Semiconductor Technology from A to Z

Photolithography

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

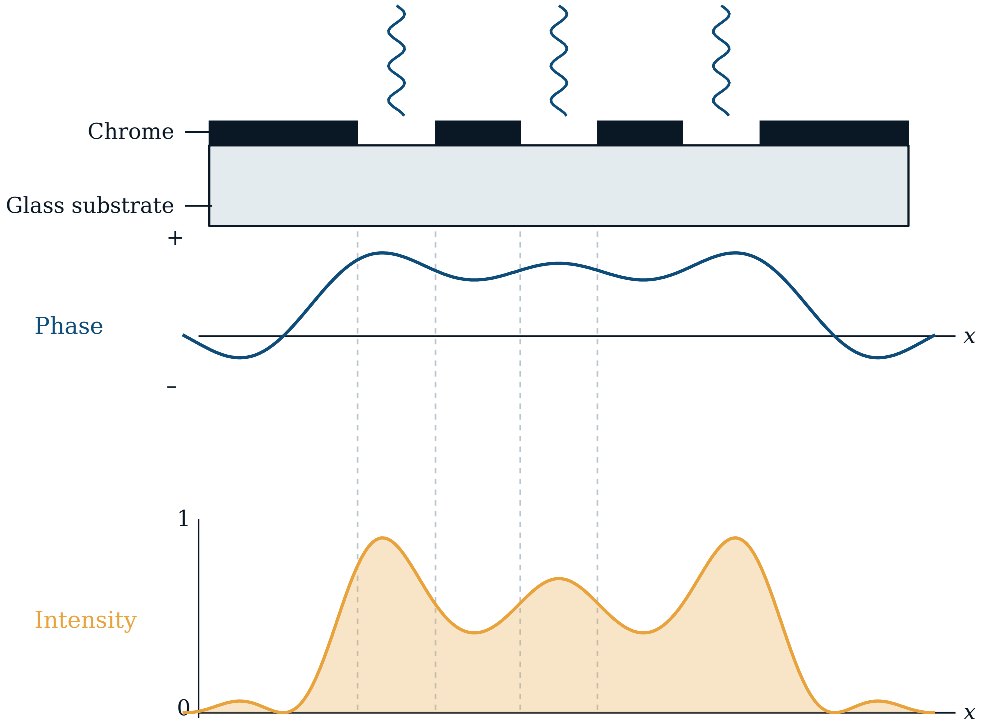
4. Resist removal

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

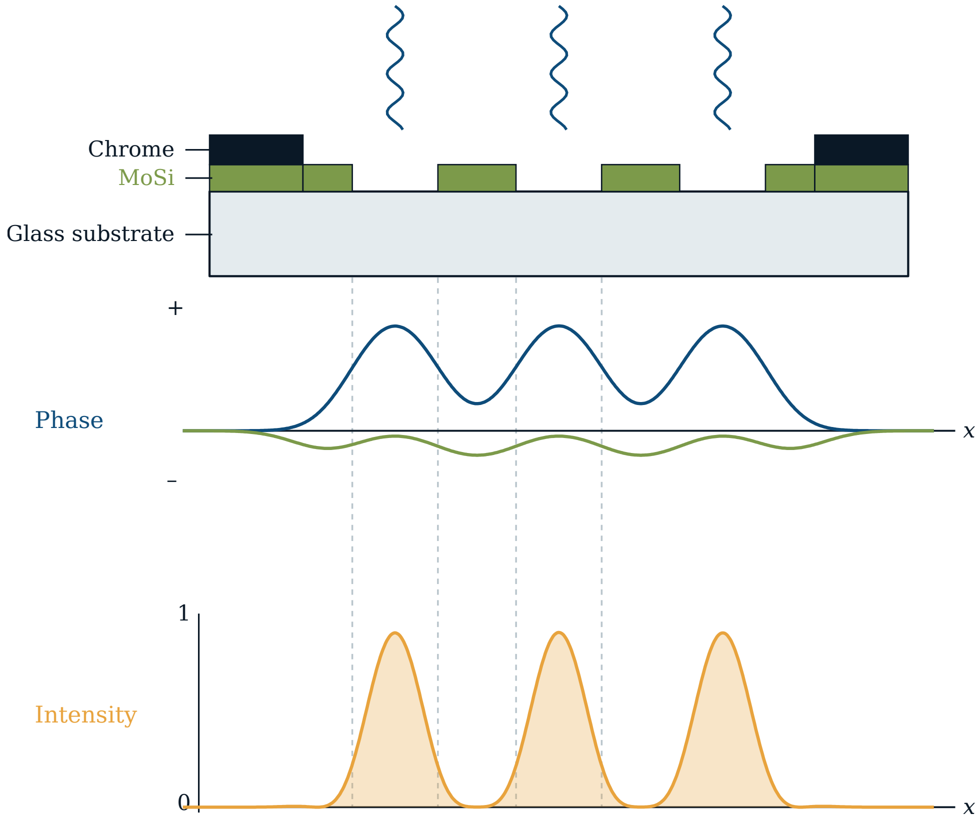


Various measures are used to reduce the intensity of the diffracted light. These are described in more detail below using the different mask types.

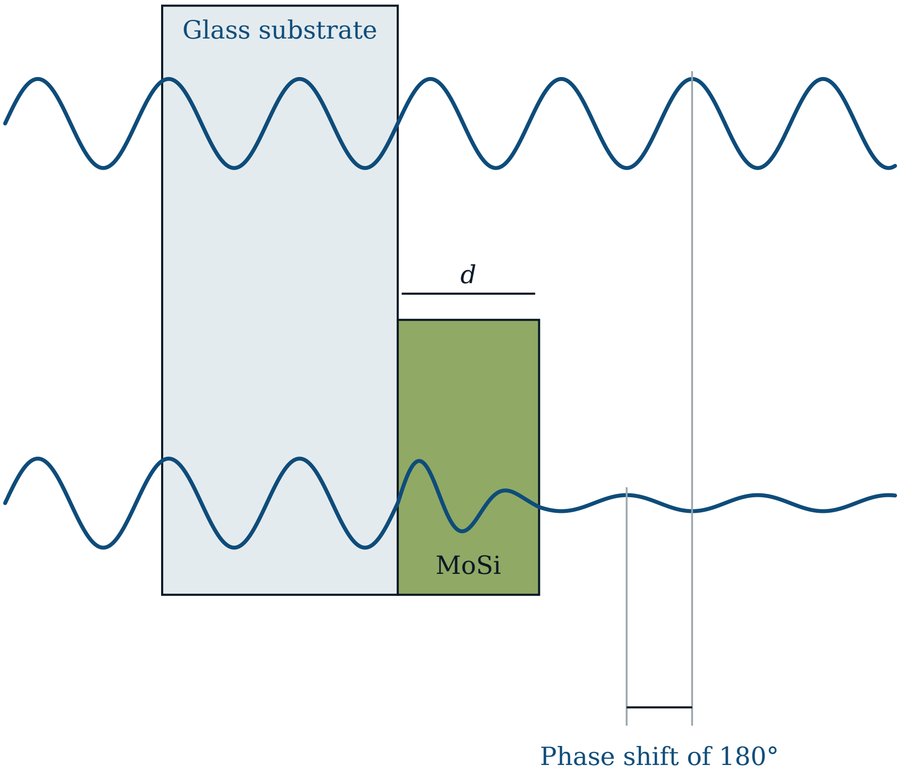
Attenuated Phase Shift Mask (AttPSM)

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

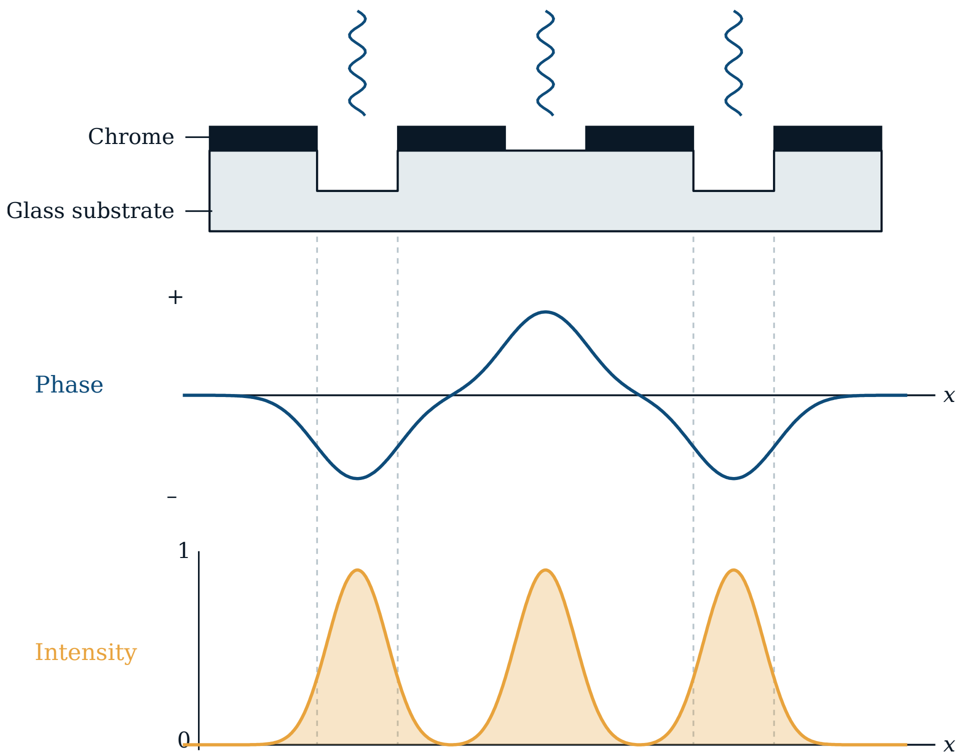
Chromeless masks have no pattern-defining coating at all. The phase shift is produced by trenches etched directly into the glass plate. Manufacturing these masks is difficult, since the etch process has to be stopped partway through the glass. In contrast to etch processes in which a layer is etched 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.

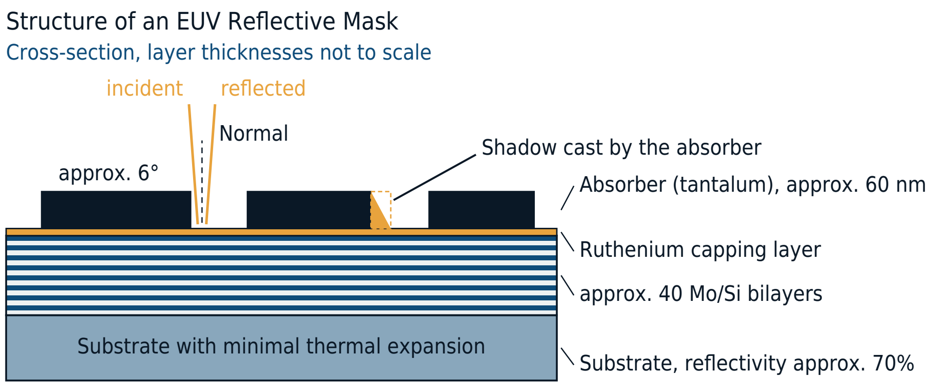
Target structure on the wafer and corresponding mask structure

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.