Semiconductor Technology from A to Z

Photolithography

Exposure and resist coating

Overview

In semiconductor manufacturing, structures are created on silicon wafers using lithographic processes. First, a radiation-sensitive film, usually a photoresist layer, is applied to the wafer, patterned, and then transferred into the underlying layer using etch processes.

Photolithography comprises the following process steps:

- applying an adhesion promoter and removing water from the wafer
- coating the wafers with resist
- stabilizing the resist layer
- exposure
- developing the resist
- curing the resist
- inspection

In some processes, such as ion implantation, the photoresist serves as a protective layer to exclude certain areas of the wafer from implantation. In this case, no transfer of the resist mask by an etch process takes place.

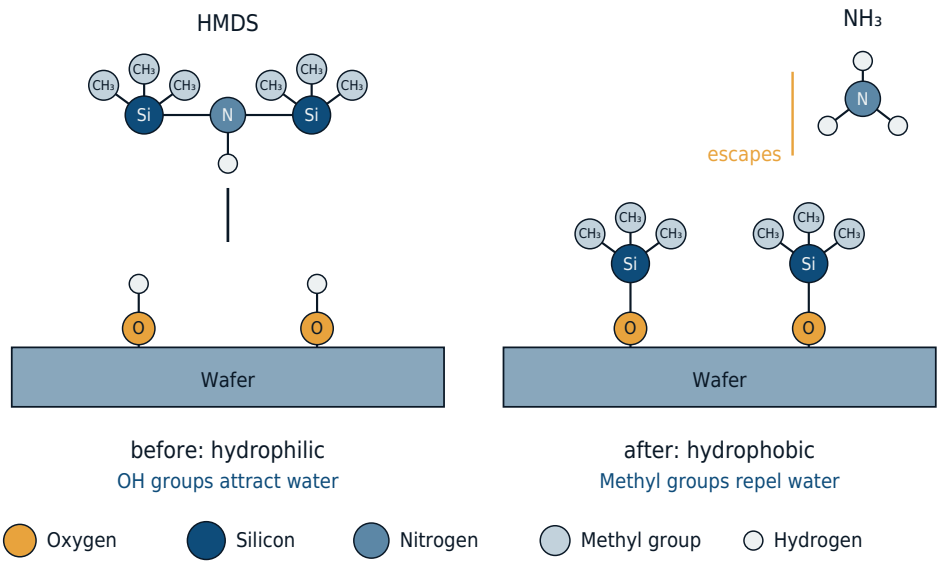
Applying an adhesion promoter

First, the wafers are cleaned and baked out (pre-bake) to remove adhering particles and any adsorbed water. The surface of the wafers is water-attracting (hydrophilic) and must be made hydrophobic – that is, water-repellent and thus resist-attracting – before the resist layer is applied. For this purpose, an adhesion promoter, usually hexamethyldisilazane (HMDS), is applied to the wafers. The wafers are exposed to the vapor of this liquid so that the wafer surfaces are wetted with it.

Due to the moisture in the ambient air, hydrogen H or hydroxide ions OH^- are always present at the wafer surface, even after baking. The HMDS splits into trimethylsilyl groups $Si(CH_3)_3$ and removes the hydrogen, forming ammonia NH_3.

Attachment of HMDS

The water-attracting surface becomes water-repelling



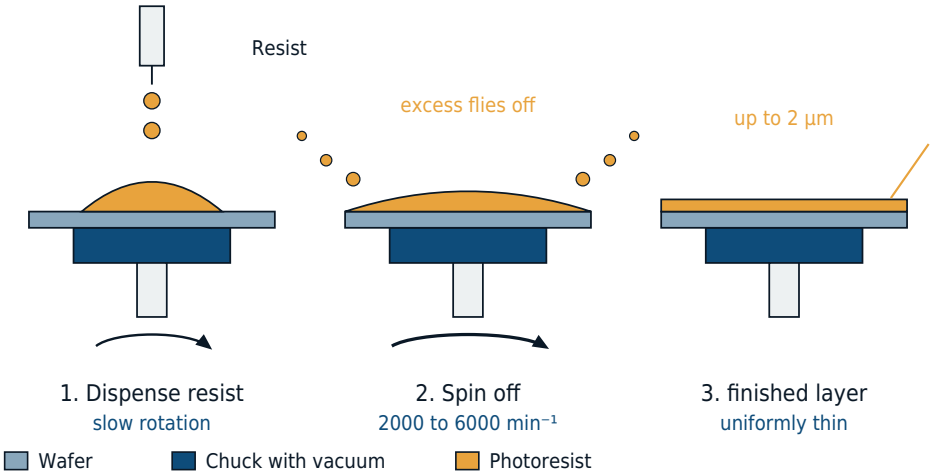
The HMDS splits into two trimethylsilyl groups. Each bonds to an oxygen atom on the surface and takes its hydrogen along; together with the nitrogen, this forms ammonia.

Coating

The wafers are coated with resist by spin coating on a rotating platen with vacuum suction (chuck). At low rotational speed, resist is dispensed onto the center of the wafer and then, at 2000–6000 revolutions per minute, spread out into a homogeneous resist layer by centrifugal force. Depending on the subsequent process, its thickness is up to 2 µm. The thickness depends on the rotational speed and the viscosity of the resist.

Resist Coating by Spin Coating

Centrifugal force spreads the resist drop into a thin layer



The layer thickness follows from spin speed and resist viscosity: the faster the chuck spins and the thinner the resist, the thinner the layer.
By far the largest part of the dispensed resist is spun off in the process.

So that the resist can flow evenly across the wafer, it contains water and solvents that soften it. To stabilize the resist layer, the wafer is subsequently baked out at about 100 °C (post-bake or soft-bake). Water and solvents are partially evaporated; a residual moisture content must be retained for exposure.

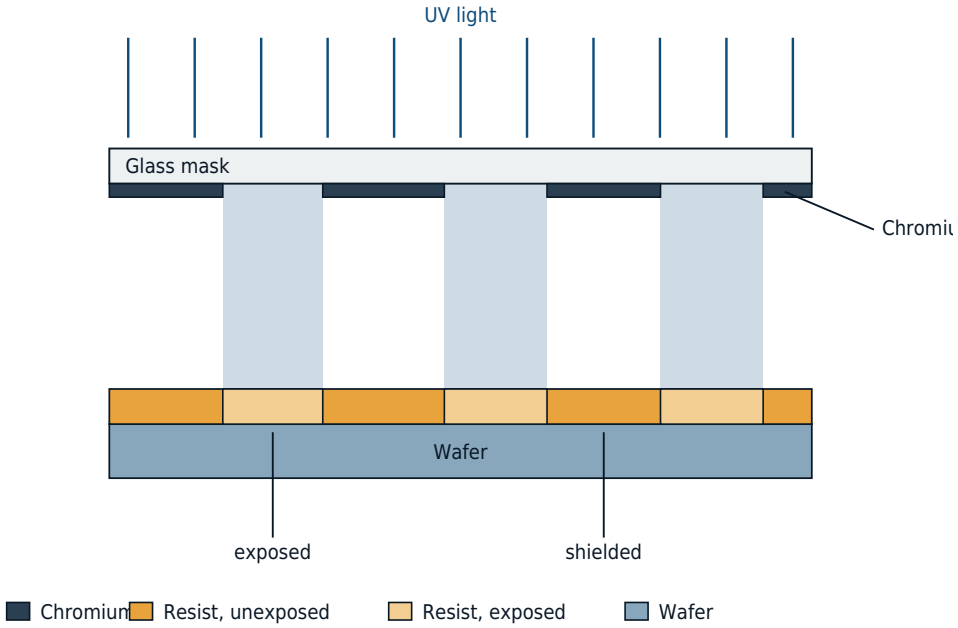
Nowadays, due to the small feature sizes, additional auxiliary layers are often used as well, which are applied to the wafer either before or after the photoresist. Among other things, these layers can serve as an anti-reflective coating or for additional planarization.

Exposure

An exposure tool contains a glass mask that is partially coated with chrome, whereby some areas of the resist-coated wafer are exposed while others are shadowed.

Exposure Through a Mask

Where chromium sits on the mask, the resist stays unexposed

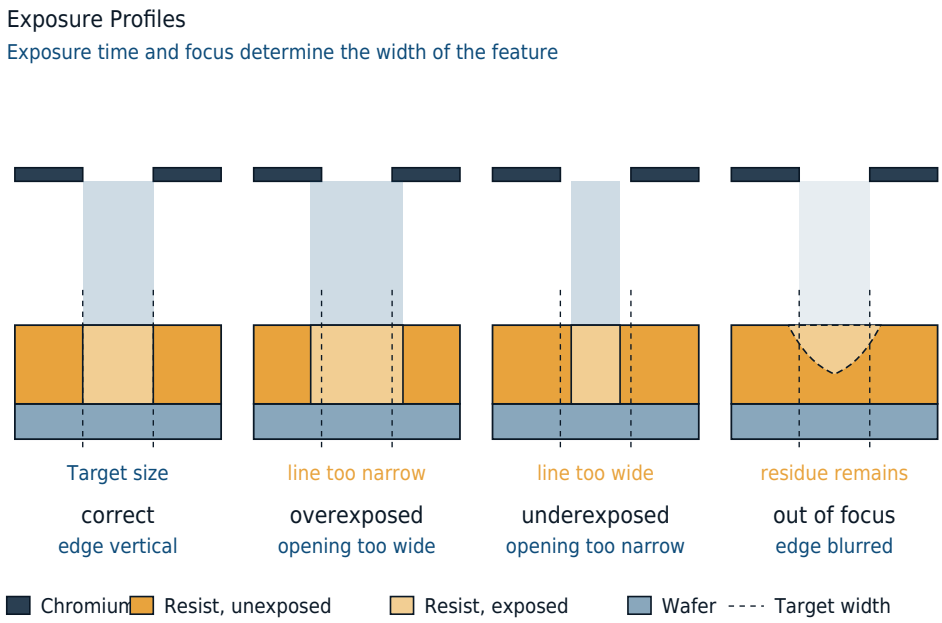


The mask carries the pattern as a chromium layer on glass. During development, depending on the resist type, either the exposed or the unexposed part is removed – leaving behind the patterned resist layer.

Depending on the type of resist, exposed portions become soluble or insoluble. Using a developer solution, the soluble areas are removed, leaving a patterned resist layer behind. In classic i-line positive resist, a nitrogen molecule N_2 is split off by the high-energy UV light. What remains is a keto-carbene, which, for energetic reasons, converts to ketene (CH_2=C=O). By absorbing moisture from the ambient air, the ketene forms a carboxylic acid. At 248 nm and below, this mechanism is replaced by chemically amplified resists, which release a photoacid (see photoresist).

The exposure time is very important so that the structures obtain the exact size. The longer the wafer is exposed to the radiation of the exposure tool, the larger the exposed areas become. A precise determination of the correct exposure duration to achieve the specified structure widths using one or more test wafers (precursors) is necessary, since the resist can behave differently depending on the ambient temperature.

With overexposure, resist lines – and thus the structures beneath them – become too small, while contact holes become too large. With too short an exposure time, the contact holes are not opened, interconnects are too wide, and may end up in contact with one another. In addition, poor focusing leads to unexposed areas, so that contact holes are not exposed during development and connections between interconnects remain, which can lead to short circuits.



The longer the exposure, the wider the exposed area: resist lines get narrower, contact holes larger. With too short an exposure or poor focus, resist remains at the bottom and the contact hole fails to open.

Depending on the subsequent process, the resist dimension – that is, the width of the resist lines or the size of the contact holes – has to be adjusted. In isotropic etch processes (etching occurs both vertically and horizontally), the mask is not transferred 1:1 into the layer to be patterned.

Exposure methods used

Depending on requirements, high-energy radiation such as UV light, x-rays, electron beams, and ion beams are used for exposure. The rule is: the shorter the wavelength of the radiation, the smaller the achievable feature widths. This relationship is described by the Rayleigh criterion:

$$CD=k_1\cdot\frac{\lambda}{NA}$$

Here, λ is the exposure wavelength, NA is the numerical aperture of the lens, and k_1 is a process factor that combines the illumination method, mask technology, and photoresist. The depth of focus decreases with the square of the aperture:

$$DOF=k_2\cdot\frac{\lambda}{NA^2}$$

Every improvement in resolution therefore comes at the cost of a smaller focus window – which is why planar wafer surfaces (see CMP) are a prerequisite for fine structures.

Overview of the wavelengths

Wavelength	Source	Use
436 nm (g-line) 365 nm (i-line)	Mercury vapor lamp	historical; the i-line is still used for non-critical layers, power semiconductors, and MEMS
248 nm	Krypton fluoride excimer laser	coarser layers, analog and power technology
193 nm	Argon fluoride excimer laser	workhorse of manufacturing, both dry and as immersion exposure
13.5 nm (EUV)	Tin plasma, laser-ignited	critical layers of all leading logic and memory processes

Originally, a fluorine laser at 157 nm was also planned as an intermediate step. It turned out to be impractical, since the calcium fluoride lenses required for it absorb moisture from the air; in addition, immersion exposure with water would not be possible, because water absorbs light at this wavelength.

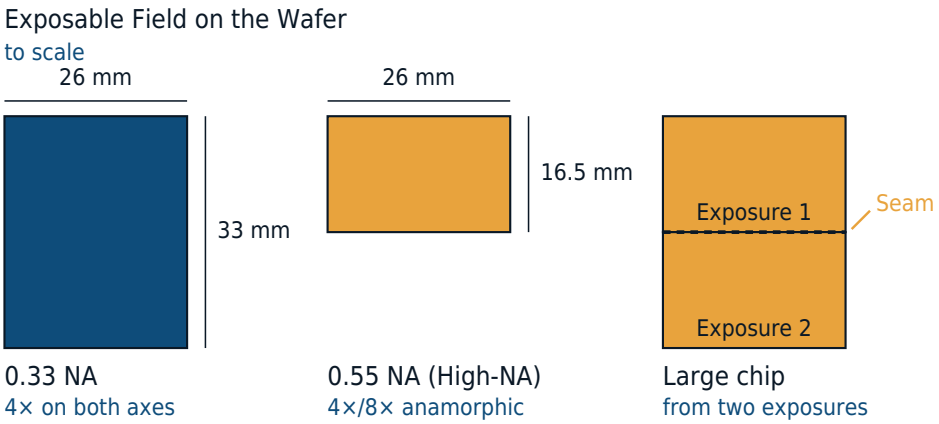
Immersion lithography

Instead of further shortening the wavelength, the numerical aperture can be increased. For this purpose, the gap between the last lens and the wafer is filled with highly pure water, whose refractive index at 193 nm is about 1.44. This allows apertures of up to 1.35, compared with about 0.93 in air. A single immersion exposure resolves structures down to about 38 nm half-pitch – this is the physical limit of this technique. Finer structures at 193 nm can only be produced by splitting a pattern across multiple exposures (see multipatterning).

EUV lithography

The jump to extreme ultraviolet at 13.5 nm has been in volume production since 2019. The radiation is generated by vaporizing tin droplets into a plasma using a high-power laser; since EUV is absorbed by air and by any glass, the entire tool operates under vacuum and images exclusively via mirrors. Today's generation of tools, with an aperture of 0.33, achieves about 13 nm half-pitch in a single exposure, thereby replacing two or more immersion steps on many layers.

Exposable field at 0.33 NA and 0.55 NA



Since 2024, the High-NA generation with an aperture of 0.55 has been added, resolving about 8 nm half-pitch. The price for this is an anamorphic optical system: the image is reduced by a factor of 4 in one direction and by a factor of 8 in the other, halving the exposable field to 26 × 16.5 mm. Large chips therefore have to be assembled from two adjoining exposures. In July 2026, High-NA-exposed layers were qualified in a production product for the first time; widespread adoption is expected for 2027 and 2028.

Other methods

X-ray and ion beam lithography have never reached production and remain confined to research. Electron beam writers, on the other hand, are indispensable – not for exposing the wafers, but for producing the photomasks.

Multipatterning

A single 193 nm immersion exposure resolves structures down to about 38 nm half-pitch. When devices dropped below this limit before EUV lithography became available, only one path remained: the pattern is split across several exposure and etch steps, each of which individually stays within the resolution limit. Together they produce a finer pattern than a single exposure could create.

Multiple exposure with two masks (LELE)

In the litho-etch-litho-etch process, the target pattern is decomposed into two sub-patterns, each with twice the feature spacing. The first is exposed and etched, followed by a second complete pass whose structures fall precisely into the gaps of the first.

The disadvantage lies in the alignment: the position of the second layer relative to the first directly affects the resulting feature width. Every alignment error

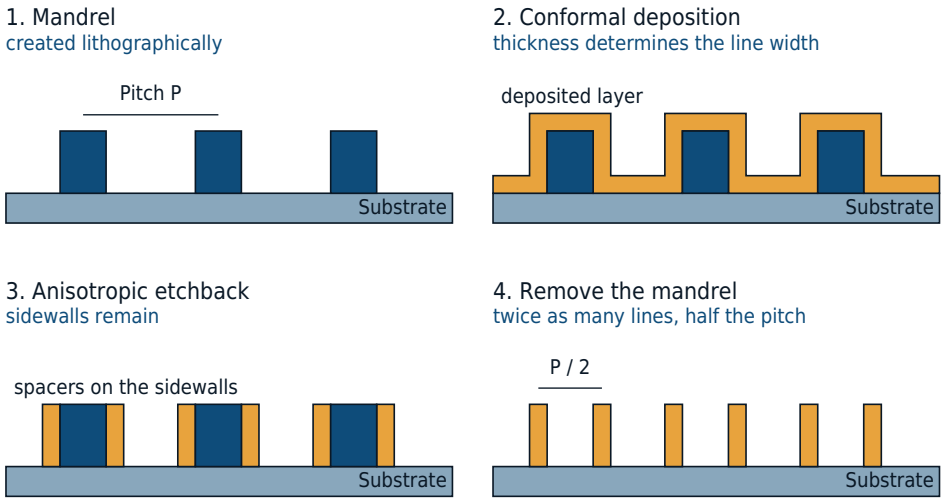
shows up as alternating too-wide and too-narrow spacings – a systematic error that makes the devices non-uniform. With three or four passes (LELELE and beyond), this problem grows, as does the number of process steps and thus the cost.

Self-aligned double patterning (SADP)

The self-aligned process avoids the alignment problem by depositing the fine structures rather than exposing them:

- An auxiliary pattern (mandrel) is created lithographically at normal resolution.
- A thin layer is deposited conformally on top of it, usually by atomic layer deposition. Its thickness determines the resulting feature widths – and it can be controlled far more precisely than an exposure step.
- An anisotropic etch step removes the layer from the horizontal surfaces, leaving lines standing at the sidewalls of the auxiliary pattern (spacers).
- The auxiliary pattern is removed selectively. What remains is twice as many lines as the auxiliary pattern had features.

Sequence of self-aligned double patterning



Since the lines are positioned by the mandrel itself, there is no alignment error between them. Repeating the process yields four structures per original one (SAQP). The price: closed loops always form around the mandrel, and the resulting pattern is inevitably regular. Both issues have to be corrected with additional masks that deliberately cut through unwanted sections (cut mask) or define line ends. The process is therefore suited to dense, regular patterns – interconnects, gate arrays, memory cell arrays – but not to arbitrary layouts.

Significance today

Multipatterning has not remained a stopgap solution. It continues to be used even with EUV, just in a different place: where EUV can complete the critical layers in a single exposure, the split is no longer needed; where structures fall below EUV's resolution, it becomes necessary again. At the same time, circuit design has changed permanently. Layouts today follow strict rules with uniform pitches and uniform preferred directions, because only such patterns can be decomposed at all – lithography dictates its constraints to circuit design.

Exposition methods

Overview

Depending on the type of irradiation, photolithography can be divided into different processes: optical lithography (photolithography), electron beam lithography, x-ray lithography, and ion beam lithography.

In optical lithography, patterned photomasks (reticles) are used. Exposure with UV light or gas lasers is carried out either at a scale of 1:1 or in a reducing scale, for example 4:1 or 10:1.

Today, optical lithography splits into two technically completely different branches:

- DUV (deep ultraviolet, 248 and 193 nm): the mask is transparent, imaging is done through a lens system, and the tool operates either in air or with a film of water between the lens and the wafer.
- EUV (extreme ultraviolet, 13.5 nm): since this radiation is absorbed by every material, the tool operates under vacuum, images via mirrors, and the mask is not a transparency but a mirror.

The exposure methods described below are arranged in order of increasing resolving power. Contact and proximity exposure are found today only in micromechanics, in research, and in education; chip manufacturing exclusively uses reducing projection exposure.

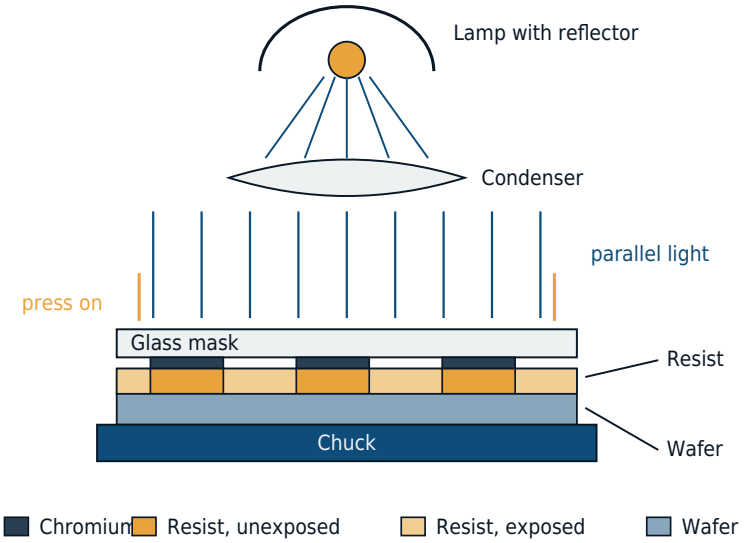
Contact exposure

Contact exposure is the oldest method used. Here, the mask is pressed directly onto the resist layer, and the structures are transferred at a scale of 1:1. Resolution-limiting scattering and diffraction effects of the light occur only at the edges of the structures. However, the achievable feature widths with this

method are small. Since all chips are exposed at once, wafer throughput with this technique is very high, and the construction of the exposure tools is relatively simple.

Contact Exposure

The mask sits directly on the resist, transfer is 1:1



Because mask and resist touch, diffraction effects occur only at the feature edges. The setup is simple and exposes all chips at once.
In return, the mask gets dirty and scratched, since it is pressed onto the resist.

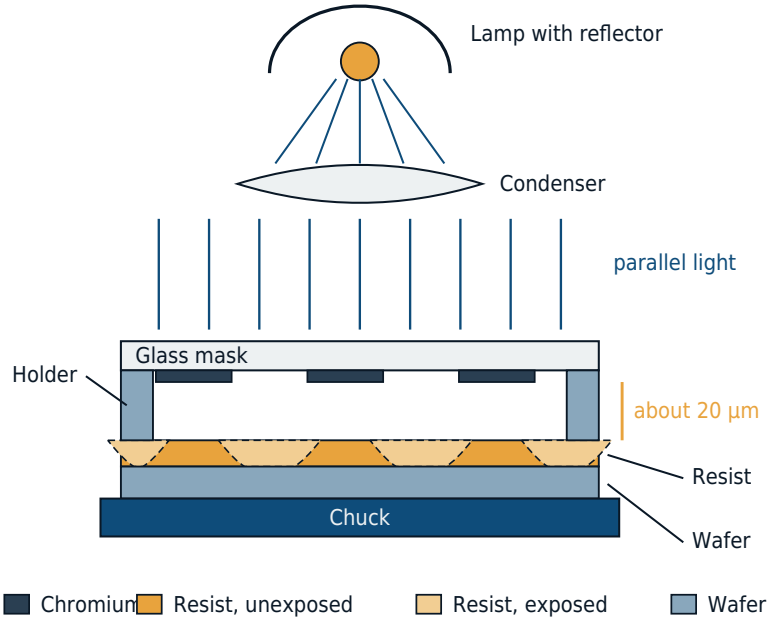
The disadvantages, however, are obvious: because the mask is pressed onto the resist, it becomes contaminated quickly, or can be scratched, as can the resist. If particles are present between the wafer and the mask, the resulting gap between mask and wafer degrades the image.

Proximity exposure

In proximity exposure, contact between the mask and the wafer is avoided by means of a spacer (about 20 µm). In this case, however, only a shadow image of the mask is projected onto the wafer, which offers a significantly worse resolution of the structures.

Proximity Exposure

Spacers hold the mask above the resist – a shadow image results



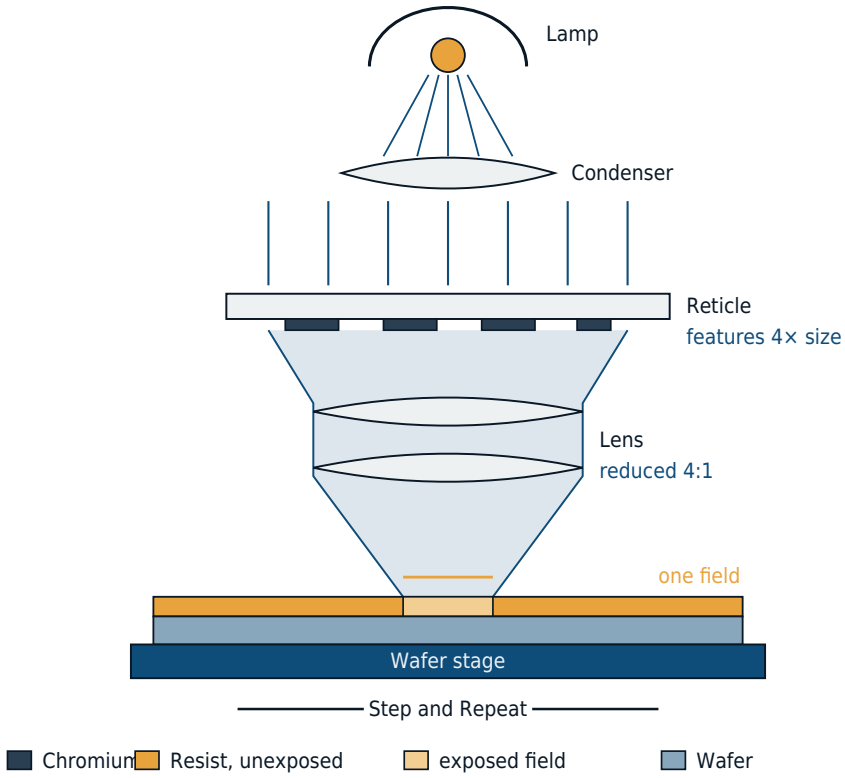
The gap protects mask and resist from scratches and from particles in between. In return, only a shadow image falls onto the wafer: the light diffracts at the chromium edges, so the features widen and their edges become blurred.

Reducing projection exposure

This is the only exposure method still used in chip manufacturing. The step-and-repeat technique is commonly used for this method. A single die – or, for small enough dies, several at once – is imaged onto the wafer via a reticle. In this way, the entire wafer surface is exposed die by die.

Reduction Projection Exposure

The lens images the reticle onto the wafer at reduced scale



Because the features on the reticle are four to ten times larger, particles and mask defects are also scaled down and often fall below the resolution limit. Exposure proceeds field by field, with the wafer stage moving between.

At high resolution, the field is no longer exposed as a whole; instead, the mask and wafer are moved synchronously in opposite directions beneath a narrow slit of light (step and scan). Only in this way can the lens be kept optically corrected across the entire field. The advantage of this method is that the structures depicted on the reticle are enlarged by a factor of 4 (by a factor of 8 in one axis for High-NA EUV tools). When the image is reduced onto the wafer, not only the structures but also any defects, such as particles, are likewise reduced in size, or even fall below the resolution limit; the resolution itself is improved by this method.

Since not the entire mask is used for a single layer, several layers can be placed on one glass plate, which reduces mask costs.

In addition, a pellicle – a thin film with an aluminum frame – can be attached to the mask, which keeps particles away from the mask. These then lie outside the depth of focus and are not imaged.

There is also mirror projection exposure (1:1), in which the wafer is exposed through a complex system of mirrors. Because mirrors are used, no chromatic

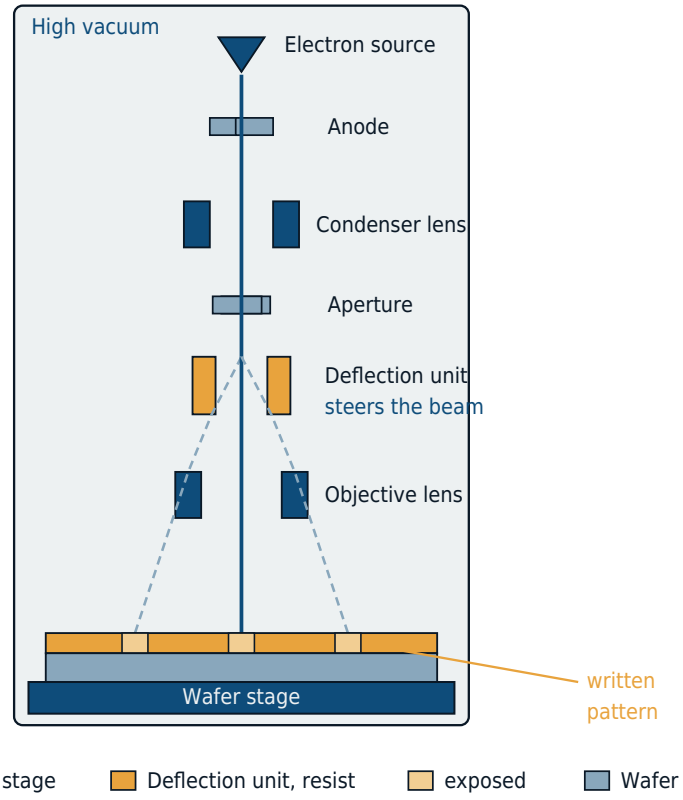
aberrations occur, as they do with lens systems, and in addition, thermally induced expansion of the wafer during processing can be compensated for. However, mirror systems produce distorted/curved images. Furthermore, due to the 1:1 imaging, the resolution is strongly limited.

Electron beam lithography

As in photomask manufacturing, a focused electron beam is scanned across the resist-coated wafer. This scanning can be carried out line by line using the raster-scan method or using the vector-scan method. However, here too every structure has to be written individually, which is very time-consuming. The advantage is that no masks are required, which saves costs. The entire apparatus is housed under high vacuum.

Electron Beam Lithography

A focused beam writes the pattern directly – no mask needed



The beam scans the pattern point by point – row by row in raster scan or targeted in vector scan. This takes time but saves the mask, which is why it is mainly used for prototypes and for making the masks themselves.

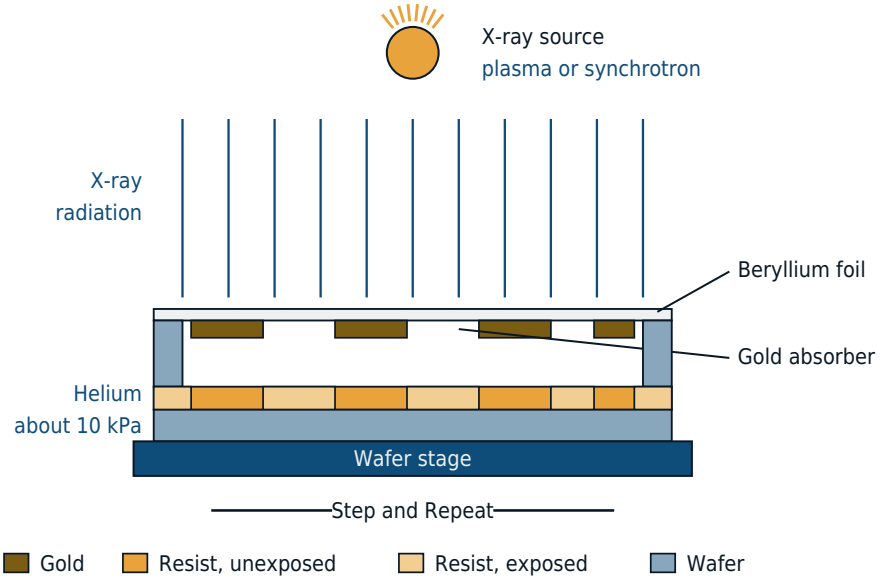
X-ray lithography

The resolution limit of x-ray lithography is about 40 nm. Imaging is carried out using the 1:1 step-and-repeat method via shadow projection, and is performed either at atmospheric pressure in air or at slightly reduced pressure in a helium atmosphere (about 10,000 Pa). The x-ray source can be either a plasma source or synchrotron radiation.

Instead of chrome-coated glass masks, thin, mechanically stable foils made of beryllium, and in some cases silicon, are used. To absorb the x-radiation, the foils are coated with heavy elements such as gold. Both the equipment and the masks are very expensive.

X-Ray Lithography

1:1 shadow casting – there are no lenses for X-rays



X-rays cannot be focused with lenses, so the pattern is transferred directly as a shadow – one to one, field by field. The resolution reaches down to about 40 nm, but the equipment and masks are very expensive.

Additional methods

Another possibility for lithography is irradiating the wafer with ions. With ions, the wafer can be patterned either through a mask or directly, as described for the electron beam method. In the case of hydrogen ions, the wavelength is 0.0001 nm. With other elements, direct doping without masking is also conceivable.

Photoresist

Photoresist

There are positive and negative resists, which are used in different applications. While with positive resist the exposed areas become soluble, with negative resist the irradiated areas become insoluble for the developer solution.

Characteristics of positive resists:

very good resolution
stable in the developer solution
can be developed in alkaline solutions
only limited resistance to etching and implantation processes
poor adhesion on the wafer

Characteristics of negative resists:

very sensitive (allows more precise patterning during exposure than positive resist)
good adhesion
resistant to etching processes and implantation
cheaper than positive resist
low resolution
xylene developer (toxic)

For patterning wafers in manufacturing, almost exclusively positive resists are used. Negative resists are mostly used as a passivation layer (photoimide layer), which can be cured using UV light. Unless stated otherwise in the text, positive resist is meant; where negative resist is used, this is explicitly mentioned.

Chemical composition

Photoresists (also called photo resist; resist = to withstand) are composed of a binder, a sensitizer, and a solvent (thinner).

Binder (proportion ~20%)
: Novolacs are frequently used as binders. These are phenolic resins (synthetic resin, plastic), which primarily determine the thermal properties of the resist.

Sensitizer (proportion ~10%)
: The sensitizer determines the photosensitivity of the resist. Sensitizers are composed of molecules that change the solubility of the resist when exposed to high-energy radiation (in positive resist, the sensitizer forms a carboxylic acid upon exposure. More on this in the chapter Development). So that the resist is not exposed by the lighting in the manufacturing

facilities, the lithography processes take place under yellow light, to which the resist is insensitive.

Solvent (proportion ~70%)
: The solvents determine the viscosity of the resist. By evaporating the solvents on a hot plate, the resist is stabilized and made resistant for subsequent processes.

A resist supplied by the manufacturer has a defined surface tension and density, a specific solids content, and a specific viscosity. Thus, when coating wafers in chip manufacturing, the resulting photoresist thickness depends exclusively on the rotational speed of the coating tool.

Chemically amplified resists

The composition described above applies to the classic resists used at the i-line. At 248 nm and below, the number of incoming photons is no longer sufficient to directly convert enough sensitizer molecules. For this reason, chemically amplified resists (CAR) are used: instead of a sensitizer, they contain a photo acid generator (PAG), which releases a strong acid upon exposure. In a subsequent annealing step (post-exposure bake), this acid cleaves protecting groups from the binder and is not consumed in the process, but instead continues to act catalytically. A single photon can thus trigger hundreds to thousands of reactions.

This amplification mechanism comes at a price: the acid diffuses during the bake step, which blurs the edge of the resulting resist structure. Diffusion length, bake temperature, and bake time are therefore process parameters that determine resolution.

Resists for EUV exposure

At 92 eV, an EUV photon carries roughly fourteen times more energy than a photon at 193 nm. For the same amount of energy delivered, correspondingly fewer photons therefore strike the area, and their arrival is a statistical process. At feature sizes of just a few nanometers, the number of photons per feature fluctuates noticeably – the result is randomly missing or extra structures, as well as a rougher edge. These stochastic defects, rather than the optics, are today the practical resolution limit of EUV lithography.

Countermeasures include higher exposure doses (which cost throughput) and resists with higher absorption. Besides further-developed CARs, tin-based metal oxide resists are being considered for this purpose, as they absorb EUV significantly more strongly than organic systems. Also under development are dry-deposited resists, which are deposited from the gas phase rather than spin-

coated.

The layer thicknesses involved are considerably lower than in classical photolithography: for EUV, they lie in the range of a few tens of nanometers, since tall, narrow resist lines would otherwise topple over due to the surface tension of the developer solution (pattern collapse). Mechanical stability and etch resistance are therefore provided by additional auxiliary layers beneath the resist.

Development and inspection

Development

The exposed wafers are processed using either dip or spray development. While in the dip etch step an entire batch can be developed in a caustic solution and then rinsed, in spray development each wafer is processed individually. As in resist coating, the wafer sits on a chuck and is continuously sprayed with developer solution while rotating slowly. Once the resist has been fully developed, the wafer is rinsed with sprayed water to stop the development process. Some advantages of spray development over the dip process are as follows:

- the smallest structures can be exposed
- the developer solution is continuously renewed: contamination is prevented
- the amount of developer solution used is considerably lower

During development, certain areas of the resist dissolve, depending on the resist type, so that a patterned wafer remains at the end. Exposure triggers a reaction in the resist in which the sensitizer is converted into an acid. During development, this carboxylic acid is converted into a water-soluble salt with sodium hydroxide (NaOH) according to the following equation:

$R\text{-}COOH + NaOH \rightarrow (R\text{-}COO)^- + Na^+ + H_2O$ (R = other unrelated substances)

Illustration of the exposed resist types after development

Since residues of the developer remain on the wafer with potassium or sodium hydroxide solutions, metal-ion-free developers such as TMAH (tetramethylammonium hydroxide) are also used. A further curing step (hard-bake) makes the resist resistant for subsequent processes, such as etching or ion implantation.

Inspection

The resist structure is now inspected. Under oblique light incidence, a microscope can be used to detect the uniformity of the resist layer, as well as poor focusing and resist accumulation. If the resist lines are too thick or too thin, the resist has to be removed and the process repeated. Likewise, the resist structure must be precisely aligned to the underlying layer, otherwise the coating and exposure processes also have to be repeated. For this purpose, there are various alignment marks for checking the alignment accuracy and the line width:

Illustration of alignment marks

The width of the resist lines is checked with a microscope: light rays strike the wafer perpendicularly and are not scattered back to the objective at edges. This makes black lines visible as boundaries of the structures, and by measuring their spacing with the aid of the microscope's magnification, the width of the lines can be calculated.

Resist removal

After the pattern beneath the resist has been transferred by etching, or after the resist has served as a masking layer during an implantation process, the resist has to be removed. This is done using strong etch solutions (remover), in a dry etch step, or with solvents. Acetone is suitable as a solvent for removing the resist layer, since it does not attack the wafer or other layers. However, an ion implantation or a dry etch process can further harden the resist layer, so that solvents can no longer attack and remove the resist.

In this case, the resist can be removed with a remover at about 80 °C in a dip process. If the resist has been heated to above 200 °C during processing, it can no longer be removed even by the remover. In that case, the resist has to be removed by ashing or burning.

In the presence of oxygen, a gas discharge is ignited by high-frequency excitation, generating excited oxygen atoms. These burn off the resist without leaving residue. However, the charged particles are strongly accelerated by the electric field and can thereby cause slight removal of the wafer surface or minor damage to the wafer. Resist ashing (stripping) can be carried out either directly after etching in the same process chamber (in-situ) or in a separate chamber (ex-situ).

Photomasks

Mask technology

The masks used in photolithography contain a pattern with which the respective layer on the wafer is patterned. The starting material for the masks is glass plates that are coated over their entire area with chrome and resist (blanks). Using a resist sensitive to electron beams, the chrome layer is patterned, which then represents the opaque areas on the glass mask.

The masks are written directly with an electron beam. The entire apparatus – the electron beam source, the focusing and deflection unit, and the blank – is housed under high vacuum (0.01–100 Pa; normal air pressure is about 100,000 Pa). The electron beam is guided across the mask under computer control and exposes the resist. With this method, structures well below 100 nm can be resolved.

A single beam, however, cannot write a high-resolution mask in a reasonable amount of time. Today's mask writers therefore work with several hundred thousand individual beams guided in parallel (multi-beam mask writer), which are moved together across the blank and switched on and off individually. Only this makes write times of a few hours per mask achievable – and only this allows arbitrarily shaped, curved structures to be written just as quickly as rectangular ones.

Correction of imaging errors

Due to wave-optical effects (e.g. diffraction), imaging errors can occur during the exposure of the wafers, which are corrected or reduced by so-called *optical proximity correction* (OPC).

For this purpose, the actual structures can be modified so that the image on the wafer corresponds to the desired pattern. In addition, additional auxiliary structures can be written onto the mask that serve only to minimize imaging errors but have no function for the circuit itself.

For the smallest structures, it is no longer sufficient to correct the mask alone. Instead, the mask and the illumination are optimized together (source mask optimization, SMO): the angular distribution of the incident light is also tailored to the respective pattern. The most far-reaching variant is inverse lithography (inverse lithography technology, ILT). Here, a drawn mask is no longer merely corrected; instead, the mask structure that produces the desired result on the wafer is calculated backward from that result itself. The outcome is freely

shaped, often curved contours that no longer bear any resemblance to a drawn template. The computation time required for this now considerably exceeds the writing time of the mask and is provided by large computer clusters.

Because a single mask exposes many thousands of wafers, every defect on it is transferred to every chip. Masks are therefore inspected after writing using high-resolution methods, and individual defects are repaired in a targeted manner – for example, by removing excess absorber material with a focused electron or ion beam.

Photomask manufacturing

In principle, the structures on the masks are produced in the same way as on the wafers. In contrast to the exposure process in wafer manufacturing, where the structures of the mask are imaged onto the resist via shadow projection, the masks are written directly with an electron beam.

1. Exposing a photosensitive resist to pattern a chrome layer on the glass substrate using a laser or electron beam

2. Developing the resist layer

3. Etching the chrome layer

4. Resist removal

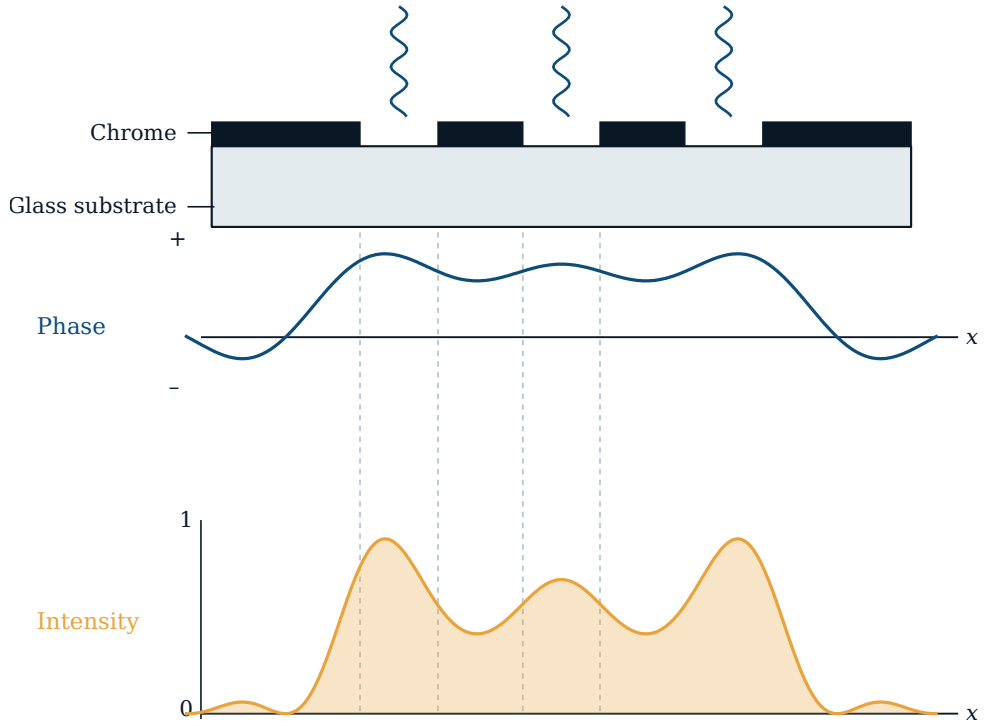
5. Attaching the pellicle

Mask types

Besides the classic chrome mask (COG, Chrome On Glass), there are further mask types that enable improved feature resolution. The main problem with the COG mask is the diffraction that light undergoes at the edges of structures. As a result, the light does not fall onto the wafer only perpendicularly, but is also deflected into the shadow region, where the photoresist is not meant to be

exposed.

Intensity profile of a Chrome On Glass mask

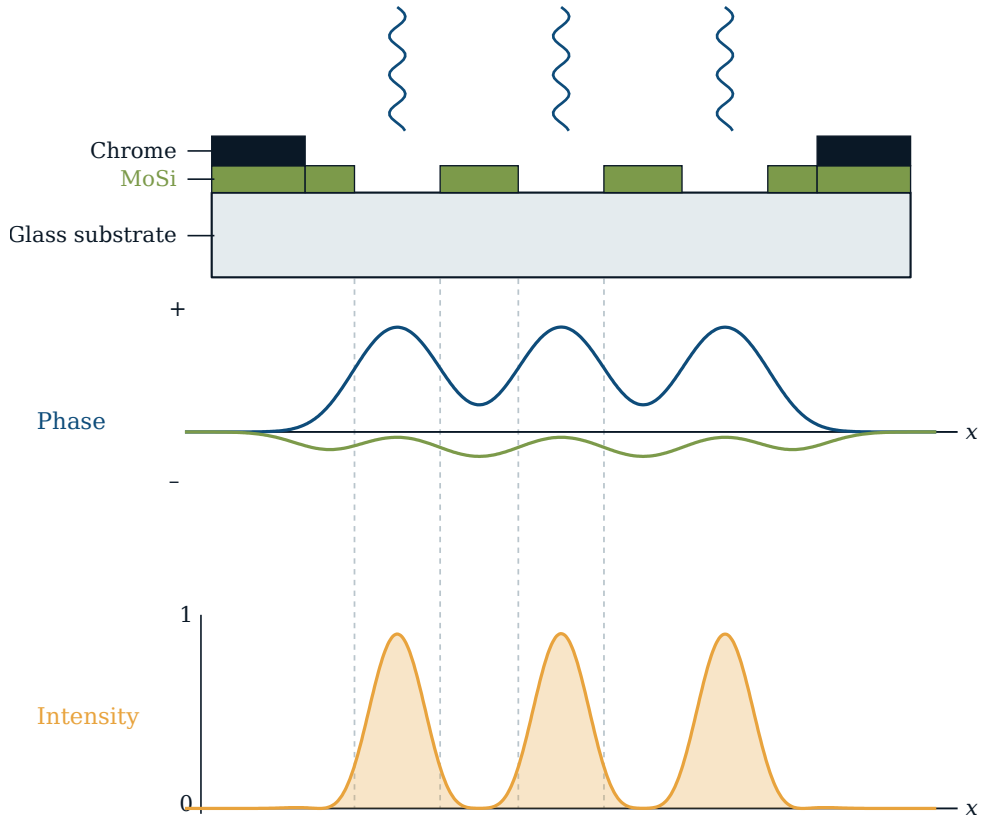


Various measures are used to reduce the intensity of the diffracted light. These are described in more detail below using the different mask types.

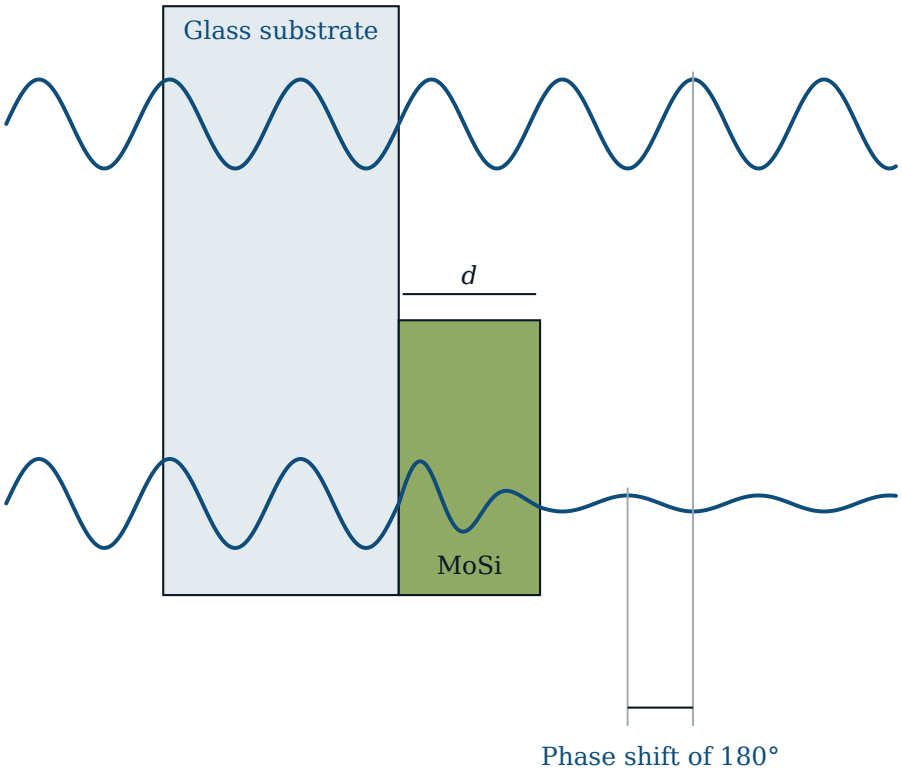
Attenuated Phase Shift Mask (AttPSM)

In the so-called halftone mask, or soft phase-shift mask, a layer of molybdenum silicide (MoSi) forms the pattern-defining part; there is no chrome layer here. The thickness of the MoSi layer is chosen such that light passing through it undergoes a phase shift of 180° relative to light that passes only through glass. This occurs because of the different speed of light in air and in MoSi. At the same time, depending on the molybdenum content in the silicon, the layer is 6 or 18 % transparent (at an exposure wavelength of 193 nm), meaning the light is attenuated. The out-of-phase light waves thus nearly cancel each other out beneath the MoSi structures, increasing the contrast between light and dark. In addition, chrome can be applied in areas not needed for exposure, to block light completely. These masks are referred to as tritone masks.

Intensity profile of an Attenuated Phase Shift Mask



Principle of phase shifting using molybdenum silicide



Chromeless phase-shift mask

Chromeless masks have no pattern-defining coating at all. The phase shift is produced by trenches etched directly into the glass plate. Manufacturing these masks is difficult, since the etch process has to be stopped partway through the glass. In contrast to etch processes in which a layer is etched all the way through and changes in the plasma indicate when the underlying layer has been exposed, here there is no signal indicating when the required depth has been reached.

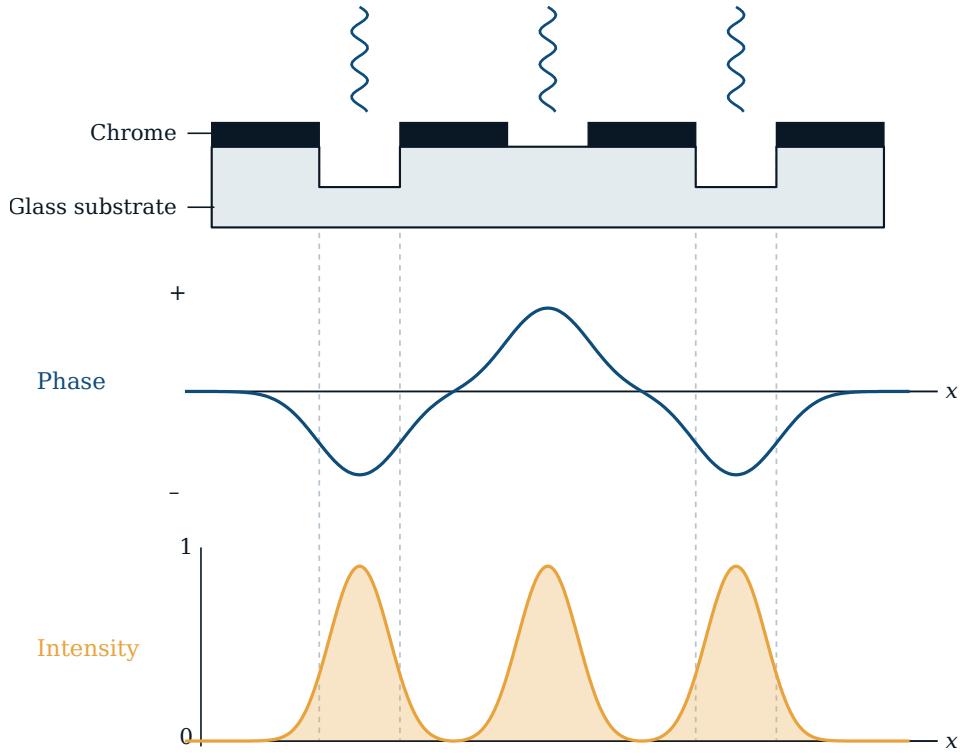
In addition, problems arise when producing large structures. Because there are no shadowing regions and the light strikes the mask everywhere with the same intensity, the desired interference – and the resulting attenuation of the light – occurs only at the edges of the structures. In the middle of larger areas, however, little or no attenuation takes place. To prevent exposure of these areas, the light intensity has to be reduced from the outset, which creates the risk of underexposing all structures.

Alternating Phase Shift Mask (AltPSM)

In the alternating phase-shift mask, trenches are etched directly into the glass substrate, as with the chromeless mask, but alternating with unetched areas. In

addition, certain areas are coated with chrome to reduce the light intensity at the corresponding locations.

Intensity profile of an Alternating Phase Shift Mask



However, this results in regions with an undefined phase shift, so that with this mask type – which enables a very high resolution – two exposures are generally required. The first mask contains the structures running in the x-direction, the second mask the structures in the y-direction.

Target structure on the wafer and corresponding mask structure

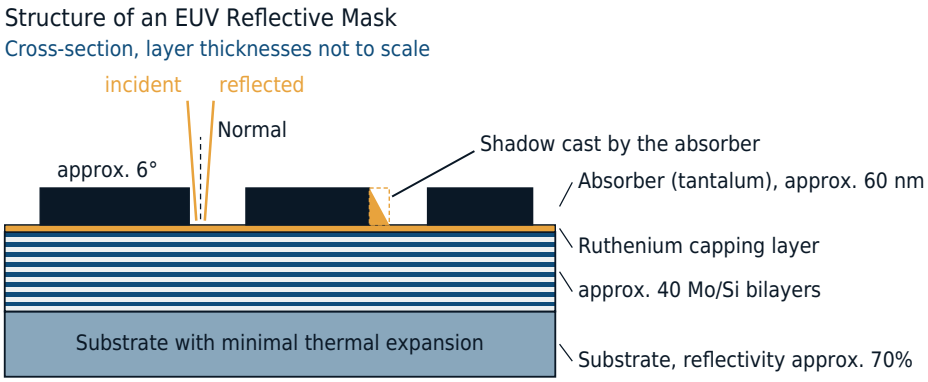
Reflective masks for EUV exposure

All mask types described so far are transmissive: light passes through the glass substrate. For EUV at 13.5 nm this does not work, because every material absorbs this radiation. The EUV mask is therefore a mirror.

The substrate is a glass with virtually zero thermal expansion, since the mask would otherwise warp under irradiation. On top of it lie about forty bilayers of

molybdenum and silicon, whose thicknesses are chosen such that the portions reflected at the individual interfaces superpose constructively. Even this multilayer mirror reflects only about 70 % of the incident radiation; since several such mirrors lie in the beam path, only a small fraction of the generated power reaches the wafer. A thin capping layer of ruthenium protects the stack, and above it an absorber, usually a tantalum compound, forms the actual pattern.

Structure of an EUV reflective mask



The mirror principle gives rise to a peculiarity that does not exist with transmissive masks: the beam has to strike at an angle so that the incident and reflected light can be separated. At this angle of incidence of a few degrees, the roughly 60 nm thick absorber casts a shadow, and this shadow differs depending on the orientation of the structure. These mask effects have to be accounted for when correcting the mask.

Protection against particles is also more difficult. A pellicle here has to be traversed twice by the radiation and must absorb almost none of it; at the same time, the membrane has to withstand the thermal load of a high-power source. For this purpose, extremely thin membranes are used, including ones based on carbon nanotubes.

Phase-shifting EUV masks, which would additionally increase contrast as they do at 193 nm, are the subject of ongoing development but are not yet established in production. At EUV, resolution is therefore gained primarily through the numerical aperture, the illumination, and the resist, rather than through mask technology.

Next-generation lithography

The transition that this chapter described as imminent for years has now taken place: EUV lithography at 13.5 nm has been in volume production since 2019

and exposes the critical layers of all leading logic and memory processes. As predicted, it operates under vacuum, images via mirrors, and uses masks with a reflective rather than a transparent surface.

At the same time, 193 nm immersion lithography has not disappeared. It continues to expose the majority of layers in a modern device, because it is faster and considerably cheaper per exposure. EUV is used only where it pays off: wherever a single EUV exposure replaces two or more immersion steps together with their intermediate processes, it has the advantage.

High-NA EUV

The current generation increases the numerical aperture from 0.33 to 0.55, thereby achieving about 8 nm half-pitch in a single exposure. This comes at the cost of an anamorphic optical system, which demagnifies by different amounts along the two axes and therefore exposes only half the field, as well as considerable expense: a tool costs roughly twice as much as a conventional EUV tool and takes months on site to be qualified. The first layers of a production product were released in July 2026, with widespread adoption expected in 2027 and 2028. Not all manufacturers are taking this path – in some cases, the 0.33 generation combined with multiple exposures is judged to be more economical.

Where the limit now lies

Optical resolution is no longer the actual bottleneck. The limiting factors are:

- Statistics: the stochastic defects caused by the low number of photons per feature (see photoresist)
- Overlay: the permissible error when aligning several layers to one another lies in the range of just a few nanometers and is further consumed by split fields and multiple exposures
- Optical power: the throughput of an EUV tool depends directly on the power of the tin plasma source and on the required exposure dose
- Cost: tools and masks are so expensive that the finest structures only pay off at very high volumes

Alternative approaches

In addition to further increasing the aperture, approaches are being pursued that work without imaging optics at all:

Nanoimprint
: A template carrying the negative of the pattern is pressed into a liquid resist layer, which subsequently cures. This has no diffraction limit and requires no expensive optics, but suffers from the problems of any contact

process: defects and wear of the template. In use for memory devices and optical components.

Direct write with electrons
: Multi-beam systems write the pattern directly onto the wafer without a mask. Attractive for prototypes, small production runs, and customer-specific devices, but still too slow for mass production.

Directed self-assembly
: Block copolymers arrange themselves into regular, fine patterns within a guide structure that is only coarsely predefined by lithography. Suitable for strictly periodic structures such as contact hole arrays, but not for arbitrary layouts.

None of these approaches replaces optical projection in logic manufacturing. Rather, the development of the past twenty years has shown that an established exposure method can be carried much further, through mask technology, illumination, computational correction, and multiple exposure, than the wavelength alone would initially suggest.

Optical Proximity Correction (OPC)

Introduction & Motivation

When the Mask No Longer Prints What It Shows

In the ideal case, a photomask contains exactly the geometry that should later appear on the wafer: a rectangle on the mask yields a rectangle in the resist. This ideal, however, only holds as long as the smallest features on the mask are considerably larger than the wavelength of the light used. Once feature size and wavelength enter the same order of magnitude, diffraction effects dominate the imaging behavior — edges are no longer imaged sharply but are systematically distorted by the optical system.

This relationship is described by the so-called k1 factor, which relates the achievable resolution to the wavelength and aperture of the exposure system. Modern processes operate at k1 values well below what classical, uncorrected optics can still image sharply — the exposure wavelength (still 193 nm in many nodes) is larger than the printed feature width itself. In such a "sub-resolution" process, simply transferring the desired geometry 1:1 is no longer sufficient.

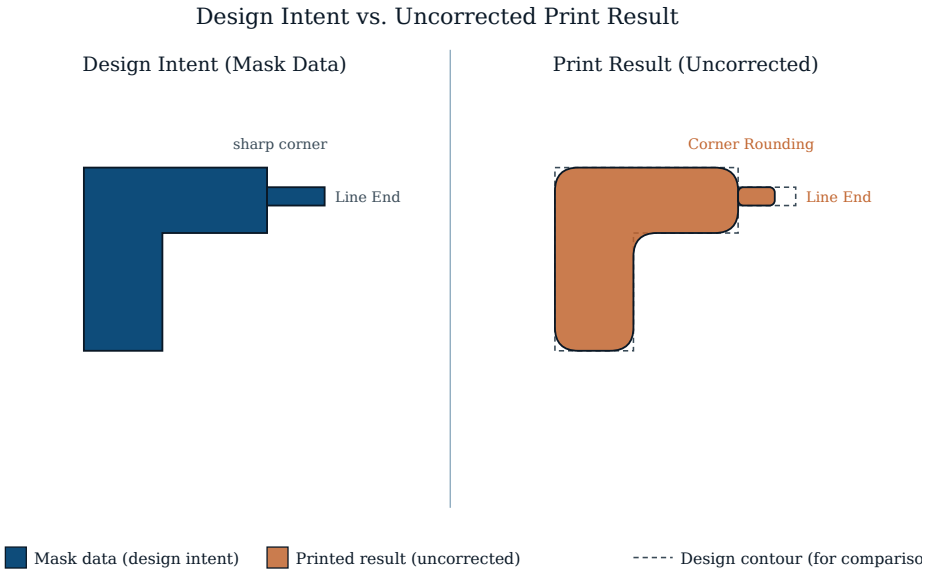
Optical Proximity Correction: Pre-Distorting the Mask

Optical Proximity Correction (OPC) solves this problem by inverting it: rather

than trying to make the optics more perfect, the mask geometry is pre-distorted so that the known, predictable imaging errors of the optical system produce the desired result on the wafer. A corner that would round off without correction is already over-drawn on the mask, so that it appears sharp again after imaging.

OPC is therefore not an after-the-fact repair step but a fixed part of the mask data flow: before actual mask fabrication, every modern layout passes through automated OPC software that adjusts the geometry chip-wide based on a calibrated model of the exposure process. Without this step, most of today's feature sizes simply could not be manufactured with the available exposure wavelength.

The diagram below shows the basic problem using a simple example: the desired layout on the left, and the result that would actually print without correction on the right.



Typical Print Errors Without Correction

Corner Rounding and Line-End Shortening

The effects shown in the previous section are the most immediate consequence of diffraction: sharp, right-angled corners in a mask structure become rounded contours in the resist, because the optical system no longer transmits the high spatial-frequency components that sharp edges would require. At convex corners, the rounding "eats into" the structure; at concave corners (for instance the inner corner of an L-shaped layout), it instead fills in the notch.

A related phenomenon occurs at the end of a trace: line-end shortening. Because

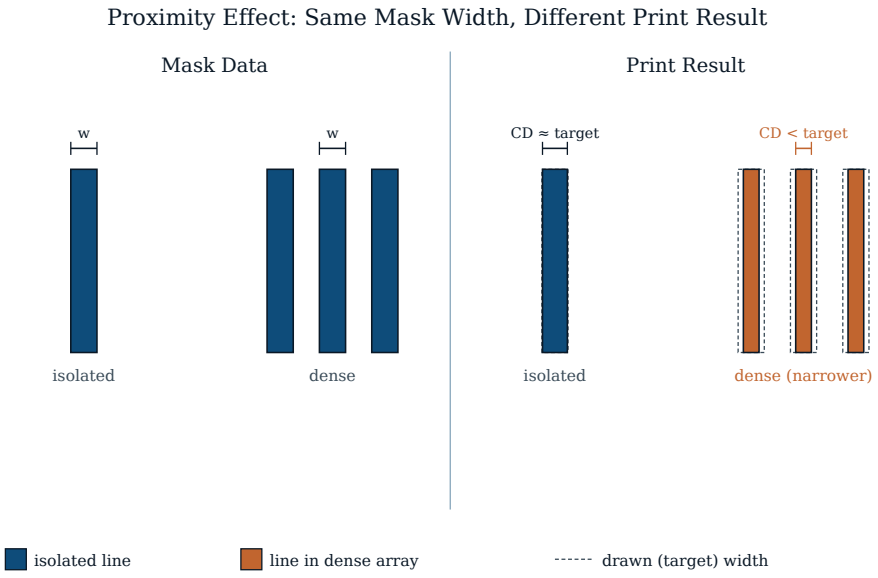
the optical system images the end of a line not as a sharp edge but as a gradually falling intensity, the resist that actually develops at the line end pulls back relative to the mask specification — a line meant to reach exactly to a certain contact may, in print, end a few nanometers short of it.

The Proximity Effect: Same Width, Different Result

The effect that gives Optical Proximity Correction its name, however, is a third, more subtle one: two lines drawn with exactly the same width on the mask print at different widths depending on how many other structures sit in their immediate neighborhood. An isolated line receives its diffraction pattern solely from its own edges. A line in the middle of a dense line array, by contrast, has its diffraction pattern overlap with those of its neighbors — the resulting intensity profile in the resist differs systematically from that of the isolated line, even though the mask geometry is identical.

This so-called iso-dense bias varies in magnitude across the pitch spectrum (center-to-center spacing of features) and is non-linear — it must be characterized separately for each process, typically by systematically exposing and measuring test structures with varying pitch. The result of this characterization feeds directly into the correction model introduced later in this chapter.

The diagram below shows the proximity effect on an isolated line compared to a line in the middle of a dense array — both with identical mask width but different print results.



Mask Bias & Serifs

Mask Bias: Correcting the Dimension

The simplest form of correction addresses the proximity effect from the previous section directly: if it is known that a line in a dense array systematically prints narrower than intended, it is simply drawn somewhat wider on the mask from the outset — the so-called mask bias. Since the magnitude of this deviation depends on the local pitch, the bias is not a fixed constant but is determined separately for each neighborhood situation in the layout, based on the previously characterized iso-dense curve: densely packed regions receive a different bias value than isolated lines or intermediate pitch ranges.

Mask bias thus corrects only the line width as a whole — it does not change the shape of a corner or a line end. These local, geometrically confined errors require a second, more targeted correction mechanism.

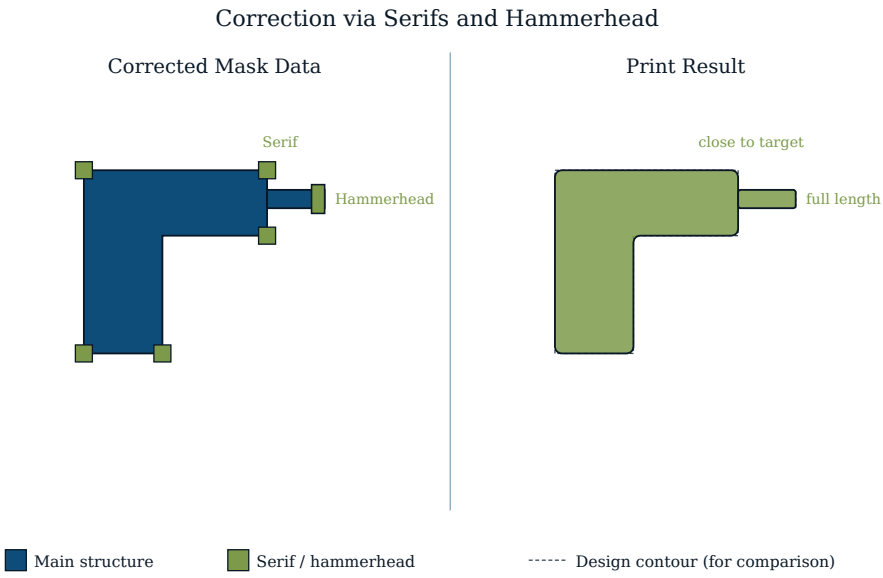
Serifs and Hammerheads: Targeted Corner Reinforcement

To counteract corner rounding, a small additional square structure is added directly to the mask at particularly sensitive corners — above all at convex outer corners — a so-called serif. The serif itself remains below the resolution limit and does not print as a standalone structure; however, it locally alters the light intensity at the edge of the main structure so that the corner appears sharper after imaging than it would without this reinforcement.

At the end of a line, an enlarged variant of the same principle corrects line-end shortening: a so-called hammerhead — a local widening of the line end perpendicular to the line direction — compensates for the systematic pullback of the resist at the line end, so that the line still reaches its intended position even after optical imaging.

Both techniques — serifs and hammerheads — can be formulated as simple geometric rules: "add a square of size Y at every convex corner with interior angle smaller than X" or "widen every line end by Z nm." This rule-based approach is the simplest form of OPC and will be contrasted with the model-based variant two sections from now.

The diagram below shows the layout geometry from Section 1 after adding serifs and a hammerhead, along with the substantially improved print result this produces.



Sub-Resolution Assist Features

The Problem: Isolated Lines Behave Differently

Serifs and mask bias correct the shape and width of a structure directly at its own contour. For a more fundamental problem, however, this is not enough: an isolated line receives its diffraction pattern — as described in Section 2 — solely from its own edges, while a line in a dense array benefits from the diffraction patterns of its neighbors. This difference in diffraction conditions affects not only the printed width but also how tolerant the respective structure is to focus and dose variations in the exposure process — its so-called process window. Isolated structures typically have a considerably smaller process window than densely packed ones, even if both print at exactly the same width after correction.

Serifs and bias alone cannot fix this deficit, since they only alter the structure itself, not its optical surroundings.

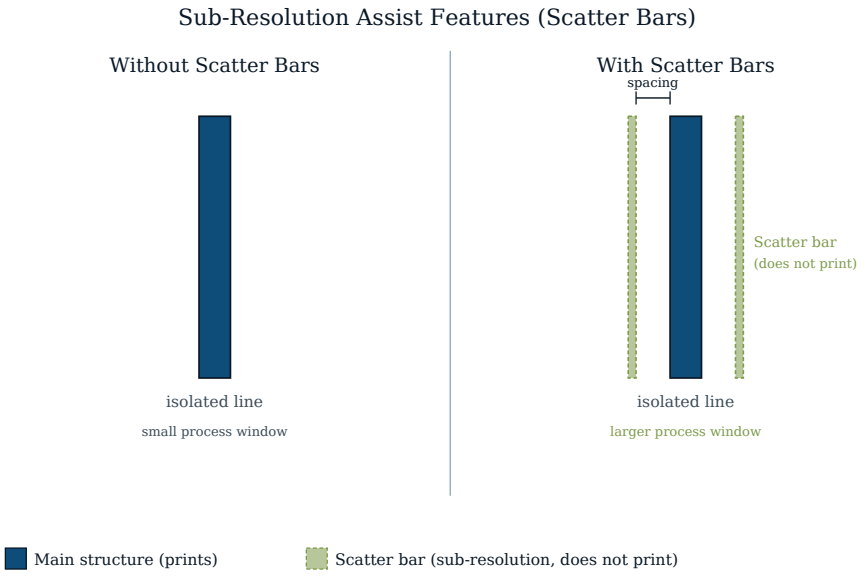
Scatter Bars: Neighbors That Never Print

Sub-resolution assist features (SRAFs), colloquially also called scatter bars, solve this problem by artificially giving the isolated line neighbors: additional, very narrow lines are drawn on the mask parallel to the actual structure — narrow enough to remain below the resolution limit of the exposure system themselves and therefore not appear as a standalone structure in the resist. They do, however, alter the diffraction pattern at the location of the main structure so that it more closely resembles that of a densely packed

environment — the isolated line now optically "looks" almost like a line within an array, with a correspondingly larger, more stable process window.

Placing SRAFs is subject to strict geometric rules: the spacing to the main structure must be large enough to avoid merging but small enough to still have an effect; the width of the scatter bars themselves must remain safely below the printing threshold, even under unfavorable process conditions. Software-assisted SRAF placement today generates these assist structures automatically based on the characterized diffraction behavior of the respective process, typically as the last step before the actual serif and bias correction.

The diagram below shows an isolated line without and with scatter bars — the assist structures themselves do not appear in the printed resist.



Rule-based vs. Model-based OPC

Rule-Based OPC: Fast but Coarse

The corrections introduced in the previous sections — mask bias, serifs, scatter bars — can be organized as a lookup table: for a given line width and a given spacing to the nearest neighboring structure, the table returns a previously characterized correction value. This rule-based OPC is computationally cheap, since only a single table lookup is needed per edge, with no elaborate simulation.

The limits of this approach show up with complex, two-dimensional geometry: near a corner, a line end, or an unusual neighborhood configuration, several proximity effects overlap simultaneously, for which no single table row can still

provide a fitting answer. Rule-based OPC always treats an edge as a whole: the entire straight edge segment is shifted by the same amount, regardless of whether an additional structure near one end creates a different correction need there than at the other end.

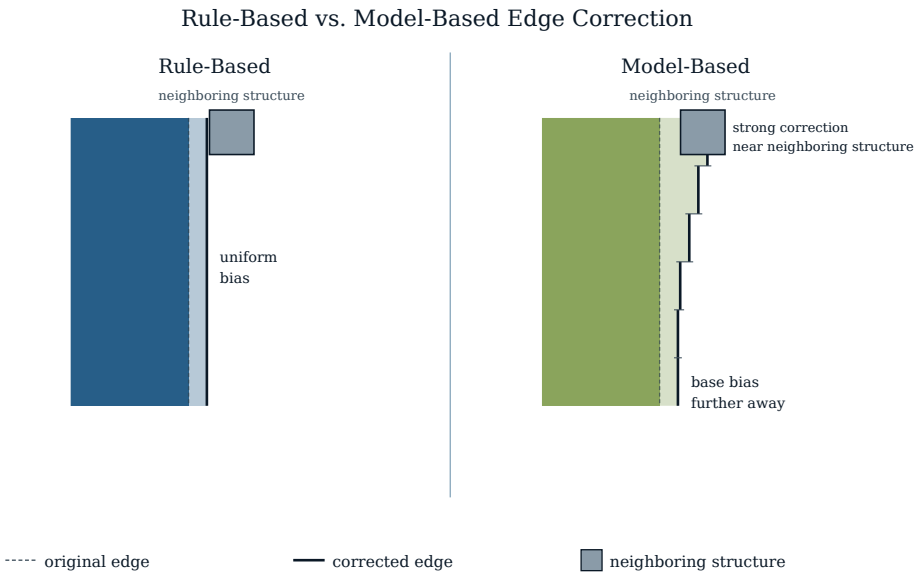
Model-Based OPC: Simulated Segment by Segment

Model-based OPC takes a fundamentally different approach. Instead of looking up table values, it uses a calibrated numerical model of the entire imaging process — optical model and resist model combined — to predict the actually expected printed contour for a given mask geometry. To do this, every edge of a polygon is subdivided into many short segments, which can then be shifted independently of one another.

The correction algorithm works iteratively: it simulates the print of the current mask geometry, compares the result against the target contour, shifts each segment individually toward a better match, and repeats this cycle until the simulated contour agrees with the target within a defined tolerance or a maximum number of iterations is reached. Because each segment responds individually, the correction near a complex 2D situation — such as near a corner — can turn out considerably stronger than further away on the same edge, where simple rule-based correction would only know a single, averaged value.

This level of detail comes at a cost: model-based OPC requires a lithography simulation across thousands of segments simultaneously, chip-wide, for every iteration — a computational effort orders of magnitude beyond that of the rule-based variant. In practice, modern OPC flows therefore combine both approaches: rule-based correction for simple, well-characterized situations, model-based refinement wherever the geometry demands it.

The diagram below compares both approaches on the same edge near a neighboring structure: on the left, the uniform shift of rule-based correction; on the right, the segment-by-segment shift of model-based correction, adapted to the local geometry.



Verification: Lithography Simulation & Process Window

Why Correction Alone Is Not Enough

An applied OPC correction is initially just a claim: the correction model predicts that the adjusted mask geometry will print the desired result under the assumed process conditions. Whether this prediction actually holds across the entire chip design — often repeated billions of times — and whether it survives realistic variation in the manufacturing process, must subsequently be checked separately. This OPC verification runs technically similar to the actual correction: a full-chip lithography simulation, but this time not to adjust the geometry, but to check the already corrected layout.

The Process Window: More Than Just the Ideal Case

Exposure tools do not hold focus and dose perfectly constant — both vary within a certain range from wafer to wafer, across the wafer surface, and even within a single exposure field. A correction that only works at the exact target focus and target dose would be worthless in practice. OPC verification therefore simulates not just the nominal operating point but an entire process window of focus and dose combinations — typically the four extreme corners plus several intermediate points — and checks whether the structure stays within the allowed tolerance at every single one of these points.

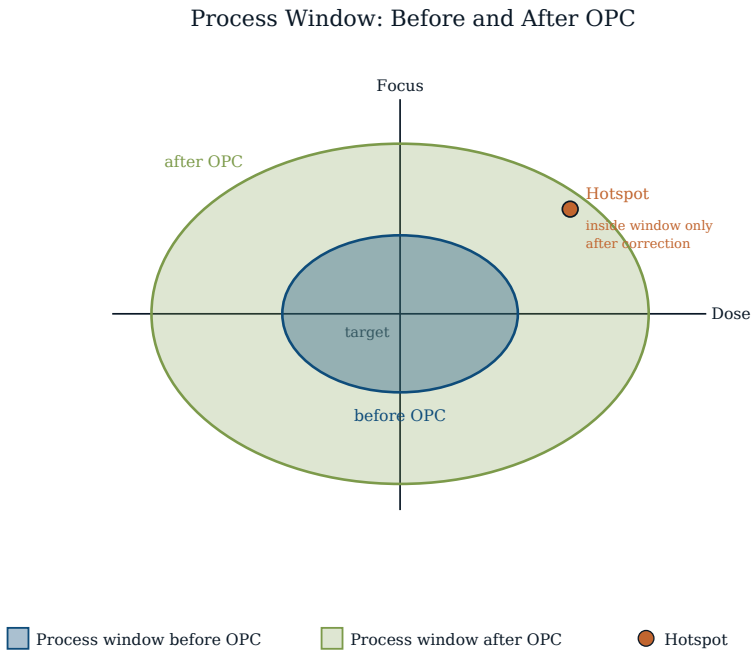
Well-executed OPC noticeably widens this process window compared to the uncorrected layout: structures that previously only printed cleanly within a narrow focus-dose range remain within specification after correction even under

larger deviations. The wider window is thus a direct, measurable quality metric for the correction itself — in addition to plain agreement at the nominal operating point.

Hotspots: Where the Correction Falls Short

Locations where the simulated contour violates a design rule at one or more points of the process window — for instance because two neighboring lines nearly touch under unfavorable focus (bridging) or a line narrows so much that it nearly breaks (necking) — are flagged as hotspots. Every hotspot found must be fixed before tape-out: through a targeted local re-correction, or if necessary through a manual layout change at that location. Verification then reruns at the affected location until even the worst-case point in the process window stays within tolerance.

The diagram below shows the principle using a process window diagram: the axes represent focus and dose deviation from the target value, the shaded area marks the combinations at which the structure still prints correctly — before and after OPC compared, with a hotspot that only falls inside the window after correction.



Limits and Outlook

The Computational Cost of Full-Chip OPC

A modern chip contains several billion polygons. If every edge is subdivided into dozens of segments for model-based correction and each segment is simulated across multiple iterations, the resulting computational effort poses a serious challenge even for large data centers. Full-chip OPC runs today routinely spread across thousands of CPU cores in parallel and still often require many hours to individual days of compute time — per mask layer, and a modern process has several dozen of those.

This data volume has also changed the mask data format itself: the classic GDSII format, originally designed for comparatively simple, uncorrected layouts, runs into practical limits with the heavily fragmented, serif-rich geometries that result from OPC. The OASIS format was specifically developed to store the substantially larger data volumes after correction more compactly, among other things through more efficient compression of repeating geometry patterns.

EUV: New Wavelength, New Effects

With the transition to EUV lithography (extreme ultraviolet, 13.5 nm wavelength), the starting situation shifts fundamentally: since the wavelength now lies far below the printed feature size, the k1 factor described in Section 1 is considerably more relaxed for many structures than with 193 nm immersion lithography — the classic proximity effects appear weaker in many cases, and the correction demand from mask bias and serifs decreases accordingly.

In exchange, EUV brings its own, novel effects that require their own correction treatment. Since no transmissive lenses exist at this wavelength, EUV lithography operates exclusively with reflective mirror optics, and the light strikes the mask not perpendicularly but at an angle — this gives rise to so-called mask 3D effects, in which the finite thickness of the mask absorber layer itself casts shadows and distorts the image asymmetrically, depending on a structure's orientation relative to the angle of incidence. Added to this is a lower photon count per exposure, leading to statistical fluctuations — so-called stochastic effects — which cannot be fully mastered by classical, deterministic OPC alone.

OPC therefore does not disappear with EUV but transforms: the fundamental principles introduced in this chapter — mask bias, serifs, SRAFs, model-based segment correction, process-window verification — remain valid but are extended with additional model components for mask-3D and stochastic effects. For structures that fall below the resolution limit of a single exposure even with EUV, multi-patterning is added on top: a layout is split across multiple masks whose OPC corrections must be coordinated with one another — a topic beyond

the scope of this chapter, but one that builds on exactly the foundations introduced here.