

Chapter 8

PATTERN GENERATION

This chapter is intended to give readers a fundamental understanding of the mask making, or pattern generation, process. In addition to that, the chapter will introduce a few resolution enhancement techniques and principles. We start out in Section 8.1 with an introduction of the overall mask-making process flow and the requirements for masks. Further explanations of the steps are given in Sections 8.2 and 8.3. Section 8.2 covers the front-end processes, including e-beam writing and resist chemistry. Section 8.3 covers the back-end processes, including inspection, repair, cleaning, and pellicle mounting. The last section, 8.4, explains the principles of a few types of resolution enhancement techniques RET such as phase-shift masks, optical proximity corrections, off-axis illumination, and subresolution assisting features. The use of these RETs do affect mask making. The implications are illustrated.

8.1. Introduction

Pattern generation or mask making is a process in which a layer of the IC circuit pattern is engraved onto an absorber-on-quartz blank; this results in a photomask. The photomask is then repeatedly used for photolithography in wafer manufacturing. In contrast to the pattern generation, photolithography is a pattern transfer step, in which the pattern on a mask is transferred, or printed, onto wafer surfaces. Comparing to wafer processing, mask making is done on quartz substrates with a single layer of absorber. There are no topography issues. However, because this is the mother pattern that will

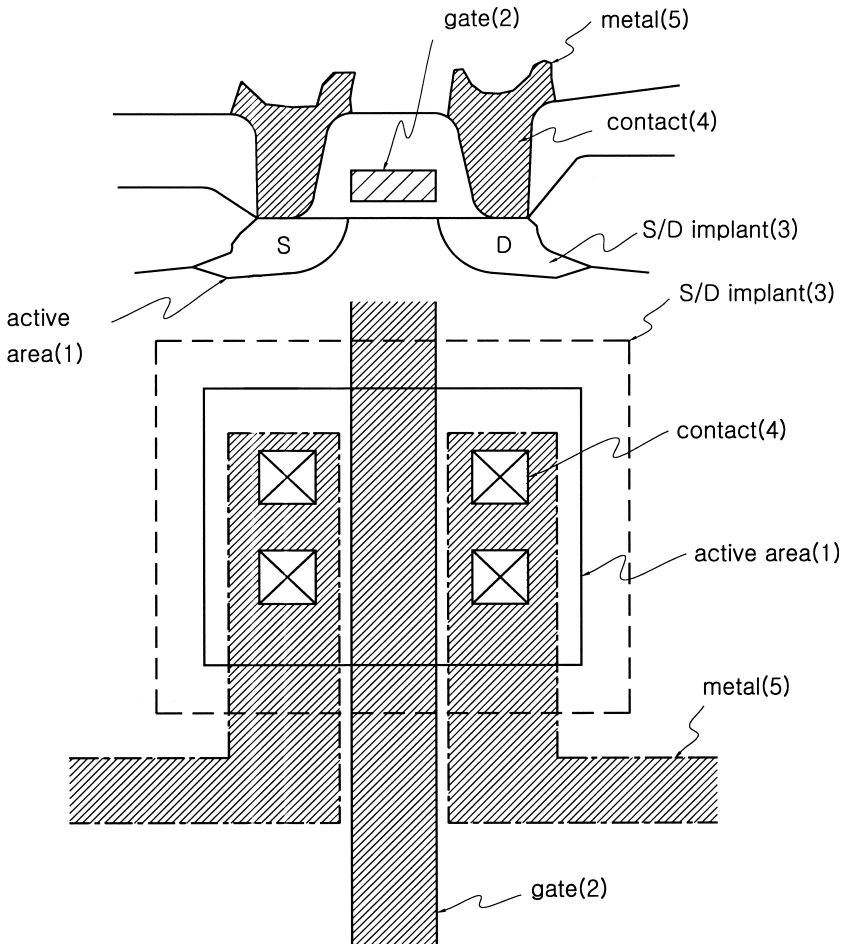


Fig. 8.1. A schematic MOS transistor structure with five masking layers. The numbers in the parentheses indicate the masking sequence.

be repeatedly printed on wafers, the quality must be impeccable, in the sense that no printable defect is allowed. Otherwise, any defect will be printed on every wafer.

An IC product design, be it a consumer product or a memory product, starts from a market survey and consolidation and a product definition. The product is then designed and verified. Once the product design is complete, the database is sent to a mask shop. The

mask shop receives the database on which the circuit is presented in the form of physical representations composed of polygons. The circuit is also split into a number of layers; each layer is drawn with a distinctive color or pattern of stripes. Figure 8.1 shows an example of a simple MOS structure with its physical layouts for each layer. Each of these layers corresponds to a mask. It is customary in the industry to refer to a complete chip on a mask or a wafer as a die. On each mask, there could be one die or a number of dies, depending on the die size and the image field size of the exposure tools. Figure 8.2 shows a single-die mask and a multidie mask. For the former, one exposure of the mask prints a die on a wafer; for the latter, one exposure prints two dies on a wafer. A complete device structure is then processed from the bottom to the top. Each mask corresponds to a photolithography step, followed by an etching or implantation step. A typical $0.25\text{-}\mu\text{m}$ logic product has about 25 layers of masks.

A typical mask-making process with a quality checking procedure is shown in Fig. 8.3. Quartz is chosen because of its low thermal expansion coefficient and high transmittance for the exposure light. The blank is about one quarter of an inch thick, to avoid deformation due to gravity. The blank is cleaned and then sputter-coated with absorber; the most commonly used material is chromium. The thickness of the Cr is around $500\text{--}1200\text{ \AA}$, depending on the technology node. On top of the Cr layer, a thin layer of chromium oxide is

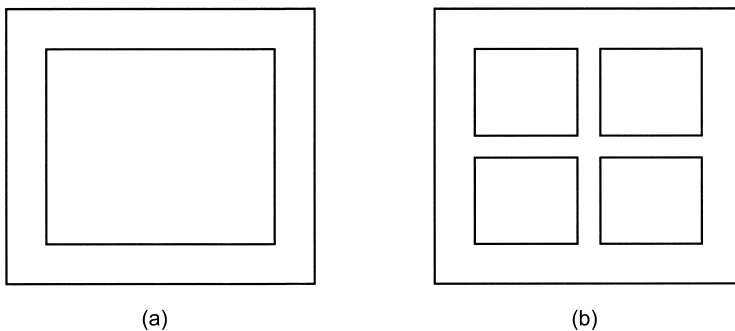


Fig. 8.2. A (a) single-die versus a (b) multidie mask: the first prints one die on wafers with one exposure, while the second prints four dies.

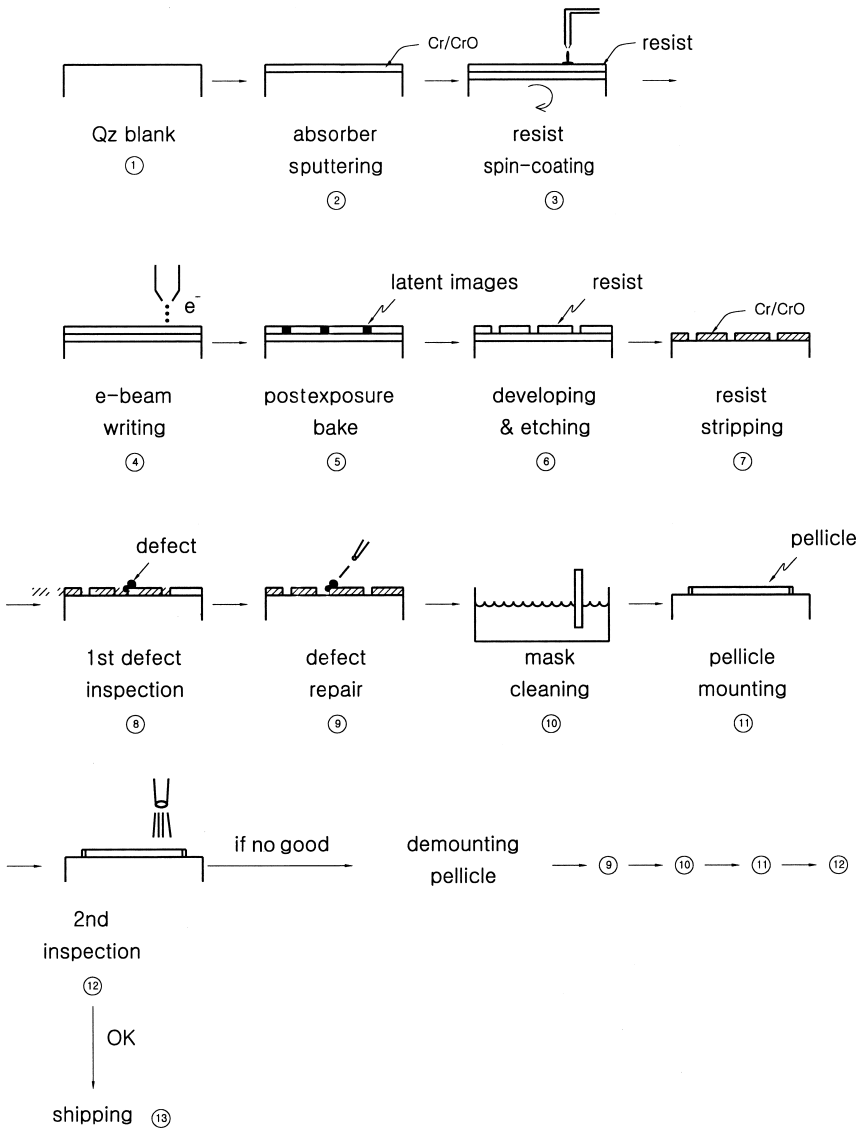


Fig. 8.3. A typical mask-making process with quality checking flow.

deposited to lower the Cr surface reflectivity, which could induce stray light, degrading wafer printing quality. The flatness of a mask blank surface is around 0.5–2 μm. There are two types of blanks: one for binary masks and the other for phase-shifting masks. For binary

masks, Cr is used as the absorber; the tone is either clear (without Cr left, light transmitting) or dark (covered with Cr, opaque). For a phase-shifting mask, a phase-shifting material (mostly MoSi_xNO) is used to shift the phase angle of the incident light so as to enhance the image contrast. Further details will be discussed later in this chapter. After the absorber is deposited, photoresist is then spin-coated onto the square substrate.

Unlike wafer manufacturers, most mask shops start their mask-making process from resist-coated blanks provided by external suppliers; that is, they leave the quartz blank polishing, cleaning, Cr sputtering, and resist coating to the suppliers. There are a couple of reasons for this. One is because the volume is often too low to be economical for an individual mask shop; the other is associated with the difficulty of resist coating on square substrates. A mask shop starts the mask-making process with e-beam writing on the substrate, using mask layer data transferred from computers. The circuit layer data must be converted to a format that is readable to the e-beam writer at this stage. After the exposure, chemical reactions take place in the bulk of the resist. The blank is often placed in a hot plate for post-exposure baking to allow the reactions to continue in a controlled ambient environment. The latent image is then developed after post-exposure baking. After that, the pattern on the resist is transferred onto the absorber, either Cr or shifter, using plasma or wet chemical etchings. Resist is then stripped off after the etching is complete.

In contrast to wafer processing, the mask has to be defect-free; otherwise, the defects get printed on every single wafer that goes through the exposure. Therefore an inspection procedure is needed to capture the possible defects. There are two approaches to inspection. One is the die-to-die approach, comparing the manufactured patterns of two neighboring dies on the same mask. The other is the die-to-database approach, comparing the manufactured mask patterns to the database. Any differences found in the comparison are considered as defects. The defect coordinates are transferred to a repair machine. The repair machine removes the extrapattern defects with an ion beam or a laser beam and fills the void defects with an ion-beam-induced or laser-induced microdeposition of carbon.

A cleaning procedure is required after the repair to clean off residues and particles left on the mask. Just as with the passivation at the end of wafer manufacturing, in mask making, a pellicle is needed to protect the mask pattern from being scratched. More important, the pellicle prevents particles from directly landing on the quartz surface. Particles on the pellicle surface are not printable on wafers as they are off-focus. A second pattern inspection is required to ensure that no particles are incorporated during mounting. Once the second inspection result is acceptable, the mask is ready for shipping.

The ultimate requirements for mask making are threefold. The first requirement is having good critical dimension uniformity (CDU) across the mask and the mean CD-to-target difference. Figure 8.4 shows the trend for CDU requirements for advanced technologies. Unlike wafer measurements, mask CDUs are often expressed in a range (maximum CD–minimum CD) instead of 3σ , mainly due to the relatively small number of measurement points. These normally range from 20 to 100 points per mask made. Nowadays, there is a demagnification factor, $4\times$ or $5\times$, for photolithography technology.

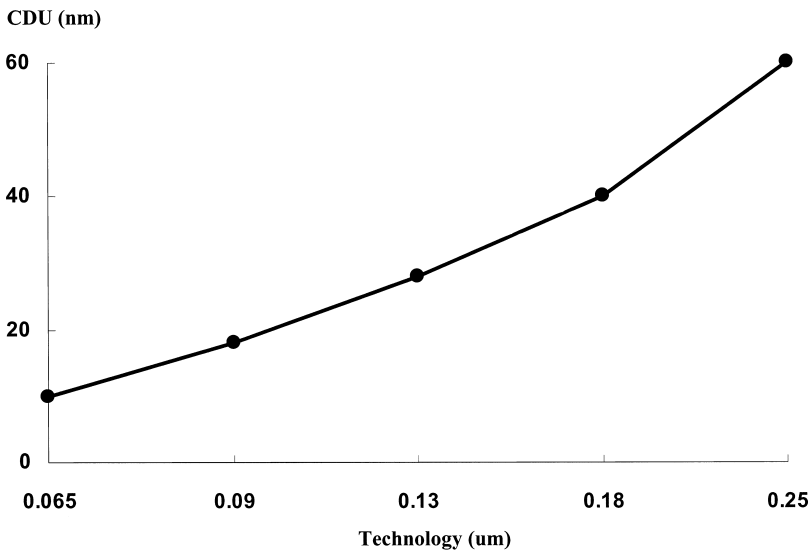


Fig. 8.4. Mask CDU requirement for various technology nodes.

This allows more forgiving specifications for mask making. However, as photolithography technology advances to a point where the to-be-printed feature sizes are close to the operating wavelength, there emerges an error factor, the mask error enhancement factor (MEEF):

$$\text{MEEF} = \frac{\partial(\text{CD}_{\text{wafer}})}{\partial(\text{CD}_{\text{mask}})/4}. \quad (8.1)$$

The perfect MEEF is unity, but a severe diffraction effect tends to enlarge it. In general, the MEEF increases with decreasing geometry and increasing pattern density. The MEEF is also related to the photolithographic process parameters and the resist system used. Figure 8.5 demonstrates that the MEEF for an advanced mask tends to increase with decreasing design pitches. One can see that the CD error on a mask can cause the same size of wafer CD error if MEEF is close to 4. This seems quite common in nanometer technologies. It is this factor that pushes mask CDU specifications. The difference between the mean CD to target is another important mask CD specification. If the CD were off target, it would be very difficult for the wafer exposure tool to shoot the correct CD on the wafer, which in turn would significantly affect device performance.

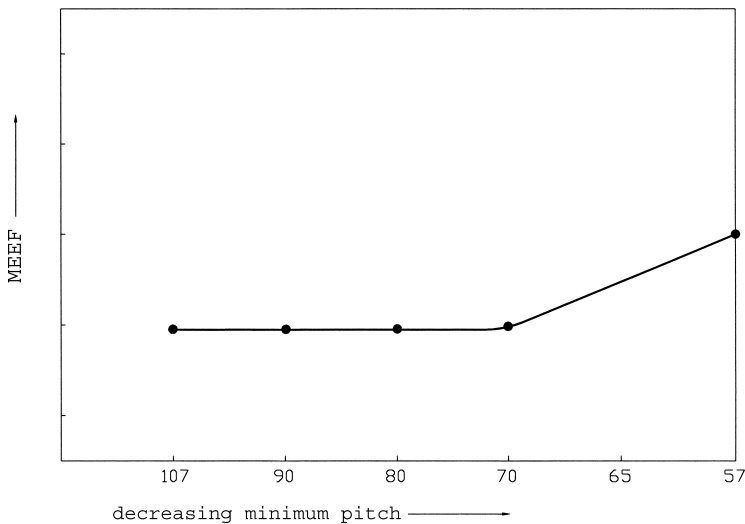


Fig. 8.5. MEEFs tend to increase with decreasing design pitches.

The second requirement for mask making is that all the masks of a whole set must be aligned with each other. All mask layers must be aligned to a reference file, which is composed of the coordinates of a number of alignment marks. Once a reference file is set up on an overlay measurement tool, it is considered the reference grid. In mask manufacturing, each mask is embedded with the same number of alignment marks, evenly located on the edges of the mask patterns. For each mask made, the positions of these marks must be measured and compared to the reference coordinates. The maximum deviation values of the x - and y -directions are considered the overlay errors. Again, the mask overlay readings are reduced by the demagnification factor of a wafer exposure system, $4\times$ or $5\times$; that is, the mask-induced wafer alignment error equals the mask overlay error divided by the demagnification factor. As long as the mask overlay errors are known, one can set up the wafer exposure so as to partly compensate for the errors.

The third requirement for mask making is that each mask must be free of printable defects. The printability of defects depends on the resolution of the exposure tool. If defects are too small to be resolved under the exposure tool, they are considered nonprintable; there is no need to repair them.

8.2. Electron Beam Writing and Resists

8.2.1. *The e-beam system*

Electron beams are often used to generate circuit patterns on masks, mainly because of their high resolution and accuracy. Electron beam lithography systems have evolved from SEM systems. Figure 8.6 shows a schematic of an electron beam system. It includes the e-gun, the column, the stage, and the control computers. The electron beam emitted from the e-gun passes downward through the column, which focuses and deflects the beam toward the substrate. The substrate is placed and fixed on a stage, the movement of which is controlled with respect to the beam position so as to control placement accuracy. The stage is mounted on an antivibration table that absorbs all

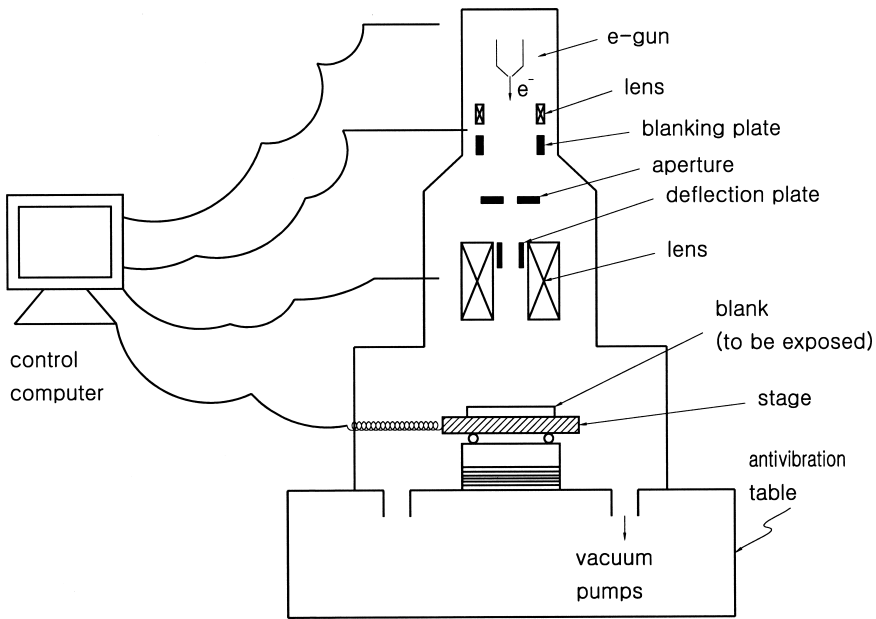


Fig. 8.6. A schematic of an e-beam system, composed of an e-gun, column, stage, and control computers.

vibrations from the environment. The control computer is the brain of the system, controlling the beam and stage movement as well as data transfer.

An e-gun is used to deliver an electron beam with a narrow electron energy range. There are two ways to extract electrons from a conducting solid surface. One is thermo-ionic emission, as shown in Fig. 8.7. This method resistively heats the solid to a very high temperature so that electrons in the solid material essentially evaporate from the solid surface into the surrounding space. The other method is field emission, as shown in Fig. 8.8. This method applies a high electrical field ($\sim 10^9$ V/m) to accelerate electrons to escape from the solid surface. For e-beam mask writing applications, the lanthanum hexaboride crystal is a commonly used electron beam source because of its relatively high brightness, low electron energy spread, and long lifetime as compared to a tungsten filament. Lanthanum hexaboride

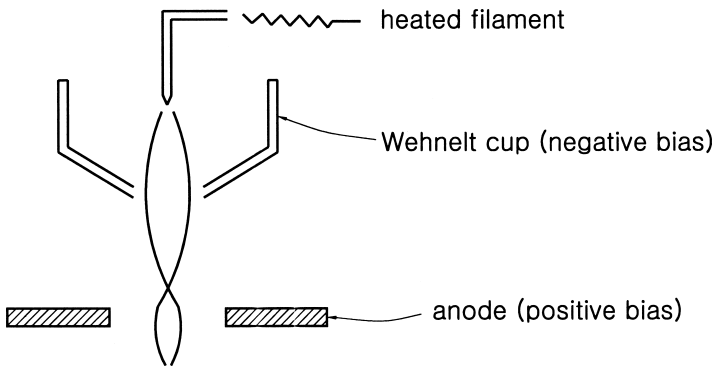


Fig. 8.7. Thermo ionic e-gun: electrons are emitted from a heated filament tip, moving toward the anode.

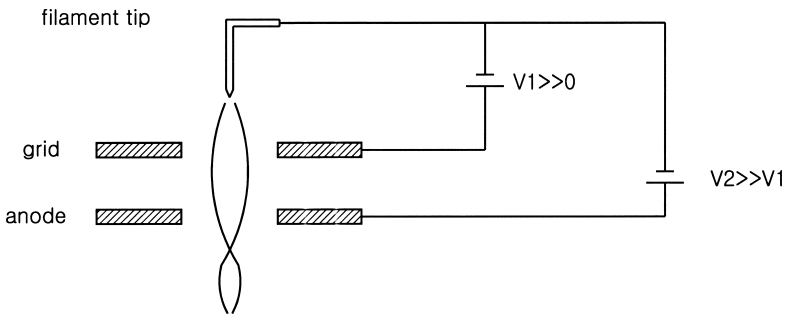


Fig. 8.8. Field emission e-gun: a strong electrical field ($> 10^9$ V/m) is applied to extract electrons from the filament tip.

can be grown in a Tammann vertical furnace by mixing La, B, and high-purity LaB_6 . The grown crystal is then further polished and machined into a tip shape, as needed.

The tip is mounted on a cathode. As the cathode is heated up, an electron stream is accelerated down the column by the positive potential of the anode plate. Electrons exit through the small hole of the Wehnelt cup (a grid). The negative potential of the cap pushes the electron stream toward the optical axis of the column, forming a cross-over point below the grid, which is the virtual source size of the

e-gun. The emission current increases with the power input to the filament; so does the brightness. The higher brightness represents a higher dosage that the beam can deliver to the exposure substrate per unit time. In particular, the higher the brightness is, the higher the exposure throughput will be. Unfortunately, as the beam current increases, the Coulomb effect (interactions) among the electrons in the beam increases. This results in electron energy spread (aberrations), which degrades the beam resolution. Refocusing through the magnetic lenses can improve the edge resolution.

The positively biased anode aperture extracts the electron beam, only allowing electrons to pass through its center hole. The voltage of this anode plate determines the energy of the electron that lands on the substrate. The electron beam can be focused or deflected electromagnetically or electrostatically, just like a conventional lens that can be used to refract the light, as illustrated in Fig. 8.9. A magnetic field, if properly set up for an electron beam, can send moving electrons into a spiral motion, hence focusing the beam. A magnetic lens is basically a copper coil embedded in an iron pole. The opening of the pole gives the magnetic field. By adjusting the current in the copper coil, one can alter the magnetic field strength.

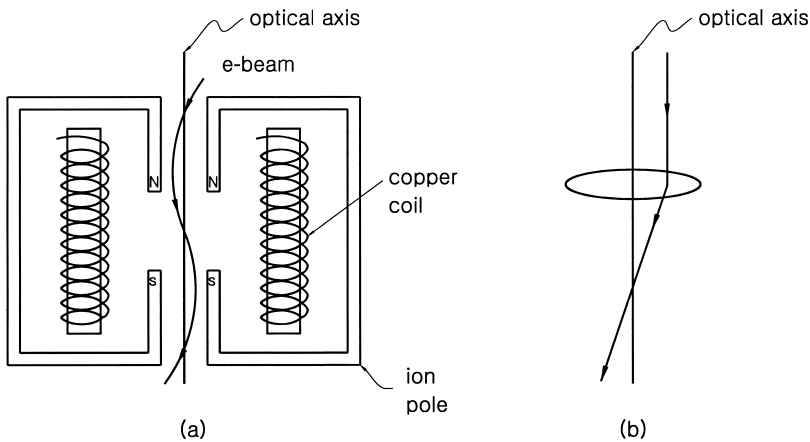


Fig. 8.9. A magnetic lens (a) versus an optical lens (b); both are capable of bending the incoming rays.

The electrical field, on the other hand, can directly deflect the beam. Important characteristics of an electron beam source include source size, brightness, and energy spread. The blanker can basically turn the beam on and off. The aperture lets the portion of the beam pass through the aperture, shaping the beam cross section.

There are various mask writing (exposure) strategies for e-beam mask writers. Each has its own pros and cons. The writing strategy basically involves varying the electron beam shapes and the ways the beam moves during writing. There are three common types of electron beams. The Gaussian beam is essentially a focused beam spot, that is, the cross-sectional area of the beam at the cross-over spot right after the Wehnelt cup. The beam size is fixed. The Gaussian beam is most often coupled with raster scan, in which the beam rasters through every location of the blank. As the beam rasters, the stage moves perpendicular to the direction of beam movement, as shown in Fig. 8.10. The beam is blanked off in areas where exposures are not intended.

A fixed shaped beam uses a large spot beam shaped by an aperture. It exposes large features such as polygons or parts of a polygon. An even more flexible and larger size of beam is the variable-shaped beam, the beam size and shape of which can be varied according to

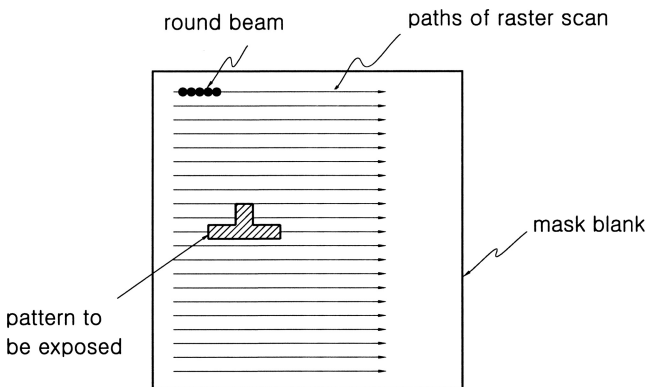


Fig. 8.10. A typical raster scan e-beam writing strategy. A fixed Gaussian round beam rasters through every location of the blank.

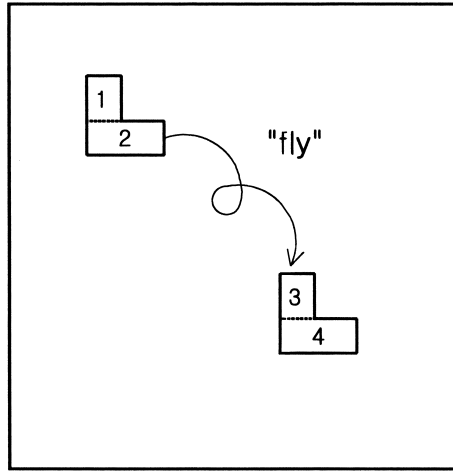


Fig. 8.11. A typical variable-shaped vector scan e-beam writing strategy. The beam shape changes when arriving at each small figure.

the figures that need to be exposed. This approach greatly improves the exposure throughput if coupled with a vector scan. A vector scan differs from a raster scan in that it does not sequentially go through every location of the blank; instead, it finishes exposing one local area and then flies to the next. Figure 8.11 shows a vector scan with variable-shaped beams. The obvious disadvantage of the vector scan is that as it flies to the next exposure area, a beam settling time is needed before the subsequent exposure resumes. This settling time is not needed for the raster scan. The variable-shaped beam coupled with a vector scan is used in most advanced e-beam mask writers. This approach saves a great deal of writing time when applied on sparse patterns such as hole patterns. As pattern density increases for dense layers, such as high-end gate layers or active layer masks, the vector scan system requires a significant increase in writing time. This is not true for the raster scan because it rasters through every location anyway.

When they impinge on the resist and substrate, charged particles, such as electrons and ions, see the resist and substrate as arrays of nuclei surrounded with electron clouds. This is different from the

situation for optical waves. As a result, for e-beam lithography, the incident electrons can be forward-scattered or backward-scattered, depending on the incident angles and the atomic number of the target atom. Owing to their light weight, electrons lose an insignificant amount of energy on scattering. When the electrons are scattered back to the resist bulk, they expose the resist in exactly the same way as the primary incident electrons from the e-beam source, as illustrated in Fig. 8.12. These scattered electrons affect (add on to) the exposure dosage of the neighboring patterns, causing the CDs of a specific feature to vary with neighboring pattern density. This is called the proximity effect. For the same target CDs in positive resists, features located next to a blanket exposure tend to be smaller than those located next to an unexposed area or dense patterns. The range of influence of the backscattered electrons tends to decrease with increased electron beam voltage; the higher the electron voltage, the smaller the affected range. On the other hand, thinner resist is less affected by the scattered electrons. As a result, the minimum achievable feature size decreases with increasing e-beam voltage and decreasing resist thickness. To obtain uniform CD across the whole

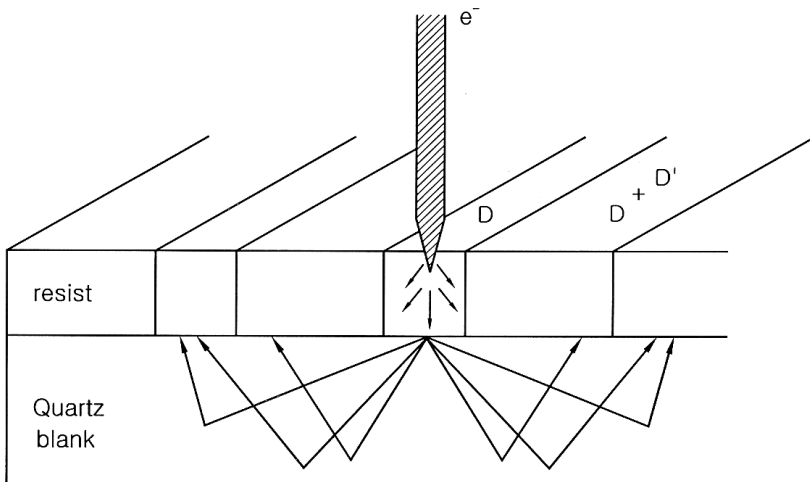


Fig. 8.12. The electron beam backscattering phenomenon.

mask, extra dosage compensation, that is, optical proximity correction, must be implemented. There are two approaches for achieving this purpose: one is to adjust the exposure energy according to adjacent pattern densities, and the other is to compensate for the dosage of the backscattered electrons with a defocused predosage.

8.2.2. The e-beam resist

In most resists, functional groups can be cleaved with an electron beam. The cleaved molecules then undergo various reactions that lead to differential solubility during the developing stage. If the exposed area turns out to be insoluble in developer, the resist is called a negative resist; conversely, if the exposed area is soluble in developer, the resist is a positive resist.

Polymethylmethacrylate (PMMA), as shown in Fig. 8.13, is one of the early successful resists developed for DUV and e-beam lithography processes. In the 1970s and 1980s, extensive research activities were concentrated on modifying polymers to suit various needs. The electron energy causes molecular chain scissions, which result in polymers with lower averaged molecular weight. The averaged molecular weights decrease with increasing electron dosage or energy density.

The developer solution is typically a mixture of methylisobutylketone (MIBK) and isopropane alcohol (IPA). PMMA has good resolution and wide process latitude but lacks dry etching resistance and requires high dosages ($50\text{--}100\ \mu\text{c}/\text{cm}^2$ at 20 keV). The high dosage

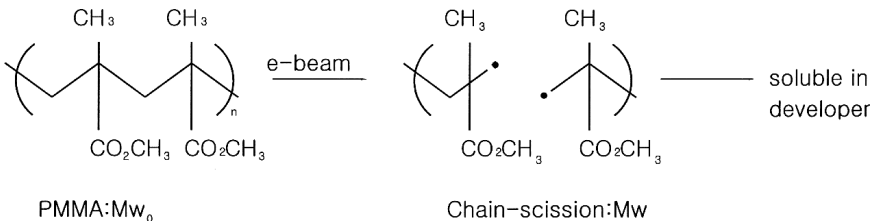


Fig. 8.13. PMMA undergoes e-beam-induced chain scission reaction.

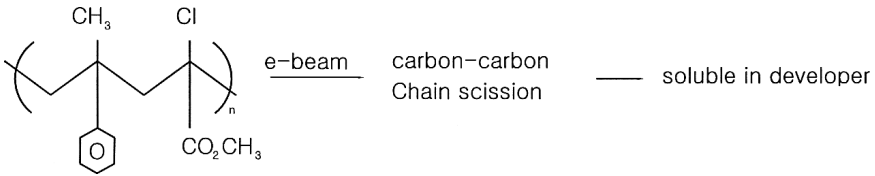


Fig. 8.15. ZEP resist structure.

0.18 μm , the CD control budget tightens significantly. Advantages offered by ZEP tend to lag behind advanced technology needs.

Chemically amplified resists (CARs) technology, as developed for 248- and 193-nm photolithography, are also adopted for e-beam lithography. CARs provide several advantages over the above mentioned e-beam resists. CARs have very good resolution and CD control and require a low dosage, around $8 \mu\text{C}/\text{cm}^2$ at 50 keV. CARs also have very good plasma etching resistance. They are abundantly available with stable quality, mainly due to the fact that they are being used for wafer photolithography production. Figure 8.16 demonstrates a typical acid catalyzed deprotection reaction of the tBOC CAR resist. Use of negative-tone CAR is gaining momentum in high-end mask production. Apart from its excellent resolution, high throughput, and plasma etching resistance, it is obviously advantageous for high pattern density layers such as high-end polygate or active layers. For these layers, the e-beam exposing area is about 60% to 80% for positive-tone resists. These translate into 40% to 20% for negative-tone resists, hence gaining throughput. Furthermore, the negative resist has much less proximity effect for isolated features than does the positive resist. For this application, a negative resist gives the potential for better CD uniformity control.

Despite the above mentioned advantages, chemically amplified resists have some lingering issues that remain to be resolved. First, owing to the action of the acid catalytic reactions, CARs are extremely sensitive to airborne bases such as ammonia. The base neutralizes the generated acids, poisoning the catalytic reaction and leading to resist scums, CDU degradation, or profile deterioration.

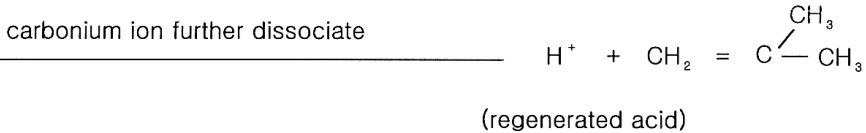
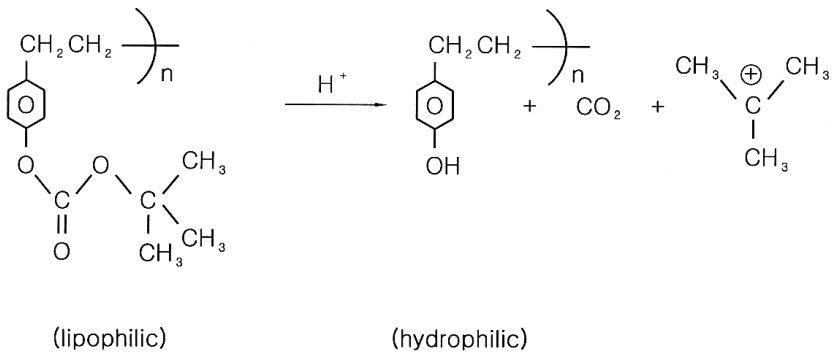


Fig. 8.16. Acid-catalyzed deprotection reaction of a positive CAR resist.

Second, CD performance tends to be affected by various delay times such as post coating delay (PCD), postexposure delay (PED), and post-PEB delay (PPD). The three delays are somewhat related to the ambient conditions. With nitrogen ambient, the CD drifts of PCD and PED are reduced, while the PPD shows marginal difference. It has also been shown that PED seems to dominate the CD drift among the three steps, even with nitrogen ambient. This is related to the fact that the generated acids are prone to base neutralization. In view of the resist stability and CD control, it is very important to keep ammonia concentration as low as possible. It can be achieved by adding chemical filters in clean room airflow paths.

8.3. Back-End Processing for Mask Making

The back-end processing, the process segment that occurs after pattern definition is done and resist is stripped, includes mask inspection, repair, cleaning, and pellicle mounting. In mask manufacturing,

there are two types of defects: soft and hard. Defects that can be removed by the mask cleaning process are called soft defects. Conversely, defects that cannot be removed with cleaning are called hard defects, and they must be repaired (removed).

8.3.1. The mask inspection

A mask inspection system, as shown in Fig. 8.17, is typically equipped with powerful computers that handle data preparation, illumination, light sources and optics, the mask handling mechanism, and the image processing and detection mechanism. The light source illuminates a reflected mirror toward the condenser lens through the mask pattern, where the light is either reflected or transmitted. Both the reflected and transmitted lights are collected and processed by computers.

There are two types of inspections: pattern inspection and particle inspection. Pattern inspection is further classified into die-to-die and die-to-database inspections. Die-to-die inspection compares

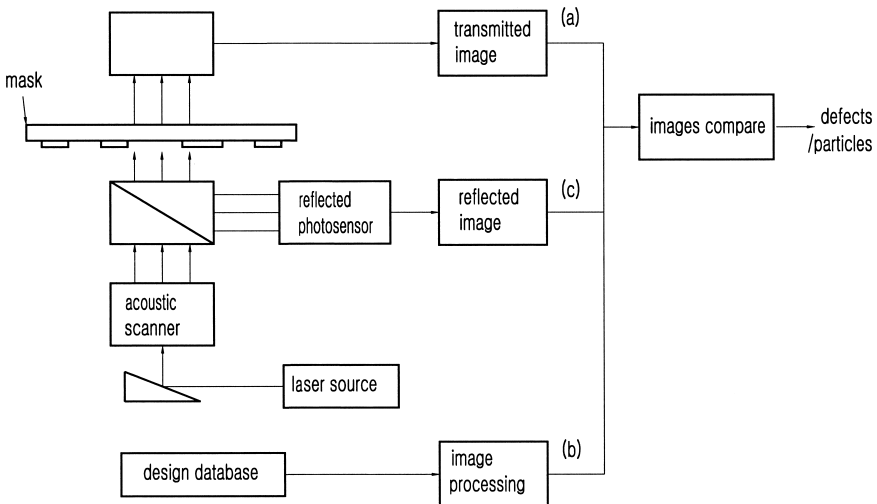


Fig. 8.17. A representative setup for a mask inspection system. (a) a , transmitted image of two neighboring die: die-to-die; (b) $a + b$, transmitted image versus processed database: die-to-database; and (c) $a + c$, transmitted image versus reflected image: particle inspection.

the transmitted images of two neighboring dies. On the other hand, die-to-database inspection compares the transmitted image of a die to its own processed image. Particle inspection (soft defect) results are revealed by the sum of the reflected and the transmitted light intensities, as demonstrated in Fig. 8.18. For a chromium surface, the reflected light accounts for 100% of the incoming light intensity. For quartz, the transmitted light is 100%. In the case of a particle on the surface, some of the light rays are refracted, and the sum of the transmitted and reflected light does not equal the incoming light intensity. The existence of the particle is therefore detected.

For die-to-die inspection, as illustrated in Fig. 8.19, the images of two neighboring dies are compared. The probability of having two

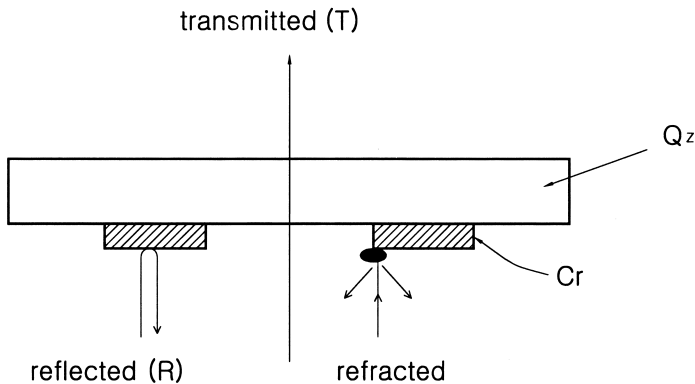


Fig. 8.18. The concept of particle inspection. In a particle that causes light refraction, $T + R < 100\%$, a defect is detected.

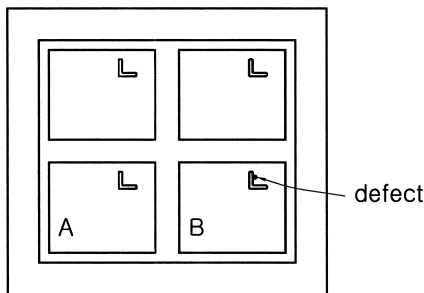


Fig. 8.19. A four-die reticle with a defect lands on the right lower corner die. Comparing A with B die, the defect can be detected.

defects at the same coordinate of two neighboring dies is nearly zero. Hence the image differences between the two dies are classified as defects. With die-to-database inspection, the transmitted light forms a die pattern image, which is then compared with the database. The database is not exactly the CAD database, but the so-called processed database, which looks very similar to aerial images of the database, as shown Fig. 8.20. The processed database images are calibrated to mask processing characteristics. The comparison shows the difference as defects. It is obvious that the defect count of pattern inspections is a result of the inspection sensitivity settings. Which differences are regarded as defects and which are regarded as false counts must be ultimately correlated to the wafer printing results. A tight sensitivity setting gives a high defect count, but many of the counted defects may be false. Appropriate inspection sensitivity often comes with experience and machine stability; however, the best approach would be determining the sensitivity with wafer printability.

Figure 8.21 demonstrates various types of mask defects. Depending on the defect types and sizes, some can be captured by the

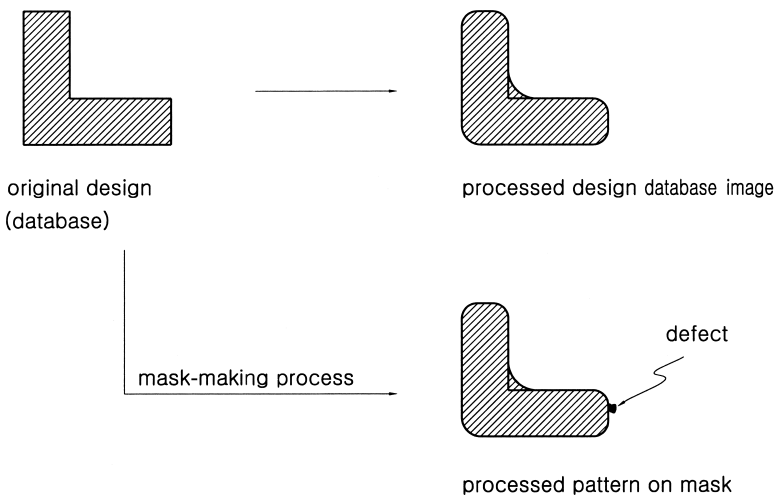


Fig. 8.20. The concept of die-to-database inspection compares the processed database image with the processed pattern on mask.

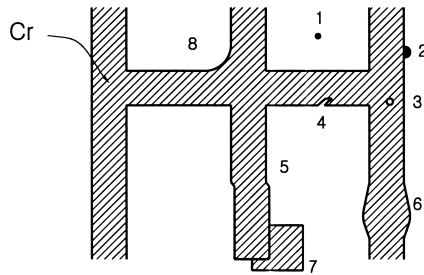


Fig. 8.21. A few examples of defects on mask: (1) pin dot; (2) extrusion; (3) pin-hole; (4) intrusion; (5) butting error; (6) CD offset; (7) extra pattern; and (8) corner rounding.

inspection machine; some cannot be captured. Machine capability is expressed in terms of capture rates; for each type and size of defects, the machine has a certain probability of capturing the defects. As the defect size approaches the machine resolution limits, the capturing probability decreases. The actual machine capability is defined as the 100% capturing rate.

8.3.2. *The mask repair*

There are two mainstream mask repair technologies: one with focused ion beams, and the other with laser beams. A focused ion beam (FIB) operates in a similar manner to a scanning electron microscope (SEM), except that the FIB uses a finely focused ion beam in lieu of an electron beam. It employs ion-beam-assisted microchemical vapor deposition for clear-tone defect filling and gas-assisted ion milling for removing opaque defects. When an ion beam hits a substrate surface, the energetic ions impart their energy to the substrate surface atoms, giving off secondary electrons, secondary ions, and sputtered atoms. For a low ion energy level, the collected secondary electrons or ions can give surface image information. With a higher level of ion energy, the incident ions can be used for ion milling an opaque spot, that is, for defect repair, as indicated in Fig. 8.22. An ion beam repair system is composed of several parts. The column holds the ion beam optics, and gallium ions are used in the ion gun. The stage holds the mask to be repaired with an interferometric positioning mechanism.

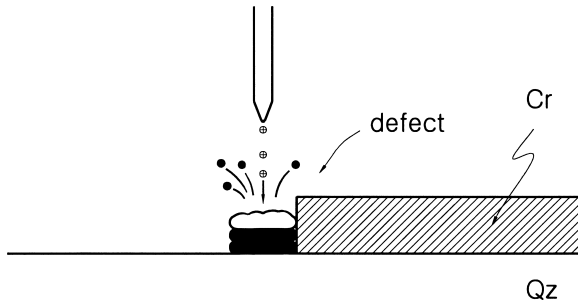


Fig. 8.22. FIB mask repair system uses Ga for milling out the defect.

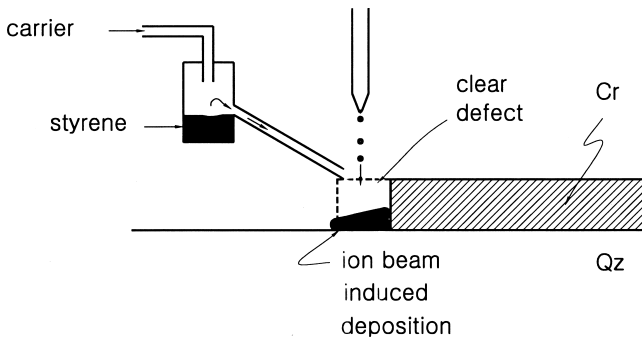
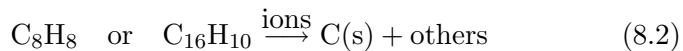


Fig. 8.23. Clear defect (missing Cr) repair uses ion-beam-induced polymer decomposition to form an opaque film.

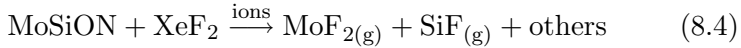
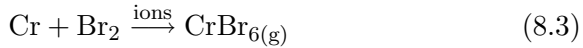
An electron gun floods electrons onto the substrate to neutralize the ions used for repair. The chemical for the microdeposition is delivered through a nozzle close to the ion gun so that the deposition proceeds right at the clear defect location, as illustrated in Fig. 8.23.

To repair a clear defect, carbon-containing polymers, such as pyrene or styrene, are often used as the reacting gas. The energetic ions result in decomposition of the carbon-containing species and formation of a carbon film, an opaque film:



To repair an opaque defect, ion milling is used. The incident ions knock off the chromium atoms to remove defects. Gas-assisted etching (GAE) can be used to enhance the defect removal rate. Halogen

gases are good candidates for Cr removal, while XeF_2 can be used for phase shifter material such as MoSiON:



The drawback of ion beam repair is that the gallium ions often get implanted into the quartz substrate, causing transmittance loss. A proper postrepair procedure, such as a wet etching, is required to retrieve the lost transmittance. During the repair operation, some of the deflected ions can hit the quartz surface, as demonstrated in Fig. 8.24, knocking the Si atoms out of the quartz surface. Such a phenomenon is called river bedding. Severe river bedding also causes transmittance loss. It shows up on wafers as printable defects. In general, the ion beam can be focused onto a beam diameter of about 1/10th of a micron, much smaller than a laser beam.

Laser beam (UV) repair technology, on the other hand, uses photon energy to initiate photolytic CVD reactions, leading to deposition of material for repair of clear-tone defects and laser ablation for repair of opaque defects. The laser energy profile incident on a substrate surface is fairly close to the Gaussian distribution. Figure 8.25 demonstrates a schematic of the laser repair scheme. The laser beam (Nd:YVO_3 of 355 nm or Nd:YLF of 349 nm) rasters on the substrate surface at the defect coordinates transferred from the inspection machine. The defect coordinate information drives the stage to the defect location through a stage control mechanism. Imaging is taken

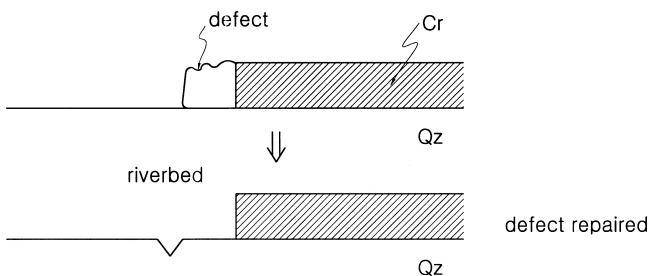


Fig. 8.24. Riverbed results from defect ions during repair.

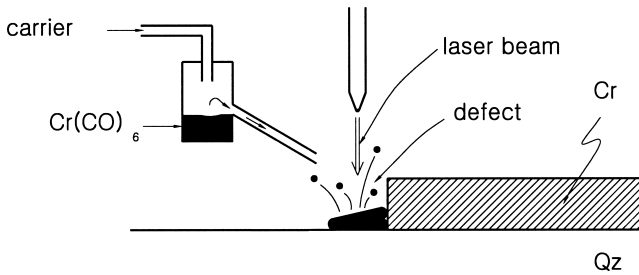


Fig. 8.25. Laser repair system removes defects via laser ablation; clear defects via laser-induced deposition of Cr.

through an optical microscope. The deposition chemical, chromium hexacarbonyl, is fed through a nozzle to the laser rastering position, where the photolytic decomposition occurs:



The reaction product, chromium, is deposited in the clear-tone defect area. It is very difficult to control the deposited film profile and the spread area for microscale chemical reactions. As long as the deposition covers the area of the defects, it is considered an acceptable repair, providing that the resulting dimension is within less than 10% of the desired value.

Repair of opaque defects employs laser ablation; the photon energy is locally absorbed in the bulk of the Cr in the defect area. The bulk temperature increases with the laser flux and dwelling time (nanosecond range). The instantaneous local temperature can rise above 1000°. The heat is also conducted vertically toward the substrate bulks (both Cr and quartz) and spread laterally. As the bulk temperature rises to the evaporation temperature, the Cr in the defect area is ablated.

The drawback of the laser repair system is that a large amount of instantaneous heat is generated, which could cause damage to the repaired edges, such as rolled up edges, or quartz damage and loss of transmittance. There have been efforts to shorten the laser pulse time from the nanosecond to the femtosecond range to shorten the effects of the generated heat. With laser beam repair technology, the laser

beam can be focused to as fine as $1\ \mu\text{m}$. This is an order of magnitude larger than an ion beam diameter. As a result, the ion beam can be used for technologies from $0.25\ \mu\text{m}$ down to $0.13\ \mu\text{m}$ and possibly beyond, while laser beam technology is used for $0.35\ \mu\text{m}$ and up. Edge misalignments (the alignment between original and the repaired line edge) are often seen during repair. For the 90-nm mask repair tool, the edge alignment accuracy is about 15-nm (3σ). Consequently, an iterative procedure is often required to achieve good edge alignment accuracy for advanced mask repair.

Is repair quality acceptable? The ultimate answer, of course, lies in the wafer printing results. However, when a mask goes to a wafer fab without confidence in the repair quality, it can be a very risky undertaking as a large number of wafers can be at risk. The conventional checking procedure is based on the operator's or engineer's educated judgment. Recent development has resulted in a subjective checking procedure. This involves looking at the repaired area with an aerial imaging system. The system includes major components of an optical column, allowing for adjusting NA, sigma, and illumination. The major difference between a scanner and an aerial image tool is that the latter looks at only a single spot, instead of the whole field. An aerial imaging system shows the intensity profile and a contour plot. To ensure good quality of the repair process, the aerial images are often calibrated against actual wafer prints to account for resist performance. The wafer results set the upper and lower limits of the aerial imaging results in terms of the CD tolerance percentage, for example, $\pm 6\%$, which are then used to judge the quality of all repaired locations on a mask.

8.3.3. Mask cleaning

Once the mask reaches the cleaning stage, it should be free of hard defects. Mask cleaning is a necessary step before pellicle mounting. It cleans out the soft defects and residues of the repair process. Soft defects adhere to the mask surface through various mechanisms such as Van der Waals forces, electrostatic forces, physical adsorption, or chemisorption. Van der Waals forces result from dipole–dipole

interaction or the electronic polarization of the surface atoms and molecules when two solid surfaces come into close contact. Van der Waals forces increase with the size and shape of the interacting molecules. Longer molecules tend to give larger forces than more spherical or symmetric molecules due to the larger extent of polarization. Flakes of resist material or chipped-off repair carbon films adhering on a mask surface often demonstrate Van der Waals forces.

Electrostatic forces are the other commonly seen forces that make particles stick to a mask surface. Their existence is attributed to the coulombic effects that occur between solid surfaces. The excess static charges on a solid surface polarize the solid surfaces of incoming solid particles, which are possibly carried by laminar fluid flow or by random motion, turning their surfaces into charges of opposite polarity, hence attracting particles onto the solid surface. Tiny particles seen on a mask surface are often visible examples of electrostatic forces. As a solid surface is exposed to ambient gas, some gas molecules could land on the solid surface and become adsorbed. Adsorption decreases with increasing substrate temperature but increases with molecule size. Precipitation, such as ammonium sulfate powder, which is seen on a dried mask surface, often results from the fact that ambient ammonia and sulfur dioxide are physically adsorbed on the mask surface, and they later react to form ammonium sulfate powders. Chemisorption is a result of adsorption, with chemical reactions occurring between the surface and the adsorbents. The binding force of chemisorption is relatively strong as compared to the other forces.

The above mentioned soft defects are supposed to be cleaned with mask cleaning procedures. Typically, a wet mask cleaning bench consists of a series of tanks containing ammonia and sulfuric acid and a de-ionized water rinse. Ammonia and hydrogen peroxide solution (typically $\text{NH}_4\text{OH}:\text{H}_2\text{O}:\text{H}_2\text{O}_2$ of 1:5:1) is also called RCA Standard Clean-1(SC-1). SC-1 can remove organic and metallic particles and flakes. Hydrogen peroxide is a strong oxidant that oxidizes the particles, and then ammonium hydroxide dissolves the oxidized products by solvation. Sulfuric acid ($\text{H}_2\text{SO}_4:\text{H}_2\text{O}_2$ of 4:1 at 80°C – 100°C) oxidizes organic (hydrocarbon) residues, such as resist flakes or pellicle glue residues, at about 100°C . There are several cycles of quick-down rinses (QDR) between the ammonia and sulfuric acid solutions

to avoid the formation of ammonium sulfate salt. Also, extensive rinse cycles are needed before the cleaning cycle is done. Failure to use enough water rinse can give rise to the formation of ammonium sulfate, white powders, on the quartz surface.

Wet cleaning recipes must be optimized in terms of defect removal efficiency and CD loss of the mask patterns, or transmittance and phase angle loss of phase shifting materials. Increases in concentrations of SC-1 and ammonium solution as well as increases in wet bench temperatures tend to increase cleaning efficiency, but they result in some negative effects such as binary mask CD shrinkage. Ammonia attacks the phase shifter material (MoSiON), causing the transmittance, phase angle, and CD changes. These variations must be characterized and routinely monitored for each wet bench to avoid mask scrap due to repetitive cleaning cycles. Generally, when a wet bench line is designed for binary mask cleaning, it may not be optimized for phase shifter masks (PSM). Furthermore, to avoid cross contamination in manufacturing, the two types of masks are cleaned in separate wet bench lines.

Following the QDR, the final step in mask cleaning is the IPA cleaning step, in which IPA is sprayed onto the mask surface and forms a thin film falling along the mask surface. Without the IPA film, the DI water dries as a film breaking down into small islands due to surface tension. Particles are easily trapped in the islands, leading to the formation of stains on the mask surface. The IPA film basically displaces the islands and drags them along the film during drying. As the IPA film dries out, the particles are removed as well.

8.3.4. Pellicle mounting

After the first mask inspection, repair, and cleaning are done, a pellicle is mounted on the mask. The pellicle is a thin polymer film about 0.8–1.5 μm thick, glued to an anodized aluminum frame, which is about 4–6 mm high. The pellicle frame is then mounted on the mask surface, as illustrated in Fig. 8.26. The primary requirement for a pellicle is its high light transmittance (>99%) and transmission uniformity (>99.9%). To meet these requirements, the composition and

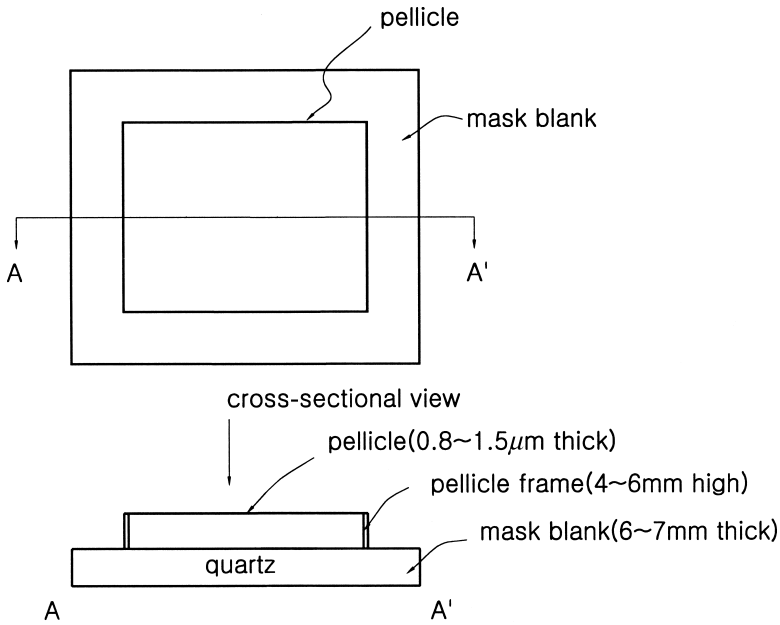


Fig. 8.26. The structure of a pellicle mounted on a mask.

thickness of the polymer must be changed as operating wavelengths shorten.

A pellicle serves different purposes. It prevents particles from landing on the mask surface. The particles would land on the pellicle surface instead. As a result, the particle is out of focus range, becoming nonprintable. A 6" mask of 6.35 mm thickness has a pellicle mounted on one side. A 5" mask is much thinner, 2.2 mm thick, and it requires pellicles mounted on both sides. Because the mask is thinner than a 6" mask, particles on either side would be printable on wafers without the pellicles. With the pellicles mounted, the masks do not need to be recleaned frequently. Most of the particles on the pellicle surface can be carefully blown away with an air gun.

A good pellicle material should possess the following characteristics:

- (a) high transmittance
- (b) high illumination endurance

- (c) excellent light transmittance uniformity
- (d) readily available and low cost.

Owing to transmittance requirements at different wavelengths, pellicle composition must vary with the operating wavelengths to have maximum transmittance. Pellicles are composed of cyclic polymers for 365 nm exposure wavelength and cyclic fluoropolymers for 248/193 nm exposure wavelength. The pellicle composition that is optimized for a long wavelength exposure may not be appropriate for a shorter wavelength exposure as the shorter wavelength may burn the pellicle. Pellicle transmittance for a fixed wavelength tends to decrease after a large number of exposures. Sometimes a so-called ghost image can be observed; that is, the circuit patterns get vaguely printed on the pellicle surface. The appearance of the ghost image indicates transmittance decay.

8.4. Resolution Enhancement Technology

Shrinking device geometry has been the never-ending game for the semiconductor industry. Photolithography is the workhorse that enables the game to move forward by pushing the resolution limits. Obvious approaches for pushing resolution limits are to use either a shorter wavelength or a larger NA, which can be realized in the equation discussed earlier:

$$\text{Min. resolvable feature sizes} = k_1 \frac{\lambda}{\text{NA}}. \quad (8.6)$$

NA of a lens corresponds to its physical size and indicates its diffracted light collecting capability, as shown in Fig. 8.27. For a resolved image, the NA corresponds to the minimal lens size in air that can collect the zero- and first-order diffracted light. Any pitches that are smaller than the minimum resolution capability would push the first-order diffracted light a farther apart, beyond the numerical aperture of the lens, and hence the image blurs. On the other hand, if a pitch is larger than the minimum resolution capability, higher-order diffracted light can be collected by the lens, which results in a better image. Improving the resolution limit by increasing the NA

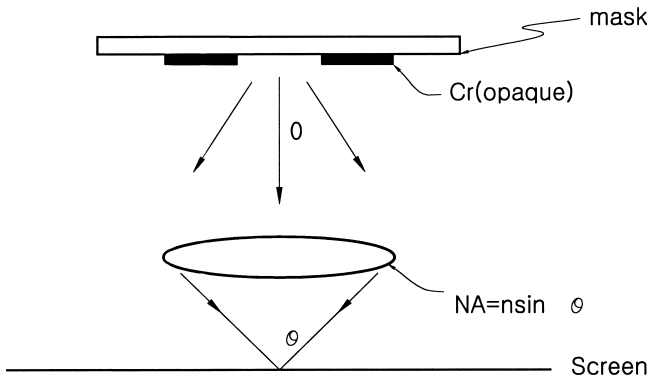


Fig. 8.27. NA of a lens equals $n \sin \theta$. It indicates the diffracted light collecting capability.

would sacrifice the depth of focus (DOF). Furthermore, maintaining superior lens quality, such as low aberration or distortion, can be very costly and difficult for the large NA. On the other hand, using a shorter wavelength would demand that several things be changed altogether, such as the resist, mask blank, and pellicle materials. Oftentimes, the market demand for the end product outruns the evolution of the NA and wavelength. As a result, with the same wavelength and numerical aperture, one often needs to push the k_1 -factor of Eq. (8.6) to achieve smaller printed feature sizes with good performance and reasonable process latitudes.

One effective way to push the k_1 -factor is to use resolution enhancement techniques (RETs). Some RETs can be realized by changing the mask patterns and structures, others, by changing the illumination approaches. Commonly used resolution enhancement techniques include phase shifting masks (PSM), off-axis illumination (OAI), subresolution assisting features (SRAF), and optical proximity correction (OPC).

8.4.1. Phase shifting mask technology

When a beam of light goes through a material with an extinction coefficient of k and refractive index of n , its intensity is attenuated

according to Lambert's law:

$$I = I_0 \exp\left(\frac{-4\pi kd}{\lambda}\right), \quad (8.7)$$

where k is the extinction coefficient of the material, d is the thickness of the material, and λ is the wavelength of the incident light. A material is called transparent if the k is zero; it is called absorbing if the k is not zero. For a chrome-on-glass mask (binary mask), the chrome is so strongly absorbing that no light can pass through, while the quartz is so transparent that all light can pass through. Both the refractive index and the extinction coefficient are functions of the material and the operating wavelengths.

Phase shifting technology takes advantage of the fact that when light passes through a material with a refractive index n , its speed slows down by a factor of $1/n$. This creates an optical path difference between the light that goes through the material and the light that does not, as shown in Fig. 8.28.

The optical path difference is,

$$\text{OPD} = (n - 1)d, \quad (8.8)$$

and the corresponding phase difference is:

$$\text{OPD} = \frac{2\pi(n - 1)d}{\lambda}, \quad (8.9)$$

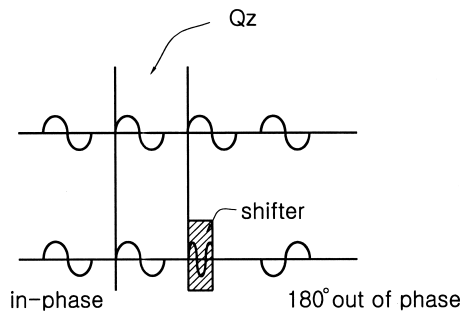


Fig. 8.28. A phase shifter with a designated n , k , and thickness can be used to shift the phase of a light wave.

where n and d are the shifter's refractive index and thickness, respectively. Now, if one desires to have a phase difference of 180° , then the required shifter thickness, d , is

$$d = \frac{\lambda}{2(n - 1)}. \quad (8.10)$$

It can be observed that the higher the refractive index or the lower the operating wavelength, the thinner the shifter is required to be achieve a desired value of phase shift. With a weak phase shifter material, such as a rim type or an attenuated PSM, the transmittance is relatively low, ranging from 5% to 15%. The purpose is to pull the attenuated transmitted light rays to the opposite phase and hence improve the contrast of the aerial images, as illustrated in Fig. 8.29.

Attenuated PSM is widely used for hole patterning. Figure 8.30 demonstrates the $0.4\text{-}\mu\text{m}$ line patterning performance improvement in terms of aerial image slope. With attenuated PSM, the slope

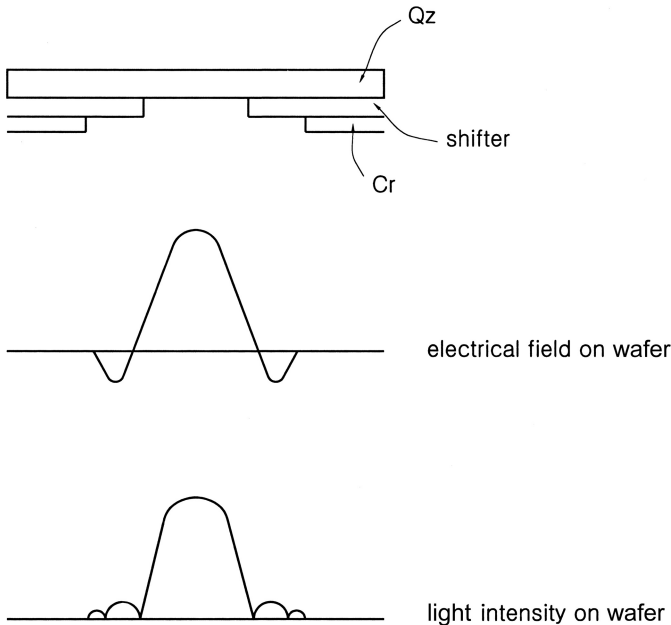


Fig. 8.29. The rim-type PSM using 5–15% of phase shifting material to enhance the edge contrast.

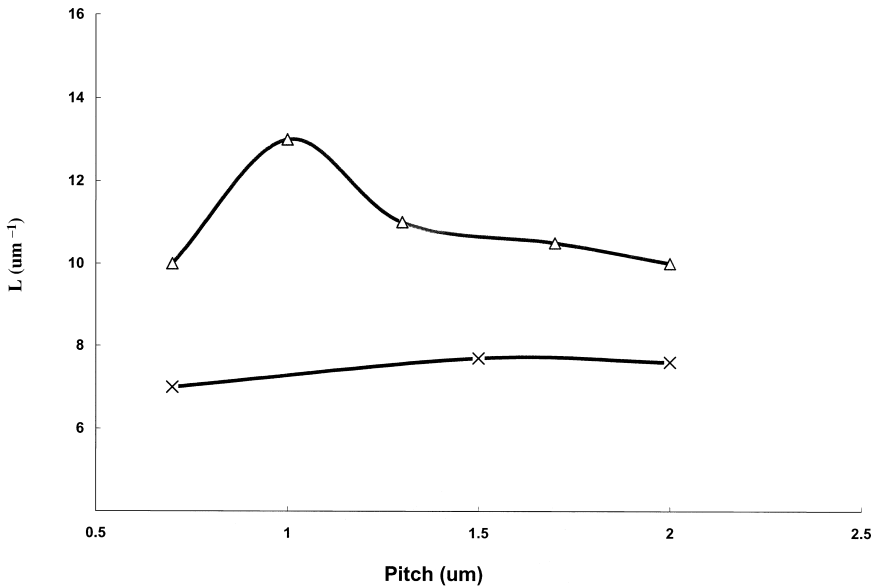


Fig. 8.30. A rim-type PSM for $0.4\text{-}\mu\text{m}$ line (Δ) results in better aerial image (L) sidewall slope than conventional masks (X).

is greatly improved compared with the conventional binary mask. As a result, the DOF can also be significantly improved. Using Eqs. (8.7) and (8.10), one can choose a proper material (n, k) with a proper thickness to meet both transmittance and phase shifting requirements.

The mask manufacturing processes for these types of masks are more complicated than those for binary masks. Figure 8.31 illustrates a typical contact hole PSM mask-making process flow. It starts out with a resist (PR) coated blank, PR/Cr/MOSiON/Qz, that is, a Cr layer on a shifter on a quartz blank. After e-beam writing, the chrome layer is etched through with plasma etching using chlorine and oxygen. Then, the resist is removed, and a second plasma etching is carried out on the shifter using fluorine chemistry, with the chrome as the etching mask. Before shifter etching, the chrome pattern dimensions are measured to check if the target CD is reached. If not, the shifter etching recipe is selected so as to have different

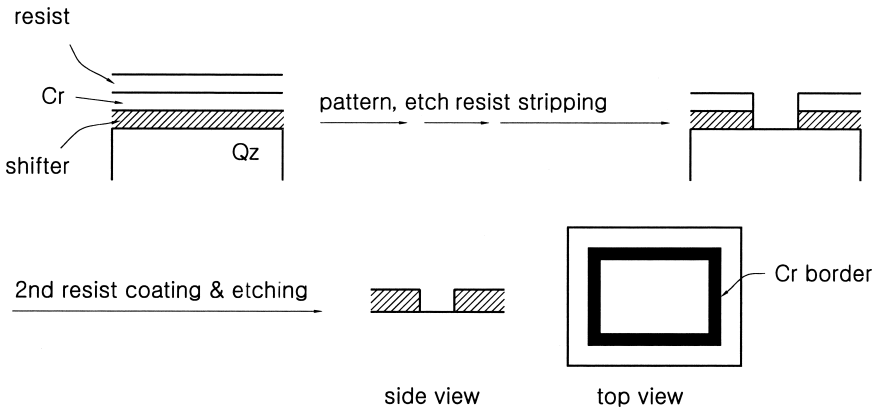


Fig. 8.31. The process flow for making attenuated PSM.

undercuts to achieve the final target CD on the etched shifter pattern. As a further complication, the shifter etching also more or less attacks quartz. The shifter etching recipe must be tuned such that it cannot have too much quartz loss because the refractive index of quartz is not the same as air; in other words, it also causes phase shifting. The defined Cr layer on the shifter then undergoes a second exposure and etching to remove the chrome layer in the pattern area, leaving chrome only in the frame area, where overlay and alignment marks are placed. The process flow for the rim-type PSM for holes is very similar to that of attenuated PSM, except that the second exposure, instead of a blanket exposure, leaves Cr around the shifter edges. The most difficult part of making a PSM is the mask repair, especially for intrusions at a line edge or a hole edge. Because this type of defect repair is done by depositing an opaque carbon film, its transmittance is nearly 0% for DUV light, and it evidently does not render any phase shifting. As a result, if the missing edge is larger than a certain percentage, the repair will not succeed. Normally, for a hole pattern, a defect is considered nonrepairable if two sides need repair and each exceeds 30% of the edge length.

An alternating phase-shifting mask is a strong phase shifting technology in that it shows more pronounced improvements in resolution and photoprocess latitude as compared to attenuated

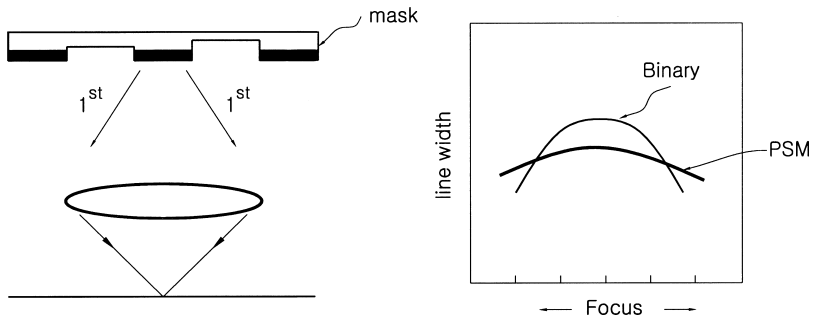


Fig. 8.32. (a) Alternating PSM image formed via two first-order diffracted beams, (b) resulting in great DOF improvement.

PSM. Figure 8.32 shows the neighboring spaces where the light can go through and take opposite phases alternately. This causes the zero-order diffraction to be absent, leading to the interference of two first-order beams for image formation. These two beams are symmetric with respect to the optical axis; hence there is very significant improvement in focus latitude. The reason is that for the conventional mask, three beams are used to reconstruct the image. As the imaging plan moves along the optical axis, the optical path differences (OPD) vary, and the image quality degrades. However, for two-beam imaging with alternating PSM, the OPD between the two symmetrical first-order diffracted beams stays the same. Obviously, the alternating PSM renders much better resolution and focus latitude. Furthermore, it is shown to give rise to smaller mask error factors (MEF) as compared to conventional masks. However, application of alternating PSM technology is relatively limited in comparison to attenuated PSM because its application is limited to periodic patterns.

Alternating PSM mask making is a lot more complicated than that of attenuated PSM, as shown in Fig. 8.33. The phase shift is created by etching through the quartz to a depth having a phase shift. First, the chrome pattern is created, as in the creation of conventional binary masks. Next, the areas that are assigned as zero-degree are exposed, and the pattern is etched into the quartz with plasma etching, using fluorine-based chemistry. After that, the 180° area is exposed and etched. The depth difference of the two areas must be

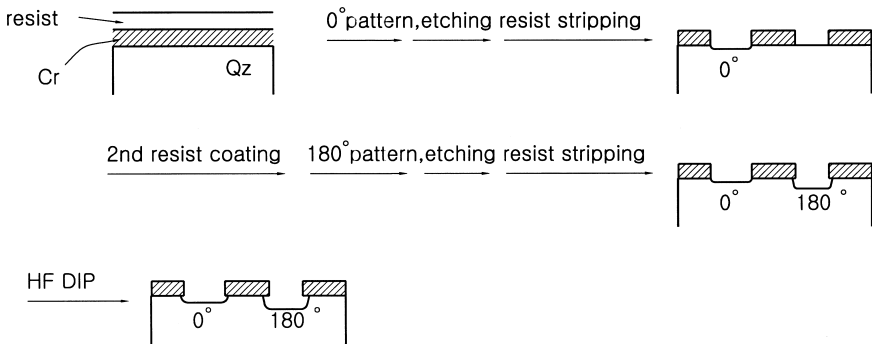


Fig. 8.33. A typical process flow for making an alternating PSM.

accurately controlled to have the desired phase shifting (difference). Next, a light HF wet etching is employed to round off the corners and eliminate some quartz defects. The most difficult part of alternating PSM manufacture is inspection and repair. This is because the quartz defects, such as quartz pits or humps, are still transparent but have different phase shiftings. Unlike in traditional inspection, in this method, special inspection techniques are needed to differentiate the phase angles of the defect area from normal ones. Quartz bumps pose the most challenging issue for repair. The repair technique must remove the quartz bump and yet leave no damage on the quartz that could lead to local transmission loss. Alternating PSM has not been as popular as its attenuated counterpart, mainly because design, software, and mask making are not as straightforward. Despite all these drawbacks, it is used for production of products with very large wafer volumes per mask set. For such circumstances, the long mask-making cycle time is relatively more acceptable as compared to the mask lifetime, and wafer print verification can be used in lieu of a mask inspection tool.

8.4.2. *Off-axis illumination*

Off-axis illumination (OAI) is a relatively cost-effective RET as it does not require mask-making process changes or exposure tool upgrades. The principle is straightforward. A fixed optical system

has a specific NA, capable of resolving a minimum pitch, p . For the case of a pitch smaller than p , the first-order diffracted light will fall out of the lens' collecting range, resulting in no image formation. If one makes the light illuminate the mask surface obliquely, with an angle of ω with respect to the optical axis, the zero-order and one of the first-order diffractions (with different intensities) will now be collected by the lens, as indicated in Fig. 8.34. With these two diffracted rays, an image can be constructed. Thus the resolution is improved. However, one can expect that the exposure intensity will be reduced and that the two beams will have unequal intensity. The asymmetry between the zero-order and first-order diffracted beams can be solved by having a second light source from the opposite direction such that the diffractions of the two lights are added and the asymmetry is canceled out; that is, the zero-order diffraction of one light source coincides with the first-order diffraction of the other. Now we have two symmetrical diffracted rays, rendering improved DOF. Figure 8.35 shows the commonly used apertures for off-axis illuminations of annular, dipole, and quadrupole types. OAI optimization is pitch- and pattern-orientation-dependent. An OAI optimized for one pitch may not be suitable for the other.

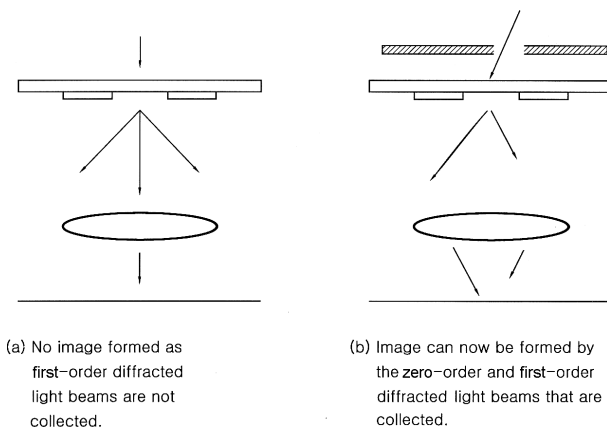


Fig. 8.34. Comparing the image formation with (a) regular illumination and (b) off-axis illumination.

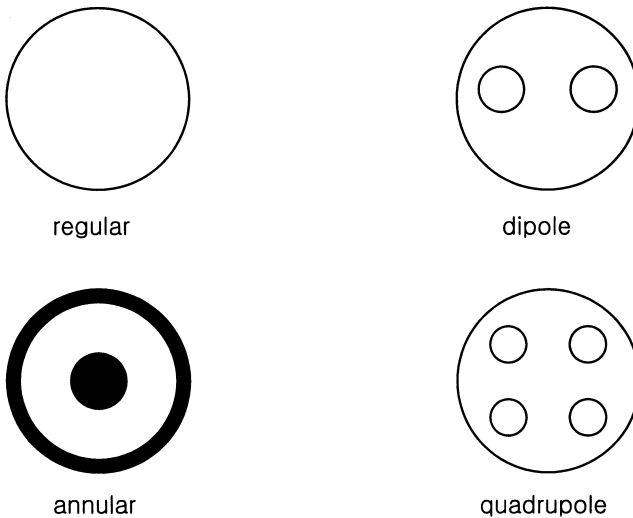


Fig. 8.35. Types of off-axis illumination.

8.4.3. *Optical proximity correction*

One common phenomenon in wafer printing is that the printed wafer dimensions tend to deviate from the designed ones as pattern proximity changes. The other phenomenon is the two-dimensional corners' rounding as the features approach the exposure tools' resolution limits. Examples are shown in Figs. 8.36 and 8.37. The impacts of these proximity effects on circuit performance are tremendous; they sometimes lead to circuit failure. One-dimensional CD variation of the gate layer directly affects the transistor current, which relates to circuit timing. If the CD is below the lower limit, it causes leaky transistors; no signals can be held. CD variations on interconnected layers, such as metal or silicided lines, result in resistance variation or RC delay variations. Smaller CDs on interconnected lines can also result in electromigration or even circuit burn out during operation. Line end shortening, if it happens on a polysilicon gate over an active area, such as that shown in Fig. 8.38, can result in total transistor failure as the source is shorted to the drain area. Optical proximity correction (OPC) is often used to solve the problem. There are two main approaches to implement OPCs: the rule-based

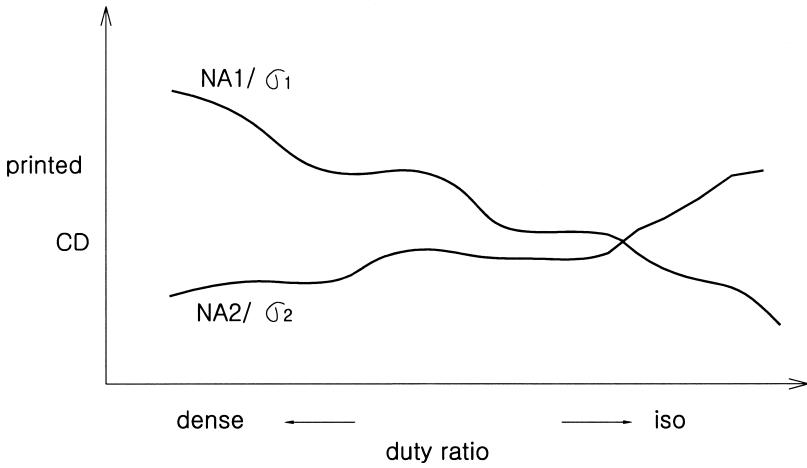


Fig. 8.36. One-dimensional optical proximity effect: printed CD varies with surrounding pattern density.

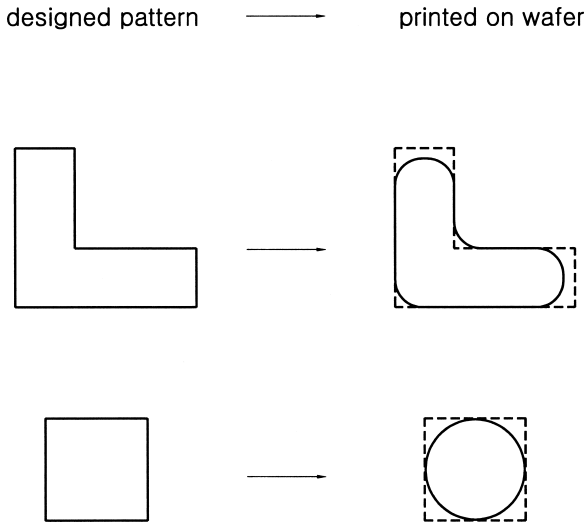


Fig. 8.37. Patterns' corner rounding caused by exposure tools' resolution limit.

approach and the model-based approach. For rule-based OPC, the correction values that are added to the features are basically taken from a lookup table derived from experimental results. Model-based OPC is based on mathematical modeling of the mask patterns. Both

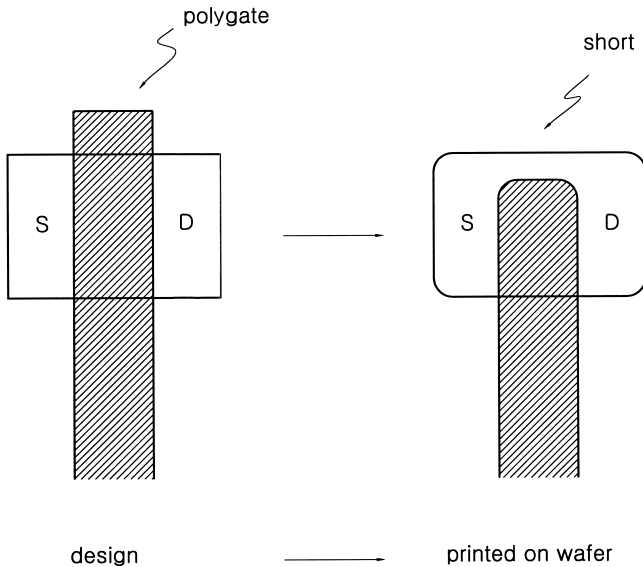


Fig. 8.38. Line end shortening effect causes gate end cap to be severely rounded off and shortened as it is printed on wafers.

methods require iterative approaches to reach the optimized condition where the residual differences between the printed patterns and the designed patterns are as small as desired.

To come up with the rules for rule-based OPC, a large number of one-dimensional and two-dimensional test patterns must be designed to simulate different layout scenarios. The mask pattern is then printed on wafers, measured, and compared with the design values. If the comparison shows inconsistencies, corrections can be added to the mask patterns until matched results (the discrepancy is within tolerance) are found.

On the other hand, instead of using a lookup table, model-based OPC uses correction values that are derived from mathematical modeling of the mask patterns. The model setup procedure starts out with a large number of designed patterns that simulate all the possible patterns and proximities. The patterns are then simulated with the model. The corrected pattern is made into a mask and printed on a wafer. The wafer results are compared with the original design. Depending on how much the residual errors are to be tolerated, the

correction values can always be refined. The finer the corrections are, the smaller the residual errors will be. However, nothing comes free; fine corrections result in increases in mask-making cost and cycle time.

8.4.4. *Subresolution assist features*

Another important category of OPC is the addition of subresolution assist features (SRAF). Line width biasing, which is often used in rule-based OPC, does ensure, to some extent, the line width uniformity across the whole design pattern. However, the overlapping process window (latitude of energy defocus window) of the different lines in different proximities can be unacceptably small. The addition of SRAF into the large empty space (around isolated lines), as illustrated in Fig. 8.39, improves the photolithography process window. For a given space, the number of assist features needed, and the distance between the main pattern and the SRAF, must be optimized in terms of wafer print results.

Because the photolithography technology evolution lags behind product design needs, RETs are extensively applied until the next generation of photolithography technology becomes available. Figure 8.40 shows resolution enhancement techniques used for various technology generations. Above $0.35\text{-}\mu\text{m}$ technology, hardly any RET techniques are used. After $0.25\text{ }\mu\text{m}$, some manual OPCs and

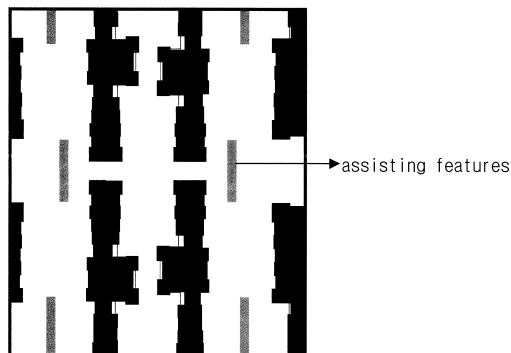


Fig. 8.39. The assisting features for improving photoprocess.

	0.35 μm	0.25 μm	0.18 μm	0.13 μm	0.09 μm
PSM	X	X	O	O	O
Serif	X	O	O	O	O
Rule base	X	O	O	X	X
SRAF	X	X	O	O	O
Model base	X	X	X	O	O

Fig. 8.40. Types of RET used in various technologies. O, used; X, not used.

hole-attenuated PSMs are used, but in general, the RETs are relatively simple. Below $0.18 \mu\text{m}$, the industry starts to see the use of the SRAF and rule-based OPC applications. After $0.18 \mu\text{m}$, extensive use of RETs becomes a must. Oftentimes, more than one RET is needed for each technology.

Figure 8.41 illustrates the percentage of masking layers with RETs in a set of masks; one can see that the percentages go

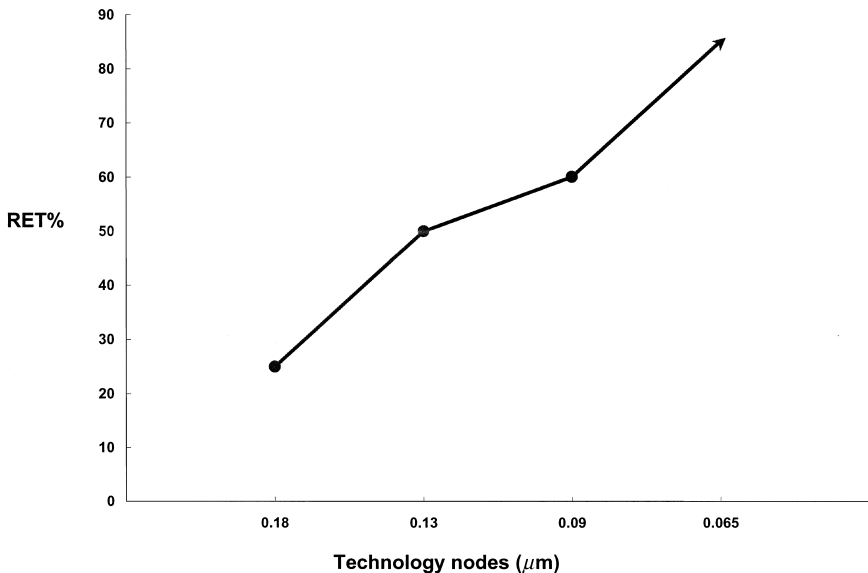


Fig. 8.41. In a set of masks, the percentage of mask layers with RET increases dramatically as technology shrinks.

straight up. This brings up the mask cost and lengthens mask manufacturing cycle time as well. The use of RETs significantly complicates mask manufacturing. Apart from the PSM making issues discussed earlier, OPC also poses some very challenging issues in mask making. The fine OPC jigs and jogs require the minimum writing grid of the electron beam to be used, leading to long writing time. Up to 1.5 times longer writing time is needed for some cases in 0.13- μm technology, as compared to that without OPC. The fine jigs and jogs also require a special mask inspection algorithm and are very difficult to repair. Furthermore, the metrology of such a pattern cannot be measured with an optical metrology tool due to its inappropriate resolution and proximity limitation. It must be measured with a CD SEM tool with a restriction that the CD measurement location must be on a long, smooth line — having no jigs and jogs within a few microns.