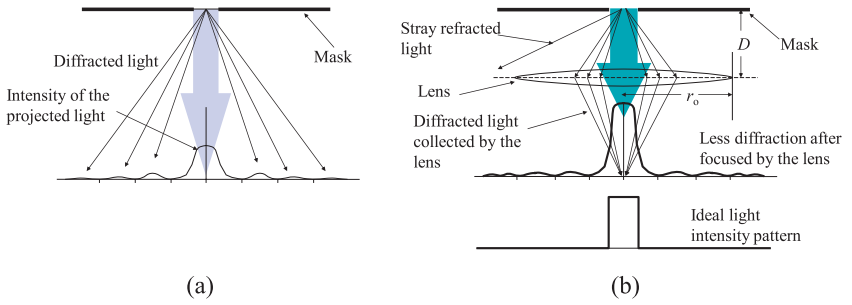


Figure 6.42 Light diffraction (a) without and (b) with a lens.

some of the diffracted light can improve resolution by reducing the light diffraction (see Fig. 6.42).

The minimum resolution an optical system can achieve is determined by the wavelength of the light and the numerical aperture of the system. The resolution can be expressed as

$$R = \frac{k_1 \lambda}{NA}, \tag{6.1}$$

where k_1 is the system constant, λ is the wavelength of the light, and NA $= 2r_o/D$ is the numerical aperture, which indicates the capability of the lens to collect the diffracted light. D is the distance between the object, mask (or reticle), and the lens, and $2r_o$ is the diameter of the lens. Equation (6.1) shows that finer resolution can be achieved with a larger lens diameter, similar to a camera with a larger lens achieving a sharper image. However, an optical system with a large lens is much more expensive, the way a camera with a large lens is more expensive than a compact camera with a small lens. There are also technological limitations when creating a high-precision lens with a very large diameter.

Equation (6.1) also illustrates that resolution improves when a shorter wavelength is used for exposure. This is why the wavelength of an exposure light source becomes shorter and shorter in photolithography processes (Fig. 6.43). However, there is a limit; when a wavelength reduces to a certain value, light leaves the UV range and becomes an x ray, as illustrated in Fig. 6.44. For x rays, most optical equations are no longer valid, including Eq. (6.1).

Another important property of an optical system is depth of focus (DOF), which is the range in which light is in focus and can achieve good resolution of the projected image. DOF can be expressed as

$$DOF = \frac{k_2 \lambda}{2(NA)^2}. \tag{6.2}$$

Equation (6.2) confirms that an optical system with a smaller NA has a larger DOF, which is why point-and-shoot cameras on cell phones have very small lenses.

The focus needs no adjustment when a picture is taken with this kind of camera, since its DOF is so large that almost everything is in focus. However, this kind of camera cannot produce a sharp image, since resolution will be poor with such a small lens. Figure 6.45 gives an illustration of the DOF of an optical system.

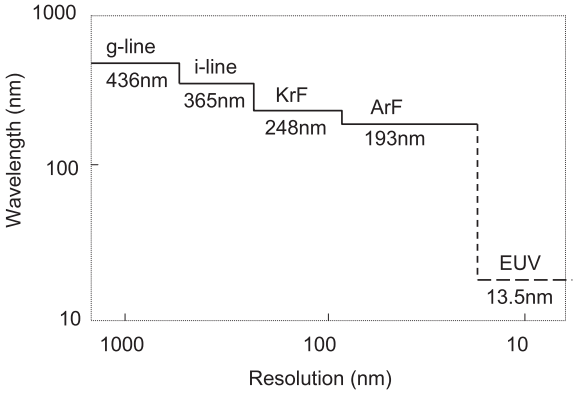
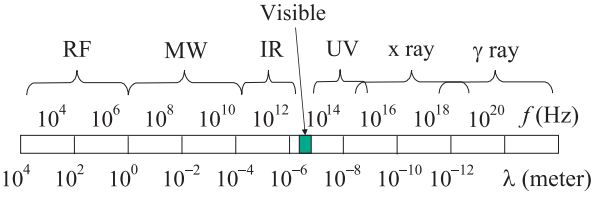


Figure 6.43 Relationship of lithography wavelength and resolution.



RF: Radio frequency; MW: Microwave; IR: infrared; and UV: ultraviolet

Figure 6.44 Wavelength and frequency of an electromagnetic wave.

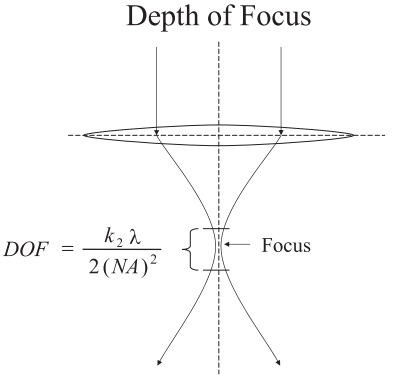


Figure 6.45 DOF of an optical system.

The larger the DOF is on a scanner system, the easier it will be to focus the exposure light in the photoresist on a wafer surface. Unfortunately, the DOF and resolution always work against each other—to improve resolution, a shorter wavelength and larger NA are needed, both of which reduce the DOF.

In advanced photolithography processes, resolution is so high that the DOF becomes very small; thus, the center of the focus must be placed exactly at the mid-plane of the photoresist to achieve the best lithography results, as shown in Fig. 6.46. Photolithography engineers in IC fabs routinely use a focus-exposure matrix (FEM) to check on focusing issues when exposing a photoresist wafer.

Because of DOF requirements in photolithography processes, wafer surfaces need to be highly planarized for the patterning of feature sizes smaller than 0.25 μm. This is one reason that CMP processes are widely used in advanced semiconductor fabs. Only a CMP process can achieve the surface planarization results required by photolithography resolution for quarter-micron (and smaller) pattern geometry.

6.4.2 Mercury lamps and excimer lasers

Since shorter wavelengths achieve higher resolution, stable, high-intensity, and short-wavelength light sources have been developed and applied in exposure systems. Ultrahigh-pressure mercury lamps and excimer lasers are widely used as light sources for steppers.

A mercury vapor lamp has many lines of radiation, of which an i-line at 365 nm is the most commonly used in stepper exposure systems for 0.35-μm feature size IC processes.

KrF excimer lasers at 248 nm have been developed as light sources of DUV scanners for 0.25-μm minimum feature size IC applications, and they have also been used for 0.18- and 0.13-μm IC fabrication. KrF scanners are still heavily used in noncritical layers such as implantations and upper metal layers of advanced nanometer technology IC fabs. Scanners using ArF excimer lasers at 193 nm are widely used for 130- to 65-nm IC chip manufacturing. Photolithography processing tools using F_2 lasers with 157-nm wavelengths have been studied. However, the development of a 157-nm scanner has been discontinued due to

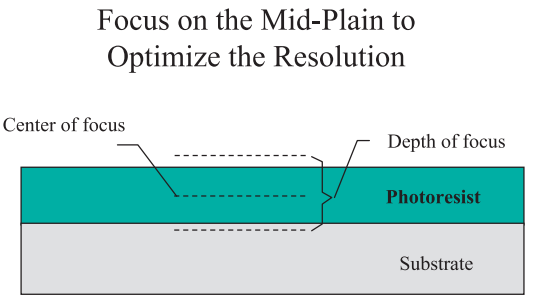


Figure 6.46 Focus light in the mid-plane of a photoresist to optimize resolution.

the development of 193-nm immersion lithography technology, which extends the resolution of ArF systems to the 45-nm technology node. Combined with double-patterning or multiple-patterning techniques, ArF systems could possibly push IC manufacturing beyond the 22-nm node. Next-generation lithography (NGL) technologies, such as EUV lithography, nanoimprint lithography (NIL), and e-beam direct write (EBDW), could start seeing use in IC manufacturing in 16-nm or smaller technology nodes by the mid-2010s. These technologies will not replace optical lithography in the near future.

Different lithography exposure wavelengths use different photoresists because a photoresist is designed to be sensitive only to certain wavelengths. ARC is also related to exposure wavelength. Different wavelengths require different dielectric deposition processes for dielectric ARCs or different spin coating processes for BARCs.

Since the 1980s, there have always been predictions that photolithography processing must be replaced by an alternative lithography process within ten years, due to the resolution limit of the optical exposure system. However, scientists and engineers working in photolithography processing constantly improve resolution, push the limit, and extend the lifetime of photolithography (see Fig. 6.47).

Because of the development of photolithography technologies such as phase shift mask (PSM), optical proximity correction (OPC), off-axis illumination, immersion lithography, and double patterning, engineers can use 193-nm optical lithography systems to pattern 22-nm technology node devices. With the development of multiple-patterning techniques, it is possible to push the limit farther and delay deployment of NGL technologies.

6.4.3 Resolution enhancement techniques

To improve photolithography resolution, several resolution enhancement techniques have been developed and applied in IC chip manufacturing that extend

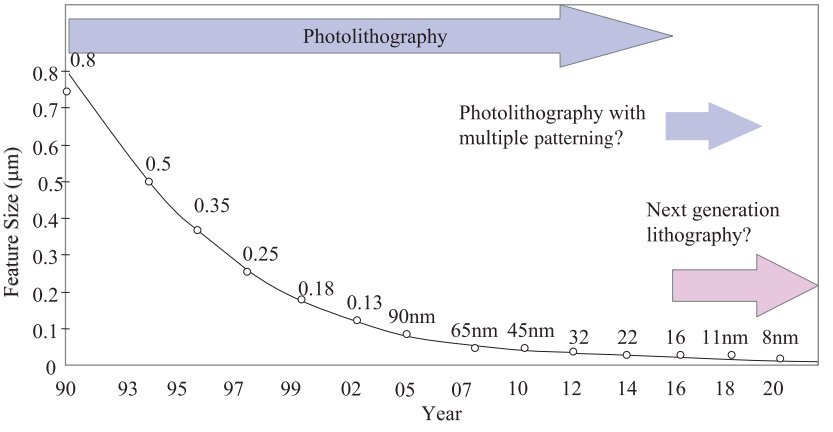


Figure 6.47 IC technology nodes and lithography trends.

optical lithography to the 22-nm technology node and possibly beyond. This section briefly describes some important resolution enhance techniques, such as PSM, OPC, and off-axis illumination.

6.4.3.1 Phase shift mask

Equation (6.1) shows that photolithography resolution can be improved by reducing k_1. One of the methods used to reduce k_1 is the PSM.

It is not difficult to transfer an isolated, small feature pattern to a photoresist; the challenge comes when the small features are densely packed—light diffraction and interference can distort the feature. To solve this problem, a PSM is introduced. The PSM has a phase shifter at every other opening (clear, transparent area) on the mask, as shown in Fig. 6.48. Although this kind of PSM is not used in IC manufacturing, it is still useful for explaining the principle of a PSM.

The thickness and dielectric constant of the dielectric layer are carefully controlled, so that $d(n_f - 1) = \lambda/2$, where d is the thickness of the dielectric, n_f is the dielectric constant, and λ is the exposure light wavelength. Due to opposite phase shift, light passing through an open area without phase shift coating will destructively interfere with light passing through an opening with phase shift coating, resulting in a sharper image in the densely packed area, as shown in Fig. 6.49.

By etching a quartz substrate instead of adding a shifter, an alternating aperture PSM (AAPSM) can be created, as shown in Fig. 6.50. The etch depth needs to be precisely controlled at $d = \lambda/[2(n - 1)]$, where d is the etch depth, λ is the wavelength of exposure light, and n is the refractive index of the quartz substrate. An etch depth of 171.4 nm is needed for a 193-nm exposure wavelength and a quartz refractive index of 1.563 at 193 nm.

Another type of PSM used in IC manufacturing in nanometer node technology is an attenuated PSM (AttPSM). An AttPSM is formed by patterning the partially transparent film deposited on the quartz substrate (see Fig. 6.51).

Molybdenum silicide (MoSi) with transmittance of 6 to 20% is commonly used for this application. Thickness of the MoSi film is carefully controlled so that the small amount of light passing through it has a phase difference 180 deg from the light passing through the quartz at the exposure wavelength (193 nm for an ArF excimer laser). Light intensity in the area covered by MoSi is lower than the photoresist exposure threshold because of the destructive interference of the light, while light intensity in the area without MoSi is higher than the photoresist

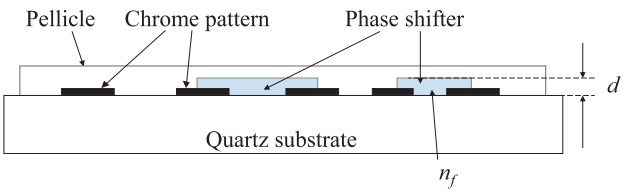


Figure 6.48 A PSM.

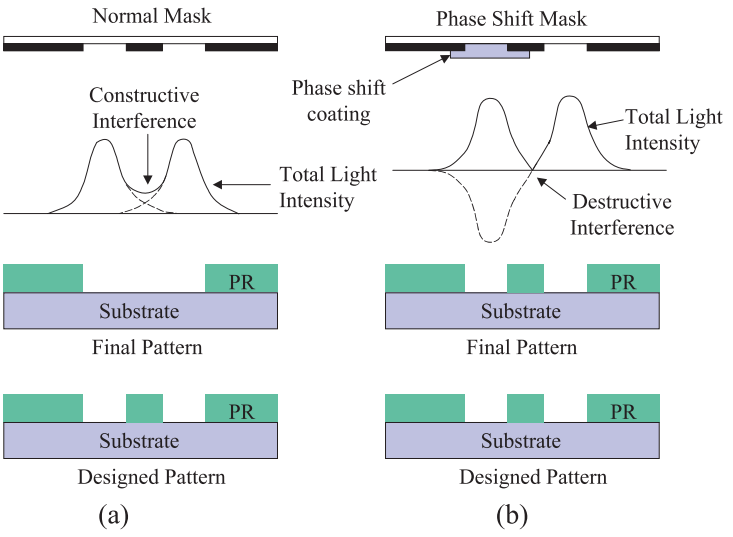


Figure 6.49 Patterning with (a) normal mask and (b) PSM.

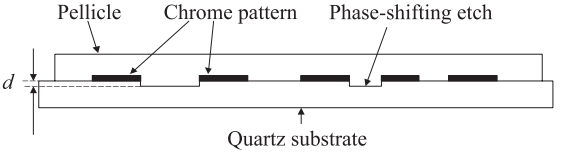


Figure 6.50 AAPSM formed by quartz etch.

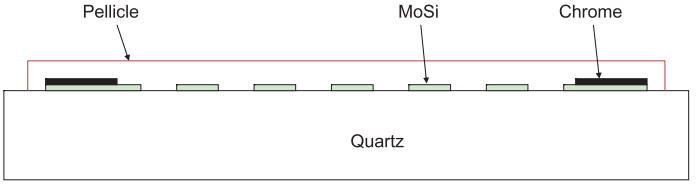


Figure 6.51 The AttPSM.

exposure threshold. This allows high-resolution subwavelength patterning (see Fig. 6.52).

6.4.3.2 Optical proximity correction

In the good old days when feature size was larger than the exposure wavelength, patterns printed on a wafer were almost the same as patterns on the mask, except for some corner-rounding effects caused by light diffraction. When feature size becomes smaller than the wavelength, the light diffraction effect becomes more profound, and patterns printed on a wafer are no longer the same as patterns on the

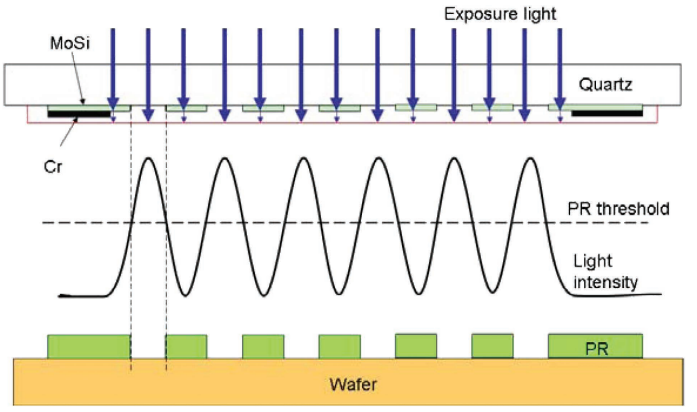


Figure 6.52 Photoresist patterning with AttSPM.

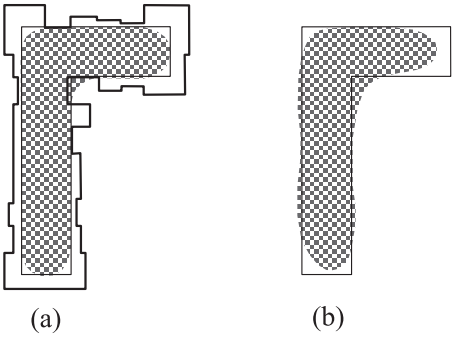


Figure 6.53 Optical lithography pattern (a) with and (b) without OPC.

mask. Therefore, to print the desired patterns on a wafer, miniature features have to be added on the mask pattern to compensate for diffraction effects. These add-on features are called OPC features (see Fig. 6.53).

6.4.3.3 Off-axis illumination

By using an aperture that forces incident light to enter at an angle (off axis) through the lens of an optical system, it is possible to collect the first diffraction order of light from the mask to effectively reduce the k_1 factor in Eq. (6.1) and improve lithographic resolution. Figure 6.54(a) illustrates an on-axis illumination system, and Fig. 6.54(b) shows an off-axis illumination system, which is also called monopole illumination, since incident light comes from a single pupil of the aperture.

A dipole illumination system uses two different illuminations with symmetric angles; this allows zero- and first-order diffraction from both poles to be collected and imaged, resulting in more balance in intensity than with monopole illumination. An illustration of diffracted orders of dipole illumination and a dipole

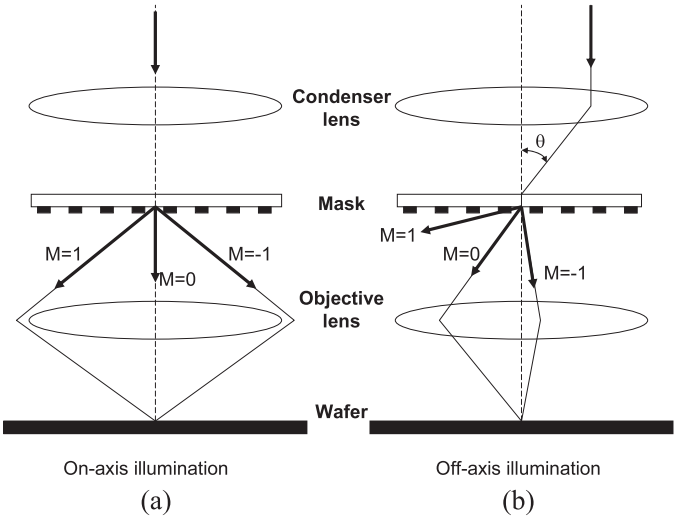


Figure 6.54 Schematics of (a) on-axis and (b) off-axis illumination.

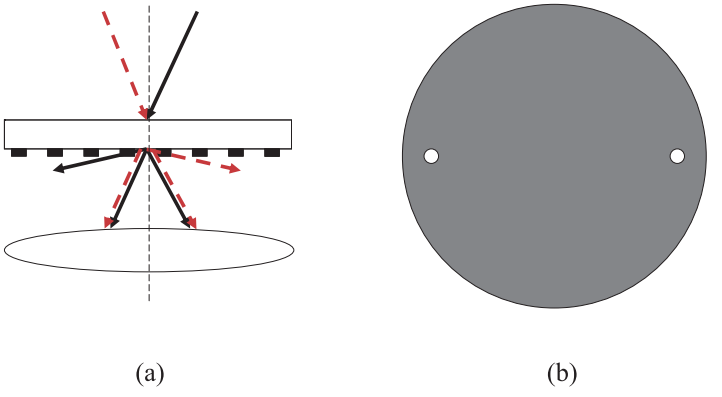
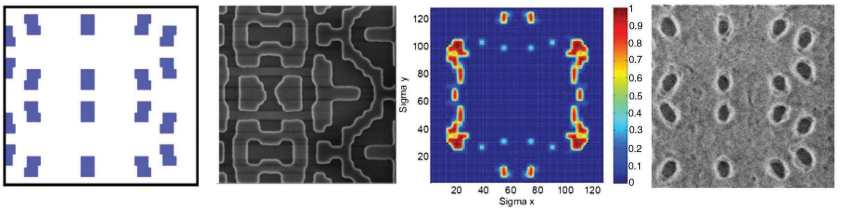


Figure 6.55 (a) Dipole illumination and (b) an aperture of dipole illumination.

illumination pupil can be seen in Figs. 6.55(a) and 6.55(b), respectively. Similarly, a quadruple illumination source can be used for optical lithography.

By optimizing both aperture features and mask patterns, called source-mask optimization (SMO), scientists and engineers can stretch the applications of optical lithography. Figure 6.56 shows an example of SMO. The pattern on the mask [Fig. 6.56(b)] is completely different from the pattern on the wafer [Fig. 6.56(d)], and the source design [Fig. 6.56(c)] is much more complicated than dipole or quadruple designs. Because it would take an excessive amount of computation to optimize both source and mask patterns, this technique is also called computational lithography. Combined with immersion technology, optical lithography can be used



(a) (b) (c) (d)

Figure 6.56 Example of SMO of contact photolithography: (a) design layout, (b) mask, (c) designed source, and (d) photoresist pattern on the wafer (K. Lai, et al., *Proc. of SPIE*).

for 22-nm technology node patterning. One of the advantages of this approach is that the mask is still binary, which means that it is easier to create than a PSM.

6.4.4 Immersion lithography

By immersing the gap between the objective lens of a microscope and a sample viewed with water or oil, resolution of the microscope image can be improved. The same idea can be applied to photolithography. By filling the gap between the objective lens and wafer surface with DI water, lithographic resolution can be improved significantly. Resolution of immersion lithography can be expressed by

$$R = \frac{k_1 \lambda}{n_{fluid} NA}. \tag{6.3}$$

The only difference between Eqs. (6.1) and (6.3) is n_{fluid}, the refractive index of the fluid between the objective lens and wafer. When the fluid is air, the refractive index is very close to 1, and Eq. (6.3) becomes Eq. (6.1). Equation (6.3) demonstrates that by adding water ($n_{fluid} = 1.44$ at 193 nm) between the objective lens and wafer of a 193-nm scanner, resolution can be improved by one technology node without a major change in the optical system.

$$DOF_{immersion} = \frac{1 - \sqrt{1 - (\lambda/p)^2}}{n_{fluid} - \sqrt{n_{fluid}^2 - (\lambda/p)^2}} \frac{k_2 \lambda}{2(NA)^2}, \tag{6.4}$$

where λ is the wavelength of the exposure light, and p is the pitch of the mask pattern. When $n_{fluid} = 1$, Eq. (6.4) becomes Eq. (6.2).

The DOF improvement factor of an immersion system, $\eta = DOF_{immersion}/DOF$, is larger than n_{fluid}, meaning that the DOF increases when applying immersion fluid between the objective lens and wafer. When applying a DI wafer, the DOF increases by at least a factor of 1.46 with 193-nm immersion lithography for large-pitch mask patterns. For smaller-pitch mask patterns, the DOF improvement factor

η increases more (see Fig. 6.57). By increasing the DOF using immersion fluid, the NA can be increased to further enhance resolution while keeping the DOF in a comfortable range. High-NA 193-nm immersion lithography is widely used in IC manufacturing for 45- to 22-nm technology nodes. Figure 6.58 shows that water is only in the gap between the objective lens and wafer in an immersion lithography system.

6.4.5 Double, triple, and multiple patterning

By patterning a single mask layer more than once, patterning resolution will improve because the k_1 factor of Eq. (6.1) is effectively reduced. If the same lithography system is used, the effective k_1 will be k_1/N, where N is the number of the patterning. For commonly used double patterning, $k_{1,\text{eff}} = k_{1/2}$. Double-patterning technology (DPT) has been used in IC chip manufacturing since the 45-nm technology node and is widely used in 32- and 22-nm technology nodes. It is also possible to pattern more than twice, such as three times (triple patterning) or four times (quadruple patterning), to further reduce effective k_1 and to pattern even

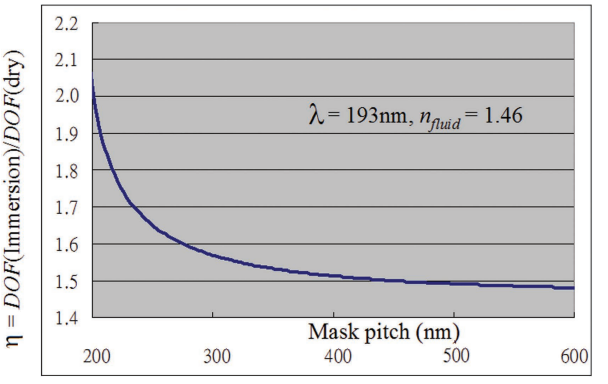


Figure 6.57 Immersion-induced DOF improvement versus mask line/space pitch.

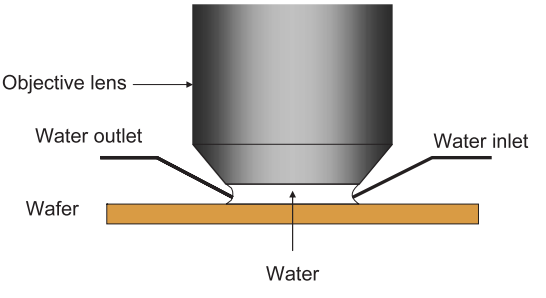


Figure 6.58 Schematic of an immersion lithograph system.

finer features. If NGL (such as EUV lithography) is not ready in time for 16-nm technology nodes and beyond, multiple-patterning technologies could be used.

There are many approaches of DPT, such as litho-freeze-litho-etch (LFLE), litho-etch-litho-etch (LELE), and self-aligned double patterning (SADP). LFLE is very attractive to IC chip manufacturers because it is regarded as the DPT with the lowest cost. By freezing the developed photoresist of the first mask and applying the second photoresist layer then exposing it with the second mask, the pitch density can be doubled using an optical lithography system with resolution that cannot pattern the final pitch density. Figure 6.59 illustrates the LFLE process. The pattern freeze of the first photoresist can be performed with a chemical process that reduces the solubility of the photoresist while it is in the developer solution so that the photoresist will not dissolve when developing the second photoresist. Ion implantation that hardens the photoresist has also been studied for this application.

The LELE (or LE^2) process, illustrated in Fig. 6.60, uses two layers of hard mask. One of the most challenging portions of this process flow is the etch selectivity of the first hard mask to the second hard mask, a process that is discussed in Chapter 9. If polysilicon is the material to be etched on the wafer surface, silicon oxide can be chosen as the material for the second hard mask, and amorphous carbon can be chosen for the material of the first hard mask. Using oxygen plasma ashing, amorphous carbon can be removed when stripping the second photoresist.

LELE double patterning can also be achieved with one hard mask. The drawback of this approach is that the CD of the final pattern is directly related to the overlay between the two masks. Any overlay error will transfer into a CD variation (see Fig. 6.61).

In a critical mask layer, such as a MOSFET gate, the CD control gate limit is plus or minus 10% of the target CD, while the overlay control is a little more relaxed at about 20% of the target CD. Therefore, this LELE double-patterning technique

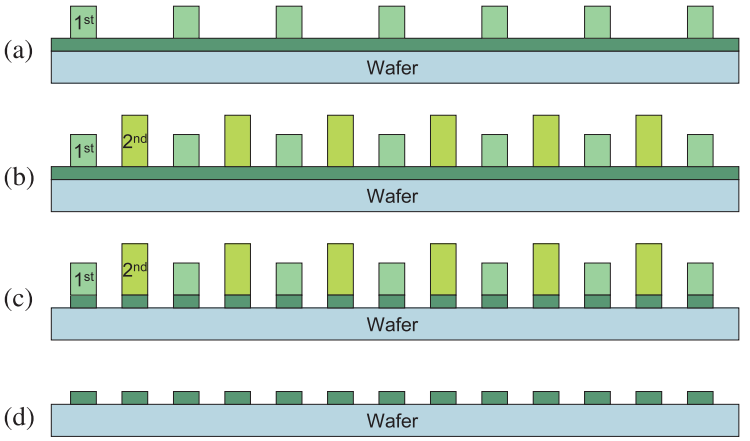


Figure 6.59 Illustrations of the LFLE process: (a) the first photoresist pattern and pattern freeze, (b) the second photoresist pattern, (c) the etch pattern, and (d) the strip photoresist.

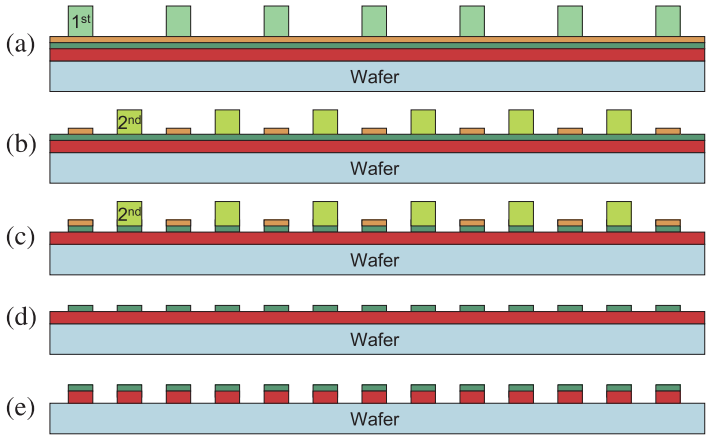


Figure 6.60 Illustrations of the LELE process: (a) first litho, (b) second litho after etch of the first hard mask layer, (c) etch of the second hard mask layer, (d) stripping of the second photoresist and the first hard mask, and (e) etch of the pattern on the wafer.

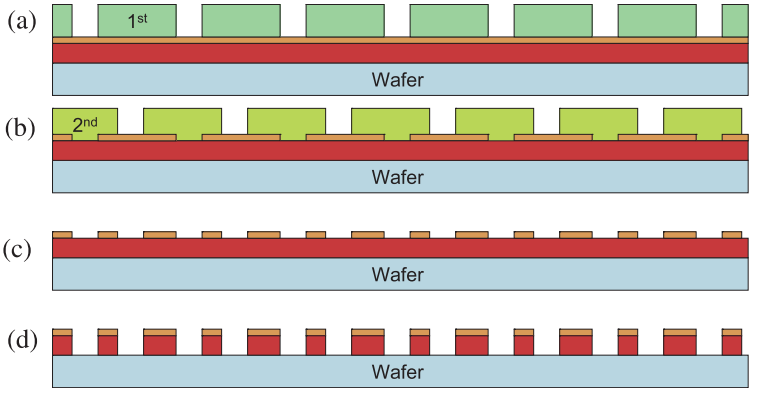


Figure 6.61 Illustrations of the LELE process using only one hard mask. Note the CD variation of the final pattern induced by an overlay error of the first and second masks.

can only be used for noncritical layers, where CD control can also be relaxed to ~20%.

Spacer (self-) aligned double patterning (SADP) is the most promising technique for patterning sub-22-nm device features, especially for NAND flash manufacturing. Figures 6.62 and 6.63 illustrate a cross-sectional view and top view of the SADP process, respectively. The first mask defines the photoresist pattern, as well as the low-temperature oxide CVD and etchback to form the spacer. After removal of the photoresist dummy pattern, the hard mask is etched for the first time. After second photomask patterning, hard mask etch, and photoresist removal, the third etch forms the designed device pattern. The pattern CD is controlled by the

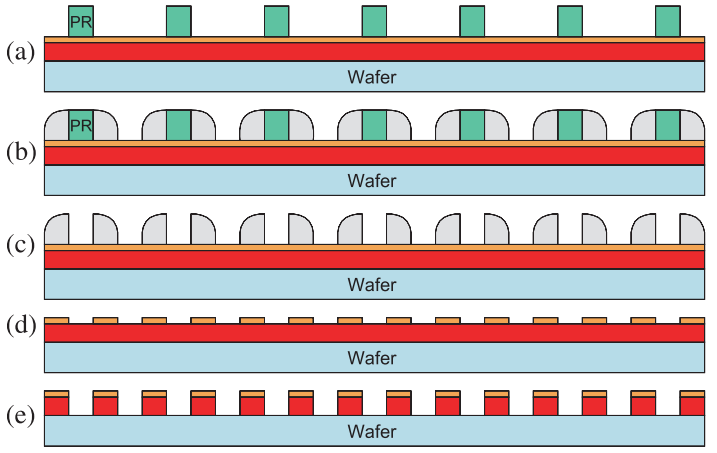


Figure 6.62 Illustrations of the SADP process: (a) first mask photoresist patterning; (b) low-temperature oxide CVD and spacer etch; (c) photoresist strip; (d) hard mask etch, spacer oxide strip, second mask pattern, and hard mask etch; and (e) final pattern etch.

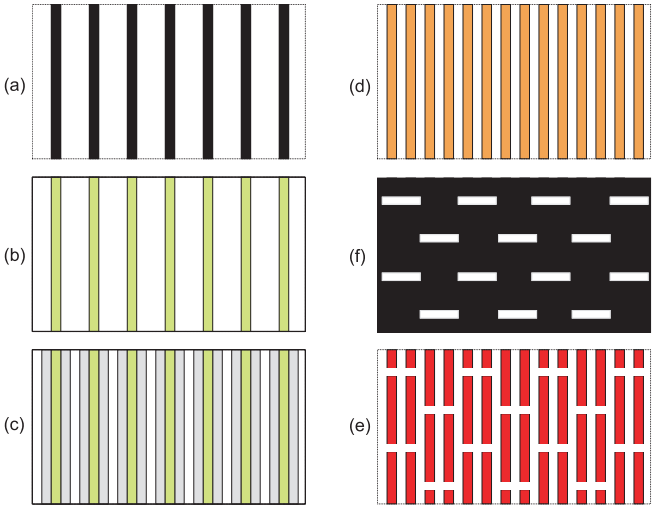


Figure 6.63 Top view of the SADP process: (a) the first mask, (b) the photoresist pattern of the first mask, (c) low-temperature oxide spacer formation, (d) hard mask etch and spacer removal, (e) the second mask, and (f) hard mask and final pattern etch.

thickness of a low-temperature CVD oxide film that forms the spacer, a thickness that can be precisely controlled with a transparent thin-film metrology system.

Because SADP requires more processing steps than other double-pattern processes (such as oxide CVD and spacer etch), it is the most expensive approach of the three DPTs discussed. However, it has many advantages, such as accurate pattern CD and space CD control, less strict overlay requirements for the second

mask, low line edge roughness (LER), etc., making it very attractive to some IC manufacturers, especially NAND flash makers. This is because NAND flash has many layers of dense line–space patterns with very tight CD requirements, which is perfect for SADP application. For advanced microprocessor manufacturers, SADP is also very attractive due to its good CD control and low LER.

Although photolithography has its limits, patterning has not reached its limit yet. It is possible to use multiple pattering such as quadruple patterning to achieve even smaller feature size. Directed self-assembly could also be used to multiply pattern frequency and increase pattern density. Fabs could even use double-patterning techniques on the tiny features formed by EUV lithography or NIL to further reduce feature size.

6.4.6 Extreme-ultraviolet lithography

One of the most promising NGL technologies for patterning sub-22-nm technology nodes is EUV lithography with a 13.5-nm wavelength. An electromagnetic radiation wave with wavelength between 1 and 50 nm is in the region between UV and x ray. It can be called EUV, vacuum UV, or soft x ray. At one time it was called soft x-ray projection lithography, but was renamed EUV lithography to avoid confusion with x-ray lithography in the late 1990s.

The basic idea of EUV lithography is that by sharply reducing wavelength λ and moderately reducing NA, lithography resolution can be improved. The process can also operate in the comfort zone of DOF > 100 nm for mass manufacturing. For example, when $k_1 = 0.25, k_2 = 1.0, \lambda = 13.5$ nm, and NA = 0.25, from Eqs. (6.1) and (6.2), resolution $R = 13.5$ nm, and DOF = 108 nm.

Due to strong absorption at short wavelengths, no material can be used to make a lens for the EUV lithography process; thus, an EUV system must use a mirror-based system. High-intensity EUV light sources have been developed for preproduction systems (~60 wafers/h), and sources with even higher intensity for full production systems (120 wafers/h) are still in development. Figure 6.64 illustrates an EUV exposure system.

To effectively reflect EUV light, multilayer coatings of thin Mo/Si pairs are needed. Figure 6.65 illustrates an EUV mask that uses a quartz substrate. Backside metal coating is needed for electrostatic chucking, which is the preferred chucking mechanism in semiconductor manufacturing because it eliminates mechanical chucking, which is prone to generating particles. About 40 pairs of Mo and Si thin films are deposited on top of the quartz to form a reflective layer of EUV light at a 13.5-nm wavelength. About 70% reflective efficiency can be achieved at a 6-deg incident angle. The absorber usually is boron-doped tantalum nitride (TaBN), and the buffer layer can be chromium nitride (CrN) with a silicon capping layer or ruthenium (Ru). The buffer layer is used to protect the multilayer from damage when pattering the absorber. The ARC usually is boron-doped tantalum oxynitride (TaBON) deposited on top of a TaBN absorbing layer. It can reduce reflection of the absorber area and enhance contrast of the multilayer area at DUV inspection wavelengths to improve the sensitivity of defect inspection using optical

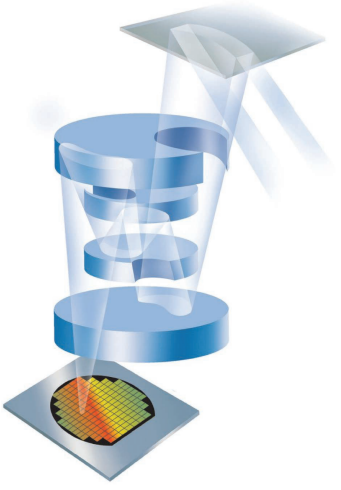


Figure 6.64 EUV lithography system (Carl Zeiss).

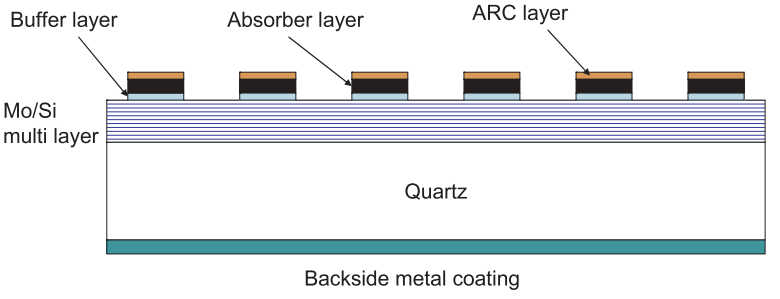


Figure 6.65 An EUV mask.

mask inspection systems. E-beam inspection and actinic inspection do not require ARC.

Ideally, an EUV mask can be inspected at the EUV wavelength of 13.5 nm, which is called actinic inspection. Theoretically, actinic inspection can capture all printable defects caused by both surface and phase defects, both of which are induced by defects embedded or below the multilayer. Since actinic EUV mask inspection systems are still in development, optical mask and e-beam mask inspections can be used for the development and pilot line production of EUV lithography. Due to an inability to capture embedded phase defects in the EUV mask, it is very likely that EUV lithography will be deployed in mask layers insensitive to phase defects, such as bitline contact layers of NAND flash and some via layers of CMOS. Because more than 90% or even 99% of the areas are covered by an absorber, embedded phase defects have very little chance of being printed to affect device performance. For memory chips such as NAND flash, the redundancy

design would allow them to withstand some defects without suffering major yield loss.

6.4.7 Nanoimprint lithography

Imprint technology is widely used to make coins, music, videos, and software compact disks. Using an imprint lithography process to generate nanometer IC features was proposed in 1996, and this alternative lithography technology was called NIL. Figure 6.66 illustrates the NIL patterning process. First, a resist is sprayed onto a wafer surface [see Fig. 6.66(a)]; then the template is pressed onto the resist [Fig. 6.66(b)]; and the resist is hardened by applying either thermal or UV energy [Fig. 6.66(c)]. After resist hardening, the quartz template or mold is released from the wafer surface [Fig. 6.66(d)], and an etch process follows to transfer the resist pattern formed by NIL to the thin film on the wafer surface [Fig. 6.66(e)].

Figure 6.67 shows a line–space pattern formed by the NIL process. Notice that the CD of the line is only 11 nm with very low LER.

The advantage of NIL is high resolution, low LER, and a low cost of ownership, especially for the 22-nm technology node. However, to apply NIL in high-volume IC production, several issues must be solved: template defects, overlay, throughput, and management of thousands of working templates. A template might be able to last for about 10,000 prints, which is only about 100 wafers. To reduce the cost of wafer printing, thousands of working templates need to be duplicated from a master template with the nanoimprint technique.

One promising application of NIL is to form pattern media of hard disk drives (HDDs). When HDD storage increases, the size of storage cells reduces. HDD

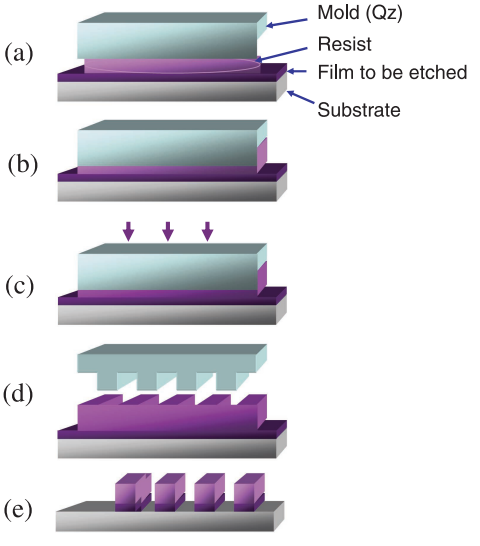


Figure 6.66 NIL patterning process: (a) resist coating, (b) imprint, (c) resist hardening, (d) template release, and (e) pattern etch.

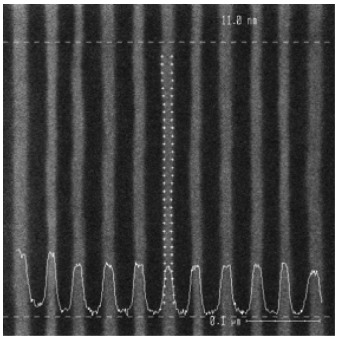


Figure 6.67 This SEM image shows the capability of NIL to form an 11-nm line pattern. (M. Malloy and L. C. Litt, *Proceedings of SPIE*).

storage almost reaches the limit of a storage cell of a nonpatterned magnetic disk, which is called media in the HDD industry. To further reduce the size of storage cells and increase storage density, cell patterning with sub-50-nm pitch is required. Because there is only one layer of HDD media, there is no overlay issue, and working-template management is also significantly simpler than multiple-layer patterning of IC processing.

6.4.8 X-ray lithography

When a wavelength is shorter than 5 nm, an electromagnetic wave becomes an x ray. An x ray has a much shorter wavelength than UV light; therefore, it can achieve higher resolution than photolithography. X-ray lithography has been in research and development since it was proposed in 1972.

Because no material can refract an x ray, and the ray can reflect at only a very small angle (~1 deg), x-ray lithography must be accomplished by a direct printing process, which is very similar to proximity printing, as shown in Fig. 6.68. The x ray passes through the transparent part of the mask and exposes the photoresist on the wafer surface. Because the wavelength is so small that the diffraction effect is almost negligible, it can achieve resolution close to that on the mask.

There are many disadvantages of x-ray lithography. For instance, UV light can be reflected by the mirror and focused by the lens, but an x ray cannot; thus, the exposure system needs to be redesigned. For a commonly used 4:1 scanner system, feature size on the reticle is four times larger than that on the wafer surface. Therefore, to pattern a 25-nm feature size on the wafer surface, feature size on the reticle must be 100 nm. Since it takes less than 100-nm chromium to block UV radiation, the pattern aspect ratio is smaller than 1:1 on the optical reticle. Because an x ray cannot be focused, the mask and wafer feature sizes must be in a 1:1 ratio. To pattern a 25-nm feature size on the wafer surface, feature size on the mask must also be 25 nm. An x-ray mask requires a layer of gold greater than 100-nm thick to block more than 90% of the x ray. Also, the pattern aspect ratio is larger than 4:1,

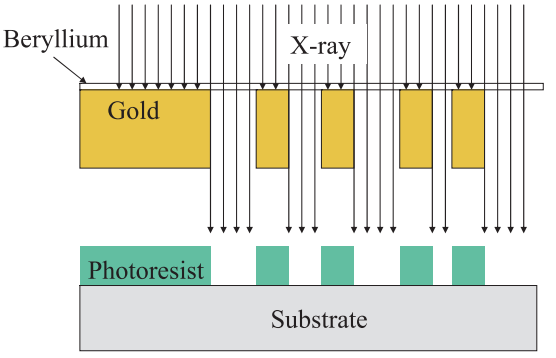


Figure 6.68 X-ray lithography process.

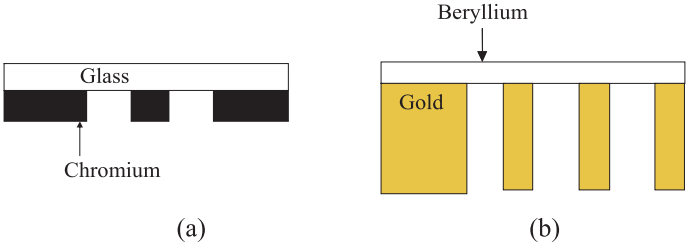


Figure 6.69 Comparison of (a) a photomask and (b) an x-ray mask.

making the x-ray mask much more difficult to produce than an optical photomask. Figure 6.69 illustrates a photomask and an x-ray mask.

Question: What is the pattern aspect ratio of an x-ray mask if that mask is intended to print 10-nm half-pitch line–space patterns on the wafers?
Answer: The aspect ratio is 100 nm:10 nm, or 10:1.

When pattern dimensions go down to the sub-10-nm region, the aspect ratio of the pattern on the x-ray mask further increases, causing it to be too high and very difficult to making the mask produce. Also, x-ray lithography requires a single-wavelength x-ray source, such as a synchrotron, which has a large footprint and is very expensive to build. X-ray lithography is no longer regarded as one of the promising candidates for NGL solutions.

6.4.9 Electron beam lithography

An electron is a very small particle and is also a wave, with its wavelength determined by its momentum, which is related to electron energy. Electron wavelength is inversely proportional to the square root of electron energy; thus, the higher the electron energy, the shorter the electron wavelength. An energetic electron beam (10 to 100 keV) has a much shorter wavelength than UV light, thus electron beam (e-beam) lithography inherently has higher spatial resolution

and a wider processing window than optical lithography. E-beam lithography is widely used in mask shops for writing design layout patterns to the resist on a mask. It is one of the most time-consuming and expensive processes in the mask manufacturing process.

An EBDW lithography system uses a fine scanning e-beam to directly write the design patterns stored in a database computer onto the electron-sensitive resist on a substrate surface, similar to a laser printer printing words or images from a computer onto a sheet of paper. Energy from the electrons changes the solubility of the resist where the electron beam hits. A positive resist becomes soluble in the development solution. Because of the nature of serial writing, throughput of EBDW is low. For a single–e-beam direct write system, as illustrated in Fig. 6.70(a), throughput will be too low to be cost effective for large-scale semiconductor wafer production. Therefore, the only option for EBDW applications in IC manufacturing is a multiple beam solution, as shown in Fig. 6.70(b).

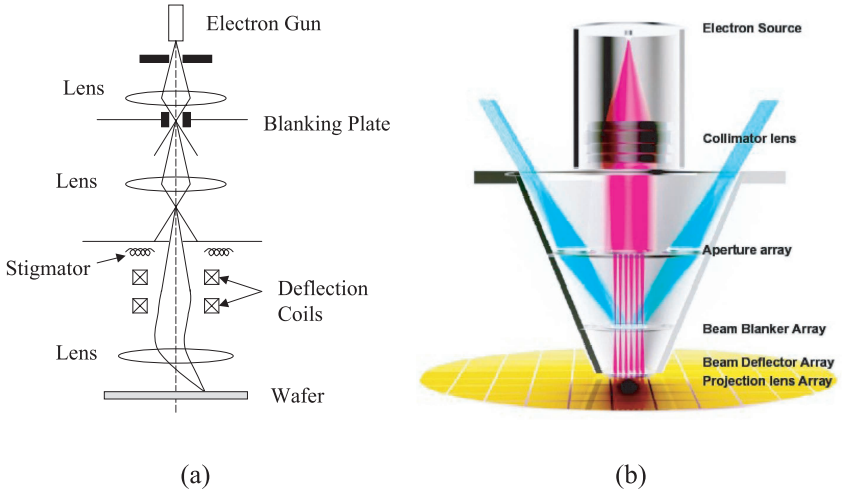


Figure 6.70 EBDW systems: (a) single beam and (b) multibeam (Mapper Lithography, B.V.).

6.4.10 Ion beam lithography

Similar to e-beam lithography, ion beam lithography also can achieve higher resolution than photolithography. An ion beam can also be used in both direct writing and projection resist exposing. One of the advantages of ion beam lithography is direct ion implantation and ion beam sputtering pattern etch, which saves some processing steps. However, throughput of ion beam lithography is very low due to the serial writing for ion beam direct writing; thus, this technology is very unlikely to be used in mass production. One application of ion beam processing is mask/reticle repairing. It can also be used for IC chip device defect detection and repair.

6.5 Safety

Safety is always one of the most important issues in semiconductor fabs. For photolithography processes, the safety concerns are mainly chemical, mechanical, and electrical.

Photolithography uses many chemicals—some of them are flammable and explosive, and some of them are corrosive and toxic. For chemicals commonly used for wet cleaning, sulfuric acid (H_2SO_4) is corrosive and causes skin burns if directly contacted; even a diluted solution can cause a skin rash. Hydrogen peroxide (H_2O_2) is a strong oxidizer, and direct contact with it can cause irritation and burns to the skin and eyes. Xylene, a commonly used solvent and developer for negative photoresists, is flammable with a flash point at 27.3 °C (about room temperature) and is explosive when its concentration reaches 1 to 7% in air. Repeated contact with xylene can cause skin irritation. Xylene vapor has a sweet, airplane-glue-like odor; exposure to it can cause eye, nose, and throat irritation, and inhaled fumes can cause headache, dizziness, loss of appetite, and fatigue. HMDS is the primer most commonly used to promote photoresist adhesion to wafer surfaces. It is flammable with a flash point at 6.7 °C and is explosive when its vapor concentration is 0.8 to 16% in air. It will vigorously react with water, alcohol, and mineral acid and release ammonia. TMAH is widely used as positive photoresist developer solution. It is poisonous and corrosive; swallowing and direct skin contact can be fatal. Contact with its dust or mist can be irritating to the eyes, skin, nose, and throat; high concentration can cause death from fluid in the lungs.

Mercury lamps supply UV light sources for i-line (365 nm) systems, which are used for large feature size patterning. Mercury is a liquid that begins to evaporate at room temperature. Mercury (Hg) vapor is highly toxic; exposure to it can cause coughing, chest pain, headache, difficulty sleeping, loss of appetite, and lung and kidney disorders. Both chlorine (Cl_2) and fluorine (F_2) are used in excimer lasers as DUV light sources. Both gases are toxic, with a greenish color and strong irritating odor. Inhaling at high concentrations can cause death.

UV light can break chemical bonds between atoms and molecules by providing energy to the binding electrons to break free from the bonds. Organic molecules are more vulnerable to damage from strong UV light because of their relatively weak bonds in comparison to the bonds of inorganic compounds. This is why UV light can be used to kill bacteria in food processing. Looking directly at a UV light source in photolithography tools can cause damage to visual cells in the eyes and result in eye injury. UV protection goggles are required when working on optical systems with UV light sources.

High-pressure mercury lamps must be carefully handled. Gloves are required during lamp replacement, since any fingerprint left on the lamp surface can cause uneven heating of the glass, possibly cracking the glass and causing an explosion.

Both mercury lamps and lasers use high-voltage electrical power, which is also a safety hazard. The power supply should always be turned off and static charges grounded before working on these parts. Tag-out and lockout is also very important to avoid a high-voltage electric power supply being turned on while people are still working on these parts.

All moving parts are potential mechanical hazards, especially robotics and slit valves.

6.6 Summary

- Photolithography is a patterning process that uses UV light to transfer a designed pattern on the mask/reticle to the photoresist temporarily coated onto the wafer surface.
- Positive photoresist becomes soluble after exposure to UV light; negative photoresist becomes insoluble due to crosslinking of a photoresist polymer. Positive photoresist is commonly used due to its higher resolution capability.
- A photoresist consists of polymer, sensitizer, solvent, and additive.
- The basic photolithography steps are: wafer cleaning, prebake and HMDS primer vapor coating, photoresist spin coating, soft bake, alignment and exposure, PEB, optical EBR (optional), development, hard bake, and pattern inspection.
- Wafer cleaning reduces contamination and improves photoresist adhesion.
- Prebake dehydrates the wafer surface, and the HMDS primer layer helps photoresist stick to the wafer surface.
- Convection ovens, infrared ovens, microwave ovens, and hot plates can be used for the baking processes. The hot plate is the most commonly used method in advanced semiconductor fabs.
- Spin coating is most commonly used for coating the photoresist.
- Photoresist thickness and uniformity are related to spin rate, spin rate ramp, photoresist temperature, wafer temperature, gas flow rate, and gas temperature.
- Soft bake drives out the majority of solvent from the photoresist and causes it to become solid.
- Overbaking during the soft bake process can polymerize the photoresist and affect its exposure sensitivity.
- Underbaking during the soft bake process can cause a blurred image due to excessive solvents and photoresist lift during etch or ion implantation processes.
- Contact printers, proximity printers, projection printers, and steppers have been used as alignment and exposure systems. A stepper has the highest resolution and is the most commonly used in advanced semiconductor fabs.
- PEB suppresses the standing wave effect due to the thermal movement of photoresist molecules.
- Developer solution dissolves the exposed positive photoresist during the development process. The development process is sensitive to temperature.
- Hard bake drives out the residue solvent from the photoresist and improves etch, implantation resist, and photoresist adhesion. Underbaking can cause

photoresist loss during the etch process. Overbaking can cause photoresist flow and affect resolution.
- Baking, coating, and developing processes are performed in-line with a track system, which is commonly integrated with a stepper.
- Shorter wavelengths have better lithography resolution than longer wavelengths. 193-nm ArF is the most commonly used wavelength of photolithography systems used in advanced IC fabs.
- The combination of immersion lithography and double- or multiple-patterning techniques extended optical lithography to the 22-nm technology node and likely will extend it to 16 nm and beyond.
- EUV lithography, EBDW lithography, and NIL are the three candidates for NGL technologies.
- EUV lithography could replace photolithography in some critical patterning processes in the near future.

6.7 Bibliography

D. G. Baldwin, M. E. Williams, and P. L. Murphy, *Chemical Safety Handbook for the Semiconductor/Electronics Industry*, 2nd ed., OME Press, Beverly, MA, 1996.

R. Bowman, G. Fry, J. Griffin, D. Potter, and R. Skinner, *Practical VLSI Fabrication for the 90s,* Integrated Circuit Engineering Corp., Scottsdale, AZ (1990).

T. H. P. Chang, D. P. Kern, and L. P. Muray, "Arrayed miniature electron beam columns for high throughput sub-100 nm lithography," *J. Vac. Sci. Technol.* **10**(6), 2743–2748 (1992).

L. de Broglie, "Recherches sur la théorie des quanta (Research on the Quantum Theory)," PhD Thesis, Paris Univ. (1924).

B. Dinu, S. Fuch, U. Kramer, M. Kubis, A. Marchelli, A. Navarra, C. Sparka, and A. Widmann, "Overlay control using scatterometry based metrology (SCOM^{TM}) in production environment," *Proc. SPIE* **6922**, 69222S-1 (2008) [doi:10.1117/12.771581].

H. S. Kim, B. H. Lee, E. Ma, F. Wang, Y. Zhao, K. Kenai, H. Xiao, and J. Jau, "After development inspection defectivity studies of an advanced memory device," *Proc. SPIE* **7638**, 76380L (2010) [doi:10.1117/12.848066].

K. Lai, A. E. Rosenbluth, S. Bagheri, J. Hoffnagle, K. Tian, D. Melville, J. Tirapu-Azpiroz, M. Fakhry, Y. Kim, S. Halle, G. McIntyre, A. Wagner, G. Burr, M. Burkhardt, D. Corliss, E. Gallagher, T. Faure, M. Hibbs, D. Flagello, J. Zimmermann, B. Kneer, F. Rohmund, F. Hartung, C. Hennerkes, M. Maul, R. Kazinczi, A. Engelen, R. Carpaij, R. Groenendijk, and J. Hageman, "Experimental result and simulation analysis for the use of pixelated illumination from source mask optimization for 22-nm logic lithography process," *Proc. SPIE* **7274**, 72740A-1 (2009) [doi:10.1117/12.814680].

M. D. Levenson, N. S. Viswanathan, and R. A. Simpson, "Improving resolution in photolithography with a phase-shifting mask," *IEEE Trans. Electron Dev.* **ED-29**(12), 1812–1846 (1982).

J. A. Liddlea, Y. Cui and P. Alivisatos, "Lithographically directed self-assembly of nanostructures," *J. Vac. Sci. Technol.* **22**(6), 3409–3414 (2004).

B. J. Lin, "The k_3 coefficient in nonparaxial 1/NA sealing equations for resolution, depth of focus, and immersion lithography," *J. Micro/Nanolith. MEMS MOEMS* **1**(1), 7–12 (2002) [doi:10.1117/1.1445798].

M. Malloy and L. C. Litt, "Step and flash imprint lithography for semiconductor high volume manufacturing?" *Proc. SPIE* **7637**, 763706-1 (2010) [doi:10.1117/12.846617].

J. A. McClay and A. S. L. McIntyre, "157-nm optical lithography: the accomplishments and the challenges," *Solid-State Tech.* **42**(6), 57–68 (1999).

M. Rothschild, M. W. Horn, C. L. Keast, R. R. Kunz, V. Liberman, S. C. Palmateer, S. P. Doran, A. R. Forte, R. B. Goodman, J. H. C. Sedlacek, R. S. Uttaro, D. Corliss, and A. Grenville, "Photolithography at 193 nm," *Lincoln Lab. J.* **10**(1), 19–34 (1997).

A. H. Shih, *Scatterometry-based critical dimension and profile metrology*, http://www.eetasia.com/ART_8800271012_480200_TA_7589d612.HTM (last accessed on 05/25/2010).

S. M. Sze, *VLSI Technology*, 2nd ed., McGraw-Hill, New York (1988).

P. van Zant, *Microchip Fabrication, a Practical Guide to Semiconductor Processes*, 3rd ed., McGraw-Hill, New York, 1997.

See http://www.semiconductorblog.com/wp-content/uploads/2010/03/IMFT25nmWLSEM.jpg (last accessed April 2010).

6.8 Review questions

1. What is photolithography?
2. What is the difference between positive and negative photoresists?
3. List the four components of photoresist and explain their functions.
4. List the steps of the photolithography process sequence.
5. Why do wafers need to be cleaned before photoresist coating?
6. What is the purpose of prebake and primer coating?
7. List two methods for primer coating. Which one is used the most in advanced IC fabs? Why?
8. What factors can affect photoresist spin coating thickness and uniformity?
9. What is the purpose of soft bake? List the consequences of over- and underbaking.

10. List the four exposure methods. Which one has the highest resolution?
11. What are two main factors that control the exposure process?
12. Explain the purpose of PEB. What can go wrong with overbaking and underbaking in PEB?
13. Name the three steps of development.
14. Explain the purpose of hard bake. What can go wrong if the photoresist is underbaked or overbaked?
15. What are the two processes that follow the photolithography process?
16. Why does a wafer need parametric measurements and defect inspection before it is sent to the next processing step?
17. Explain why a high-intensity, short-wavelength light source is sought out.
18. Why does subquarter-micron IC manufacturing need a CMP process?
19. Explain how the immersion technique improves photolithography resolution.
20. An engineer working on double-patterning process development stated: "Lithography has a limit, patterning does not!" Do you agree with this statement? Explain your answer.
21. List at least two alternative lithography technologies that might replace photolithography in the near future.
22. In your opinion, which method is the most promising candidate for NGL?

Chapter 7
Plasma Basics

Objectives

After finishing this chapter, the reader will be able to:

- define plasma
- list the three basic components of plasma
- list the three key collisions in plasma and identify their importance
- list the benefits of using plasmas in CVD and etch processes
- identify major differences between plasma-enhanced CVD and plasma etch processes
- name at least two high-density plasma systems
- define and explain the mean free path and its relationship to pressure
- define and explain the effect of magnetic fields on plasma
- describe ion bombardment and its relationship to plasma processes.

7.1 Introduction

Plasma processes are widely used in semiconductor processing. For example, all patterned etching in IC fabrication is either a plasma etch or dry etch process. Plasma-enhanced CVD (PECVD) and high-density plasma CVD (HDP-CVD) processes are widely used for dielectric depositions. Ion implanters use plasma sources to generate ions for wafer doping and to provide electrons to neutralize ions on wafer surfaces. Plasmas are also used in physical vapor deposition (PVD) processes, in which metal is sputtered from a target surface by ion bombardment and deposited on a wafer surface. Remote plasma source systems are commonly used in process chamber cleaning, film stripping, and film deposition processes. This chapter covers the basic properties of plasma and its applications in semiconductor processes, focusing mainly on etch and CVD.

7.2 Definition of Plasma

Plasma is widely defined in the semiconductor industry as an ionized gas with equal numbers of positive and negative charges. More precisely, plasma is a quasi-neutral gas of charged and neutral particles that exhibit a collective behavior. Readers can find more information about plasma in the References section at the end of this chapter.

7.2.1 Components of plasma

Plasma consists of neutral atoms or molecules, negative charges (electrons), and positive charges (ions). In a plasma, the electron concentration is about equal to the ion concentration, $n_e = n_i$, which keeps the plasma quasi-neutral.

The ratio of electron concentration to the total concentration is defined as the ionization rate:

$$\text{ionization rate} = n_e/(n_e + n_n),$$

where n_e is electron concentration, n_i is ion concentration, and n_n is neutral concentration. Ionization rate is mainly determined by electron energy; it is also related to the gas species, since different gases require different ionization energies. For example, the sun is a big ball of plasma. At the edge of the sun, ionization rate is very low, or $n_e \ll n_n$, due to the relatively low temperature (about 6000 °C). At the center of the sun, the temperature is so high (~10,000,000 °C) that almost all gas molecules are ionized. Thus, at $n_n \ll n_e$, the ionization rate is almost 100%.

Plasmas used in semiconductor processing have very low ionization rates. For example, the ionization rate of plasma generated in a PECVD reactor with two parallel plate electrodes is about one in one million to one in ten million, or less than 0.0001%. In a plasma etch chamber with two parallel plate electrodes, the ionization rate is a little higher. About one in ten thousand atoms or molecules are ionized, meaning that the ionization rate is less than 0.01%. Even in the two most commonly used high-density plasma sources, inductively coupled plasma (ICP) or electron cyclotron resonance (ECR), ionization rates are still very low, usually from 1 to 5%. High-density plasma sources with almost 100% ionization rates are still in research and development and not used in IC production yet.

The ionization rate of a plasma reactor is mainly determined by electron energy, which is controlled by the applied power. It is also related to pressure, electrode spacing, and the processing gas species.

7.2.2 Generation of plasma

An external energy source must be applied in order to generate plasma. There are several ways to generate plasma in semiconductor processes. A hot filament with potential dc bias is commonly used to generate plasma in ion source and plasma flooding systems in ion implanters. Most PVD systems use a dc power supply to generate plasma. The most commonly used plasma sources in semiconductor processing are rf plasma sources, which are the main topic of this chapter.

In most PECVD and plasma etch chambers, plasmas are generated by rf power applied between two parallel plate electrodes in a vacuum chamber, as shown in Fig. 7.1. This is called a capacitively coupled plasma source, since the two parallel electrodes are exactly like the electrodes of a capacitor.

When rf power is applied to the electrodes, a varying electric field is established between them. If the rf power is high enough, a free electron can be accelerated by the varying electric field until it gains enough energy to collide with an atom or

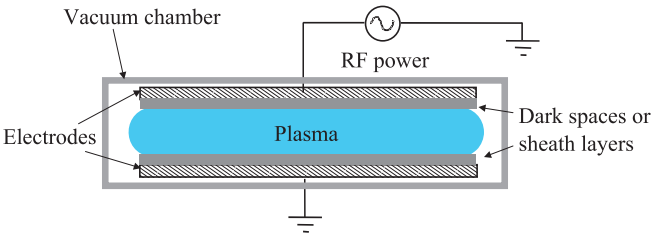


Figure 7.1 Schematic of a capacitively coupled plasma source.

molecule inside the chamber, generating an ion and another free electron. Due to the cascading of ionization collisions, the entire chamber fills quickly with equal numbers of electrons and ions—this is plasma.

In plasma, some electrons and ions are continually lost and consumed by collisions with electrodes and chamber walls, and also by recombination collisions between electrons and ions. When the generation rate of electrons through ionization collisions is equal to the loss rate of electrons, the plasma is said to be stabilized.

Other plasma sources such as dc plasma, ICP, ECR, and microwave (MW) remote plasma sources all have very similar plasma generating processes.

Question: Without the first free electron inside a chamber, there is no way to start plasma. Where does the first free electron come from?
Answer: It is usually generated by a cosmic ray. It also can be generated by heat (thermal electrons) or by native radioactive decay.

7.3 Collisions in Plasma

There are two kinds of collisions inside plasma: elastic and inelastic. Elastic collisions are the most frequent collisions in plasma. Since there is no energy exchange between colliding particles in an elastic collision, it is not important. There are many kinds of inelastic collisions happening simultaneously in plasma: collisions between electrons and neutrals, collisions between ions and neutrals, collisions between ions and ions, and collisions between electrons and ions. In plasma, any possible collision imaginable can happen. Since different collisions have varying probabilities of occurrence, the importance of each type of collision is not the same. For plasmas used in semiconductor processing, three collisions are very important: ionization, excitation-relaxation, and dissociation collisions.

7.3.1 Ionization

When an electron collides with an atom or molecule, it can transfer part of its energy to the orbital electron confined by the nucleus of that atom or molecule. If the orbital electron gains enough energy to break free from the constraint of the nucleus, it becomes a free electron (see Fig. 7.2). This process is called electron

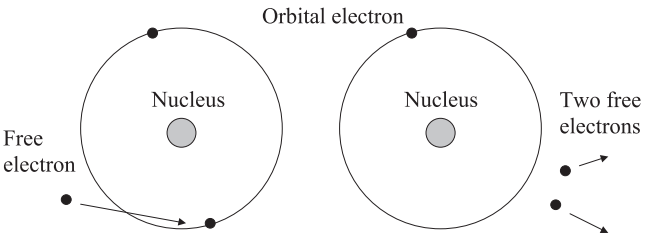


Figure 7.2 Ionization collision before and after electron impact.

impact ionization. Ionization collision can be expressed as

$$e^- + A \rightarrow A^+ + 2e^-,$$

where e^- represents an electron with a negative charge, A represents a neutral atom or molecule, and A^+ represents a positive ion. Ionization is very important because it generates and sustains the plasma.

7.3.2 Excitation-relaxation

Sometimes an orbital electron does not get enough energy from the impact electron to break free from the constraint of the nucleus. The transferred energy to the orbital electron from the collision will cause it to jump to a higher energy level of the orbit. This process is called excitation. An excitation collision can be expressed as

$$e^- + A \rightarrow A^* + e^-,$$

where A^* is the excited state of atom or molecule A, indicating that it has an electron in the higher energy level orbit.

The excited state is not stable and has a short lifetime. An electron in an excited orbit does not stay at that higher energy level for very long and falls back to the orbit with the lowest possible energy level, or ground state. This process is called relaxation. The excited atom or molecule quickly relaxes back to its ground state and releases the extra energy it gained from the electron impact in the form of a photon, which is the light emission. The relaxation process can be expressed as

$$A^* \rightarrow A + h\nu \text{ (photons)},$$

where $h\nu$ is the energy of the photon, h is Planck's constant, and ν is the frequency of the light emission, which determines the color of light emitted from the plasma. Different atoms or molecules have different orbital structures and energy levels; therefore, the light emission frequencies are different. This is why different gases glow in various colors within the plasma. The glow of oxygen is grayish-blue, nitrogen is pink, neon light is red, a fluorine glow is orange-red, etc.

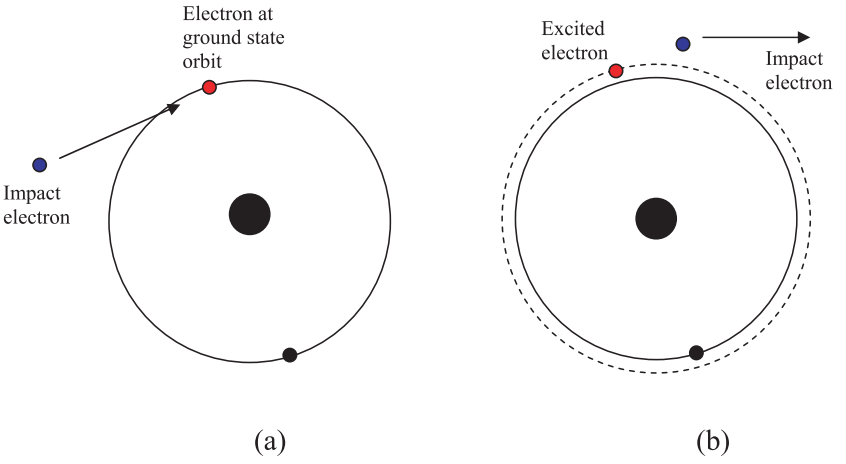


Figure 7.3 (a) Before and (b) after excitation collision.

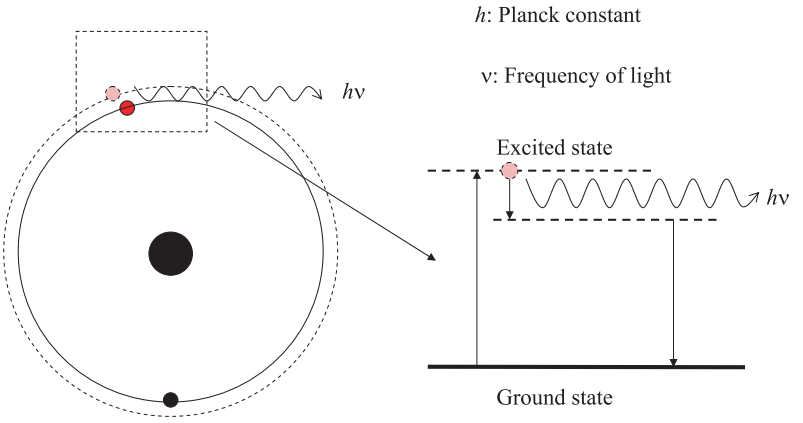


Figure 7.4 Relaxation process.

The excitation-relaxation process is illustrated in Figs. 7.3 and 7.4. Detection of a change in light emission is widely used in semiconductor processes to determine the endpoints of both etch and CVD chamber cleaning processes. This is covered in greater detail in Chapters 9 and 10.

7.3.3 Dissociation

When an electron collides with a molecule, it can break the chemical bond and generate free radicals if the energy transferred by the impact to the molecule is higher than the molecule's bonding energy. This is called a dissociation collision

and can be expressed as

$$e^- + AB \rightarrow A + B + e^-,$$

where AB is a molecule, and A and B are free radicals generated by a dissociation collision. Free radicals are molecular fragments with at least one unpaired electron. This makes them chemically very reactive, since they have a strong tendency to grab an electron from other atoms or molecules to form a stable molecule. Free radicals can enhance chemical reaction in both etch and CVD processes. Figure 7.5 illustrates the dissociation collision process.

For example, in both oxide etch and CVD chamber cleaning processes, fluorocarbon gas such as CF_4 is used in plasma to generate free fluorine radical F:

$$e^- + CF_4 \rightarrow CF_3 + F + e^-.$$

In the PECVD oxide process, silicon precursor silane (SiH_4) and oxygen precursor nitrous oxide (N_2O, also known as laughing gas) are used to generate free radicals:

$$e^- + SiH_4 \rightarrow SiH_2 + 2H + e^-,$$
$$e^- + N_2O \rightarrow N_2 + O + e^-.$$

Free radicals such as F, SiH_2, and O are chemically very reactive. This is why plasma can enhance chemical reactions in both CVD and etch processes.

Question: Why are dissociation collisions unimportant in aluminum and copper sputtering processes?

Answer: Aluminum and copper sputtering processes only use argon, which is a noble gas. Unlike other gases, noble gases exist in the form of atoms instead of molecules. Thus, there is no dissociation process in argon plasma.

Question: Is there any dissociation collision in plasma sputtering PVD processes?

Answer: Yes. Both argon (Ar) and nitrogen (N_2) are used in titanium nitride (TiN) sputtering deposition processes. In plasma, N_2 is dissociated to generate a free radical N, which reacts with titanium or tantalum to from TiN

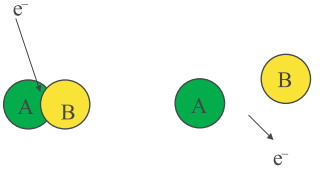


Figure 7.5 Dissociation collision.

on the target surface. Ar^+ ions sputter TiN molecules from the target surface and deposit them on a wafer surface. Sputtering deposition of tantalum nitride (TaN) is very similar to the deposition of TiN.

7.3.4 Other collisions

Other collisions inside plasma, such as recombination, charge-exchange, pitch-angle scattering, and neutral-neutral collisions, etc., are not important in PECVD and plasma etch processes. Some collisions are the combination of two or more different kinds of collisions.

Table 7.1 shows possible collision processes in PECVD silane plasma. It can be seen that some are dissociation, some are a combination of dissociation and excitation, and some are a combination of dissociation and ionization collisions.

Question: Which one of the collisions in Table 7.1 is most likely to happen? Why?

Answer: The collision that requires the least energy is the most likely to happen. It is much easier for an electron to fall to a lower energy level than climb to a higher one. Under conditions of field strength, pressure, and temperature, an electron is more likely to be accelerated for a short distance without a collision to attain enough energy required for that first collision to occur (2.2 eV). It is least likely for an electron to travel a much longer distance without a collision to gain enough energy (15.3 eV) needed for the last reaction in Table 7.1.

7.4 Plasma Parameters

The main plasma parameters are mean free path (MFP), thermal velocity, magnetic field, and the Boltzmann distribution.

Table 7.1 Possible plasma collision processes.

Collisions	Byproducts	Energy of formation
$e^- + SiH_4 \rightarrow$	$SiH_2 + H_2 + e^-$	2.2 eV
	$SiH_3 + H + e^-$	4.0 eV
	$Si + 2H_2 + e^-$	4.2 eV
	$SiH + H_2 + H + e^-$	5.7 eV
	$SiH_2^* + 2H + e^-$	8.9 eV
	$Si^* + 2H_2 + e^-$	9.5 eV
	$SiH_2^+ + H_2 + 2e^-$	11.9 eV
	$SiH_3^+ + H + 2e^-$	12.32 eV
	$Si^+ + 2H_2 + 2e^-$	13.6 eV
	$SiH^+ + H_2 + H + 2e^-$	15.3 eV

7.4.1 Mean free path

The MFP is defined as the average distance a particle can travel before it collides with another particle. MFP (or λ) can be expressed by the following equation:

$$\lambda = \frac{1}{\sqrt{2} n \sigma}, \tag{7.1}$$

where n is particle density, and σ is the collision cross section. Higher particle density causes more collisions, which shorten the MFP. Large particles collide more easily with other particles and therefore have shorter MFPs. As shown in Eq. (7.1), MFP is determined by chamber pressure, which determines particle density. Gas (or gases) in the chamber can also affect MFP because different gas molecules have different sizes or cross-sectional areas.

Question: What is the MFP for a molecule with diameter 3 Å and density 3.5×10^{16} cm^{-3} (the density of an ideal gas at 1 torr or 1 mm Hg)?
Answer: Based on Eq. (7.1), $\lambda = 1/[\sqrt{2} \times \overline{3.5} \times 10^{16} \times \pi \times (3 \times 10^{-8}/2)^2] = 0.029$ cm.

Figure 7.6 shows that the MFP is shorter when the gas density is higher [Fig. 7.6(a)] and longer when gas density is lower [Fig. 7.6(b)]. Large particles have a larger cross section and sweep through more space. Therefore, they are more likely to collide with other particles and have shorter MFPs than average sized or smaller particles. Changing pressure changes particle density, therefore affecting MFP:

$$\lambda \propto 1/p.$$

As pressure reduces, the MFP increases; as pressure decreases, particle density reduces, and therefore the frequency of collision decreases. An approximation of the MFP of a gas molecule in air is

$$\text{MFP (cm)} \approx 50/p \text{ (mtorr)}.$$

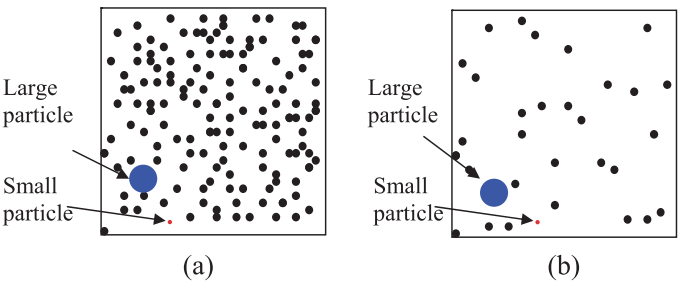


Figure 7.6 (a) High-pressure short MFP and (b) low-pressure long MFP.

For an electron, the MFP can be twice as long, since it has a much smaller size:

$$\lambda_e \text{ (cm)} \approx 100/p \text{ (mtorr)}.$$

PECVD processes normally operate at 1 to 10 torr; therefore, in PECVD chambers the electron MFP λ_e = 0.01 to 0.1 cm. Etch process pressures are much lower, from 3 mtorr to 300 mtorr. Therefore, in an etch chamber, electron MFP λ_e varies from 0.33 to 33 cm.

The MFP is one of the most important concepts of plasma. It can be controlled by chamber pressure, and it can affect processing results, sometimes significantly, especially for plasma etch processes. When pressure in a plasma processing chamber changes, the MFP changes. Likewise, ion bombardment energy and ion direction change, which can affect rate and profile in etch processes, and film stress in PECVD processes. Plasma shape will also change due to electron MFP changes. At higher pressure, plasma is more concentrated near the electrodes, while at lower pressure, plasma can spread throughout the chamber. Pressure can affect plasma uniformity and change the etch rate or deposition rate across a wafer.

Question: Why is a vacuum chamber necessary to generate stable plasma?

Answer: At atmospheric pressure (760 torr or 760 mm Hg), the MFP of an electron is very short. It is very difficult for an electron to acquire enough energy to ionize gas unless it is in an extremely strong electric field. This creates plasma in the form of arcing (lightning) instead of a steady-state glow discharge.

7.4.2 Thermal velocity

In plasma, electrons, ions, and neutrals constantly move due to external electric power and thermal movement. Electrons have the lightest weight and smallest size; therefore, they are more likely to pick up energy from an external power supply compared to ions and neutrals. In plasma, electrons always move faster than ions and neutrals.

The mass ratio of an electron compared to the mass of the lightest hydrogen ion is 1 to 1836. In PECVD, etch, and PVD processes, the most commonly used ions are oxygen, argon, chlorine, and fluorine, all of which are all significantly heavier than a hydrogen ion. Therefore, these ions are much heavier than an electron by at least a factor of 10,000. However, the electric force that provides energy to the plasma is the same for both electrons and ions, since it is only related to electrical charge and electric field:

$$F = qE,$$

where F is the force on a charged particle, q is the electrical charge (electron is negative and ion is positive), and E is the electric field generated by an external

power supply such as rf, dc, or microwave. The acceleration of the charged particle can be expressed as

$$a = F/m = qE/m,$$

where m is the mass of the charged particle. Since the mass of electrons is less than a factor of 1/10,000 compared with ions, electrons can be accelerated more than 10,000 times faster than ions! This is similar to a motorcycle, which can accelerate much faster than an 18-wheel truck. It is not very hard to imagine how fast a motorcycle would start up using a powerful engine from an 18 wheeler, or to visualize how slow an 18 wheeler powered by a motorcycle engine would creep away from a traffic stop.

Most etch and CVD plasma sources use rf power. Radio frequency power generates a varying electric field, which changes directions rapidly. Electrons can be accelerated rapidly and start collision processes such as ionization, excitation, and dissociation in the positive cycle of an rf field, and repeat those processes in the negative cycle. Because ions are too heavy to respond quickly enough to the varying electric field, electrons absorb most of the rf energy due to their light weight and quick response. This process is very similar to two vehicles traveling on a road with stop signs at every intersection. The motorcycle can start up and slow down very quickly. The big truck starts very slowly and has to slow down gradually. It is easy to see that the average speed of the motorcycle would be much higher than the big truck on this kind of road.

When subjected to a lower rf power, ions pick up a little more energy than they would with a higher rf power. Lower rf power allows more time for ions to respond; therefore, it can accelerate ions to a higher energy and provide more energy for ion bombardment.

In any case, electrons always have a much higher temperature than ions or neutrals in plasma. Thermal velocity can be expressed as

$$v = (kT/m)^{1/2}, \tag{7.2}$$

where $k = 1.38 \times 10^{-23}$ J/K is the Boltzmann constant, T is temperature, and m is the mass of the particle. For plasma generated by rf power within two parallel electrodes (or capacitively coupled plasma), electron temperature T_e is about 2 eV, where one electron volt (1 eV) is equivalent to 11,594 K or 11,321 °C. Based on Eq. (7.2), electron thermal velocity can be calculated as

$$v_e \approx 4.19 \times 10^7 T_e^{1/2} \ (T \text{ in } eV) \approx 5.93 \times 10^7 \text{ cm/sec} = 1.33 \times 10^7 \text{ mph}.$$

An electron in plasma moves faster than an orbiting space shuttle.

In contrast, the thermal velocity of an argon ion (Ar^+) with $T_{Ar} \approx 0.05$ eV is $v_i = 3.46 \times 10^4$ cm/sec = 774 mph. The ions move at the speed of an airplane, much slower than the electrons.

7.4.3 Magnetic field

In a magnetic field, a charged particle experiences a magnetic force that can be expressed as

$$F = qv \times B, \tag{7.3}$$

where q is the charge of the particle, v is the velocity of the particle, and B is the magnetic field strength. Since a magnetic force is always perpendicular to particle velocity, a charged particle will spiral around the magnetic field line. This movement is called gyromotion.

The gyromotion of a charged particle in a magnetic field (see Fig. 7.7) is a very important property of plasma and has many applications in semiconductor processing. A number of capacitively coupled plasma etch chambers have magnetic coils to generate a magnetic field and cause electron gyromotion. This helps to generate and maintain a higher-density plasma at low pressure. One of the most commonly used high-density plasma sources is the ECR plasma source, which uses a magnetic field and microwave power source. When the microwave frequency is equal to the electron gyrating frequency, the microwave power resonantly couples to the electrons to generate high-density plasma at very low pressure.

The ion implanter is another processing tool that uses magnets. The strong magnetic field generated by a dc current in the analyzer magnetic coils of an ion implanter can bend the trajectories of the high-energy ions. Because ions with different charge/mass (q/m) ratios have different trajectories in a magnetic field, they exit the magnetic field at different places. This allows fabs to precisely select the desired ions and terminate the undesired ones.

The frequency of a charged particle cycling around a magnetic field line is called the gyrofrequency Ω and can be expressed as

$$\Omega = \frac{qB}{m}. \tag{7.4}$$

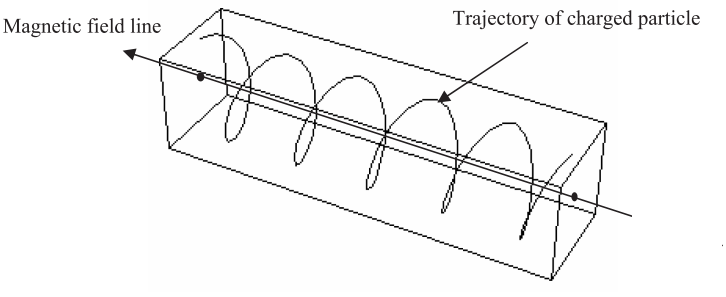


Figure 7.7 Gyromotion of a charged particle in a magnetic field.

For a charged particle with a fixed charge and specific mass, the gyrofrequency is determined only by magnetic field strength B. The gyrofrequency for an electron is Ω_e (MHz) = $2.8B$ (gauss or G).

The radius of the gyration is called the gyroradius ρ and can be expressed as

$$\rho = v_\perp/\Omega,$$

where $v_\perp$ is the particle speed perpendicular to the magnetic field line. For an electron, the gyroradius is ρ_e (cm) = $2.38T_e^{1/2}/B$, where T_e is the electron temperature expressed in electron volts eV, and B is in gauss. Similarly, the ion gyroradius can be expressed by ρ_i (cm) = $102\ (AT_i)^{1/2}/ZB$. Here, A is the atomic weight of the ion, and Z is the number of the ionization charge of the ion; both values are integer numbers. Ion mass $m_{ion} = Am_p$, where m_p is the mass of the proton, which is equal to 1.67×10^{-27} g. Ion charge q = Ze, where $e = 1.6 \times 10^{-19}$ C is the charge of an electron.

Question: In an argon sputter etch chamber, the electron temperature is $T_e \approx 2$ eV, argon ion temperature is $T_i \approx 0.05$ eV, and a magnetic field $B = 100$ G. For an argon ion, $A = 40$ and $Z = 1$. What are the electron and ion gyroradii?

Answer: Electron gyroradius $\rho_e = 2.38 \times 2^{1/2}/100 = 0.034$ cm; and argon ion (Ar^+) gyroradius $\rho_i = 102 \times (40 \times 0.05)^{1/2}/100 = 1.44$ cm.

Question: In an ion implanter with an analyzer magnetic field $B = 2000$ G and argon ion (Ar^+) energy $E_{Ar} = 200$ keV, what is its gyroradius?

Answer: $\rho_i = 102 \times (40 \times 200,000)^{1/2}/2000 = 144$ cm.

7.4.4 Boltzmann distribution

In thermal equilibrium plasma, both electron and ion energy follow the Boltzmann distribution, as shown in Fig. 7.8. The average electron energy in a capacitively coupled plasma source is about 2 to 3 eV. The ion energy in bulk plasma is mainly determined by the chamber temperature, ~200 to 400 °C, or 0.04 to 0.06 eV.

Figure 7.8 demonstrates that most electrons have energy around the average value, 2 to 3 eV. Very few electrons have energy high enough for ionization, which is approximately 15 eV. This explains why the ionization rate is so low in parallel plate plasma sources.

Question: Assume that the electron temperature in a plasma source averages 1 keV (~11,600,000 °C, the temperature in the core of the sun). What is the ionization rate of this plasma?

Answer: Almost 100%.

7.5 Ion Bombardment

Since electrons move much faster than ions, anything close to the plasma, including the chamber wall and electrodes, will be charged negatively as soon as the plasma

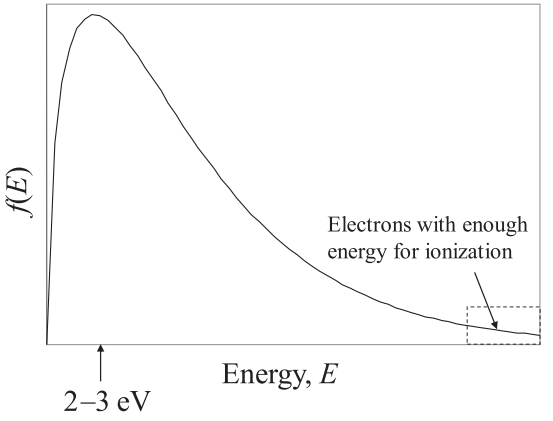


Figure 7.8 Electron energy distribution.

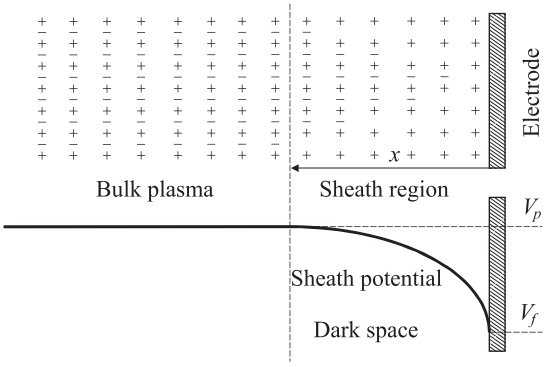


Figure 7.9 Sheath potential nears the plasma surface.

is initiated. The negatively charged electrodes will repel negative electrons and attract positive ions; therefore, in the vicinity of the electrodes, there are fewer electrons than ions.

The difference between the negative and positive charges creates an electric field in that region, called the sheath potential, as illustrated in Fig. 7.9. Light emission from this region is less intense than from the bulk plasma, since fewer electrons are present and fewer excitation-relaxation collisions occur; thus, a dark space can be observed near the electrodes.

The sheath potential accelerates ions toward the electrodes, causing ion bombardment. If a wafer is placed over an electrode, ions accelerated by the sheath potential will bombard its surface.

Ion bombardment is a very important property of plasma. Any conducting materials close to plasma will be subject to ion bombardment. Ion bombardment

can affect etch rate, selectivity, and profiles in plasma etch processes. It also can affect deposition rate and film stress in PECVD and HDP-CVD processes.

Ion bombardment has two parameters. One is ion energy, the other is ion flux. Ion energy is related to external power supply, chamber pressure, electrode spacing, and processing gases. Ion flux is related to plasma density; it is also determined by external power supply, chamber pressure, electrode spacing, and processing gases.

In rf plasma systems, rf frequency also affects ion energy. At a high frequency, such as 13.56 MHz, electrons pick up the majority of the energy, and ions are left "cold." At a much lower frequency, such as 350 kHz, although electrons still acquire most of the energy, ions have a better chance to pick up some energy from the rf power due to the slower change rate of the electric field. By using the earlier analogy of the two vehicles, this would be akin to increasing the distance between stop signs from one every block to one every kilometer. In this case, the motorcycle (electron) still has the advantage at an average speed because of its fast acceleration and quick stops, but the average speed of the truck (ion) will be significantly increased because this scenario enables the big truck to reach and sustain a high speed.

Question: Why is 13.56 MHz the most commonly used frequency for rf systems?
Answer: Because the government regulates applications of rfs under an international treaty, it is necessary to regulate the applications of rfs to avoid interference with different applications. If some rf interferes with an air traffic control radio signal, it could have disastrous consequences. Industrial manufacturing, medical applications, and scientific research have 13.56 MHz assigned to them. Such rf generators have been commercially available for a long time and are more cost effective than other rf sources such as 2 MHz, 1.8 MHz, etc.

7.6 Direct-Current Bias

In an rf system, the potential of an rf hot electrode changes rapidly. Plasma potential also changes very quickly. Because electrons move faster than ions, anything close to the plasma will be charged negatively; plasma always has a higher potential than anything close to it.

A plasma potential curve (solid curve) is shown in Fig. 7.10. When the rf potential (dashed curve) is in its positive cycle, plasma potential is higher than the rf hot electrode potential. When the rf potential goes negative, plasma potential will not follow it to the negative side. It must maintain a higher potential than the ground potential. When the rf potential returns to positive again, plasma potential rises again. Thus, the plasma potential cycles entirely in a range higher than the ground potential. Therefore, on average, there is a dc potential difference maintained between bulk plasma and electrodes. This difference is called the dc bias. The energy of a bombarding ion is determined by the dc bias, which is about 10 to 20 V in a PECVD chamber with parallel plate electrodes. The dc bias mainly depends on rf power; it is also related to the chamber pressure and processing gas

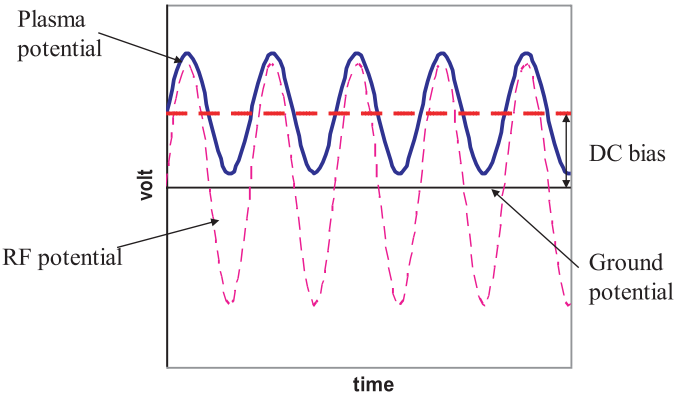


Figure 7.10 dc bias and its relation to rf power.

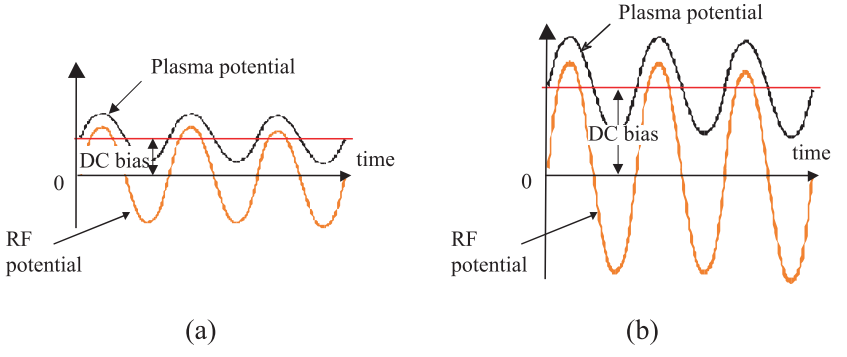


Figure 7.11 dc biases with (a) lower and (b) higher rf power.

species. When rf power increases, the amplitude of the rf potential will increase correspondingly, and the plasma potential and dc bias increases (see Fig. 7.11).

Plasma potential depends on the rf power, pressure, and spacing between the electrodes. For a symmetrical system with two electrodes having the same area (shown in Fig. 7.12), the dc bias is about 10 to 20 V. Most PECVD systems have this type of structure. Since rf power also affects plasma density, a capacitively coupled (parallel plate) plasma source cannot independently control ion energy and flux.

Figure 7.13 shows the voltage in an asymmetric electrode plasma source. In this case, the two electrodes have different areas. Current continuity induces a self-bias when the negative bias voltage builds on the smaller electrode. Sheath voltages depend on the ratio of the electrode areas, as shown in Fig. 7.14. In an ideal case (no collisions in the sheath region), voltages and electrode areas have the following relation:

$$V_1/V_2 = (A_2/A_1)^4, \tag{7.5}$$

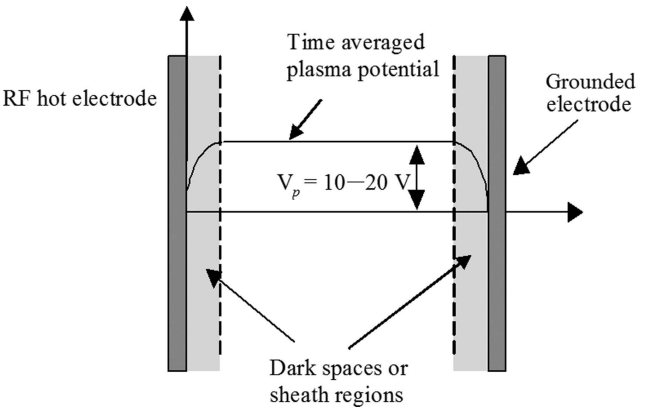


Figure 7.12 Plasma potential in a symmetric electrode system.

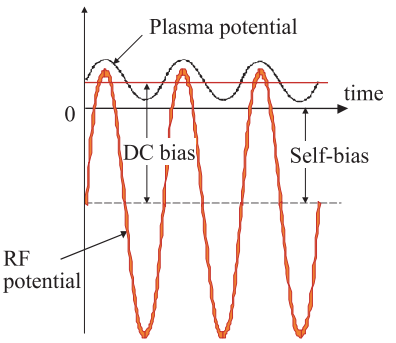


Figure 7.13 dc potential in an asymmetric electrode rf system.

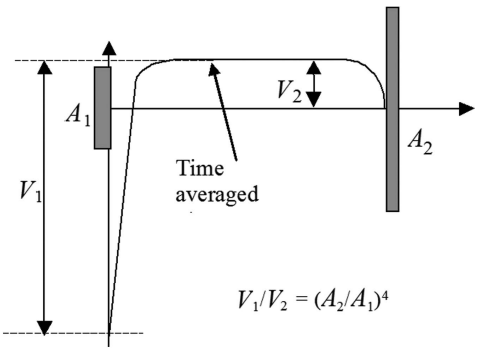


Figure 7.14 Plasma potential in an asymmetric electrode system.

where V_1 is the dc bias, which is the potential difference between the bulk plasma and the rf hot electrode; and V_2 is the plasma potential (plasma to ground). The self-bias equals $V_1 - V_2$ (rf hot to ground). The smaller electrode has a larger sheath voltage and thus receives more energetic ion bombardment.

Most etch chambers use asymmetric electrodes, and a wafer is always placed on the smaller rf hot electrode to receive more energetic ion bombardment during the entire rf cycle. Since etch processes require more ion bombardment, self-bias on the smaller electrode can help increase ion bombardment energy. For an rf system with symmetric parallel electrodes, the amount of ion bombardment on both electrodes is about the same. Therefore, if the wafer receives heavy ion bombardment, the other electrode will receive the same amount of bombardment, which is very undesirable, as this can cause particle contamination and short lifetimes of the processing chamber parts.

Question: If an electrode area ratio is 1:3, what is the difference between the dc bias and the self-bias, compared to the dc bias?

Answer: The dc bias is V_1, and the self-bias is $V_1 - V_2$; therefore, the difference is:

$$[V_1 - (V_1 - V_2)]/V_1 = V_2/V_1 = (A_1/A_2)^4 = (1/3)^4 = 1/81 = 1.23\%.$$

In this case, dc bias V_1 is significantly larger than the plasma potential V_2, and it is very close to the self-bias. Many people measure the potential difference between the rf hot electrode and the grounded electrode and call the result the dc bias, when in fact it is the self-bias. Of course, they are close enough to ignore plasma potential V_2, which is very hard to measure. Ion bombardment energy is determined by dc bias.

Question: Can a fine metal probe be inserted into the plasma to measure plasma potential V_2?

Answer: Yes, it can. However, the metal probe will be negatively charged by the fast-moving electrons and form a sheath potential between its surface and the bulk plasma when it comes close to the plasma. Therefore, measurement results are determined by theoretical models of the sheath potential, which have not been fully developed yet.

7.7 Advantage of Plasma Processes

Several qualities of plasma are very helpful in semiconductor processes. Ion bombardment is crucial for sputtering deposition, etch, and CVD film stress control. Free radicals from disassociated electron molecules greatly enhance chemical reaction rates in both CVD and etch processes. The glows emitted from plasma due to their excitation and relaxation reliably measure optical endpoints for plasma etch and plasma chamber cleaning processes.

7.7.1 Chemical vapor deposition

The major benefits of using plasma in CVD processes are:

- high deposition rates at relatively lower temperatures
- control of deposited film stress with ion bombardment
- process-chamber dry cleaning with fluorine-based plasma
- excellent gap-fill capabilities with a high-density plasma source.

Free radicals generated in plasma by dissociation processes can dramatically increase chemical reaction rates, thus significantly increasing deposition rates, especially at the relatively low temperatures required for interconnection processes after the first aluminum metallization.

7.7.1.1 Plasma-enhanced chemical vapor deposition

Comparison of silane-based silicon oxide plasma-enhanced CVD (PECVD) and low-pressure CVD (LPCVD) processes is a good example for demonstrating the advantages of using PECVD in lower-temperature (<450 °C) processes. Interlayer dielectric (ILD) deposition requires lower temperatures after the first layer of metallization.

Some dissociation collisions in the silane oxide PECVD process are

$$e^- + SiH_4 \rightarrow SiH_2 + 2H + e^-$$
$$e^- + N_2O \rightarrow N_2 + O + e^-.$$

Both SiH_2 and O are free radicals with two unpaired electrons; they are very reactive. On a heated wafer surface, they react very quickly and form silicon dioxide:

$$SiH_2 + 2O \rightarrow SiO_2 + \text{other volatiles}.$$

In comparison, the LPCVD silane oxide deposition process uses SiH_4 and O_2. SiH_4 thermally dissociates to SiH_2 when it comes close to a heated wafer surface. SiH_2 then chemisorbs on the wafer surface, reacts with oxygen, and forms silicon dioxide:

$$SiH_4 \rightarrow SiH_2 + H_2,$$
$$SiH_2 + O_2 \rightarrow SiO_2 + \text{other volatiles}.$$

Without plasma, the chemical reaction rate of this CVD process at ~400 °C is very low, which causes a very low deposition rate. Thus, low-temperature LPCVD processes for ILD applications must be performed in a batch system to achieve reasonable throughput. Table 7.2 compares the properties of PECVD and LPCVD processes for ILD silane oxide deposition at 400 °C.

Table 7.2 Comparison of PECVD and LPCVD silane oxide processes.

Processes	LPCVD (150 mm)	PECVD (150 mm)
Chemical reaction	$SiH_4 + O_2 \rightarrow SiO_2 + \cdots$.	$SiH_4 + N_2O \rightarrow SiO_2 + \cdots$.
Process parameters	p = 3 torr, T = 400 °C	p = 3 torr, T = 400 °C, and **rf = 180 W**
Deposition rate	**100 to 200 Å/min**	**≥8000 Å/min**
Processing systems	Batch system	Single-wafer system
Wafer-to-wafer uniformity	Difficult to control	Easier to control

7.7.1.2 Stress control

Film stress is the force at the interface of two layers caused by a mismatch of the two materials. Ion bombardment of PECVD can be used to control CVD film stress. For dielectric thin films, especially silicon oxide films, compressive stress is favored. When heated, silicon expands faster than silicon oxide. If the film stress is compressive at room temperature, when the wafer is heated up for the next process step, the substrate will expand faster and therefore relieve compressive stress in the oxide film. If the oxide film has tensile stress at room temperature, it becomes even more tensile when heated. High tensile stress can cause film cracking, and in extreme cases it can even break a wafer.

Ion bombardment can hammer molecules and pack them densely into the film, which tends to make film stress become more compressive. Increasing rf power can increase both ion bombardment energy and ion flux, an increase that normally makes PECVD film more compressive. One advantage of PECVD is that by controlling rf power, the film stress can be independently controlled without major effects on other deposition characteristics, such as deposition rate and film uniformity. This is discussed in detail in Chapter 10.

7.7.1.3 Chamber cleaning

During CVD processes, deposition occurs not only on the wafer surface, but also on the processing kits, chamber walls, etc. The film deposited on these parts must be periodically cleaned to maintain stable processing conditions and prevent particle contamination of the wafers. Most CVD chambers use fluorine-based chemicals for cleaning processes.

Fluorocarbon gases such as CF_4, C_2F_6, and C_3F_8 are commonly used to clean silicon oxide CVD chambers. In plasma, these gases are dissociated and release free fluorine radicals. The chemical reaction can be expressed as

$$e^- + CF_4 \rightarrow CF_3 + F + e^-$$
$$e^- + C_2F_6 \rightarrow C_2F_5 + F + e^-.$$

Atomic fluorine is one of the most reactive free radicals; with ion bombardment from plasma, it can react with silicon oxide rapidly and form the gaseous compound SiF_4, which can be easily pumped out of the processing chamber:

$$F + SiO_2 \rightarrow SiF_4 + O + \text{volatile}.$$

SF_6 and NF_3 are commonly used fluorine-source gases for tungsten CVD chamber cleaning. Fluorine radicals can react with tungsten to form volatile tungsten hexafluoride, WF_6, which can be evacuated from the chamber with a vacuum pump.

Plasma chamber cleaning processes can be automatically endpointed (to prevent overcleaning) by monitoring the characteristic fluorine light emissions from the plasma. This is further discussed in Chapter 10.

7.7.1.4 Gap fill

When gaps between metal lines shrink to 0.25 μm with an aspect ratio of 4:1, most CVD methods cannot fill them without voids. The best known method for filling such narrow gaps without voids is the HDP-CVD process, shown in Fig. 7.15. The HDP-CVD process is covered in Chapter 10.

7.7.2 Plasma etch

Compared to wet etch, the advantages of plasma etch are anisotropic etch profile, an automatic endpoint, and less chemical consumption. Plasma etch also has a reasonably high etch rate, good selectivity, and good etch uniformity.

7.7.2.1 Etch profile control

Before plasma etch processes were widely used in semiconductor manufacturing, most wafer fabs used wet chemical etching to achieve pattern transfer. However, a wet etch is an isotropic etch (it etches in all directions at the same rate). Undercutting caused by an isotropic etch limits its application when feature size is smaller than 3 μm.

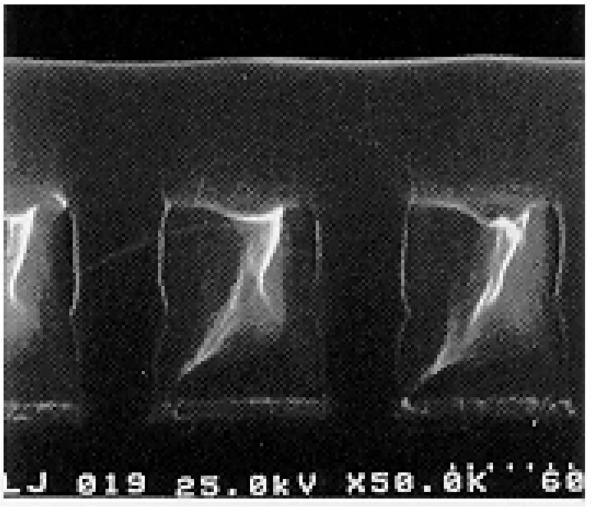


Figure 7.15 HDP-CVD oxide fills a 0.25-μm (aspect ratio 4:1) gap without voids.

Ions constantly bombard a wafer surface in plasma processes. With ion bombardment, by either a lattice-damaging mechanism or sidewall blocking mechanism, plasma etch processes can achieve an anisotropic etch profile. By reducing the etch processing pressure, ion MFP can be increased. This, in turn reduces ion collisional scattering and helps achieve better profile control.

7.7.2.2 Etch rate and selectivity

Ion bombardment associated with plasma helps break chemical bonds between surface atoms. These atoms are exposed to free radicals generated in the plasma. The combination of physical ion bombardment and chemical free radicals dramatically enhances the chemical reactions needed for etch processes. The etch rate and selectivity of the process is determined by the processing requirement. Since both ion bombardment and free radicals play an important role in etch processes, and rf power controls both ion bombardment and free radicals, rf power is the most important variable controlling etch rate. Increasing rf power can significantly increase etch rate. However, as discussed in detail in Chapter 9, this increase can affect etch selectivity.

7.7.2.3 Endpoint

Without plasma, etch endpoint is determined by time, or by an operator's visual inspection. In a plasma process, when the etch process has removed the material that requires etching and starts to remove the material underneath (the endpoint), chemical components in the plasma change due to the change in etch byproducts, signified by a variation in light emission. By detecting the change in light emission with an optical sensor, the endpoint can be automatically determined. This is a very useful tool in plasma etching for IC production.

7.7.2.4 Chemical use

Compared to wet etching, plasma etching uses significantly smaller amounts of chemicals, thus reducing the costs for chemicals and waste disposal.

7.7.3 Sputtering deposition

Compared with evaporation-deposited metal thin films, metal thin films deposited by plasma sputtering have a higher film quality, fewer impurities, and higher conductivity. They also have better uniformity, process control, and process integration capabilities. It is easier to deposit metal alloy films with sputtering deposition than with evaporation processes.

7.8 Plasma-Enhanced Chemical Vapor Deposition and Plasma Etch Chambers

7.8.1 Processing differences

CVD processes add materials to substrate surfaces, and etch processes remove materials from them. Therefore, etch processes operate at a lower pressure. Low pressure and high pumping rates help increase ion bombardment and remove etch

byproducts from an etch chamber. PECVD processes normally operate at a higher pressure (from 1 to 10 torr) than etch processes (from 30 to 300 mtorr).

7.8.2 Chemical vapor deposition chamber design

PECVD processes deposit thin films on wafer surfaces. Ion bombardment is very useful in controlling film stress. For PECVD chambers, an rf hot electrode (called a face plate, showerhead, etc.) has about the same area as the grounded electrode where the wafer is placed, as illustrated in Fig. 7.16. The electrode has very little self-bias. Ion bombardment energy is about 10 to 20 eV, mainly determined by the rf power.

7.8.3 Etch chamber design

If an etch system has both rf hot and grounded electrodes of the same size, both electrodes will receive approximately the same amount of ion bombardment. Etch processes depend heavily on ion bombardment to remove materials from the wafer surface. Ion bombardment can physically dislodge material off of the substrate surface and, more importantly, break chemical bonds so molecules on the surface react easily with etchant radicals. The easiest way to increase ion bombardment on a wafer is to increase the rf power. This increases both ion bombardment energy and flux. Unfortunately, it also increases ion bombardment on other electrodes and shortens lifetimes because of particle contamination.

By designing an rf hot electrode (called a chuck or cathode) to be smaller than a grounded electrode (chamber lid), and taking advantage of the self-bias, the plasma potential of the wafer side will be much higher than the chamber lid side (see Fig. 7.17). Thus, the most energetic ion bombardment occurs on the wafer side, while the chamber lid receives minimal bombardment. Ion bombardment energy on the wafer side varies from 200 to 1000 eV, and on the lid side is about 10 to 20 eV. Ion bombardment energy is mainly determined by rf power. It also relates

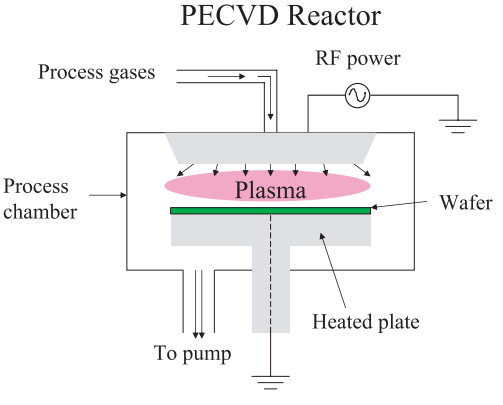


Figure 7.16 Schematic of a PECVD chamber.

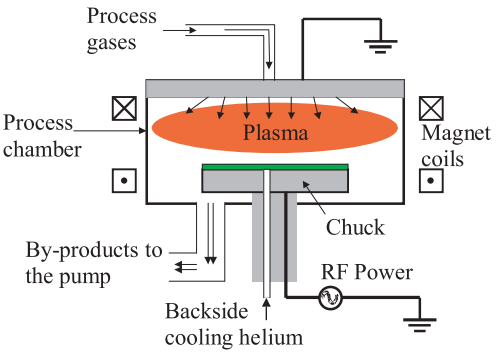


Figure 7.17 Schematic of a plasma etch chamber.

to chamber pressure, electrode spacing, processing gases, and applied magnetic field.

Plasma etch chambers operate at much lower pressure than PECVD chambers. At low pressure, the electron MFP is long. If the MFP is at the same order of magnitude as the electrode spacing or chamber dimension (~10 cm), it will not take many collisions before electrons are lost by hitting electrodes or chamber walls. Since ionization collisions are necessary for generating and maintaining plasma, it is difficult to generate plasma when pressure is so low.

Magnetic fields cause electrons to move in a spiral path. This helical path forces electrons to travel much longer distances and increases the chance of ionization collisions. A magnetic field can help generate and maintain plasma at lower pressure (<100 mtorr). Increasing the magnetic field can effectively increase plasma density, especially at low pressure. Since it increases electron density near the electrode surface, increasing the magnetic field can reduce dc bias.

Heavy ion bombardment generates a large amount of heat and can raise wafer temperature if the wafer is not properly cooled. For patterned etch, wafers are coated with a thin layer of photoresist as a pattern mask. The photoresist will reticulate if wafer temperature is above 150 °C. Therefore, a chamber designed for patterned etch requires a cooling system to prevent photoresist reticulation. Since chemical etch rates are sensitive to wafer temperature, even some blanket etch processing chambers [such as spin-on-glass (SOG) etchback chambers] need a wafer cooling system to regulate wafer temperature and control etch rate. Because etch processes require low operation pressure, and low pressure is not conducive for heat transfer, pressurized helium is normally introduced on the backside of a wafer to transfer heat from the wafer to the water-cooled pedestal (also called a chuck, cathode, etc.). An electrostatic chuck (E-chuck) is needed to prevent the wafer from being blown away from the pedestal due to the high backside pressure. Helium, which has very high thermal conductivity (second only to hydrogen), can provide a conductive path for heat transfer from the backside of the wafer to the water-cooled pedestal. For some old etch chambers, clamp rings were used to hold the wafer onto the pedestal. Because an E-chuck has much better performance with

plasma uniformity and particle contamination, it replaced the clamp ring in most advanced plasma systems.

Argon-sputtering etch chambers were widely used in dielectric thin-film processes for tapering the corners of gaps prior to gap filling and thin-film surface planarization. Since the sputtering etch rate is relatively insensitive to wafer temperature, it does not need the helium backside cooling system with a clamp ring or E-chuck.

7.9 Remote Plasma Processes

Some processes require free radicals to enhance chemical reactions but do not benefit from ion bombardment because it can cause plasma-induced damage. Remote plasma systems have been developed to meet these processing demands.

Figure 7.18 shows a remote plasma system. Plasma is created in a remote chamber by either microwave or rf power. Free radicals generated in the plasma flow into the processing chamber for either etch or deposition processes.

7.9.1 Photoresist strip

A remote plasma photoresist strip process uses O_2 and H_2O to remove photoresist right after etching. A remote plasma photoresist strip system (as shown in Fig. 7.19) can be easily integrated within an etch system. A wafer can remain in the same mainframe to perform in-situ etch/strip processing in sequence. This can improve both processing throughput and yield, since photoresist and residue etchants are stripped before the wafer is exposed to the atmosphere, where moisture in the air can react with residue etchants and cause corrosion on the wafer.

7.9.2 Remote plasma etch

Some etch processes, such as a local oxidation of silicon (LOCOS) nitride strip, wine glass contact hole, etc., do not require an anisotropic etch; therefore, ion

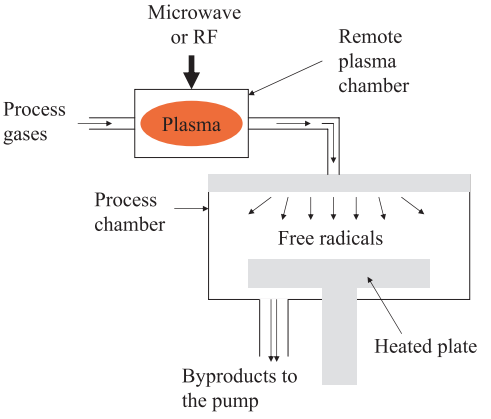


Figure 7.18 Remote plasma system.

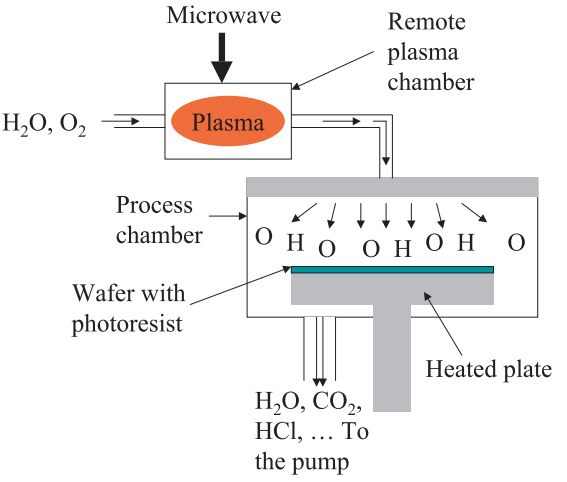


Figure 7.19 Remote plasma photoresist strip system.

bombardment is not necessary in these cases. Remote plasma etch systems are designed as dry chemical etching systems that can compete with wet etching processes for these applications. There was a trend toward replacing all wet processes in IC fabs with dry processes; however, this improvement proved to be impossible because of the wide application of CMP in advanced IC chip manufacturing processes.

7.9.3 Remote plasma cleaning

Since plasma in a processing chamber always generates ion bombardment along with free radicals, it can damage the parts inside a chamber, thus increasing cost of ownership. Another issue is that the most commonly used gases for CVD chamber cleaning processes are perfluorocarbon (PFC) gases, such as CF_4, C_2F_6, and C_3F_8. Because PFC gases contribute to global warming and ozone depletion, there is a strong political will to limit the use of these gases. Remote plasma cleaning was developed to solve these problems.

A remote plasma source uses microwave power to create a stable high-density plasma in a small cavity upstream from the processing chamber. Free radicals generated by the plasma flow into the processing chamber and react with deposited films, cleaning the chamber (shown in Fig. 7.20).

The most commonly used gas for remote plasma cleaning is NF_3. In microwave plasma, more than 99% of NF_3 dissociates. In comparison, less than 10% of carbon tetrafluoride or tetrafluoromethane (CF_4) dissociates in rf plasma. By using NF_3 microwave remote plasma cleaning processes, global warming PFC gases released from the semiconductor industry could be cut by more than 50%, and processing kit lifetimes could be significantly prolonged.

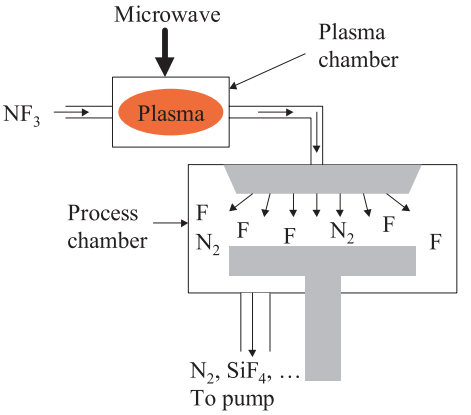


Figure 7.20 Remote plasma cleaning process.

7.9.4 Remote plasma chemical vapor deposition

Much research and development effort has been made to apply remote plasma CVD (RPCVD) processes for deposition of epitaxial silicon and epitaxial silicon germanium as semiconductor substrates, and silicon dioxide, silicon oxynitride, and silicon nitride for gate dielectric materials. Integrated with rapid thermal processes, RPCVD could also be used to deposit high-κ dielectrics such as TiO_2 and Ta_2O_5 for deep submicron devices in the future. RPCVD could also find applications for ILD0 barrier nitride deposition for sub-0.13-μm devices, since thermal budget limitations rule out the application of LPCVD nitride, and plasma-induced damage would limit the application of PECVD nitride, especially for large wafers.

7.10 High-Density Plasma

A plasma source that can generate HDP at low pressure (approximately a few millitorr) is highly desired for both etch and CVD processes. For etch processes, lower pressure gives longer ion MFP and less ion collisional scattering, enhancing etch profile control. HDP also provides more free radicals, which accelerate etching processes. For CVD processes, HDP can help achieve excellent gap-fill capabilities by in-situ, simultaneous deposition/etchback/deposition processing.

A conventional capacitively coupled plasma source cannot generate HDP. In fact, it is very difficult to generate plasma, even in a magnetic field when chamber pressure is a few millitorr. At that low pressure, electron MFP would be about the same or even longer than the electrode spacing, so there would not be enough ionization collisions. Therefore, different mechanisms are needed to generate HDP at very low pressure.

Another drawback of a capacitively coupled plasma source is that both ion flux and energy are directly related to rf power; therefore, this source cannot independently control them. To achieve better etch and CVD processing control

as feature size continuously shrinks, a plasma source capable of independently controlling both ion flux and energy is needed.

There are two types of HDP sources commonly used in the semiconductor industry: the ICP (also called transformer-coupled plasma, or TCP) source, and the ECR plasma source. Both can generate HDP at a few millitorr, with the added ability of independently controlling ion bombardment flux and energy.

7.10.1 Inductively coupled plasma

The mechanism of an ICP source is very similar to a transformer, explaining why it is also called a TCP source. The inductive coils shown in Fig. 7.21(b) perform like the initial coils of a transformer. When an rf current flows into the coils, a changing magnetic field is generated, which in turn generates a changing electric field through inductive coupling, as shown in Fig. 7.21(a). The inductively coupled electric field accelerates electrons and causes ionization collisions. Since the electric field is in an angular direction, electrons are accelerated in an angular direction. This allows electrons to travel a long distance without collision with the chamber wall or electrode. That is why an ICP system can generate HDP at the low pressure of a few millitorr.

The ICP design is very popular in the semiconductor industry. Systems using ICP include HDP dielectric CVD systems; silicon, metal, and dielectric HDP etching systems; native oxide sputtering cleaning systems; and ionized metal plasma PVD systems.

A bias rf system is added to the ICP chamber to generate self-bias and control ion bombardment energy. Since ion bombardment from HDP generates a considerable amount of heat, a helium backside cooling system with an E-chuck is needed for better wafer temperature control. Figure 7.21(b) illustrates an ICP chamber. In the ICP system, ion flux, which is mainly determined by plasma density, is controlled by the rf power source; ion bombardment energy is mainly controlled by the bias rf power.

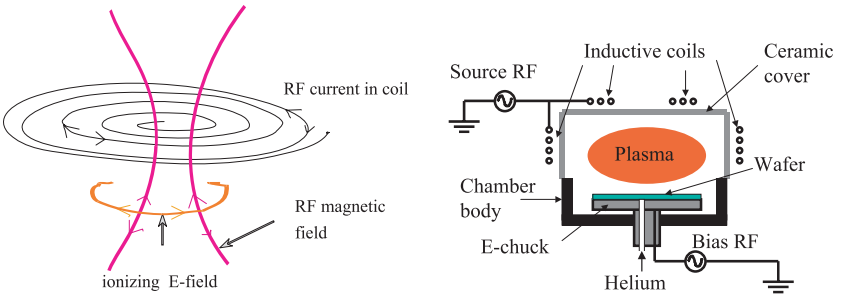


Figure 7.21 (a) Inductive coupling and (b) an ICP chamber.

7.10.2 Electron cyclotron resonance

Charged particles tend to rotate in a magnetic field. The frequency of the rotation, called gyrofrequency or cyclotron frequency, is determined by magnetic field strength. From Eq. (7.4), electron gyrofrequency is

$$\Omega_e \text{ (MHz)} = 2.8B \text{ (gauss)}.$$

In a magnetic field, when the frequency of the imposed microwave power equals the gyrofrequency of the electron, $\omega_{MW} = \Omega_e$, ECR occurs. Electrons can acquire energy from microwaves, growing more and more energized. Electrons then collide with other atoms or molecules, and the ionization collisions generate more electrons. These electrons are also resonant with the microwaves, gaining energy and creating even more electrons by ionization collisions. Since electrons are spiraling around the magnetic field line, as shown in Fig. 7.22(a), they cannot escape to the chamber wall or to the electrodes without many collisions, even if the MFP is longer than the chamber dimensions. This is how an ECR system can generate high-density plasma at low pressure.

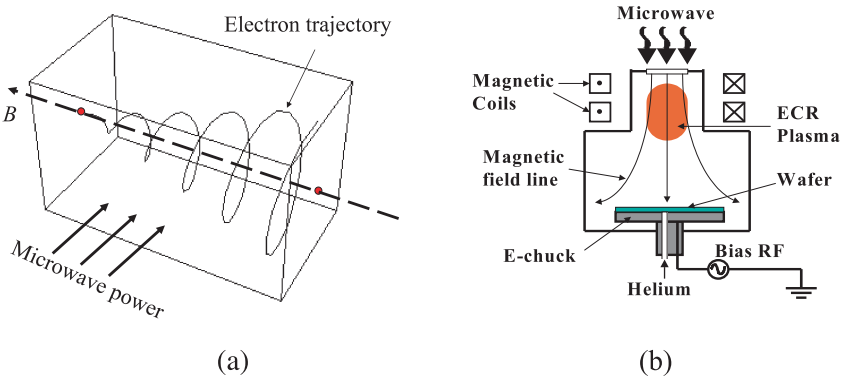


Figure 7.22 (a) ECR and (b) ECR chamber.

Like an ICP system, the ECR system also has a bias rf system to control ion bombardment energy, and an E-chuck with a helium backside cooling system to control wafer temperature [as shown in Fig. 7.22(b)]. Ion bombardment flux is mainly controlled by microwave power. One advantage of an ECR system is that, by changing the electric current in the magnetic coils, the position of resonance can be adjusted. Thus, process uniformity can be optimized by adjusting the current in the magnetic coils, which control plasma position.

7.11 Summary

- Plasma consists of ions, electrons, and neutrals.
- Three important collisions are ionization, excitation-relaxation, and disassociation.

- MFP is the average distance a particle can travel before it collides with another particle. It is inversely proportional to pressure.
- Free radicals from dissociation collisions enhance chemical reactions in CVD, etch, and dry cleaning processes.
- Plasma always has higher potential than the electrodes. Higher plasma potential causes ion bombardment.
- Increasing rf power increases both ion bombardment flux and ion bombardment energy in capacitively coupled plasmas.
- Low-frequency rf power gives ions more energy, leading to heavier ion bombardment.
- Etch processes need much more ion bombardment than PECVD processes. Etch chambers use a magnetic field to increase plasma density at low pressure.
- Capacitively coupled plasma sources cannot generate HDP.
- HDP at low pressure is desirable for both etch and CVD processes.
- ICP and ECR are the two most commonly used HDP sources.
- Both ICP and ECR plasma sources can independently control ion bombarding flux and energy.

7.12 Bibliography

B. Chapman, *Glow Discharge Process*, John Wiley & Sons, New York (1980).

F. F. Chen, *Introduction to Plasma Physics and Controlled Fusion, Volume 1: Plasma Physics*, 2nd ed., Plenum Press, New York (1984).

S. Dushman, *Scientific Foundations of Vacuum Technique*, J. M. Lafferty, Ed., John Wiley and Sons, New York (1962).

M. A. Lieberman and A. J. Lichtenberg *Principles of Plasma Discharges and Materials Processing*, John Wiley & Sons, New York (1994).

7.13 Review Questions

1. List the three components of plasma.
2. Which component of plasma moves the fastest?
3. True or false: the ionization rate of a conventional PECVD chamber is $\sim 100\%$.
4. List three important collisions in plasma and explain why they are important.
5. How can a PECVD process achieve a high deposition rate at a lower temperature?
6. What is MFP? How does it change with pressure?
7. How does the dc bias change when rf power increases?
8. List the importance of plasma ion bombardment for etching, PECVD, and sputtering PVD processes.
9. What are the major differences between plasma etch and PECVD processing chambers?
10. In an etch chamber, the wafer normally sits on which electrode? Why?

11. Why does an etch chamber need a helium backside cooling system with an E-chuck?
12. What are the advantages of an E-chuck over a clamp ring?
13. When the etch rate goes awry in a plasma etch system, why do fabs always check the rf system first?
14. Can a capacitively coupled plasma source generate HDP?
15. List the two most commonly used HDP systems.

Chapter 8
Ion Implantation

Objectives

After finishing this chapter, the reader will be able to:

- list at least three dopants commonly used in IC chip fabrication
- identify at least three doped areas from a CMOS chip cross section
- describe the advantages of ion implantation over diffusion
- describe the major components of an ion implanter
- explain the channeling effect and list at least two ways to minimize this effect
- explain the relation of ion range to ion species and energy
- explain the requirements of postimplantation annealing
- identify safety hazards related to ion implantation processes.

8.1 Introduction

One of the most important properties of semiconductor materials is that their conductivity can be controlled by adding dopants. Semiconductor materials such as silicon, germanium, and III–VI compounds (e.g., gallium arsenate) are doped with either n- or p-type dopants in IC fabrication. There are two methods used to dope semiconductors: diffusion or ion implantation. Before 1970, a diffusion process was used in IC fabrication. Currently, doping processes are mainly performed by ion implantation.

Ion implantation is an adding process in which dopant atoms are forcefully added into a semiconductor substrate by means of energetic ion beam injection. It is the dominant doping method in the semiconductor industry and is commonly used for various doping processes in IC fabrication. Figure 8.1 shows the relationship of the ion implantation process to other processes in IC manufacturing flow.

8.1.1 Brief history

Pure single-crystal silicon has a high resistivity; the purer the crystal, the higher the resistivity. Its conductivity can be improved by adding dopants such as boron (B), phosphorus (P), arsenic (As), and antimony (Sb). Boron is a p-type dopant with only three electrons in the outermost orbit (the valence shell). When a boron atom replaces a silicon atom in a single-crystal silicon lattice, a hole is created

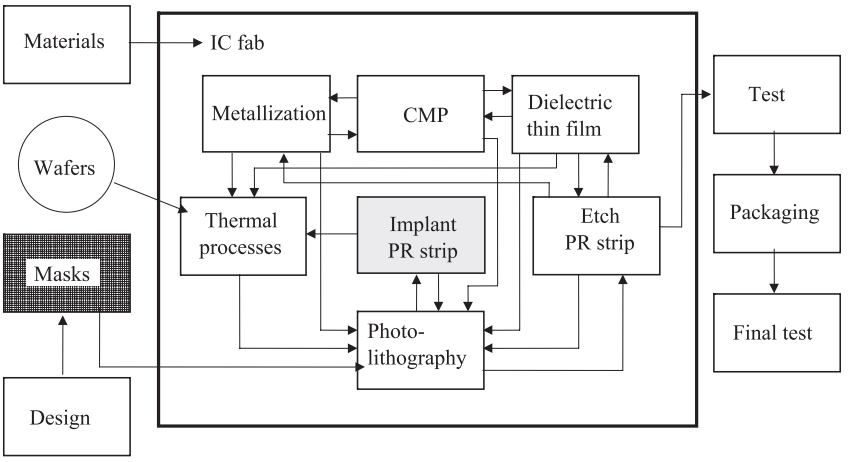


Figure 8.1 An advanced IC fabrication process flow.

in that single-crystal silicon. Holes can carry electric currents as positive charges. Phosphorus, arsenic, and antimony atoms have five electrons in their valence shells, thus they can cause electrons in single-crystal silicon to conduct electrical current. Because an electron has a negative charge, P, As, or Sb are called n-type dopants, and semiconductors with these dopants are called n-type semiconductors.

Before the mid-1970s, doping was accomplished with a diffusion process in high-temperature furnaces. The area where the furnaces were located was called the diffusion bay, and the furnaces were called diffusion furnaces, whether or not they were used for diffusion or other processes such as oxidation or annealing. Currently, few diffusion doping processes are performed in advanced IC fabs; furnaces are mainly used for oxidation and annealing processes. However, as mentioned in Chapter 2, the furnace area in an IC fab is still called the diffusion bay, and furnaces are still called diffusion furnaces for historical reasons.

The diffusion process usually takes several steps. Normally, a dopant oxide layer is deposited on the wafer surface in a predeposition process. Then an oxidation process converts the dopant oxide into silicon dioxide and forms a high-concentration dopant in the silicon substrate near the silicon–silicon dioxide interface. A high-temperature drive-in process diffuses dopant atoms into the substrate, to a depth required by the device design. All three processes, predeposition, oxidation, and drive-in, are high-temperature processes that take place in furnaces. After dopant drive-in, the oxide layer is stripped by a wet etching process. Figure 8.2 illustrates the diffusion doping process.

The physics of thermal diffusion are well known, and the processing tool is relatively simple and inexpensive. However, the diffusion process has some major limitations. For instance, dopant concentration and junction depth cannot be independently controlled, since both of them are strongly related to diffusion temperature. Another major disadvantage is that the dopant profile is always

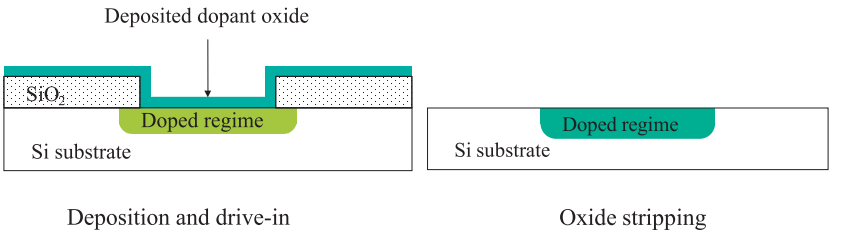


Figure 8.2 The diffusion process.

isotropic due to the nature of the diffusion process. These problems drove the search for an alternative doping method.

The idea of using ion implantation to dope semiconductors was first proposed by one of the three inventors of the first transistor, William Shockley, at Bell Laboratories in 1954. Shockley also held the patent for ion implantation (U.S. Patent 2787564). The physics and technology of a high-energy ion beam was developed during World War II, driven by research and manufacture of the first nuclear bomb. Accelerator and isotope separation technologies were directly employed in the design of the ion implanter. The introduction of ion implantation in the mid-1970s revolutionized the IC fabrication process.

Prior to the mid-1970s, semiconductor doping always used a diffusion process, which required silicon dioxide hard masks. In that time, bipolar transistors dominated IC production; when a metal-oxide semiconductor (MOS) transistor was made, it was the slower pMOS instead of the faster nMOS. This is because the p-type dopant, boron, diffuses much faster than n-type dopants, such as phosphorus and arsenic, in single-crystal silicon, making it easier to form a heavily doped p-type source/drain (S/D) than to form a heavily doped n-type S/D.

For pMOS made with a diffusion doping process, S/Ds were formed by boron diffusion with a silicon dioxide mask because boron diffuses much slower in silicon dioxide than it does in silicon. After S/D diffusion, the gate area was etched and cleaned, the thin gate oxide was grown, and the metal gate was formed. If the metal gate was misaligned with the S/D, as illustrated in Fig. 8.3, the transistor did not work properly. Oversized gates were required to ensure complete gate coverage over the S/D. The gate alignment issue presented a great challenge for shrinking feature size with this type of MOSFET manufacturing technology.

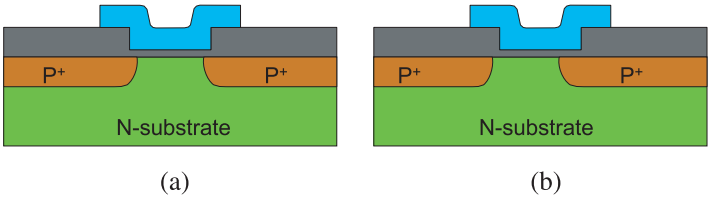


Figure 8.3 Gate alignment with S/D: (a) well aligned and (b) misaligned.

The application of ion implantation solved the gate alignment problem by using a self-aligned S/D doping process. In this case, the gate oxide is grown, and polysilicon is deposited, patterned, and etched. After photoresist stripping, high-current ion implantation is used to form the S/D. Because the polysilicon gate structure and field oxide block the ions, the S/D is always aligned with the polysilicon gate, as shown in Fig. 8.4.

It is not difficult to form heavy n-type junctions with ion implantation processes. Therefore, nMOS quickly replaced the slower pMOS after introduction of the ion implantation process. Because energetic dopant ion bombardment damages the single-crystal structure of the silicon substrate after ion implantation, a high-temperature (>1000 °C) annealing process is required to repair the damage and activate dopant atoms. The implantation annealing temperature is higher than the melting point of aluminum; therefore, different gate material must be used. Polysilicon and polysilicon-silicide stack (called polycide) have been used as gate materials for a long time, since the introduction of ion implantation. However, the transistor is still called an MOS, and no one calls it a POS (polysilicon-oxide semiconductor).

8.1.2 Advantages of implantation

Ion implantation has much better control over the doping process than the diffusion process. For example, ion implantation can independently control both dopant concentration and junction depth. Dopant concentration can be controlled by the combination of ion beam current and implantation time, and junction depth can be controlled by ion energy. Ion implantation processes can dope with a wide range of dopant concentrations, from 10^{11} to 10^{17} atoms/cm^2. The diffusion process is a high-temperature process and requires silicon dioxide hard masks. Before the diffusion process, a thick oxide layer needs to grow, then be patterned and etched to define the areas needing to be doped. The ion implantation process is a room-temperature process, and a layer of photoresist can block the energetic dopant ions. The thickness of the photoresist can be determined by the ion species and ion energy. Therefore, compared with diffusion processes, ion implantation has a lower cost because it does not require oxide growth or etch processes. Of course, the wafer holder of an ion implanter must have a cooling system that removes heat generated by energetic ions to prevent photoresist reticulation.

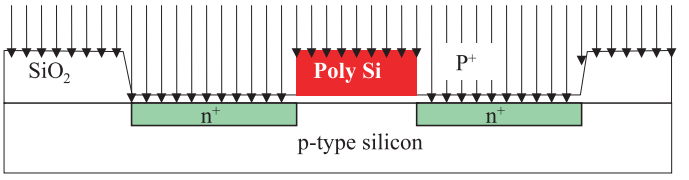


Figure 8.4 The self-aligned S/D with the ion implantation process.

The mass analyzer of an implanter selects exactly the ion species that needs to be implanted and generates a very pure ion beam, thus ion implantation has less possibility of elementary contamination. The ion implantation process always operates in a high vacuum, which is an inherently clean environment. Ion implantation is an anisotropic process, and dopant ions are implanted into the silicon mainly in the vertical direction. The doped region very closely reflects the area defined by a photoresist mask. Diffusion is an isotropic process; dopant always diffuses laterally underneath the silicon dioxide hard mask. For small feature sizes, it is very difficult to use diffusion processes to form a dopant junction. Figure 8.5 gives a comparison of the diffusion and ion implantation doping processes. Table 8.1 summarizes the advantage of ion implantation over diffusion in the doping processes.

8.1.3 Applications

The main application of ion implantation is doping the semiconductor substrate. A silicon wafer needs to be doped in order to change its conductivity in designated areas to form junctions such as wells and S/Ds for a CMOS IC. For a bipolar IC, doped junctions are needed to form buried layers, emitters, collectors, and bases.

Other applications of ion implantation are preamorphous implantation and buried layer implantation. Preamorphous implantation with silicon or germanium creates an amorphous layer on the substrate surface. It allows easier junction depth and profile control in the subsequent dopant implantation process. Germanium is a heavier atom than silicon, thus it has better damaging effects and is more favored for this application. There are two methods to make silicon-on-insulator (SOI)

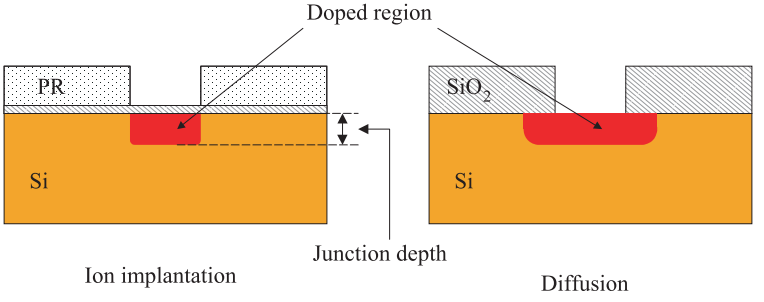


Figure 8.5 Illustration comparing of ion implantation and diffusion doping profiles.

Table 8.1 Comparison of ion implantation and diffusion.

Diffusion	Ion implantation
High temperature, hard mask	Low temperature, photoresist mask
Isotropic dopant profile	Anisotropic dopant profile
Cannot independently control the dopant concentration and junction depth	Can independently control dopant concentration and junction depth
Batch process	Both batch and single-wafer processes

wafers for high-end electronic devices; both involve ion implantation processes. One involves the implantation of high-dose oxygen ions into a silicon wafer and annealing the wafer to form a buried oxide layer underneath a thin single-crystal silicon layer. Another way is to use hydrogen implantation of the first wafer and bond it with the second wafer with a silicon dioxide layer on the surface. Because of the complete isolation between neighboring transistors, IC chips made on SOI wafers have lower leakage, higher noise immunity, higher radiation resistance, and better system reliability than IC chips made on bulk silicon wafers.

In advanced CMOS IC chips, the polysilicon gate for the nMOS is heavily n-type doped, and the polysilicon gate for the pMOS is heavily p-type doped. Silicide on top of a polysilicon structure shorts the p–n junction of the polysilicon and forms a local interconnection. Nitrogen can be implanted into the n-type doped polysilicon to form a barrier and prevent p-type dopant boron from diffusing into n-type doped polysilicon and causing device malfunction. Table 8.2 summarizes ion implantation applications, and Tables 8.3 through 8.7 list facts about phosphorus, arsenic, antimony, boron, and germanium.

8.2 Ion Implantation Basics

8.2.1 Stopping mechanisms

When ions bombard and penetrate the silicon substrate, they collide with lattice atoms. The ions then gradually lose their energy and eventually stop inside the silicon. There are two stopping mechanisms. Nuclear stopping occurs when the implanted ions collide with the nuclei of lattice atoms, are scattered significantly by the collision, and then transfer energy to the lattice atoms. In this hard collision, lattice atoms absorb enough energy to break free from the lattice binding energy, causing disorder and damage of the crystal structure. Another stopping mechanism occurs when incident ions collide with electrons of the lattice atoms. The incident ion path in the electronic collision is almost unchanged, energy transfer is very minimal, and crystal structure damage is negligible. This soft collision is called electronic stopping. The total stopping power, which is the energy loss of the ion per unit distance it travels inside the substrate, can be expressed as

$$S_{\text{total}} = S_{\text{n}} + S_{\text{e}},$$

where S_{n} is the nuclear stopping power, and S_{e} is the electronic stopping power. Figure 8.6 illustrates the stopping mechanisms, and Fig. 8.7 shows the relation between stopping power and ion velocity.

Table 8.2 Summary of ion implantation applications.

Applications	Doping	Preamorphous	Buried oxide	Poly barrier
Ions	n-type: P, As, Sb p-type: B	Si or Ge	O	N

Table 8.3 Facts about phosphorus.

Symbol	P
Atomic number	15
Atomic weight	30.973762
Discoverer	Hennig Brand
Place of discovery	Germany
Discovery date	1669
Origin of name	From the Greek word phosphoros, meaning bringer of light (an ancient name for the planet Venus)
Density of solid	1.823 g/cm^3
Molar volume	17.02 cm^3
Velocity of sound in substance	N/A
Electrical resistivity	10 $\mu\Omega \cdot$ cm
Refractivity	1.001212
Reflectivity	N/A
Melting point	44.3 °C
Boiling point	277 °C
Thermal conductivity	0.236 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	N/A
Applications	n-type dopant in diffusion, ion implantation, epitaxial growth, and polysilicon deposition. Dopant of CVD silicate glass (PSG and BPSG).
Main sources	P (red), PH_3, $POCl_3$

Table 8.4 Facts about arsenic.

Symbol	As
Atomic number	33
Atomic weight	74.9216
Discoverer	Known since ancient times
Place of discovery	Not known
Discovery date	Not known
Origin of name	From the Greek word arsenikon, meaning yellow orpiment
Density of solid	5.727 g/cm^3
Molar volume	12.95 cm^3
Velocity of sound in substance	N/A
Electrical resistivity	30.03 $\mu\Omega \cdot$ cm
Refractivity	1.001552
Reflectivity	N/A
Melting point	614 °C
Boiling point	817 °C
Thermal conductivity	50.2 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	N/A
Applications	n-type dopant in diffusion, ion implantation, epitaxial growth, and polysilicon deposition.
Main sources	As, AsH_3

Ion energy of the ion implantation process ranges from ultralow energy (0.1 keV) for ultrashallow junction (USJ), to high energy (1 MeV) for deep well implantation, which is located in region 1 of Fig. 8.7. From the far-left side of

Table 8.5 Facts about antimony.

Symbol	Sb
Atomic number	51
Atomic weight	121.760
Discoverer	Known since ancient times
Place discovered	Not known
Discovery date	Not known
Origin of name	From the Greek words anti and monos, meaning not alone (the origin of the symbol Sb comes from the Latin word stibium)
Density of solid	6.697 g/cm^3
Molar volume	18.19 cm^3
Velocity of sound in substance	3420 m/sec
Electrical resistivity	40 $\mu\Omega \cdot$ cm
Refractivity	1.001212
Reflectivity	55%
Melting point	630.78 °C
Boiling point	1587 °C
Thermal conductivity	24 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	11×10^{-6}K^{-1}
Applications	n-type dopant in ion implantation
Main sources	Sb

Table 8.6 Facts about boron.

Symbol	B
Atomic number	5
Atomic weight	10.811
Discoverer	Sir Humphrey Davy, Joseph-Louis Gay-Lussac, and Louis Jaques Thénard
Place discovered	England, France
Discovery date	1808
Origin of name	From the Arabic word buraq and the Persian word burah
Density of solid	2.460 g/cm^3
Molar volume	4.39 cm^3
Velocity of sound in substance	16200 m/sec
Electrical resistivity	$>10^{12}$ $\mu\Omega \cdot$ cm
Refractivity	N/A
Reflectivity	N/A
Melting point	2076 °C
Boiling point	3927 °C
Thermal conductivity	27 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	6×10^{-6}K^{-1}
Applications	p-type dopant in diffusion, ion implantation, epitaxial growth, and polysilicon deposition. Dopant of CVD silicate glass (BPSG).
Main sources	B, B_2H_6, BF_3

the figure, it is clear that the main stopping mechanism is nuclear stopping for low-energy and high-atomic-number ion implantation processes. The electronic

Table 8.7 Facts about germanium.

Symbol	Ge
Atomic number	32
Atomic weight	72.61
Discoverer	Clemens Winkler
Place discovered	Germany
Discovery date	1886
Origin of name	From the Latin word Germania, meaning Germany
Density of solid	5.323 g/cm^3
Molar volume	13.63 cm^3
Velocity of sound in substance	5400 m/sec
Electrical resistivity	~50000 μΩ · cm
Refractivity	N/A
Reflectivity	N/A
Melting point	938.25 °C
Boiling point	2819.85 °C
Thermal conductivity	60 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	6×10^{-6} K^{-1}
Applications	Ge, GeSi, and semiconductor substrate; Ge ions for preamorphous implantation.
Main sources	Ge, GeH_4

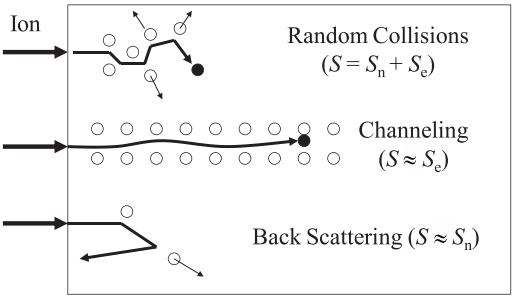


Figure 8.6 Different stopping mechanisms.

stopping mechanism is more important for high-energy, low-atomic-number ion implantation.

8.2.2 Ion range

When an energetic ion penetrates a target, it gradually loses its energy through collisions with electrons and nuclei of the atoms in the substrate, and eventually stops inside the substrate. Figure 8.8 shows the trajectory of an ion inside a substrate and the definition of an ion's projected range.

Generally, the higher the ion energy, the deeper the ion can penetrate into the substrate. However, even with the same implantation energy, ions do not stop exactly at the same depth in the substrate because each ion has different collisions

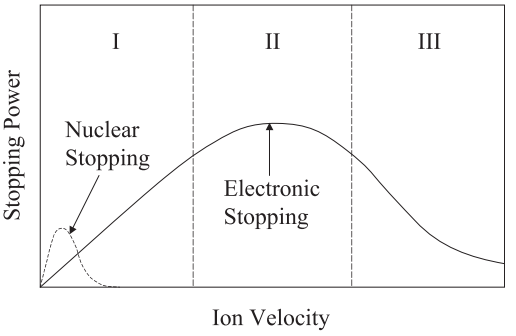


Figure 8.7 Relationship between stopping power and ion velocity (S. A. Cruz).

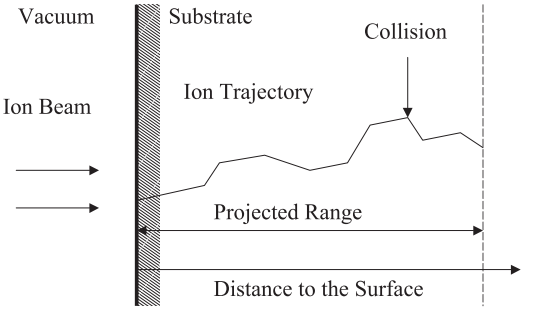


Figure 8.8 Ion trajectory and projected range.

with different atoms, causing each ion to stop at a different place. The projected ion range always has a distribution, as illustrated in Fig. 8.9.

Higher-energy ion beams can penetrate deeper into the substrate and therefore have longer projected ion range. Smaller, lighter ions have smaller collision cross sections; therefore, with the same ion energy, they can penetrate deeper into substrate and mask materials than larger, heavier ions. Figure 8.10 illustrates the projected range at different energy levels for boron, phosphorus, arsenic, and antimony ions in silicon substrate.

Ion projection range is an important parameter for ion implantation because it indicates the necessary ion energy for certain dopant junction depths. It also gives information on the required implantation barrier thickness for the ion implantation process. Figure 8.11 shows the required thickness of different barrier materials for 200-keV dopant ions. The figure shows that a boron ion requires the thickest masking layer because for the four dopant ions, boron has the lowest atomic number, smallest atom size, and longest projection range, and can penetrate deeper into the materials than any of the other three dopant ions. At high energy levels, for atoms with low atomic numbers such as boron, the main stopping mechanism is electronic stopping. Nuclear stopping is the main stopping mechanism for high-

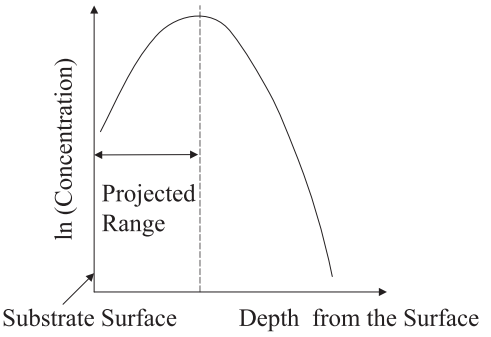


Figure 8.9 Distribution of projected ion range.

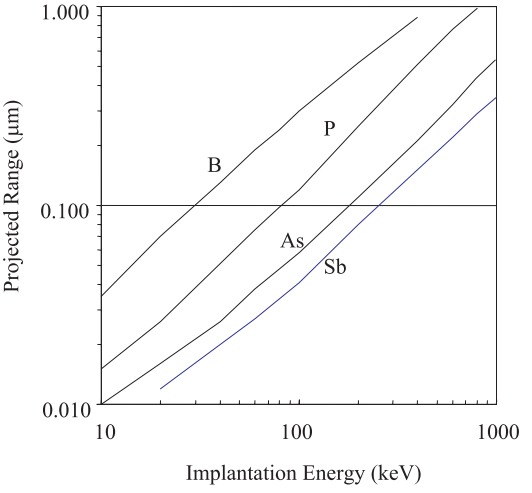


Figure 8.10 Projected range of dopant ions in silicon.

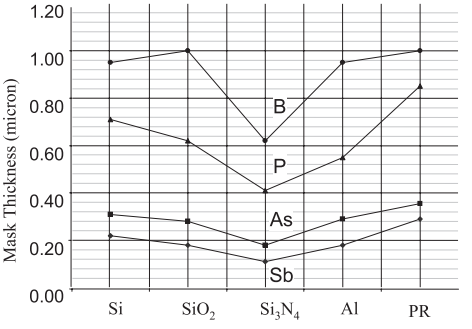


Figure 8.11 Required barrier thicknesses for 200-keV dopant ions.

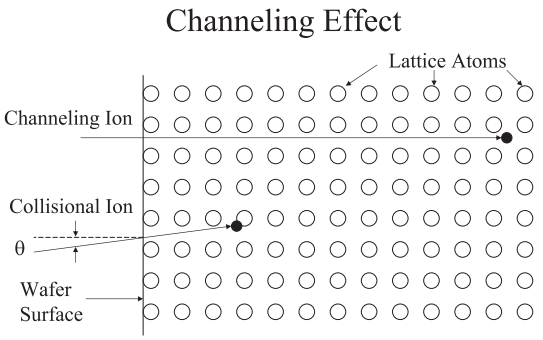


Figure 8.12 The channeling effect.

atomic-number dopant atoms. Similarly, the dopant ion with the highest atomic number, antimony, has the highest stopping power and shortest projection range, and therefore requires the thinnest masking materials.

8.2.3 Channeling effect

The projected range of an ion in an amorphous material always follows Gaussian distribution, also called normal distribution. In single-crystal silicon, lattice atoms are arranged in an orderly fashion, and many channels can be seen at certain angles. If an ion is implanted at those same angles into single-crystal silicon, it can travel a long distance along the channel, as illustrated in Fig. 8.12. This effect is called the channeling effect.

The channeling effect allows some ions to penetrate very deeply into a single-crystal substrate and causes a "tail" on the normal dopant distribution curve, as shown in Fig. 8.13. It is an undesired dopant profile, which could affect performance of the microelectronic device. Therefore, several methods have been developed to minimize this effect.

Question: The channeling effect allows an ion to penetrate deep into a single-crystal silicon substrate with relatively low energy. Why is this effect not used by fabs to create deep dope junctions?

Answer: If all ions of the ion beam are vertically implanted perfectly into the substrate, the channeling effect could be quite useful to form deep junctions with relatively low ion energy. However, ions always repel each other due to Coulombic forces; thus, an ion beam cannot be perfectly parallel. This means that many ions impact the wafer surface at small tilt angles and begin to have numerous nuclear collisions with

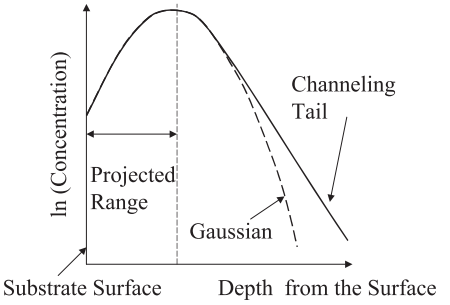


Figure 8.13 Channeling effect on the projection range distribution.

lattice atoms immediately after penetrating the substrate. The result is that only some ions channel deep into the substrate, but many ions are stopped due to normal Gaussian distribution.

One way to minimize the channeling effect is ion implantation on a tilted wafer, typically with a tilt angle of $\theta = 7$ deg. By tilting the wafer, ions impact the wafer at an angle that prevents them from channeling, as shown in Fig. 8.12. Incident ions have nuclear collisions immediately after entering the silicon substrate, effectively reducing the channeling effect. Most ion implantation processes use this technique, and most wafer holders used by ion implanters have the ability to adjust the tilting angle of the wafers.

Tilting a wafer can cause the photoresist to create a shadowing effect made by the photoresist, as shown in Fig. 8.14. This problem can be solved by rotating the wafer and/or a small amount of dopant diffusion during postimplantation annealing. If the tilt angle is too small, the dopant concentration in silicon can form a hump distribution due to the channeling effect, as shown in Fig. 8.15.

Another method widely used to handle the channeling problem is implantation through a thin layer of screen silicon dioxide. Thermally grown silicon dioxide is an amorphous material. Passing implantation ions will collide and scatter with silicon and oxygen atoms in the screen layer before they enter the single-crystal silicon substrate. Due to the collisional scattering, the pitch angle of the ions to the silicon crystal will be distributed in a wider range, thus reducing their chances of channeling. The screen oxide also can prevent the silicon substrate from touching and being contaminated by the masking photoresist. In some cases, both screen oxide and wafer tilting are used to minimize the channeling effect in ion implantation processes.

The problem with a screen layer is that some atoms in that layer can attain enough energy from energetic ions to implant themselves in the silicon. This is called the recoil effect. For a silicon dioxide screen layer, recoiling oxygen atoms can become implanted in the silicon substrate and form a high-oxygen-concentration region in the substrate near the silicon oxide interface, this oxygen-rich region can degrade carrier mobility and introduce deep-level traps. Therefore, in some implantation processes, a screen oxide cannot be used. In some cases,

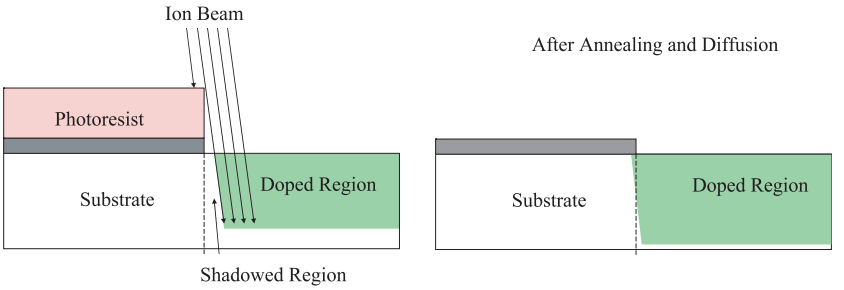


Figure 8.14 Shadowing effect and diffusion treatment.

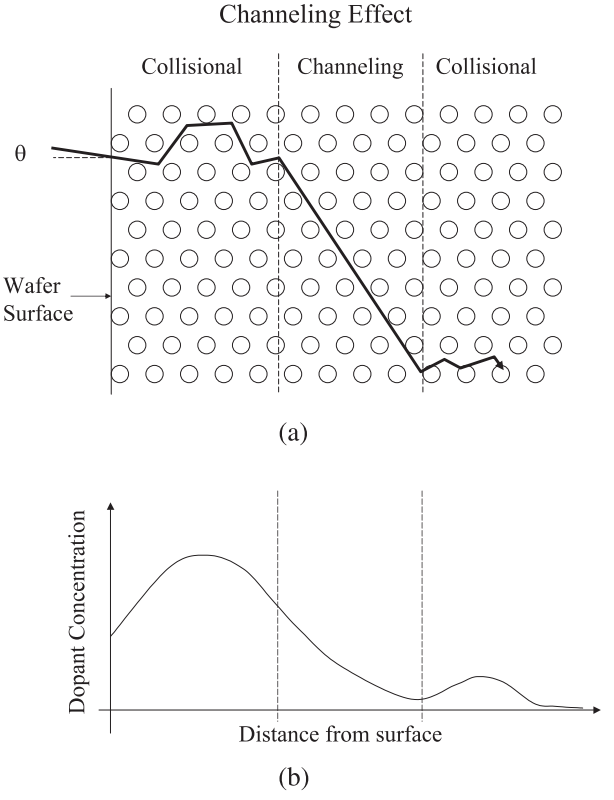


Figure 8.15 (a) Illustration and (b) graph of the postcollision channeling effect.

postimplantation oxidation and sacrificial oxide strips are needed to remove the oxygen-rich silicon layer. During the oxidation process, crystal damage caused by implantation can be annealed, and the oxide layer grows into the silicon substrate to consume the oxygen-rich region. Oxide stripping helps remove surface defects

and the high-oxygen-concentration layer. However, this technique is impracticable for USJ implantation because it consumes doped silicon substrate, and there will be no extra silicon for USJ to consume.

High-current silicon or germanium ion implantation can heavily damage a single-crystal lattice structure and create an amorphous layer near the wafer surface. By using Si or Ge amorphous implantation processes, the channeling effect can be completely eliminated because, in an amorphous substrate, a dopant junction profile formed by ion implantation always follows Gaussian distribution, which is very predictable, repeatable, and controllable. The tradeoff for this preamorphous implantation solution is that it increases production costs by adding an extra ion implantation step. Also, as feature size shrinks, the thermal budget for annealing also decreases. For advanced nanometer nodes, the thermal budget may not fully anneal and recover the crystal damage caused by the preamorphous implant, and the end-of-range defect could cause junction leakage.

8.2.4 Damage and annealing

In an ion implantation process, ions gradually lose their energy by colliding with lattice atoms, transferring their energy to these atoms during the process. The transferred energy is high enough (typically about 25 eV) for these atoms to break free from the lattice binding energy. These freed atoms will also collide with other lattice atoms while traveling inside the substrate, knocking them free from the lattice by transferring enough energy to them. These processes continue until none of the freed atoms have enough energy to free other lattice atoms. One energetic ion can cause thousands of displacements of lattice atoms. Damage caused by one energetic implanted ion is shown in Fig. 8.16.

Damage created by one ion can be quickly self-annealed due to thermal movement of atoms inside the substrate at room temperature. However, in an ion implantation process, the number of ions is so great that substantial lattice damage always occurs in a single-crystal substrate. This causes the layer close to the surface to become amorphous, and the self-annealing process cannot repair the crystal damage in a short period of time. Damage effect is related to the dose, energy, and mass of the ion species. Damage increases when dose and ion energy increase. If the implantation dose is high enough, the substrate crystal structure will be completely destroyed and become amorphous near the substrate surface within ion range.

To achieve device requirements, lattice damage must be repaired in an annealing process to restore the single-crystal structure and activate the dopant. Only when dopant atoms are close to the single-crystal lattices can they effectively provide electrons or holes as majority carriers to help carry electric current. In a high-temperature process, atoms have random thermal movement until they move into single-crystal lattice locations, where the lowest free energy is found. Since the undamaged substrate underneath is single-crystal silicon, silicon and dopant atoms in the damaged amorphous layer reconstruct the single-crystal structure by falling into the lattice grids and becoming bonded by the lattice energy.

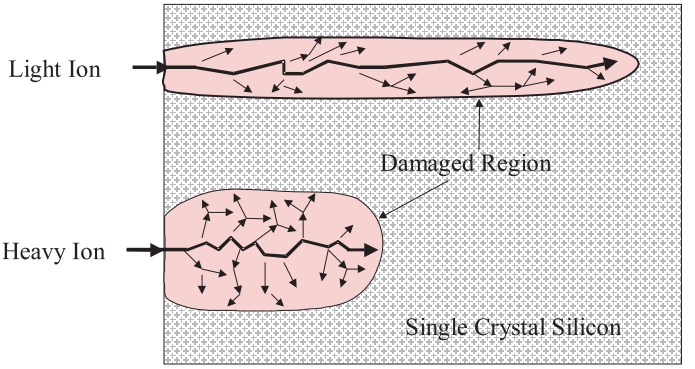


Figure 8.16 Damage caused by one ion.

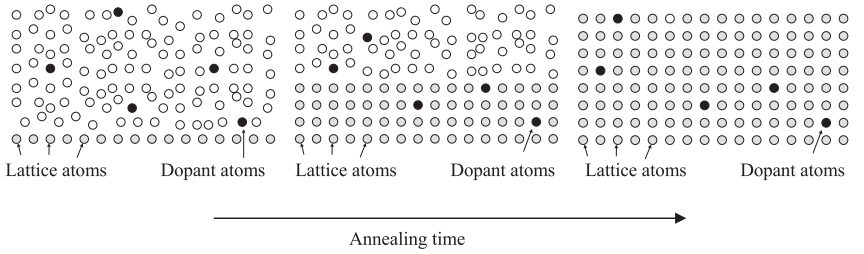


Figure 8.17 Postimplantation annealing process.

Figure 8.17 illustrates crystal recovery and dopant activation in a thermal annealing process.

During high-temperature annealing, single-crystal lattice recovery, dopant atom activation, and dopant atom diffusion happen simultaneously. When IC feature size shrinks deep into the submicron level, there is less room for dopant atoms to diffuse; thus, minimizing diffusion during the thermal annealing process becomes very important. In amorphous silicon, dopant atoms have free thermal movement without any restrictions and thus diffuse much faster than in single-crystal lattices because the lattice binding energy severely restricts movement of the dopant atoms. It has been found that at a lower temperatures, the diffusion process outpaces the annealing process. When the temperature is high enough, such as >1000 °C, the annealing process is faster than the diffusion process. This is because annealing activation energy (~5 eV) is higher than diffusion activation energy (3 to 4 eV).

Furnaces have been used for postimplantation annealing processes. A furnace is considered a batch system, which can process more than 100 wafers with temperature ranges from 850 to 1000 °C and a processing time of about 30 min. Since the furnace annealing process takes a long time, the diffusion of dopant atoms can be quite substantial; thus, for devices with small feature sizes,

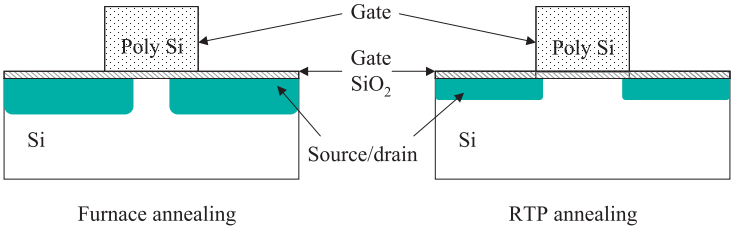


Figure 8.18 Dopant diffusion in furnace and rapid thermal annealing processes.

it becomes unacceptable. Only for some noncritical implantations with large feature sizes, such as well implantation processes, are furnaces still used for postimplantation annealing and dopant drive-in. For critical doping processes such as S/D implantation annealing, furnaces induce too much dopant diffusion and cause intolerable performance degradation for nanometer microelectronic devices.

Rapid thermal processing (RTP) has been developed to anneal the damages caused by implantation while minimizing dopant diffusion to meet the requirements of dwindling IC devices. An RTP system is a single-wafer system with a temperature ramp rate of up to 250 °C/sec, and good temperature and temperature uniformity control at about 1100 °C. With an RTP system, the rapid thermal annealing (RTA) process can operate up to 1150 °C, at which temperature implantation damage can be annealed in less than 20 sec. RTA systems can process a wafer in about a minute—from chamber entry to temperature ramp up, annealing, wafer cool-down, and wafer exit.

Question: Can a furnace temperature be ramped up and cooled down as quickly as an RTP system?

Answer: No. A furnace has very large thermal capacity, and a great deal of heating power is required to rapidly ramp up its temperature. This is very difficult to achieve without large temperature oscillation due to temperature overshoot and undershoot.

Because of the slow temperature ramp rate (usually less than 10 °C/min), it takes a long time to ramp up the processing tube from an idle temperature (normally between 650 to 850 °C) to a required annealing temperature, such as 1000 °C. While the temperature ramps up, some damage begins to anneal. Wafers must be pushed in and pulled out of a furnace very slowly to prevent wafer warping due to thermal stress caused by sudden temperature changes. Because a furnace is always at a high temperature when idling, wafers at both ends of a wafer boat have different annealing times due to the slow pushing and pulling, which causes wafer-to-wafer (WTW) nonuniformity.

RTA systems can ramp up wafer temperature from room temperature to 1100 °C very quickly (within 10 sec), while maintaining precise control of wafer temperature and temperature uniformity within the wafer. At about 1100 °C, a single-crystal lattice can be recovered in 10 sec with minimum dopant diffusion.

The RTA process has better WTW uniformity control than the furnace annealing process.

With the continued decrease in device feature size, even RTA processes are not fast enough to achieve dopant activation while keeping dopant diffusion in a tolerable range. Other annealing techniques such as spike, laser, and flush annealing have been developed and applied in IC manufacturing.

8.3 Ion Implantation Hardware

An ion implanter is a very large piece of equipment. It is one of the largest processing systems found in a semiconductor fab. It has several subsystems, including gas, vacuum, electrical, and control systems, and most importantly, the beam line system, shown in Fig. 8.19.

8.3.1 Gas system

Ion implanters use many dangerous gases and vapors to generate dopant ions. Examples are: flammable and poisonous gases such as arsine, phosphine, and diborane; corrosive gases such as boron trifluoride; and harmful vapor-form solid materials such as boron and phosphorus. To reduce the risk of these hazardous gases leaking into a fab, a special gas cabinet is placed inside the enclosure of an ion implanter to store these chemicals close to the ion source where they will be used. For some poisonous gases, subatmospheric cylinders are required for safety.

8.3.2 Electrical system

To generate high-energy ions, high-voltage dc power is needed to accelerate the ions. Normally, up to 200-kV dc power supply systems are equipped in an implanter. To generate ions in an ion source, either a hot filament or rf plasma system is used. A hot filament system requires a large current and a few hundred volts of bias power supply, and a rf ion source requires about a thousand watts

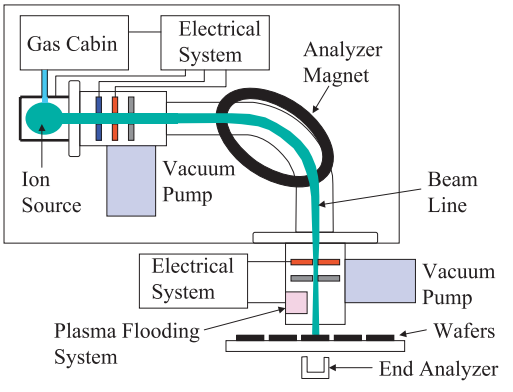


Figure 8.19 An ion implanter.

of rf power supply. Analyzer magnets need a high current to generate magnetic fields strong enough to bend the trajectories of energetic ions, and help select the correct ion species to create an ultrapure ion beam. Power supply systems need to be calibrated accurately. The voltage and current of the power supply must be very stable to ensure high and consistent process yields.

8.3.3 Vacuum system

A beam line must operate entirely in high-vacuum conditions to minimize collisions between energetic ions and neutral gas molecules along the ion trajectory. Collisions can cause ion scattering and loss, as well as implantation of unwanted species due to charge exchange collisions between ions and neutral atoms that cause elementary contamination. The pressure of an ion beam line should be low enough that the MFP of the ion is much longer than the length of the ion trajectory from the ion source to the wafer surface. Combinations of cryo-, turbo, and dry pumps are used to achieve 10^{-7}-torr high vacuum in a beam line system.

Because of the dangerous gases used in ion implantation processes, the exhaust of an implanter vacuum system must be kept separate from other process exhaust systems. Exhaust gases need to pass through a burn box and a scrubber before they can be safely released into the atmosphere. In a burn box, flammable and explosive gases are neutralized by oxygen in a high-temperature flame. In a scrubber, flushing water dissolves the corrosive gases and the burnt dust that pass through it.

8.3.4 Control system

To achieve design requirements, an ion implantation process needs to precisely control ion beam energy, current, and ion species. The process also needs to control mechanical parts, such as the robot for wafer loading and unloading, and control wafer movement to achieve uniform implantation across a wafer. To maintain system pressure, throttle valves are controlled according to a pressure setting point.

The center of the control system is a CPU board. Different control boards collect signals from the systems within the implanter and send them to that CPU board, which processes the data and sends instructions to the implanter's systems via the control board.

8.3.5 Beam line system

The ion beam line system is the most important part of an ion implanter. It consists of an ion source, extraction electrodes, mass analyzer, postacceleration electrodes, plasma flooding system, and end analyzer. Figure 8.20 illustrates a beam line system of an ion implanter.

8.3.5.1 Ion source

Dopant ions are generated in an ion source through ionization discharge of the atoms (or molecules) of dopant vapors or gaseous dopant chemical compounds. A hot filament ion source is one of the most commonly used ion sources. In this

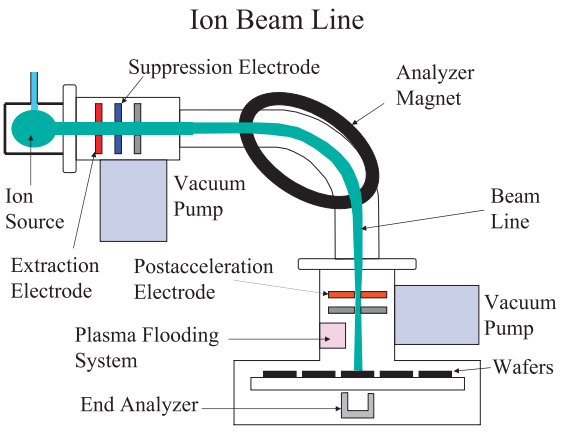


Figure 8.20 The beam line system of an ion implanter.

type of ion source, large electrical currents flow through the filament, heat the tungsten filament, and cause thermal electron emission from the red-hot filament surface. Thermal electrons are accelerated by the arc power supply voltage with enough energy to dissociate and ionize the dopant gas molecules and atoms. Figure 8.21 shows a hot filament ion source. The magnetic field in the ion source forces electrons into gyromotion, which helps the electrons travel longer distances and increases their probability of colliding with dopant molecules, generating more dopant ions. The negatively biased anticathode plate expels electrons from its vicinity and reduces electron loss by collision with the wall along the magnetic field line.

Other types of ion sources such as rf and microwave sources have also been developed and applied to ion implantation processes. The rf ion source uses inductively coupled rf power to ionize dopant ions. The microwave ion source uses ECR to generate plasma and ionize dopant ions. Figure 8.22 shows the schematics of rf and microwave ion sources.

8.3.5.2 Extraction system

An extraction electrode with negative bias draws ions out of the plasma in an ion source and accelerates them to sufficiently high energy, about 50 keV. It is necessary for ions to have high enough energy for the analyzer magnetic field to select the correct ion species. Figure 8.23 shows an extraction system. When dopant ions accelerate toward extraction electrodes, some of the ions pass through the slit and continue to travel along the beam line. Some ions hit the extraction electrode surface, which generates x rays and excites some secondary electrons. A suppression electrode with sufficiently lower electrical potential (up to 10 kV) than an extraction electrode is used to prevent these electrons from being accelerated back toward the ion source, causing damage there. All electrodes are shaped with a narrow slit, through which ions are extracted as a collimated ion flux forming an ion beam.

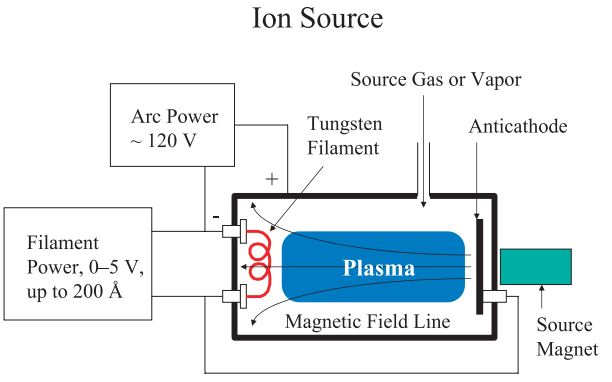


Figure 8.21 Hot filament ion source.

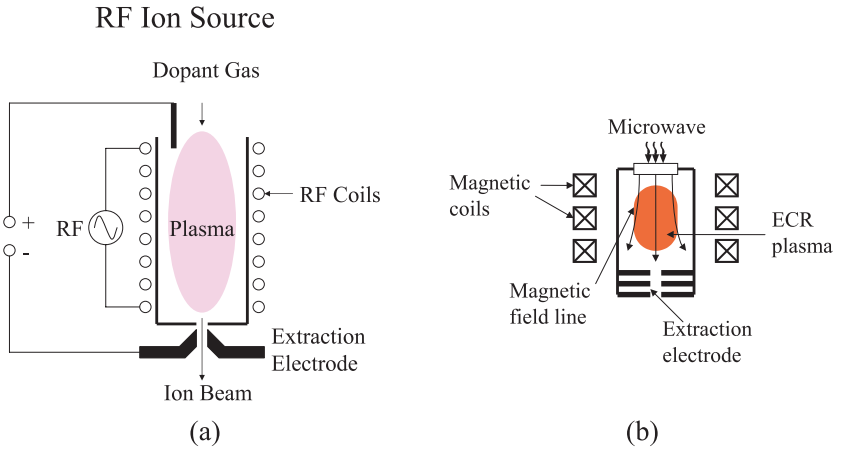


Figure 8.22 Schematics of (a) the rf and (b) microwave ion source.

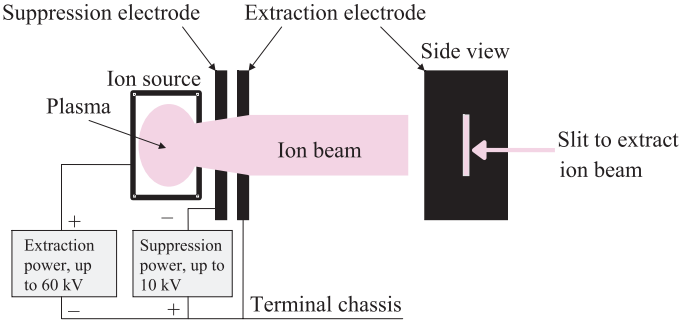


Figure 8.23 Schematics of an extraction system.

Ion beam energy after extraction is determined by the difference between the ion source and the extraction electrode. The extraction electrode potential is the same as the terminal frame and is often called the system ground potential. Notice that the potential difference between system ground and real ground (implanter cover plates) can be as high as −50 kV, which can cause a lethal electrical shock through arcing discharge without direct contact!

8.3.5.3 Mass analyzer

In a magnetic field, moving charged particles start rotating due to magnetic force, which is always perpendicular to the movement direction of a charged particle. For fixed magnetic field strength and ion energy, the gyroradius of the charged particle is only related to its mass/charge ratio, or m/q. This property was used to separate isotopes to generate enriched uranium-235 (^{235}U) from uranium-238 (^{238}U) for making the atomic bomb. In almost every ion implanter, a mass analyzer is applied to precisely select the desired ion species and weed out the unwanted ion species. Figure 8.24 illustrates the mass analyzer of an ion beam.

For example, BF_3 is commonly used as a boron dopant source. In plasma, combinations of dissociation and ionization collisions create a variety of ions. Because boron has two isotopes—^{10}B (19.9%) and ^{11}B (80.1%)—it also has several ionization states, which further increase the number of ion species, as shown in Table 8.8.

For p-well implantation, $^{11}B^+$ is preferred because, at the same energy level, it can penetrate deeper into silicon substrate due to its lighter weight. For shallow junction implantation, the $^{11}BF_2^+$ ion is the most likely species to be selected due to its large size and heavy weight. At the lowest energy level an implanter can remain stable, the $^{11}BF_2^+$ ion has the shortest ion range of these boron-containing ions; therefore, it can create the shallowest p-type junction. A small amount of fluorine integrated into the silicon substrate can bind with dangling silicon bonds at the silicon–silicon dioxide interface, which can help minimize the interface state charge and improve device performance.

Before ions enter the analyzer, their energy is determined by the potential difference between the ion source and the extraction electrode, the potential difference normally being set at approximately 50 kV. The energy of the single

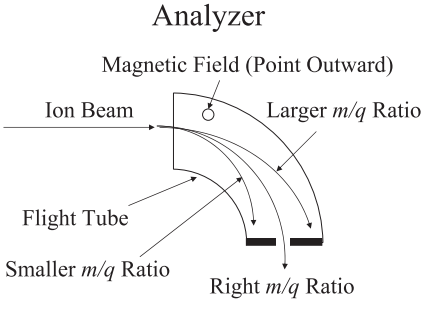


Figure 8.24 Mass analyzer of an ion implanter.

Table 8.8 Possible ion species from BF_3 plasma.

Ions	Atomic or molecular weight
^{10}B	10
^{11}B	11
^{10}BF	29
^{11}BF	30
F_2	38
$^{10}BF_2$	48
$^{11}BF_2$	49

charged ions after extraction is 50 keV. With knowledge of the ion m/q ratio and ion energy, the required magnetic field strength for ion trajectory passing through the narrow slit can be calculated. By adjusting electrical current in the magnetic coils, a mass analyzer can precisely select the desired dopant ions.

Question: $^{10}B^+$ is lighter than $^{11}B^+$, thus it can penetrate deeper than $^{11}B^+$ when both ions have the same energy. Why is $^{10}B^+$ not selected for p-well implantation?

Answer: Since only one in five boron atoms are ^{10}B and the rest are ^{11}B, in the plasma the $^{10}B^+$ ion concentration is only about one quarter of the $^{11}B^+$ ion. If a $^{10}B^+$ ion is selected, the ion beam current will be about one-fourth of the ion current of the $^{11}B^+$ beam. To achieve the same level of dopant concentration would take four times the amount of time to implant using a $^{10}B^+$ ion beam as with a $^{11}B^+$ ion beam, which causes lower throughput and is not cost effective.

Question: In plasma, phosphorus vapor can be ionized and form different ions. P^+ and P_2^{++} are two of them. Can a mass analyzer separate these two?

Answer: If P^+ and P_2^{++} ions have the same energy, a mass analyzer cannot separate them because they have exactly the same m/q ratio; therefore, they will have the same ion trajectories. When they implant into the substrate, P_2^{++} will not penetrate as deeply as P^+ because it has shorter ion range due to its larger size and heavier mass. This could cause energy contamination, which induces unwanted dopant concentration profiles and affects device performance. Fortunately, most P_2^{++} ions have about twice the amount of energy of P^+ ions after preacceleration by the extraction potential because they are doubly charged. With the same m/q ratio, higher energy ions have larger gyroradii and therefore will hit the outer wall of the flight tube of the mass analyzer. Their trajectory is similar to the ion beam with the larger m/q ratio shown in Fig. 8.24.

8.3.5.4 Postacceleration

After the mass analyzer selects the correct ion species, ions pass into the postacceleration section, where beam current and final ion energy are controlled.

Ion beam current is controlled by a pair of adjustable vanes, and ion energy is controlled by postacceleration electrode potential. Ion beam focus and beam shapes are also controlled in this section by the defining apertures and electrodes. Figure 8.25 shows a postacceleration assembly.

For the high-energy ion implanters used in well and buried layer implantation processes, several stages of high-voltage acceleration electrodes are required, connected in a series along the beam line to accelerate ions to several megaelectron volts. For ion implanters used for USJ applications (especially for p-type with boron implantation), an electrode in postacceleration is connected in a reverse way so that the ion beam is decelerated instead of accelerated when passing the electrode. This can generate a pure ion beam with energy as low as 0.1 keV.

In some implanters, after postacceleration, an electrode is used to bend the ion beam at a small angle, such as 10 deg, which helps dispose of energetic neutral particles. Neutral particles will not respond to the electric field generated by the electrode; they will keep moving straight forward while the ions' trajectories bend and move toward the wafer, as shown in Fig. 8.26. Some implantation systems bend the ion beam twice, in an S-shaped trajectory, to achieve even higher energy purity.

8.3.5.5 Charge neutralization system

When ions implant into the silicon substrate, they carry a positive charge to the wafer surface. If positive charges accumulate on the wafer, they can cause a wafer charging effect; positively charged wafer surfaces tend to expel positive ions and

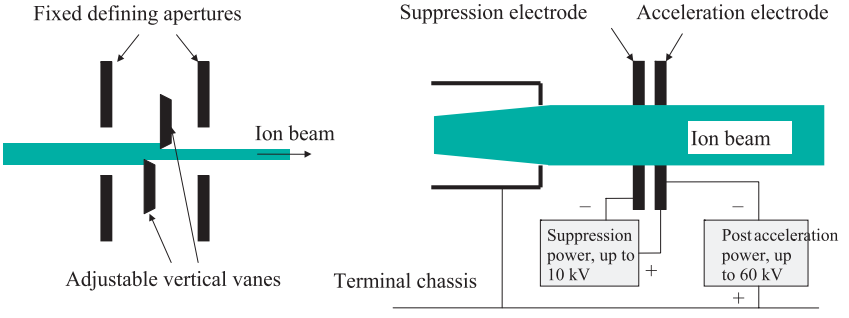


Figure 8.25 Beam current control and postacceleration assembly.

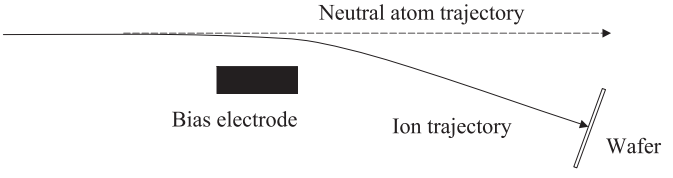


Figure 8.26 Bent ion beam trajectory.

cause beam blowup. This results in nonuniform ion implantation and nonuniform dopant distribution across the wafer, as shown in Fig. 8.27.

When the surface charge concentration is too high, charge-induced electric fields can be high enough to break down the thin gate-oxide layer and significantly impact IC chip yield. When positive charges accumulate to a certain level, they discharge in the form of arcing and cause defects on the wafer surface.

To manage the wafer charging problem, large numbers of electrons, which have negative charges, are required to neutralize positive ions on the wafer surface. There are several approaches for wafer neutralization: plasma flooding systems, electron guns, and electron showers are all used for providing electrons to neutralize ions and minimize the wafer charging effect. Figure 8.28 shows a plasma flooding system.

In a plasma flooding system, thermal electrons are emitted from a hot tungsten filament surface and accelerated by a dc power supply. They collide with neutral atoms in the chamber and generate plasma by means of many ionization collisions with electrons and ions. Electrons in the plasma are attracted into the ion beam and

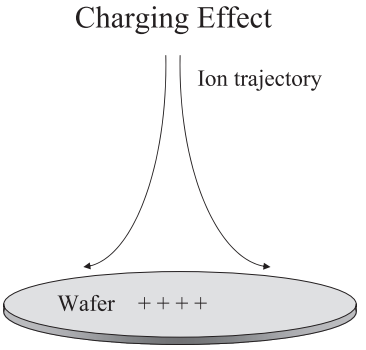


Figure 8.27 Nonuniform implantation caused by the wafer charging effect.

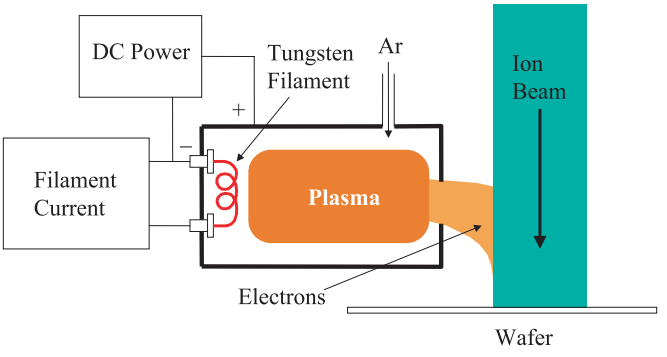


Figure 8.28 A plasma flooding system.

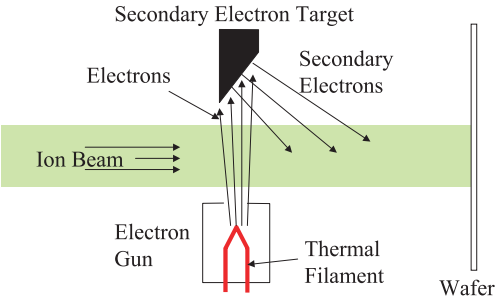


Figure 8.29 Electron gun system.

flow to the wafer surface with ions to neutralize the wafer and minimize the wafer charging effect.

Figure 8.29 illustrates an electron gun system. Thermal electrons emitted from the hot filament accelerate to a secondary electron target and collide into that target with high enough energy to generate large amounts of secondary electrons, which are knocked loose from the target surface. These electrons flow with the ions in the beam line and neutralize ions on the wafer surface.

8.3.5.6 Wafer handler

The most important task for a wafer handler is to achieve uniform implantation across a wafer. The diameter of an ion beam is about 25 mm (~1 in.); it needs to move either the ion beam or wafer—or for some implanters, both—to scan the ion beam uniformly across the wafer, which can have a diameter up to 300 mm.

In a spin wheel system, the wheel spins at a very high rate. An arc band with the ion beam diameter as its bandwidth is implanted with ions when a wafer passes through the ion beam. The center of the wheel swings back and forth, allowing the ion beam to scan uniformly across the entire wafer for all wafers on the spin wheel. Figure 8.30 illustrates a spin wheel wafer-handling system.

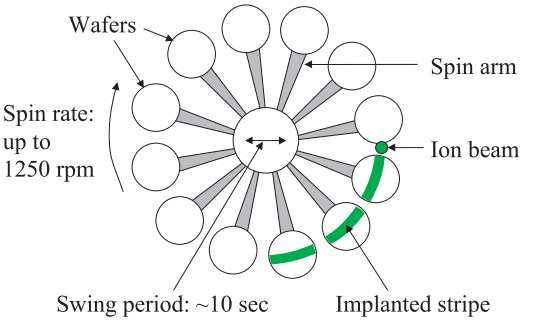


Figure 8.30 Schematic of a spin wheel wafer-handling system.

A spin disk is very similar to a spin wheel, the difference being that instead of swinging the entire disk, the spin disk scans the ion beam to achieve uniform ion implantation across the wafer. Figure 8.31 illustrates a spin disk system.

Another type of wafer-handling system used in ion implanters, which combines ion beam scanning and wafer movement, is shown in Fig. 8.32(a). By varying the applied bias potential between the scanning electrodes, the ion beam can scan back and forth in the x direction, while the wafer moves up and down in the y direction, driven by a step motor. The entire wafer can be uniformly implanted in this way. This type of scanning technique can be used in single-wafer implantation systems.

Some single-wafer implantation systems use wide-band or ribbon beams instead of scanning beams in the x direction while moving the wafer up and down for uniform ion implantation, as shown in Fig. 8.32(b). Some other systems use wide-band beams placed in the y direction and swing the wafer back and forth in the x direction to achieve uniform implantation. Single-wafer ion implantation systems have become mainstream in advanced nanometer IC manufacturing for critical ion implantation processes such as USJ S/D formation, including S/D extension (SDE) implantation and S/D implantations.

A wafer holder must have a cooling system to remove heat generated by the energetic ion bombardment and control wafer temperature. Otherwise, wafer

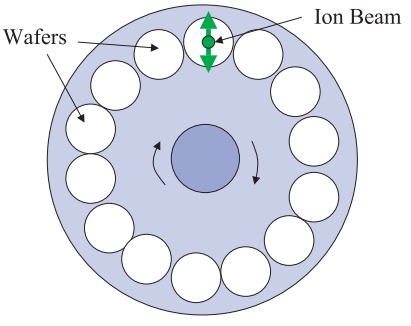


Figure 8.31 Schematic of spin disk wafer-handling system.

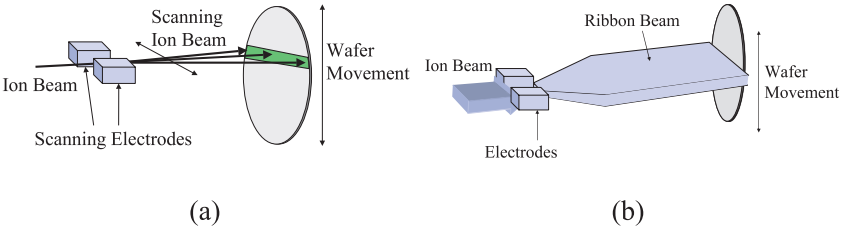


Figure 8.32 Single-wafer ion implantation systems: (a) scanning ion beam and (b) ribbon beam.

temperature can be high enough to cause photoresist reticulation. Wafer holders are usually water cooled, and temperatures are controlled below 100 °C.

8.3.5.7 Beam stop

At the end of a beam line, a beam stop or end station is needed to absorb ion beam energy; it also serves as an ion beam detector for the beam current, beam energy, and beam shape measurement. A water-cooled metal plate is used to remove heat generated by energetic ion bombardment and to block the x-ray radiation generated by the sudden stopping of the energetic ions on the target surface.

Figure 8.33 illustrates a beam stop. At the bottom of the ion beam stop, there is an array of ion detectors, which can be used to measure ion energy and energy spectrum, beam current, and beam shape. In an ion beam, many electrons travel with the ion beam. The major source of these electrons is a charge neutralization system, such as an electron flooding system, electron gun, or some other electron source, that generates a large number of electrons to neutralize the wafer surface and minimize the wafer charging effect. If these electrons enter the ion beam stop and hit the Faraday detector, current reading can be reduced and the accuracy of the beam current measurement can be affected. Permanent magnets are used to generate a magnetic field to prevent electrons from entering the beam stop due to their small gyroradii. The magnetic field can also prevent secondary electrons emitted from the graphite surface from backstreaming to the postacceleration electrode and causing damage there.

8.4 Ion Implantation Process

The three main issues in ion implantation processing are: dopant type, which is determined by ion species; junction depth, which is determined by ion energy; and dopant concentration, which is controlled by the combination of ion current and implantation time.

8.4.1 Device applications

In IC chip manufacturing, many ion implantation processes are involved to make billions of tiny transistors on a silicon wafer surface. Due to the different requirements of dopant concentration and junction depth, ion energy and beam currents for these implantation processes are quite different. In an advanced semiconductor fab, different types of implanters are employed to meet these requirements.

Table 8.9 lists a few examples of low-energy, high-current ion implantation applications in 32-nm CMOS ICs. Symbols and numbers in the right column are dopant isotopes or molecules, ion energy in keV, and the dosage in ions/cm^2. For example, pMOS S/D implantation, B11/0.4/1E15, indicates that it uses ions of boron 11 (^{11}B) at 0.4 keV with 1×10^{15} ions/cm^2 dosage. nMOS SDE, As2_150/2/5E14, indicates that it uses ions of a gaseous arsenic molecule (As_2) of atomic weight 150, or two ^{75}As, with 2 keV and 5×10^{14} ions/cm^2 dosage.

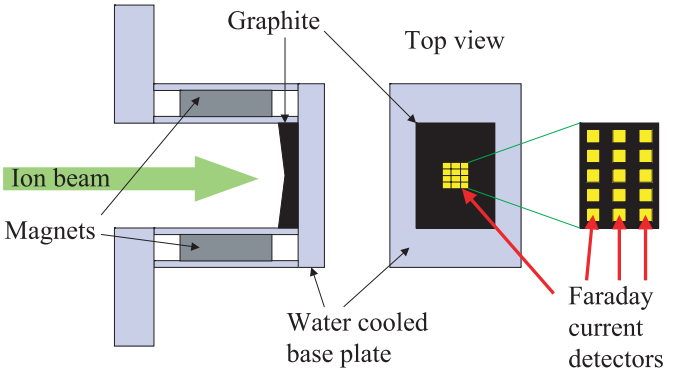


Figure 8.33 Schematic of a beam stop.

Table 8.9 Examples of 32-nm CMOS ion implantation.

Implantation step	Implantation conditions
pMOS S/D extension	B11/0.4/1E15
pMOS S/D	B11/2/3E15
nMOS S/D extension	As2_150/2/5E14
nMOS S/D	P31/4/3.5E15

Fig. 8.34 illustrates a well implantation process, which is a high-energy ion implantation process because it needs to form deep junctions called wells or tubs on which MOSFETs can be built. An nMOS is built on p-wells, and a pMOS is built on n-wells. Well implantation usually requires a high-energy ion implanter.

Antipunch-through implantation is also called mid-well implantation and is used to suppress the punch-through effect, which can cause transistor breakdown. Large angle tilt (LAT, normally 35 to 45 deg) implantation, or halo implantation, is also used for punch-through suppression for many IC devices.

Threshold implantation is also called V_T adjustment implantation, which is a low-energy, low-dosage implantation process. Threshold implantation determines at what voltage a MOSFET can be turned on or off. This is called threshold voltage, or V_T, which can be expressed as

$$V_T = \Phi_{ms} - Q_i/C_{ox} - Q_d/C_{ox} + 2\Phi_f \tag{8.1}$$

where Φ_{ms} is the potential difference between the gate material and the semiconductor substrate. In the case of polysilicon gates, the potential difference is controlled by poly-dope implantation. Q_i is the interface charge, which is determined by preoxidation clean and gate oxidation processes; $Q_d = -2(k_{Si}eN_c\Phi_j)^{1/2}$ is the depletion charge and is determined by V_T adjustment implantation, which controls majority carrier concentration N_c. $C_{ox} = k_{ox}/t_{ox}$ is the unit gate capacitance and is determined by gate dielectric material k_{ox} and gate dielectric thickness t_{ox}. Φ_f is the

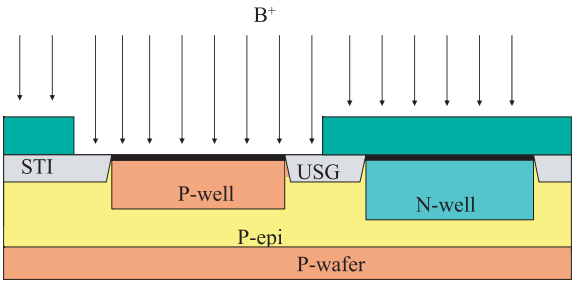


Figure 8.34 Well implantation.

Fermi potential of the substrate (silicon for most cases). Threshold voltage is one of the most important characteristics of MOSFETs, and V_T adjustment implantation is one of the most critical ion implantation processes.

Some old electronic devices require a 12-V dc power supply, many electronic circuits need 5 or 3.3 V, and advanced IC chips operate at 1.0 V. Ultralow-power IC chips can even operate at 0.4 V. These operation voltages must be higher than threshold voltages to guarantee that these transistors can be turned on or off; however, they cannot be so high that they break down the gate oxide and destroy the transistors. Figure 8.35 shows the threshold implantation of a CMOS IC chip. V_T adjustment implantations are performed with well implantations using the same implanter in the same processing sequence with significantly lower energy, as shown in Fig. 8.35.

Polysilicon needs to be heavily doped to reduce resistivity. This can be achieved either by introducing dopant gases to the CVD reactor with the silicon precursor for in-situ doping during the deposition process, or by high-current ion implantation. For advanced CMOS chips, ion implantation is preferred because it can dope pMOS and nMOS polysilicon gates separately. Polysilicon gates of pMOS are heavily p-type doped, and polysilicon gates of nMOS are heavily n-type doped. This allows better device property control. It also generates p–n junctions on the polysilicon lines that form local interconnections for connecting neighboring gates at the junctions of the p- and n-type MOSFET of the CMOS circuit. These p–n junctions must be shorted by the silicide formed on the top of the polysilicon lines in a later salicide process. Otherwise, they can cause intolerably high resistance between neighboring gates.

Poly-dope implantations need two masks, one for nMOS and another for pMOS. To reduce production cost, a poly-counter doping process has been developed and applied in IC production. It uses an implanter to heavily dope n-type dopants on polysilicon layers of entire wafers without a mask. It then patterns the wafer to expose the pMOS polysilicon layer and heavily dope it with p-type dopant. The p-type dopant concentration is so high that the polysilicon reverses from n- to p-type. Plasma doping systems, which can achieve much higher dopant concentrations than high-current ion implanters, have been developed and used

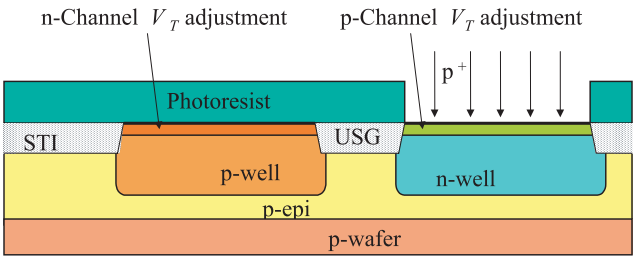


Figure 8.35 V_T adjustment implantation.

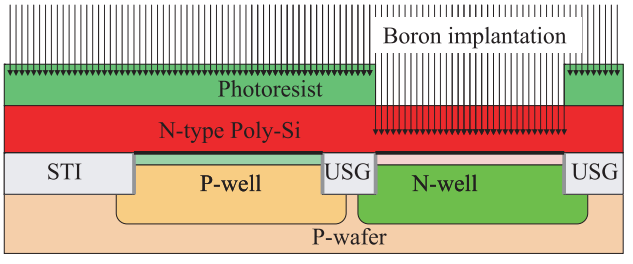


Figure 8.36 Poly-dope implantation.

for this application. Figure 8.36 shows a heavy p-type (boron) poly-counter dope process.

When transistor size shrinks, the effects of dopant diffusion in polysilicon gates can become evident and affect performance of microelectronic devices. It is very important to suppress dopant diffusion, especially to prevent boron in pMOS polysilicon gates from diffusing into nMOS polysilicon gates; otherwise, transistor characteristics can change. A diffusion blocking implantation is introduced, where a high dosage of nitrogen is implanted into the polysilicon to trap boron atoms and prevent them from diffusing too far from the polysilicon.

Lightly doped drain (LDD) is a low-energy, low-current implantation process by which a shallow junction is formed near the gate to reduce vertical electric field near the drain. LLD is required for submicron MOSFETs in order to suppress the hot electron effect, which can cause device degradation and affect chip reliability. The hot electron effect (or hot carrier effect) occurs when electrons tunnel through ultrathin gate oxide from drain to gate due to acceleration of the vertical component of the electric field induced by S/D bias and short channel. Because the dopant concentration of this implantation has increased with diminishing feature size, for sub-0.25-μm devices, the dosage has become high enough that the process can no longer be called LDD. It has been renamed SDE implantation, which provides a buffer for high-concentration S/D dopants. SDE has the shallowest junction depth in an IC chip and requires a low-energy implanter. Figure 8.37 illustrates the SDE implantation process for CMOS IC chip processing.

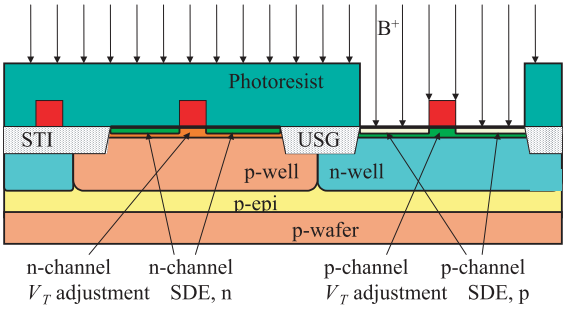


Figure 8.37 SDE implantation.

S/D implantation is a high-current, low-energy implantation process. It is the last implantation process of a MOSFET in CMOS IC chip fabrication processes. One significant difference between S/D and SDE implantation is that the dosage of S/D implantation is higher, and S/D is performed after sidewall spacer formation. Sidewall spacers keep the heavily doped S/D slightly apart from the channeling area directly underneath the gate and help suppress the hot electron effect. Figure 8.38 illustrates the S/D implantation process.

S/D implantation uses a high-current ion beam to heavily dope silicon. The masking photoresist receives a high concentration of dopant atoms after S/D implantation. This can create difficulties for the photoresist dry strip process. Oxygen radicals can oxidize and remove the photoresist, which is mainly hydrocarbon polymer. However, most dopant oxides, such as phosphorus and boron oxides, are solids instead of gases. They tend to stay on the wafer surface and cause defective residue called scum. A wet process is required after dry strip to remove these residues. This process is called descuming. In some fabs, the masking photoresist for S/D implantation is stripped in a wet process with a strong oxidizer solution, such as H_2O_2.

For the older technology with LOCOS isolation, ion implantation is normally used to form a p-type doped isolation region before the thick field oxide growth. This process is called isolation implantation or channel stop implantation. It is also used to form guarding rings around the active regions; these rings help to electrically isolate neighboring transistors.

For CMOS processes, almost every type of implantation process needs to be performed twice, once for pMOS and once for nMOS. A state-of-the-art CMOS IC device requires up to 20 implantation processes to create the multiple tiny transistors. For bipolar and BiCMOS IC chips, ion implantation is widely applied for buried layer doping, isolation formation, and base, emitter, and collector formations.

Unique DRAM ion implantation applications are contact implantations that reduce contact resistance between the polysilicon plugs and the silicon substrate, and between two polysilicon plugs. It uses high-current p-type implantation to heavily dope the silicon or polysilicon at the bottom of a contact hole after contact

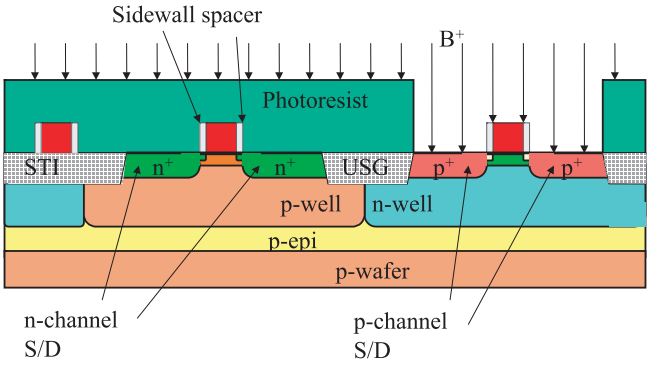


Figure 8.38 S/D implantation.

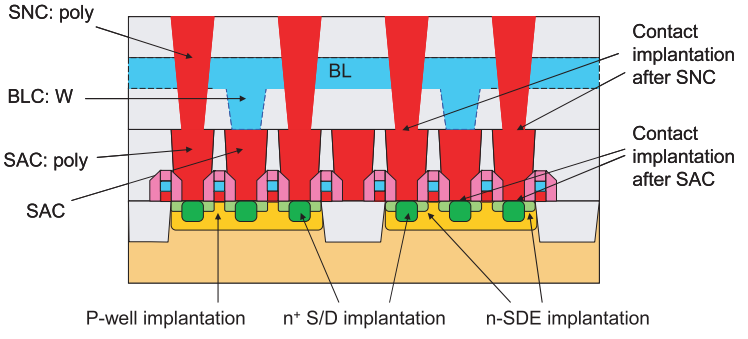


Figure 8.39 Ion implantation applications in DRAM cell array transistors and connections.

etches. Figure 8.39 illustrates the applications of ion implantations in cell array MOSFET formation and contact implantations for connection.

Since the ion implantation process is directly related to microelectronic devices, some background in device physics is helpful to better understand the process. Readers are strongly recommended to acquire information about semiconductor devices from the books by Streetman and Banerjee (entry level) and *Physics of Semiconductor Devices* by Sze (advanced level) given in the Bibliography section at the end of this chapter.

8.4.2 Other ion implantation applications

As device dimensions continue to shrink, soft error problems (due to trace amounts of natural background α-decay) become more and more pressing, especially for memory chips. Every α-particle can generate more than one million electron–hole pairs in the silicon substrate. Capacitance of a memory capacitor or transistor of a memory circuit must be large enough that stored data will not be overwritten by the surging electrons coming from the α-particle-induced electron–hole pairs

when α-decay happens. Figure 8.40 illustrates the process of electron–hole pair generation by an α-particle.

Capacitance can also decrease when device dimension shrinks. One approach to solve this problem is by using a silicon-on-insulator (SOI) substrate. Figure 8.41 shows a MOSFET circuit built on a SOI substrate.

Figure 8.41 shows that each transistor is created on its own silicon substrate tub and is completely isolated from neighboring transistors and the bulk silicon substrate. Therefore, the possibility of crosstalk, latch up, and soft error are completely eliminated. SOI-based IC chips can work in extreme conditions, such as high-radiation environments, in which regular IC chips cannot work.

One approach for making a SOI substrate is to use high-energy, high-current ion oxygen implantation (up to 10^{18} ions/cm^2) to form an oxygen-rich layer in the silicon substrate. A high-temperature annealing process causes a chemical reaction between silicon and oxygen atoms and forms a buried silicon dioxide layer, while the implant-damaged silicon lattices near the surface recover the single-crystal structure. This process is called separation by implantation of oxygen, or SIMOX (see Fig. 4.26).

Another technique for making SOI wafers also involves high-current ion implantation. The Smart Cut™ (trademarked by Soitec) uses hydrogen ion implantation to generate a hydrogen-rich layer in an initial wafer with oxide growth on its surface, which bonds with a second wafer in a thermal process. During the

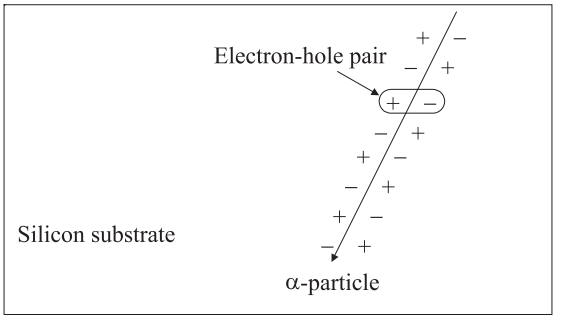


Figure 8.40 α-particle-induced electron–hole pairs.

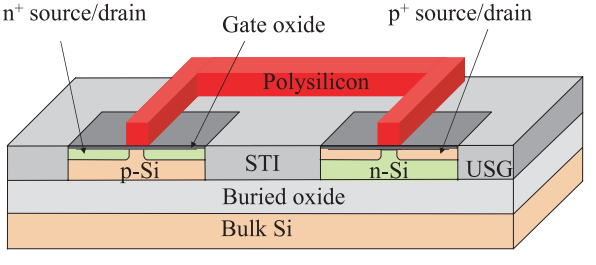


Figure 8.41 MOSFET on the SOI substrate.

bonding process, the hydrogen-rich layer splits the first wafer, leaving thin layers of silicon and oxide on the surface of the second wafer (see Fig. 4.27). Because SOI wafers made by Smart Cut™ technology have a lower cost, especially for the thick buried oxide (BOX) layer, they dominate the SOI wafer market.

Ion implantation has been used to harden photoresist to improve its resistance during etch processes in IC fabs. Ion implantation is also used as one possible way to freeze the first photoresist pattern in a litho-freeze-litho-etch (LFLE) style of double-patterning technology (see Fig. 6.59).

In pattern media research and development in the hard disk drive (HDD) industry, ion implantation is one candidate for isolating magnetism patterns; other candidates include metal etch, nonmagnetic material deposition, and planarization. Figure 8.42(a) illustrates the patterning process based on etch, and Fig. 8.42(b) shows the patterning process based on ion implantation. It can be seen that the ion-implantation-based process potentially could have a lower cost than the etch-based process; however, there are still plenty of issues that need to be solved.

Another possible application of ion implantation is EUV lithography mask manufacturing. It is known that ion implantation can cause degradation of reflectivity of the multilayer. By using ion implantation damaging patterns instead of TaBN absorber patterns, shadowing effects from the TaBN absorber can be avoided during EUV lithography due to the reflective nature of EUV masks. If EUV lithography can be implemented in mass production for IC manufacturing, the technology node most likely will become 10 nm, which means the clear gap

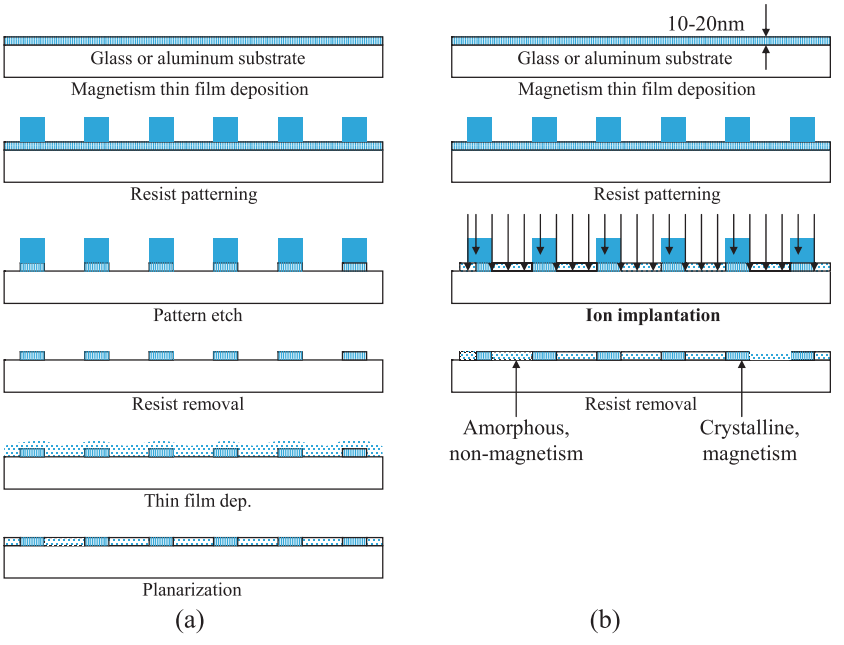


Figure 8.42 Pattern media processing sequences: (a) etch versus (b) ion implantation.

would be only 40 nm on the mask, the shadowing effect from the 6-deg incident angle would become very significant for EUV masks with absorber patterns, and ion implantation would be needed. Figure 8.43 shows these two EUV masks.

Ion implantation can also be used in solar panel manufacturing. By using external hard masks, dopant ions can be implanted into a designated area without photolithography processes, potentially saving solar cell manufacturing costs and driving down solar panel prices.

8.4.3 Processing issues

There are many processing challenges for ion implantation processes, such as wafer charging, contamination control, and process integration.

8.4.3.1 Wafer charging

Wafer charging can cause gate oxide breakdown. For example, the dielectric strength of silicon dioxide is about 10 MV/cm. Theoretically, when the surface charge is 2.2×10^{13} ions/cm^2, the electric field inside the oxide can be high enough to break down the 100-Å gate oxide if there is absolutely no electron leaking. Since electrons can always tunnel through thin dielectric layers due to the quantum effect, 100-Å oxide could survive dosages of up to 6.2×10^{18} ions/cm^2 as long as the gate voltage is lower than the breakdown voltage. For 100 Å, gate voltage is 10 V.

As transistor geometry continues to shrink, the thickness of the gate oxide becomes thinner and thinner. It can be as thin as 12 to 15 Å for 65-nm generation devices, which require even better charge neutralization techniques. Wafer charging can be monitored by several mechanisms: test wafers with capacitors, erasable programmable read-only memory (EPROM), and transistor structures are the most commonly used tools. These structures can be built on wafers specially made for the charging test. In-situ charging sensors are developed and applied in implanters next to the wafer to monitor charging conditions on the wafer surface. Figure 8.44 shows an antenna capacitor charging test structure—the ratio of the polysilicon pad area to the thin oxide area is called the antenna ratio, which can be as high

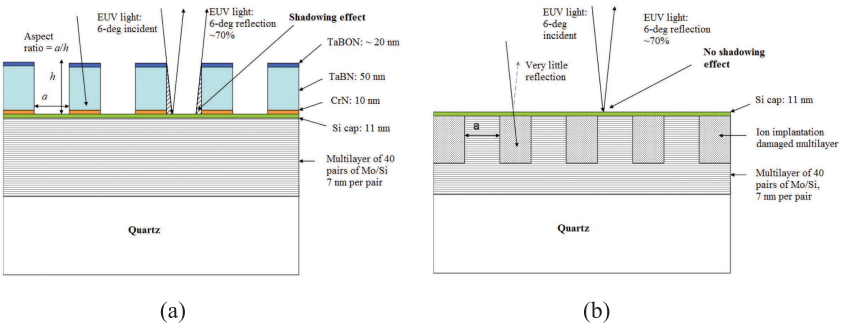


Figure 8.43 (a) A normal EUV mask with absorber patterns that have shadowing effects, and (b) an EUV mask with ion implantation patterns that do not have shadowing effects.

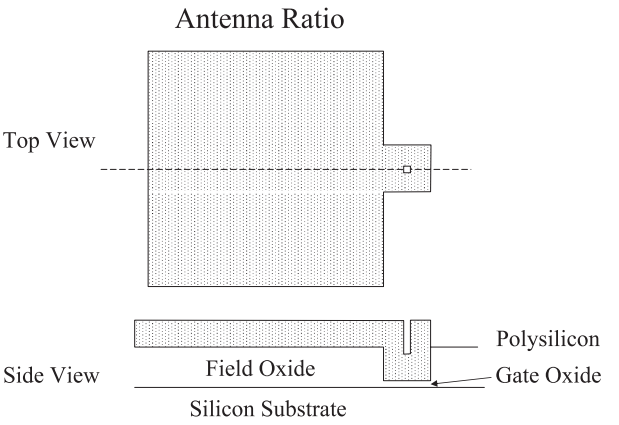


Figure 8.44 An antenna-style capacitor.

as 100,000:1. The larger the antenna ratio, the easier it is to breakdown the thin gate oxide.

The main factors that affect wafer changing are the beam current, beam scan width, disk or spin wheel radius, and spin rate. Reducing the ion beam current obviously will reduce the charging effect. The tradeoff is that it will also reduce wafer throughput. Increasing the beam scan width, disk or spin wheel radius, and spin rate effectively reduces the dosage per wafer per pass, thus reducing the charging effect. By using a large beam scan width, large disk or spin wheel radius, high spin rate, and an electron flooding system (which flows electrons to the wafer surface with the ion beam), advanced implanters have successfully addressed the charging issue for several generations of nanometer IC devices.

8.4.3.2 Particle contamination

Particle contamination is always an area of concern in almost every IC processing step. Large particles on a wafer surface can block ion beams, especially for low-energy ion implantation processes such as threshold, SDE, and S/D implantations, and cause incomplete dopant junction, which can be very harmful to IC yield. When device dimensions decrease, the size of killer particles also decreases. Figure 8.45 illustrates a killer particle causing incomplete junction implantation by shielding part of a substrate from the incident ion beam.

For many implantations still using spin wheel batch implanters, large particles can destroy patterns on a wafer's surface, similar to when a high-speed missile collides with the wall of a building. Because wafers require 15 to 20 different ion implantation processes in advanced IC chip fabs, the effects of particle-induced yield loss is accumulative. Therefore, it is extremely important to minimize the addition of killer particles in each implantation process.

Particles can be measured by scanning the wafer surface with a laser beam, and then using a photomultiplier tube to collect, transfer, and amplify the particle-induced scattered light signal. The number and location of particles of a certain size

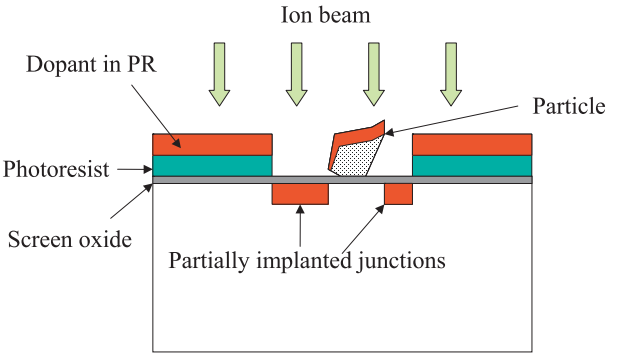


Figure 8.45 Effect of particle implantation.

or larger are measured immediately before and immediately after the implantation process; the difference in particle count is called particle adding. Locations of the added particles are also recorded, as the signatures of particles of different sizes are very useful in determining the source or sources of the particles.

Particles can be mechanically introduced by the wearing down of moving parts, typically valves and seals, clamps, and loading robots. They can also be introduced by processing. For example, arsenic, phosphorus, and antimony vapors can recondense along the beam line, and the flaking residues can reach the wafer surface during the pump-down cycle. Energetic ion sputtering is also a major particle source. Aluminum and carbon sputtered from the beam line and beam stop also can contribute to particle adding in implantation processes. A silicon wafer itself can be a particle source whenever there is wafer chipping or breakage. Photoresist film is brittle, and improper EBR in photolithography processes can leave photoresist on the edges of wafers. During wafer transfer and handling, robot fingers and wafer holder clamps can crack the photoresist on the edge, chip it off, and generate particles.

Both implanter design improvement and routine PM contribute to the minimization of particle adding in ion implantation processes. Statistic process control (SPC) methods can be applied to identify major contamination sources and help improve process control.

8.4.3.3 Elemental contamination

Elemental contamination is the coimplantation of other elements within an intended dopant. For example, both double-charged molybdenum $^{94}Mo^{++}$ and boron fluoride ions $^{11}BF_2^+$ have the same mass/charge ratio (AMU/e = 49); therefore, a mass analyzer cannot separate them, and $^{94}Mo^{++}$ can implant into a silicon wafer along with $^{11}BF_2^+$ to cause heavy metal contamination. This is the reason ion sources cannot use standard stainless steel, which contains molybdenum. Other materials such as graphite and tantalum are usually used.

If there is a small air leak, nitrogen can slip into an ion source chamber; $^{28}N_2^+$ ions have the same mass/charge ratio as silicon ions $^{28}Si^+$, which are used for

preamorphous implantation. Similarly, outgassing of an ion source chamber wall can release carbon monoxide; when it becomes ionized, its ion also has the same mass/charge ratio (AMU/e = 28).

A mass analyzer also has difficulties separating ions with very close mass/charge ratios, due to insufficient resolution. $^{75}As^{+}$ ions can contaminate $^{74}Ge^{+}$ or $^{76}Ge^{+}$ ions in germanium preamorphous implantations, and $^{30}BF^{+}$ ions can contaminate $^{31}P^{+}$ ion implantation.

Other elemental contamination can be caused by the sputtering of beam line and wafer holder materials, such as aluminum and carbon, which can implant these ions into the wafer. Aluminum and carbon in the silicon substrate can cause device degradation.

8.4.4 Process evaluation

Dopant species, junction depth, and dopant concentration are the most important factors for ion implantation. Dopant species can be determined by the mass analyzer of an implanter, and dopant concentration is determined by ion beam current multiplied by implantation time. In ion implantation process monitoring, a four-point probe is commonly used to measure sheet resistance on a silicon surface. Sheet resistance R_s is defined as $R_s = \rho/t$ for ion implantation processes, where resistivity ρ is determined by dopant concentration, and thickness t is determined by the doping junction depth, which is controlled by dopant ion energy. Therefore, sheet resistance measurements can give information about dopant concentration, since junction depth can be estimated by knowing the ion energy, ion species, and substrate materials.

8.4.4.1 Secondary ion mass spectroscopy

By using a primary heavy ion beam to bombard a sample surface and collect a mass spectrum of sputtered secondary ions at different times, fabs can measure the dopant species, dopant concentrations, and the depth profile of the dopant concentration. Secondary ion mass spectroscopy (SIMS) is a standard measurement of ion implantation because of its ability to measure all factors critical to the ion implantation process. However, it is destructive; the sputtered spot size is large and very slow. SIMS is widely used in laboratories as the golden tool for off-line ion implanter tuning and early ion implantation process development. It is not used as an in-line monitoring system for ion implantation process control. Figure 8.46(a) illustrates how SIMS works; the depth of the profile can be calibrated with sputtering time using profilometer measurements. Figure 8.46(b) shows a SIMS measurement result of 1-keV ^{11}B ion implantation on a silicon wafer.

8.4.4.2 Four-point probe

A four-point probe (see Fig. 5.37) is the most commonly used tool employed to measure sheet resistance. Since sheet resistance is related to dopant concentration, this measurement provides useful information for ion implantation process control. By applying a certain amount of current between two of the pins and measuring the

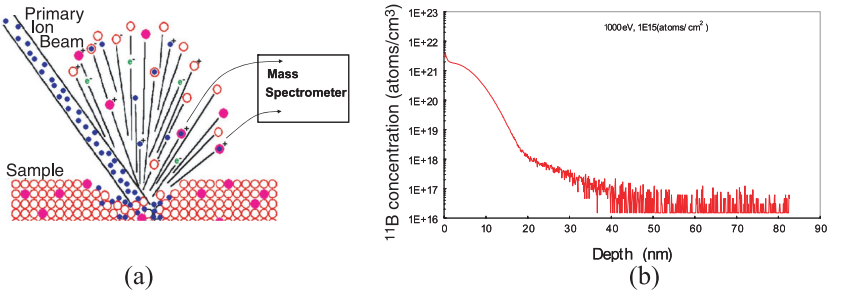


Figure 8.46 (a) SIMS and (b) an example of SIMS results for ^{11}B implantation (Advanced Ion Beam Technology, Inc.).

voltage difference between the other two pins, sheet resistance can be calculated. Four-point probe measurement is performed after the annealing process activates the dopants. Since the four-point probe makes direct contact with the wafer surface, it is only used on test wafers for process development, qualification, and control. During the measurement, the probes must come in contact with the silicon surface with enough force that the pins break through the thin (~10 to 20 Å) native oxide and make solid contact with the silicon substrate. Figure 8.47 shows an example of a four-point probe measurement of an ion implantation wafer after annealing.

8.4.4.3 Thermal wave

Another common ion implantation process monitoring technique is the thermal wave probe system. In a thermal wave system, an argon pump laser generates thermal pulses on the wafer surface, and an He-Ne probe laser measures the dc reflectivity R and reflectivity modulation ΔR induced by the pump laser at the same spot. The ratio of the two, $\Delta R/R$, is called a thermal wave (TW) signal, which is related to crystal damage, and the crystal damage in turn is a function of the implant dose. Figure 8.48 illustrates a thermal wave system.

A thermal wave measures the degree of crystal damage inflicted by ion implantation. The measurement is performed immediately after ion implantation, prior to the annealing process. This is one advantage of a thermal wave over a four-point probe, which needs annealing to active dopants prior to the measurement. Unlike a four-point probe, which requires firm contact with a wafer surface for reliable measurement, a thermal wave probe measures the wafer without physical contact. Therefore, thermal wave measurements are nondestructive and can be used to measure production wafers. One disadvantage of thermal wave measurements is that they have low sensitivity at low dosages; for example, with arsenic and phosphorus implantation dosage around 10^{12} ion/cm^2, 10% dosage change only causes a 2% change of TW signal. Another disadvantage is the drift of the TW signal over time due to room temperature annealing (or ambient annealing). Thus, a thermal wave measurement needs to be taken as soon as implantation is finished. Wafer heating caused by the laser beam during measurement also accelerates relaxation of the damage, which can change the reflectivity of the substrate. This

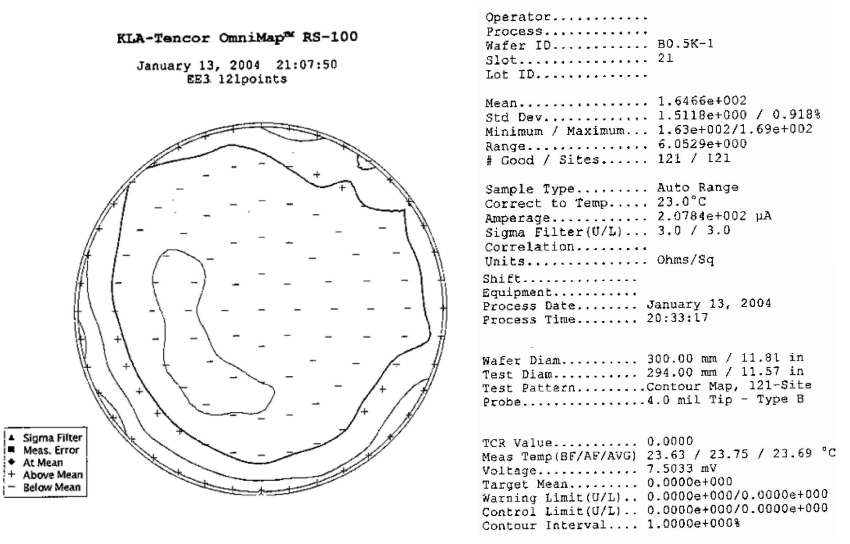


Figure 8.47 Example of a four-point probe measurement result (Advance Ion Beam Technology, Incorporated).

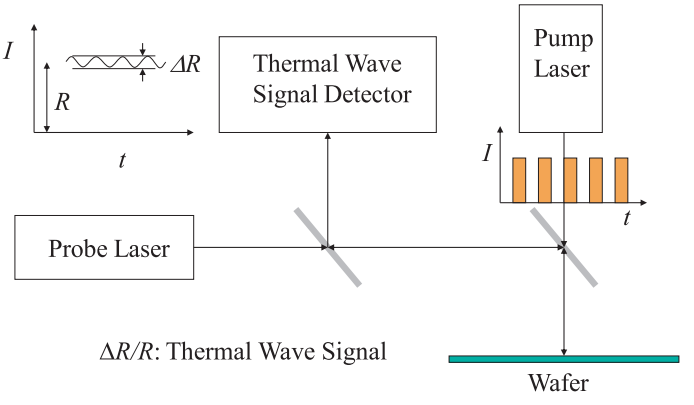


Figure 8.48 Thermal wave system.

is a standard case: the measurement process disturbs what is supposed to be measured. Therefore, inherently the process does not have very high measurement accuracy. Many factors can affect thermal wave measurement, such as ion beam current, ion beam energy, wafer patterns, and thickness of the screen oxide. The primary advantage of the thermal wave is that it can measure production wafers. It gives process engineers a very useful tool for better process control by measuring production wafers immediately after the implantation process without the long wait that other process monitors require.

8.4.4.4 Optical-electrical sheet resistance measurement

Optical-electrical sheet resistance measurement uses a pulse laser to illuminate semiconductor substrates and generate electron–hole pairs. The electron–hole pairs disperse to sensor electrodes that can detect voltage change caused by diffusion of the carriers. Diffusion rate is related to sheet resistance, and the measured voltage ratio V_1/V_2 is almost linearly related to sheet resistance. Figure 8.49 illustrates the processes of laser pulse and carrier diffusion, which allow measurement of the voltage ratio of the first and second sheet resistance R_s using this optical-electrical system.

8.5 Safety

All ion implanters use hazardous solids and gases, which are toxic, poisonous, flammable, explosive, or caustic. Very high electrical voltage (normally up to 250,000 V) is applied to produce the desired processing results.

8.5.1 Chemical hazards

Both solid and gaseous dopant sources are used in ion implantation processes. Antimony, arsenic, and phosphorus are the common solid sources, and arsine, phosphine, and boron trifluoride are the commonly used gases.

Antimony (Sb) is a brittle, silvery-white metallic element used as an n-type dopant in implantation processes. It is poisonous; direct contact with solid antimony can cause irritation of the skin and eyes. Antimony dust is extremely toxic; direct contact with it can cause severe irritation of the skin, eyes, and lungs. It can also cause heart, liver, and kidney damage.

Arsenic (As) is poisonous; direct contact with solid arsenic can cause irritation of the skin and eyes, and it also can cause discoloration of the skin. Arsenic dust is very toxic; direct contact can cause irritation of the skin and lungs. It can also cause nose and liver damage, and possibly lung and skin cancer.

Red phosphorus (P) is a commonly used solid material in implantation processes as an n-type dopant and can also be found on the side of matchboxes in the form of a dark stripe. It is flammable and can ignite with friction. Direct contact can be irritating to the skin, eyes, and lungs.

Arsine (AsH_3) is commonly used as an arsenic source gas; it is one of the most poisonous gases used in the semiconductor industry. At 0.5 to 4 ppm, a garlic-like

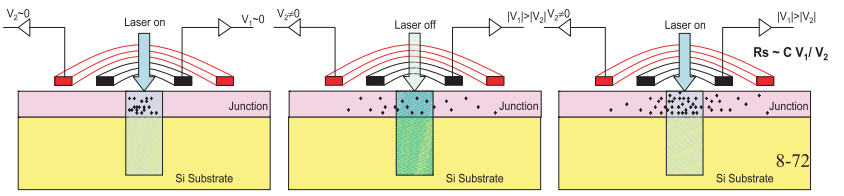


Figure 8.49 Optical-electrical measurement.

odor can be detected; at 3 ppm it can cause an immediate danger to life and health (IDLH). This is the reason that the smell of garlic is the most feared smell inside an IC fab. It can cause irritation to the nose and eyes at low concentrations. Exposure at 500 ppm for a few minutes can be fatal. Arsine is also flammable and becomes explosive when its concentration in air is 4 to 10%.

Phosphine (PH_3) is commonly used as a phosphorus source gas. It is flammable and becomes explosive when its concentration in air is higher than 1.6%. Phosphine is a poisonous gas with a rotten fish odor, which is detectable at 0.01 to 5 ppm; IDLH limit is 50 ppm. With exposure at low concentration, it can be irritating to the eyes, nose, and lungs. Exposure at 10 ppm can cause headache, breathing difficulties, cough and chest tightness, loss of appetite, stomach pain, vomiting, and diarrhea.

Boron trifluoride (BF_3) is commonly used as a boron source gas. It is corrosive because it forms hydrogen fluoride acid when it comes in contact with water. Exposure to it can cause severe skin, eye, nose, throat, and lung irritation; it also can cause a build up of fluid on the lungs.

Decaborane ($B_{10}H_{14}$) is used for molecular ion implantation for USJ formation. It is a solid with low vapor pressure at room temperature. It is toxic and when absorbed through skin can affect the central nervous system. The Occupational Safety and Health Administration (OSHA) permissible exposure limit (PEL) for decaborane is 0.05 ppm ($0.3 mg/m^3$).

Octadecaborane ($B_{18}H_{22}$) and carborane ($C_2B_{10}H_{12}$ or CBH) are two other promising dopants for forming boron USJ with high throughput materials for molecular ion implantation. Both of them are solid materials with very low vapor pressure at room temperature.

8.5.2 Electrical hazards

Contact with high voltage and current flow can cause electrical shock, skin burns, muscle and nerve damage, heart failure, and death. About 1 mA of current across the heart can be fatal. Statistics show that death rates from contact with a 250-V ac power line is about 3%. Predictably, this percentage increases dramatically when the voltage of the electrical power supply is higher than 10 kV.

The spark breakdown voltage of air is about 8 kV/cm. For an implanter with an acceleration electrode powered at 250 kV, the breakdown distance is about 31 cm. With a sharp point, breakdown distance can be even longer. Therefore, safety interlocks are required for ion implanters to prevent electrical power from ramping up when the implanter is not fully shielded.

Because high voltage can generate large amounts of electrostatic charge, which can cause a dangerous electrical shock if contacted without full discharge, grounding bars are required for making contact and discharging all parts before any contact is made.

Ion implanters are fully enclosed systems and are so big that workers can easily be hidden inside them without others noticing. Therefore, before entering a system, it is very important to work with a partner and tag out of the system, to make sure

others know someone is working inside and not to start the equipment and raise the voltage. Employees must always take a door key when entering an implanter so that other employees cannot lock the door for system start up.

8.5.3 Radiation hazards

When an energetic ion beam hits a wafer, slits, beam stop, or anything along the beam line, energy loss from the ion emits out in the form of x-ray radiation. A safety interlock is needed to prevent systems from being turned before the system is completely shielded.

Electrons generated by ion collisions with neutral atoms along the beam line and by secondary electron emission from the solid surface can also be accelerated by the electric field due to dc voltage of the accelerating electrodes. Suppression electrodes are used to prevent these electrons from accelerating to a high energy, becoming backstreamed to bombard the ion source and other beam line parts, and causing x-ray radiation and parts damage.

8.5.4 Mechanical hazards

A spin wheel and spin disk can rotate up to 1250 rpm; the velocity of the wafer can be as fast as 90 m/sec (~220 mph) while in full-speed spin. In the event of malfunction, these systems can release huge amounts of energy and cause extensive damage. Constant monitoring of the vibration magnitude of the spin wheel or disk is necessary to make certain that the spin wheel or disk can be stopped before failure happens. Spin and scan motors are very powerful; any attempt to adjust something when these motors are still active could cost fingers and arms.

8.6 Recent Developments and Applications

As the minimum feature size of devices continues to reduce, the junction depth of the channel and S/D of the MOSFET will become shallower and shallower. USJ ($x_j \leq 0.05$ μm) formation raises a big challenge for both ion implantation and annealing, especially for p-type USJ, since it requires low- to high-current, pure boron ion beams at very low energy. Requirements for the USJ are shallow junctions, low sheet resistance, low contact resistance, compatibility with silicide with low diode leakage current, minimal impact on channel profile, and compatibility with polysilicon or high-κ/metal gates. Other requirements are low cost, good within-wafer and wafer-to-wafer uniformity, low added particle count, and reliable transistors and contacts.

In addition to developing single-wafer implanters with high-current and ultralow energy (as low as 0.1 keV) with high energy purity, scientists and engineers have also searched for other approaches for USJ formation. Molecular ion implantation is one of them. In regular ion implantation, BF_3 is commonly used for p-type shallow junction implantation (not B, because the BF_2^+ ion is larger and heavier). Decaborane ($B_{10}H_{14}$), octadecaborane ($B_{18}H_{22}$), and carborane ($C_2B_{10}H_{12}$ or CBH) are the large molecules used for these studies. There are several benefits to

using large molecules for USJ formation. It is very difficult to achieve high energy purity for 0.1-keV ion beams because the ion beam must be accelerated to about 5 keV for a magnetic mass spectrum analyzer to effectively isolate the desired ion species, then be decelerated by 4.9 keV. The power supply voltages for acceleration and deceleration must be very accurate and stable to achieve energy purity. Because large molecules are significantly bigger and heavier, they require a higher energy to achieve the same junction depth as BF_2^+ or B^+ ions. It is much easier to achieve high energy purity for a 1-keV (or higher) energy ion beam than it is for a 0.1-keV ion beam. These larger molecules also consist of many boron atoms; for example, CBH has ten boron atoms, thus CBH molecular ion implantation can achieve ten times higher throughput than BF_2^+ or B^+ ion implantation at the same beam current. Large-molecule implantation also causes more lattice damage, thus it has less tunneling effect and better junction profile control.

A plasma immersion ion implantation (PIII) or plasma doping (PLAD) system has been developed for targeting low-energy, high-dosage applications such as USJ and deep trench. The PIII system usually has a plasma source power for generating high-density plasma to ionize dopant gas, and a bias power to accelerate dopant ions into a wafer surface, as shown in Fig. 8.50. The most commonly used dopant gas in PLAD is B_2H_6, used for boron doping. Source power can be rf or MW. B_2H_6 can quickly dope a wafer with a very high dose; that dosage can be so high that even at the highest beam current, a beam-line-based implanter would require such a long implantation time that throughput requirements would not be met. The tradeoff of PLAD is that it cannot select ion species or precisely control ion flex or dosage. Therefore, its main applications are in high-dose, noncritical layers. It has been widely used in DRAM chip manufacturing for polysilicon counter-doping of periphery devices. It also could be used in contact implantations of array devices of DRAM.

In plasma immersion systems, dopant ions bombard the wafer and are implanted into the substrate, dopant ion flux is mainly controlled by MW power, and ion

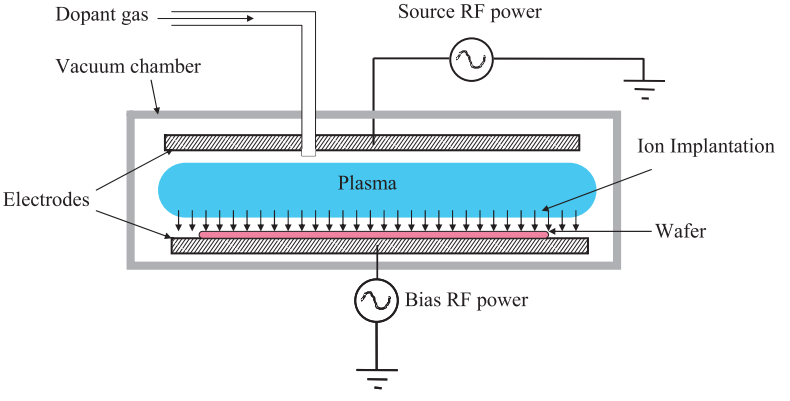


Figure 8.50 Plasma immersion ion implantation (or plasma doping) system.

energy is mainly determined by the bias rf power. Current through the magnets can affect resonance position, thus it can be used to control the position of plasma, which in turn controls dopant uniformity.

PIII is a low-energy process; its ion energy is less than 1 keV. Therefore, it can be applied for USJ formation for sub-0.1-μm device applications. A plasma immersion system cannot select specific ion species; this is a disadvantage compared with standard ion implantation. Other disadvantages are that ion flux can be affected by plasma position and chamber pressure, and that ion energy is distributed in a wide range instead of in a sharp narrow distribution like the ion implanter. Thus, PIII has difficulties precisely controlling dopant concentration and junction depth.

8.7 Summary

- Dopants commonly used in IC chip fabrication are boron for p-type and phosphorus, arsenic, and antimony for n-type.
- CMOS processes require many ion implantations, such as well/threshold implantations, LDD or SDE implantations, poly-dope implantations, and S/D implantation processes, one each for pMOS and nMOS.
- In addition to the above implantations, DRAM chip manufacturing has unique contact ion implantation.
- Ion implantation processes can precisely and independently control dopant junction depth and dopant concentration. Ion implantation is an anisotropic process, which is required for submicron processes.
- An ion implanter consists of gas, electrical, vacuum, control, and beam line systems.
- A beam line system consists of an ion source, extraction electrodes, mass analyzer, postacceleration assembly, charge neutralization system, wafer handler, and beam stop.
- Formation of abnormally deep junctions caused by long ion travel ranges in single-crystal channels is called the channeling effect. Wafer tilt and screen oxide are two commonly used methods to minimize this effect.
- For the same ion species, the higher the ion energy, the longer the ion range. At the same energy level, the lighter the ion, the longer the ion range.
- Implantation causes crystal structure damage; postimplantation thermal annealing is needed to recover the single-crystal structure and activate the dopant.
- Ion implantation processes use many hazardous solids and gases, most of which are poisonous, flammable, explosive, or corrosive. Other safety hazards are high electrical voltage, x-ray radiation, and moving parts.

8.8 Bibliography

D. G. Baldwin, M. E. Williams, and P. L. Murphy, *Chemical Safety Handbook for the Semiconductor/Electronics Industry*, 2nd ed., OME Press, Beverly, MA (1996).

S. A. Cruz, "On the energy loss of heavy ions in amorphous materials," *Radiat. Eff.* **88**, 159 (1986).

M. I. Current, *Basics of Ion Implantation*, Ion Beam Press, Austin, TX (1997).

T. Roming, J. McManus, K. Olander, and R. Kirk, "Advanced in ion implanter productivity and safety," *Solid State Technol.* **39**(12), 69-74 (1996).

W. Shockley, *Forming Semiconductor Devices by Ion Bombardment*, U.S. Patent No. 2787564 (1954).

B. G. Streetman and S. Banerjee, *Solid State Electronic Devices*, Prentice Hall, Upper Saddle River, NJ (1999).

S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed., John Wiley & Sons, Inc., New York (1981).

S. M. Sze, *VLSI Technology*, 2nd ed., McGraw-Hill Companies, Inc., New York (1988).

J. F. Ziegler, *Ion Implantation - Science and Technology*, Ion Implantation Technology Co., Yorktown, NY (1996).

See http://www.webelements.com for further information on the elements in this chapter.

8.9 Review Questions

1. List the advantages of ion implantation over diffusion in doping processes.
2. Name at least three implantation processes for CMOS IC fabrication.
3. What is unique DRAM implantation? What is the purpose of this implantation?
4. What major changes did ion implantation bring to IC fabrication processes?
5. What are the two ion stopping mechanisms?
6. The most important factors for a doping process are dopant concentration and dopant junction depth. Which systems in an ion implanter can be used to control these two factors?
7. When two ions with the same energy and incident angle implant into single-crystal silicon, will they stop at the same depth in the silicon? Explain your answer.
8. Describe the relationships between the ion projection range, ion energy, and ion species.
9. Why does a wafer need annealing after ion implantation?
10. List the advantages of RTA over furnace annealing.
11. What ion implantations require spike and laser annealing?
12. Why does an ion implanter beam line require high vacuum during processing?
13. What is the most poisonous gas used in IC fabrication? What does it smell like? Is it a p-type or n-type dopant?

14. Why does an ion implanter need a high-voltage power supply?
15. Before entering an ion implanter, why is the use of a grounding bar required?
16. Explain the reason for wearing double-layer gloves when wet cleaning a beam line tube.
17. Why are wafers tilted during the ion implantation process?
18. List the energy and current requirements of well implantation and S/D implantation processes. Explain your answer.
19. If the dc current in analyzer magnets is not correct, what could be a possible processing problem?
20. Compare four-point probe and thermal wave measurements and list their advantages and disadvantages.
21. What are the main differences between ion implantation and PIII?
22. What other industry or industries could use ion implantation besides the semiconductor industry?

Chapter 9
Etch

Objectives

After finishing this chapter, the reader will be able to:

- list at least three materials that require etching in IC chip fabrication
- identify at least three etch processes from a CMOS chip cross section
- describe both wet and dry etch processes and identify their differences
- describe the plasma etch processing sequence
- name the most commonly used etchants and identify their safety hazards.

9.1 Introduction

Etch is a process that removes materials from a wafer surface to achieve the requirements of IC design. There are two types of etch processes: pattern and blanket. Pattern etch selectively removes materials from designated areas and transfers the patterns of the photoresist or hard mask on the wafer surface to films underneath. Blanket etch removes all or part of the film on the surface to achieve the desired processing results. This chapter covers both etch processes, with an emphasis on pattern etch. Figure 9.1 illustrates a MOSFET gate patterning process. First, a photolithography process with gate mask defines the photoresist pattern on the polysilicon film on the wafer surface, as shown in Fig. 9.1(a). A pattern etch process transfers the pattern of the photoresist to polysilicon film underneath [see Fig. 9.1(b)]. Figure 9.1(c) shows the photoresist after being stripped by either a dry or wet process (or a combination of both), which finishes the gate patterning.

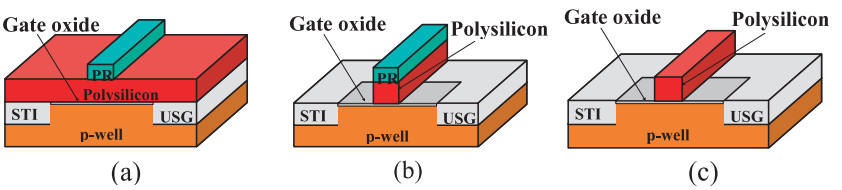


Figure 9.1 Pattern etch processes of a polysilicon gate: (a) photolithography, (b) etch polysilicon, and (c) strip photoresist.

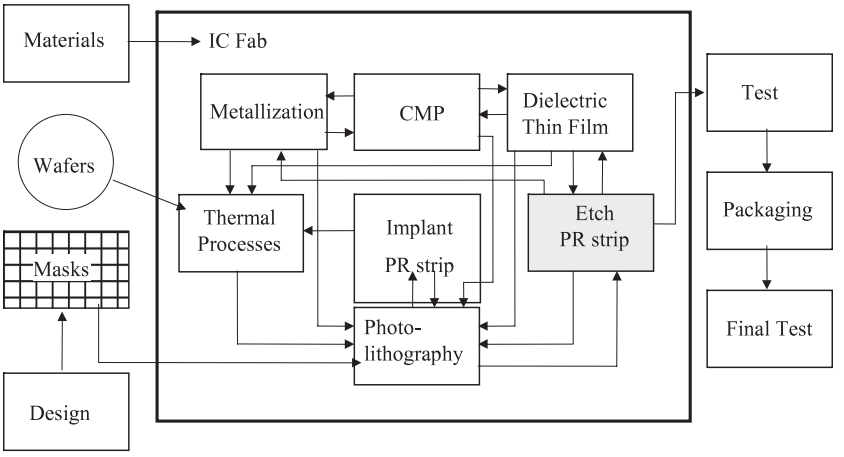


Figure 9.2 Advanced IC fabrication process flow.

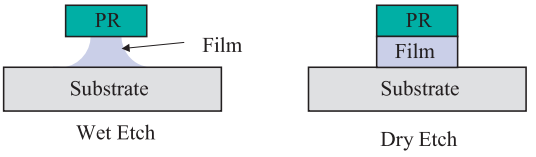


Figure 9.3 Wet and dry etch profiles.

A combination of photolithography and wet etch processes has been used by the printing industry for a long time and is still used to make printed circuit boards (PCBs). These technologies were adopted by the semiconductor industry for transistor and IC manufacture in the 1950s. Device and circuit designs are transferred to a wafer, by either etch or ion implantation, through the pattern defined on the photoresist on a wafer surface by means of the photolithography process. Figure 9.2 shows pattern etch in an IC chip manufacturing flow.

Before 1980, wet processes dominated semiconductor fabs. Chemical solutions were used to dissolve the materials not covered by photoresist to achieve pattern transfer. After 1980, when the minimum feature size became smaller than 3 μm, wet etch processes were gradually replaced by dry (plasma) etch processes. The reason for this is that wet etch always has an isotropic etch profile, which can cause CD loss, as shown in Fig. 9.3.

In advanced semiconductor fabs, almost all pattern etch processes are plasma etch processes. However, wet etch processes are still widely used for film stripping and thin-film quality control. Figure 9.4 illustrates the layers of some etch processes from the cross section of a CMOS IC chip with aluminum metallization technology.

Many etch processes (including both pattern and blanket etches) are involved in IC chip processing. For example, single-crystal silicon etch is needed to

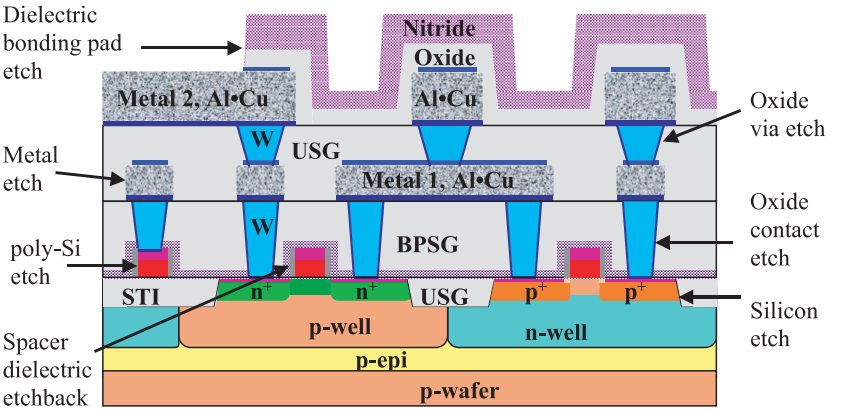


Figure 9.4 Etch processes in a CMOS IC chip with polysilicon gate and aluminum metal interconnection.

form shallow trench isolation (STI), and polysilicon etch defines gates and local interconnections. Oxide etch defines contact and via holes, and metal etch forms interconnections. There are also some blanket etch processes: the nitride stripping process after oxide chemical mechanical polishing (CMP) stops on the nitride layer (not shown in Fig. 9.4) during STI formation; dielectric anisotropic etchback to form sidewall spacers; and titanium stripping after titanium silicide formation (not shown in Fig. 9.4).

Figure 9.5 shows the cross section of an advanced CMOS IC with selective epitaxial S/D, gate-last high-κ and metal gates (HKMG), and copper/low-κ interconnection. Plasma etching of single-crystal silicon is required to form STI, and wet etching of single-crystal silicon is required to form the selective epitaxial S/D. Also, a dry etch of polysilicon is required to form the dummy polysilicon gate, and a wet etch of polysilicon is required to strip the dummy gate to create space for the HKMG. Note that there is no metal etch process for copper/low-κ interconnection. This connection uses a dielectric trench etch instead.

Topics covered in this chapter are etch basics; wet and dry etch; chemical, physical, and reactive ion etch; and etch processes for silicon, polysilicon, dielectric, and metal. Future trends of etch processes are also discussed in this chapter.

9.2 Etch Basics

9.2.1 Etch rate

Etch rate is the measure of how fast material is removed during the etch process. It is a very important characteristic, since it directly affects throughput of the etch process. Etch rate is defined as thickness change caused by the etch process divided by etch time. To calculate etch rate, film thickness must be measured before and

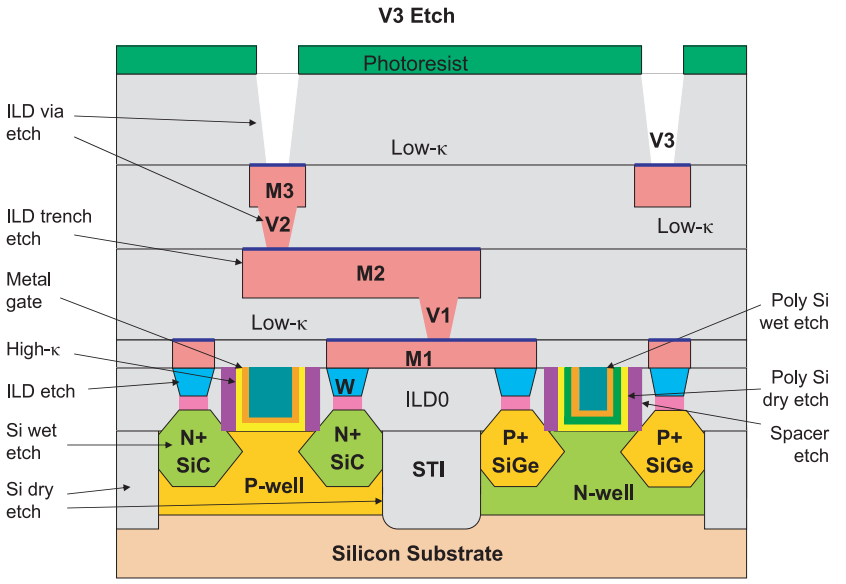


Figure 9.5 Etch processes of an advanced CMOS IC with selective epitaxial S/D, high-κ metal gate, and copper metal interconnection.

after the etch process, and etch time needs to be recorded.

$$\text{etch rate} = \frac{\text{thickness before etch} - \text{thickness after etch}}{\text{etch time}}.$$

Because etch rate is measured on blanket films, it is also called blanket etch rate. For pattern etch, the etch rate can be determined by a cross-section scanning electron microscope (SEM) measurement, which can directly measure the removed film thickness.

Question: A thermal oxide film thickness is 5000 Å; after a 30-sec plasma etch, the thickness becomes 2400 Å. What is the etch rate?
Answer: Etch rate = (5000 Å − 2400 Å)/0.5 min = 2600 Å/0.5 min = 5200 Å/min.
Question: A borophosphosilicate glass (BPSG) contact hole etch profile is shown in Fig. 9.6. What is the etch rate?
Answer: Etch rate = 4500 Å/(45/60) min = 4500Å/0.75 min = 6000 Å/min.

9.2.2 Uniformity

It is very important to have uniform etch rate across a wafer or a good within-wafer (WIW) uniformity, and high repeatability or good wafer-to-wafer (WTW) uniformity. Etch-rate uniformity is calculated from etch rates measured at certain points of a wafer. If these values are $x_1, x_2, x_3, x_4, \ldots, x_N$, N is the total number of

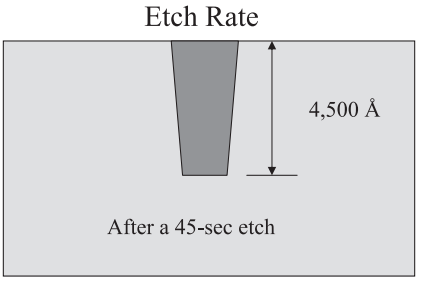


Figure 9.6 A contact profile.

data points. The mean value (or average value) of the measurement is

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}.$$

The standard deviation of the measurement is

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N - 1}}.$$

The standard deviation nonuniformity (in percentages) is defined as

$$NU(\%) = (\sigma/\bar{x}) \times 100.$$

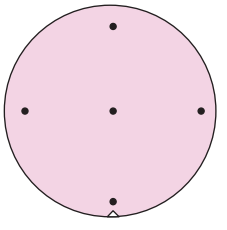
Maximum-minus-minimum nonuniformity is defined as

$$NU_M = \frac{x_{max} - x_{min}}{\bar{x}} \times 100.$$

Question: What is the maximum-minus-minimum NU_M for the five-point measurement?

Before etch:
3500, 3510, 3499, 3501, 3493 Å.

After a 60-sec etch:
1500, 1455, 1524, 1451, 1563 Å.



Answer: Etch rates are 2000, 2055, 1975, 2055, and 1930 Å/min. The average etch rate is: $\bar{x} = 2003$ Å/min, and the maximum-minus-minimum nonuniformity: $NU_M = (2055 - 1930)/(2 \times 2003) = 3.12\%$.

It is very important to have a clear definition of uniformity in the tool specifications, especially when working with the vendors or customers, because different definitions give different results.

9.2.3 Selectivity

The pattern etch process involves three kinds of materials: photoresist, the film that needs to be etched, and the film underneath. During the etch process, all three can be etched by chemical reaction of the etchants and by ion bombardment in a plasma etch process. The difference in the etch rates is characterized as selectivity.

Selectivity S is the ratio of the etch rates between different materials, especially the material that requires etching compared to the material that should not be removed.

$$S = \frac{ER_1}{ER_2}.$$

For example, in the gate etch illustrated in Fig. 9.3, photoresist provides the etch mask, and polysilicon is the material that needs to be etched. In a plasma etch process, inevitably photoresist will be etched; therefore, it is necessary to have high enough polysilicon-to-photoresist selectivity to prevent excessive photoresist loss before the etch process is finished. Underneath the polysilicon is the ultrathin gate oxide (15 to 100 Å, depending on device requirement); the process must have a very high polysilicon-to-oxide selectivity to prevent etching through the gate oxide layer during the polysilicon overetch step.

9.2.4 Profile

One of the most important characteristics of etch is the etch profile, which affects all of the deposition processes that follow. Etch profile is measured by a cross-section SEM. Figure 9.7 shows different types of etch profiles.

A perfect vertical profile is a theoretically ideal profile, since it can transfer images from the photoresist to the underside of the film without any CD loss. However, in many cases (especially for contact and via etch), an anisotropic tapered profile is preferred because the tapered contact and via hole create larger incoming angles and are easier for the following tungsten CVD processes to fill without voids. A pure chemical etch process has an isotropic profile, which causes undercuts beneath the photoresist as well as CD loss. Reactive ion etch (RIE)

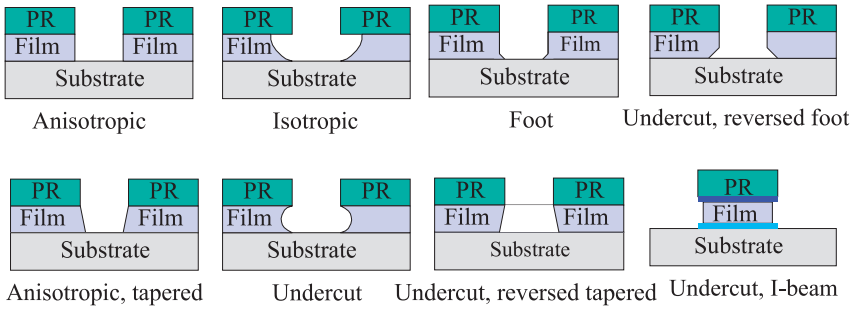


Figure 9.7 Different etch profiles.

combines both physical and chemical etches. Undercut profiles are caused by too many chemical etch components in the RIE process, or excessive ion scattering to the sidewalls. Undercut profiles are undesirable, since it is extremely difficult for the following deposition processes to fill the gaps or holes without voids. An I-beam profile is caused by incorrect etch chemistry for the middle layer of sandwiched film.

9.2.5 Etch bias

Etch bias is the difference between the CD of an etched pattern and the CD of a photoresist pattern. In Fig. 9.1, the anisotropic etch profile has the same CD as the photoresist, or zero etch bias. The undercut I-beam etch profile has CD loss, or negative etch bias. It is possible to have positive etch bias, which occurs when the CD of a photoresist pattern is smaller than the CD of an etched pattern.

9.2.6 Loading effects

In plasma pattern etch processes, etch rate and etch profile are related to the etched pattern. This phenomena is called the etch loading effect. There are two types of loading effects: macroloading and microloading.

9.2.6.1 Macroloading

The etch rate of a wafer with a larger open area is different from a wafer with a smaller open area; this WTW etch-rate difference is called the macroloading effect. Macroloading mainly affects the batch etch process and has minimal effect on the single-wafer process.

9.2.6.2 Microloading

For contact and via hole etch processes, smaller holes usually have a lower etch rate compared to larger holes, as shown in Fig. 9.8. This is called the microloading effect. Etchants have a difficult time passing through the smaller holes to reach the film that needs to be etched, and etch byproducts are harder to diffuse out. These two factors slow down the etch process.

Reducing process pressure can minimize the microloading effect. At lower pressure, the MFP is longer. Therefore, it is easier for etchants to pass through the tiny holes to reach the film. It is also easier for etch byproducts to escape from the tiny holes and be removed.

The etch pattern profile in an isolated pattern area is usually wider than the densely packed area due to photoresist sputtering deposition on the sidewall. Lack of sidewall ion bombardment from the ion scattering from neighboring patterns causes buildup of photoresist on the pattern sidewall and results in a wider profile. Figure 9.9 illustrates the profile microloading effect.

9.2.7 Overetch

During thin-film etching—including polysilicon, dielectric, and metal etch—the etch rate within the wafer is not perfectly uniform, and neither is the film thickness. Therefore, while some parts of the film have already been etched away, other parts

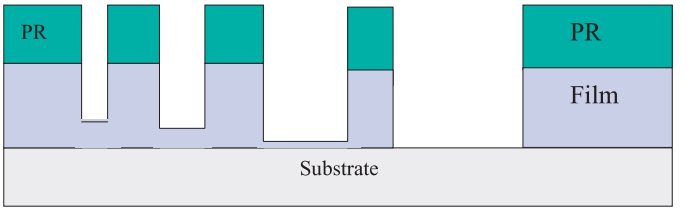


Figure 9.8 The microloading effect.

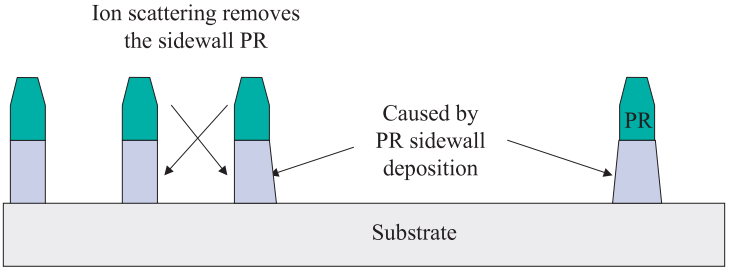


Figure 9.9 Profile microloading effect.

still have some film left behind that needs to be removed. The process that removes the leftover film is called the overetch, and the etch process that removes the bulk of the film is called the main etch.

During the overetch process, selectivity between the etched film and the substrate material must be high enough to prevent excessive substrate material loss. By using different etch conditions from the main etch process, etch selectivity between the etched film and substrate can be significantly improved during the overetch process. For a plasma etch process, overetch can be automatically triggered by an optical endpoint detector, since the chemical composition of the plasma changes when etchants in the main etch process start to etch the substrate film. For example, during the polysilicon gate etch process illustrated in Fig. 9.1, the main etch step removes the bulk polysilicon film and does not bother with selectivity over silicon dioxide. When the polysilicon has been etched away and the etchants in plasma start to etch the silicon dioxide, the radiation intensity of the oxygen line in the plasma increases. This can trigger an electrical signal to end the main etch and switch to overetch, which the high selectivity of polysilicon to silicon dioxide requires.

Figure 9.10 illustrates the main and overetch processing steps. Δd is the film thickness variation due to thickness nonuniformity. $\Delta d'$ is the maximum substrate thickness loss. The minimum requirement of film-to-substrate selectivity during the overetch processing step is: $S > \Delta d/\Delta d'$ if the etch rate is uniform in the illustrated area.

Question: For an IC chip, polysilicon film thickness is 3000 Å, and the thickness nonuniformity is 1.5%. The polysilicon etch rate is 5000 Å/min, and the

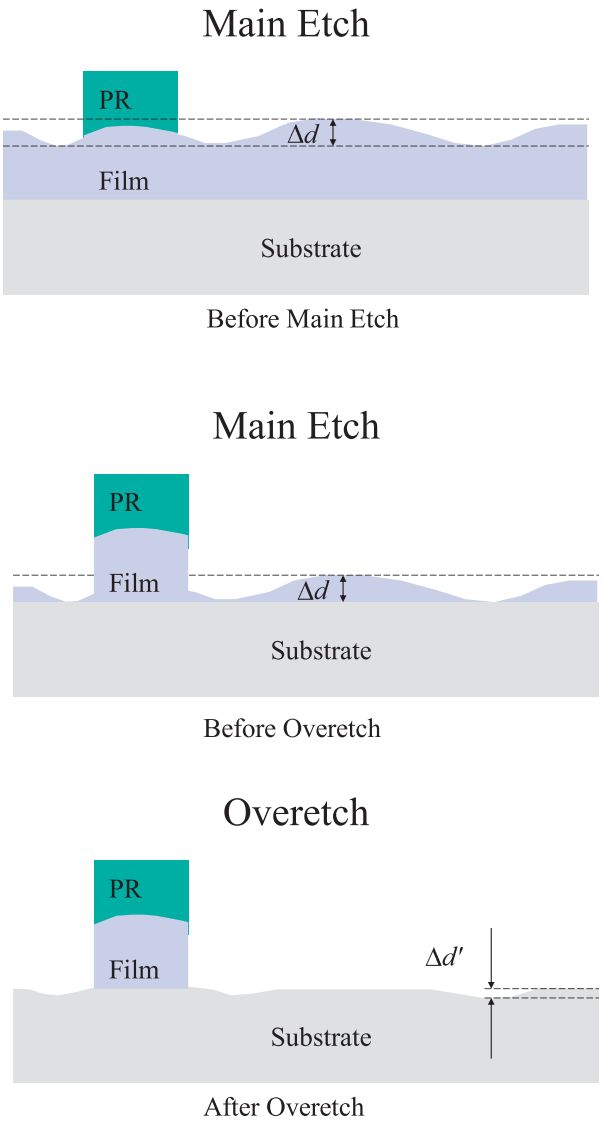


Figure 9.10 Main etch and overetch.

etch-rate nonuniformity within the wafer is 5%. If only 5 Å of gate oxide ($t_{ox} \sim 40$ Å) is allowed to be lost, what is the minimum poly-to-oxide selectivity during the overetch step?

Answer: Consider the worst possible scenario: the highest etch rate (5000 + 5000 × 5% = 5250 Å/min) on the thinnest film (3000 − 3000 × 1.5% = 2955 Å), and the lowest etch rate (5000 − 5000 × 5% = 4750 Å/min) on the thickest film (3000 + 3000 × 1.5% = 3045 Å). The etch time

difference between the two (3045/4750 – 2955/5250 = 0.0782 min) is the overetch time. Since 5 Å of oxide cannot be lost during overetch, the maximum oxide etch rate is 5/0.0782 = 64 Å/min. The minimum poly-to-oxide selectivity is 5000/64 = 78.2.

9.2.8 Residue

After the etch process is finished, there can be some unwanted leftover materials either on the sidewall or the wafer surface. These unwanted leftovers are called residues. Residue can be caused by insufficient overetch of a film with complex surface topography or nonvolatile etch byproducts. Figure 9.11 shows a formation of residue called a stringer on the sidewall due to insufficient overetch of a film with step-like topography. In polysilicon etch processes, stringers are killer defects because they can cause short-circuiting between the polysilicon lines.

Adequate overetching removes most stringers. Sufficient ion bombardment can help remove surface residue, or the right amount of chemical etch can scoop out nonvolatile etch byproducts, such as copper chloride generated during the Al-Cu alloy etch process (illustrated in Fig. 9.12). Organic residue can be cleaned up by an oxygen plasma ashing process, which is also used to strip the remaining photoresist. Wet chemical clean processing can help remove the inorganic residue.

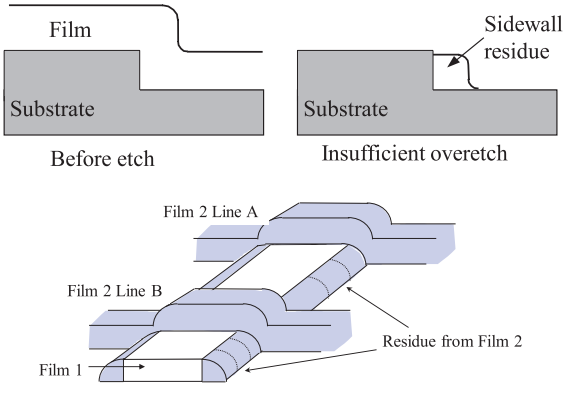


Figure 9.11 Residue caused by insufficient overetching and step topography.

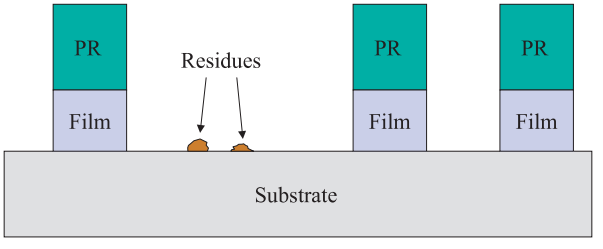


Figure 9.12 Surface residue caused by nonvolatile etch byproducts.

9.3 Wet Etch Process

9.3.1 Introduction

Wet etch is a process that uses chemical solutions to dissolve materials on the wafer surface and achieve device and circuit requirements. The byproducts of wet etch chemical reactions are gases, liquids, or materials that are soluble in the etchant solution. It has three basic steps—etch, rinse, and dry—as shown in Fig. 9.13.

Wet etching was widely used for patterning processes in semiconductor fabs before 1980, when feature size was larger than 3 μm. Wet etching normally has a high etch rate, which is controlled by etchant temperature and concentration. Wet etch processes have very good selectivity. For instance, hydrofluoric acid (HF) etches silicon dioxide very quickly, while it hardly etches silicon at all if used alone. Therefore, using HF to etch a silicon dioxide layer grown on a silicon wafer can achieve very high selectivity of silicon substrate. Compared to dry etch, wet etch uses much less expensive processing equipment because it does not require the vacuum, rf power, and complex gas delivery systems. However, after feature size shrank to less than 3 μm, it became very difficult to continue using wet etch for pattern etch processing because of its isotropic etch profile (shown in Fig. 9.14). It is impossible to wet etch a densely packed pattern with a feature size less than 3 μm; therefore, plasma etch has gradually been replacing wet etch for patterning since the 1980s because it can etch patterns with anisotropic profiles.

Fabs tried very hard to eliminate all wet processes; however, when CMP and electrochemical deposition became widely used in advanced IC fabrications, it became impossible to get rid of all wet processes. Wet etch processes are still widely used in IC fabs for thin-film stripping due to their high selectivity. Wet etch is also applied in blanket thin-film etch processing to determine deposited thin film quality.

9.3.2 Oxide wet etch

Hydrofluoric acid (HF) is commonly used for silicon dioxide wet etch. Since 1:1 HF (49% HF in H_2O) etches oxide too quickly at room temperature, it is very difficult to control the oxide etch process with 1:1 HF. HF is further diluted either

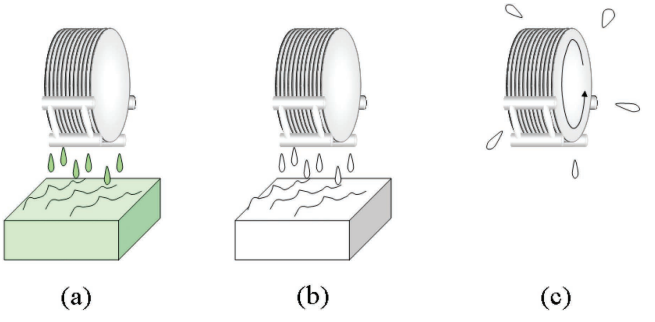


Figure 9.13 Wet etch processing steps: (a) wet etch, (b) rinse, and (c) dry.

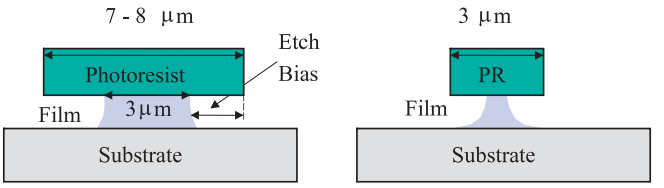


Figure 9.14 Wet etch profiles.

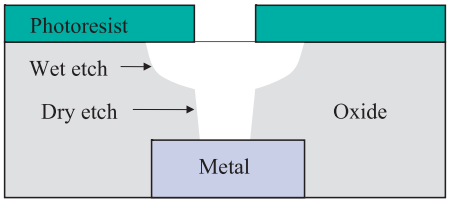


Figure 9.15 Wine-glass contact.

in water or in some buffer solution, such as ammonium fluoride (NH_4F), to slow down the oxide etch for better control of etch rate and uniformity. 6:1 buffered solution of HF, 10:1 HF, and 100:1 in H_2O are commonly used solutions for oxide wet etch.

The chemical reaction of oxide wet etch is

$$SiO_2 + 6\,HF \rightarrow H_2SiF_6 + 2\,H_2O.$$

H_2SiF_6 is water soluble, so an HF solution can easily etch away silicon dioxide. This is why HF solution cannot be kept in a glass container, and HF etch experiments cannot be performed in glass beakers or test tubes.

Some IC fabs still use combined HF oxide wet etching and plasma oxide etching to create so-called wine-glass contact holes (as shown in Fig. 9.15), which are easier for PVD aluminum to fill.

The 6:1 buffered oxide etch (BOE) and 100:1 diluted HF (DHF) etch are used daily in semiconductor fabs, even those with the most advanced technology. They are used to monitor CVD oxide film quality by comparing the wet etch rate of CVD oxide to thermally grown silicon dioxide, called the wet etch rate ratio (WERR). 10:1 HF is also used to strip native oxide from the silicon wafer surface before the thermal oxidation process.

HF is corrosive, and the unique hazard of HF is that skin or eye contact may not be immediately noticed, until up to 24 h later when severe pain can be felt as HF starts to attack the bone. HF reacts with calcium in the bone to form calcium fluoride (which is eventually neutralized). Calcium-containing solution injections can be used to treat HF attacks to prevent or reduce bone loss. Common sense requires that employees treat all clear liquid in fabs as HF—never assume it is

water! If direct contact with HF is suspected, the area affected should be washed thoroughly and immediately, and medical attention should be sought.

9.3.3 Silicon etch

Single-crystal silicon etch is used for forming an isolation block between neighboring transistors, and polysilicon etch forms gates and local interconnections. Single-crystal and polycrystalline silicon can be etched isotropically with a mixture of nitric acid (HNO_3) and hydrofluoric acid (HF) through a complex chemical reaction. At first, HNO_3 oxidizes silicon on the surface to form a thin layer of silicon dioxide, which blocks the oxidation process. HF then reacts with the silicon dioxide and dissolves it to expose the silicon underneath, which is oxidized again by HNO_3. Then the oxide is etched away by the HF, and the process repeats again and again. The overall chemical reaction can be expressed as

$$Si + 2HNO_3 + 6\,HF \rightarrow H_2SiF_6 + 2HNO_2 + 2\,H_2O.$$

The mixture of potassium hydroxide (KOH), isopropyl alcohol (C_3H_8O), and water can selectively etch to the different orientations of single-crystal silicon. With 23.4 wt% KOH, 13.3 wt% C_3H_8O, and 63.3 wt% H_2O at 80 to 82 °C, the etch rate along ⟨100⟩ planes is about 100 times higher than along ⟨111⟩ planes. The hexagon-shaped groove shown in Fig. 9.16 can be wet etched by this type of anisotropic single-crystal silicon mechanism.

Nitric acid is very corrosive, and it becomes an oxidizer at concentrations higher than 40%. Direct contact can cause severe burns to the skin and eyes and can leave a bright yellow stain on the skin. Nitric acid vapor has a sharp odor and can be irritating to the throat when exposure is at low concentrations. If nitric acid vapor is inhaled at high concentrations, it can cause choking, coughing, and chest pain. In more serious cases, it can cause severe breathing difficulty, bluish skin color, and even death within 24 h due to accumulation of fluid in the lungs.

KOH is corrosive and can cause serious burns. It is harmful if ingested, inhaled, or touched. If the solid or solution comes into contact with the eyes, serious eye damage can result.

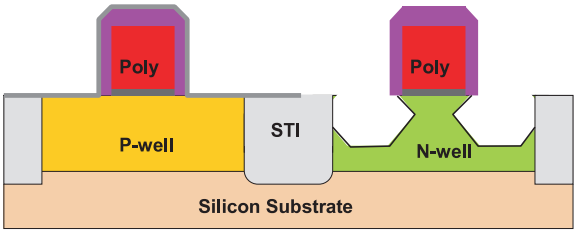


Figure 9.16 The anisotropic KOH silicon etch forms cavities for selective epitaxial SiGe pMOS S/D.

9.3.4 Nitride etch

Nitride is widely used for isolation formation processes. Figure 9.16 shows an isolation process used in the 1970s for bipolar transistor-based IC chip manufacturing, which involved silicon nitride, silicon oxide, and single-crystal silicon etches.

Hot phosphoric acid (H_3PO_4) is most commonly used to etch silicon nitride. At 180 °C with 91.5% H_3PO_4 concentration, the nitride etch rate is about 100 Å/min. This silicon nitride etch process has very good selectivity to thermally grown silicon dioxide (> 10:1) and to silicon (> 33:1). Increasing H_3PO_4 concentration to 94.5% and the temperature to 200 °C increases the nitride etch rate to 200 Å/min, while the selectivity to silicon dioxide drops to about 5:1, and selectivity to silicon reduces to around 20:1.

Question: HF can also be used to etch silicon nitride. However, in the isolation formation process (as shown in Fig. 9.17), both patterned nitride etch and the nitride strip cannot use HF. Why not?

Answer: HF etches silicon nitride at a much lower rate than it etches silicon dioxide. Therefore, using HF for nitride pattern etching can cause excessive pad oxide loss and severe undercuts. If HF is used for the nitride strip, it will very quickly etch away the pad oxide and the isolation oxide!

The chemical reaction of the silicon nitride etch is

$$Si_3N_4 + 4\,H_3PO_4 \rightarrow Si_3(PO_4)_4 + 4NH_3.$$

Both byproducts, silicon phosphate [$Si_3(PO_4)_4$] and ammonia (NH_3), are water soluble. This process is still used in isolation formation processes to strip the nitride—either after the field oxide is grown in the local oxidation of silicon (LOCOS) process, or after undoped silicate glass (USG) polishing and annealing in the STI process.

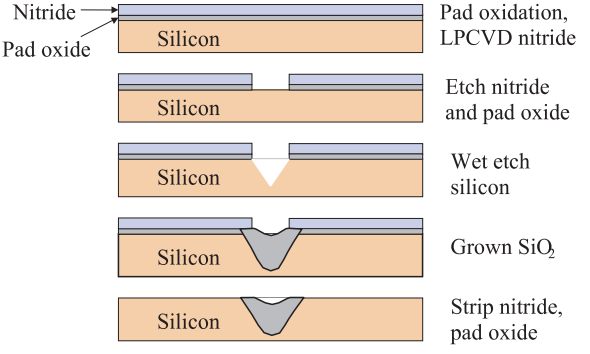


Figure 9.17 Isolation formation processing steps.

Phosphoric acid is an odorless liquid; it is corrosive and can severely burn the skin and eyes by direct contact. Exposure to its mist can cause irritation to the eyes, nose, and throat at a low concentration, while it can cause coughing and burning of skin, eyes, and lungs at a high concentration. Long-term exposure can also cause tooth erosion.

9.3.5 Metal etch

Aluminum can be etched by a wide variety of acidic formulations. One of the most commonly used formulations is the mixture of phosphoric acid (H_3PO_4, 80%), acetic acid (CH_3COOH, 5%), nitric acid (HNO_3, 5%), and water (H_2O, 10%). The etch rate for pure aluminum is about 3000 Å/min at 45 °C. The etch mechanism of aluminum etching is very similar to silicon etching: HNO_3 oxidizes the aluminum to form aluminum oxide, and H_3PO_4 dissolves Al_2O_3. Both oxidization and oxide dissolution proceed simultaneously.

In advanced IC fabs, wet processing is no longer used for aluminum pattern etching; however, some small fabs and research laboratories still use this process.

The most commonly used metal wet etching process in advanced semiconductor fabs is nickel stripping after nickel silicide formation, as shown in Fig. 9.18. The 1:1 mixture of hydrogen peroxide (H_2O_2) and sulfuric acid (H_2SO_4) is commonly used to selectively etch away nickel (Ni) while keeping silicon oxide and nickel silicide intact. The etching mechanism is similar to other metal wet etching processes. Hydrogen peroxide oxidizes nickel to form nickel oxide, while sulfuric acid simultaneously reacts with nickel oxide to form water-soluble $NiSO_4$.

Acetic acid (CH_3COOH, solutions of 4 to 10% in water make up vinegar) is a corrosive and flammable liquid with a strong vinegar odor. Direct contact with acetic acid can cause a chemical burn. High concentrations of acetic acid vapors can cause coughing, chest pain, nausea, and vomiting. Hydrogen peroxide (H_2O_2) is an oxidizer. Contact can cause irritation and burns to the skin and eyes. High concentrations of its vapor can cause severe irritation of the nose and throat, and fluid in the lungs. H_2O_2 is unstable; it self-decomposes while in storage. Sulfuric acid (H_2SO_4) is corrosive and can cause the skin to burn with direct contact. Even

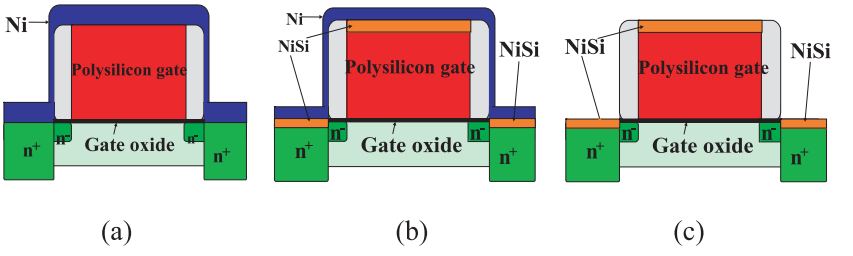


Figure 9.18 Self-aligned silicide (salicide) processes: (a) nickel deposition, (b) nickel silicide annealing, and (c) nickel wet stripping.

the diluted solution can cause a skin rash. Highly concentrated mist can cause severe chemical burns to the skin, eyes, and lungs.

9.4 Plasma (Dry) Etch

9.4.1 Introduction

Dry etch processes use gaseous chemical etchants to react with materials to be etched to form volatile byproducts, which will be removed from the substrate surface. Plasma generates chemically reactive free radicals that can significantly increase the chemical reaction rate and enhance the chemical etch. Plasma also causes ion bombardment to the wafer surface. Ion bombardment can either physically remove materials from the surface or break the chemical bonds between atoms on the surface, significantly accelerating the chemical reaction rate for the etch process. This is why most dry etch processes are plasma etch processes.

After the 1980s when feature size became smaller than 3 μm, plasma etch processes gradually replaced wet etching for pattern etch processes. The reason for this was that isotropic etch profiles of wet etch processes could no longer meet the requirements of such a small geometry. Plasma etch is anisotropic due to ion bombardment associated with the presence of plasma. Therefore, it has a much smaller etch bias and CD loss compared to the wet etch process. Table 9.1 compares wet and dry etch processes.

9.4.2 Plasma review

Plasma is ionized gas with an equal number of negative and positive charges. It also consists of ions, electrons, and neutral atoms or molecules. The three important collisions in plasma are ionization, excitation-relaxation, and disassociation. These collisions generate and sustain the plasma, cause the glow of gas discharge, and create chemically reactive free radicals that enhance chemical reactions, respectively.

MFP is the average distance a particle can travel before it collides with another particle. Reducing pressure increases MFP, increases ion bombardment energy, and reduces collision-induced ion scattering, helping to achieve vertical etch profiles.

Plasma always has a higher potential than electrodes because the faster moving electrons always charge the electrodes negatively at the start of the plasma. Higher

Table 9.1 Comparison of wet and dry etch processes.

	Wet Etch	Dry Etch
Etch bias	Unacceptable for < 3 μm	Minimum
Etch profile	Isotropic	Anisotropic to isotropic, controllable
Etch rate	High	Acceptable, controllable
Selectivity	High	Acceptable, controllable
Equipment cost	Low	High
Throughput	High (batch)	Acceptable, controllable
Chemical usage	High	Low

plasma potential causes ion bombardment because positively charged ions are accelerated to the lower potential electrodes by the sheath potential. Increasing rf power increases both ion bombardment flux and ion bombardment energy in capacitively coupled plasmas. Increasing rf power also increases free radical concentration.

Since etch is a removal process, low operation pressure is desired. The MFP is longer at a lower pressure, which is good for both ion bombardment and byproduct removal. Some etch chambers also use magnetic coils to generate a magnetic field for increasing plasma density at low pressure (<100 mtorr). As a removal process, the plasma etch process requires much more ion bombardment than does a PECVD process. Therefore, in most etch processes, a wafer is put on the smaller electrode to take advantage of the self-bias and receive more energetic ion bombardment.

High-density plasma at a low pressure is desired for both etch and CVD processes. However, commonly used capacitively coupled plasma sources cannot generate high-density plasma. ICP and ECR systems are the two most commonly used high-density plasma sources. By using a separated bias rf system, both ICP and ECR plasma sources can independently control both flux and energy of ion bombardment.

9.4.3 Chemical, physical, and reactive ion etches

There are three kinds of etch processes: pure chemical etch, pure physical etch, and RIE, which is something in-between.

Wet etch is a pure chemical etch process. Another example of pure chemical etch is the remote plasma photoresist strip. For pure chemical etch, there is no physical bombardment; materials are solely removed by chemical reaction. The etch rate of a pure chemical etch can be very high or fairly low, depending on the process. Pure chemical etch always has an isotropic etch profile, which prevents it from being applied to pattern etching when the feature size is smaller than 3 μm. Since it normally has very good etch selectivity, the pure chemical etch process is used for strip processes, such as photoresist, silicon nitride, pad oxide, screen oxide, and sacrificial oxide strips. Remote plasma (RP) etch processes, which use plasma to generate free radicals in a remote chamber and flow them into the processing chamber to react with the wafer, are pure chemical etch processes.

Argon sputtering is a purely physical etch. It is widely used in dielectric sputtering etchback to taper the openings for easier gap fill in the subsequent deposition. It is also used in precleaning to remove native oxide before metal PVD processes to reduce contact resistance. Argon is an inert gas. Therefore, there is no chemical reaction during the process. The material is physically dislodged from the surface by energetic argon ions, similar to being hit with a hammer. Etch rate of a pure physical etch is very low and mainly depends on the flux and energy of the ion bombardment. Since ion bombardment hits and removes anything in its way, selectivity for a pure physical etch is very poor. The direction of ion bombardment is mostly perpendicular to the wafer surface in plasma etch processes. Therefore, a pure physical etch is an anisotropic etch process, which etches mainly in the vertical direction.

The name RIE can be slightly misleading. The proper name of this type of etch process should be an ion-assisted etch, since the ions in this etch process are not necessarily reactive. For example, in many cases argon ions are used to increase ion bombardment. As an inert atom, argon is not chemically reactive at all. Reactive species in most etch processes are neutral free radicals, which have much higher concentrations than ions in semiconductor etch processing plasmas. This is because the activation energy of ionization is significantly higher than the activation energy of dissociation, and the species concentration is exponentially related to the activation energy. However, the term RIE has been used in the semiconductor industry for such a long time that no one will likely change it. Figure 9.19 shows a schematic of an early experiment with ion-assisted etch, along with the results.

The experiment starts with a XeF_2 gas flow, achieved by opening the shutoff valve. XeF_2 is an unstable gas because xenon is a noble gas and does not like to bond with any other atom. It is always used as a free fluorine carrier for dry chemical etch processes. When XeF_2 hits the heated single-crystal silicon sample, it disassociates and releases two free fluorine radicals. A free fluorine radical has only one unpaired electron, which makes it seek out an electron from other atoms; this makes it chemically very reactive. Fluorine reacts with silicon on the sample surface, forming volatile silicon tetrafluoride (SiF_4) and chemically dry etching the silicon. Measurement results in Fig. 9.19 show that the etch rate is very low for this purely chemical etch process.

Next, the argon ion gun is turned on. The silicon etch rate increases dramatically with the combination of the physical ion beam bombardment and the free fluorine radical chemical etch. After termination of the XeF_2 gas flow, silicon is etched solely by argon ion sputtering. This is a pure physical etching process with very low etch rate, even lower than a pure chemical etch with only XeF_2 gas flow.

From Fig. 9.19, it can be seen that the combination of both XeF_2 gas flow and argon ion bombardment has the highest etch rate, which is much higher than the sum of the two individual etch rates. This is because the argon ion bombardment breaks the chemical bonds of the silicon atoms on the surface. Silicon atoms

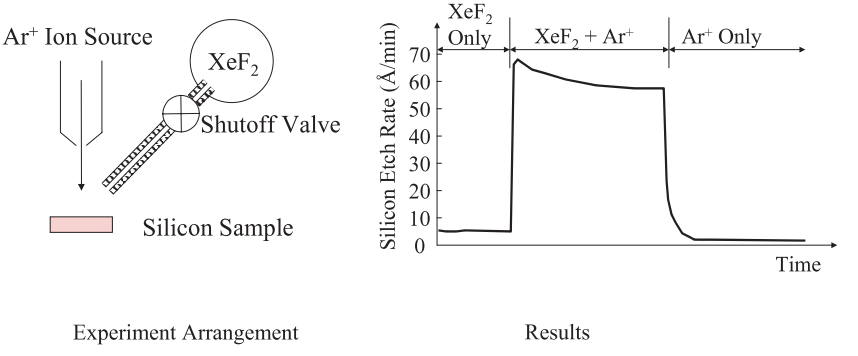


Figure 9.19 The ion-assisted etch experiment (Coburn and Winters, *J. Applied Physics*).

with dangling bonds on the surface bond with free fluorine radicals and form silicon tetrafluoride much more easily than do silicon atoms without broken bonds. Because ion bombardment is mainly perpendicular to the wafer surface, vertical etch rate is significantly higher than horizontal etch rate. This is the reason RIE has an anisotropic etch profile.

Almost all pattern etch processes in advanced semiconductor fabs are RIE processes. RIE has a reasonable and controllable etch rate and etch selectivity. It also has an anisotropic and controllable etch profile. Table 9.2 summarizes these three types of etch processes.

9.4.4 Etch mechanisms

In plasma etch processes, first the etchants are introduced to the vacuum chamber. After pressure is stabilized, rf power is used to strike discharge plasma glow. Some etchant molecules are dissociated in the plasma by the impact of collisions with electrons, which generate free radicals. The free radicals then diffuse across the boundary layer, reach the wafer surface, and are adsorbed on the surface. With the help of ion bombardment, these free radicals react with surface atoms or molecules very quickly and form gaseous byproducts. The volatile byproducts desorb from the surface, diffuse across the boundary layer, move into the convection flow, and are pumped out of the chamber. This plasma etch process sequence is illustrated in Fig. 9.20.

Plasma etch can achieve an anisotropic etch profile due to ion bombardment associated with the plasma. There are two anisotropic mechanisms: damaging and blocking; both are related to ion bombardment.

For damaging mechanisms, energetic ion bombardment breaks the chemical bonds between the atoms on a wafer surface. Atoms with dangling bonds on the surface are vulnerable to the etchant free radicals. They bond easily with the etchant radicals to form volatile byproducts, which can be removed from the surface. Because ions mainly bombard in a perpendicular direction, etch rate in the vertical direction is much higher than etch rate in the horizontal direction. Therefore, ion bombardment can achieve an anisotropic etch profile. Etch processes with a damaging mechanism are RIE processes that are closer to the physical etch side. The damaging mechanism for anisotropic etch is illustrated in Fig. 9.21.

Dielectric etch processes (mainly silicon oxide, silicon nitride, and low-κ dielectric etch processes) normally use a damaging mechanism. They are the RIE

Table 9.2 Three types of etch processes.

	Chemical etch	RIE	Physical etch
Examples	Wet etch, strip, PR etch	Plasma pattern etches	Argon sputtering
Etch rate	High to low	High, controllable	Low
Selectivity	Very good	Reasonable, controllable	Very poor
Etch profile	Isotropic	Anisotropic, controllable	Anisotropic
Endpoint	By time or visual	Optical	By time

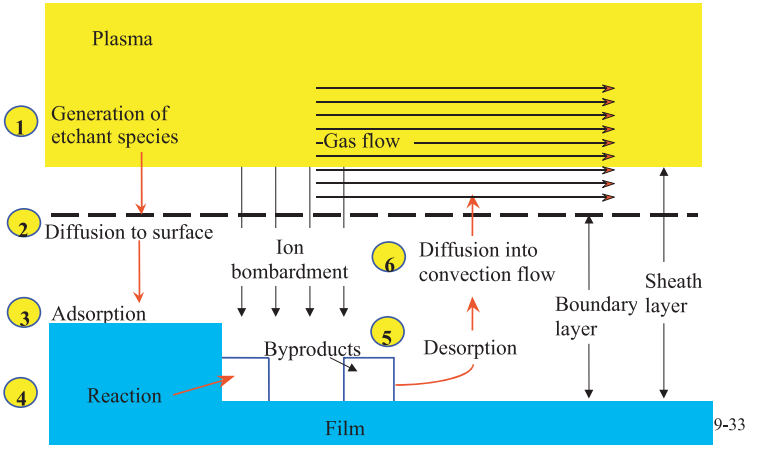


Figure 9.20 Plasma etch sequence.

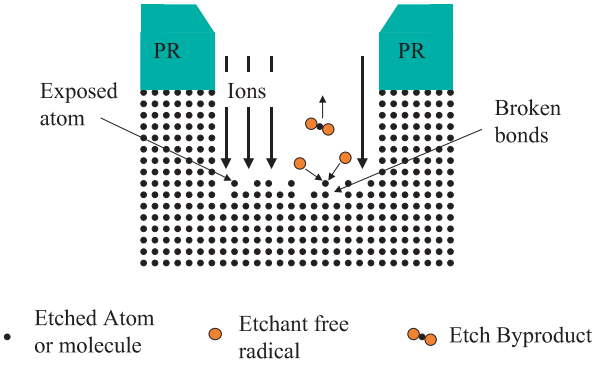


Figure 9.21 Damaging mechanism for anisotropic etch.

processes that are closer to the physical etch side. To improve the anisotropic profile with a damaging mechanism, ion bombardment needs to increase. By using heavy ion bombardment at high rf power and low pressure, a nearly vertical etch profile can be achieved. However, this can also cause plasma-induced device damage, especially for a polysilicon gate etch. Therefore, another anisotropic etch mechanism with less ion bombardment is needed to meet the requirements of these etch processes.

During the development of the single-crystal silicon etch process, someone forgot to strip the photoresist used for the silicon dioxide hard mask patterning before the silicon etch process. (The normal procedure requires stripping of the photoresist before silicon etch to prevent contamination). What happened next led to the discovery of another anisotropic etching mechanism, the blocking mechanism. During the plasma etch process, the ion bombardment sputtered some

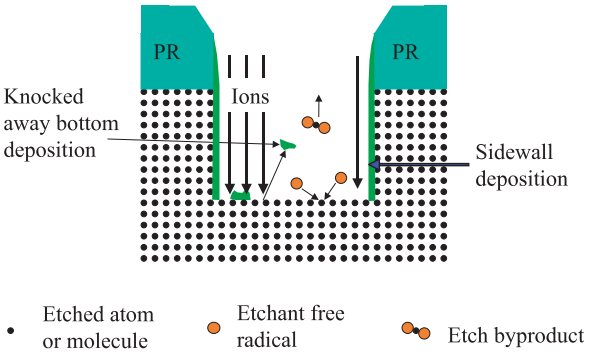


Figure 9.22 Blocking mechanism for anisotropic etch.

photoresist into the gap. The photoresist coating on the sidewall blocked the etch in the horizontal direction, while photoresist deposited on the bottom was constantly removed by ion bombardment from the plasma and exposed the surface to the etchants. Thus, the etch process is now mostly in the vertical direction, as shown in Fig. 9.22.

This mechanism has been used to develop a variety of anisotropic etch processes, in which chemical deposition during etch processing protects the sidewall and blocks etching in the horizontal direction. Etch processes using a blocking mechanism generally require less ion bombardment than etch processes using a damaging mechanism. Single-crystal silicon etch, polysilicon etch, and metal etch processes use this mechanism and are the RIE processes that are closer to the chemical etch side. The sidewall deposition needs to be cleaned by either (or both) dry and wet cleaning processes. Table 9.3 summarizes the two anisotropic etch mechanisms and their applications in the IC processes.

9.4.5 Plasma etch chamber

The plasma process was first applied to etch carbonaceous materials, such as photoresist, in an oxygen plasma, a process also known as plasma stripping or plasma ashing. In this case, oxygen free radicals generated in the plasma by electron dissociation collisions react rapidly with carbon and hydrogen in carbonaceous material to form volatile CO, CO_2, and H_2O. This process effectively removes the carbonaceous material from the surface. A barrel system with an etching tunnel used for this application is illustrated in Fig. 9.23.

This concept was extended in the late 1960s to silicon etch and silicon compound etch that used fluorine-containing gas such as CF_4 as the etchants and SiF_4 as the gaseous etch byproducts.

Another type of dry etch system is a downstream or remote plasma system, which generates plasma in a remote chamber. Etchant gases are allowed to flow through the plasma chamber and dissociate in the plasma, and the free radicals are allowed to flow into the processing chamber to react and etch the materials on the wafer. Figure 9.24 shows the schematics of a downstream etch processing system.

Table 9.3 The two anisotropic etch mechanisms and their applications.

	RIE		
Pure chemical etch	**Blocking mechanism**	**Damaging mechanism**	**Pure physical etch**
No ion bombardment Photoresist strip	Light ion bombardment Single-crystal silicon etch	Heavy ion bombardment Oxide etch	Only ion bombardment
Silicide metal strip Nitride strip	Polysilicon etch Metal etch	Nitride etch Low-κ dielectric etch	Sputtering etch

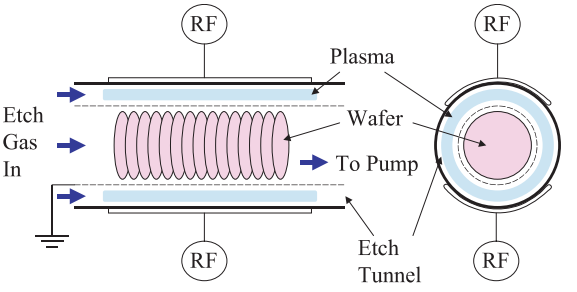


Figure 9.23 Barrel etch system.

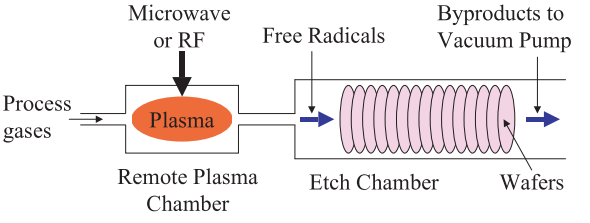


Figure 9.24 Downstream plasma etch system.

Barrel and downstream systems are designed as isotropic etch systems. To achieve a directional etch profile, different etch systems were developed. The parallel plate plasma etch system is one of those approaches. It operates at about 0.1 to 10 Torr, and the wafer sits on the grounded electrode, as shown in Fig. 9.25. Since both rf hot electrodes and grounded electrodes have similar areas, there is very little self-bias on the wafer. Both electrodes receive about the same amount of ion bombardment due to the dc bias of the rf plasma.

By increasing ion bombardment, etch rate can be increased, and more importantly, the directional etch profile can be improved. To increase ion bombardment, rf power needs to increase, and pressure must be reduced. For a parallel plate plasma etch system with equal-area electrodes, increasing rf power increases ion bombardment and etch rate on both the wafer surface and the other

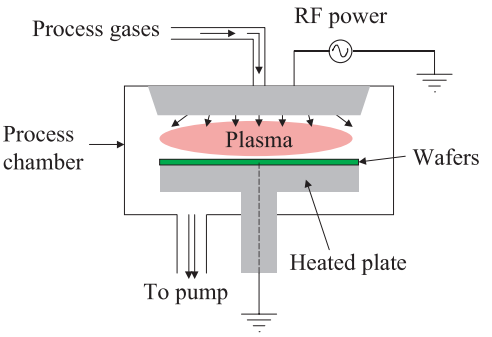


Figure 9.25 Parallel plate plasma etch system.

electrode, the chamber lid. This is very undesirable, since it causes shorter lifetimes for parts inside the chamber and increases particle contamination.

By designing an etch system with a wafer on the smaller, rf hot electrode, and taking advantage of the self-bias, the wafer surface can receive significantly high-energy ion bombardment, while the ion-bombardment grounded energy on the chamber lid is much lower. Ion energy to bombard the wafer is the sum of the plasma dc bias and the self-bias. Ion energy to bombard the chamber lid is the dc bias, which is much lower than the self-bias when the area of the wafer-side electrode is less than one half of the chamber lid area. This type of etch system is the batch RIE system and has been the most popular etch system since the 1980s. Figure 9.26 illustrates this system.

With device dimensions shrinking, requirements of etch uniformity (especially WTW uniformity) become higher. Single-wafer processing tools, which have better WTW process control, started dominating the etch process. Figure 9.27 gives a schematic of a single-wafer magnetically enhanced RIE (MERIE) system.

Reducing process pressure increases the MFP. Therefore, ions have higher energy for bombardment and less scattering due to collisions. Both are good for anisotropic pattern etch profiles. However, at low pressure, the MFP of an electron can be too long to have enough ionization collisions to generate and sustain the plasma. A magnetic field is commonly used to force electrons into a helical movement. This causes electrons to travel much longer distances and increases their chances of ionization collision. Increasing the magnetic field increases plasma density, especially at low pressure. However, it also increases electron density near the wafer surface, which reduces plasma sheath potential (or dc bias) and ion bombardment energy. Without a magnetic field, electrons in the sheath region are repelled to the bulk plasma because the wafer is negatively charged by the fast moving electrons as soon as the plasma begins. With a magnetic field, electrons are not easily expelled due to the helical motion, as shown in Fig. 9.28; this increases electron density and reduces dc bias.

Heavy ion bombardment generates a large amount of heat and raises wafer temperature if the wafer is not properly cooled. For pattern etch, a wafer is coated with a thin layer of photoresist as the pattern mask, which can be burned if the

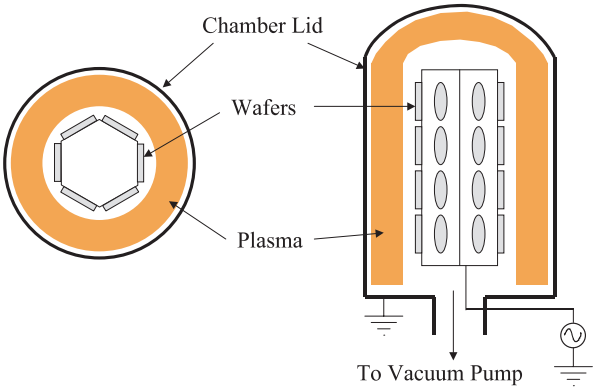


Figure 9.26 Batch RIE system.

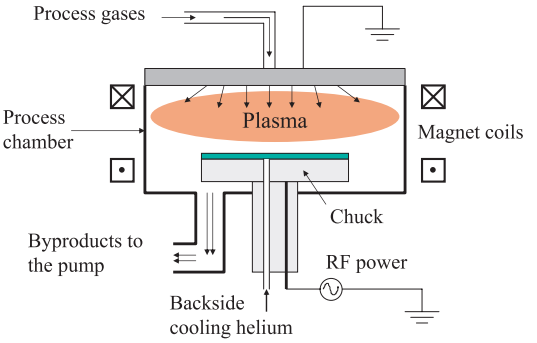


Figure 9.27 Single-wafer MERIE system.

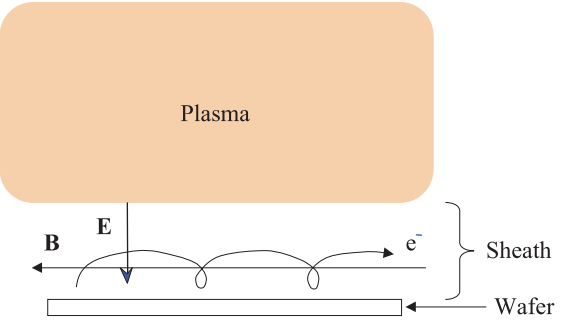


Figure 9.28 Electron helical movement in the sheath layer.

wafer temperature is more than 150 °C. Also, the chemical etch rate is sensitive to the wafer temperature. Therefore, a cooling system is required for the pattern etch chamber to prevent photoresist reticulation, and control wafer temperature and etch rate. Since an etch process needs low pressure and low pressure is not very good for heat transfer, pressurized helium at the backside of the wafer is used to remove the heat from the wafer. To prevent high-pressure backside helium from blowing the wafer off the pedestal, either a clamp ring (which mechanically holds the wafer on the pedestal) or an electrostatic chuck (E-chuck, which uses electrostatic force to hold the wafer) is required. Figure 9.29 gives the schematics of both a clamp ring and E-chuck.

The E-chuck became more popular in the 1990s because it had better etch uniformity and less particle contamination. It has better etch uniformity because it does not have the shadowing effect of the clamp ring on the wafer edge. It also has better temperature uniformity due to uniform cooling across the wafer, and no center bowing effect associated with a clamp ring. Since the E-chuck does not have mechanical contact (as the clamp ring does), it can contribute to the reduction of particles in the etch process. Helium, which has very high thermal conductivity, is commonly used to provide a conductive path for heat transfer from the wafer to a water-cooled pedestal. Some facts of helium are listed in Table 9.4.

Question: Hydrogen has a higher thermal conductivity than helium. Why not use hydrogen for wafer cooling in etch processes?

Answer: For safety reasons. Because hydrogen is explosive and flammable, whereas helium is inert, helium is much safer. As the second-best thermal conducting gas, helium can meet the cooling requirements of plasma etch systems.

Since there are always some depositions in the plasma etch process, a routine plasma dry cleaning process is used to remove most of the deposited materials left inside the chamber. However, after several thousand micron film etches, the deposition film eventually builds up, increasing the risk of particle contamination. Therefore, scheduled preventive maintenance (PM) chamber wet cleanings are needed to manually remove deposited materials from the surface of parts inside the chamber and chamber wall. Some etch chambers are designed with a liner

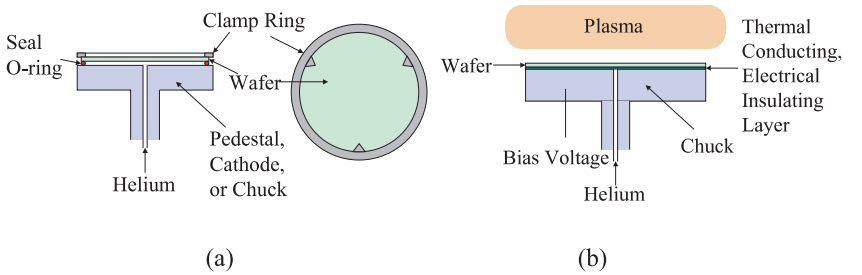


Figure 9.29 (a) Clamp ring and (b) E-chuck.

Table 9.4 Facts about helium.

Symbol	He
Atomic number	2
Atomic weight	4.002602
Discoverer	Sir William Ramsay, and independently by N. A. Langley and P. T. Cleve
Place of discovery	London, England, and Uppsala, Sweden
Discovery date	1895
Origin of name	From the Greek word helios, meaning sun. Its line radiation (a yellow line at 587.49 nm) was first detected from the solar spectrum during the solar eclipse of 1868 in India by French astronomer Pierre-Jules-César Janssen.
Molar volume	21.0 cm^3
Velocity of sound in substance	970 m/sec
Refractive index	1.000035
Melting point	0.95 K or −272.05 °C
Boiling point	4.22 K or −268.78 °C
Thermal conductivity	0.1513 $Wm^{-1}K^{-1}$
Applications	Cooling gas and carrier gas in CVD and etch processes

inside the chamber. During PM, technicians only need to replace the liner and send the dirty one out to a specialized shop for cleaning and preparation for reuse. This practice can sharply reduce system downtime for wet cleaning and improve production throughput.

As the minimum feature size continues to shrink, pattern etch processes are required to operate at even lower pressures to reduce ion scattering for better etch profile and tighter CD control. A capacitively coupled plasma source cannot generate and sustain plasma at a low pressure of a few mtorr, because the electron MFP is too long to receive enough ionization collisions. ICP and ECR are two types of plasma sources commonly used in the semiconductor industry to generate high-density plasma at low pressure for deep submicron feature pattern etch. Figure 9.30 illustrates the two systems.

One of the most important advantages of these high-density plasma systems is that ion bombardment flux and energy can be independently controlled by the source and bias rf power, respectively. In capacitively coupled plasma, both ion flux and energy are related to rf power.

9.4.6 Endpoint

For wet etch, in most cases the endpoint of the etch process is determined by time, which is calculated the by premeasured etch rate and required etch thickness. Visual inspection by an operator is also used for determining the endpoint, since there is no adequate solution for automatic endpoint determination. Wet etch rate is very sensitive to etchant temperature and concentration, which can be slightly different from station to station and lot to lot. Therefore, postetch visual inspection by an operator is always recommended.

One of the advantages of plasma etch is that it can use an optical system to automatically determine the endpoint of an etch process. At the end of an

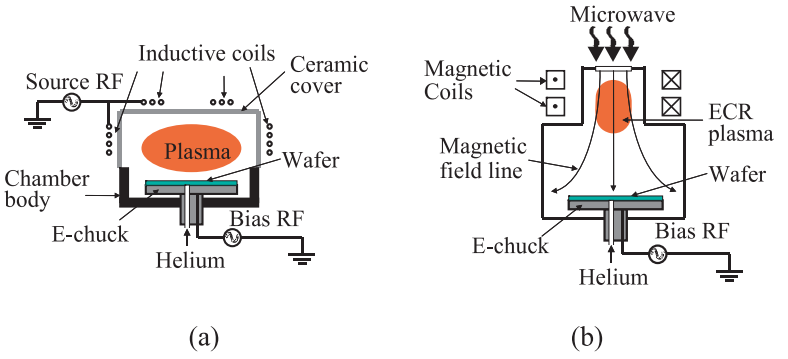


Figure 9.30 (a) ICP and (b) ECR high-density plasma etch systems.

etch process, chemical components change in the plasma, leading to a change in color and intensity of light emission from the plasma. By monitoring a certain wavelength of light emission with a spectroscope and detecting the signal change (due to the chemical changes at the end of the etch process), an electrical signal can be sent to a computer controling the etch system to automatically stop the etch process. Table 9.5 lists some of the wavelengths of chemicals used for etch endpoint detection.

For example, at the end of an aluminum etch process, the AlCl line intensity reduces due to a lack of etch byproduct $AlCl_3$ because most of the aluminum has been already etched out. This change in line intensity can be the signal to end the etch process.

Some other methods, such as pressure change, bias voltage change, mass spectrometry, etc., can also be used for endpoint detection. However, changes in pressure and bias voltage during the process are very unfavorable from a processing point of view, since they can affect process repeatability. Mass spectrometry measures the concentration of different chemical species in the processing chamber that change at the end of the etch process. Therefore, it can be used for endpoint detection. Since mass spectrometry requires a vacuum chamber and pump system, it is not cost effective compared to an optical endpoint system. It is rarely used in RIE process endpoint systems. For downstream or remote plasma etch systems, there is no plasma inside the processing chamber, therefore there is no light emission glow to provide an optical system endpoint, and mass spectrometry becomes the system of choice.

9.5 Plasma Etch Processes

9.5.1 Dielectric etch

Silicon compound-based dielectrics, such as silicon oxide, silicon nitride, and silicon oxynitride, have been widely used in IC chip manufacturing since the early 1960s. Dielectric etch is used mainly to create contact and via holes for interconnections between different layers of conductors. The etch process

Table 9.5 Wavelengths of the chemicals used for etch endpoint.

Film	Etchant	Wavelength (Å)	Emitter
Al	Cl_2, BCl_3	2614	AlCl
		3962	Al
Poly Si	Cl_2	2882	Si
		6156	O
Si_3N_4	CF_4/O_2	3370	N_2
		3862	CN
		7037	F
		6740	N
SiO_2	CF_4 and CHF_3	7037	F
		4835	CO
		6156	O
PSG, BPSG	CF_4 and CHF_3	2535	P
W	SF_6	7037	F

that creates the contact holes for connection between the first metal layer and silicon S/D and polysilicon gates is called a contact etch. The contact etch needs to etch the ILD0 layer, which normally is doped silicate glass, either PSG or BPSG.

Via etch is very similar to contact etch. It etches intermetal dielectric (IMD) or interlayer dielectric (ILD), which is mainly USG, fluorinated silicate glass (FSG), a low-κ dielectric such as SiCOH, or an even lower-κ dielectric of porous SiCOH, depending on the technology nodes. Via etch ends on a metal surface, while contact etch ends on a silicon or silicide surface.

Other dielectric etch processes are hard mask etch and bonding pad etch. Both LOCOS and STI need to etch nitride and pad oxide hard masks. For the LOCOS process, a silicon nitride layer serves as the oxidation mask, and in the STI process, it serves as a hard mask of silicon etch and the stop layer of the oxide CMP process. There are also silicon oxide hard masks for silicon alignment mark etching, and hard mask etching for copper, gold, and platinum etches. Bonding pad etch removes passivation nitride and oxide stacked film that opens the metal pad for wire bonding or contact bump formation.

Dielectric etch processes generally use fluorine chemistry with heavy ion bombardment. A damaging mechanism is employed to achieve an anisotropic etch profile. The most commonly used gases for dielectric etches are carbon fluorides, such as CF_4, CHF_3, C_2F_6, and C_3F_8. Some oxide etch systems also use SF_6 as the fluorine source gas. In normal conditions, carbon fluorides are very stable and will not react with silicon dioxide or silicon nitride. In plasma, carbon fluorides dissociate and generate very reactive free fluorine radicals. These radicals chemically react with silicon dioxide or silicon nitride to form volatile silicon tetrafluoride, which can be readily pumped away from the surface. Some facts about fluorine are listed in Table 9.6.

Table 9.6 Facts about fluorine.

Symbol	F
Atomic number	9
Atomic weight	18.9984032
Discoverer	Henri Moissan
Place of discovery	France
Discovery date	1886
Origin of name	From the Latin word fluere, meaning to flow
Molar volume	11.20 cm^3
Velocity of sound in substance	N/A
Refractive index	1.000195
Melting point	53.53 K or −219.47 °C
Boiling point	85.03 K or −187.97 °C
Thermal conductivity	0.0277 W m^{-1} K^{-1}
Applications	Free fluorine is the main etchant for silicon oxide and silicon nitride etching processes. WF_6 is widely used as a W precursor for WCVD. FSG is widely used in 90-nm technology node logic IC. F_2 excimer lasers and CaF_2 have been used in 157-nm lithography research and development.

Chemical reactions for a plasma etch of silicon dioxide and silicon nitride are

$$\mathrm{CF_4} \overset{\text{plasma}}{\rightarrow} \mathrm{CF_3} + \mathrm{F},$$
$$\mathrm{F} + \mathrm{SiO_2} \overset{\text{plasma}}{\rightarrow} \mathrm{SiF_4} + \mathrm{O},$$
$$\mathrm{F} + \mathrm{Si_3N_4} \overset{\text{plasma}}{\rightarrow} \mathrm{SiF_4} + \mathrm{N}.$$

Argon is commonly used in dielectric etch to increase ion bombardment. This helps increase etch rate and achieves the anisotropic etch profile by breaking the strong Si–O and Si–N bonds. Oxygen can be added to react with carbon and free more fluorine radicals, increasing the etch rate. However, introducing oxygen can affect the dielectric etch selectivity to the silicon and photoresist. Hydrogen can be added to improve selectivity over silicon.

In BPSG contact etching, when a contact hole reaches the silicide gate and local interconnection, contact with the S/D is only about halfway complete. While the BPSG etch process continues to reach the silicide S/D contact, overetch on the gate/local interconnection must be minimized. Therefore, the contact etch requires a very high oxide-to-silicide selectivity to prevent excessive loss of silicide during metal polycide contact. The required selectivity is $S \geq t/\Delta t$, as shown in Fig. 9.31.

For dielectric etch processes, the fluorine-to-carbon (F/C) ratio plays a very important role in etch selectivity. When F/C < 2, the polymerization process can occur, and a layer of Teflon®-like polymer can be deposited inside the processing chamber. For CF_4, the F/C ratio starts with 4. In plasma, CF_4 dissociates to CF_3 and F; F is consumed in the etching process, while CF_3 can continue to dissociate into CF_2. This sequence reduces the F/C ratio in the processing chamber. The polymerization process starts when many CF_2 radicals link into a

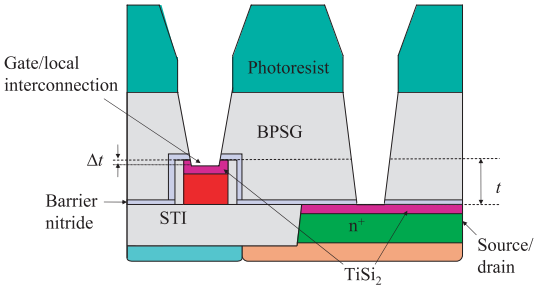


Figure 9.31 Contact etch.

long chain. Ion bombardment associated with dc bias can suppress polymerization by physically removing the deposited polymer before it builds up and forms a continuous film. Figure 9.32 shows the relationship between F/C ratio, dc bias, and the polymerization process.

For oxide etching, especially contact etch, the F/C ratio is chosen in an etching region close to the boundary of the polymerization region. While fluorine etches oxide, oxygen comes into play when fluorine replaces oxygen to bond with silicon. Oxygen reacts with carbon in carbon fluorides to form CO and CO_2. This frees more fluorine radicals to maintain the F/C ratio in the etching regime. When the etch process reaches the silicon or silicide surface, fluorine is consumed while carbon is not because there is no oxide as an etch byproduct, since the etched film does not contain oxygen. The F/C ratio decreases at the silicon or silicide surface, and the process quickly moves into the polymerization regime and begins depositing polymer, which stalls the etch process on the silicon or silicide surface. This provides a mechanism for achieving very high oxide-to-silicon/silicide etch selectivity. It allows oxide etch to continue for another few thousand angstroms to reach the S/D, while contact to the silicide gate/local interconnection is already reached (shown Fig. 9.31). Polymer deposited on the silicon or silicide surface can be removed by either an oxygen plasma ashing process (which also removes photoresist), or by a wet cleaning process.

Since dielectric etch mainly uses the damaging mechanism, this process is very physical, with heavy ion bombardment to the wafer surface when the wafer is in the etching region. The wafer must be very close to the boundary of the polymerization regime to achieve high selectivity to silicon or silicide. At the same time, the larger grounded electrode (the chamber lid) is in the polymerization region. It encounters much less energetic ion bombardment than the wafer surface. Thus, polymerization is always happening in the dielectric etch chamber during contact and via etch processes. Polymers deposited in the plasma etch chamber need to be removed with a routine O_2/CF_4 plasma cleaning process to prevent particle contamination caused by cracking of the polymer film. For etch chamber cleaning, it can be necessary to put a dummy wafer on the chuck to protect it from ion bombardment damage. A seasoning step always follows the cleaning process to intentionally deposit a thin layer of polymer inside the chamber to prevent loose residue falling from the

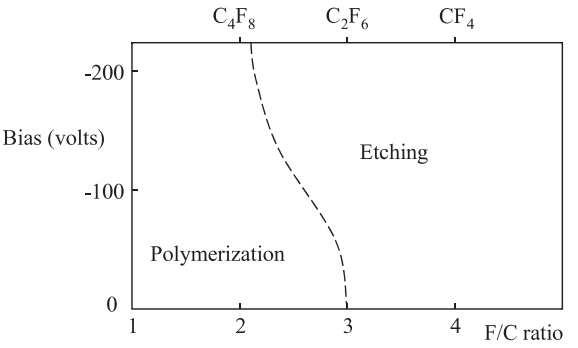


Figure 9.32 Relationships between F/C ratio, dc bias, and polymerization (From *Plasma Etching and Reactive Ion Etching*, J. W. Coburn).

chamber wall, and to keep the process in similar processing condition and prevent "first-wafer" effects.

Densities of wordlines (WLs) and contact holes are quite different from SRAM in logic devices, DRAM cell arrays, and NAND flash cell arrays, as shown in Fig. 9.33. The line/space ratio of a NAND flash WL is about 1:1, DRAM WL is about 1:2, and SRAM gate CD and gap is about 1:3. Figure 9.33 shows that NAND flash does not require a contact hole between WLs. In fact, it only needs contact holes to connect to bit or source lines for a string of 32 (32-bit) or 64 (64-bit) WLs. However, DRAM and SRAM need a contact between gates.

For DRAM applications, because spacing between the array transistor gates or WLs is significantly smaller than that of logic devices at the same technology nodes, there is little room for contact holes between WLs. To avoid shorting between the contact and WL in such a dense WL pattern, a self-aligned contact (SAC) process has been developed. As shown in Fig. 9.34, the SAC process etches BPSG, with a high selectivity to silicon nitride, which forms a hard mask of the WL stack and sidewall spacer that fully covers WLs. The SAC process allows high contact hole density of DRAM applications. After polysilicon deposition and CMP, it provides landing pads for bitline contact (BLC) and storage node contact (SNC). Therefore, SAC is also called landing pad contact (LPC) after etching, and landing pad poly (LPP) after polysilicon CMP.

For copper metallization, more dielectric etches are added in IC manufacturing processes. Besides via etch, there are trench etches. Two approaches are used: via first and via last. The via first process is illustrated in Fig. 9.35. It starts with via mask lithography to define the via holes, then etches the via holes and stops in the middle of the etch-stop layer (ESL). After being stripped, the photoresist is used to fill the via holes to protect them during trench etching. Trench mask lithography defines the photoresist patterns, and trench etch transfers the patterns to a low-κ dielectric ILD film. The ESL is broken through during the cleaning process and is called the cap layer because it is on top of any previous metal layer. ESL is formed with a thin layer of silicon nitride (SiN), silicon oxynitride (SiON), or silicon carbon nitride (SiCN).

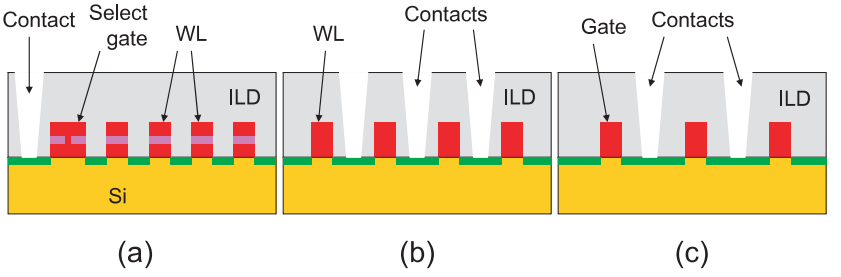


Figure 9.33 Wordlines and contact hole densities of different memory devices: (a) NAND flash, (b) DRAM, and (c) SRAM.

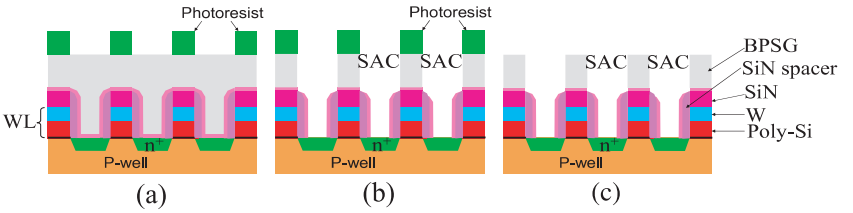


Figure 9.34 DRAM SAC process: (a) SAC mask lithography, (b) SAC etch, and (c) photoresist strip.

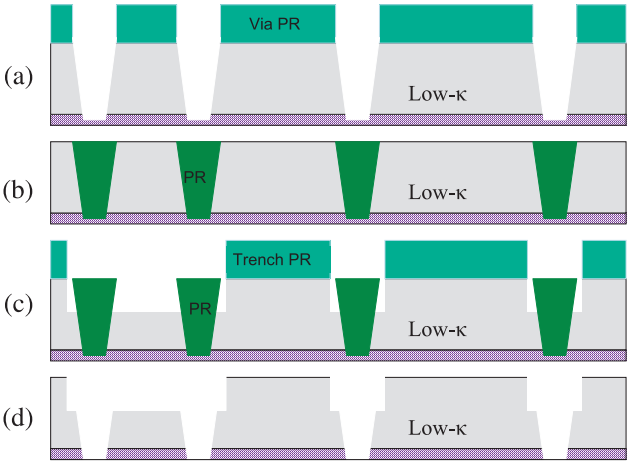


Figure 9.35 A via-first dual damascene process: (a) via etch, (b) photoresist fill and etchback, (c) trench etch, and (d) ESL breakthrough.

Figure 9.36 illustrates a trench-first dual damascene process. After trench mask lithography, trenches are etched onto ILD film, as shown in Fig. 9.36(a). For advanced nanometer technology node devices, ILD film is a low-κ dielectric film. After via mask lithography [illustrated in Fig. 9.36(b)], via holes are etched on the ESL, which is broken through after the photoresist cleaning step.

Because many low-κ and ultralow-κ (ULK) dielectric films are organosilicate glass (OSG) or porous OSG with large amounts of carbon and hydrogen integrated into the films, they can be damaged by oxygen radicals in photoresist ashing and stripping processes. To protect low-κ or ULK films from damage during photoresist striping processes, a metal hard mask such as titanium nitride (TiN) is commonly used for ILD etch. After ILD deposition, the metal hard mask layer is deposited.

Figure 9.37 illustrates a copper metallization process with TiN and TEOS oxide hard masks. BARC is a spin-on material coated onto the wafer surface before photoresist spin-coating. Tetraethoxysilane or tetraethylorthosilicate (TEOS) is a widely used precursor for silicon oxide CVD processes. TEOS is also commonly known in IC fabs as undoped silicon oxide deposited by a PECVD process using TEOS as precursor. The PECVD TEOS process is described in detail in Chapter 10. Trench masks are used to pattern photoresist first, then BARC and hard masks are etched with the photoresist pattern, as shown in Fig. 9.37(a). TiN and TEOS hard masks can prevent low-κ or ULK films from being damaged by oxygen radicals during the two photoresist stripping processes. After photoresist ashing, a via mask is patterned, and via holes are etched more than halfway, as shown in Fig. 9.37(b). A metal hard mask is used to etch the low-κ dielectric trenches and via holes simultaneously (with high selectivity to ESL), so the via holes stop on the ESL, as illustrated in Fig. 9.37 (c). The TiN hard mask stays on the wafer surface after copper deposition, shown in Fig. 9.37(d), and eventually is removed by the metal CMP of copper metallization, shown in Fig. 9.37(e).

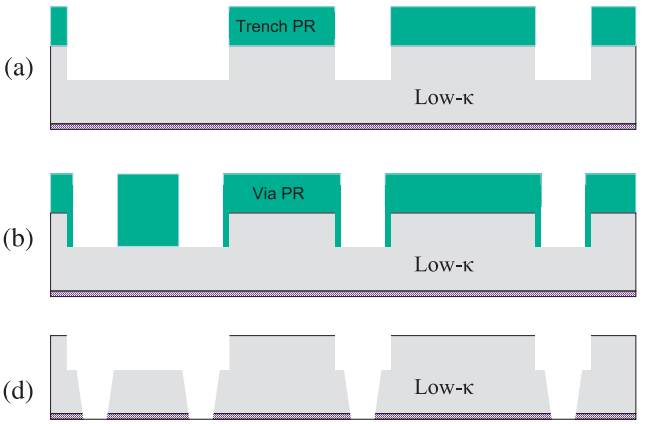


Figure 9.36 A trench-first dual damascene process: (a) trench etch, (b) via mask lithography, (c) via etch, and (d) ESL breakthrough.

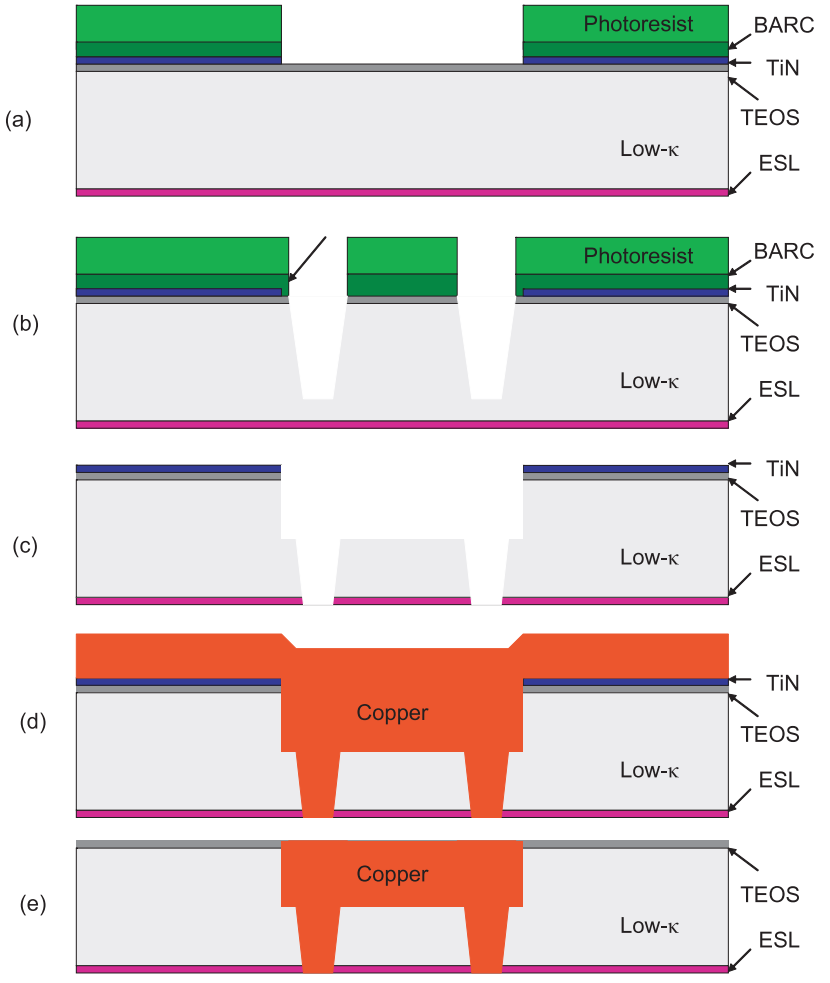


Figure 9.37 Copper metallization with a metal hard mask: (a) etch bottom antireflective coating (BARC) and metal hard mask with trench mask, (b) via etch halfway with a via mask, (c) trench and via etch with metal hard mask and ESL breakthrough, (d) copper deposition, and (e) metal CMP to remove the metal hard mask.

The main etch processes of low-κ or ULK dielectrics can use CF_4, CHF_3, or C_4F_8 chemistry, with argon for ion bombardment. Other fluorocarbon gases, such as C_4F_6 (hexafluorobutadiene) and c-C_5F_8 (octafluorocyclopentene), can also be used with O_2 in low-κ and ULK dielectric etching. Carbon monoxide (CO) can also be used to control the F/C ratio. Reducing the F/C ratio in the etchant gas can also help improve etch selectivity of the low-κ dielectric to the photoresist. Table 9.7 summarizes dielectric etch processes.

Table 9.7 Summary of dielectric etch processes.

Name of etch	Hard mask	Contact	Via (Al-Cu)	Via/trench (Cu/low-κ)	Bonding pad
Materials	Si_3N_4 or SiO_2	PSG or BPSG	USG or FSG	Low-κ or ULK	Nitride and oxide
Etchants	CF_4, CHF_3, etc.	CF_4, CHF_3, etc.	CF_4, HF_3, etc.	CF_4, CHF_3, CO, etc.	CF_4, CHF_3, etc.
Underlying layer	Si, Cu, Au,	Polysilicon or silicide	TiN/Al-Cu	ESL/Cu	Metal
Endpoints	CN, N, or O	P, O, and F	O, Al, and F	CN, O, and F	O, Al, and F

9.5.2 Single-crystal silicon etch

Single-crystal silicon etch is required for forming STI, which has been used in advanced IC chips since the mid-1990s, when IC technology nodes went to a half-micron. STI has replaced LOCOS field oxide as the isolation between neighboring microelectronic devices. This is because STI does not have "bird's beak," and the surface topography of STI is much flatter than that of LOCOS.

Single-crystal silicon etch is also needed for deep trench capacitor formation in DRAM chip manufacturing. The capacitor packing density is increased by forming the capacitor in a vertical direction deep in the silicon substrate, as shown in Fig. 9.38. Deep trench capacitor DRAM is widely used as embedded DRAM in system-on-chip (SOC) IC chips because it is more compatible with standard CMOS processes.

Figure 9.39(a) shows a stack capacitor DRAM with planar array nMOS, which is commonly used for DRAM with technology nodes larger than 100 nm. Figure 9.39(b) shows a stack capacitor DRAM recessed gate array nMOS, which is widely used in DRAM with sub-100-nm technology. In Fig. 9.39, BL and BLC in the cell array area are marked with a dashed outline because they are not in the cross section. Instead, they are behind and in front of the cross section, sandwiched between the SNC. To avoid shorting between the SNC polysilicon plug and tungsten BL, a silicon nitride liner is deposited on the sidewall of the SNC, as shown in Fig. 9.39(b). SN stands for storage node.

By using single-crystal silicon etch with a reversed cell gate mask, recessed gate electrodes of cell array nMOS can be created, which can help reduce the short channel effect due to shrinking of the gate feature size. It can be seen that the stack capacitor DRAM with a planar array transistor only needs one single-crystal silicon etch that forms the STI, while the recessed gate array transistor DRAM needs two single-crystal silicon etches, one for STI and another for recessed gate.

Buried wordline (bWL) DRAM technology (illustrated in Fig. 9.40) has been developed based on deep trench DRAM and recessed gate transistor technology. For recessed gate array transistors, the gate electrode is buried under the wafer surface. Conducting wires with polysilicon/metal/nitride stacks above the wafer surface serve as WLs of the DRAM cells. By placing the cell array nMOS and WL under the wafer surface, the bWL DRAM reduces a masking step because

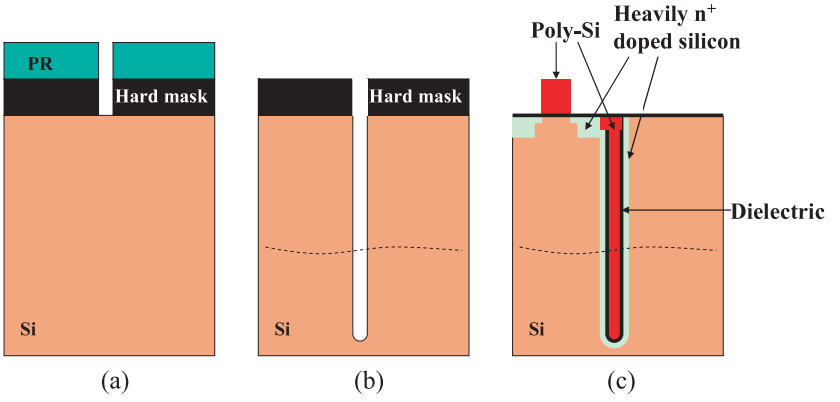


Figure 9.38 Deep trench capacitor DRAM formation: (a) hard mask etch, (b) single-crystal silicon etch, and (c) deep trench capacitor DRAM formation.

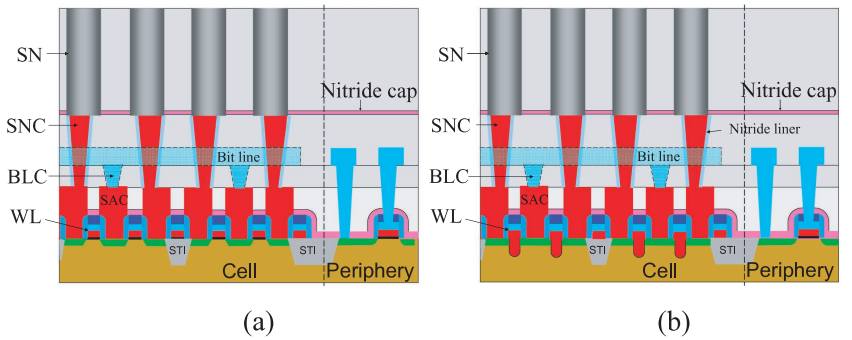


Figure 9.39 (a) Cross section of a stack capacitor DRAM with planar array nMOS and (b) with recessed gate array nMOS.

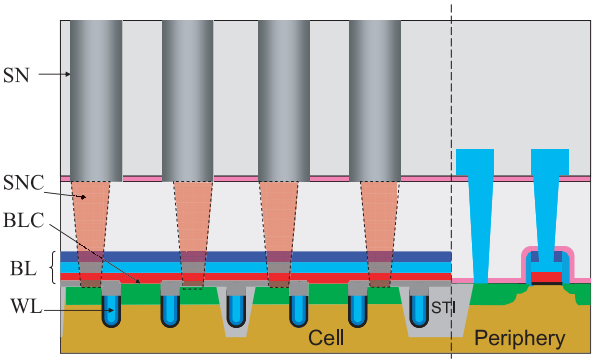


Figure 9.40 Buried WL DRAM.

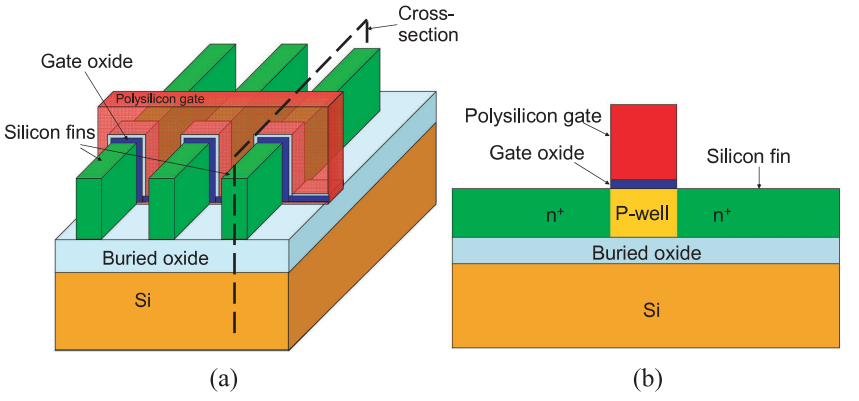


Figure 9.41 (a) Three FinFETs and (b) the cross section of one of them.

WL patterning is no longer required. By removing the topography of the wafer surface due to the dense WL pattern, another masking layer is reduced, since the SAC layer is not needed. The tradeoff is the challenge of WL trench etching, which must be well controlled for CD, depth, and profile. It also must achieve nearly the same etch rate and etch profile in both single-crystal silicon and in the STI oxide. Both deposition and etchback of the metal layers, TiN and W, in the trench also need to be well controlled. Another challenge is that bWL DRAM array transistors are nMOS with metal gates, usually titanium nitride (TiN), not the polysilicon gates commonly used in both planar and recessed gate array nMOS. Also, for bWL DRAM, polysilicon and metal stacks (usually polysilicon/WN/W stack) are used as gate electrodes for the periphery MOSFETs, which are also used as bitline stacks. Therefore, it is very challenging to form bitline contacts that connect bitlines to the array nMOS while still maintaining the integrity of the gate oxide of the periphery MOSFETs.

Single-crystal silicon etch is also needed for making 3D devices such as FinFET, which is regarded by many as one of the most promising candidates for next-generation devices when the widely used planar MOSFETs can no longer scale. Figure 9.41(a) illustrates three FinFETs with a shared gate electrode built on an SOI substrate. Figure 9.41(b) shows the cross section of the FinFET along the dashed surface of Fig. 9.41(a). It can be seen from the cross section that the FinFET looks just like a planar MOSFET made with an SOI substrate. Figure 9.41(a) shows that the silicon fins can be formed with a single-crystal silicon etch. It is significantly easier to form silicon fins with SOI wafers than with bulk wafers because the buried oxide can be used as an etch endpoint, and the height of the fin is easier to control because it is the thickness of the silicon on the buried oxide.

FinFET can also be made from a bulk silicon wafer, which requires better control of the single-crystal silicon etch process, such as CD, depth, and profile. The height of the silicon fin is controlled by the amount of recess of the STI oxide, as shown in Fig. 9.42. The CD and height of the silicon fins can be measured by atomic force microscope (AFM) and scatterometry.

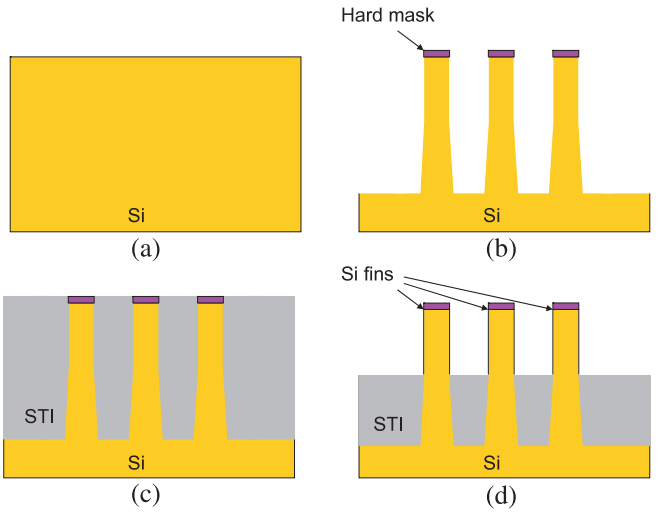


Figure 9.42 Fin formation with a bulk silicon wafer: (a) bulk silicon substrate, (b) trench etch, (c) STI oxide fill and CMP, and (d) STI oxide recess.

Single-crystal silicon etch normally employs silicon dioxide (or in many cases, both silicon dioxide and silicon nitride) hard masks instead of photoresists to avoid contamination, as illustrated in Fig. 9.42(b). It uses a blocking mechanism, with HBr as the main etchant and O_2 as a sidewall passivation agent. In plasma, HBr dissociates and releases free bromine radicals, which can react with silicon to form volatile silicon tetrabromide ($SiBr_4$). Oxygen oxidizes silicon on the sidewalls to form silicon dioxide, which protects silicon from the free bromine radicals. At the bottom of the trench, ion bombardment keeps the oxide from growing, thus the etch process continues only in the perpendicular direction. Some facts about bromine are listed in Table 9.8.

The main chemical reaction of single-crystal silicon plasma etching processes is

$$\mathrm{HBr} \overset{\text{plasma}}{\rightarrow} \mathrm{H} + \mathrm{Br},$$
$$\mathrm{Br} + \mathrm{Si} \overset{\text{plasma}}{\rightarrow} \mathrm{SiBr_4}.$$

Oxygen is used to improve the selectivity over oxide hard masks. It is also used to react with etch byproduct $SiBr_x$ to form and deposit $SiBr_xO_y$ on the trench sidewall. This protects the sidewall and restricts the etch to mainly the vertical direction because $SiBr_xO_y$ deposited on the trench bottom is constantly removed by ion bombardment. Fluorine source gases such as SiF_4 and NF_3 can be used to improve etch profile control on the trench sidewall and bottom. Fluorine is also needed to achieve an etch rate similar to single-crystal silicon and silicon oxide, required by bWL DRAM.

Single-crystal silicon etch takes two steps—breakthrough and the main etch. A brief breakthrough step removes the thin native oxide on the silicon surface

Table 9.8 Facts about bromine.

Symbol	Br
Atomic number	35
Atomic weight	79.904
Discoverer	Antoine J. Balard
Place of discovery	France
Discovery date	1826
Origin of name	From the Greek word bromos, meaning stench
Molar volume	19.78 cm^3
Velocity of sound in substance	No data
Resistivity	$>10^{18} \mu\Omega \cdot cm$
Refractive index	1.001132
Melting point	−7.2 °C
Boiling point	59 °C
Thermal conductivity	0.12 $W\ m^{-1}\ K^{-1}$
Applications	Free bromine as the main etchant for single-crystal silicon etching processes
Source	HBr

with heavy ion bombardment and fluorine chemistry. The main etch step etches silicon with HBr and O_2 (diluted to 30% in He) chemistry. After the etch process is finished, the wafer requires wet cleaning processes to remove sidewall deposition. A significant difference in single-crystal silicon etch compared with other plasma etch processes is that it does not have an underlying layer. Therefore, it cannot be endpointed by an optical signal and usually endpoints by time.

Single-crystal silicon etch chambers always acquire some depositions of complex compounds of silicon, bromine, oxygen, and fluorine on their walls. These must be removed with a routine plasma cleaning process with fluorine chemistry to control particle contamination. Similar to other cleaning processes, a seasoning step is needed after the cleaning process.

9.5.3 Polysilicon etch

Polysilicon gate etch is the most important etch process because it defines the gates of MOSFET. Polysilicon gate etch has the smallest CD of all the etching processes. In the good old days, when people talked about *N*-micron technology, it indicated that the gate CD was *N* μm.

When feature size shrinks deep into the nanometer range, gate CD and technology nodes no longer remain the same number. Technology node is mainly determined by the pitch of gate patterns or portion of the gate pitch. The definition of a technology node is also different for different devices. The NAND flash technology node is half-pitch: 20-nm NAND flash has 40-nm WL pitch, usually with a 20-nm gate CD and 20-nm gap CD. DRAM technology nodes are about one-third WL pitch; a 33-nm DRAM has about a 99-nm WL pitch with ~30-nm gate CD. Technology nodes of CMOS logic devices are defined as about one-quarter of SRAM gate pitch because there is contact between the gates. For example, a 22-nm SRAM device (published in 2008 by Haran) has a gate pitch of 90 nm and gate CD

of 25 nm. Figure 9.43 illustrates cross sections of array transistors of NAND flash, DRAM, and SRAM.

Figure 9.44 shows Intel's six-transistor (6T) SRAM size scaling timeline and top-view SEM images of unit cells of the 6T SRAM from 90- to 22-nm technology nodes after polysilicon gate etch. The SRAM layout had a revolutionary change starting with the 65-nm node, which was quite different from the layout of the 90-nm node. After the 45-nm node, a double-patterning technique must be applied in gate patterning. As the technology node continued to shrink, the CDs of MOSFET gates had trouble continuing to shrink, and IC designers started to reduce the gaps between the gates.

MOSFETs with polysilicon gates require polysilicon etch to form gate patterns. MOSFETs with HKMG must etch polysilicon, whether it is used as a gate-first approach or gate-last technique. In fact, the polysilicon gates of the Intel SRAMs with 45-, 32-, and 22-nm technology nodes shown in Fig. 9.43 will be removed after ILD0 CMP and replaced by metal stacks to form the metal gates in a gate-last processes. Because the gaps between gates are so narrow for 32- and 22-nm SRAMs, trench-style contact, with etch processes similar to the SAC of DRAM, have been developed and applied to form the contact plug.

Figure 9.45 illustrates a polysilicon etch process that forms CMOS gates and local interconnections. This process uses photoresist as the etch mask, and the etch stops on gate and STI oxides.

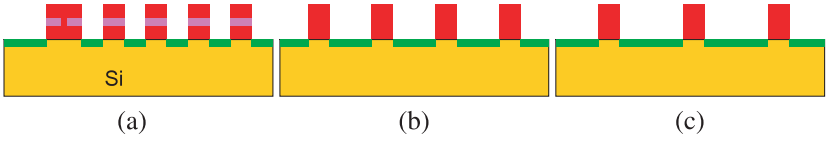


Figure 9.43 Cross sections of array transistors for different memory devices with the same technology node: (a) NAND flash, (b) DRAM, and (c) SRAM.

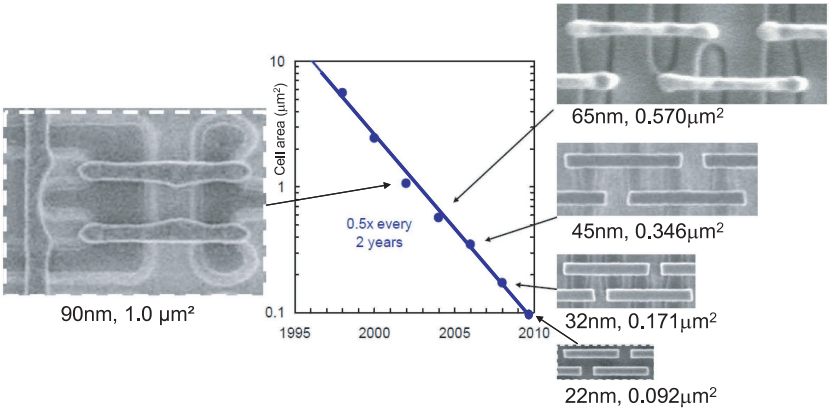


Figure 9.44 Intel SRAM size scaling (Intel).

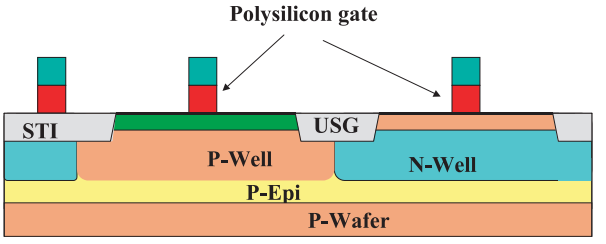


Figure 9.45 Polysilicon gates and local interconnection etches.

Cl_2 is the most popular main etchant for polysilicon etch processes. In plasma, Cl_2 molecules are dissociated to generate free chlorine radicals, which are very reactive and can react with silicon, forming gaseous silicon tetrachloride ($SiCl_4$). Table 9.9 lists some facts about chlorine.

Chlorine tends to combine with photoresist materials and deposit a thin polymer layer on the sidewall. This can help achieve an anisotropic etch profile and reduce CD bias (polysilicon CD loss or gain compared to the CD of photoresist). HBr can also be used as a secondary etchant and sidewall passivation promoter. O_2 can be used to improve selectivity over oxide.

One of the biggest challenges for polysilicon gate etching is the high selectivity to silicon dioxide, since underneath the polysilicon is the ultrathin gate oxide. For a 45-nm device, thickness of the gate oxide is about 12 Å, which is equivalent to only two layers of a silicon dioxide molecule! Since neither the etch rate nor polysilicon film thickness is perfectly uniform, some parts of the polysilicon may already be etched away, while for some other parts the etch is still going on (shown in Fig. 9.46). The thin gate oxide cannot afford to be lost, otherwise the etchant that removes polysilicon will also remove the single-crystal silicon underneath the

Table 9.9 Facts about chlorine.

Symbol	Cl
Atomic number	17
Atomic weight	35.4527
Discoverer	Carl William Scheele
Place of discovery	Sweden
Discovery date	1774
Origin of name	From the Greek word chloros, meaning pale green
Molar volume	17.39 cm^3
Velocity of sound	206 m/sec
Resistivity	$>10^{10}$ μΩ·cm
Refractive index	1.000773
Melting point	−101.4 °C or 171.6 K
Boiling point	−33.89 °C or 239.11 K
Thermal conductivity	0.0089 W m^{-1} K^{-1}
Applications	Used as the main etchant for polysilicon and metal etching processes; and for polysilicon, epitaxy silicon deposition chamber cleaning.
Sources	Cl_2, HCl

gate oxide, causing silicon pit defects. Therefore, the selectivity to oxide must be high enough during the polysilicon overetch process.

Figure 9.46 shows requirements of overetch for a polysilicon etch process. The etch process in the higher-etch-rate portions of the wafer have already reached the gate oxide [left side of Fig. 9.46(a)], while in the lower-etch-rate portion there is still a remaining thin layer of polysilicon needing to be removed [right side of Fig. 9.46(a)]. Assume the etch rate has 3% nonuniformity. Thus, for a perfectly uniform 50-nm polysilicon film, the remaining polysilicon thickness is about 1.5 nm. By only allowing ~0.5 Å loss of gate oxide, the overetch selectivity should be higher than 30:1. For the LOCOS case illustrated in Fig. 9.46(b), residue polysilicon on the sidewall of the field oxide needs to be etched away, while in the active area the polysilicon is already etched out. Selectivity is determined by the ratio of the sidewall residue thickness and the tolerated gate oxide loss. If the sidewall residue thickness is 1500 Å and the maximum gate oxide loss is 30 Å, selectivity of the polysilicon-to-silicon dioxide during the overetch needs to be at least 1500:30 = 50:1.

Polysilicon etch has three steps: breakthrough, main etch, and overetch. The breakthrough step removes the thin native oxide (~10 to 20 Å) with high ion bombardment and sometimes uses fluorine chemistry. The main etch step removes polysilicon from designated areas and forms the gate and local interconnect patterns. In the overetch step, the etch condition is changed to remove the polysilicon residue while minimizing the gate oxide loss. The main step etches polysilicon at a high rate. The selectivity over silicon dioxide is not a concern, since the etch process has not reached the oxide yet. As soon as the etchants start to etch the gate oxide, oxygen will be released from the film and penetrate the plasma. Optical spectroscopic sensors monitoring oxygen line radiation detect an intensity increase of the oxygen radiation line and trigger the endpoint to stop the main etch process and start the overetch process. The overetch process flows oxygen, reduces rf power, and reduces Cl_2 flow to improve polysilicon-to-oxide selectivity. Some fabs have a phororesist trimming step before breakthrough to remove any residue photoresist in the bottom of the gap between the two photoresist patterns. It also can reduce gate CD and improve device speed.

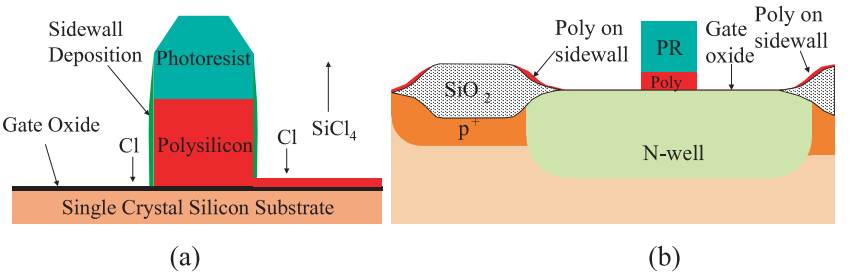


Figure 9.46 Polysilicon overetch requirement.

For advanced technology node IC chips, in order to pattern the fine patterns with 193-nm ArF optical lithography, photoresist patterns alone cannot meet the requirements of etch bias. Dielectric hard masks on top of the polysilicon are commonly used. This dielectric layer sometimes serves as an ARC. Above the dielectric hard mask, a spin-on BARC layer is added prior to photoresist coating. These are commonly called trilayers. After photoresist patterning by a photolithography process, BARC can be etched with oxygen plasma, and the dielectric hard mask can be etched with fluorine chemistry and argon ion bombardment, which is similar to a breakthrough step. The main etch of polysilicon can use fluorine chemistry to endpoint (by means of an optical system monitor) the thickness variation of the polysilicon. The overetch chemistry is usually HBr with O_2, which allows for very high selectivity of the gate oxide.

Fluorine can be used to etch polysilicon. Some polysilicon etch processes use SF_6 and O_2 chemistry. Because fluorine etches silicon dioxide faster than chlorine does, it has a lower selectivity to oxide; therefore most fabs prefer to use chlorine chemistry for the main etch step.

In some DRAM gate (WL) processes, tungsten (W) or tungsten silicide (WSi_2) is deposited on top of the polysilicon to reduce the resistance of the WL and local interconnections. The etch process for this silicide–polysilicon stack film requires one more step to etch W or WSi_2, first with fluorine chemistry, then switching to chlorine chemistry to etch the polysilicon.

9.5.4 Metal etch

Metal etch has been used to form the interconnection lines for an IC to connect transistors and circuit units. For some CMOS IC chips with mature technology (and even advanced DRAM and flash chips), the metal layer consists of three layers: TiN ARC, aluminum copper alloy, and titanium nitride plus titanium or a titanium tungsten welding layer. The TiN ARC layer reduces reflection from the aluminum surface and helps improve lithography resolution. The Al–Cu alloy is the main metal material for carrying the electric current and forming global interconnections. The Ti, TiN/Ti, or TiW layer can help reduce contact resistance between the Al–Cu and tungsten plug. This layer also prevents the copper in aluminum from diffusing into the silicate glass, reaching the silicon substrate, and causing device damage.

The most commonly used chemical for metal etch is chlorine. In plasma, Cl_2 dissociates and generates the free radicals Cl, which react to TiN, Al, and Ti and form the volatile byproducts of $TiCl_4$ and $AlCl_3$:

$$\mathrm{Cl_2} \xrightarrow{\text{plasma}} \mathrm{Cl + Cl},$$
$$\mathrm{Cl + Al} \xrightarrow{\text{plasma}} \mathrm{AlCl_3},$$
$$\mathrm{Cl + TiN} \xrightarrow{\text{plasma}} \mathrm{TiCl_4 + N},$$
$$\mathrm{Cl + Ti} \xrightarrow{\text{plasma}} \mathrm{TiCl_4}.$$

Metal etch normally uses Cl_2 as the main etchant, and BCl_3 is commonly used for sidewall passivation. It is also used as a secondary Cl source and a heavy ion source to provide the BCl_3^+ ion with ion bombardment. In some cases, argon is also used to increase ion bombardment. N_2 and CF_4 can be employed to improve sidewall passivation.

Question: Why does no one use fluorine to etch a TiN/Al-Cu/Ti metal stack?

Answer: The main byproduct of an aluminum-fluorine reaction is AlF_3, which has very low volatility. At normal etch conditions, ~100 mtorr, <60 °C, AlF_3 is a solid. Therefore, fluorine cannot be used in a plasma etch process to pattern the TiN/Al-Cu/Ti metal stack.

It is very important that metal etch processes have good profile control, residue control, and metal corrosion prevention. During the metal etch process, a small amount of copper in the aluminum can cause a residue problem because $CuCl_2$ has very low volatility and will stay on the wafer surface. It can be removed by physical ion bombardment that dislodges it from the surface. It can also be removed by chemical etching, which cuts $CuCl_2$ off of the surface. Since both $CuCl_2$ particles and the wafer surface are negatively charged by the plasma, $CuCl_2$ can be pushed away from the surface by electrostatic force. It is very important for the metal etch process to strip the photoresist in situ before the wafer is exposed to moisture in the atmosphere. Otherwise, chlorine residue in the photoresist and sidewall deposition will react with H_2O to form HCl and cause metal corrosion problems.

For the gate-first approach of a HKMG process, the stack film of a dielectric hard mask, the polysilicon, and TiN must be gate etched, as shown in Fig. 9.47. The etching process is similar to a normal polysilicon gate etching process: first a breakthrough etch is required to pattern the dielectric hard mask, then the main etch process removes bulk polysilicon with fluorine plasma, and then the metal etch step uses Cl or HBr to etch TiN with high selectivity to the capping layers and high-κ dielectric layer. For polysilicon gate etch, oxygen can be added to improve etch selectivity to silicon oxide. However, for a metal gate etch, adding oxygen to the plasma can result in the oxidation of TiN, which forms TiO_2 and causes metal gate loss.

TiN etch is also needed to pattern hard masks of ULK dielectric in copper low-κ interconnection formation, as shown in Fig. 9.37(a).

9.5.5 Photoresist strip

After the etch process, the photoresist has served its purpose and needs to be stripped. Photoresist can be stripped with either a wet or dry process. Oxygen is most commonly used for dry stripping. Water vapor (H_2O) is usually added to plasma to provide an extra oxidation agent (HO) to remove photoresist, and hydrogen radicals (H) to remove the chlorine on the sidewall and in the photoresist. For the metal etch process, it is very important to strip the photoresist right after the etch process, before the wafer is exposed to moisture in the atmosphere. This is

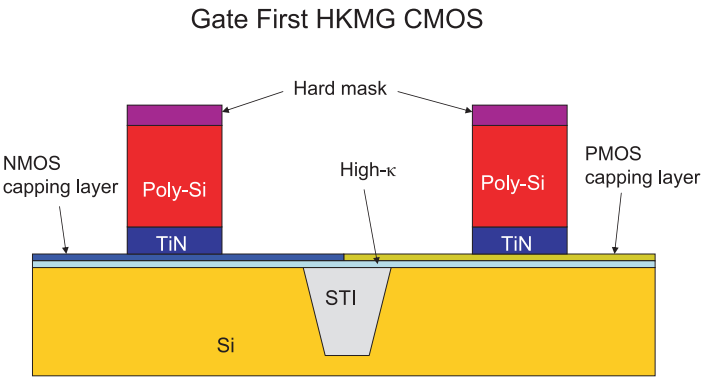


Figure 9.47 Gate etch of the gate-first approach for a HKMG device.

because the chlorine absorbed on the sidewall deposition and photoresist will react with moisture and form hydrochloric acid, which can etch aluminum and cause metal corrosion. The basic chemistry involved in the photoresist stripping process is

$$\begin{aligned} O_2 &\overset{\text{plasma}}{\rightarrow} O + O \\ H_2O &\overset{\text{plasma}}{\rightarrow} 2H + O, \\ H + Cl &\rightarrow HCl, \\ O + PR &\rightarrow H_2O + CO + CO_2 + \cdots. \end{aligned}$$

Figure 9.48 shows a schematic of a photoresist stripping chamber with a remote plasma source. It can be put in line with the etch chambers to strip the photoresist in situ in the same mainframe.

9.5.6 Dry chemical etch processes

Dry chemical etch can be achieved by either using thermally unstable chemical gases, such as XeF_2 and ozonator-generated O_3, or by using a remote plasma source, which generates free radicals in a remote plasma chamber and flows them into the processing chamber. Since these unstable gases are expensive and difficult to store, remote plasma processes are more commonly used in IC fabrication.

By using a remote plasma source and taking advantage of plasma-generated chemically active free radicals that are free from ion bombardment on the wafer surface, a dry chemical etch can be applied to thin-film stripping and wine-glass contact etch processes. The advantage of a dry chemical etch over a wet one is that it can be placed in line with another RIE chamber on one system. It can process wafers in situ and improve production throughput. One example is the wine-glass contact shown in Fig. 9.15. The remote plasma etch chamber can be placed in line with the RIE chamber in the same mainframe. First, the wafer is in the remote plasma etching chamber for isotropic etch, then it is transferred

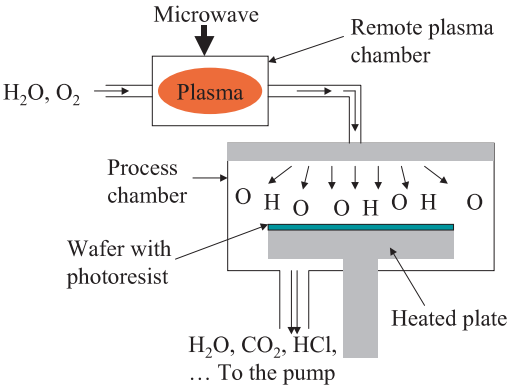


Figure 9.48 A remote plasma photoresist stripping system.

to the RIE chamber for anisotropic etch. Another application for remote plasma etch processes is silicon nitride layer stripping in the LOCOS process, and both silicon nitride and polysilicon layer stripping in poly-buffered LOCOS (PBL) processes.

9.5.7 Blanket dry etch processes

Blanket plasma etch removes material from the entire wafer surface. It does not have a photoresist pattern on the wafer surface. The main applications of blanket etch processes are etchback, spacer formation, and film strip.

Argon sputtering etchback is a pure physical etch process; it removes materials by physically dislodging tiny pieces of the material from the surface with energetic ion bombardment. It is widely used in dielectric thin-film applications to taper gap openings. It increases accessibility to CVD precursor molecules and improves gap-fill abilities. Ar sputtering is also widely used to remove native oxide from wafer surfaces prior to metal deposition.

RIE etchback combines both the physical and chemical etch processes. It can be used in line with a dielectric CVD tool for sidewall spacer formation so that the CVD chamber can deposit a conformal dielectric film on the pattern, as shown in Fig. 9.49(a). The RIE etch chamber can etch back the conformal dielectric film from the sidewall spacer, as shown in Fig. 9.49(b). RIE etch can also be applied with a tungsten CVD tool to form tungsten plugs. In this case, a WCVD chamber deposits tungsten into contact holes or via holes and all across the wafer surface. An RIE etch chamber with fluorine plasma had been used to remove bulk tungsten from the surface. The RIE process has also been used in photoresist or SOG etchback processing for dielectric planarization.

9.5.8 Plasma etch safety

There are some safety issues in plasma etch processes, as they involve certain chemicals that are corrosive and toxic, such as Cl_2, BCl_3, SiF_4, and HBr. Inhalation

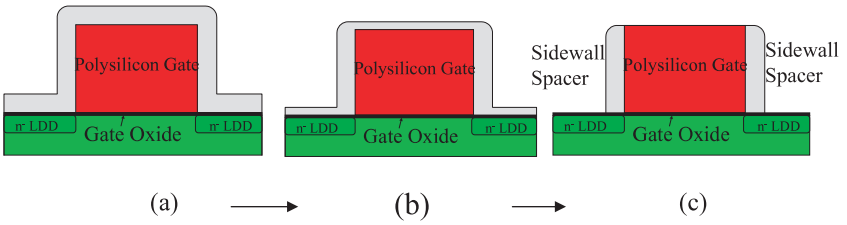


Figure 9.49 RIE etchback for sidewall spacer formation: (a) CVD conformal dielectric film, (b) etchback dielectric film, and (c) formation of sidewall spacer.

of these gases at a high concentration (>1000 ppm) can be fatal. Carbon monoxide (CO) is a colorless and odorless gas. It is flammable and can cause flash fires. It is toxic and harmful if inhaled because it combines with hemoglobin in the blood and decreases the amount of oxygen delivered to tissues, causing blood damage, difficulty breathing, and even death.

Radio frequency power can cause electric shock, and at high power can be lethal. All moving parts, including robot and vacuum valves, are mechanical hazards and can cause injury. Lock out and tag out are important for preventing someone from turning on a system from the cleanroom side while others are still working on the system in a gray area.

9.6 Process Trends

For wet etch, etch rate is mainly determined by temperature and etchant concentration. Increasing temperature increases etch rate due to the higher chemical reaction rate and higher etchant and etch byproduct diffusion rates. Increasing etchant concentration also increases etch rate. Selectivity is mainly determined by the chemistry of the etchant and the materials involved in the etching process. Wet etch normally has very good selectivity. For wet etch, the etch profile is always isotropic, and there is no way to control it. Etch-rate uniformity is mainly determined by the uniformity of the etchant solution temperature and concentration. Solution and wafer agitation can help to improve etch uniformity. The endpoints of wet etch processes are determined by time and detected by the visual inspection of an operator.

For plasma etch, especially the commonly used parallel plate electrode rf plasma, etch rate is most sensitive to rf power. Increasing rf power increases both ion bombardment flux and ion bombardment energy. This significantly enhances physical etch rate and damaging effects. Increasing rf power also increases free-radical concentration, an increase that enhances the chemical etch effect. Therefore, if the etch rate of a plasma etch system is outside of specifications, the rf system should be checked first. The rf system includes the source, cables and connections, and the rf match. Increasing rf power causes the RIE process to become more physical. It normally reduces etch selectivity, especially the selectivity to photoresist, due to physical sputtering. The relationship between etch rate and selectivity to rf power is illustrated in Fig. 9.50.

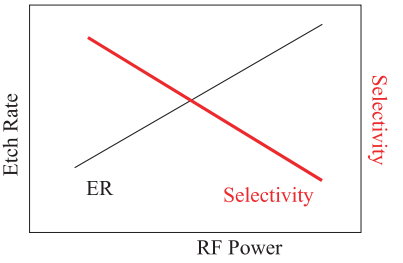


Figure 9.50 Etch process trends with rf power.

Pressure mainly controls etch uniformity and profile. It can also affect etch rate and selectivity. Changing pressure changes the MFP of electrons and ions. It affects the uniformity of the plasma and etch rate. By increasing pressure, the MFP becomes shorter, causing more collisions for the ions. Ion energy decreases, and ion collisional scattering increases, increasing the chemical etch components in the RIE process. Normally, increasing the pressure increases etch rate if the etch process is chemically dominant and reduces etch rate if the etch is physically dominant.

Increasing the magnetic field increases plasma density. Therefore, this increase also increases the ion bombardment flux, which increases the physical etch component. However, it decreases ion energy by reducing the sheath bias, and increases free radical concentration, an increase that can cause etch to be more chemical. Normally, at low pressure, when the magnetic field is weak, physical etch improvement outweighs chemical etch. Increasing the strength of the magnetic field causes the etch to become more physical. When magnetic field strength attains a certain value, the etch becomes more chemical due to the lesser ion energy caused by the reduction of dc bias. Figure 9.51 summarizes etch trends when rf power, pressure, and magnetic field strength B increases.

If the etch chamber is leaking, the etching rate to the photoresist increases due to the presence of oxygen in the plasma. The selectivity to photoresist degrades, and particle count increases. Insufficient photoresist hard bake during the photolithography process can also cause high photoresist etch rates and excessive photoresist loss during the etch process.

Since each etch process requires different chamber designs, uses different chemicals, and operates at different conditions, processing trends can be quite different from one another. Normally, a tool vender provides the basic processing information, including process trends and troubleshooting guides for the tools.

9.7 Recent Developments

To create better anisotropic etch profiles and reduce CD loss, lower pressure is desired because it can increase MFP and reduce ion scattering. Increasing plasma density can increase ion bombardment flux; for a certain amount of ion bombardment requirements, increasing ion bombardment flux can reduce the

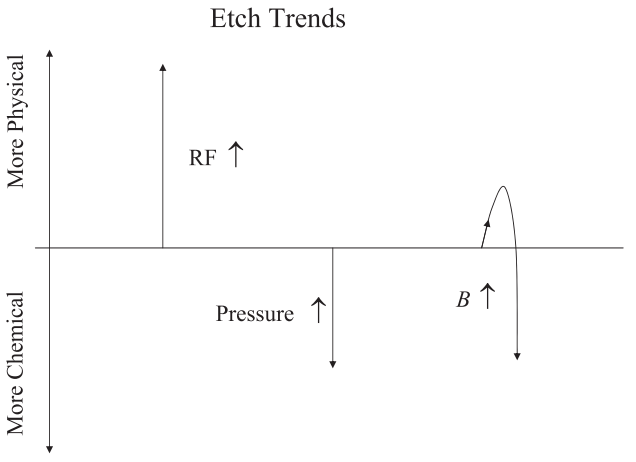


Figure 9.51 Summary of etch trends.

required ion energy, a method that is a favorite for reducing device damage. Plasma chambers with higher density at lower pressures are the desired chamber designs for achieving higher etch rates and better profile control. ICP and ECR etch chambers have been developed to meet these requirements. Both can generate high-density plasma at a low pressure and can independently control plasma density and ion bombardment energy, a feature that is very important for etch process control. For both ICP and ECR plasma sources, the ionization rate is not very high, about 1 to 5%. In advanced 300-mm fabs, ICP etch systems are widely used to etch conductive materials such as silicon, polysilicon, and metal layers, while capacitively coupled plasma etch systems are commonly used for dielectric etch processes. A helicon plasma source (illustrated in Fig. 9.52) can achieve almost 100% ionization rate at a low pressure of a few millitorr. It could very well be one of the candidates for future etch chamber designs.

Another concern is plasma uniformity control, especially for larger wafer sizes. Plasma uniformity and position control will become more important in the future. For the mature technology node, high throughput and low cost of ownership are key factors to competing with the existing in-production systems. If a reliable etch system can be manufactured with a low cost of ownership that can save money for customers in the long run, it would be possible for IC manufacturers to replace the existing systems and introduce the new low-cost systems. The key is that yield and up-time should be the same or higher than existing systems, throughput should be higher, and the consumable lower, so that customers can justify replacement if one year of operation savings can pay off the cost of the equipment.

New materials have recently been added to IC chip manufacturing processes, such as HKMG and ULK dielectrics. Etch of these new materials is one of the challenges for etch process developers. New device structures have also been applied in advanced IC chip manufacturing, such as 3D FinFET, trigate devices,

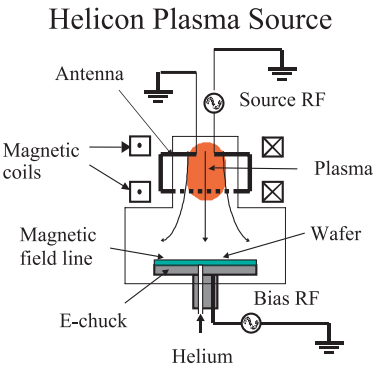


Figure 9.52 Etch chamber with a helicon plasma source.

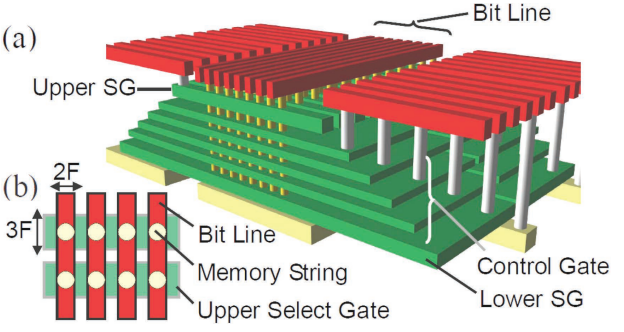


Figure 9.53 (a) Birds-eye view and (b) top view of a 3D flash memory array (*IEDM Technical Digest*).

vertical devices, etc. For 3D FinFET, a single-crystal silicon etch process becomes more challenging, especially for a FinFET built on bulk silicon wafers. For advanced bWL DRAM, an etch process with about 1:1 etch selectivity of single-crystal silicon and silicon oxide is required. One of the approaches for increasing the storage density of NAND flash is to adapt a 3D stacked cell structure, as shown in Fig. 9.53, which illustrates four layers of NAND flash cells used to form a string of 4-bit flash memory. Real applications probably require 16-bit strings. The etch process of holes for memory strings can be very challenging, and an etch process for contact holes of WLs that have different depths would be even more challenging. 3D packaging such as through-silicon via (TSV) processes provide more challenges and opportunities for etch processes. Different from submicron and nanometer pattern etch processes in front-end-of-line (FEoL) processes, TSV etch processes have significantly larger CDs (50 to 5 μm), and require significantly higher etch rates to achieve the required reasonable throughput.

9.8 Summary

- The four materials that need to be etched in IC chip fabrication are single-crystal silicon, polysilicon, dielectric (silicon oxide and nitride), and metals (TiN, Al–Cu, Ti, W, and WSi_2).
- Four major etch processes are silicon, poly, dielectric, and metal etches.
- Wet etch uses a chemical solution to dissolve material that needs to be etched.
- Dry etch uses chemical gases to remove material from the substrate surface by physical etch, chemical etch, or a combination of both.
- Wet etch has high selectivity, high etch rate, and low equipment cost. The isotropic etch profile limits its applications in pattern etch to feature sizes larger than 3 μm.
- Wet etch is still widely used in advanced semiconductor fabs for thin-film stripping and dielectric thin-film quality control.
- In plasma etch processes, etchant flows into the chamber and dissociates in the plasma, and free radicals diffuse across the boundary layer and are adsorbed on the surface. The free radicals react with an atom or molecule on the surface with the help of ion bombardment. Volatile byproducts desorb from the surface, diffuse across the boundary layer, and are pumped out of the chamber in the convection flow.
- There are two anisotropic etch mechanisms—damaging and blocking. Dielectric etches normally use a damaging mechanism. Silicon, polysilicon, and metal etches use a blocking mechanism.
- Dielectric etch uses fluorine chemistry. CF_4, CHF_3, and Ar are commonly used for the dielectric etch processes. CF_4 is the main etchant, and CHF_3 is a polymer precursor used to improve etch selectivity to photoresist and silicon. Ar is used to enhance ion bombardment. O_2 can be used to increase etch rate, and H_2 can be used to improve etch selectivity to photoresist and silicon.
- For low-κ and ULK dielectric etching, CO can be added to improve etch process control.
- Silicon etch uses HBr as a main etchant, and O_2 and fluorine compounds are used for sidewall deposition.
- Polysilicon can be etched with Cl_2 or SF_6. O_2 is normally used to improve the selectivity to oxide, and HBr is used to help with sidewall protection.
- Metal etch uses Cl_2, BCl_3 and N_2 to improve sidewall passivation.
- Copper metallization does not require metal etching; it requires dielectric trench etches.
- Photoresist stripping after ULK dielectric etch becomes very challenging, and a metal hard mask (TiN) is commonly used.

9.9 Bibliography

D. G. Baldwin, M. E. Williams, and P. L. Murphy, *Chemical Safety Handbook for the Semiconductor/Electronics Industry*, 2nd ed., OME Press, Beverly, MA (1996).

R. Bowman, G. Fry, J. Griffin, D. Potter, and R. Skinner, *Practical VLSI Fabrication for the 90s*, Integrated Circuit Engineering Corp., Raleigh, NC (1990).

J. W. Coburn, *Plasma Etching and Reactive Ion Etching*, AVS Monograph Series, M-4, American Institute of Physics, Inc., New York (1982).

J. W. Coburn, H. F. Winters, "Ion- and electron-assisted gas-surface chemistry: an important effect in plasma etching," *J. Appl. Phys.* **50**, 3189–3196 (1979).

Y. Fukuzumi et al., "Optimal integration and characteristics of vertical array devices for ultrahigh-density, bit-cost scalable flash memory," *IEDM Tech. Dig.*, 449–452 (2007).

S. K. Ghandhi, *VLSI Fabrication Principles*, 2nd ed., Wiley-Interscience Publication, John Wiley & Sons, Inc., New York (1994).

B. S. Haran et al., "22-nm technology compatible fully functional 0.1 μm^2," *IEDM Tech. Dig.*, 625–628 (2008).

D. M. Manos and D. L. Flamm, *Plasma Etching, An Introduction*, Academic Press, San Diego, CA (1989).

I. Morey and A. Asthana, "Etch challenges of low-κ dielectric," *Solid State Technol.* **42**(6), 71–78 (1999).

T. Schloesser et al., "A 6F2 buried wordline DRAM cell for 40 nm and beyond," *IEDM Tech. Dig.*, 809–812 (2008).

See http://www.webelements.com for further information on the elements in this chapter.

9.10 Review Questions

1. What is the pattern etch process?
2. Define etch selectivity.
3. What are the differences between wet etch and RIE?
4. What is the chemical most commonly used to wet etch silicon dioxide? What is the special safety concern of that particular chemical?
5. Why is wet etch preferred for thin-film stripping processes?
6. Explain the two anisotropic etch mechanisms.
7. Name the etch process that has the smallest critical dimension.
8. Why does the polysilicon gate etch process require high polysilicon-to-oxide selectivity?
9. Why does polysilicon etch prefer chlorine over fluorine as the main etchant?
10. How does the F/C ratio affect oxide etch processes?
11. Why isn't fluorine used as the main etchant for Al–Cu metal etches?

12. Explain the benefit of a low-pressure, high-density plasma source in pattern etch processes.
13. What metallic materials have been used for hard masks in low-κ and ULK dielectric etch?
14. After low-κ or ULK dielectric etch, is a metal etch process required to remove the metal hard mask after photoresist stripping?

Chapter 10
Chemical Vapor Deposition and Dielectric Thin Films

In the semiconductor industry, electrically insulating materials are commonly called dielectrics. Dielectric thin-film processes are adding processes because they add a thin layer of dielectric materials to the wafer surface. While the majority of dielectric thin-film processing is CVD processing, spin-on dielectric processes are also used in IC manufacturing. The main issues with dielectric thin-film processes are filling gaps without voids, deposition of uniform film with high throughput, and making the final surface as planarized as possible.

Objectives

After finishing this chapter, the reader will be able to:

- list two dielectric thin-film materials used in IC chip fabrication
- identify at least four applications of dielectric thin films from a CMOS chip cross section
- describe the CVD processing sequence
- list the two deposition regimes and describe their relation to temperature
- name the two most commonly used silicon precursors for dielectric CVD.

10.1 Introduction

There are two types of dielectric thin films widely used in the semiconductor industry: thermally grown and deposited thin films. Thermally grown dielectric thin films are discussed in Chapter 5. In this chapter, we cover deposited dielectric thin films. The fundamental difference between thermally grown and deposited thin films is that grown film consumes silicon from the substrate, while deposited film does not. Figure 10.1 shows the difference between silicon dioxide films. For thermally grown silicon dioxide, oxygen comes from the gas phase, and silicon comes from the silicon substrate. This process consumes silicon from the substrate, and the film grows into the substrate. For CVD oxide, both silicon and oxygen come from the gas phase, so there is no consumption of silicon substrate.

The quality of thermally grown oxide film is much better than that of deposited film. Therefore, the gate oxide of a MOSFET always uses thermally grown silicon

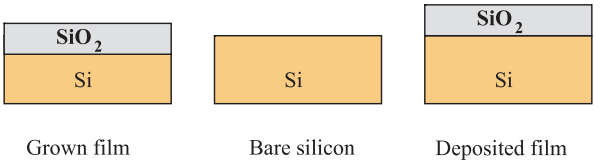


Figure 10.1 Thermally grown and deposited thin films.

dioxide or nitridated thermal oxide. There are many applications of deposited dielectric thin films in IC chip fabrication, including electrical isolation, ion implantation blocking, dopant source, ARC, hard mask, capping layer, ESL, and circuit passivation.

The main role of dielectric thin film is as a dielectric layer for electrical insulation in multilevel metal interconnections. Both CVD dielectrics and the combination of spin-on and CVD dielectrics are used for this application. It is also used in STI as electrical insulation between neighboring transistors. It is used to form spacers on the sidewalls of polysilicon or polycide (polysilicon/silicide stack) gate that are needed to form an LDD or SDE and the diffusion buffer. A passivation dielectric seals the IC chip to protect the tiny circuit from chemical damage due to moisture and mobile ions. It is used as an ESL for dual damascene copper metallization. It also can be used as a cap barrier layer for low-κ or ULK dielectrics. It is also used as a passivation layer to protect the IC chip from mechanical damage during testing and packaging. The CVD or spin-on dielectric ARC is used to minimize reflection from the wafer surface to meet the resolution requirements of photolithography processes. The commonly used metallic ARC can no longer meet the resolution requirements once pattern feature size becomes smaller than 0.25 μm.

Applications of dielectric thin films in a CMOS IC with Al–Cu interconnection are illustrated in Fig. 10.2. Different companies use different acronyms for dielectric layers in the interconnection applications. Many companies use ILD, which stands for interlayer dielectric. For the dielectric between the polysilicon and first metal layer, some fabs call it a premetal dielectric (PMD), while some others call it ILD0. Some fabs call the dielectric layers between the metal layers an intermetal dielectric (IMD), while others call them ILD-*X* (*X* from 1 to whatever the number of metal layers minus one).

In the first edition of this book, PMD and IMD were the only acronyms used for dielectrics because at that time most interconnections in IC chips were aluminum–copper alloys (Al–Cu). For IC chips with Al-Cu interconnections, the deposition conditions of PMD and IMD are quite different. IMD layers used USG and were deposited at around 400 °C due to the limitations of deposition temperature of Al–Cu alloys. PMD normally used a doped oxide, either PSG or BPSG. Deposition temperature was limited by the thermal budget, which was the diffusion tolerance of the dopants in the activation area defined by the device design. However, device structure has since been changed. For example, fabs are manufacturing IC chips with metal gates with both gate-first and gate-last approaches,

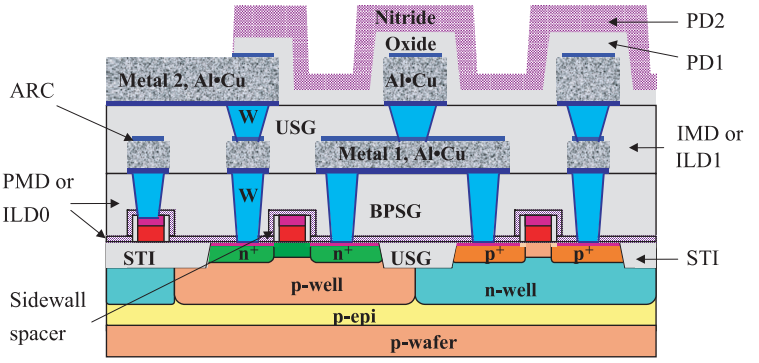


Figure 10.2 Applications of dielectric thin film in a CMOS circuit with Al–Cu interconnection.

thus it no longer makes sense to call the dielectric between a gate and the first metal wire a PMD. Also, for the bWL DRAM, metals (TiN and W) are used for array transistor gate electrodes, thus PMD is no longer a correct description of the dielectric. Therefore, in this updated edition, the acronym ILD is used to describe the dielectric layers for the interconnection, either with Cu or with Al–Cu alloys.

Figure 10.3 illustrates a cross section of a CMOS chip with five layers of Cu/low-κ interconnection. More advanced IC chips can have up to 11 metal layers, which need 12 ILD layers. Thus, for an advanced IC fab, many dielectric CVD tools are dedicated to ILD deposition. For an N-layer metal interconnection IC chip with STI, the minimum number of dielectric layers is:

$$\underset{\text{dielectric layer}}{} = \underset{\text{STI}}{1} \quad \underset{\text{spacer}}{+1} \quad \underset{\text{ILD}}{+N+1} \quad \underset{\text{PD}}{+1} = N+4 \tag{10.1}$$

Some advanced IC chip devices use double or even triple spacers, and embedded DRAM in SOC devices requires dielectric deposition between the two capacitor electrodes. For these chips, there are more dielectric deposition steps than $N + 4$.

10.2 Chemical Vapor Deposition

CVD is a process in which a gaseous chemical precursor (or precursors) has a chemical reaction on the wafer surface and deposits a solid byproduct as a layer of thin film. Other byproducts are gases and leave the surface.

CVD processes are widely used in the semiconductor industry for various thin-film depositions, such as epitaxial silicon, polysilicon, dielectric thin film, and metal thin-film depositions.

Epitaxial and polysilicon depositions are the high-temperature CVD processes discussed in Chapter 5, and the metal thin-film CVD process is covered in Chapter 11. Dielectric CVDs are covered in detail in this chapter. Table 10.1 summarizes some of the important CVD thin films used in IC manufacturing.

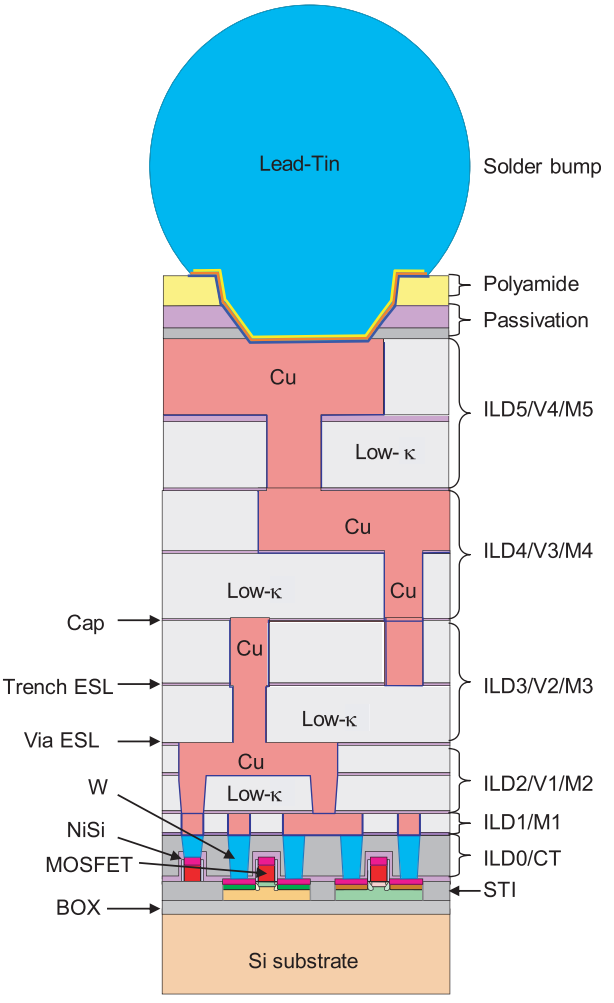


Figure 10.3 Dielectric layers in a copper/low-κ interconnection application.

10.2.1 Chemical vapor deposition process description

CVD is a sequential process, which includes several steps:

- gas or vapor phase precursors are introduced into the reactor
- precursors diffuse across the boundary layer and reach the substrate surface
- precursors adsorb on the substrate surface
- adsorbed precursors migrate on the substrate surface
- chemical reaction occurs on the substrate surface
- solid byproducts form nuclei on the substrate surface
- nuclei grow into islands
- islands merge into the continuous thin film

Table 10.1 CVD thin films and precursors used in the IC industry. Note that Cu (hfac) (vtms) stands for hexaflouroacetyl-acetonato-copper-vinyl-trimethylsilane.

	FILMS	PRECURSORS
Semiconductor	Si (poly)	SiH_4 (silane)
		$SiCl_2H_2$ (DCS)
	Si (epi)	$SiCl_3H$ (TCS)
		$SiCl_4$ (Siltet)
Dielectrics	Oxide	SiH_4, O_2
		SiH_4, N_2O
		$Si(OC_2H_5)_4$ (TEOS), O_2
		TEOS
		TEOS, O_3 (ozone)
	Oxynitride	SiH_4, N_2O, N_2, NH_3
		SiH_4, N_2, NH_3
	Si_3N_4	SiH_4, N_2, NH_3
		$C_8H_{22}N_2Si$ (BTBAS)
	Low-κ	3MS (trimethylsilane),
		4MS (tetramethylsilane), etc., and O_2
	ULK	DEMS (diethoxymethylsilane) and $C_6H_{10}O$ (cyclohexene oxide)
	W (Tungsten)	WF_6 (tungsten hexafluoride), SiH_4, H_2
Conductors	WSi_2	WF_6 (tungsten hexafluoride), SiH_4, H_2
	TiN	$Ti[N(CH_3)_2]_4$ (TDMAT)
	Ti	$TiCl_4$
	Cu	Cu (hfac) (vtms)

- other gaseous byproducts desorb from the substrate surface
- gaseous byproducts diffuse across the boundary layer
- gaseous byproducts flow out of the reactor.

Figure 10.4 illustrates the sequential process, which includes the precursor gases being introduced, precursors diffusing across the boundary layer, byproducts diffusing across the boundary layer, and byproducts flowing out of the reactor.

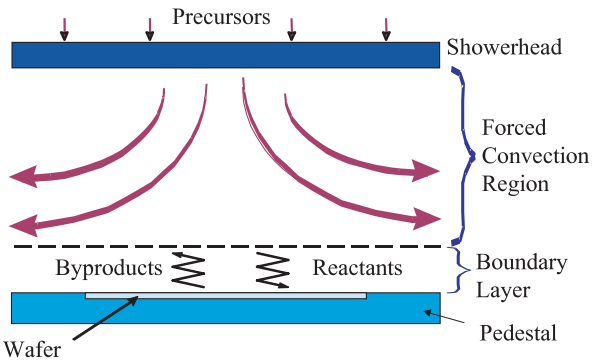


Figure 10.4 CVD process sequence.

Figure 10.5 shows that precursors reach the substrate surface by diffusing across the boundary layer, adsorbing, and migrating on the substrate surface. The ability of the precursor to migrate on the surface is called surface mobility, which is very important for thin-film step coverage and gap-fill properties. This is covered in a later section of this chapter. When precursors have chemical reactions on the surface, they form into a solid material and release gaseous byproducts. The first few molecules of the solid material form nuclei on the surface, and additional chemical reactions cause the nuclei to grow into islands. The islands grow, merge, and finally form a continuous thin film on the wafer surface.

Figure 10.5 can also be used to illustrate the physical vapor deposition (PVD) process. The major difference between CVD and PVD is at the nucleation step. CVD has a chemical reaction on the substrate surface, while PVD does not.

10.2.2 Chemical vapor deposition reactor types

There are three types of CVD reactors commonly used in the semiconductor industry: atmospheric pressure (APCVD), low-pressue (LPCVD), and plasma-enhanced (PECVD).

10.2.2.1 Atmospheric pressure chemical vapor deposition

Atmospheric pressure equals 760 torr at sea level at 0 °C. Figure 10.6 illustrates an APCVD system. It has three zones: two nitrogen buffer zones and a processing zone in between. Nitrogen curtains in both buffer zones separate the processing gases and prevent them from leaking into the atmosphere. Heater elements heat the wafer carriers or disks, which in turn heat the wafers. A conveyor belt constantly transports wafers into the processing zone. Precursor chemical gases react on the heated surface and deposit a thin film. The deposited wafers are removed from the carrier. After several depositions, the thin-film-coated carriers are sent to a clean shop to remove the coated thin film before being reused. The conveyor belt is also coated with a layer of deposited thin film as it passes through the processing zone. Therefore, it needs constant cleaning to maintain a stable processing condition. The APCVD process is controlled by temperature, processing gas flow rate, and belt speed.

The APCVD process has been used to deposit silicon oxide and silicon nitride. The APCVD O_3-TEOS oxide process is widely used in the semiconductor industry, especially in STI and ILD0 applications.

Question: A semiconductor manufacturer has its research and development laboratory on the coast near sea level and one of its manufacturing fabs on a high-altitude plateau. It was discovered that APCVD processes developed in the laboratory could not be directly transferred to that particular fab. Why?

Answer: Atmospheric pressure is related to altitude. On a high-altitude plateau, the atmospheric pressure is significantly lower than that at sea level. The earlier APCVD reactor did not have a pressure-control system, so

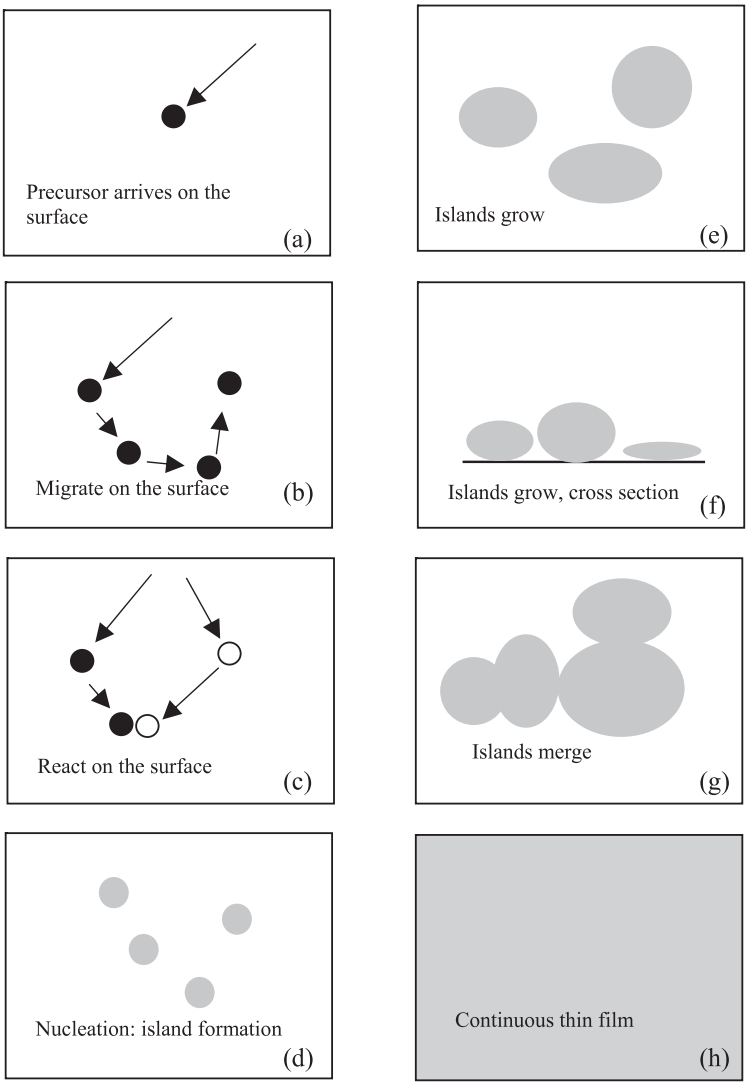


Figure 10.5 Deposition processes.

processing conditions were quite different between the laboratory and the manufacturing fab. A process that worked fine in one location might not work well in the other because of the pressure difference.

10.2.2.2 Low-pressure chemical vapor deposition

LPCVD operates at a low pressure, from 0.1 to 1 torr. The LPCVD reactor is very similar to an oxidation furnace. It has three heating zones, and wafers are

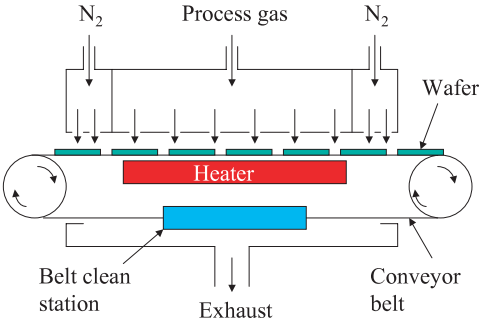


Figure 10.6 The APCVD system.

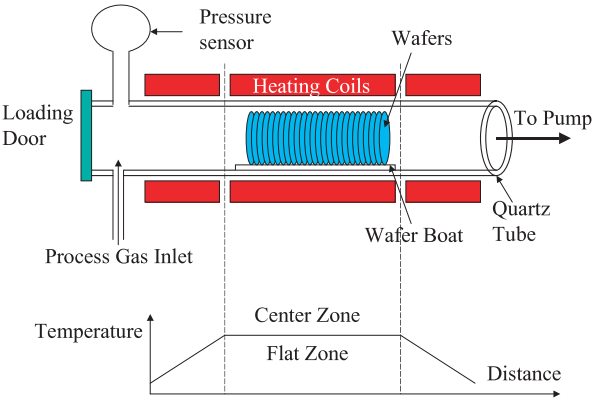


Figure 10.7 The LPCVD system.

processed in the center zone or flat zone. LPCVD systems require vacuum systems to control pressure inside the reactors. LPCVD reactors operate in surface-reaction-limited regions, which are covered in a later section. The deposition process is mainly controlled by wafer temperature and is relatively insensitive to gas flow rate. Therefore, wafers can be loaded vertically with very little spacing required. The large wafer load improves throughput and reduces wafer processing costs compared with APCVD processes. Figure 10.7 illustrates the LPCVD system.

LPCVD processes have been used to deposit oxide, nitride, and polysilicon. The majority of polysilicon and amorphous silicon are deposited in LPCVD reactors. Silicon nitride layers used for oxidation masks in LOCOS processes and for CMP hard stop layers in STI processes are deposited in LPCVD processes. LPCVD is also commonly used to deposit diffusion barrier silicon nitride layers to block diffusion of the dopant atoms in doped oxide from diffusing through the thin gate oxide layer and into activation areas.

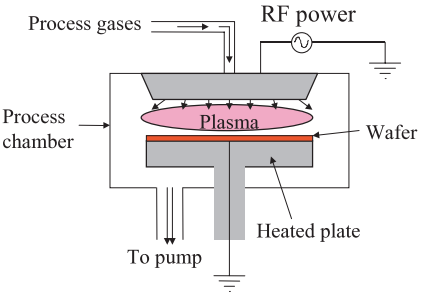


Figure 10.8 The PECVD reactor.

The LPCVD reactor operates at a high temperature, usually higher than 650 °C. Therefore, it cannot be used for the IMD depositions after the first metal layer deposition.

10.2.2.3 Plasma-enhanced chemical vapor deposition

The operating pressure of PECVD usually is from 1 to 10 torr. Because free radicals generated from the plasma dramatically increase the chemical reaction rate, PECVD can achieve high deposition rates at relatively low temperatures, which is very important for ILD depositions. Figure 10.8 gives an illustration of a PECVD reactor.

Another advantage of PECVD is that deposited film stress can be controlled by rf power without having a major effect on deposition rate. PECVD processes are widely used for oxide, nitride, low-κ, ESL, and other dielectric thin-film depositions.

10.2.3 Chemical vapor deposition basics

10.2.3.1 Step coverage

Step coverage is a measurement of the deposited film reproducing the slope of a step on a substrate surface. It is one of the most important specifications of the CVD process. Definitions of sidewall step coverage, bottom step coverage, conformality, and overhang are given in Fig. 10.9.

Step coverage is determined by both the arriving angle and precursor surface mobility. The arriving angle is illustrated in Fig. 10.10. It can be seen that corner A has the largest arriving angle (270 deg), and corner C has the smallest. Therefore, corner A will have more precursor atoms or molecules when they diffuse across the boundary layer. If these precursors react immediately after being adsorbed on the wafer surface without migration, corner A will have more deposition than corner C and will form overhangs as shown in Fig. 10.10.

The arriving angle can also be modified by controlling etch processes. For example, most fabs etch contact holes with a tapered angle instead of a straight hole (shown in Fig. 10.11) to create a larger arriving angle, and to make it easier for tungsten CVD processes to fill the contact hole.

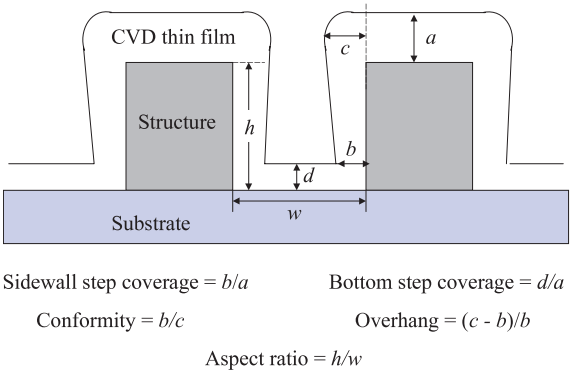


Figure 10.9 Step coverage and conformality.

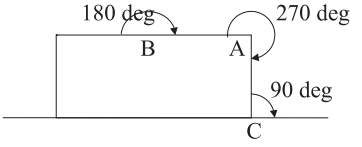


Figure 10.10 Arriving angles.

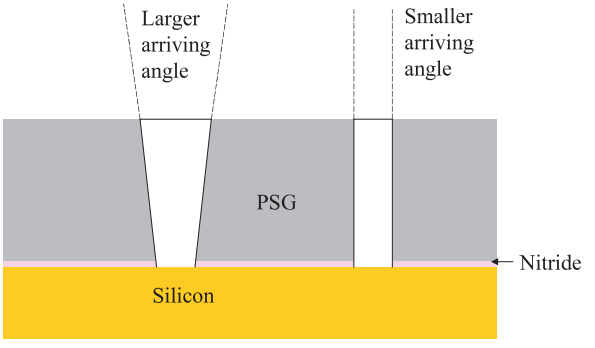


Figure 10.11 Arriving angles for tapered and straight contact holes.

Overhangs are very undesirable. If a deposited film starts with overhangs caused by the arriving angle effect and low surface mobility, as the film thickness increases, the overhang will grow faster than the film due to the larger arriving angle. Very soon the overhangs seal the top of the gap to form the void (also known as the key hole) between the polysilicon pattern, as shown in Fig. 10.12.

These voids can have processing gases sealed inside, and the diffusion of these gases inside the IC chip can cause problems in later processes, causing yield issues

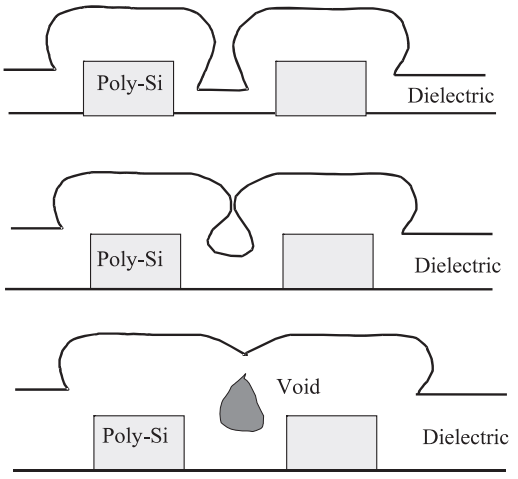


Figure 10.12 Void formation process.

or reliability issues during chip operation in an electronic system. Therefore, for most CVD processes, voids are unwanted, and void-free gap fill is required to ensure high yield and reliability of IC chips.

By reducing process pressure, the MFP of precursors increases. When the MFP is longer than gap depth h (shown in Fig. 10.9), there are very few collisions inside the gap. Thus, the precursors have little chance to go backward and reach corner A (Fig. 10.11) from inside the gap. Therefore, the MFP being longer than the gap depth effectively reduces the arriving angle and improves step coverage. This is another advantage of the LPCVD process over the APCVD process, especially for the silane-based silicon oxide CVD process, since silane has very low surface mobility.

Figure 10.13 shows that surface mobility can significantly affect step coverage. After precursors adsorb on the surface, and if the precursors have enough energy to break the adsorption bond with the surface, they can leave the surface and hop along it. If they move along the surface rapidly, migration of the precursor can smooth out the arriving angle effect. Thus, precursors with high surface mobility can achieve very good step coverage and good conformality.

Surface mobility is mainly determined by precursor chemistry, which is discussed in a later section. It is also related to wafer temperature, since heat can provide energy to the precursors that is necessary for them to break free and migrate on the wafer surface. Increasing wafer temperature can improve deposited film step coverage. For PECVD processes, ion bombardment on the wafer surface can also provide energy for the precursors to break free from the substrate surface and increase surface mobility. Therefore, increasing the rf power in a PECVD reactor can improve deposited film step coverage with an effect similar to raising the temperature.

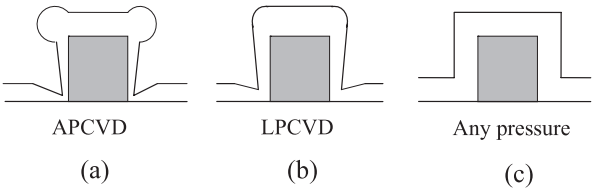


Figure 10.13 Relationship of step coverage to pressure and surface mobility: (a) high pressure, low mobility; (b) low pressure, low mobility; and (c) high mobility.

10.2.3.2 Gap fill

It is very important for CVD processes to fill gaps without voids. For example, in CMOS and DRAM ILD0 deposition processes, there is zero tolerance for any voids between the gates. This is because the voids in ILD0 can cause electrical shorts between the contact plugs, as shown in Fig. 10.14.

For tungsten CVD or copper deposition processes, voids inside contact or via plugs are also undesired; they can cause high interconnect electrical resistance. They can also induce chip reliability issues due to void migration and out-diffusion of the trapped processing gases and byproducts, such as WF_6, H_2, F, and HF, that can corrode metal wires or damage devices.

For the ILD deposition process, voids above the metal surface are intolerable. They can cause trouble for later processes, especially for the dielectric CMP process. However, if there are no contact plugs between the metal lines, voids can be tolerated when they are located inside the gap below the top metal surface. Voids can effectively reduce the dielectric constant between the metal lines, in turn reducing parasitic capacitance. Because voids always trap some processing gases and byproducts, most fabs prefer void-free gap fill (or hole fill) for CVD processes.

Question: How can voids help reduce κ-value between metal lines?

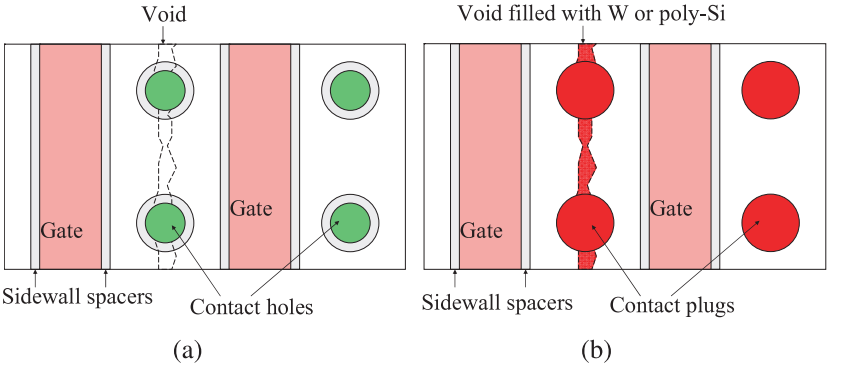


Figure 10.14 Voids that caused plug-to-plug shorts: (a) after contact hole etch and (b) after contact plug formation.

Answer: The dielectric constant (or κ-value) in a void is close to 1, which is significantly lower than the 4.0 to 4.2 of CVD silicon oxide, or 2.5 of ULK dielectric film. Therefore, a gap sealed with dielectric film with voids in the gap will have a lower effective dielectric constant than a gap filled with dielectric film without voids. The purposely formed void is called an air gap, which has the lowest achievable κ-value.

When a deposited film hangs over a pattern, it can cause voids if the film keeps growing. Different methods have been developed to alleviate this issue. One of them is argon ion sputtering etch, which chips the corner of overhangs and tapers the gap opening to increase the arriving angle, so that the subsequent deposition process can fill the gap without voids. This approach is called dep/etch/dep, short for deposition, etchback, and deposition, as illustrated in Fig. 10.15. This process is widely used for IMD processes in IC manufacturing.

If CVD precursors have very high surface mobility, the CVD film will have good step coverage and conformality, as shown in Fig. 10.13(c). The film can also grow up conformally to fill the gap without voids, as shown in Fig. 10.16. The O_3-TEOS oxide CVD and the tungsten CVD processes belong to this case.

In high-density-plasma (HDP) CVD processes, deposition and sputtering etch happen simultaneously inside the same processing chamber. Heavy ion bombardments due to low pressure (a few millitorr) and the high plasma density

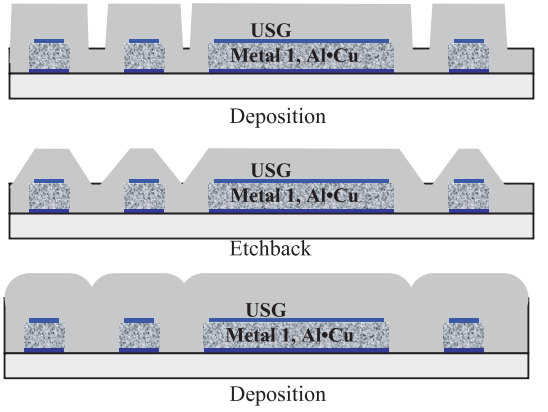


Figure 10.15 Dep/etch/dep gap-fill processes.

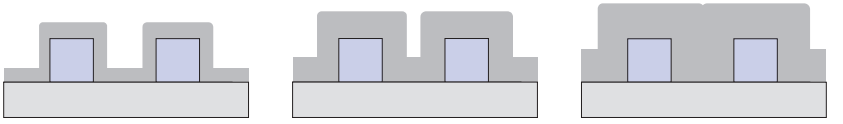


Figure 10.16 Conformal film deposition and gap fill.

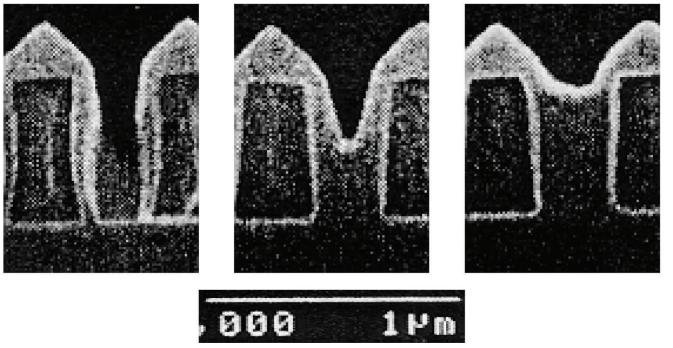


Figure 10.17 Gap fill of HDP-CVD. (Applied Materials, Incorporated).

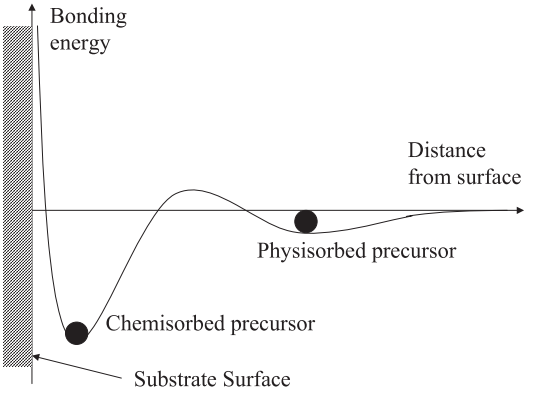


Figure 10.18 Relationship of bonding energy to chemical and physical adsorption.

constantly chip off the corner to keep the gap open and tapered, allowing a larger arriving angle and bottom-up deposition. Figure 10.17 shows a gap fill with the HDP-CVD process.

10.2.4 Surface adsorption

When precursors reach the substrate surface after diffusion across the boundary layer, they are adsorbed by the surface. There are two kinds of adsorption—chemisorption and physisorption (see Fig. 10.18).

10.2.4.1 Chemisorption

Chemisorption is the shortened term for chemical adsorption. In this case, an actual chemical bond is formed between an atom on the surface and an atom in the adsorbed precursor molecule. The chemisorbed atoms or molecules are held to

the surface with energy that exceeds 2 eV. Because of the strong chemical bonds, the chemisorbed precursors have very low surface mobility.

For most dielectric CVD processes (especially ILD and passivation dielectric processes), deposition temperature cannot be higher than 450 °C because at this temperature aluminum begins to react with silicon dioxide. Even with copper interconnections, ILD deposition temperature is still limited to around 400 °C. At 400 °C (~0.06 eV), the heat alone cannot provide enough energy to cause the chemisorbed precursors to break free from their chemical bond and leave the substrate surface. In PECVD processes, ion bombardment provides enough energy (10 to 20 eV) to break the chemical bonds and free chemisorbed precursors from the surface, causing surface migration.

10.2.4.2 Physisorption

Physisorption is the shortened term for physical adsorption. In this case, the adsorbed molecules are held to the surface with forces much weaker than a chemical bonding force. Physisorption involves energies less than 0.5 eV per molecule. The nature of forces involved in physisorption varies from long-range Van der Waals forces to dipole–dipole forces (of which hydrogen bonding is a special case).

Both thermal energy at 400 °C and ion bombardment provide enough energy to cause significant amounts of physisorbed precursors to break free and leave the surface. The physisorbed precursors can move about the surface, so they have much higher surface mobility than that of chemisorbed molecules.

10.2.5 Chemical vapor deposition precursors and their adsorption

The most commonly used silicon source gases for dielectric CVD processes are silane (SiH_4) and TEOS [$Si(OC_2H_5)_4$]. For low-κ dielectrics, trimethylsilane [$(CH_3)_3SiH$] or 3MS is a commonly used precursor. Diethoxymethylsilane ($C_5H_{14}Si$) or DEMS can be used as a precursor as well, and cyclohexene oxide ($C_6H_{10}O$) or CHO can be used as a porogen for ULK dielectric deposition.

Silane is one of the most commonly used silicon source gases for dielectric thin-film deposition. The main purpose of the silane PECVD process is passivation dielectric depositions. It is used for the ILD0 barrier nitride layer and dielectric ARC layer depositions. This process is also used in HDP-CVD oxide processes. Silane is widely used in LPCVD polysilicon and epitaxial silicon deposition processes, as well as tungsten CVD processes for the tungsten nucleation step. It is also employed as a silicon source for tungsten silicide deposition.

Silane is very reactive, pyrophoric (ignites itself), explosive, and toxic. If a silane gas line is opened without thoroughly purging it beforehand, oxygen and moisture will react with the residual silane inside the gas line. This can cause fire or a minor explosion and forms fine silicon dioxide particulates that dust the gas line. This can induce significant processing downtime and increase production cost for replacement of the dusted double-layer silane line.

Since the silane molecule is perfectly symmetrical in a tetrahedral structure (shown in Fig. 10.19), it can neither chemisorb nor physisorb to the substrate surface. However, since silane is chemically very reactive, it can easily dissociate from the heat and plasma. The molecular fragments formed by pyrolysis or plasma dissociation—SiH_3, SiH_2, or SiH—are chemically very reactive free radicals and can easily form chemical bonds with surface atoms and readily chemisorb on the substrate surface. Silane-based precursors have very low surface mobility, explaining why silane-based dielectric CVD films always have overhangs at the upper corners of patterns and normally have poor step coverage, especially for the APCVD process.

TEOS is a large organic molecule with an ethoxy group (OC_2H_5) bonded to the silicon atom at each corner of its tetrahedron [Fig. 10.20(a)]. Different from a silane molecule, the TEOS molecule is no longer perfectly symmetric due to the large size of its ethyl group. It can form hydrogen bonds with surface atoms and physisorb on the substrate surface. Therefore, the TEOS precursor has high surface mobility, and TEOS-based CVD film generally has good step coverage and

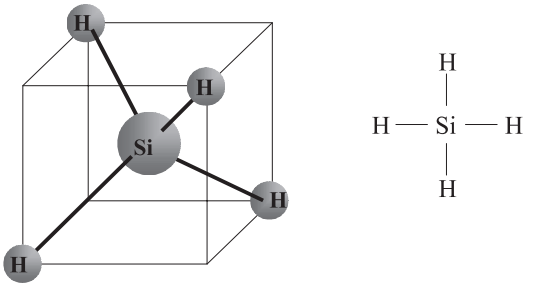


Figure 10.19 Silane molecular structure.

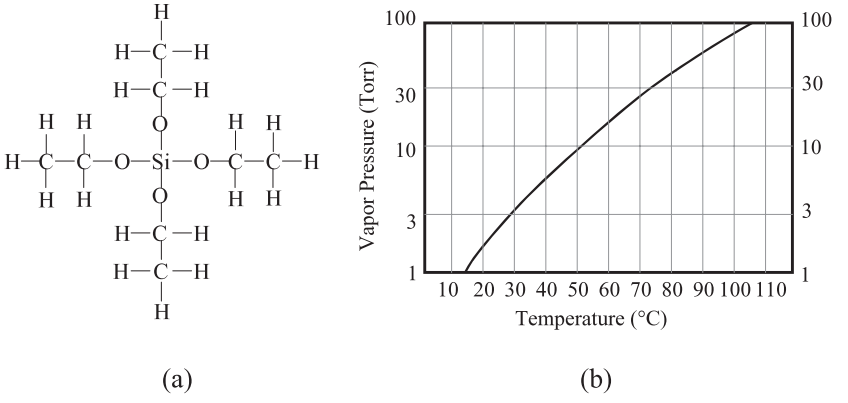


Figure 10.20 (a) TEOS molecule and (b) vapor pressure.

conformality. TEOS is widely used for oxide deposition. Its applications include STI, sidewall spacers, and ILD. Some fabs even use PE-TEOS (PECVD silicon oxide with TEOS and O_2) film for passivation oxide deposition. In IC fabs using Al–Cu interconnections, most dielectric CVD processes are TEOS-based oxide processes. For advanced IC fabs with copper low-κ interconnections, PE-TEOS film is commonly used as the capping layer for low-κ or porous low-κ dielectric films. Because PE-TEOS film is widely used in fabs, some call it TEOS for short.

TEOS is a liquid at room temperature and has a boiling point at the sea level of 168 °C. As a reference, the boiling point of water (H_2O) at sea level is 100 °C. TEOS vapor pressure is shown in Fig. 10.20(b).

To use TEOS or any liquid chemical in CVD processes, special delivery systems to vaporize the chemical and deliver the vapor into the processing reactor are required. Figure 10.21 shows three of those systems: boiler, bubbler, and injection.

All three vapor delivery systems are used in IC manufacturing for liquid chemical vapor delivery. For TEOS processes, the injection system is receiving more attention, since it can precisely and independently control the TEOS flow rate into the processing chamber.

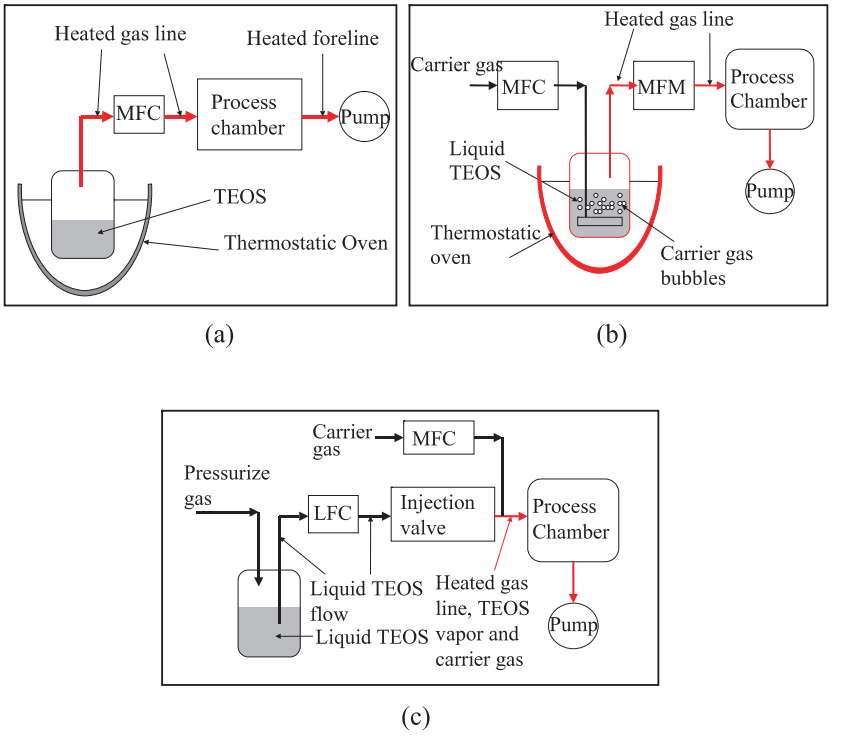


Figure 10.21 TEOS vapor delivery systems: (a) boiler, (b) bubbler, and (c) liquid injector.

3MS is a commonly used silicon precursor for low-κ dielectric film deposition. DEMS can be used as precursors, and CHO can be used as porogens. DEMS and CHO can be used to deposit porous ULK dielectric film for the ILD application of cutting-edge IC chips. Figure 10.22 shows a 3MS molecule, a DEMS molecule, and a sketch of the CHO molecule.

10.2.5.1 Sticking coefficient

The sticking coefficient is the probability that an atom or a molecule will form a chemical bond with surface atoms and chemisorb on the substrate surface in one surface collision. It can be calculated by comparing the results of the theoretically calculated deposition rate on a flat surface with 100% sticking coefficient, and the measured actual deposition rate on that surface.

The lower the sticking coefficient, the higher the surface mobility. Silane has a very low sticking coefficient due to its symmetric molecular structure. To the fragments of silane, SiH_3 is unstable, and SiH_2 and SiH have very high sticking coefficients. Both TEOS and WF_6 have very low sticking coefficients; therefore, they do not easily chemisorb on the surface and usually have to hop along the substrate surface. They have high surface mobility. This is the reason that TEOS and WF_6 CVD films always have very good step coverage and conformality.

From the SEM images in Fig. 10.23, it can be seen that the TEOS oxide film has better step coverage than the silane oxide film. That is why most silicon oxide depositions are TEOS-based processes in advanced semiconductor fabs: because they have much better gap-fill capability.

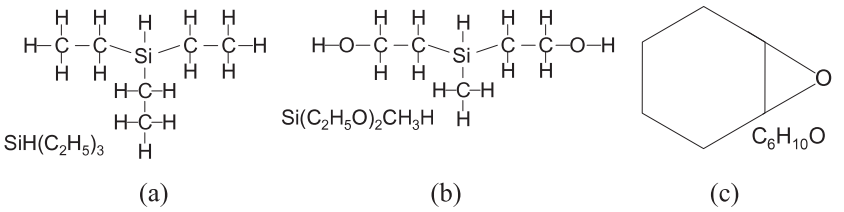


Figure 10.22 Molecular structures of (a) 3MS, (b) DEMS, and (c) CHO.

Table 10.2 Sticking coefficients for various precursors.

Precursors	Sticking coefficient
SiH_4	3×10^{-4} to 3×10^{-5}
SiH_3	0.04 to 0.08
SiH_2	0.15
SiH	0.94
TEOS	10^{-3}
WF_6	10^{-4}

Question: TEOS cannot be used as the silicon source gas for silicon nitride deposition. Why?

Answer: In the TEOS molecule, a silicon atom is bonded with four oxygen atoms. It is almost impossible to strip all of the oxygen atoms and have silicon bond only with nitrogen. Therefore, TEOS is mainly used for oxide deposition, and nitride deposition uses silane as the silicon source gas.

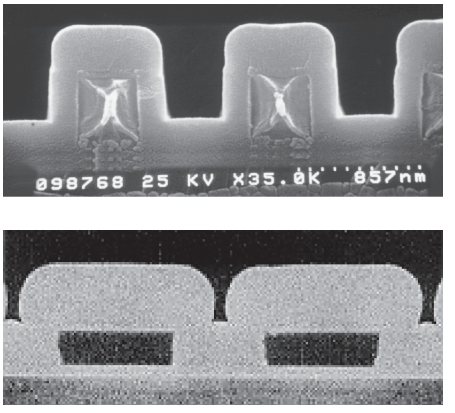


Figure 10.23 Step coverage of TEOS oxide and silane oxide (Applied Materials, Incorporated).

10.2.6 Chemical vapor deposition kinetics

10.2.6.1 Chemical reaction rate

The chemical reaction rate can be expressed as the Arrhenius equation:

$$CR = A \exp(-E_a/k_B T), \tag{10.2}$$

where A is a constant, E_a is the activation energy (as illustrated in Fig. 10.24), k_B is the Boltzmann constant, and T is the substrate temperature. Low activation energy making E_a means a low chemical reaction barrier, which makes it easier for a reaction to happen.

External energy sources such as heat, rf power, or UV radiation are needed for chemical precursors to overcome activation energy barriers and achieve a chemical reaction.

Because the chemical reaction rate is exponentially related to temperature, it is very sensitive to changes in temperature. Changing temperatures can dramatically alter chemical reaction rates. For a CVD process, the deposition rate (DR) is related to the chemical reaction rate (CR), the precursor diffusion rate (D) in the boundary layer, and the precursor adsorption rate (AR) on the substrate surface.

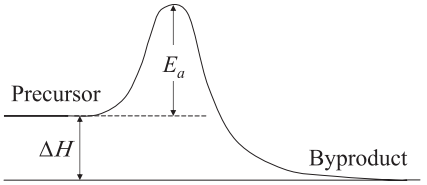


Figure 10.24 Chemical activation energy.

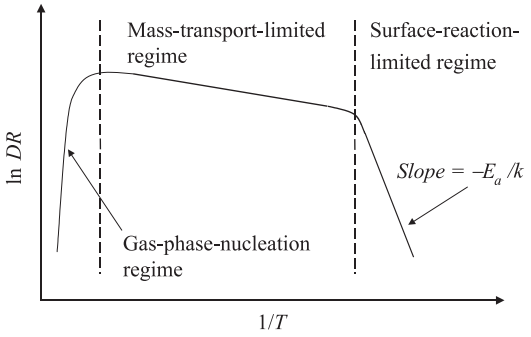


Figure 10.25 Deposition rate regimes.

From Fig. 10.25, it can be seen that the deposition rate has three regimes when the temperature changes. At low temperature, the chemical reaction rate is low, and the deposition rate is very sensitive to temperature. This is called a surface-reaction-limited regime. At higher temperature, the deposition is much less sensitive to temperature. This is the mass-transport-limited regime. When the temperature increases further, the deposition rate sharply decreases due to gas phase nucleation. This is a very undesirable deposition regime, since chemical precursors react in mid-air, generating huge amounts of particles and contaminating the wafer and reactor. Gas phase nucleation (or the homogenous nucleation regime) can be used to generate nanoparticles, but it must be avoided for all CVD processes in IC production. For silane-based dielectric PECVD processes, if the pressure is too high (>10 torr), the gas-phase-nucleation regime can begin and create particle problems.

10.2.6.2 Surface-reaction-limited regime

In a surface-reaction-limited regime, the chemical reaction rate cannot match precursor diffusion and adsorption rates; precursors pile up on the substrate surface and wait their turn to react. In this case, the deposition rate is mainly determined by the chemical reaction rate on the substrate surface:

$$DR = CR[B][C][] \ldots,$$

where CR is the chemical reaction rate defined in Eq. (10.2), and $[B]$, $[C]$, etc., are the concentrations of the adsorbed precursors.

In a surface-reaction-limited regime, deposition rate is very sensitive to temperature because it is mainly determined by the chemical reaction rate.

Some LPCVD processes, such as polysilicon and amorphous silicon deposition processes, operate in the surface-reaction-limited regime due to low deposition temperatures. It is vital to precisely control the processing temperature, since a one-percent difference in temperature can result in a five- to ten-percent difference in deposition film thickness and cause WTW uniformity problems.

10.2.6.3 Mass-transport-limited regime

When the surface chemical reaction rate is high enough, the chemical precursors react immediately when they adsorb on the substrate surface. In this case, the deposition rate is no longer determined by the surface reaction rate but by how fast chemical precursors can diffuse across the boundary layer, and reach and adsorb on the surface:

$$\text{deposition rate} = D\frac{dn}{dx}[B][C][]\ldots,$$

where D is the diffusion rate of the precursors in the boundary layer, dn/dx is the gradient of the precursor concentration in the boundary layer, and $[B][C]$, etc., are precursor concentrations on the substrate surface determined by precursor adsorption rates.

In a mass-transport-limited regime, deposition rate is not very sensitive to temperature, and deposition is mainly controlled by gas flow rates. The relationship between deposition rate and temperature is illustrated in Fig. 10.26, which also indicates temperature sensitivity of the two regimes.

10.2.6.4 Chemical vapor deposition reactor deposition regime

In a surface-reaction-limited regime, precursors wait on the substrate surface to react, and the deposition rate is mainly determined by the chemical reaction rate,

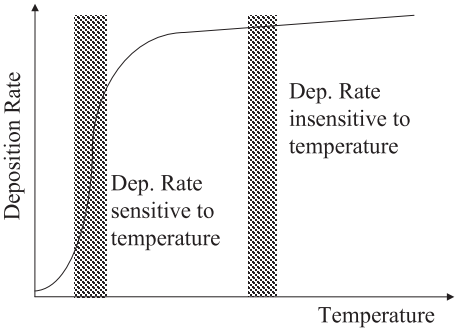


Figure 10.26 Relationship of deposition rate and temperature.

which is very sensitive to temperature. In a mass-transport-limited regime, the surface chemical reaction rate is high enough that the surface waits for chemical precursors to diffuse across the boundary layer and chemisorb on the surface. In this regime, deposition does not depend on the chemical reaction rate but is determined by the precursor diffusion and chemisorption rates.

Most single-wafer process reactors are designed to operate in mass-transport-limited regimes because it is easier to control the gas flow rate than wafer temperature. Plasma processes are commonly employed to generate chemically reactive free radicals, which can dramatically increase the chemical reaction rate and help processes achieve mass-transport-limited regimes at relatively low temperatures. Unstable chemicals such as ozone are also used to achieve high chemical reaction rates at lower temperatures and make the processes less sensitive to temperature.

10.3 Applications of Dielectric Thin Films

Commonly used dielectric thin films in the semiconductor industry are silicon-based compounds such as silicon oxide, silicon nitride, and silicon oxynitride.

Table 10.3 Properties of silicon dioxide and silicon nitride.

Oxide (SiO_2)	**Nitride** (Si_3N_4)
High dielectric strength, $> 1 \times 10^7$ V/cm	High dielectric strength, $> 1 \times 10^7$ V/cm
Lower dielectric constant, $\kappa = 3.9$	Higher dielectric constant, $\kappa = 7.0$
Not a good barrier for moisture and mobile ions (Na^+)	Good barrier for moisture and mobile ions (Na^+)
Transparent to UV	Conventional PECVD nitride opaque to UV
Can be doped with P and B	–

Both silicon oxide and silicon nitride are good electrical insulators with very high dielectric strength (breakdown voltage). Dielectric constants of CVD oxide and nitride are slightly higher than their stoichiometric values listed in Table 10.3. Silicon oxide has a lower dielectric constant than nitride. Therefore, interconnects with silicon oxide between the metal lines have a smaller parasitic capacitance and shorter RC delay time than those with silicon nitride. This is the reason that ILD mainly uses silicon oxide instead of nitride. Because nitride is a much better mobile-ion and moisture barrier layer than oxide, it is generally used for the final layer of passivation dielectric applications, and a barrier layer for doped oxide in ILD0 applications. Some devices such as those using EPROM require UV transparent passivation layers. Silicon oxynitride, which has properties somewhat in between silicon oxide and silicon nitride, is commonly used for this application. For advanced IC chips in nanometer technology, low-κ dielectrics such as PECVD OSG or SiCOH are commonly used as ILD material. To further reduce RC delay, ULK dielectrics, mainly porous SiCOH, were developed and used in ILD applications.

10.3.1 Shallow trench isolation

As IC device dimensions shrank to deep submicron levels, STI gradually replaced LOCOS as the isolation between neighboring transistors. Undoped silicate glass (USG) is used for STI trench fill. PE-TEOS, O_3-TEOS, HDP-CVD oxide, and spin-on dielectrics are used, depending on the trench geometry. Figure 10.27 shows a cross-section SEM image of a STI after CVD oxide trench fill, and Fig. 10.28 illustrates the STI formation processes.

Requirements of dielectric CVD thin films for STI applications are void-free gap fill and a dense film with little shrinkage. It uses higher deposition temperatures to help achieve better step coverage and higher film quality for STI applications.

10.3.2 Sidewall spacer

Sidewall spacers can be found in most MOS transistor-based IC chips when gate dimensions are smaller than 2 μm. They are mainly used to form the LDD or SDE to suppress the hot carrier effects. They also can provide a diffusion buffer for the dopant atoms in a S/D. A sidewall spacer is needed in self-aligned silicide formation to prevent shorting between the S/D and the gate. It is also needed for selective epitaxial growth (SEG) of SiGe for pMOS S/D and SiC for nMOS S/D. If

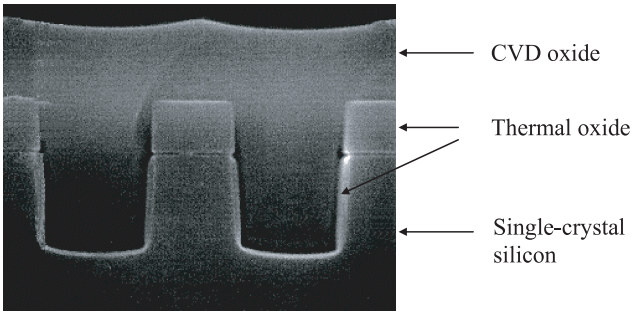


Figure 10.27 CVD oxide trench-fill application in STI process (Applied Materials, Incorporated).

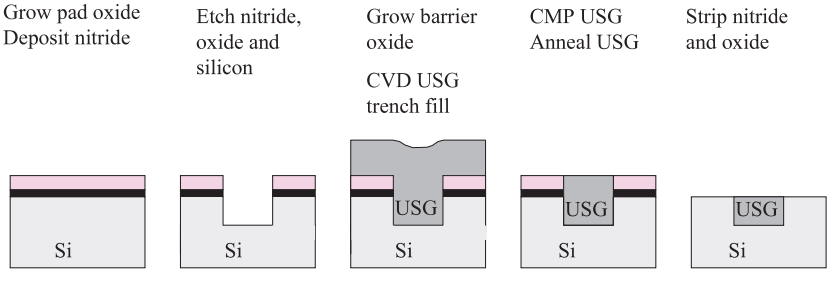


Figure 10.28 Processing steps of STI.

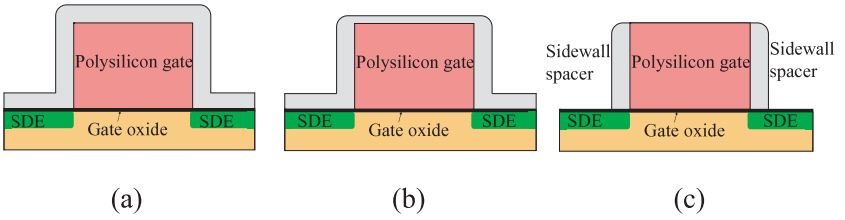


Figure 10.29 Formation of a sidewall spacer: (a) conformal dielectric film deposition, (b) dielectric etch back, and (c) spacer formation.

the spacer layer cracks, SEG film will grow on the polysilicon gate. Figure 10.29 illustrates sidewall spacer formation processes.

A sidewall spacer can use LPCVD nitride with thermal growth oxide as an etch endpoint. It can also use a LPCVD nitride liner as an etch stop to etchback TEOS-based CVD oxide film. In some devices, trilayer spacers are used.

O_3-TEOS USG film is used for oxide deposition due to its almost perfect conformality and sidewall step coverage. Requirements for spacer dielectric film are good conformality and high selectivity to polysilicon during etchback. Deposition temperature limitation for spacer dielectrics is the thermal budget, which is determined by device design.

10.3.3 Interlayer dielectric-0

ILD0 is the first dielectric layer deposited on a wafer surface after transistors have been created. The requirements of ILD0 are low dielectric constant, mobile-ion barriers, void-free gap fill, and planarization of the surface. Temperature limitations for ILD0 deposition and thermal flow are determined by the thermal budget of the device.

ILD0 is normally a phosphorus-doped silicate glass (phosphosilicate glass, PSG) or phosphorus- and boron-doped silicon oxide (borophosphosilicate glass, BPSG). To prevent phosphorus and boron from diffusing into the activation region (S/D), it is necessary to deposit a barrier layer before PSG or BPSG deposition. USG (~1000 Å) was used as a barrier layer when device geometry was still in the micron range. CVD nitride (<100 Å) is used now.

There are two important reasons to dope phosphorus with silicon oxide: gettering mobile sodium ions (Na^+) and reducing silicate glass thermal flow temperature.

Sodium is located in column IA of the periodic table, which means that it has only one electron in the outermost shell. It is very easy for sodium to lose that electron and become an ion. Sodium ions are very small and mobile. When they accumulate in the gate oxide of a MOSFET, they can change the threshold voltage of the MOSFET and cause it to turn on or off randomly. Figure 10.30 illustrates sodium ions turning on a normally off nMOS. It takes only a tiny amount of sodium to substantially damage a MOSFET and cause circuit malfunction. Therefore, it is extremely important to control mobile ion contamination. Since sodium is almost

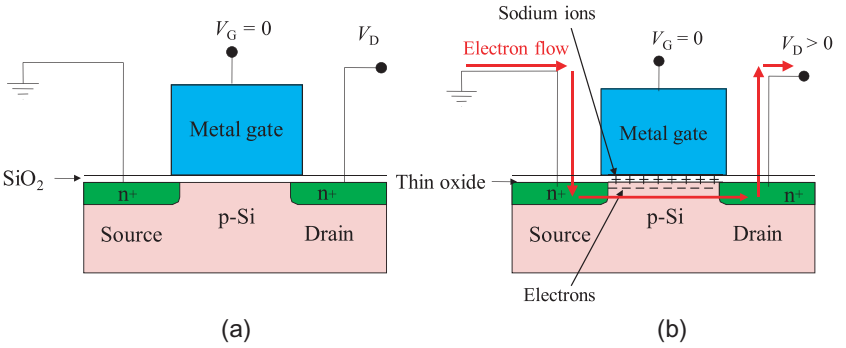


Figure 10.30 Effect of sodium ions on an nMOS: (a) normal off mode with V_G = 0 V and (b) turned on by Na^+, even at V_G = 0 V.

everywhere, and it is very difficult to completely eliminate sodium, a sodium barrier layer right above the gate of the MOSFET is required. PSG and BPSG are commonly used because they can trap sodium ions and prevent them from diffusing into gate oxide and causing device damage. Some facts about sodium are summarized in Table 10.4.

Question: Silicon nitride is a better sodium barrier layer than silicon oxide. Why isn't nitride used for ILD0?

Answer: Because silicon nitride has a higher dielectric constant than oxide, using nitride as the dielectric between the conducting lines can cause a much longer RC time delay and significantly affect circuit speed. A thin layer of nitride (to <100 Å) is used as a diffusion barrier layer in ILD0 applications to prevent diffusion of phosphorus and boron from PSG or BPSG diffusing into the S/D.

When USG is heated to >1500 °C, it becomes soft and starts to thermally flow. Since the melting point of silicon is 1414 °C, the wafer will melt before the USG starts to flow. It is a known fact from the glass industry that PSG can flow at significantly lower temperatures. As deposited, the silicate glass surface is rough and full of hills and valleys, which can cause resolution problems in photolithography processes due to DoF, and serious step coverage problems for the subsequent metal deposition process. At high temperature, glass becomes soft and viscous. It will respond to surface tension force, which creates a smoother topography on the glass surface.

As shown in Fig. 10.31, the higher the phosphorus concentration in PSG, the better the reflow results that can be achieved. The θ in the figure is the reflow angle. The smaller the reflow angle, the better the surface planarization.

If the phosphorous concentration is too high (>7 wt%), the PSG surface will be highly hygroscopic. P_2O_5 will react with moisture (H_2O) and form phosphoric acid (H_3PO_4) on the PSG surface, possibly leading to aluminum corrosion. It can also cause photoresist adhesion problems during the photolithography process

Table 10.4 Facts about sodium.

Symbol	Na
Atomic number	11
Atomic weight	22.989770
Discoverer	Sir Humphrey Davy
Place of discovery	England
Discovery date	1807
Origin of name	From the English word soda (the origin of the symbol Na comes from the Latin word natrium)
Density of solid	0.968 g/cm^3
Molar volume	23.78 cm^3
Velocity of sound in substance	3200 m/sec
Electrical resistivity	4.7 $\mu\Omega \cdot$ cm
Reflectivity	N/A
Melting point	97.72 °C
Boiling point	882.85 °C
Thermal conductivity	140 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	71×10^{-6}K^{-1}
Applications	Major contaminant, needs to be strictly controlled
Main removal agent	HCl
Barrier materials used	Silicon nitride and PSG

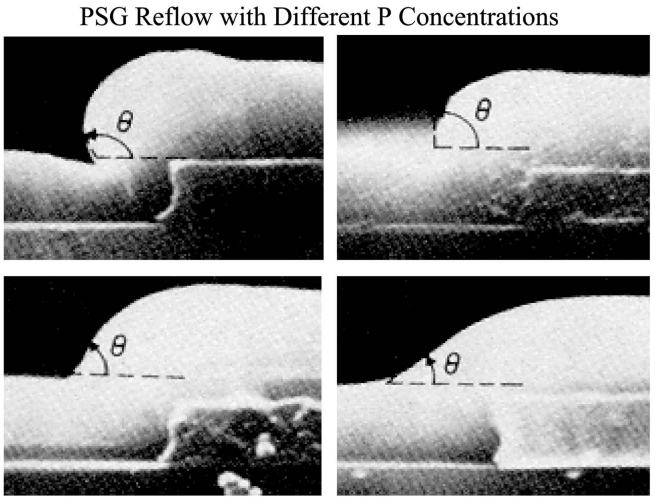


Figure 10.31 PSG reflow at 1100 °C in N_2 ambient for 20 min with different phosphorus concentrations (A.C. Adams).

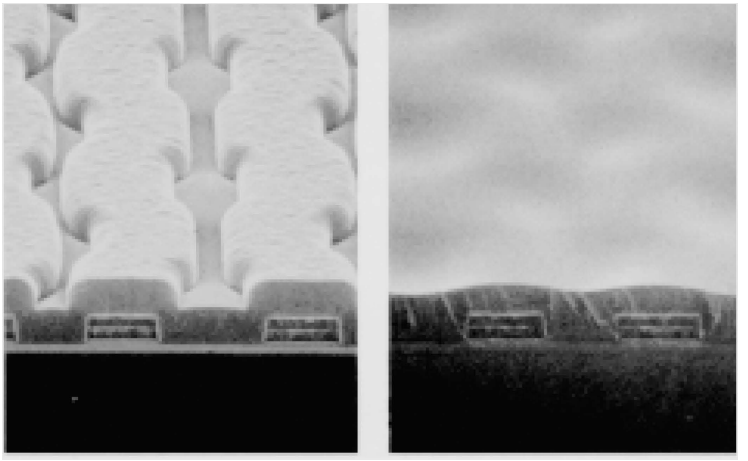
for contact hole masking because photoresist will not stick well to hygroscopic surfaces.

When device dimensions shrink, thermal budget limitations require lower reflow temperatures. Boron has also been used along with phosphorus to dope silicate glass to further reduce reflow temperatures without too much phosphorus. BPSG

can flow at 850 °C, as shown in Fig. 10.32. The first four of 4 × 4 indicates 4-wt% boron, and the second four indicates 4-wt% phosphorus. BPSG is widely used in IC chips with feature sizes from 2 to 0.25 μm. If the boron concentration in BPSG is too high, B_2O_3 can react with moisture (H_2O) and form boric acid (H_3BO_3) crystals on the BPSG surface, possibly causing defects such as particle contamination. The upper limit of BPSG dopant concentration is about 5 × 5.

As feature size continues to reduce, thermal reflow can no longer meet the planarization requirements of nanometer photolithography, and there is no more thermal budget for thermal reflow. CMP processing has been used to replace reflow for ILD0 planarization. Since no reflow is required, boron is no longer needed in the film, so PSG is used again for ILD0 application. Table 10.5 summarizes the development of ILD0 deposition and planarization processes.

LPCVD nitride had been used as a ILD0 barrier layer. Due to thermal budget concerns, PECVD nitride has been more widely used because PECVD nitride can be deposited at significantly lower temperatures. ILD0 was deposited with



(a) (b)

Figure 10.32 4 × 4 BPSG reflow at 850 °C, 30 min in N_2 ambient: (a) as deposited and (b) after reflow (Applied Materials, Inc.).

Table 10.5 Development of ILD0 processes.

Dimension	PMD	Planarization	Reflow temperature
>2 μm	PSG	Reflow	1100 °C
2 to 0.35 μm	BPSG	Reflow	850 to 900 °C
0.25 μm	BPSG	Reflow + CMP	750 °C
180 nm and beyond	PSG or USG	CMP	-

silane-based CVD processes long ago. O_3-TEOS-based CVD processes began to be commonly used for ILD0 deposition in semiconductor fabs after the 1990s.

10.3.4 Interlayer dielectric-1 and up

For advanced IC chips with multilayer interconnections, more than half of the dielectric thin-film processes are ILD processes. ILD normally uses USG, low-κ, or ULK dielectrics, depending on the technology node. Different processing sequences can be used for gap filling and planarization, depending on gap geometry and the manufacturer's preference. The requirements of ILD are a low dielectric constant, void-free gap fill, and planarization of the surface. ILD deposition temperature does not exceed 450 °C because of the presence of metal interconnections. Normally, the ILD deposition temperature is about 400 °C.

For metal line spacing >0.6 μm, deposition/sputter etchback/deposition is commonly used for gap filling. TEOS-based oxide deposition and argon sputter etch are employed for this process.

Spin-on glass (SOG) processes, including PECVD liner deposition, liquid glass spin-on, SOG cure, SOG etchback, and PECVD cap layer deposition, are used in semiconductor manufacturing. Figure 10.33 shows the application of the SOG process for ILD gap fill (SOG1) and planarization (SOG2) applications. Since the curing temperature is limited by the presence of metal lines and cannot exceed 450 °C, SOG film quality is not good and needs to be etched away as much as possible from the wafer surface, leaving only SOG in the gap for gap fill and planarization. Spin-on dielectrics (SOD) with processes similar to SOG were also intensively developed to compete with CVD-based dielectric thin films for low-κ ILD applications. Due to reliability issues, SOD lost the competition.

An ozone-TEOS-based process can deposit nearly perfect conformal oxide film with excellent gap-fill ability. This process is used to fill >0.35-μm gaps and >2:1 aspect ratio. Since ozone-TEOS oxide film has poor quality and unfavorable tensile stress, it needs a PECVD barrier layer and cap layer, similar to a SOG process. It does not require etchback, and all three layers can be deposited in the same processing chamber in one processing sequence.

As device dimensions continue to shrink, metal gaps will become smaller and smaller. At the same time, the thickness of the metal cannot be reduced accordingly. Otherwise, resistance of the metal line will increase accordingly. Void-free filling of narrow (<0.25-μm) and high-aspect-ratio (>3:1) gaps poses a great challenge to dielectric thin-film deposition processes. By doing dep/etch/dep simultaneously in the same processing chamber, the HDP-CVD process can fill gaps 0.20-μm wide with a 4:1 aspect ratio without voids. Figure 10.34 shows an SEM cross section of HDP-CVD USG filling a 0.25-μm, 4:1 aspect-ratio gap. Before SEM analysis, a diluted HF decoration etch is performed. Because HDP-CVD USG and PECVD USG have different etch rates, the boundary between the two layers shows up after the wet etch from the chip cross section.

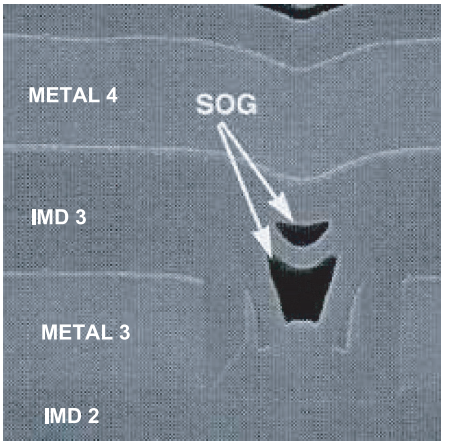


Figure 10.33 Spin-on glass for ILD gap fill (lower SOG) and planarization (upper SOG) applications (Integrated Circuit Engineering Corp.).

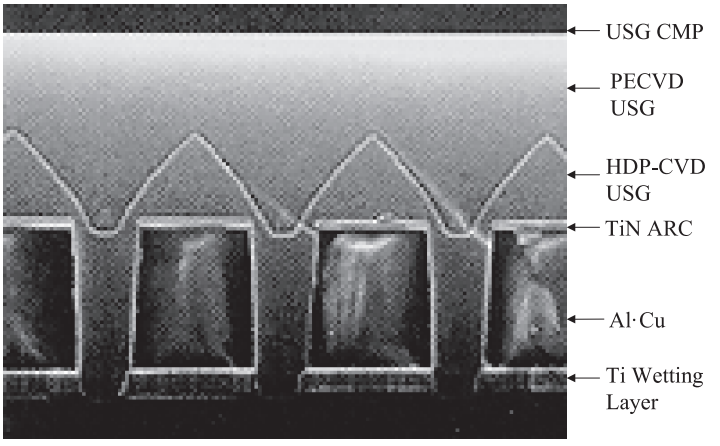


Figure 10.34 HDP-CVD, PECVD, and CMP for 0.25-μm, 4:1 aspect-ratio IMD gap fill (Applied Materials, Inc.).

Since HDP-CVD deposits and sputter etches thin film at the same time, the net deposition rate is not very high. Thus, it is mainly used for gap fill, and a PECVD film with a higher deposition rate is used to cap it. Dielectric CMP planarizes the PECVD cap layer to meet the planarization requirement for photolithography processes.

With low-κ copper interconnections, there is no metal gap to fill, and film is deposited on a flat surface after the metal CMP, thus there is no planarization issue. The requirements of dielectric thin films become simpler: high throughput (which means high deposition rate), good uniformity, and stable low-κ dielectric

properties, including dielectric constant, thermal conductivity, dielectric strength, and mechanical strength.

10.3.5 Passivation dielectrics

Most fabs use plastic packaging instead of ceramic packaging to reduce costs for the back end of IC chip processing. Plastic is not a good barrier for moisture and mobile ions. Therefore, a good barrier layer deposited at a low temperature (~400 °C) with high dielectric and mechanical strength is required for the final passivation layer of an IC chip. PECVD silicon nitride satisfies all of these requirements and is commonly used for passivation layers. Since the passivation dielectric is the last layer of a chip, the higher dielectric constant of the silicon nitride does not affect the speed of the circuits. For chips using ceramic packaging, oxide passivation is adequate, because ceramic (aluminum oxide) is a very good barrier against moisture and mobile ions.

Nitride does not stick very well to aluminum lines due to stress mismatch. Therefore, an oxide layer is always deposited prior to nitride deposition to buffer the stress and help silicon nitride adhesion. In-situ PECVD silane oxide and nitride processes are commonly used for passivation dielectric depositions. The deposition temperature of passivation dielectrics is also limited by the presence of metal interconnection wires; therefore, 400 °C is the most commonly used deposition temperature.

10.4 Dielectric thin-film characteristics

In this section, dielectric thin-film characteristics, including refractive index, thickness, and stress and metrology tools for their measurement, are discussed.

10.4.1 Refractive index

The definition of refractive index is

$$\text{refractive index}, n = \frac{\text{speed of light in vacuum}}{\text{speed of light in film}}. \tag{10.3}$$

For SiO_2, $n = 1.46$, and for Si_3N_4, $n = 2.01$. Refractive index is related to the wavelength of the light used for measurement. The refractive indexes mentioned in this chapter are measured at 633 nm, the wavelength of red light emitted from a He-Ne laser, which is commonly used in laser pointers. A prism separates sunlight into a colorful spectrum. This occurs because the refractive index of the prism material—quartz—is a function of the wavelength of the light. Light with different wavelengths has different refractive angles entering and leaving the prism, giving the light a colorful spectrum.

The refractive index and refractive angle illustrated in Fig. 10.35 can be described by the law of refraction:

$$n_1 \sin \theta_1 = n_2 \sin \theta_2,$$

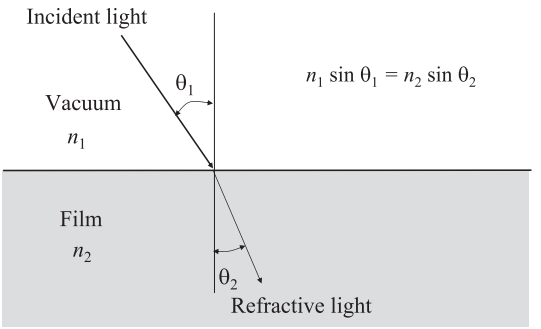


Figure 10.35 Refractive index and refractive angle.

where n_1 is the refractive index of the first dielectric material, usually air, with a dielectric index very close to 1. θ_1 is the incident angle, n_2 is the refractive index of the second dielectric material, and θ_2 is the refractive angle. This equation can be used to measure dielectric index by shining a laser beam into the dielectric material and measuring the refractive angle. However, it is impossible to measure the refractive index of a dielectric film by applying this equation when film thickness is too thin.

For silicon compound dielectric thin films, refractive index measurements can provide useful information on the chemical composition and physical conditions of the film. For a silicon- or nitrogen-rich oxide, the refractive index will be higher than 1.46 of the SiO_2, and it will be lower than 1.46 when it is oxygen rich. For nitride, silicon-rich film will have a higher refractive index than 2.01, and nitrogen- or oxygen-rich film will have a lower value, as shown in Fig. 10.36.

Question: If a process reactor is leaking during the silicon nitride deposition process, how will the nitride refractive index change?

Answer: The chamber leak introduces oxygen from the atmosphere into the processing chamber. Oxygen is fairly reactive and will react with silicon precursor gases, becoming integrated into the deposited film. This will create oxygen-rich nitride or even oxynitride. From Fig. 10.36, a decrease in refractive index can be predicted.

In dielectric thin-film thickness measurements, refractive index and film thickness are always considered together. Therefore, it is necessary to know the refractive index first before measuring film thickness. Ellipsometry and prism couplers are two of the most common metrology tools used to measure refractive index in semiconductor processing fabs.

10.4.1.1 Ellipsometry

The polarization state of light changes when a light beam is reflected from a film surface, as illustrated in Fig. 10.37. By monitoring this change, information can be attained about the refractive index and thickness of dielectric thin films. The

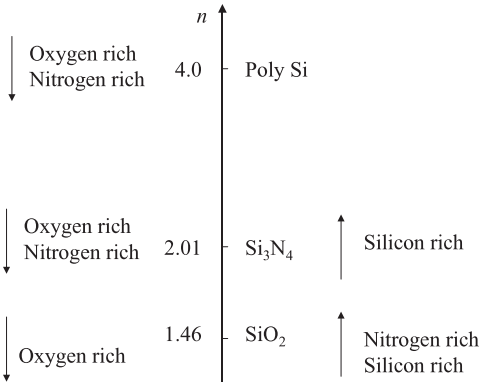


Figure 10.36 Refractive index and its importance for dielectric thin films.

change in polarization of the **p** and **s** components of a light beam during reflection can be determined. The fundamental equation of ellipsometry can be expressed as

$$\rho = r_p/r_s = \tan\Psi e^{I\Delta},$$

where ρ is a complex amplitude reflection ratio, and r_p and r_s are the Fresnel reflection coefficients. Both thickness and refractive index can be calculated from the parameters Ψ and Δ. An approximate value of film thickness is needed in advance because the measured ellipsometric quantities are the periodic functions of the thickness. Ellipsometry can also be used to measure dielectric thin-film thickness if its refractive index is known.

Question: Is a 0.98 measurement result possible for a thin-film refractive index?

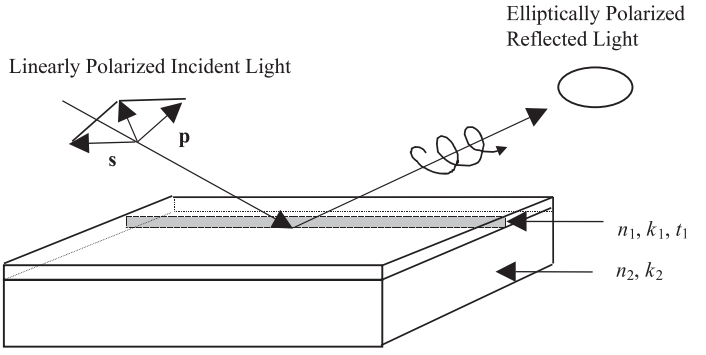


Figure 10.37 Ellipsometry system.

Answer: Based on the definition in Eq. (10.3), a refractive index cannot be less than 1, which is the refractive index of vacuum. The measurement must be incorrect.

10.4.1.2 Prism coupler

In a prism coupler, a wafer with a dielectric thin film is brought into contact with the base of a prism by a pneumatically operated coupling head. There is a small air gap between the film and prism base. A laser beam shining through the prism is totally reflected at the prism base and deflected to a photodetector, as shown in Fig. 10.38.

At certain discrete values of incident angle θ (called mode angles), photons can tunnel through the small air gap into the dielectric thin film and enter into a guided optical propagation mode. The loss of photons at the mode angles causes sudden light intensity drops on the photodetector, as illustrated in Fig. 10.39.

The first mode angle approximately determines the thin-film refractive index, and the difference between the mode angles determines film thickness, allowing an independent measurement of the dielectric thin-film refractive index and thickness.

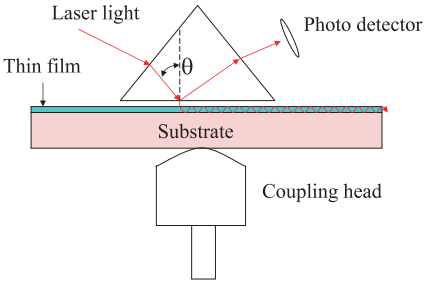


Figure 10.38 Prism coupler.

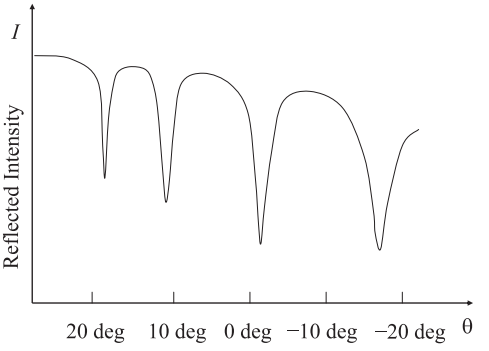


Figure 10.39 Intensity (I) of reflected light versus incidence angle (θ).

Prism coupler measurement does not require advanced knowledge of film thickness. However, it does require a minimum film thickness (typically 3000 to 4800 Å) to support at least two modes to achieve good results. The thicker the film, the more angular modes it can support, and the better the measurement results are. Prism couplers can also be used to measure refractive indices and thicknesses of stacked films.

10.4.2 Thickness

Thickness measurement is one of the most important measurements for dielectric thin-film processes. Film deposition rate, wet etch rate, and shrinkage all depend on thin-film thickness measurements.

10.4.2.1 Color chart

After deposition of the dielectric thin film, a wafer will have different colors on its surface, depending on film thickness, refractive index, and the angle of light.

Reflected light from a dielectric thin-film surface (light 1), and the reflected light from the interface of a dielectric thin film and the substrate (light 2), have the same frequencies but different phases, as shown in Fig. 10.40. There is interference between the two reflected light rays that can cause both constructive and destructive interference at different wavelengths, since the refractive index is a function of wavelength. Table 10.6 lists the color chart of oxides with different thicknesses.

The color seen on the wafer is determined by constructive interference frequency, which relates to the phase difference of the two reflected lights. The thicker the film, the larger the phase shift:

$$\Delta\phi = 2tn(\lambda)/\cos\theta, \tag{10.4}$$

where t is the thin-film thickness, $n(\lambda)$ is the thin-film refractive index, and θ is the incident angle. When the phase shift $\Delta\phi$ is larger than 2π, the color pattern repeats itself.

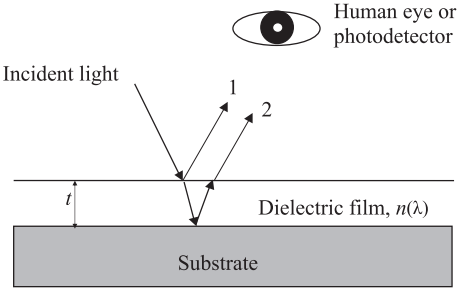


Figure 10.40 Reflection light rays and phase difference.

Table 10.6 Color chart of silicon oxide.

Thickness (Å)	Color	Thickness (μm)	Color
500	Tan	1.0	Carnation pink
700	Brown	1.02	Violet-red
1000	Dark violet to red-violet	1.05	Red-violet
1200	Royal blue	1.06	Violet
1500	Light blue to metallic blue	1.07	Blue-violet
1700	Metallic to very light yellow-green	1.10	Green
2000	Light gold or yellow slightly metallic	1.11	Yellow-green
2200	Gold with slight yellow-orange	1.12	Green
2500	Orange to melon	1.18	Violet
2700	Red-violet	1.19	Red-violet
3000	Blue to violet-blue	1.21	Violet-red
3100	Blue	1.24	Carnation pink to salmon
3200	Blue to blue-green	1.25	Orange
3400	Light green	1.28	Yellowish
3500	Green to yellow-green	1.32	Sky blue to green-blue
3600	Yellow-green	1.40	Orange
3700	Green-yellow	1.45	Violet
3900	Yellow	1.46	Blue-violet
4100	Light orange	1.50	Blue
4200	Carnation pink	1.54	Dull yellow-green
4400	Violet-red		
4600	Red-violet		
4700	Violet		
4800	Blue-violet		
4900	Blue		
5000	Blue-green		
5200	Green		
5400	Yellow-green		
5600	Green-yellow		
5700	Yellow to yellowish (at times appears light gray or metallic)		
5800	Light orange or yellow to pink		
6000	Carnation pink		
6300	Violet-red		
6800	Bluish (appears between violet-red and blue-green, overall looks grayish)		
7200	Blue-green to green		
7700	Yellowish		
8000	Orange		
8200	Salmon		
8500	Dull, light red-violet		
8600	Violet		
8700	Blue-violet		
8900	Blue		
9200	Blue-green		
9500	Dull yellow-green		
9700	Yellow to yellowish		
9900	Orange		

Using a color chart is a convenient way to measure film thickness. Although it is no longer used for thickness measurements in advanced IC fabs, it is still a useful tool for detecting obvious nonuniformity problems of deposited film.

Question: If some beautiful color rings appear on a wafer with a CVD dielectric layer, what is the conclusion?

Answer: Color change indicates dielectric thin-film thickness change; thus, the film with the color rings must have a problem with thickness uniformity, which is most likely caused by a nonuniform thin-film deposition process.

Question: Why does the color of a thin film change when seen at a different angle on a wafer?

Answer: As is seen from Fig. 10.29, when looking at a wafer from a different angle, incident angle θ changes. Equation (10.4) shows that the phase shift also changes; thus, the wavelength for constructive interference changes, which causes color change. It is important to hold a wafer straight when using a color chart to measure thin-film thickness. Tilting the wafer can cause the film thickness to seem thicker than it actually is.

10.4.2.2 Spectroreflectometry

Spectroreflectometry measures the reflected light intensity at different wavelengths. Thin-film thicknesses can be calculated from the relation of the reflected light intensity and wavelength of the light, as shown in Fig. 10.41. A photodetector is much more sensitive than human eyes in detecting the spectrum of intensity and wavelength; therefore, spectroreflectometry can obtain much higher resolution and accuracy for thickness measurements.

Thickness can be calculated by the following equation:

$$\frac{1}{\lambda_m} - \frac{1}{\lambda_{m+1}} = \frac{1}{2nt}. \tag{10.5}$$

Here, λ_m and λ_{m+1} are the m'th and $(m + 1)$'th constructive interference wavelengths, respectively; n is the film refractive index; and t is the film thickness. From Eq. (10.5) it can be seen that using the wrong refractive index can cause an incorrect thickness measurement result.

Figure 10.42 shows a schematic of a spectroreflectometry system. UV light is used for the measurement because, for dielectric thin films, the refractive index is insensitive to wavelengths in the UV range.

Question: Many advanced thin-film thickness measurement tools allow users to choose the refractive index of a film. If the PE-TEOS USG film refractive index is chosen to measure O_3-TEOS USG film thickness, what will be the effect on the measurement result?

Answer: Since nt is always coupled together, as shown in Eq. (10.5), inputting an incorrect refractive index n will cause the wrong thickness measurement result. Because O_3-TEOS USG is a porous film with many tiny holes ($n_{hole} \sim 1$) in it, a O_3-TEOS USG film normally has a lower refractive index, about 1.44, which is slightly lower than the 1.46 of PE-TEOS USG. Therefore, the measured O_3-TEOS film thickness will be slightly thinner than its actual value.

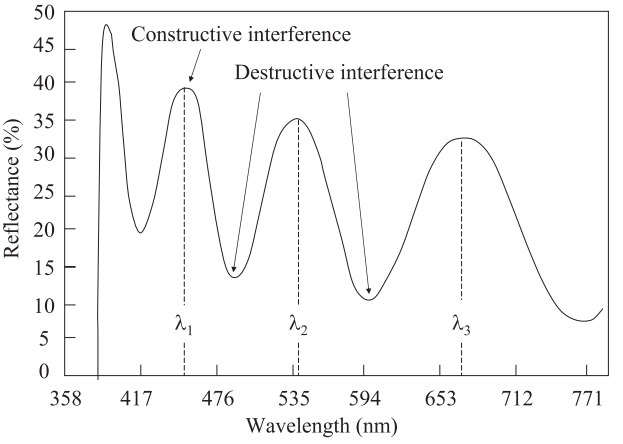


Figure 10.41 Relationship of reflectance and wavelength.

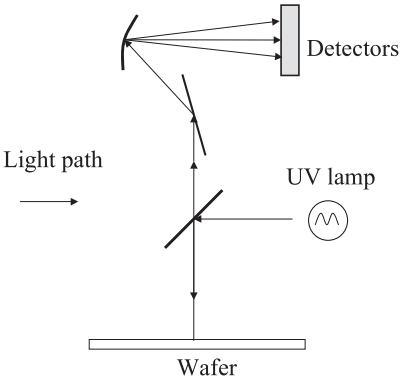


Figure 10.42 Schematic of a spectroreflectometry system.

10.4.2.3 Deposition rate

Deposition rate is defined as

$$\text{deposition rate} = \frac{\text{deposited film thickness}}{\text{deposition time}}.$$

Deposition rate is one of the most important characteristics of any deposition process. It affects process throughput, while directly relating to cost of ownership. The most commonly used unit for deposition rate is Å/min. The higher the deposition rate, the higher the throughput, and the lower the cost of ownership. For example, assume a fab has 20 CVD systems, and chamber cleaning and other overhead take 50% of the total deposition process. If there were a way to double

the deposition rate while keeping all of the other factors unchanged, the fab could reduce down to five CVD systems.

Question: An LPCVD silane process deposits 10,000-Å oxide film in 50 min. What is the deposition rate?

Answer: Deposition rate = 10,000/50 = 200 Å/min.

10.4.2.4 Wet etch rate

Wet etch rate and wet etch rate ratio (WERR) are commonly used for dielectric thin-film quality control. A film with better quality has a lower WERR.

Hydrofluoric acid (HF) can etch silicon oxide and silicon nitride. Since 1:1 HF (49% HF in H_2O) etches oxide too quickly to control, HF is diluted to 6:1 and buffered with ammonium fluoride (NH_4F) to slow down the wet etch rate. This procedure is called a 6:1 BOE, and it is commonly used to qualify dielectric thin-film processes. Another popular oxide wet etch uses 100:1 HF solution.

The definition of wet etch rate is:

$$\text{wet etch rate} = \frac{\text{thickness before wet etch} - \text{thickness after wet etch}}{\text{wet etch time}}.$$

The wet etch rate of 6:1 BOE at 22 °C for thermally grown silicon dioxide is about 1000 Å/min. CVD oxide has a higher wet etch rate because it is not as dense. The wet etch rate is very sensitive to HF solution concentration and temperature. To eliminate measurement uncertainties caused by the variations of temperature and concentration, a wafer with a thermally grown silicon dioxide film is etched alongside wafers with CVD thin films as the reference. This technique is called the WERR to thermal oxide:

$$\text{WERR} = \frac{\text{thickness change of the CVD film}}{\text{thickness change of the thermal oxide film}}. \tag{10.6}$$

The WERR is preferred over the wet etch rate itself because using the ratio has eliminated the major measurement uncertainties associated with exact concentration and temperature.

Question: A O_3-TEOS CVD oxide film thickness changed from 4500 Å before 6:1 BOE, to 2400 Å after it, while the reference thermal oxide thickness changed from 2000 to 1550 Å. What is the WERR?

Answer: WERR = (4500 – 2400)/(2000 – 1550) = 2100/450 = 4.67.

10.4.2.5 Shrinkage

Heating up a wafer and cooling it back to room temperature is called a thermal cycle. During a thermal cycle, atoms inside a thin film have some thermal

movement, which finishes at a lower potential position and causes the film to be denser. Therefore, after a thermal cycle, film thickness decreases. The amount of shrinkage indicates the original film quality. Film with better quality has less shrinkage:

$$\text{shrinkage} = \frac{\text{thickness change after thermal cycle}}{\text{thickness before thermal cycle}}. \tag{10.7}$$

Shrinkage is one of the major concerns for oxide films in STI applications because the trench fill oxide requires high-temperature annealing. Too much oxide shrinkage can cause stress-induced defects in silicon substrates, especially at trench corners.

Question: After 1100 °C anneal, HDP-CVD USG film thickness changes from 3500 to 3460 Å. What is the shrinkage of the film?
Answer: Shrinkage = (3500 – 3460)/3000 = 40/3500 = 1.14%.

10.4.2.6 Uniformity

Uniformity, or to be precise, nonuniformity, is a very important factor of the thin-film deposition because it can affect product yield. If a CVD tool has a uniformity issue and deposits a film with a thickness variation larger than the control limit, it can cause underetch where the film is too thick, and overetch where the film is too thin. Uniformity can be calculated by multipoint thickness measurements. A clear definition of uniformity is important, since different definitions give different results on uniformity, even with the same sets of measurement data.

Measurement results are $x_1, x_2, x_3, x_4, \ldots, x_N$, where N equals the total number of data points. In an IC fab, $N = 9$, 25, 49, or 121 are typical sample numbers. $N = 49$ is used for 200-mm wafers and $N = 121$ for 300-mm wafer. The mean value (or average value) of the measurement is

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}.$$

The standard deviation of the measurement is

$$\sigma = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N - 1}}.$$

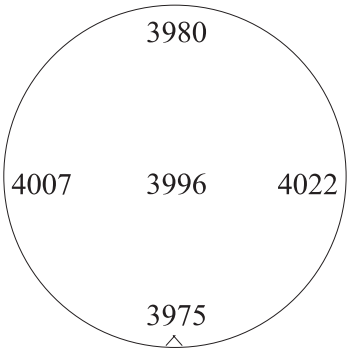
The standard deviation nonuniformity (in percentages) is defined as

$$NU(\%) = (\sigma/\bar{x}) \times 100. \tag{10.8}$$

The definition of maximum-minus-minimum nonuniformity (in percentages) is:

$$NU_{\max-\min} = [(x_{\max} - x_{\min})/(2\bar{x})] \times 100 \tag{10.9}$$

Question: Calculate the NU and $NU_{\max-\min}$ for a five-point thickness (in Å) measurement.



Answer: The average thickness is: $\bar{x} = 3996$ Å, and the standard deviation is $\sigma = 18.8$ Å.

The standard deviation nonuniformity is:

$$NU = 18.8/3966 = 0.47\%,$$

and the maximum-minus-minimum nonuniformity is

$$NU_{\max-\min} = (4022 - 3975)/(2 \times 3996) = 0.59\%.$$

10.4.3 Stress

This section defines stress and details the cause and nature of stress in dielectric films. Stress arises due to a mismatch between different materials, such as between the substrate (wafer) and the film (dielectrics, metal, etc.). There are two kinds of stresses, intrinsic and extrinsic. Intrinsic stress develops during the film nucleation and growth process. Extrinsic stress results from differences in the coefficients of thermal expansion between the film and substrate. Stress can be either tensile (noted as positive) or compressive (noted as negative), as shown in Fig. 10.43. High stress on a dielectric film, whether it is compressive or tensile, can cause film cracking, metal line spiking, or void formation. In some cases, extremely high tensile stress can even break the wafer.

For PECVD dielectric thin films, intrinsic film stress can be controlled by rf power. Film stress is kept at a compressive 10^9 dyne/cm^2 (-100 MPa). Ion bombardment from the plasma can pack molecules densely inside the film, thus increasing film density and compressive stress. It is important to control dielectric thin-film stress for dielectric CVD processes because this stress can affect defect densities of both dielectrics and metals, and it can affect any dielectric CMP processes that follow.

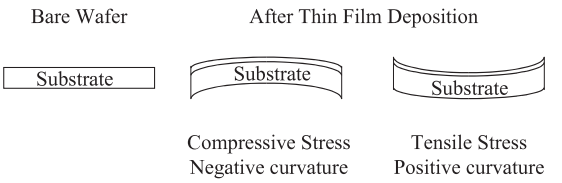


Figure 10.43 Illustration of compressive and tensile stress.

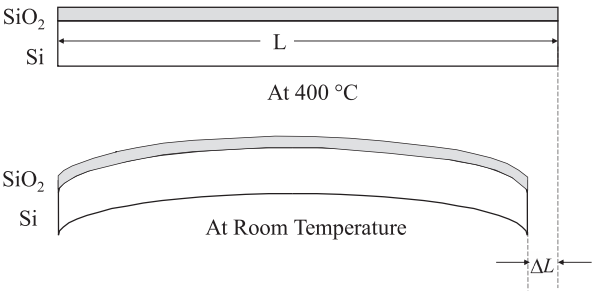


Figure 10.44 Illustration of thermal stress.

Figure 10.44 shows thermal stress on an oxide film due to the different thermal expansion rates between the oxide and silicon substrate. The thermal expansion of a material can be expressed as

$$\Delta L = \alpha \Delta T L \tag{10.10}$$

where ΔL is the change of dimension, ΔT is the change of temperature, and α is the coefficient of thermal expansion. Table 10.7 lists the coefficients of thermal expansion of some commonly used materials in IC processing. All of them are measured at 300 K (or 22.85 °C), which is close to room temperature.

Assume that there is no stress between an oxide film and a silicon substrate right after oxide film deposition at 400 °C. When the wafer has cooled down to room temperature, the silicon shrinks more than the oxide, and the substrate compresses the film. Due to a higher thermal expansion rate of the silicon, the wafer curvature becomes more concave, as shown in Fig. 10.43. Because silicon contracts more than oxide, oxide film has a compressive stress.

Question: Why does an oxide film prefer compressive stress, while metal films prefer tensile stress?

Answer: If an oxide film has tensile stress at room temperature, when the wafer is heated up for the next process, the silicon substrate expands more and causes the film stress to become more tensile. This can crack the oxide film. If an oxide film stress starts as compressive, when wafer temperature increases, the film stress becomes less compressive.

Table 10.7 Coefficients of thermal expansion (in 10^{-6} °C^{-1}).

$\alpha(SiO_2) = 0.5 \times 10^{-6}$ °C^{-1}	$\alpha(W) = 4.5 \times 10^{-6}$ °C^{-1}
$\alpha(Si) = 2.5 \times 10^{-6}$ °C^{-1}	$\alpha(Al) = 23.2 \times 10^{-6}$ °C^{-1}
$\alpha(Si_3N_4) = 2.8 \times 10^{-6}$ °C^{-1}	$\alpha(Cu) = 17 \times 10^{-6}$ °C^{-1}

Therefore, oxide films always favor compressive stress. To the contrary, metal films such as tungsten and aluminum favor tensile stress at room temperature because they have higher thermal expansion rates than silicon and silicon oxide, and expand more when temperature increases, which can make the film stress less tensile.

Stress is calculated by measuring the change of the wafer curvature before and after thin-film deposition. The most commonly used unit for film stress is MPa = 10^6 Pa (1 MPa = 10^7 dynes/cm^2):

$$\sigma = \frac{E}{1-\nu}\frac{h^2}{6t}\left(\frac{1}{R_2} - \frac{1}{R_1}\right) \tag{10.11}$$

where σ is the film stress (Pa), E is Young's modulus of the substrate (Pa), ν is Poisson's ratio of the substrate, h is the substrate thickness (μm), and t is the film thickness (μm). R_1 is the wafer curvature radius before deposition (μm), and R_2 is the wafer radius curvature after deposition (μm). Figure 10.45 shows a schematic of a stress measurement system.

Question: When using recycled dummy wafers to test thin-film deposition processes, what factors need to be taken into consideration?

Answer: Recycled wafers normally require a polishing process in which their thickness is reduced and becomes thinner than the standard thickness (725 μm for the 200-mm wafer, and 775 μm for the 300-mm wafer). Deposition rate, nonuniformity, and WERR measurements are not

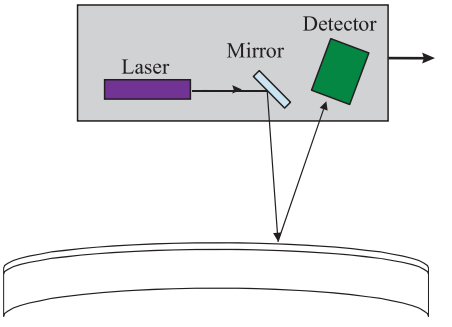


Figure 10.45 Laser scanning stress measurement tool.

affected by using recycled wafers. It is evident from Eq. (10.11) that the stress measurement is very sensitive to wafer thickness h. Therefore, it is necessary to adjust the system constant regarding the wafer thickness to achieve an accurate reading. It is very important to change that constant back to the standard value, since other people using the same stress gauge may not be aware of the change and acquire false readings on their measurements when using a normal wafer, possibly causing unnecessary process downtimes.

10.5 Dielectric Chemical Vapor Deposition Processes

Two processes are widely used for dielectric thin-film deposition: thermal CVD and PECVD. In thermal processes, heat alone provides the energy needed for chemical precursors to overcome the activation energy barrier and achieve chemical reactions. Both APCVD and LPCVD are thermal CVD processes. In PECVD processes, both heat and rf provide the energy required for chemical reactions.

The commonly used silicon precursors are silane, TEOS, and 3MS. Silane is used to deposit both nitride and oxide; TEOS is used to deposit oxide, either doped or undoped; and 3MS is used to deposit low-κ dielectric SiCOH and porous SiCOH. 3MS can also be used to deposit low-κ barrier SiCN layers.

10.5.1 Thermal silane chemical vapor definition process

Silane is one of the most commonly used silicon source gases for dielectric CVD processes. It has been used for silicon dioxide deposition with both the APCVD and LPCVD process:

$$SiH_4 + 2O_2 \underset{\text{heat}}{\rightarrow} SiO_2 + 2H_2O.$$

APCVD uses diluted silane (3% in nitrogen), and LPCVD uses pure silane. Compared with APCVD silane oxide, LPCVD silane oxide has higher throughput due to larger wafer load and better step coverage from the longer MFP of the precursors. Both APCVD and LPCVD silane oxide processes have been replaced by the TEOS-based oxide process because of the latter's superior gap-filling abilities.

LPCVD silane nitride is commonly employed as an oxidation mask for LOCOS formation. It is also used as an oxide CMP hard stop for STI processes. When thermal budget allows, fabs use LPCVD nitride as the PMD barrier layer to prevent dopants from PSG or BPSG diffusing into the activation areas:

$$3\ SiH_4 + 4\ NH_3 \underset{\text{heat}}{\rightarrow} Si_3N_4 + 12\ H_2.$$

The LPCVD nitride process is covered in detail in Chapter 5.

10.5.2 Thermal tetraethoxysilane chemical vapor deposition process

TEOS oxide film has better step coverage and conformality due to the high surface mobility of the physisorption of a TEOS molecule on a substrate surface. At high temperatures, TEOS will dissociate and form silicon dioxide:

$$\mathrm{Si(OC_2H_5)_4} \underset{\text{heat}}{\rightarrow} \mathrm{SiO_2} + \text{volatile organics.}$$

This thermal TEOS oxide deposits at temperatures higher than 700 °C in a LPCVD reactor. Deposition rate is very low, about 20 Å/min. This process has been used to deposit TEOS for sidewall spacers. By using TEOS along with TMB [trimethylborate, $\mathrm{B(OCH_3)_3}$] and TMP [trimethylphosphite, $\mathrm{P(OCH_3)_3}$], the reaction was used to deposit BPSG for PMD applications. This reaction temperature is too high to deposit USG for ILD applications.

10.5.3 Plasma-enhanced chemical vapor deposition silane process

For PECVD silane oxide, most fabs use nitrous oxide ($\mathrm{N_2O}$, laughing gas) instead of $\mathrm{O_2}$ as an oxygen source gas because it is compatible with silane at high pressure. By flowing silane and $\mathrm{N_2O}$ into a vacuum chamber (less than 10 torr) and striking plasma with rf power, the gases react and deposit silicon oxide on the heated wafer surface. Both the heat from the wafer surface and the energetic electrons inside the rf plasma provide the necessary energy for chemical reactions to occur. PECVD oxide is not a pure silicon dioxide $\mathrm{SiO_2}$; it always has a small amount of hydrogen in the deposited film:

$$\mathrm{SiH_4 + N_2O} \overset{\text{plasma}}{\underset{\text{heat}}{\rightarrow}} \mathrm{SiO_xH_y + H_2O + N_2 + NH_3} + \cdots.$$

In plasma, electrons dissociate both silane and nitrous oxide via dissociation collisions:

$$\mathrm{e^- + SiH_4 \rightarrow e^- + SiH_2 + H_2},$$
$$\mathrm{e^- + N_2O \rightarrow e^- + O + N_2}.$$

Both O and $\mathrm{SiH_2}$ are free radicals with two unpaired electrons and are chemically very reactive. They are easily chemisorbed on a wafer surface and can chemically react. Plasma-generated free radicals take the deposition process into the mass-transport-limited regime at a relatively low temperature. Deposition rate is high and mainly depends on the gas flow rate. Normally, $\mathrm{N_2O}$ is overflowed; therefore, the deposition rate is controlled by the silane flow rate. The PECVD silane oxide process has only two precursor gases—silane as the silicon source and $\mathrm{N_2O}$ as the oxygen source—so it is easier to explain the deposition process.

Question: Can silane be overflowed and the nitrous oxide flow rate be used to control the deposition rate?

Answer: Theoretically it can be done, but practically no one should try this because it is very dangerous and not cost effective. Overflowing silane will create a big safety hazard of fire and explosions, and silane is more expensive than nitrous oxide.

The PECVD silane oxide process has three main steps. They are stabilization (~5 sec), deposition (depending on the required film thickness), and chamber pump down. In the stabilization step, both pressure and gas flow rate are stabilized so that they do not change during the deposition process. During the deposition step, rf power is turned on to strike the plasma and start deposition. After deposition is finished, both rf power and process gases are turned off, which stops deposition. The chamber is pumped down to prepare for the next process. Sometimes a plasma purge step with rf power and N_2O flow is added to consume all of the residue silane before chamber pump down. Table 10.8 gives a sample process recipe for PECVD silane oxide (step 2 of the processing sequence).

In the table, sccm stands for standard cubic centimeter per minute. The expected deposition rate is 1 μm/min (10,000 Å/min), refractive index is 1.46, and film stress is compressive 100 MPa (10E9 dyne/cm^2).

Question: Increasing the silane flow rate alone increases the deposition rate. How do the refractive index and film stress change in this case?

Answer: By increasing silane flow, the SiH_4 to N_2O ratio increases, which causes more silicon to be deposited into the film. Figure 10.36 shows that the refractive index increases. While the deposition rate increases, rf power and pressure are kept constant. This counts as the same amount of ion bombardment on a thicker film, meaning that the film will receive less ion bombardment per unit thickness. It will make the film less dense and cause less compressive film stress.

Therefore, to increase deposition rate while keeping the same stress and refractive index, silane flow rate, N_2O flow rate, and rf power must increase at the same time.

10.5.3.1 Passivation

Silicon nitride is a very good barrier layer for moisture and mobile ions and is widely employed as the final passivation layer. PECVD nitride is used in this process because passivation dielectric deposition requires a relatively low deposition temperature (<450 °C) at a reasonably high deposition rate. In the

Table 10.8 PECVD silane oxide process.

Step 2	Deposition
Pressure (torr)	3.0
Temperature (°C)	400
rf (W)	250
SiH_4 flow (sccm)	120
N_2O flow (sccm)	2400

PECVD nitride process, silane is commonly used as the silicon source, ammonia is used as the main nitrogen source, and nitrogen is used as the carrier gas and secondary nitrogen source:

$$SiH_4 + N_2 + NH_3 \overset{\text{plasma}}{\underset{\text{heat}}{\rightarrow}} SiN_xH_y + H_2 + N_2 + NH_3 + \cdots.$$

The passivation dielectric nitride requires good step coverage, high deposition rate, good uniformity, and stress control.

PECVD silane passivation dielectric deposition processes have seven steps, including stabilization 1, oxide deposition, pumping, stabilization 2, nitride deposition, plasma purging, and chamber pumping.

PECVD nitride has about 20 atomic percent (different from weight percent) of hydrogen. Hydrogen bonds with both nitrogen and silicon. Since Si–H bonds absorb UV light, conventional nitride is opaque to UV transmission.

Question: Where does the hydrogen in PECVD passivation dielectric nitride come from?

Answer: Both silicon precursor SiH_4 and nitrogen precursor NH_3 have hydrogen atoms, so naturally many of them are integrated into the film due to the low deposition temperature of passivation dielectric nitride.

By using silane and nitrogen without ammonia, nitride can be deposited with much less hydrogen concentration and high UV transmission rate. However, the deposition rate will be significantly lower due to the lack of nitrogen free radicals. A nitrogen molecule is very stable; it is much harder to dissociate nitrogen molecules than to dissociate ammonia molecules in plasma.

EPROM requires a UV transparent passivation layer to allow UV light to reach the floating gate. UV light can erase memory by exciting stored electrons in the floating gate to tunnel across the interpoly dielectric and drain out of the grounded upper gate. Oxynitride (SiO_xN_y) is commonly used for the EPROM passivation dielectric. By using silane, nitrogen, ammonia, and nitrous oxide, silicon oxynitride can be deposited:

$$SiH_4 + N_2 + NH_3 + N_2O \overset{\text{plasma}}{\underset{\text{heat}}{\rightarrow}} SiO_xN_y + H_2O + N_2 + \cdots.$$

Oxynitride has properties between oxide and nitride. Its refractive index is about 1.7 to 1.8, it is UV transparent, and it is a fairly good barrier layer for moisture and mobile ions. The oxynitride process requires monitoring of the refractive index more frequently than other dielectric CVD processes. This is because the barrier property and UV transmission rate of oxynitride film are sensitive to the silicon, nitrogen, and oxygen concentrations in the film. Refractive index measurements can provide this information.

10.5.3.2 Interlayer dielectric-0 barrier layer

Most fabs use TEOS-based PSG or BPSG for ILD0 applications. A diffusion barrier layer is needed to prevent dopant atoms (phosphorus and boron) from diffusing into activation areas and damaging the transistors. Both USG (1000 Å) and nitride (<300 Å) can be used for this application. As device dimensions have continued to shrink, most fabs prefer to use thinner LPCVD nitride as the barrier layer. When device dimensions shrink even farther, thermal budget limitations will rule out the application of LPCVD for ILD0 deposition; thus, PECVD nitride will start to be used for this application.

The PECVD ILD0 nitride barrier layer usually is deposited at around 550 °C, a deposition temperature significantly lower than the 700 °C required by LPCVD nitride processes. Therefore, the PECVD nitride process has a much lower thermal budget. It is higher than the 400 °C used for PECVD passivation nitride. At higher temperatures, PECVD nitride film has higher deposition rates, lower hydrogen concentration, and better film quality.

ILD0 nitride has also been applied as a stress liner to stress the gate and form a favorite strain in the channel of a MOSFET. An nMOS channel requires a tensile strain, and a pMOS channel requires a compressive strain. Both strains can be achieved in PECVD nitride by adjusting the rf power of the deposition process. A dual stress liner can be formed by depositing the first tensile strain nitride liner all over the wafer. Then, a mask defines the pMOS area to selectively remove the first nitride liner from the pMOS area. After photoresist is removed and the wafer is cleaned, a compressive stress liner is deposited. Depending on the device requirement, the second stress liner can either stay on top of the tensile stress liner in nMOS regions, or the compressive stress liner can be removed from the nMOS region with a masking step.

10.5.3.3 Dielectric antireflective coating

To achieve high resolution for photolithography processes, an ARC layer is required to reduce reflection from the shiny aluminum and polysilicon surfaces. With reflectivity of about 30 to 40%, titanium nitride (which can be sputter deposited with aluminum–copper alloy in a cluster PVD tool) is a commonly adapted ARC layer for IC chips with Al–Cu interconnections. For polysilicon patterning, a dielectric ARC layer is used to meet the resolution requirements of photolithography.

Figure 10.46 illustrates the dielectric ARC. When the phase difference of the reflected light on the photoresist–ARC interface (light 1), and the reflected light on ARC–metal interface (light 2) equals half of the wavelength,

$$\Delta\phi = 2nt = \lambda/2,$$

the two reflection rays will create destructive interference with each other and sharply reduce the reflection light intensity inside the photoresist. In the figure,

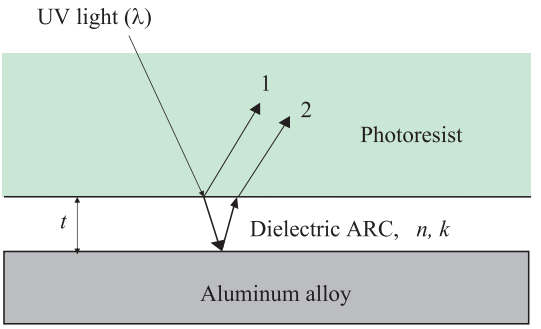


Figure 10.46 Dielectric ARC.

n is the refractive index of the ARC film, t is thickness of the ARC film, and λ is the lithography light wavelength. By controlling absorption coefficient k, the intensities of the two reflection rays can be exactly the same, so there will be complete destructive interference between the two and no reflected light in the photoresist at all. This significantly improves lithography resolution, especially when feature size is smaller than a quarter-micron.

Silane PECVD silicon-rich oxynitride film with a high refractive index ($n \sim 2.2$) has been developed for this application:

$$SiH_4 + N_2O + He \overset{\text{plasma}}{\underset{\text{heat}}{\rightarrow}} SiO_xN_y + H_2O + N_2 + NH_3 + He + \cdots.$$

SiH_4 flow rate controls the deposition rate, SiH_4/N_2O ratio determines the refractive index n, pressure and rf power control absorption rate k, and helium flow controls film uniformity. Normally, the film thickness is about 300 Å. For different wavelengths, i-line (365 nm), DUV (248 nm), and 193 nm, requirements for dielectric ARC film are different; therefore, different process recipes are needed for these applications.

10.5.4 Plasma-enhanced chemical vapor deposition tetraethoxysilane process

A thermal TEOS process can deposit oxide film with excellent step coverage and gap-filling ability. However, the deposition temperature is so high (>700 °C) that it is limited to ILD0 applications. In the 1980s, requirements of multilevel metal interconnections strongly drove the development of low-temperature TEOS oxide deposition processes for ILD applications. The PECVD TEOS oxide process uses plasma to dissociate oxygen molecules to generate free oxygen radicals, which dramatically enhance TEOS oxidation rate and help to achieve a high oxide deposition rate at relatively low temperatures (~400 °C). Because most TEOS precursors are physisorbed on the oxide surface with high surface mobility,

PE-TEOS oxide film has very good step coverage and conformality:

$$Si(OC_2H_5)_4 + O_2 \overset{\text{plasma}}{\underset{\text{heat}}{\rightarrow}} SiO_2 + \text{other volatiles.}$$

PE-TEOS USG was the dielectric most used for ILD applications in the 1990s. In a cluster tool, PE-TEOS chambers can be used with argon sputter etch chambers to complete an in-situ PE-TEOS dep/etch/dep processing sequence for gap fill and planarization. PE-TEOS USG is also widely used as barrier and cap layers for spin-on glass and O_3-TEOS USG processes in ILD applications.

PE-TEOS can also be used to deposit PSG and BPSG for ILD0 applications. By reacting phosphorus source gases such as TMP or TEPO [$PO(OCH_3)_3$], PSG can be deposited:

$$PO(OCH_3)_3 + Si(OC_2H_5)_4 + O_2 \overset{\text{plasma}}{\underset{\text{heat}}{\rightarrow}} SiO_2 + \text{other volatiles.}$$

With both phosphorus and boron (TMB or TEB) sources, a similar process can deposit PE-TEOS BPSG for PMD application. However, most fabs prefer nonplasma CVD processes such as O_3-TEOS CVD BPSG or PSG for ILD0 applications due to the concern of plasma-induced gate damage.

The PE-TEOS process is also employed to deposit fluorinated silicate glass (FSG). By using TEOS and oxygen along with a fluorine precursor gas such as SiF_4 or FTES [$FSi(OC_2H_5)_3$, fluorotriethyloxysilane], FSG can be deposited on a heated wafer surface in the rf plasma chamber. FSG has a lower dielectric constant (κ = 3.5 to 3.8) compared to USG's κ = 4.0 to 4.2 because the electronegative nature of fluorine decreases the polarizability of the overall SiO_xF_y network and lowers the dielectric constant. By using FSG, the parasitic capacitance between metal lines can be reduced, since the capacitance is proportional to the dielectric constant. Therefore, using FSG can reduce the RC time delay, minimize crosstalk, and reduce power consumption. It was used for ILD applications before maturity of the low-κ dielectrics. The chemical reaction of the FSG CVD process can be expressed as:

$$\underset{\text{(FTES)}}{FSi(OC_2H_5)_3} + \underset{\text{(TEOS)}}{Si(OC_2H_5)_4} + O_2 \overset{\text{plasma}}{\underset{\text{heat}}{\rightarrow}} \underset{\text{(FSG)}}{SiO_xF_y} + \text{other volatiles.}$$

The requirements for FSG films are stability, device compatibility, cost effectiveness, and easy process integration.

Figure 10.47 compares silicon dioxide, FSG, and silicon tetrafluoride. When fluorine is integrated into the silicate glass, the dielectric constant is reduced. The more that fluorine is integrated into the film, the lower the dielectric constant will be. However, the film will become more unstable, and fluorine tends to outgas from the film. If the fluorine concentration is too high, this can cause outgassing during the next thermal cycle when the wafer is heated up for deposition or metal

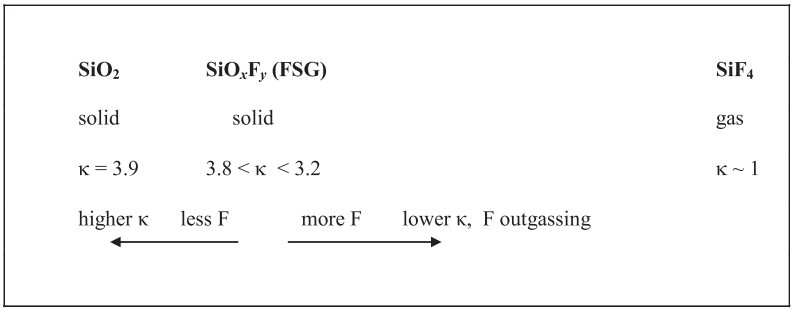


Figure 10.47 Demonstration of FSG processing trends.

annealing. HF and F_2 being released from the film can cause device or circuit damage due to corrosion.

FSG film has better gap-fill abilities than USG film due to the etch effect of the combination of plasma-induced free fluorine radicals and ion bombardment, especially on the upper corner of the gap. Because of the larger arriving angle at the upper corner, this area receives more ion bombardment and more free fluorine radicals, which keep an overhang from forming. By using PE-TEOS SiF_4 chemistry, a void-free 0.65-μm gap fill without an etchback can be achieved, as can a 0.35-μm void-free gap fill with an etchback.

10.5.5 Dielectric etchback process

Sputtering and planarization etchback processes are widely used with dielectric CVD to achieve gap fill and planarization. Unlike the deposition process, etchback is a removal process. In many IC fabs, dielectric etchback processes are performed in the thin-film area, along with the dielectric CVD process. Therefore, etchback processes are covered in detail in this chapter instead of the etch chapter (Chapter 9).

In many cases, sputtering etch chambers are placed on the same mainframe with the dielectric CVD chambers of a cluster tool to operate the in-situ integrated dep/etch/dep process, as shown in Fig. 10.48.

The sputtering etch process is a purely physical etch process. It operates at low pressure (~30 mtorr) in argon plasma. Argon ions bombard the wafer surface and dislodge the dielectric thin film on the surface by breaking chemical bonds by means of energy and momentum transfer from the bombarding ions. The dielectric fragments are negatively charged by the electrons from the plasma, and expelled from the wafer surface, since the wafer is also charged negatively by the plasma. The dislodged dielectric fragments then cross the boundary layer into the convection flow and are pumped out from the chamber.

Since the sputtering etch is a purely physical etch (which solely depends on argon ion bombardment, with no chemical reaction with dielectric thin films), it has a very low etch rate. It chips off the step corners much faster than it removes materials from the surface. The step corner facets are about 45 deg in slope,

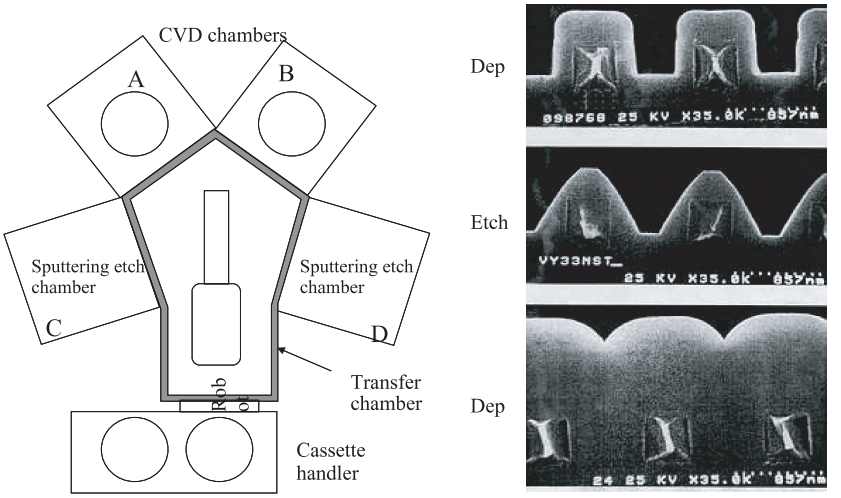


Figure 10.48 (a) Cluster tool and (b) SEM images of the PE-TEOS dep/etch/dep process (Applied Materials, Inc.).

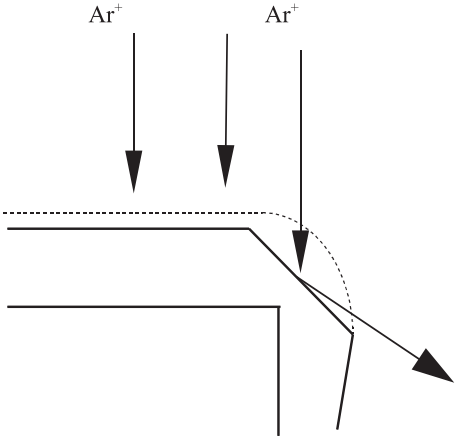


Figure 10.49 Argon sputtering chops the corner of a gap.

tapering the gaps and making it easier for the ensuing deposition step to fill the gap without voids.

Question: Why do sputtering etch processes usually use argon as a processing gas?
Answer: Gap-fill sputtering etch processes need to chip off the step corners of the gaps with purely physical bombardment. Therefore, an inert

gas is required. Since argon is a heavy and inexpensive noble gas, it became the gas of choice for almost all physical sputtering processes. The atomic weight of argon is 40, compared with silicon's 28 and helium's 4. Heavy ions are favored for achieving better bombardment results. Argon is the third most abundant component in the earth's atmosphere (~1%) [only less abundant than nitrogen (78%) and oxygen (20%)]. It can be purified directly from condensed air, making it more affordable than other heavier inert gases such as krypton, xenon, and radon. Radon is a radioactive gas and cannot be used in semiconductor processes.

A sample of an argon sputter etch process recipe is listed in Table 10.9. The expected blanket etch rate on thermally grown silicon dioxide is about 250 Å/min for this process. Thermally grown oxide is used for process qualification because of its WIW and WTW uniformity. For CVD oxide, the sputter etch rate is slightly higher. Like most plasma etch processes, the etch rate of the sputter etch process is mainly determined by rf power because both ion bombardment energy and ion density will increase when rf power increases. Increasing rf power can significantly increase the etch rate. Etch uniformity is very sensitive to the pressure.

A schematic of an etch chamber is given in Fig. 10.50. Since sputter etch operates at a low pressure and the electron MFP is long, it is difficult to generate and maintain the plasma. An external magnetic field is introduced to force electrons to spiral around the magnetic field lines and travel much longer distances. This increases the probability of ionization collisions between electrons and the neutral argon atoms. The sputtering etch process is a blanket etch process, and there is no need to protect the photoresist. The sputtering etch process is a purely physical etch process, and the sputter etch rate is not sensitive to wafer temperature, so the sputtering etch chamber does not need a wafer cooling system.

By using CF_4 and O_2, a sputter etch chamber can also be used for dielectric planarization etchback, as shown in Fig. 10.51. In this case, oxide is etched by a combination of both fluorine chemical etch and ion bombardment sputtering. Because chemical etch tends to etch isotropically, the dielectric thin-film surface after etchback will be smoother than when it is deposited. To enhance chemical etch effect, the planarization etchback process normally operates at a much higher pressure than does a sputtering etchback process. Increasing pressure can reduce MFP, increase ion collisions and scattering, and reduce ion energy. This reduces ion bombardment and the physical sputtering etch effect.

Table 10.9 Sample argon sputter etch recipe.

Pressure (mtorr)	30
rf power (W)	300
B-field (G)	50
Ar (sccm)	50

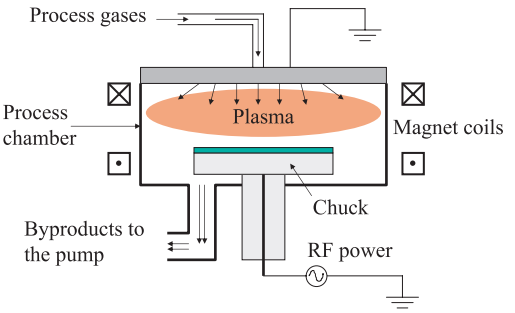


Figure 10.50 A sputtering etch chamber.

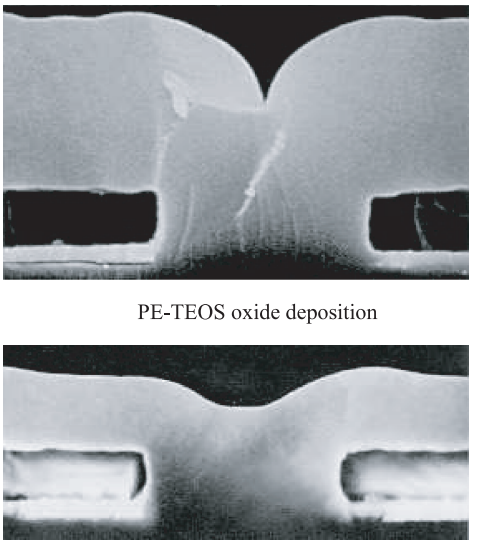


Figure 10.51 SEM images of the planarization etchback process (Applied Materials, Inc.).

10.5.6 Ozone-tetraethoxysilane process

Ozone is a very unstable molecule, so thermal energy can easily dissociate it and release chemically reactive free oxygen from the reaction:

$$O_3 \rightarrow O_2 + O.$$

At 25 °C, the half-life of ozone is about 86 h, which means that after 86 h, half of the ozone molecules will be dissociated and gone. Because chemical reaction rates increase exponentially with temperature [see Eq. (10.2)], the dissociation rate of ozone is much faster, and its half-life is much shorter at higher temperatures. At

400 °C, the half-life of ozone is less than a millisecond. Therefore, ozone can be used as a carrier of oxygen free radicals at room temperature and can generate free oxygen at higher temperatures at the wafer surface to enhance chemical reactions without the presence of plasma inside the processing chamber. Using ozone to react with TEOS oxidizes the TEOS and deposits silicon oxide. O_3-TEOS oxide film has excellent conformality and gap-fill abilities, which can fill very narrow gaps, and it is commonly employed for deep submicron IC chip dielectric thin-film deposition. O_3-TEOS oxide can be deposited at both atmospheric (APCVD) and subatmospheric (SA-CVD) pressures.

10.5.6.1 Ozonator

Since ozone has a very limited lifetime, it must be generated on site and consumed right away. In nature, ozone is generated in lightning during a thunderstorm. An ozonator generates ozone in a very similar way. In a high-pressure corona discharge, oxygen molecules are dissociated by energetic electrons, and released as free oxygen radicals:

$$O_2 \overset{\text{plasma}}{\rightarrow} O + O.$$

The free oxygen radicals react with an oxygen molecule to form ozone in a three-body collision:

$$O + O_2 + M \rightarrow O_3 + M \ (M = O_2, N_2, Ar, He, \text{etc.}).$$

The third party in the collision is necessary to meet the requirements for conservation of both energy and momentum.

Figure 10.52 shows an ozone generation process. Mixing approximately one-percent nitrogen in the oxygen flow into the ozonator can improve and stabilize ozone yield. Several ozone cells are connected in a series to repeat the process and maximize ozone concentration. Since ozone is unstable and tends to dissociate, reducing ozone cell temperature can also help improve ozone yield by reducing ozone dissociation.

After passing several ozone cells, ozone concentration can reach up to 14 wt%, depending on the oxygen flow rate, rf power, and nitrogen concentration. Generally, at the same rf power and nitrogen concentration, a higher flow rate will have lower ozone concentration. At the same flow rate and nitrogen concentration, higher rf power will have higher ozone yield.

Ozone concentration can be monitored by its UV absorption property. From Beer–Lambert's law,

$$I = I_0 \exp(-ACL),$$

where I is the UV intensity measured with ozone, I_0 is the UV intensity measured without ozone flow, A is the UV absorption coefficient of ozone, L is the absorption

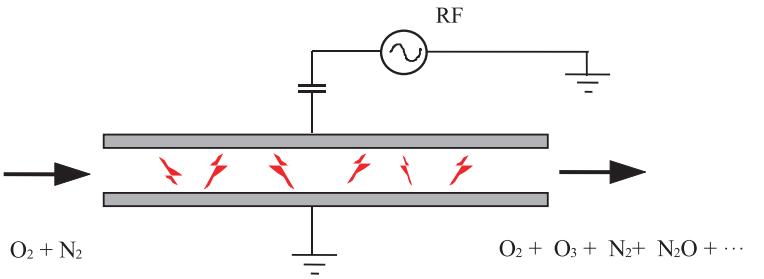


Figure 10.52 Ozone generation in an ozone cell.

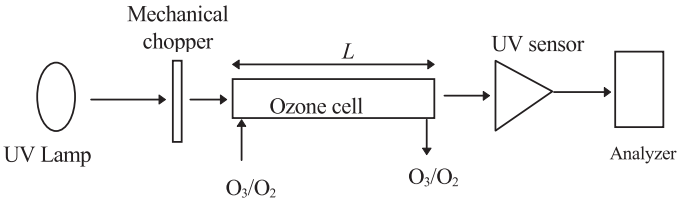


Figure 10.53 Zone concentration monitoring system.

cell length, and C is the ozone concentration. Figure 10.53 shows an ozone monitoring system.

The main goal for an ozonator is to generate stable, repeatable, and high-concentration ozone flow at high flow rates to maintain maximum ozone output.

10.5.6.2 Ozone-tetraethoxysilane undoped silicate glass process

The main applications of O_3-TEOS USG films are STI trench fill and ILD layering of IC chips with Al–Cu interconnections. Using ozone to react with TEOS in the vicinity of 400 °C deposits silicon oxide at reasonably high deposition rates, 1500 to 2000 Å/min, depending on gas flow rate, temperature, and pressure:

$$\text{TEOS} + O_3 \xrightarrow[\text{heat}]{} SiO_2 + \text{volatile organics}\,.$$

Because it has almost perfect step coverage and conformality, O_3-TEOS USG film can fill very small gaps without voids, as shown in Fig. 10.54. It is often used for ILD applications in multimetal layer interconnections.

In the O_3-TEOS deposition process, higher O_3:TEOS ratio means more free oxygen to react and oxidize TEOS. Therefore, the higher the O_3:TEOS ratio, the better the deposited oxide film quality. To increase the O_3:TEOS ratio, both ozone concentration and ozone flow rate need to be increased (both of which are limited by ozonator capability); another way is to reduce TEOS flow rate, causing a lower deposition rate.

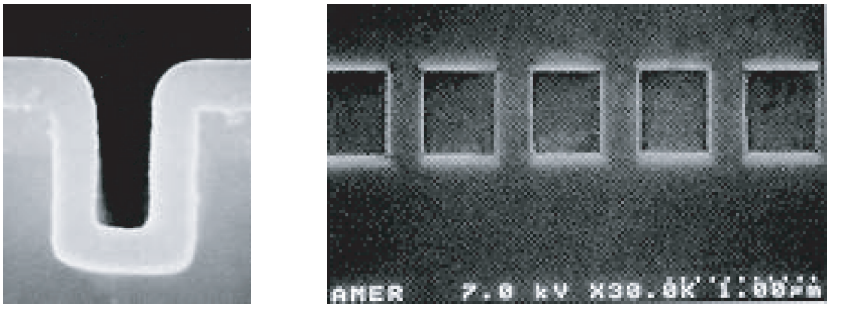


Figure 10.54 O_3-TEOS USG step coverage and gap fill (Applied Materials, Inc.).

Because O_3-TEOS oxide film is porous and absorbs moisture, it needs a dense PECVD cap film to seal it from the atmosphere. Since it has unfavorable tensile stress, it requires a compressive PECVD film to buffer the stress to avoid metal hillocks. O_3-TEOS USG film for ILD applications is always sandwiched between two PECVD USG films.

For subatmospheric pressure O_3-TEOS processes, all three layers are deposited in the same processing chamber, with the same processing steps. First, a PE-TEOS USG thin film (about 1000 Å) is deposited as a stress buffer layer, then pressure is gradually ramped up to prevent deposited film on the wall of the process reactor from flaking and causing particle contamination. After stabilization of TEOS flow and pressure, O_3-TEOS USG film is deposited to fill the gap by turning on the ozonator and flowing ozone into the processing chamber. After gradually pumping down the chamber and stabilizing the pressure, oxygen flow, and TEOS flow, rf power is turned on to strike the plasma, and a thick layer of PE-TEOS USG cap film is deposited to seal the porous O_3-TEOS USG film. Either CMP or etchback is employed to planarize the PE-TEOS USG cap layer, depending on the IC processing requirements.

10.5.6.3 Ozone-tetraethoxysilane phosphosilicate glass and borophosphosilicate glass processes

Both O_3-TEOS BPSG and PSG are commonly used for ILD0 applications. Figure 10.55 shows that O_3-TEOS BPSG fills 0.25-μm gaps with aspect ratios of 4:1 in ILD0 applications. Reacting ozone with TEOS, TEB, and TEPO will form BPSG. The reaction can be expressed as:

$$O_3 \rightarrow O_2 + O,$$
$$O + TEB + TEPO + TEOS \rightarrow BPSG + \text{volatile organics}.$$

10.6 Spin-On Glass

The SOG process and photoresist coating and baking processes are very similar, which is one important reason that fabs adopted this technology, since

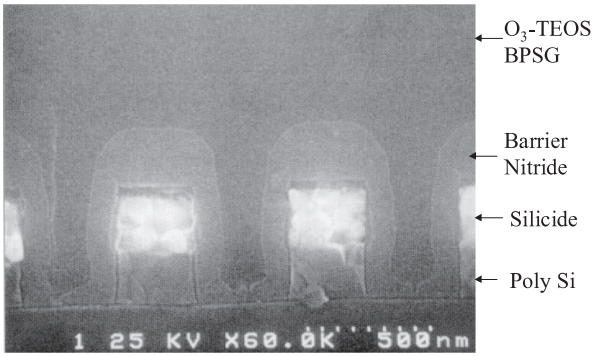


Figure 10.55 O_3-TEOS BPSG fills 0.25-μm gaps with aspect ratios of 4:1 (Applied Materials, Inc.).

manufacturers like familiar technology. Some IC producers have used a SOG process for ILD gap fill and planarization of IC chips with Al–Cu interconnections for a long time.

There are two kinds of SOGs commonly used in the semiconductor industry: one is silicate, the other is siloxane. Both of them have Si-O bonds, as shown in Fig. 10.56. Solvents used for SOG are alcohol, ketone, etc.

SOG processes require several processing tools and different process steps, such as PECVD/SOG spin coating/SOG curing/SOG etchback/PECVD. SOG is always sandwiched between two PECVD layers, as shown in Fig. 10.57.

To begin, a thin layer of PECVD film is deposited in a PECVD reactor as a liner or barrier layer; then liquid SOG is uniformly spun onto the wafer surface in a spinner and forms a thin film a few thousand angstroms thick. Film thickness is determined by the spin rate and viscosity of the liquid, just as with the spin coating of a photoresist. The liquid may need to be spun twice to achieve the required uniformity. Surface tension of the liquid forces SOG to flow into the narrow gaps and fills them without voids. After prebake on a hotplate, some solvents are driven out of the SOG, and Si–O bonds start to crosslink. Then the wafers are placed into a baking furnace to bake at 400 to 450 °C, which drives out most solvents, making it solid silicate glass. The thickness of the film shrinks about 5 to 15% after SOG curing processes. Since the quality of the glass is not very good, in most cases etchback is required to remove the majority of SOG from the surface and leave SOG only in the gaps. A PECVD cap layer is deposited to prevent both outgassing and moisture absorption of the SOG.

In some cases, fabs prefer to deposit the cap PECVD oxide without etchback to save a major processing step in order to improve throughput and reduce production cost. This is called SOG without etchback, and the possibility of SOG outgassing is much higher than that of SOG with etchback.

The disadvantages of SOG are its inherently complicated process integration, vulnerability to particulate contamination, film cracking or peeling, and residue

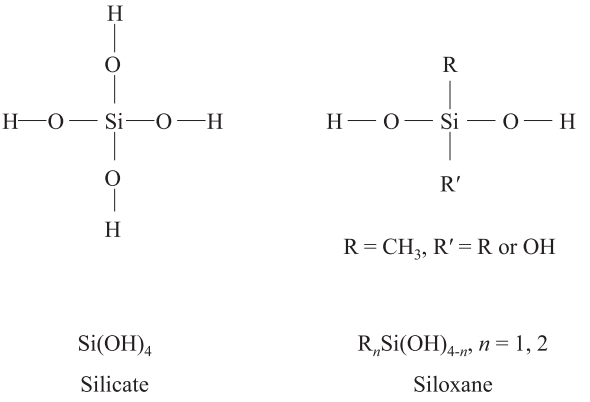


Figure 10.56 Two common SOGs (without solvent): (a) silicate [$Si(OH)_4$] and (b) siloxane [$R_nSi(OH)_{4-n}$, n = 1, 2].

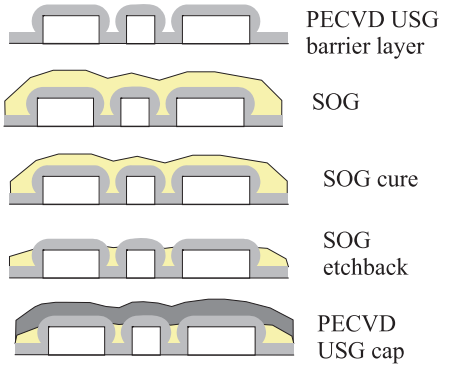


Figure 10.57 SOG processing sequence.

solvent outgassing. The last three problems can be solved by carefully controlling the process.

10.7 High-Density Plasma Chemical Vapor Deposition

While the dep/etch/dep processing sequence can fill small gaps, it requires two chambers: a CVD chamber and an etch chamber. Wafers need to be transferred back and forth between these two processing chambers. When a gap becomes narrower due to feature size, it may need an additional dep/etch cycle to become filled without voids. When the gaps become even smaller, and aspect ratios become even larger, many dep/etch cycles will be needed to fill the gaps, making it impractical for mass production because throughput will be too low. Therefore, a tool that can deposit and sputter etch simultaneously would be very helpful for deep submicron gap filling.

Question: With smaller feature sizes, the metal line width and gap between the metal lines become smaller. However, the metal line height does not shrink accordingly, causing a larger gap aspect ratio. Why does the metal height not shrink accordingly to keep the same aspect ratio and make it easier for dielectric gap fill?

Answer: Metal line resistance $R = \rho l/wh$. With smaller feature size, metal line length l and width w shrink accordingly. If metal line height also reduces accordingly, metal line resistance increases accordingly, which, of course, is unacceptable. Therefore, the metal interconnection line must retain the same height. By reducing feature size, more chips can be placed on a wafer. However, this also raises more challenges for dielectric deposition processes.

To achieve a high sputtering etch rate, the processing chamber must operate at a low pressure (<30 mtorr, the lower the better) to provide the long MFP required for ions to accelerate. At that low pressure, PECVD deposition rate is too low to match the sputtering etch rate in a standard capacitively coupled plasma chamber with two parallel plate electrodes, due to low plasma density. Keep in mind that the deposition rate must be higher than the etch rate to achieve net deposition. Therefore, different HDP sources are needed for these in-situ dep/etch/dep process reactors. Two types of HDP sources have been developed and applied in semiconductor processing: ICP and ECR, as illustrated in Fig. 10.58.

Since HDP-CVD is an in-situ dep/etch/dep process, the net deposition rate is not very high. Therefore, HDP-CVD is normally only used for gap fill, while the cap layer is deposited by the PECVD process, which has significantly higher deposition rates.

Since some of the dislodged oxide fragments that sputtered off the gap corner redeposit at the gap bottom, deposition rate of the film on the bottom is usually about three times higher than that on the sidewall in HDP-CVD processes. The result is that the sidewall films never meet, and the gap is filled from the bottom up without a seam, as shown in Fig. 10.59. This is an attractive feature with respect to subsequent integration with CMP because in CMP processes a seam provides a weak point and polishes faster than the other areas, creating a defect on the surface.

HDP-CVD oxide processes use silane as a silicon precursor and oxygen as an oxygen precursor. Argon is added for enhanced sputtering etch effect. It can also deposit PSG and FSG by reacting with phosphine and silicon tetrafluoride:

$$\begin{aligned}
&\text{USG } SiH_4 + O_2 + Ar \rightarrow \text{USG} + H_2O + Ar + \cdots,\\
&\text{PSG } SiH_4 + PH_3 + O_2 + Ar \rightarrow \text{PSG} + \text{volatiles},\\
&\text{FSG } SiH_4 + SiF_4 + O_2 + Ar \rightarrow \text{FSG} + \text{volatiles}.
\end{aligned}$$

Question: Why is silane used instead of TEOS as the silicon source gas in HDP-CVD oxide processes?

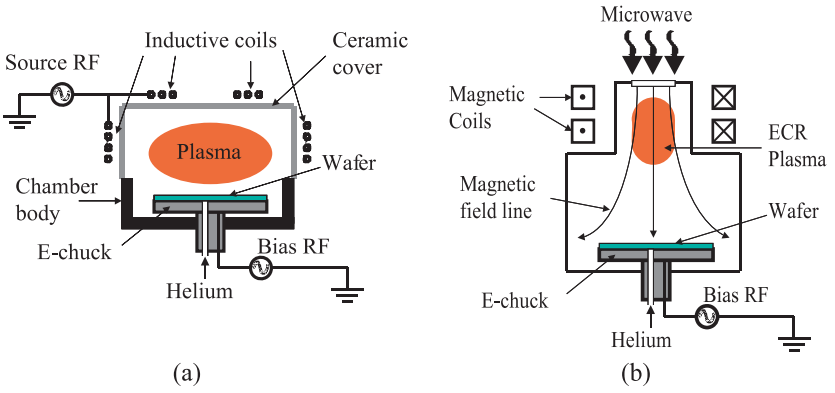


Figure 10.58 Schematics of (a) ICP and (b) ECR chambers.

Answer: For HDP-CVD, step coverage is no longer an important factor for gap fill because heavy ion bombardment always keeps the gap tapered, and deposition is from the bottom up. Using silane can save the cost and hassle related to the vapor delivery system of liquid TEOS sources.

In HDP-CVD, the deposition rate is controlled by silane flow, and the refractive index is determined by the SiH_4/O_2 flow ratio. Film stress is mainly controlled by the bias rf power, source rf, and helium backside pressure, which controls the wafer temperature. The pressure and source rf power can also affect film uniformity.

10.8 Dielectric Chemical Vapor Deposition Chamber Cleaning

During the dielectric CVD process, a dielectric thin film can be deposited not only on the wafer surface but also on everything inside the chamber, particularly the wafer chuck, showerhead, and chamber wall. It is very important to clean the chamber routinely to prevent the thin film on those surfaces from cracking and causing particulate contamination. For dielectric CVD tools, it is not unusual for chambers to be undergoing cleaning processes more than half of the time, rather than running deposition processes!

For PECVD tools, rf plasma cleaning is a popular way to clean the chamber, and remote plasma cleaning processes are receiving more attention.

10.8.1 Radio-frequency plasma cleaning

One of the advantages of dielectric PECVD over LPCVD and APCVD is that it can use plasma-generated free fluorine radicals for chamber dry cleaning. A plasma cleaning process is a plasma etch process in which both chemical and physical processes help remove dielectric film on the processing kits and chamber walls. Both silicon oxide and silicon nitride processes use carbon fluoride compounds such as CF_4, C_2F_6, and C_3F_8 as a fluorine precursor gas because they are stable and easy to handle. In some cases, NF_3 is used to generate more free fluorine radicals.

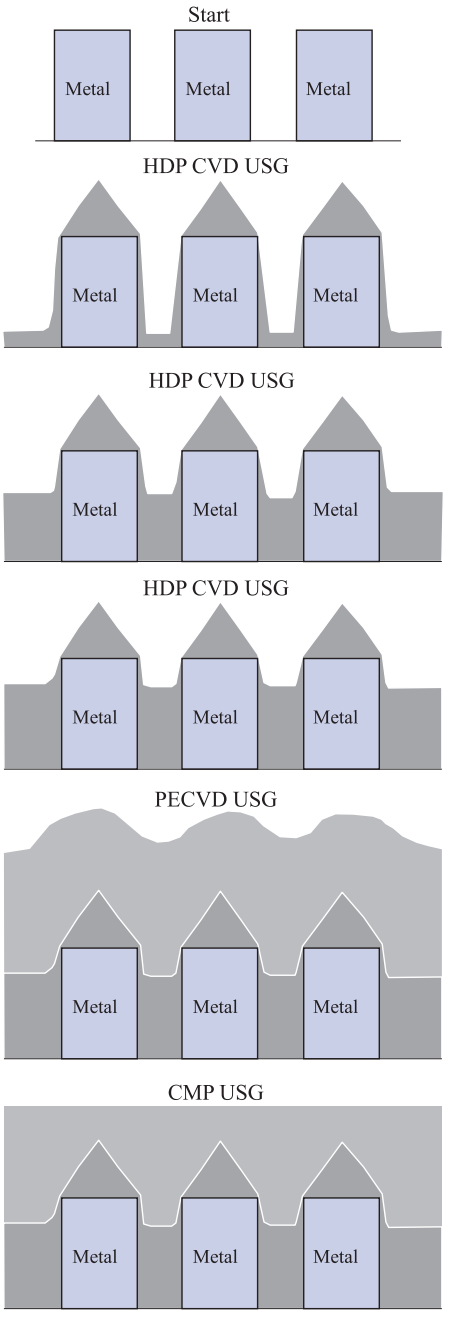


Figure 10.59 IMD HDP-CVD gap fill and planarization processes.

In plasma, carbon fluoride is dissociated to release the free fluorine needed for removing silicon oxide and silicon nitride:

$$CF_4 \overset{\text{plasma}}{\rightarrow} CF_3 + F,$$
$$F + SiO_2 \underset{\text{heat}}{\overset{\text{plasma}}{\rightarrow}} SiF_4 + O,$$
$$F + Si_3N_4 \underset{\text{heat}}{\overset{\text{plasma}}{\rightarrow}} SiF_4 + N.$$

In plasma cleaning processes, oxygen source gases such as N_2O and O_2 are used with carbon fluoride to react with carbon to form CO and CO_2, freeing more fluorine radicals and increasing the F/C ratio. This can prevent carbon fluoride polymerization, and increases cleaning efficiency.

It is very important to have a F/C ratio above 2, otherwise polymerization can occur. In plasma, CF_4 dissociates into CF_3 and F. CF_3 will continue to dissociate into CF_2 and F. The linking of CF_2 radicals into a long chain is called polymerization (as shown in Fig. 10.60). In this case, white Teflon®-like polymers are deposited inside the processing chamber, and it is necessary to open the chamber and do a wet clean to mechanically scrub the polymer off the chamber. This tool downtime can significantly affect throughput.

Chamber cleaning processes require six steps. Pressure stabilization, plasma cleaning, pump down, pressure set up for deposition, seasoning, and pump down. In the seasoning step, about 1000-Å oxide film is intentionally deposited on the wafer chuck and everywhere else inside the chamber. This is a very important step that can help keep the same deposition condition for each wafer, achieve good WTW uniformity, and improve production repeatability. It also can help reduce chemical contamination by covering fluorine residue, as well as reduce particulate contamination by sealing the loose fragments of the thin-film residue.

A sample recipe for 200-mm-wafer silane PECVD chamber cleaning is given in Table 10.10.

It takes about 90 sec to clean the chamber after 1-μm USG deposition. In the silicon oxide cleaning process, an etch chemical reaction releases oxygen from the

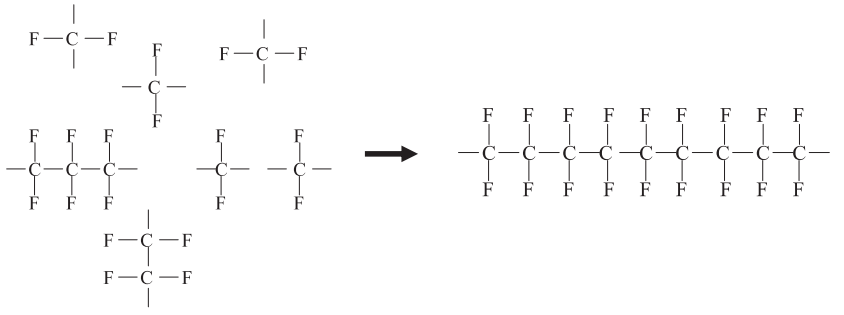


Figure 10.60 Carbon fluoride polymerization process.

Table 10.10 200-mm-wafer silane PECVD chamber cleaning recipe.

Step 2	Clean
Pressure	5 torr
rf	1000 W
CF_4	1200 sccm
N_2O	400 sccm

film and causes extra carbon oxidation. Since there is no extra oxygen released from the film, the silicon nitride cleaning needs a higher N_2O flow rate (about 600 sccm) to react with carbon, free more fluorine, and keep the right F/C ratio to prevent polymerization. It takes about 2 min to clean a nitride chamber after 1-μm nitride film deposition. By comparison, it takes about 1 min to deposit 1-μm oxide, and about 85 sec to deposit 1-μm nitride.

Question: For the 1200-sccm CF_4 flow into the processing chamber, what percentage is actually used for the oxide cleaning process?

Answer: Less than 3%! More than 97% of CF_4 simply flows into the chamber and is pumped away. CF_4 is a stable gas and cannot be removed by a scrubber; it is considered safe to release into the atmosphere. However, carbon fluoride gases are classified as global warming gases, and demands for limiting their use are increasing.

In the plasma, excitation-relaxation collisions between electrons, atoms, and molecules cause glow. Since different gases have different atomic structures, the colors of light emission in plasma can provide information about the chemical components inside the chamber and can be used to monitor the cleaning process.

When the cleaning process begins, the plasma generates free fluorine atoms, and initially the fluorine increases. During the main cleaning process, fluorine atoms are busy etching silicon oxide or silicon nitride, so the free fluorine concentration in the plasma remains low, and the intensity of the fluorine line (704 nm) also stays low. When approaching the end of the cleaning process, silicon oxide is gradually removed, and the free fluorine concentration in plasma increases; the fluorine line intensity also increases. When oxide is completely removed, the fluorine concentration becomes a constant, which indicates the endpoint, as shown in Fig. 10.61.

10.8.2 Remote plasma cleaning

Since the rf plasma cleaning process uses high rf power to generate plasma inside a processing chamber, ion bombardment associated with plasma causes processing kit damage and increases the cost of ownership for IC manufacturers. In rf plasma cleaning, most carbon fluoride gases are not dissociated and are released into the atmosphere. Carbon fluoride gases are considered global warming gases. Global warming is mainly caused by increased carbon dioxide in the atmosphere due to fossil fuel consumption, which always accompanies development and industrialization. Carbon dioxide and carbon fluoride gases absorb the inferred

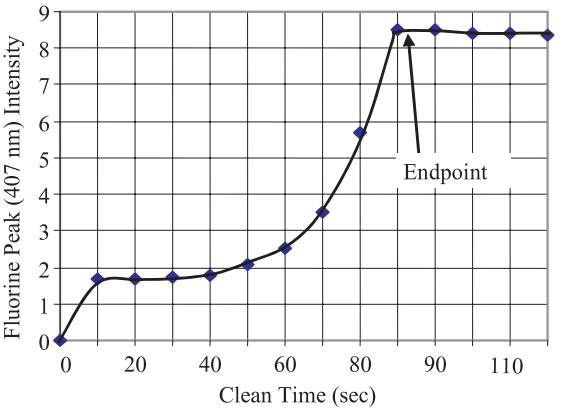


Figure 10.61 Endpoint trace of a cleaning for 1.8-μm PE-TEOS film (Applied Materials, Inc.).

radiation from sunlight more efficiently than nitrogen and oxygen. Increasing the amounts of carbon dioxide and carbon fluoride gases in the atmosphere can increase the heat absorption and raise the temperature of the earth.

Carbon fluoride gases are also believed to be responsible for holes in the ozone layer of the upper atmosphere. The ozone layer blocks UV radiation from solar rays. Areas under the holes have abnormally high UV radiation intensity, which can cause sunburn and skin cancer. Due to the long lifetimes of carbon fluoride gases in the atmosphere (half-life >10,000 years), they can cause long-term environmental damage. Therefore, there are strong demands to reduce the release and use of carbon fluoride by the semiconductor industry. Remote plasma cleaning fills this demand.

By using MW power, which has a much higher frequency than rf, a stable plasma can be generated at high pressure. Normally, NF_3 is used as a fluorine source gas. In high-density MW plasma, more than 99% of NF_3 is dissociated and releases three free fluorine atoms. The free fluorine atoms then flow into the CVD processing chamber and react with the deposited film on the processing kits and chamber wall to form gaseous SiF_4 and other gaseous byproducts, and eventually are pumped out of the chamber. Since there is no plasma inside the processing chamber, there is no ion bombardment on the chamber body or parts inside the processing chamber, prolonging their lifetime. Figure 10.62 illustrates a chamber cleaning process with a MW remote plasma source.

The advantages of MW remote plasma cleaning are longer lifetimes of the processing chamber parts, lower cost of ownership, and much less fluoride gas release. Disadvantages are a less-mature technology compared to the production-proven rf plasma cleaning process, higher cost for extra equipment, and the use of hazardous, expensive NF_3 gas. Another disadvantage is that the existing optical endpoint system for rf plasma cleaning cannot be used for remote plasma cleaning because there is no plasma-induced light emission in the processing chamber.

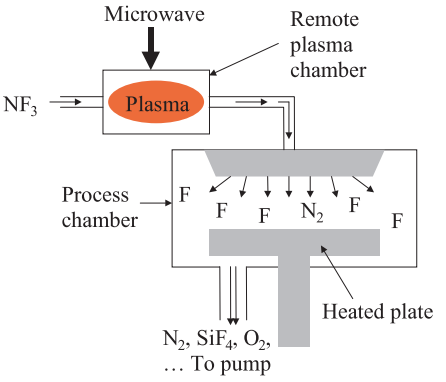


Figure 10.62 Remote plasma CVD chamber cleaning process.

Remote plasma cleaning requires different types of measurement methods, such as Fourier transform infrared (FTIR) spectroscopy, to achieve an automatic processing endpoint. FTIR systems measure chemical bonding concentrations to detect chemical component changes and determine the cleaning endpoint.

10.9 Process Trends and Troubleshooting

10.9.1 Silane plasma-enhanced chemical vapor deposition processing trends

Many silane-based thin-film processes are PECVD processes. For silane processes, due to the higher diffusion rate of the precursor in the boundary layer, increasing the process temperature always increases the deposition rate [see Fig. 10.63(a)]. Increasing temperature will also improve deposited film step coverage and quality.

For PECVD films, increasing rf power always increases compressive stress [see Fig. 10.63(b)] because increasing rf power will increase both ion density and energy of ion bombardment. Increasing rf power increases the deposition rate when rf power is low and reduces the deposition rate when rf power is high. At very low rf power, the deposition rate can be very low at relatively low temperatures (~400 °C) because there are not enough free radicals to enhance the chemical reaction rate. When rf power increases, the deposition rate increases dramatically due to the enhancement of the surface chemical reaction from the free radicals generated by the plasma. When rf power further increases, deposition will be well into the mass-transport-limited regime, and the deposition rate will no longer be related to the chemical reaction rate on the surface. Deposition rate stops increasing and starts to decrease because the heavier ion bombardment associated with increasing rf power reduces the precursor adsorption rate on the surface.

From Fig. 10.64, it can be seen why most PECVD processes operate at the high-rf-power side of the curve because in that regime, deposition rate is not very sensitive to the rf power. Therefore, rf power can be used to control the film stress and WERR without having a major effect on deposition rate. Deposition rate is too sensitive to the rf power if the processes operate on a low-rf power regime.

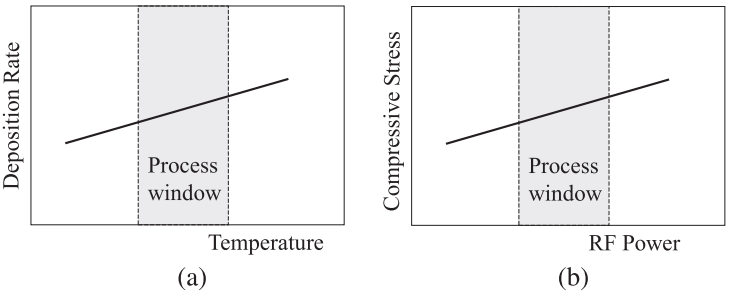


Figure 10.63 Silane PECVD processing trends: (a) temperature-to-deposition rate, and (b) rf power to film stress.

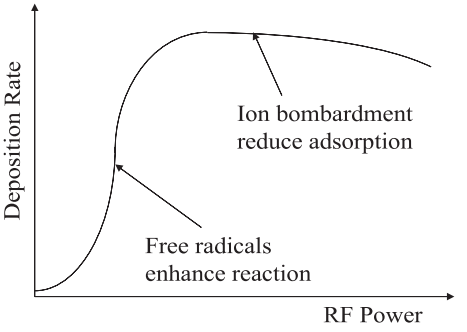


Figure 10.64 Relationship of deposition rate to rf power.

For the silane process, refractive index is mainly determined by the chemical composition of the film. Silicon-rich film has a higher refractive index, and oxygen-rich film has a lower refractive index, as shown in Fig. 10.65.

Since silane PECVD processes operate in the mass-transport-limited regime, the deposition rate is mainly determined by the gas flow rate, especially the silane flow rate. By increasing silane flow alone, the deposition rate increases while film compressive stress is reduced due to less ion bombardment per unit of film thickness. Refractive index increases because the film becomes more silicon rich; the WERR increases as well because the film is less dense due to less ion bombardment per unit thickness film.

10.9.2 Plasma-enhanced tetraethylorthosilicate trends

There are some similarities in processing trends between PE-TEOS and PE-silane processes. When increasing rf power from a very low level, the deposition rate increases very rapidly at first, then slows down and eventually starts to decrease when rf power is high enough (shown in Fig. 10.64). At the same time, film stress becomes more compressive due to increasing ion bombardment.

Similar to the PE-silane process, the PE-TEOS process also operates in the mass-transport-limited regime, and deposition rate is mainly controlled by TEOS

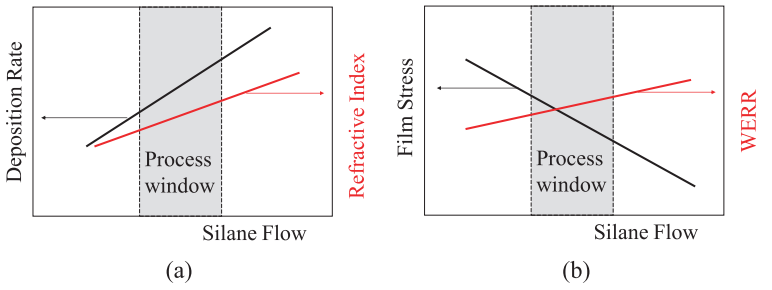


Figure 10.65 PECVD silane processing trends: (a) silane flow rate versus deposition rate and refractive index, and (b) silane flow rate versus compressive stress and WERR.

flow rate. Since TEOS vapor always flows into the processing chamber with a carrier gas such as helium, it is important to have a precise and independent control of the net TEOS flow rate. By increasing TEOS flow rate alone, the deposition rate will increase linearly, stress will become less compressive, and the wet etch rate will increase, as shown in Fig. 10.66.

Question: For silane processes, increasing silane flow rate increases the refractive index significantly because the deposited film becomes richer in silicon. Why does the refractive index almost not change when the TEOS flow rate increases, as shown in Fig. 10.66?

Answer: In a TEOS molecule, the silicon atom is already bonded with four oxygen atoms. Therefore, increasing TEOS flow does not make the deposited film silicon rich. For TEOS oxide film, the refractive index is very stable and only changes slightly due to film density change. Denser film has a slightly higher refractive index.

Different from the silane process, increasing temperature for a TEOS process at 400 °C always reduces the deposition rate, no matter whether it is a PE-TEOS or an O_3-TEOS process. The difference in this trend is caused by the different adsorption mechanisms of TEOS and silane. Silane uses chemisorption, and TEOS uses physisorption. For a chemisorption precursor such as silane, the bonding energy between the precursor and surface atom is very strong, and a temperature change near 400 °C has no significant effect on the adsorption rate. For a physisorption precursor like TEOS, the bonding energy is low, and varying the temperature near 400 °C has some strong effects on the adsorption rate. At higher temperatures, more physisorbed precursors can receive enough energy to break the weak bonds and leave the surface, thereby decreasing the adsorption rate and causing lower deposition rates, as shown in Fig. 10.67.

10.9.3 Ozone-tetraethoxysilane trends

O_3-TEOS film always has tensile stress, and there is no way to control the film stress, since its deposition is a thermal process. The O_3-TEOS film refractive index is about 1.44, due to the porosity of the film.

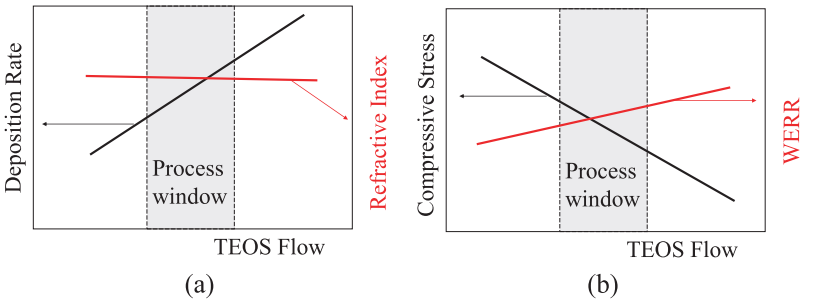


Figure 10.66 PE-TEOS processing trends: (a) TEOS flow rate versus deposition rate and refractive index, and (b) TEOS flow rate versus compressive stress and WERR.

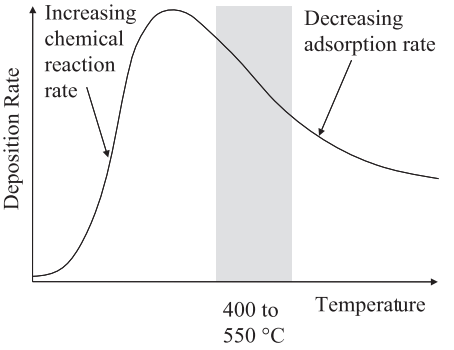


Figure 10.67 TEOS processing trend: deposition rate versus temperature.

In the O_3-TEOS deposition process, a higher O_3:TEOS ratio is always desired, since a higher ratio means more free oxygen to react and oxidize the TEOS; the higher the O_3:TEOS ratio, the better the oxide film quality. O_3-TEOS USG film quality can be monitored by measuring the WERR related to the thermally grown SiO_2, a measurement that normally is around 4 to 5. Film quality can also be monitored by the shrinkage measurement. O_3-TEOS film shrinkage is normally about 4 to 5% for the 450 °C thermal cycle. Shrinkage can reach 7% for >1000 °C STI annealing. Increasing the O_3:TEOS ratio can reduce both WERR and film shrinkage. One way to increase the O_3:TEOS ratio is to increase both ozone concentration and ozone flow rate, which is limited by the ozonator capability; another way is to reduce TEOS flow rate, causing a lower deposition rate.

At high pressure, the O_3-TEOS process has better step coverage and gap-fill abilities. This is because higher pressure reduces the MFP, allowing TEOS molecules to have more collisions in the boundary layer, causing them to hop along the substrate surface more frequently to eliminate the arriving angle effect.

Increasing the deposition temperature for the O_3-TEOS process increases chemical reaction rates and improves film quality. Increasing deposition temperature can

also increase surface mobility, thus improving film step coverage and gap-fill abilities. Since a TEOS molecule is physisorbed on the substrate surface, increasing temperature reduces the adsorption rate and increases desorption rate, reducing deposition rate. The relation of temperature and deposition rate is illustrated in Fig. 10.63(a).

Question: For both PE- and O_3-TEOS processes, the maximum deposition rate can be achieved at about 250 °C. Why do ILD TEOS processes normally operate about 400 °C, and ILD0 and STI processes deposit at even higher temperatures (~550 °C)?

Answer: At higher deposition temperatures, higher film quality and better step coverage can be achieved, as well as void-free gap fill. In these cases, the concern of yield outweighs the concern of throughput.

10.9.4 Troubleshooting guide

It is impossible to learn troubleshooting skills solely from classroom training and reading books. Troubleshooting skills must be learned by hands-on experience. However, a troubleshooting guide that outlines previous fab experiences can be very helpful to engineers and technicians on the job, and can also benefit those who plan on working in the semiconductor industry in the future.

Processing problems can happen when something in the process goes wrong either suddenly or gradually; often a process can fail, then return to normal, then gradually go wrong again. In those cases, an investigation should be performed to find out what has been changed between the good process and the bad process, or vice versa. For instance, a PE-TEOS process operated normally for a while, then suddenly the deposition rate seemed decreased because measurement results showed the film thickness to be thinner than the normal value. In this case, the metrology tool should be checked first to make sure that the system constants used for measurements are the correct ones. If someone measured the nitride film thickness with that tool and did not change the refractive index back to PE-TEOS oxide, the measured oxide film thickness would be significantly thinner than it actually is, since refractive index and film thickness are always coupled together. If nothing is wrong with the metrology tool, then check whether the process recipe has been changed. Inappropriate changes in processing recipes, such as shorter deposition time, lower TEOS flow rate, and higher wafer temperature, will cause the same kind of problem in thickness measurement.

If a process always has problems at the end of each shift, someone should work cross-shifts to find out what is being changed during the shift change. Some part of the process must be different between the shifts; that change puts the process back to normal at the start of the shift. Most likely that difference is also the source of the problem, as the process gradually goes wrong at end of the shift.

If a process always has problems at the beginning of each shift and gradually returns back to normal after a few wafer depositions, check whether the tool idle time is too long. Long idle times can cause the first-wafer effect, especially for

TEOS processes. The first-wafer effect is mainly due to the cooler processing kit temperature, which causes higher viscosity of the TEOS vapor flow and lower deposition rates. Thus, running a seasoning deposition and plasma cleaning cycle process during wafer loading and transfer can heat up the processing kits and effectively reduce the first-wafer effect for the TEOS process.

Since most dielectric CVD processes operate in the mass-transport-limited regime, deposition rate is mainly determined by gas flow rate, especially the flow rate of silicon source gases such as silane and TEOS. Therefore, it is likely that the deposition rate problems are related with silane or TEOS flow.

For PECVD film, stress is mainly determined by ion bombardment. Therefore, if the film stress is not of the desired value (~100 MPa = 10^9 dyne/cm^2, compressive), anything related to ion bombardment should be on the checklist, most notably rf power system and wafer grounding. Gas flow rate and pressure also can affect film stress.

Refractive indices for PE-TEOS and O_3-TEOS processes are very stable, 1.46 and 1.44, respectively. For silane processes, the refractive index can vary in a wide range, depending mainly on the chemical composition of the film. If the refractive index is out of specification range, the gas flow ratio (such as SiH_4/N_2O or SiH_4/NH_3) should be checked first. For the nitride process, a chamber leak can cause low refractive index.

Uniformity is determined by flow pattern. If a nonuniformity pattern is thin or thick at the center, it can by adjusted by varying the spacing of the wafer and showerhead, or by changing the carrier gas flow rate, such as helium flow for TEOS processes and nitrogen flow for nitride process. The relationship between spacing and profile is illustrated in Fig. 10.68. If the nonuniformity is side to side, it could be a wafer leveling or centering problem. Valve leaks can also cause side-to-side nonuniformity. Tests with different wafer orientations can help confirm or eliminate this possibility.

If the particle counts are gradually increasing and out of specification range, check whether the chamber needs a scheduled wet clean. For particle

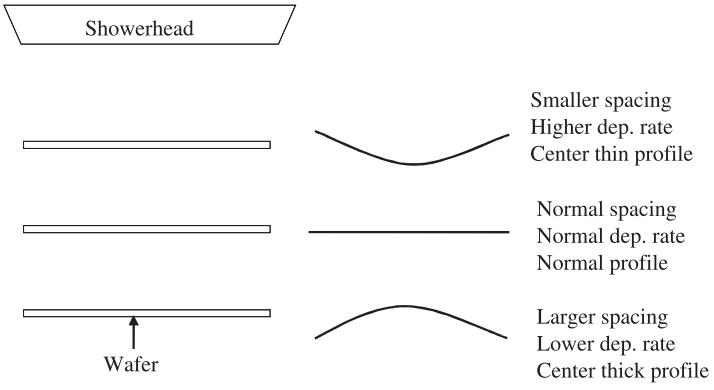


Figure 10.68 Wafer spacing and film thickness profiles.

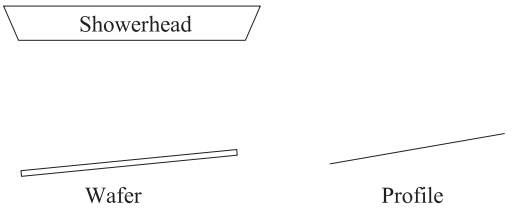


Figure 10.69 Wafer position and film thickness profile.

contamination problems, a partition method should be employed. At first, wafers should be run through the loading, transfer, and processing chamber one by one to check whether it is a mechanical particle problem. Then gases should be flowed into each processing chamber individually to check whether there is a dusted gas line. After eliminating all possible mechanical particle sources, the deposition process can be run to see whether there are process-induced particles. Improper cleaning processes, incorrect pressure, rf power changes, or a leaking process chamber can all cause particle contamination during deposition processes. The possible sources of particles should be eliminated one by one.

Troubleshooting procedures are very tool dependent, since different tools run different processes. The equipment providers normally have hardware and process troubleshooting guides for to each type of tool.

10.10 Recent Developments

In the first decade of the twenty-first century, IC device development placed new demands on dielectric thin films. The isolation trench became narrower, and aspect ratio became so high (up to 8) that it was no longer shallow. HDP-CVD oxide, O_3-TEOS oxide, flowable CVD, and spin-on oxide were used for void-free STI trench fill. Sidewall spacers are still mainly formed with the nitride/TEOS oxide combination.

As device dimensions continue to grow smaller, the thickness of gate oxides will also decrease. Eventually, the thickness will become too thin to work reliably, even with 1 V of power supply (see Example 3.6 in Chapter 3). Therefore, high-κ dielectric materials have been developed for gate dielectrics starting with the 45-nm technology node. Hafnium oxide (HfO_x) is the most commonly used high-κ gate dielectric, and atomic layer deposition (ALD) has been used to form it.

While gate dielectrics are requiring high-κ materials, the κ-value of ILD layers has gone in the other direction. To further reduce κ-value, PECVD porous low-κ dielectric deposition processes have been developed. The combination of low-κ dielectric and copper metallization reduces the RC delay to meet the demands of increasing IC speed.

The most significant difference between the dual damascene Cu process and the Al–Cu process (Fig. 10.70) is that the dual damascene process does not need to etch metal. Because copper is very difficult to dry etch, dual damascene has become the process of choice for copper metallization. In advanced CMOS ICs,

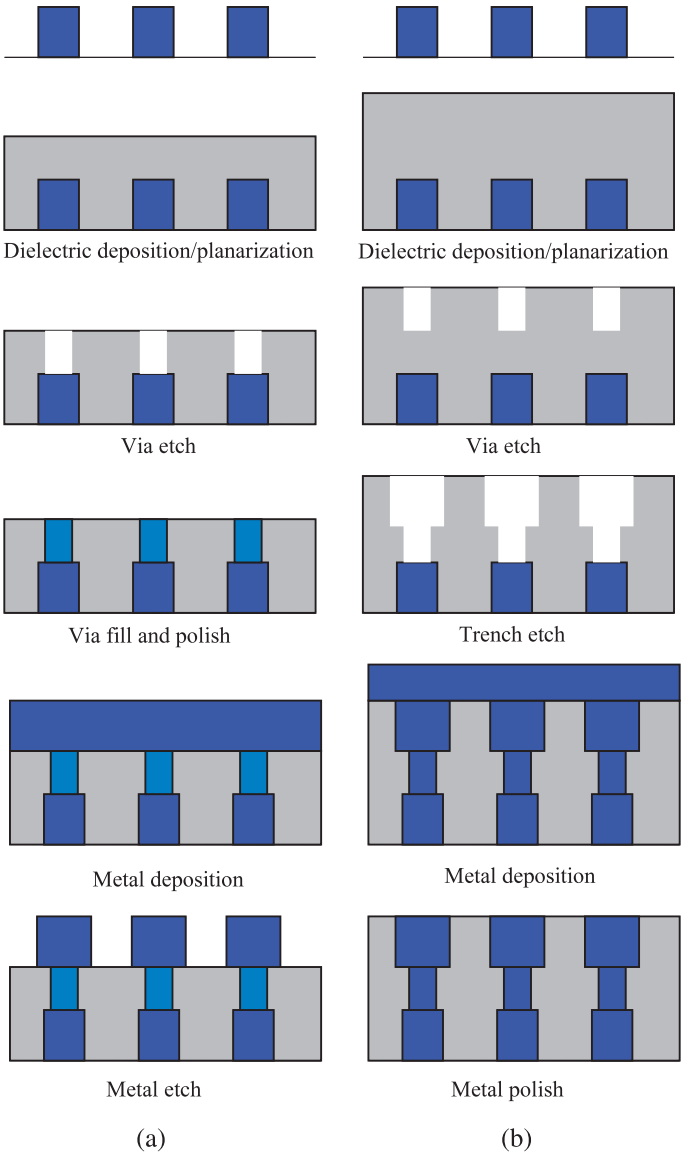


Figure 10.70 (a) Al–Cu interconnection and (b) dual damascene copper interconnection processes.

copper interconnects have already replaced Al–Cu interconnects and become mainstream. Even in memory chips, such as DRAM and NAND flash, copper interconnects are being applied in high-speed products. For Cu interconnects, ILD and passivation dielectric thin-film depositions have become relatively simple, with no more gaps to fill and no more planarization issues. Requirements for those thin-film depositions are high deposition rate, good uniformity, stable low-κ value, and capability of process integration. Of course, for STI and ILD0, there are still the great challenges of filling even narrower gaps and planarizing dielectric layer surfaces.

10.10.1 Low-κ dielectrics

3MS and other silicon-containing organic compounds are used as organosilicate precursors in PECVD processes to deposit low-κ (~2.9 to 2.7) SiCOH film for ILD applications. To further reduce RC delay and improve circuit clock speed, ULK dielectrics based on porous SiCOH (κ = 2.5 to 2.2) have been developed. Using precursors such as DEMS (diethoxymethylsilane) for the base film and a porogen such as CHO (cyclohexene oxide or $C_6H_{10}O$), composite OSG-organic film can be deposited with organic (C_xH_y) pockets. With UV and visible light post-treatment, which outgases the organic and transforms the organic pockets into air pockets, porous ULK dielectric film can be formed for ILD applications in cutting-edge IC chips, as shown in Fig. 10.71. Pore size is approximately a few nanometers.

Dielectric barrier materials with lower κ-values have also been developed for copper/low-κ interconnect applications to replace silicon nitride barrier materials, which have a dielectric constant of κ = 7. Silicon carbon nitride (SiCN) with a κ-value of approximately 4.8 has been developed and used for this application. Amorphous silicon carbide (α-SiC:H) with a κ-value of 2 to 3.6 has also gained much attention as a strong candidate for this application. The α-SiC:H can be deposited with the PECVD process, with tetramethylsilane [4MS or $Si(CH_3)_4$] or trimethylsilane [3MS or $SiH(CH_3)_3$] as a precursor. It can help to further reduce the overall dielectric constant of the ILD layer and improve IC speed.

10.10.2 Air gap

The lowest dielectric constant is 1.0, which can only be achieved in vacuum. Gases usually have dielectric constants just slightly higher than 1. At an atmospheric pressure of 1, the dielectric constant of air is 1.00059, which is the lowest κ-value achievable in ILD for metal interconnection of an IC chip.

There are two approaches to forming air gaps, one is void formation with nonconformal PECVD film, and the other uses sacrificial material between metal wires. Void formation is discussed in earlier sections of this chapter.

Figure 10.72 is an example of void-formed air gaps in a 25-nm NAND flash device. Figure 10.72(a) shows the air gaps between WLs of the 25-nm NAND flash, and Fig. 10.72(b) shows the air gaps between BLs of the same device. By using air gaps, the parasitic capacitance between WLs and BLs have been reduced by 25 and 30%, respectively.

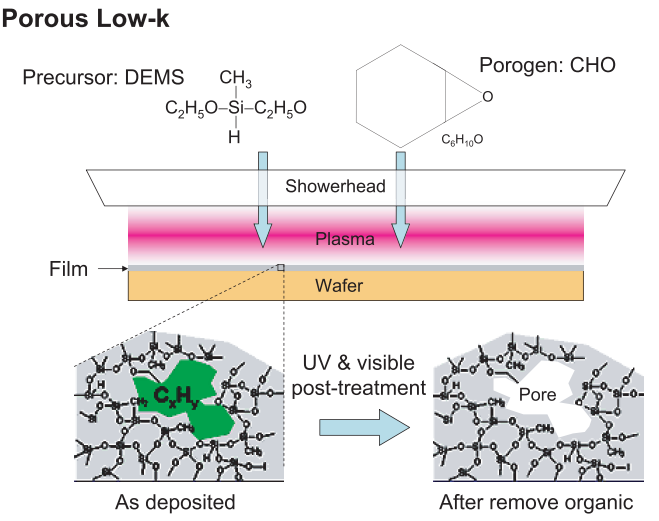


Figure 10.71 Deposition and post-treatment for forming porous low-κ dielectric film.

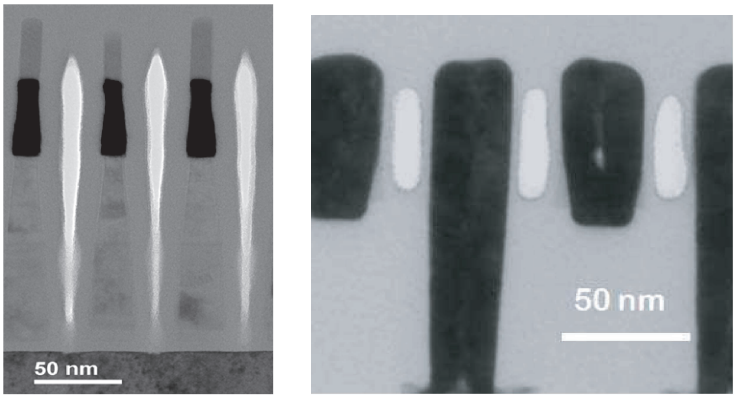


Figure 10.72 Cross-section TEM images of Intel/Micron 25-nm node NAND flash: (a) air gaps between WL and (b) BL (K. Prall and K. Parat, *IEDM*, 102–105, 2010).

The sacrificial method also has several approaches. Figure 10.73 shows one of them. It starts with an ILD stack of ESL/low-κ/ESL/TEOS. After copper CMP, TEOS can be stripped with BOE, as shown in Fig. 10.73(a). A protective liner is deposited, followed by organic coating, curing, and etchback [see Fig. 10.73(b)]. A porous cap layer is deposited, and organic outgassing with UV or visible radiation creates air gaps between the copper lines.

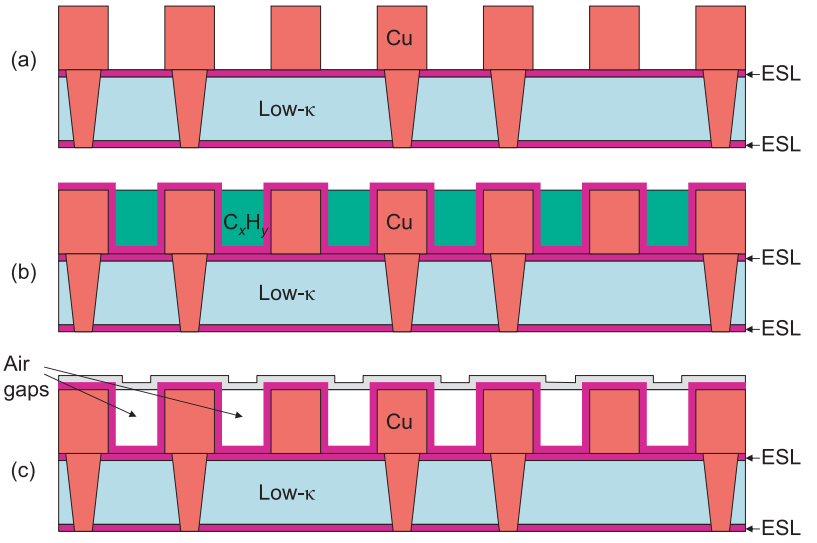


Figure 10.73 Air gap formation with sacrificial layers: (a) after oxide strip, (b) after organic etchback, and (c) after organic outgassing.

10.10.3 Atomic layer deposition

ALD is a deposition process that allows users to deposit a very thin film in the form of atomic layers in multiple deposition circles. ALD can be used to deposit compound semiconductors such as GaAs, InP, GaP, AlN, GaN, InN, etc. It can also be used to deposit high-κ dielectrics, such as Al_2O_3, TiO_2, ZrO_2, HfO_2, Ta_2O_5, La_2O_3, etc. Metallic nitrides, including those used in the metal gates of MOSFETs such as TiN, TaN, Ta_3N_5, NbN, and MoN, also can be deposited with the ALD process.

The ALD process is performed in a sealed reactor. The first processing gas flows into the chamber, and the precursor molecules adsorb on the substrate surface. The chamber is then purged, and the first processing gas is removed from the chamber, leaving behind only those molecules adsorbed on the wafer surface. The second processing gas then flows into the chamber to react with the first process molecules on the surface to form a molecular layer of compound material on the surface. After all of the first precursor molecules are consumed, the chemical reaction is self-terminated, and a purge process removes the second processing gas and byproducts of the chemical reaction from the chamber. The ALD cycle, illustrated (clockwise) in Fig. 10.74 (first precursor, purge, second precursor, and purge), deposits a molecular layer of the compound material onto the wafer surface. Multiple circles can be performed until the required compound film thickness is reached.

HfO_2 is one of the candidates for the high-κ gate dielectrics of MOSFETs. In this case, $HfCl_4$ can be the first precursor, H_2O can be the second precursor, and HCl is the byproduct of the chemical reaction that deposits the HfO_2 on the surface.

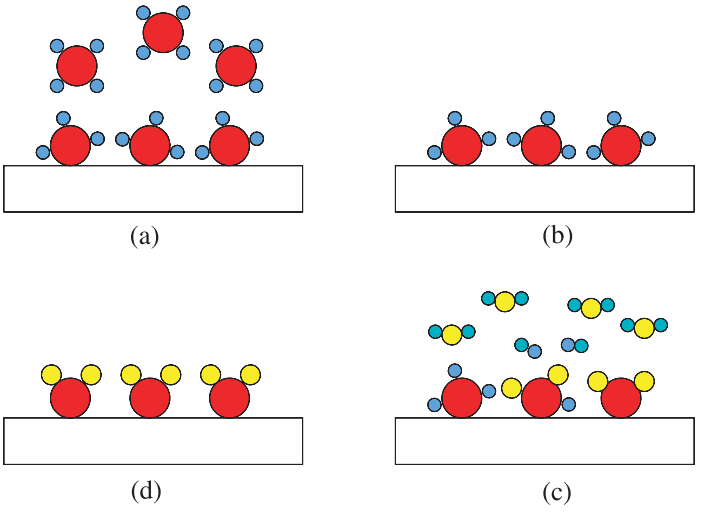


Figure 10.74 An ALD circle: (a) first precursor pulse, (b) purge, (c) second precursor pulse and reaction, and (d) purge.

The chemical reaction can be expressed as

$$HfCl_4 + 2H_2O \rightarrow HfO_2 + 4HCl.$$

The advantages of ALD processes are formation of stoichiometric films with large area uniformity and 3D conformality, and precise control of film thickness. It is possible for ALD to deposit a film at low temperatures, and the deposition process is gentle enough for sensitive substrates. A disadvantage of the ALD process is low throughput and low use of precursors.

Because of low deposition rates (typically less than 10 nm/min compared with ~800 nm/min of PECVD oxide), ALD processes are mainly applied to thin films with thicknesses less than 10 nm. ALD processes can be used to deposit high-κ gate dielectrics of MOSFETs, high-κ dielectrics of DRAM storage capacitors, metallic metal compounds of MOSFET metal gates, and some barrier layers of copper/low-κ interconnection.

The high-κ gate dielectric HfO_2 of a MOSFET can be deposited with hafnium tetrachloride ($HfCl_4$), TEMAHf $\{Hf[N(C_2H_5)(CH_3)]\}_4$, or tetrakisethylmethyl-aminohafnium as hafnium precursors, and O_3 or H_2O as the oxygen precursors. High-κ dielectric Al_2O_3 of DRAM storage node capacitors can be deposited with TMA [$Al(CH)_3$ or trimethylaluminium] as an aluminum precursor, and H_2O or O_3 as an oxygen precursor.

10.10.4 High-κ dielectrics

The basic components of a DRAM cell are a transistor and a capacitor. DRAM fabrication processing requires a dielectric material with an even higher dielectric

constant for the capacitor to shrink. Al_2O_3 ($\kappa = 11.5$) and ZrO_2 ($\kappa \sim 25$) have been used in this application in advanced DRAM manufacturing. TiO_2 (κ up to 85), Ta_2O_5 ($\kappa \sim 25$), and $Ba_{0.6}Sr_{0.4}TiO_3$ or BST (κ up to 600) can also be employed for this application. ALD and CVD processes can be used to deposit these high-κ dielectrics for DRAM capacitor applications.

The thickness of gate oxide in 65-μm devices is about 15 Å; this is too thin to withstand any future scaling down without significantly increasing gate leakage due the quantum tunneling effect. By switching to materials with dielectric constants significantly larger than the 3.9 of silicon dioxide, gate dielectric thickness can be increased, while the effective oxide thickness (EOT) can be further reduced without excessive gate leakage. Gate dielectrics of MOSFETs made a dramatic change starting at the 45-nm technology node when Intel introduced high-κ dielectric gate oxides and metal gate electrodes in 2007. The industry grew away from thermally grown silicon dioxide, which had been used for about 50 years as a gate dielectric, and adapted hafnium-based high-κ gate dielectrics. ALD processes are used to deposit HfO_2 for this application.

To conduct electric current between the S/D, a conductive channel must be formed underneath the gate by reversion of the minority carrier there. Minority carrier reversion can be achieved by charging a metal–oxide interface with the applied gate voltage. Therefore, capacitance of an MOS gate capacitor must be large enough to hold the charges. When feature size reduces, the area of the capacitor electrode reduces quickly, as does the gate capacitance. To retain enough capacitance, the distance between the two electrodes (which is the thickness of the gate dioxide layer) must be reduced. For 0.18-μm IC chips, the gate oxide is about 35 Å. It will be approximately 25 Å for 0.13-μm minimum-feature-size IC chips and about 15 Å for 90-nm devices. A problem will occur when the gate oxide thickness further reduces, as gate-to-substrate leakage will significantly increase, and IC chip reliability will be affected. Nitridated silicon oxide with $\kappa \sim 5$ has been used in 90-, 65-, 45-, and even 32/28-nm devices. High-κ dielectrics such as hafnium dioxide (HfO_2, $\kappa \sim 25$) have been developed and implemented in HKMG CMOS manufacturing, starting from the 45-nm technology node. With HfO_2 as the gate dielectric, gate dielectric thickness can be increased significantly, and gate leakage can be reduced exponentially. Gate capacitance is determined by EOT, which can be expressed as:

$$\text{EOT} = \frac{\kappa_{SiO_2}}{\kappa_x} t_{ox}.$$

Here, $\kappa_{SiO_2} = 3.9$ is the dielectric constant of silicon dioxide, which is used as the standard κ-value in the IC industry to determine high- or low-κ of dielectric materials. X can be HfO_2 or another of the high-κ dielectrics listed in Table 10.11. Gate leakage is determined by the gate dielectric thickness, t_{ox}. The high-κ gate dielectric allows fabs to use thicker gate dielectric layers to reduce gate leakage while still keeping the gate capacitance large enough for device performance. It allows fabs to continue traditional scaling, which requires reducing EOT.

Many other high-κ dielectrics such as titanium dioxide (TiO_2, κ ~ 80), tantalum pentoxide (Ta_2O_5, κ ~ 26), and zirconium dioxide (ZrO_2, κ ~ 25) have also been researched and could be used in IC chip manufacturing in the future. Figure 10.75 shows suitable elements with oxides that could be used for gate dielectrics of silicon MOSFETs. Table 10.11 lists some dielectrics and their κ-values.

10.11 Summary

- Applications of dielectric thin film are STI, high-κ gate dielectrics, sidewall spacers, ILD, and passivation dielectrics; ILD applications are most prevalent.

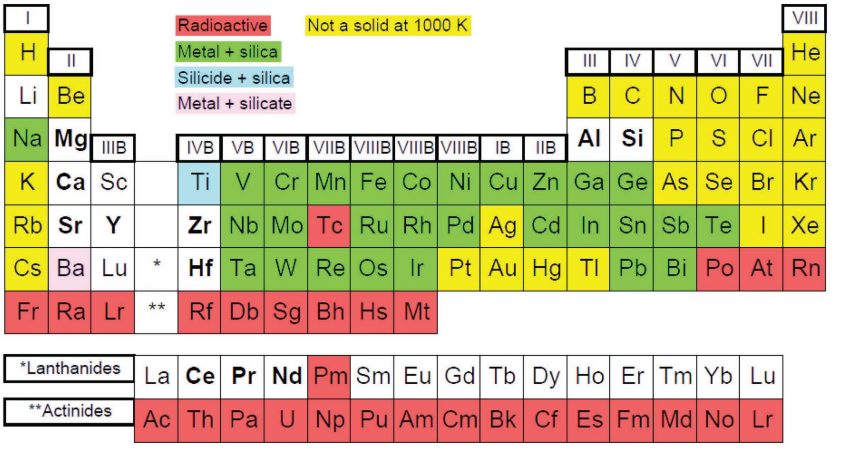


Figure 10.75 Elements in the table with bold letters have oxides that are suitable for gate dielectric applications of silicon MOSFETs. Elements in a shaded box are not suitable due to radioactivity or instability of their oxides on silicon substrates (Osten et al.).

Table 10.11 Dielectric materials and their κ-values.

Material	*k*
SiO_2	3.9
Si_3N_4	7.8
Y_2O_3	15
La_2O_3	30
Pr_2O_3	31
Dy_2O_3	14
Ta_2O_5	26
TiO_2	80
HfO_2	25
ZrO_2	25
Gd_2O_3	12–13
CeO_2	26
Sc_2O_3	>10
BST	>200

- Silicon oxide and silicon nitride are the two most commonly used dielectric materials in ILD applications.
- Low-κ and porous ULK dielectrics are widely used in advanced Cu interconnections.
- Basic CVD processing sequence is as follows: introduction of precursor, precursor diffusion and adsorption, chemical reaction, gaseous byproducts desorption, and diffusion.
- There are two deposition regimes: surface-reaction-limited and mass-transport-limited. Most dielectric CVD reactors are designed to operate in the mass-transport-limited regime because it is less sensitive to temperature variations.
- Film step coverage and conformality are determined by the arriving angle and surface mobility of precursors.
- ILD0 uses PSG or BPSG; deposition and reflow temperatures are determined by the thermal budget.
- ILD mainly uses USG, FSG, SiCOH, and porous SiCOH, depending on the process technology nodes. Deposition temperature is limited by the presence of metal interconnections.
- A passivation dielectric is mainly formed with silicon oxide/nitride stack films.
- Silane and TEOS are the two silicon sources most commonly used for dielectric CVD processes.
- 3MS is widely used for low-κ OSG deposition.
- O_2, N_2O, and O_3 are the most frequently used oxygen precursors for silicon oxide depositions.
- NH_3 and N_2 are the most common nitrogen source gases for nitride depositions.
- Fluorine chemistry is used for dielectric CVD chamber dry cleaning. CF_4, C_2F_6, C_3F_8, and NF_3 are commonly used fluorine source gases.
- The argon sputtering etch process is widely used along with the dielectric CVD process for gap-fill applications. The CF_4/O_2 etchback process is used for dielectric layer planarization.
- Both ellipsometry and prism couplers can be used for measurement of a film's refractive index, which can provide information about the film's chemical composition and physical condition.
- Both ellipsometry and spectroreflectometry can be used for dielectric thin-film thickness measurement. A correct refractive index must be used for an accurate thickness measurement.
- Compressive stress (~100 MPa) is favored for dielectric thin films.
- Most dielectric CVD processes are operated in the mass-transport-limited regime, and the deposition rate is mainly controlled by the silicon source gas flow rate.
- Radio frequency power can be used to control film stress and WERR for PECVD dielectric thin films. The higher the rf power, the more compressive the film stress, and the lower the WERR.
- Film refractive indexes are mainly controlled by the flow rate ratio for silane processes and are determined by the film density for TEOS processes.

- HDP-CVD processes uses silane and oxygen to deposit oxide, and use argon for in-situ sputtering to achieve high-aspect-ratio gap fill.
- ICP and ECR are the two most commonly used HDP sources in the semiconductor industry.
- HfO_2 has been used as high-κ gate dielectric for advanced HKMG MOSFETs.
- High-κ gate dielectrics and metal gate electrodes of HKMG MOSFETs usually are deposited by ALD processes.
- ALD has excellent film uniformity and step coverage. It can control film composition and film thickness very well. However, its deposition rate is very low, thus it is limited to ultrathin-film applications.

10.12 Bibliography

A. C. Adams and C. D. Capio, "Planarization of phosphorus-doped silicon dioxide," *J. Electrochem. Soc.* **128**, pp. 423–429 (1981).

C. Bencher, C. Ngai, B. Roman, S. Lian, and T. Vuong, "Dielectric antireflective coatings for DUV lithography," *Solid State Technol.* **40**(3), pp. 109–114 (1997).

K. Cherkaoui, M. Modreanu, A. Negara, P. Hurley, A. Groenland, and S. H. Lo, "High dielectric constant (high-κ) materials for future CMOS processes," *Electron Dev. Lett.* **18**, 209 (1997).

C. C. Chiang, M. C. Chen, C. C. Ko, Z. C. Wu, S. M. Jang, and M. S. Liang, "Physical and barrier properties of plasma-enhanced chemical vapor deposition α-SiC:H films from trimethylsilane and tetramethylsilane," *Japan. J. Appl. Phys.* **42**, pp. 4273–4277 (2003).

"Dielectric PE- and SA-CVD Processes," training manual of Applied Materials, Inc. Santa Clara, CA (1997).

L. Favennec, V. Jousseaume, V. Rouessac, J. Durand and G. Passemard, "Ultra low κ PECVD porogen approach: matrix precursors comparison and porogen removal treatment study," *Mater. Res. Soc. Symp. Proc.* **863**, B3.2.1 (2005).

H. D. B. Gottlob, T. Echtermeyer, M. Schmidt, T. Mollenhauer, J. K. Efavi, T. Wahlbrink, M. C. Lemme, M. Czernohorsky, E. Bugiel, A. Fissel, H. J. Osten, and H. Kurz, "0.86-nm CET gate stacks with epitaxial Gd2O3 high-κ dielectrics and FUSI NiSi metal electrodes," *IEEE Electron Dev. Lett.* **27**(10), pp. 814–816 (2006).

E. Korczynski, "Low-k dielectric costs for dual-damascene integration," *Solid State Technol.* **42**(5), pp. 43–51 (1999).

H. J. Osten, J. P. Liu, P. Gaworzewski, E. Bugiel, P. Zaumseil, "High-κ gate dielectrics with ultra-low leakage current based on praseodymium oxide," *IEDM Tech. Dig.*, pp. 653–656 (2000).

H. J. Osten, J. Dabrowski, H. J. Müssig, A. Fissel, V. Zavodinsky, "High-κ dielectrics: the example of Pr_2O_3," in *Challenges in Process Simulation*, J. Dabrowski and E. R. Weber, Eds., pp. 259–293, Springer Verlag, Berlin (2004).

J. E. J. Schmitz, *Chemical Vapor Deposition of Tungsten and Tungsten Silicides For VLSI/ULSI Applications*, Noyes Publications, Park Ridge, NJ (1992).

S. Sivaram, *Chemical Vapor Deposition Thermal and Plasma Deposition of Electronic Materials*, Van Nostrand Reinhold, New York (1995).

S. M. Sze, *VLSI Technology*, 2nd ed., McCraw-Hill, New York (1988).

D. Wang, S. M. Chandrashekar, S. P. Beaudoin and T. S. Cale, "Nonuniformity in CMP Processes Due to Stress," 2nd Intl. Conf. Chem.-Mech. Polish Plan. ULSI Multi. Intercon., Santa Clara, CA (1997).

D. L. W. Yen and G. K. Rao, "SOG without etchback," *Proc. VMIC*, pp. 85 (1988).

See http://www.webelements.com for further information on the elements in this chapter.

10.13 Review Questions

1. Describe the CVD process sequence. What are the major differences between CVD and PVD?
2. List at least three CVD dielectric thin-film layers in an IC chip.
3. What is the basic difference between a thermally grown oxide and a CVD oxide?
4. Of APCVD, LPCVD, PECVD, and HDP-CVD processes, which one has the highest processing pressure? Which has the lowest?
5. Describe the three deposition regimes at different temperatures. What is the relationship between deposition rate and temperature in each regime?
6. List the most commonly used silicon source gases for dielectric CVD.
7. List three commonly used oxygen precursor gases.
8. List three precursor gases used for silicon nitride deposition.
9. What type of chemistry is used for dielectric CVD chamber plasma cleaning?
10. Explain why compressive stress is favored for ILD films.
11. For PECVD processes, how does the dielectric film stress change when rf power increases?
12. List two reasons that PSG is used for ILD0. What is the reason for using BPSG?
13. List at least two applications of PE-TEOS USG film.
14. Why are both silicon oxide and nitride used as a passivation dielectric?
15. What are the differences between HDP-CVD and PECVD processes?
16. List two commonly used HDP sources.
17. If the silane flow rate is increased in a PECVD silane oxide deposition process, what are the effects on deposition rate, refractive index, and film stress?

18. Increasing temperature has different effects on deposition rates for PECVD silane USG and PE-TEOS USG processes. Explain.
19. Using the philosophy behind Fig. 10.47, predict the changes for OSG (SiCOH) film properties (such as dielectric constant κ and film stability) after increasing carbon concentration in the film.
20. What is the purpose of porogen in porous low-κ dielectric deposition?
21. What are the main advantages and disadvantage of ALD compared to CVD?
22. What materials can be used as precursors for HfO_2 deposition?
23. What is the standard κ-value when discussing high- and low-κ dielectrics?
24. Low-κ dielectrics are mainly applied in ILDs of metal interconnects. What about high-κ dielectrics?
25. How is stress controlled in an ILD0 stress liner? Which MOSFET needs tensile stress and which one needs compressive stress?

Chapter 11
Metallization

Various types of conductors are applied in IC chip manufacturing. Metals with high conductivity are widely used for the interconnections that form microelectronic circuits. Metallization is an adding process that deposits metal layers on a wafer surface.

Objectives

After finishing this chapter, the reader will be able to:

- explain the device application of metallization processes
- list at least five metals commonly used in IC chip manufacturing
- describe the advantages of copper interconnections over Al–Cu interconnections
- list three different methods used in metal deposition
- describe the sputtering deposition process
- explain the purpose of high vacuum in metal deposition processes
- list the metals used in high-κ, metal gate MOSFETs.

11.1 Introduction

Metals such as copper and aluminum are very good conductors. They are widely used to make conducting lines for transporting electrical power and electric signals. On an IC chip, miniature metal lines are used to connect billions of transistors made on the surface of a semiconductor substrate.

Requirements for metallization are: low resistivity for low power consumption and high IC speed, smooth surfaces for high-resolution patterning processes, high electromigration resistance for achieving high chip reliability, and low film stress for good adhesion to underlying substrates. Other requirements are stable mechanical and electrical properties during subsequent processing, good corrosion resistance, and relatively easy deposition, etch, or CMP.

It is very important to reduce resistance of the interconnect lines because IC device speed is closely related to the RC delay time, which is proportional to the resistivity of the conductor used to form the metal wires. Metal lines with low resistivity have shorter RC times and faster clock speed.

Although copper (Cu) has lower resistivity than aluminum (Al), technical difficulties such as adhesion, diffusion problems, and difficulties with dry etching

hampered its applications in IC chip manufacturing. Al interconnections dominated metallization applications for a long time. In the 1960s and 1970s, pure Al and Al–Si alloys were used as metal interconnection materials. In the 1980s, when device dimensions shrank, one layer of metal interconnection was no longer enough to route all of the transistors, and multilayer interconnections became popular. Soon there was no room for widely tapered contact and via holes when device density increased. Metallization on tapered contact holes is not good for multilayer interconnections because it leaves a very rough surface topography. It is difficult to uniformly deposit a thin-film layer over a rough surface, and it is difficult to precisely pattern the rough layer. To increase packing density, near-vertical contact and via holes are needed; these holes are too narrow and have an aspect ratio that is too high for PVD Al alloys to fill without voids. Tungsten (W) was introduced and became widely used to fill the contact and via holes, and served as a plug to connect the different metal layers. Titanium (Ti) and titanium nitride (TiN) barrier/adhesion layers are required prior to tungsten deposition to prevent diffusion and film peeling of the tungsten. Figure 11.1 illustrates a cross section of a CMOS IC chip with an Al–Cu alloy interconnection and tungsten plug.

Question: Can all dimensions of a metal line be reduced according to the shrinkage of minimum feature size to connect every transistor in one layer of metal?

Answer: Because metal line resistance $R = \rho l/wh$, if a metal line shrinks in every dimension (length l, width w, and height h) according to the shrinkage of the device dimensions by a factor of 1.4 (on the technology node), metal line resistance R will increase by factor of 1.4. This shrinkage will affect device performance and slow down circuit speed due to increasing power consumption and heat generation. Line resistance will be the same if only the width and length of the metal line are reduced while keeping the height unchanged. However, keeping height unchanged will cause high-aspect-ratio metal stacks, which are difficult to etch. It also forms narrow gaps, which are hard to fill without voids for dielectric deposition processes.

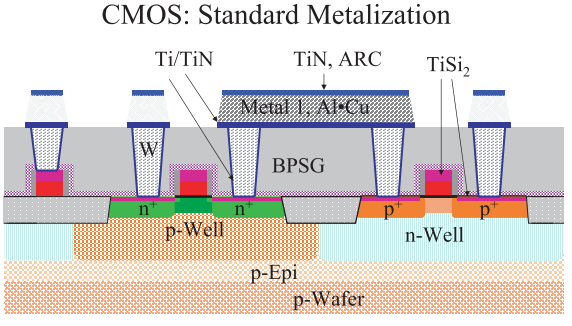


Figure 11.1 Cross section of an IC chip with Al–Cu interconnection.

In the late 1990s, development of the CMP process paved the way for using copper in IC interconnection applications with damascene or dual damascene processes. Tantalum (Ta) or tantalum nitride (TaN) is used as the barrier layer, and silicon nitride (SiN) or silicon carbon nitride (SiCN) has been used as barrier and cap layers to isolate the copper layer and prevent it from diffusing through the low-κ dielectric layer into the silicon substrate, causing interference with the normal operations of the transistors. Silicon nitride is also used as an etch-stop layer for the dual damascene dielectric etching process. Figure 11.2 illustrates a CMOS chip cross section with a copper interconnection.

Figure 11.3 illustrates the position of metallization in IC manufacturing processes. This chapter covers the metals used in IC chip manufacturing, their applications, and their deposition processes.

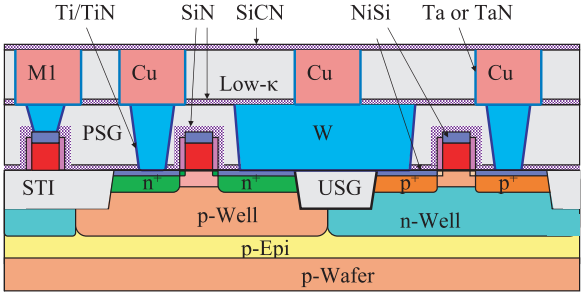


Figure 11.2 Cross-section of an IC chip with Cu/low-κ interconnection.

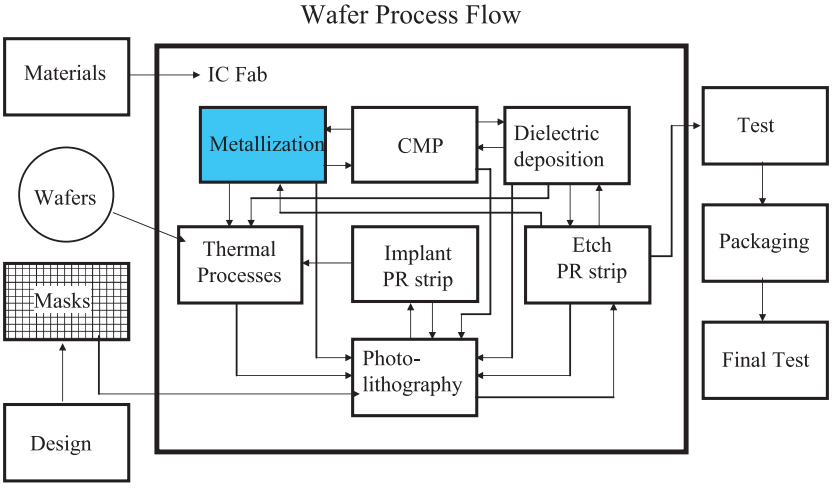


Figure 11.3 Metallization in an IC fabrication process flow.

11.2 Conducting Thin Films

Polysilicon, silicides, aluminum alloys, titanium, titanium nitride, and tungsten are the most commonly used conductors in IC chip manufacturing for gate materials, and for location and global interconnections. Copper has been used for interconnections in IC production since the late 1990s.

11.2.1 Polysilicon

Polysilicon is the most common material used to form gates and local interconnections. Polysilicon has replaced aluminum as gate material since the introduction of ion implantation technology in the mid-1970s. Polysilicon has high temperature stability, which is necessary for self-aligned S/D implantation and postimplantation high-temperature annealing processes. An aluminum gate cannot sustain the high temperature (>1000 °C) required for postimplantation annealing processes.

Polysilicon normally is deposited with an LPCVD process; either SiH_4 or SiH_2Cl_2 can be used as a silicon precursor. Deposition temperatures range from 550 to 750 °C, and the polysilicon can be heavily doped with boron, phosphorus, or arsenic, either *in situ* during deposition or *ex situ* by an ion implantation process. The polysilicon deposition process is discussed in detail in Chapter 5.

11.2.2 Silicides

Even heavily doped polysilicon has a high resistivity of approximately a few hundred μΩ·cm. When device dimensions shrink, resistance of the polysilicon local interconnection will also increase, which causes more power consumption and longer RC time delay. To reduce the resistance and increase device speed, silicides (which have much lower resistivity than polysilicon) have been developed and used along with polysilicon to form a polycide stack. Titanium silicide ($TiSi_2$) and tungsten silicide (WSi_2) are the two commonly used silicides in IC devices that use mature technologies.

Titanium silicide is formed in a self-aligned silicide (salicide) process. First, the wafer surface is cleaned in chemical solutions to remove contaminants and particles. Then, argon sputtering in a vacuum chamber removes the native oxide from the wafer surface. A layer of titanium is deposited by a sputtering process on the wafer surface, with silicon exposed at the S/D and on top of the polysilicon gate. A thermal annealing process, performed in a rapid thermal processing (RTP) system, allows titanium to chemically react with silicon to form titanium silicide on the top of polysilicon and on the surface of the S/D. Because titanium does not react with silicon dioxide and silicon nitride, silicide is only formed at the places where silicon directly comes into contact with titanium. After a wet etch process strips the unreacted titanium with a mixture of hydrogen peroxide (H_2O_2) and sulfuric acid (H_2SO_4), the wafer is annealed once more to fully silicide the titanium and increase its grain size, improving conductivity and reducing contact resistance.

Tungsten silicide processing is quite a different process. First, a thin film of tungsten silicide is deposited on a polysilicon surface by a thermal CVD process

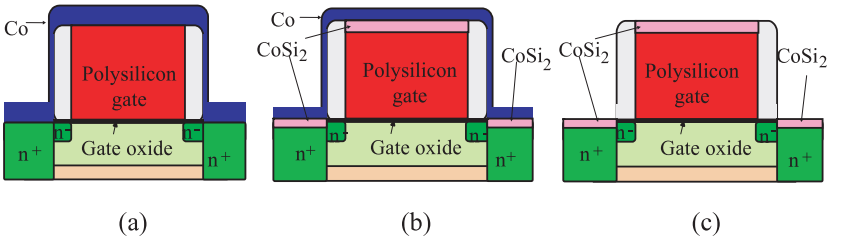


Figure 11.4 Self-alignment cobalt silicide formation process steps: (a) cobalt deposition, (b) silicide annealing, and (c) cobalt wet striping and second annealing.

with WF_6 as the tungsten precursor and SiH_4 as the silicon precursor. Then, a polycide stack is etched in a multistep process in which fluorine chemistry is used to etch the tungsten silicide and chlorine chemistry is used to etch the polysilicon. After photoresist stripping, a RTA process follows to increase tungsten silicide grain size and improve conductivity.

As device dimensions grow smaller, the feature size of a gate can become so small that it is narrower than a titanium silicide grain size, which is about 0.2 μm. Cobalt silicide ($CoSi_2$) has been used in gate and local interconnection applications for CMOS devices from 180- to 90-nm technology nodes. Figure 11.4 illustrates the process steps of self-aligned cobalt silicide formation. Titanium silicide was used before cobalt silicide in earlier technology nodes. Replacing the cobalt (Co) with titanium (Ti), the $TiSi_2$ formation process is almost the same as the steps shown Fig. 11.4.

With shrinking device dimensions, the annealing temperature of $CoSi_2$ (typically ~750 °C) and its annealing time of ~30 sec take away too much of the precious thermal budget. A nickel (Ni) silicide process has been developed and applied to CMOS devices since the 65-nm technology node. Nickel silicide (NiSi) processing is also very similar to the process shown in Fig. 11.4. The main advantage of NiSi processing is that it requires a significantly lower annealing temperature, typically less than 500 °C. NiSi can be used for CMOS devices down to 10-nm technology node.

Because NiSi is thermally unstable, it tends to further react with silicon to form $NiSi_2$, which can cause nickel silicide to grow into the silicon substrate, commonly called NiSi piping or NiSi encroachment, inducing junction leakage and causing yield loss. Many methods have been studied to solve this killer defect. A popular solution is to add platinum (Pt) to a nickel target that can be used to sputter NiPt alloy on the wafer surface to form NiPt silicide.

11.2.3 Aluminum

At one time, aluminum was the most common metal used for interconnections, which route thousands or millions of transistors built on a single wafer surface. Copper interconnections became mainstream technology in the first decade of the twenty-first century. Aluminum is the fourth-best metal for electrical conduction

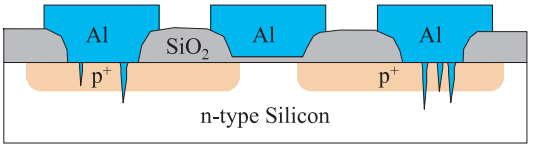


Figure 11.5 Junction spiking.

(2.65 μΩ·cm), following silver (Ag, 1.6 μΩ·cm), copper (Cu, 1.7 μΩ·cm), and gold (Au, 2.2 μΩ·cm). Of these four, it is the only metal that can be easily dry etched to form tiny metal interconnection lines. It was also used as a gate material before the introduction of ion implantation in the mid-1970s.

Silicon will dissolve in aluminum. At the S/D area, where aluminum metal lines come into direct contact with silicon, silicon can dissolve into the aluminum, and the aluminum will diffuse into the silicon to form aluminum spikes. Aluminum spikes can punch through the doped junction and short the S/D junction with the substrate, causing excessive device leakage and reliability issues. This effect is called junction spiking and is illustrated in Fig. 11.5.

The saturation concentration of dissolved silicon in aluminum is about 1%. Therefore, adding about 1% of Si into Al can effectively prevent further dissolving of the silicon. Thermal annealing at 400 °C forms Si–Al alloys at the silicon-aluminum interface that also can help prevent aluminum-silicon interdiffusion and junction spiking.

Metallic aluminum is a polycrystalline material that consists of many small monocrystalline grains. When electric current flows through an aluminum wire, a stream of electrons constantly bombards the grains. The smaller grains will start to move after a certain amount of time, similar to small rocks on the bottom of a creek being flushed down a stream during flood season. This effect is called electromigration. Figure 11.6 shows the electromigration process.

Electromigration can cause serious problems in aluminum wires. When a few grains start to move due to electron bombardment, the metal line begins to tear apart at certain points and causes higher current density in the line at these points. This aggravates the bombardment and causes further aluminum grain migration. The high current density and high resistance generate a large amount of heat and eventually break down the metal line. Electromigration can affect IC chip reliability because it can break the aluminum line and causes an open loop in the tiny circuit after the chip is employed in an electronic system. For old houses with aluminum wiring, this can become a safety hazard because it can cause fire due to the intense heat generated at the tearing point of the aluminum line, normally at the contact point.

It was discovered that when a small amount of copper is added to aluminum to form an Al–Cu alloy, electromigration resistance can be significantly improved, since copper can act like glue between the aluminum grains and prevent them from migrating due to electron bombardment. An Al–Si–Cu alloy was used after this discovery, and it could still be used by some older fabs because IC manufacturers normally are very reluctant to replace a production-proven process.

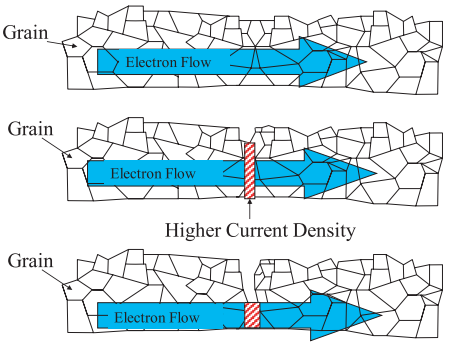


Figure 11.6 Electromigration.

By then end of the 1990s, the most commonly used metal for interconnection was the Al–Cu alloy. Because the aluminum no longer made direct contact with silicon, silicon was no longer required in the aluminum alloy. Copper concentration varied from 0.5 to 4%, depending on the process requirements and manufacturer's specifications. High copper concentration improves electromigration resistance. However, higher copper concentrations can make metal etch processes more difficult. Chapter 9 revealed that the main etchant for Al–Cu alloy etch is chlorine. One of the byproducts of Al–Cu alloy etch is copper chloride, which has a very low volatility and tends to stay on the wafer surface to form unwanted residues, commonly called scum. High copper concentrations in Al–Cu alloy can worsen the scum issue. Heavy ion bombardment in the plasma etch process can help dislodge the nonvolatile copper chloride, or a little isotropic chemical etch could scoop it out from the surface and prevent scum from forming. The etch residues can cause defects on the wafers and affect IC chip yield. An extra wet descumming process is often required to remove these residues.

Both CVD and PVD processes can be used to deposit aluminum. Because PVD aluminum has a higher quality and lower resistivity, it is the most popular method used in the IC industry. Thermal evaporation, electron beam evaporation, and plasma sputtering can be used for Al PVD. Magnetron sputtering deposition is the most popular PVD process for aluminum alloy deposition in semiconductor fabs. Electron beam evaporation and thermal evaporators are no longer used in advanced IC fabs. They are still used by universities and research laboratories for education and research applications. Some less-advanced fabs could still use them for IC chip manufacturing that does not require state-of-the-art technology.

Aluminum CVD is a thermal CVD process with aluminum organic compounds such as dimethylaluminum hydride [DMAH, $Al(CH_3)_2H$] as precursors. Compared with PVD aluminum, CVD aluminum has much better step coverage and gap-fill abilities, making it a very attractive candidate to replace tungsten CVD processes for contact/via hole filling. The quality of CVD aluminum film is poorer and resistivity is higher than with PVD films. It is not very difficult to deposit aluminum in a CVD process. However, it is not easy to deposit Al–Cu alloy film in a CVD

Table 11.1 Facts about aluminum.

Symbol	Al
Atomic number	13
Atomic weight	26.981538
Discoverer	Hans Christian Oersted
Place of discovery	Denmark
Discovery date	1825
Origin of name	From the Latin word alumen, meaning alum
Density of solid	2.70 g/cm^3
Molar volume	10.00 cm^3
Velocity of sound in substance	5100 m/sec
Hardness	2.75
Electrical resistivity	2.65 μΩ·cm
Reflectivity	71%
Melting point	660 °C
Boiling point	2519 °C
Thermal conductivity	235 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	23.1 ×10^{-6} K^{-1}
Etchants (wet)	H_3PO_4, HNO_4, CH_3COOH
Etchants (dry)	Cl_2, BCl_3
CVD precursor	$Al(CH_3)_2H$
Main applications in IC manufacturing	To form Al–Cu alloy in metal interconnections. Used as a bulk gate material in gate-last HKMG of MOSFET.

process, so the applications of aluminum CVD processes are limited. Without copper in the aluminum, electromigration can cause serious problems in device reliability. Some facts about aluminum are summarized in Table 11.1.

11.2.4 Titanium

There are several applications that use titanium: silicide formation, titanium nitridation, wetting layers, welding layer, metal gates, and metal hard masks.

Titanium silicide is commonly used in IC fabrication for chips with 250nm or larger technology modes. It has low resistivity and can be formed on the polysilicon gate and the S/D in a self-aligned silicide (salicide) process. In the salicide process, titanium is deposited on the wafer surface by a sputtering deposition process. It is then annealed in a thermal process to form titanium silicide ($TiSi_2$). Titanium can also be deposited by a CVD process, which uses $TiCl_4$ and H_2 to react and deposit titanium at 650 °C, and can form titanium silicide in situ in the regions where Ti directly contacts with Si.

Titanium is also widely used as a welding layer for tungsten and aluminum to reduce contact resistance. This is because titanium can scavenge oxygen atoms and prevent them from bonding with tungsten and aluminum atoms to form high-resistivity WO_4 and Al_2O_3. It is also used with titanium nitride as a diffusion barrier layer for tungsten plugs and for local interconnection to prevent tungsten from diffusing into the silicon substrate. Figure 11.7 shows an application of titanium and titanium nitride in the interconnection application of an IC chip with Al–Cu interconnection.

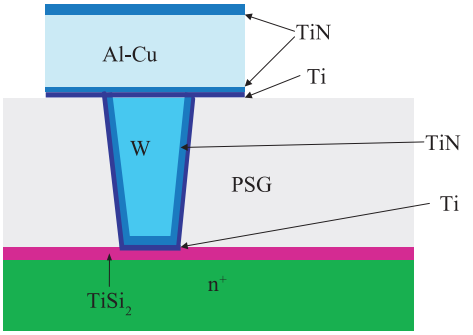


Figure 11.7 Titanium applications of IC chips with Al–Cu interconnection.

Titanium can also be used in titanium nitride formation; by depositing titanium and annealing it at a high temperature in the nitrogen/ammonia ambient, TiN can be formed through a chemical reaction on the surface. Of course, most people prefer PVD titanium nitride for its easy process control and lower temperature requirement.

Titanium normally is deposited with a magnetron plasma sputtering process. It also can be deposited by a CVD process with $TiCl_4$ as the precursor and reacts with hydrogen at high temperatures. Some facts of titanium are listed in Table 11.2.

11.2.5 Titanium nitride

Titanium nitride is widely used in IC manufacturing for barrier layers, adhesion layers, and antireflection coating (ARC). A thin layer (50 to 200 Å) of TiN is also needed for barrier and adhesion layers for tungsten. It prevents tungsten from diffusing into the oxide and silicon, and helps tungsten stick to silicon oxide surfaces. TiN also has a much lower reflectivity than the Al–Cu alloy and has been used as an ARC to improve photolithography resolution in metal patterning processes. A TiN layer on top of the aluminum alloy also can prevent hillocks and help control electromigration.

In advanced HKMG MOSFET, TiN is used as a metal gate electrode. Gate-first HMMG MOSFET uses TiN as a gate electrode for both nMOS and pMOS, as shown in Fig. 11.8(a). Gate-last HKMG MOSFET uses TiN for nMOS and to react with TiAl to form TiAlN for nMOS gate electrodes [see Fig. 11.8(b)]. TiN is also used as gate electrodes of cell array transistors of bWL DRAM.

For copper/low-κ and copper/ULK interconnections (especially for porous ULK dielectrics), TiN has been used as metal hard masks for low-κ dielectric dual damascene etch processes to protect the OSG film from attacking oxygen radicals during photoresist ashing. Otherwise, oxygen radicals can react with methyl or ethyl groups in the OSG film and increase the dielectric constant of the film.

TiN has also been used as the electrodes of storage node capacitors in DRAM chips, which require very good film conformality and excellent step coverage of both sidewalls and bottoms of high-aspect-ratio storage node holes.

Table 11.2 Facts about titanium.

Symbol	Ti
Atomic number	22
Atomic weight	47.867
Discoverer	William Gregor
Place of discovery	England
Discovery date	1791
Origin of name	Named after the Titans, the sons of the earth goddess in Greek mythology
Density of solid	4.507 g/cm^3
Molar volume	10.64 cm^3
Velocity of sound	4140 m/sec
Hardness	6.0
Electrical resistivity	40 μΩ·cm
Melting point	1668 °C
Boiling point	3287 °C
Thermal conductivity	22 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	8.6 ×10^{-6} K^{-1}
Etchants (wet)	H_2O_2, H_2SO_4
Etchants (dry)	Cl_2, NF_3
CVD Precursor	$TiCl_4$
Main applications in IC manufacturing	Wetting layers for contact and Al–Cu metallization. Barrier layer for W. Reacts with nitrogen to form TiN in reactive sputtering.

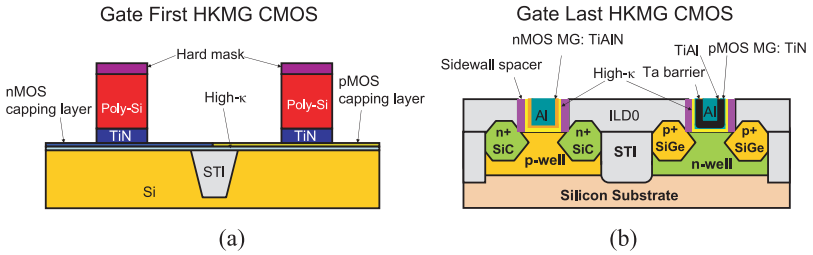


Figure 11.8 Applications of different metal layers in HKMG CMOS: (a) gate-first and (b) gate-last. (MG stands for metal gate.)

Titanium nitride can be deposited by PVD, CVD, and ALD processes. PVD TiN is deposited in a sputtering chamber with a titanium target, which is capable of depositing titanium in situ with TiN. Since a thin layer of titanium (50 to 100 Å) is always required for trapping oxygen and reduces contact resistance before titanium nitride deposition, in-situ Ti/TiN deposition can improve process throughput.

Reactive sputtering is the most commonly used method to deposit TiN in PVD processes. It uses argon and nitrogen as processing gases. In plasma, parts of both gases are ionized, and some nitrogen molecules are dissociated; these molecules generate chemically reactive free nitrogen radicals. Titanium atoms sputtered off from the target surface by the argon ions react with nitrogen when they pass through the argon-nitrogen plasma, and titanium nitride is formed and deposited

on the wafer surface. Some titanium atoms can pass through the plasma and be deposited on the wafer surface, where they react with nitrogen radicals and form a titanium nitride there. Nitrogen radicals can also react with the titanium target and form a titanium nitride layer on the target surface. Ar^+ ions sputter the TiN molecules off the target surface and deposit them on the wafer surface.

CVD processes have been developed and applied to deposit conformal TiN films in IC chip manufacturing. They can be either high-temperature (~700 °C) CVD with precursors of $TiCl_4$ and NH_3, or low-temperature (~350 °C) metal organic CVD (MOCVD) with precursors of tetrakis-dimethylamino-titanium {TDMAT or $Ti[N(CH_3)_2]_4$}. TiN film deposited at a high temperature has better quality, lower resistivity, and better step coverage. However, it cannot be used for via applications, since the processing temperature is so high that it will melt the aluminum lines already on the wafer. MOCVD TiN has much better film conformality and step coverage compared with PVD TiN and has been widely used in advanced CMOS logic IC fabs for contact barrier/glue layer deposition. As deposited, MOCVD TiN has many organic groups integrated in the film. To densify the film, plasma treatment usually follows. N_2-H_2 plasma bombards the TiN film, helping to fully nitridize the film and remove residual methyl groups. MOCVD TiN usually needs several deposition-treatment cycles.

For HKMG applications, especially gate-last approaches, the ALD process has been used to deposit TiN due to its excellent step coverage. The ALD process has also been used for DRAM storage node TiN depositions.

In some old fabs, titanium nitride is still formed by the nitridation of the titanium surface with ammonia in a thermal annealing process.

11.2.6 Tungsten

Tungsten is the metal of choice for filling contact holes and forming the plugs that connect the metal layer and silicon surface. It is also applied to fill via holes between the different metal layers for Al–Cu interconnections, which are used in memory chips and mature (0.25 μm or earlier technology nodes) CMOS logic chips. When IC device dimension shrank to the submicron level, contact holes between the interconnection layers became smaller with higher aspect ratio, and it was impossible for PVD aluminum to fill them without voids. Hence, tungsten CVD (WCVD) processes were developed. With excellent step coverage and gap-fill abilities, it became the film most used to fill high-aspect-ratio contact or via holes, as illustrated in Fig. 11.9. Because WCVD has higher resistivity (8.0 to 12 μΩ·cm) than PVD aluminum–copper alloys (2.9 to 3.3 μΩ·cm), it is only used for local interconnections and plugs for connection between different layers.

Question: Why is a void in a contact hole unacceptable?

Answer: Because a void in a contact hole will cause high contact resistance and high current density due to the small metal cross section in the contact hole. Excessive heat generated there can damage the IC device very quickly.

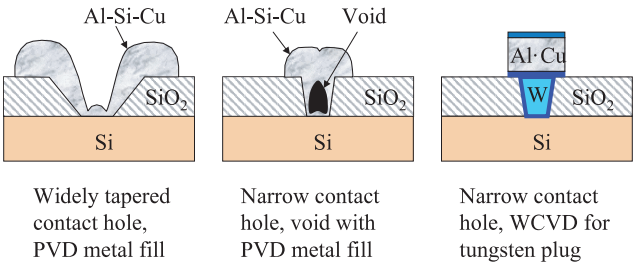


Figure 11.9 Evolution of the contact metallization processes.

Tungsten has been widely used in DRAM chips to form BLs and BL contact plugs. Tungsten or tungsten silicide is also added to polysilicon to form WLs in DRAM. Polysilicon plugs have long been used for DRAM FEoL contacts, such as self-aligned contact (SAC) and storage node contact (SNC). Due to smaller feature sizes, tungsten is being employed to form DRAM FEoL contact plugs. Because most DRAM chip do not need to operate at high clock speeds (2 GHz or higher) as CMOS logic ICs do, some advanced DRAM chips still use Al–Cu interconnections instead of Cu interconnections to reduce fabrication costs. Tungsten plugs are widely used in DRAM metal contact plugs between the metal layer and devices in the periphery area, and in via plugs between the metal layers, which in most cases are limited to two or three layers. For bWL DRAM, tungsten is added to TiN to form WLs and gate electrodes of the array transistors. It is also used with polysilicon to form the BLs and gate electrodes of periphery MOSFETs. NAND flash memory devices use tungsten as well to form BLs and source lines.

Tungsten is deposited during the CVD process, with WF_6 as the tungsten precursor. WF_6 can react with SiH_4 at 400 °C to deposit tungsten, which is used as the nucleation layer. A WF_6 and H_2 reaction at about 400 °C is applied for bulk tungsten deposition, which fills the narrow contact/via holes. Since tungsten does not stick very well to silicon dioxide surfaces, a layer of TiN is always needed to help tungsten adhere to the oxide surface. TiN and Ti stacks can prevent tungsten from diffusing into the silicon dioxide layer and cause heavy metal contamination; it can also block tungsten from contacting the silicon and reacting with it to form tungsten silicide and cause junction loss. Figure 11.10 shows an SEM photograph of a tungsten plug and TiN/Ti barrier/adhesion layer. Some facts of tungsten are summarized in Table 11.3.

11.2.7 Copper

Copper has much lower resistivity (1.7 μΩ·cm) than an aluminum–copper alloy (2.9 to 3.3 μΩ·cm). It also has significantly higher resistance to electromigration. It is highly desirable to use copper in an IC chip as the metal interconnection wire because copper can reduce power consumption and increase IC speed. However, copper has poor adhesion to silicon dioxide. It also has a very high diffusion rate in silicon and silicon dioxide, and copper diffusion can cause heavy metal

Table 11.3 Facts about tungsten.

Symbol	W
Atomic number	74
Atomic weight	183.84
Discoverer	Fausto and Juan Jose de Elhuyar
Place of discovery	Spain
Discovery date	1783
Origin of name	From the Swedish words tung sten, meaning heavy stone. W comes from wolfram, named after the tungsten mineral wolframite.
Density of solid	19.25 g/cm^3
Molar volume	9.47 cm^3
Velocity of sound in substance	5174 m/sec
Hardness	7.5
Reflectivity	62%
Electrical resistivity	5 μΩ·cm
Melting point	3422 °C
Boiling point	5555 °C
Thermal conductivity	170 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	4.5 ×10^{-6} K^{-1}
Etchants (wet)	KH_2PO_4, KOH, and $K_3Fe(CN)_6$; boiling H_2O
Etchants (dry)	SF_6, NF_3, CF_4, etc.
CVD precursor	WF_6
Main applications	Contact plugs for CMOS logic devices. WLs, BLs, and contact plugs for DRAM and flash memory chips.

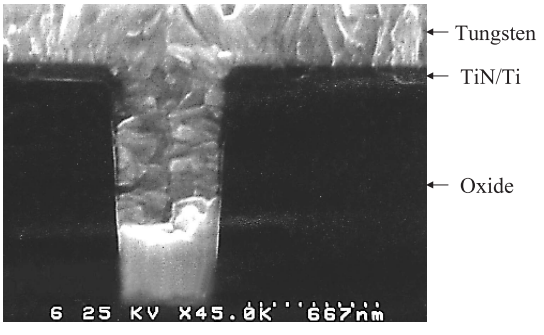


Figure 11.10 Tungsten plug and TiN/Ti barrier/adhesion layer (Applied Materials, Inc.).

contamination and device malfunction. Copper is very difficult to dry etch because copper–halogen compounds have very low volatility. After feature sizes dipped below 3 μm, all patterned etching processes have been dry etch (RIE) processes. Lack of an effective method to anisotropically etch copper became one of the important reasons that copper was not as popular as aluminum before the 0.18-μm technology node. Some basic facts about copper are listed in Table 11.4.

Technology developments, including high-aspect-ratio dielectric contact/via and trench etch processes, barrier layer deposition, copper deposition, and more importantly copper chemical mechanical polishing (CMP), led to the maturity of

Table 11.4 Facts about copper. Note that Cu(hfac)(vtms) stands for hexaflouroacetyl-acetonato-copper-vinyl-trimethylsilane.

Symbol	Cu
Atomic number	29
Atomic weight	63.546
Discoverer	
Place of discovery	Copper has been used by human beings since ancient times, long before any written history.
Discovery date	
Origin of name	From the Latin word cuprum, meaning the island of Cyprus
Density of solid	8.92 g/cm^3
Molar volume	7.11 cm^3
Velocity of sound in substance	3570 m/sec
Hardness	3.0
Reflectivity	90%
Electrical resistivity	1.7 μΩ·cm
Melting point	1084.77 °C
Boiling point	5555 °C
Thermal conductivity	400 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	16.5×10^{-6} K^{-1}
Etchants (wet)	HNO_4, HCl, H_2SO_4
Etchants (dry)	Cl_2, needs low pressure and high temperature
CVD precursor	Cu(hfac)(vtms)
Main applications in IC manufacturing	Main metal for interconnections

dual damascene metallization technology. The fact that the dual damascene process does not require a metal etch step paved the way for copper metallization of IC interconnections in the late 1990s.

Copper deposition requires two steps: deposition of barrier and seed layers that cover the sidewall and bottom of the trenches and via holes, and bulk layer deposition by electrochemical plating (ECP) that completely fills the via holes and trenches. An annealing process follows bulk copper deposition to increase grain size and improve electrical conductivity before the metal CMP process, which removes copper and the barrier metal layer from the wafer surface, leaving the copper inside the trenches and via holes to form metal interconnects.

11.2.8 Tantalum

Tantalum is used as a barrier layer prior to copper deposition to prevent copper from diffusing across silicon oxide into the silicon substrate and causing device damage. Tantalum is a better barrier material for copper than other barrier materials such as titanium and titanium nitride. Tantalum is deposited with a sputtering process. Some facts about tantalum are listed in Table 11.5.

11.2.9 Cobalt

Cobalt is mainly used to form cobalt silicide ($CoSi_2$) for CMOS logic devices with 180- to 90-nm technology nodes. Cobalt silicide is still used for advanced flash memory devices and is deposited with a sputtering process. Some facts about cobalt are listed in Table 11.6.

Table 11.5 Facts about tantalum.

Symbol	Ta
Atomic number	73
Atomic weight	180.9479
Discoverer	Anders Ekeberg
Place of discovery	Sweden
Discovery date	1802
Origin of name	From the Greek word Tantalos, meaning father of Niobe due to its close relation to niobium in the periodic table.
Density of solid	16.654g/cm^3
Molar volume	7.11 cm^3
Velocity of sound in substance	3400 m/sec
Hardness	3.0
Reflectivity	90%
Electrical resistivity	12.45 μΩ·cm
Melting point	2996 °C
Boiling point	5425 °C
Thermal conductivity	57.5 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	6.3×10^{-6} K^{-1}
Etchants (wet)	2:2:5 mixture of HNO_3, HF, and H_2O
Etchants (dry)	Cl_2
Main applications in IC manufacturing	Used as a copper barrier layer. Also used as a barrier layer in pMOS of gate-last HKMG MOSFETs to block TiAl from reacting with TiN. TaN is also used as a copper barrier layer. TaBN is widely used as an absorber of EUV masks.

11.2.10 Nickel

Nickel (see Table 11.7) is required to form nickel silicide (NiSi) for CMOS logic devices with 65-nm and smaller technology nodes because NiSi formation needs significantly lower temperatures than other salicide formations. This makes it suitable for devices with smaller feature sizes and lower thermal budgets. It is deposited by means of a sputtering process.

Nickel is a magnetic material. Nickel–iron (NiFe) alloys can be used to form a magnetic memory cell in magnetoresistive random access memory (MRAM).

11.3 Metal Thin-Film Characteristics

Thickness measurements of metal thin films are different from dielectric thin-film measurements. It has been very difficult to directly and precisely measure the thickness of opaque thin films, such as metal films; the measurement needed to be performed on a test wafer in a destructive way. Recently, some nondestructive methods, such as laser acoustic measurements and x-ray reflectometry, have been developed and applied in metal thin-film metrology and processing controls.

Conducting films have polycrystalline structures. A metal's conductivity and reflectivity are related to grain size; metal films with larger grain sizes have higher conductivity and lower reflectivity. These properties are related to the deposition process. For instance, a higher deposition temperature provides higher mobility to deposited atoms on the substrate surface and forms larger grains in the deposited

Table 11.6 Facts about cobalt.

Symbol	Co
Atomic number	27
Atomic weight	180.9479
Discoverer	Georg Brandt
Place of discovery	Sweden
Discovery date	1735
Origin of name	From the German word kobald, meaning goblin or evil spirit
Density of solid	8.900 g/cm^3
Molar volume	6.67 cm^3
Velocity of sound in substance	4720 m/sec
Hardness	6.5
Reflectivity	67%
Electrical resistivity	13 μΩ·cm
Melting point	1768 K or 1495 °C
Boiling point	3200 K or 2927 °C
Thermal conductivity	100 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	13.0 ×10^{-6} K^{-1}
Etchants (wet)	H_2O_2 and H_2SO_4
Etchants (dry)	-
Main applications in IC manufacturing	To form cobalt silicide for salicide applications.

Table 11.7 Facts about nickel.

Symbol	Ni
Atomic number	28
Atomic weight	58.693
Discoverer	Axel Fredrik Cronstedt
Place of discovery	Sweden
Discovery date	1751
Origin of name	From the German word kupfernickel, meaning devil's copper or Saint Nicholas' (old Nick's) copper
Density of solid	8.908 g/cm^3
Molar volume	6.59 cm^3
Velocity of sound in substance	4970 m/sec
Hardness	4.0
Reflectivity	72%
Electrical resistivity	7.2 μΩ·cm
Melting point	1728 K or 1455 °C
Boiling point	3186 K or 2913 °C
Thermal conductivity	91 W m^{-1} K^{-1}
Coefficient of linear thermal expansion	13.4 ×10^{-6} K^{-1}
Etchants (wet)	Mixture of H_2O_2 and H_2SO_4
Etchants (dry)	Cl_2
Main applications in IC manufacturing	To react with silicon to form NiSi for reducing the resistance of S/D contacts and local gate interconnections.

film. Sheet resistance is routinely measured and monitored for metal deposition process control in IC fabs.

11.3.1 Thickness and deposition rate

Metal films such as aluminum, titanium, titanium nitride, and copper are opaque films. Optically based techniques such as reflectospectrometry and ellipsometry commonly used in dielectric thin-film measurements cannot be used to measure metal film thicknesses. A destructive process is required to precisely measure an actual metal film thickness by cross-section SEM, or by measuring the step height with a profilometer after removing part of the deposited film.

For SEM measurement, a test wafer needs to be cut after metal thin-film deposition, and the cut samples are placed on the stage of the SEM system. An energetic electron beam scans across the sample, and the bombardment causes a secondary electron emission from the sample. Since different materials have different yield rates of secondary electron emissions, SEM can precisely measure the metal film thickness by measuring the intensity of the secondary electron emission. For example, a TiN film thickness can be measured on the top and sidewall of the contact hole from the SEM image shown in Fig. 11.10. Cross-section SEM can also detect whether there are voids in the plugs, and whether the tungsten plugs have good contact with the substrate. However, the process is expensive, destructive, and very time consuming; it is difficult to measure the uniformity of film across an entire wafer.

Question: Why are SEM images always in black and white?
Answer: SEM photographs are taken from the intensity of a secondary electron emission, which only emits strong or weak signals, and these transfer to bright and dim in a photographic image. SEM images with beautifully colored patterns are artificially painted after the photograph has been taken and the image has been analyzed.

Profilometer measurements can provide information about the thickness and uniformity of thick films (>1000 Å). A profilometer requires a patterned etch process prior to the measurement. First, a layer of metal film is deposited on the wafer surface, then a photolithography process patterns the photoresist in certain areas. After a wet etch process, most of the metal film on the wafer is removed, and the photoresist is stripped, leaving metal steps on certain parts of the wafer. Then the wafer is put in a stylus profilometer, which can sense and record the microscopic surface profile with a stylus probe (see Fig. 11.11).

Deposition rate is determined by the measured film thickness divided by the deposition time. In a metal CVD process, deposition rate can be affected by gas flow rate and processing temperature. In a magnetron sputtering deposition process, deposition rate is mainly controlled by the bias power. For an evaporator process, deposition rate is mainly determined by the filament current.

Ultrathin (50 to 100 Å) titanium nitride film is almost transparent. Therefore, reflectospectrometry and ellipsometry can be used to measure its thickness.

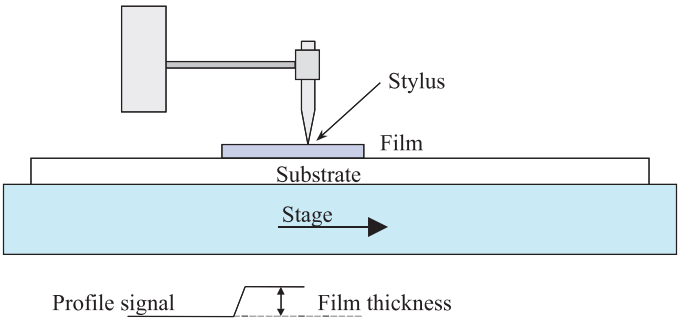


Figure 11.11 Schematic of a profilometer.

A four-point probe is commonly used to indirectly monitor metal film thickness by assuming that the resistivity of a metal film is constant over the entire wafer surface.

The acoustic method is a technique that can directly measure opaque thin-film thicknesses without direct contact with the film. Therefore, it is capable of measuring metal film thickness and uniformity on production wafers, an advantage that is very helpful for metal deposition process control.

The basic principle of the acoustic measurement method can be illustrated in Fig. 11.12. A laser beam is shot onto the thin-film surface; a photodetector is then used to measure the intensity of the reflected light. A very short laser pulse (~0.1 ps or 10^{-13} sec) from a pump laser is focused and shot onto the same spot, about 10×10 μm, which heats up the surface from 5 to 10 °C in that short time. Thermal expansion of the material at that spot causes a sound wave, which propagates through the film at the speed of sound of that particular material. When the acoustic wave reaches the interface of different materials, part of the wave is reflected from that interface, while the other part continues to propagate in the material underneath. The reflected sound wave, or echo, can cause reflectivity changes when it reaches the thin-film surface. The acoustic wave will echo back and forth in the film until it is damped out. The time between peaks Δt of the reflectivity change indicate the time the sound wave traveled back and forth in the films. When the speed of sound (V_s) of the material is known, the film thickness can be calculated by

$$d = V_s \Delta t/2.$$

The decay rate of the echo is related to film density. This method can also be used to measure film thickness for each film in a multilayer structure.

Figure 11.12(a) shows the measurement of a TiN film on a thick TEOS oxide layer, which is treated as the substrate. Figure 11.12(b) shows the time between peaks of reflectivity change $\Delta t \approx 25.8$ ps. The speed of sound in TiN film is $V_s = 9500$ m/s = 95 Å/ps; therefore, the TiN film thickness can be calculated as $d = 1225$ Å.

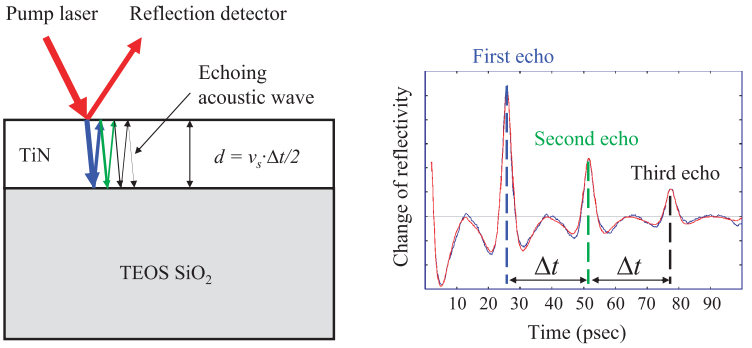


Figure 11.12 Acoustic method for metal thin-film measurement [Fig. 11.12(b) courtesy of Rudolph Technologies, Inc.].

X rays can reflect on a sample surface when the incident angle is very small. The reflectivity of an x ray is determined by the incident angle, surface roughness, film thickness, and film density. Due to interference between the reflection from the film surface and reflection from the interface of the film and substrate, reflectivity of an x ray oscillates with the changing incident angle. By varying the glancing incident angle and measuring the intensity of the reflected x ray, a spectrum of oscillating x ray intensity can be acquired as a function of incident angle. By fitting with the theoretical model, x-ray reflectometry (XRR) can be used to measure film thickness and film density, which provides information on film composition. Figure 11.13(a) illustrates the principle of the XRR measurement, and Fig. 11.13(b) gives an example of XRR measurement results of TaN/Ta film on Si substrate, where $q_z = 4\pi \sin\theta/\lambda$.

11.3.2 Uniformity

Uniformity (actually nonuniformity) of the thickness, sheet resistance, and reflectivity are routinely measured during process development and for process maintenance. It can be calculated by measuring the sheet resistance and reflectivity at multiple locations on a wafer in the patterns illustrated in Fig. 11.14.

The more measurement points are taken, the higher the precision of the measurement. However, more measurement points require longer measurement time, which means lower throughput and higher production cost because, in an IC manufacturing fab, time is money—a lot of money.

49-point measurement, 3σ standard deviation nonuniformity is the most commonly used nonuniformity definition for process qualification in 200-mm semiconductor fabs. To attain better control of the larger wafer size, more measurement points are needed. A 300-mm wafer usually needs a 121-point measurement. It is very important to clearly define nonuniformity because, for the same set of measurement data, different definitions can cause different nonuniformity results. For

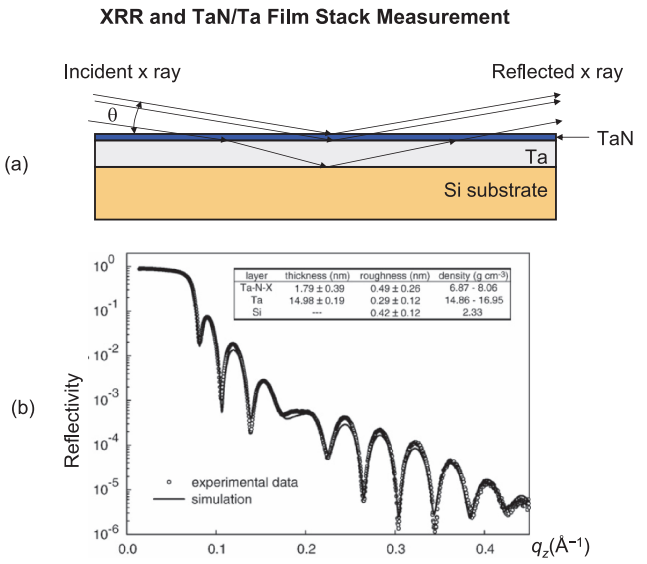


Figure 11.13 (a) XRR measurement principle and (b) an example of measurement results. [Fig. 11.13(b), from R. J. Matyi)].

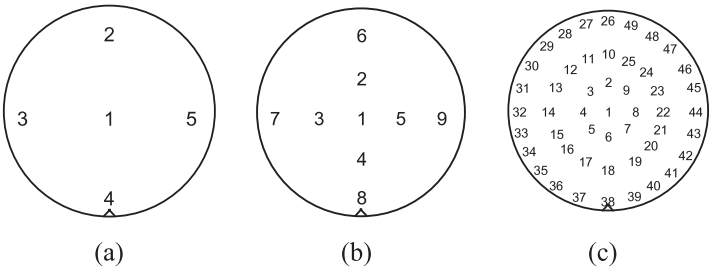


Figure 11.14 Mapping patterns for uniformity measurement: (a) 5 point, (b) 9 point, and (c) 49 point.

production wafers, sometimes less-time-consuming 5- and 9-point measurements are used for process monitoring and control, determined by processing requirements and metrology tool throughput.

11.3.3 Stress

Stress is due to material mismatch between a film and substrate. There are two types of stress, compressive and tensile. If stress is too high, whether compressive or tensile, it can cause serious problems. For metal thin films, high compressive stress can cause hillocks, which can short metal wires between different layers;

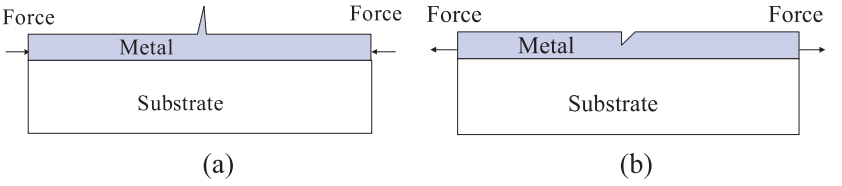


Figure 11.15 Different defects caused by high film stresses. (a) Hillock caused by compressive stress and (b) crack caused by tensile stress.

high tensile stress can cause films or interconnection lines to crack or peel. Figure 11.15 illustrates thin film defects caused by high film stresses.

There are two causes of stress: intrinsic and thermal. Intrinsic stress is caused by film density and determined by ion bombardment in the plasma sputtering deposition process. When atoms on a wafer surface are bombarded by energetic ions from plasma, they are squeezed densely together while forming the film. This type of film likes to expand, thus the film stress is compressive. Higher deposition temperature increases the mobility of the adatoms, also increasing film density and causing less tensile stress. Intrinsic stress is also related to the deposition temperature and pressure. Thermal stress is due to wafer temperature change and the different thermal expansion rates of the film and substrate. Aluminum has a very high thermal expansion rate, $\alpha_{Al} = 23.6 \times 10^{-6}\ K^{-1}$, compared to silicon, $\alpha_{Si} = 2.6 \times 10^{-6}\ K^{-1}$. For aluminum film deposited at a higher temperature such as 250 °C, when the wafer is cooling down to room temperature, the aluminum film shrinks more than the silicon substrate due to its higher thermal expansion rate. In this case, the wafer pulls the aluminum film and causes tensile stress. In fact, at room temperature, a little tensile stress is favorable for aluminum thin films, because the stress becomes less tensile when the wafer is heated during the succeeding metal annealing (~450 °C) and dielectric deposition (~400 °C) processes.

Question: Why does silicon oxide film favor compressive stress at room temperature?

Answer: Because silicon oxide has a lower thermal expansion rate ($\alpha_{SiO_2} = 0.5 \times 10^{-6}\ K^{-1}$) than the silicon substrate. If it has tensile stress at room temperature, it becomes more tensile in later processes when the wafer is heated up, which can crack the oxide film and cause hillocks on the aluminum line.

Stress can be measured from the change of wafer curvatures before and after thin-film deposition, described in detail in Chapter 10. The normal procedure for metal thin-film stress measurement is: measurement of wafer curvature, deposition of thin film with known thickness, and second wafer curvature measurement.

11.3.4 Reflectivity

Reflectivity is an important property of thin metal films. For a stable metallization process, reflectivity of the deposited film should be a constant. A change of reflectivity during the process indicates drifts of the processing conditions. Reflectivity is a function of film grain size and surface smoothness and needs to be controlled. The larger the grain size is, the lower the reflectivity. The smoother the metal surface is, the higher the reflectivity. Reflectivity measurement is an easy, quick, and nondestructive process, and it is frequently performed in the metal bays of semiconductor fabs.

Reflectivity is very important to photolithography processes because it can cause a standing wave effect due to interference between incoming light and reflecting light. This can affect photolithographic resolution by creating wavy grooves on the sidewall of the photoresist stack due to periodic overexposure and underexposure. ARC is required for metal patterning processes, especially for aluminum patterning, because aluminum has very high reflectivity—180 to 220% relative to silicon.

Reflectivity can be measured by focusing a light beam on a film surface and then measuring the intensity of the reflected beam. Reflectivity measurement results use the value relative to silicon.

11.3.5 Sheet resistance

Sheet resistance is one of the most important characteristics of conducting materials, especially for conducting films. It is commonly used to monitor the conducting thin-film deposition process and deposition chamber performance. For conducting films with known conductivity, a sheet resistance measurement is widely used to determine the film thickness, since this method is much faster than SEM and profilometer thickness measurements. Resistivity is one of the most fundamental properties of a material. For a conducting thin film, resistance can be calculated by the production of film sheet resistance and film thickness.

Sheet resistance R_s is a defined parameter. A four-point probe is a measurement tool that measures voltages and currents, and calculates sheet resistance. By measuring sheet resistance, film resistivity ρ can be calculated if the film thickness is known, or film thickness can be determined if its resistivity is known.

For the conducting line shown in Fig. 11.16(a), resistance can be calculated as

$$R = \rho \frac{L}{A},$$

where R is the resistance, ρ is the resistivity of the conductor, L is the length of the conducting line, and A is the area of the line cross section. If the wire cross-section is a rectangle, as shown in Fig. 11.16(b), its area can be simply expressed as the

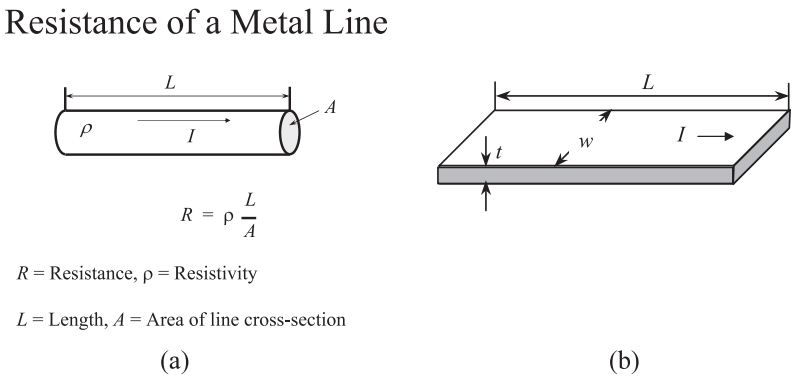


Figure 11.16 Conducting lines with (a) round and (b) rectangular cross sections.

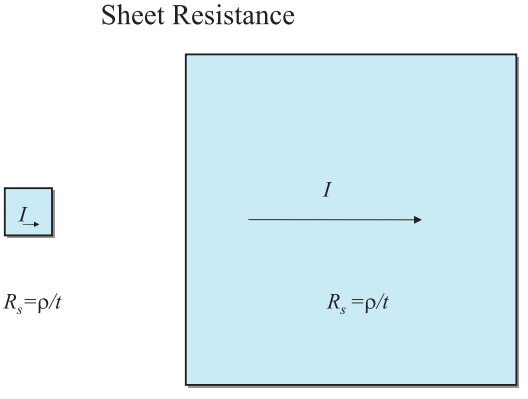


Figure 11.17 Sheet resistance is the same for these two thin-film squares.

product of the width and thickness (wt). The line resistance can be expressed as

$$R = \rho \frac{L}{wt}.$$

For a square sheet, the length is equal to the width, $L = w$, thus the two cancel each other. Therefore, resistance of a square conducting sheet, defined as sheet resistance, can be expressed as

$$R_s = \rho/t.$$

The unit of sheet resistance is ohms per square ($\Omega/\square$). The square symbol $\square$ here only denotes that the value is the resistance of a square; the size of the square does not matter. If the metal film thickness is perfectly uniform, sheet resistance of the square sheet with one micron at each side will be the same as a square sheet with one inch at each side, as shown in Fig. 11.17.

Question: For two conducting lines patterned from the same thin metal film with the same length-to-width ratios (as shown in Fig. 11.18), are their line resistances the same?

Answer: Yes. Both lines are consistent with the same number of square sheet resistors in serial connection. Since each square sheet has the same sheet resistance, their line resistances are the same.

Resistivity is related to the film material, grain size, and structure. For a particular metal film, the larger the grain size, the lower the resistivity.

The four-point probe shown in Fig. 11.19 is the most popular tool used to measure sheet resistance. A certain amount of current is applied between two of the pins, and voltage is measured between the other two pins. Sheet resistance equals the ratio of the voltage to the current and multiplies a constant, which depends on the pins being used. Typically, the distances between the probes are $S_1 = S_2 = S_3 = 1$ mm; if a current I is applied between P_1 and P_4, $R_s = 4.53\ V/I$, where V is the voltage between P_2 and P_3. If the current is applied between P_1 and P_3, $R_s = 5.75\ V/I$, where V is measured between P_2 and P_4. These two equations are derived with the assumption of a infinite film area, which is not quite true for thin-film measurements on a wafer.

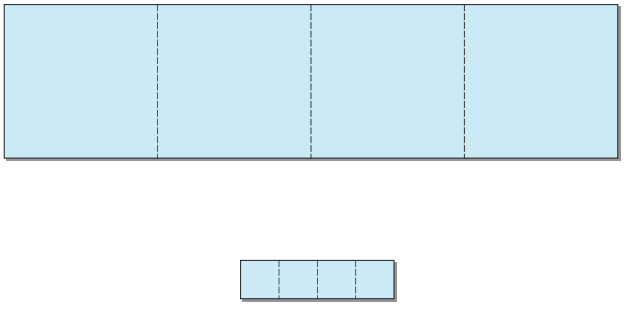
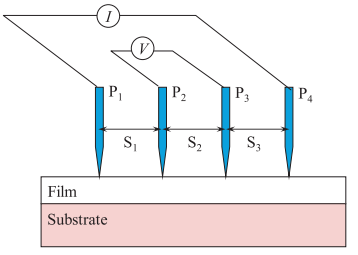


Figure 11.18 Line resistance of two conducting lines.



For a typical four-point probe, $S_1 = S_2 = S_3 = 1$ mm,
If current is applied between P_1 and P_4, $R_s = 4.53\ V/I$
If current is applied between P_1 and P_3, $R_s = 5.75\ V/I$

Figure 11.19 A four-point probe.

An advanced four-point measurement system runs four measurement procedures automatically in sequence, with both configurations and reverse current for each configuration to minimize the edge effect and achieve optimized results. Since four-point probes make direct contact with the film, they cannot be used on a product wafer. They are only used on test wafers for process development, tool qualification, and process control.

Sheet resistance measurement can provide important information about the film resistance, which is affected by film thickness, grain size, alloy concentration, and the incorporation of impurities, such as oxygen. Since process development usually sets up the correlation between the sheet resistances of metal layers and the product yield, these measurements are always closely monitored in the IC manufacturing process.

The resistivities of Al–Cu alloys and Cu films are well known and stable at certain deposition conductions. Therefore, sheet resistance measurements can provide a quick and convenient way to monitor film thickness and uniformity. However, this does not work with titanium nitride, because TiN resistivity is very sensitive to several process variables, such as nitrogen/argon flow ratio and processing temperature. Measured sheet resistance cannot be used to calculate TiN film thicknesses, since there is little confidence that the assumed film resistivity is the right value.

11.4 Metal Chemical Vapor Deposition

11.4.1 Introduction

Metal CVD is widely used to deposit metal in IC processing. CVD metal films have very good step coverage and gap-fill abilities and can fill tiny contact holes to create connections between metal layers. CVD metal thin films normally have poorer quality and higher resistivity than PVD metal thin films. Therefore, they are mainly used for contact and via plugs and local interconnections, not for global interconnections.

The most popular CVD metals are tungsten, tungsten silicide, and titanium nitride. This section covers these metal CVD processes. Most metal depositions are thermal processes; external heat from the heating elements provides the free energy needed for chemical reaction. In some cases, remote plasma sources are used to generate free radicals and increase the chemical reaction rate. A schematic of a metal CVD system is given in Fig. 11.20. The rf unit in the system is used for the plasma dry clean of the processing chamber.

A normal processing sequence for metal CVD deposition is as follows:

- wafer slides into the chamber
- slip valve closes
- pressure and temperature are set, with secondary process gas(es)
- all process gases flow in; deposition starts
- main processing gas is terminated (secondary processing gas remains on)
- all processing gases are terminated

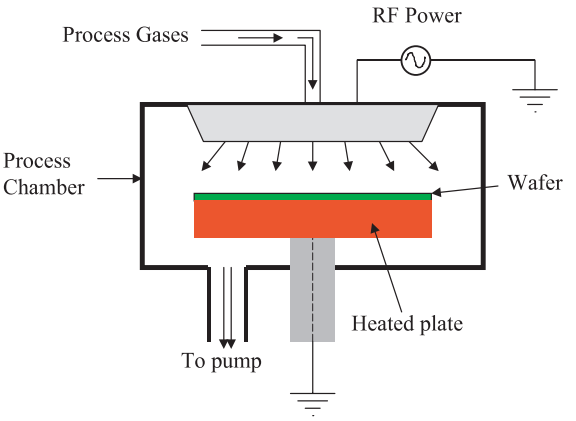


Figure 11.20 A metal CVD system.

- chamber is purged with nitrogen
- slip valve opens, and robot pulls wafer out.

During the metal deposition process, the metal thin film is deposited not only on the wafer surface, but also anywhere inside the processing chamber. A periodic cleaning process is needed to prevent the film inside the chamber from cracking and flaking, which generates particles. A plasma-enhanced chemistry is employed to routinely dry clean the chamber. For metal CVD chambers, fluorine- and chlorine-based chemicals are used in chamber plasma cleaning processes to dry etch films deposited inside the process chamber.

A processing sequence for plasma cleaning is as follows:

- chamber pumps down
- pressure and temperature are set with cleaning (etchant) gas(es)
- rf turns on; plasma and cleaning process is started
- rf turns off; chamber is purged
- set up pressure and temperature, with secondary process gas(es)
- main processing gas flows in; seasoning layer is deposited
- termination of main processing gas; secondary processing gas(es) remains on
- All processing gases are terminated
- chamber is purged with nitrogen
- chamber is ready for next deposition.

The seasoning step is very important. A thin layer of metal film intentionally deposited inside the chamber can effectively prevent loose residue from flaking and generating particles. It also can seal the remaining etchant gas and prevent it from interfering with the deposition process, as well as help prevent the first-wafer effect.

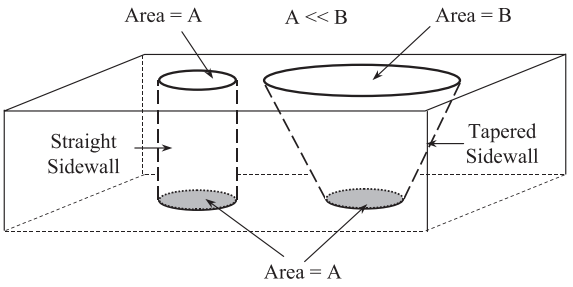


Figure 11.21 Vertical and tapered contact holes.

11.4.2 Tungsten

Tungsten has been extensively used in the semiconductor industry for metallization since the 1980s. Smaller feature sizes no longer have room for tapered contacts for the contact/via holes. Narrow and near-vertical contact/via holes must be used in metallization processes; however, PVD metal cannot fill these holes without voids. Thus, tungsten CVD processes have been developed. CVD tungsten film has almost perfect step coverage and conformity and can fill tiny holes without voids. It has been widely used as a plug to connect conducting layers in semiconductor fabrication processes. Figure 11.21 illustrates the vertical and tapered contact holes and clearly shows that using vertical contact holes can increase the packing density of IC wiring.

Because tungsten has lower resistivity than silicide and polysilicon, it can also be used for local interconnection applications. A local interconnection tungsten CVD process is different from the plug CVD process, since interconnection processing requires lower resistivity, while the main concern for tungsten plug application is gap-fill capability.

Tungsten hexafluoride (WF_6) is always used as a tungsten precursor for tungsten CVD processes. Silane (SiH_4) and hydrogen (H_2) are used to react with WF_6 and reduce fluorine to deposit tungsten. WF_6 is corrosive; when it reacts with water, it forms HF. The special safety hazard of HF is that it cannot be felt when direct contact is made with skin, and it can cause severe pain and damage once it reaches the bone and reacts with calcium. Double-layer rubber gloves are always required for tungsten CVD chamber wet cleaning. SiH_4 is pyrophoric (ignites itself), explosive, and toxic. H_2 easily leaks due to its small molecular size, and it is flammable and explosive.

For contact holes between metal and silicon, a silicon reduction process can be used to deposit tungsten:

$$2\,WF_6 + 3\,Si \rightarrow 2\,W + 3\,SiF_4.$$

It can be used as a nucleation layer to form selective tungsten. In this case, tungsten only deposits on the nucleation layer, which is on the silicon surface and not on the silicon dioxide surface during bulk deposition with H_2/WF_6 chemistry.

One of the advantages of selective tungsten is that it has good contact with silicon and thus a lower contact resistance. Tungsten can be selectively deposited in contact holes and therefore does not need etchback or polishback of bulk tungsten on the wafer surface. However, this process consumes silicon from the substrate and can cause junction loss. It also introduces fluorine into the substrate. Tungsten will react with silicon to form tungsten silicide during subsequent processes and will further consume the substrate silicon and cause junction loss. Another problem of the selective tungsten CVD process is that it is difficult to achieve perfect selectivity, since there are always some unpredictable nucleation sites on the silicon oxide surface. Selective tungsten processes are not commonly used in IC applications.

Most fabs use a blanket tungsten CVD process with a titanium/titanium nitride barrier/glue layer, which reduces contact resistance, prevents tungsten from reacting with silicon, and helps tungsten stick to silicate glass. The blanket tungsten CVD process has two steps: nucleation and main deposition. First, a thin layer of tungsten is deposited on the glue layer by using SiH_4/WF_6 chemistry:

$$2\,WF_6 + 3\,SiH_4 \rightarrow 2\,W + 3\,SiF_4 + 6\,H_2.$$

The nucleation step can strongly affect the uniformity of the subsequent film. After the nucleation step, H_2/WF_6 chemistry is used to deposit bulk tungsten to fill the contact/via holes:

$$WF_6 + 3\,H_2 \rightarrow W + 6\,HF.$$

During tungsten CVD processes, Ar is used for backside purging to prevent deposition on the edge and backside of the wafer. N_2 is used to improve film reflectivity and conductivity. Figure 11.22 illustrates the CVD tungsten seed and bulk layers.

Question: Plasma can generate free radicals and enhance the deposition rate of CVD processes. However, most tungsten CVD processes are thermal processes, not PECVD. Why?

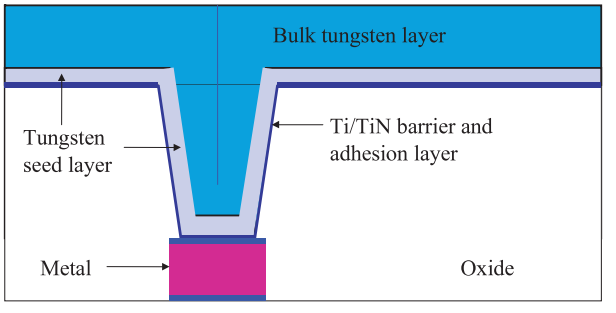


Figure 11.22 CVD tungsten seed and bulk layers.

Answer: In plasma, WF_6 dissociates and generates fluorine free radicals. With the help of ion bombardment (which always comes along with plasma), fluorine radicals can etch silicate glass during tungsten CVD and create many defects.

After each deposition, the CVD chamber must be cleaned with fluorine chemistry. NF_3 is used to provide free fluorine radicals in a plasma process, and Ar ions are used to provide heavy bombardment. Ion bombardment helps to damage bonds between tungsten atoms and significantly improves the etching effect:

$$6\,F + W \rightarrow WF_6.$$

After the cleaning process, a thin layer of tungsten is deposited in the chamber in the seasoning step to reduce particle contamination and eliminate the first-wafer effect.

The bulk tungsten CVD process fills contact/via holes and also coats a layer of tungsten over the entire wafer surface. Bulk tungsten on the wafer surface must be removed so that only tungsten is left behind in the contact/via holes. (These serve as plugs to connect different conducting lines between different layers.) Tungsten etchback processes with fluorine chemistry were extensively used in the industry until tungsten CMP processes become widely adapted by IC fabs in the late 1990s.

Question: By patterning and etching bulk tungsten film, a metal interconnection can be formed directly with tungsten wires, allowing many processing steps [such as tungsten CMP, metal stack (Ti/Al–Cu/TiN) PVD, metal wire patterning, and metal wire etch] to be skipped. Why do fabs not do this to save processing costs?

Answer: CVD tungsten has a much higher resistivity (8 to 11 μΩ·cm) than a PVD Al–Cu alloy (2.9 to 3.3 μΩ·cm). Tungsten can be used for local interconnections, not for global interconnections, because highly resistant interconnects will slow down circuit speed and increase power consumption, which are unacceptable consequences for most CMOS logic devices. However, DRAM chips are very cost sensitive, and many of them indeed use tungsten to fill BL and contact holes, then pattern and etch the tungsten film to form the BL directly. DRAM clock speed is not that high, so tungsten wire BLs are acceptable, and this low-cost approach helps drive down cost. Also, using a tungsten BL allows designers to put BLs below the storage node capacitors, and allow higher temperature processes such as LPCVD of polysilicon and silicon nitride.

Figure 11.23 shows applications of conducting layers in a recessed-gate stacked capacitor DRAM. In the figure, BLC is the bitline contact, and SN is the storage node. It is evident that the WL is formed by a polysilicon-tungsten stack. A thin layer of tungsten nitride (WN) is deposited between the polysilicon and tungsten. SAC is formed by polysilicon. Both BLCs in the array and periphery areas are

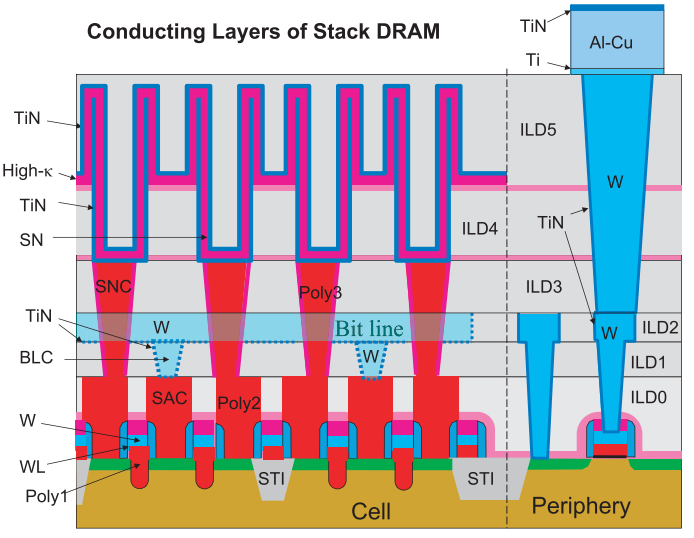


Figure 11.23 Tungsten layers in a recessed-gate stacked DRAM.

formed by tungsten with a Ti/TiN barrier/glue layer. SNC is formed by polysilicon. The SN capacitor is formed by a TiN/high-κ/TiN stacked layer. High-κ can be an Al_2O_3 film with silicon oxide on both sides. The metal contact is filled with tungsten using Ti/TiN as a barrier/glue layer.

Some DRAM manufacturers implement bWL technology for advanced DRAM chip fabrication. A bWL DRAM uses tungsten for WL and a tungsten/polysilicon stack for BL. It also uses tungsten with a TiN/Ti liner to form SNC plugs.

After tungsten deposition, bulk tungsten on a wafer surface needs to be removed by either tungsten etchback or tungsten CMP. The advantage of the tungsten etchback process is that it can be performed in situ with a tungsten CVD process in a single mainframe. However, tungsten CMP has much better process quality control and can significantly improve chip yield. It quickly replaced the etchback process when CMP technology matured and became more broadly accepted.

11.4.3 Tungsten silicide

Tungsten silicide (WSi_2) is used for gate and WL applications for DRAM chips with mature technology. Both SiH_4 and SiH_2Cl_2 (dichlorosilane, or DCS) are employed as silicon source gases, and WF_6 is employed as the tungsten precursor. SiH_4/WF_6 chemistry requires a lower processing temperature, typically at 400 °C, while DCS/WF_6 chemistry requires a higher processing temperature, about 550 to 575 °C.

For SiH_4/WF_6 chemistry,

$$WF_6 + 2\,SiH_4 \rightarrow WSi_2 + 6\,HF + H_2.$$

This process is very similar to the nucleation step in tungsten CVD processes. The main difference is the ratio of the SiH_4/WF_6 flow rate; when the ratio is lower than 3, the chemical replacement reaction deposits silicon-rich tungsten instead of tungsten silicide. To ensure tungsten silicide (WSi_x, with x between 2.2 to 2.6) deposition, the SiH_4/WF_6 flow ratio must be larger than 10.

The DCS/WF_6 chemistry can be expressed as

$$2\,WF_6 + 7\,SiH_2Cl_2 \rightarrow 2\,WSi_2 + 3\,SiF_4 + 14\,HCl.$$

DCS/WF_6-based processing requires a higher deposition temperature, has a higher tungsten silicide deposition rate, and has better film step coverage than SiH_4/WF_6-based processes. It also has much lower fluorine concentration in the film, and less film peeling and cracking problems due to its lower tensile stress. DCS/WF_6 silicide processing is gradually replacing silane-based processes.

The advantages of WSi_x over $TiSi_2$ are fewer process steps and the ease of integration with polysilicon deposition in one processing tool. However, tungsten silicide has higher resistivity than titanium silicide, and it only forms silicide on the gate, while the salicide process can form titanium silicide on the gate and S/D at the same time.

11.4.4 Titanium

There are two major applications of titanium in IC chip manufacturing. It is needed to reduce contact resistance before the titanium nitride barrier/glue layer deposition because direct contact between TiN and Si can cause high contact resistance. Titanium also reacts with silicon to form titanium silicide. For barrier layer applications, most choose PVD titanium instead of CVD titanium because PVD film has better quality and lower resistivity.

For titanium silicide processes, CVD titanium has some advantages. It has better step coverage than PVD titanium; this is important because Ti is deposited after gate etch, and the wafer surface is not planarized. At high temperatures (~600 °C), CVD Ti can react with Si to form $TiSi_2$ simultaneously during Ti deposition. The process can be expressed as

$$TiCl_4 + 2\,H_2 \rightarrow Ti + 4\,HCl,$$
$$Ti + Si \rightarrow TiSi_2.$$

11.4.5 Titanium nitride

Titanium nitride (TiN) is extensively used as a barrier/glue layer for tungsten plugs. Although the quality of CVD TiN film is not as good as PVD, and it has higher resistivity than PVD TiN, it has much better sidewall step coverage than PVD TiN films, 70 versus 15%. A thin layer of CVD TiN (~75 to 200 Å) is applied for contact/via holes as the barrier and glue layer for a tungsten plug after thin layers of PVD Ti deposition, as shown in Fig. 11.24.

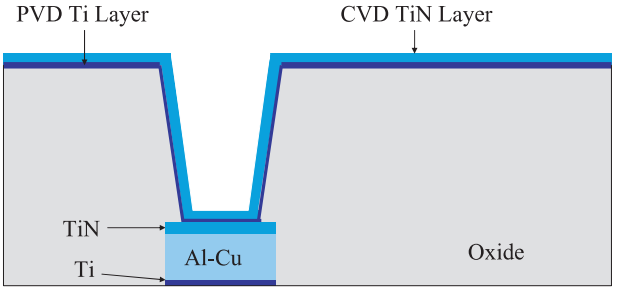


Figure 11.24 PVD and CVD titanium nitride layers.

TiN can be deposited from the reaction of $TiCl_4$ and NH_3 at 400 to 700 °C:

$$6\,TiCl_4 + 8\,NH_3 \rightarrow 6\,TiN + 24\,HCl + N_2.$$

TiN film deposited at higher temperatures has better film quality and a lower chlorine concentration. For low-temperature processes at 400 °C, Cl concentration can be as high as 5% in film; even high-temperature processes have ~0.5% Cl in film. Chlorine causes aluminum corrosion, which affects IC chip reliability. One of the possible byproducts, ammonium chloride (NH_4Cl), is a solid and can cause particle contamination.

Metal organic CVD (MOCVD) processes have been developed and widely used for TiN deposition. Metal organic compounds such as tetrakis-dimethylamido-titanium {$Ti[N(CH_3)_2]_4$, TDMAT} and tetrakis-diethylamido-titanium {$Ti[N(C_2H_5)_2]_4$, TDEAT} can be used as titanium nitride precursors, which dissociate at low temperatures (<450 °C) and deposit TiN with excellent step coverage. TDMAT is the more commonly used precursor. Its deposition temperature is about 350 °C and pressure is about 300 mtorr. The reaction can be expressed as:

$$Ti[N(CH_3)_2]_4 \rightarrow TiN + \text{organics}.$$

As deposited, a TiN thin film (~100 Å) is not as dense as a high-temperature deposited film, and has higher resistivity. It also has a high concentration of carbon and hydrogen. A postdeposition N_2-H_2 plasma treatment combined with chemical reactions from N and H free radicals can help to scavenge the C and H in the film and reduce their concentration. Ion bombardment of plasma can densify the film and reduce its resistivity. RTP annealing with N_2 ambient around 450 °C can also be used to densify the film and reduce resistivity.

TDMAT is a very poisonous liquid, and swallowing it can be fatal. $TiCl_4$ is highly irritating to the skin, eyes, and mucous membranes. Acute exposure can result in skin burns and marked congestion of the mucous membranes. It can also cause cornea damage. Ammonia is corrosive with a sharp, penetrating odor. Direct contact with liquid ammonia can cause severe chemical burns on skin and eyes. Contact and inhalation of ammonia vapors at low concentrations

(~25 ppm) can instigate irritation of the skin, eyes, nose, throat, and lungs. At high concentrations (~5000 ppm), it can cause severe eye irritation, chest pain, and fluid in the lungs.

11.4.6 Aluminum

Aluminum CVD processes to replace tungsten plugs and reduce interconnection resistance have been in research and development for a long time. Aluminum can be deposited at a relatively low temperature with aluminum organic compounds, such as dimethylaluminum hydride [$Al(CH_3)_2H$, or DMAH] and tri-isobutyl-aluminum [$Al(C_4H_7)_3$, or TIBA]. These compounds can decompose and deposit aluminum in a thermal process in a vacuum chamber. At about 350 °C, DMAH dissociates and deposits aluminum; the chemical reaction is expressed as

$$Al(CH_3)_2H \rightarrow Al + \text{volatile organics}.$$

DMAH is pyrophoric (igniting on contact with air), expensive, and often has high viscosity, which makes it difficult to handle. CVD aluminum film has very good trench and hole fill ability, and has lower resistivity than that of CVD tungsten film. Unlike tungsten film, which requires an etchback or a CMP process to remove the bulk tungsten deposition on the wafer surface, CVD aluminum film does not need to be removed from the surface for aluminum metallization. Because it is difficult to integrate a small amount of Cu into CVD Al, a layer of PVD Al–Cu is needed on top of CVD Al. Theoretically, CVD Al/PVD Al–Cu processing can save a process step, improve overall throughput, and reduce the processing cost for Al metal interconnections. Because copper metallization technology matured rapidly and replaced Al–Cu metallization in advanced CMOS logic ICs, the CVD Al process did not have a chance to be widely used for metal interconnections in IC manufacturing.

By using the cluster tools shown in Fig. 11.25, the entire ALD/CVD metallization process for gate-last HKMG integration can be performed in situ with a multichamber configuration. The integrated processing sequences are: wafer loading, preclean, HfO_2 ALD, TiN ALD, Ta ALD, cool down, and wafer unloading, performed in the cluster system shown in Fig. 11.25(a). After a pattern etch process removes the Ta protection layer from nNMOS, the photoresist is removed, and the wafer is cleaned, the cluster tool shown in Fig. 11.25(b) can be used for TiAl ALD, TiN ALD, and Al CVD. For planar MOSFET with gate-last HKMG process, CVD Al is commonly used to fill the gate trench after deposition of the work-function metals [see Fig. 11.8(b)]. For HKMG FinFET, CVD W is used to fill the gate trench.

11.5 Physical Vapor Deposition

11.5.1 Introduction

PVD is a deposition process that vaporizes solid materials, either by heating or by sputtering, and recondenses the vapor on the substrate surface to form a solid

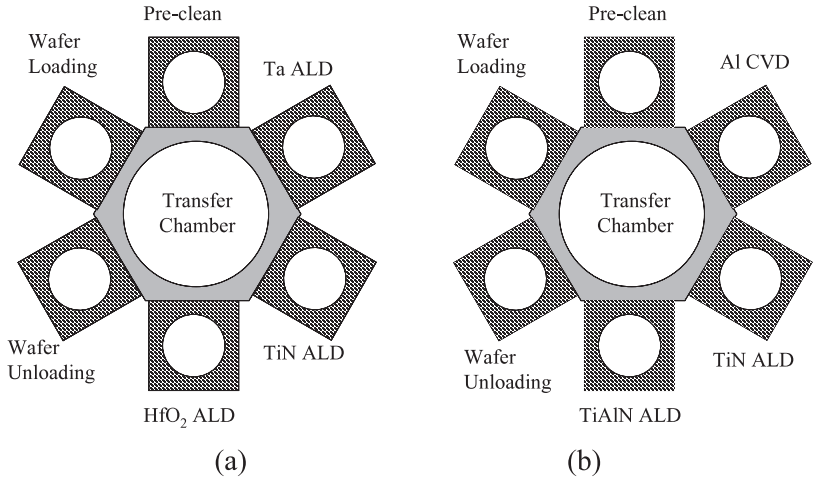


Figure 11.25 Cluster tools used for gate-last HKMG metallization.

thin film. It plays a very important role in metallization processes in semiconductor manufacturing.

Comparing CVD and PVD processes, the PVD process uses solid sources, while CVD processes use gaseous or vapor precursors. A CVD process relies on chemical reaction on the substrate surface; the PVD process does not. CVD film has better step coverage, and PVD film has better quality, lower impurity concentration, and lower resistivity. Figure 11.26 compares CVD and PVD processes.

For metallization processes in IC manufacturing, PVD processes are used to deposit Ni layers for NiSi formation, Ta barrier layers, TiN glue layers, Ta and TaN barrier layers, copper seed layers, Al–Cu layers, and TiN ARC layers. CVD processes are commonly employed to deposit TiN layers for barrier/adhesion, and tungsten for plugs.

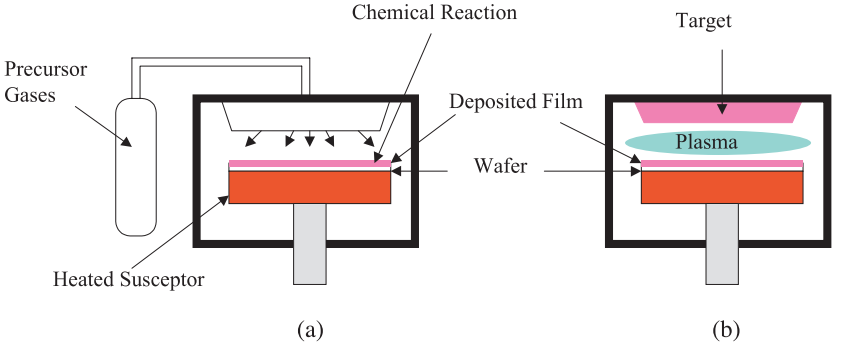


Figure 11.26 Comparison of (a) CVD and (b) PVD processes.

Two methods are used in metal PVD processes, evaporation and sputtering. Sputtering is the dominant PVD process in semiconductor fabrication processes because it can deposit high-purity, low-resistivity metal films with excellent uniformity and reliability.

11.5.2 Evaporation processes

In the early years of IC manufacturing, only aluminum was used for the metallization process, and thermal evaporation was generally used for Al deposition. Thermally evaporated Al films always have high mobile-ion concentration, which can affect transistors and circuits. An electron beam evaporator was developed to deposit high-purity Al and Al alloys.

11.5.2.1 Thermal evaporation

Aluminum has relatively low melting (660 °C) and boiling points (2519 °C), and it is relatively easy to vaporize under low pressure. In the early years of IC processing, thermal evaporators were widely used to deposit aluminum thin film to form gates and interconnections. A schematic of a thermal evaporation system is illustrated in Fig. 11.27. Throughout the process, the system needs to be under high vacuum, about 10^{-6} torr, to minimize residual oxygen and moisture. These residues can react with aluminum to form high-resistivity aluminum oxide and significantly increase the sheet resistance of the film.

Flowing a large amount of electric current through a tungsten filament heats up the filament by resistive heating, $P = RI^2$. A red-hot tungsten filament heats up the aluminum charge, melts it, and vaporizes it in the vacuum chamber. When aluminum vapor reaches the wafer surface, it recondenses and forms a thin layer of aluminum film on the surface.

In a filament evaporation system, a shutter mechanism is placed between the filament and wafers. At the beginning of the deposition process, the filament is heated to just above the metal melting point to melt all of the metal charge while the shutter is closed. After the temperature is stabilized and volatile impurities are

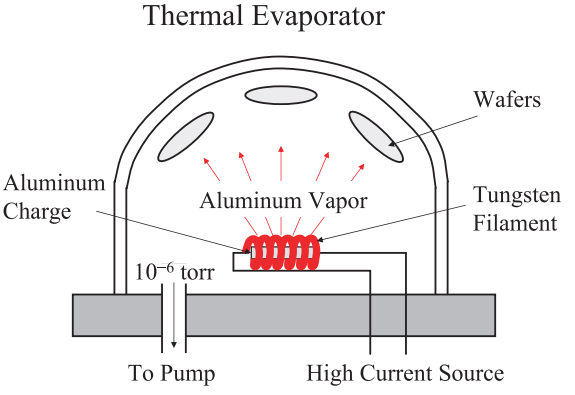


Figure 11.27 A thermal evaporator.

driven away from the charge surface by the heat, the current ramps up to raise the temperature and evaporate the metal. Then the shutter is opened to allow the metal vapor to emit, reach the wafer, condense on the surface, and deposit metal thin film on the wafer.

For thermal evaporation deposition processes, the deposition rate of aluminum is related to the heating power, which is controlled by the electric current; usually, a higher current has a higher deposition rate. One important safety issue for the thermal evaporator is electrical shock. The high current (~10 A) used by an evaporator can cause a fatal electric shock in the case of direct contact. In fact, a 1-mA electric current across a human heart can be lethal.

Aluminum thin film deposited with a thermal evaporator always has a trace mount of sodium from the tungsten filament; it is high enough to shift the threshold voltage of MOSFETs and affect IC device reliability. It also has a low deposition rate and poor step coverage. It is also very difficult to precisely control the proper proportions for alloyed films such as Al-Si, Al–Cu, and Al–Cu-Si. It is no longer used for metallization processes in VLSI and ULSI chip manufacturing. It could still be used for wafer backside gold coating processes for facilitating the die attachment. Some university and college research and education laboratories still use thermal evaporation systems because they are a simple tool and are easy to operate with relatively low maintenance cost.

11.5.2.2 Electron beam evaporation

To replace filament heating, which can cause contamination and poor step coverage, electron beam (e-beam) heating technology was developed to evaporate metals for IC metallization. A beam of electrons, typically with energy of about 10 keV and currents up to several amperes, is directed at the metal in a water-cooled crucible in a vacuum chamber, and heats the metal to an evaporation temperature. During the evaporation deposition process, the outer portion of the charge does not melt and remains in a solid state, minimizing film contamination from the trace amounts of impurities inside the graphite or silicon carbide crucible. Figure 11.28 illustrates a typical e-beam evaporating system.

To improve deposition uniformity and step coverage, planetary wafer holding systems are used. By using multiple electron guns and crucibles, it is possible to simultaneously evaporate different metals, and deposit metal alloys such as Al-Si, Al–Cu, Al-Si-Cu, etc. Heating the wafers with IR lamps can raise the wafer temperature, increase adatom surface mobility, improve film step coverage, form large grains, and reduce resistivity.

Although e-beam evaporation has better processing results than thermal evaporation, it cannot match the metallization process results of sputtering deposition. It is rare to see an e-beam evaporator in an advanced semiconductor fab. The e-beam evaporator also can cause device damage due to x-ray radiation generated when the energetic electrons hit the metal charge. Some less advanced semiconductor fabs are still using this tool for metallization processes due to its high-throughput and low equipment cost.

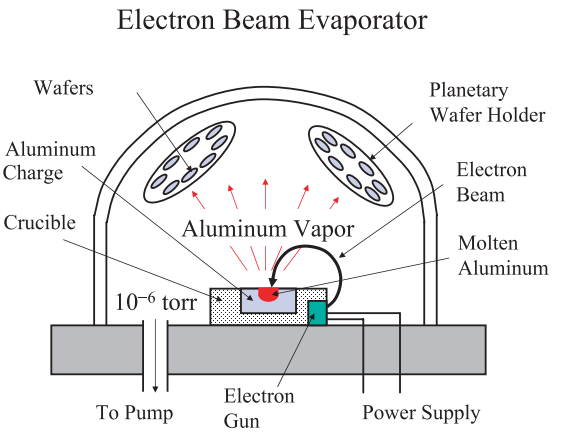


Figure 11.28 E-beam evaporator.

11.5.3 Sputtering

Sputtering deposition is the most common PVD process used for metallization in the IC industry. In fact, it has almost become a synonym for PVD in advanced semiconductor fabs. It involves energetic ion bombardments, which physically dislodge atoms or molecules from a solid metal surface and redeposit them on a substrate surface to form a metal thin film. Argon is the most frequently used gas for sputtering processes because it is inert, relatively heavy, abundant (~1% of atmosphere), and cost effective.

By applying electric power between the two electrodes under low pressure, a free electron between the electrodes is accelerated by the electric field and becomes more energetic as it gains energy from the electric field. When it collides with a neutral argon atom, it can knock one of its orbiting electrons out of the range of the attractive force of the nuclei to become a free electron. This is called an ionization collision, which generates a free electron and a positively charged argon ion. When a neutral argon atom loses a negatively charged electron in a collision, it becomes a positively charged argon ion. The two free electrons repeat this process to generate even more free electrons and ions, while electrons and ions are constantly lost due to collisions with electrodes, chamber walls, and electron ion recombinations. When their generation rate equals their loss rate, a steady state is reached, and stable plasma is generated.

While the negatively charged electrons are accelerated to a positively biased electrode called an anode, the positively charged argon ions accelerate toward a negatively biased cathode plate, called the target. The target plate is made from metal that is to be deposited on the wafer. When these energetic argon ions hit the target surface, atoms of the target material are physically dislodged from the surface by the momentum transfer from the impacting ions and are thrown into the vacuum in the form of a metal vapor. Figure 11.29 illustrates the sputtering process.

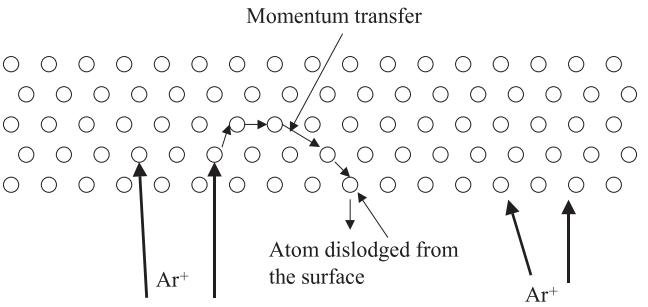


Figure 11.29 The sputtering process.

After the atoms or molecules leave the target surface, they travel inside the vacuum chamber in the form of a metal vapor. Eventually, some of them reach the wafer surface, adsorb on the surface, and become adatoms. The adatoms migrate on the wafer surface until they find nucleation sites and rest there. Other adatoms stop and recondense around the nucleation sites to form grains, which are in a single-crystal structure. When the grains grow and touch other grains, they form a continuous polycrystalline thin metal film on the wafer surface. The border between the grains is called a grain boundary, which can scatter electron flows and cause higher resistivity. Grain size is mainly determined by surface mobility, which is related to many factors, such as: wafer temperature, substrate surface condition, baseline pressure (contamination), and the final annealing temperature. Higher temperatures result in higher surface mobility and large-sized grains. Grain size has a strong impact on film reflectivity and sheet resistance. A metal film with a larger grain size has fewer grain boundaries for scattering electron flow and therefore has lower resistivity.

Question: High wafer temperatures increase grain size and improve conductivity, electromigration resistance, and film step coverage. Why do fabs not always use high wafer temperatures in PVD processes?

Answer: A metal film with large grain size is hard to etch with smooth sidewalls. Therefore, it is common practice to deposit metal films with smaller grain sizes at lower temperatures, and anneal metal films at higher temperatures after metal etch and photoresist strip. The annealing process forms larger grain sizes and reduces film resistivity. For Al–Cu alloy metal interconnects, processing temperature is limited to ~400 °C after deposition of the first layer of Al–Cu film.

The simplest sputtering system is the dc diode sputtering system, as illustrated in Fig. 11.30. In this system, a wafer is placed on a grounded electrode, and the target is a cathode, the negatively biased electrode. When a high-power dc voltage (several hundred volts) is applied to the system under low pressure, argon atoms are ionized by the electric field, and argon ions are accelerated and bombard the target, sputtering the target material from the surface.

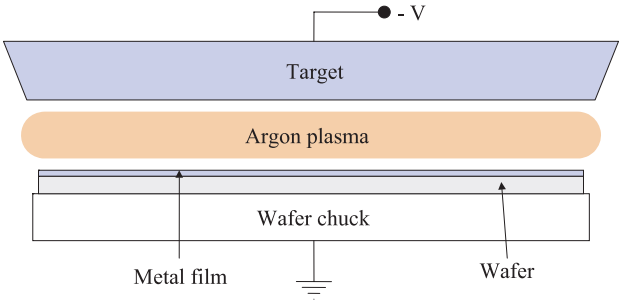


Figure 11.30 A dc sputtering system.

Other basic types of sputtering systems include dc triode, rf diode, dc magnetron, rf triode, and rf magnetron. The dc magnetron is the most popular method for PVD metallization processes because it can achieve high deposition rates, good film uniformity, good step coverage, high film quality, and easy process control. High deposition rates allow single-wafer sputtering deposition processes, which have several advantages over batch systems with planetary wafer holders. They have better WTW uniformity, higher system reliability, and lower particle contamination levels.

In a magnetic field, magnetic force causes gyromotion of charged particles. Because an electron is a very lightweight particle, it has a very small gyroradius and will be constrained near the magnetic field line. In a magnetic field, electrons have to travel much longer distances by spinning around the magnetic field lines; this gives electrons more chances for ionization collisions. Therefore, a magnetic field can help increase plasma density, especially at low pressure. In a dc magnetron sputtering system, a rotating magnet is placed on top of the metal target. The magnetic field generates higher plasma density and causes more sputtering near the magnet, where the magnetic field strength is stronger. By adjusting the location of the magnets, uniformity of the deposited film can be optimized. Magnetron sputtering creates a ring of erosion grooves on the target surface after a certain number of watt-hours of sputtering operation, as illustrated in Fig. 11.31.

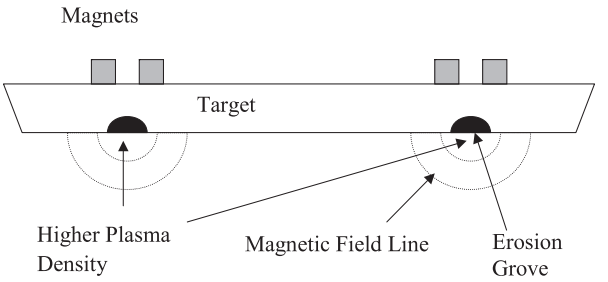


Figure 11.31 Magnetron sputtering system.

Another advantage of the sputtering process is that it can be easily used to deposit metal alloy films by using alloy targets with proper metal proportions. The sputtering deposition process can also be employed to deposit metal oxide or nitride, such as titanium nitride and tantalum nitride, in a reactive sputtering process by using oxygen or nitrogen with argon.

Normally, a shield or liner is installed inside a PVD chamber to protect the chamber wall and chamber parts from deposition of metal thin films, as shown in Fig. 11.32. The shield also serves as an anode of the dc discharge system while the wafer is placed on a floated chuck. After a certain amount of deposition, the metal-coated shield is replaced by a clean one during routine PM. The dirty one is sent out for cleaning and prepared for reuse.

With an rf power supply, a sputtering system can also be employed to deposit dielectric thin film by sputtering the nonconducting target. Radio frequency power can be applied to the backside of an insulating target and capacitively coupled to the plasma. Since electrons are much more mobile than ions in rf plasma, the surface of the insulating target accumulates net negative charges and dc voltage, up to a thousand volts. A negatively charged surface expels electrons and attracts positive ions, causing positively charged argon ions in the plasma to bombard the dielectric target surface and sputter molecules off the target surface to recondense on the wafer surface.

Argon is inert, relatively heavy, and inexpensive due to its abundance in the atmosphere. In fact, argon is the third most abundant elementary material in the atmosphere (about 1%), after only nitrogen (78%) and oxygen (20%). It is the most commonly used gas in sputtering processes, for both deposition and etch. It has also been applied in RIE and ion implantation for amorphous implantation. Table 11.8 lists some facts of argon.

11.5.4 Basic metallization processes

In an advanced wafer fab, the metallization process is normally performed in one integrated process sequence using a cluster tool with multiple chambers. Figure 11.33 shows a photograph of a cluster system used in advanced metallization processes.

Metal PVD chambers must achieve a high-baseline vacuum to minimize contamination, especially for aluminum deposition. It is necessary to achieve

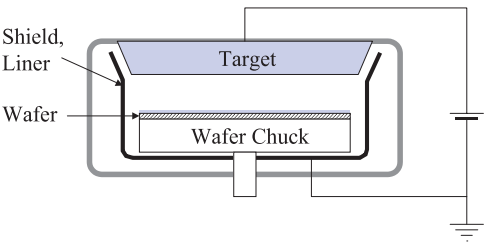


Figure 11.32 Sputtering system with a shield.

Table 11.8 Facts about argon.

Symbol	Ar
Atomic number	18
Atomic weight	39.948
Discoverer	Sir William Ramsay, Lord Rayleigh
Place of discovery	Scotland
Discovery date	1894
Origin of name	From the Greek word argos, meaning inactive
Molar volume	22.56 cm^3
Speed of sound in substance	319 m/sec
Refractive index	1.000281
Electrical resistivity	N/A
Melting point	−189.2 °C
Boiling point	−185.7 °C
Thermal conductivity	0.01772 W m^{-1} K^{-1}
Main applications in IC manufacturing	PVD, sputtering etch, dielectric etch, and ion implantation



Figure 11.33 Cluster tool for integrated metallization process (Applied Materials, Inc.).

ultrahigh vacuum (UHV, $<10^{-9}$ torr) baseline pressure to minimize contamination and improve conductivity. Combinations of dry, turbo, and cryo-pumps are used to achieve the required vacuum level for PVD systems.

Since a metal target always grows a thin layer of native oxide when it is exposed to the atmosphere, a burn-in process is required to condition the target before it processes production wafers for a new target, or after opening a chamber for maintenance. By sputtering the target with argon ions, the native oxide and defects generated during the target manufacturing process can be removed. Several dummy wafers are needed to protect the wafer chuck during the burn-in process.

A PVD process for Al–Cu metallization involves several steps: degas, predeposition sputtering clean, barrier layer deposition, bulk aluminum alloy layer deposition, and ARC layer deposition. For copper metallization, the PVD process is used to deposit Ta barrier and Cu seed layers. For contact and via processes, a

combination of PVD and CVD processes is used. The process starts with degassing and a predeposition sputtering clean, then Ti/TiN barrier/glue layers PVD and TiN CVD, and finally tungsten CVD to fill the holes.

11.5.4.1 Degas

Before starting the PVD process, it is very important to heat a wafer to a temperature high enough to drive out gases and moisture adsorbed on the wafer surface. Otherwise, outgassing of the adsorbed gasses and moisture during the deposition process can cause serious contamination and result in a high resistivity of the deposited metal film.

11.5.4.2 Preclean

Before metal deposition, a preclean step is required to remove the native oxide on the metal surface to reduce contact resistance. It is very important to remove this native oxide, since metal oxides such as aluminum and tungsten have very high resistivity. Without a precleaning process, contact resistance can be much higher than it should and can affect IC chip performance and reliability.

The argon sputtering process is commonly used for precleaning processes. In argon plasma, argon ions are accelerated by external power, in most cases by rf power. They bombard the wafer surface and blast off the ultrathin native oxide layer from the metal surface. This process exposes the metal under high vacuum and prepares the wafer for metal deposition. The precleaning process can also remove polymer residues from the bottom and sidewall of the contact/via holes. Carbon fluoride polymer residues are common byproducts of the silicate glass etching process when carbon fluoride etchant gases are used. Ion bombardment in the precleaning process always chops away some oxide from the upper corners of the contact/via holes and tapers the opening prior to the metal deposition process to improve step coverage and plug fill. Figure 11.34 illustrates the argon sputtering preclean process.

Both capacitively coupled plasma sources and inductively coupled plasma (ICP) sources are used for the preclean process. Advantages of an ICP plasma source are higher plasma density at lower pressure, and independent control of ion energy and flux.

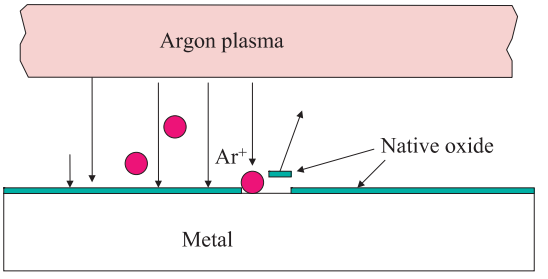


Figure 11.34 Argon sputtering preclean process.

11.5.4.3 Titanium physical vapor deposition

Before tungsten CVD and aluminum alloy PVD processes are implemented, a thin layer of titanium is always deposited to reduce contact resistance because titanium can trap oxygen and prevent it from bonding with tungsten or aluminum to form high-resistivity tungsten and aluminum oxides. For titanium deposition, larger grain size with low resistivity is desired. Therefore, a wafer is heated to approximately 350 °C during the deposition process to improve surface mobility of the titanium adatoms; this step also helps to improve step coverage.

For interconnection metal applications, titanium PVD with a standard magnetron chamber can meet the process requirements. However, for submicron contact/via applications, good bottom step coverage of the titanium film across the entire wafer is needed to reduce contact resistance. Standard magnetron systems can no longer meet this demand, since they always cause uneven step coverage from a wafer's center to its edge. Thus, collimating devices and metal ionization techniques have been developed for contact/via Ti/TiN sputtering processes.

A collimating system allows metal atoms or molecules to move mainly in a vertical direction, so that they can reach the bottom of the narrow contact/via holes. Although this slows down the deposition rate (because collimators stop many metal atoms and molecules from reaching the wafer surface), it improves bottom step coverage over the entire wafer. To compensate for the decrease in deposition rate, a higher dc power is applied to the electrodes to increase the sputtering effect. During this process, collimator holes near the center receive more deposition than those on the edge and begin to clog more quickly. Therefore, the magnetic system of a magnetron is designed to start the process with a center-thick deposition profile. The collimated sputtering system is illustrated in Fig. 11.35.

The rf current in inductively coupling coils can ionize metal atoms through an inductive coupling mechanism. The positive metal ions impact the negatively charged wafer surface mostly in a vertical direction, helping to improve bottom step coverage and reduce contact resistance. Ionized metal plasma systems have been used for copper metallization processes to deposit tantalum or tantalum nitride

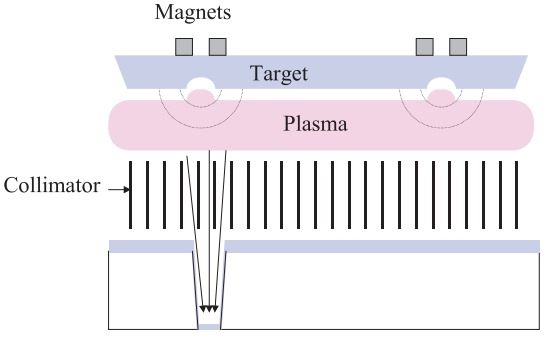


Figure 11.35 Collimated sputtering system.

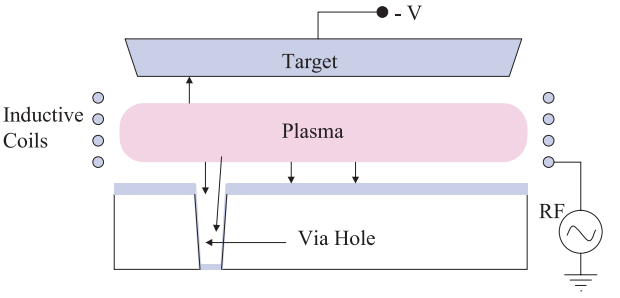


Figure 11.36 Ionized metal plasma system.

barrier layers and copper seed layers into high-aspect-ratio trench and via holes. An ionized metal plasma system is illustrated in Fig. 11.36.

The inductively coupled coils of an ionized metal plasma system are located inside the vacuum chamber and are made of the same material as the metal target because it will also be sputtered by ion bombardment.

11.5.4.4 Titanium nitride physical vapor deposition

Titanium nitride PVD uses a reactive sputtering process. By flowing nitrogen with argon into the processing chamber, some nitrogen molecules become dissociated by electron impact collisions in the plasma. Dissociated free nitrogen radicals have three unpaired electrons, which make them very reactive. Nitrogen radicals can react with sputtered Ti atoms to form and deposit TiN on the wafer surface. They also can react with a titanium target to form a thin layer of TiN on the target surface. Argon ions bombard the target surface and sputter titanium nitride from the target surface, then redeposit it on the wafer surface.

There are many applications involving titanium nitride, and all require different deposition processes. For TiN used as barrier and glue layers for tungsten, low resistivity and good step coverage (especially good bottom step coverage) are required. Therefore, collimated systems or ionized metal plasma systems are needed. For a TiN layer underneath Al–Cu, the main concern is lower resistivity. A standard magnetron system can meet the requirement. For ARC applications, low reflectivity is the major concern, and the coating can be deposited in a standard magnetron chamber. These three applications of titanium nitride are shown in Fig. 11.37.

TiN is also used for metal gate electrodes in HKMG processes. For gate-first HKMG, PVD TiN can be used. However, it is not suitable for gate-last HKMG because PVD nitride step coverage does not deposit film into narrow gate trenches very well. CVD TiN or ALD TiN can be used for gate-last HKMG.

Titanium nitride can be deposited in situ with titanium in the same PVD chamber. First, argon is used to sputter titanium from the target and deposit it onto the wafer surface. Then nitrogen flows into the processing chamber with argon for reactive sputtering, which deposits the titanium nitride. For integrated Ti and TiN in-situ deposition processes, a target cleaning process is necessary after the

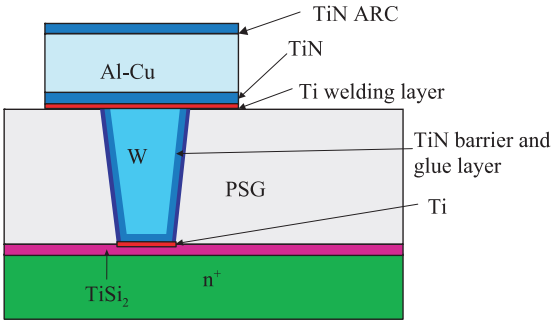


Figure 11.37 Three applications of TiN.

production wafer deposition process. Argon is used to sputter away the titanium nitride layer from the target surface to prepare it for the next titanium deposition process. A titanium dummy wafer is required during the cleaning process to protect the wafer chuck from being coated with titanium thin film. A silicon dummy wafer is not as good as a titanium dummy wafer because both titanium and titanium nitride stick well to a titanium wafer. This reduces the risk of particle contamination due to film cracking and flaking from the dummy wafer surface.

For a PVD chamber dedicated to TiN deposition, a periodic titanium coating is required to prevent flaking of the TiN layer off of the chamber parts and possibly causing particle contamination.

Nitrogen is the most commonly used gas in semiconductor fabs. It is used for nitride formation processes, such as silicon nitride and titanium nitride. It is the most abundant gas species in the earth's atmosphere (78%). Nitrogen is widely used as a purge gas because N_2 molecules are very stable, and it can be used as an inert gas as well for low-temperature (<700 °C) processes. Nitrogen is also used in pneumatics systems in processing tools. Some facts about nitrogen are listed in Table 11.9.

11.5.4.5 Aluminum–copper physical vapor deposition

Al–Cu alloy deposition requires an ultrahigh-baseline vacuum to achieve low film resistivity. At the start of the chamber pump-down process, air begins to be evacuated from the vacuum chamber. Residue gas inside the vacuum chamber, such as N_2 and O_2, is mainly from the atmosphere. When the chamber pressure reduces down to the millitorr range, the majority of residue gases are no longer the leftovers from the atmosphere, but the outgassing of adsorbed gases from the chamber wall. Moisture is one of these residual gases and is the most difficult to eliminate. If moisture is present during the aluminum sputtering deposition process, aluminum atoms will react with the residual H_2O to from aluminum oxide (Al_2O_3), which is a very good insulator. Integration of a tiny amount of oxygen in aluminum alloy film can significantly increase the film's resistivity. Therefore, to achieve high-quality, low-resistivity aluminum film deposition, the processing chamber must achieve a very high baseline vacuum to minimize the amount of moisture inside

Table 11.9 Facts about nitrogen.

Symbol	N
Atomic number	7
Atomic weight	14.007
Discoverer	Daniel Rutherford
Place of discovery	Scotland
Discovery date	1772
Origin of name	From the Greek words nitron genes, meaning nitre and forming, and from the Latin word nitrum (nitre is a common name for potassium nitrate)
Molar volume	13.54 cm^3
Speed of sound in substance	333.6 m/sec
Refractive index	1.000298
Electrical resistivity	N/A
Melting point	−209.95 °C
Boiling point	−195.64 °C
Thermal conductivity	0.02583 W m^{-1} K^{-1}
Main applications in IC manufacturing	As a purge gas for all IC manufacturing equipment, CVD precursor and carrier gas, and PVD reactive gas.

the chamber. A cluster tool is usually equipped with a staged vacuum system using a combination of pumps—including dry, turbo, and cryo-pumps—to achieve UHV in the aluminum PVD chambers. A cryo-pump can help a PVD chamber reach up to 10^{-10}-torr base pressure by freezing residual gases in a freeze trap.

There are two types of aluminum processes, standard and hot aluminum. A standard process is used in the deposition of aluminum alloy over a tungsten plug, normally after titanium and titanium nitride wetting layers. The main concerns with this process is uniformity of the deposited film and the consistency of the process during the lifetime of the target. Al–Cu alloy film deposited at a higher temperature has larger grain size, better electromigration resistance (EMR), and lower film resistivity. The problem is that films with larger grain sizes are difficult to etch with smooth pattern sidewalls. A standard aluminum process is normally operated at around 200 °C and can deposit metal films with smaller grain sizes, making it easier to achieve good interconnect line pattern etching. Larger grain sizes with high EMR and lower resistivity can be achieved later with a metal annealing process after photoresist stripping.

Some IC fabs have attempted to replace tungsten plugs with aluminum plugs for contact and via hole filling. Hot aluminum processing was developed for this application, which allows aluminum to fill contact and via holes. This reduces resistance between the conducting layers, since PVD aluminum is a much better conductor (2.9 to 3.3 μΩ·cm) than CVD tungsten (8 to 11 μΩ·cm). Hot aluminum processing consists of several steps. First, either titanium or both titanium and titanium nitride are deposited to lower contact resistance, and to prevent junction spiking for contact with silicon. A thin aluminum film is deposited at a low temperature (<200 °C) as the seed layer for the hot aluminum deposition. Then a thicker aluminum layer is deposited at a higher temperature (450 to 500 °C). Hot aluminum can diffuse into the contacts/via holes due to the high adatom surface

mobility. Hot aluminum fills the holes without voids while keeping the aluminum surface relatively planarized due to thermal flow. It also forms interconnections between the conduction layers with aluminum plugs.

High-pressure aluminum reflow processes have also been developed to fill the small contact/via holes with aluminum. Another approach for aluminum metallization is the combination of aluminum CVD and PVD processes.

11.6 Copper Metallization

Copper has a lower resistivity than an aluminum–copper alloy. It also has better electron migration resistance because copper atoms are much heavier than aluminum. Copper has always been an attractive choice for metal interconnections in IC chips. By using copper to replace the aluminum alloy, interconnection resistance can be significantly reduced due to copper's lower resistivity (1.8 to 1.9 μΩ·cm for copper versus 2.9 to 3.3 μΩ·cm for aluminum–copper alloy). The high EMR of copper allows higher current density in copper interconnection lines. Both properties help increase IC speeds. However, the high diffusion rates of copper in both silicate glass and silicon, and device reliability degradation from copper contamination in silicon substrates, hampered copper applications from 1960 to 1970. After the 1980s, when plasma dry etch replaced wet etch for pattering etch processing, lack of volatile inorganic copper compounds made application of copper for IC metal interconnections even more difficult.

In the 1990s, development of the CMP process paved the way for copper interconnections by applying the dual damascene process, which does not require a metal etching step. There are several other applications for the dual damascene copper interconnection process. A barrier layer of copper can be deposited into high-aspect-ratio via holes to prevent copper diffusion. This barrier layer provides good sidewall and bottom step coverage, good adhesion to dielectrics, and low contact resistance. Other applications are the deposition of copper film with high-quality, low-resistivity, and void-free filling of high-aspect-ratio trench and via holes. Finally, there are defect-free copper polishing and post-CMP cleaning processes. A copper interconnection processing sequence is illustrated in Fig. 11.38.

11.6.1 Precleaning

All metallization processes require a precleaning process to remove thin native oxide and possible polymer depositions from dielectric etch on the metal surface and bottom of the via holes. It is a very important step because insufficient cleaning can result in high contact resistance, induced by either the presence of native oxide between the metal layers or voids in the via plugs.

Like other metallization processes, argon sputtering etch is predominantly used for predeposition cleaning to remove native oxide and any other materials on the metal by physical ion bombardment. Radio frequency power is used to generate argon plasma that is either inductively coupled, called the source rf, or capacitively coupled, called bias rf. Source rf mainly controls ion flux, whereas bias rf controls ion bombardment energy. After a sputtering clean removes native oxide at the

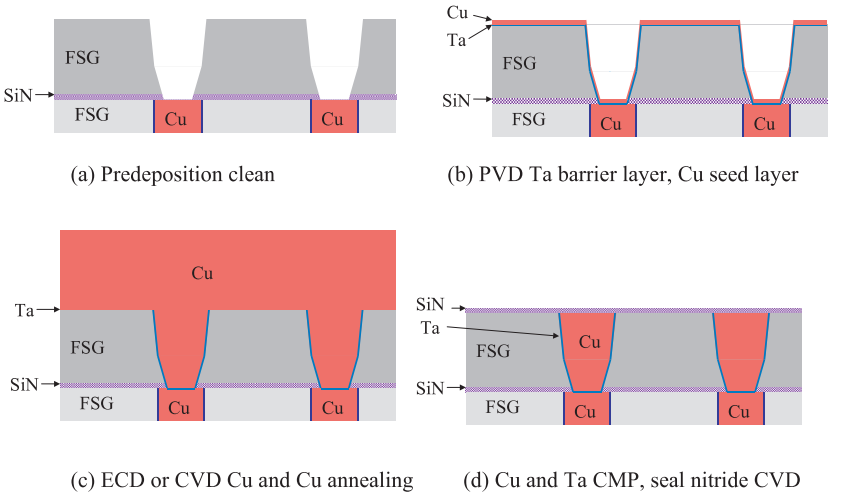


Figure 11.38 Dual damascene copper interconnection process.

bottom of the metal surface, the wafer is transferred to a PVD chamber in an UHV environment, which prevents the metal surface from reoxidizing. Because barrier and seed layer depositions require a low wafer temperature, an electrostatic chuck (E-chuck) system is desired for the sputtering etch chamber, since it can effectively cool down the wafer during the sputtering etch process; this eliminates the cooling time required for systems without E-chuck. It is difficult to transfer heat effectively in the UHV environment required for PVD processes.

At the beginning of copper metallization, an argon sputtering preclean process is used to remove copper oxide before barrier and seed layer deposition. As technology nodes continue to shrink, low-κ and porous low-κ dielectrics have been integrated with copper to form interconnections. Fabs started to worry about possible contamination issues due to Ar sputtering preclean because this process can sputter copper from the bottom to the sidewall of a via and penetrate the pores causing shorts between copper wires. Also, copper has a high diffusion rate in silicate glass; thus, it can diffuse into the silicon substrate later and cause reliability problems for the IC chip due to device malfunction from heavy metal contamination. Therefore, reactive preclean techniques were developed. By striking plasma with hydrogen (H_2) and helium (He) gases, free hydrogen radicals (H) can be generated at low pressures. Hydrogen radicals can diffuse into the via holes, react with copper oxide, and replace the copper atoms to bond with oxygen. This process forms moisture, which is driven away by the heated wafer surface, and effectively removes native oxide from the copper surface at the bottom of the via. The chemical reaction of hydrogen reduction in copper precleaning processes can be expressed as

$$4\,H + CuO_2 \rightarrow Cu + 2\,H_2O.$$

Reactive precleaning has some advantages, but it also introduces hydrogen, which is flammable and explosive, into the system. Since the wafer heats up during this process, it needs extra time to cool down after precleaning; this cooldown time can affect throughput. Most likely, engineers will continue to use the production-proven Ar sputtering etch for precleaning processes in copper metallization until it no longer works.

11.6.2 Barrier layer

Several barrier materials such as Ti, TiN, Ta, TaN, W, WN, etc., have been studied for adoption as diffusion barrier layers for copper metallization processes to prevent copper from diffusing into the silicon substrate and damaging microelectronic devices. Both tantalum and tantalum nitride can be used for this application. Most fabs use a few hundred angstroms of tantalum or tantalum nitride, or a combination of both films, as a copper barrier layer.

Requirements for the tantalum barrier layer are low contact resistance, good bottom and sidewall step coverage, and film quality integrity on the trench/via sidewalls. An ionized metal plasma chamber is customarily used to achieve high bottom step coverage and reduce contact resistance. Although high deposition temperatures improve deposited film step coverage, quality, and electrical conductivity, low temperatures are desired for barrier layer deposition because they deposit film with smoother surfaces. At lower temperatures, adatoms have lower surface mobility, causing smaller grain size and smoother film surfaces, which are important for copper seed layer deposition. To effectively remove heat generated on the wafer surface due to ion bombardment during deposition, an E-chuck system is desired for the barrier layer deposition chamber. Conductivity of the metal films can be improved by a thermal annealing process after deposition of copper seed and bulk layers.

11.6.3 Copper seed layer

Before bulk copper deposition (either by ECP or by CVD), a copper sputtering deposition process is always needed to deposit a thin seed layer (500 to 2000 Å). The process provides nucleation sites for bulk copper grain and film formation. Without this conducting surface, copper atoms will not stick to the wafer surface very well when they migrate to the surface. If this seed layer is not present, there could be either no deposition or deposition with poor uniformity.

An ionized metal plasma system is used for this application, since it has good bottom step coverage. Deposition is performed at a low temperature to achieve a smooth film surface, which is important for bulk copper deposition with ECP. A rough seed layer surface in the trench and via sidewalls can cause voids during ECP bulk copper deposition. As in the barrier layer deposition, an E-chuck system is preferred in the PVD chamber for copper seeding.

Since copper has low ionization energy, copper vapor can be easily ionized. After initial argon plasma bombardment, argon flow can be turned off, and plasma is almost solely formed with copper vapor. At low pressure, the MFP of a copper

ion is much longer than the via depth (a few thousand angstroms); therefore, copper ions can be thrown to the bottom of the high-aspect-ratio trench and via holes to achieve reasonably good bottom and sidewall coverage and smooth film surface. As feature sizes shrink, via holes could become so narrow that PVD copper can no longer achieve the requirements for a seed layer due to its poor step coverage; CVD or ALD copper processes could be used for this application. This is discussed in a later section.

11.6.4 Copper electrochemical plating

ECP is an old technology still used for metal plating processes in a wide range of industries including hardware, glass, automotive, and electronics. It is extensively used to plate copper layers on both sides of fiber plastic boards for printed circuit board fabrication. ECP processes were developed for copper deposition in IC interconnection processes at the end of the 1990s and became the process of choice in the early 2000s. One of the advantages of ECP is that it is a low-temperature process. This makes it compatible with the low-κ polymeric dielectrics that could be used in future interconnection processes. Other names used for ECP are electrochemical deposition (ECD) and electroplating deposition (EPD).

The copper ECP process involves the following steps. First, a wafer is placed on a plastic wafer holder. A cathode holds the wafer with a conducting seal ring and immerses it into plating solution containing sulfuric acid (H_2SO_4), copper sulfate [$Cu(SO_4)$], and other additives. An electric current flows from the anode (which is a pure copper plate) to the cathode. In the solution, the $Cu(SO_4)$ dissociates to copper (Cu^{2+}) and sulfate (SO_4^{2-}) ions. Copper ions in the solution flow to the wafer surface carrying the electrical current that follows the applied electric field. When they reach the wafer surface, copper ions start to adsorb on the surface, nucleate, and deposit copper film on the copper seed layer (or strike layer). At the same time, some copper atoms on the anode surface are ionized, leave the plate surface, and dissolve in the ECP solution. Figure 11.39 shows the copper ECP process.

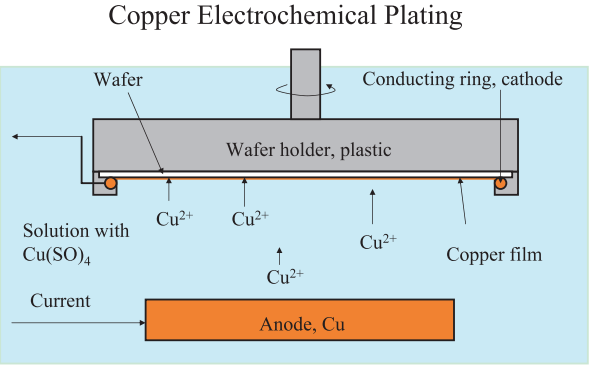


Figure 11.39 Copper ECP process.

During the ECP process, a copper-sulfate solution flows into the high-aspect-ratio trench and via holes due to surface tension at the liquid–solid interface. Sulfuric acid in the solution dissolves the native oxide from the surface of the copper seed layer and exposes the copper underneath, similar to a pre-PVD sputtering etch. The applied electric current drives the copper ions toward the sidewalls and bottoms of the via holes. With the presence of the copper seed layer, a conformal film on the sidewalls and bottoms of the via holes is deposited. The consumed copper ions in the sulfate solution inside the trench and vias are constantly resupplied by the diffusion of copper ions. Thus, the conformal copper film eventually fills the trenches and via holes without voids (see Fig. 11.40).

Increasing the driving current between the two electrodes increases the copper deposition rate. However, high deposition inside the trench and via holes will consume all of the copper ions present there. When copper ions outside the trench and via holes diffuse into them (driven by the concentration gradient), they are more likely to deposit on the gap corners and form overhangs. This can cause trouble in via filling and contribute to void formation. To achieve better gap fill, a pulse current with large forward amperage and small reversed amperage is used. During a reversed current pulse, copper is removed from the wafer surface, reducing the overhang of the gap. Therefore, an ECP process with a pulse-driven current has effects similar to the dep/etch/dep process described in Chapter 10. Some additives such as inhibitors are used to reduce deposition on the corner of the trench and via holes, and other additives such as accelerators are used to accelerate the deposition on the bottom. Thus, the ECP process can deposit copper film from the bottom up to further improve via and trench filling abilities.

Some fabs have attempted to develop electrochemical processes which—after copper fills the via holes and trenches—can reverse the electrodes, making the copper-plated wafer the anode, and the copper plate the cathode. The majority of bulk copper film on the wafer surface can be etched away with an electrochemical etching process.

After copper ECP processing, the wafer is rinsed with deionized (DI) water to remove chemical solutions on the surface and avoid metal corrosion. Then the

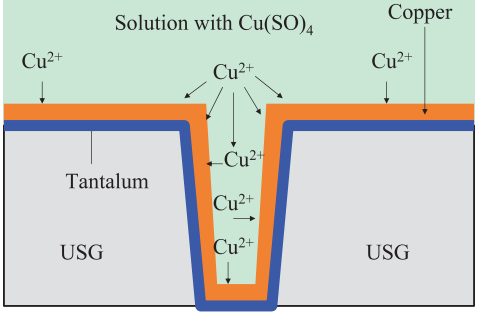


Figure 11.40 ECP via fill.

wafer is spin dried, and a robot removes it from the ECP tool (called a plater). An ECP copper wafer must be annealed in a thermal furnace to densify the film and reduce resistivity before being sent to the CMP bay to remove the bulk copper film on its surface. In an approximately 30-min thermal process at around 250 °C, copper grain size, film density, and conductivity can be increased. ECP processes can provide acceptable films with excellent trench and via fill abilities for copper dual damascene processes in deep submicron IC metallization applications.

Some challenging issues for ECP processes are determining and controlling three aspects—the copper ion concentration in the copper sulfate solution, the effects of additives, and the repeatability of trench and via filling as device dimensions continue to shrink. For small trench and via holes, the liquid surface tension tends to create bubbles inside the via holes that prevent deposition and cause voids.

11.6.5 Copper chemical vapor deposition

Another approach to filling trench and via holes is copper CVD. Frequently used copper precursors are bis-hexafluoroacetyl-acetonate copper [$Cu(hfac)_2$] and hexaflouroacetyl-acetonato-copper-vinyl-trimethylsilane [Cu(hfac)(vtms)]. $Cu(hfac)_2$ is a solid at room temperature, and it sublimes at low heat (35 to 130 °C). Its chemical structure is shown in Fig. 11.41. $Cu(hfac)_2$ can deposit copper by the hydrogen reduction reaction:

$$Cu(hfac)_2 + H_2 \rightarrow Cu + 2\ H(hfac).$$

Copper deposited in this reaction requires a temperature of 350 to 450 °C to achieve low resistivity. This is a little too high for process integration with the polymeric low-κ dielectric needed in IC interconnection applications.

Cu(hfac)(vtms) is a liquid at room temperature. It can deposit copper at a low temperature (<200 °C) with high quality, low resistivity, and good gap and via filling abilities. The chemical reaction is expressed as

$$2\ Cu(hfac)(vtms) \rightarrow Cu + Cu(hfac)_2 + 2\ (vtms).$$

This reaction is reversible; therefore, vtms can be used to dry clean the deposition chamber. By removing copper deposited on the chamber wall and parts

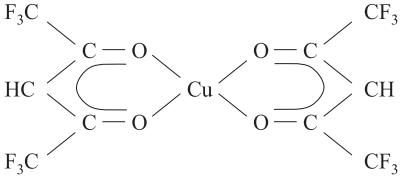


Figure 11.41 Chemical structure of $Cu(hfac)_2$.

inside the chamber, particle contamination caused by film cracking and flaking can be prevented. Because byproducts of the process are stable, they can be easily stripped and recycled for future applications. The Cu(hfac)(vtms) process is the most promising copper CVD process. However, it faces tough competition from the production-proven copper ECP process.

One possible application of copper CVD is depositing a conformal copper seed layer in narrow trenches and via holes, which is an increasingly difficult achievement using PVD processes with shrinking feature size. However, PVD engineers are still attempting to develop a copper seed PVD process that will meet next-generation copper/ULK interconnection requirements. It is quite possible that copper CVD will never become mainstream in the IC manufacturing process.

11.7 Safety

Metal sputtering PVD processes do not use any dangerous chemicals; their processing gases are argon and nitrogen, both of which are considered safe gases. However, metal CVD processes use a variety of chemicals with different chemical hazards.

WF_6, SiH_4, and H_2 are widely used in tungsten and tungsten silicide deposition. WF_6 is corrosive; silane is pyrophoric, explosive, and toxic; and hydrogen is flammable and explosive. TDMAT is commonly used for titanium nitride deposition and is very poisonous. DMAH is a precursor used for aluminum deposition; it is pyrophoric and explosive.

Other safety issues are electrical and mechanical in nature, such as electrical shock from dc and rf power sources, moving parts, and hot surfaces. Detailed information about safety issues related to the processes and processing tools should be obtained from the equipment providers.

11.8 Summary

- Metallization processes are mainly used in the formation of metal interconnections.
- For mature IC products using Al–Cu interconnections, the most commonly used metals are Al, W, Ti, and TiN.
- For advanced IC chips with copper metallization, the most frequently used metals are Cu, Ta, and/or TaN.
- Metals can be deposited with CVD, PVD, ALD, and EPD processes.
- Higher deposition temperatures can increase metal grain size, which reduces film resistivity.
- Silicides are used to reduce the resistance of local interconnections and contact resistance.
- CVD tungsten can fill high-aspect-ratio contact and via holes and is commonly used as metal plugs in interconnection processes.
- Titanium is used to reduce contact resistance.

- Titanium nitride is frequently used as a tungsten diffusion barrier layer and adhesion layer. It is also used as an ARC to improve resolution for metal patterning processes. TiN has been used as hard masks for ULK dielectric etching. It is also used as metal gates for HKMG MOSFETs.
- In a sputtering process, argon ions bombard a metal target, dislodging metal atoms or molecules from the surface as vapor. These atoms or molecules travel to the substrate surface, adsorb, and deposit on the surface as a thin film.
- A dc magnetron sputtering system is the most commonly used sputtering system in metal PVD processes.
- A high-baseline vacuum is required in metal PVD processes to reduce contamination and resistivity of deposited films. For aluminum PVD, UHV baseline pressure (10^{-9} torr) is needed to minimize moisture concentration inside the PVD chamber.
- Titanium nitride can be deposited with a reactive sputtering process by using nitrogen with argon in the plasma.
- CVD titanium nitride is commonly used in submicron IC interconnection processes due to its excellent sidewall step coverage.
- For contact and via application, either a collimated or ionized metal plasma system is used for Ti and TiN PVD to improve bottom step coverage and reduce contact resistance.
- Poor adhesion to silicon dioxide, high diffusion rates in silicon and silicon dioxide, device degradation due to deep-level acceptors caused by copper contamination, and hard-to-dry etching due to lack of simple volatile chemical compounds prevented copper applications in IC chip metallization procesess before the 1990s.
- Improvement in copper barrier deposition, copper deposition, and more importantly, the copper CMP process, paved the way for copper metallization, which is the future of IC metal interconnections.

11.9 Bibliography

D. G. Baldwin, M. E. Williams and P. L. Murphy, *Chemical Safety Handbook for the Semiconductor/Electronics Industry*, 2nd ed., OME Press, Beverly, MA, 1996.

J. Baliga, "Depositing Diffusion Barriers," *Semicond. Intl.* **20**(3), pp. 76–81 (1997).

A.E. Braun, "Copper Electroplating Enter Mainstream Processing," *Semicond. Intl.* **22**(4), pp. 58–66 (1999).

C. Y. Chang and S.M. Sze, *ULSI Technologies*, McGraw-Hill, New York (1996).

A. V. Gelatos, C. J. Mogab, R. Marsh, E. T. T. Kodas, Jain, A, and M. J. Hampden-Smith, "Selective chemical vapor deposition of copper using (hfac) copper(I) vinyltrimethylsilane in the absence and presence of water," *Thin Solid Films* **262**(1–2), pp. 52–59 (1995).

D. James, "Intel pushes lithography limits, co-optimizes design/layout/process at 45 nm," *Solid State Technol.*, 30–33 (March 2007).

J. A. Kittl, W. T. Shiau, D. Miles, K. E. Violette, J. C. Hu, and Q. Z. Hong, "Salicides and alternative technologies for future ICs: Part 1," *Solid State Technol.* **42**(6), pp. 81–92 (1999).

J. A. Kittl, W. T. Shiau, D. Miles, K. E. Violette, J. C. Hu, and Q. Z. Hong, "Salicides and alternative technologies for future ICs: Part 2," *Solid State Technol.* **42**(8), pp. 55–62 (1999).

X.W. Lin and Dipu Pramanik, "Future interconnection technologies and copper metallization," *Solid State Technol.* **41**(10), pp. 63–78 (1998).

R. J. Matyi, L. E. Depero, E. Bontempi, P. Colombi, A. Gibaud and M. Jergel, "The international VAMAS project on x-ray reflectivity measurements for evaluation of thin films and multilayers: preliminary results from the second round-robin," *Thin Solid Films* **516**(22), 7962–7966 (2008).

K. A. Mistry et al., "A 45-nm logic technology with high-k & metal gate transistors, strained silicon, 9 Cu interconnect layers, 193-nm dry patterning, and 100% Pb-free packaging," *Proc. IEDM*, 247–250 (2007).

C. Ryu, H. Lee, K. W. Kwon, A. L. S. Loke, and S. Wong, "Barriers For Copper Interconnections," *Solid State Technol.* **42**(4), pp. 53–56 (1999).

L. Shon-Roy, A. Wiesnoski, and R. Zorich, *Advanced Semiconductor Fabrication Handbook*, Integrated Circuit Engineering Corp., Scottsdale, AZ (1998).

See http://www.webelements.com for further information on the elements in this chapter.

11.10 Review Questions

1. List the four metals commonly used in IC chip manufacturing.
2. Why does aluminum always alloy with copper? Why does it sometimes also alloy with silicon?
3. What is the advantage of an e-beam evaporator over a filament evaporator?
4. Why is tungsten used for metal plugs to connect different layers of conductors?
5. List at least four applications of titanium nitride.
6. What is the metrology tool used to measure sheet resistance?
7. Describe the sputtering deposition process.
8. Explain why a dc sputtering system cannot be used to deposit dielectric materials.
9. In what application does Ti and TiN PVD need a collimated sputtering system?
10. Why does an aluminum alloy sputtering chamber require a UHV?
11. What could happen to deposited film if the argon gas line has a small leak?
12. If the power in a dc magnetron sputtering system is increased, how will the deposition rate change?

13. If the deposition temperature in a dc magnetron sputtering system is increased, how will the metal film sheet resistance change?
14. List the reasons that hampered the application of copper in IC fabrication before 1990s.
15. Compare copper metallization and the standard aluminum alloy-tungsten metallization; what are the main differences?
16. List three methods used in copper deposition.
17. For HKMG processes, which approach uses CVD Al, gate-first or gate-last?
18. What are work-function metals for pMOS and nMOS?
19. Why do fabs use Ta in gate-last HKMG MOSFETs?
20. How is it that DRAM can use tungsten wires while CMOS logic ICs cannot?

Chapter 12
Chemical Mechanical Polishing

CMP is a removal process that strips small portions of film deposited on a wafer by a combination of chemical reaction and mechanical polishing, thus making the surface smoother and more planarized. It is also used to remove bulk dielectric film on the surface to form STI on the silicon substrate, and remove bulk metal film from the wafer surface to form metal interconnection lines or plugs in the dielectric film. This chapter covers CMP processes.

Objectives

After finishing this chapter, the reader will be able to:

- list the applications of CMP
- describe the necessity of dielectric planarization
- describe the basic structure of a CMP system
- list the differences between oxide CMP slurry and metal CMP slurry
- describe the oxide CMP process
- explain the metal polishing process
- describe the importance of post-CMP cleaning
- describe CMP applications in copper metallization
- list two CMP processes required in gate-last high-κ, metal gate MOSFET fabrication processes.

12.1 Introduction

After wafers have been sawed from a single-crystal silicon ingot, there are many processing steps to prepare the flat, shiny, and defect-free wafer surfaces for IC processing. Along with wafer edge rounding, lapping, and etching, there is always a CMP process used in the final steps of wafer manufacturing that planarizes the wafer and virtually eliminates all defects from the surface caused by the wafer sawing process. However, the initial reaction of the industry to CMP processing was not encouraging. ("You're doing *what* to my wafers?!")

Traditionally, direct contact with the frontal surface of a product wafer in semiconductor fabs was strictly forbidden. The reason is simple and straightforward: any direct contact can create defects and particles that reduce the yield of IC chips

and cause lower profit margins for the IC fab. In the case of CMP, a wafer is not only held face down and pushed forcefully against a rotating polishing pad, but it also endures this process in the presence of alkaline or acidic polishing slurry, which contains large amounts of silica or alumina particles. To the initial surprise of skeptics, the CMP process not only planarizes wafer surfaces (as it was designed to accomplish), but it also reduces defect density and improves IC chip yield.

As the technology matured, most IC companies gradually adopted the CMP process as a standard routinely performed in semiconductor fabs. A brief overview of CMP development history, advantages of CMP, and applications of CMP processes are given in this section.

12.1.1 Overview

After the 1980s, more than two metal layers were required to connect the increasing number of transistors in IC chips. One of the great challenges was planarization of the dielectric layers between the metal layers. It is very difficult to achieve high-resolution photolithography on a rough surface due to the DOF requirements of optical systems. A rough dielectric surface can also cause problems in metallization because metal PVD processes usually do not have good sidewall step coverage. Thinner metal lines on the sidewall have a higher current density and are more vulnerable to electromigration issues.

Figure 12.1 shows a brief summary of the IC manufacturing flow and makes it clear that CMP is a very important part of the process. Wafers come to CMP from a thin-film process, either dielectric or metal thin-film deposition. In most cases, wafers are transported out of the CMP bay into a photo bay or thin-film bay. Dielectric CMP is followed by either photolithography or metal thin-film deposition, and metal CMP is only followed by dielectric thin-film deposition.

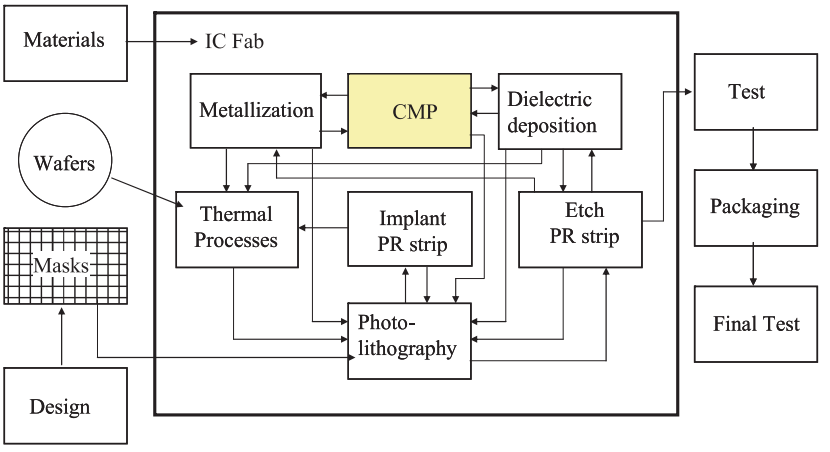


Figure 12.1 CMP applications in IC manufacturing flow.

Several methods, such as thermal flow, sputtering etchback, photoresist etchback, and SOG etchback, have been used for dielectric planarization. Dielectric CMP processing, pioneered by IBM in the mid-1980s, was developed for dielectric planarization applications. In fact, some fabs would prefer that the acronym CMP stand for chemical mechanical planarization.

Tungsten has been used to form metal plugs to connect different layers of conductors. CVD tungsten fills narrow contact and via holes and also covers the surface of the entire wafer. To remove bulk tungsten film from the wafer surface and form tungsten plugs, fluorine-based plasma etchback processes were developed and widely used in IC fabrication. Later, tungsten CMP (WCMP) processes were developed for bulk tungsten removal. WCMP replaced tungsten etchback for plug formation processes in a very short time because of its superior performance in yield improvement, giving fab managers enough justification to make the switch due to cost savings.

Figure 12.2 illustrates that many CMP processing steps are required to fabricate a CMOS IC chip with Al–Cu interconnections. It starts with two dielectric CMP processes, STI dielectric and ILD0. Then, every metal layer requires two CMP processes, a dielectric CMP and a WCMP. For the four-layer metal IC chip illustrated in Fig. 12.2, at least eight CMP processing steps are required: five dielectric and three tungsten CMP processes.

Figure 12.3 shows a cross section of an advanced CMOS chip with HKMG, selective epitaxial S/D, contact bottom silicide, and copper/ULK interconnection,

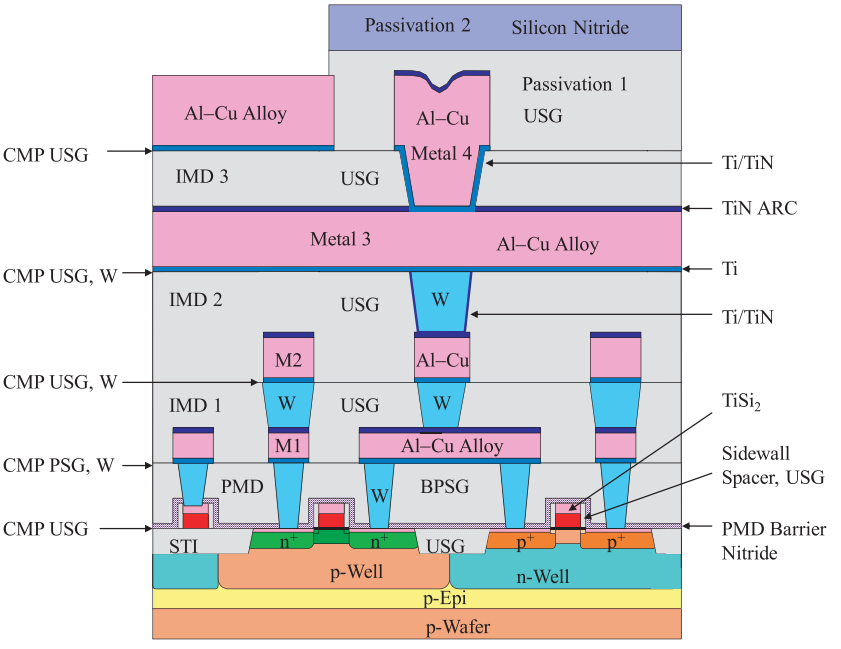


Figure 12.2 CMP applications in CMOS IC chip processing.

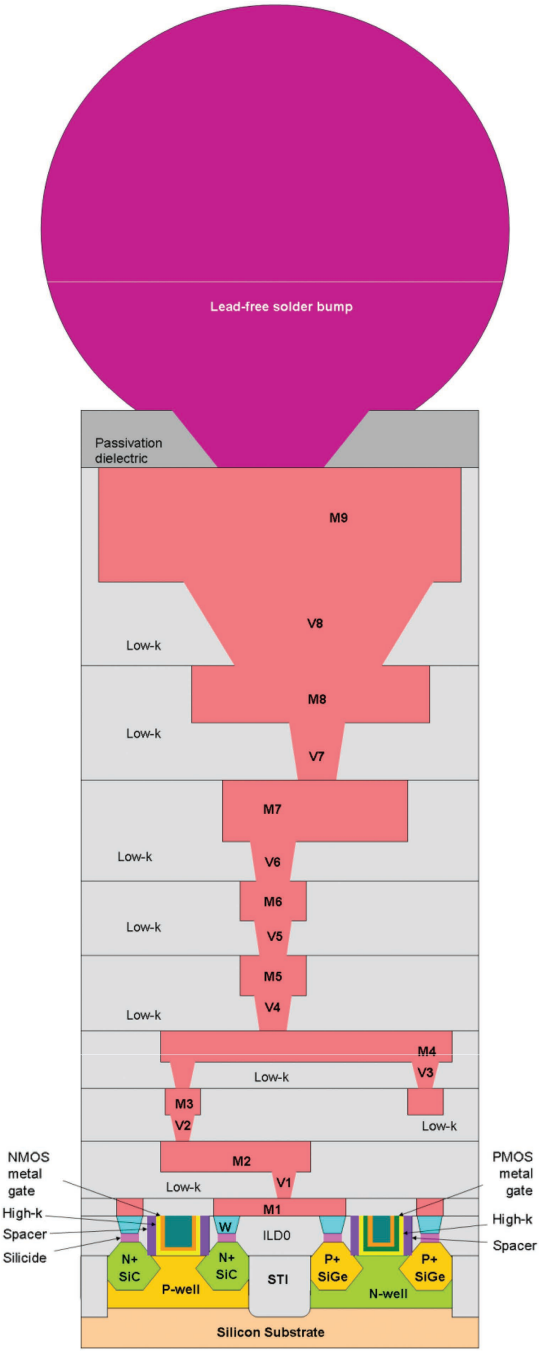


Figure 12.3 CMP applications in HKMG CMOS with Cu interconnection.

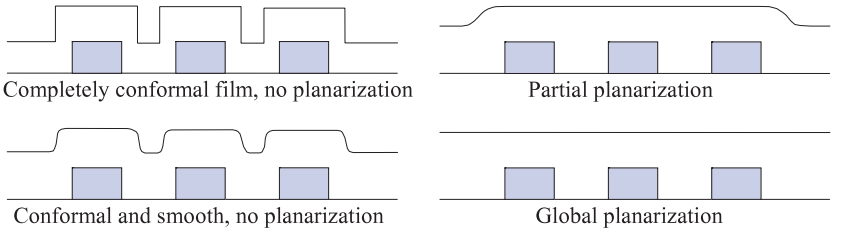


Figure 12.4 Definition of planarization.

which requires more metal CMP processes than dielectric CMP processes. This nine-metal-layer chip requires two dielectric CMP processes, STI and ILD0. The metal needs to be polished 11 times: MG CMP, WCMP, and one Cu CMP for each metal interconnection layer.

12.1.2 Definition of planarization

Planarization is a process that removes surface topological features and smoothes and flattens the surface. The degree of planarization indicates the flatness and smoothness of a wafer surface, particularly for a dielectric thin-film surface after deposition over a patterned wafer surface. The definition of planarization is illustrated in Fig. 12.4.

The degrees of planarity are defined in Table 12.1 and Fig. 12.5, where R is resolution. Smoothing and local planarization can be achieved by thermal flow and etchback processes. Feature sizes smaller than 0.35 μm require global planarization, which can only be achieved by CMP.

12.1.3 Other planarization methods

Thermal flow has long been used for ILD0 planarization. By heating a wafer to a high temperature between 800 to 1000 °C, the doped silicate glass (either PSG or

Table 12.1 Degrees of planarity.

Planarity	R (μm)	θ
Surface smoothing	0.1 to 2.0	>30 deg
Local planarization	2.0 to 100	30 to 0.5 deg
Global planarization	>100	<0.5 deg

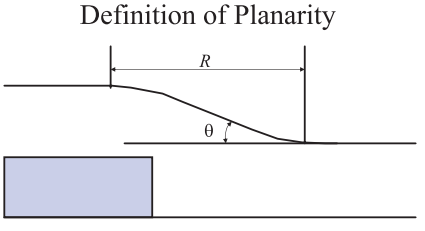


Figure 12.5 Thin-film surface topography.

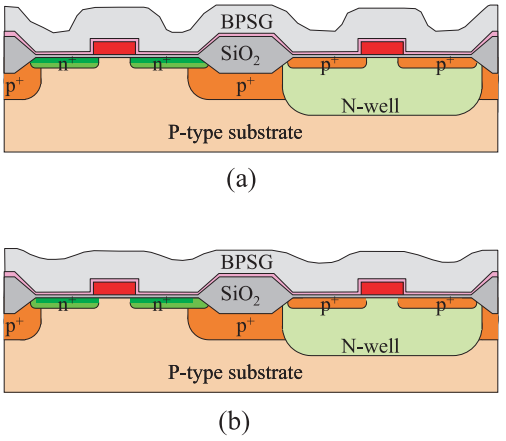


Figure 12.6 Planarization by thermal flow in CMOS IC chip processing: (a) as BPSG is deposited and (b) after reflow.

BPSG, depending on the allowed reflow temperature) becomes soft and starts to flow due to surface tension, as shown in Fig. 12.6.

There are several limitations in thermal reflow planarization. Planarization is mainly determined by reflow temperature and dopant concentration. Higher temperatures give better planarization results. However, high temperatures also cause degradation of transistor performance due to excessive dopant diffusion. To lower the reflow temperature, higher dopant concentrations are needed. These higher concentrations can cause metal corrosion issues if the phosphorus concentration is too high (>7 wt%) due to the phosphoric acid formed when phosphorus oxide (P_2O_5) reacts with moisture (H_2O). They also can cause surface defects due to crystallization of the boric acid formed when boron oxide (B_2O_3) reacts with moisture, if the boron concentration in BPSG is too high.

Because the reflow temperature is far higher than the melting point of aluminum, it cannot be used to planarize the dielectric after the first aluminum alloy layer. Another planarization method is needed for the ILD.

Argon sputtering etchback (ion mill) has been developed and used for ILD planarization. In the sputtering etch process, energetic argon ions bombard the wafer surface, chip off the corners of the gap, and taper the gap openings. This allows subsequent CVD processes to fill the gap, leaving a reasonably planarized surface. Reactive ion etchback processes with CF_4/O_2 chemistry can further planarize the dielectric surface. Figure 12.7 illustrates this etchback planarization process. A detailed description of this dep/etch/dep/etch planarization method is given in Chapter 10.

Photoresist etchback is another method employed to planarize the dielectric surface. After the deposition of the dielectric layer, a layer of photoresist is spin coated on the wafer surface. Due to surface tension, the liquid photoresist can fill the gaps and retain a very flat surface. After the baking process, the

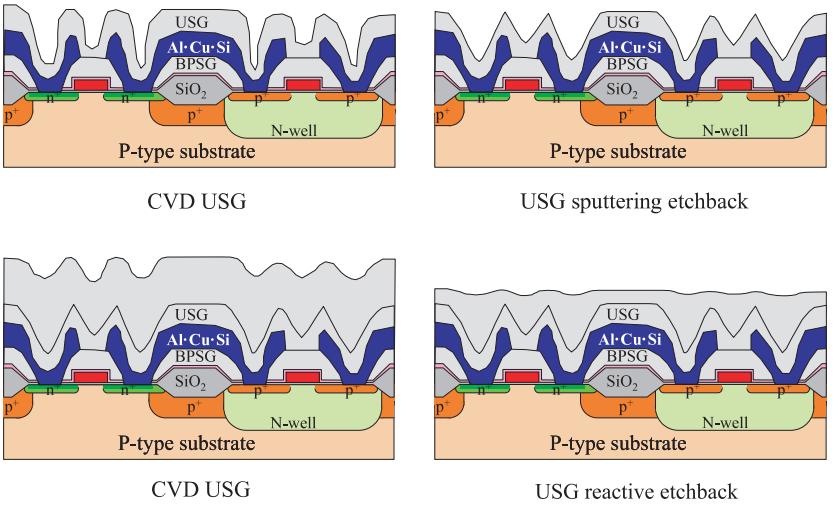


Figure 12.7 Dep/etch/dep/etch gap fill and planarization methods.

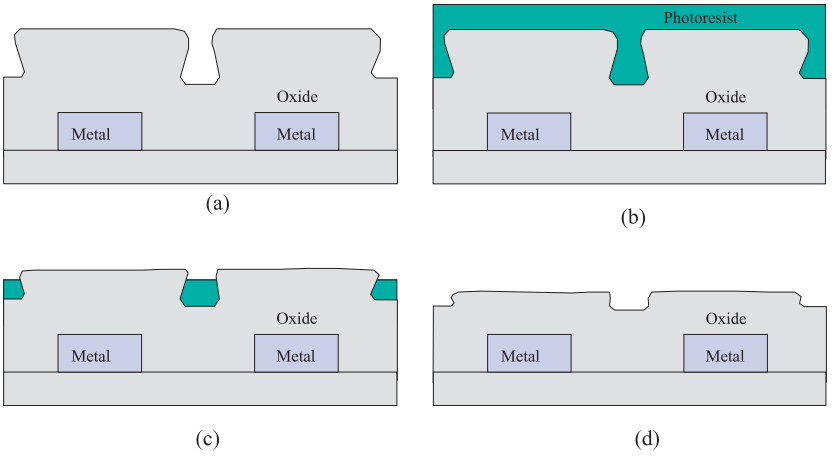


Figure 12.8 Photoresist etchback planarization process: (a) as oxide film is deposited, (b) photoresist coating, (c) photoresist and oxide etch, and (d) after etchback.

photoresist becomes a solid thin film with a planarized surface coated on the wafer surface. Using a plasma etch process with CF_4/O_2 chemistry, silicon oxide can be anisotropically etched by fluorine radicals, and the photoresist can be anisotropically etched by oxygen radicals. By adjusting the gas flow ratio of CF_4 and O_2, theoretically, close to 1:1 etch selectivity of silicon oxide and photoresist can be achieved. Therefore, after etchback, the oxide surface can be planarized. Figure 12.8 shows the photoresist etchback process.

When fluorine radicals start to etch silicon oxide, oxygen replaced by fluorine is released from the etched oxide film. These extra oxygen radicals help to etch the photoresist and cause a higher photoresist etch rate. This is the reason that photoresist etchback cannot achieve a very high degree of planarization the way it is designed. However, after photoresist etchback, the dielectric film surface is significantly flatter than when it was first deposited. In some cases, fabs repeat the photoresist etchback process one or more times to achieve the desired planarization result.

By using SOG to replace photoresist, SOG etchback processes can help both the gap fill and planarization of ILD. One advantage that the SOG etchback process has over photoresist etchback is that a small portion of SOG can stay on the wafer surface to fill narrow gaps between the metal lines. PECVD USG liner and cap layers are used in SOG processes, and ILD with a sandwich structure of USG/SOG/USG can fill gaps and has a reasonable planarized surface. In some cases, two SOG coating, curing, and etchback processes are used to meet gap-fill and planarization requirements. Figure 12.9 shows an SEM photograph of an IC chip with SOG gap fill and planarization.

12.1.4 Necessity of chemical mechanical polishing

As device dimensions reduce, the resolution of photolithography becomes higher. Using Eq. (6.1), $R = K_1\lambda/\text{NA}$, resolution can be improved by increasing the NA of an optical system or by reducing the exposure light wavelength λ. From Eq. (6.2),

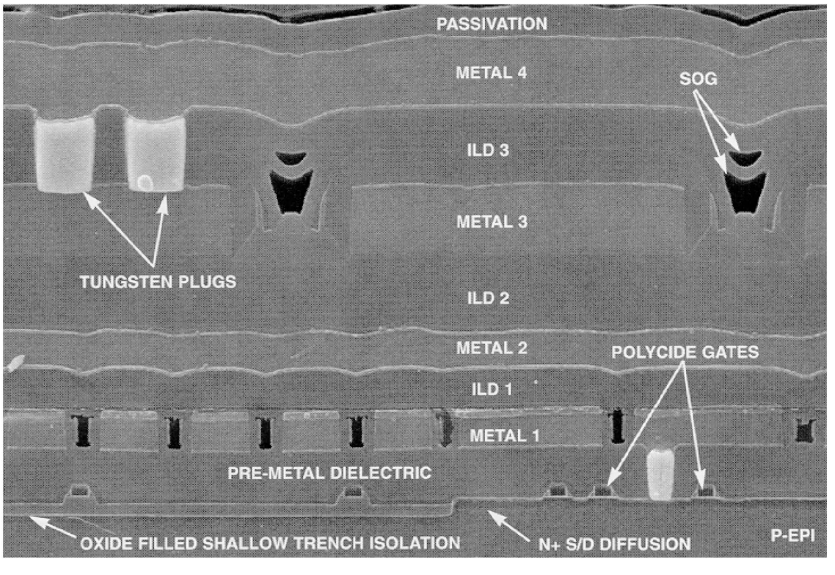


Figure 12.9 SOG gap fill and planarization of ILD. (Integrated Circuit Engineering Corporation)

DOF = $K_2\lambda/2(\text{NA})^2$, it is obvious that both approaches reduce the DOF of an optical system. With a simple calculation using Eqs. (6.1) and (6.2), the DOF is found to be about 2083 Å for 0.25-μm resolution, and 1500 Å for 0.18-μm resolution. Assume that $K_1 = K_2$, $\lambda = 248$ nm (DUV), and NA = 0.6. Therefore, when the patterning feature size is smaller than one-quarter of a micron, or 250 nm, the surface topography must be controlled to less than 2000 Å to ensure the required photolithography resolution. For feature sizes larger than 0.35 μm, other planarization methods can be used to meet the DOF requirements of photolithography. When the feature size is smaller than 0.25 μm, the required planarization level can only be achieved by using CMP processes.

12.1.5 Advantages of chemical mechanical polishing

CMP results in a planarized wafer surface, which allows high resolution of the photolithography process. A planarized surface also eliminates the high resistance and electromigration problems of metal lines caused by the sidewall thinning associated with poor step coverage of metal PVD processes (shown in Fig. 12.10).

A planarized surface can also eliminate the requirements of excessive exposure and development to clear the thicker photoresist regions due to dielectric steps. This improves the resolution of via holes and metal line patterning processes, as shown in Fig. 12.11.

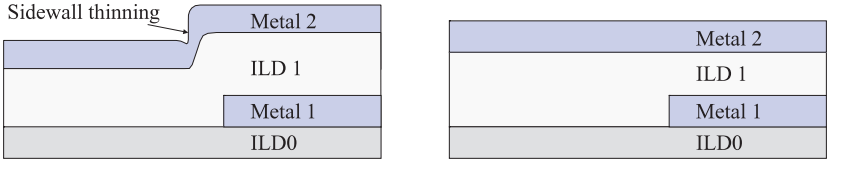


Figure 12.10 (a) Sidewall thinning of a metal line caused by a nonplanarized surface. (b) No such issue with CMP-planarized ILD1.

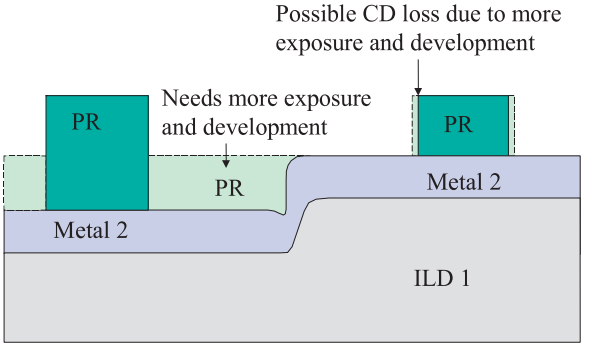


Figure 12.11 Overexposure and overdevelopment due to a nonplanarized surface.

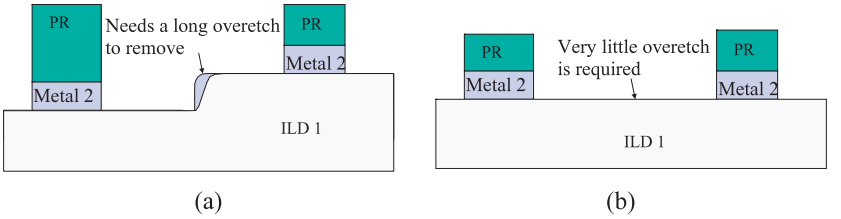


Figure 12.12 Surface planarization and overetch requirements (a) without CMP and (b) with CMP.

A planarized surface allows more uniform thin-film deposition, which helps to sharply reduce any required overetch time. This reduces the chance of undercut or substrate loss associated with long overetch steps during the etch processes, as shown in Fig. 12.12.

Planarization with CMP can minimize defects and improve yield by reducing the processing problems in thin-film deposition, photolithography, and etch processes. The application of CMP also widens IC chip design parameters.

The CMP process can effectively reduce defect density. Inherently, the CMP technique can remove asperities, stringers, and foreign particles from a wafer surface. However, CMP can introduce defects of its own, such as scratches, residues, delamination, dishing and erosion, etc. Large foreign particles can cause scratches, and too much downward force can cause delamination. If the pattern densities in the chip vary from one area to another area, dishing and erosion can happen during the CMP process. CMP systems need to be in conjunction or integrated with appropriate post-CMP cleaning processes to ensure that wafer surfaces are free of slurry residues and other impurities.

12.1.6 Applications of chemical mechanical polishing

CMP processes are routinely performed in IC chip manufacturing. They are used to remove bulk oxide film in STI formation processes. A detailed STI description and illustration is given in Chapter 13. CMP processes are also used to planarize ILD layers in Al–Cu interconnection processes. This allows higher lithography resolution in contact and metal patterning, and easier metal deposition because CMP processes can effectively remove dielectric steps. CMP processing replaced the RIE tungsten etchback process and since the late 1990s has been widely used to remove bulk tungsten and TiN/Ti glue/barrier layers from the wafer surface to form tungsten plugs. Figure 12.13 shows the applications of CMP processes from the cross section of a CMOS chip with a traditional aluminum alloy interconnection.

In DRAM manufacturing processes, polysilicon is frequently used to form contact plugs. After polysilicon deposition fills the contact holes, a polysilicon CMP process removes the polysilicon on the surface and keeps the polysilicon in the contact holes as conducting plugs in DRAM array areas. For recessed-gate

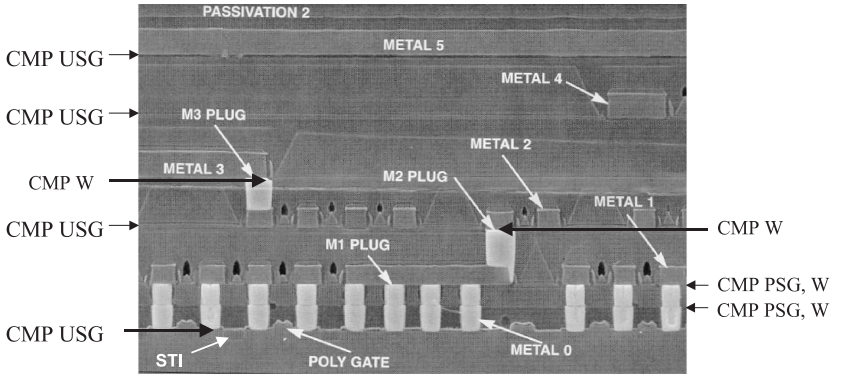


Figure 12.13 CMP applications in IC chip fabrication (Integrated Circuit Engineering Corp.).

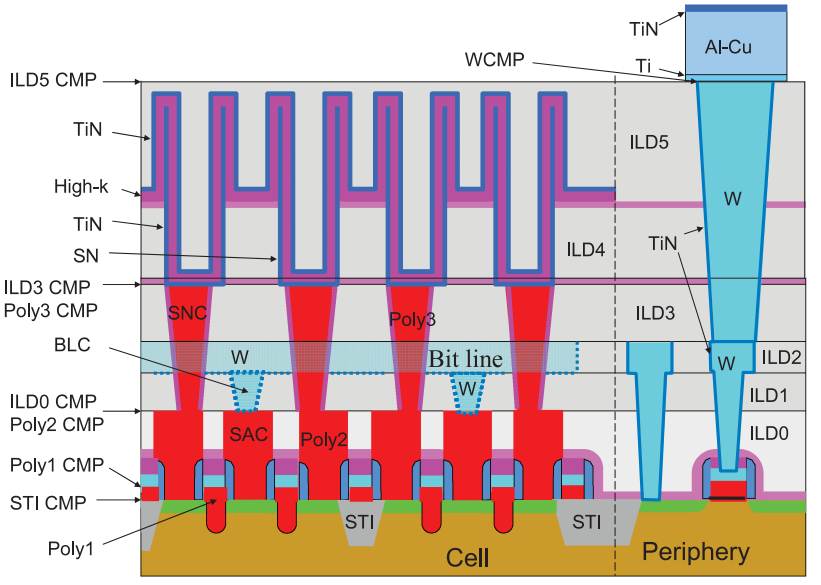


Figure 12.14 CMP applications in a recessed-gate stacked capacitor DRAM.

DRAM, after poly1 fills the recessed trench of the array transistor, a polysilicon CMP process is needed to planarize the polysilicon surface before WL metal deposition. Figure 12.14 illustrates CMP applications in recessed-gate DRAM.

One of the most important applications of CMP processes are copper interconnections. Because copper is very difficult to dry etch, dual damascene has become the process of choice for copper metallization in IC chip manufacturing. The term damascene originates from the ancient Middle East in today's Damascus,

capital of Syria. People there invented a process to decorate sword surfaces with gold inscriptions by using diamond to cut trenches into the steel sword surface, polishing gold into the trenches, washing away the gold on the surface, and leaving gold in the trenches. After this process, the sword was decorated with inscriptions on its surface. This technique has been used in the jewelry industry to place gold inscriptions on gemstone surfaces and is called a damascene process. In fact, the tungsten plug formation process is a damascene process.

For copper applications, the dual damascene process is used. It uses two dielectric etch processes, one for via etch and one for trench etch. After dielectric etches, metal layers (PVD Ta/PVD Cu/ECP Cu) are deposited into via holes and trenches. After metal annealing, a metal CMP process removes the copper and tantalum barrier layer from the wafer surface and leaves copper lines and plugs embedded inside the dielectric layer. Figure 12.15 illustrates the dual damascene copper metallization process.

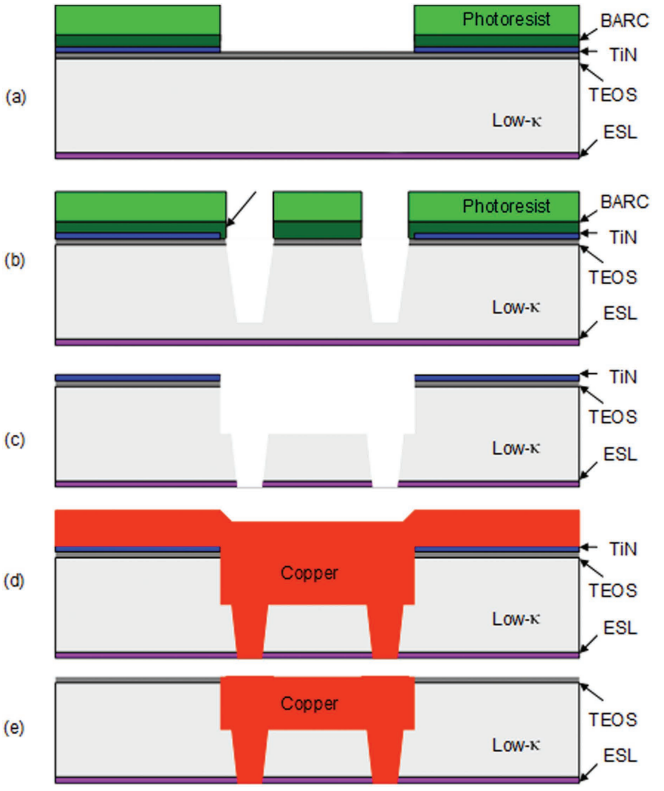


Figure 12.15 Copper metallization with a metal hard mask: (a) etch BARC and metal hard mask with trench mask, (b) etch via halfway with via mask, (c) etch trench and via with metal hard mask and ESL breakthrough, (d) metal deposition, and (e) metal CMP.

12.2 Chemical Mechanical Polishing Hardware

12.2.1 Introduction

A CMP system consists of a polishing pad, a rotating wafer carrier (which holds the wafer face down against the polishing pad), and a slurry dispenser unit. Figure 12.16 illustrates the most commonly used CMP system, with a polishing pad affixed on a rotating polishing table or platen.

CMP processes in the IC industry use water-based polishing slurries with abrasive particles and chemical additives. Slurry is dispensed onto the polishing pad surface, and the wafer is pressed face down against the polishing pad. Both the platen and wafer carrier rotate, usually in the same direction. The combined effect of mechanical abrasion and chemical etch removes material from the wafer surface. Protruding areas receive more mechanical abrasion and are removed faster than recessed areas; this helps planarize the wafer surface.

12.2.2 Polishing pad

A polishing pad is made of a porous, flexible polymer material, such as cast and sliced polyurethane or urethane-coated polyester felt. The pad properties directly affect the quality of the CMP process. The polishing pad materials must be durable, reproducible, and compressible at the processing temperature. The primary processing requirement is high topographic selectivity to achieve surface planarization.

The main qualities of polishing pads are hardness, porosity, fillers, and surface morphology. A harder polishing pad allows for higher removal rates and better within-die (WID) uniformity, while a softer pad allows for better within-wafer (WIW) uniformity. Extremely hard pads cause wafer scratches more easily. The hardness of a pad can be controlled by changing the pad chemical compositions or by changing the cellular structure. Porous cells in the pad absorb polishing slurry and help transfer slurry to the wafer surface, especially at the contact points of the pad and wafer. This is similar to holes in a bath sponge helping deliver liquid soap to the skin. Filler materials can be added to the polymer to improve

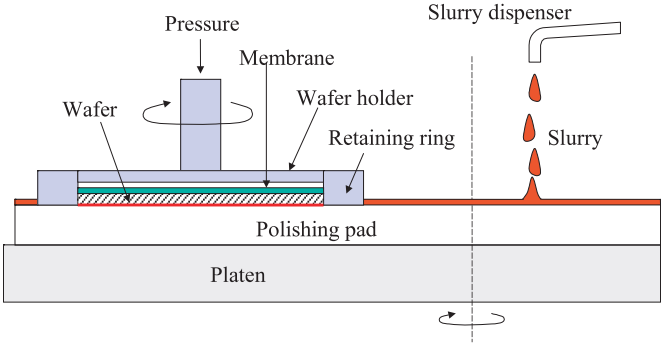


Figure 12.16 Schematic of a CMP system.

its mechanical properties and tailor the pad properties to meet specific process requirements. Polishing pad surface roughness determines the conformality range. Polishing pads with smoother surfaces have a shorter conformality range, which results in poorer topographical selectivity and causes fewer planarization polishing effects. Pads with rougher surfaces have a longer conformality range and better planarization polishing results. Figure 12.17 illustrates the pattern removal effect of rough and smooth polishing pads.

During processing, the pad becomes smoother due to polishing. Therefore, it is necessary to have a reconditioning process to recreate the rough pad surface. Most CMP tools have an in-situ pad conditioner for each polishing pad. The conditioner resurfaces the polishing pad, removes the used slurry, and supplies the surface with fresh slurry. Figure 12.18 shows the position and movement of the pad conditioner on a polishing pad.

The slurry-free polishing pad is composed of four layers: a microreplicated abrasive, a rigid layer, a resilient layer, and a self-adhesive backing layer. Since the abrasive particulates come from the pad surface, only DI water or a basic solution needs to be applied to the pad surface during the CMP process. Advantages of the slurryless pad are that it dramatically simplifies the requirements of slurry storage, delivery, and mixture processes, and it can be used in existing CMP systems with limited retrofitting. Some CMP systems are designed with a roll of slurryless polishing pads that roll during CMP processing and do not require a conditioning system.

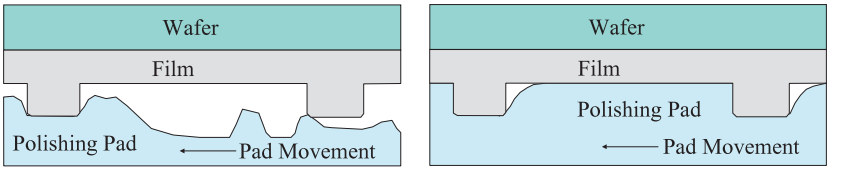


Figure 12.17 Rough pad (left) and smooth pad (right).

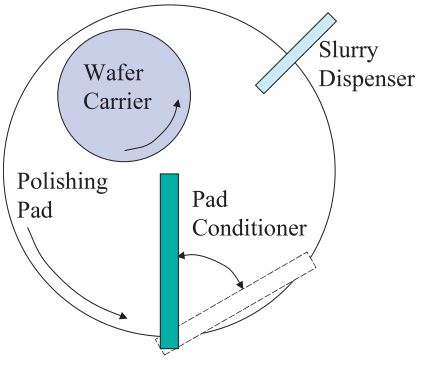


Figure 12.18 Polishing pad and pad conditioner.

12.2.3 Polishing head

A polishing head is also called a wafer carrier and consists of a polishing head body, retaining ring, carrier membrane, and a downward force driving system, as shown in Fig. 12.19. Carrier membranes or carrier films are made of polyurethane or rubber-like materials with a tubular structure. The main purpose of a carrier membrane is to support and cushion wafers in the carrier. Otherwise, the metal polishing head with its downward force can cause wafer damage. A flexible carrier membrane conforms to the wafer's backside so that pressure is applied uniformly. The membrane can compensate for wafer deformations, such as bow or warp due to thin-film or thermal stresses, and help to improve the uniformity of the CMP process. The important parameters of a carrier membrane are its porosity and compressibility. An evenly compressible membrane is important to achieve a uniform CMP process. Membranes must be kept clean because slurry particles can reside in them and cause wafer damage.

The retaining ring is made of plastic and prevents the wafer from slipping out of the wafer carrier. This ring can last for several thousand wafer polishings. The wafer is held to the wafer carrier by a vacuum chuck; the pressurized carrier chamber delivers the downward force to the wafer, and pushes the wafer to the polishing pad. A pneumatic system is used to control the position of the retaining ring to independently control the polishing rate near the wafer edge and contributes to the reduction of edge exclusion due to the shadowing effect. Figure 12.20 illustrates the schematic of a polishing head.

12.2.4 Pad conditioner

A pad conditioner uses a rotating disk with diamond coating that sweeps across the surface of a polishing pad to increase roughness of the pad surface and remove the used slurry. During this process, the polishing pad maintains the surface roughness required for achieving good planarization results. Most CMP tools have in-situ pad conditioners. Some CMP conditioners have a stainless steel plate coated with nickel-plated diamond grits on the surface. Diamond CMP conditioners have also been used and are made from stainless steel plates coated with CVD silicon, then layered with diamond grits coated by CVD diamond film. Figure 12.21 illustrates the surface of these pad conditioners.

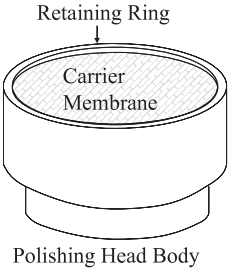


Figure 12.19 Schematic of a polishing head.

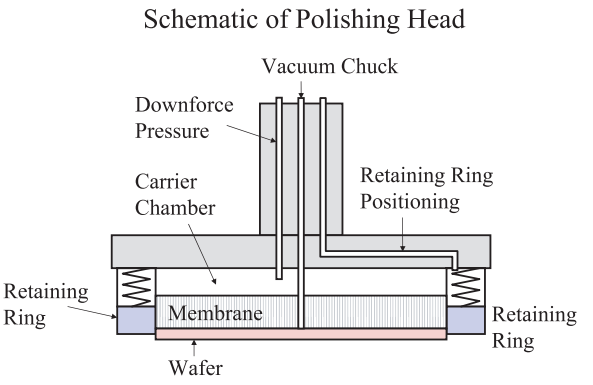


Figure 12.20 Cross section schematic of a polishing head.

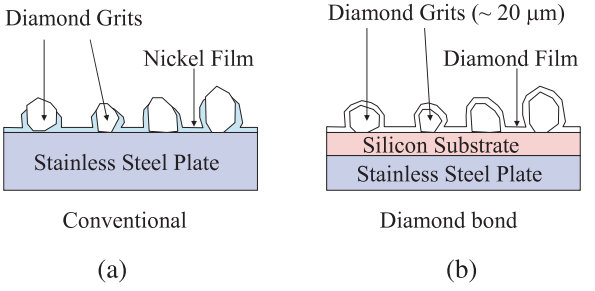


Figure 12.21 Surface of (a) nickel-plated conditioner and (b) diamond-plated conditioner.

A pad conditioner roughens the polishing pad surface while the pad is polishing the wafer. This ensures that roughness of the pad remains unchanged during the polishing process and retains consistent WTW process uniformity.

12.3 Chemical Mechanical Polishing Slurries

Slurry plays a very important role in the CMP process. Particulates in slurry mechanically abrade the wafer surface and remove surface materials. Chemicals in a slurry solution react with surface materials or particulates and dissolve the materials or form chemical compounds that can be removed by abrasive particles. Additives in CMP slurries help to achieve the desired polishing results. In fact, CMP slurries work like toothpaste. Toothpaste provides particulates to abrade the unwanted coating from the surface of the teeth during brushing, while chemical reactions from additives kill gems, remove tartar, and form a protective layer on the teeth.

Slurries used in CMP processes are water-based chemicals with abrasive particles and chemical additives. Different polishing processes require different slurries. Slurries can impact the removal rate, selectivity, planarity, and uniformity

of CMP processes. Therefore, slurries always are engineered and formulated for specific applications. There are two types of slurries in CMP processes: one for oxide removal and one for metal removal. Oxide slurries are alkaline solutions with suspended silica. Metal slurries are acidic solutions with suspended alumina particles. Additives are used in the slurries to control the pH value, which affects chemical reactions in CMP processes and helps to achieve optimized processing results.

Components of slurries are stored in different bottles, such as DI water with particulates in one bottle, additives for pH control in another, and oxidants for metal oxidation in a third bottle. They are flowed into a mixer where the components of the slurries are mixed according to the ratio determined by the processing requirements. A slurry delivery system is illustrated in Fig. 12.22, where LFC stands for the liquid flow controller.

12.3.1 Oxide slurry

The most commonly used dielectric in IC processing is silicon oxide. STI formation needs oxide CMP, and ILD0 needs CMP to planarize the surface before contact photolithography. Slurry for an oxide CMP process was developed from the experiences of the optical industry, which requires a fine polish of silicate glass to make lenses and mirrors for optical equipment. Oxide slurry is a colloidal suspension of fine fumed silica (SiO_2) particles in water with base additives. KOH is used to adjust the pH, and sometimes NH_4OH is used. A diluted basic solution of KOH (usually <1%) is frequently used to control the slurry pH at 10 to 12.

The pH value represents the acidity or alkalinity of an aqueous solution, which ranges from 0 to 14. Neutral solutions such as water have pH value equals to 7. Solutions with pHs lower than 7 are acids; the lower the pH, the stronger the acid. Alkaline solution has pH higher than 7. The higher the pH, the higher the alkalinity, as shown in Fig. 12.23.

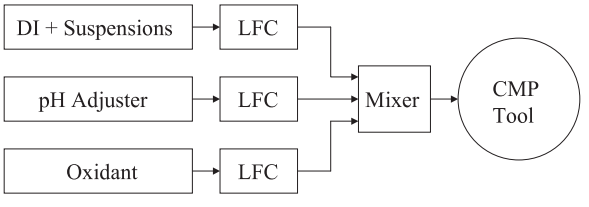


Figure 12.22 Slurry delivery system.

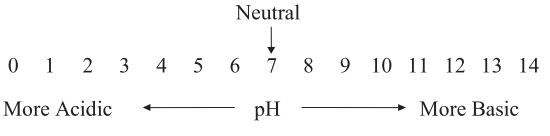


Figure 12.23 Relationship of pH values and acidity or alkalinity.

The fine fumed silica particles that are suspended in solution are abrasives for CMP. Oxide polishing slurries contain approximately 10% solids. With proper temperature control, these slurries can have a shelf lifetime of up to one year.

Fumed silica particles can be formed in a vapor phase hydrolysis of silicon tetrachloride in a hydrogen-oxygen flame. The chemical reaction can be expressed as

$$2\,H_2 + O_2 \rightarrow 2\,H_2O,$$
$$SiCl_4 + 2\,H_2O \rightarrow SiO_2 + 4\,HCl \uparrow .$$

The overall reaction can be expressed as

$$SiCl_4 + 2\,H_2 + O_2 \rightarrow SiO_2 + 4\,HCl \uparrow .$$

At about 1800 °C, these reactions form molten spheres of silicon dioxide. The size of the silica spheres varies from 5 to 20 nm, depending on process parameters. The molten spheres can collide and fuse with one another to form branched, 3D, chain-like aggregates. Figure 12.24 illustrates the fumed silica formation process.

The pH of a slurry can significantly affect dispersion of the silica particles. The particles attain a surface charge, and the sign and magnitude of the charge depend on the pH of the solution. Up to a pH of 7.5, the viscosity of silica slurry in an aqueous medium is high enough to prevent silica particles from dispersing. At a pH higher than 7.5, the silica particles attain sufficient charges to generate electrostatic repulsion, which effectively disperses the slurry. When the pH is higher than 10.7, the particles dissolve and form silicate.

Another type of silica abrasive used in oxide CMP processes is colloidal silica, which is also called precipitated silica. Colloidal silica can be prepared in an alkali silicate solution. At near-neutral pH conditions, silica nucleation forms colloidal silica particles of about 1 to 5 nm in diameter. If the pH is kept just slightly

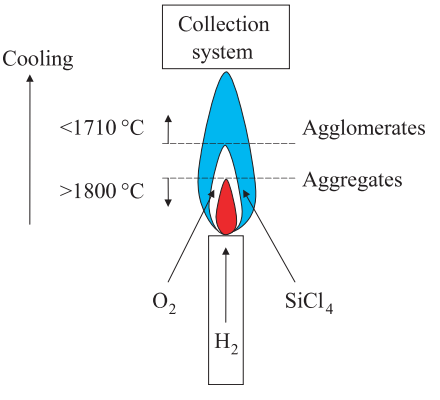


Figure 12.24 Fumed silica formation.

alkali, the colloidal silica particles will not fuse together, and they gradually grow larger in size, 100 to 300 nm. Some particles grow even bigger, and large-particle count (LPC) is an important factor of slurry. Very large particles (>1 μm) need to be filtered out because they tend to generate scratch defects in CMP processes. Figure 12.25 shows the differences between fumed silica and colloidal silica particulates.

For STI oxide CMP processes, high selectivity to nitride is required so that the process can stop on the silicon-nitride polish stop layer, which is also the hard mask layer for STI etching. Slurry with a cerium oxide (CeO_2 or ceria) abrasive has been developed and implemented in IC fabrication. This slurry contains about 5% CeO_2 abrasive in a weak alkaline aqueous media. By using CeO_2 abrasive with resin, the slurry can achieve high topographic selectivity and help to avoid the dishing effect in large field oxide areas, as illustrated in Fig. 12.26. Slurry with a CeO_2 abrasive can achieve better selectivity to nitride than slurry with silica abrasive particles. It also has less dishing effect in STI oxide CMP processes.

12.3.2 Metal polishing slurry

Metal CMP processing is similar to a metal wet etch process. At first, the oxidant in the slurry reacts with metal to form an oxide on the metal surface; then the oxide is removed, exposing the metal surface for oxidation, and the oxide is removed again. Metal polishing slurries are pH-adjusted suspensions of alumina (Al_2O_3). The slurry pH controls the two mechanisms attributed to metal removal: corrosive wet etching and oxidation passivation.

In metal CMP processes, oxidant in the slurries oxidizes metal surfaces. In varied conditions, different types of metal oxides can be formed, and each of them has a different solubility. This causes two competing removal mechanisms. If

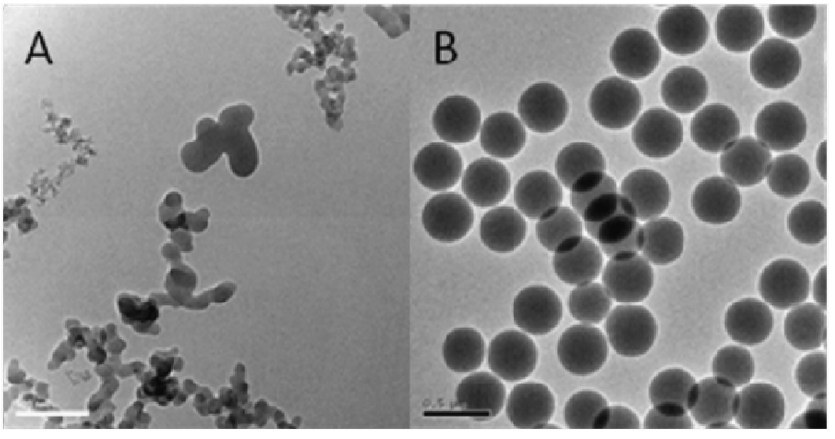


Figure 12.25 SEM images of silica abrasive particles: (a) fumed silica and (b) colloidal silica (Fujimi Corp.).

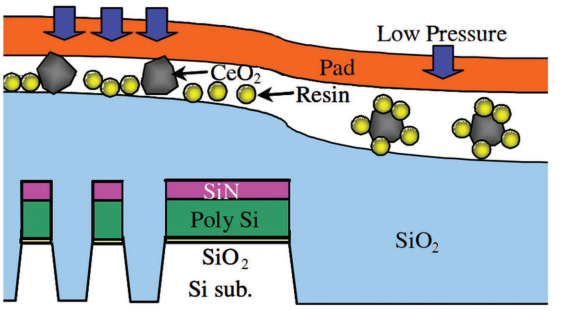


Figure 12.26 STI CMP with CeO_2 abrasive and resin additive (Y. Matsui, et al.).

an oxidation process mainly generates oxide ions that are soluble in the slurry solution, the wet etch will dominate the metal removal process. This is not a favorable process for planarization applications because wet etch is an isotropic process with no topographic selectivity. If the metal oxide is insoluble, the oxide will coat the metal surface and block any further oxidation processes. The fine alumina particles in the polishing slurry mechanically abrade the passivation oxide layer, expose the metal surface, and allow metal oxidation and oxide abrasion processes to be repeated. This chemical mechanical removal process has a high surface topographic selectivity, which is favorable for surface planarization. Additives are used in metal CMP slurries to control the pH and achieve the delicate balance between wet etch, passivation, and oxide removal to optimize the metal CMP processing results.

12.3.3 Tungsten slurry

Tungsten can be passivated by the formation of WO_3 due to chemical reaction with an acidic solution (pH lower than 4) during the tungsten CMP process. For a higher pH, soluble $W_{12}O_{41}^{10-}$, WO_4^{2-}, and $W_{12}O_{39}^{6-}$ ions are formed in the solution, and the tungsten is wet etched at a high etch rate. Figure 12.27 illustrates a potential pH diagram, called a Pourbaix diagram, for tungsten and indicates the passivation and wet etch regime at different pH values and potential. When pH < 2, tungsten is in the passivation regime.

In the presence of an oxidant, such as potassium ferricyanid $[K_3Fe(CN)_6]$, ferric nitrade $[Fe(NO_3)_3]$, and hydrogen peroxide (H_2O_2), the range of pH in which tungsten becomes passivated can be extended to 6.5. By adjusting the slurry pH, both low wet etch rates and high CMP removal of tungsten film can be achieved simultaneously.

Tungsten slurries are quite acidic, with pH levels from 4 to 2. Compared to oxide slurries, tungsten slurries have lower solid contents and much shorter shelf lifetimes. Alumina particles are not in colloidal suspension in a slurry at that pH value. Therefore, tungsten slurries require mechanical agitation prior to and during delivery to the CMP tools.

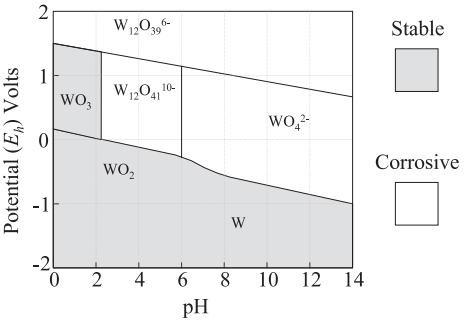


Figure 12.27 Pourbaix diagram for tungsten.

12.3.4 Aluminum and copper slurries

Aluminum slurries are water-based acidic solutions, with H_2O_2 or amines and H_2O_2 as oxidants, and alumina as abrasives. They also have very limited shelf lifetimes, since an H_2O_2 molecule is unstable and tends to dissociate to H_2O, releasing a free oxygen radical. Aluminum CMP has been used in replacement gate or gate-last processes of HKMG to form metal gate electrodes for advanced MOSFETs since the 45-nm technology node.

Figure 12.28 illustrates the potential pH diagram for copper. It can be seen that at $5 < \text{pH} < 13$, copper is in the passivation regime. To achieve consistent polishing process results, a colloidally stable slurry is needed. A colloidally stable alumina suspension can be achieved at a pH just below 7. Therefore, this suspension has only a small window in which copper slurries can achieve both electrochemical passivation and colloidally stable suspension of aqueous alumina particulates.

Copper slurries are acidic solutions with alumina as abrasives. Different oxidants can be used, such as hydrogen peroxide (H_2O_2), ethanol (HOC_2H_5) with nitric acid (HNO_4), ammonium hydroxide (NH_4OH) with potassium ferri- and ferrocyanide, or nitric acid with benzotriazole.

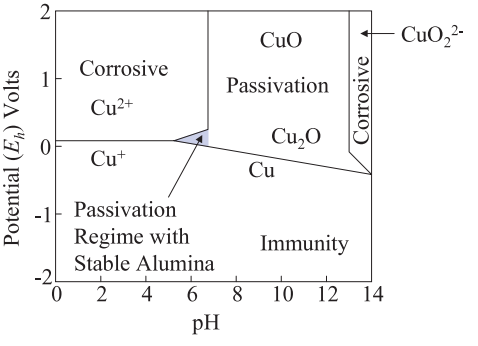


Figure 12.28 Potential pH diagram for copper.

Question: Oxide slurries use silica, while metal slurries use alumina as abrasives. Can oxide slurries use alumina and metal slurries use silica?

Answer: Silica particles form chemical bonds with atoms on the surface of silicate glass. This helps to chemically remove silicon oxide by tearing atoms or molecules from the glass surface with the silica particles, and dissolving the particulates in a high-pH solution. Alumina cannot form chemical bonds with oxide film and is not soluble in high-pH aqueous solutions. Therefore, alumina only provides mechanical abrasion for oxide removal, causing a low oxide removal rate. Both silica and alumina can be used in metal slurry as abrasives. However, using silica can cause a higher oxide removal rate and worsen metal-to-oxide polish selectivity.

By adding aluminate ions Al $(OH)_4^-$ into a slurry, colloidal stability can be significantly improved. Copper CMP using slurries with aluminate-modified colloidal silica abrasives can achieve high removal rates and good planarization and dishing performance.

12.4 Chemical Mechanical Polishing Basics

12.4.1 Removal rate

Mechanical material removal rate R was found by Preston in research related to glass polishing processes. The Preston equation can be expressed as

$$R = K_p \cdot p \cdot \Delta v,$$

where p is the polishing pressure, which is determined by downforce divided by the contact area; K_p is the Preston coefficient, which is related to a specific process and must be determined empirically; and Δv is the relative velocities of the wafer and polishing pad. The Preston equation works very well for bulk film polishing processes. Because protruding portions on a rough surface have a higher polishing pressure than other parts of the surface, the Preston equation reveals that the removal rate of the protruding parts would be higher than that of the rest of the surface. This helps remove surface topography and planarize the surface. Figure 12.29 shows the higher polishing pressure of the protruding parts of the wafer.

Since CMP processes are never purely mechanical processes, the Preston equation does not always describe the process accurately. Chemical interactions always play a crucial role in removal processes, especially in metal CMP processes.

Polish rate can be determined by measuring the film thickness change before and after CMP processing, and dividing that value by the CMP time. For dielectric CMP processes, the removal rate can be monitored in situ by optical reflective interferometry, described in Chapter 10. An optical reflective interferometry system is built into the CMP system to detect the endpoint of a polishing process.

Removal rate of a CMP process is several thousand angstroms per minute. It is mainly determined by the downforce pressure, stiffness of the polishing pad, and

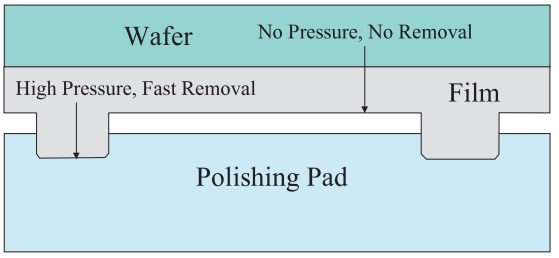


Figure 12.29 Topographic selectivity of a mechanical polish.

amount of applied polish slurry. Different films also have different polishing rates. For instance, different silicate glasses, such as SOG film, PECVD oxide film, and O_3-TEOS oxide film, have different polishing rates. Doped oxide film has different polishing rates from undoped oxide film, etc.

If the removal rate begins to drop gradually during a routine process, the most likely source of the problem is pad surface deterioration. If pad conditioning is adequate, the pad could need replacement.

12.4.2 Uniformity

For 200-mm wafers, 49-point, 3σ standard deviation measurements are used as the definition of uniformity for CMP process qualifications, and for changes in film uniformity before and after CMP processing is monitored. For 300-mm wafers, 121-point measurements might be required. For production wafers, only the uniformity after CMP processing is monitored.

Both WIW and WTW uniformity can be affected by the polishing pad condition, downforce pressure distribution, relative speeds of the wafer and the polishing pad, retaining ring position, and shape of the wafers. By using a harder pad and lower pressure (as low as 2 psi), good global uniformity within less than 3% nonuniformity can be achieved.

12.4.3 Selectivity

Removal selectivity is the ratio of removal rates of different materials. In CMP processing, a large ratio is desired for film that should be removed and material that should not be removed. Removal selectivity is a very important factor in CMP processing. It can significantly affect CMP-induced defects, such as erosion or dishing, and can also be very important for endpoint detection. Slurry chemistry is the primary factor that affects the removal selectivity of a CMP process. For oxide CMP in STI formation processes, high oxide-to-nitride selectivity is very important to ensure that the polishing process stops on the nitride surface. Oxide-to-nitride selectivity is between 3 and 100, depending on slurry type, pad stiffness, downforce pressure, and pad rotation speed. Because only the oxide layer is polished, selectivity is not important in ILD planarization processes in Al-Cu interconnections. For tungsten CMP processes, selectivity of oxide and titanium

nitride is very important. Tungsten-to-TEOS oxide selectivity is very high, from 50 to 200.

For a particular metal CMP slurry, reactivity of the chemical oxidant is critical to control the removal rate and selectivity. This makes the choice of oxidant one of the most critical factors in metal slurry engineering. Selectivity is also related to pattern density. For example, the tungsten-to-oxide removal rate ratio in tungsten CMP processes can be as high as 150:1, based on blanket film polishing results. In reality, the ratio can be much smaller, depending on the pattern density of each material; the higher the pattern density, the lower the removal selectivity. This loss of selectivity can lead to erosion of the tungsten and oxide films, as illustrated in Fig. 12.30.

IC design layout can directly affect erosion problems. Designing opening areas on less than 30% of a chip's surface can help to solve erosion problems. Dummy patterns are a very important part of design for manufacturing (DFM), because dummy pattern placement can significantly affect dishing and erosion issues, and impact product yield.

12.4.4 Defects

CMP processing can remove many defects from a wafer's surface, thereby helping to improve product yield. However, it also can introduce defects such as scratches, residual slurry, particles, erosion, and dishing.

Large foreign particles and hard polishing pads can cause scratches on wafer surfaces. Tungsten can fill scratches in the oxide surface caused by oxide CMP processing and can form microscopic tungsten wires after tungsten CMP that cause short circuits or crosstalk, reducing IC yield.

Improper downforce pressure, worn polishing materials, inadequate pad conditioning, particle surface attraction, and slurry drying can cause slurry residue on wafer surfaces. This can cause defects due to contamination and reduce the IC yield. Post-CMP cleaning is very important to remove slurry residue and improve process yield.

Erosion is mainly caused by the pattern-density-induced selectivity degradation, as shown in Fig. 12.30. It can result in incomplete interconnections in the next layer of metal interconnections because it increases the depth of the via holes. This can cause incomplete via hole etch and create open loops between the different layers in the next dual damascene interconnection, as shown in Fig. 12.31.

The dishing effect happens at a bigger open area, such as at large metal pads or STI oxide in the trenches. Because more material is removed from the center of a

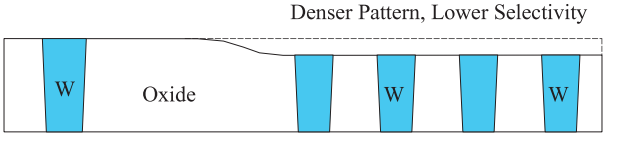


Figure 12.30 Erosion caused by high pattern density.

larger pattern, and the cross-sectional view looks like a dish (shown in Fig. 12.32), this is called the dishing effect.

Both dishing and erosion effects are related to removal selectivity. For example, in tungsten CMP processing, if tungsten-to-oxide selectivity is too high, dishing and recessing effects can occur at the tungsten plugs and pads during tungsten overpolishing, after the many tungsten layers have been removed. If selectivity is not high enough, both oxide and tungsten will be polished away during the overpolishing step, causing erosion problems. Oxide CMP with a high selectivity of oxide to nitride can cause oxide dishing during the oxide overpolishing step in STI formation, as shown in Fig. 12.33.

Dishing and erosion can be measured by an atomic force microscope (AFM). An AFM usually has a microscopic silicon or silicon nitride cantilever with a sharp-tipped probe at the end. The tip of the probe has a radius on the order of nanometers and can scan the proximity of a sample surface without directly touching it, due to the force between the tip atoms and surface material atoms. Interaction can deflect the cantilever along the sample surface. By recording the deflection variation caused by the surface variation, AFMs can measure surface roughness in the nanometer range. A silicon nanotube whisker or carbon nanotube

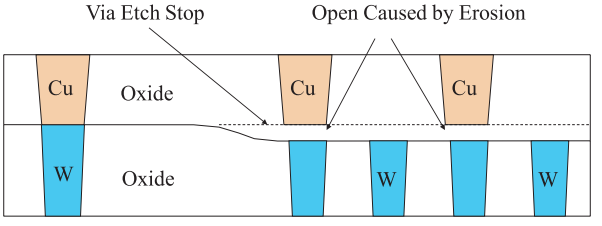


Figure 12.31 Erosion caused by open circuit problems.

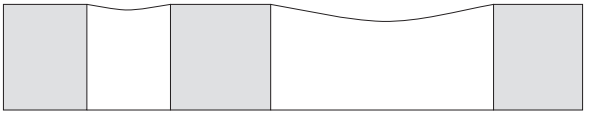


Figure 12.32 The dishing effect.

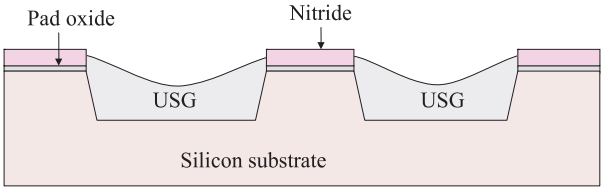


Figure 12.33 Dishing and recessing effects of the STI USG.

whisker can be grown on the AFM tip to help improve measurement resolution of the tiny patterns in the nanometer range. Figure 12.34 shows a cantilever and probe tip without and with a nanotube whisker on an AFM system, and Fig. 12.35 illustrates the AFM measurement sequence.

By scanning a smaller area with dense scanning lines, an AFM image can be formed that shows a 3D profile of a microscopic pattern. AFMs can also be used to measure pattern CD, height and profile of photoresists, and etched patterns. Because they are very slow, AFMs are mainly used for research and development, and engineering troubleshooting. They are also used as golden standard tools to calibrate CD measurement systems such as scatterometry systems.

Particles and defects can be measured by light scattering. Because particles and defects have irregular surface topographies, they scatter incident light while smooth surfaces reflect it. By detecting the scattered light, particles and defects can be monitored on the wafer surface. Figure 12.36 illustrates the idea of particle detection by scattering light.

Because the intensity of the scattered light is very weak, an elliptical mirror is used to collect it. An elliptical mirror collects all of the light scattered from one focus and reflects it to another focus. Particle detection systems are designed so that a laser beam scans the wafer surface vertically at one focus of an elliptical mirror, and a photodetector is placed at another focus. This design allows the user to collect the most scattered light by moving the wafer to detect tiny particles and defects, and then map their position on the wafer surface. Figure 12.37 illustrates this type of particle detection system.

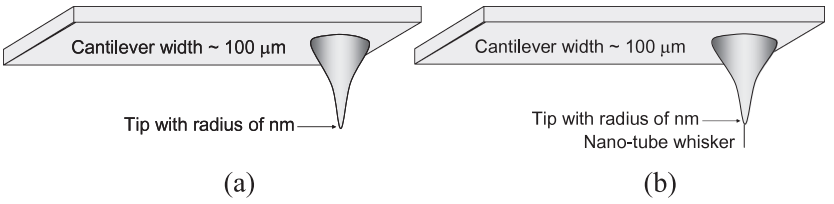


Figure 12.34 Cantilever and probe tip of the AFM: (a) conventional tip and (b) tip with a nanotube whisker.

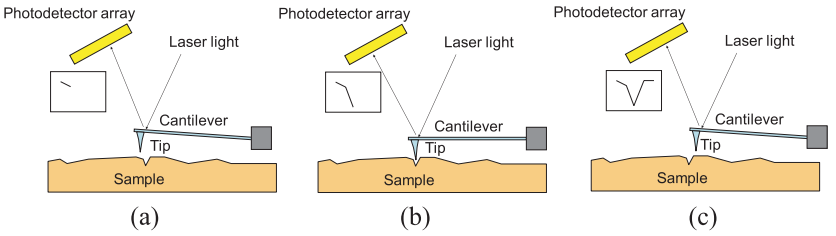


Figure 12.35 AMF measurement sequence.

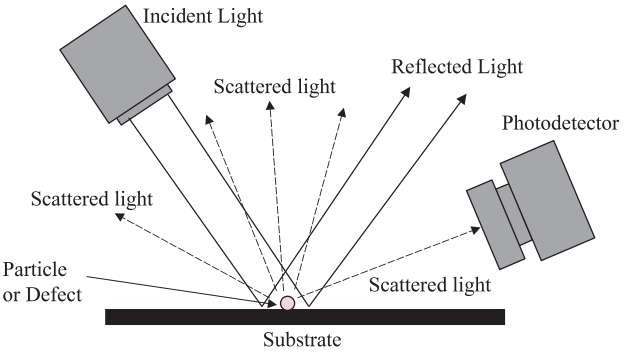


Figure 12.36 Light-scattering particle measurement.

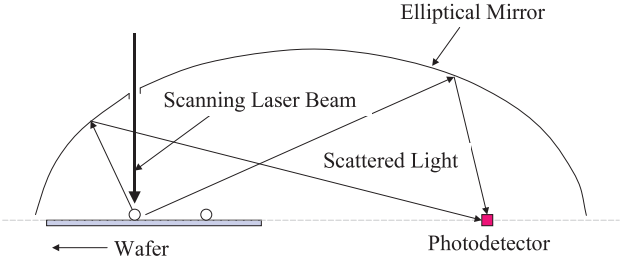
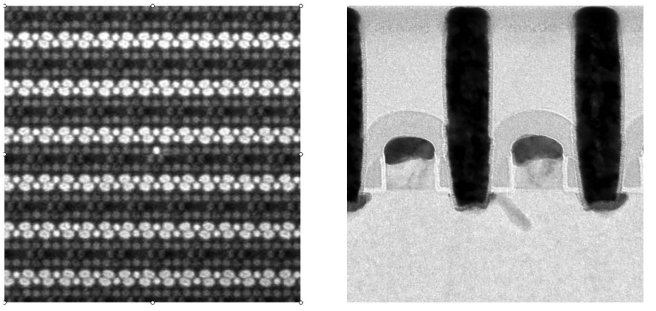


Figure 12.37 Light-scattering particle detector with an elliptical mirror.

SEMs are also widely used for defect detection. The SEM uses an energetic electron beam to scan across the wafer surface and collect the secondary electron emission signals for detection of microscopic features. It is widely used to review physical defects captured by optical inspection and helps to determine the type of defect by looking at it with a high-resolution image.

Because an electron beam charges up the surface to affect the secondary electron emission, an SEM signal of the metal contact plugs connected to the ground can have a higher contrast than a signal from those not connected to the ground. This is called voltage contrast, and it is unique to electron beam inspection (EBI), which can be used to capture electrical defects such as open circuits of contact and via plugs (Fig. 12.31) and junction leakage, which can be caused by junction spiking (see Fig. 11.5). Since the development of 90-nm technology nodes, EBI has been widely implemented after WCMP because it can capture electrical or voltage contrast defects that optical inspections cannot. Figure 12.38(a) shows an nMOS leakage captured by EBI as a bright voltage contrast defect, and Fig. 12.38(b) is the cross-section transmission electron microscope (TEM) image of that defect. The TEM confirmed that the bright tungsten plug came into contact with a leaking n+/p-well junction, and the leak was caused by nickel silicide diffusion along the dislocation fault line that shorted the nMOS S/D junction to the substrate.



(a) (b)

Figure 12.38 (a) Top-down SEM image of an nMOS leak captured as a bright voltage contrast (BVC) defect by post-WCMP EBI. (b) Cross-section TEM image of that BVC defect.

One of the most important advantages of the CMP process is that it can reduce defect density induced by rough surface topography. The benefits of defect reduction by CMP processes far outweigh the defects introduced by them.

12.5 Chemical Mechanical Polishing Processes

There are two types of CMP processes. One is for planarization, which removes part of the film (usually about 1 μm) and planarizes the thin-film surface. Another one is a bulk removal process in which bulk film on the surface is removed by a polishing process, and only parts of the film are left behind to fill the trenches or holes.

For Al-Cu interconnections, the most commonly used CMP processes are oxide and tungsten CMP. The majority of oxide CMP processes are planarization processes, such as ILD CMP processes. Only STI oxide CMP is a bulk removal process, which removes the oxide from the wafer surface and leaves it only in the trenches to form separations between neighboring microelectronic devices. Tungsten CMP processes are bulk removal processes, which remove bulk tungsten from the wafer surface and leave small amounts of tungsten in the contact or via holes to form plugs of interconnection between different conducting layers.

Question: Dual damascene copper CMP belongs to what type of process?
Answer: Copper CMP is a bulk removal process that removes bulk copper and leaves copper only in the trenches and via holes to form metal interconnections.

12.5.1 Oxide chemical mechanical polishing

Silicon oxide CMP processes have been developed and applied in the optical industry for a long time to fine grind and polish glass surfaces for lenses and mirrors. In the mid-1980s at IBM, an early oxide CMP process was developed

by combining the knowledge and experience of glass polishing and bare silicon wafer polishing.

The chemical interaction between the silica particles in slurry and the oxide thin-film surface during the silicon oxide CMP process can be described as follows. First, hydroxyls are formed on both oxide film and silica particle surfaces when contact is made with the water-based slurry. Then, hydrogen bonds are formed between hydroxyls on the oxide surface and silica particle surfaces in the slurry. With help from heat generated by mechanical polishing, molecular bonds between the surfaces can be formed. Mechanical removal of the particles bonded to the wafer surface also tears away the atoms or molecules from the wafer surface, significantly helping the removal process. Figure 12.39 illustrates this oxide polishing mechanism.

No oxide polishing effects are observed in nonaqueous solution, highlighting the importance of surface hydroxylation in silicate glass polishing processes. No oxide polishing occurs without silica abrasives, demonstrating that particles in the slurry are the primary polishing mechanism. A silica slurry has a very low oxide polishing ability except at very high pH (>10) when silica dissolves and enters the solution as a silicate anion. A high concentration of hydroxyl ions in the slurry significantly increases the removal rate of silicate glass.

12.5.2 Tungsten chemical mechanical polishing

Tungsten is widely used as a plug to connect different levels of metal lines. Since CVD tungsten films have excellent gap-fill abilities, removing bulk tungsten film

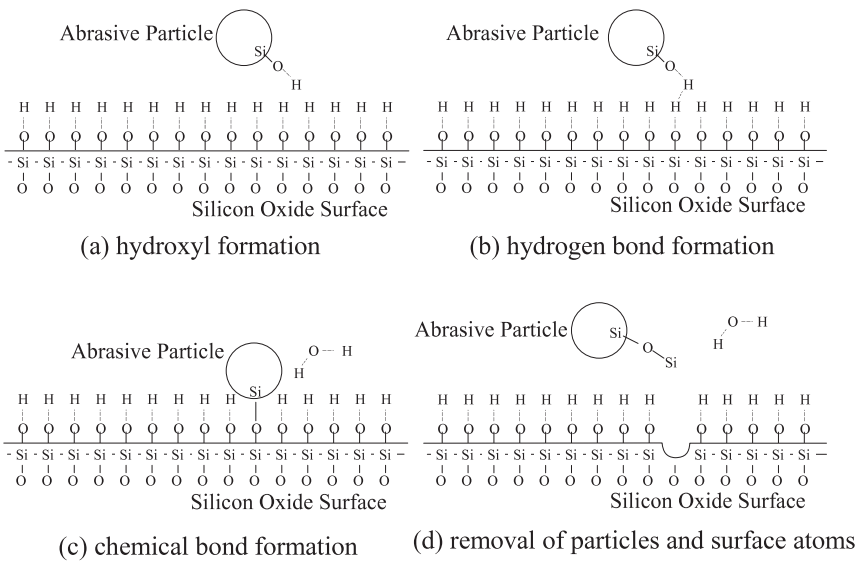


Figure 12.39 Oxide polishing mechanism with silica particles.

and leaving tungsten in the contact/via holes can form tungsten plugs. For IC fabs still using Al–Cu interconnections, the most frequently used metal CMP processes are tungsten CMP processes. These fabs include very advanced nanotechnology memory fabs of both DRAM and flash, and CMOS logic fabs with technology nodes from 0.8 to 0.25 μm. Before tungsten CMP processing was widely adapted by IC fabs, fluorine-based RIE etchback processes were commonly used to remove bulk tungsten on the wafer surface after tungsten CVD processing.

One advantage of the tungsten etchback process is that it can be performed in situ with the tungsten CVD process in a cluster tool. However, tungsten etchback processes always cause the recessing of tungsten and the Ti/TiN barrier/adhesion layer due to the aggressive fluorine chemical etch of Ti/TiN, and this affects chip yield, as shown in Fig. 12.40. The ex-situ tungsten CMP process quickly replaced the tungsten etchback process because it significantly improves yield.

There are always two competing removal mechanisms in metal CMP processes. One is wet etch, in which the oxidant oxidizes the metal and forms a metal oxide that is soluble in the slurry solution. This is a pure chemical process. The other removal mechanism is a combination of chemical and mechanical processes. In this case, the oxidant oxidizes the metal surface and forms a strong metal oxide, which protects the metal surface and stops further oxidation. Mechanical abrasions from particulates in the slurry remove the passivation metal oxide and expose the metal surface to repeated oxidation and oxide removal processes. Figure 12.41 illustrates the two removal mechanisms in the metal CMP process.

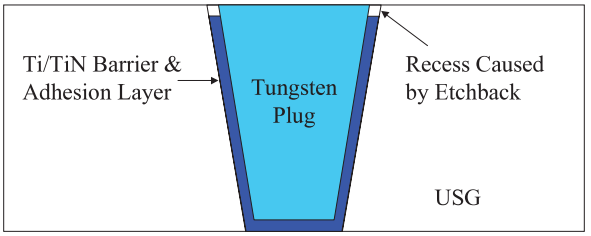


Figure 12.40 Ti/TiN recessing due to tungsten etchback.

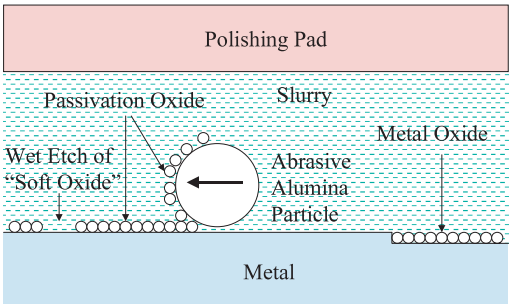


Figure 12.41 Metal polishing mechanisms.

Fine alumina powder is used in tungsten CMP slurry, and potassium ferricyanide $[K_3Fe(CN)_6]$ is used as both an etchant and oxidant. The wet etch chemistry can be expressed as

$$W + 6Fe(CN)_6^{-3} + 4H_2O \rightarrow WO_4^{-2} + 6Fe(CN)_6^{-4} + 8H^+,$$

in which the ferricyanide serves as an electron sink that oxidizes and dissolves the tungsten as a WO_4^{-2} ion in the slurry solution. The competing passivation oxidation reaction can be expressed as

$$W + 6Fe(CN)_6^{-3} + 3H_2O \rightarrow WO_3 + 6Fe(CN)_6^{-3} + 6H^+.$$

This reaction forms the passivating oxide WO_3.

In tungsten CMP processes, these two competing processes are controlled by the local pH at the tungsten–slurry interface. This is accomplished by using additives, such as potassium hydrogen phosphate (KH_2PO_4), which adjusts the pH to between 5 and 6. To improve polishing planarity, a weak organic base such as ethylenediamine is used to further adjust the pH closer to 7 (a neutral value), which increases passivation and reduces wet etch.

Because potassium ferricyanide is highly toxic, and used slurry disposal causes severe environmental problems, many fabs prefer to use ferric nitrate $[Fe(NO_3)_3]$ as an oxidizer. Tungsten CMP uses a two-step polishing process. The first step is removal of bulk tungsten with slurry pH < 4, and the second step is removal of the titanium nitride/titanium stacked barrier/adhesion layer with slurry pH > 9.

12.5.3 Copper chemical mechanical polishing

It is very difficult to pattern copper with plasma etching due to the lack of volatile inorganic copper compounds. Since a dual damascene process does not require a metal etch, it is the perfect candidate for copper metallization. Copper CMP is one of the most challenging processes in dual damascene interconnection processes for copper applications.

Hydrogen peroxide (H_2O_2) or nitric acid (HNO_3) can be used as an oxidant in copper polishing slurry, and alumina particulates are used for abrasion. Because copper oxide CuO_2 is porous and cannot form a passivation layer to stop further copper oxidation on the surface, an additive that can enhance the passivation effect is needed. Ammonia (NH_3) is one additive used in copper CMP slurry. Other additives such as ammonium hydroxide [NH_4OH], ethanol, or benzotriazole can also be used as complexing agents to reduce the wet etch effect.

In dual damascene copper metallization processes, both bulk copper and barrier tantalum layers need to be removed by a CMP process. Because copper slurry cannot effectively remove tantalum, the lengthy overpolishing step for tantalum removal can cause copper recessing and dishing effects, as shown in Fig. 12.42.

To solve this problem, a two-slurry polishing approach is adopted. In this approach, the first slurry is used primarily to remove the bulk copper layer, while

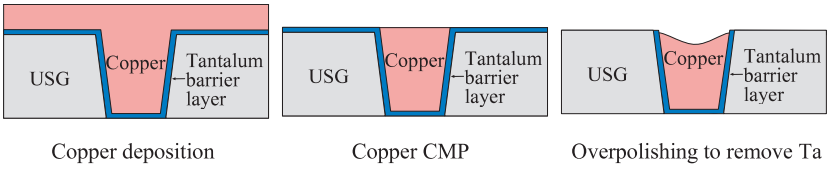


Figure 12.42 Dishing and recessing caused by overpolishing.

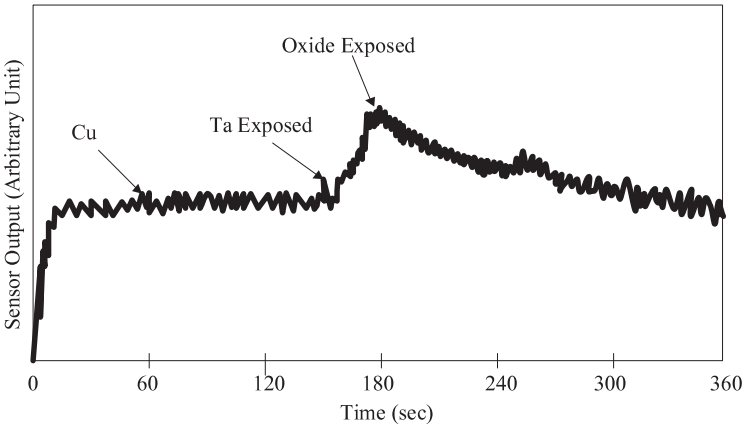


Figure 12.43 Motor current output during a copper CMP process (Aplex).

the second slurry is used to remove the tantalum barrier layer. It is very important to remove all copper from the surface with the first slurry before using the second one, since the second slurry has a very low copper removal rate. The extent of copper loss with overpolishing can be significantly reduced compared to the one-slurry approach. The two-slurry CMP process with separate copper and barrier polishing steps can help reduce copper dishing and oxide erosion effects. The use of multiple polishing platens greatly simplifies multislurry CMP processing.

12.5.4 Chemical mechanical polishing endpoint detection

The endpoints of CMP processes can be detected by monitoring the motor current or by optical measurement. When the CMP process comes to an end, the polishing pad reaches and begins to polish the underside, causing the friction force to change. To maintain a constant pad rotation rate, the current of the polishing head rotary motor changes. By monitoring the change in motor current, the endpoint of the CMP process can be detected. Figure 12.43 illustrates current change during the copper CMP process.

This endpoint detection process is described in detail as follows. During processing, the rotary motor current is close to constant while polishing copper.

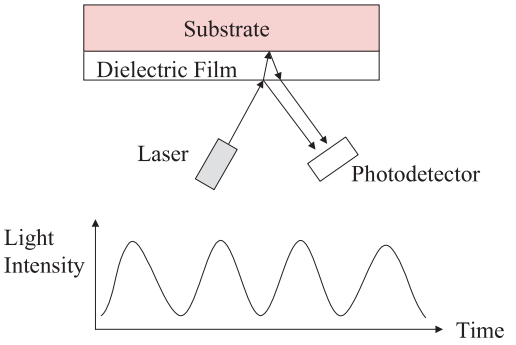


Figure 12.44 Dielectric CMP endpoint.

Noise in the signal comes from the rotation frequency of the motor. When copper is removed and the tantalum barrier layer is exposed, polishing friction increases. To keep the rotation rate, the current of the rotary motor increases. When the tantalum barrier layer is removed, the pad starts to polish oxide. The current starts to gradually drop while tantalum is gradually removed from the surface. Sensing current change as the endpoint for the copper CMP process allows for an in-situ, two-slurry polishing process. In that process, the copper CMP stops when the endpoint monitor senses the motor current increasing. Then the wafer is transferred to another pad that uses a different slurry to effectively remove tantalum.

Another commonly used endpoint technique for CMP processes is the optical endpoint. For dielectric CMP, either the film thickness itself or the film thickness change can be detected in situ with spectroreflectometry. Endpoint can be achieved by either monitoring the thickness change or measuring the film thickness. Reflected lights from dielectric surfaces and dielectric–substrate interfaces interfere with each other. Depending on the film refractive index, film thickness, and incident angle of the light, the interference status can vary from constructive to destructive. Constructive interference gives a brighter reflection, and destructive interference causes a dimmer reflection. When polishing dielectric thin films, a change in film thickness causes periodic changes of the interference states between constructive and destructive. This causes repeated high and low intensities of the detected reflected light. If a single-wavelength light source such as a laser is used, dielectric film thickness changes can be monitored by the change in reflected light. By using a light source with a wide spectrum, such as a UV lamp or multiple-wavelength lasers, the type and thickness of the dielectric film can be directly monitored. Figure 12.44 illustrates dielectric thin-film CMP endpoint detection.

For metal CMP, the change of reflectivity can be used for the processing endpoint. Metal surfaces have high reflectivity; when metal film is removed, the reflectivity significantly reduces, indicating the endpoint. Figure 12.45 illustrates a metal CMP endpoint.

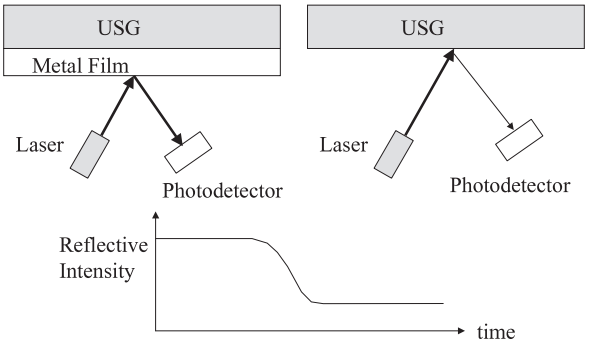


Figure 12.45 Metal CMP endpoint.

12.5.5 Post-chemical-mechanical-polishing wafer cleaning

Post-CMP wafer cleaning is an inseparable part of the CMP process. Immediately after the CMP process, wafers must be thoroughly cleaned, otherwise many defects are left on the wafer surface that are associated with the polishing process and polishing slurries. Post-CMP cleaning must effectively remove both residual slurry particles and other chemical contaminants introduced during the CMP process by slurries, pads, and conditioning tools. The post-CMP cleaning process usually includes poly-vinyl-alcohol (PVA) brush cleaners with DI water, and mechanical scrubbing to remove CMP slurry particles. Most brush cleaning involves the application of DI water through rinse nozzles. Higher cleaning efficiency can be achieved by increasing the DI water volume or brush pressure, or by application of ultrasonic waves. A PVA brush is made of porous polymers, which allows chemicals to penetrate and be delivered to the wafer surface, as illustrated in Fig. 12.46(a). Double-sided scrubbers have also been used in post-CMP clean processes [see Fig. 12.46(b)].

Some slurry particles are chemically bonded to atoms on the wafer surface, which happens when slurry dries on the wafer surface. Chemical additives, such

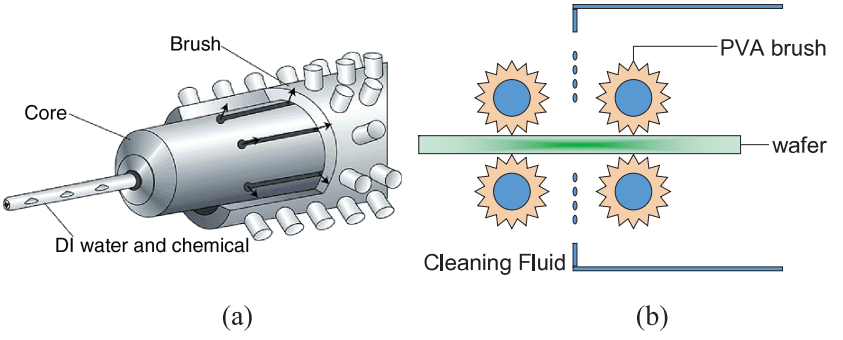


Figure 12.46 (a) A slip-on-the-core (SOTC) design PVA brush (Lam Research Corp.). (b) Illustration of a double-sided scrubber.

as ammonium hydroxide (NH_4OH), hydrofluoric acid (HF), or surfactants can be necessary to efficiently remove these bonded particles by weakening the particle-to-surface bonds or breaking the bonds. The additives also help particles diffuse away from the surface and prevent new particles from forming in the vicinity of the wafer. Residual slurry should never be allowed to dry on the wafer surface because slurry particles can develop a strong chemical bond if the wafer dries. DI water with chemical additives can be used to reduce the adhesion between the particles and wafer. A basic solution is used to adjust the wafer and particle surface charges so that electrostatic repulsion keeps particles from redepositing on the surface. Acidic solutions can be used to oxidize and dissolve organic or metal particles. Figure 12.47 illustrates particle removal processes with acidic and alkaline chemicals.

After oxide CMP processing, silica particles from slurry either adhere to the oxide surface or are embedded in it. It is very important to keep the wafer wet while transferring the wafer to the cleaning station. An alkaline chemical, usually NH_4OH, is used for postoxide CMP cleaning. The alkaline solution negatively charges both silica particles and oxide surface; thus, the electrostatic force expels particles from the surface. For particles strongly bound to the surface with molecular bonds, HF is used to remove such particles by either breaking the bonds or dissolving the silica particles and part of the oxide surface. Megasonics (ultrasound with frequency in the few hundred hertz to megahertz range) is frequently used to form microbubbles in the chemical solution that release shock waves at implosion that help to dislodge particles.

Tungsten slurries are much harder to remove than oxide slurries. DI wafers with NH_4OH additives are commonly used in post-tungsten CMP cleaning processes. The use of ferric nitrate [$Fe(NO_3)_3$] as the oxidant results in high Fe^{3+} ion concentration in the solution. The Fe^{3+} ion interacts with OH^- to form $Fe(OH)_3$ particulates during cleaning with DI water and NH_4OH, and the particles can grow to larger than 1 μm. $Fe(OH)_3$ particulates can cause high surface defect density and can contaminate the brush (called brush loading). Defects caused by $Fe(OH)_3$ particles can be reduced by using 100:1 HF cleaner.

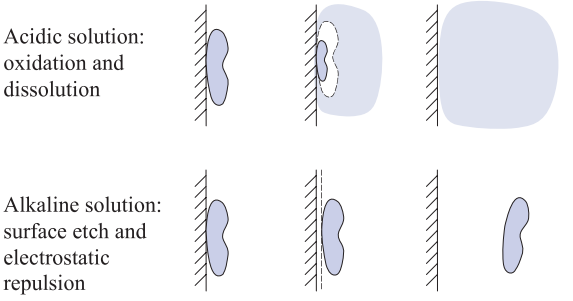


Figure 12.47 Particle removal mechanisms with acidic and basic chemical solutions.

After the cleaning process, a DI water rinsing process follows. Then, the water must be completely removed from the wafer surface without leaving any residue. Wafer drying must be a physical removal process without water evaporation, since evaporative drying is likely to leave dissolved chemicals in the DI water as contaminants.

Spin drying is the most commonly used technique in both single-wafer and batch spin dryers. The centrifugal force from spinning drives water toward the wafer edge and off of the wafer. Ultraclean drying air or nitrogen flow helps drive residual water from the center, where centrifugal force is very low.

Another drying method is vapor drying, which uses high vapor pressure of an ultrapure solvent, usually isopropyl alcohol (IPA, C_3H_8O), to displace the water from the wafer surface.

CMP tools are connected with wet cleaning tools. These integrated systems allow the dry-in dry-out CMP, post-CMP cleaning, and drying process sequence, and improve process yield.

12.5.6 Process issues

The main concerns with CMP processes are polishing rate, planarization capability, WID uniformity, WIW uniformity, WTW uniformity, removal selectivity, defects, and contamination control.

Polishing rate is determined by downforce pressure, pad hardness, pad condition, and applied slurry. Different films also have different polishing rates. Planarization capability is determined by the stiffness and surface condition of the polishing pad. Uniformities can be affected by the polishing pad condition, downforce pressure, relative speeds of the wafer and the polishing pad, and the curvature of the wafers, which is related to film stress. Downforce pressure distribution is the most important factor for controlling CMP uniformity. Removal selectivity is controlled by slurry chemistry and is related to pattern density, which is determined by design layout. There are many different types of defects that are related to many different processing parameters. For example, a copper CMP process can have corrosion on Cu pads that connect to pMOS, and dendrite growth on Cu pads that connect to nMOS. This is because in an acidic solution, copper ions tend to leave pMOS-contacting metal surfaces and move toward nMOS-connecting metal surfaces, especially if light shines on the wafer surface to create a 0.6- to 0.7-V photovoltage between the p and n junctions. Figure 12.48 illustrates the photoelectrochemical process of dendrite-corrosion defect formation.

One way to eliminate this defect is running copper CMP and post-CMP processes in the dark, or sealing the process system so that photovoltage between the p–n junctions can be avoided.

Contamination control is one of the most important issues related to CMP processes. Because slurries contain large amounts of particles and alkali ions, the CMP bay has a much higher probability of mobile-ion and particle contamination than other processing areas in an IC fab. Therefore, some fabs isolate their CMP

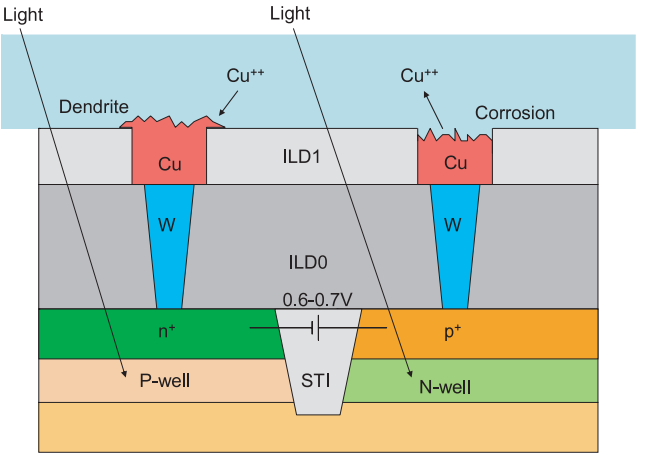


Figure 12.48 Dendrite and corrosion on p^+/n-well metal.

bay from other processing areas to avoid cross-contamination. Wafer cassettes need to be changed when product wafers are transferred in and out of the CMP bay. Personnel movement between the CMP bay and other areas without a gown change is restricted. A good rule of thumb is that anything belonging to the CMP bay stays in the CMP bay; anything not belonging to the CMP bay should be kept out.

Copper CMP tools must be dedicated tools used only for copper polishing processes to avoid copper contamination of the silicon wafer, since copper contamination can cause unstable performance of MOSFETs and ruin IC chips.

If slurry is spilled, it is very important to immediately wash and clean the area thoroughly before the slurry dries. Dried slurry leaves huge amounts of tiny particulates, which easily become airborne. Good housekeeping is needed to avoid slurry spillage and residue build up. Strict procedures and disciplines are required when changing polishing pads and carrier films, and the fresh pads usually need three to five dummy wafer polishings to season them before being ready to polish production wafers.

12.6 Recent Developments

The CMP process is one of the fundamental processes in IC chip manufacturing, along with etch and CVD. Multilayer Al–Cu interconnections need oxide and tungsten CMP processes to planarize ILD layers and form tungsten plugs, respectively. Copper interconnections have become a mainstream technology in CMOS logic IC since the development of the 130-nm technology node. Copper CMP processes have been widely used in dual damascene processes to form copper interconnections in advanced IC fabs.

Interconnections with a combination of copper metallization and low-κ dielectrics raise new challenges for CMP processes. Because low-κ dielectric

materials have less mechanical strength compared to silicon oxide, it is easier for low-κ dielectrics to crack and delaminate during their interface with copper. A low-downforce CMP process has been sought for copper and barrier metal CMP to avoid damaging low-κ dielectrics.

For DRAM applications, polysilicon and oxide-nitride-oxide dielectrics have been used for storage capacitor formation. To shrink the size of the capacitors, high-κ dielectric materials such as Al_2O_3 have been used. DRAM manufacturers are also looking for copper interconnections as a cost-cutting solution. Copper CMP processes involved with these low-κ dielectric materials can also be applied in advanced DRAM fabrication.

The most significant development in CMOS IC is the implemention of HKMG in MOSFETs; these revolutionized gate dielectrics for the first time in the 50-year history of MOSFETs. The gate-last HKMG process requires two CMP processes. The first is poly-open CMP, which polishes ILD0 to expose polysilicon dummy gates. After dummy gate removal and cleaning, high-κ dielectrics and multiple metal layers are deposited to fill the gate trenches to form HKMG. Dummy gate removal can help to further increase silicon strain in the channel, and improve device speed. This is one of the advantages of gate-last over gate-first HKMG. The second process is metal CMP, which removes bulk metal layers on the wafer surface and keeps them only in the gate trenches; this finishes the replacement metal gate process. Figure 12.49 shows the CMP applications in gate-last HKMG CMOS formation.

To avoid high-downforce-induced copper delamination during copper CMP processing, low-downforce CMP processes have been developed. Electrochemical mechanical polishing (ECMP) processes have been introduced to remove copper from wafer surfaces, which are positively biased and serve as anodes. Some companies even introduced downforce-free electrochemical polishing processes to meet the requirements of copper CMP for copper/ULK interconnections.

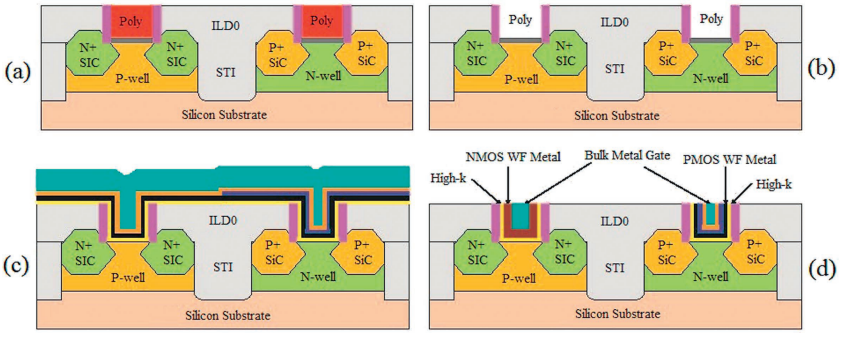


Figure 12.49 Gate-last HKMG processes: (a) ILD0 CMP or poly-open CMP, (b) polysilicon dummy gate removal, (c) high-κ and metal layer deposition, and (d) metal CMP to form HKMG MOSFETs.

12.7 Summary

- The main applications of CMP processes are dielectric planarization and bulk film removal for STI, tungsten plugs, and dual damascene copper interconnections.
- Multilevel metallization needs planarized dielectric surfaces for both high-resolution photolithography in via hole and metal line patterning, and for easier metal deposition.
- For 0.25-μm and smaller patterned features, CMP processes must prepare the planarized surface required by high-resolution photolithography processes due to the small DOF.
- The advantages of CMP processes are: possibility of high-resolution photolithography patterning due to the planarized surface; higher yield; lower defect density due to the removal of rough surface topologies; and more flexibility for IC design.
- A basic CMP system consists of a rotating wafer carrier, a polishing pad on a rotating platen, a pad conditioner, and a slurry delivery system.
- Oxide CMP slurries are alkaline solutions at $10 < \text{pH} < 12$, with colloidal suspension of silica abrasives. Metal CMP slurries are acidic solutions at $4 < \text{pH} < 7$ with alumina abrasives.
- Important factors for CMP processes are polish rate, planarization capability, selectivity, uniformities, defects, and contamination controls.
- Polishing rate is related to downforce pressure, pad stiffness, pad surface condition, relative speeds of the pad and wafer, and slurry type.
- CMP uniformity is determined by downforce pressure distribution, pad stiffness, and pad surface condition.
- Removal selectivity is determined by slurry chemistry.
- In oxide CMP processing, silica abrasive particles in slurry form chemical bonds with the silicon oxide surface, and polishing slurry mechanically abrades the bonded silica along with bonded surface materials from the surface.
- There are two metal removal mechanisms in a metal CMP process: wet etch and passivation/abrasion. In wet etch, an oxidant forms metal oxides that are soluble in slurry. In passivation/abrasion, an oxidant oxidizes metal to form a passivation oxide, which stops metal oxidation and is removed by abrasive particles in the polishing slurry.
- Post-CMP cleaning is a very important step for reducing defects and improving yield. DI water with ammonium hydroxide (NH_4OH) is commonly used for post-CMP cleaning. For oxide CMP processes, HF is used to remove silica particulates that form chemical bonds with surface molecules and cannot be removed by NH_4OH solution. For metal processes, oxidants such as surfactants and nitric acid are used to oxidize and dissolve metal particles that cannot be removed by NH_4OH solution.
- CMP-related defects are scratches, residues, delamination, metal corrosion, dielectric film cracks, etc. These defects can be captured by optical inspection.

- Post-WCMP and post-Cu-CMP electron beam inspections can effectively capture electrical defects such as contact-plug open circuits and junction leakage.
- CMP processes are required for gate-last HKMG CMOS manufacturing.

12.8 Bibliography

T. Ashizawa, "Novel cerium oxide slurry with high planarization performance for STI," *Proc. 4th CMP Symp.* **S-6** (1999).

I. Belov, J.Y. Kim, P. Watkins, M. Perry, and K. Pierce, "Polishing slurries with aluminate-modified colloidal silica abrasive," *MRS Proc.* **867**, W6.9.1 (2005).

A. E. Braun, "Slurries and pads face 2001 challenges," *Semicond. Intl.* **21**(13), 65–74 (1998).

C. Y. Chang and S. M. Sze, *ULSI Technologies*, McGraw-Hill, New York (1996).

P. Feeney, "CMP for metal-gate integration in advanced CMOS transistors," *Solid State Technol.*, 14 (Nov. 2010).

M. A. Fury, "CMP processing with low-k dielectric, *Solid State Technol.* **42**(7), 87–96 (1999).

R. R. Jin, S. H. Ko, B. A. Bonner, S. Li, T. H. Osterheld, and K. A. Perry, "Advanced front-end CMP and integration solutions," *Proc. CMP-MIC*, 119–129 (2000).

Y. Matsui, Y. Tateyama, K. Iwade, T. Mishioka, and H. Yano, "High-performance CMP slurry with CeO_2/resin abrasive for STI formation," *ECS Trans.* **11**, 277–283 (2007).

K. Mistry et al., "A 45-nm logic technology with high-k + metal gate transistors, strained silicon, 9 Cu interconnect layers, 193-nm dry patterning, and 100% Pb-free packaging," *IEDM Tech. Dig.*, 247–250 (2007).

F. W. Preston, "The theory and design of plate glass polishing machine," *J. Soc. Glass Technol.* **11**, pp. 214–256 (1927).

W. S. Rader, T. Holt, and K. Tamai, "Characterization of large particles in fumed silica-based CMP slurry," *MRS Proc.* **1249** (2010).

S. Reabke Selinidis, D. K. Watts, J. Saravia, J. Gomez, C. Dang, R. Islam, J. Klain, and J. Farkas, "Development of a copper CMP process for multilevel, dual inlaid metallization in semiconductor devices," *Proc. Electrochem. Soc.* **98**(7), 9 (1998).

C. Sainio and D. J. Duquette, "Electrochemical characterization of copper in ammonia-containing slurries for chemical mechanical planarization of interconnects," *Proc. Electrochem. Soc.* **98**(7), 126–133 (1998).

R. K. Singh, D. W. Stockbower, C. R. Wargo, V. Khosla, M. Vinogradov and N. V. Gitis, "Post-CMP cleaning applications: challenges and opportunities," *Proc. 13th Intl. CMP-MIC*, 355 (2008).

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

"Trends and future developments for diamond CMP pad conditioners," *Industrial Diamond Rev.* **1**(4), 16–21 (2004).

H. Xiao, L.E. Ma, Y. Zhao, and J. Jau, "Study of devices leakage of 45-nm node with different SRAM layouts using an advanced ebeam inspection systems," *Proc. SPIE* **7272**, 72721E-1 (2009) [doi:10.1117/12.813885].

12.9 Review Questions

1. What were the earliest applications of CMP processes in the semiconductor industry?
2. What are the two main applications of CMP processes in IC chip fabrication in Al–Cu interconnections?
3. Name the other dielectric planarization methods used before CMP processes became widely applied in IC manufacturing.
4. Why does the fabrication of IC chips with feature sizes smaller than 0.25 μm require CMP processes?
5. Compared with other planarization methods, what are the benefits of CMP processes?
6. Why do polishing pads need reconditioning?
7. What abrasive particles are commonly used in oxide slurry?
8. What kind of particle is normally used in metal CMP slurry?
9. Why do oxide slurries need a high pH?
10. Describe the two competing removal mechanisms in the metal CMP process.
11. What are erosion and dishing effects?
12. What metrology tool can be used to measure dishing and erosion?
13. How can particles and defects on the wafer surface be measured?
14. What inspection system can be used after WCMP to capture electrical defects such as contact open circuits and junction leakage? Can an optical inspection be used to capture these defects? Explain your answer.
15. Explain the importance of post-CMP cleaning.
16. Describe the two particle removal mechanisms in a wet chemical cleaning process.
17. If slurry spills and dries in the fab, what problems can it cause?
18. Why is a copper CMP tool only used for a copper polishing process?

19. When silica particulates form molecular bonds with oxide surfaces, they cannot be removed by NH_4OH. What is the chemical usually used to remove these silica particles? Try to describe the removal process.
20. What two CMP processes are needed in advanced gate-last HKMG CMOS fabrication?

Chapter 13
Process Integration

IC manufacturing involves many processing steps that have been described in the previous chapters. To fabricate a functional chip, every processing step must be well integrated with the other processing steps. This chapter describes the integration of these steps in CMOS IC chip fabrication.

Objectives

After finishing this chapter, the reader will be able to:

- list the three methods used for isolation formation
- describe the three-well formation processes
- explain the purpose of threshold adjustment implantation
- describe the sidewall spacer process and explain its applications
- explain the advantages of high-κ gate dielectrics over SiON gate dielectrics
- describe the gate-first and gate-last approaches of metal gate formation
- name at least three silicides used for gate and local interconnections
- name three metals used in a traditional aluminum interconnection process
- list the basic steps for a copper metallization process
- identify the material most commonly used as a final passivation layer for an IC chip
- describe the main differences of CMOS, DRAM, and NAND flash processes.

13.1 Introduction

It takes more than 30 masks and several hundreds of processing steps to finish the fabrication of an advanced CMOS IC chip. Every step is related to all of the other steps. For CMOS processes, the steps can be classified as front-end of line (FEoL), mid-end of line (MEoL) and back-end of line (BEoL). FEoL includes active area (AA) formation, well implantations, gate patterning, and S/D formation that create the transistors. MEoL includes self-aligned silicide, contact hole pattern and etch, and tungsten deposition and CMP that form the contacts between the devices and the metal wires. BEoL steps form the interconnections and passivation. Traditional aluminum interconnections include metal stacked (Ti/TiN/Al-Cu/TiN) PVD and etch; dielectric CVD; dielectric planarization; and via hole pattern and etch. For copper interconnections, BEoL includes via hole patterning and etch;

trench patterning and etch; barrier layer (Ta or TaN) and copper seed layer PVD; bulk copper plating and annealing; and metal (Cu/Ta) CMP.

For flash processes, FEoL includes the AA, WL, contact to BL/source line (CB1), source line, contact to BL (CB2), and BL. The BEoL includes via and metal layers.

For DRAM processes, two major competing technologies coexisted for a long time; one is stacked capacitor technology and the other is deep trench capacitor technology. Stacked capacitor DRAM dominates the DRAM market, while deep trench capacitor DRAM is widely used as embedded DRAM for system-on-chip applications because it is more compatible with CMOS and logic processes. More-advanced DRAM processes have also been developed; one of them in mass production is the buried wordline (bWL) technology.

13.2 Wafer Preparation

Single-crystal silicon wafers with ⟨100⟩ orientation are frequently used for CMOS IC chip manufacturing. Bipolar and BiCMOS chips use wafers with ⟨111⟩ orientation. Wafers used in IC processing are doped wafers, either n- or p-type, with typical substrate dopant concentrations of 1×10^{15} atoms/cm^3. Before the mid-1970s, pMOS IC chips were made with n-type wafers. After the introduction of ion implantation processes in the mid-1970s, nMOS IC chips were manufactured with p-type wafers. Although both n- and p-type wafers can be used for CMOS IC fabrication, more fabs use p-type wafers than n-type wafers. The reason is mainly historical. Driven by the demands of digital logic IC chips with low power consumption, high noise immunity, and high thermal stability, the CMOS process was developed based on the nMOS process in the late 1970s. Naturally, p-type wafers, which were used to make nMOS ICs, were chosen as the substrate for early CMOS IC fabrication.

The simplest nMOS IC process had five masking steps: activation area, gate, contact, metal, and bonding pad. Early CMOS IC processing required eight masking steps: n-well (for p-type substrate), activation, gate, n-S/D, p-S/D, contact, metal, and bonding pad. Figure 13.1(a) illustrates an nMOS chip cross section, and Fig. 13.1(b) shows a cross section of an early CMOS chip.

Bipolar transistors and BiCMOS chips always require wafers with an epitaxial silicon layer to form a heavily doped buried layer. Certain power devices need substrates with resistance so high that only wafers made with a floating zone (FZ) method can meet the requirement. When the clock frequency of CMOS IC chips is not very high, the chips do not need an epitaxial layer. However, silicon wafers with an epitaxial silicon layer are required for high-speed CMOS chips. Silicon wafers made by the Czochralski (CZ) method always have some oxygen, and oxygen can reduce carrier lifetime and slow down device speed. The epitaxial silicon layer can create an oxygen-free substrate and help to achieve high device speed.

RCA cleaning processes are commonly used to remove contaminants from the silicon wafer surface before epitaxy. Anhydrate HCl dry cleaning helps to remove

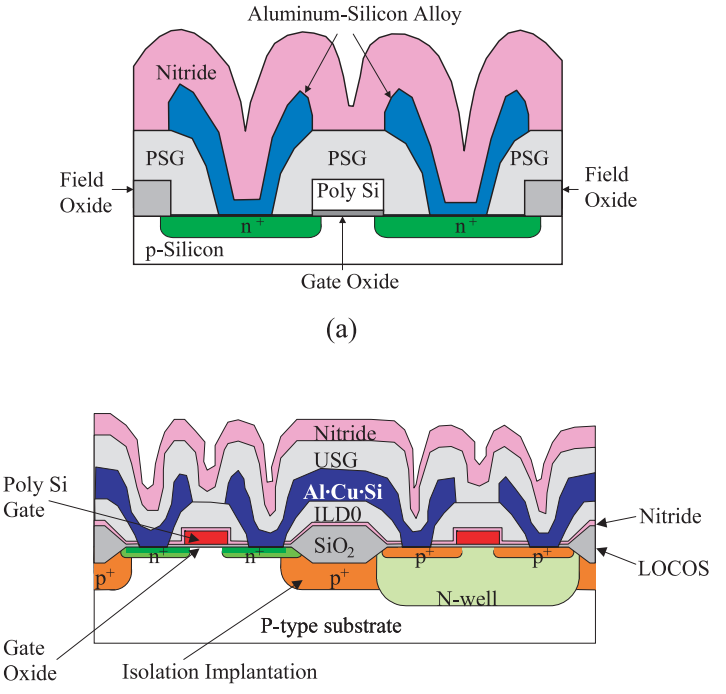


Figure 13.1 Cross section of (a) nMOS and (b) early CMOS chips.

mobile ions and native oxide. Epitaxial silicon layer growth usually is a high-temperature (>1000 °C) CVD process with silane (SiH_4), DCS (SiH_2Cl_2), or TCS ($SiHCl_3$) as the main processing gas. Hydrogen is commonly used in the epitaxy growth process as a secondary processing gas or carrier and purge gas. Arsine (AsH_3) or phosphine (PH_3) is used as the n-type dopant gas, and diborane (B_2H_6) is used as the p-type dopant gas.

Advanced CMOS IC chips normally use p-type ⟨100⟩ single-crystal silicon wafers with p-type epitaxial layers.

13.3 Isolations

Blanket field oxide, local oxidation of silicon (LOCOS), and STI are the three isolation techniques that can be used in IC fabrication. p-type doped junctions also have been used to electronically isolate neighboring transistors.

13.3.1 Blanket field oxide

Blanket field oxide was used in the early years of the IC industry. It is a simple and straightforward process. Blanket field oxide can be formed by growing an oxide

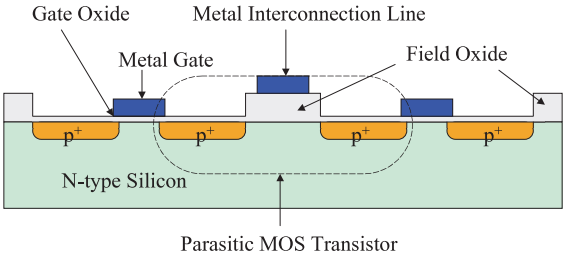


Figure 13.2 pMOS IC with blanket-field oxide isolation.

layer with a desired thickness on a flat silicon surface, then patterning and etching the activation windows on the oxide layer. The thickness requirement of a field oxide is determined by the field threshold voltage, noted as V_{FT}, which needs to be high enough ($V_{FT} \gg V$) to prevent crosstalk between the neighboring transistors. While the circuit power supply voltage can switch on and off MOS transistors on the IC chips ($V > V_T$), it cannot turn on the parasitic MOS transistors and cause chip malfunction. Figure 13.2 illustrates a pMOS IC chip with a blanket-field oxide as isolation. The thickness of the blanket-field oxide normally is about 10,000 to 20,000 Å.

13.3.2 Local oxidation of silicon

Blanket field oxides were used until about 1970. Although the process is simple, it has some disadvantages. One is that the active window has a high oxide step with a sharp edge, which is very difficult to cover in the subsequent metal deposition process. Another disadvantage is that the channel stop doping must occur before oxidation, requiring the field oxide to align with the isolation doping region. This requirement causes difficulties in feature size reduction.

LOCOS has been used in IC production since the1970s. One of the advantages of LOCOS is that the silicon dioxide is grown after channel stop implantation; thus, the field oxide is self-aligned with the isolation doping area. By using channel stop implantation, field threshold voltage V_{FT} can remain unchanged, while the field oxide thickness decreases. Compared to blanket-field oxide, the step height between the activation area and LOCOS field oxide is lower, and the sidewall is sloped. This makes sidewall coverage much easier for the later metal or polysilicon layer deposition. The thickness of LOCOS oxide is 5,000 to 10,000 Å. The LOCOS isolation process is illustrated in Fig. 13.3.

LPCVD silicon nitride is used as an oxidation mask, which only allows thick silicon dioxide (LOCOS) to grow at designated areas. The activation areas where transistors will be built are covered by nitride and do not grow the thick oxide. Pad oxide is needed to relieve the strong tensile stress of the LPCVD nitride. Plasma etch with fluorine chemistry is employed for the nitride patterned etch, and hot phosphoric acid is used to strip the nitride layer after LOCOS formation.

One of the major disadvantages of LOCOS is the “bird’s beak.” Silicon dioxide grows in all directions, causing encroachment in a lateral direction underneath the

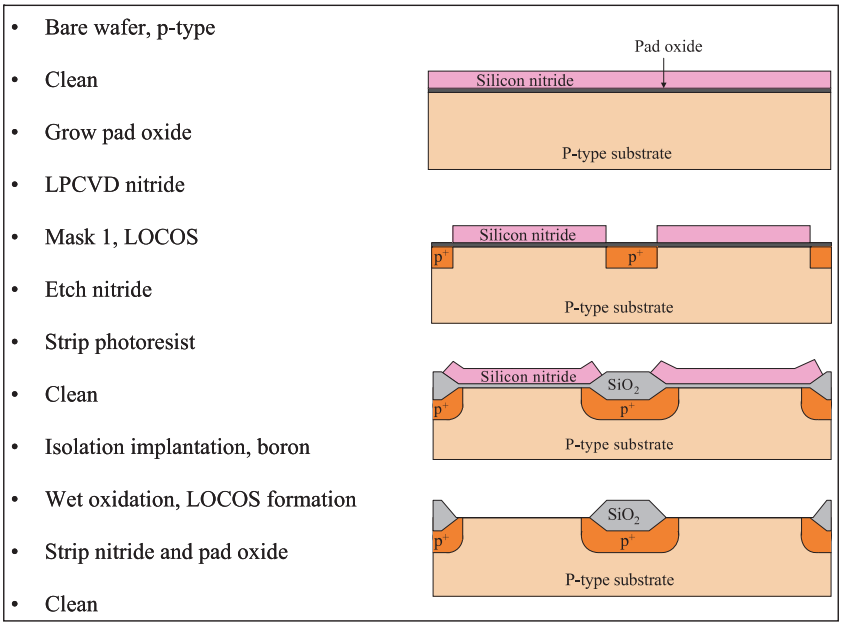


Figure 13.3 LOCOS isolation formation.

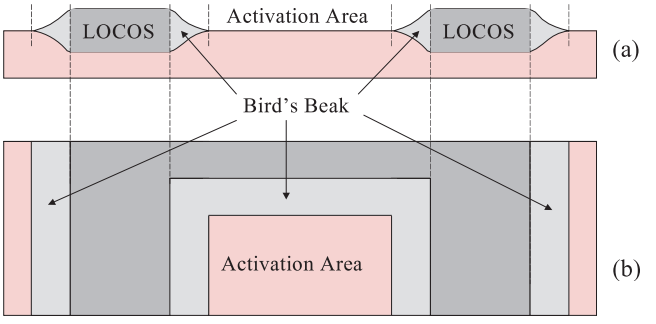


Figure 13.4 Bird's beak of LOCOS: (a) cross section view and (b) top view.

silicon nitride layer, as shown in Fig. 13.3. Bird's beak is caused by the isotropic diffusion of oxygen inside the silicon dioxide layer during the thermal oxidation process. The dimension of LOCOS encroachment is about the same as the oxide thickness on both sides for 5000-Å oxide; the bird's beak is about 0.5 μm on both sides. This requires a considerable amount of silicon surface area and creates difficulties when increasing the transistor packing density, as shown in Fig. 13.4.

To reduce the bird's beak of LOCOS, several improvements have been made. Poly-buffered LOCOS (PBL) is the most popular. Depositing a polysilicon buffer layer before LPCVD silicon nitride can reduce bird's beak from 0.1 to 0.2 μm. PBL

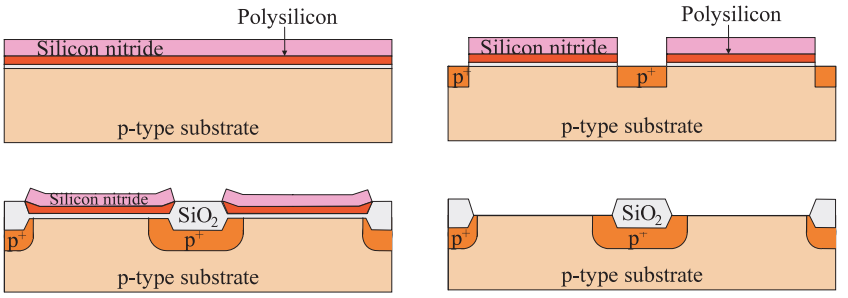


Figure 13.5 Processing steps of PBL formation.

reduces bird's beak due to the consumption of the lateral diffusion oxygen by the polysilicon layer. The process is illustrated in Fig. 13.5.

13.3.3 Shallow trench isolation

LOCOS and improved PBL work very well until the feature size reduces to less than a half micron. Just like a very thick wall to separate neighboring houses would take up too much real estate, the bird's beak of LOCOS takes up too much precious space from the silicon surface that could be used to build transistors. Keep in mind that transistors are surrounded by isolation oxide. In real life, narrow fences are used to separate neighboring houses; in IC fabrication, their equivalent is STI.

To reduce oxide encroachment, silicon etching with a nitride mask and oxidation of the trench was researched. The STI process was then developed with CVD oxide trench fill instead of thermal oxidation. An early STI process is illustrated in Fig. 13.6.

In early STI processes, an LPCVD silicon nitride layer is used as a hard mask for single-crystal silicon etching and the oxide etching stop layer. Before CVD oxide fills the trench etch on the silicon surface, a thin layer of barrier oxide is grown, and boron ions are implanted at the bottom of the trench to form a channel stop junction. Channel stop implantation can help reduce the required isolation oxide thickness, which reduces the depth of the trench. Photoresist etchback planarization is applied to remove CVD oxide from the wafer surface. By choosing a suitable CF_4/O_2 ratio, an approximately 1:1 etch rate ratio of photoresist and oxide can be achieved. Etchback processing stops at the nitride layer; this endpoint occurs automatically by the appearance of a C-N line radiation signal in the glow of the plasma. After stripping the nitride with hot phosphoric acid, oxide left inside the trenches serves as the isolation blocks for electrically isolating the neighboring devices.

Although STI has several advantages over LOCOS, it did not immediately replace LOCOS. The processing sequence of LOCOS has fewer steps and was production proven. LOCOS continued to be used in IC fabrication until the mid-1990s, when the minimum feature size became smaller than 0.35 μm, and the bird's beak of LOCOS was no longer tolerable. The smaller geometry also required

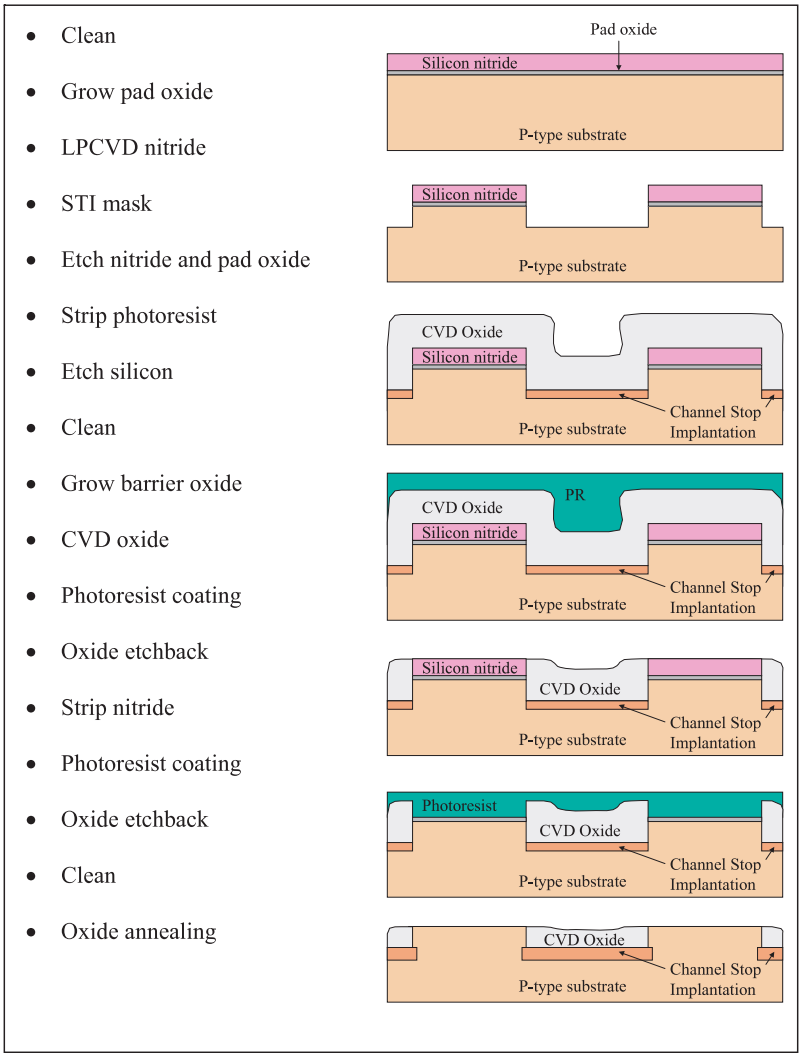


Figure 13.6 An early STI process.

highly planarized surfaces to ensure the photolithography resolution because of the DOF requirement. LOCOS always has a step of 2500 Å or higher between the active area and oxide surface, a step that is too large to precisely pattern 0.25-μm features. STI processes with oxide CMP have been developed and applied in IC fabrication. An advanced STI processing flow is illustrated in Fig. 13.7.

Channel stop ion implantation is no longer necessary when the power supply voltage of an IC device reduces to 1.8 V or lower. The trench fill process can also be achieved with thermal CVD processing with O_3-TEOS chemistry at atmospheric

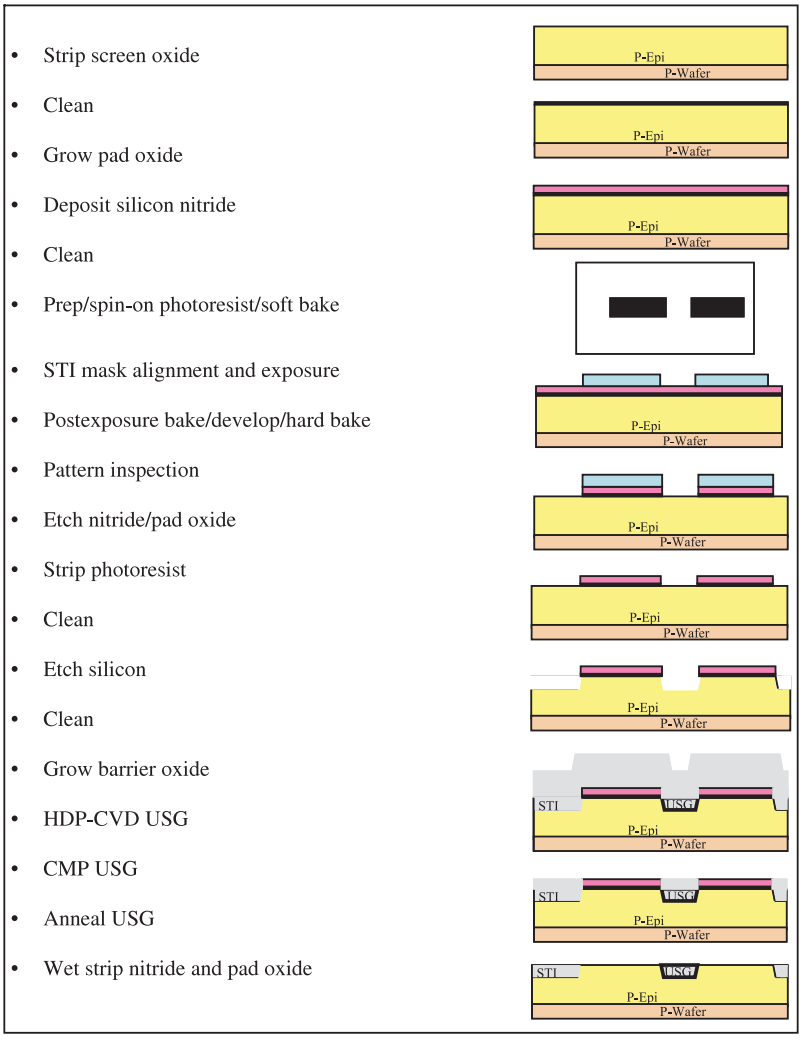


Figure 13.7 Advanced STI processing flow.

or subatmospheric pressure. Oxide deposited with a high-density plasma (HDP) process does not require thermal annealing because the oxide is very dense due to the heavy ion bombardment during deposition. Oxide deposited with O_3-TEOS chemistry has to anneal at more than 1000 °C in the oxygen ambient to densify the film.

STI formation involves many processes, such as oxidation, nitride deposition, nitride/oxide etching, silicon etching, oxide CVD, oxide CMP, oxide annealing, and nitride stripping. The main challenges of STI processes as device feature sizes continue to shrink are single-crystal silicon etching, oxide CVD, and oxide CMP.

Recently, STI oxide has also been used to induce stress to strain silicon in the active area to improve device speed.

13.3.4 Self-aligned shallow trench isolation

For flash memory devices, the most commonly used isolation technique is the self-aligned STI. Figure 13.8 shows a layout of NAND flash memory with self-aligned STI. The process starts with wafer cleaning, followed by gate oxide growth, floating gate polysilicon, and nitride hard mask depositions, as shown in Fig. 13.9(a). After patterning with an AA mask, a sequential process etches nitride, polysilicon, gate oxide, and silicon substrate, as shown in Fig. 13.9(b). After wafer cleaning, CVD oxide fills the gap [Fig. 13.9(c)], and a CMP process removes oxide from the wafer surface, endpointing at the nitride layer [Fig. 13.9(d)]. Stripping of the nitride layer finishes the formation of the self-aligned STI, as shown in in Fig. 13.9(e).

13.4 Well Formation

13.4.1 Single well

Early CMOS IC processing only required a single well, either n- or p-, depending on the type of wafers used. Well formation is accomplished by high-energy, low-current ion implantation and thermal anneal/drive-in. Single-well processing steps are illustrated and listed in Fig. 13.10.

A masking step indicates a photolithography process, which includes photoresist coating, alignment and exposure, photoresist development, and pattern inspection. Processing steps for p-well formation on an n-type wafer are almost the same as the n-well formation on a p-type wafer. Figure 13.11 shows a CMOS with n- and p-wells.

13.4.2 Self-aligned twin wells

A twin-well structure allows IC designers more flexibility in designing CMOS IC circuits. To save a photolithography masking step, a self-aligned twin-well process has been developed and used in the well formation process. The self-aligned twin well is achieved by using LPCVD silicon nitride. Nitride is a very dense layer. It can prevent oxidation on p-wells by blocking oxygen diffusion, and it also prevents n-type ions from penetrating to the p-well during ion implantation. The thick oxide growth on an n-well region can block p-well boron ion implantation. A pad oxide layer is needed to relieve the high tensile stress of the LPCVD nitride, otherwise the high stress can break down the wafers. Nitride can be stripped by hot phosphoric acid with a high selectivity to oxide. The processing steps of a self-aligned twin well are listed and illustrated in Fig. 13.12.

The advantage of a self-aligned twin well is that it eliminates a photolithographic masking step, reducing production cost and improves chip yield. The disadvantage of a self-aligned twin well is that the silicon wafer surface is no longer flat. When silicon dioxide grows on an n-well region, the oxide layer grows into the silicon

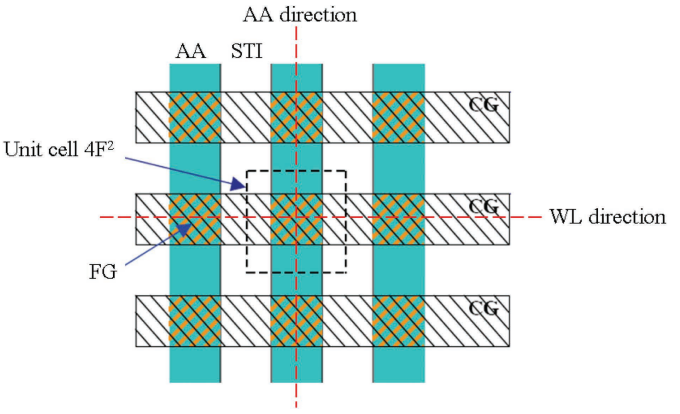


Figure 13.8 Layout NAND flash with self-aligned STI. FG indicates the floating gate, and CG indicates the control gate.

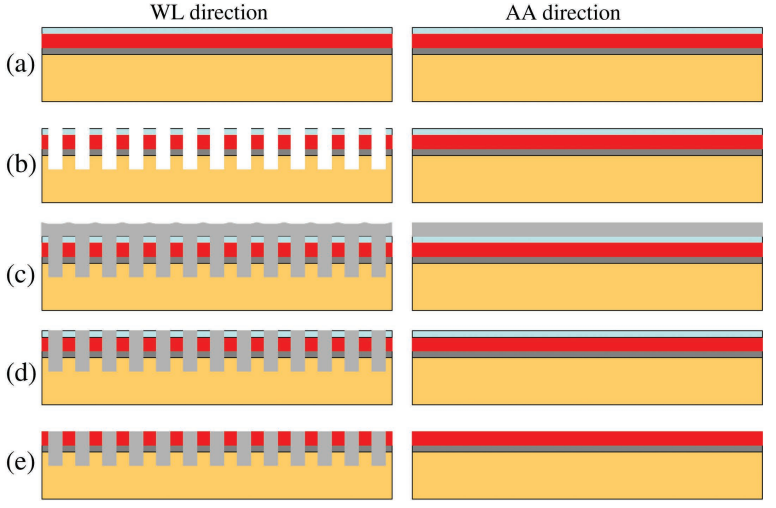


Figure 13.9 Self-aligned STI process of a NAND flash: (a) gate oxidation, polysilicon, and nitride hard mask deposition; (b) pattern etch of nitride hard mask, polysilicon, gate oxide, and silicon substrate; (c) CVD oxide; (d) oxide CMP; and (e) stripping of the nitride hard mask.

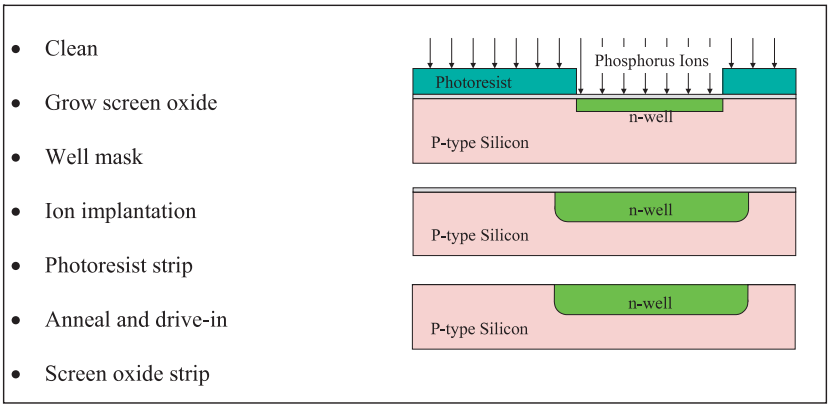


Figure 13.10 Step formation of n-wells.

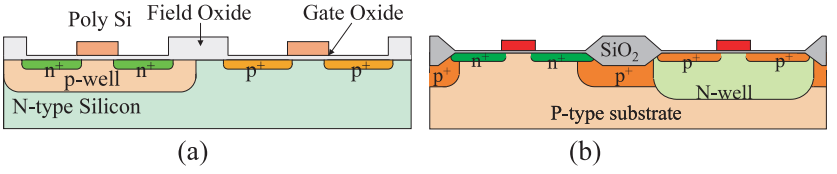


Figure 13.11 CMOS with (a) p-well and (b) n-well.

substrate. Therefore, the n-well is always lower than the p-well in a self-aligned twin well.

Twin-well structures have better control of substrates and allow more freedom for CMOS IC design. In twin-well formation process, the n-well is always implanted before the p-well. Phosphorus has a lower diffusion rate in single silicon than boron has. If a p-well is implanted first, boron in the p-well can diffuse out of control during n-well annealing and dopant drive-in.

13.4.3 Twin wells

The p- and n-wells formed by self-aligned twin-well processing are not on the same level, which can affect photolithography resolution due to the DOF problem. Double-photomask twin wells are frequently used in advanced CMOS IC chip manufacturing. Their processing steps are listed and illustrated in Fig. 13.13. Both well implantation processes use high-energy, low-current implanters. Furnaces are used for well implantation annealing and drive-in processes.

13.5 Transistor Formation

13.5.1 Metal gate process

Before the mid-1970s, an MOS transistor was formed by a process in which the S/D and gates were not self-aligned. First, the sources and drains were formed

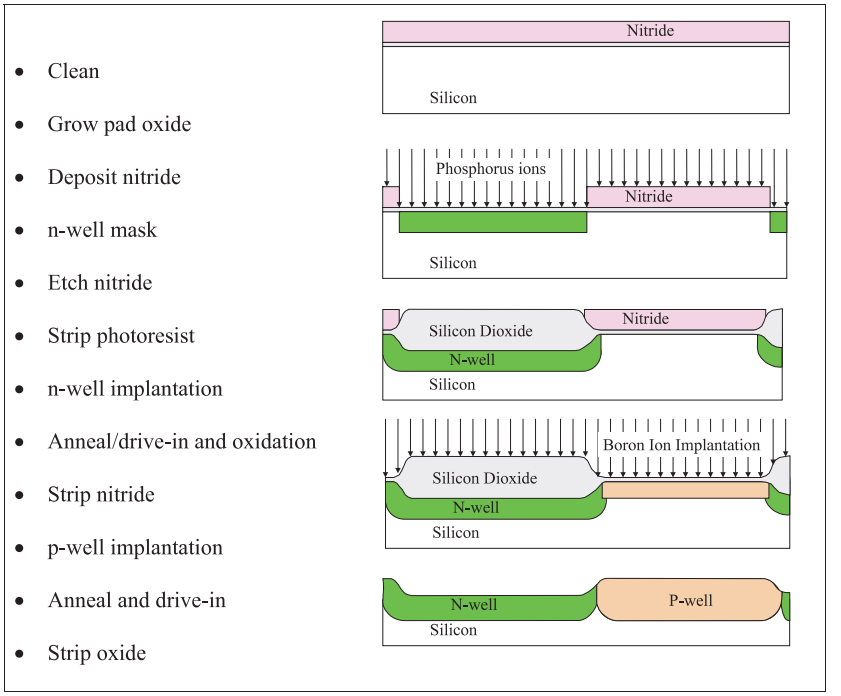


Figure 13.12 Self-aligned twin-well process.

by a diffusion process with thermally grown silicon dioxide as a diffusion mask. Then, oxide in the gate area was etched away, and a thin layer of gate oxide was grown. A third mask defined the contact holes, and a fourth mask formed the metal gates and interconnections. A final mask defined the bonding pad. After etching of the bonding pads and photoresist stripping, the wafers were ready for testing and packaging. Table 3.1 lists the pMOS transistor process without self-alignment, and Figs. 3.30 and 3.31 illustrate the processing steps. A gate usually is designed wider than the distance between the S/D to ensure full coverage of the S/D by the gate. This makes it very difficult to scale down the device feature size. Currently, only educational laboratories manufacture MOSFETs using this process.

13.5.2 Self-aligned source/drain process

With the introduction of ion implantation, processes for manufacturing MOSFETs have changed to self-aligned S/D processes. Polysilicon has replaced aluminum as the gate material because aluminum alloys cannot sustain the high temperature required by the postimplantation thermal annealing. The process starts with an activation mask with an etched window on the field oxide to define the area where the transistors will be built. After wafer cleaning, gate oxide growth, and polysilicon deposition, a gate mask defines the gate and local interconnection.

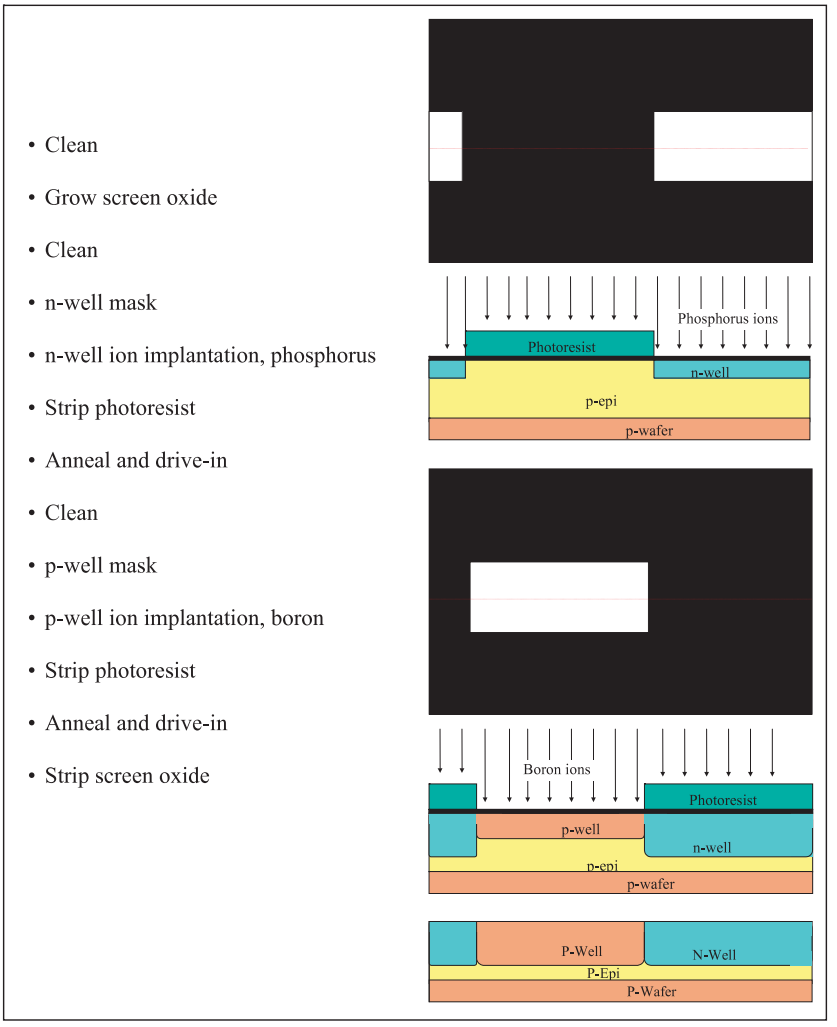


Figure 13.13 Double-photomask twin-well formation.

After S/D ion implantation and thermal annealing, MOSFETs are formed. Figure 13.14 illustrates the self-aligned S/D process for an nMOS transistor.

Self-aligned gate processes are the basic transistor formation processes in IC fabrication. Almost all transistor formation processes for advanced IC chips are developed based on this process.

13.5.3 Lightly doped drain

When a gate width is smaller than 2 μm, the vertical component of the electric field induced by the bias voltage between the S/D can be high enough to accelerate

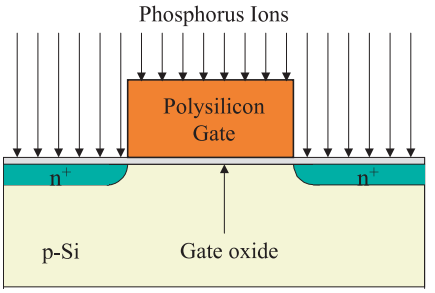


Figure 13.14 Formation of an nMOS with self-aligned S/D.

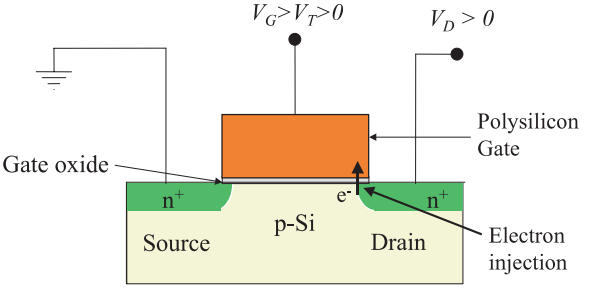


Figure 13.15 Hot electron effect of a MOS transistor.

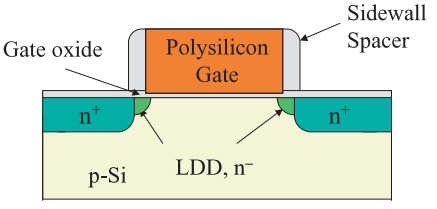


Figure 13.16 LDD of a MOS transistor.

electrons tunneling through the thin gate oxide layer. This is called the hot electron effect, which can affect transistor performance due to gate leakage, and cause reliability problems for IC chips due to the trapping of electrons in the gate oxide. Figure 13.15 illustrates the hot electron effect of a MOS transistor.

The most widely used method to suppress the hot electron effect is called a lightly doped drain (LDD) or source-drain extension (SDE), as shown in Fig. 13.16.

LDD/SDE junctions can be formed by using low-energy, low- to medium-current ion implantation. It is a shallow junction with very low to medium dopant concentration extended just underneath the gate. By depositing conformal dielectric layers and etching them back vertically, sidewall spacers can be formed on both sides of the polysilicon gate. High-current, low-energy ion implantation forms a heavily doped S/D, which is kept apart from the gate by the sidewall

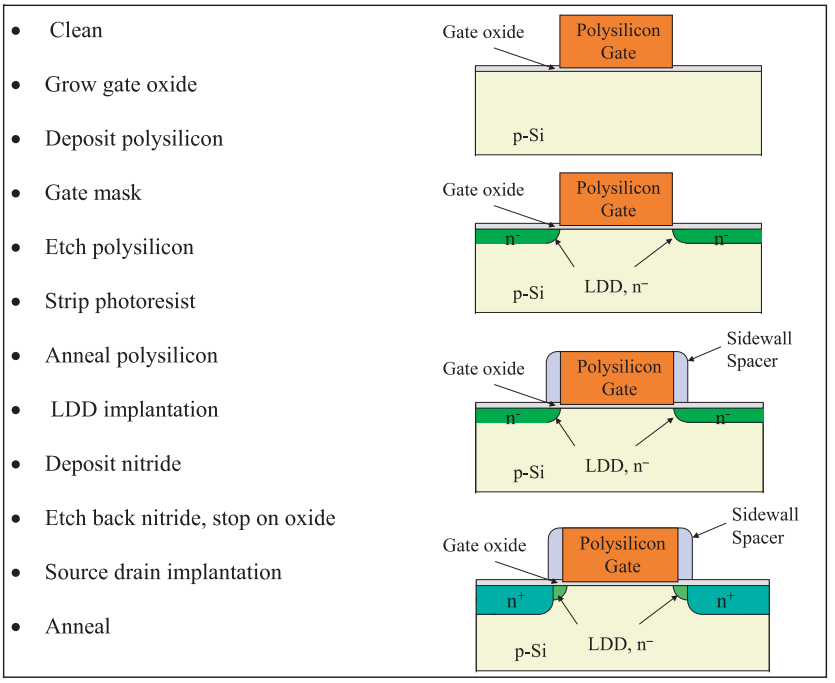


Figure 13.17 Processing steps for a MOS transistor with LDD.

space. This reduces the vertical component of the S/D bias-induced electric field, and reduces the available electrons for tunneling, thus suppressing the hot electron effect. Transistors with LDD can be made using the processing steps listed and illustrated in Fig. 13.17.

When feature sizes reduce to sub-0.18 μm, the dosage of this ion implantation process is no longer considered to be light; therefore, the process is referred to as SDE implantation.

13.5.4 Threshold adjustment

Threshold adjustment implantation controls the threshold voltage of MOSFETs. This ion implantation process ensures that the power supply voltages of electronic systems can turn on and off MOSFETs in IC chips. The threshold adjustment implantation is a low-energy, low-current implantation process and is usually performed before gate oxide growth. Figure 13.18 illustrates this implantation process.

For CMOS IC chips, two threshold adjustment implantation processes are needed: one for p-type and one for n-type. With feature size shrinking, well implantation depth can be achieved with high-energy ion implantation, and the drive-in process is no longer required. Thus, V_T-adjust implantation can be performed with well implantation.

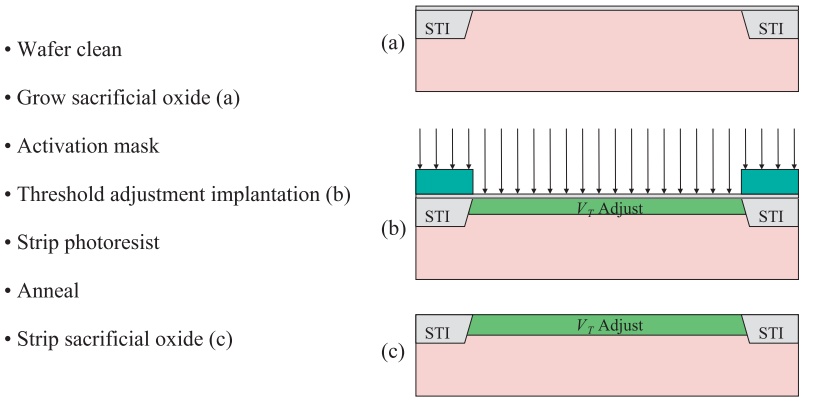


Figure 13.18 Threshold adjustment implantation process.

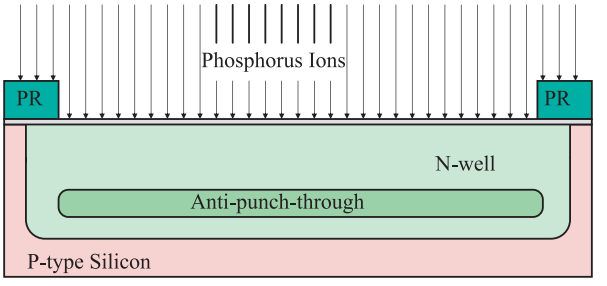


Figure 13.19 Anti-punch-through implantation process.

13.5.5 Anti-punch-through

Punch-through is an effect where depletion regions of the S/D short each other under the influence of both gate substrate bias and S/D bias. Anti-punch-through implantation, which is a medium-energy, low-current implantation process, protects transistors against this effect. Anti-punch-through implantation is normally performed with well implantation. Figure 13.19 illustrates the anti-punch-through implantation process.

Another implantation process commonly used to prevent punch-through effect is halo implantation. It is a low-energy, low-current implantation process with a large incident angle, such as 45 deg. The halo junction formed in this implantation process can help to suppress punch-through effect. Figure 13.20 illustrates the halo implantation process.

13.6 Metal-Oxide-Semiconductor Field-Effect Transistors with High-κ and Metal Gates

As device feature sizes continue to shrink, gate oxide thicknesses will become too thin for MOS transistor to operate reliably, even at 1 V. High-κ dielectrics

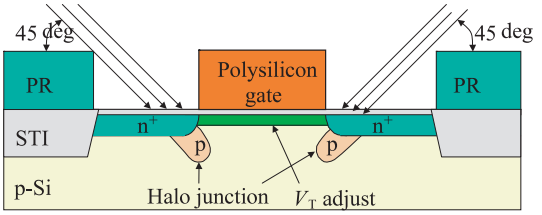


Figure 13.20 Threshold adjustment implantation process.

started to replace standard silicon dioxide or silicon oxynitride as gate dielectric materials when device dimensions reached 45 nm. By using a high-κ dielectric, the thickness of gate dielectrics can be increased to prevent tunneling leakage and dielectric breakdown. It also helps to maintain a large enough gate capacitance while gates sizes shrink. The gate capacitance must be able to hold enough charge to cause carrier inversion underneath the gate dielectric, which forms a minority carrier channel to turn on the MOS transistor.

To improve device speed, metal will likely be used again as a gate electrode for the MOS transistor because metals such as aluminum have significantly lower resistivity than polysilicon and silicide. Also, when a gate is biased to turn on a MOSFET, polysilicon forms an inversion layer at the polysilicon–oxide interface, which acts just like a dielectric layer, effectively increasing the thickness of the gate oxide when the MOSFET is turned on, and slowing down the device speed. Using metal gates to replace polysilicon gates solves the poly-inversion layer issue.

There are several high-κ metal gate (HKMG) integration approaches, such as gate-first, gate-last, and hybrid methods (nMOS gate-first and pMOS gate-last methods).

13.6.1 Gate-first process

Figure 13.21 shows a gate-first process of an nMOS on SOI substrate. This process consists of a thin layer (~4 Å) of silicon dioxide growth, high-κ dielectric deposition, followed by deposition of cap dielectric layers [Fig. 13.21(a)]. The cap layer of nMOS is different from the cap layer of pMOS for controlling the work functions and achieving desired threshold voltages. A thin metal layer, usually titanium nitride (TiN), is then deposited, followed by polysilicon and hard mask deposition [Fig. 13.21(b)]. After patterning and etching the hard mask, the photoresist is stripped, and a gate pattern is etched with the hard mask [Fig. 13.21(c)]. The rest of the process is very similar to normal SiON-polysilicon gate processes, with SDE implantation, sidewall spacer formation, S/D implantations [Fig. 13.21(d)], RTP, silicide formation [Fig. 13.21(e)], ILD0 CVD and CMP, contact etch and cleaning [Fig. 13.21(f)], glue layer and tungsten deposition, and WCMP to form the contact plugs [Fig. 13.21(g)]. For the 32-nm technology node, the high-κ dielectric is hafnium dioxide (HfO_2)-based material, and metal gate usually uses titanium nitride (TiN).

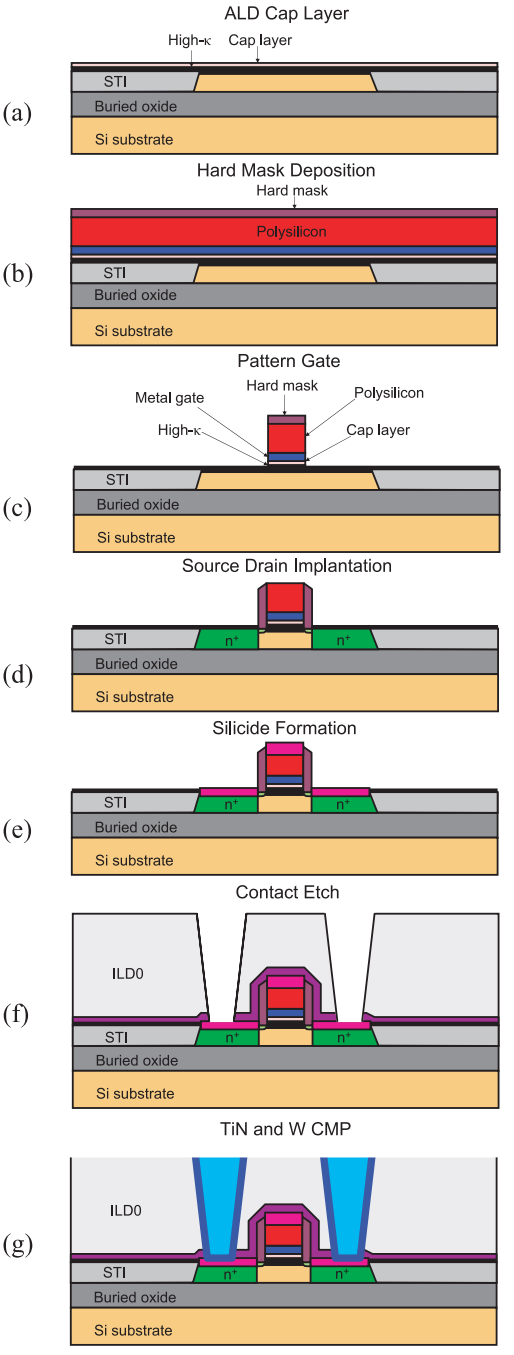


Figure 13.21 HKMG gate-first processes.

Since the gate-first HKMG process is very similar to a regular SiON-polysilicon process, many tools can be shared by both processes. Gate-first HKMG also has fewer processing steps compared to gate-last HKMG, thus total cost can be lower, which is very important for cost-sensitive fabs. In gate-first HKMG processes, high-κ and metal materials must be able to sustain the high-temperature annealing processes; this makes the choice of candidates very limited.

The main challenges for a gate-first process are polysilicon and oxide stripping, high-κ dielectric deposition, and metal gate formation. Polysilicon and oxide stripping must remove all of the oxide on the silicon surface and avoid undercutting caused by overetch. A hafnium-oxide-based dielectric is the most promising of the high-κ dielectric materials. A thin layer (<10 Å) of silicon oxide or silicon oxynitride grown on the silicon surface prior to high-κ dielectric deposition is needed to prevent high interstate charge. High-κ gate dielectrics are formed by ALD. The requirements for high-κ dielectric thin-film deposition are excellent bottom coverage and uniform deposition across the wafer.

13.6.2 Gate-last process

The gate-last process of high-κ metal gates is the first HKMG process used in IC products in the 45-nm technology node. Although it has more processing steps compared to gate-first approaches, it has many advantages, such as a wider range of choice of high-κ and metal gate materials, because HKMGs are formed after S/D and silicide annealing. For the gate-first approach, HKMG materials must be able to sustain the high-temperature annealing processes. Figure 13.22 illustrates the gate-last HKMG processing steps.

The first half of the process is similar to a regular CMOS process, with gate oxidation, polysilicon deposition and gate patterning [Fig. 13.22(a)], spacer formation, and selective epitaxy formation of nMOS S/D [Fig. 13.22(b)] and pMOS S/D [Fig. 13.22(c)]. Advanced CMOS devices use selective epitaxy growth to form the S/D to achieve stronger channel strain and better junction profile control. Silicide formation and ILD deposition [Fig. 13.22(d)] finish the first half of the gate-last HKMG process. Dielectric CMP opens the polysilicon [Fig. 13.22(e)], and wet etch selectively removes the dummy polysilicon gate and gate oxide [Fig. 13.22(f)]. High-κ gate dielectric ALD and metal ALD form the HKMG [Fig. 13.22(g)], and bulk metal deposition and CMP are the last of the gate-last HKMG process [Fig. 13.22(h)].

Different metals are needed for nMOS and pMOS to adjust the work function and threshold voltage. One method deposits TiN for pMOS, then deposits a protective layer of tantalum (Ta). Using a photoresist mask to protect pMOS and remove Ta from nMOS, a layer of TiAl alloy is deposited that reacts with the TiN underneath to form TiAlN after thermal annealing. Because of the Ta barrier layer, there is no chemical reaction between TiN and TiAl on the pMOS side. TiAlN works well for nMOS, while TiN works well for pMOS.

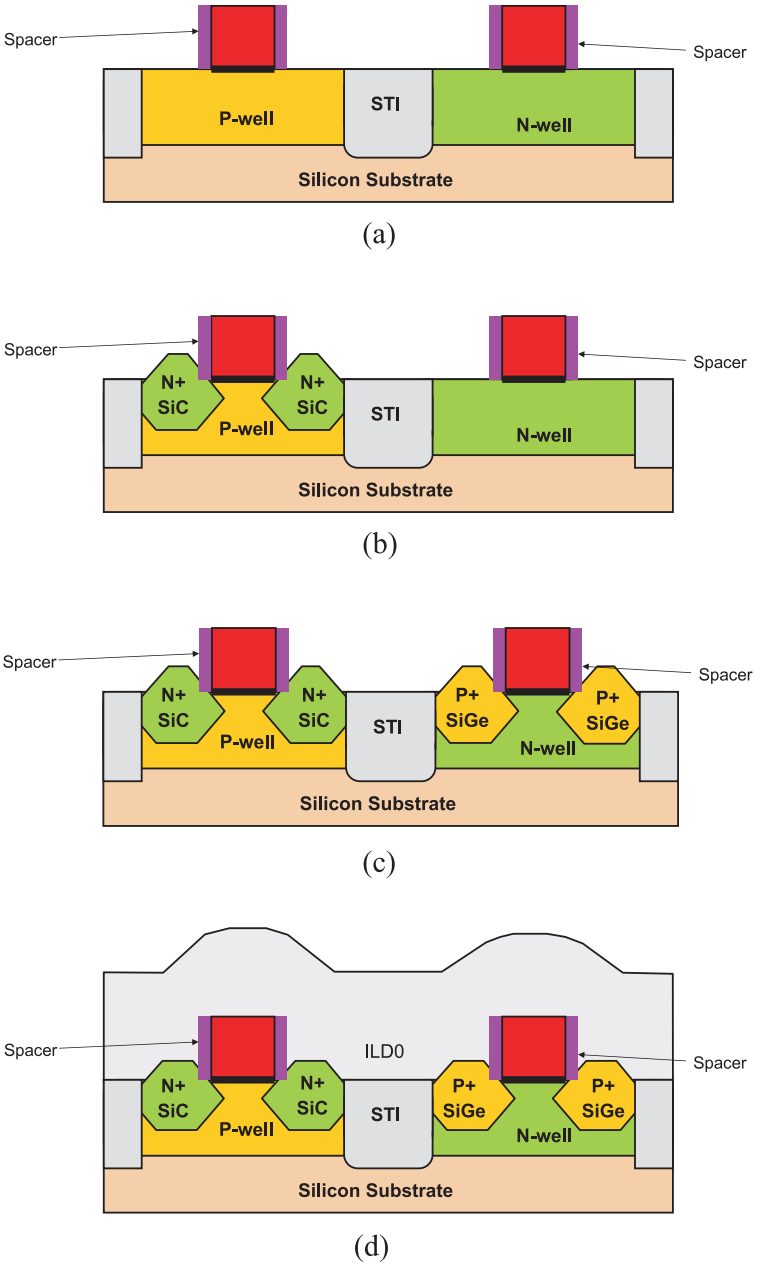
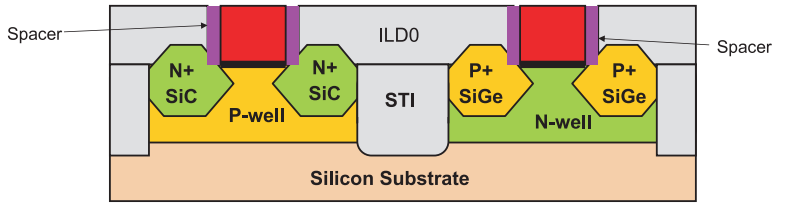
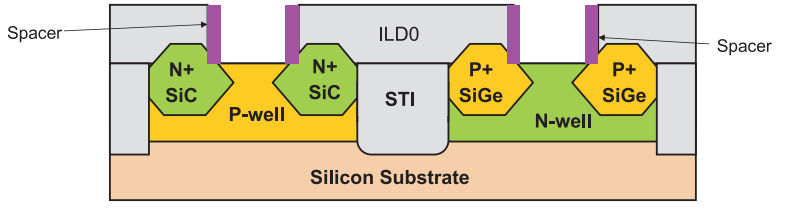


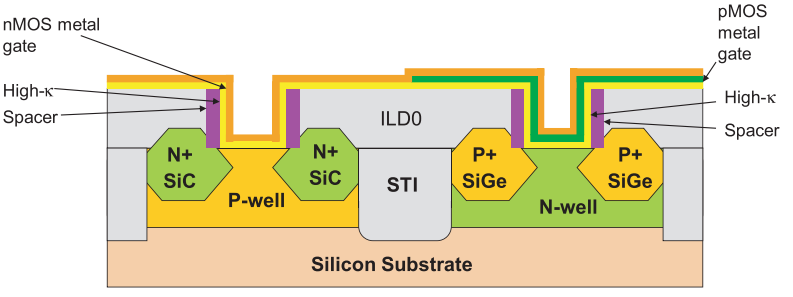
Figure 13.22 Gate-last HKMG CMOS process flow.



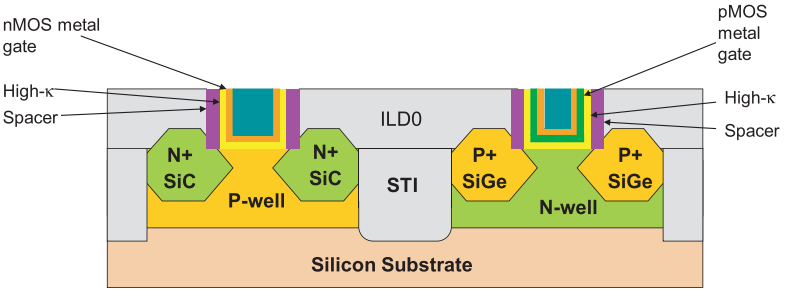
(e)



(f)



(g)



(h)

Figure 13.22 (*continued*)

13.6.3 Hybrid high-κ metal gates

One big advantage of the gate-last HKMG process is the enhancement of channel strain by removal of the dummy gate, a step that is more significant for pMOS performance improvement when selective epitaxial SiGe is used to form the pMOS S/D. Therefore, a hybrid approach that combines both gate-first and gate-last processes has been developed. Figure 13.23 illustrates a hybrid HKMG process. The nMOS uses gate-first integration, while pMOS uses gate-last integration.

13.7 Interconnections

After MOS transistors are formed on the wafer surface, the wafer process is approximately half done. The front-end process is finished, while the back-end process has just begun. For an advanced IC chip, multiple metal layers with dielectric material in between are required to route millions or billions of transistors. Silicide is commonly used to improve conductivity of local interconnections and reduce contact resistance. For many devices that do not require more than a gigahertz (GHz) of clock frequency (such as most DRAM and flash memory devices), tungsten and aluminum alloys are still widely used in metallization processes, along with undoped silicate glass (USG) as the dielectric. At the turn of the century, interconnection technology made a transition from traditional tungsten and aluminum alloy interconnections to copper interconnections for high-speed (clock frequency >1 GHz) CMOS logic devices.

13.7.1 Local interconnections

A local interconnection is the connection between neighboring transistors. It is formed by polysilicon or a polysilicon-silicide stack. Tungsten silicide and tungsten/tungsten nitride are extensively used in DRAM local interconnection applications. For flash memory devices, cobalt silicide is widely used for local interconnections. For CMOS logic devices, frequently used silicides for local interconnects are titanium silicide (>180 nm), cobalt silicide (250 to 90 nm), and nickel silicide (65 nm and beyond).

Tungsten silicide used in DRAM chips is deposited by a CVD process, with WF_6 as a tungsten precursor and SiH_4 as a silicon precursor. Titanium silicide is formed by sputtering titanium onto a silicon surface, then thermally annealing the wafer to induce chemical reactions between the titanium and silicon. Cobalt silicon is frequently used in flash memory devices and has been used in CMOS devices for 250- and 90-nm technology nodes. For 65-nm technology nodes, silicide materials had to make another major change, and nickel silicide became widely used. A process flow for tungsten silicide gates and local interconnections is listed and illustrated in Fig. 13.24.

Titanium silicide is used in local interconnection processes in some less-advanced flash memory devices and CMOS IC chips. It is formed by a self-aligned silicide (salicide) process. Compared with tungsten silicide, titanium silicide has lower resistivity.

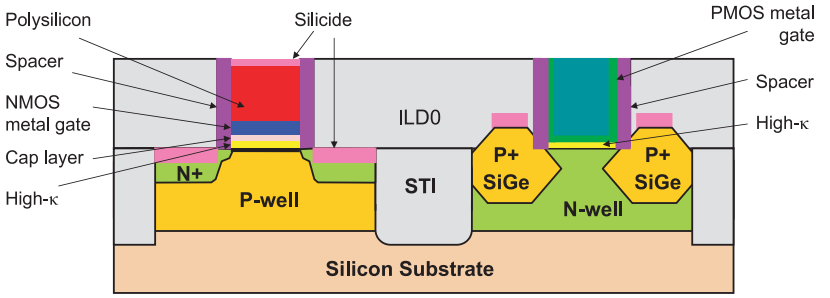


Figure 13.23 Hybrid HKMG CMOS with gate-first integration for nMOS and gate-last integration for pMOS.

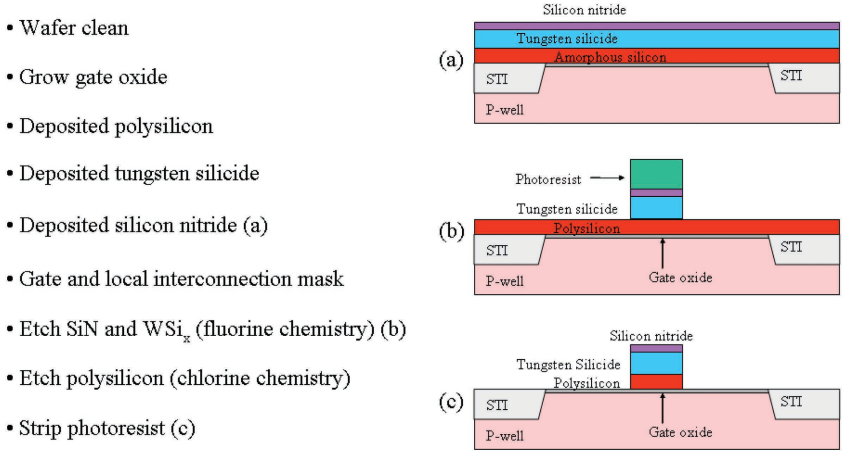


Figure 13.24 Tungsten silicide gate and local interconnection formation.

The grain size of low-resistance, C-54 phase titanium silicide is about 0.2 μm. When gate widths become smaller than 0.2 μm, titanium silicide cannot be used. Therefore, cobalt silicide was introduced into local interconnection applications. Cobalt silicide has low resistivity, and it can also be formed in a self-aligned silicide process. Both $TiSi_2$ and $CoSi_2$ formation require ~750 °C annealing, which is too high for devices with feature sizes 65 nm or smaller; thus, nickel silicide processing has been developed and implemented in CMOS IC manufacturing. The annealing temperature of NiSi is about 450 °C, significantly lower than the 750 °C required by $CoSi_2$. Figure 13.25 lists and illustrates nickel silicide processing steps.

Tungsten has been used to form local interconnections, which can significantly reduce resistance and improve device speed. Tungsten local interconnections use damascene processes similar to the tungsten plug process. First, trenches are etched in the silicate glass layer; then titanium and titanium nitride, and diffusion barrier

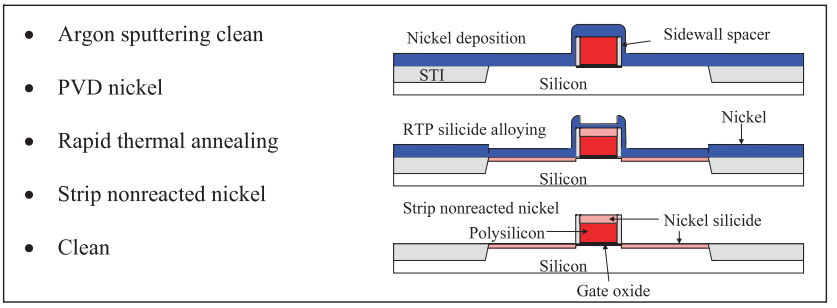


Figure 13.25 Nickel silicide formation.

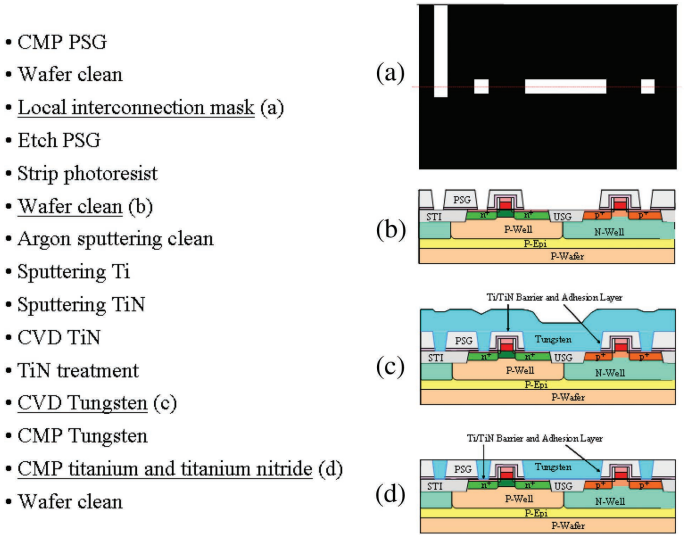


Figure 13.26 Tungsten local interconnection processing steps.

and tungsten adhesion layers, are deposited. A WCVD process fills trenches and contract holes with tungsten, and a CMP process removes the bulk tungsten from the wafer surface, leaving tungsten in the trenches and contact holes. A tungsten local interconnection process is listed and illustrated in Fig. 13.26.

13.7.2 Early interconnections

The steps of early global interconnection processing are oxide CVD, oxide etch, metal PVD, and metal etch. Oxide etching forms contact or via holes, and metal etching forms interconnection lines. Figure 13.27 lists and illustrates the early interconnection process.

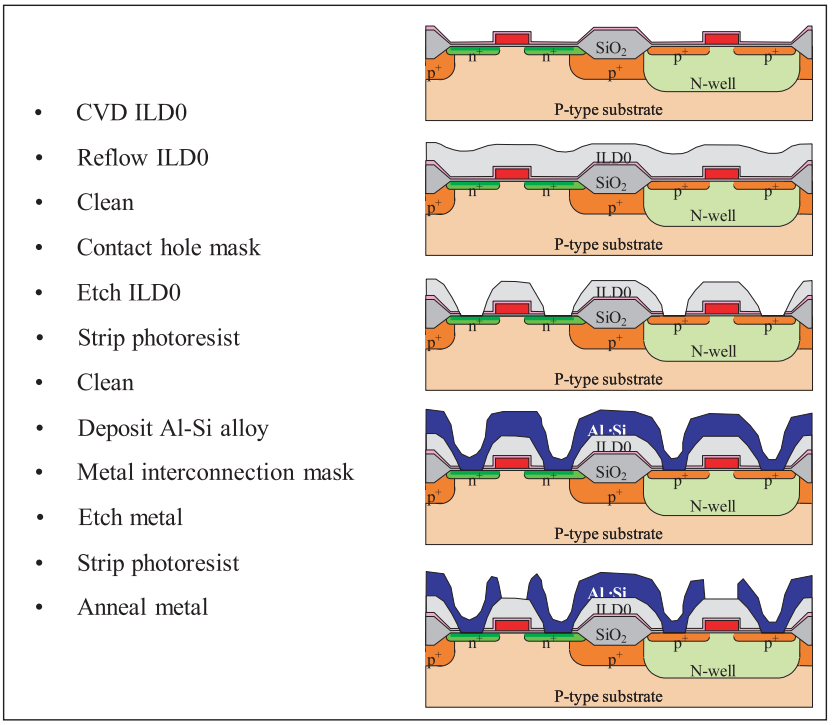


Figure 13.27 Early aluminum interconnection process.

13.7.3 Aluminum alloy multilevel interconnections

When the number of transistors increased, one layer of metal interconnections could not route all of the transistors on the chip; thus, multilayer metal interconnections began to be used. The earlier interconnection processes always left a rough surface, which caused problems for photolithography and metal deposition. As device feature size scaled down, there was no room for widely tapered contact holes, allowing PVD aluminum to cover the bottom of the holes. The tungsten CVD process was introduced to fill the narrow contact and via holes. The basic interconnection processing steps are dielectric CVD, dielectric planarization, dielectric etch, tungsten CVD, bulk tungsten removal, metal stacked PVD, and metal etch. A commonly used dielectric is silicate glass, both doped (such as PSG or BPSG for ILD0) and undoped (USG for IMD or ILDx). Dielectric planarization can be achieved with thermal flow (only for ILD0), etchback, and CMP. Bulk tungsten removal also can be achieved by both etchback and CMP. CMP became more popular in dielectric planarization and bulk tungsten removal applications in the late 1990s. The dielectric etch forms the contact and via holes. A frequently used metal stack includes a titanium welding layer, aluminum–copper alloy as the main conductor, and titanium nitride as an ARC. Metal etch defines the metal interconnection lines. Figure 13.28 lists and illustrates