Ch intro

2.1 SILICON MANUFACTURE

In this chapter, we will consider only silicon-based (Si) technologies. Although other

compound materials in groups III through V, such as gallium arsenide (GaAs) and aluminum

gallium nitride (AlGaN), are also used to produce VLSI chips, silicon is still the most

popular material, with excellent cost–performance trade-off. The mineral quartz consists entirely of silicon dioxide, also known as silica. Ordinary sand is chiefly composed of tiny grains of quartz and is therefor also silica.

Despite the abundance of its components, elemental silicon does not exist in nature. The element can be synthesized artificially by heating silica and carbon in an electric boiler. The carbon reacts with the oxygen in the silica, resulting in nearly pure molten silicon. As this cools, a slew of minute crystals forms and coalesces to form a fine-grained grey solid. Because it contains a large number of crystals, this type of silicon is called poly-crystalline. This metallurgical-grade polysilicon is unsuitable for semiconductor manufacturing due to impurities and a disordered crystal structure.

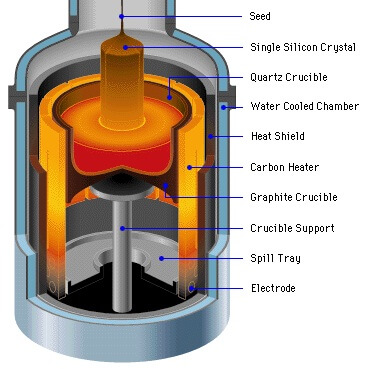
Metallurgical-grade silicon can be further refined to produce an extremely pure semiconductor-grade material. Purification begins with the conversion of the crude silicon into a violent compound, usually trichlorosilane. After repeated distillation, the extremely pure trichlorosilane is reduced to elemental silicon using hydrogen gas. The final product is exceptionally pure, but still poly-crystalline. Practical integrated circuits can only be fabricated from single-crystal material, so that the next step consists of growing a suitable crystal.

2.1.1 Crystal Growth

The principles of crystal growth are straightforward and well-known. Assume a few sugar crystals are added to a saturated solution that then evaporates. Sugar crystals act as seeds for the deposition of more sugar molecules. The crystals eventually grow to be quite large. Even in the absence of a seed, crystal growth would occur, but the result would be a swarm of small intergrown crystals. By suppressing unwanted nucleation sites, the use of a seed allows the growth of larger, more perfect crystals.

In principle, silicon crystals can be grown in much the same manner as sugar crystals. In practice, no suitable solvent exists for silicon, and the crystals must be grown from the molten element at temperatures in excess of 1400°C. the resulting crystals are at least a meter in length and ten centimeters in diameter, and they must have a nearly perfect crystal structure to be useful to the semiconductor industry. These requirements make the process technically challenging.

The usual method for growing semiconductor-grade silicon crystal is called *Czochralski process*. This process, illustrated in figure 2.1, use a silica crucible charged with pieces of semi-grade polycrystalline silicon. An elective furnace raises the temperature of the crucible until all the silicon melts. The temperature is then reduced slightly and a small seed crystal is lowered into the crucible. Controlled cooling of the melt causes layers of silicon atoms to deposit upon the seed crystal. The rod holding the seed slowly rises so that only the lower portion of the growing crystal remains in contact with the molten silicon. In this manner, a large silicon crystal can be pulled centimeter-by-centimeter from the melt. The shaft holding the crystal rotates slowly to ensure uniform growth. The high surface tension of molten silicon distorts the crystal into a cylindrical rod rather than expected faceted prism.



**Figure 2.1** czochralski process for growing silicon crystal.

The czochralski process requires careful control to provide crystals of the desired purity and dimensions. Automated systems regulate the temperature of the melt and the rate of crystal growth. A small amount of doped polysilicon added to the melt sets the doping concentration in the crystals. In addition to the deliberately introduced impurities, oxygen from the silica crucible and carbon from the heating elements dissolve in the molten silicon and become incorporated into the growing crystal. These impurities subtly influence the electrical properties of the resulting silicon. Once the crystal has reached its final dimensions, it is lifted from the melt and is allowed to slowly cool to room temperature. The resulting cylinder of monocrystalline silicon is called an *ingot*.

Since integrated circuits are formed upon the surface of a silicon crystal and penetrate this surface to no great depth, the ingot is customarily sliced into numerous thin circular sections called wafers. That are 400μm to 600μm thick. Each wafer yields hundreds or even thousands of integrated circuits. The larger the wafer, the more integrated circuits it holds and the grater the resulting economies of scale.

2.2 Wafer Manufacturing

A series of mechanical processes are used to create wafers. The ingot's two tapered ends are sliced off and discarded. The remaining material is ground into a cylinder, the diameter of which determines the size of the resulting wafers. After grinding, no visible indication of crystal orientation remains. The crystal orientation is determined experimentally, and a flat stripe is ground down one side of the ingot. Each wafer cut from it will have a facet, or flat, that clearly identifies the crystal orientation.

After grinding the flat, the manufacturer uses a diamond-tipped saw to cut the ingot into individual wafers. Approximately one-third of the precious silicon crystal is reduced to worthless dust in the process. The sawing process leaves scratches and pockmarks on the surfaces of the resulting wafers. Because the tiny dimensions of integrated circuits necessitate extremely smooth surfaces, each wafer must have one side polished. Mechanical abrasives are used first, followed by chemical milling. The resulting mirror-bright surface has a dark grey color and the characteristic silicon near-metallic luster.

The finished wafers' thickness and diameter must correspond to the mechanical strength and other physical qualities of the material the wafers will be used to make. The finished product must be sturdy enough to support its weight without shattering when handling the products of the wafers' diverse applications. As more material is added in the fabrication of the slice, the diameter of the wafer increases, as does the weight of the wafer. When enough weight has been added, the diameter cannot be increased because it compromises the slice's strength.

If the weight is not properly placed, a small amount of pressure will be enough to break down the wafer. Before they can be used in manufacturing, sliced wafers must be processed. Machines and abrasive chemicals are used to smooth out the rough surface of the wafer. The flawless surface facilitates printing circuit layouts on the wafer surface.

The names and functions of each component of a completed wafer:

1. **Edge Die(chips)**: It is considered as the production loss. The chips along the edge of a wafer. Larger wafers have less chip loss.

2. **Scribe Lines**: Between the functional portions, there are narrow, non-functional areas where a saw can securely cut the wafer without destroying the circuits. These thin areas are the scribe lines

3. **Chip**: a little piece of silicon that has electronic circuit patterns

4. **Flat Zone**: edge of a wafer that is yanked off flat to aid in wafer orientation and type identification.

5. **Test Element Group (TEG)**: a prototype pattern that displays the actual physical features of a chip (diodes, circuits, capacitors, transistors, and resistors) so that it may be tested to know if it works adequately.

All silicon wafers are extremely useful components, just that there are different variations used for different purposes. It is therefore imperative to know about the different types of Silicon Wafers. There are mainly two varieties of silicon wafers commonly used today. These are undoped silicon wafer and Doped silicon wafer. Undoped Silicon Wafers, also known as Intrinsic or Float Zone (FZ) do not have any dopants in them. They are made of strictly pure crystalline silicon. This type of silicon wafer is recognized as the ideal semiconductor. Doped silicon wafers are formed by introducing dopants (certain impurities) into the silicon crystal during the formation process. When boron is added into the mixture, a P-type doped silicon wafer is produced. P-type silicon wafers have numerous positively-charged holes. To produce an N-type doped silicon wafer, elements like phosphorus, arsenic, or antimony will be added. N-type silicon wafers have a negatively-charged electron in them. The quantity of dopant discovered in the wafer will ascertain if it is degenerate or extrinsic. To be degenerate means there's a higher concentration of dopants in it, while to be extrinsic means it has little or moderate dopants.

2.2 PHOTOLITHOGRAPHY

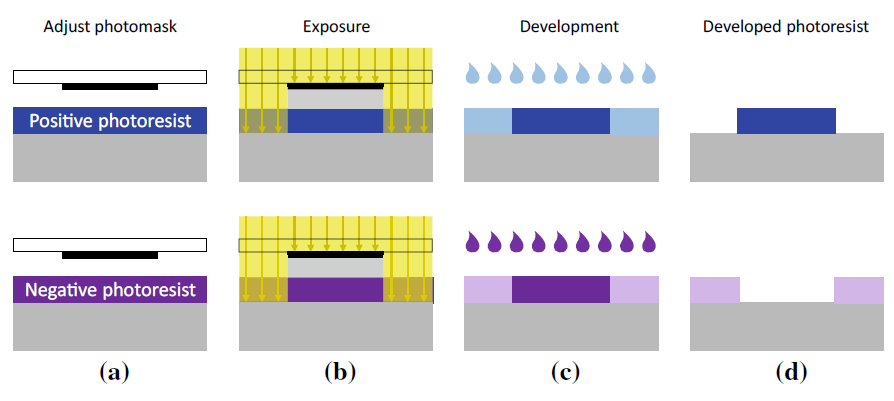
photolithography is used in all structuring process steps. Its purpose is to transfer a two-dimensional image of the required structures onto the wafer surface, so that subsequent processing (e.g.implanation, etching) can be applied to a restricted area. The wafer is first coated with a thin radiation sensitive film, called a *photoresist* or *resist*. A photomask is then exposed to light to project a black and white image of the desired structure on the photoresist. The solubility of the photoresist changes with reference to. a specific fluid, called a *developer*, at the regions exposed to light. In the *developing*

*process,* the soluble areas of the photoresist are removed from the wafer with this fluid. The insoluble parts of the photoresist remain.

2.2.1 Photoresist

Photoresist is deposited on the wafer by means of spin coating. Liquid coating material is applied to the center of the wafer, which is rapidly rotated. The centrifugal force causes the fluid to be distributed on the wafer. The solvent for setting the viscosity in the fluid is vaporized to produce a layer of constant thickness. Spraying the resist onto the wafer is another method that is sometimes used. Photoresists are radiation-reactive polymers. Positive and negative photoresists are available. A positive photoresist is a type of photoresist in which the portion of the photoresist that is exposed to light becomes soluble to the photoresist developer. The unexposed portion of the photoresist remains insoluble to the photoresist developer (see Fig. 2.2, top row). The effect is the opposite with negative photoresists where the portion of the photoresist that is exposed to light becomes insoluble to the photoresist developer. The unexposed portion of the photoresist is dissolved by the photoresist developer (see Fig. 2.2, bottom row). In each case, the remaining

photoresist serves as a mask for the next process step.



**Figure 2.2** Schematic representation of the photolithography with positive resist (above), where the

exposed regions are removed, and negative resist (bottom), where the exposed regions are kept.

2.2.2 Photomasks and Exposure

A *photomask* is used to expose a wafer. A photomask is a sheet of glass, on which

a black-and-white image of the structures to be processed is applied to an opaque

layer made of chromium. When this photomask is exposed, a shadow is cast on the

exposed wafer to produce the desired image. There are two types of exposure: *direct*

*exposure* and *projection exposure*.

**Direct Exposure**

With direct exposure, the photomask is placed above and in close proximity to the

wafer, either in direct contact with the photoresist (*contact exposure*) or close to it

(*Proximity exposure*). The result is in both cases a 1:1 (full scale) image produced by simple shadowing. Hence, the structure sizes on the photomask must correspond with the structures on the wafer (so-called “1X photomasks”).

Contact exposure is unsuitable for volume manufacture as it can cause damage to and soiling on the resist and the photomask. Proximity exposure differs in that there

is no contact between photomask and resist. However, the distance between mask and resist in proximity exposure, which cannot be reduced indefinitely but must lie between 10 and 40 μm, can cause resolution degrading.

If the dimensions of the structures being imaged are similar in size to the exposure

wavelength, significant diffraction phenomena may occur. Direct exposure therefore

approaches its limit for structure sizes to be imaged at this order of magnitude. The

limit is even bigger (approximately 3 μ m) for proximity exposure. Hence, the structure

sizes required in state-of-the-art semiconductor technologies that are smaller

than this value cannot be imaged with direct exposure; this process is therefore not

deployed any more today.

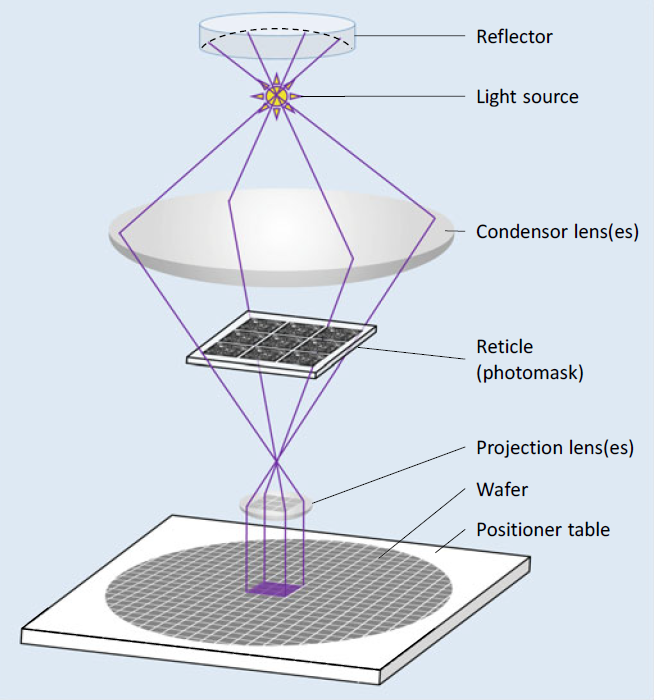
**Projection Exposure**

*Projection exposure* was developed to overcome the above-mentioned direct exposure

issues. Here, the pattern on the photomask is projected through lenses onto a wafer coated with photoresist, as illustrated in Fig. 2.3. This system brings with it two fundamental benefits: (i) photomask and wafer are physically separated and (ii) the image can be optically reduced in size during the projection thus improving imaging accuracy. Among the most commonly used photomasks are 4X, 5X and 10X masks, in other words, photomasks with imaging ratios of 4, 5 and 10 to 1. A small part of a wafer can only be exposed in one step with this method, as the sizes of the photomasks and lenses are limited.

Photomasks therefore have structures for one or a few chips. These types of photomasks are called *reticles*. A Reticle is a special type of photomask where the data for only part of the final exposed area is present. A Reticle is loaded into a Stepper or Scanner system where multiple exposures are made to cover the full patterned area. However, the term “reticle” has become a synonym for “photomask” as state-of-the-art photomasks are seldom designed for an entire wafer.

A wafer is exposed in many single steps in the so-called “step-and-repeat technique”. The wafer is transported on a positioner table under the projection optics. This type of exposure equipment is called a *wafer stepper* or simply a *stepper*. Figure 2.3 shows the stepper principle of operation.



**Figure 2.3** Exposing a wafer with step-and-repeat technology. Using one reticle, multiple exposures

are made to cover the full patterned area

innovations have pushed the boundaries of optical lithography so that feature sizes of 193 nm (argon-fluoride laser). When we talk about chips with such small feature of approximately 10 nm can be patterned with the widely used exposure wavelength sizes, we are entering the field of *nanoelectronics*.

technology is being continuously developed to improve imaging accuracy, and such

Ultraviolet light is used for higher optical resolution. Exposure and photomask

In addition to these developments, there are other approaches for downscaling structure sizes. The optical resolution can be further improved by using ultraviolet light of shorter wavelengths. Mirror systems are needed for projections using light of shorter wavelengths as the materials available for lenses are increasingly opaque.

In addition to this approach, systems for directly exposing wafers with electron beams are used as well. However, this is a very time-consuming process as all structures are individually “written” on the wafer. It is nonetheless an economical technique for small lot sizes, as there is no need for photomasks, which are expensive to produce.

2.2.3 Alignment and Alignment Marks

Progressive structural processes that interact with one another are used to construct integrated devices.

This means that a reticle's (i.e., a photomask's) position for a specific layer must always be precisely aligned with the structures that are already present on the wafer. For instance, when the devices have been installed, the contact holes need to be precisely positioned where the connecting areas are.

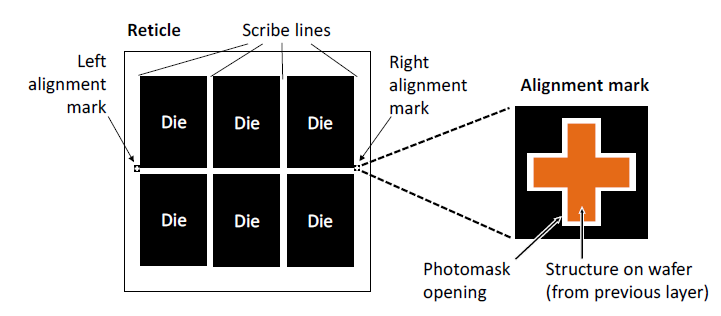
Before each exposure in the step-and-repeat procedure, the wafer's position and attitude with regard to the exposure reticle must be modified. To do this, alignment marks on the reticle are automatically identified by optical methods (Fig. 2.4). These marks, which include geometric shapes like crosses, build structures on the wafer in accordance with the actual process step, just like all other geometric elements on the reticle. If alignment marks' effects on these structures can be seen visually, the wafer position can be aligned with the help of alignment marks on subsequent photomask(s).

An opening in the photomask in the shape of a cross serves as the alignment mark in Fig. 2.4 to the right. The wafer underneath is modified so that the cross-shaped structure, which is also present, is situated in the center of the wafer.

However, some structural processes, like ion implantation doping techniques, for instance, leave no reliable optical traces. But, this is not a problem. As mechanical tolerances result in variations in every adjustment, adjustments are made in as many process stages as possible using the same alignment mark on the wafer to prevent the deviations from building up. When a structure can no longer be detected, an alignment mark from a later process stage is used. The "newer" alignment mark is then used in the subsequent phases.

During the alignment procedure, the correct angular orientation of the wafer must be checked in addition to the correct location of the alignment marks. As a result, the alignment marks are placed at two distant points on the reticle to ensure accuracy. the wafer attitude is correct.

The alignment marks are placed outside the chips because they are not functional chip structures. The chips have a 50-100μ m clearance from one another to facilitate (subsequent) chip dicing. This clearance is known as a "sawing trench," "saw street," or "saw street clearance." "scribe line". In this clearance space, the alignment marks are placed (Fig. 2.4, left).



**Figure 2.4** Alignment marks on a reticle (photomask) with six chip structures**.**

2.2.4 Reference to Physical Design

The extremely high effort required to create the photomasks derived from the layout design to perform exposure is the price for the degree of miniaturization currently achieved with structure sizes of the order of nanometers. Modern photomasks Semiconductor processes can be quite costly.

It's important to remember that a single design flaw can render photomasks and wafers totally worthless. In the event of a fault, this financial hit is accompanied by a significant delay in development time. Troubleshooting and a new production run can take up to six months. When one considers the delays in bringing the product to market, the financial losses become even more severe.

The critical conclusion here is that an IC chip's physical design must be absolutely flawless! Every effort must be made to detect design risks, and appropriate and effective preventative measures must be implemented. In this regard, automated verification algorithms, which we will discuss in Chapters 3 and 5, play an important role in physical design.

2.3 Imaging Errors

The structures on the photomasks are determined by layout data. So far, we've discussed how photolithographic images and subsequent targeted process steps are used to convert these mask structures to wafer structures. When the structures created "on" and "in" the wafer are compared to the original layout structures, various types of deviations become apparent. We then go over these unavoidable deviations and how we can deal with them in the layout design.

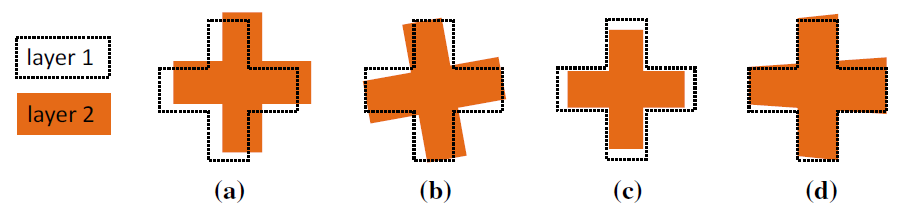
There are three categories of imaging errors: (1) *overlay errors,* (2) *diffraction effects* and (3) *edge shifts*. Overlay errors and diffraction effects occur during exposure,while edge shifts occur in subsequent structuring process steps.

2.3.1 Overlay Errors

The wafer and photomask (reticle) cannot be positioned with absolute accuracy relative to the exposure device due to mechanical tolerances and measurement inaccuracies that can occur during alignment. As a result, structures on the photomask are not precisely mapped to the wafer during exposure, as specified by the layout template.

Exaggerated representations of possible exposure faults are shown in Fig. 2.5, where we show (a) displacements and (b) rotations that can occur relative to the required positions. To adjust the depth of focus, the wafer and photomask are moved along the optical axis. The layers will be scaled relative to one another if the distances between the masks, lenses, and wafer are changed during this focusing step (c). The perspective will be distorted if the photomask is tilted with respect to the optical axis (d).

Overlay errors are also caused by the fact that wafers and photomasks, like all materials, expand when heated. If wafer exposures take place at different temperatures, spacings between structures on the wafer will change. As a result, displacements (see Fig. 2.5a) occur, the magnitude of which is determined by where they occur on the exposed wafer. To reduce this effect, keep the exposure temperatures as constant as possible throughout the manufacturing run.



**Figure 2.5** Possible overlay errors between two layers: **a** displacement, **b** rotation, **c** scaling,

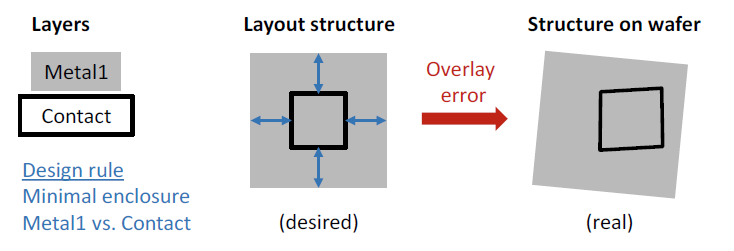
**d** perspective distortion.

This cause of overlay errors cannot be completely eliminated because it is not always possible to keep these temperatures constant. In high temperature steps of the process, the wafer can become irreversibly deformed. The severity of these deformations varies from wafer to wafer. This effect may result in displacements and scaling (see Fig. 2.5a, c).

The nature and extent of these overlay defects are unpredictable. Their effects are cumulative and can only be limited to specific areas by efforts that concentrate on specific devices and the manufacturing process. Obviously, the maximum allowable overlay error in a semiconductor process should not be greater than the minimum feature size. It should usually be much lower than this value.

Let us take a look at a typical overlay fault to see how it can affect a layout. To ensure proper electrical connection, contacts in chip fabrication must always be completely covered with metal. Because contacts and metallic interconnect layouts use different photomasks, this "design rule" for full coverage must also be met when overlay errors occur.

How overlay errors are handled in layout design is shown in Fig. 2.6. For creating the layout, an enclosure design rule is specified that requires all structures in the "Contact" layer to be covered by structures in the "Metal1" layer and to overlap on all sides by a minimum value. This is the same as the maximum allowable overlay error, which is the deviation that can occur in the worst-case scenario when all overlay errors are superimposed. This design rule is met by the layout structure in the centre of Fig. 2.6. The right-hand side of the figure depicts a possible fabrication situation. For clarity, the assumed overlay error is exaggerated here.

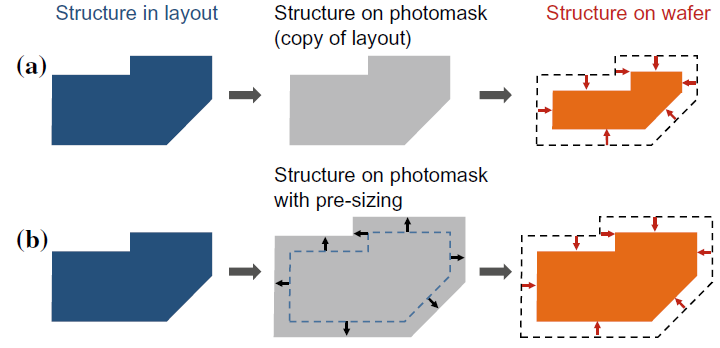


**Figure 2.6** Handling of unavoidable overlay errors in physical design. Under all circumstances, contact holes must be completely covered by metal; this necessitates a design rule for a sufficiently large "minimal enclosure" between metal and contact layer that accounts for possible shifting, rotation, scaling, and perspective distortion.

2.3.2 Edge Shifts

Some technology layers experience effects that cause the graphics elements on the processed wafer to be enlarged or shrunk in comparison to the associated graphics elements in the layout. (Please keep in mind that we are not discussing scaling, in which the dimensions of the elements are changed by a certain factor, as with a "Zoom.") The changes caused by these enlarging/shrinking effects are additive: the structure's boundary lines are shifted outwards by a specific value (positive shift) or inwards by a specific value (negative shift) (negative shift). We call this category of imaging errors edge shifts because the individual structures in the layout are typically modelled as polygons (i.e., as geometrical elements bounded by edge strips).

Figure 2.7 depicts a simple example of a process step with a negative edge shift. The structure on the wafer contracts in relation to the element on the photomask (shown with red arrows in the figure on the right). These edge shifts have layer-specific sizes that are defined for each semiconductor process. As a result, the effect can be mitigated by pre-sizing the photomask geometry. In particular, the prepared layout data are automatically modified in a layout to mask preparation process that is part of the layout post process. If a value k edge shift occurs during the process, the edges of the layout geometries are shifted by a value k before the new data is transferred to the photomask. This operation is shown in the bottom row (b) in Fig. 2.7.



**Figure 2.7** Edge shift inwards in the wafer process; **a** without pre-sizing, **b** with pre-sizing

2.3.3 Diffraction Effects

Diffraction phenomena occur at the structural edges of the chromium layer on the photomasks due to the wave properties of light, limiting the optical resolution of photolithography. The smaller the feature size (for a constant exposure wavelength), the greater the impact of diffraction effects on imaging accuracy. In Fig. 2.8, we show these effects with an L-shaped layout structure.

The layout element is transferred to the photomask (in grey) unchanged in the top row. As the feature size decreases, the shape of the area exposed in the photoresist (blue) deviates more from the shape of the photomask opening and thus from the desired layout structure (moving from left-to-right in the figure).

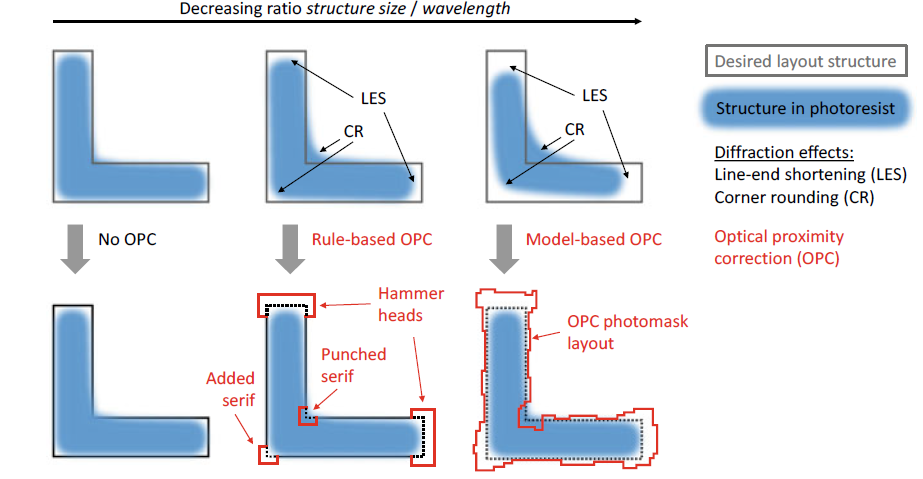
The small fillets at the corners are insignificant as long as the wavelength of light is greater than the feature size (see Fig. 2.8, top left). However, when the feature size to wavelength ratio falls below one (so-called sub-wavelength lithography), significant line-end shortening occurs. Corner rounding has also increased significantly (see Fig. 2.8, top center and right).

These diffraction effects can be corrected by slightly enlarging the photomask opening where it is underexposed and slightly shrinking it where it is overexposed. This procedure, known as optical proximity correction, is another example of a preventive measure (OPC).

As long as the diffraction effects are of the same order of magnitude as those shown in the middle of the figure, these corrective measures can be defined using simple rules based on structural shapes. If the line width is roughly the same as the feature size, "hammer heads" are attached to the ends of "thin" lines. Square elements, so-called “serifs”, are added (this means more light) to the outer corners, or “punched” (means less light, hence also called “jogs”) from the inner corners. These

measures, known as *rule*-*based OPC*, are illustrated in Fig. 2.8, bottom center. In addition, the change in line widths, caused by interference from many parallel lines, can be corrected by rule-based OPC (not shown in Fig. 2.8).

The extent of imaging errors increases as feature sizes shrink in relation to light wavelength (see Fig. 2.8, top right). In this case, rule-based OPC is ineffective because the exposure result is increasingly influenced by surrounding features. Corrections to the photomasks must then be calculated individually for each structure. The algorithms used in this case are based on models that describe wave-optical effects. A model-based OPC result is shown in Fig. 2.8, bottom right.



**Figure 2.8** Diffraction effects in photolithography (top row) and possible corrective measures using optical proximity correction (OPC, bottom row). The imaging errors increase as the ratio of feature size to optical wavelength decreases (from left to right).

2.3.4 Reference to Physical Design

Imaging errors are classified into two types: deterministic and stochastic. The former can be predicted ahead of time, but the latter cannot. As a result, they are handled differently and have a different impact on the layout design process.

**Deterministic Imaging Errors**

Edge shifts and diffraction effects are examples of deterministic imaging errors because their type and scope are known ahead of time. As a result, in these cases, as described above, we can take preventive corrective action. When the layout design is finished but before the photomasks are produced, the layout data are graphically altered in an automated layout post processing step. The modifications in question are: I edge shifts to compensate for technology edge shifts, and (ii) OPC measures, both performed during the layout to mask preparation process.

The goal of these corrective measures is to update the graphics elements in the layout so that the resulting structures appear correctly on the wafer. This strategy eliminates needless complications in layout design. For starters, it saves labor because the layout designer no longer has to perform these adjustment tasks. For another, the layout is easier to "read," which saves time and money while also improving the quality of the layout results. A word of caution, however. Pre-emptive operations do not handle and neutralize all edge shifts that occur in the semiconductor process. Some layout features will differ from those created on the wafer.

Because of the highly complex nature of the layout patterns, calculations for model-based OPC are very CPU intensive. The required computational overhead can be so large that it is best to include the required computer time in the project schedule. Fortunately, the calculations are independent of one another and can be performed in parallel, which speeds up the process.

As previously stated, the smallest technologically achievable feature size is a result of the photolithography radiation's wave-optical properties.

Even if many technological measures can reduce this boundary further, there will always be an accuracy boundary for a given process. This technological constraint is implemented in the layout design by defining design rules that specify specific minimum dimensions for features within a layer. These minimum dimensions apply to I the width of the geometrical elements to ensure that they can be exposed (i.e., the features on the wafer do not "disappear") and (ii) the spacing between two neighboring geometrical elements to ensure that they can be safely separated (i.e., the features on the wafer do not "merge"). These design rules are known as *minimum width* and

*minimum spacing* rules respectively.

**Stochastic Imaging Errors**

Overlay errors are stochastic defects, which means they cannot be predicted accurately. One can only ensure, as described above, that the sum of all deviations does not exceed a specified limit by adhering to device and process tolerances. This maximum permissible overlay error is used to derive additional key design rules that should be considered in physical design.

Rules for overlay errors describe specific minimum dimensions that refer to

features on *different layers*. The rules prescribe minimum values

(1) for two overlapping geometrical elements (to ensure the features also overlap

on the wafer),

(2) for the case where one geometrical element is enclosed by another (so that one

structure covers another on the wafer, as shown in Fig. 2.6), or

(3) for spacing between two geometrical elements (to ensure there is a clearance

between features on the wafer, or at least no contact between them).

These respective design rules are known as (1) *extension* and *intrusion* rules, (2)

*enclosure* rules, and (3) *spacing* rules.

2.4 Applying and Structuring Oxide Layers

One of silicon's major advantages over other semiconductors is that it forms a very stable intrinsic oxide: silicon dioxide (SiO2). Silicon dioxide, which we will refer to as "oxide" in the following for simplicity, has many beneficial properties.

Oxide is a good electrical insulator and a dielectric in capacitive applications. It is mechanically stable and thus appropriate for a strong layer structure. It is simple to create from a process standpoint, and it serves as a masking layer for many process steps, as we will see. It is also visible. Because alignment features can be detected beneath the oxide, this is a useful factor in fabrication. This also enables a wide range of applications, including LEDs (silicon can emit light), solar cells, and photodiodes (light can penetrate silicon from outside). Different processes are available for producing and structuring oxide layers. We will examine these now.

2.4.1 Thermal Oxidation

Thermal oxidation is used to create oxide on the wafer surface. Once formed, an oxide layer can only grow if oxygen atoms diffuse through it until they reach the silicon beneath. As a result, as the thickness of the oxide layer increases, the rate of oxide growth slows. According to analysis, oxide grows from the original silicon surface by approximately 44% into the silicon (the same amount of silicon used) and by approximately 56% outwards (Fig. 2.9). There are two different thermal oxidation processes: *dry oxidation* and *wet oxidation*.

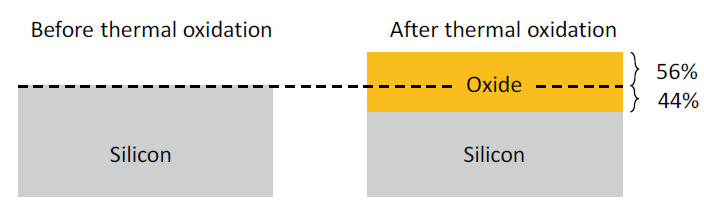
**Dry oxidation**

Wafers are heated in an oxidation furnace and exposed to pure oxygen (O2) at temperatures ranging from 1000 to 1200 °C (approximately 2000 °F). According to the formula Si + O2 → SiO2, the oxide grows slowly and produces good quality oxide with few vacancy defects. This method is used to create ultra-thin *gate oxides* (*GOX*) in field-effect transistors and dielectrics in capacitors.

**Wet oxidation**

In this process, the oxygen first flows through boiling water. The wafer is thus exposed to steam, as well. The reaction takes place as per the formula Si + 2H2O→SiO2 +2H2 at 950–1000 °C (approx. 1800 °F) and is much faster than dry oxidation. It is more difficult to control however and produces a lower quality oxide. Wet oxidation is therefore more widely used to produce the *field oxide* (*FOX*). This is the first thick oxide layer that is created directly on the silicon. It is used to laterally insulate

regions from each other. In old processes, the field oxide was only produced where there were no devices in the silicon. These “non-active” regions were also called “field” regions, and is where the name derives from.



**Figure 2.9** Thermal oxidation of silicon (Si) to silicon dioxide (SiO2)

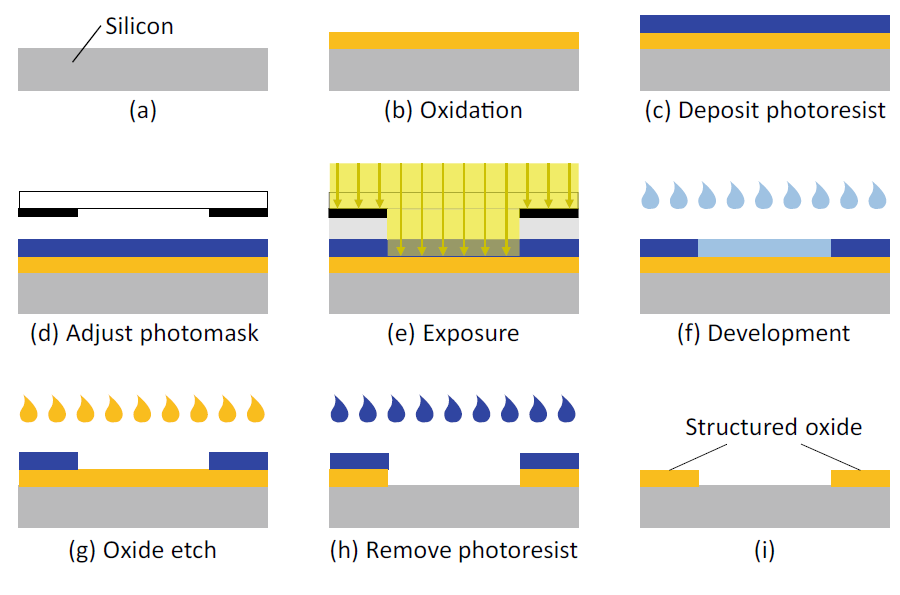
2.4.2 Oxidation by Deposition

The silicon is obtained from the wafer surface and "consumed" in the previously discussed thermal oxidation. If the silicon surface is covered by other layers, however, the oxide for additional oxide layers must be deposited. Outside silicon, as well as oxygen, must be added. There are numerous deposition methods available, the discussion of which is beyond the scope of this book. This oxidation process is used to electrically isolate metallization layers from one another.

2.4.3 Oxide Structuring by Etching

Etching is a chemical removal process that is frequently repeated for different materials in chip fabrication. It is critical that etchants that remove material be used selectively with respect to the substance, i.e., not etching away other substances (or at least only very slightly).

Figure 2.10 shows how etching can be used to structure an existing oxide layer. A photomask is used to expose and develop an etch-resistant photoresist. There are two types of etching that can be used: wet etching and dry etching.



**Figure 2.10** Structuring an oxide layer

**Wet Etching**

A fluid chemical etching agent dissolves and removes oxide during wet etching. This is a straightforward and widely used method, and the fast etching rate can be adjusted. The disadvantage of wet etching is that it is isotropic, meaning that it acts in all directions. This causes undesirable lateral etching beneath the photoresist. Because of these so-called undercuts, the oxide openings are always larger than the photoresist openings. This results in an edge shift (Fig. 2.11, left). The undercut etching rate is slightly lower than the vertical etching rate because the etching agent cannot circulate as easily under the photoresist and is thus more highly saturated. The lateral undercut is typically 80% of the etching depth.

Due to this undercutting effect, wet etching is no longer suitable for imaging typical feature sizes in advanced processes. As a result, in these processes, wet etching is only used to dissolve and remove entire layers.

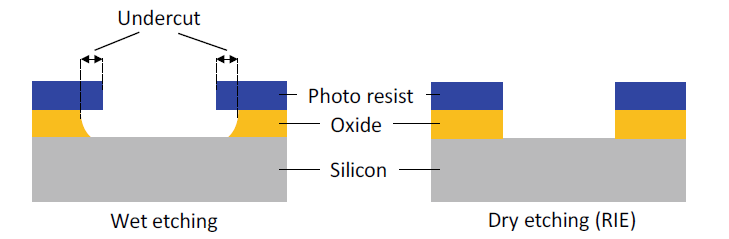
**Dry Etching**

Reactive ion etching (RIE) is a popular dry-etching method. In general, the etching agent is ionised and used as a gas plasma. An electrically alternating field causes the ions to oscillate. The field is parallel to the wafer surface. Chemically active ions oscillate in this direction and etch material only vertically. The main advantage of RIE is that there is no edge shift in this process (Fig. 2.11, right).

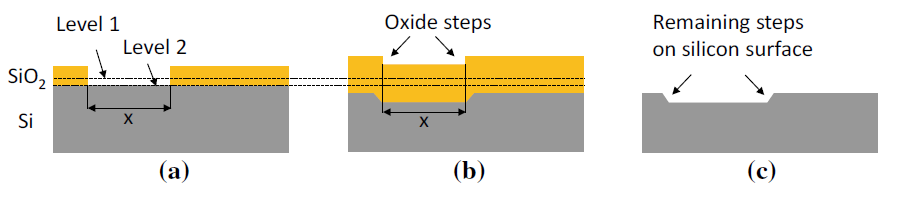
In this process, the etching effect is a combination of a physical effect (where the material to be etched is bombarded with particles in a specific direction) and a chemical effect, i.e., etching. The RIE process can produce extremely fine structures. Furthermore, the trenches formed can be much deeper than they are wide.

**Oxide Steps**

If the residual oxide is not removed following oxidation and targeted etching, an oxide step m (or simply a "step") is formed that does not disappear during subsequent processing. The term "step" refers to the small step-like elevation produced on the surface of the chip and should not be confused with a process "step" (i.e., a stage). Figure 2.12 depicts the formation of steps.



**Figure 2.11** A comparison of wet and dry etching (RIE, reactive ion etching).

****

**Figure 2.12** The formation of oxide steps.

Figure 2.12a depicts the state of the field oxide after structured etching (the first thick oxide layer). Level 1 represents the height of the silicon surface prior to the thermal formation of this oxide layer. If another thermal oxidation is performed, the oxide in the opening beginning at level 2 grows faster, lowering the step height. However, because the surrounding oxide layer has a "head start," a step at the surface will remain (Fig. 2.12b).

As the oxide growth begins in the opening at the lower level 2, a step forms below that grows in height over time. As a result, if the wafer is bright etched, a step that removes the entire oxide layer remains on the wafer surface (Fig. 2.12c).

If the above-mentioned thermal oxidation followed by structuring is repeated many times, additional oxide steps will be formed on the wafer surface. As a result, the wafer surface will become more uneven. This irregular surface makes accurate focusing during photolithography exposure more difficult, resulting in poor imaging of photomask structures on the wafer. This effect prevents process technology from progressing to smaller feature sizes, which means that multiple thermal oxidation followed by structuring cannot be used in cutting-edge technology nodes.

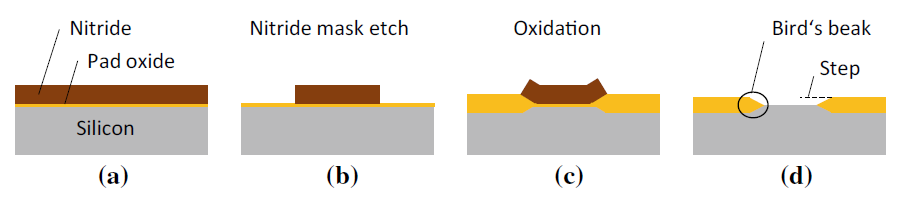
2.4.4 Local Oxidation

To ameliorate the oxide-steps problem described above, the *local oxidation* method (*LOCOS*—local oxidation of silicon) was developed to produce the field oxide.

The field oxide is not structured by masked etching in local oxidation; rather, the oxide layer is only allowed to grow in regions where the field oxide is required.

A material deposition procedure is masked rather than a material removal procedure. Specifically, the regions where the field oxide openings should be are protected from oxidation. Silicon nitride (Si3Ni4) acts as the protective layer.

The procedure is depicted in Figure 2.13. Due to the poor adhesion of silicon nitride (abbreviation: "nitride") to silicon, a thin oxide layer, the pad oxide, is first produced thermally as a bonding agent. The pad oxide layer is then covered with a nitride layer (Fig. 2.13a).

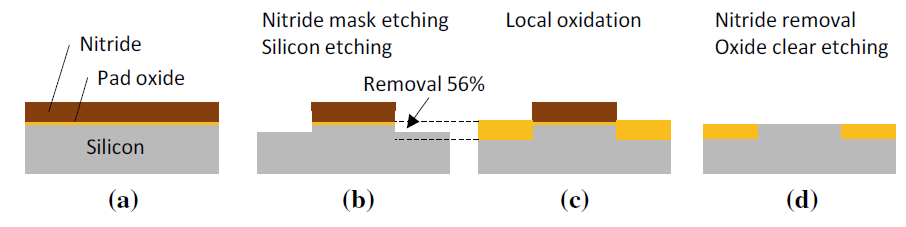


**Figure 2.13** Local oxidation of silicon (LOCOS)

The nitride mask is then etched with a photolithographically structured photoresist (Fig. 2.13b shows the result of this etching). The regions where the field oxide is to grow are now bare. Given that the thermal oxidation is isotropic, the field oxide grows at the nitride boundary in the horizontal direction as well. It pushes its way under the nitride, raising it at the edges (Fig. 2.13c). As the oxide spreads sideways under the nitride, the cross-section of the oxide layer is tapered. Finally, the nitride layer is

chemically dissolved. The residual field oxide is peaked at its boundary—which is also known as a “bird’s beak” (Fig. 2.13d). As a result, the remaining oxide step (d), is only about half as high as the step when structured by etching (cf. Fig. 2.12a).

Various LOCOS technique extensions have been developed to avoid this (minor) oxide step as well. Figure 2.14 depicts an idealized representation of the underlying principle of such an extension. The basic idea here is to first reduce the height of the silicon surface to be oxidized, so that after thermal oxidation, the original height is restored (Fig. 2.14c). The required drop here is 56% of the desired oxide thickness (Fig. 2.14b). The height can be reduced in two ways: I the silicon is masked by the nitride layer and etched, or (ii) oxide steps are built by locally oxidizing twice, as shown in Fig. 2.12. In the latter method, the first oxide layer leaves a step at the desired height, after it—the layer—has been fully removed by bright etching. The desired oxide layer is then obtained with a second oxidation.



**Figure 2.14** Schematic representation of the extensions to the LOCOS process (idealized)

The local oxidation described here offers many benefits over thermal oxidation

structured by etching:

* No silicon is consumed inside the oxide openings.
* The oxide step produced by the LOCOS method is only about half as high as the
* one produced with thermal oxidation (that is structured by etching).

The oxide step is inclined and not steep. This has the advantage that layers

protruding over the edge of the field oxide (polysilicon, metal) cover the edge better.

* The height of the oxide can be reduced by LOCOS extensions to produce an almost

flat surface.

2.4.5 Reference to Physical Design

Several factors that affect oxide structures built on the surface of a chip have been discussed in the preceding sections. We will then look at how they are dealt with during the physical design process.

**Edge Shifts**

If the field oxide is structured by wet etching, undercutting occurs, causing the oxide

openings to be *greater* than the masking structure (see Fig. 2.11, left).

If the field oxide is produced by local oxidation, the bird’s beak effect appears

(see Fig. 2.13d), which causes the oxide openings to be *smaller* than the masking

structure.

Both effects are deterministic, which means that the size of the undercut and the length of the bird's beak can be calculated. The resulting edge shifts can thus be accounted for by appropriate pre-emptive corrections in the post-process when the masks are generated. Because the methods used to address these issues differ between semiconductor processes, the field-oxide openings in the layout may or may not correspond to the actual topography on the chip. We recommend that the reader consult the semiconductor process documentation (known as the process design kit, or PDK) for assistance with this issue.

Through contacts (i.e., contacts and vias) are now exclusively dry etched in processes, resulting in no edge shifts. Edge shifts caused by wet etching of through contacts in legacy processes are typically compensated for by pre-sizing in the layout post process. As a result, through contacts always show the true dimensions in all semiconductor processes in the layout.

**Oxide Steps with Through Contacts**

When etching through contacts, oxide steps are always produced regardless of the etching process. This causes significant issues in the fabrication of metallization layers in modern processes, despite the fact that effective countermeasures exist. These topics will be covered in greater depth in Section 2.8, when we discuss metal coating.