Semiconductor Technology from A to Z

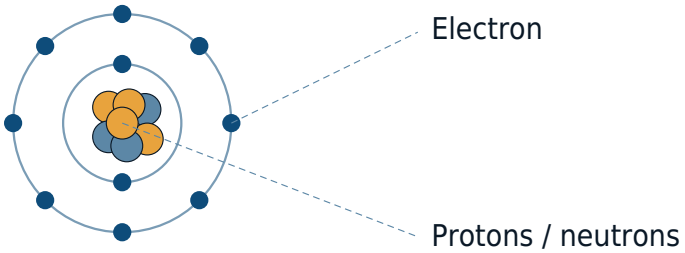
The atomic structure

The Atomic Model

An atom [Gr. átomos: indivisible] is the smallest building block of matter that cannot be further divided by chemical means. There are various types of atoms, each composed of a specific number of electrons, protons, and neutrons. Positively charged protons and uncharged neutrons form the atomic nucleus, which is orbited by negatively charged electrons at certain distances. An atom occurring in nature is electrically neutral, meaning it contains exactly as many protons as electrons. The number of neutrons, however, can vary.

According to the Bohr atomic model, the electrons are located on so-called shells, which represent different energy levels and are arranged concentrically around the atomic nucleus. There are a maximum of seven shells, which can hold varying numbers of electrons; the electrons in the outermost shell are also referred to as valence electrons.

Highly simplified representation of a neon atom

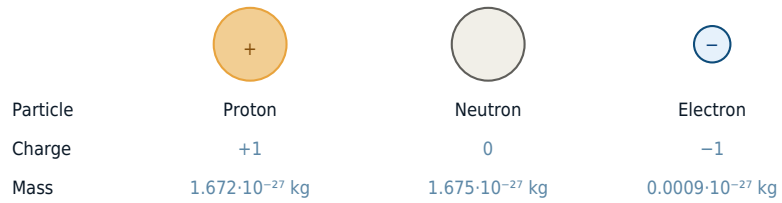


By exchanging electrons with other atoms, atoms can achieve a more stable state, which is why different types of bonds form between atoms.

Orders of Magnitude

Mass
: The mass of an atom is determined mainly by the atomic nucleus, since the mass of protons and neutrons, at $1.67 \cdot 10^{-27}$ kg, is roughly 1800 times greater than that of the electrons in the electron shell ($9.11 \cdot 10^{-31}$ kg).

Building blocks of the atom



Particle	Proton	Neutron	Electron
Charge	+1	0	−1
Mass	$1.672 \cdot 10^{-27}$ kg	$1.675 \cdot 10^{-27}$ kg	$0.0009 \cdot 10^{-27}$ kg

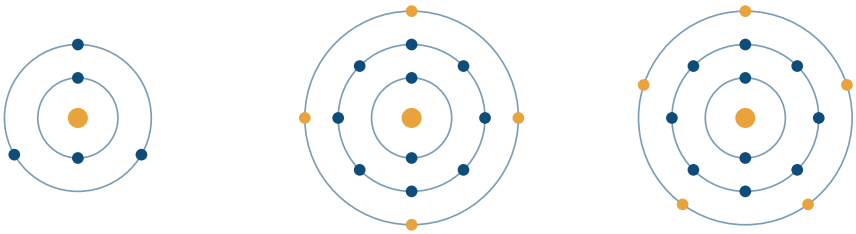
Dimensions

The diameter of the electron shell is 0.1–0.5 nm (0.0000000001 m, or 1 ångström); the diameter of the atomic nucleus is smaller by a further factor of 100,000. To illustrate this, imagine a pinhead at the center of a football field representing the atomic nucleus. The distance to the corner flags would then correspond to the distance at which the electrons orbit the nucleus.

Density

Within the nucleus of an atom, protons and neutrons are packed extremely tightly. If Earth were compressed to the same density, its radius would shrink from its original 6,700,000 m to just 100 m.

Atoms of major importance in the semiconductor industry



Boron
5p·6n·5e
3 valence electrons

Silicon
14p·14n·14e
4 valence electrons

Phosphorus
15p·15n·15e
5 valence electrons

The elements, the periodic table

The Elements

An element consists of many identical atoms (meaning atoms with the same number of protons, i.e. the same atomic number) and is a substance that cannot

be broken down further by chemical means. The mass of an element is determined solely by the number of protons and neutrons, since the mass of the electrons is negligibly small. Hydrogen, with one proton and no neutrons, has a mass number of 1; the next heavier element, helium, has a mass number of 4 (2 protons + 2 neutrons). The mass number gives the number of particles in the nucleus. Multiplying it by the atomic mass unit $u = 1.660 \cdot 10^{-27}$ kg gives approximately the mass of an atom – for helium, $6.64 \cdot 10^{-27}$ kg.

Chemical elements are usually named after the initial letters of their Latin or Greek names (hydrogen [H] from Latin hydrogenium, lithium [Li] from Greek lithos).

The Periodic Table of the Elements

The periodic table of the elements arranges all chemical elements in order of increasing proton number (nuclear charge, atomic number), grouped according to their chemical properties into periods as well as main and subgroups.

The period indicates the number of electron shells, while the main group indicates the number of electrons in the outermost shell (1 to 8 electrons). Groups 1 and 2, as well as 13–18, form the main groups; groups 3–12 form the subgroups.

The first element with one shell (period 1) and one outer electron (group 1) is hydrogen, H. The next element, helium, He, has only one electron shell just like hydrogen and is therefore also in period 1. Since the first shell is already completely filled with two electrons, helium is not in group 2 but in group 18 (the noble gas group).

To accommodate further electrons, a new shell must be started. Lithium is therefore found in group 1, period 2 (two electrons in the first shell, one valence electron in the second shell). A shell can hold a maximum of $2n^2$ electrons, where n stands for the period.

After the first two valence electrons in the outermost shell have been filled in groups 1 and 2, starting from the fourth period further inner shells are first completed with electrons before the outermost shell is fully filled with electrons across groups 13 to 18.

The Periodic Table of the Elements

Tap an element for details.

Important elements in semiconductor technology

Element	Particles	Note
B: Boron	5p, 6n, 5e	3 outer electrons: used for p-type doping of silicon

N: Nitrogen	7p, 7n, 7e	Stable N_2 molecule: protective and purge gas, protective layers on the wafer
O: Oxygen	8p, 8n, 8e	Highly reactive: oxidation of silicon, insulating layers (SiO_2), among others
F: Fluorine	9p, 10n, 9e	The most reactive element: used together with other substances for etching (e.g. HF, CF_4)
Si: Silicon	14p, 14n, 14e	Base material in semiconductor technology
P: Phosphorus	15p, 16n, 15e	5 outer electrons: used for n-type doping of silicon

Elements on the left side of the periodic table are metals. These tend to give up valence electrons in order to reach a stable electron configuration (the noble gas configuration). On the right side of the periodic table are the nonmetals, which take up additional electrons to reach the noble gas configuration. Between them are the metalloids, such as silicon and germanium.

Chemical bonds

Tendency to Form Bonds

Electrons in the outermost shell can detach from the atom (for example, by adding energy in the form of heat) and be exchanged with other atoms. Compounds formed from several elements are called molecules. The reason for this tendency to form bonds is to reach an outermost shell fully occupied with eight electrons: the so-called noble gas configuration. Substances that have already reached a full outer shell do not form compounds (a few exceptions exist, such as xenon–fluorine compounds).

A distinction is mainly made between three different types of bonds, which are explained in more detail below.

The Covalent Bond

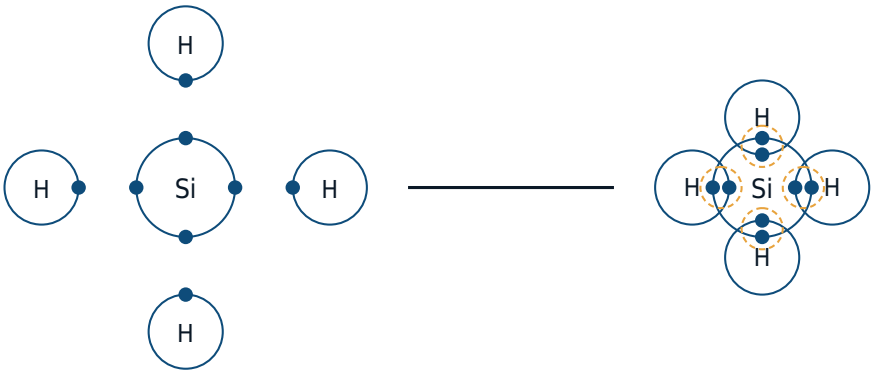
Nonmetals form this type of bond to complete their electron octet. For example, two fluorine atoms (each with seven outer electrons) can fill their electron octet by mutually sharing one electron. The distance between the two atomic nuclei represents a compromise between the attraction of the nuclei to the bonding electrons and the repulsion between the two nuclei. The reason for covalent bonds is nature's tendency to reach the lowest possible energy state. Because combining several atoms into a molecule gives the electrons "more room", which corresponds to a lower energy, covalent bonding occurs in the first place.

The tendency to reach a fully occupied outer shell means that elemental fluorine never occurs as single atoms, but always as fluorine molecules: F_2. The same is

true for nitrogen (N_2), oxygen (O_2), chlorine (Cl_2), bromine (Br_2) and iodine (I_2). Because of the shared electron pairs, this bond is also called an electron pair bond or covalent bond.

The following figure shows a covalent bond using silane (SiH_4) as an example (only the outermost shell of the silicon atom is shown). Through this bond, the silicon atom reaches a full outer shell of eight electrons, while the hydrogen atoms reach their first shell, already fully occupied with two electrons.

Covalent bond



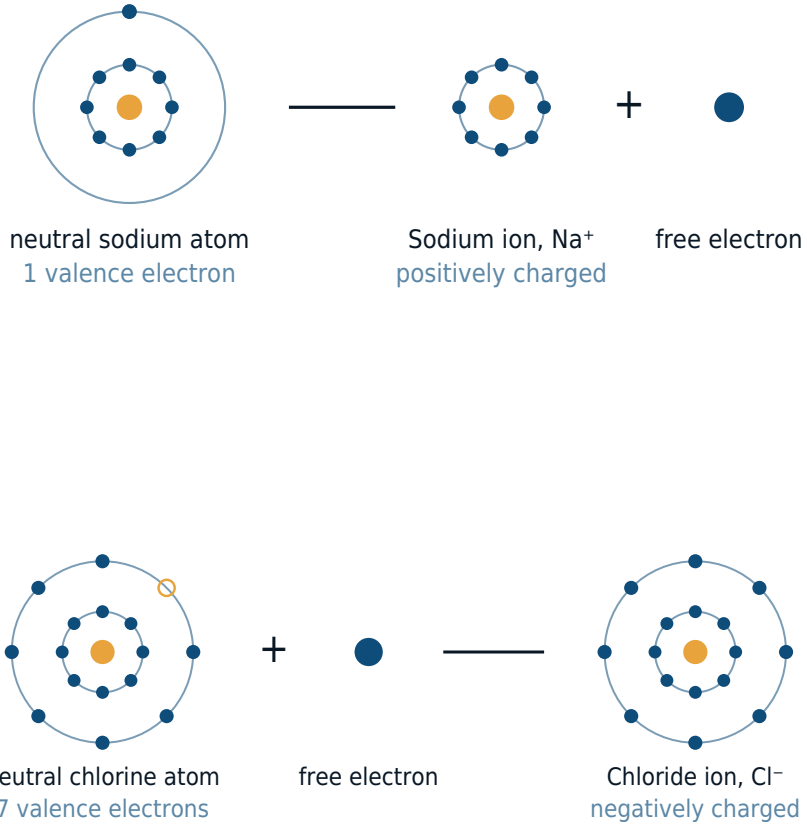
Four shared electron pairs complete the bond

The Ionic Bond

Ionic bonds form when metals combine with nonmetals. Metals tend to give up electrons in order to reach a completely filled outer shell, while nonmetals can take up additional electrons. An example of an ionic bond is sodium chloride (NaCl, table salt).

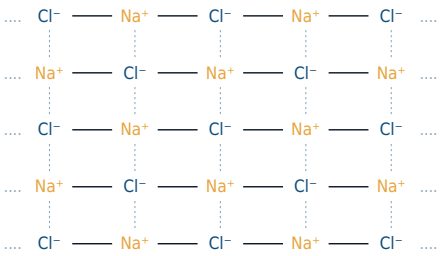
The sodium atom gives up its valence electron (leaving it with more protons than electrons, so it becomes positively charged), while chlorine takes up an electron and becomes singly negatively charged. The two atoms attract each other because of their opposite charges. A charged atom is called an ion; a distinction is made between cations (positive charge) and anions (negative charge).

Principle of the ionic bond



Since atoms always occur in very large numbers, the attractive and repulsive forces cause them to arrange themselves into a regular ionic lattice. Substances that form such a lattice in the solid state are called salts.

Crystal lattice of the ionic bond in sodium chloride



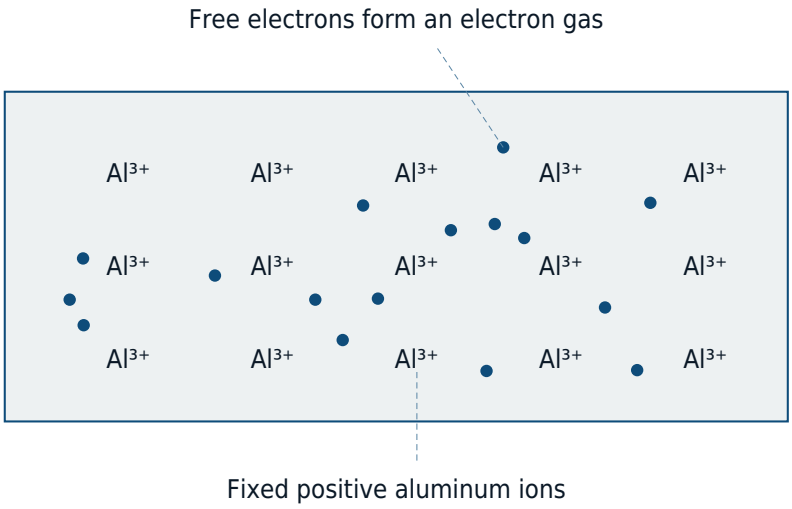
The Metallic Bond

Metals form this type of bond to reach the stable noble gas configuration. To do so, each metal atom gives up its outer electrons: positively charged metal ions

(atomic cores) are formed along with free electrons, between which strong attractive forces exist. The metal ions repel one another, just as the electrons repel one another.

Because the attractive and repulsive forces act in all directions in space, the atomic cores arrange themselves into a regular lattice. The freely moving electrons occupy the spaces in between as a so-called electron gas, which holds the positive metal ions together. Because of these freely moving electrons, metals conduct electric current very well.

Metallic bond



The physical and chemical properties of compounds depend on the type of bond. Stronger attractive forces, for instance, mean higher melting and boiling points, while the number of free electrons affects conductivity.

Weak Bonds

Weak bonds are not true bonds within molecules, but rather interactions between molecules that are usually caused by shifts in charge within a molecule. This creates positive and negative centers of charge, through which molecules attract one another.

The strongest of the weak bonds is the hydrogen bond; in the water molecule (H–O–H) it results from the bond angle between the oxygen and hydrogen atoms. This causes the electric charge within the molecule to be distributed asymmetrically (negative at the oxygen, positive at the hydrogen), so that neighboring molecules attract one another. There are also other bonds, such as the van der Waals bond, which are weaker still.

The bond energy of chemical bonds ranges from a few hundred to a few thousand kilojoules per mole (kJ/mol). For hydrogen bonds it is up to 100 kJ/mol, and for van der Waals forces it is 0.5–5 kJ/mol.

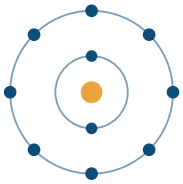
Noble gases

Noble Gases and the Electron Octet

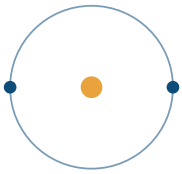
Noble gases are the elements in the eighth main group of the periodic table: helium, neon, argon, krypton, xenon, radon (from top to bottom).

What makes these elements special is that they have eight electrons in their outermost shell. This electron octet represents an energetically very stable state that all elements strive to reach. The chapter on bonds explains how elements reach a full outer shell of eight electrons. Because of the stable noble gas configuration, these elements undergo almost no reactions (a few xenon–fluorine compounds, among others, do exist).

Noble gases



Neon: 8 valence electrons
on the second shell



Helium: 2 valence electrons
on the first shell

The neon atom has 10 electrons: two in the first shell and eight valence electrons in the second, outermost shell. Exception: the helium atom, which has only one shell, already reaches the noble gas state with two electrons (electron duet).

As inert gases [Lat. inert: idle, sluggish], nitrogen (e.g. as a purge gas in furnace processes) and argon (e.g. in sputtering processes), among others, are used in semiconductor manufacturing.

Conductors – Insulators – Semiconductors

Conductors

Conductors are, in general, substances that have the property of transmitting different forms of energy. In the following, the focus is on electrical conductivity.

Metals

The conductivity of metals is based on the free electrons that exist as an electron gas within the metallic bond. Even a small amount of energy is enough to detach sufficient electrons from the atoms to achieve conductivity.

Metallic bond: fixed atomic cores and free valence electrons (electron gas)

Conductivity depends, among other things, on temperature. As temperature rises, the atomic cores vibrate more strongly, hindering the movement of the electrons and consequently increasing resistance. The best conductors, gold and silver, are used relatively rarely due to their high cost (gold, for example, is used to contact finished chips). The alternatives used in semiconductor technology for wiring the individual components of a chip are aluminium and copper.

Salts

In addition to metals, salts can also conduct electric current. Here, however, there are no free electrons. Conductivity is instead based on ions, which detach from the lattice structure and become freely mobile when a salt melts or dissolves (see the chapter on bonds).

Insulators

Insulators have no free charge carriers in the form of electrons or ions. Insulators are also referred to as non-conductors.

Covalent bond

The covalent bond is based on shared electron pairs between nonmetals. Elements with nonmetallic character all tend to take up electrons, meaning no free electrons are available that could enable conductivity.

Ionic bond

In the solid state, ions are arranged within a lattice structure. The particles are held together by electrical forces. There are no free charge carriers available for current flow. Consequently, substances composed of ions can be either conductors (when dissolved) or insulators.

Semiconductors

Semiconductors are solids whose conductivity lies between that of conductors and insulators. Through the exchange of electrons between atoms of the same kind, in order to complete the electron octet, these atoms arrange themselves into a lattice structure. Unlike metals, conductivity increases with rising temperature – up to a certain point.

As the temperature rises, bonds break and electrons are released. At the location where the electron used to be, a so-called defect electron (also called a hole) remains.

Section of a silicon crystal

The flow of electrons is based on the intrinsic conductivity of semiconductors. The so-called band model illustrates why semiconductors behave this way.

The Band Model

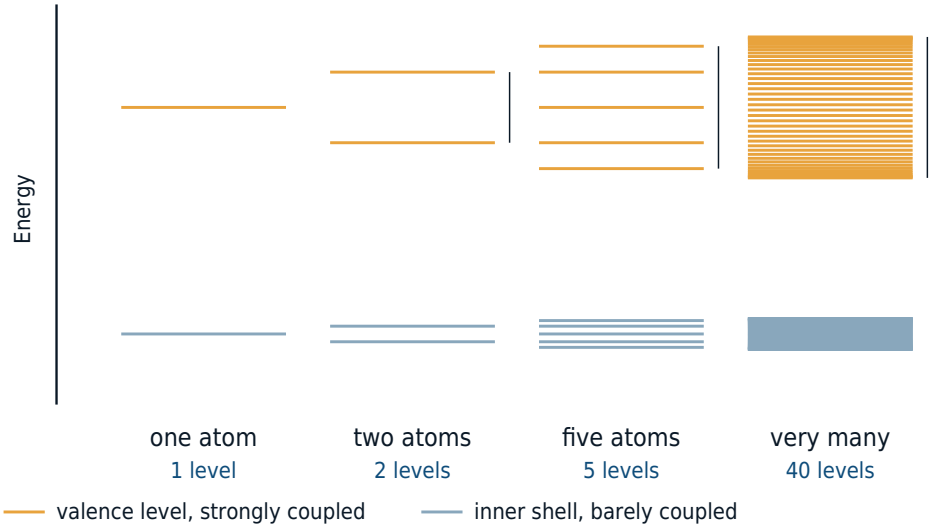
The band model is an energy diagram used to describe the conductivity of conductors, insulators, and semiconductors. The model consists of two energy bands (the valence band and the conduction band) and the band gap. The valence electrons – which serve as charge carriers – are located in the valence band; in the ground state, the conduction band is not occupied by electrons. Between the two energy bands lies the band gap, whose width, among other things, influences conductivity.

The Energy Bands

Looking at a single atom, according to the Bohr atomic model there are sharply separated energy levels that can be occupied by electrons. When several atoms are located next to one another, they interact with each other and the discrete energy levels fan out. In a silicon crystal there are approximately 5×10^{22} atoms per cubic centimetre, so that the individual energy levels are no longer distinguishable from one another and instead form broad energy ranges.

Splitting of Energy Levels

The more atoms interact, the more closely spaced the levels become

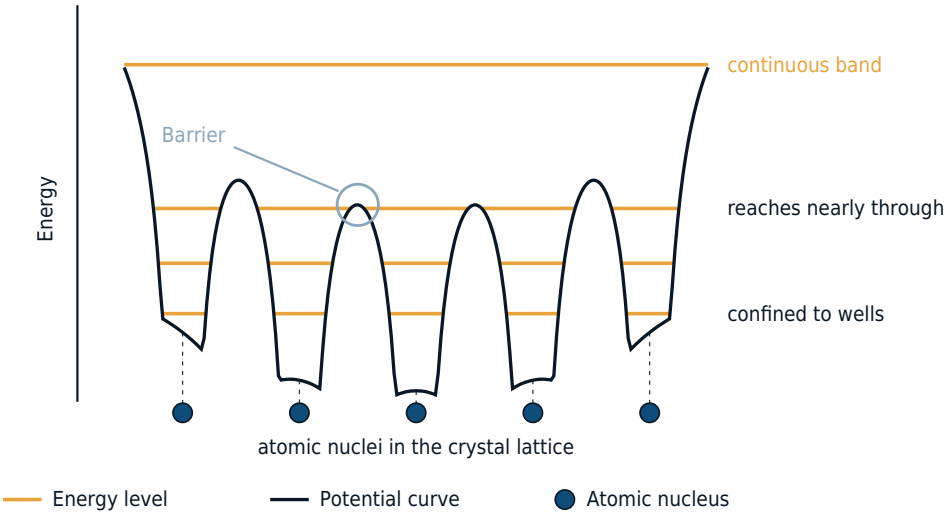


Each additional atom adds one more level. The overall width barely grows any further, so the spacing keeps shrinking – at 10^{22} atoms per cubic centimeter, the levels are no longer distinguishable.

The width of the energy bands depends on how strongly the electrons are bound to the atom. The valence electrons at the highest energy level interact strongly with those of neighbouring atoms and can be detached from the atom relatively easily; with a very large number of atoms, a single electron can no longer be assigned to a specific atom. As a result, the energy bands of the individual atoms merge into a continuous band, the valence band.

Energy Bands from Interacting Atoms

The potential wells of neighboring atoms overlap



Deeply bound electrons stay trapped in their well – their levels remain separate. The valence electrons, by contrast, sit above the barriers between the nuclei and no longer belong to a single atom: a band forms.

The Band Model in Conductors

In conductors, the valence band is either not completely filled with electrons, or the filled valence band overlaps with the empty conduction band. As a rule, both conditions apply simultaneously, so electrons can move either within the partially filled valence band or within the two overlapping bands. In conductors, there is no band gap between the valence band and the conduction band.

The Band Model in Insulators

In insulators, the valence band is completely filled with electrons due to the bonds between atoms. The electrons cannot move because they are "locked in" between the atoms. To achieve conductivity, electrons would have to move from the fully occupied valence band into the conduction band. This is prevented by the band gap that lies between the valence band and the conduction band.

This gap can only be overcome with a very large amount of energy (if at all) – according to the laws of quantum physics, no electron is allowed to reside within the band gap itself.

The Band Model in Semiconductors

Semiconductors also have this band gap, but compared to insulators it is so small that, even at room temperature, electrons pass from the valence band into

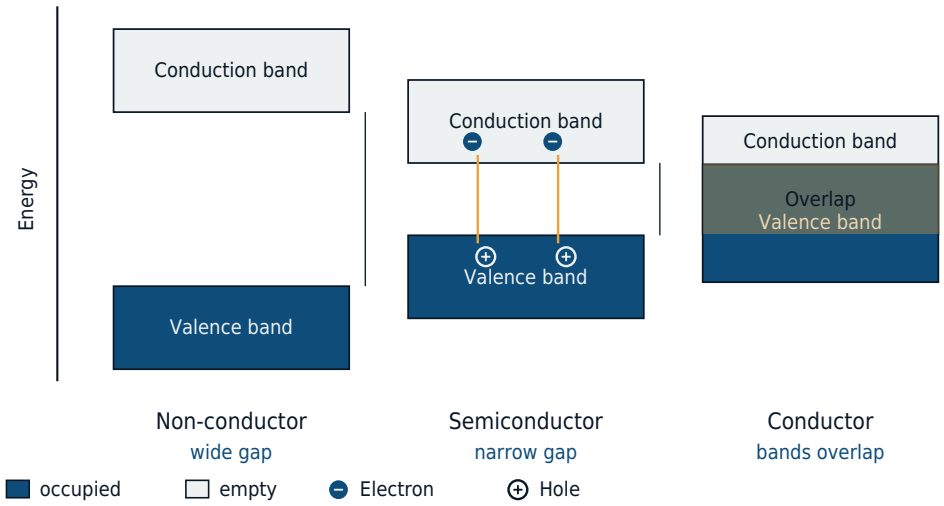
the conduction band. These electrons can then move freely and act as charge carriers. In addition, each electron leaves behind a hole in the valence band, which can be filled by other electrons in the valence band. This results in migrating holes in the valence band, which can be regarded as positive charge carriers.

Electron–hole pairs always occur together, meaning there are just as many negative as positive charges, and the semiconductor crystal as a whole remains neutral. A pure, undoped semiconductor is referred to as an intrinsic semiconductor; there are approximately 10^{10} free electrons and holes per cubic centimetre (at room temperature).

Since electrons always adopt the energetically most favourable state, they fall back into the valence band without any energy supply and recombine with holes. At a given temperature, an equilibrium is established between electrons being raised into the conduction band and electrons falling back. As temperature increases, the number of electrons able to cross the band gap increases as well. Thus, the conductivity of semiconductors increases with rising temperature.

The Band Model

The width of the band gap determines conductivity



In a semiconductor, the heat available at room temperature is already enough to raise individual electrons across the gap. Each leaves behind a hole in the valence band – both contribute to the current.

Since the width of the band gap corresponds to a specific energy and therefore a specific wavelength, attempts are made to specifically alter the band gap in order to obtain certain colours in light-emitting diodes. This can be achieved, among other methods, by combining different materials. Gallium arsenide (GaAs) has a band gap of 1.4 electron volts (eV, at room temperature) and therefore emits red light.

The intrinsic conductivity of silicon is of little interest for the functioning of devices, since it depends very strongly on the amount of energy supplied. It therefore also changes with operating temperature, and a conductivity comparable to that of metals is only reached at very high temperatures (several hundred degrees Celsius). In order to specifically influence the conductivity of semiconductors, foreign atoms can be incorporated into the regular silicon lattice, thereby adjusting the concentration of charge carriers – electrons and holes. This process is called *doping*.

Doping: n- and p-semiconductors

Doping

Doping means introducing foreign atoms into a semiconductor crystal in order to specifically alter its conductivity. Two of the most important substances used to dope silicon are boron (3 valence electrons = trivalent) and phosphorus (5 valence electrons = pentavalent). Others include aluminium and indium (trivalent), as well as arsenic and antimony (pentavalent).

The dopant elements are incorporated into the lattice structure of the semiconductor crystal; the number of outer electrons determines the type of doping. Elements with 3 valence electrons are used for p-type doping, while pentavalent elements are used for n-type doping. The conductivity of a deliberately impurified silicon crystal can be increased in this way by a factor of 10^6.

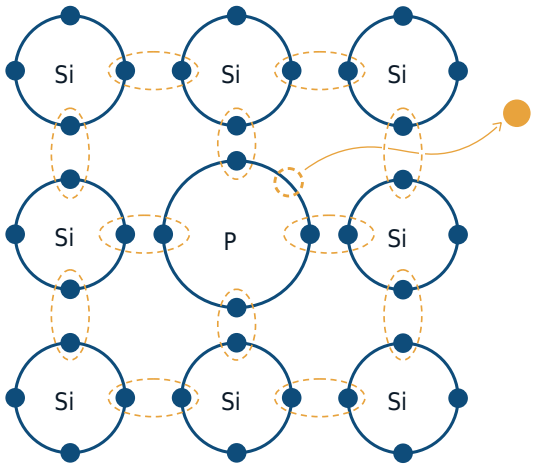
n-Type Doping

The pentavalent dopant element has one more outer electron than the silicon atoms. Four of its outer electrons can bond with a silicon atom each, while the fifth remains freely mobile and serves as a charge carrier. This unbound electron requires far less energy to be raised from the valence band into the conduction band than the electrons responsible for the intrinsic conductivity of silicon. The dopant element that releases an electron is referred to as an electron donor (from Latin donare = to give).

By releasing negative charge carriers, the dopant elements become positively charged and remain fixed in the lattice; only the electrons are able to move. Doped semiconductors whose conductivity is based on free (negative) electrons are called n-conducting or n-doped. While holes (along with an equal number of electrons) can be spontaneously generated within the crystal at any time, the number of free electrons introduced by the donors now predominates, which is

why these are referred to as majority charge carriers. Holes, on the other hand, are called minority charge carriers.

n-type doping with phosphorus



Four bonding electrons, one electron as a free charge carrier

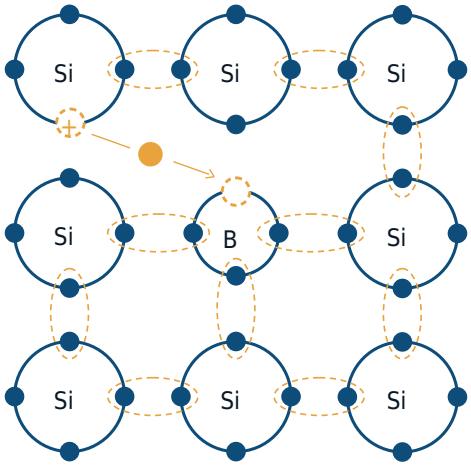
Arsenic is used as an alternative to phosphorus, since its diffusion coefficient is lower. This means that the dopant diffuses less during subsequent process steps, so the doping remains in the location where it was originally introduced.

p-Type Doping

In contrast to the free electron produced by doping with phosphorus, trivalent dopant elements have exactly the opposite effect. They can accept an additional outer electron, thereby leaving a hole behind in the valence band of the silicon atoms. As a result, the electrons in the valence band become mobile. The holes move in the direction opposite to the movement of the electrons. When indium is used as the dopant element, the energy required for this is only 1 % of the energy needed to raise the valence electrons of silicon atoms into the conduction band.

By accepting an electron, the dopant element becomes singly negatively charged; such dopant atoms are called electron acceptors (from Latin acceptare = to accept). Here too, the dopant element remains fixed within the crystal lattice; only the positive charge moves. These semiconductors are referred to as p-conducting or p-doped because their conductivity is based on positive holes. Analogous to n-doped semiconductors, holes are the majority charge carriers here, while free electrons are the minority charge carriers.

p-type doping with boron



Four bonding electrons, one electron as a free charge carrier

Externally, doped semiconductors are electrically neutral. The terms n-type and p-type doping refer only to the majority charge carriers.

N-doped and p-doped semiconductors behave in an approximately similar way with respect to current flow. As the number of dopant elements increases, so does the number of charge carriers in the semiconductor crystal. Even a very small amount of dopant elements is sufficient for this. Weakly doped silicon crystals contain only 1 foreign atom per 1,000,000,000 silicon atoms, whereas at the highest doping levels the ratio of foreign atoms to silicon atoms is, for example, 1 to 1,000.

Band Diagram of Doped Semiconductors

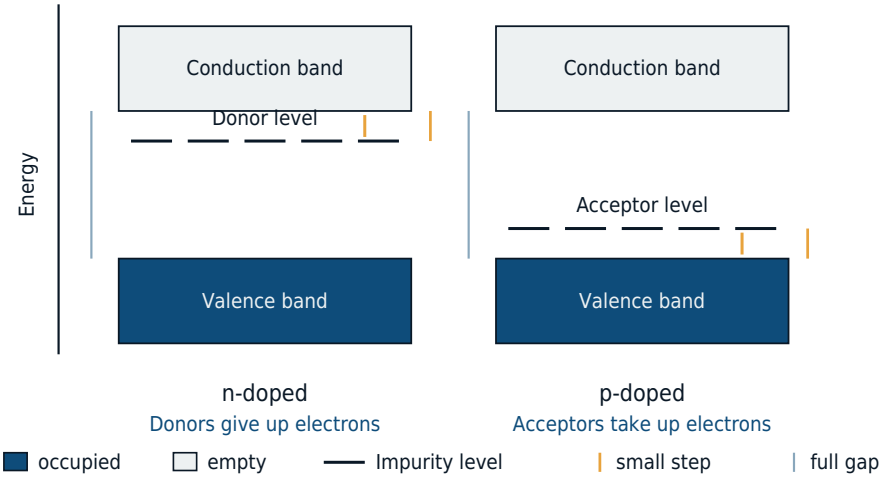
In n-doped semiconductors, introducing a dopant element with five outer electrons makes an unbound electron available within the crystal, which can therefore be raised into the conduction band with comparatively little energy. As a result, n-doped semiconductors exhibit a donor level close to the conduction band edge, and the band gap that needs to be overcome is very small.

By introducing a trivalent dopant element, p-doped semiconductors provide a vacant site that can already be filled by an electron from the valence band with only a small amount of energy. In p-doped semiconductors, an acceptor level is therefore located close to the valence band edge.

Band diagram of doped semiconductors

Band Diagram for Doped Semiconductors

The impurity level sits close to the band edge – the step becomes small



Without doping, an electron would have to cross the entire band gap. With impurities, only a fraction of that is needed – which is why conductivity rises markedly already at room temperature.

The p-n junction

The p-n Junction in Equilibrium

The p-n junction is the transition region where n-doped and p-doped semiconductor crystals meet. In this region there are no free charge carriers, because the free electrons of the n-conductor and the free holes of the p-doped crystal recombine with one another near the contact surface of the two crystals, i.e. the electrons fill the vacant holes. This movement of charge carriers (diffusion) occurs as a result of a concentration gradient: since there are only a few electrons in the p-crystal and only a few holes in the n-crystal, the majority charge carriers (electrons in the n-region, holes in the p-region) migrate into the oppositely doped semiconductor crystal. The crystal lattice at the interface must not be interrupted; simply "pressing together" a p-doped and an n-doped silicon crystal does not result in a functional p-n junction.

Due to the migrated free charge carriers, the regions near the interface become positively charged (n-crystal) or negatively charged (p-crystal). The more charge carriers recombine, the larger this depletion or space-charge region (SCR) becomes, and with it the voltage difference between the n-crystal and the p-crystal. At a certain level of this potential difference, the recombination of holes and electrons comes to a stop, since the charge carriers can no longer overcome the electric field. In silicon, this limit is approximately 0.7 V (see band model of

a p-n junction).

p-n junction with no applied voltage

A p-n junction corresponds to an electrical component that, when a voltage is applied, conducts current in one direction (forward direction) and blocks it in the other direction (reverse direction): a diode.

The p-n Junction with an Applied Voltage

If a positive voltage is applied to the n-crystal and a negative voltage to the p-crystal, the internal electric field and the field generated by the voltage source point in the same direction. The field at the p-n junction is thereby reinforced. The oppositely charged free charge carriers are attracted by the poles of the voltage source, which enlarges the depletion layer and prevents any current from flowing.

If the applied voltage across the semiconductor crystals is reversed, the electric field generated by the voltage source opposes the internal field and weakens it. Once the internal field has been completely cancelled out by the external field, new charge carriers continuously flow from the voltage source to the depletion layer and are able to recombine there without interruption: current flows.

p-n junction with an applied voltage

Because of this behaviour, the diode can be used as a rectifier: to convert alternating current into direct current. Regions where p-doped and n-doped semiconductor crystals are in contact with one another occur in many electronic components within semiconductor technology.

Transistors and Memory

MOSFET

A transistor is an electronic semiconductor device used to switch or amplify current. In a field-effect transistor (FET), a gate electrode controls the current flow between the source and drain terminals without being part of the current path itself. The most common design is the MOSFET: with no voltage applied to the gate, the p-doped substrate separates source and drain like a blocked p-n junction; the transistor is off. Only a positive gate voltage attracts electrons to the surface and forms a conducting n-channel there – the same mechanism already described under doping, here made electrically controllable. How the

MOSFET is actually fabricated is covered in the chapter Construction of a n-channel FET.

Bipolar Transistor

The second fundamental transistor type is the bipolar transistor. It consists of two p-n junctions connected back to back, with the layer sequence n-p-n or p-n-p; its terminals are called emitter, base and collector. Unlike the MOSFET, both types of charge carriers – electrons and holes – contribute to the current flow here, hence "bipolar". A small voltage at the thin base layer opens the junction to the emitter and allows a much larger current to flow from emitter to collector. Bipolar transistors are faster than MOSFETs but require more area. Details are covered in the chapter Construction of an NPN bipolar transistor.

FinFET

As feature sizes shrink, a flat, planar gate eventually can no longer reliably shut off the channel. In a FinFET, the channel is instead built as a thin, upright silicon "fin" that the gate wraps around on three sides – source, drain and the basic switching function remain identical to the MOSFET. This multi-sided access reduces leakage currents and allows the threshold voltage to be defined more sharply. How a FinFET is built and fabricated in detail is described in the chapter Construction of a FinFET.

DRAM Cell

Transistors are used not only for switching but also for storing data. The DRAM cell combines a single transistor with a capacitor for this purpose (1T1C cell): the capacitor stores one bit as an electric charge, and the transistor connects it to the bitline via the wordline when reading or writing. Because the charge slowly leaks away, it must be refreshed every few milliseconds – hence "dynamic". How the capacitor and access transistor are fabricated is shown in the chapter DRAM Fabrication: The Trench Capacitor Process.

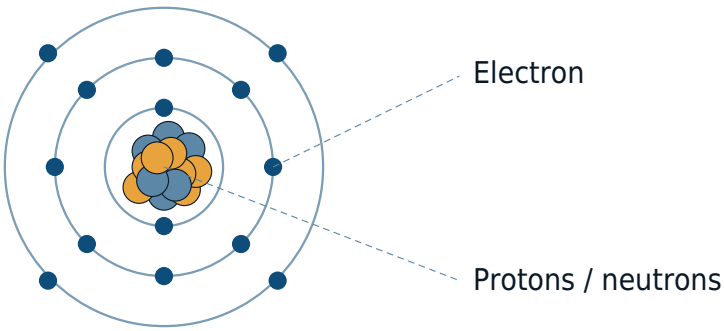
Silicon

Properties of Silicon

Silicon is the chemical element with atomic number 14 in the periodic table; it belongs to the 4th main group and the 3rd period. Silicon is a classic semiconductor, meaning its conductivity lies between that of conductors and insulators. In nature, silicon (from Latin silex/silicis: pebble) occurs exclusively as an oxide: as silicon dioxide (SiO_2) in the form of sand and quartz, or as silicate (compounds of silicon with oxygen, metals, and others). Silicon is therefore, quite literally, as abundant as sand on a beach, making it an extremely inexpensive starting material whose value is only determined through processing. Other semiconductors, such as germanium or the compound semiconductor gallium arsenide, offer in some respects substantially better electrical properties than silicon: charge carrier mobility, and the resulting switching speeds, are significantly higher in germanium and GaAs. However, silicon has decisive advantages over other semiconductors.

This dominant position applies to integrated circuits, not to semiconductor technology as a whole. In power electronics, silicon carbide and gallium nitride have secured a firm place for themselves, since they can withstand higher voltages, higher temperatures, and faster switching. And for anything intended to emit light, silicon is unsuitable: its band transition releases energy as heat rather than as light, which is why light-emitting diodes and lasers are made from compound semiconductors instead.

Bohr atomic model of silicon



Oxide layers can be produced very easily on a silicon crystal; the resulting silicon dioxide is a high-quality insulator that can be deposited on the substrate in a targeted manner. For the semiconductors mentioned above, germanium and GaAs, producing similarly good insulating layers is, by contrast, very costly. The ability to precisely alter the conductivity of pure silicon through doping is another reason this metalloid is so significant. Other materials are, in some cases, highly toxic, and compounds formed with these elements are not as durable and stable as those based on silicon. A prerequisite for using silicon in semiconductor manufacturing, however, is that it be present in an ultra-pure, single-crystal form. This means that the silicon atoms in the crystal lattice are arranged in a perfectly regular pattern, with no undefined foreign atoms present in the crystal.

In addition to the single-crystal form, there is also polysilicon (poly = many) and amorphous silicon (a-Si). While single-crystal silicon, in the form of wafers, is the foundation of microelectronics, polycrystalline silicon is deposited as a layer on the wafer in various areas to fulfil specific tasks (e.g. as a masking layer, as the gate in a transistor, and others). It can be produced easily and patterned with ease. Polysilicon is made up of many individual single crystals arranged irregularly with respect to one another. Amorphous silicon has no regular lattice structure at all, but rather a disordered one. In chip manufacturing it plays only a minor role; outside of it, however, this is by no means the case: because it can be deposited over large areas and at low temperatures onto glass, the switching elements behind every flat-panel display are built from it, and it is also used in thin-film solar cells.

Production of raw silicon

Starting Material: Silicon Dioxide

The silicon used in semiconductor manufacturing is obtained from quartz. In this process, oxygen — which combines with silicon extremely readily even at room temperature, and which in quartz is likewise present in combination with silicon as silicon dioxide — must be removed. This is done in an arc furnace using carbon, at temperatures that reach up to about 2000 °C in the hottest zone of the furnace. The oxygen separates from the silicon and reacts with the carbon (C) to form carbon monoxide (CO):

$SiO_2 + 2C \rightarrow Si + 2CO$

This equation summarizes what actually takes place in the furnace over several stages. In the cooler zone, silicon carbide initially forms; this then reacts with

additional silicon dioxide in the hotter zone to produce silicon. Silicon carbide is therefore not an unwanted by-product but a necessary intermediate product; the ratio of carbon to quartz is set so that it is fully consumed again by the end of the process. At these temperatures, the carbon monoxide is gaseous and can easily be separated from the liquid silicon.

The raw silicon obtained in this way (metallurgical-grade silicon) is approximately 98 to 99 % pure. The remainder consists of impurities such as iron, aluminium, calcium, phosphorus, and boron. For use in semiconductor devices, this is far too impure by many orders of magnitude — these substances must be removed in further processes.

Purification of the Raw Silicon

Purification is carried out indirectly via a gaseous compound: the silicon is converted into trichlorosilane, which is purified by distillation, and solid silicon is then deposited from it again. The process is known as the Siemens process, named after its developer, and to this day still supplies the majority of the world's ultra-pure silicon.

Using the trichlorosilane process, many impurities are removed by distillation. The raw silicon reacts with hydrogen chloride (HCl) at around 300 °C to form gaseous trichlorosilane ($SiHCl_3$) and hydrogen (H_2):

$$Si + 3HCl \rightarrow SiHCl_3 + H_2$$

The impurities, which also combine with the chlorine present, only pass into the gaseous state at higher temperatures. This allows the trichlorosilane to be separated out. Only carbon, phosphorus, and boron, which have similar condensation temperatures, cannot be filtered out by this process.

The trichlorosilane process can be reversed, so that the purified silicon is subsequently obtained in polycrystalline form. This takes place with the addition of hydrogen in a quartz vessel containing heated, thin silicon rods (silicon seed rods), at around 1100 °C:

$$SiHCl_3 + H_2 \rightarrow Si + 3HCl$$

and

$$4SiHCl_3 \rightarrow Si + 3SiCl_4 + 2H_2$$

The silicon deposits onto the silicon seed rods, which grow over several days into rods roughly 150 to 200 mm thick. These are subsequently broken up and processed further as chunks.

The process is extremely energy-intensive, since the rods must be kept at temperature the entire time while the chamber wall is cooled. As an alternative, fluidized-bed deposition has become established: here, small silicon granules grow within a bed through which gas flows, which requires significantly less

energy and operates continuously rather than in batches. However, its purity does not match that of the Siemens process, which is why this material is used primarily in the solar industry.

This is precisely where an important distinction lies: silicon suitable for solar applications must reach roughly six nines of purity (99.9999 %), whereas silicon suitable for electronics must reach nine to eleven nines. Using the Czochralski process, this polysilicon could already be converted into a single crystal; however, its degree of purity is still not high enough for the fabrication of semiconductor devices.

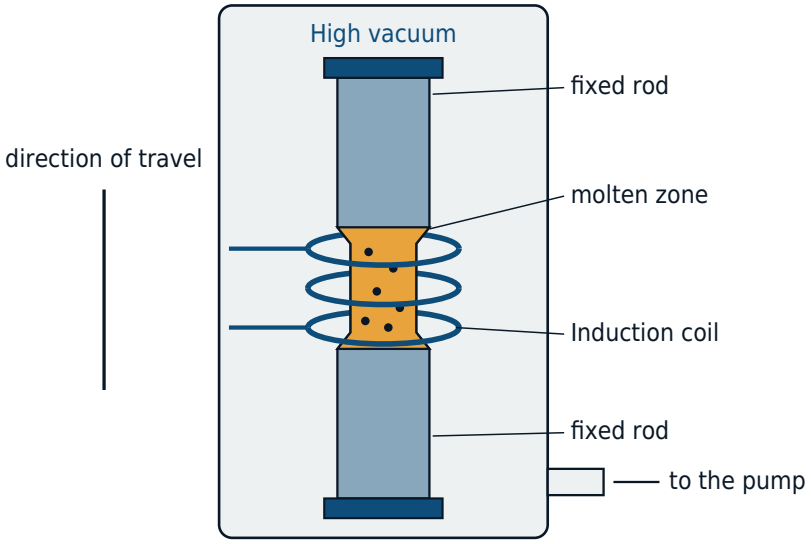
Zone Refining

To further increase purity, zone refining is used. A coil carrying a high-frequency alternating current is placed around the silicon rods. This melts the silicon inside the rods, and the impurities sink downward due to their high solubility in the melt (the surface tension of the silicon prevents the melt from flowing out). By repeating this process multiple times, the level of impurities in the silicon is reduced to the point where it can be further processed into a single crystal.

Illustration of zone refining

Zone Refining of a Silicon Rod

Surface tension holds the melt between the fixed ends



Impurities remain in the melt and travel along with the zone.

All processes are carried out under vacuum to prevent further contamination.

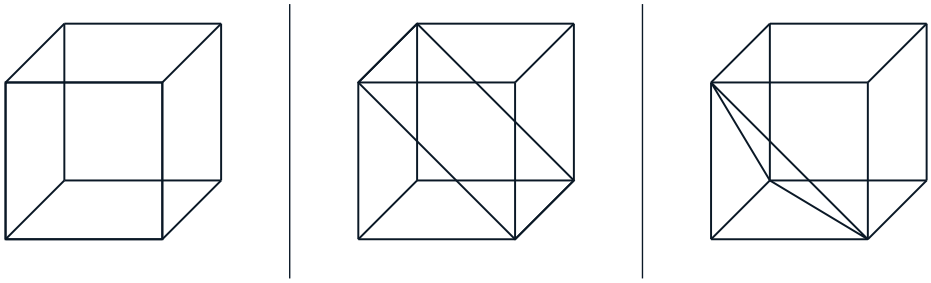
At the end of these processes, the silicon prepared for semiconductor manufacturing has a purity of more than 99.9999999 %, corresponding to less than one foreign atom per 1 billion silicon atoms.

Fabrication of the single crystal

The Single Crystal

A single crystal (monocrystal), as required in semiconductor manufacturing, is a regular arrangement of atoms. In addition, there is polycrystalline silicon (composed of many small single-crystal grains) and amorphous silicon (disordered structure). Depending on the orientation of the lattice in space, silicon wafers exhibit different surface structures, which influence various device parameters, such as charge carrier mobility.

Crystal orientations



Chip manufacturing uses almost exclusively (100)-oriented silicon. The reason lies at the interface with the gate oxide: unsaturated bonds remain there after oxidation, and their number is lowest for this orientation. The (111) orientation has the densest arrangement of bonds and therefore the most defects; however, it oxidizes the fastest.

Crystal orientation is also of particular importance in micromechanics. For instance, microchannels with vertical walls can be produced on (110) silicon, whereas (100) orientation results in sidewall angles of 54.74°.

Czochralski Crystal Growing Process

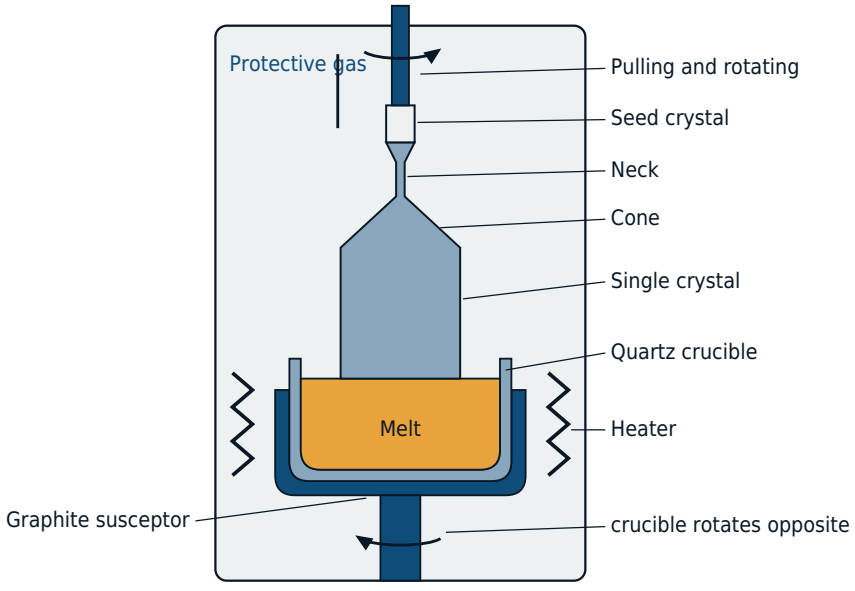
The polycrystalline silicon obtained after zone refining is melted in a quartz crucible just above the melting point of silicon. Dopants (e.g. boron or phosphorus) can now be added to the melt in order to achieve the desired electrical properties of the single crystal.

A seed crystal (a single crystal) mounted on a rotating rod is brought into contact with the surface of the silicon melt. This seed determines the orientation of the crystal. Upon contact between the seed and the melt, silicon deposits onto the seed and adopts its crystal structure. Because the crucible temperature is only slightly above the melting point of silicon, the deposited silicon solidifies immediately on the seed, and the crystal grows.

The seed is slowly pulled upward while being continuously rotated, maintaining constant contact with the melt. The crucible rotates in the opposite direction to the seed crystal. Constant temperature control of the melt is essential to ensure uniform growth. The diameter of the single crystal is determined by the pulling speed, which ranges from 2 to 25 cm/h. The faster the pulling speed, the thinner the crystal becomes. The entire apparatus is enclosed in a protective gas atmosphere to prevent oxidation of the silicon.

Czochralski Crystal Growth

The seed crystal sets the orientation of the crystal



Diameter is set via pulling speed and melt temperature.

A significant disadvantage of this process is that the melt becomes increasingly enriched with dopants during the process, since the dopants dissolve more readily in the melt than in the solid. As a result, the dopant concentration is not

constant along the length of the silicon rod. In addition, impurities or metals can dissolve out of the crucible and become incorporated into the crystal lattice.

The advantages of this process are its low cost and the ability to produce larger wafers than is possible with the float-zone process (see below).

Crucible-Free Float-Zone Growing

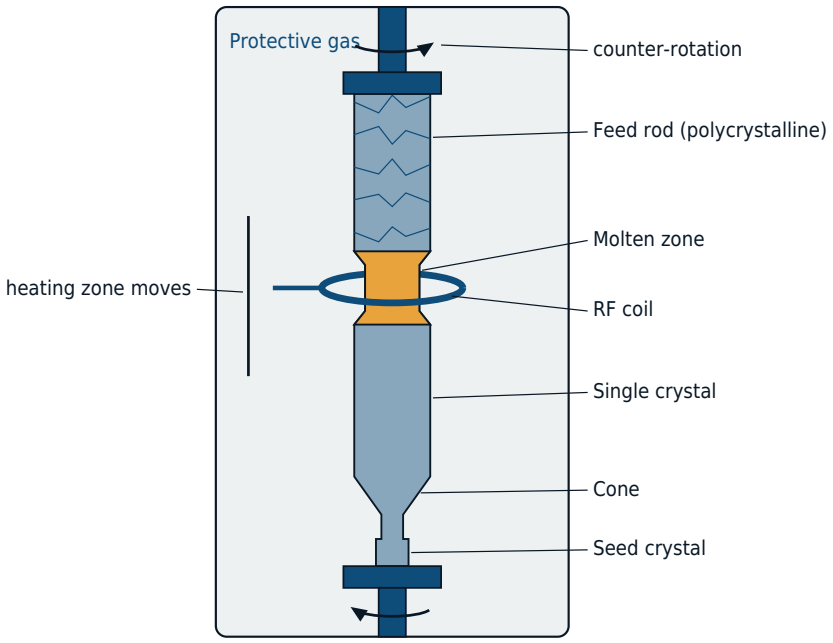
In contrast to the Czochralski crystal growing process, crucible-free float-zone growing does not melt the entire polysilicon rod; instead, as with zone refining, only a small region (a few millimetres) is melted.

Here too, a seed crystal, brought into contact with the end of the polycrystalline silicon rod, determines the crystal structure. The polycrystal is melted and adopts the structure of the seed. The heating zone is slowly moved along the rod, and the polycrystalline silicon rod gradually transforms into a single crystal.

Since only a small region of the polycrystalline silicon is melted, very few impurities can accumulate (as with zone refining, the foreign atoms migrate to the end of the silicon rod). Doping is achieved here by adding dopants (e.g. diborane or phosphine) to the protective gas that flows around the apparatus.

Float-Zone Process

Only a few millimeters of the rod are molten

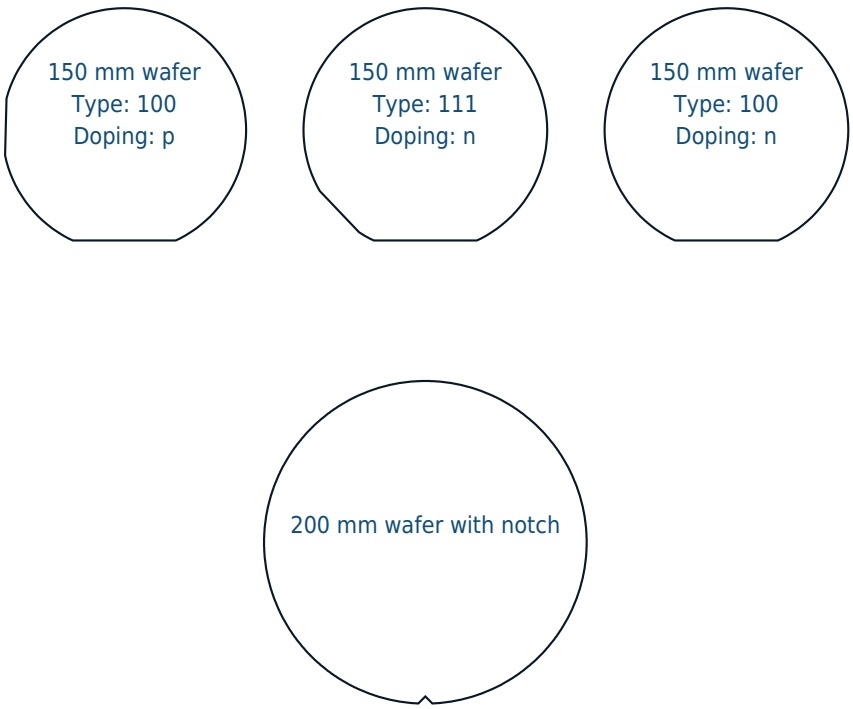


No crucible means no contact with foreign material; doping is introduced via the protective gas.

The wafer

Wafer Slicing and Surface Finishing

The single-crystal rod is first ground down to the desired diameter and then, depending on crystal orientation and doping, given one or two flats. The larger flat is used to precisely align the wafers during production. The second flat identifies the wafer type (crystal orientation, p-/n-doping), but is not always present. For wafers 200 mm in diameter and larger, so-called notches are used instead of flats. These are tiny notches at the edge of the wafer that likewise allow the wafer to be aligned, but take up far less of the wafer's valuable surface area.



Sawing

Historically, the single-crystal rod was cut using an internal diameter (ID) saw, whose cutting edge is coated with diamond fragments. It cuts precisely, but only one wafer at a time, and up to 20 % of the crystal rod is lost to the thickness of the saw blade. Given today's rod diameters and wafer prices, this is no longer acceptable; the standard method is therefore wire sawing, in which several hundred wafers are cut from the rod in a single pass. A long wire, wetted with a

slurry of silicon carbide grains and a carrier fluid such as glycol or oil, is guided over rotating rollers. The silicon crystal is lowered into the wire grid and thereby sliced into wafers. The wire moves back and forth at about 10 m/s and is typically 0.1–0.2 mm thick. Increasingly, instead of a slurry of loose grains, a wire with diamond grains fixed firmly into its surface is used. It cuts faster, produces less waste, and eliminates the need to dispose of spent slurry.

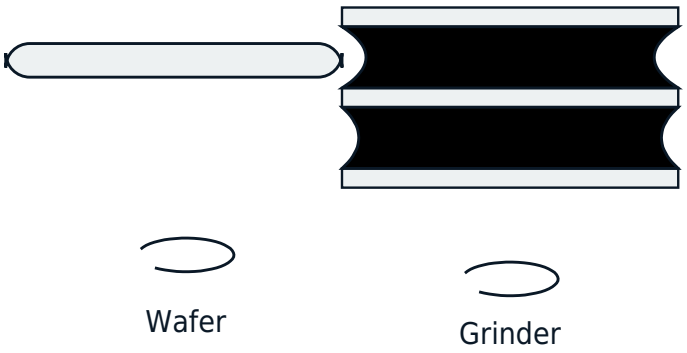
After sawing, the wafers have a roughened surface and, due to the mechanical stress, lattice damage within the crystal. To refine the surface, the wafers pass through several process steps.

Lapping

Using granular abrasives (e.g. aluminium oxide) on a rotating steel plate, 50 µm (0.05 mm) of the wafer surface is removed. The grain size is reduced in stages, but the surface is once again damaged by the mechanical treatment. The flatness after lapping is about 2 µm.

Rounding the wafer edge

In later processes, the wafers must not have any sharp edges, since deposited layers could otherwise flake off. For this reason, the edge of the wafers is rounded off using a diamond grinding tool.



Etching

In a dip-etch step using a mixture of hydrofluoric, acetic, and nitric acid, a further 50 µm is removed. Since this is a chemical process, the surface is not damaged. Crystal defects are finally eliminated.

Polishing

This is the final step toward the finished wafer. At the end of the polishing step,

the wafers have a remaining unevenness of less than 3 nm (0.000003 mm). For this, the wafers are treated with a mixture of sodium hydroxide solution (NaOH), water, and silicon dioxide particles. The silicon dioxide removes a further 5 µm from the wafer surface, while the sodium hydroxide removes oxide and eliminates processing marks left by the silicon dioxide particles.

This is followed by a final cleaning and a measurement of each individual wafer for flatness, thickness, and particles. A considerable proportion of wafers subsequently also receive an epitaxial layer: a thin, especially pure, and precisely doped layer of silicon grown onto the polished surface. The actual devices are then formed in this layer, while the wafer beneath serves only as a carrier.

Historical Development of Wafer Size

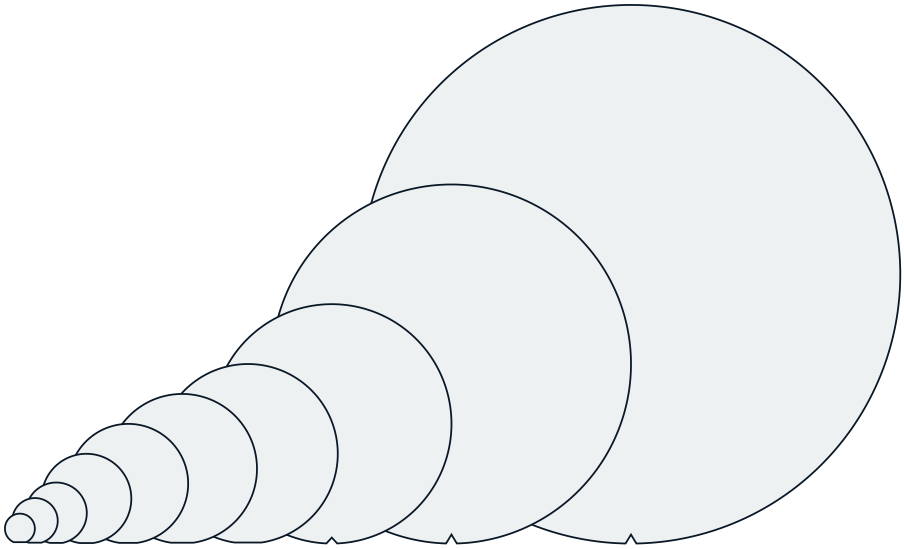
The manufacture of integrated circuits on silicon wafers began in the mid-1960s using wafers with a diameter of 25 mm. Since then, the diameter has grown in steps; since the late 1990s, the 300 mm wafer has been the largest format used in volume production. Its area is roughly 140 times that of the first 25 mm wafers.

With larger wafers, throughput in chip manufacturing increases considerably, allowing manufacturing costs to be reduced accordingly. For example, at the same feature size, more than twice as many chips can be produced on a 300 mm wafer as on a 200 mm wafer.

Typical wafer data:

Type [mm]	Diameter [mm]	Thickness [µm]	Primary flat [mm]	Bow [µm]
150	150 ± 0.5	~700	55 - 60	25
200	200 ± 0.5	~800	*Notch*	35
300	300 ± 0.5	~900	*Notch*	45

Overview of wafer sizes: 25, 38, 51, 75, 100, 125, 150, 200, 300, 450 [mm] (to scale)



The Step to 450 mm That Never Happened

The progression was supposed to continue with 450 mm. Around 2008, several major manufacturers and equipment suppliers joined forces in a shared program; initial equipment and test wafers were produced, and the transition was announced for the early 2010s. It never happened: the program was shut down in the mid-2010s, and 300 mm has remained the largest format to this day.

The reasons for this already lie in the technical difficulties that were being tackled at the time:

- Bow increases with diameter. To prevent stacked wafers from touching each other during transport, spacings, supports, and grippers would have had to be redesigned.
- Film stresses deform a large wafer more easily. It cannot simply be made thicker, since material costs and natural frequency work against this.
- The larger area also has to be processed. Compared with 300 mm, the cost per square centimeter would have had to fall by a factor of 2.25 for the economics to work out at all.
- A 450 mm crystal takes more than twice as long to grow. During this time, the probability of defects being incorporated into the lattice also increases.
- Almost every piece of equipment on the production line would have had to be redeveloped – at an investment level that individual manufacturers could no longer bear on their own.

The decisive point, however, was an economic one. The benefit of a larger wafer lies in spreading the fixed cost per piece of equipment over more chips. This advantage becomes smaller the more expensive the individual piece of

equipment becomes – and lithography equipment became dramatically more expensive during the same period. The effort required for the transition would therefore have increased, while the return from it decreased. Instead, the industry directed its investments toward smaller feature sizes and the third dimension: more devices per unit area rather than more area per wafer.

Why Are Wafers Round?

The question often arises as to why wafers are round, since microchips are, after all, rectangular. This inevitably results in wasted area on the wafer – area where no complete chips fit, and which must ultimately be discarded at the end of semiconductor manufacturing.

Having explained the two manufacturing processes - the crystal growth process and zone pulling - this question can easily be answered.

A silicon wafer for microchip manufacturing must be a single crystal. This is only possible with the processes mentioned, and these inherently produce a circular shape.

Even though it would be technically possible to subsequently shape the round silicon crystals into an angular form (e. g. by sawing), the round shape of silicon wafers nevertheless offers several advantages despite the rectangular microchips.

- Straightening the round silicon rods would place additional stress on the material and would inevitably lead to crystal defects that would affect chip quality.
- Round wafers are considerably more robust. Angular wafers could hardly be transported and processed without damage.
- Uniform processing during chip manufacturing using radially symmetric processes (CMP, spin-on, etching) is considerably simpler.
- A narrow edge region would always have to be discarded even with rectangular wafers, since the wafers must be held during processing. Deposited layers would flake off and generate additional particles if they extended all the way to the outermost edge.

As wafer size increases, the amount of wasted area also continues to decrease.

Rectangular wafers, on the other hand, are found in the manufacture of solar cells. Polycrystalline wafers are mostly used here, which can be cast into rectangular shapes. Manufacturing is comparatively simple, so angular wafers can be processed as well. The corners are usually additionally chamfered. For today's common single-crystal solar cells, it is exactly the other way around: they are made from crystals pulled in round form, which are trimmed to a square shape before sawing – with rounded corners, because cutting away the

remaining edge would cost more material than the gain in area would be worth.

Doping techniques

Doping

Doping means introducing a foreign substance into the semiconductor crystal in order to deliberately alter its conductivity through an excess or a deficiency of electrons. In contrast to doping during wafer fabrication itself, where the entire wafer is doped, the doping techniques described on this page allow silicon wafers to be doped partially. The introduction of the foreign substance can be achieved using various methods: diffusion, implantation (and alloying).

Diffusion

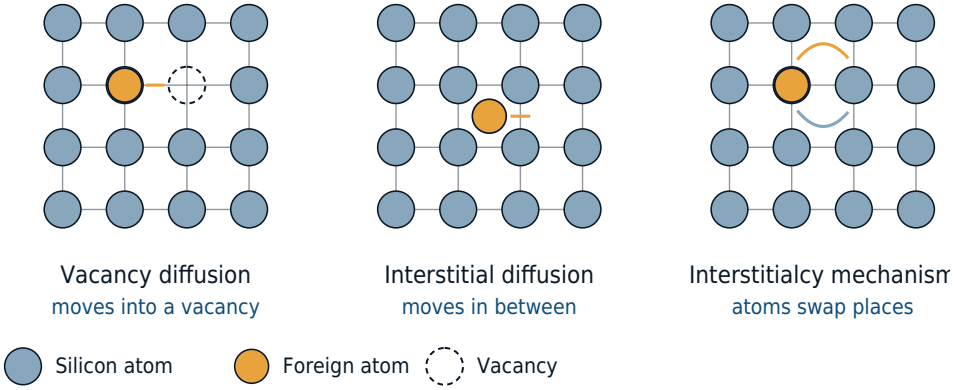
Diffusion is barely used anymore for targeted doping today; it has been superseded by ion implantation. Nevertheless, understanding it remains essential, because it never stops acting: every subsequent high-temperature step causes dopants that have already been introduced to keep migrating. This limits how much heat a wafer may still be subjected to after doping, and is the reason why furnace processes have increasingly been replaced by rapid, second-scale heating.

Diffusion means the spontaneous spreading of a substance within another substance due to a difference in concentration; for example, a drop of ink in a glass of water becomes evenly distributed after a certain amount of time. In a silicon crystal, one finds a solid lattice of atoms through which the dopant must move. This can happen in three ways:

- Vacancy diffusion: the foreign atoms occupy empty sites in the crystal lattice, which can always occur. Similar to hole conduction, an interplay arises between occupied lattice sites and vacancies.
- Interstitial diffusion: the foreign atoms move between the silicon atoms within the crystal lattice.
- Site exchange: foreign atoms located in the crystal lattice swap lattice sites with silicon atoms.

Diffusion in the crystal lattice

Three paths by which the dopant migrates through the lattice



All three paths occur side by side. Which one dominates depends on the dopant and on the temperature – boron and phosphorus mainly migrate via vacancies, smaller atoms more often between lattice sites.

The dopant can continue to move within the semiconductor crystal until either a concentration gradient has been equalized, or the temperature has been lowered far enough that the atoms can no longer move.
The speed of the diffusion process depends on several factors:

- dopant
- concentration difference
- temperature
- substrate
- crystal orientation of the substrate (see fabrication of single crystals)

Diffusion with an Exhaustible Source

Diffusion with an exhaustible source means that only a limited amount of dopant is available. The longer the diffusion process continues, the lower the concentration at the surface becomes; in return, the penetration depth into the substrate increases. The diffusion coefficient of a substance indicates how quickly it moves within the crystal. Arsenic, with a low diffusion coefficient, penetrates the substrate more slowly than, for example, phosphorus or boron.

Diffusion with an Inexhaustible Source

Here, the dopant is available in unlimited quantity. The concentration at the surface therefore remains constant throughout the process, since particles that have penetrated into the substrate are continuously replenished.

Diffusion Methods

In the following processes, the wafers are placed inside a quartz tube that is heated to a specific temperature along its entire length.

Diffusion from the Gas Phase

A carrier gas (nitrogen, argon) is enriched with the desired dopant (likewise in gaseous form, e.g. phosphine (PH_3) or diborane (B_2H_6)) and passed over the silicon wafers, where concentration equalization can take place.

Solid-Source Diffusion

Discs onto which the dopant has been applied are placed between the wafers. As the temperature in the quartz tube rises, the dopant diffuses out of the source discs into the atmosphere. A carrier gas then distributes the dopant evenly throughout the quartz tube, allowing it to reach the surface of the wafers.

Diffusion with a Liquid Source

Boron tribromide (BBr_3) or phosphorus oxychloride ($POCl_3$) serve as liquid sources. A carrier gas is passed through the liquid and thereby transports the dopant to the wafers, where it then reaches the surface of the silicon discs.

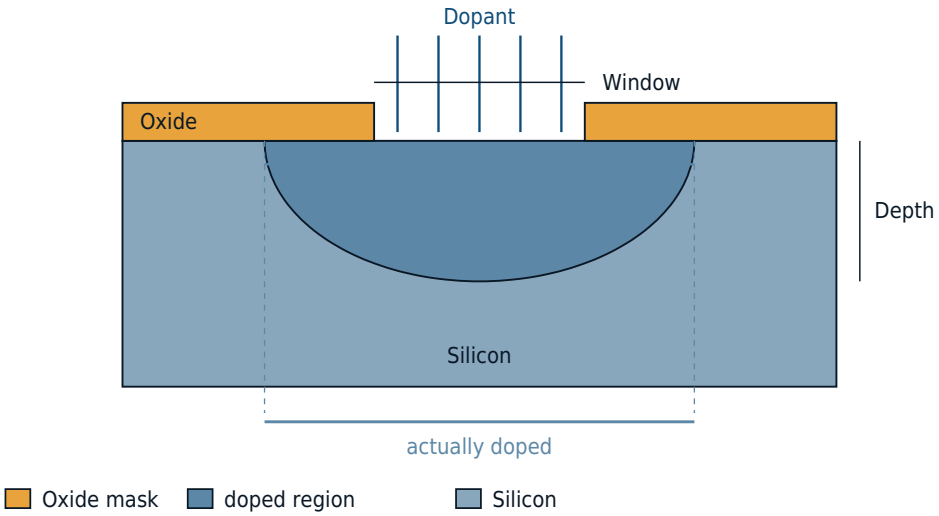
Since the entire wafer is not meant to be doped, certain regions are masked with silicon dioxide. Dopants cannot penetrate this oxide, so no doping takes place at these locations. To avoid stress or even breakage of the wafers, the tube is heated in steps (10 °C per minute) up to approximately 900 °C, at which point the dopant is introduced to the wafers. To initiate the diffusion process, the temperature is then raised to approximately 1200 °C.

Characteristics:

- Since many wafers can be processed simultaneously, this method is quite cost-effective
- If foreign substances from earlier doping steps are already present in the crystal, renewed exposure to heat can cause them to diffuse out again
- Over time, dopants accumulate inside the quartz tube and are additionally transported to the wafers by the carrier gas during subsequent doping steps
- Dopants spread within the crystal not only vertically but also laterally, so the doping windows always end up doped over a larger area than intended

Diffusion through a window in the mask

The dopant also migrates sideways and undercuts the oxide



The lateral spread is roughly eight tenths of the penetration depth. As a result, the doped region always ends up wider than the window – this has to be accounted for when designing the masks.

Ion implantation

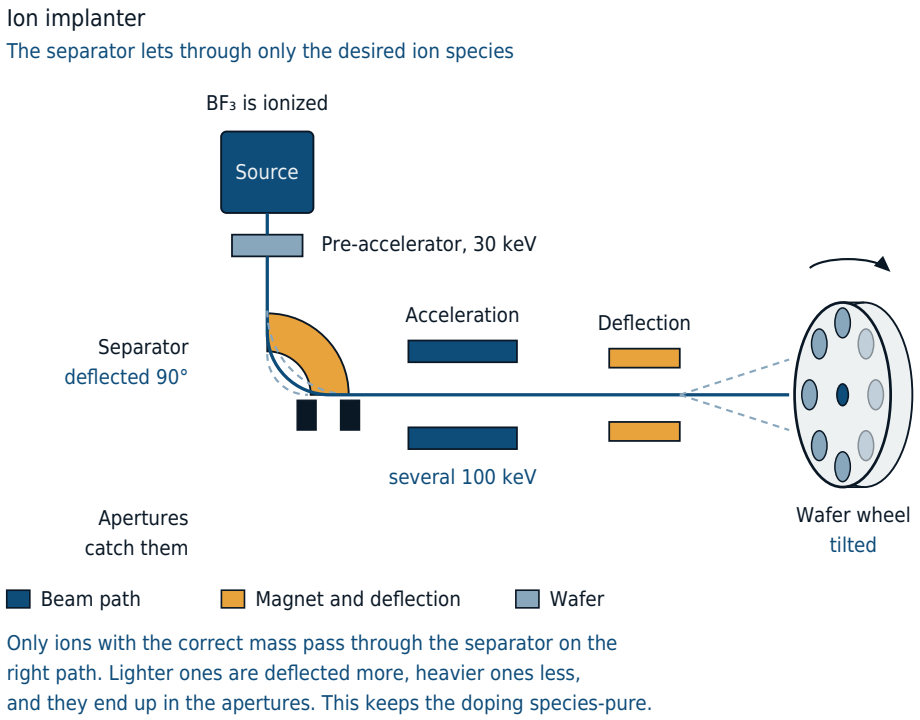
In the ion implantation charged dopants (ions) are accelerated in an electric field and irradiated onto the wafer. The penetration depth can be set very precisely by reducing or increasing the voltage needed to accelerate the ions. Since the process takes place at room temperature, previously added dopants can not diffuse out. Regions that should not be doped, can be covered with a masking photoresist layer.

An implanter consists of the following components:

- ion source: the dopants in gaseous state (e.g. boron trifluoride BF_3) are ionized
- accelerator: the ions are drawn with approximately 30 kiloelectron volts out of the ion source
- mass separation: the charged particles are deflected by a magnetic field by 90 degrees. Too light/heavy particles are deflected more/less than the desired ions and trapped with screens behind the separator
- acceleration lane: several 100 keV accelerate the particles to their final velocity (200 keV accelerate bor ions up to 2.000.000 m/s)
- Lenses: lenses are distributed inside the entire system to focus the ion beam
- distraction: the ions are deflected with electrical fields to irradiate the desired location
- wafer station: the wafers are placed on large rotating wheels and held into

the ion beam

Illustration of an ion implanter



Penetration depth of ions in the wafer

In contrast to diffusion processes the particles do not penetrate into the crystal due to their own movement, but because of their high velocity. Inside the crystal they are slowed down by collisions with silicon atoms. The impact causes damage to the lattice since silicon atoms are knocked from their sites, the dopants themselves are mostly placed interstitial. There, they are not electrically active, because there are no bonds with other atoms which may give rise to free charge carriers. The displaced silicon atoms must be re-installed into the crystal lattice, and the electrically inactive dopants must be activated.

Recovery the crystal lattice and activation of dopants

Right after the implantation process, only about 5 % of the dopants are bond in the lattice. In a high temperature process at about 1000 °C, the dopants move on lattice sites. The lattice damage caused by the collisions have already been cured at about 500 °C. Since the dopants move inside the crystal during high temperature processes, these steps are carried out only for a very short time.

Channeling

The substrate is present as a single crystal, and thus the silicon atoms are regularly arranged and form "channels". The dopant atoms injected via ion implantation can move parallel to these channels and are slowed only slightly, and therefore penetrate very deeply into the substrate. To prevent this, there are several possibilities:

- wafer alignment: the wafers are deflected by about 7° with respect to the ion beam. Thus the radiation is not in parallel direction to the channels and the ions are decelerated by collisions immediately.
- scattering: on top of the wafer surface a thin oxide is applied, which deflects the ions, and therefore prevents a parallel arival

Characteristic:

- the reproducibility of ion implantation is very high
- the process at room temperature prevents the outward diffusion of other dopants
- spin coated photoresist as a mask is sufficient, an oxide layer, as it is used in diffusion processes, is not necessary
- ion implanters are very expensive, the costs per wafer are relatively high
- the dopants do not spread laterally under the mask (only minimally due to collisions)
- nearly every element can be implanted in highest purity
- previous used dopants can deposit on walls or screens inside the implanter and later be carried to the wafer
- three-dimensional structures (e.g. trenches) can not be doped by ion implantation
- the implantation process takes place under high vacuum, which must be produced with several vacuum pumps

There are several types of implanters for small to medium doses of ions (10^{11} to 10^{15} ions/cm^2) or for even higher doses of 10^{15} to 1017 ions/cm^2.

The ion implantation has replaced the diffusion mostly due to its advantages.

Doping using Alloy

For completeness it should be mentioned that besides ion implanation and diffusion there is an alternative process: doping using alloy. Since this procedure, however, brings disadvantages with it such as cracks in the substrate, it is not used in today's semiconductor technology any more.

Oxidation

Overview

Application of Oxide Layers

Oxide is used in semiconductor manufacturing in a wide variety of fields:

- for isolation (e.g. between metallization layers)
- as a scatter layer (e.g. ion implantation)
- as an adaptation layer (e.g. LOCOS technique)
- for planarization (e.g. to soften edges)
- as a masking layer (e.g. diffusion)
- as alignment marks (alignment points in photolithography)
- as a protective layer (against mechanical damage)

Properties of oxide layers

In combination with silicon, oxide appears as silicon dioxide (SiO_2). It can be deposited on the wafer in very thin, highly uniform layers.

SiO_2 is very resistant and in semiconductor manufacturing can only be wet-etched by hydrofluoric acid (HF). Water and other acids do not attack the oxide; due to the risk of contamination by metal ions, alkaline solutions (KOH, NaOH, etc.) cannot be used. These, however, play a role in micromechanics for anisotropic wet etching – there they attack the silicon while the oxide remains virtually untouched, which is exactly why it serves as a mask. The removal process with HF proceeds according to the following reaction:

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

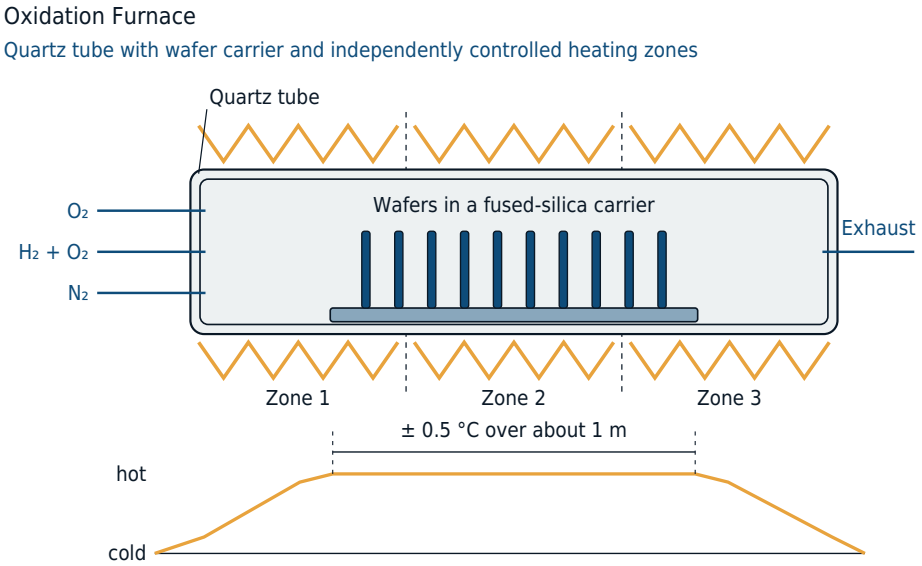
Furthermore, oxide is well suited for circuit functions, as it meets a wide variety of requirements (e.g. as gate oxide, field oxide, or interlayer dielectric).

What matters here is less the oxide itself than the interface to the silicon. It can be thermally grown with such a low defect density that a transistor switches reliably across it. No other semiconductor has its own oxide with this property – this is the main reason why silicon has prevailed over materials that would actually be more favorable electrically.

Fabrication of oxide layers

Thermal oxidation

In thermal oxidation, silicon wafers are oxidized at about 1000 °C in an oxidation furnace. This furnace essentially consists of a quartz tube in which the wafers are placed on a carrier (boat) made of quartz glass, several independently controllable heating coils, and various gas supply lines. Quartz glass has a very high melting point (well above 1500 °C) and is therefore very well suited for high-temperature processes. To avoid wafer warping or cracking, the quartz tube is heated in very small steps (max. 10 °C per minute). Thanks to the separate heating coils, the temperature along the entire tube – over a length of about 1 m – can be controlled precisely to within ± 0.5 °C.



Process gases: O_2 for dry, $H_2 + O_2$ for wet oxidation, N_2 for purging.
Process temperature about 800 to 1200 °C, ramp rate at most 10 °C per minute.

The oxygen then flows as a gas over the wafers and reacts at the surface to form silicon dioxide. A glass-like layer with an amorphous structure is formed. Depending on the process gas, different oxidations take place; a thermal oxidation naturally has to occur on a silicon surface. Thermal oxidation is divided into dry and wet oxidation, the latter of which can in turn be divided into wet oxidation and H_2-O_2 combustion.

Dry oxidation

The oxidation process takes place under a pure oxygen atmosphere. Silicon reacts with oxygen to form silicon dioxide:

$Si + O_2 \rightarrow SiO_2$

This process usually takes place at 1000–1200 °C. To produce very stable and thin oxides, the oxidation is carried out at about 800 °C.

Characteristics of dry oxidation

- slow oxide growth
- high density
- high breakdown voltage (for electrically highly stressed oxides, e.g. gate oxide)

Wet oxidation

In wet oxidation, the oxygen is passed through a bubbler vessel containing water (about 95 °C), so that in addition to oxygen, water is also present in the quartz tube in the form of steam. This results in the following reaction equation:

$Si + 2H_2O \rightarrow SiO_2 + 2H_2 \uparrow$

This process takes place at 900–1000 °C. It exhibits fast oxide growth at low temperatures and is used, among other things, for producing masking layers and field oxides. The quality of the resulting layer is lower than that of dry oxidation.

Comparison of growth rate for dry and wet oxidation

Temperature	Dry oxidation	Wet oxidation
900 °C	19 nm/h	100 nm/h
1000 °C	50 nm/h	400 nm/h
1100 °C	120 nm/h	630 nm/h

H_2-O_2 combustion

In H_2-O_2 combustion, high-purity hydrogen is used in addition to high-purity oxygen. The two gases are fed into the quartz tube separately and burned at the inlet opening. To avoid an explosive (Knallgas) reaction with the highly flammable hydrogen, the temperature must be above 500 °C; the gases then react in a silent combustion. This process allows the production of fast-growing oxide layers with only minor contamination. This makes it possible to produce both thick oxides and thin layers at comparatively low temperatures (900 °C). The low temperature also allows for the thermal treatment of wafers that have already been doped (see doping by diffusion).

In all thermal oxidation processes, oxide growth is higher on (111)-oriented substrates than on (100)-oriented substrates (see the single crystal). In addition,

a very high concentration of dopants in the substrate significantly increases the growth rate.

Course of the oxidation process

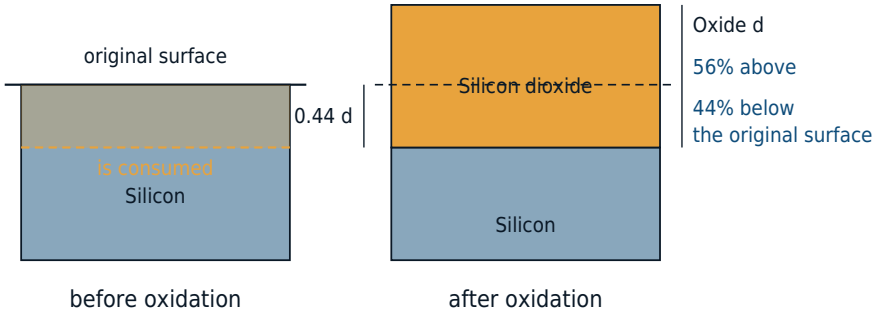
At the beginning, the oxygen reacts at the wafer surface to form silicon dioxide. An oxide layer is now present on the substrate, through which subsequent oxygen atoms must first diffuse in order to react with the silicon. The growth rate depends on the reaction time between silicon and oxide only at the beginning; beyond a certain thickness, the oxidation rate is determined by how quickly the oxygen diffuses through the already grown silicon dioxide. As the oxide thickness increases, growth therefore slows down. Mathematically, this is described by a linear-parabolic model according to Deal and Grove: initially, thickness grows roughly proportional to time, but later only proportional to the square root of time. Thick oxides are therefore disproportionately time-consuming to produce. Since the resulting layer is amorphous, not all bonds of the silicon atoms are intact; there are partly dangling bonds (free electrons and holes) at the Si-SiO_2 interface. This results in an overall slightly positive charge at this interface. Since this charge can adversely affect devices, efforts are made to keep it as low as possible. This can be achieved, for example, by using a higher oxidation temperature, or by using wet oxidation, which likewise causes only a very low charge.

Volume increase

In thermal oxidation, silicon is consumed through the reaction with oxygen to form silicon dioxide. The ratio of the grown oxide layer to the consumed silicon is 2.27; that is, the oxide grows into the substrate by 45 % of the oxide thickness.

Growth Behavior of Oxide on Silicon

The oxide takes up more space than the silicon it forms from



A silicon layer of thickness 1 becomes oxide of thickness 2.27.
The interface moves inward in the process: the oxide grows into the substrate.

Segregation

Dopants present in the substrate can be incorporated either into the silicon crystal or into the oxide, depending on which material dissolves the dopant more readily. This so-called segregation coefficient k is calculated as follows:

$$k=\frac{Solubility\ of\ the\ dopant\ in\ Si}{Solubility\ of\ the\ dopant\ in\ SiO_2}$$

If k is greater than 1, the dopants are incorporated at the surface of the substrate; if k is less than 1, the dopants accumulate in the oxide.

Where thermal oxidation stands today

The classic role of thermal oxide, forming the gate dielectric, has been discontinued. At layer thicknesses of just a few atomic layers, so much current tunnels through the oxide that it is no longer suitable as an insulator. Its place has been taken by hafnium dioxide, which is deposited via atomic layer deposition and, for the same effect, may be thicker. Even there, however, it is not possible to do entirely without thermal oxide: between the silicon and the hafnium dioxide lies an intermediate layer of silicon dioxide only a few atomic layers thick, since a low-defect interface to the silicon cannot be produced any other way.

What has remained for thermal oxidation are the tasks for which its particular characteristic – growing directly out of the substrate itself – is especially useful: the pad oxide beneath nitride masks, sacrificial oxides to protect the surface from implantation damage, the lining of freshly etched trenches in trench isolation, and the oxides used in power electronics, where thickness and breakdown strength continue to matter.

The equipment technology has also changed. The furnace described above processes an entire batch over several hours. For thin layers, this is too imprecise and thermally too demanding; there, a single wafer is heated to temperature within seconds using lamps and cooled down again just as quickly. This keeps the amount of heat introduced so low that existing doping profiles barely shift at all.

Oxidation by vapor deposition

In thermal oxidation silicon is used up to form oxide. If the silicon surface is covered by other films, the oxide layer has to be created in deposition processes since thermal oxidation needs a bare silicon surface in either case. In deposition processes, oxygen and silicon are added in gaseous states. There are two important processes for oxidation by vapor deposition: the silane pyrolysis and the TEOS deposition. A detailed description of these processes can be found in the chapter deposition.

Silane pyrolysis

Pyrolysis means a cleavage of chemical compounds - in this case the gas silane SiH_4 and highly purified oxygen O_2 - by heat. Since the toxic silane tends to self-ignition at ambient air above a concentration of 3 % it has to be diluted with nitrogen or argon below 2 %. At about 400 °C silane reacts with oxygen to form silicon dioxide and hydrogen:

$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2 \uparrow$

The quality of the silicon dioxide is low. As an alternative, a high-frequency excitation via a plasma at about 300 °C can be used for the dioxide deposition. This produces a somewhat more stable oxide.

TEOS deposition

Tetraethylorthosilicate, or TEOS ($SiO_4C_8H_{20}$) for short, used in this method contains the two required elements silicon and oxygen. Under vacuum, the compound, which is liquid at room temperature, already transitions into gaseous form. The gas is transferred into a heated quartz tube and cleaved there at about 750 °C.

$SiO_4C_8H_{20} \rightarrow SiO_2 + \text{byproducts}$

The silicon dioxide deposits on the wafers, while the byproducts (e.g. H_2O – water vapor) are extracted. The uniformity of this oxide is determined by the pressure in the quartz tube and the process temperature. It is electrically stable and very pure.

These properties result in a division of labor: wherever the interface to the silicon matters, thermal oxidation is used; wherever the oxide is to be placed on something other than bare silicon, or wherever the temperature is limited, deposition is used. Above the first metal layer, only deposition is possible, since the interconnects cannot withstand furnace temperatures.

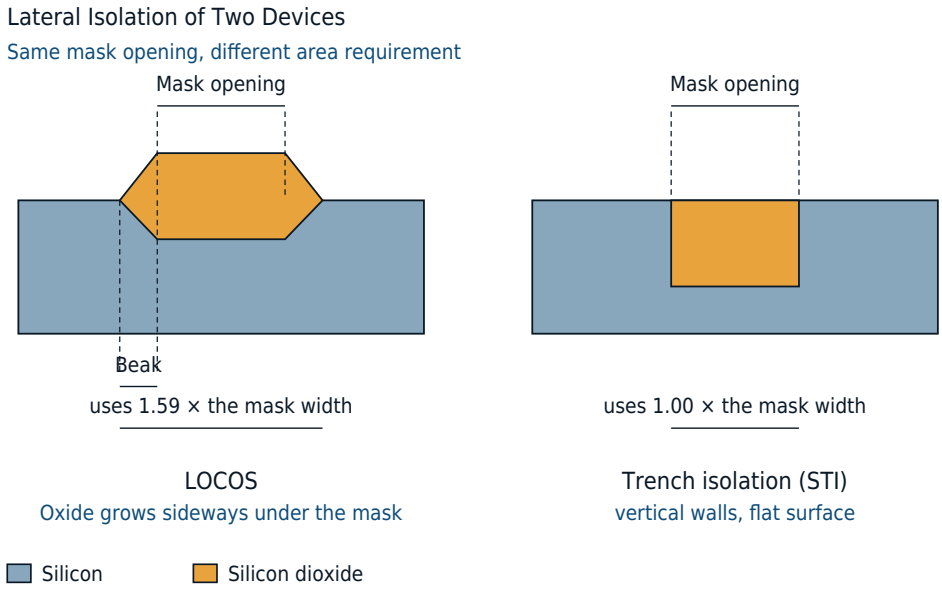
LOCOS technology

Shallow Trench Isolation

Local oxidation reaches its limits with small structures: the bird's beak does not shrink just because the devices do. It occupies a fixed length, determined by the thickness of the pad oxide and nitride, and therefore consumes an ever-growing share of the available area. Below a feature size of roughly half a micrometre, the process became unusable for this reason.

It has been replaced by shallow trench isolation (STI). Instead of growing the oxide, a trench is etched into the silicon and filled with oxide. The isolation is therefore exactly as wide as the mask opening, and the surface remains flat.

Comparison of the two methods



Process Flow

Trench isolation is one of the very first steps in the entire fabrication process; it

is created before any device is formed.

1. A thin pad oxide is applied to the wafer, followed by a layer of silicon nitride. The nitride later serves as a stop layer during polishing.
2. Using a resist mask, the nitride, pad oxide, and the underlying silicon are anisotropically etched. The trenches are deliberately given slightly sloped walls so that they can subsequently be filled more easily.
3. A brief thermal oxidation rounds off the sharp edges at the trench rim and repairs the etch damage on the trench wall. Sharp edges would locally intensify the electric field and distort the threshold voltage of the neighbouring transistor.
4. The trench is filled with oxide. Because it is narrow and deep, processes are used that leave no voids behind.
5. The excess oxide is removed by chemical mechanical polishing until the nitride is exposed.
6. The nitride is removed wet-chemically in hot phosphoric acid, and the pad oxide in hydrofluoric acid. What remains are flat silicon surfaces surrounded by oxide.

The price for this more compact isolation is the number of process steps: whereas LOCOS only required masking and oxidation, this method requires etching, filling, polishing, and two wet processes. Without chemical mechanical polishing, the process would not be possible at all – only with it can the fill material be cleanly recessed down to the level of the silicon surface.

Further Variants

The depth of the trenches depends on what needs to be separated. A few hundred nanometres are sufficient to separate neighbouring transistors. If, on the other hand, entire circuit blocks need to be isolated from one another – for example in power and analog technology, or to protect sensitive analog areas from interference originating in the digital part – trenches several micrometres deep are created (deep trench isolation).

In devices with vertically oriented channels, even trench isolation loses its role as the area-determining factor: there, it is no longer a trench separating two adjacent regions, but rather the devices simply stand free on their own.

Very large-scale integration

In semiconductor technology, structures are created using exposure and etching processes. This creates steps at which photoresist can accumulate, thereby reducing the resolving power in photolithography. With an isotropic etch characteristic (material is removed both vertically and horizontally), resist masks must be adjusted so that undercut structures have the correct dimensions

at the end of the etching process.

At these steps, problems also occur during metallization, since the interconnects are narrowed there, resulting in damage caused by electromigration.

To achieve a high packing density – that is, to fit as many devices as possible into as small an area as possible – steps and unevenness must be avoided. The first answer to this was the LOCOS technique: LOCal Oxidation of Silicon. It shaped the structure of integrated circuits for two decades and can still be found today in power and analog technology.

For fine structures, however, it is no longer sufficient. Why this is the case becomes apparent from its characteristic side effect, described in the following section; the trench isolation used today avoids it.

Bird's beak

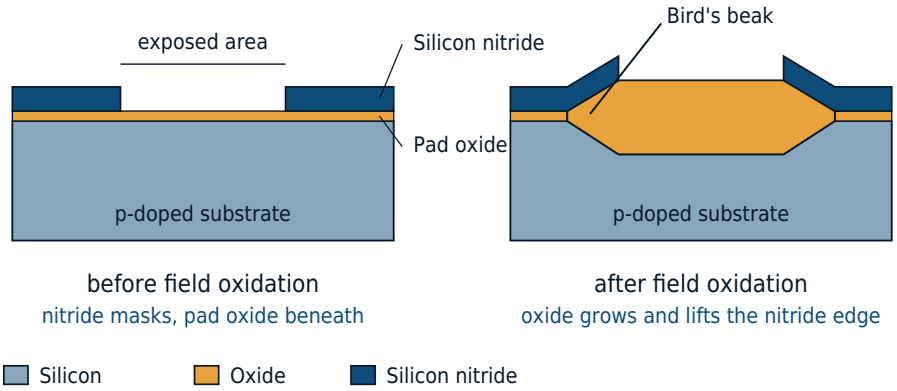
The LOCOS process utilizes the different rates of oxidation of silicon and silicon nitride to locally mask the wafer surface.

A silicon nitride layer is used to mask the areas where no oxide should grow; an oxide layer forms only on the regions free of nitride. Since silicon and silicon nitride have different coefficients of thermal expansion, a thin oxide layer – the pad oxide – is deposited between the nitride mask and the substrate to prevent strain caused by temperature changes.

For lateral isolation of transistors, a field oxide (FOX) is then grown on the bare silicon surface. While a silicon dioxide layer forms on the exposed silicon during field oxidation, the pad oxide causes a lateral diffusion of oxygen underneath the nitride mask, resulting in slight oxide growth at the edge of the mask. This oxide extension has the shape of a bird's beak, whose length depends on the oxidation process as well as on the thickness of the nitride and the pad oxide.

LOCOS: Before and After Field Oxidation

The oxide grows only where the nitride leaves the surface exposed



The oxygen diffuses sideways through the pad oxide, under the nitride mask.

(Source: Consiglio Nazionale delle Ricerche)

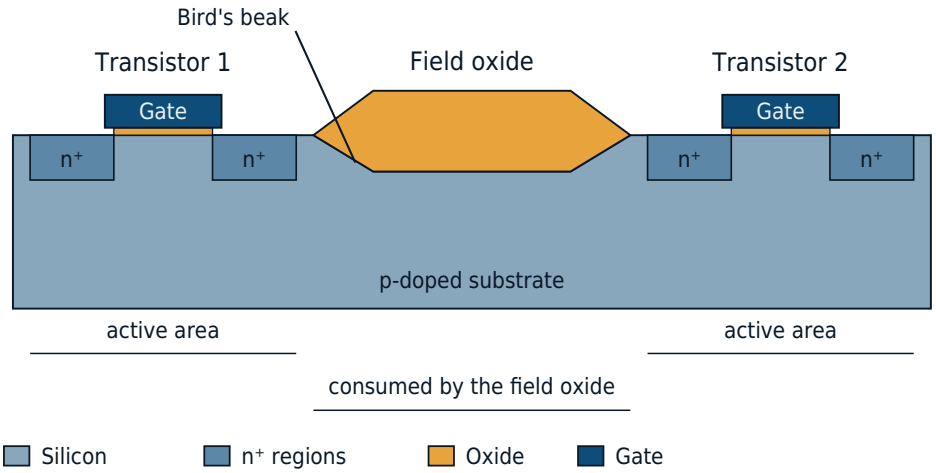
Besides this effect, which can consume up to 1 µm of the area intended for devices, wet oxidation also causes the so-called white ribbon or Kooi effect. Here, nitride from the masking layer reacts with hydrogen to form ammonia NH_3, which diffuses to the silicon surface and causes nitridation there. This nitride must be removed prior to gate oxidation, since it would otherwise act as a mask.

Despite these negative effects, the LOCOS process is a suitable method for enabling high packing density. Because of the reduced unevenness, without the formation of edges and steps, the resolving power in photolithography is improved. The field oxide can be etched back slightly; this slightly reduces the grown oxide, but the length of the bird's beak decreases, and the surface is again flattened somewhat more. This is referred to as fully recessed LOCOS.

Nevertheless, the bird's beak is the point at which the process ultimately fails. Its length depends on the thickness of the pad oxide and the nitride, not on the size of the structures. It therefore does not shrink as devices become smaller. As long as the isolation regions were several micrometers wide, an extension of a few tenths of a micrometer hardly mattered; at structures below half a micrometer, it consumes most of the area it is meant to isolate. A thinner pad oxide would shorten it, but would then transfer the stress from the nitride mask to the silicon unabated, creating crystal defects.

LOCOS Isolating Two Transistors

The field oxide separates the active areas from each other



Without the field oxide, both transistors would be connected through the substrate. The bird's beak costs area that is then unavailable for devices.

Plasma, the fourth aggregation state of a material

Plasma state

In many semiconductor manufacturing processes a plasma is used, be it in sputtering, deposition processes, or dry etching. An important point here is that the wafer remains cold in the plasma. This may sound contradictory at first, since the electrons in the plasma are extraordinarily hot, at several tens of thousands of degrees. However, they are so light that upon collision they hardly transfer any energy to the heavy gas particles – the gas itself, and with it the wafer, therefore remain close to room temperature. Such a plasma, in which the electrons and the gas do not have the same temperature, is called a non-equilibrium plasma. Only for this reason can wafers that already carry a metallization be processed in a plasma at all.

Plasma is also referred to as the fourth state of matter. A state of matter is a qualitative condition of substances that depends on temperature and pressure. The three states solid, liquid, and gaseous are also encountered in everyday life. If the temperature is low, every atom in a substance is fixed rigidly at one point. Attractive forces prevent them from moving. At absolute zero (-273.15 °C) substances also do not undergo any reaction. As the temperature increases, the particles begin to oscillate, and the bonds between the atoms become more unstable. Once the melting point is reached, a substance transitions from the first to the second state of matter: ice (solid) turns into water (liquid).

The attractive forces in liquid substances are still present, but the particles can shift position and no longer have fixed places as in the solid state; the particles adapt, for example, to a given shape. If the temperature is increased further, the bonds break apart completely, and the particles move independently of one another. At the boiling point, a substance transitions from the second to the third state of matter: water (liquid) turns into water vapor (gaseous).

While the volume of solids and liquids is constant, gaseous substances completely fill the available space; the particles distribute themselves evenly throughout the entire volume.

Every substance has a very specific melting point and boiling point. Silicon melts at 1414 °C and transitions into the gaseous state at 2900 °C. If even more energy is supplied to a substance, collisions between the particles knock electrons out of the outermost electron shell. Free electrons and positively

charged ions are now present in the space: the plasma state has been reached.

Plasma generation

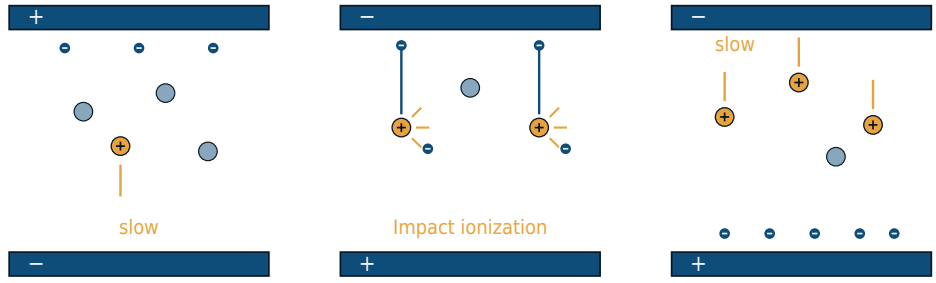
A plasma in semiconductor technology is usually generated by a high-frequency voltage; argon, for example, serves as the gas. The gas is located in a high-frequency field between two oppositely charged plates (electrodes) and is ionized there. This requires electrons that knock further electrons out of the argon atoms. These initial electrons can be generated in different ways:

- electrons are emitted from a thermionic cathode
- electrons are pulled out of the negative electrode by a very high voltage
- in every gas, brief free electrons occur through collisions of the particles

Since the electrons are much lighter than the ions, they are immediately attracted to the positively charged electrode; the heavy ions move slowly toward the negative electrode. However, before they reach it, the polarity of the electrodes is reversed; the electrons are accelerated toward the other electrode and, on their trajectory, knock further electrons out of atoms through collisions. Typical frequencies for plasma generation are 13.56 megahertz and 2.45 gigahertz, meaning the voltage at the electrodes is reversed 13.56 million or 2.45 billion times per second, respectively.

How the Plasma Ignites

The electrodes are reversed millions of times per second



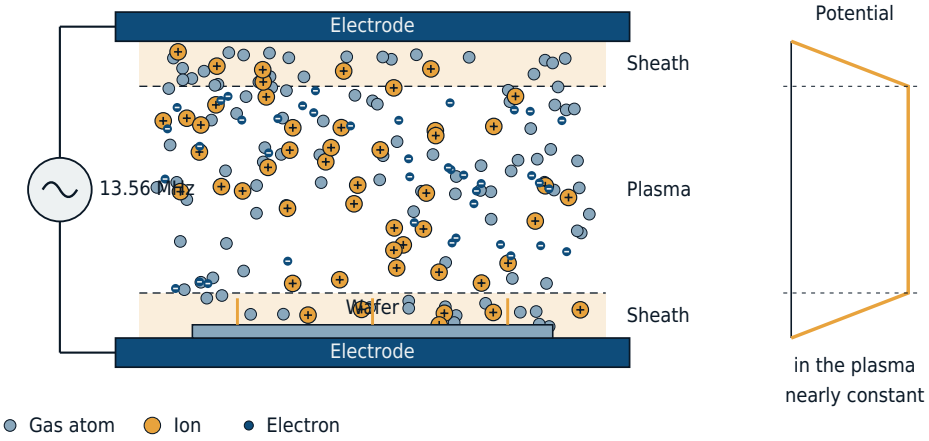
The light electrons follow the alternating field, the heavy ions barely do.
Each polarity reversal creates new charge carriers, until the plasma ignites.

Inside the plasma, electrons and ions are roughly balanced, so it is electrically neutral overall. The situation is different immediately in front of the electrodes: the light, fast electrons reach there much more often than the sluggish ions,

which cannot follow the rapid voltage changes. As a result, the electrode charges negatively, and an electron-depleted boundary layer (sheath) forms in front of it. Almost the entire voltage drops across this sheath – it is this voltage that accelerates the ions toward the wafer and thus accounts for the directional component of dry etching.

Plasma Generation Between Two Electrodes

Charges balance out in the bulk, but not in front of the electrodes



The electrons follow the alternating field, the heavy ions do not.
This depletes electrons in front of the electrodes; almost the entire voltage drops across this sheath and accelerates the ions onto the wafer.

Plasma generation takes place under vacuum; the plasma produced is not heated, which is important for many processes. Depending on the gas and the equipment in which the plasma is generated, it can be used in deposition, sputtering, etching, or also in ion implantation. Due to the rapid oscillations of the positive ions in the high-frequency field, they are very energetic. The plasma does not contain only positive ions and free electrons, since other particles are also created through collisions; the state of the plasma changes continuously. Electrons are partially recaptured by the ions and knocked out again; however, these additional particles play no role in the further use of the plasma. Depending on the particle density in the plasma (10^8–10^{12} particles per cm^3), the degree of ionization is 0.001–10 %, meaning the majority of the particles are uncharged.

The simple arrangement with two plates has one drawback: the same voltage determines both how many ions are generated and how hard they strike the wafer. The two cannot be set independently of one another. Equipment for fine structures therefore generates the plasma differently – via a coil that couples in an alternating field from outside, or via microwaves – and applies the voltage at

the wafer electrode independently of this. Ion density and ion energy are then two separate, independently adjustable parameters.

Chemical vapor deposition

Silicon vapor phase epitaxy

Epitaxy means "on top" or "arranged upon," and represents a process in which a layer is created on top of another layer and inherits its crystal structure. If the deposited layer is of the same material as the substrate, one speaks of homoepitaxy; with two different materials, of heteroepitaxy. The most significant case of homoepitaxy is the deposition of silicon on silicon; in heteroepitaxy, a different material grows on top, such as silicon-germanium on silicon or gallium nitride on silicon and sapphire. A crystalline base is required: nothing single-crystalline grows on an amorphous layer such as silicon dioxide. Silicon-on-insulator wafers (SOI) are therefore not created by epitaxy, but rather by bonding two wafers together and removing one of them down to a thin layer.

Homoepitaxy

Depending on the process, the wafers can already be delivered by the manufacturer with an epitaxial layer (e.g. in CMOS technology), or the chip manufacturer has to carry this out themselves (e.g. in bipolar technology).

As gases for generating the layer, pure hydrogen is used in combination with silane (SiH_4), dichlorosilane (SiH_2Cl_2), or silicon tetrachloride ($SiCl_4$). At about 1000 °C, the gases cleave off silicon, which deposits on the wafer surface. The silicon adopts the structure of the substrate and, for energetic reasons, grows plane by plane in succession. To prevent the silicon from growing polycrystalline, there must always be a shortage of silicon atoms present, i.e. slightly less silicon must always be available than could actually grow. With silicon tetrachloride, the reaction proceeds in two steps:

$SiCl_4 + H_2 \rightarrow SiCl_2 + 2HCl$
$2\ SiCl_2 \rightarrow Si + SiCl_4$

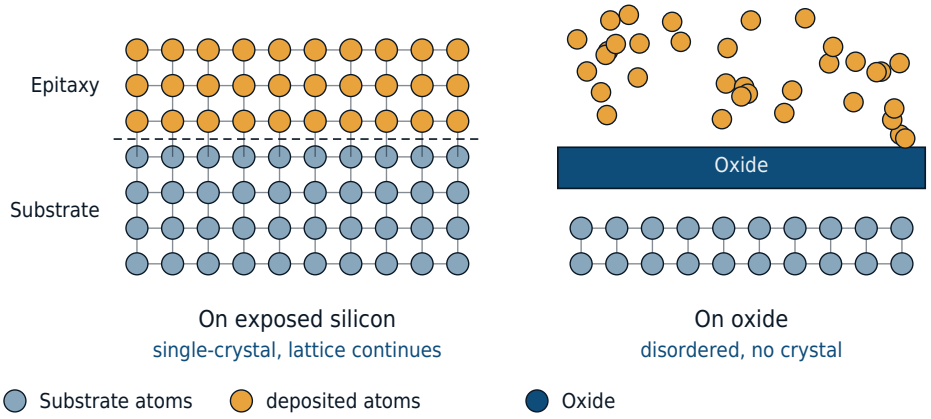
In order for the silicon on the substrate to adopt the crystal structure, the surface must be absolutely clean; this makes use of the equilibrium reaction. Both reactions can also proceed in the opposite direction, depending on the ratio of the gases used. If there is only little hydrogen present in the atmosphere, silicon is removed from the wafer surface due to the high chlorine concentration – as in the trichlorosilane process used to purify silicon. Only with an increasing concentration of hydrogen is growth achieved.

With $SiCl_4$, the growth rate is about 1–2 µm per minute. Since single-crystalline silicon only grows on the cleaned surface, certain areas can be masked with oxide, on which polycrystalline silicon then grows. Compared to single-crystalline silicon, however, this is etched away very easily by the backward-running reaction. If diborane (B_2H_6) or phosphine (PH_3) are added to the process gases, doped layers can be produced, since the doping gases decompose at the high temperatures and the dopants are incorporated into the crystal lattice.

The process for producing homoepitaxial layers is carried out under vacuum. The process chamber is first heated to 1200 °C so that the native oxide, which always forms on the silicon surface, is volatilized. As mentioned above, too low a hydrogen concentration leads to a back-etching of the wafers. This is exploited prior to the actual process to clean the wafer surface in this way. By changing the gas concentrations, deposition then takes place in an epitaxy reactor.

Epitaxy: the layer adopts the crystal structure

It grows plane by plane, continuing the lattice of the surface below



This is exactly what selective epitaxy relies on: at the right chlorine level, the disordered material is continuously etched away again, while the single-crystal material remains.

Selective epitaxy and strained layers

The circumstance described above – that single-crystalline silicon only grows on exposed silicon, while polycrystalline material is removed again by the backward-running reaction – is no longer merely a side effect today, but the basis of a process in its own right. With a suitably chosen chlorine content, the layer grows exclusively in the opened windows and not at all on oxide or nitride – this is referred to as selective epitaxy.

Its most important application lies not in adding to the silicon, but in straining it. If a silicon-germanium alloy is grown into the source and drain regions of a transistor instead of silicon, this alloy requires more space than the lattice allows, due to the larger germanium atoms. As a result, it presses on the channel located between them. In a lattice compressed in this way, holes move noticeably more easily, and the transistor switches faster. For the opposite case – tensile strain on the channel, favorable for electrons – silicon-carbon is used instead. Since the mid-2000s, these strained layers have been part of the standard structure of every high-performance transistor.

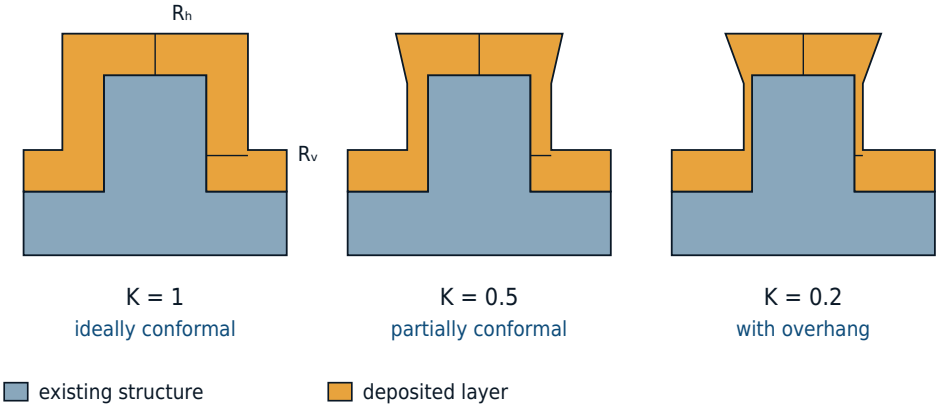
CVD process: Chemical Vapor Deposition

Semiconductor technology often requires layers that cannot be produced from the silicon substrate itself. For silicon nitride and silicon oxynitride, for example, the layer has to be created through thermal decomposition of gases that contain all the required materials, such as silicon. This is the principle underlying chemical vapor deposition, or CVD for short. The wafer merely serves as the base surface and does not react with the gases. Depending on pressure and temperature, the CVD process is divided into different methods, whose resulting layers differ in density and edge coverage. If the layer growth on vertical edges is exactly as high as on horizontal surfaces, the deposition is conformal.

The conformity K is the ratio of the layer growth on vertical surfaces R_v to the growth on horizontal surfaces R_h: $K = R_v : R_h$. If the deposition is not ideally conformal, K is noticeably less than 1 (e.g. $R_v : R_h = 1:2 \rightarrow K = 0.5$). High conformities can only be achieved through high process temperatures and low pressures. A conformity of exactly 1 is only achieved by atomic layer deposition, because there it is not surface attachment but a self-limiting reaction that determines the layer thickness.

Conformality of a Deposition

Ratio of growth on vertical and horizontal surfaces



R_h is the growth on horizontal, R_v that on vertical surfaces.
The smaller K is, the more readily the layer grows together at the edge and seals in a void beneath it.

APCVD: Atmospheric Pressure CVD

APCVD is a CVD deposition at normal pressure (atmospheric pressure) which is used for producing doped and undoped oxides. The resulting oxide has a low density, and edge coverage is moderate due to the low temperature. The major advantage is throughput: without a vacuum, wafers can be processed in a continuous flow, and the deposition rate is high. For thick, non-critical oxides, the process therefore remains economical; for thin or conformal layers, it is not an option.

The process gases used are silane SiH_4 (highly diluted with nitrogen N_2) and oxygen O_2, which are thermally decomposed at about 400 °C and react with each other.

$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2$ (T = 430 °C, p = 10^5 Pa)

By adding ozone O_3, better conformity can be achieved, since it increases the mobility of the depositing particles. The oxide is porous and electrically unstable and can be densified through a high-temperature step.

To prevent the formation of edges, which could cause problems when depositing further layers, so-called phosphorus silicate glass (PSG) is used for layers serving as interlayer dielectric. For this purpose, phosphine PH_3 is additionally added to the two gases SiH_4 and O_2, so that the resulting oxide contains 4–8 % phosphorus. A high phosphorus content significantly improves flow properties;

however, phosphoric acid can form, which corrodes aluminum (interconnects).

Since annealing steps affect preceding processes (e.g. doping), only brief annealing using powerful argon lamps (several hundred kilowatts, <10 s, T = 1100 °C) is used to reflow the PSG, instead of long furnace processes.

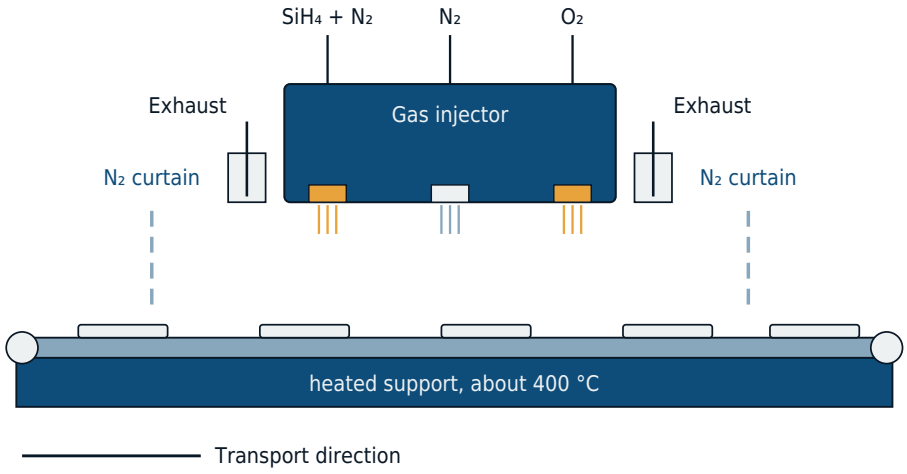
Analogous to phosphosilicate glass, boron can also be added at the same time (borophosphosilicate glass, BPSG, 4 % B and 4 % P).

The purpose of these doped glasses was to reflow at high temperature, thereby leveling out steps. This task has since been taken over by chemical mechanical polishing, which requires no thermal load and also planarizes over large areas. Doped glasses are still used today mainly as interlayer dielectric beneath the first metal layer, where the phosphorus additionally binds alkali ions and keeps them away from the transistor.

Illustration of a horizontal reactor for APCVD deposition

Horizontal Reactor for APCVD

The wafers pass under the gas injector at atmospheric pressure



At atmospheric pressure no load lock is needed: the wafers pass through continuously, which is what gives the process its high throughput.

Silane and oxygen are fed in separately and only brought together above the wafer, so they don't react inside the injector itself.

LPCVD: Low Pressure CVD

In the LPCVD process, the reaction takes place under vacuum, allowing the deposition of thin layers such as silicon nitride (Si_3N_4), silicon oxynitride (SiON), silicon dioxide (SiO_2), and tungsten (W). LPCVD processes enable very good

conformities of nearly 1; the reason for this is the low pressure of only 10–100 Pa (normal air pressure is about 100,000 Pa), which means the particles do not experience directional movement but instead spread out through mutual collisions, so that edges on the wafer surface are covered evenly by all gas particles. The process temperature of up to 900 °C also contributes to the high conformity. In contrast to the APCVD process, the density and stability are very high.

The reactions for Si_3N_4, SiON, SiO_2, and tungsten proceed according to the following reaction equations:

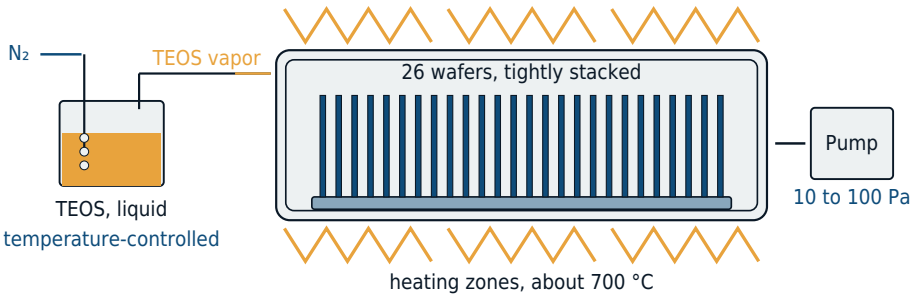
a) Si_3N_4 (850 °C): $4NH_3 + 3SiH_2Cl_2 \rightarrow Si_3N_4 + 6HCl + 6H_2$
b) SiON (900 °C): $NH_3 + SiH_2Cl_2 + N_2O \rightarrow SiO_xN_y$ + byproducts
c) SiO_2 (700 °C): $SiO_4C_8H_{20} \rightarrow SiO_2$ + byproducts
d) Tungsten (400 °C): $WF_6 + 3H_2 \rightarrow W + 6HF$

In contrast to the gaseous precursors used for Si_3N_4, SiON, and tungsten, liquid tetraethyl orthosilicate (TEOS) is used to produce SiO_2; in addition, there are other liquid sources such as ditertiarybutylsilane (DTBS, $SiH_2C_8H_{20}$) or tetramethylcyclotetrasiloxane ($Si_4O_4C_4H_{16}$).

A tungsten layer can only be produced on silicon. Therefore, if no silicon is available as a base surface, silane has to be added.

LPCVD System for TEOS Layers

TEOS is liquid at room temperature and must first be vaporized



The low pressure lets the gas molecules travel far and reach every gap. That's why the wafers can stand close together, and the edges are covered evenly: the conformality is close to 1.

PECVD: Plasma Enhanced CVD

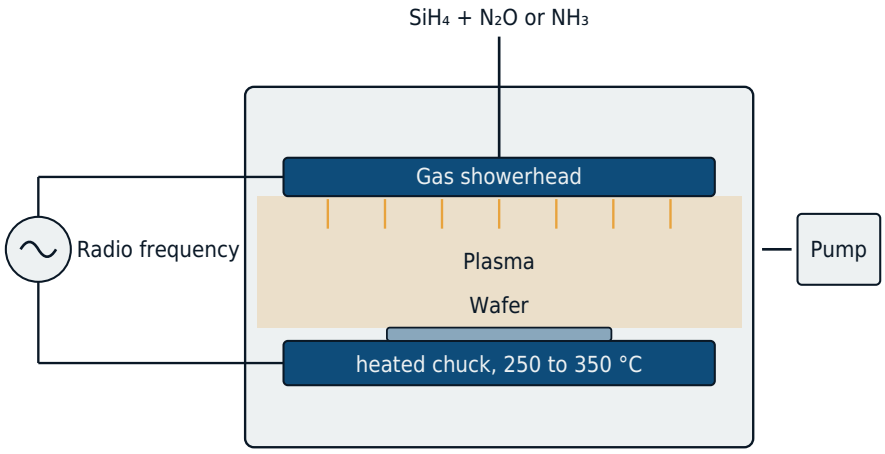
The plasma-enhanced CVD process takes place at 250–350 °C. Since the temperature is too low to decompose the gases, a high-frequency voltage is used to bring the gas into the plasma state. In the plasma state it is very energetic

and then deposits onto the wafer. In particular, aluminum interconnects cannot be exposed to high temperatures (aluminum melts at 660 °C, but the thermal budget of finished aluminum interconnects is already exhausted at about 450 °C), which is why the PECVD process is used above all here for depositing SiO_2 and Si_3N_4. Instead of dichlorosilane SiH_2Cl_2, as used in the LPCVD process, silane SiH_4 is employed, since it decomposes more readily at low temperatures. The conformity is not as ideal as with the LPCVD process (about 0.6–0.8), but the deposition rate, at 500 nm per minute, is significantly higher.

This combination – low temperature at a high rate – accounts for the process's significance today. Everything deposited after the first metal layer is produced by PECVD: the interlayer dielectrics between the wiring levels, the nitride layers that serve as etch stops and as diffusion barriers against copper, the hard masks used for patterning, and the final passivation of the finished chip. The carbon-containing low-k dielectrics are also produced this way, by adding an organosilicon compound to the process gas.

PECVD System

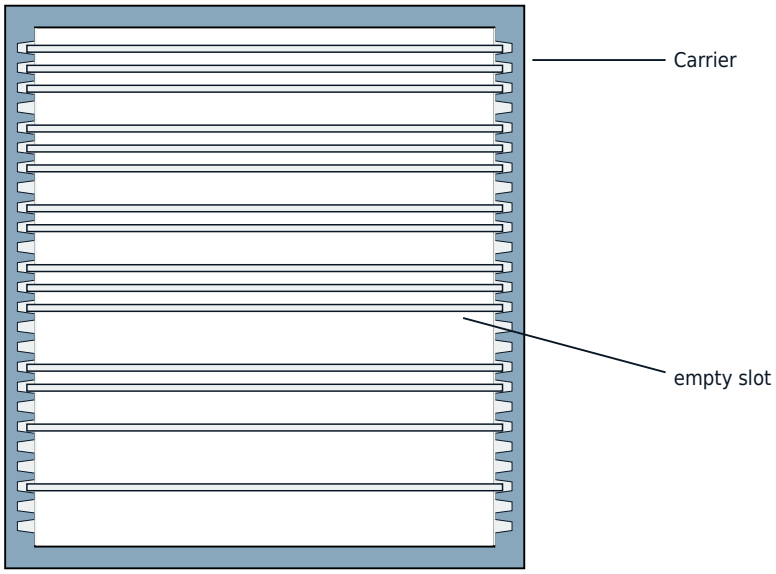
The plasma supplies the energy that would otherwise have to come from temperature



At 250 to 350 °C, the gases would not decompose on their own. That's why the process can also be used over finished aluminum interconnects – where a furnace process would already be far too hot.

Wafer Carrier

Side view: the wafers lie horizontally, stacked on top of each other



The carrier holds 25 wafers; here 15 of them are loaded. The slot is taller than the wafer is thick and widens toward the opening, so the wafer slides in easily.

ALD: Atomic Layer deposition

Atomic Layer Deposition (ALD) is a CVD deposition process for producing thin layers. It is a cyclical process in which several gases are introduced into the process chamber alternately.

Each gas reacts in such a way that the respective surface atoms become saturated, meaning the reaction is self-limiting. The other gas can then continue to react at this new surface. Alternating with the reactive gases, the chamber is purged with inert gases such as argon or nitrogen. A simple ALD cycle, which usually lasts only a few seconds, can proceed, for example, as follows:

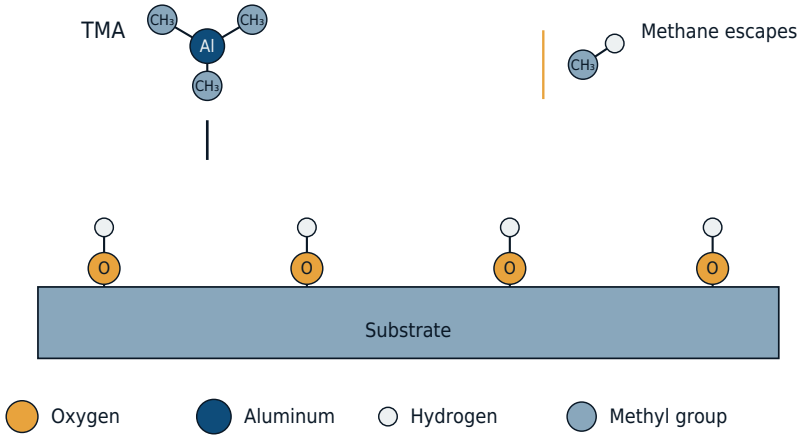
1. self-limiting reaction with the first gas, which saturates the surface atoms
2. purge step with an inert gas
3. self-limiting reaction with a second gas, which reacts at the new surface
4. purge step with an inert gas

A specific example is the deposition of an aluminum oxide layer Al_2O_3, in which trimethylaluminum C_3H_9Al (TMA) and water H_2O are used as the reactive gases.

In the first step, a methyl group CH_3 of the aluminum compound removes near-surface hydrogen atoms from OH groups, forming methane CH_4. The remaining molecules bond to the now unsaturated oxygen atoms.

1. Trimethylaluminum meets the surface

The methyl groups take the hydrogen from the OH groups

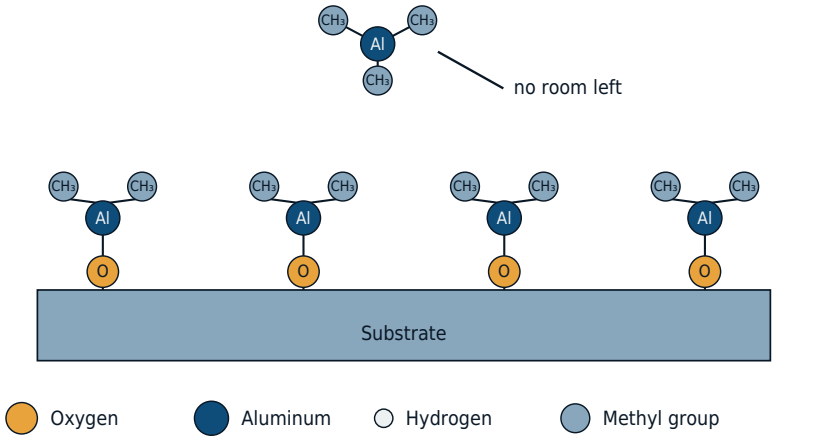


Each methyl group binds one hydrogen atom and leaves the surface as methane.

Once these atoms are saturated, no further attachment of TMA can occur.

2. The surface is saturated

No room left for more TMA – the reaction limits itself

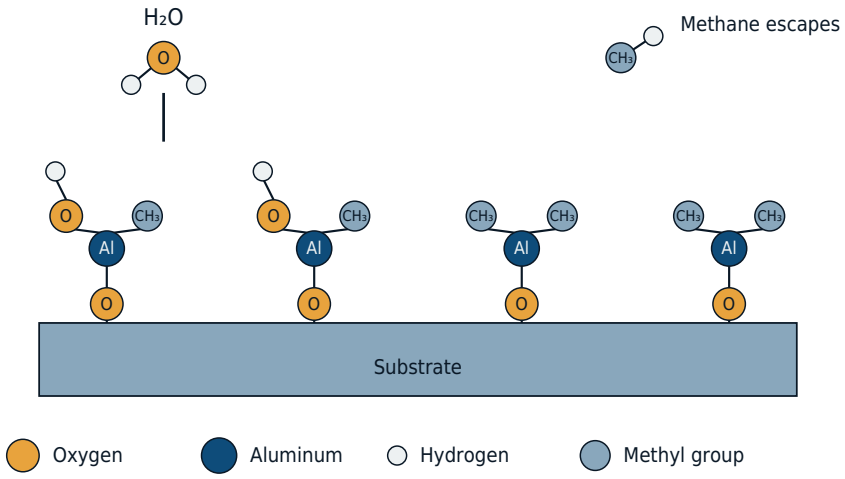


This is exactly what makes the process so precise: the layer grows only one layer at a time.

The process chamber is purged, and water vapor is subsequently introduced into the chamber. One hydrogen atom from each H_2O molecule removes the CH_3 groups of the previously deposited layer, forming methane.

3. Water vapor follows the purge

The water replaces the methyl groups, again forming methane

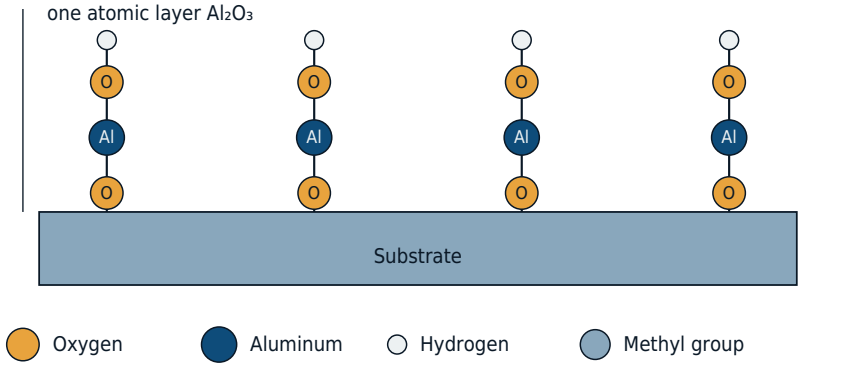


One hydrogen from the water binds the methyl group, leaving the oxygen behind.

What remains are OH^- ions, which now bond with the aluminum, so that once again near-surface hydrogen atoms are available to react with TMA.

4. One layer of aluminum oxide has grown

The new OH groups are the starting point for the next cycle



About a tenth of a nanometer grows per cycle – the thickness follows from the number of cycles.

Atomic layer deposition offers substantial advantages over other deposition techniques, which is why it is an important process for applying various layers. Even 3-dimensional structures (e.g. trenches) can be coated very uniformly. Both insulating and conductive films are possible on various substrates (semiconductors, polymers, etc.). Through the number of cycles, an exact layer thickness is easy to achieve. Since the reactive gases are not introduced into the chamber simultaneously, nucleation cannot occur before the actual deposition

takes place. The quality of the layer is therefore very high.

What ALD is used for today

The price for this precision is speed: only about one-tenth of a nanometer grows per cycle, and a cycle takes seconds. The process is therefore unsuitable for thick layers. Wherever only a few atomic layers matter, however, it is unrivaled:

- Gate dielectric: The silicon dioxide beneath the gate could not be thinned any further without current tunneling through it. It was replaced by hafnium dioxide, which, for the same effect, may be thicker. It is deposited via ALD – otherwise a layer of just a few atomic layers cannot be produced uniformly enough across the entire wafer.
- Spacer layers in multipatterning: In the self-aligned process, the thickness of a deposited layer determines the resulting feature width. It must therefore be adjustable more precisely than any exposure step could ever be.
- Barriers and seed layers: in deep, narrow contact holes, which no other process can line uniformly.
- Storage capacitors: the capacitors of a DRAM cell are narrow, deep trenches whose walls must be completely coated.

If the temperature must remain low, the second reaction step is assisted by a plasma (plasma-enhanced ALD). The radicals from the plasma react more readily than the neutral gas, allowing the cycle to proceed even well below the temperature otherwise required.

Gap fill

One task arises with every deposition process, and none of the methods described so far solves it on its own: filling narrow, deep gaps. Such gaps occur wherever an insulator has to be introduced between structures that already exist – between the trenches of trench isolation, between closely spaced interconnects, between the fins of a transistor.

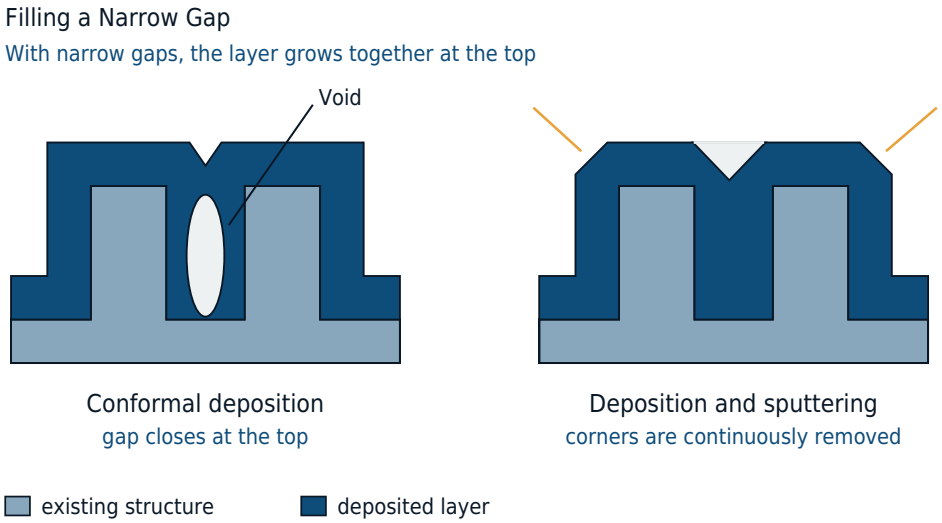
Why a void forms

A conformal layer grows at the same rate on every surface, including the two opposing gap walls. If the gap is narrower than twice the layer thickness, the two fronts meet. That alone would be unproblematic if they met everywhere at the same time. In reality, the layer grows faster at the gap opening, because the opening is accessible from more directions than the bottom. The gap therefore closes at the top first, leaving a void behind underneath.

Such a void is not merely a cosmetic problem. If it is exposed during a later step, for example during polishing or while etching a contact hole, the void fills with etch solution or with metal – in an unfavorable case, this creates a short circuit

between two interconnects.

Depositing and sputtering at the same time



The solution is to continuously remove the corners at the gap opening again while the layer grows. For this purpose, a sputter component is superimposed on the deposition: argon ions are accelerated from a dense plasma onto the wafer and knock material out there. This is the same process used in sputtering for coating – only here the ion beam does not strike a target, but the growing layer itself.

Because sputtering is most effective at oblique incidence, it attacks the edges at the gap opening much more strongly than the horizontal surface at the bottom. This creates a sloped surface at the top edge, referred to as a facet. The opening remains open, and the gap fills from the bottom upward.

Processes of this kind are known as HDP-CVD (high density plasma CVD). The decisive process parameter is the ratio of deposition rate to sputter rate, referred to as the deposition/sputter ratio, or D/S for short: too little sputtering allows the void to form, too much removes the structures themselves. The sputter component is typically 10 to 20 percent of the net deposition rate.

When even this is no longer enough

As depth-to-width ratios continue to increase, even this process reaches a limit. The next step returns to the oldest principle of all: a liquid flows into any gap by itself, regardless of its shape. For this purpose, a silicon-containing precursor is applied that is initially liquid, fills the gap completely, and is only afterward converted into solid silicon dioxide through treatment with water vapor and

temperature. The resulting layer is less dense than a deposited oxide, but it fills gaps that are inaccessible to any directional process.

Physical deposition methods

MBE: Molecular beam epitaxy

The growth behavior in molecular beam epitaxy (MBE) corresponds to that of silicon vapor phase epitaxy. However, in this case the silicon is deposited onto the wafer in a different way.

The process takes place under ultra-high vacuum (UHV, 10^{-8} Pa); the wafer to be processed is held upside down at the top of the process chamber and heated to about 600–800 °C so that the native oxide volatilizes.

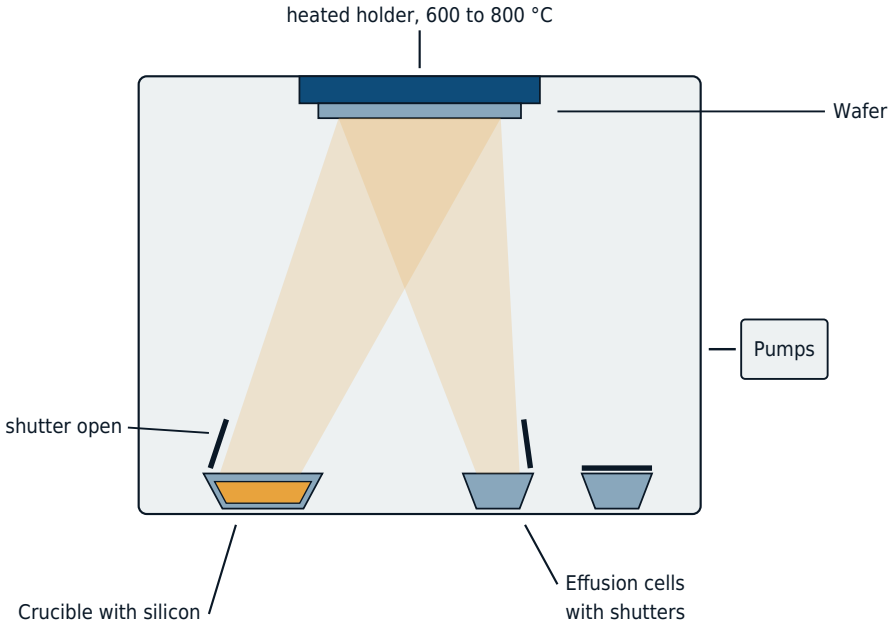
Silicon is evaporated using an electron beam in a crucible and then deposits onto the wafer. Dopants can be evaporated in an effusion source and thus reach the wafer together with the silicon vapor. Through targeted temperature control and the use of shutters, the particle beam can be precisely dosed. This process also makes it possible to grow layers of different materials on top of one another that would otherwise not be possible due to differing atomic sizes. This allows a layer of silicon and germanium to be produced, which is needed for high-frequency applications in bipolar technology.

This is precisely where the strength of the process lies: because the atoms arrive individually and without a chemical reaction, the composition can be changed from one atomic layer to the next. For layer sequences made of compound semiconductors, as needed in laser diodes and high-frequency devices, this is indispensable. In silicon manufacturing, on the other hand, MBE plays no role – there it is too slow.

However, the effort required to generate the UHV in this epitaxy process is very high. To reach this pressure, which is less than one-trillionth of normal atmospheric pressure, several pumps must be combined and the process chamber pumped down for a long time. Only one wafer can be processed at a time, and growth amounts to only about 1 µm per hour.

MBE System

The atoms fly individually through the ultra-high vacuum onto the wafer



Each source can be switched on individually; growth is about one micrometer per hour. This is the only way to change the composition from one atomic layer to the next. The pressure is around 10^{-8} Pa.

Evaporating

In evaporation, metallic layers, such as aluminum, can be produced on the wafers. The material is placed in a crucible made of a high-melting-point metal such as tantalum and heated there until it transitions into the gaseous state. The metal vapor strikes the wafer perpendicularly, so that edges are not covered well; the resulting layer is polycrystalline. As an alternative to melting in a crucible, the metal can also be evaporated using an electron beam. Compared to thermal evaporation, this electron beam method makes it possible to achieve a very precise deposition rate. Since edge coverage is not very good, these two processes are mostly used for full-area backside coating for later electrical contacting.

In the manufacturing of integrated circuits, evaporation has largely been displaced by sputtering, which provides denser layers and better edge coverage. It has been retained where poor edge coverage is precisely what is desired: in the lift-off process, a resist mask is first patterned, then the metal is evaporated, and afterward the resist together with the metal lying on top of it is removed. Because the sidewalls of the resist mask are barely covered, the film breaks cleanly at the edge. This makes it possible to pattern metals that are difficult to etch.

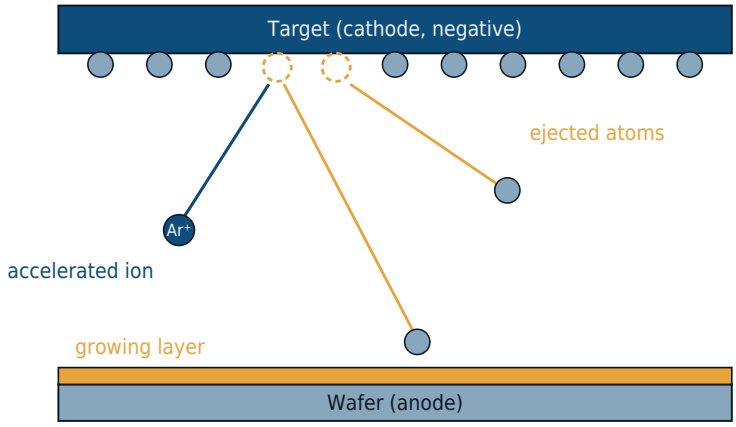
Schematic illustration of an evaporation system

Sputtering

In sputtering, ions (usually argon) are accelerated onto a target and strike out atoms or molecules there; the target consists of the material of the layer to be deposited. The mean free path of these particles is a few millimeters long, i.e. they collide frequently, which means that vertical surfaces on the wafer are also covered well. The deposited particles form a porous layer, which can be densified by annealing. Sputtering can be divided into passive (inert) and reactive sputtering.

Sputtering

Argon ions knock atoms out of the target, which deposit on the wafer



The atoms fly off in all directions and collide with gas particles along the way. That's why even vertical surfaces are well covered – the layer stays porous at first and is later densified by annealing.

In passive sputtering, only the material of the target is deposited on the wafers; depending on the target material, highly pure layers can be produced, since a precise mixing ratio of the substances in the target is possible. In reactive sputtering, a reactive gas (e.g. oxygen O_2) is added to the gas in the process chamber, which combines with the sputtered material from the target and then deposits on the wafer. This makes it possible to produce insulating layers from a metal target (e.g. aluminum Al):

$4\ Al + 3\ O_2 \rightarrow 2\ Al_2O_3$

To produce metallic layers, DC sputtering is used. Here, the ions are accelerated

onto the target with up to 3 kilovolts (kV) and discharge there. Since these charges always have to be dissipated, only a conductive material can serve as the target here. For insulating layers, reactive sputtering must be used. If an insulating layer is to be produced directly from the target, radio-frequency sputtering (RF sputtering) is used.

In RF sputtering, the voltage is applied to one electrode each behind the target (cathode) and the wafer (anode). Due to the high-frequency voltage, electrons are attracted to the target during the positive half-wave, causing the target to charge negatively. This negatively charged target attracts ions, which knock particles out of it. In magnetron sputtering, magnets are additionally placed behind the target so that electrons are deflected into circular paths and can thus ionize argon atoms more frequently, causing a considerable increase in the deposition rate. Since the anode is connected to the process chamber, its potential difference relative to the plasma, averaged over time and referenced to the surface, is substantially smaller than that of the cathode, which is why the ions migrate only toward the target and not toward the wafer.

To increase edge coverage, BIAS sputtering is used, in which a negative voltage is applied to the substrate. As a result, particles of the deposited layer are removed here as well, just as at the target, and the surface is planarized. However, care must be taken that no removal of the substrate itself occurs. This so-called back-etching is also the principle behind most plasma etching systems.

Sputtering into narrow structures

The particles knocked out are uncharged and fly off in all directions. Over a narrow, deep contact hole, this means: most of them strike the edge of the opening, and only a little reaches the bottom. The opening closes up before the bottom is covered.

A remedy is to ionize the sputtered particles themselves. In a dense plasma between the target and the wafer, they lose an electron and can then be accelerated straight down by a voltage applied to the wafer. This directs the material flow and allows it to reach the bottom of deep holes as well. In this way, the tantalum barrier and the copper seed layer are deposited in the damascene process – the seed layer must be present without gaps, on the sidewalls just as much as on the bottom, otherwise the copper will not grow continuously during the subsequent electroplating.

Sputtering is therefore well suited for producing metallic layers with good conformity and very good reproducibility. The effort involved is low, and the reduced pressure of about 5 Pa is fairly easy to generate.

Metallization

Requirements on metallization

Requirements on metallization

In metallization, contacts are made to the doped regions of semiconductor devices, and these are connected by interconnects. From there, the connections are routed to the edge of the individual chips in order to establish the connection to the package or for testing the circuit with probes during manufacturing.

What requirements must metallization layers meet in order to be used in semiconductor technology?

- good adhesion to silicon dioxide
- high current-carrying capacity, low electrical resistance
- low contact resistance between metal and semiconductor
- a process for depositing the conductive layer that is as simple as possible
- low susceptibility to corrosion for a long chip lifetime
- good contactability
- the possibility of multilayer wiring to save chip area
- high purity of the material
- resistance to electromigration at high current densities
- compatibility with chemical mechanical polishing, which is used to planarize every layer
- no diffusion into the surrounding insulation layers or into the silicon

The last three points have only become important with small structures, and it is precisely these that have driven the change of material. Aluminum meets the classic requirements well and was the standard material for decades. However, as interconnect cross-sections shrink, current densities increase, and aluminum migrates under this stress. Since the late 1990s, copper has therefore been the material of choice for wiring in high-performance circuits. This does not mean aluminum has disappeared: bond pads, power and analog devices, as well as non-critical layers, continue to be made with it.

Aluminum technology

Aluminum and aluminum alloys

Aluminum and its alloys were the standard material for on-chip wiring for decades and continue to be used wherever the smallest structures are not critical: for bond pads, in power and analog technology, and on non-critical upper layers. Aluminum offers:

- good adhesion to SiO_2 and interlayer dielectrics such as BPSG and PSG,
- good contactability when wiring to the package (e.g. with gold and aluminum wires),
- a low resistivity of 2.7 µΩ·cm for pure aluminum, and about 3 µΩ·cm for the common alloys with silicon and copper,
- and it can be patterned very well by dry etching.

The last point was long a decisive advantage: aluminum forms volatile compounds with chlorine and can therefore be deposited over the full area like any other layer, masked with resist, and etched. This simple process is precisely what is not available for copper.

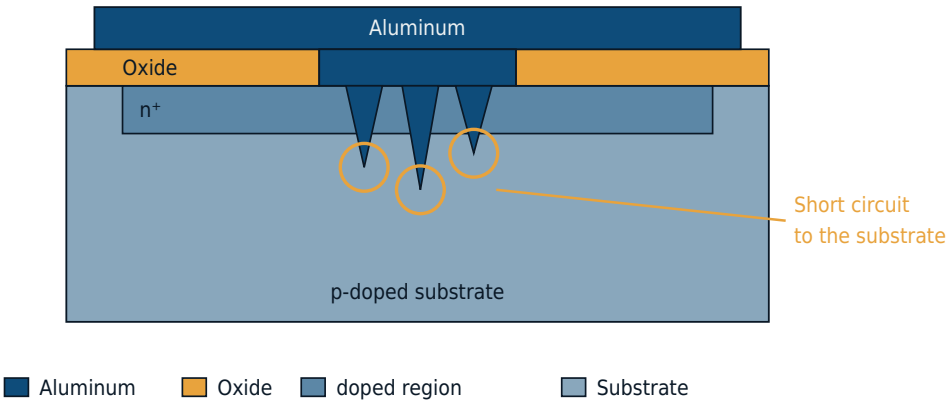
However, aluminum meets the requirements for electrical load capacity and corrosion resistance only partially. The following sections describe the three effects at which aluminum technology reaches its limits: the diffusion of silicon into the metal, electromigration, and the growth of hillocks. Metals such as silver or copper are considerably better in this respect, but cannot be dry-etched – for these, a different patterning process first had to be found.

Diffusion in silicon

The use of pure aluminum can lead to a diffusion of silicon atoms into the metal. Silicon dissolves into solid aluminum well below its melting point; at the usual sintering temperatures of 400–450 °C, which are used to improve the contact after the metal has been deposited, this occurs to a considerable extent. The semiconductor reacts with the aluminum metallization, and the resulting material loss, caused by the outdiffused atoms, leaves pits at the contact area between silicon and aluminum. The aluminum fills these pits. This creates "spikes," which can, under certain circumstances, lead to short circuits if they extend through the doped regions into the silicon crystal.

Spike Formation

The aluminum spikes pierce through the doped region

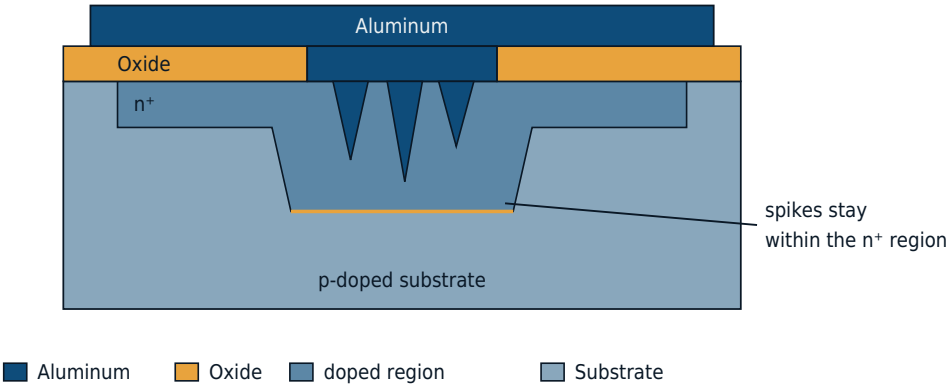


During sintering, silicon dissolves into the aluminum, leaving pits that the metal fills in. If the spikes reach deeper than the doped region, the junction is short-circuited.

The size of these spikes depends on the temperature at which the aluminum is deposited onto the silicon. There are several ways to prevent these spikes. At the location of the contact hole, a deep ion implantation, the contact implantation, can be introduced. This ensures that the spikes do not extend into the substrate.

Contact Implantation

The more deeply doped region catches the spikes



The deep implant beneath the contact hole moves the junction further down. It costs an additional process step and changes the electrical properties of the region.

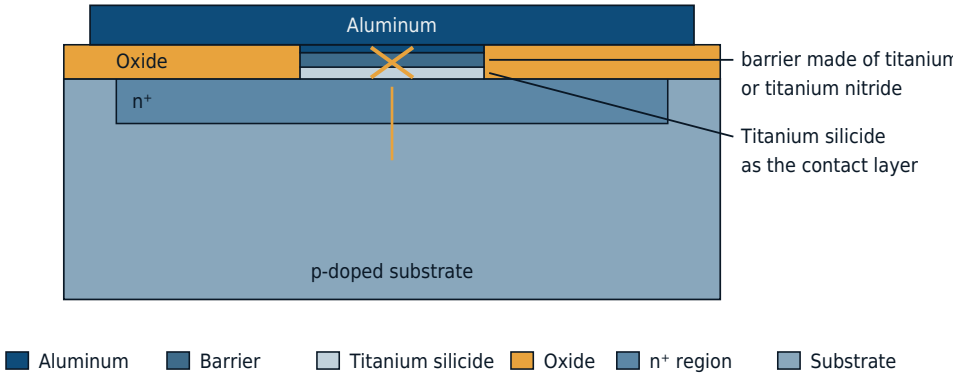
The disadvantage, however, is that an additional process step has to be introduced, and the electrical properties change due to the enlargement of the doped region.

Instead of pure aluminum, an aluminum-silicon alloy containing about 1–2 % silicon can also be used. The aluminum is now already mixed with silicon, and silicon atoms no longer diffuse out of the wafer into the aluminum. With very small contact holes, however, silicon can precipitate at the contact area, resulting in an increased contact resistance.

For high-quality contacts, a separation between aluminum and silicon is required. For this purpose, a barrier of various materials, such as titanium, titanium nitride, or tungsten, is deposited on the silicon. To prevent an increase in contact resistance at the interface between titanium and silicon, a contact layer of titanium silicide must also be applied here.

Barrier Layer Between Aluminum and Substrate

Aluminum and silicon no longer touch



The blocked arrow represents the suppressed diffusion: the barrier holds back the silicon, and the silicide keeps the contact resistance low. Without this interlayer, the transition from titanium to silicon would be too high-resistance.

This barrier is the forerunner of what became standard practice in copper technology: there, a diffusion barrier surrounds every single interconnect. Spike formation itself is thus a historical problem – today, the contact to the silicon is made via a silicide and a tungsten, cobalt, or ruthenium plug, and aluminum no longer touches the silicon at all.

Electromigration

At high current densities (current flow per area), the flowing electrons transfer a small amount of momentum to the ion cores with every collision. Individually, this impulse is insignificant, but in total this "electron wind" slowly pushes the atoms out of their positions in the direction of current flow. Especially at locations with a small interconnect cross-section, the current density is increased; as the atoms shift, the cross-section decreases further, and the current density rises even more. Such constrictions are found in particular at

edges over which the interconnects run. In extreme cases, the aluminum interconnects break due to this material transport.

Damage to interconnects caused by electromigration

(Source: Max-Planck-Institut für Metallforschung Stuttgart)

Electromigration limits how much current an interconnect can carry continuously. Since the cross-section decreases with every shrink, but the current to be switched does not decrease to the same extent, this has tended to become more important over the years. Copper withstands roughly a hundred times the lifetime of aluminum under the same load, because its atoms are more strongly bound. In this case, the material transport shifts to the interfaces: copper atoms migrate preferentially along the top of the interconnect, where it borders the capping layer. A thin cobalt layer on the interconnect, instead of a purely dielectric cap, binds the surface and significantly extends the lifetime.

Hillocks

Electromigration causes material to be shifted and accumulated at locations of lower current density. These so-called hillocks can break through overlying layers and thus cause a short circuit with another metallization layer; in addition, moisture can penetrate through the resulting cracks and lead to corrosion. Hillocks can, however, also arise due to differing coefficients of thermal expansion of the materials. The materials expand differently with temperature changes, creating stresses between the layers. This problem can be solved using compensating layers that have an "intermediate" coefficient of expansion (e.g. titanium, titanium nitride).

Further negative effects that can occur during metallization:

- Stray exposure: due to unevenness, incident light rays during exposure can be reflected in such a way that areas are exposed uncontrollably. Reflection is prevented with the help of an anti-reflective coating (ARC); on metal this is usually titanium nitride, while on other layers silicon oxynitride or an organic resist is used
- Poor edge coverage: at edges, increased layer growth can occur, while at corners growth is reduced. To counteract this, edges can be rounded before metal deposition:

The design of the interconnects therefore has to be planned precisely to prevent these problems. By adding a small amount of copper, the lifetime of aluminum interconnects can be substantially increased, although patterning of aluminum

mixed with copper becomes considerably more difficult. To protect against corrosion, the surfaces are passivated with silicon dioxide or silicon nitride. This passivation is the actual protection: the vast majority of devices are encapsulated in plastic. Ceramic packages remain reserved for applications that must be hermetically sealed, such as in aerospace. How metallization layers are deposited onto the wafer is described in more detail in the chapter Deposition.

Copper technology

Copper technology

Starting at feature sizes below 250 nm, aluminum, even with copper additions, can barely meet the requirements needed for use in integrated circuits any longer. The resistivity of copper, at 1.7 µΩ·cm, is about a third lower than that of aluminum at 2.7 µΩ·cm. For the same cross-section, less voltage therefore drops across a copper interconnect, it heats up less, and signals travel faster. Stress and electromigration are also considerably more pronounced in aluminum: copper has the higher melting point and more strongly bound atoms, so it tolerates significantly higher current densities. Given the continuous miniaturization, a switch to copper was therefore unavoidable.

Copper, however, has the negative property of contaminating almost everything it comes into contact with. In silicon it is extraordinarily mobile and creates defects there that trap charge carriers and render devices unusable. The areas and equipment in manufacturing where copper is processed must therefore be strictly separated from the others, and every copper interconnect is enclosed by a diffusion barrier. In addition, copper corrodes very easily and, like aluminum, must therefore be sealed with a passivation layer.

While in aluminum technology the vias between layers are made with tungsten, in copper technology the metal itself is used for this purpose (only the very first layer, at the contact to the doped silicon regions, requires separation with tungsten). This not only eliminates the negative thermoelectric effects that arise at the junctions between different metal layers, but also the additional process steps needed to deposit multiple layers. Unlike aluminum, however, copper cannot be patterned by dry etching, because its halogen compounds are not volatile at process temperature.

The "ordinary," subtractive process for patterning a layer, as is also used for aluminum, essentially proceeds through the following steps:

- deposit the layer to be patterned
- apply photoresist, expose, and develop it

- transfer the resist mask into the underlying layer in an etch step
- remove the resist
- deposit the passivation layer

For copper, an additive process is used instead: the so-called damascene process.

Damascene process

In the damascene process, the contact holes (VIA = Vertical Interconnect Access) of the individual metallization layers are etched into interlayer dielectrics that already exist and serve for isolation or passivation, and the trenches, in which the copper interconnects will later run, are patterned as well. Copper can then be deposited into the openings using an electrochemical process (electroplating). CVD or PVD depositions with reflow techniques are also possible. The copper is then planarized in a CMP process until the surface is level.

A distinction is made between single- and dual-damascene processes, and within the latter between VFTL (VIA First Trench Last) and TFVL (Trench First VIA Last). The two dual-damascene processes are explained in more detail below.

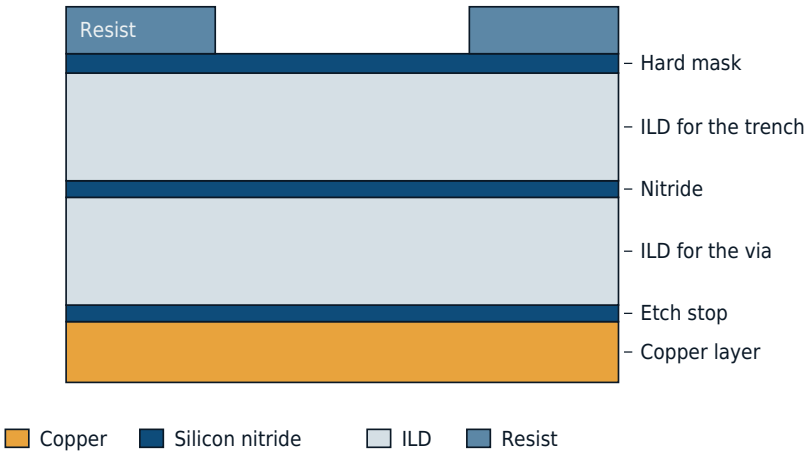
Trench First VIA Last

On the wafer (here on an already existing copper layer), various materials are deposited that serve as isolation, passivation, or protective layers. Silicon nitride SiN can be used as an etch stop and to protect against etch gases. As the interlayer dielectric (ILD), materials with a low k-value, such as silicon dioxide SiO_2, are used. A resist mask is then patterned on top.

1. The wafer is coated with resist, which is patterned in a lithography process.

1. Resist Mask for the Trench

The resist is patterned on top of the layer stack

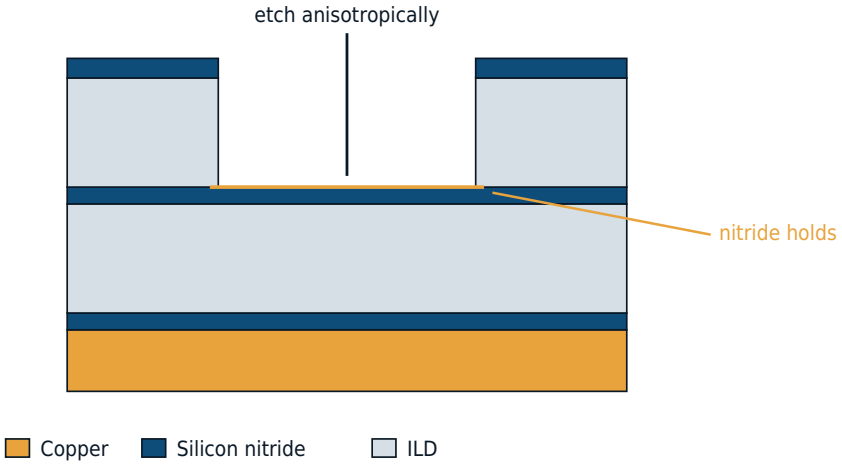


Two roughly equally thick ILD levels are stacked, separated by a nitride layer: the via will later sit in the lower one, the trench in the upper one. The hard mask protects the ILD during resist removal and serves as the CMP stop.

2. In an anisotropic etch process, the hard mask (SiN) and the ILD layer are opened down to the first etch stop (also SiN).

2. Etch the Trench

Hard mask and upper ILD are opened down to the separating nitride



The etch runs through the hard mask and upper ILD and stops on the separating nitride. The trench is therefore exactly one ILD level deep.

The photoresist is removed, and the trench for the interconnect is fully patterned.

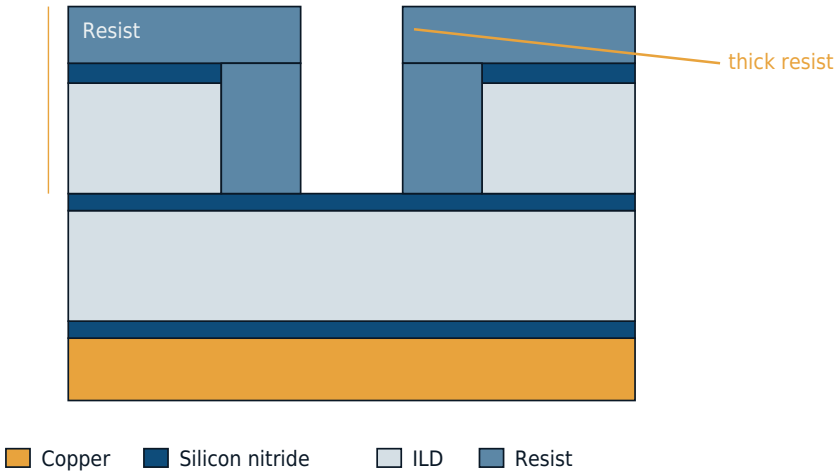
The hard mask at the surface is needed to protect the ILD layer from the plasma during resist removal. This is necessary because the ILD is chemically similar in composition to the photoresist and can therefore be

attacked by the same process gases. In addition, the hard mask serves as a CMP stop during the final planarization of the copper.

3. Next, resist is applied and patterned again.

3. Resist Mask for the Via

The resist fills the trench and must therefore be applied thickly

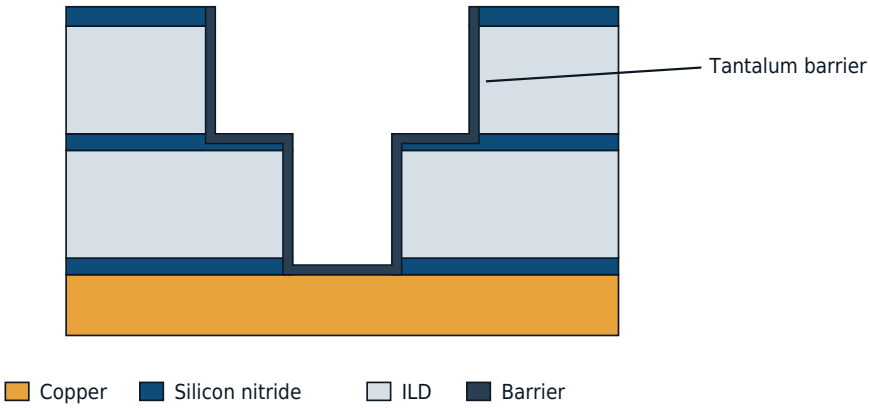


Because the resist also fills the trench, it is noticeably thicker here than in the first step. Exposing a small contact hole through it is the actual drawback of this process.

4. The contact holes (VIA) are then patterned in an anisotropic etch process.

4. Etch the Via and Deposit the Barrier

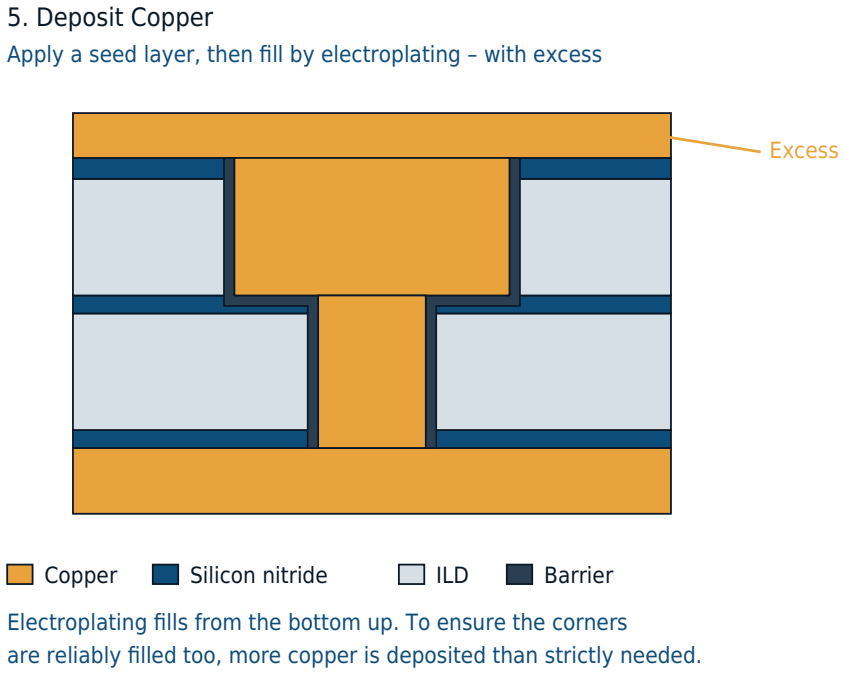
Through nitride, lower ILD, and etch stop, down to the copper



The via passes through the separating nitride, the lower ILD, and the etch stop. This is opened last and with low energy, so no copper gets knocked loose. The barrier lines both openings.

In a low-energy etch process, the etch stop layer is then opened so as not to sputter out any underlying copper, which could otherwise diffuse into the ILD layer. The photoresist is then removed, and a thin tantalum barrier layer is deposited, which prevents the copper deposited afterward from penetrating into the ILD.

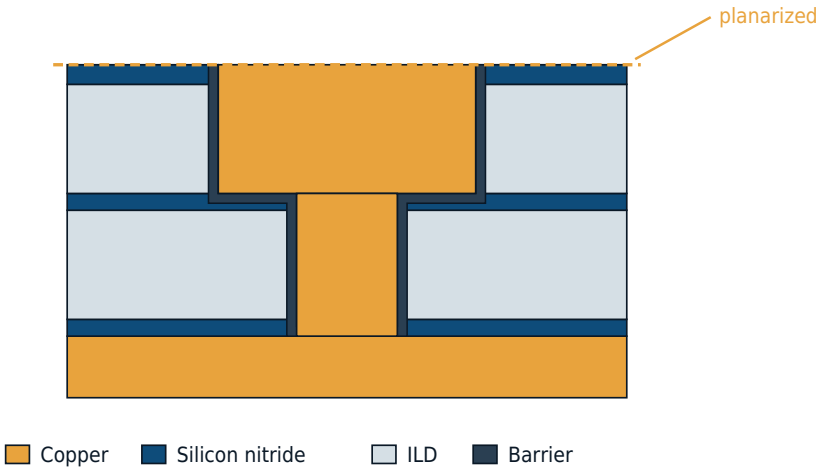
5. A thin copper seed layer is deposited, and the structures are filled with copper in an electroplating process.



6. Finally, the copper is planarized by chemical mechanical polishing.

6. Chemical Mechanical Polishing

The excess is removed, with the hard mask serving as the stop



The interconnect and the via have formed in a single sequence – that's the core of the dual-damascene process. The next level again begins at step 1.

The biggest disadvantage of the TFVL process is the thick resist layer that has to be applied after etching the trenches (step 3). The tiny contact holes are very difficult to produce in such a thick resist layer. The TFVL process is therefore only used for the uppermost metallization layers, where the dimensions of the structures are not as critical as in the lowest layers.

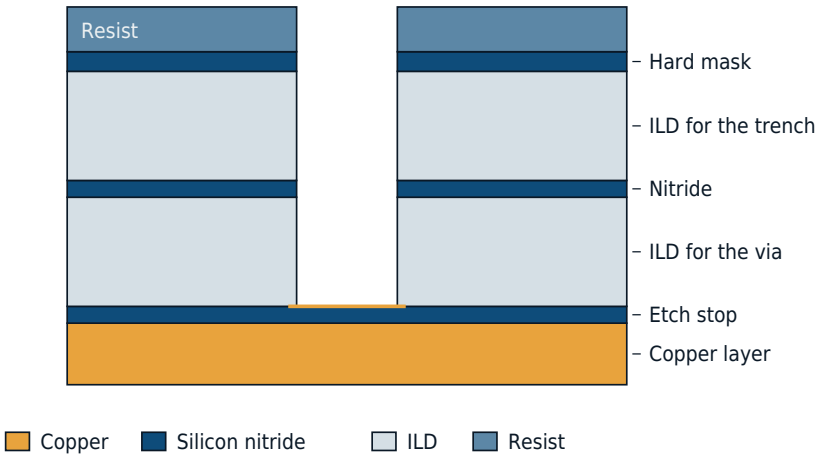
VIA First Trench Last

The VFTL process is similar to the TFVL process, except that the contact holes are patterned first and the trenches afterward.

1. First, a resist mask for the VIAs is patterned, and the contact holes are opened in an anisotropic etch step down to the lowest etch stop. It is important that the etch stop is not etched through, so that the underlying copper is not sputtered out and does not diffuse into the ILD.

1. Etch the Via First

A narrow resist mask; the hole extends through both ILD levels

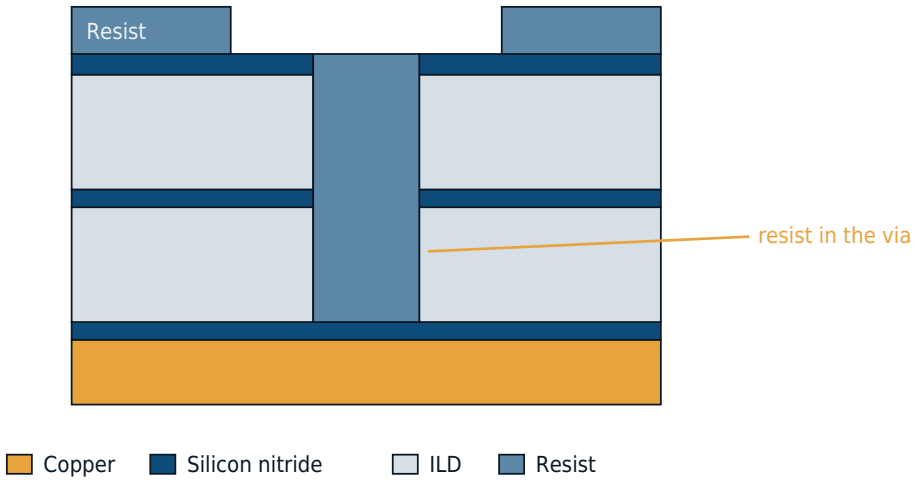


The hole goes straight through both ILD levels. Its lower part later becomes the via, while the upper part is widened into the trench in the next step.

2. The photoresist is then removed, and a new resist mask is patterned for the trenches; in the process, the already opened VIAs are also filled with resist.

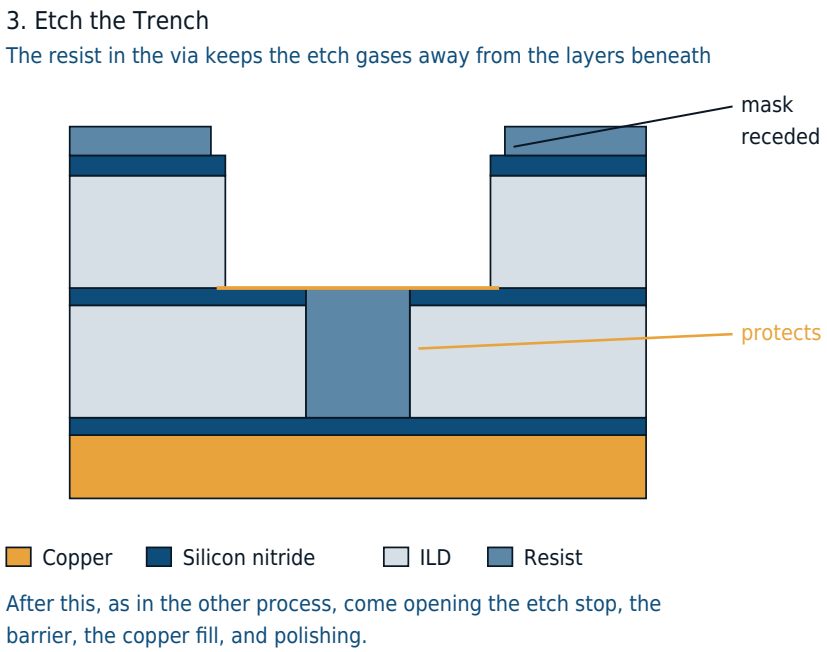
2. Wide Resist Mask for the Trench

The resist simultaneously fills the via and protects it



The resist plug in the hole is the key to this process: without it, the subsequent etch would deepen the via further and break through the etch stop.

3. The lowest etch stop is protected from the etch gases during the subsequent trench etch by the resist plug in the VIA.



Analogous to the TFVL process, the lowest etch stop is then opened, and a tantalum barrier and the copper seed layer are deposited.

4. Finally, copper is deposited and planarized by CMP.

In the single-damascene process, the layers for VIAs and trenches are deposited and patterned separately. This requires two copper processes (each with deposition, patterning, and planarization).

Barrier and liner

The tantalum barrier mentioned above is in reality a stack of layers. Tantalum nitride is applied to the dielectric first, which stops copper diffusion; on top of it comes metallic tantalum, on which the copper seed layer adheres well and grows evenly. Both conduct considerably worse than copper and contribute practically nothing to the current flow.

As long as the interconnects are wide, this does not matter much. At the lowest layers of today's processes, however, the barrier can no longer be made arbitrarily thin without losing its blocking effect – it then takes up a considerable portion of the cross-section. The answer to this is a cobalt liner instead of tantalum, on which copper deposits better, and, for the finest layers, abandoning copper altogether in favor of metals that can do without a barrier. More on this in the section Limits of copper wiring.

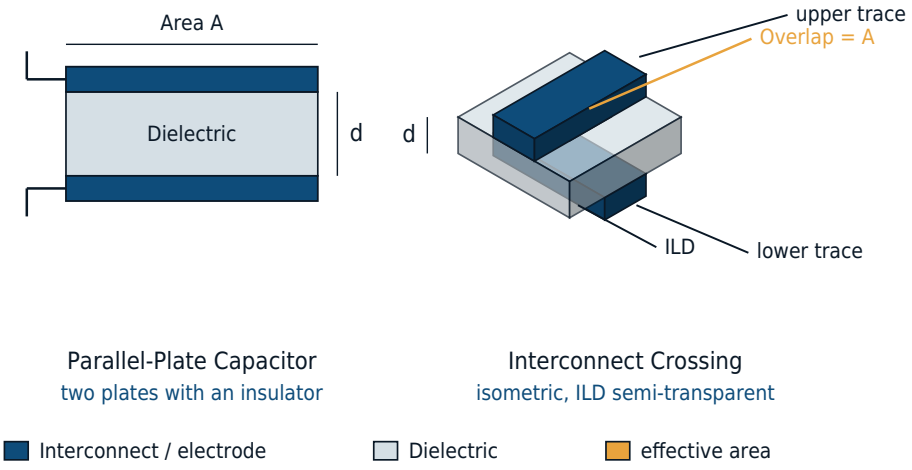
Low-k technology

Due to the continuous shrinking of structures on microchips – on the one hand to reduce power consumption, and on the other to achieve maximum switching speeds – the interconnects used to wire the individual devices move ever closer together, both vertically and laterally. To isolate the interconnects from one another, layers such as silicon dioxide SiO_2 have to be deposited as ILD.

Wherever interconnects run parallel to each other or cross on metallization layers stacked one above another, parasitic capacitances (capacitors) arise. The interconnects form the conductive electrodes, and the SiO_2 lying between them forms the dielectric.

Parallel-Plate Capacitor and Interconnect Crossing

Two crossing interconnects form a capacitor



Both arrangements obey the same formula: capacitance grows with the overlapping area and falls with the spacing. Because both shrink with every generation, only a smaller ε_r helps – a low-k material.

The capacitance C of a capacitor is calculated as follows:

$$C=\frac{\epsilon_r\epsilon A}{d}$$

Here, d stands for the distance and A for the area of the electrodes, i.e. of the overlapping interconnects. ε_0 denotes the absolute permittivity of vacuum, and ε_r (in English often κ (kappa), or simply k) the relative permittivity of the insulator (here SiO_2).

The magnitude of the parasitic capacitance influences electrical properties such as the maximum switching speed or the power consumption of the chip, which is why efforts are made to keep C as small as possible. In theory this is possible by decreasing ε_0, ε_r, and A, or by increasing d. Since, as explained above, d keeps

getting smaller, A is dictated by electrical requirements, and ε_0 is a physical constant, it follows that the capacitance of a capacitor can essentially only be lowered by decreasing ε_r.

Dielectrics with a *low* ε_r are therefore needed: low-k.

The classic dielectric, SiO_2, has a permittivity of about 4. Low-k refers to materials with a value of $\varepsilon_r < 4$; ultra-low-k materials (ULK) with $\varepsilon_r < 2.4$ have been state of the art in manufacturing for years. The permittivity indicates the polarization (displacement of charges within an insulator) in the dielectric, and is the factor by which the charge of a capacitor increases compared to empty space, or by which the electric field inside the capacitor is weakened.

To reduce the permittivity, there are basically two approaches:

- reducing the polarizability of bonds within the dielectric
- reducing the number of bonds

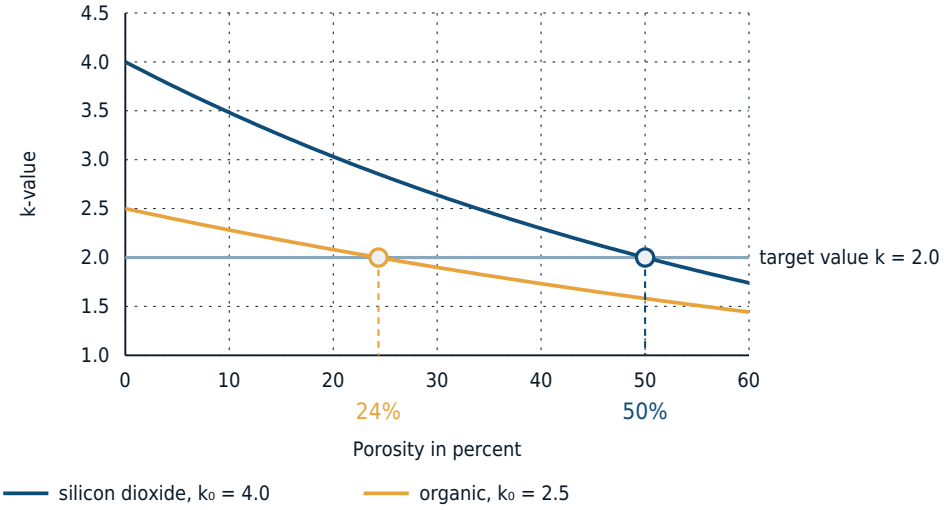
Polarizability can be lowered using materials with fewer polar groups; possibilities include fluorinated (FSG, ε_r about 3.6) or organic (OSG) silicon oxides. However, on its own this is no longer sufficient given ever-shrinking feature sizes, which is why the trend is toward porous layers. Due to porosity, "empty space" is then present within the ILD, which, in the case of air, has a permittivity of about 1. This lowers ε_r for the entire layer. The pores can be created by mixing the ILD material with polymers, which are then driven out of the layer in an annealing step.

However, several problems arise that have to be overcome in order to use these new materials in semiconductor manufacturing.

Pores in the material reduce its density, resulting in lower mechanical stability. In the case of SiO_2, about 50 % porosity would have to be introduced into the material to achieve a k-value of 2.0. If an organic material is assumed whose k-value without pores is 2.5, a porosity of about 22 % is sufficient.

k-Value as a Function of Porosity

The lower the starting value, the fewer pores are needed



Calculated using the logarithmic mixing rule k = k₀^(1−p):
A material with a low starting value needs markedly fewer pores.
However, pores make the layer mechanically softer and harder to process.

(Source: Semiconductor International)

Likewise, process gases or copper from the interconnects can more easily penetrate the porous layer and damage it, causing the permittivity to rise again. To counteract this, the pores must be distributed as evenly as possible and must not be interconnected. To prevent copper from diffusing into the ILD, a thin diffusion barrier has to be deposited in an additional process step; however, care must be taken that this material does not increase the permittivity.

Just like the photoresist used in semiconductor manufacturing, the organic silicon oxides are also composed of hydrocarbon groups (CH). If the resist is removed after patterning the dielectric using an oxygen plasma, the plasma also attacks the dielectric. Here too, additional protective layers (SiN, as described in the section Damascene process) have to be applied to prevent the insulating layer from being damaged.

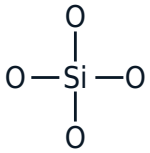
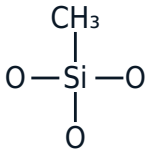
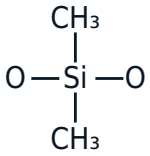
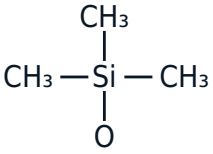
When no material helps anymore: air gaps

At the end of this development lies the elimination of the dielectric altogether. If the material between two closely spaced interconnects is deliberately removed and the resulting gap is bridged at the top by a deposition with poor coverage, a cavity remains – with a permittivity of 1, the best possible insulator. Such air gaps are used only at selected locations where the capacitance is especially problematic, since they further weaken the structure mechanically and make

heat dissipation more difficult.

This is precisely where the real difficulty of the whole low-k development lies: every step toward a smaller k-value makes the layer more porous and thus softer. Yet it still has to withstand chemical mechanical polishing and later endure the forces that act on the chip during package assembly and with every temperature change in operation.

Overview of various organic silicon oxides

Chemical formula	Chemical structure	k-value
SiO_2		4.0
$SiO_{1.5}CH_3$		3.0
$SiO(CH_3)_2$		2.7
$SiO_{0.5}(CH_3)_3$		2.55

Limits of copper wiring

The switch to copper bought the wiring roughly two decades of headroom. In the meantime, however, wiring has become a bottleneck again – for a reason that is not apparent from the material itself: copper gets worse the narrower the interconnect becomes.

Resistivity is no longer a constant

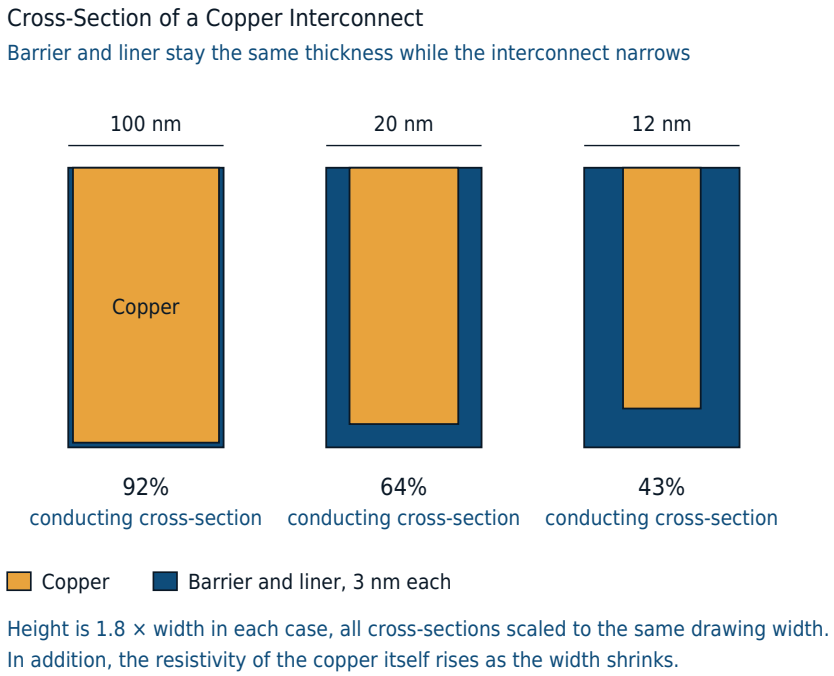
The value of 1.7 µΩ·cm applies to an extended piece of copper. It is based on the fact that electrons travel a certain mean free path between two collisions; in

copper, at room temperature, this is about 40 nm. If the interconnect is narrower than this path length, the electrons no longer collide predominantly within the crystal, but instead at the sidewalls and at the grain boundaries. Both types of collision add to the normal resistance, and the effect grows the tighter the confinement becomes. At interconnect widths around 20 nm, the effective resistivity is already several times higher than the textbook value.

The barrier eats into the cross-section

On top of this comes a purely geometric problem. Every copper interconnect has to be enclosed by a diffusion barrier, and its thickness cannot be reduced indefinitely without it losing its blocking effect. As the interconnect becomes narrower, the barrier thus remains nearly the same thickness and claims an ever-larger share of the cross-section – a share that contributes nothing to the current flow.

Cross-section of a copper interconnect at three widths



Both effects act in the same direction and reinforce one another: the core becomes smaller, and at the same time the copper within it conducts worse. As a result, the resistance of an interconnect rises far more steeply than the mere shrinking of the cross-section alone would suggest.

Ways out of the problem

Other metals
: Cobalt and ruthenium have a higher resistivity than copper, but a shorter electron mean free path – so they degrade less severely at small dimensions. Above all, they require only a very thin barrier, or none at all. Below a certain width, an interconnect made of these metals therefore conducts better overall than one made of copper, even though the material itself is inferior. They are already used in the lowest layers and in the contacts.

Less capacitance instead of less resistance
: Since the signal delay depends on the product of resistance and capacitance, it also helps to lower the capacitance – through porous dielectrics and air gaps.

Moving the power supply to the backside
: A considerable portion of the wiring does not serve signal transmission at all, but power delivery. If these lines are moved to the backside of the wafer, which up to now has served only as a carrier, and routed through the substrate to the top, this relieves the fine layers on the front side in two ways at once: the signal lines gain more room, and the supply lines on the backside can be made considerably wider, and thus lower in resistance.

None of these approaches solves the underlying problem; they merely shift it. Unlike with transistors, where shrinking makes the devices faster, shrinking makes the wiring worse – and its share of delay and power loss grows with every generation.

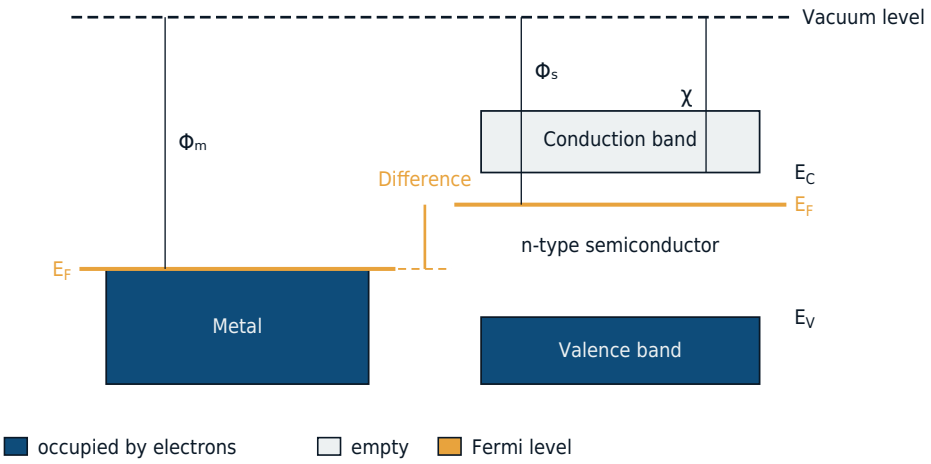
Metal semiconductor junction

n-type semiconductor contact

Whether and in which direction electrons flow at the contact is determined by the work function of the two materials – the energy required to release an electron from the Fermi level out of the material. If the work function of the metal is greater than that of the n-type semiconductor (as is the case, for example, with aluminum on silicon), the Fermi level of the metal lies energetically lower: upon contact, electrons flow from the silicon into the metal, since they can occupy energetically more favorable states there. In the opposite case – work function of the metal smaller than that of the n-type semiconductor – no barrier forms; instead, an ohmic contact is established right from the start.

Band Diagram Before Contact

The metal has the larger work function, so its Fermi level sits lower



Φ is the work function, χ the electron affinity. Because Φ_m is larger than Φ_s, the metal's Fermi level sits lower. On contact, electrons therefore flow from the semiconductor into the metal until both levels are equal.

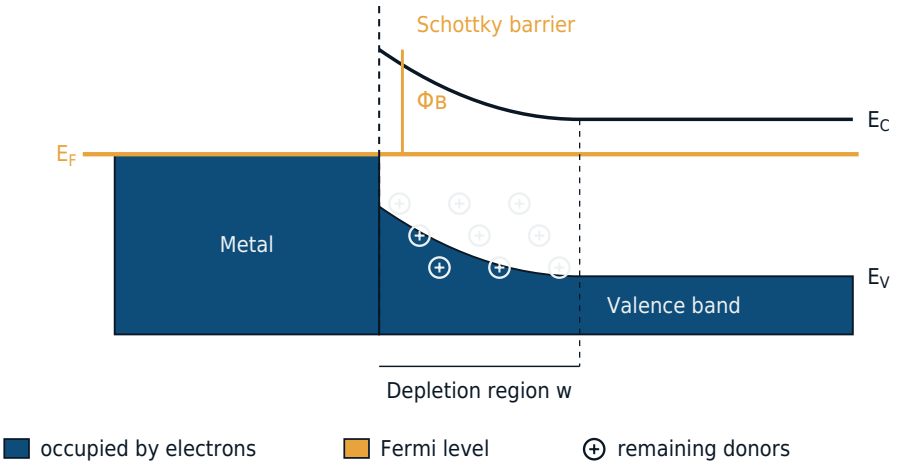
Thus, the probability of electrons being present in the conduction band of the semiconductor decreases: the distance between the conduction band edge and the Fermi level – which describes the highest energy state still occupied by electrons – increases at the interface.

Due to the departed negative charge carriers, positive donors (e.g. phosphorus ions) remain behind, and a space charge region is formed. The bending of the conduction band illustrates the voltage barrier (Schottky barrier) that the remaining electrons in the n-type conductor must overcome in order to flow into the metal.

When metal and semiconductor come into contact, the Fermi levels align through diffusion processes; in the region of the interface, the Fermi level is constant.

Band Diagram After Contact

The Fermi levels align, and the bands bend



occupied by electrons Fermi level remaining donors

The departed electrons leave behind positive donor cores. Their space charge bends the bands – parabolically, as required by Poisson's equation. The more heavily doped, the narrower w becomes.

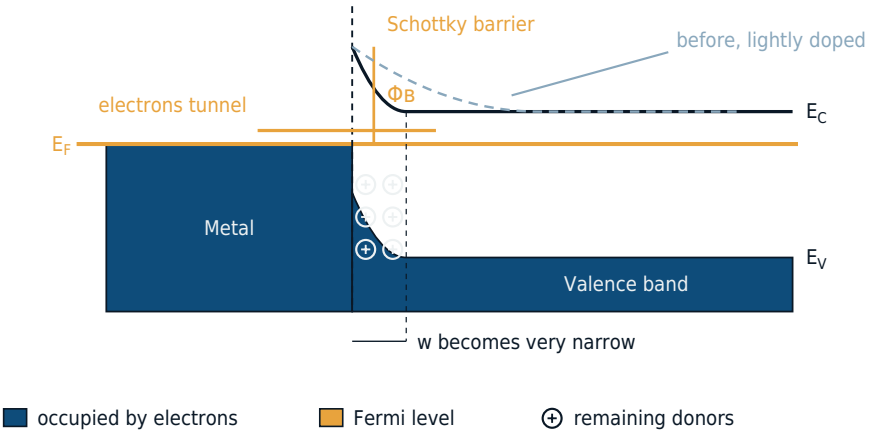
The width *w* of the depletion zone depends on the strength of the doping. The electrons that have migrated out of the semiconductor generate a negative space charge in the metal, which is confined to the surface region.

This metal-semiconductor contact exhibits a nonlinear current-voltage characteristic, known as a Schottky diode. Electrons can overcome this barrier either through thermal energy from outside or by "tunneling" through it under an applied electric field (according to quantum theory, a particle can cross a region in which it cannot exist for energetic reasons by, to put it in simplified terms, briefly borrowing energy to overcome the barrier and then returning that energy: the tunneling effect). This effect can also be observed with aluminum. Since aluminum always oxidizes at the surface, two aluminum surfaces placed against each other would have an insulating effect. However, a current flow can be observed, which is based on the tunneling effect.

Depending on the application, one may want to produce this diode effect or prevent it. To create an ohmic contact, i.e. a contact without this potential barrier, the contact area can be heavily doped (n^+ doping) so that the depletion zone becomes very thin and the metal-semiconductor contact exhibits a linear current-voltage relationship as a result of the tunneling effect.

Band Diagram After n⁺ Doping

The barrier stays the same height but becomes thin enough for electrons to tunnel through



The heavy doping compresses the space charge into just a few nanometers. The barrier is therefore unchanged in height, but so thin that the electrons tunnel through it – the contact becomes ohmic.

Since aluminum is incorporated into silicon as an electron acceptor (it accepts electrons), forming a p-doping at the interface, an ohmic contact results with p-doped silicon. In an n-doped region, however, the aluminum causes a doping reversal, so that a p-n junction is formed here: a diode. There are two ways to avoid this:

- the n-doped region is doped so heavily that the aluminum only weakens it but does not reverse it
- an intermediate layer of titanium, chromium, or palladium prevents the re-doping of the n-doped regions

To improve the contacts, metal silicides (metals combined with silicon) can also be applied to the contact surface.

In contrast to the diode at the p-n junction, where the switching speed relies on the diffusion of electrons, Schottky diodes have very short switching times. They are therefore suitable as protection diodes to absorb voltage spikes.

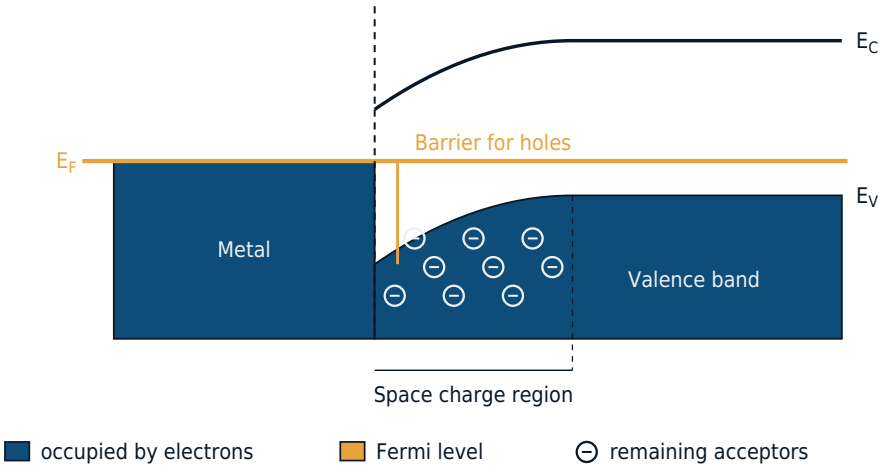
p-type semiconductor contact

In metal-p-type semiconductor contacts, an exchange of charge carriers between the metal and the semiconductor results in a downward band bending; holes in the semiconductor recombine with electrons from the metal. Due to the reduction of the hole concentration, a negative space charge region forms in the semiconductor crystal. The distance between the valence band edge and the Fermi level – representative of the highest occupied states by electrons – increases at the interface, and the resulting potential barrier at the valence

band edge prevents further movement of the holes, which – complementary to electrons – seek to occupy the energetically highest states.

Band Diagram After Metal–p-Type Semiconductor Contact

The bands bend downward, the barrier sits at the valence band



occupied by electrons Fermi level remaining acceptors

Holes from the semiconductor recombine with electrons from the metal. What remains are negatively charged acceptor cores, whose space charge pulls the bands downward. This increases the gap from the valence band to the Fermi level.

Without an external voltage, the diffusion processes come to a halt. Here too, the Fermi levels align with each other in thermodynamic equilibrium.

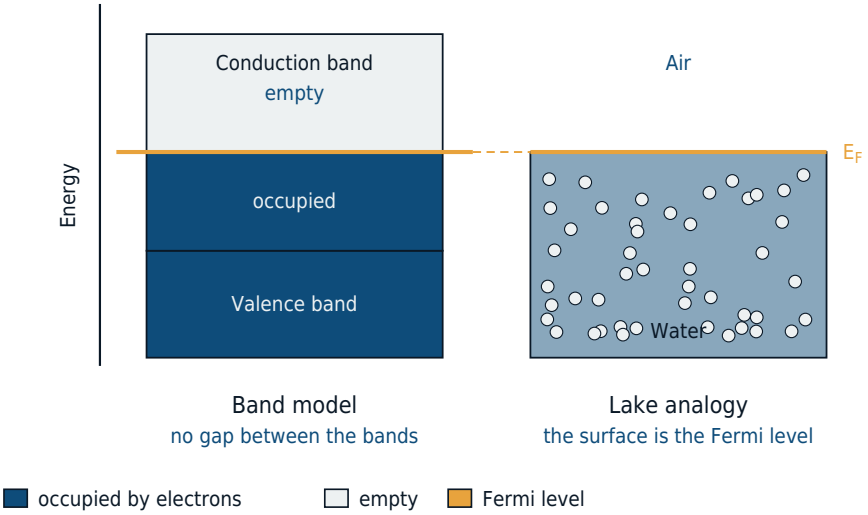
Contacting doped semiconductors

After the transistors have been fabricated in the silicon substrate, they have to be connected to one another by means of electrical contacts. On the one hand, the gate electrode is contacted to control the transistor; on the other hand, the doped source and drain regions, through which the current flows, have to be addressed. Problems arise here at the contact areas where the metallization meets the silicon, since, depending on the doping type of source and drain, there is either a lack of electrons (p-doped) or a surplus of electrons (n-doped).

The Fermi level plays an important role here. The Fermi level is the energy level up to which electrons are still present at absolute zero temperature (-273.15 °C). In conductors, electrons occupy the valence band as well as the energetically higher conduction band, so the Fermi level lies at the level of the conduction band. As an illustration, consider the surface of a lake: the water molecules beneath it represent the electrons, which extend up to the surface – the Fermi level.

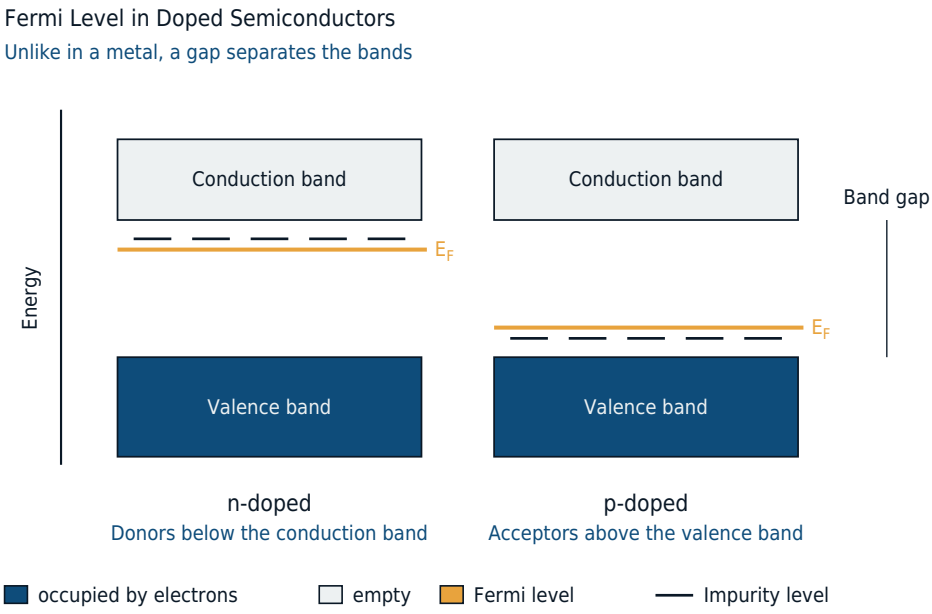
Fermi Level in Metals

The two bands border each other with no gap; everything up to the Fermi level is occupied



At absolute zero, all states up to the Fermi level are occupied and all those above are empty. Because it sits in the middle of the conduction band for a metal, a tiny amount of energy is enough to lift electrons into free states – the metal conducts.

In doped semiconductors, foreign atoms are present in the crystal lattice as donors or acceptors. In n-doped semiconductors, the Fermi level lies close to the conduction band edge, since the donor atoms can provide free electrons even with only a small input of energy. Correspondingly, in a p-doped semiconductor the Fermi level lies close to the valence band edge, since electrons from the valence band of the silicon crystal can easily be captured by the acceptor atom (see chapter Doping).

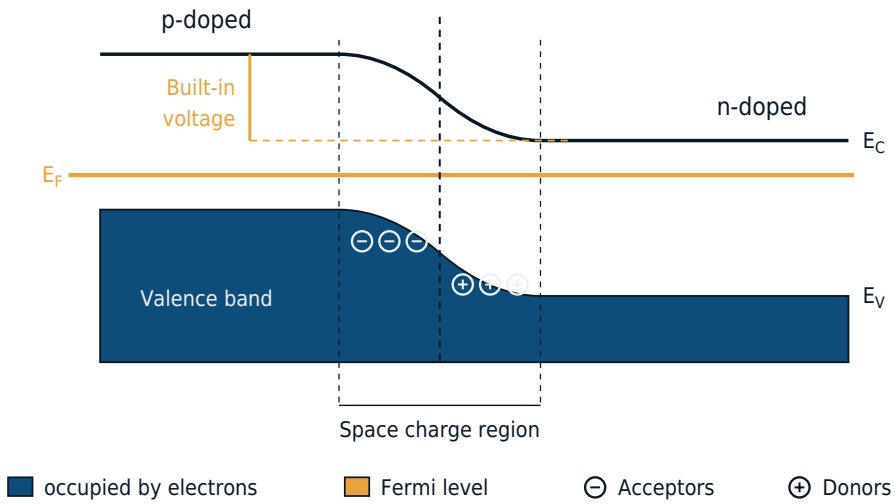


Band model of a p-n junction

From the fact that the Fermi level must be constant (otherwise electrons would flow to locations with a lower Fermi level, occupy free states there, and thereby raise the Fermi level again), a bending of the bands also results at the p-n junction. This illustrates the space charge region that forms as a result of the departed majority charge carriers and the remaining fixed dopant atoms; in other words, the potential barrier that, in the equilibrium state (without an applied voltage), prevents further diffusion of electrons and holes into the respective other crystal. In silicon, the diffusion voltage needed to overcome this potential drop is about 0.7 V.

Band Diagram at the p-n Junction

The Fermi level is constant, which is why the bands bend



Electrons and holes migrate across the junction and recombine. What remains are fixed dopant atoms, whose space charge bends the bands. The resulting potential barrier stops the diffusion; for silicon, about 0.7 V must be overcome for this.

Wiring

Multilayer wiring

The wiring can take up more than 80 % of the chip area in an integrated circuit, which is why techniques have been developed to stack the wiring in several layers on top of one another. This allows the total length of the interconnects to be reduced by up to 30 % with just one additional layer. In a modern processor, the interconnects add up to a total length of several tens of kilometers.

Insulation layers are deposited between the wiring layers, and the individual layers are connected to one another through contact openings (VIA, vertical interconnect access). Today, about ten to twenty wiring layers are common.

These layers are not all built the same way, but are graded according to their function:

Local wiring
: The lowest layers connect neighboring transistors over short distances. They have the finest dimensions, the highest resistance per unit length, and are produced using the most demanding lithography processes.

Intermediate layers

BPSG reflow

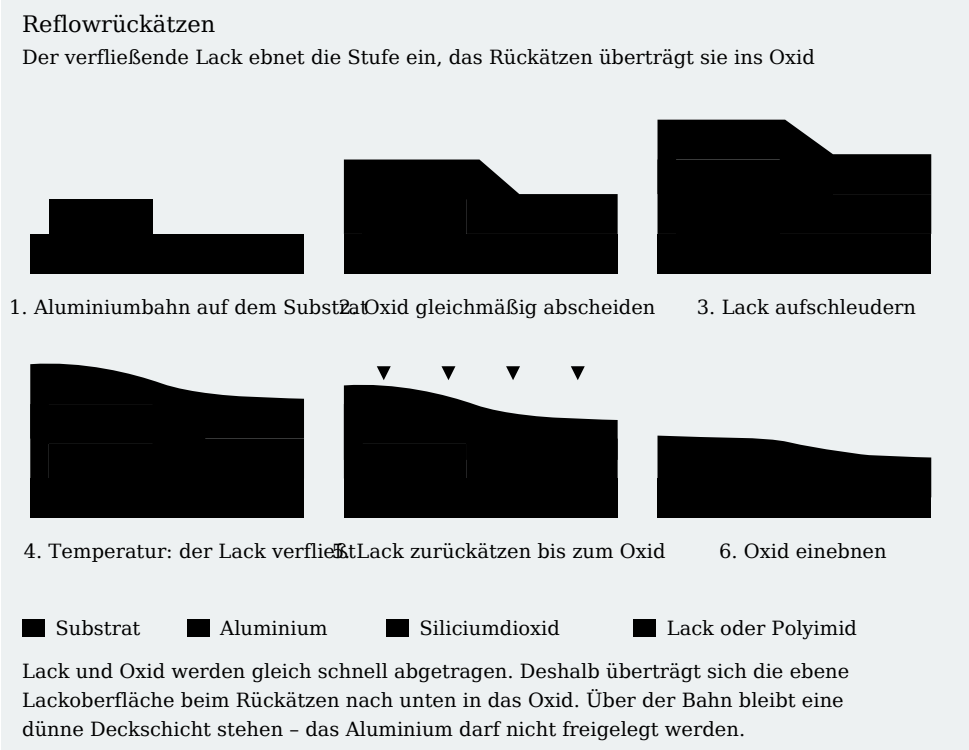
In the reflow technique, layers of doped glasses are deposited onto the wafer. Widely used are phosphosilicate glass (PSG) and borophosphosilicate glass (BPSG). In a high-temperature step, the glasses flow and form a level surface: for PSG and BPSG, this happens at about 900 °C. However, this technique is not suitable for planarizing a wiring layer, since the aluminum would melt under these high temperatures.

Its range of application is therefore narrow: the process can only be used before the first metal layer, as long as no metal is yet present on the wafer. Even there, it has largely been superseded, because the temperature alters the doping profiles already introduced, and because the reflow only planarizes locally – the surface remains uneven across the entire wafer. The chemical mechanical polishing process accomplishes both better.

Reflow back etching

A silicon dioxide layer is deposited onto the wafer, at least as thick as the highest step on the wafer. A resist or polyimide layer is then spin-coated onto the oxide layer and thermally treated for stabilization (see photolithography); the temperature causes the layer to flow.

In a dry etch process, the resist or polyimide and the silicon dioxide are removed at equal etch rates, leaving behind a planarized oxide surface.



In addition to the technique using resist or polyimide, so-called spin on glass (SOG) can also be applied to the wafer. Likewise deposited by spin coating, this directly produces a planarized layer, which is stabilized by a thermal treatment. In this case, the preceding silicon dioxide layer is not required. However, these techniques do not provide uniformity across the entire wafer, but only level out steps locally.

This is precisely where they failed. They smooth the area surrounding an individual step, but leave the large-scale height differences across the wafer unchanged – and it is precisely these that interfere with the exposure of fine structures. Both processes have therefore been superseded by chemical mechanical polishing and are today only of historical interest.

Chemical mechanical polishing

In chemical mechanical polishing (also chemical mechanical planarization, CMP for short), a uniform surface is achieved across the entire wafer, unlike with reflow techniques. This is especially important with regard to lithographic processes, which require as planar a surface as possible for correct exposure. Likewise, a wafer surface without topography is advantageous for all subsequent layers.

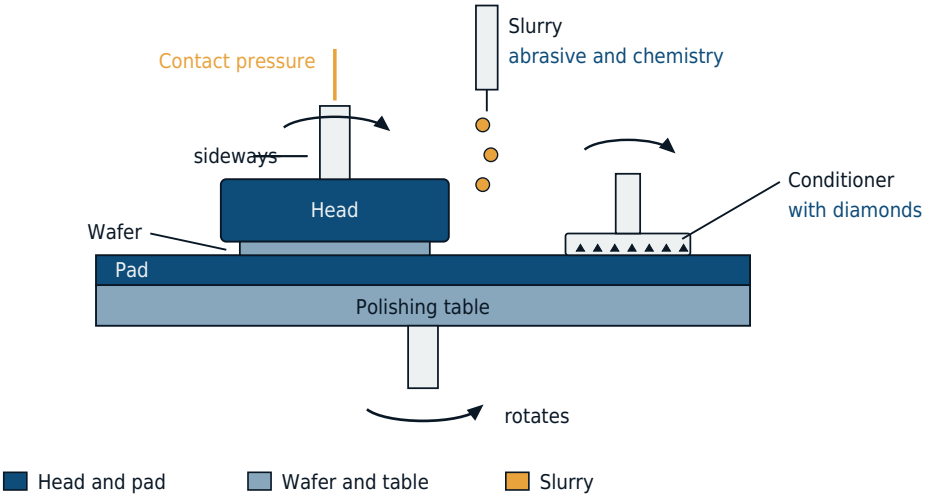
For this purpose, the wafer is held with its active side facing down in a chuck

with vacuum suction (head) and pressed onto a polishing surface (pad, usually made of polyurethane) on the polishing table. The head and the polishing table rotate, while the head can simultaneously perform horizontal movements. A solution (slurry) of abrasives and chemical substances serves as the polishing medium between the table and the wafer; under pressure, these substances alter the surface and thus support the polishing process. To better distribute the slurry and to condition the polishing cloth, the pad can be roughened using a steel disc studded with diamonds (dresser/conditioner). This is done either during the polishing step (in-situ) or before/after it (ex-situ).

The CMP process is usually carried out in two or three stages, on pads with different surfaces and using different slurries. For this purpose, the wafers are transferred to the next pad after each polishing step. Afterward, a cleaning step removes particles and slurry residues.

CMP Tool

The wafer is pressed face-down onto the rotating pad



Head and polishing table rotate in opposite directions, and the head also moves sideways across the pad. Only this combined motion ensures that every point on the wafer is removed at the same average rate.

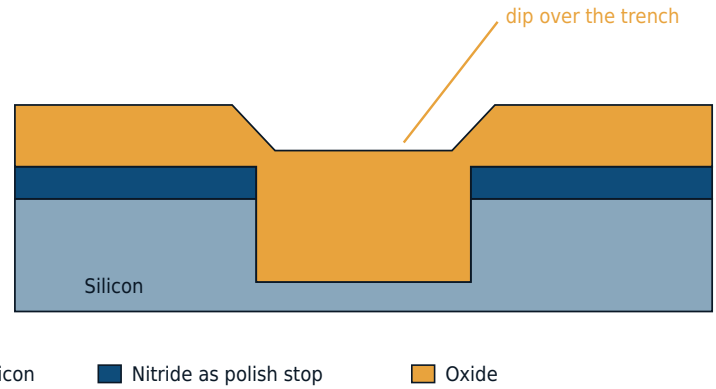
Typically, the CMP process is used after depositing the TEOS for shallow trench isolation, to remove the oxide until only the isolation between the active areas of the transistors remains. Likewise, the interlayer dielectric between the transistor level and the first metal layer (first contact) is polished back to the required thickness using CMP. In this oxide, the contacts to the source and drain regions are subsequently made using tungsten. Here too, chemical mechanical polishing serves to remove the metal on the surface. As described in the chapter Damascene process, the copper wiring layers are likewise

planarized in a CMP process.

The following describes the CMP process with two polishing steps in the STI area. After the first polishing step, the oxide on the active area and over the trenches is planarized. In the second polishing step, the remaining oxide is then removed in a selective process down to the passivation layer. It is important here that the oxide be completely removed from the areas where the transistors will later be fabricated, since otherwise the nitride, which protects the underlying silicon during polishing, cannot be removed by wet chemistry.

1. Trench overfilled with oxide

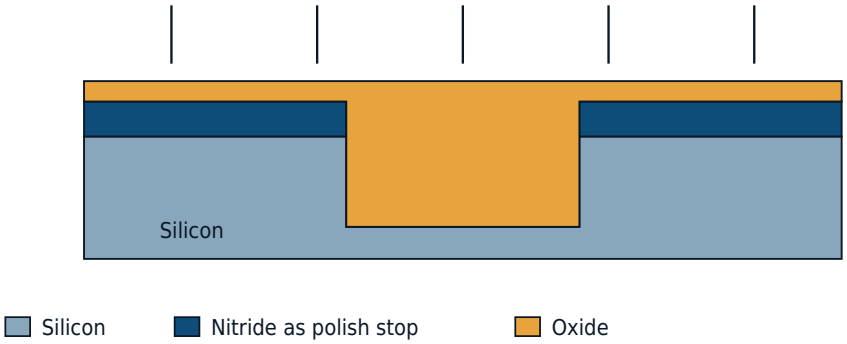
The oxide follows the topography and sits highest over the nitride



Significantly more oxide is deposited than the trench can hold – otherwise a dish would remain after polishing.

2. First Polishing Step

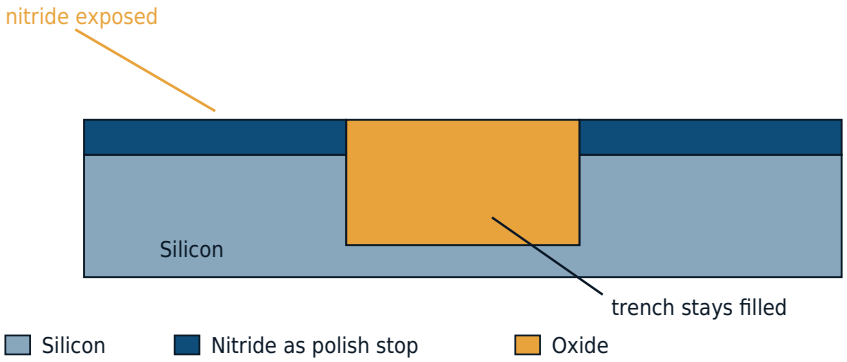
The surface is flat, with a residual layer still over the nitride



The first step removes material aggressively and levels the surface. It deliberately stops short of the nitride, so that it isn't attacked.

3. Second Polishing Step

Selective down to the nitride – no oxide remains on the active areas



If oxide remained on the nitride here, the nitride could no longer be removed wet-chemically afterward – so this step must be complete.

Dishing and erosion

The process has a peculiarity that extends all the way into circuit design: it removes soft areas faster than hard ones. Over a wide copper area, the polishing pad is pushed into the trench and hollows it out (dishing); in areas with many closely spaced interconnects, the entire region is lowered relative to its

surroundings (erosion). Both create exactly the kind of unevenness that polishing is supposed to eliminate, and neither depends on the process itself but rather on how the layout looks.

The countermeasure is unusual: metal structures with no electrical function are inserted into empty areas of the layout, serving only to ensure that the metal density is uniform across the chip. These fill structures are generated automatically and, on some layers, make up a considerable portion of the pattern.

Even though this process may seem rather crude, it is nevertheless capable of producing a surface that is planar to within a few nanometers. It is by no means a special, occasional step anymore: in a modern process flow, a wafer is polished several dozen times.

Contacting the metallization layers

To connect the metal layers, contact holes are etched into the insulation layers with very high anisotropy, so that edges at the contact holes are avoided. The contact holes have to be filled in such a way that, on the one hand, optimal contacting is ensured and, at the same time, the surface remains planar.

Tungsten has proven suitable for filling the contact holes. With the addition of silane, a thin layer of tungsten is deposited from tungsten hexafluoride as a nucleation seed in a CVD process; silicon tetrafluoride and hydrogen fluoride are exhausted as byproducts:

$4\ WF_6 + 3SiH_4 \rightarrow 4W + 3SiF_4 + 12HF$

With the addition of hydrogen to tungsten hexafluoride, the contact holes are then filled:

$WF_6 + 3H_2 \rightarrow W + 6HF$

The advantage of this process is that deposition from the gas phase fills even narrow, deep holes evenly from the inside out, whereas an evaporated or sputtered metal would close off the opening at the top and enclose a cavity. Tungsten is therefore deposited over the full area and subsequently polished back using CMP until it remains only in the holes.

The next metal layer can then be deposited on top, patterned, and planarized. When copper is used as the metallization, tungsten is only needed for the first contact to the silicon substrate. The connection of the individual copper layers is made with the copper itself.

The contact as a bottleneck

As dimensions shrink, the problem shifts. The resistance of a contact is made up

of the resistance of the fill metal, that of the barrier, and the transition resistance to the silicon. The first two components increase rapidly as the cross-section shrinks, because tungsten requires a comparatively thick adhesion layer of titanium and titanium nitride, which itself conducts poorly. In the lowest layers of modern processes, tungsten is therefore being replaced by cobalt or ruthenium: both have a higher resistivity than tungsten, but require only a very thin adhesion layer, or none at all. For the smallest contacts, the fill therefore conducts better overall, even though the material itself conducts worse.

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:

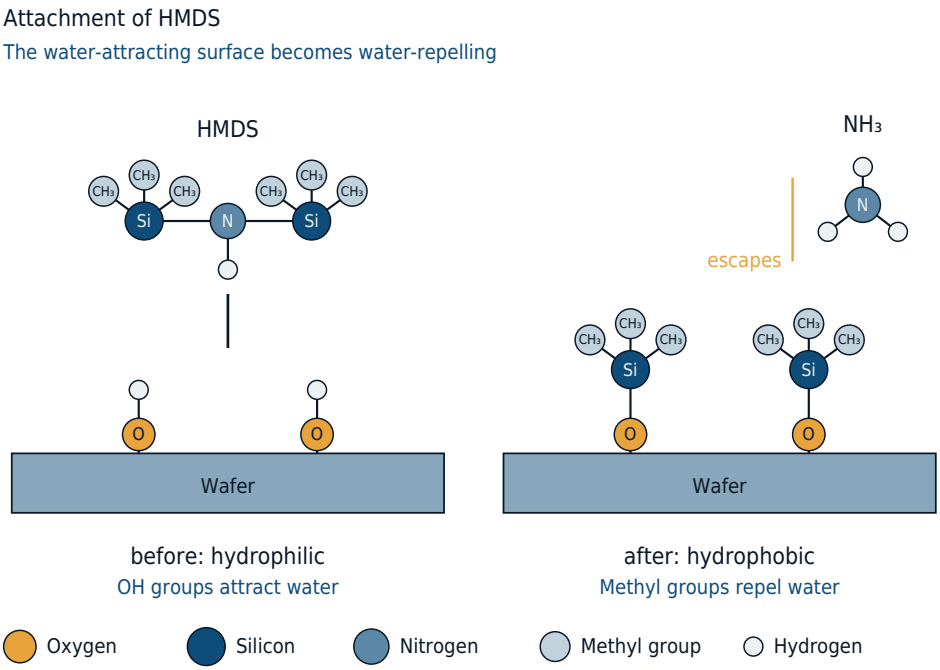
- applying an adhesion promoter and removing water from the wafer
- coating the wafers with resist
- stabilizing the resist layer
- exposure
- developing the resist
- curing the resist
- inspection

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.

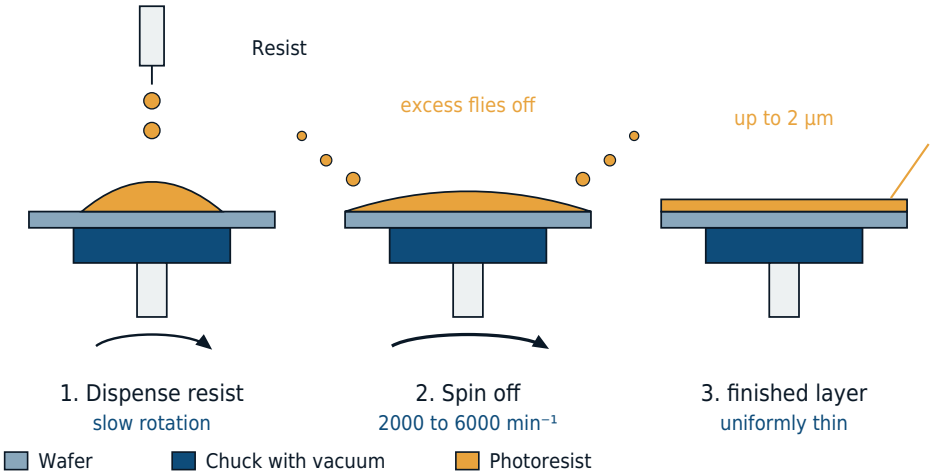


Coating

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.

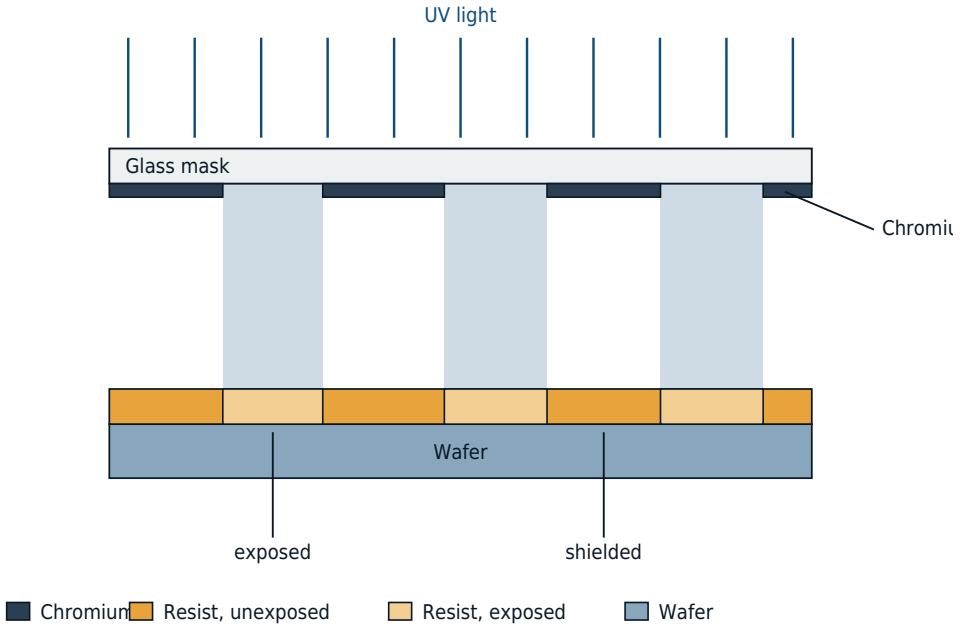
Nowadays, due to the small feature sizes, additional auxiliary layers are often used as well, which are applied to the wafer either before or after the photoresist. Among other things, these layers can serve as an anti-reflective coating or for additional planarization.

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.

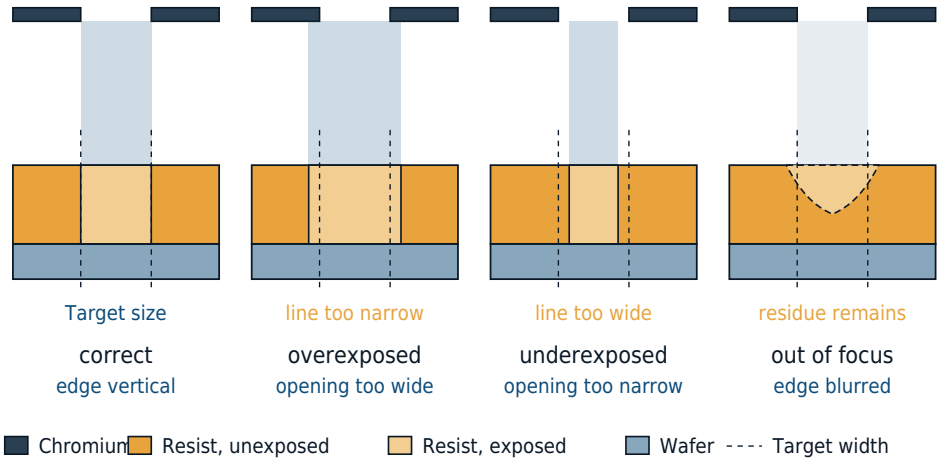
Depending on the type of resist, exposed portions become soluble or insoluble. Using a developer solution, the soluble areas are removed, leaving a patterned resist layer behind. In classic i-line positive resist, a nitrogen molecule N_2 is split off by the high-energy UV light. What remains is a keto-carbene, 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.

Exposure Profiles

Exposure time and focus determine the width of the feature



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.

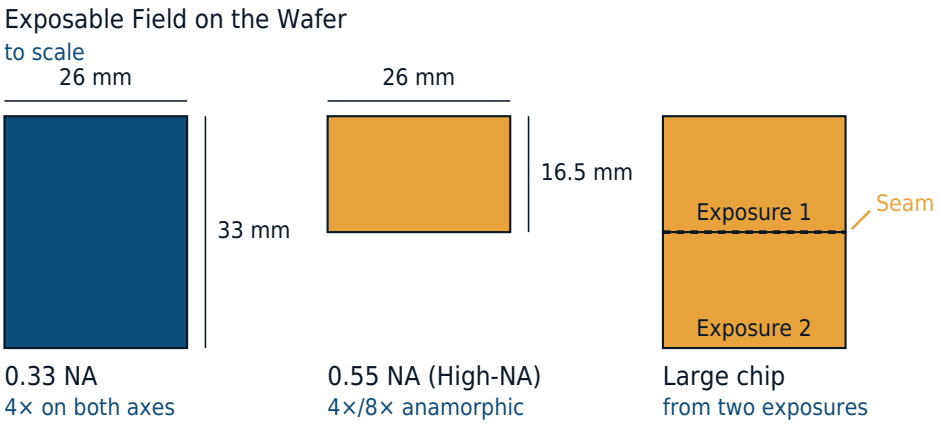
Immersion lithography

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

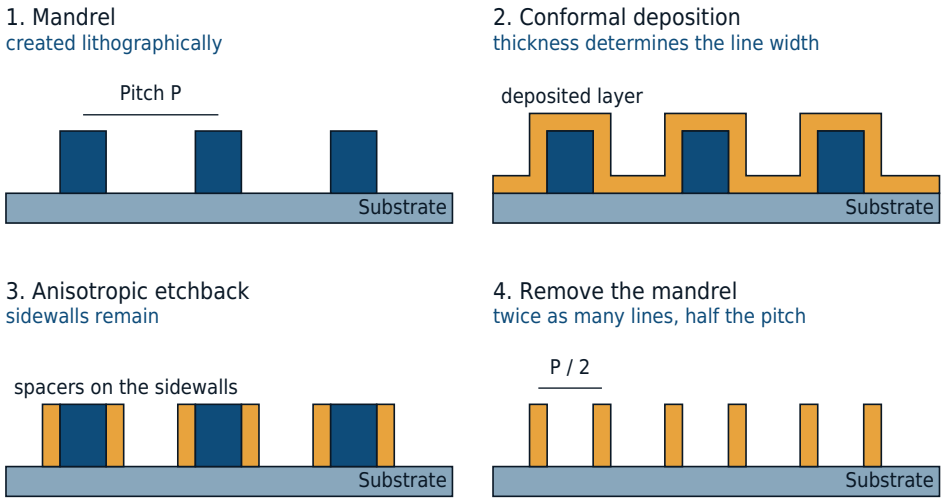
shows up as alternating too-wide and too-narrow spacings – a systematic error that makes the devices non-uniform. With three or four passes (LELELE and beyond), this problem grows, as does the number of process steps and thus the cost.

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.

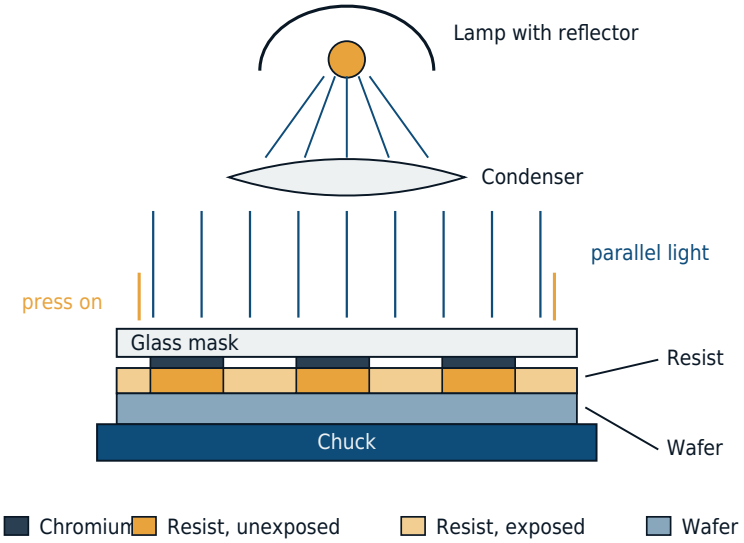
Contact exposure

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.

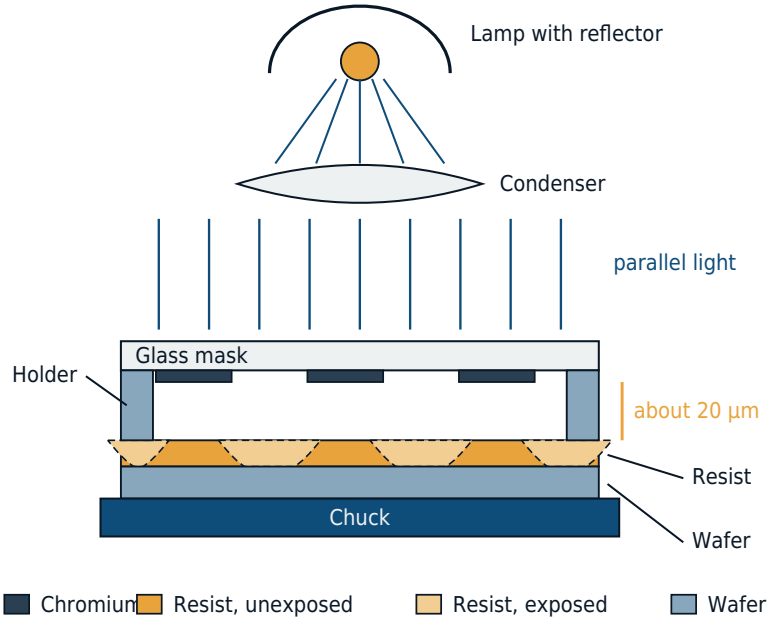
The disadvantages, however, are obvious: because the mask is pressed onto the resist, it becomes contaminated quickly, or can be scratched, as can the resist. If particles are present between the wafer and the mask, the resulting gap between mask and wafer degrades the image.

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



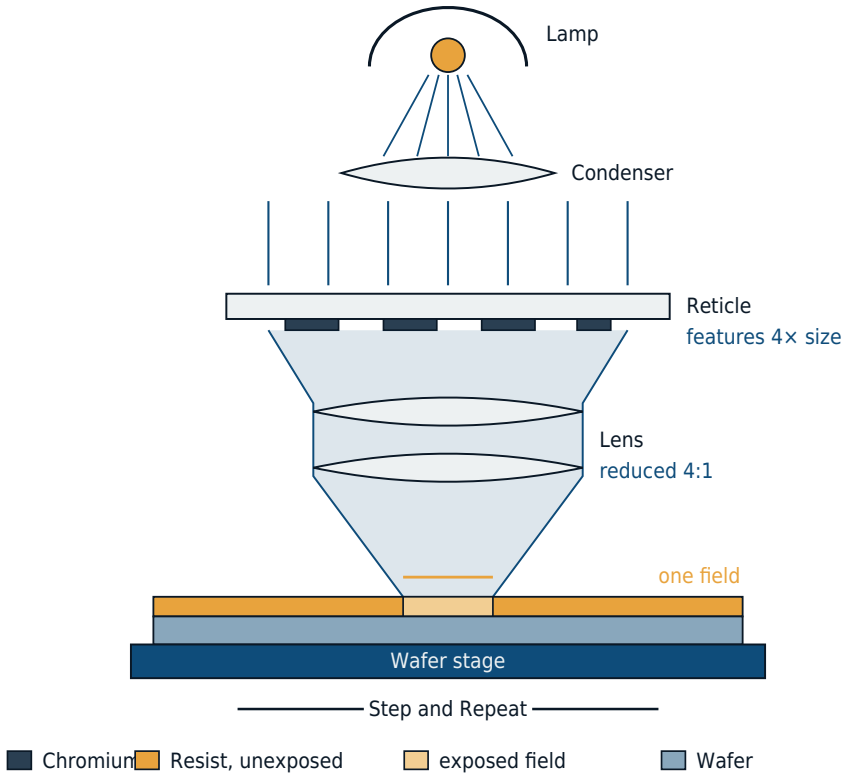
The gap protects mask and resist from scratches and from particles in between. In return, only a shadow image falls onto the wafer: the light diffracts at the chromium edges, so the features widen and their edges become blurred.

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

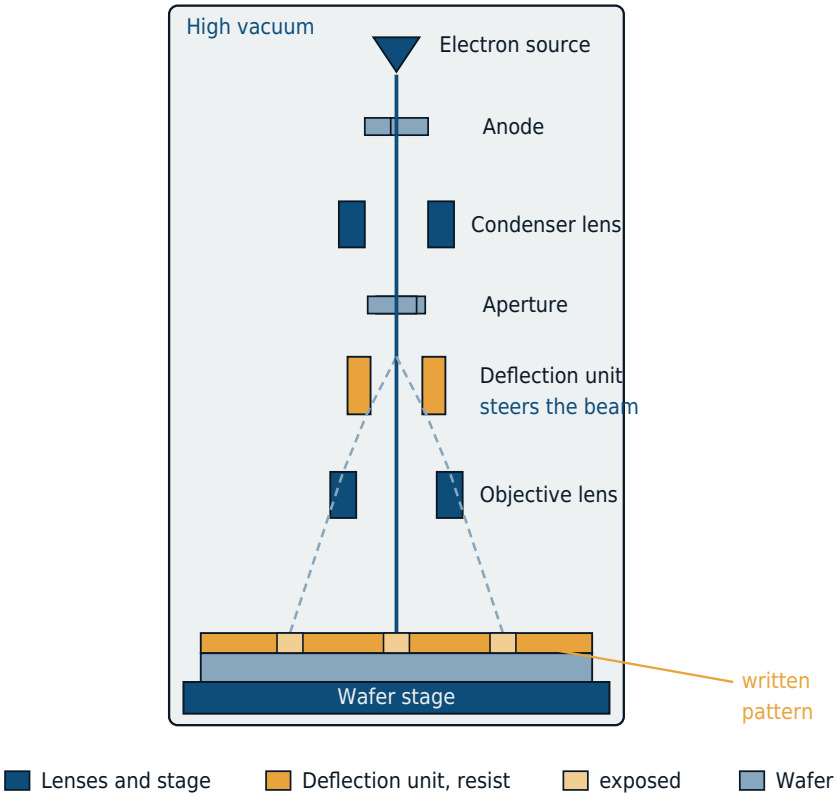
aberrations occur, as they do with lens systems, and in addition, thermally induced expansion of the wafer during processing can be compensated for. However, mirror systems produce distorted/curved images. Furthermore, due to the 1:1 imaging, the resolution is strongly limited.

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

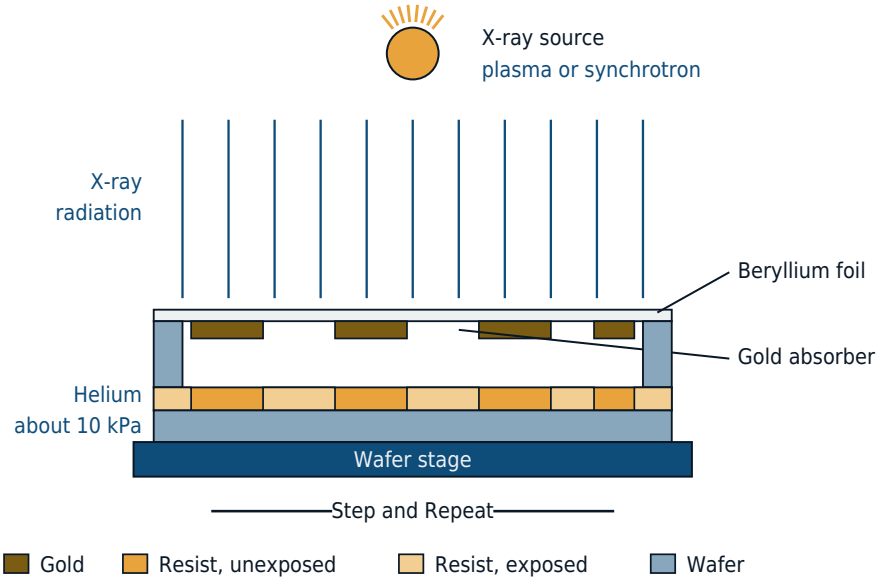
The resolution limit of x-ray lithography is about 40 nm. Imaging is carried out

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

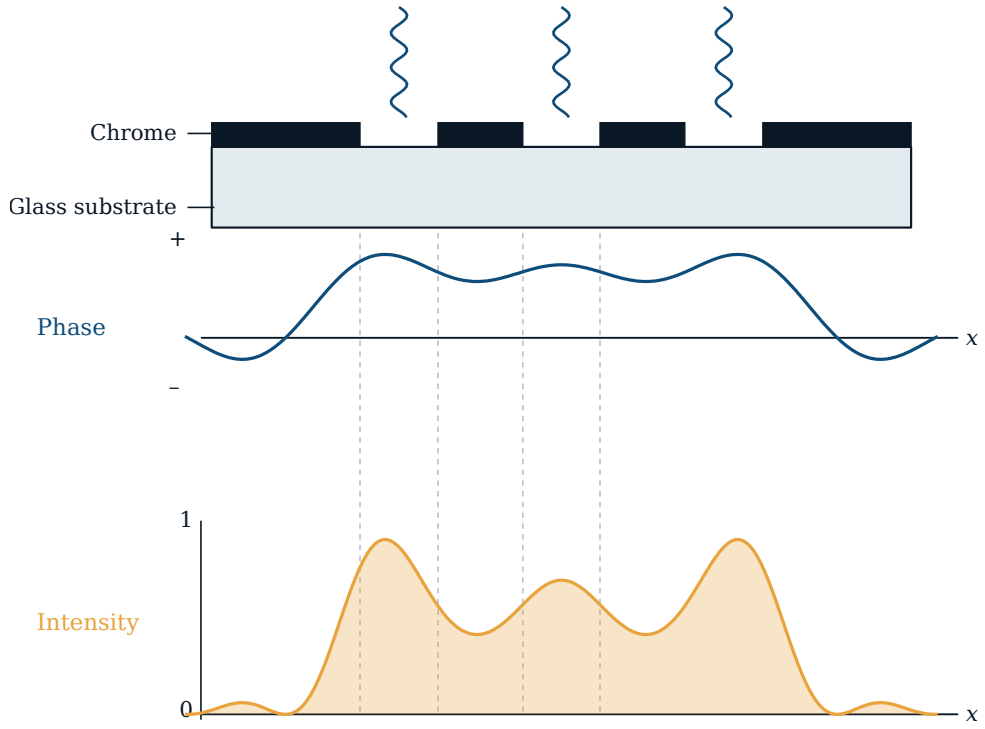
4. Resist removal

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

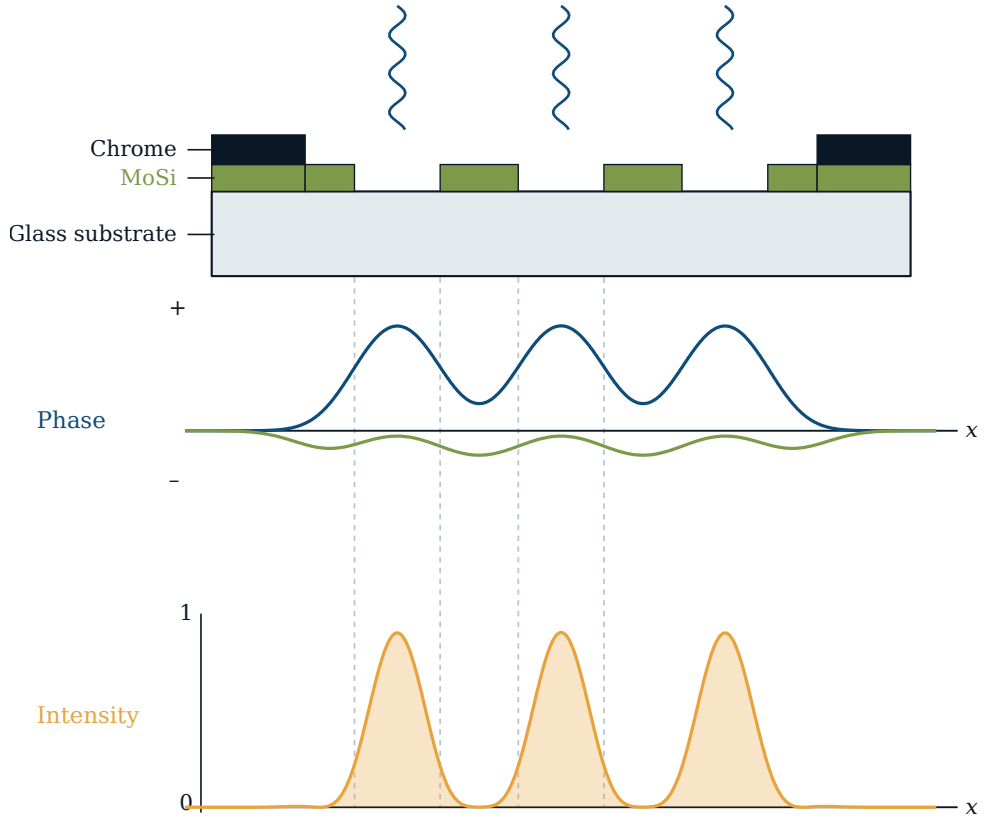


Various measures are used to reduce the intensity of the diffracted light. These are described in more detail below using the different mask types.

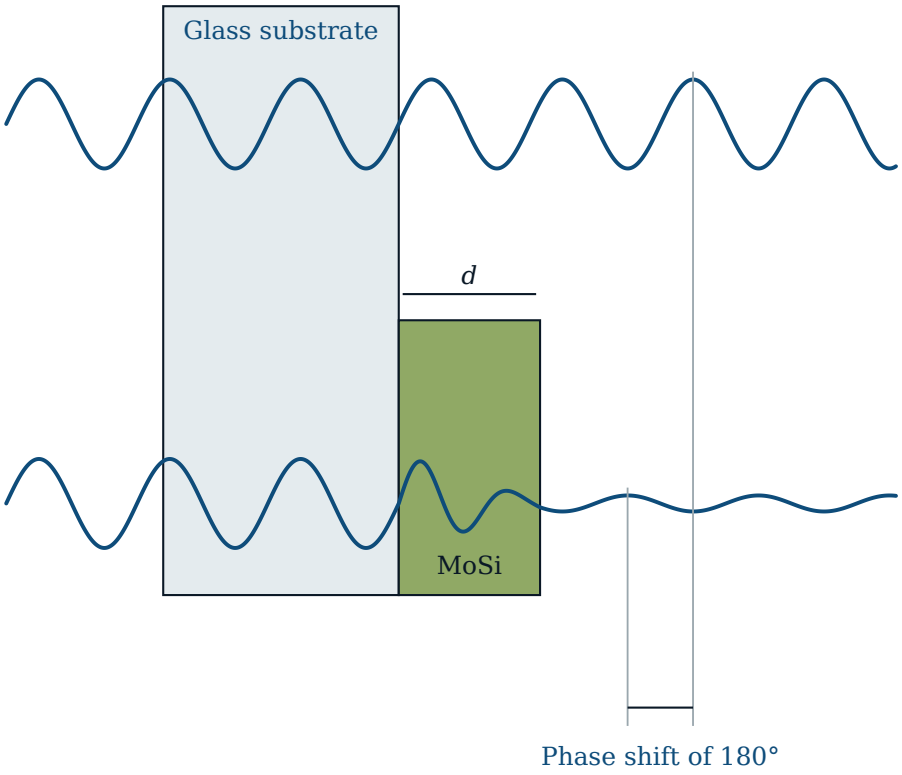
Attenuated Phase Shift Mask (AttPSM)

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.

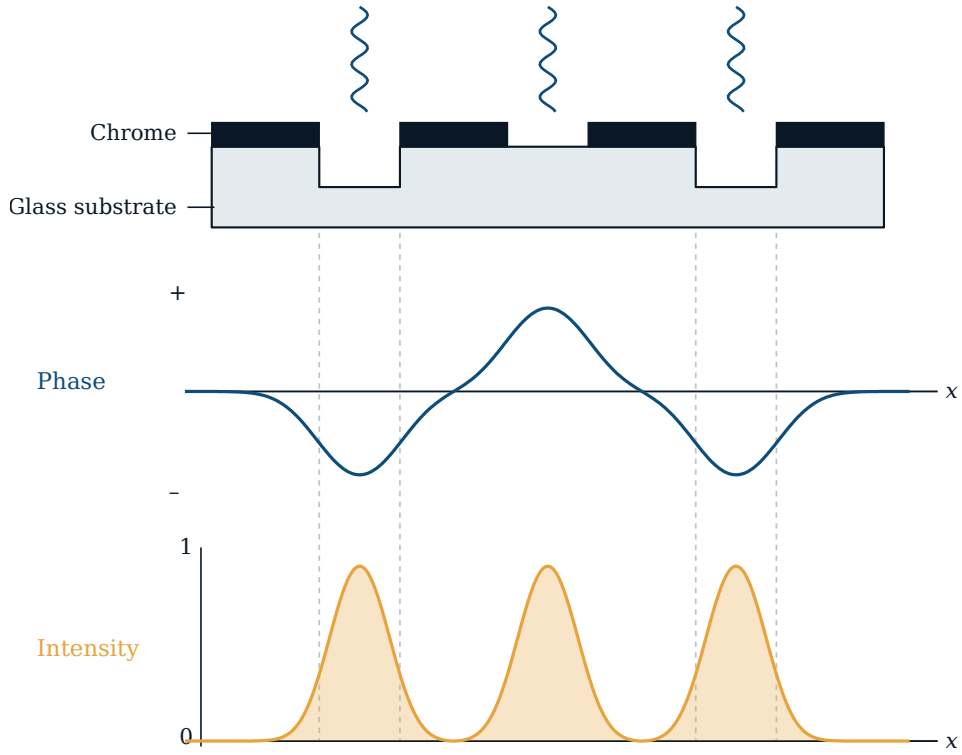
In addition, problems arise when producing large structures. Because there are no shadowing regions and the light strikes the mask everywhere with the same intensity, the desired interference – and the resulting attenuation of the light – occurs only at the edges of the structures. In the middle of larger 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

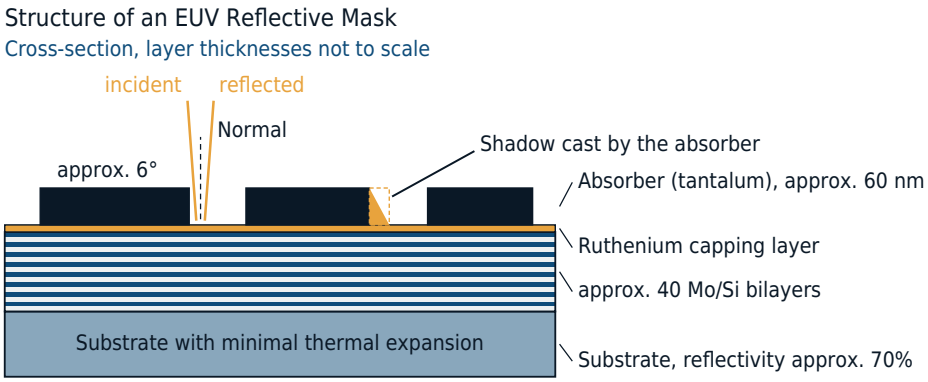
Reflective masks for EUV exposure

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.

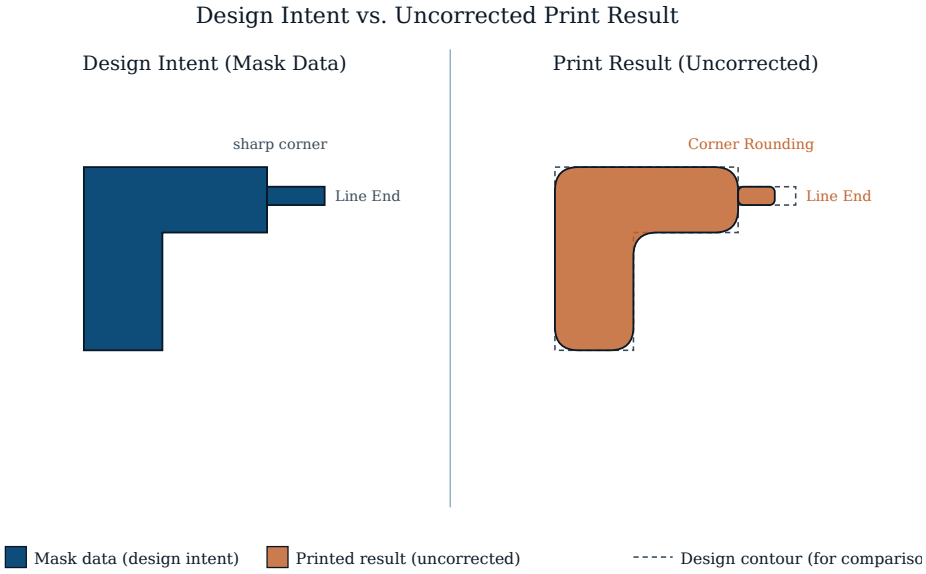
Optical Proximity Correction: Pre-Distorting the Mask

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

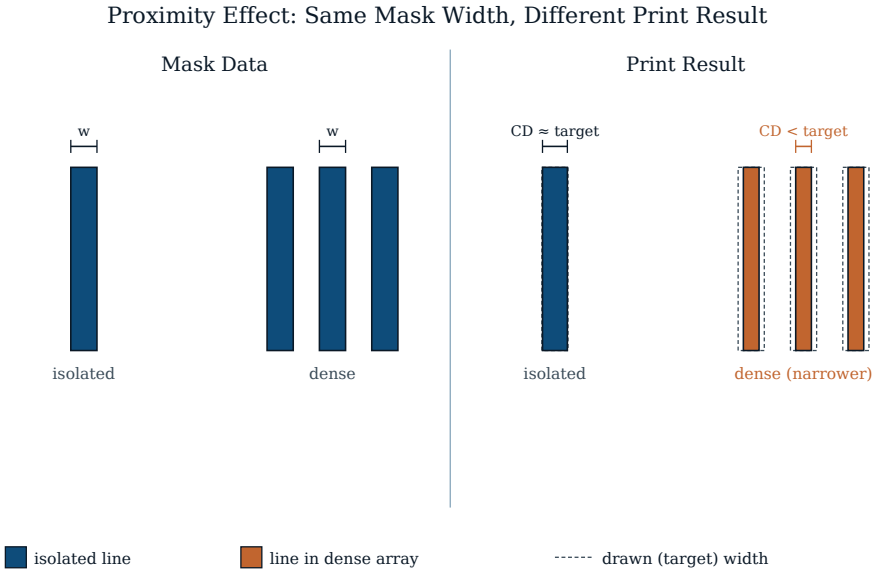
back relative to the mask specification — a line meant to reach exactly to a certain contact may, in print, end a few nanometers short of it.

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.

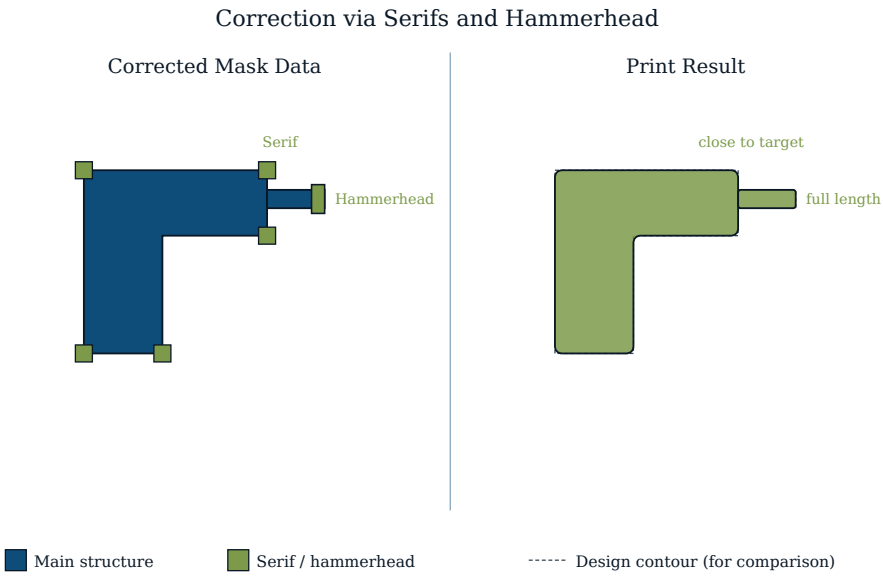
Serifs and Hammerheads: Targeted Corner Reinforcement

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.

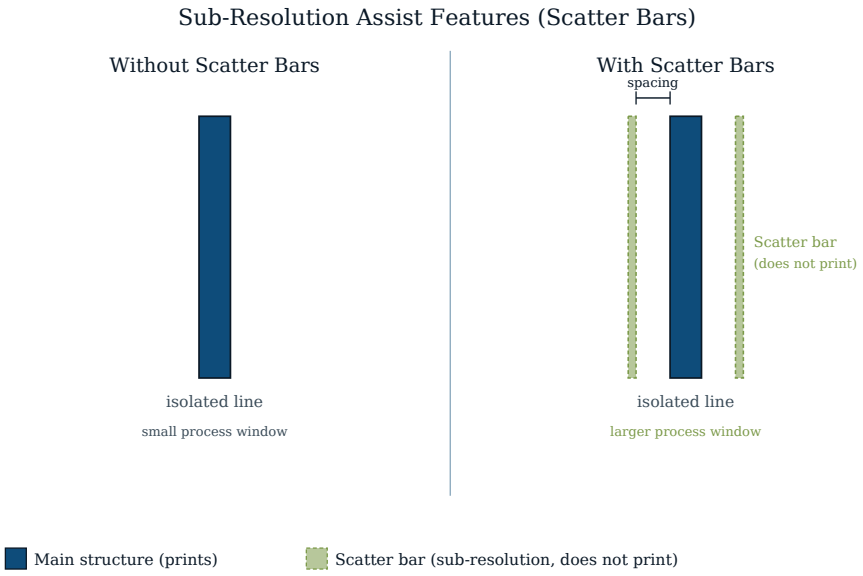
Scatter Bars: Neighbors That Never Print

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.

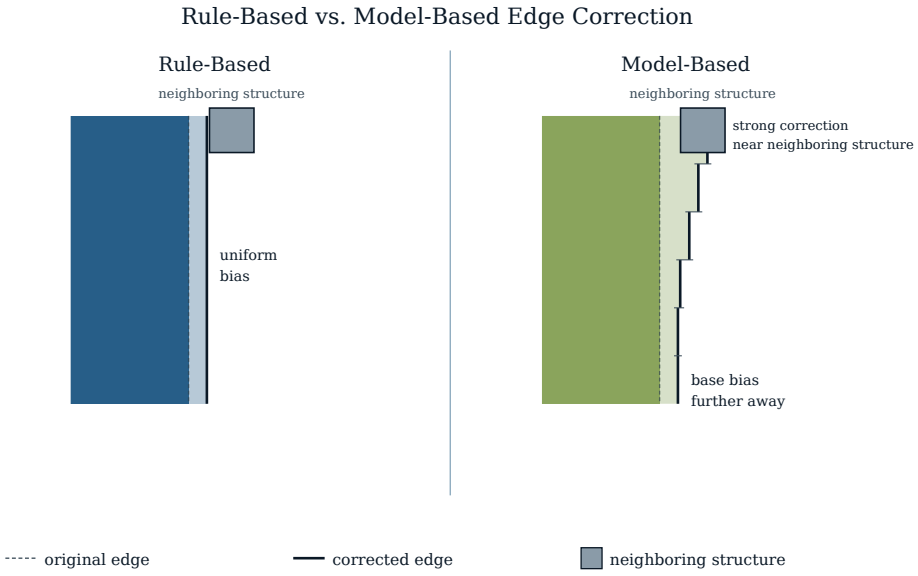
Model-Based OPC: Simulated Segment by Segment

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

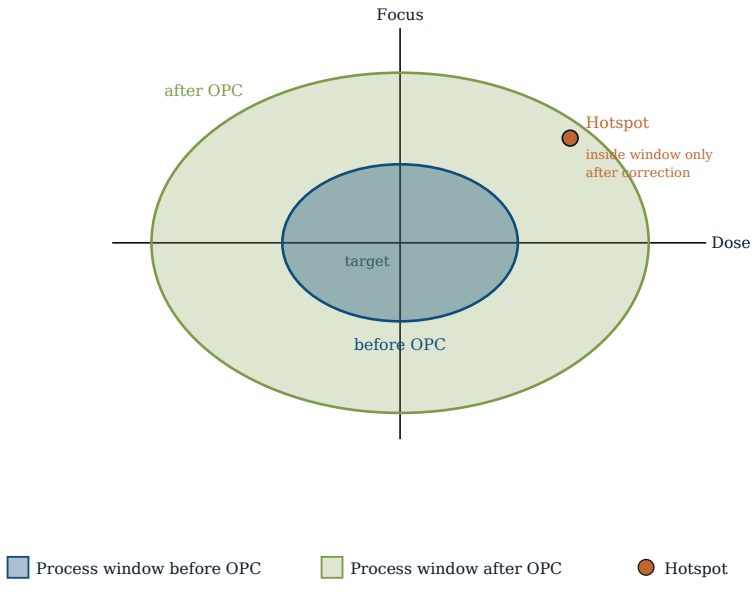
larger deviations. The wider window is thus a direct, measurable quality metric for the correction itself — in addition to plain agreement at the nominal operating point.

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.

Process Window: Before and After OPC



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.

Etch processes

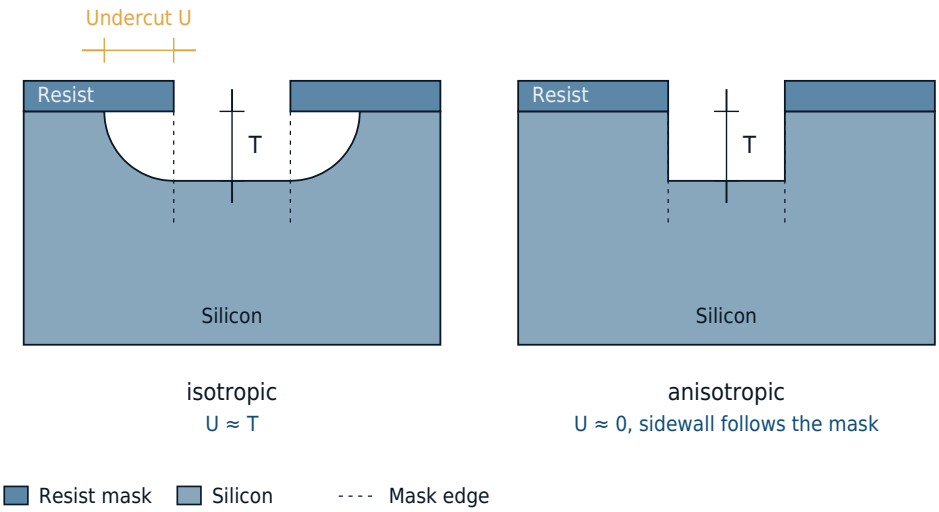
Overview

In semiconductor manufacturing, a wide variety of layers have to be etched. This may be either to remove a layer over the full wafer area, or to transfer a patterned photoresist film into an underlying layer. Etch technology can be divided into wet chemical etching and dry etching. In addition, a distinction is made between isotropic and anisotropic processes, as well as between the chemical and physical character of the etch process.

In an isotropic etch process, etching occurs in all directions. This means layers are removed not only in thickness but also in their lateral extent. In anisotropic etching, the layer is removed in only one direction. Depending on the process, either an isotropic or an anisotropic etch process may be desired.

Isotropic and Anisotropic Etching

With ideal isotropy, the undercut is about as large as the etch depth



Wet chemical etching mostly attacks without direction: the solution reaches the area under the mask just as it reaches the trench bottom – the structure becomes wider than the mask opening as a result. Anisotropic processes like RIE etch directionally instead, transferring the mask geometry almost unchanged.

An important parameter of etch processes is the selectivity. This indicates the removal ratio between two layers. If the selectivity is 2:1, one layer is removed

by the etch process twice as fast as the other.

The wet chemistry described below, however, is not used solely for etching layers, but is also employed in other processes:

- Wet etching: for the full-area removal of doped and undoped oxide layers
- Wafer cleaning
- Resist removal
- Backside processing: removal of layers that formed on the wafer backside during furnace processes
- Polymer removal: removal of byproducts that form during plasma etching and deposit on the wafers
- Edge bevel removal: targeted removal of layers at the very edge of the wafer, where they would later flake off and generate particles

Because of its generally isotropic etch profile, wet etching is hardly ever used for patterning. An exception is micromechanics. Owing to the lattice structure of single-crystal silicon, wet chemical etch solutions can produce well-defined structures, such as those with sidewall angles of 90° or 54.74°.

However, the loss of its patterning role does not mean that wet chemistry has become less important – quite the opposite. The smaller the structures have become, the more process steps have been added, and the more sensitive the device has become to contamination. A modern process flow contains more wet chemical steps than ever before; it is just that these are now predominantly cleaning, removal, and pre-treatment steps rather than patterning steps. At the same time, the equipment technology has shifted from batch processing to single-wafer processing.

Wet etching

Principle

The principle of wet etching is the conversion of the solid material of the layer into liquid compounds with the help of a chemical solution. The selectivity is very high, since the chemicals used can be precisely tailored to the layers present; for most solutions, it exceeds 100:1.

The process proceeds in three sub-steps, which must occur one after another:

1. The reactive species travel from the solution to the surface of the layer to be etched.
2. The chemical reaction takes place at the surface, forming a soluble compound.
3. The reaction product detaches from the surface and is transported away into

the solution.

Whichever of these steps is the slowest determines the behavior of the entire process. If the actual reaction is the bottleneck, the process is called reaction-limited: the etch rate then depends strongly on temperature and increases roughly exponentially with it, but the process etches uniformly across the entire wafer. If, on the other hand, the transport to and from the surface is the bottleneck, the process is diffusion-limited. The temperature dependence is then weaker, but any non-uniformity in the flow directly affects the removal rate – the edges of the wafer are etched faster than the center.

For practical purposes, this means: reaction-limited processes are controlled through very precise temperature control, while diffusion-limited processes are controlled through movement of the solution, i.e. through circulation, rotation of the wafer, or ultrasound. In narrow trenches and contact holes, every process becomes diffusion-limited, because the solution there is hardly exchanged at all. This is one of the reasons why wet etching is not suitable for fine structures.

Requirements

The following requirements have to be met by the chemical solutions:

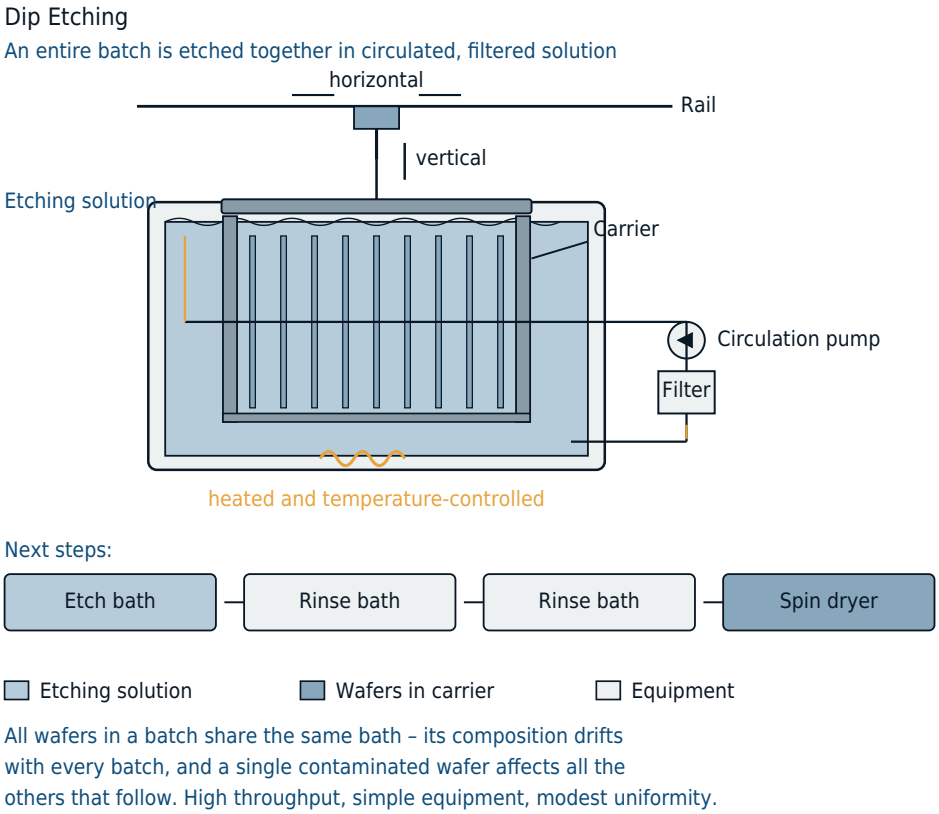
- the masking layer must not be attacked
- the selectivity must be high, both with respect to the masking and to the underlying layer
- the etch process must be able to be stopped by dilution with water
- no gaseous reaction products may form, since bubbles can locally shadow certain areas
- constant etch rate over a long period
- the reaction products must dissolve directly, so that no particles contaminate the etch solution
- the solution must fully wet the surface; otherwise water-repellent areas and narrow trenches will not be reached
- no introduction of alkali or heavy metal ions, which would enter the oxide as mobile charges and render devices unusable
- good environmental compatibility and easy disposal

The last requirement is the most difficult one in practice. The most effective etch solutions are also the most hazardous, and some of the chemicals used cannot be replaced by more benign alternatives. The effort required for exhaust air, wastewater, and occupational safety therefore accounts for a considerable share of the operating costs of a wet chemical facility.

Dip etching

In the dip etch process, an entire batch of wafers is processed at once in a bath containing the etch solution. Filters and circulation pumps keep particles away from the wafers. Since the concentration of the etch solution decreases as more wafers are processed, it often has to be renewed.

The etch rate, i.e. the material removal per unit time, has to be precisely known in order to achieve reproducible etch results. Precise temperature control of the etch solution is required, since the etch rate of most chemical solutions increases with temperature.



The lifting mechanism can transport the batch both horizontally and vertically. After the wafers have been etched, the etching process is stopped by rinsing in several dip baths; the wafers are then dried in a spin dryer.

The advantages of dip etching are the high wafer throughput and the relatively simple construction of the etch equipment. However, the uniformity of layer removal is low.

A fundamental problem of the process is the bath itself: every wafer in a batch sees the same solution, which changes over time. Its composition shifts with every batch processed, dissolved material accumulates, and a single

contaminated wafer can affect the entire batch as well as all subsequent ones. The concentration is therefore continuously monitored and replenished, and the service life of a bath is limited.

For critical steps, dip etching has therefore largely been replaced by single-wafer processing. It has, however, persisted wherever long etch times at high temperatures are required and the uniformity requirement is moderate – the classic example being the removal of silicon nitride in hot phosphoric acid, which, depending on layer thickness, takes an hour or longer and would be uneconomical to perform wafer by wafer.

Spray etching

Spray etching is comparable to spray development in photolithography. Owing to the rotation of the wafer on the chuck, combined with a continuous supply of fresh etch solution, the uniformity is very good. Bubbles cannot form here because of the fast rotation; however, the disadvantage here too is that each wafer has to be processed individually.

As an alternative to sequential processing, spray etching of several batches at once is also possible in a drum, in which the wafers rapidly rotate around a central spray fixture. Immediately afterward, the wafers are dried under rotation in a hot nitrogen atmosphere.

What started out as a disadvantage is today the norm: single-wafer processing has established itself for nearly all critical wet chemical steps. The reason lies in control. Every wafer sees only fresh chemistry, the process time can be set to the second, and any error always affects only a single wafer rather than an entire batch.

A modern single-wafer tool has several pivoting nozzles that apply different media to the rotating wafer one after another – etch solution, rinse water, drying agent – without the wafer ever leaving the process chamber. Consumption per wafer is orders of magnitude lower than that of a bath, which considerably reduces the costs for chemicals and disposal.

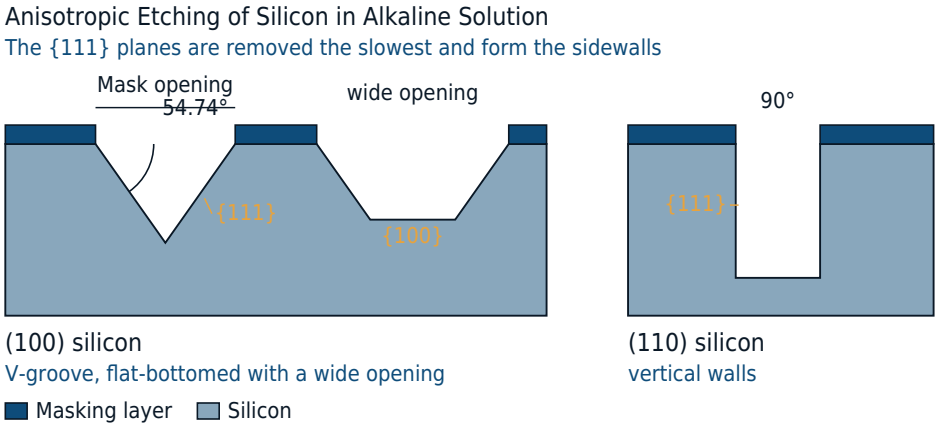
This also makes possible something that is fundamentally impossible in a bath: spatially confined treatment of a wafer. A nozzle at the edge can be used to strip only the outermost bevel, while a supply from below can treat only the backside, with the front side protected by a stream of nitrogen.

Anisotropic etching of silicon

Although the particles in an etch solution can etch the layer on the wafer in every direction, there are also processes that allow an almost anisotropic etch profile to be obtained during wet etching. These make use of the different etch rates on the various crystal planes.

The cause lies in the number of bonds with which a silicon atom is held in the respective plane of the crystal. In the {111} plane, the density of bonds is highest, and it is correspondingly removed most slowly. The {100} and {110} planes are etched considerably faster – the ratio in potassium hydroxide reaches more than 100:1, depending on concentration and temperature. As a result, the etch proceeds in depth until it reaches {111} planes; these remain standing and form the sidewalls of the structure. The shape of the result is therefore not determined by the etch time, but by the crystal orientation of the wafer.

Etch profiles depending on crystal orientation



On (100)-oriented wafers, the {111} planes intersect the surface at 54.74°, producing V-shaped trenches and pits with a pyramidal cross-section. If the mask opening is wide or the etch time short, a flat bottom remains and the structure is trapezoidal. On (110)-oriented wafers, the {111} planes stand perpendicular to the surface, so that trenches with vertical walls and a very high depth-to-width ratio can be etched.

Etch solutions

Etching is carried out with potassium, sodium, or lithium hydroxide (KOH, NaOH, LiOH), with tetramethylammonium hydroxide (TMAH), or with an EDP solution (a mixture of water, pyrazine, catechol, and ethylenediamine). In every case, the reaction is driven by the OH^- group (hydroxide group) of these

substances:

$$Si + 2OH^- + 2H_2O \rightarrow SiO_2(OH)_2^{2-} + 2H_2$$

The choice of solution is less a question of etch rate than one of compatibility with the rest of the manufacturing process. Potassium hydroxide etches with the strongest anisotropy but introduces potassium ions: alkali ions are mobile in the gate oxide, shift the threshold voltage of transistors, and render a device unusable. In a production line where integrated circuits are also fabricated, potassium hydroxide is therefore ruled out. TMAH contains no metal ions and is compatible with CMOS manufacturing; its anisotropy is lower, but it barely attacks silicon dioxide. EDP is largely avoided today because of its toxicity. TMAH, as a processing solution, is by no means harmless either, being highly toxic through skin contact – in a much more dilute form, namely as a 2.38 % solution, the same substance is the standard developer used in photolithography.

Limiting the etch depth

Since the etch stops on its own only at {111} planes, the depth has to be defined in some other way if a membrane of a defined thickness is to be produced. Two approaches are common: a layer very heavily doped with boron, which is barely attacked by the lye any longer, or electrochemical control, in which a p-n junction is placed under voltage and the etch comes to a halt at it.

Applications

Anisotropic wet etching plays no role in patterning integrated circuits. Outside of that field, however, it is a standard process:

- Micromechanics: membranes for pressure sensors, proof masses for accelerometers, nozzles for inkjet print heads
- Solar cells: a brief treatment in dilute lye produces a dense array of small pyramids on (100) silicon. This texture causes incident light to strike the surface multiple times and considerably lowers the reflectance.
- Optics and packaging: V-grooves for holding and aligning optical fibers, alignment structures for assembly

Etch solutions for isotropic etching

Different etch solutions are available for the respective materials. In the semiconductor industry, oxide layers are etched with hydrofluoric acid HF:

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

The solution is buffered with NH_4F to keep the HF concentration constant. In a mixture of a 40 % NH_4F solution and 49 % hydrofluoric acid (ratio 10:1), the

etch rate on thermal oxide is 50 nm/min. TEOS oxides are etched considerably faster, at about 150 nm/min, and PECVD oxides at about 350 nm/min. The selectivity relative to crystalline silicon, silicon nitride, and polysilicon is much greater than 100:1.

The fact that the etch rate depends so strongly on the origin of the oxide is not a side effect but a useful tool: conversely, the etch rate in buffered hydrofluoric acid can be used to assess the density – and thus the quality – of a deposited oxide layer.

A brief treatment in highly diluted hydrofluoric acid, the HF dip, removes the native oxide that forms on any silicon surface in air within a short time. It therefore precedes many processes that require a clean contact to the silicon, such as epitaxy, gate oxidation, or metal deposition. Afterward, the surface is saturated with hydrogen and thus water-repellent – water visibly beads off it. Since new oxide forms again immediately, the subsequent step must follow promptly.

Silicon nitride is etched with hot phosphoric acid H_3PO_4 at about 160 °C. Here, however, the selectivity to SiO_2 is very low, at 10:1. For polysilicon, the selectivity to nitride is essentially determined by the concentration of the phosphoric acid. To increase the selectivity relative to oxide, dissolved silicon is added to the solution so that it becomes saturated with respect to silicon dioxide and thus barely attacks it any longer.

Crystalline and polycrystalline silicon are first oxidized with nitric acid (HNO_3), after which the silicon dioxide is etched with hydrofluoric acid:

1. $3Si + 4HNO_3 \rightarrow 3SiO_2 + 4NO + 2H_2O$
2. $SiO_2 + 6HF \rightarrow H_2SiF_6 + 2H_2O$

Aluminum is etched at about 50 °C with a mixture of phosphoric, nitric, and acetic acid. The nitric acid oxidizes the aluminum, the phosphoric acid dissolves the resulting oxide, and the acetic acid lowers the surface tension so that the solution fully wets the structure and the hydrogen generated does not remain adhered as bubbles. Titanium is etched with a solution of ammonia water NH_4OH, hydrogen peroxide H_2O_2, and water (ratio 1:3:5). Since this solution also attacks silicon once the peroxide is consumed, its bath lifetime is short. The same mixture, in a more dilute form, serves as a cleaning solution for wafer cleaning.

Handling hydrofluoric acid

Among process chemicals, hydrofluoric acid occupies a special position and deserves an explicit warning. It is only a weak acid and initially causes little pain on the skin, but penetrates deep into the tissue. There, the fluoride binds the

body's calcium, which can lead to chemical burns reaching down to the bone and to disturbances of the heart rhythm. Symptoms often appear only hours after contact, by which time the damage has already occurred. Handling it therefore requires appropriate protective equipment, a readily available calcium gluconate gel, and proper training – in manufacturing, hydrofluoric acid is handled exclusively in closed systems.

Assessment

In general, wet chemical etching is well suited for removing entire layers on a wafer. The selectivity between the layer to be etched and the one beneath it is usually very good, so there is little risk of removing the wrong layer. In addition, the removal rate per unit time is very high, and in dip etching many wafers can be etched simultaneously. For small structures, however, wet etching cannot be used, since its isotropic etch profile is unsuitable for this purpose. In this case, layers are removed anisotropically using dry etch processes.

For very fine structures, a second reason comes into play that has nothing to do with the etch profile anymore: at the end, the liquid has to be removed from the wafer again, and it is precisely during this step that the exposed structures can be destroyed.

Drying the wafers

At the end of every wet chemical step comes rinsing with ultrapure water, followed by drying. What sounds like a minor step is, for fine structures, the most critical part of the entire process.

Why drying is a problem

As long as liquid remains between two neighboring fins or lines, its surface tension pulls them toward each other. The force grows the more closely spaced and the taller the structures are. Once two lines touch, they adhere to each other through bonds between their surfaces and do not spring back apart. This structural collapse – known as *stiction*, from *static friction* – is the most common cause of failure for slender lines at the end of a wet process. Affected above all are free-standing elements in micromechanics and tall, narrow resist lines in photolithography.

Drying by simple spin-off or evaporation makes the problem worse, because the receding liquid front then runs directly through the structure. All of the following methods therefore aim either to defuse the transition from liquid to dry, or to avoid it altogether.

Methods

Spin drying
: The wafer rotates rapidly, the liquid is flung outward, and a stream of nitrogen finishes the drying. Simple and fast, but unsuitable for fine structures and prone to leaving behind dry spots.

Isopropanol drying
: The rinse water is displaced by isopropanol, whose surface tension is only about one-third that of water. The capillary forces are correspondingly lower.

Marangoni drying
: The wafer is slowly withdrawn from the water while isopropanol vapor is deliberately supplied at the water surface. The resulting difference in surface tension along the interface pulls the water film away from the wafer, so that it emerges dry from the bath and no liquid has to evaporate on it.

Supercritical carbon dioxide drying
: The liquid is replaced by carbon dioxide, which is then brought into the supercritical state through pressure and temperature. In this state, there is no longer an interface between liquid and gas, and thus no surface tension either; the carbon dioxide can be released without leaving any residue. This is the most elaborate but also the gentlest method – it is used wherever all the others reach their limit.

The same reasoning underlies the development of dry alternatives to individual wet steps: wherever a layer can be removed with a gas instead of a liquid, the problem is avoided from the outset.

Wafer cleaning

Cleanroom

Semiconductor manufacturing takes place in cleanrooms in order to protect the highly complex circuits of the semiconductor devices from contamination that could affect their functionality. Cleanrooms are classified according to the size of the particles and their number per cubic foot (= cbf; 1 foot = 30.48 cm) or, according to DIN/ISO, per 1 m^3:

Permitted number of particles per 1 m^3

Class ↓	≥ 0.1 µm	≥ 0.2 µm	≥ 0.3 µm	≥ 0.5 µm	≥ 1.0 µm	≥ 5.0 µm
ISO 1	10	2				
ISO 2	100	24	10	4		
ISO 3	1,000	237	102	35	8	

ISO 4	10,000	2,370	1,020	352	83	
ISO 5	100,000	23,700	10,200	3,520	832	29
ISO 6	1,000,000	237,000	102,000	35,200	8,320	293
ISO 7	10,000,000	2,370,000	1,020,000	352,000	83,200	2,930

For example, in a class 3 cleanroom there may be a maximum of 1,000 particles with a diameter ≥0.1 µm, 237 particles ≥0.2 µm, 102 particles ≥0.3 µm, 35 particles ≥0.5 µm, and 8 particles ≥1 µm. In an operating room, the cleanroom class is 2 or 3; in a volume of 1 m^3 of city air, there are 400 million particles of size 5 µm.

The air in cleanrooms for microelectronics is cleaned through ultra-fine filters and then introduced through the ceiling. Air is extracted through holes in the floor, so that this laminar flow carries particles downward and away through the floor. To prevent contamination from entering the cleanroom from outside, a slight overpressure is generally maintained there. To protect the wafers as effectively as possible against particles, they are transported from one production tool to the next in transport boxes. In modern cleanrooms, these are so-called FOUPs (Front Opening Unified Pods), which dock directly onto the loading stations of the tools and are opened in such a way that the wafers pass from the FOUP into the tool without being exposed to the cleanroom air.

Since the wafers inside the FOUPs are completely shielded from the ambient cleanroom air, a different cleanroom class can prevail inside the transport boxes than in the cleanroom itself. This makes it possible to operate the large production areas at a relatively high cleanroom class of ISO 5, while a cleanroom class of ISO 1 or ISO 2 prevails inside the FOUPs. This allows enormous cost savings, since only a small volume needs to meet the highest cleanliness class.

Illustration of a class 1 cleanroom

(Click to enlarge, 700 KB · source: unknown)

In the illustration, the production area is only the region between the pink-colored ventilation system beneath the roof and the yellow-colored floor area. Below that is the basement housing the supply equipment (pumps, chemical tanks, etc.). Some cleanrooms are surrounded by a so-called grey room in which the tools themselves are located, so that, separated by a wall, only the control panels and the load ports for bringing wafers into the tools are accessible from within the cleanroom.

Personnel wear special cleanroom suits that do not shed particles. Depending on requirements, the suit covers the entire body. The head is covered either with a

simple hairnet or a hood and a face mask, or with a full mask. In addition, there are special low-particle shoes, undergarments, and gloves. At the entrance to the cleanroom there are often air showers that blow particles off the suit once more before entry.

Types of contamination

Despite the cleanroom, there are various types of contamination, which are mainly caused by personnel in the production line, the ambient air, chemicals (gases, solutions), and the equipment itself:

- Microscopic contamination: e.g. particles from the ambient air or from gases
- Molecular contamination: e.g. hydrocarbons from oil in the pump systems
- Ionic contamination: e.g. hand sweat
- Atomic contamination: e.g. heavy metals from solutions, abrasion from solid parts

Microscopic contamination

Microscopic contamination consists of particles that attach to the wafer surface. Sources of this contamination include the ambient air, personnel clothing, abrasion from moving parts in process equipment, inadequately filtered liquids such as etch and cleaning solutions or developer, or etch residues remaining after dry etch processes.

Microscopic contamination causes shadowing effects when particles on the wafer surface prevent exposure of the photoresist; in contact exposure, poor resolution results when relatively large particles are located between the mask and the wafer. In addition, they can shield accelerated ions from the wafer surface during implantation processes or dry etching.

Particles can also become enclosed by layers, creating unevenness. Subsequent layers can crack open at these locations, or resist can accumulate there, resulting in areas that are not fully exposed due to the excessive resist thickness.

Molecular contamination

Molecular contamination results from resist and solvent residues on the wafers or from oil mist deposits originating from vacuum pumps. This contamination

can be present on the wafer surface as well as diffuse into the layers. Contamination adhering to the surface partly hinders the adhesion of subsequently deposited layers, considerably so in the case of metallizations (thin interconnects). Diffusion into oxides degrades the electrical robustness of these layers.

Alkaline and metallic contamination

The main source of this contamination is humans, who constantly release salts through the skin as well as through breath. In addition, insufficiently deionized water introduces (alkali) ions such as sodium or potassium onto the wafers. Heavy metals present in etch solutions can also lead to contamination. Radiation-based processes in equipment (implanters, dry etching) can sputter material off the chamber walls, which then deposits onto the wafers.

Ionic contamination affects, for example, the electrical properties of MOS transistors, since their charge alters the threshold voltage – that is, the voltage above which the transistor becomes conductive. Heavy metals such as iron or copper provide free electrons, causing the power consumption of diodes to increase. Metals can also form recombination centers for free charge carriers, so that insufficient free charge carriers remain available in the circuit for proper operation.

Cleaning techniques

The wafers are rinsed with ultrapure water after every wet chemical process step, but wafers are also freed from contamination after other processes. There are various cleaning techniques for removing the different types of contamination. The consumption of ultrapure water in a semiconductor fab is extremely high, amounting to several million liters per year. In ultrapure water, almost no contamination remains; 1–2 ppm (parts per million = number of contaminant particles per million water molecules) is permitted (example, as of 2002). The water is usually purified directly on site in treatment plants.

One method of wafer cleaning is the ultrasonic bath, in which the wafers are placed into a solution of water together with ultrasonic cleaning agents and wetting agents. Ultrasonic excitation loosens particles from the surface, while metals and molecular contaminants are partially bound by the cleaning agent. Particles that do not adhere strongly can also be blown off with nitrogen.

Solvents such as acetone or ethanol are suitable for removing organic contamination such as grease, oil, or skin flakes. However, these can leave carbon residues behind.

Ionic contamination (ions of potassium, sodium, etc.) is removed by rinsing with

deionized water. Cleaning with rotating brushes together with a cleaning liquid is also possible. However, on patterned wafers, particles accumulate at edges, and the brushes can damage the wafer surface. Contact holes or other recesses can be rinsed out using high-pressure cleaning at about 50 bar. This method, however, does not remove ionic or metallic contamination.

Cleaning with water and various cleaning agents is often not sufficient on its own. Contaminants are frequently removed using aggressive etch solutions, or the surface is deliberately removed to a minimal extent. Mixtures of hydrogen peroxide with sulfuric acid or ammonia, or Caro's acid/piranha solution (sulfuric acid with hydrogen peroxide), dissolve organic contamination through oxidation at about 90 °C.

A mixture of hydrochloric acid and hydrogen peroxide converts alkali metals into readily soluble chlorides (salts), while heavy metals form complexes and go into solution. Native oxide can be removed with hydrofluoric acid, whereas a deliberately deposited oxide layer using hydrogen peroxide can instead be used to protect the surface.

In addition to rinsing the wafers after every wet chemical process step, wafers undergoing a full cleaning pass through a series of cleaning steps performed one after another. The order of the cleaning processes is important here, since some can partially interfere with each other's cleaning effect. A cleaning sequence might, for example, proceed as follows:

- blowing off particles with nitrogen
- cleaning in an ultrasonic bath
- removing organic contamination with Caro's acid (H_2SO_4–H_2O_2)
- removing finer organic contamination with an NH_4–H_2O_2 solution
- removing metallic contamination with hydrochloric acid and hydrogen peroxide
- drying the wafers in a spin dryer under a hot nitrogen atmosphere

After every cleaning step, the wafers are rinsed with ultrapure water, and, where necessary, native oxide is removed with hydrofluoric acid. Depending on the current wafer surface, the cleaning sequences differ, since the solutions can attack certain layers. As structures continue to shrink, cleaning becomes increasingly difficult as well. Not only are small openings harder to reach, but surface tension and capillary effects can also cause structures to topple over, destroying the circuit.

Dry Etching

Overview

General

In wet chemical etching, layers are generally removed isotropically, yet an anisotropic etch profile is particularly important for small structures. Dry etch processes are well suited for this, offering sufficient selectivity. These processes enable reproducible, uniform etching of most layers used in semiconductor manufacturing. Besides anisotropic etch profiles, isotropic ones can also be achieved. Despite the high equipment costs and sequential single-wafer processing, dry etching has established itself over wet chemical etching.

Key parameters in dry etching

Dry etch processes are described by a small number of key parameters, which are also the target parameters of every process development effort.

Etch rate r
: The etch rate describes the material removed per unit time and is expressed, for example, in nanometers per minute or Ångström per second.

$$r=\frac{Etch\ removal\ \Delta z}{Etch\ time\ \Delta t}$$

Anisotropy factor f
: The anisotropy factor describes the ratio of the horizontal etch rate r_h to the vertical etch rate r_v.

$$f=1-\frac{r_h}{r_v}$$

For pattern transfer, strongly anisotropic processes are desired, i.e. etching only in the vertical direction, so that the resist mask is not undercut. For anisotropic etch processes, $f \to 1$, and correspondingly, for isotropic processes, $f \to 0$.

Selectivity S_{jk}
: The selectivity S_{jk} between material j and material k describes the ratio of the etch rates of two materials, e.g. of the layer to be patterned (j) and the resist mask (k).

$$S_{jk}=\frac{r_j}{r_k}$$

Whether a high or a low selectivity is desired depends on the particular process. When patterning layers, the value should be as high as possible, i.e. the layer to be patterned is removed faster than the resist mask. In reflow etchback, the selectivity has to be 1 in order to planarize topographies uniformly.

Uniformity U
: Uniformity describes how evenly the material removal occurs – across a single wafer, from wafer to wafer, and from batch to batch. It is expressed as the spread of the etch rate relative to its mean value.

$$U=\frac{r_{max}-r_{min}}{r_{max}+r_{min}}\cdot 100\,\%$$

Aspect ratio A
: The aspect ratio is the ratio of etch depth z to feature width b. It describes how slender a trench or hole is.

$$A=\frac{z}{b}$$

As the aspect ratio increases, the etch rate decreases, because the etch species have more difficulty reaching the bottom of the structure and the reaction products are less readily transported away. This effect is known as aspect ratio dependent etching (ARDE), or RIE lag. Today it is the decisive limitation when etching deep structures, such as the memory holes in 3D NAND flash, which have aspect ratios exceeding 50:1. Typical profile defects in such etches include bowing (bulging widening in the upper region), twisting (deep holes tilting relative to one another), and microtrenching (excessive removal at the bottom corners).

Loading effect
: The etch rate depends on the total exposed area: a large area to be etched consumes more reactive species, so the etch rate decreases. Macro-loading affects the entire wafer or batch, while micro-loading affects the immediate vicinity of a single structure.

Since the etch rate depends on the state of the equipment, the gas flow, and the loading, an etch process in manufacturing is not controlled by a fixed time, but by endpoint detection. For this purpose, a spectrometer monitors the optical emission of the plasma (optical emission spectroscopy, OES) – when a layer is etched through, the intensity of characteristic lines changes because different reaction products are formed – or an interferometer measures the remaining layer thickness. Once the endpoint has been detected, a short, defined overetch time follows to reliably remove residues across the entire wafer.

Dry etch processes

In dry etch processes, gases are excited by high-frequency alternating fields, typically at 13.56 MHz and 2.45 GHz. At a pressure between 0.1 Pa and 100 Pa, the mean free path of the particles is a few millimeters to centimeters.

Modern tools separate the generation of the plasma from the acceleration of the ions: a high-frequency generator (e.g. 13.56 MHz or 60 MHz, coupled in inductively or via microwaves) sets the charge carrier density and thus the etch rate, while a second, low-frequency generator (e.g. 400 kHz to 2 MHz) at the wafer electrode sets the ion energy and thus the physical component of the etch. This allows both quantities to be adjusted largely independently of one another. In addition, the plasma is frequently pulsed to reduce charge buildup and profile defects in deep structures.

There are generally three different types of dry etching, which are described in more detail on the following page:

- Physical dry etching: physical removal of the wafer surface by accelerated particles
- Chemical dry etching: reaction between gas and the wafer surface
- Chemical-physical dry etching: physical etching with a chemical component

A more recent development is atomic layer etching (ALE). Here, the chemical activation of the surface and the actual removal step are separated in time and repeated cyclically, so that only one or a few atomic layers are removed per cycle. The amount removed is thus controlled not by the etch time but by the number of cycles – a prerequisite for devices whose layer thicknesses are only a few nanometers, such as gate-all-around transistors.

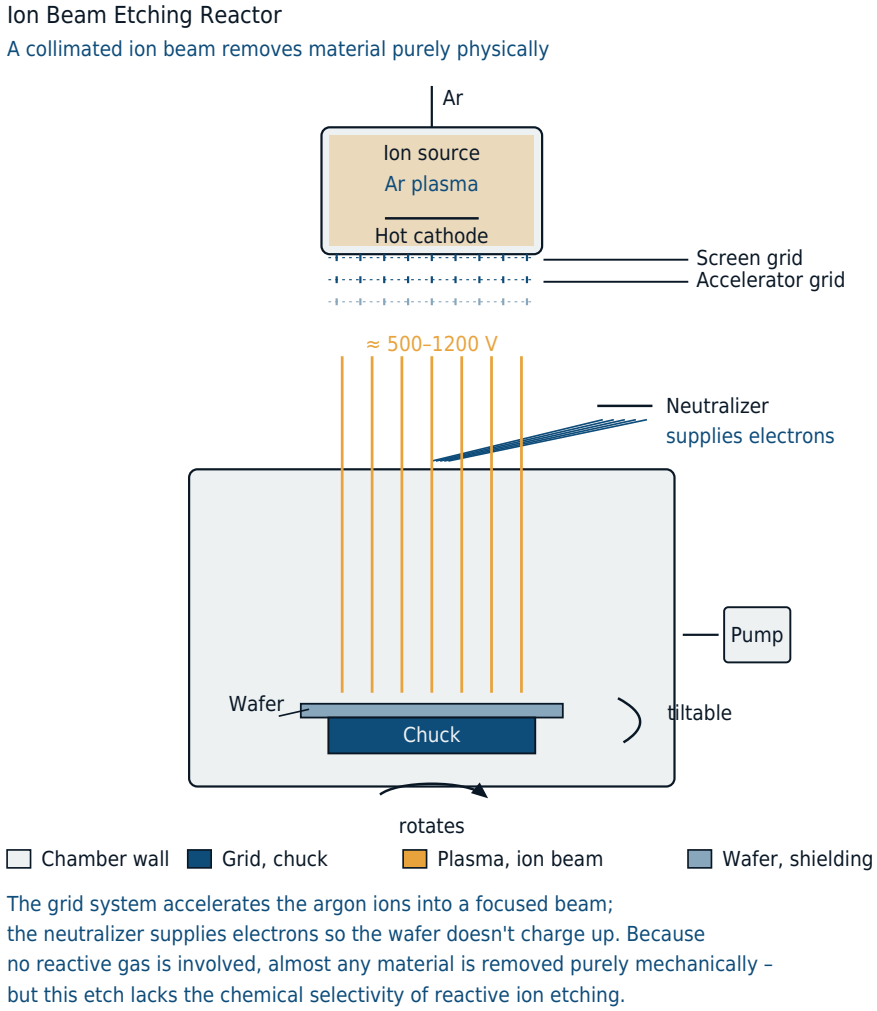
Dry etch processes

Ion beam etching

Ion beam etching (IBE) is a purely physical dry etch process. Here, argon ions are directed onto the wafer as a collimated ion beam with an energy of 1–3 keV. Due to the energy of the particles, they knock material out of the surface. The ions are generated in a source separated from the process chamber and extracted through a grid system; the working pressure is below 0.1 Pa, so that the ions retain their direction all the way to the wafer. The wafer is held perpendicular or tilted relative to the ion beam in the process chamber, and the etch proceeds in a completely anisotropic manner. The selectivity is only low, since the accelerated ions do not distinguish between materials. The gas and the

material knocked out are extracted by the pump system; however, particles also deposit on the chamber walls and on vertical edges of the wafer itself, since the removed material does not transition into the gaseous state.

To prevent particle deposition, a reactive gas is introduced into the chamber in addition to argon. This gas reacts with the argon ions, resulting in a physical etch process with a chemical character. The gas then reacts partly with the surface, but also with the material knocked out by the physical component, to form a gaseous reaction product. If the reactive gas is already added within the ion source, this is called reactive ion beam etching (RIBE); if it is introduced only in front of the wafer, it is called chemically assisted ion beam etching (CAIBE).



With this process, almost all materials can be etched. Due to the perpendicular irradiation, removal at vertical edges is very low (high anisotropy).

Because of its low selectivity and low etch rate (low ion density in the beam), ion beam etching no longer plays a role in patterning silicon and silicon compounds.

It is, however, indispensable for materials that do not form volatile reaction products and are therefore hardly etchable by chemical means:

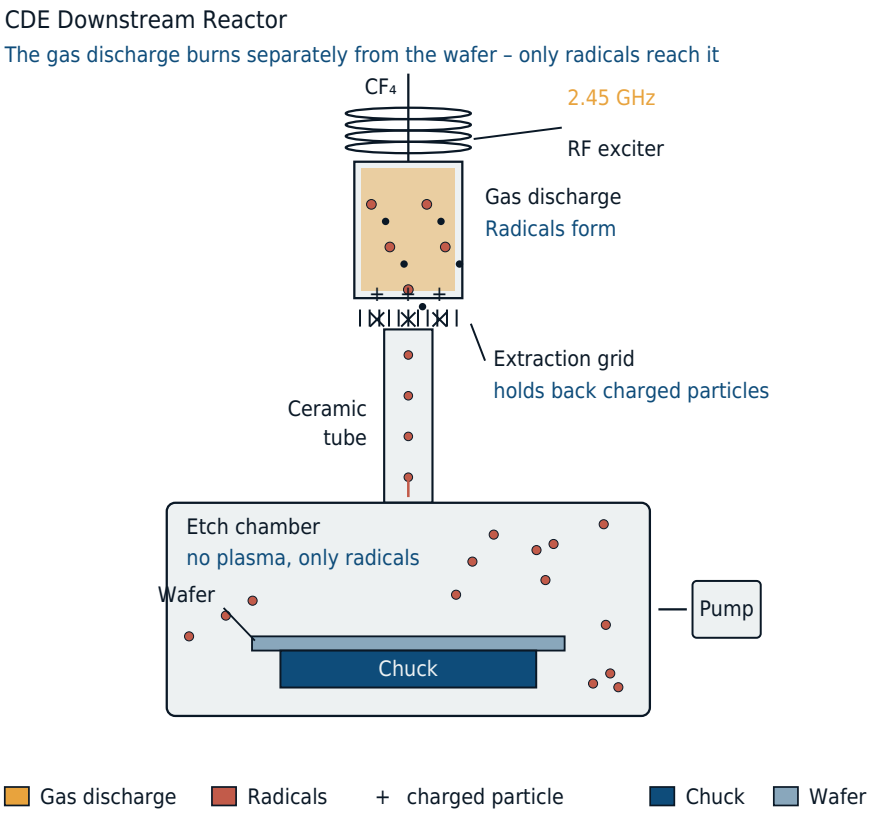
- magnetic layer stacks, such as the tunnel elements of MRAM memory cells made of CoFeB and MgO, as well as read heads and magnetic field sensors
- noble metals such as platinum, iridium, and ruthenium
- piezoelectric and ferroelectric layers such as PZT
- lithium niobate and similar crystals in integrated photonics

The tiltable, rotating substrate holder is a significant advantage here: the angle of incidence can be used to set the sidewall angle of the structure and to remove redeposited material from the sidewalls – for layer stacks made up of many different materials, a result that cannot be achieved with plasma chemical processes. Since insulating substrates would charge up under ion bombardment, the source operates with a neutralizer that adds electrons to the beam.

Plasma etching

Plasma etching is a purely chemical etch process (chemical dry etching, CDE). Its advantage is that the wafer surface is not damaged by accelerated ions. Because of the freely moving gas particles, the etch profile is isotropic, which is why plasma etching is mostly used for removing entire layers at a high etch rate.

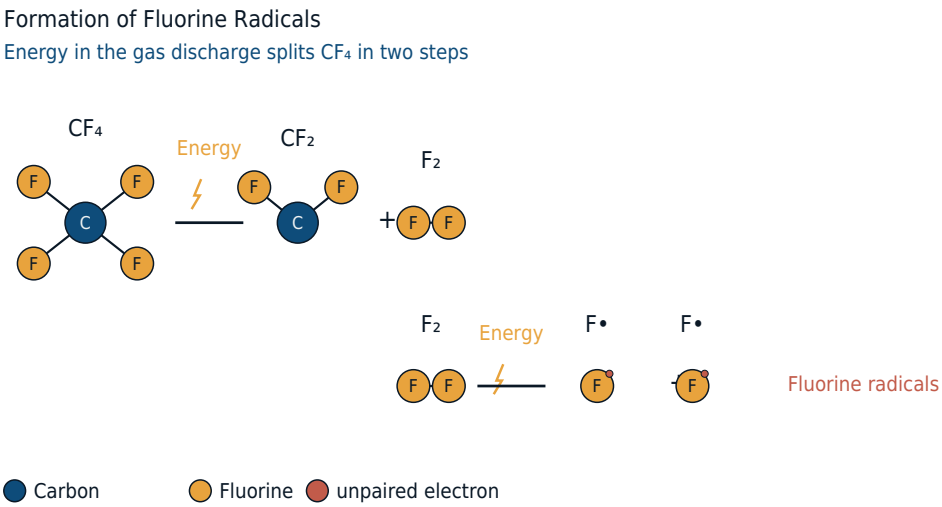
One type of equipment used for plasma etching is the downstream reactor. Here, a plasma is generated by impact ionization through the application of a high-frequency voltage (e.g. 2.45 GHz). The space where the gas discharge occurs is separated from the wafer, and the gas reaches the wafer via a ceramic tube.



In the gas discharge zone, various species are formed through collisions between the gas molecules, including radicals. Radicals are molecules with a broken bond, made highly reactive by an unpaired outer electron. As a neutral gas, tetrafluoromethane CF_4 (also known as carbon tetrafluoride), for example, is introduced and broken down by the alternating voltage into CF_2 and a fluorine molecule F_2. Fluorine can likewise be split off from CF_4 by adding oxygen:

$$2CF_4 + O_2 \rightarrow 2COF_2 + 2F_2$$

The fluorine molecule can then be split into two individual fluorine atoms by the energy present in the gas discharge zone: fluorine radicals (radicals are uncharged particles; a single fluorine atom has nine protons in its nucleus and nine electrons in its electron shell).



In addition to the neutral radicals, further particles are formed, some of them charged ($CF4^+$, $CF3^+$, $CF2^+$, ...), which are all carried together through the gas tube to the etch chamber. Charged particles are either intercepted by an extraction grid or recombine back into neutral molecules on the way. The fluorine radicals, too, are partly recombined again. Nevertheless, enough radicals reach the etch chamber, where they react with the wafer surface. Neutral particles do not take part in the reaction and, like the resulting reaction products, are extracted by the pump system.

Examples of layers etched using the CDE process:

Silicon: $Si + 4F \rightarrow SiF_4$
Silicon dioxide: $SiO_2 + 4F \rightarrow SiF_4 + O_2$
Silicon nitride: $Si_3N_4 + 12F \rightarrow 3SiF_4 + 2N_2$

Present-day relevance

Purely chemical dry etching is unsuitable for patterning layers, because the isotropic profile undercuts the mask. As a remote or downstream process – the plasma burns separately from the wafer, and only long-lived neutral radicals reach it – it fulfills clearly defined tasks in manufacturing, however:

- Resist removal (ashing): oxygen radicals burn off the photoresist after etching or implantation into CO_2 and H_2O.
- Chamber cleaning: nitrogen trifluoride NF_3 is dissociated in an upstream plasma source; the fluorine radicals remove deposits from process chambers without the chambers having to be opened.
- Selective etchback: radical processes based on NH_3 and NF_3 remove native oxide and thin nitride layers without causing damage – even inside contact

holes, which a directional etch can only reach with difficulty.
- Channel release: in nanosheet or gate-all-around transistors, silicon-germanium is removed isotropically and selectively against silicon in order to release the channel layers. Here, the isotropy that is a nuisance in patterning is deliberately exploited.

Reactive ion etching

In reactive ion etching (RIE), the etch characteristics – selectivity, etch profile, etch rate, uniformity, reproducibility – can be precisely adjusted through the gases used and the process parameters (generator power, pressure, electrode spacing, gas flow). Both an isotropic and an anisotropic etch profile are possible. This makes RIE, a chemical-physical etch process, the most important etch process for patterning various layers in semiconductor manufacturing.

Inside the process chamber, the wafers sit on an electrode fed with an alternating voltage (RF electrode). Impact ionization generates a plasma containing free electrons and positively charged ions. When the RF electrode is at a positive voltage, electrons accumulate on it and, due to the work function, cannot leave it during the positive half-cycle; the electrode thus charges up negatively to as much as 1000 V (bias voltage). The slow ions, which could not follow the fast alternating voltage, now move toward the negatively charged electrode carrying the wafers.

If the mean free path of the ions is long, the particles strike the wafers at nearly perpendicular incidence due to their velocity. Material is thereby knocked out of the surface by the accelerated ions (physical etching), while some particles also react chemically with the substrate. Vertical sidewalls are not struck, so no removal occurs there and the etch profile remains anisotropic. The selectivity is not very high because of the physical removal component, and in addition the wafer surface is damaged by the accelerated ions. This damage later has to be repaired by thermal treatment.

The chemical component of the etch occurs through the reaction of free radicals with the wafer surface and with the physically removed material, so that this material, unlike in ion beam etching, cannot redeposit on the chamber walls or on the wafers. As the pressure is increased, the mean free path of the particles decreases, so that many collisions occur between particles on their way to the wafer, continuously changing their direction. Directional etching of the wafer surface then no longer occurs; the etch process takes on a more chemical character, the etch profile becomes isotropic, and the selectivity increases.

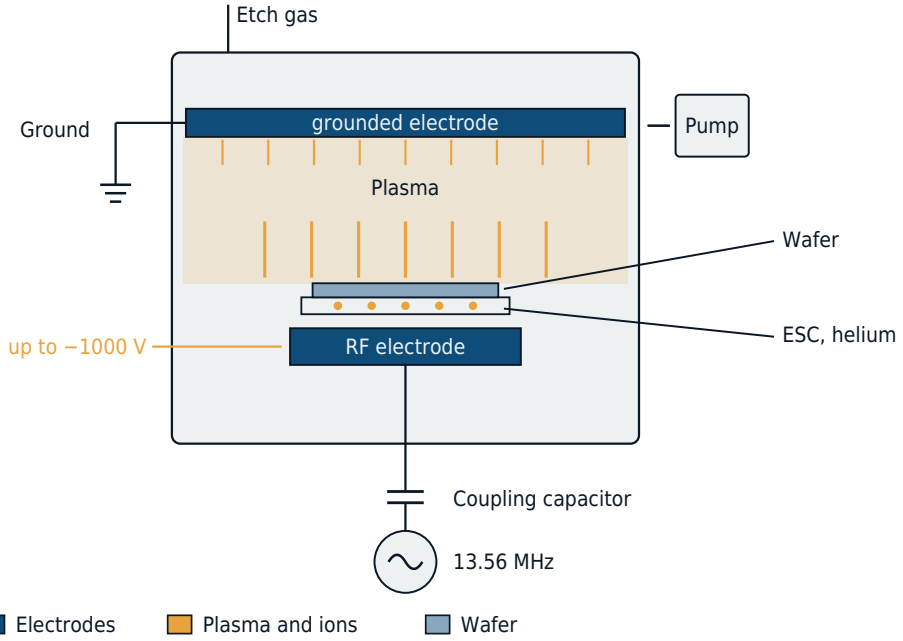
The hexode design etched several wafers simultaneously on a six-sided carrier. Such batch tools are today only of historical interest, because uniformity and endpoint can no longer be controlled with sufficient precision across many

wafers.

Today, etching is carried out exclusively on a single-wafer basis in a parallel-plate configuration. The wafer rests on an electrostatic chuck (ESC), which holds it flat and thermally couples it to the temperature-controlled electrode via helium on the wafer backside. This allows wafer temperatures to be set from about −100 °C up to over 100 °C – temperature is one of the most effective control parameters for selectivity and profile.

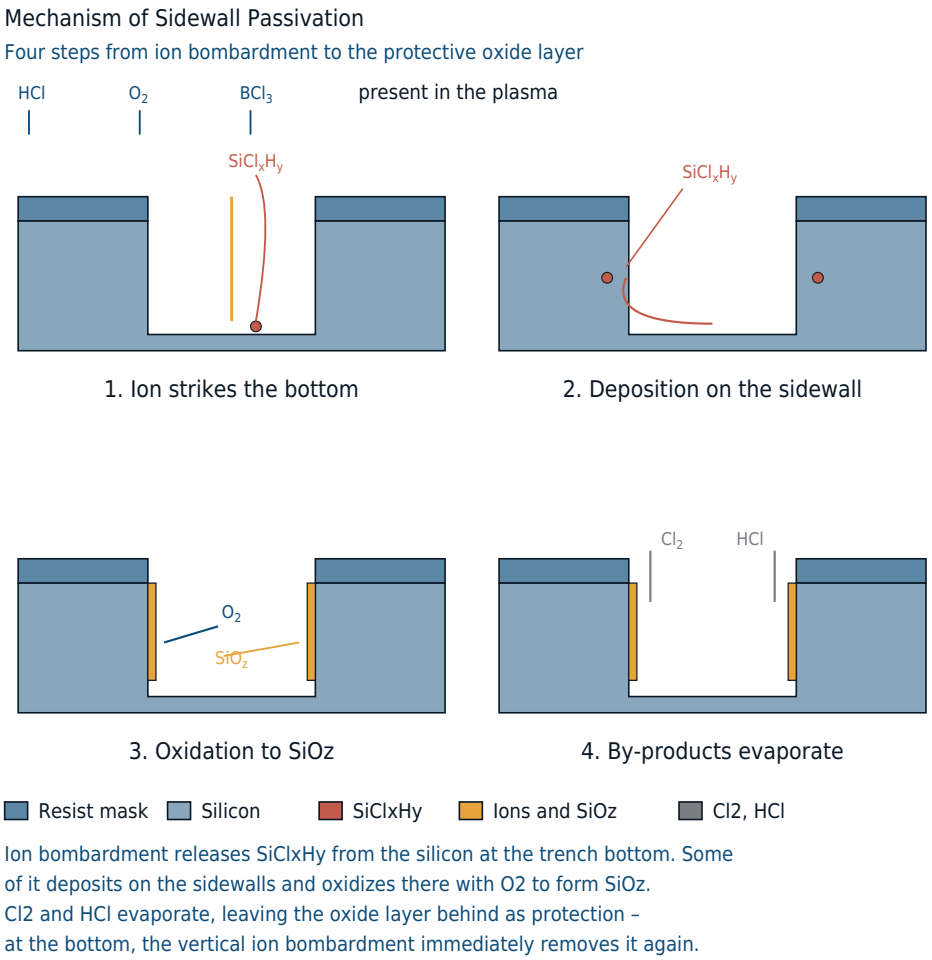
Reactive Ion Etching

Parallel-plate design: the wafer sits on the RF electrode



Because the lower electrode is smaller than the upper one, it charges negatively relative to the plasma. This accelerates the positive ions perpendicular to the wafer, removing material directionally – the etch profile becomes anisotropic.

An anisotropic etch profile is achieved in silicon etching through passivation of the sidewalls. Here, oxygen in the process chamber reacts with silicon released from the wafer surface to form silicon dioxide, which grows on the vertical sidewalls. An oxide layer on horizontal surfaces is removed by ion bombardment, allowing the etch to proceed further into the depth.



The etch rate depends in every case on the pressure, the power of the RF generator, the process gases, the gas flow rate, and the wafer or electrode temperature.

Anisotropy increases with rising RF power, decreasing pressure, and decreasing temperature. The uniformity of the etch is determined by the gases, the electrode spacing, and the electrode material. If the electrode spacing is too small, the plasma is not distributed evenly within the chamber, leading to non-uniformity. As the spacing increases, the etch rate decreases, since the plasma is spread over a larger volume. Carbon was formerly used as the electrode material: since fluorine and chlorine gases also remove carbon, the electrode causes a uniform loading of the plasma, so that the wafer edge is not stressed more than the wafer center. Today, electrodes made of silicon, silicon carbide, or quartz are common; the chamber parts facing the plasma are coated with yttrium oxide or aluminum oxide to prevent particles and metallic contamination.

Further developments of reactive ion etching

The classic RIE reactor couples plasma density and ion energy through the same electrode: more power simultaneously means more etch rate and more damage. Today's generations of equipment decouple these two quantities and extend the process in several directions.

High-density plasma sources
: In inductively coupled plasma (ICP, also TCP) tools, a coil above a dielectric window generates the plasma; in ECR tools, a microwave in a magnetic field does so. The ion energy is set separately via a bias generator at the wafer electrode. The result is high etch rates at low ion energy, i.e. less damage, at pressures well below 1 Pa.

Multi-frequency tools
: For etching dielectrics, capacitively coupled reactors operate with two or three generators of different frequencies on the same electrode, in order to set ion density, ion energy, and energy distribution separately. For deep structures, bias voltages of several kilovolts are used.

Pulsed plasma
: If the supplied power is pulsed in the kilohertz range, charge buildup in deep structures decays between pulses. This reduces profile defects such as bowing and notching and improves selectivity.

Deep etching (DRIE, Bosch process)
: In deep reactive ion etching, very short etch steps using SF_6 alternate cyclically with passivation steps using C_4F_8. This produces trenches with aspect ratios above 30:1 and depths of several hundred micrometers. The cyclic sequence leaves a wavy sidewall (scalloping). The main applications are micromechanical devices (MEMS) and through-silicon vias.

Cryogenic etching
: At wafer temperatures around −100 °C, reaction products condense on the cold sidewalls and passivate them, while ion bombardment keeps the bottom clear. This yields smooth, vertical sidewalls without scalloping. For several years, this process has also been used in volume production for etching the memory holes in 3D NAND flash: at depths of around 10 µm and aspect ratios above 50:1, the low temperatures allow etch chemistries that do not work at room temperature, and roughly double the etch rate of conventional dielectric processes.

Atomic layer etching (ALE)
: Surface activation and removal are split into separate, self-limiting half-steps and repeated cyclically. The removal per cycle is on the order of one atomic layer, controlled via the number of cycles. ALE is used wherever individual nanometers determine functionality, such as in releasing channels and etching gate structures.

Selectivity and etch rate can be strongly influenced by the etch gases used. For silicon and silicon compounds, chlorine and fluorine are primarily employed.

Etch gases used in dry etching

Processes do not necessarily run with a single gas mixture and constant etch parameters throughout. Native oxide on polysilicon, for example, can first be removed with a high etch rate and low selectivity. The polysilicon is then etched with high selectivity to the underlying layer. Real-world recipes therefore consist of several steps with different gas mixtures, powers, and pressures.

The following table gives an overview of gases used in dry etching.

Layer	Process gases	Remark
SiO_2, Si_3N_4	CF_4, O_2	F etches Si, O_2 removes carbon (C)
	CHF_3, O_2	CHF_3 acts as a polymer, increased selectivity against Si
	CHF_3, CF_4	
	CH_3F, CH_2F_2	improved selectivity of Si_3N_4 against SiO_2
	C_2F_6 / SF_6	
	C_3F_8	increased etch rate compared to CF_4
	C_4F_6, C_4F_8 + O_2, Ar	standard chemistry for deep structures (contact holes, memory holes in 3D NAND); the carbon-rich molecules form a polymer layer on the mask and sidewall, O_2 controls its thickness
	BCl_3, Cl_2	no contamination by carbon (C)
	$SiCl_4$, Cl_2 / HCl, O_2 / $SiCl_4$, HCl	
Poly-Si	HBr / Cl_2 / O_2	improved selectivity against photoresist and SiO_2; standard chemistry for gate structures
	SF_6	high etch rate, good selectivity against SiO_2
	NF_3	high etch rate, isotropic
	HBr, Cl_2	
	HBr, NF_3, O_2	higher selectivity against SiO_2
monocrystalline Si	BCl_3, Cl_2 / HBr, NF_3	
	SF_6 / C_4F_8 SF_6, O_2	deep etching: cyclic in the Bosch process, or at low temperatures in the cryogenic process
	Cl_2	isotropic etching
Al alloys	BCl_3	only low etch rate, additionally binds residual moisture and native Al_2O_3
	BCl_3 / Cl_2 / CF_4	anisotropic etching
	BCl_3 / Cl_2 / CHF_3	improved sidewall passivation
	BCl_3 / Cl_2 / N_2	higher etch rate, no carbon contamination

W, TiN, Ti	SF_6 / NF_3	etchback of tungsten in contact holes
	Cl_2, BCl_3	barrier and metal gate layers

Historically, bromine-containing chlorofluorocarbons such as CF_3Br were also used for monocrystalline silicon. As ozone-depleting substances, they are banned under the Montreal Protocol and have been replaced in manufacturing by HBr mixtures.

Metallization: why copper is not etched

Etching of aluminum alloys has become less important with the transition to copper interconnects. Copper does not form compounds with chlorine or fluorine that are volatile at process temperature and can therefore practically not be dry-etched. Instead, in the damascene process, the trench is first etched into the dielectric, copper is then deposited, and the excess is removed by chemical mechanical polishing. Aluminum continues to be patterned for bond pads as well as in power and analog technology.

Climate impact of the process gases

The perfluorinated etch and cleaning gases are chemically extremely stable – precisely the property that makes them useful in the process. In the atmosphere, they are therefore barely broken down and act as very potent greenhouse gases. According to the sixth IPCC report, the 100-year global warming potentials are around 7,400 for CF_4, around 12,400 for C_2F_6, around 14,600 for CHF_3, around 17,400 for NF_3, and around 25,000 for SF_6. Based on current estimates, CF_4 remains in the atmosphere for several tens of thousands of years.

On the manufacturing side, the exhaust of every tool is treated individually (point-of-use abatement), either thermally, catalytically, or in a plasma torch. Reactive gases such as NF_3 are broken down by more than 99 %, but CF_4 only partially; moreover, it is also formed as a decomposition product of other fluorine compounds. The contribution is therefore reduced primarily at the process level itself: by replacing C_2F_6 with NF_3 for chamber cleaning, lower gas flows, shorter etch times, and processes such as cryogenic etching with a higher etch rate. Regulatorily, these substances are covered by the EU regulation on fluorinated greenhouse gases; in addition, a restriction on PFAS is being discussed under REACH, with exemptions and long transition periods so far for semiconductor manufacturing.

Wafer Test

Fundamentals of Wafer Testing

After all front-end process steps are complete, the wafer contains a large number of finished but still unseparated dies. Before the wafer is diced, every single die is electrically tested while still on the wafer – this step is known as wafer test or probe test.

The test serves two purposes: first, it identifies defective dies so they are not unnecessarily processed further in the subsequent assembly steps. Second, it provides early electrical characterization data, such as threshold voltages or leakage currents, which can be used for process control.

A test system essentially consists of three components:

- Wafer prober: a precision tool that positions the wafer and moves it die by die under the test needles
- Probe card: the actual contacting unit with fine needles that land on each die contact pads
- Tester: the electronic test equipment that generates test patterns and evaluates the circuit response

Since modern chips can have several hundred to several thousand contact pads, the probe card must feature a correspondingly large number of precisely positioned needles that reliably make contact without damaging the sensitive chip surface.

The Probe Card

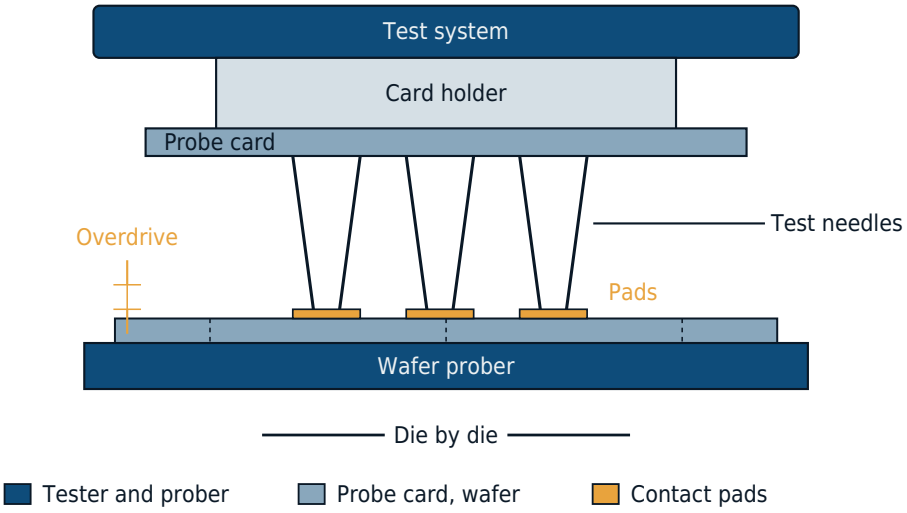
The probe card forms the mechanical and electrical interface between tester and wafer. Older designs use free-standing needles made of tungsten or a beryllium-copper alloy, mounted radially on a ring. For very fine pad pitches, MEMS-based probe cards are increasingly used today, whose contact tips are manufactured using lithographic processes from silicon or metal – a method that allows significantly higher positioning accuracy.

When the needles land on the pads, a small but deliberate force (overdrive) is applied that breaks through the native oxide layer on the pad surface, ensuring a low-resistance electrical contact. However, excessive overdrive damages the pads and can harm underlying layers – tuning this force is therefore a critical

process parameter.

Structure of the Wafer Test

The needles land on the pads before the wafer is diced



The needles are set down with a small force, the overdrive. It breaks through the native oxide layer on the pads and ensures contact. Too much overdrive damages the pads and the layers beneath them.

To increase throughput, modern test systems frequently contact multiple dies simultaneously (multi-site testing). This allows several chips to be tested in parallel, substantially reducing the test time per wafer.

Yield Mapping and Redundancy

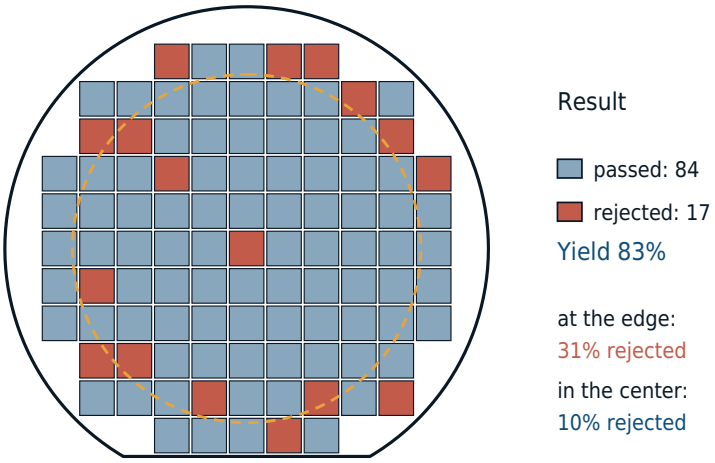
The result of each individual test is stored with its exact position in a so-called wafer map – a digital map that records, for every die on the wafer, whether it passed the test (pass) or not (fail). This map is referred to as a yield map and accompanies the wafer through all subsequent stages of production.

The yield map serves several important functions:

- During the subsequent dicing and assembly processes, only functional dies are further processed based on the map, while defective ones are discarded
- Spatial patterns in the yield map (such as clustered failures near the wafer edge) provide clues to systematic process issues
- For memory devices (DRAM, NAND flash), the map enables the targeted use of redundant circuit blocks: if a die contains individual defective memory cells, these can be electrically replaced by built-in spare circuits (redundancy), allowing the die to remain usable overall

Yield Map of a Wafer

Every die carries its test result – failures accumulate at the edge



The map accompanies the wafer through further processing: during dicing, only the passing dies are processed further. Spatial patterns like this edge loss point to systematic process problems.

The average yield of a wafer – that is, the proportion of functional dies relative to the total number – is one of the most important metrics in semiconductor manufacturing, as it directly determines the production cost per chip.

Final Test and Binning

The Final Functional Test

Once the die has received its finished package through assembly and encapsulation, it undergoes one last, comprehensive functional test – the final test. Unlike the wafer test, which typically offers only limited test coverage due to a restricted number of needles and contact quality, the finished, packaged chip can be tested far more comprehensively and under more realistic electrical conditions through its regular pins.

Tests performed include:

- Complete functional verification of all logic and analog circuit blocks
- Electrical parameters such as operating voltage, current consumption and signal timing
- Behavior under varying temperature and voltage conditions

- For memory devices: a complete cell test across the entire address range

Only after passing the final test is a chip considered ready for sale. Defective components are sorted out and typically scrapped, since repair at this stage of production is no longer economically viable.

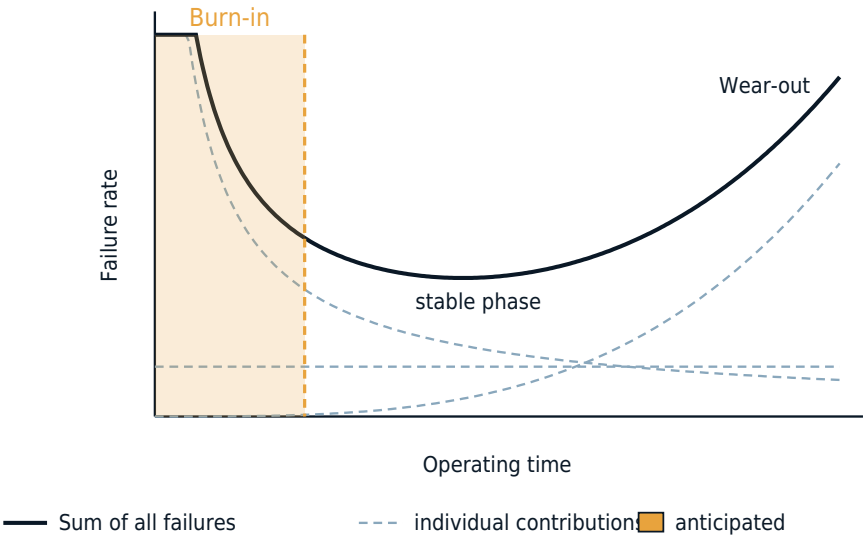
Burn-in

For applications with particularly strict reliability requirements (such as automotive or aerospace electronics), a so-called burn-in process is often performed before the final test itself. During burn-in, chips are stressed for a defined period – ranging from a few hours to several days – under elevated temperature and, in some cases, elevated operating voltage.

The rationale lies in the so-called bathtub curve of electronic component failure rates: an unusually high proportion of early failures typically occurs within the first operating hours of a component, most often caused by hidden manufacturing defects such as material flaws or contamination that the final test alone would not catch. After this early-failure phase, the failure rate drops to a low, stable level before rising again toward the end of the component lifetime due to wear-related effects.

Bathtub Curve of the Failure Rate

Burn-in anticipates the early failures



Hidden manufacturing defects show up in the first hours of operation. Under elevated temperature and voltage, these components already fail at the manufacturer – only what survives the steep initial region gets shipped.

The burn-in process anticipates these early failures: components with hidden defects fail during burn-in itself and are sorted out before ever reaching a

customer. The surviving portion of chips consequently exhibits significantly higher reliability during subsequent field use.

Binning

Even within a single wafer, and indeed within the very same die design, unavoidable process variations lead to slight differences in electrical characteristics – for instance in the maximum achievable clock frequency of a processor or in power consumption. Rather than discarding chips with slightly inferior characteristics, they are instead sorted according to their actual performance and marketed as different product variants. This practice is known as binning.

A well-known example is processor families in which technically identical chips are sold under different model names and at different prices depending on the maximum clock frequency they achieve. Chips that fail to meet the highest requirements are placed into a lower performance segment (bin) rather than being discarded as scrap.

Binning allows manufacturers to economically capture the natural variation inherent in manufacturing, rather than treating it purely as yield loss. Following binning, chips receive their final marking – typically via laser engraving of the package – and are packaged for shipment to device manufacturers.

Dicing

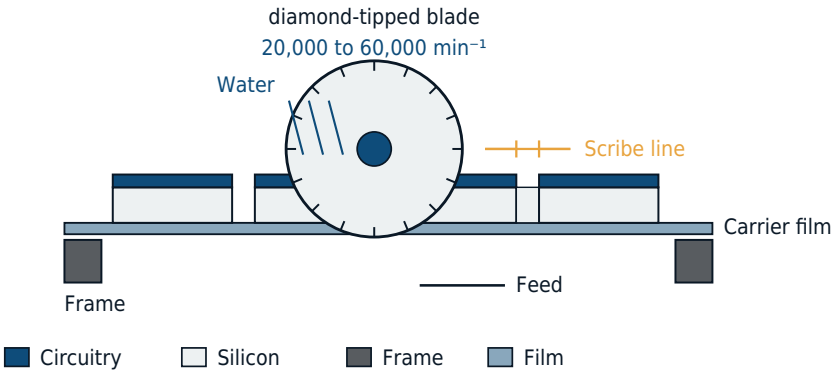
From Wafer to Individual Chip

After the wafer test, the wafer still exists as a single, continuous disc containing hundreds to several thousand individual dies, separated by narrow, unpopulated strips – the so-called scribe lines or dicing lanes. These strips contain no functional circuit structures and are deliberately reserved for the subsequent separation step.

During dicing (also called wafer sawing or die separation), the wafer is cut into individual dies along these scribe lines. This step marks the transition from the front-end of manufacturing, where all dies were still processed together on the wafer, to the back-end, where each die is handled individually from this point onward.

Dicing – Sawing the Wafer

The saw cuts along the scribe lines, the film holds the dies



The scribe lines carry no circuitry – their width limits how thick the blade can be. As they get narrower with miniaturization, laser methods are gaining importance. The film keeps the dies in place after cutting.

Before the actual separation, the wafer is typically mounted onto a thin, elastic carrier film (dicing tape) stretched over a metal frame. This film keeps the individual dies in their original position after sawing, preserving the spatial correspondence to the previously generated yield map.

Sawing Methods

The most widely used method is mechanical sawing with a thin, diamond-embedded circular blade (blade dicing). The blade rotates at very high speed (typically 20,000 to 60,000 revolutions per minute) and is continuously cooled and flushed with deionized water during cutting to dissipate frictional heat and remove sawing debris from the cut.

The width of the scribe lines limits how narrow the saw blade can be and how precisely the cut must be guided. As available scribe line width continues to shrink with ongoing miniaturization, alternative separation methods are gaining increasing importance:

- Laser dicing: a focused laser beam separates the wafer either by direct ablation of material or by deliberately introducing internal micro-cracks (stealth dicing), along which the wafer is subsequently broken apart mechanically
- Plasma dicing: the wafer is completely etched through along the scribe lines using an anisotropic dry etch process; this method produces particularly smooth, damage-free edges and is well suited for very thin wafers

Each method has specific advantages and disadvantages regarding cutting speed, mechanical stress on the material, and achievable edge quality, which is why the choice of separation method depends on the specific product and wafer layer stack.

Challenges and Quality Assurance

During mechanical sawing, fine chipping can occur at the edge of the cut – small fractures of the brittle silicon material. These micro-cracks can propagate further under mechanical or thermal stress during the component later operation, potentially compromising reliability. Sawing parameters – particularly feed speed, rotational speed and cooling – must therefore be carefully matched to the specific layer system of the wafer.

Particular challenges arise with wafers containing sensitive layer stacks, such as components with low-k dielectrics in the wiring layers. These porous materials are considerably more mechanically fragile than classic silicon dioxide and are more prone to delamination – the separation of individual layers – during sawing.

After separation, the quality of the cut edges is frequently inspected optically. In addition, the previously generated yield map remains valid: since the dies remain in position thanks to the carrier film, the map can be used to unambiguously determine which of the now-separated dies were already identified as functional during the wafer test, and therefore proceed to the next

assembly step, die attach.

Die Attach

Mechanical Mounting of the Die

After dicing, the individual, already identified functional dies remain on the carrier film. In the first step of the actual assembly process, each die is lifted from the film (die pick) and mechanically mounted onto its final carrier material – this process is referred to as die attach or die bonding.

Depending on the component type, different carrier materials (substrates) are used:

- Leadframe: a stamped or etched metal grid, usually made of copper or a copper alloy, which later forms the outer leads of the package
- Laminate or ceramic substrate: multilayer carrier boards with integrated wiring, as used for example in ball-grid-array packages (BGA)
- Direct chip attach: in some applications, the die is mounted directly onto a printed circuit board or another system without first receiving its own package

The die is precisely picked up by a gripping tool (die bonder), aligned, and placed onto the substrate with a defined contact force. Accurate positioning is important, since in the subsequent bonding step the die contact pads must line up as closely as possible with the corresponding connection points on the substrate.

Bonding Materials

Various materials are used to actually attach the die to the substrate, with the choice depending on the electrical, thermal and mechanical requirements of the specific component:

- Adhesives (epoxy resins): by far the most commonly used method. The adhesive is either dispensed as a paste or applied in pre-formed film form (die attach film) between die and substrate, and subsequently cured thermally. Electrically conductive adhesives additionally contain fine silver particles to establish, in addition to the mechanical bond, an electrical connection between the backside of the die and the substrate
- Solder joints: particularly for components with high heat dissipation requirements, such as power semiconductors, the die is bonded directly to the substrate using a solder material (often gold- or tin-based). Solder joints offer significantly better thermal conductivity than adhesives

- Eutectic bonding: the die is bonded directly to a thin eutectic metal layer (frequently a gold-silicon alloy) on the substrate under the application of heat, without requiring an additional paste-form bonding material

The choice of bonding material significantly influences how efficiently the heat generated by the chip during operation can be dissipated through the substrate – a particularly critical factor for high-performance processors and power semiconductors.

Mechanical Stress and Quality Assurance

A key challenge in die attach is mechanical stress arising from differing thermal expansion coefficients between die and substrate. Silicon and common substrate materials such as copper or ceramic expand to different degrees with temperature changes. During curing of the bonding material, and later during operation of the finished component, this can generate stress within the die that, in the worst case, leads to cracking.

To limit this stress, bonding materials with a certain degree of elasticity are used, capable of compensating for small expansion differences without damaging either the bond itself or the die. In addition, the thickness of the bonding material layer is carefully controlled: a layer that is too thin provides insufficient compliance, while a layer that is too thick can worsen the thermal connection.

After the bonding material has cured, the quality of the die attach bond is inspected, among other things through optical inspection for voids (cavities within the bonding material) and through shear strength testing, in which the mechanical adhesion of the die is checked on a sample basis. Only after successful inspection does the die proceed to the next process step, bonding.

Bonding

Electrical Contacting of the Die

After the die has been mechanically attached to the substrate through die attach, it still needs to be electrically connected to the substrate connection structures. This process is referred to as bonding. Electrical contacting is achieved in about 90% of cases via wire bonding, while the mechanical attachment of the chip – as described in the previous chapter – is mainly carried out by bonding with epoxy resins.

In wire bonding, thin wires made of gold or aluminum are used, which are cold-

welded to the contact pads of both chip and substrate. Three common variants are distinguished:

- Thermosonic ball-wedge bonding
- Ultrasonic wedge-wedge bonding
- Thermocompression ball-wedge bonding

In all three methods, a small ball first forms at the end of the wire, created by locally melting the wire tip. Due to the strong surface forces of the molten gold, the melt solidifies into a ball whose larger diameter compared to the bond wire ensures a more reliable mechanical and electrical connection.

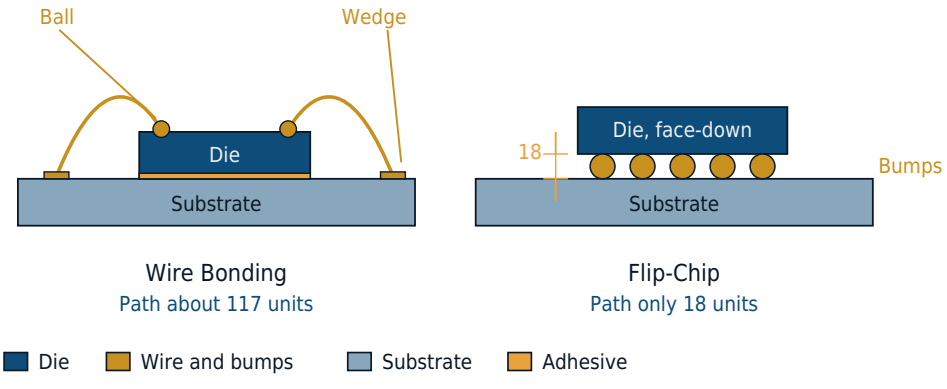
Flip-Chip Technology

In addition to classic wire bonding, flip-chip technology offers an alternative method that requires no long bond wires at all and represents the most compact form of connection between chip and substrate.

To create a flip-chip connection, small solder bumps must be present on the chip and/or the substrate. To form the connection, the chip is placed face-down, but with very precise alignment, onto the substrate. During a subsequent reflow process, the bumps melt and simultaneously form a mechanical and electrical connection between chip and substrate.

Wire Bonding and Flip-Chip

The length of the connection path determines inductance and heat



The bond wire is roughly 6 times as long as the path across a bump. Short paths mean less parasitic inductance and better heat dissipation – both favor flip-chip technology.

The key advantage of flip-chip technology lies in the very short connection paths

from chip to substrate via the solder bumps – in contrast to the comparatively long wire bonds used in classic wire bonding. This reduces parasitic inductance and capacitance in the connection, which is electrically advantageous at very high signal frequencies.

Since a flip-chip connection typically leaves an air gap between chip and substrate, it is usually necessary to additionally introduce a so-called underfiller. This fills the remaining gap between chip and substrate and compensates for mechanical stress arising from the differing thermal expansion coefficients of the two materials.

Requirements and Heat Dissipation

Regardless of the bonding method chosen, the electrical connections between chip and substrate must meet a number of fundamental requirements: low electrical resistance, high reliability over the component entire service life, and good mechanical adhesion even under thermal cycling.

An increasingly important aspect of bonding is heat dissipation. As the power dissipation of modern chips continues to rise, removing the resulting heat becomes a very important consideration throughout the entire packaging and interconnection process. Short conduction paths and materials with good thermal conductivity are indispensable for this. This is precisely where flip-chip technology offers advantages over classic wire bonding, since the short connection paths via the bumps are also more favorable for heat dissipation than long bond wires.

Once bonding is complete, the chip is fully electrically connected to its substrate. The next step in assembly is to protect this sensitive connection, as well as the die itself, from external influences through suitable encapsulation.

Encapsulation

Purposes of Encapsulation

After bonding, the die is fully electrically connected to its substrate but still lies exposed and unprotected. In the final major step of assembly, the die is therefore embedded in a protective molding compound – this process is referred to as encapsulation or molding.

As part of the overall packaging process, encapsulation fulfills several fundamental purposes:

- Protecting the circuit from external influences such as moisture, dust and

mechanical damage
- Routing the sensitive on-chip contacts out to robust, manageable external connections
- Supplying the circuit with electrical power via the package leads
- Dissipating the power loss generated during operation
- Distributing signals and data between the chip and the surrounding circuitry

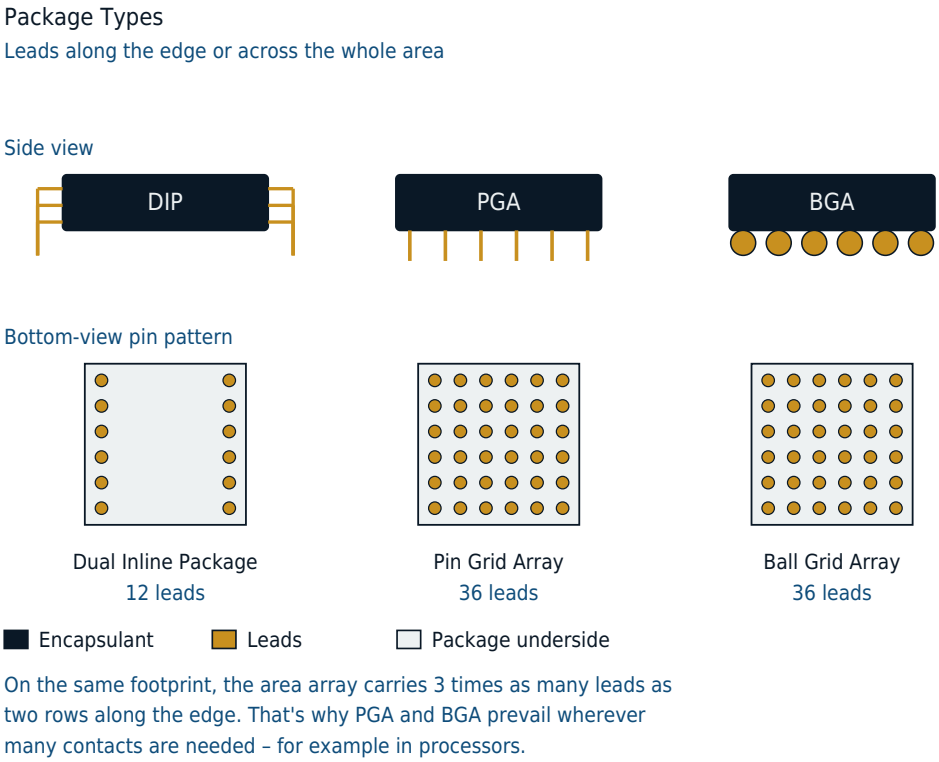
The trend in packaging technology is toward increasingly smaller and thinner packages (chip size package, CSP), in which the finished package is only marginally larger in length and width than the actual chip, typically by a factor of around 1.2.

Materials and Methods

The most widely used method is molding, in which the die, together with its bond wires or bumps, is injected into a thermoset plastic compound. This molding compound cures under heat and subsequently forms the actual, solid package of the component.

Well-known package types produced this way include, for example:

- DIP – Dual Inline Package
- PGA – Pin Grid Array, commonly used especially for processors
- BGA – Ball Grid Array



For components with particularly high heat dissipation requirements, such as power semiconductors, packages made from highly thermally conductive materials such as metal or ceramic are used instead of plastic. However, these materials are considerably more expensive than plastic, which is why mass production predominantly relies on plastic packages.

In the flip-chip technology described earlier, a classic full package is often omitted entirely: here, the underfiller already introduced during the bonding step takes over part of the protective function, while the chip itself remains exposed on the substrate.

Heat Dissipation and Final Quality Inspection

As the power dissipation of modern chips continues to increase, heat dissipation is one of the most important considerations when selecting package type and molding material. Short conduction paths and materials with good thermal conductivity are indispensable for this. For particularly high-performance components, additional heat dissipation measures are therefore frequently integrated into the package design, such as exposed metal areas on the underside of the package (exposed die pad) or dedicated thermal paths within the substrate.

After the molding compound has cured, the package is given a unique marking, typically via laser engraving, which carries manufacturer information, type designation and, where applicable, the result of the preceding binning step.

This concludes the actual assembly process. The fully encapsulated chip subsequently undergoes the final functional test described earlier, before being packaged for shipment to device manufacturers.

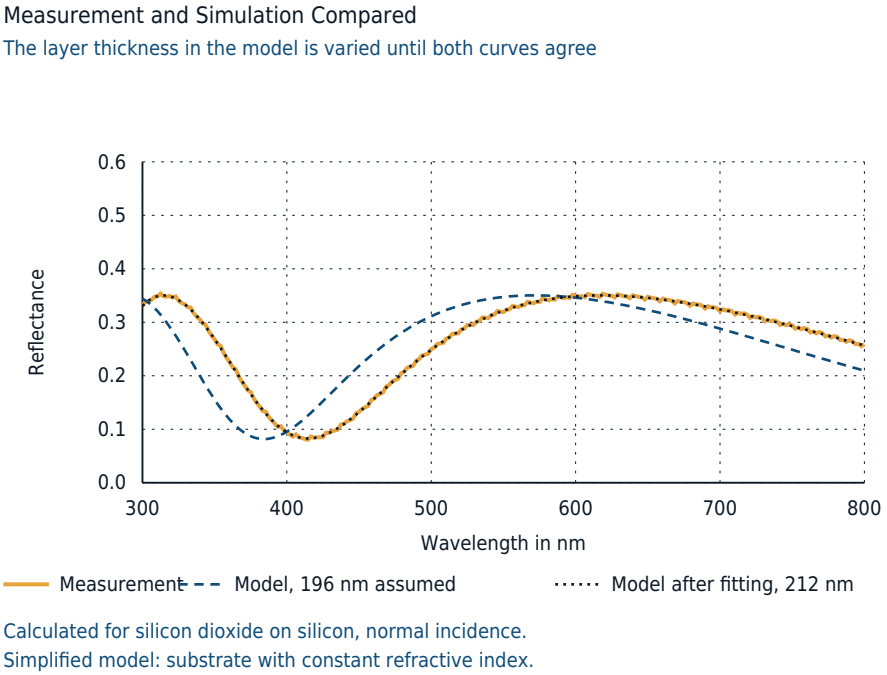
Film thickness measurement

Evaluation of the Measurement

With these optical measurement methods, layer thickness is determined only indirectly; the optical parameters of the layers being measured must be sufficiently well known beforehand. Using these values, a model of the layer stack on the wafer is created and a measurement is simulated. The result is then compared with the actual measurement, and the layer thicknesses (as well as other optical indices) of the materials defined in the model are varied until the simulation and the measurement match as closely as possible.

Comparison of simulation and measurement in an optical scatterometry measurement

(Source: n&k Technology)



The more parameters that are varied in the model, the easier it becomes to

achieve a match with the measurement, but the less certain the result also becomes. Typically, a parameter known as goodness of fit (GOF) indicates how well the simulation matches the measurement (0–100 %).

Metrology

Oxide layers are transparent films. If light is irradiated onto the wafer and reflected, various properties of the light wave are changed which can be detected with metering devices. If there are multiple layers stacked on each other they must differ in optical properties to allow a determination of the materials.

To monitor the film thickness across the entire wafer, several measuring points are measured (e.g. 5 points on 150 mm, 9 points on 200 mm, 13–21 points on 300 mm wafers). Not only must the values of the individual points be compared with the specified target value, but also the thickness of the points relative to one another, since homogeneity is also important for subsequent processes. If the deposited layer is too thick or too thin, material has to be removed (e.g. by chemical mechanical polishing) or additional material has to be deposited (by repeating the corresponding process).

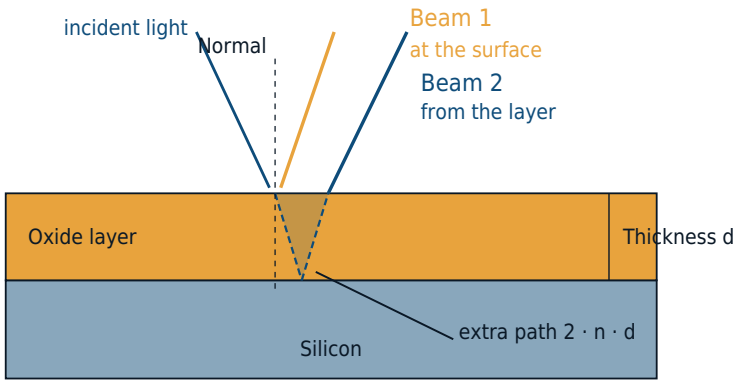
Interferometry

If light waves interfere with each other, they can be amplified, weakened, or even canceled out. This phenomenon is used in semiconductor manufacturing for measuring transparent layers.

Light rays incident on the wafer are partially reflected at a transparent layer, while part of the rays also penetrates the material and is reflected at the layer underneath.

Light Path in Film Thickness Measurement

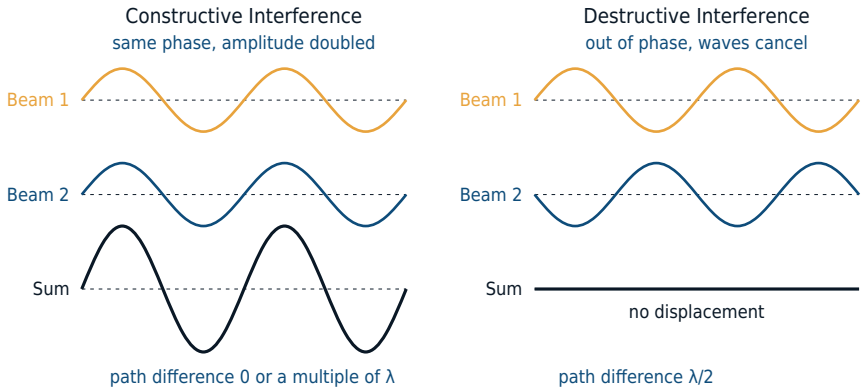
Part is reflected at the top, part only at the interface below



Both beams leave the layer parallel to each other and combine in the measuring instrument.
Beam 2 has traveled the longer path: twice the layer thickness,
multiplied by the refractive index n of the layer.

Superposition of Two Light Waves

The path difference determines whether they reinforce or cancel



In an actual measurement there is a smooth transition between the two extremes:
the strength of the reflected light reveals the thickness of the layer.

A spectrum of different wavelengths is irradiated onto the wafer. Depending on the thickness of the layer being penetrated, the reflected rays interfere differently, resulting in a characteristic interference pattern for each film thickness and material. Using a photometer, the film thickness can be determined from the reflected light.

Interference measurements are possible for layers whose thickness corresponds to roughly one quarter of the irradiated light wavelength or more.

Ellipsometry

Ellipsometry is the determination of optical properties through the change in polarization of light. Linearly polarized light (the light waves have a specific oscillation) is irradiated onto the wafer at a fixed angle.

Illustration of possible polarizations of light (left linear, right circular):

(Source: Optics Group, University of Glasgow)

Upon reflection at the wafer surface or at the interface between two layers, the light is repolarized. This change can then be measured with an analyzer. From the known optical properties of the layers (e.g. refractive angles and absorption coefficient), the irradiated wavelength, and the polarization, the film thickness can be determined.

In contrast to measurement by interference, ellipsometry is suitable for thin layers whose thickness is less than one quarter of the irradiated light wavelength.

Particle inspection

Metrology

Particles are among the most critical sources of defects in semiconductor manufacturing. A single sub-micrometer particle, if it lands on an active structure, can already cause a short circuit, an open line, or another electrical defect, rendering an entire chip non-functional. Since modern feature sizes are now in the single-digit nanometer range, particle contamination on the wafer must be monitored continuously.

Particles can originate from a wide variety of sources: from the ambient air of the cleanroom, from equipment components (e.g. flaking coatings, seal abrasion), from process gases and chemicals, or even from the wafer itself (e.g. crystal defects that emerge at the surface). To narrow down the cause and ensure cleanroom quality as well as equipment performance, wafers are checked for particle contamination after critical process steps.

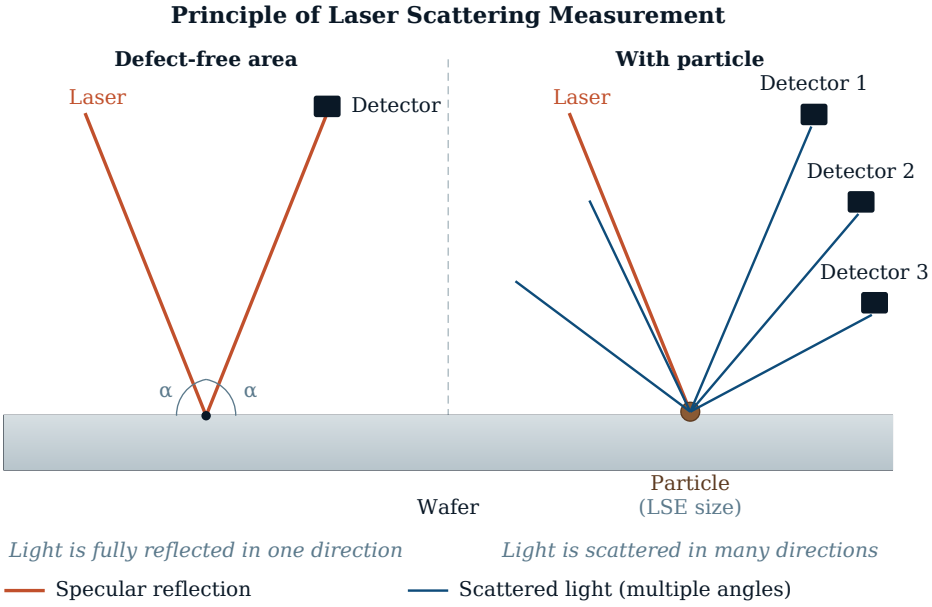
The most common method for particle detection uses the interaction of light with the wafer surface, similar to some film-thickness measurement techniques. However, since particles do not form large, homogeneous layers but rather represent localized point defects, this method does not rely on the interference

or polarization of reflected light, but instead on the scattering of light at small objects.

Laser Scattering Measurement

In laser scattering measurement, the wafer surface is systematically scanned with a focused laser beam. When the beam hits a flat, defect-free area, the light is reflected in a largely directional manner. If, however, a particle, a scratch, or another topographic irregularity is present at that location, the light is scattered in almost all directions.

Principle of scattered-light measurement at a particle



One or more detectors positioned outside the direct reflection path capture this scattered light. The intensity of the scattered light can be used to approximate the particle size – however, this is typically not reported as the actual physical dimension, but as the so-called Latex Sphere Equivalent (LSE) size, i.e. the size of a reference sphere (usually polystyrene latex) that would produce the same scattered-light intensity.

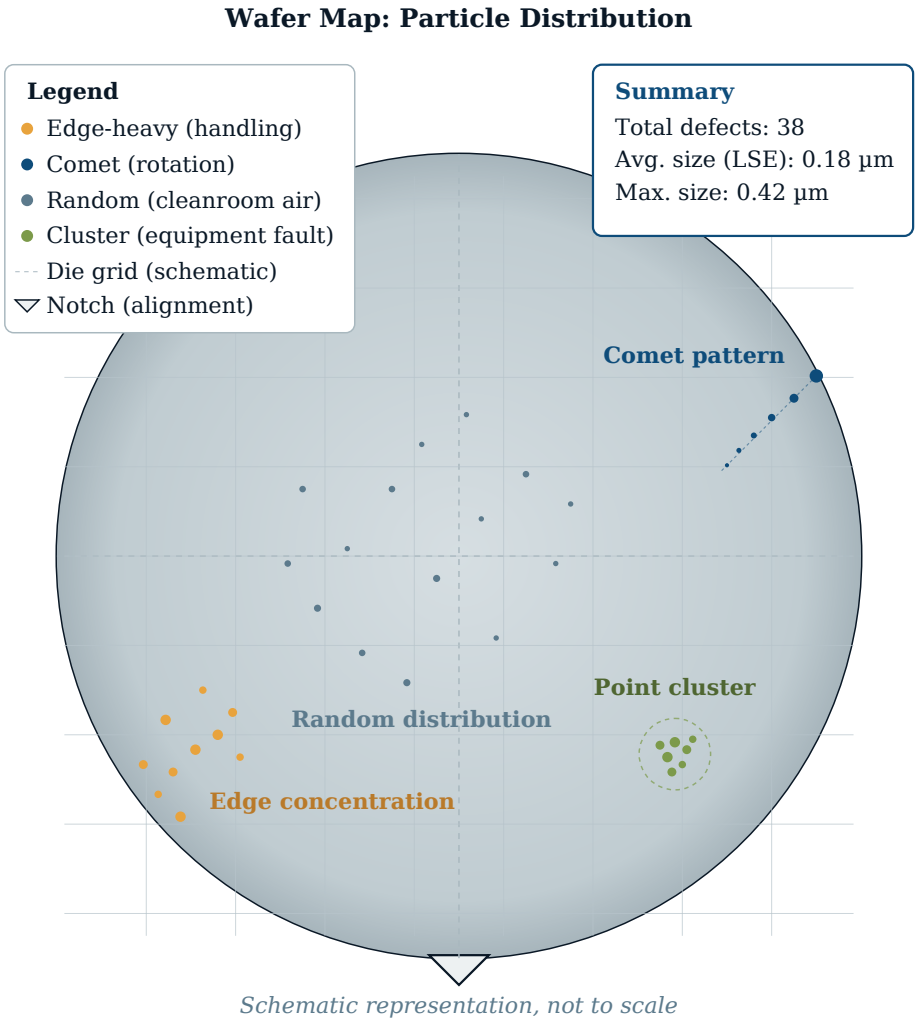
Because different materials (metal, photoresist, crystal fragments, etc.) scatter light to differing degrees, the real particle size often deviates from the measured LSE size. Modern systems therefore frequently use several wavelengths and detection angles simultaneously (multi-channel or multi-angle scattering) to allow initial conclusions about the particle type in addition to its size.

Modern inspection systems achieve a detection limit of roughly 20–30 nm (LSE), although actual sensitivity depends heavily on the wafer surface (film structure, roughness, reflectivity) and the wavelength used. Patterned wafers with existing device structures are inherently more difficult to inspect than unstructured (blanket) wafers, since structure edges themselves scatter light and can therefore be falsely identified as particles.

Wafer Map and Classification

The result of a particle measurement is usually displayed as a so-called wafer map: a graphical top-down view of the wafer on which every detected defect is marked as a point at its exact x/y coordinate. In addition to the position, the LSE size and a preliminary classification code are stored for each defect.

Example of a wafer map with particle distribution



The spatial distribution of the defects already provides important clues about the root cause. Typical patterns include, for example:

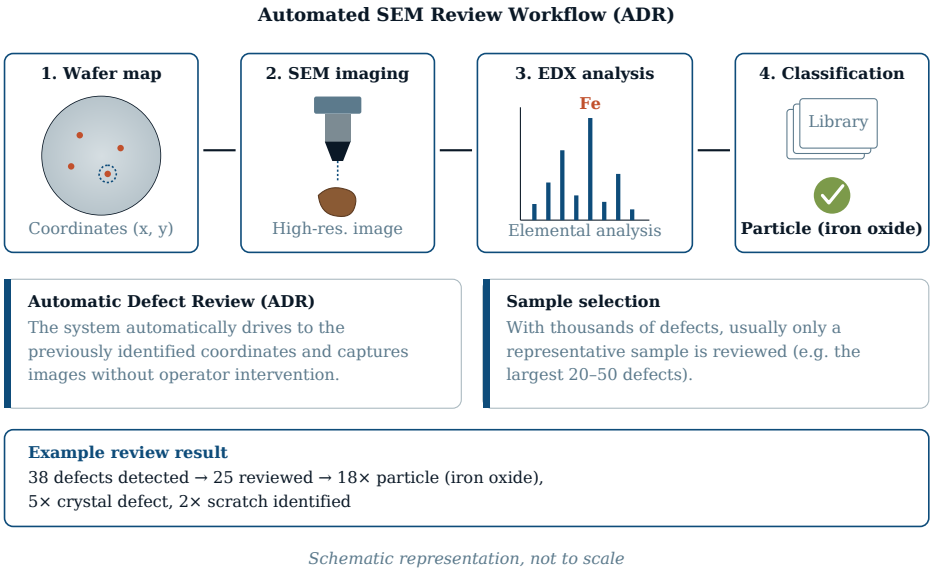
- Edge-heavy concentration: often indicates handling problems or uneven process conditions at the wafer edge (e.g. during spin-coating or CMP processes).
- Comet-shaped patterns: often result from a single, larger particle that is dragged across the wafer during a rotational process (e.g. resist spinning), leaving a trail of further, smaller defects behind it.
- Random, uniform distribution: usually points to contamination from the ambient air of the cleanroom.
- Localized clusters at specific coordinates: can indicate a faulty piece of equipment that touches the wafer at the same location every time (e.g. a wafer handler or chuck).

The inspection software automatically assigns a preliminary class to each defect based on its scattered-light signature, size, and shape (so-called Automatic Defect Classification, ADC). However, this automatic pre-classification is not always reliable, especially for new or unknown defect types, which is why critical or conspicuous defects must be examined more closely in a further step.

Review with the Scanning Electron Microscope (SEM)

Since optical scattered-light measurement only provides position and an approximate size, but no reliable information about the defect's shape and material, selected particles are subsequently examined more closely in a so-called defect review. For this purpose, the wafer map is imported into a scanning electron microscope (SEM), which automatically drives to the previously determined coordinates (Automatic Defect Review, ADR).

Workflow of the automated SEM review



Upon reaching the target location, the SEM captures a high-resolution image of the defect. Since modern inspection systems can detect several thousand defects per wafer, it is not practically possible to review every single defect. Instead, a representative sample is usually driven to automatically (e.g. the largest 20–50 defects, or a random selection per class).

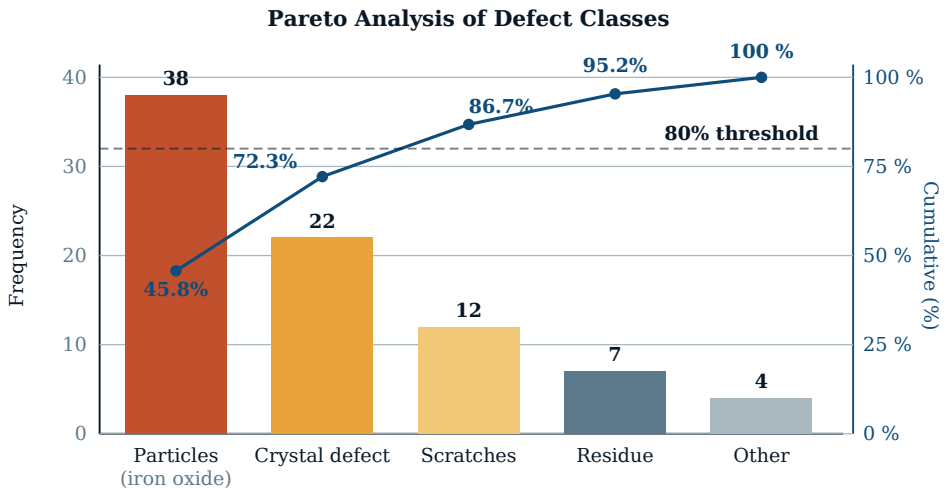
In addition to plain image capture, many SEM systems feature an energy-dispersive X-ray detector (Energy Dispersive X-ray Spectroscopy, EDX). This allows the chemical composition to be determined directly at the defect location, which enables conclusions about its origin: if, for example, a metal is detected that is not used in any of the current process steps, this points to contamination from an equipment component (e.g. worn seals, chamber coating).

Based on shape, size, and material composition, the defect is then finally classified (e.g. particle, scratch, crystal defect, residue). This classification is usually compared against historical reference images (a so-called defect library), either manually by a technician or automatically via image recognition.

Evaluating the Measurement

The results of particle measurement are not only considered for the individual wafer, but are evaluated statistically across multiple wafers, lots, and a longer time period. A key metric here is the total defect count per wafer (Total Defect Count) above a defined size threshold, which is compared against control limits (Upper Control Limit, UCL). If this limit is exceeded, the process is considered “out of control” and must be investigated.

Pareto evaluation of the defect classes across multiple wafers



Schematic representation, not to scale

A common method for prioritization is Pareto analysis: the various defect classes are displayed sorted by frequency, so that the class with the largest share is immediately apparent. Since experience shows that the majority of defects can usually be traced back to a few main causes, troubleshooting can be focused specifically on the most frequent classes instead of tracking down every individual cause separately.

Of particular importance is also the correlation between defect density and the later electrical yield of the chips. Not every defect actually leads to a device failure – a particle lying on an uncritical area (e.g. in the scribe line between the chips) has no effect on functionality. By comparing defect positions with the later test results of the individual chips (yield map), it is possible to statistically determine which defect classes and sizes are actually yield-relevant. This so-called defect-to-yield correlation is one of the most important tasks of failure analysis in semiconductor manufacturing, as it allows resources to be focused specifically on the truly critical defect sources instead of pursuing every detected particle class with equal intensity.

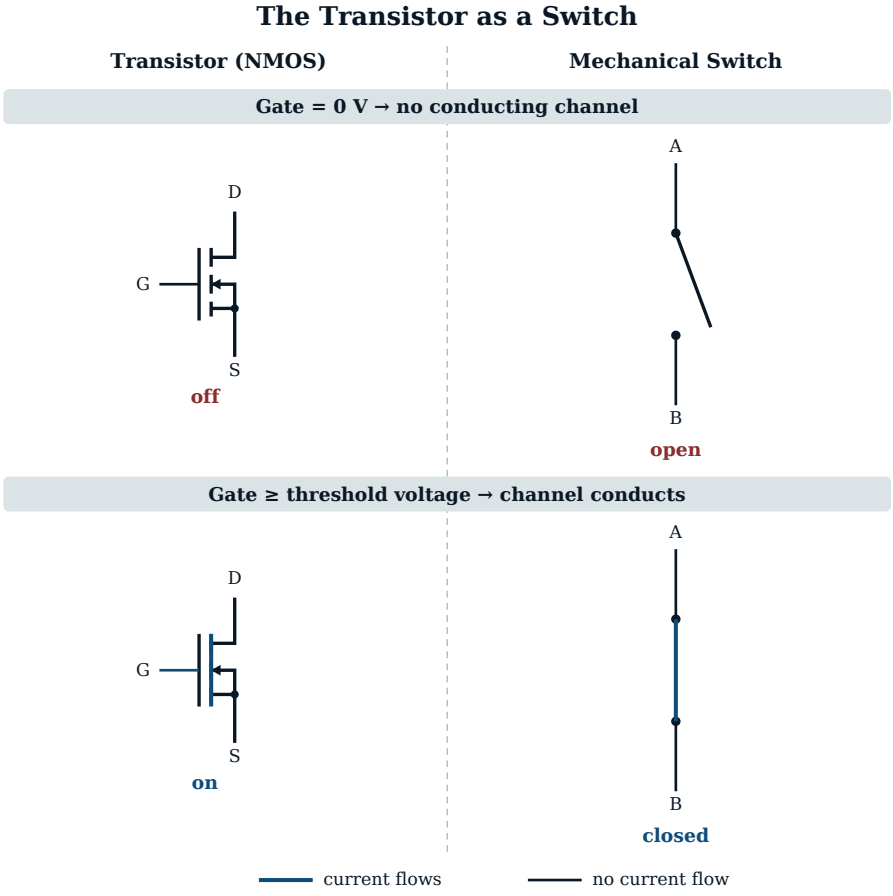
In the long term, the insights gained feed into a continuous improvement program: equipment with elevated particle loads is specifically maintained or requalified, process parameters are adjusted, and cleanroom conditions are improved as needed, in order to reduce defect density and, ultimately, manufacturing cost per functional chip.

From transistor to logic gate

The Transistor as a Switch

A MOSFET can be reduced to its simplest function: a voltage-controlled switch. Current either flows between source and drain or it doesn't, depending on the voltage applied to the gate. This makes it fundamentally different from a mechanical switch — nothing moves, there's no wear, and switching takes only a fraction of a nanosecond.

In an enhancement-mode NMOS transistor, no conducting channel exists between source and drain without gate voltage. Only once the gate voltage exceeds the threshold voltage does the channel form, and the transistor conducts. Think of the gate as a switch's control input: 0 V at the gate → switch open; enough positive voltage at the gate → switch closed.



NMOS and PMOS Compared

Alongside the NMOS transistor there's its counterpart, the PMOS transistor. It behaves as a mirror image: while the NMOS conducts at a high gate level, the PMOS blocks at exactly that point — and instead conducts at a low level. This complementarity isn't a footnote; it's the core principle behind all of CMOS technology ("Complementary MOS").

The NMOS is typically used as the connection to ground (0 V), the PMOS as the connection to the supply voltage. This interplay lets both transistor types together switch an output cleanly between High and Low, without ever creating a direct short between supply and ground.

NMOS and PMOS Compared

Gate voltage	NMOS	PMOS
Low (0 V)	**off**	**on**
High (VDD)	**on**	**off**

The Inverter — the First Gate

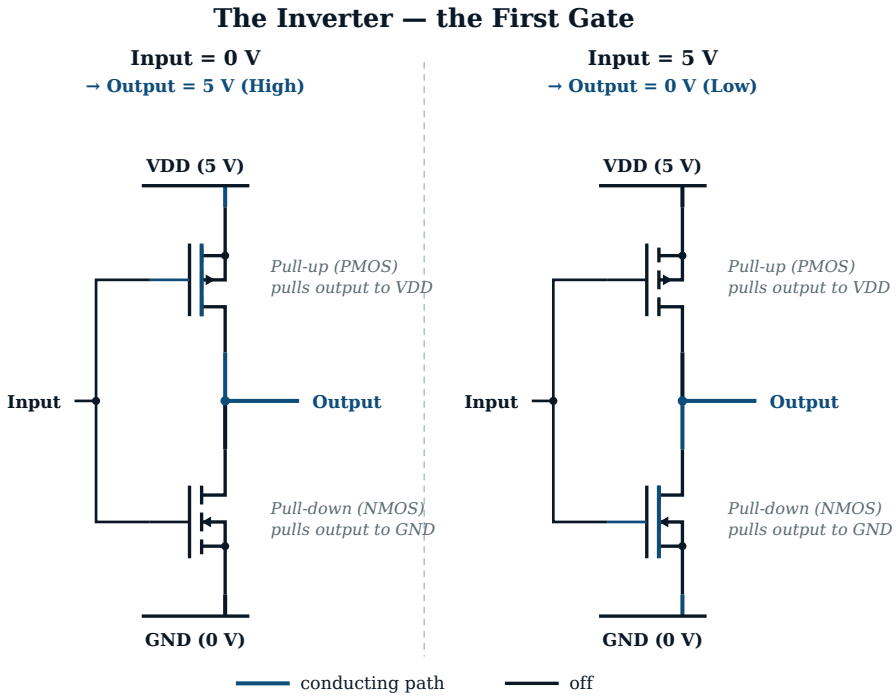
Wiring a PMOS and an NMOS together as follows creates the simplest of all logic gates: the inverter (NOT gate).

- The PMOS sits between the supply voltage (VDD) and the output.
- The NMOS sits between the output and ground (GND).
- Both gates are tied together and form the input.

Apply Low (0 V) at the input, and the NMOS blocks while the PMOS conducts — the output is connected to VDD through the PMOS and sits at High. Apply High, and the picture flips: the PMOS blocks, the NMOS conducts, the output is pulled to ground and sits at Low. The output is thus always the opposite of the input — hence the name inverter.

Notably, in neither stable state does current flow from VDD to GND, since exactly one of the two transistors is always blocking. That's why CMOS circuits draw very little power at rest.

This means the two transistors in the inverter also take on the classic roles of pull-up and pull-down: the PMOS actively pulls the output to VDD (pull-up), the NMOS actively pulls it to GND (pull-down). Since the two are never active at once, there's never a short circuit — and the output always has a clearly defined level in every state, instead of floating.



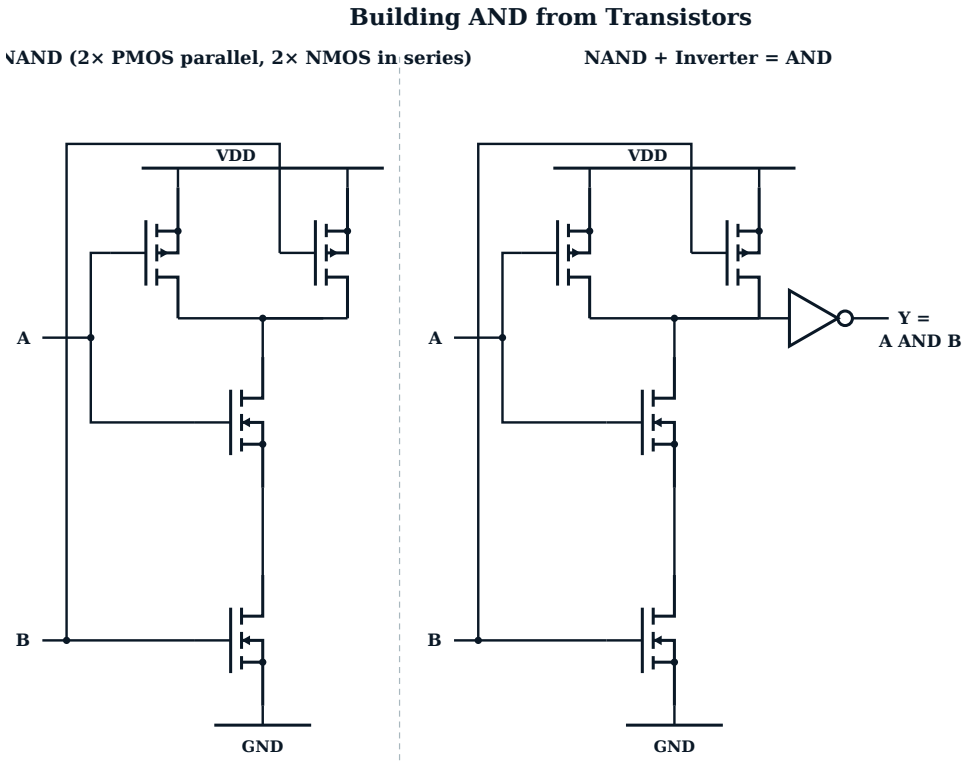
Building AND from Transistors

A handful of switches is all it takes to build an AND function from two individual switches. The principle: two NMOS transistors are wired in series between the output and ground. Only when *both* inputs are High do *both* transistors conduct, clearing the path to ground — pulling the output Low. If even one input is Low, the series path is broken.

In parallel, two PMOS transistors sit between the output and the supply voltage. If both inputs are High, both PMOS block, and neither pulls the output up. If at least one input is Low, the corresponding PMOS conducts and pulls the output High.

This circuit is, at first, a NAND gate (output Low only when both inputs are High

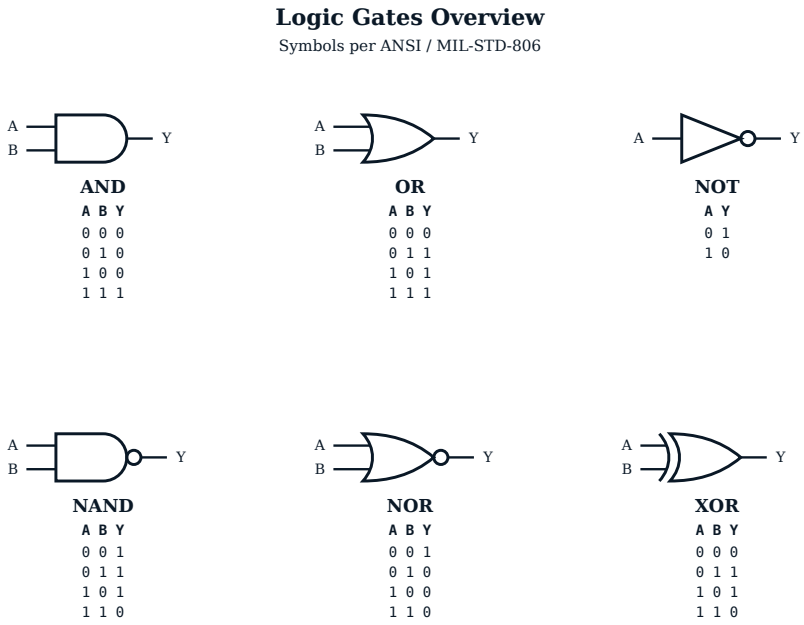
— the inverse of AND). Only a following inverter (see "The Inverter") turns it into a true AND gate. That's no coincidence: in CMOS technology, NAND and NOR gates are the "natural" building blocks, since they arise directly from series and parallel arrangements of transistors — AND and OR only appear once an extra inverter is added at the output.



Logic Gates Overview

All the common logic gates can be derived from the basic circuits covered so far. The table below shows the gate symbol and truth table for the most important gates:

Gate	Output = 1 when …
AND	both inputs are 1
OR	at least one input is 1
NOT	input is 0
NAND	not both inputs are 1
NOR	no input is 1
XOR	the inputs differ

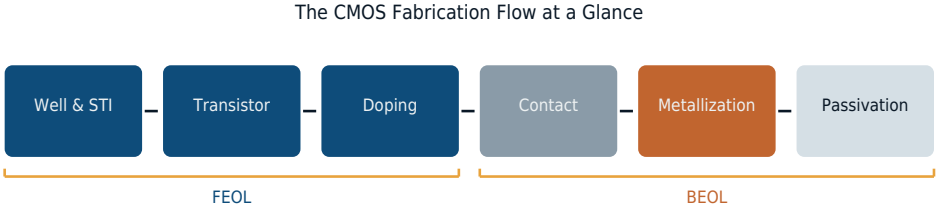


The CMOS Fabrication Flow

Front-End and Back-End of Line

A finished chip is the result of often several hundred individual process steps, which can be roughly divided into two major phases: Front-End of Line (FEOL) and Back-End of Line (BEOL). The FEOL builds the actual transistors – from the well through the gate to the doping. The BEOL then wires these transistors together into a working circuit across several metal layers.

The individual detail chapters on this site – for example on FET fabrication or FinFET technology – each cover a section of this flow in great depth. This chapter serves a different purpose: it places the individual steps into the overall process and shows how FEOL and BEOL interlock.



Wafer and Wells

Every CMOS fabrication process starts with the silicon wafer – a thin, highly pure disc of single-crystal silicon. The first step is forming the wells: regions with deliberately introduced n- or p-type doping, in which the complementary transistor types will later be built. PMOS transistors grow in an n-well, while NMOS transistors form in the surrounding p-doped substrate (or a p-well).

To keep neighboring transistors from interfering with each other, they are separated by Shallow Trench Isolation (STI): narrow, oxide-filled trenches that provide electrical isolation while also clearly defining the active areas in which the transistors will later be built.

Transistor Fabrication in the FEOL

On the prepared active areas, the actual transistor is then built: gate oxide, gate electrode, spacers, and the characteristic structure of source, drain, and channel. This flow is already described step by step in detail in the FET fabrication series and is not repeated here.

What matters for this overview is mainly the timing: every step up to and including gate patterning belongs to the FEOL. Only after that does the transition to the BEOL begin.

Doping and Silicidation

After gate patterning comes doping of source and drain – usually by ion implantation, in which dopant atoms are shot into the wafer at high energy. Spacers along the sides of the gate ensure that this implantation is offset precisely from the channel, creating a cleanly defined junction.

A thin metal layer (often nickel or cobalt) is then deposited on the exposed silicon surfaces – source, drain, and gate – and annealed. This forms a low-resistance metal silicide known as salicide (self-aligned silicide). It significantly reduces contact resistance and also marks the end of the FEOL.

Contacts

With the transistor complete, the transition into the BEOL begins. First, an insulating layer (usually an oxide) is deposited over the entire structure and planarized. Tiny contact holes are then etched into this layer, landing precisely on source, drain, and gate, and are subsequently filled with tungsten.

These contacts form the first electrical connection between the transistor and the wiring layers above – the actual BEOL.

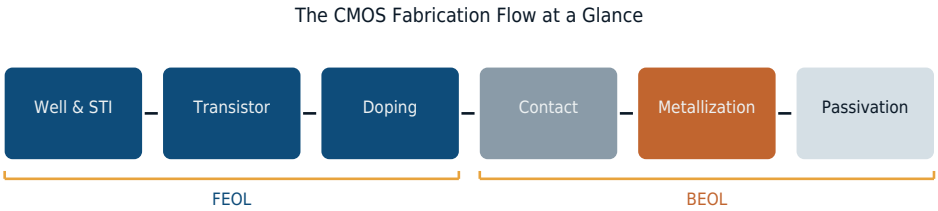
Metallization in the BEOL

From here on, a pattern repeats across multiple layers: trenches and vias are etched into an oxide layer and then filled with copper using the so-called damascene process. Excess copper is removed again by chemical-mechanical polishing (CMP), leaving a flat surface for the next layer.

Modern processes use eight to fifteen metal layers this way – from very fine lower layers for local wiring to wide upper layers for power delivery. The process concludes with a passivation layer, which protects the chip from moisture and mechanical damage.

The Flow at a Glance

The following overview summarizes the entire flow from the well to the finished wiring once more and assigns the individual steps to the two main phases:



For a deeper look at individual steps, see the corresponding detail chapters: transistor fabrication in the FET series, the FinFET variant in the FinFET fabrication chapter, and layout and verification topics in the chapters on Design Rules and Layout versus Schematic.

The SRAM Cell

Vom Schaltbild zum Chip: EDA

All the circuit diagrams in this section — from the single transistor to the SRAM cell — were created using Electronic Design Automation (EDA): the software toolchain used to design integrated circuits today. An EDA schematic like the one shown here is more than an illustration — it's the input for the next design steps: simulation verifies the electrical behaviour, place & route generates the physical layout on the wafer, and verification checks that the two match. An SRAM cell is a particularly good example of this, because it's repeated

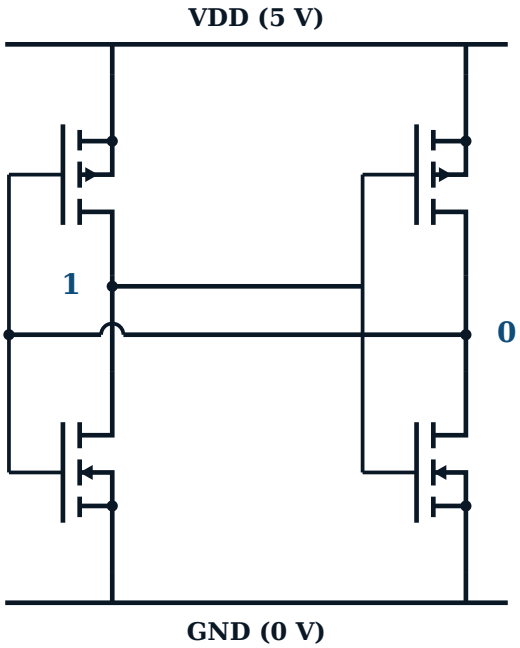
identically millions of times on a modern chip — exactly the kind of regular, highly optimised structure that makes EDA tools indispensable.

Storing a single bit sounds like a trivial task at first — but it's actually one of the most common circuit blocks of all. Every processor cache, every register file consists of millions to billions of such cells. The SRAM cell (Static Random Access Memory) doesn't store its bit as charge on a capacitor like DRAM, but as the stable switching state of two cross-coupled inverters. As long as the supply voltage is present, the stored value is retained — no refresh needed, but also without DRAM's high packing density, since a single bit takes six transistors instead of one.

Two Cross-Coupled Inverters as the Storage Core

At the heart of the cell are two CMOS inverters (familiar from chapter 3: one PMOS as pull-up, one NMOS as pull-down), whose inputs and outputs are cross-connected — the output of one inverter drives the input of the other. This feedback creates two stable states: if node 1 is High, that forces the second inverter to hold node 0 Low, which in turn confirms the first inverter's node 1 stays High. The state stabilises itself, with no clock or refresh needed — hence "static". Flipping it requires writing against this feedback from outside with sufficient drive strength.

Storage core: two cross-coupled inverters



The Access Transistors

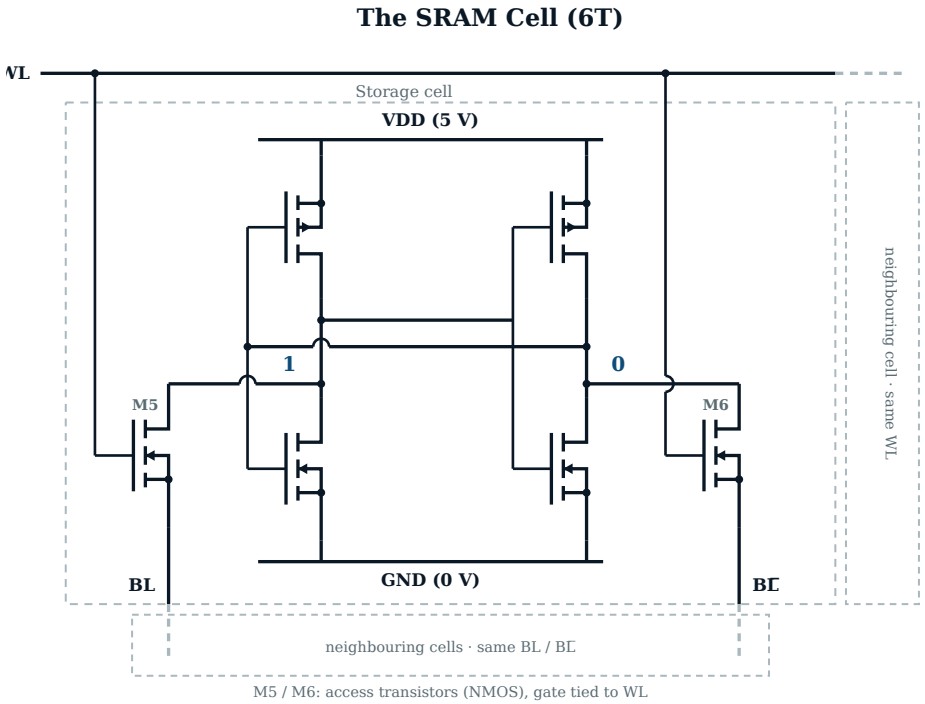
The two stored nodes 1 and 0 aren't connected directly to the cell's input/output lines, but each through its own NMOS access transistor to the bit lines BL and BL̄. Their gates are tied together to the word line (WL). When WL is Low, both access transistors are fully off — the cell is disconnected from the bit lines and holds its state in isolation. Only when WL is pulled High do both transistors turn on simultaneously, connecting node 1 to BL and node 0 to BL̄. This one shared word line is later used to address a whole row of cells in the memory array at once.

Writing

To write, BL and BL̄ are driven externally, low-impedance, to an inverted pair of levels (e.g. BL = 5 V, BL̄ = 0 V for a "1"), while WL is activated. The external drivers are deliberately made stronger than the cell's own feedback loop — they override the current state through the access transistors and force the inverters into the new, desired state. Once WL falls Low again, the two inverters hold the freshly written value on their own.

From a Single Bit to an Array

A cell's word line (WL) isn't wired individually — it runs through an entire row of neighbouring cells, and every cell in that row shares the same WL. Perpendicular to it, each bit-line pair (BL/BL̄) runs through every cell in a column. This forms a grid: only the cell at the intersection where WL is active *and* the matching bit lines are present is actually addressed during a write or read — every other cell in the same column stays isolated by its own, inactive word line, even though it hangs off the same bit lines. This row/column addressing lets a comparatively small number of lines individually address an array of millions of cells.



Reading

Before every read access, both bit lines are first precharged to a shared mid-level potential. Activating WL then pulls one of the two lines slightly toward the stored state through the cell's access transistors, while the other stays close to the precharge level — a small voltage differential builds up between BL and BL̄. A downstream sense amplifier detects even this tiny difference and amplifies it to a clean digital level. Importantly, the read itself only weakly disturbs the stored state — unlike DRAM, the cell's content is left unchanged.

The DRAM Cell

A Capacitor Instead of Six Transistors

Where the SRAM cell holds its bit as a stable switching state of two cross-coupled inverters, the DRAM cell (Dynamic Random Access Memory) takes a fundamentally different approach: it stores information as electric charge on a single capacitor. A single NMOS transistor is enough to selectively connect this capacitor to the bit line or isolate it from it – hence the name 1T1C cell (one transistor, one capacitor).

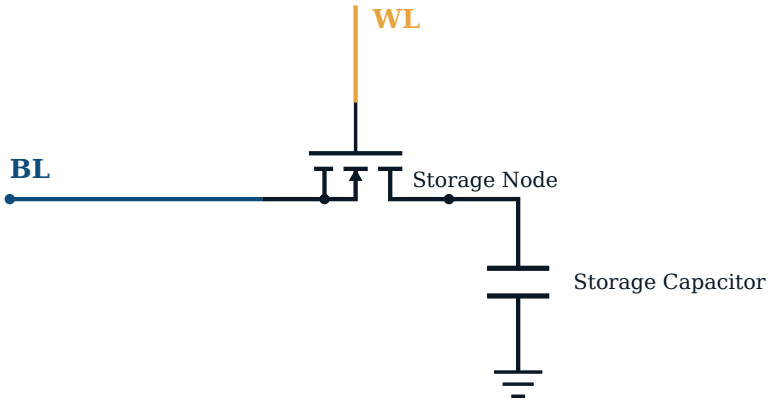
This radical reduction to two components instead of six makes DRAM significantly denser and cheaper per bit than SRAM – main memory is therefore built almost exclusively from DRAM. The price for this: the stored charge is not stable. It slowly leaks away through parasitic paths, which is why every cell must be refreshed periodically to avoid losing its content – hence “dynamic”.

Access Transistor, Word Line, and Bit Line

As with the SRAM cell, the word line (WL) and bit line (BL) handle addressing – but with a single access transistor instead of two. Its gate connects to the word line, and its source-drain path connects the storage capacitor to the bit line. When WL is low, the transistor is fully off, and the capacitor is electrically isolated – aside from leakage, the stored charge has nowhere to go. Only when WL is activated does the access transistor open and connect the capacitor to the bit line.

To write, the bit line is driven low-impedance to the desired level from outside while WL is active – the capacitor charges or discharges accordingly through the open access transistor. Once WL falls low again, the capacitor is isolated once more and holds its charge – for now.

The DRAM Cell (1T1C)



Leakage and Charge Decay

The storage capacitor of a DRAM cell is tiny – typically just a few femtofarads. Even the smallest leakage currents through the off-state access transistor,

through the capacitor dielectric, or through the pn junction to the substrate are enough to noticeably drain the stored charge within a few milliseconds. Without countermeasures, the stored value would simply vanish – a state no circuit, however clever, can tolerate.

This charge problem is exactly what fundamentally distinguishes DRAM from SRAM: while the cross-coupled inverters of an SRAM cell actively defend their state against disturbances, a DRAM cell is purely passive – it cannot maintain its own content.

The Refresh Cycle

To compensate for charge loss, the memory controller reads each cell at regular intervals and immediately writes the detected value back – a process known as refresh. Typical DRAM devices must refresh every cell roughly every 64 milliseconds, with some modern, denser cells needing it even more often. Since a single read access already activates an entire row of the array at once, refresh can be performed efficiently row by row rather than cell by cell.

This refresh costs time and energy: while a row is being refreshed, it is unavailable for regular read or write access. In very large memory devices with billions of cells, refresh overhead accounts for a noticeable share of total bandwidth and power consumption.

The Sense Amplifier

During a read, the tiny storage capacitor briefly discharges onto the bit line, which was previously precharged to a mid-level voltage – the resulting voltage change is minimal, often just a few tens of millivolts. A sense amplifier at the end of each bit line detects this tiny signal and amplifies it into a clean digital level.

This amplification serves two purposes at once: it not only delivers the read value to the outside world, but also drives the original capacitor, via the activated word line, back to its full level. Every read access is therefore automatically also a refresh of the currently active row – unlike SRAM, where reading barely affects the stored state, reading a DRAM cell is destructive and makes this “rewrite” mandatory.

DRAM versus SRAM

The fundamentally different storage principles lead to very different characteristics, each suited to different applications:

	SRAM	DRAM

	SRAM	DRAM
Transistors per bit	6	1 (+ 1 capacitor)
Cell size	large	very small
Refresh required	No	Yes (~every 64 ms)
Access speed	very fast	slower
Cost per bit	high	low
Typical use	cache, registers	main memory (RAM)

In practice, the two complement each other: fast but expensive SRAM cells in the cache levels close to the processor core, and dense, inexpensive DRAM cells for high-volume main memory.

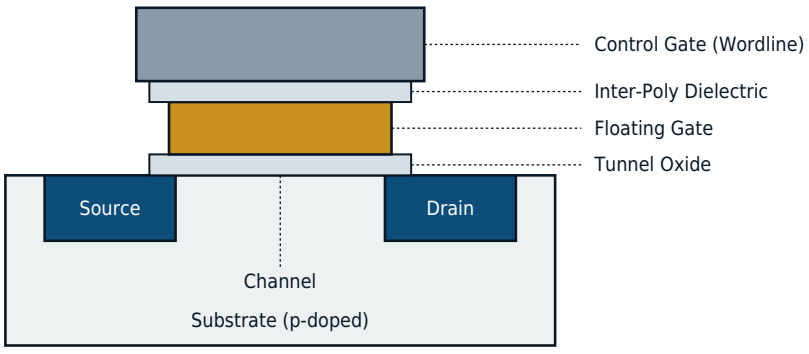
The Flash Memory Cell

Structure of the Floating-Gate Cell

A flash memory cell is essentially a MOSFET with an additional, electrically isolated gate electrode – the floating gate. It sits between the regular control gate and the channel, fully enclosed by silicon dioxide: separated from the channel below by the thin tunnel oxide (a few nanometers), and isolated from the control gate above by a thicker interpoly oxide.

Because the floating gate has no external electrical contact, charge once placed on it remains stored for years – this is the basic principle of non-volatile storage.

The Floating-Gate Cell
Structure in cross-section



The floating gate is completely surrounded by oxide and therefore electrically isolated. The stored charge remains in place for years.

Programming and Erasing

To program a cell (placing charge on the floating gate), either hot-electron injection or Fowler-Nordheim tunneling is used, depending on the cell type: with hot-electron injection, a high drain-source field accelerates electrons strongly enough that some gain sufficient energy to overcome the oxide barrier and become trapped on the floating gate.

Fowler-Nordheim tunneling instead uses a very high electric field across the tunnel oxide, allowing electrons to tunnel through the barrier quantum-mechanically – this mechanism is typically used for erasing, where charge is removed from the floating gate again. The stored charge measurably shifts the cell's threshold voltage, allowing "0" and "1" to be distinguished during readout.

NOR vs. NAND Architecture

The way individual cells are wired into an array distinguishes two basic architectures. In NOR flash, each cell is individually connected in parallel to the bitline and wordline, similar to a NOR gate – this enables fast, random access to individual bytes, but costs area.

In NAND flash, several cells (typically 32–128) are connected in series between the bitline and source line, as in a NAND gate; this saves considerable area and enables today's high storage densities, but only allows page-wise rather than random access. This is why NAND flash dominates in USB drives, SSDs, and memory cards, while NOR flash is used where fast direct access to individual addresses is required, such as in firmware storage.

Multi-Level Cell (MLC/TLC)

Instead of distinguishing only between two charge states (0/1), the amount of charge on the floating gate can be graded more finely. An SLC cell (single-level cell) stores 1 bit, an MLC cell stores 2 bits across four distinguishable charge levels, and a TLC cell stores 3 bits across eight levels.

This multiplies the storage density per cell but reduces the margin between levels – the cells become more sensitive to charge loss and require more elaborate error correction, and the number of possible write/erase cycles also decreases.

Comparison: DRAM, SRAM, and Flash

The three memory types differ fundamentally in volatility and intended use. DRAM stores charge in a capacitor and must be refreshed continuously, losing

its content immediately on power loss – but it is fast and serves as main memory.

SRAM stores its state in a bistable circuit of six transistors, is even faster but more area-intensive, and is used for example as cache memory. Flash, by contrast, retains its content without power but is significantly slower to write than DRAM and SRAM, making it suited for persistent data storage rather than main memory.

Design Rules

Introduction & Motivation

Why Design Rules?

Every integrated circuit exists as a geometric layout: a stack of overlapping layers of diffusion, polysilicon, contacts, and multiple metal layers that together form transistors and their interconnects. For this layout to be manufactured correctly and reliably on the wafer, it must obey a set of geometric constraints known as design rules. They translate the physical limits of a fabrication process into concrete requirements for chip design.

These limits arise from several sources at once. Lithography can only image structures cleanly down to a certain resolution; below that, edges blur or lines break. Masks cannot be aligned to each other with perfect precision during exposure, so a safety margin for misalignment must be built in between successive layers. Etch and deposition processes vary from wafer to wafer and across the wafer surface, so minimum spacings must also guard against electrical shorts caused by process variation. On top of this come effects such as latchup in CMOS well structures or charge damage from the so-called antenna effect, which require additional, electrically motivated rules.

Design rules are therefore always a trade-off: generous spacings improve manufacturing yield and reliability but cost chip area and thus money. Tight, aggressive rules allow higher packing density but reduce yield and increase the risk of fabrication defects. Each foundry publishes a complete rule set for its specific process, which is checked automatically against the finished layout by a Design Rule Check (DRC) tool before a mask is allowed to be exposed.

Basic Rule Types

Basic Rule Types

Although individual rule sets differ substantially between foundries and technology nodes, nearly all design rules can be traced back to a small number of geometric primitives. These are usually defined either on a single layer (one-layer rule) or between two layers (two-layer rule):

Width specifies the minimum width a structure on a given layer may have — for example a polysilicon gate line or a metal trace. Below this width, lithographic and etch fidelity is no longer guaranteed; the structure could break or become electrically too resistive.

Spacing describes the minimum gap between two structures on the same layer. It prevents neighboring traces from merging or shorting due to process variation, while also accounting for the optical resolution limit of the exposure.

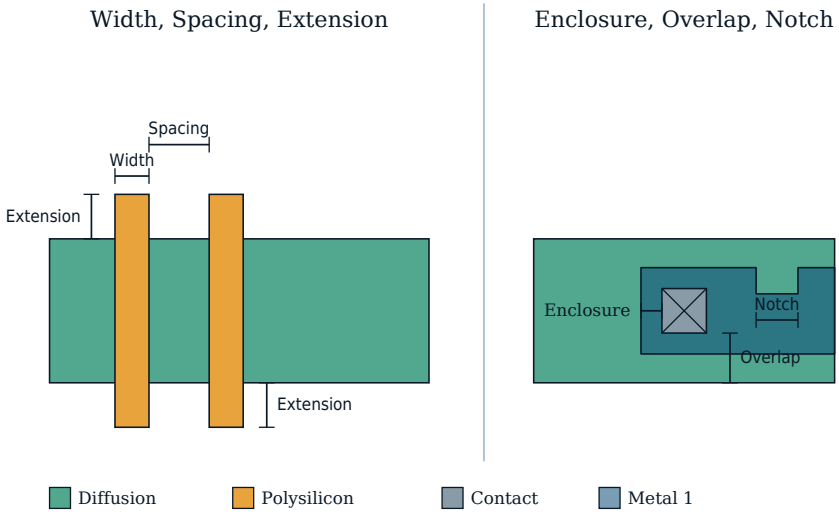
Enclosure requires one layer to surround another by a minimum margin — classically at a contact: the metal must extend past the underlying contact on all sides by a fixed amount, so that misalignment cannot leave an exposed, uncontacted edge.

Overlap is closely related to enclosure but is expressed from the lower layer's perspective: diffusion must overlap the contact so that it lands entirely on conductive material.

Extension requires a structure to project beyond the edge of another — for instance the polysilicon gate, which must extend past the diffusion edge on both sides so that source and drain remain reliably separated during ion implantation, even with slight misalignment.

Notch limits the minimum size of a re-entrant corner or narrow indentation in a structure. Very sharp or narrow notches cannot be imaged crisply by lithography and round off in the real process.

The diagram below illustrates these six basic concepts on a single, simplified layout excerpt: a polysilicon gate over a diffusion area with one contact and the overlying first metal layer.



Lambda-based vs. Absolute Rules

Scalable Rules: the Lambda Concept

In the early 1980s, as integrated circuits first began to be designed on a large scale outside individual semiconductor manufacturers, the Mead-Conway approach introduced an influential concept: instead of specifying every design rule in absolute units such as micrometers, all spacings and widths were expressed as multiples of a single reference quantity called lambda (λ). A typical rule might read, for example, "minimum poly width = 2λ" or "minimum diffusion-to-contact spacing = 3λ".

The advantage of this approach lay in portability: a layout designed entirely in lambda units could, in principle, be ported to a different fabrication process with finer or coarser resolution simply by redefining the numeric value of λ — without adjusting any geometry in the layout individually. For teaching and academic designs, where circuits were often fabricated at multiple foundries (for instance through multi-project-wafer programs), this offered a substantial practical benefit.

Why Modern Processes Use Absolute Rules

As technology nodes continued to scale, however, the lambda model ran into its limits. Not all structures in a modern process scale uniformly: transistor gates typically shrink faster than contact spacings, and metal layers at different levels often have entirely different minimum widths and spacings depending on their

role in the interconnect hierarchy. A single scaling factor can no longer capture these asymmetries.

On top of this come physical effects that do not scale linearly with feature size — such as antenna ratios, latchup spacings, or density requirements for chemical-mechanical planarization (CMP). For these reasons, modern foundries publish their design rules almost universally in absolute units (micrometers or nanometers), each with its own rule set often comprising several hundred individual rules per process variant. The lambda concept remains valuable for teaching the underlying principle of design rules, but plays essentially no role in commercial chip development today.

Example Rule Set

A Typical, Simplified Rule Set

To put the concepts introduced in Section 2 into concrete numbers, the table below shows a simplified, illustrative rule set typical of a generic single-poly, multi-metal sub-micron CMOS process. The values are modeled on common teaching design kits and are not the rules of any specific real foundry process — real rule sets comprise several hundred individual rules depending on the technology node and naturally differ from manufacturer to manufacturer.

Layer / Rule	Meaning	Minimum Value
Poly Width	min. width of a poly gate line	0.35 µm
Poly Spacing	min. spacing between two poly lines	0.35 µm
Poly Extension	gate overhang past diffusion edge	0.25 µm
Diffusion Width	min. width of a diffusion area	0.4 µm
Diffusion Spacing	min. spacing between two same-type diffusion areas	0.5 µm
Contact Size	fixed contact hole size	0.4 × 0.4 µm
Contact Spacing	min. spacing between two contact holes	0.4 µm
Diffusion Overlap of Contact	diffusion margin around contact	0.15 µm
Metal-1 Width	min. width of metal-1	0.5 µm
Metal-1 Spacing	min. spacing between two metal-1 traces	0.5 µm
Metal-1 Enclosure of Contact	metal-1 margin around contact	0.15 µm
Via-1 Size	fixed via size (metal-1 to metal-2)	0.4 × 0.4 µm
Metal-2 Width / Spacing	min. width / spacing of metal-2	0.5 / 0.5 µm

Even in this highly simplified excerpt, a typical pattern emerges: contacts and vias are usually square and fixed in size (not freely scalable), while enclosing

layers such as diffusion and metal must maintain an additional safety margin. The enclosure values are deliberately smaller than the width and spacing values of their respective layer — they primarily compensate for mask-to-mask misalignment rather than the resolution limit of the lithography itself.

In practice, such a rule set is never looked up by hand; it is encoded into a machine-readable technology file (e.g. in LEF/DEF or a vendor-specific format), against which automated tools such as layout editors and the Design Rule Checker (see Section 8) continuously verify the layout.

Latchup and Well Rules

The Latchup Mechanism

In a CMOS process, NMOS and PMOS transistors are not electrically isolated from one another but share the same substrate: PMOS transistors sit inside an N-well, while NMOS transistors sit directly in the P-substrate. This arrangement inevitably creates two parasitic bipolar transistors: a vertical PNP transistor (emitter = PMOS P^+ source/drain, base = N-well, collector = P-substrate) and a lateral NPN transistor (emitter = NMOS N^+ source/drain, base = P-substrate, collector = N-well).

The collector of each parasitic transistor doubles as the base of the other — together with the lateral resistances of the well (R_{well}) and the substrate (R_{sub}), the PNP and NPN form a feedback four-layer structure electrically identical to a thyristor (SCR). If a voltage spike, a substrate current, or crosstalk momentarily turns on one of the two base-emitter junctions, this feedback can reinforce itself: the resulting low-resistance path between supply and ground — latchup — persists until the supply voltage is removed, and can destroy the chip thermally.

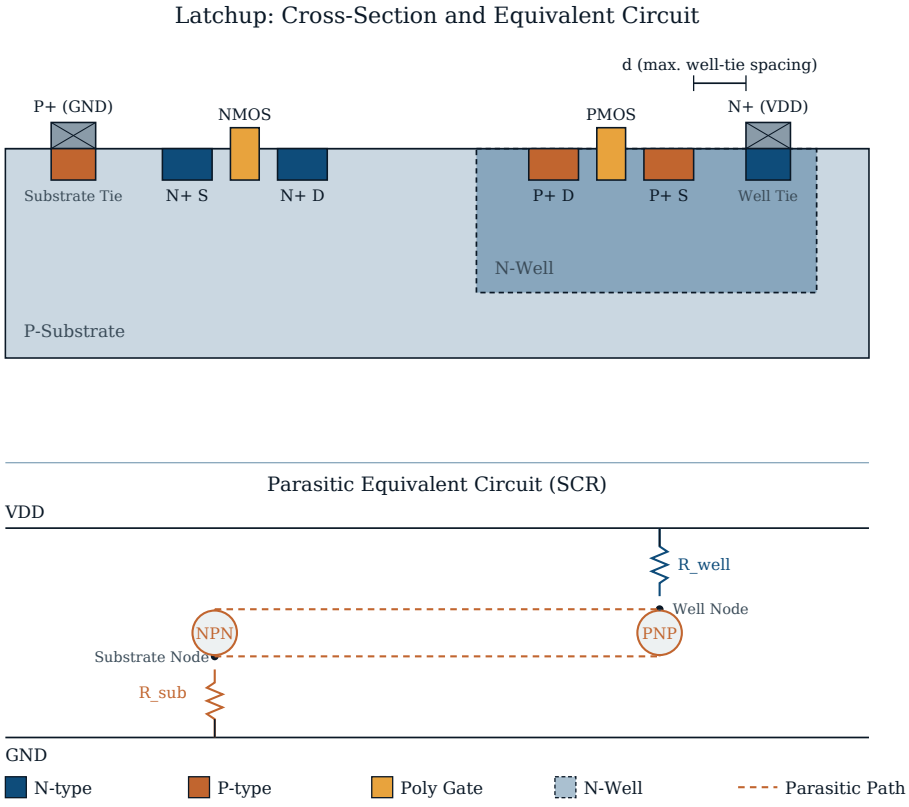
Well and Substrate Contacts as a Countermeasure

The current gain of the parasitic transistors can be effectively limited through layout: the lower the lateral resistances R_{well} and R_{sub}, the smaller the voltage drop needed to turn on the respective other transistor. Design rules therefore specify a maximum distance between every source/drain diffusion and the nearest well or substrate contact — usually set considerably tighter than pure fabrication tolerances alone would require.

In practice, these contacts are often implemented as a continuous guard ring: an N^+ ring around especially latchup-sensitive PMOS structures (tied to VDD), or a P^+ ring around NMOS structures (tied to GND), which diverts the base current along the shortest possible path before it can turn on the parasitic transistor. Such rings are particularly important around I/O drivers and other circuit blocks

that carry high or fast-switching currents.

The diagram below shows the cross-section of a simple inverter with substrate and well contacts, along with the overlaid parasitic thyristor path.



Antenna Rules

The Plasma Charging Effect

Many etch and deposition steps in semiconductor fabrication — particularly reactive ion etching (RIE) of polysilicon and metal — operate using a plasma. During these steps, conductive structures on the wafer surface that are not yet electrically connected to anything else charge up, because the electron and ion currents arriving from the plasma do not hit all surfaces symmetrically. A long polysilicon or metal line that is still incompletely wired acts like an antenna: it collects charge from the plasma during etching, which — as long as the line terminates only at a transistor gate and is not yet connected to source/drain or a stabilizing lower layer — must discharge across the thin gate oxide.

At only a few nanometers thick, the gate oxide is the most sensitive structure in the entire device. If the accumulated charge is enough to build up a sufficiently high electric field across the oxide, a Fowler-Nordheim tunneling current flows, damaging the oxide or, in extreme cases, causing breakdown. Such damage is

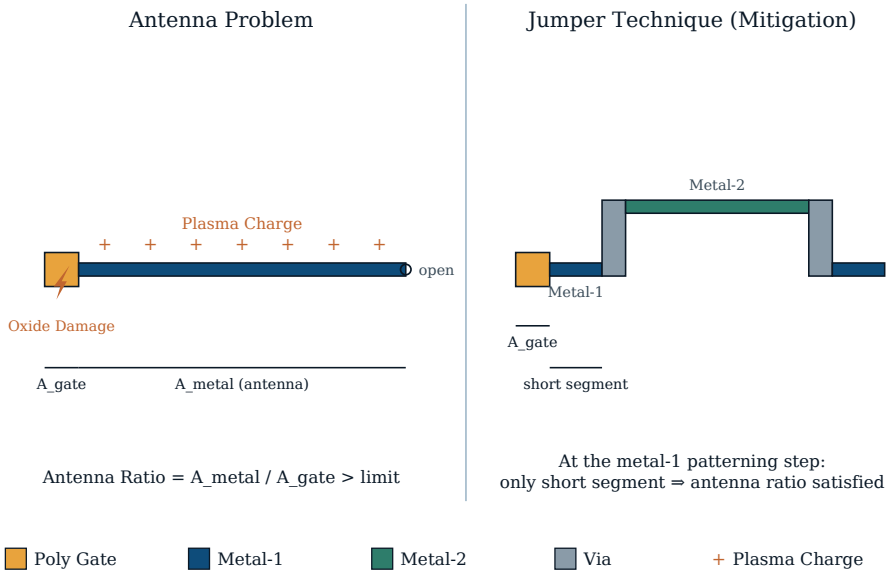
often not immediately apparent as a failure but shows up later as premature device aging or elevated leakage currents, which makes debugging considerably harder.

Antenna Ratio as a Design Rule

To limit this risk, foundries define a maximum allowed antenna ratio for every conductive layer: the ratio of the area of all conductor segments electrically connected to a gate (poly, metal-1, metal-2, and so on, each counted up to the fabrication step at which it is patterned) to the area of the connected gate oxide itself. If a net connection exceeds this limit, the DRC reports an antenna violation.

Two common remedies exist in practice. The antenna diode adds an extra, reverse-biased diode to the substrate on the at-risk line, which bleeds off accumulated charge in a controlled way without affecting circuit function. The jumper technique instead breaks the long line on its original layer and routes it onward through a higher metal level: since that higher level is only patterned in a later, temporally separate fabrication step, the area connected to the gate at any given point in time is much smaller — keeping the antenna ratio within the allowed limit at every individual fabrication step.

The diagram below shows the antenna problem on the left, on a long line connected directly to the gate, and the mitigation via a metal jumper on the right.



Density Rules (CMP)

Why Pattern Density Is a Design Rule

Modern processes with multiple metal layers need a surface that is as flat as possible between layers — only then can the next layer be exposed sharply, without focus deviations caused by height differences degrading the lithography. This planarization is handled by chemical-mechanical planarization (CMP), in which a rotating polishing pad wetted with slurry mechanically and chemically removes material from the wafer's oxide or metal surface.

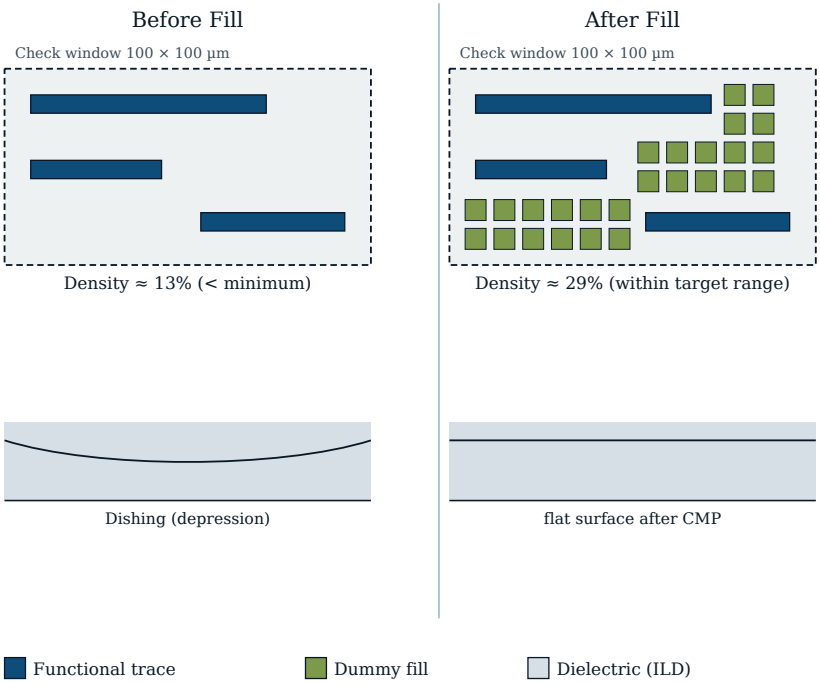
CMP, however, does not polish at the same rate everywhere: the removal rate depends on the local pattern density of the underlying layer. In large, mostly empty areas — for example above a wide, sparsely patterned trace — the pad polishes in deeper than intended, an effect known as dishing. Conversely, in densely packed areas with many narrow, closely spaced structures, erosion occurs, in which the surrounding dielectric is also removed more than intended. Both effects create height variations across the wafer surface that propagate into subsequent fabrication steps, where they can cause focus and yield problems.

Fill Structures as a Countermeasure

To prevent this, design rules specify a minimum and maximum local pattern density for each layer, typically measured over a sliding check window of fixed size (e.g. 100 µm × 100 µm) that is swept across the entire layout. If the density in a window falls below the minimum, automated fill generators insert additional, electrically non-functional dummy structures into the gaps — small, evenly distributed metal or polysilicon shapes that are not connected to the circuit's nets (floating fill) and exist solely to satisfy the density requirement.

These fill structures are themselves subject to design rules — for instance a minimum spacing to functional traces, to limit parasitic capacitance or unwanted coupling. Fill generation today is almost always fully automated and runs as the last step before tape-out, once the actual circuit layout has already been finalized.

The diagram below shows a check window before and after fill insertion, along with a schematic view of the resulting surface topography after the CMP step.



Design Rule Check (DRC)

Automated Rule Checking

A complete rule set is of little use unless it is consistently checked against every layout — for modern chips with billions of individual structures, manual inspection is out of the question from the start. This task is handled by the Design Rule Check (DRC): a software tool that checks the layout against the rules stored in a technology file (the so-called DRC deck) and reports every violation found as a flagged error.

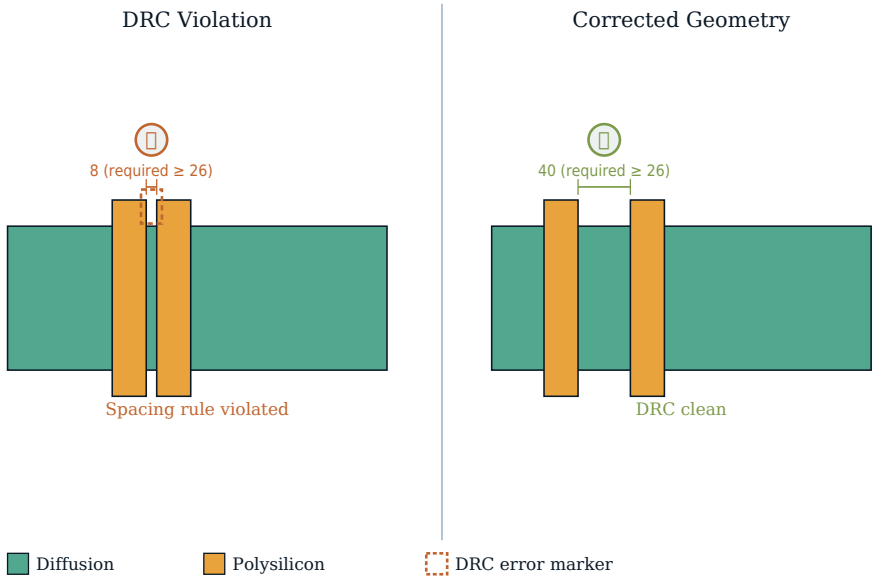
At its core, a DRC tool works with Boolean and geometric operations on the individual layers: it computes intersections, unions, and spacing measurements between polygons on the same or different layers, and compares the results against the limits stored in the rule set. A width rule, for instance, is checked by measuring every structure on the relevant layer at its narrowest point; a spacing rule by determining the minimum distance between neighboring polygons.

Role in the Design Flow

In practice, DRC does not run just once at the end but repeatedly throughout the entire layout process: many layout editors offer real-time checking that flags violations with color highlighting as they are drawn, supplemented by full batch runs over the entire design ahead of major milestones. Every violation found

must be resolved before the next fabrication step — a layout is considered "DRC clean" only once a run over the complete design reports no remaining open violations. Only after that do further checks typically follow, such as layout-versus-schematic (LVS) verification, which confirms that the drawn layout actually implements the same circuit as the schematic.

The diagram below shows a typical spacing violation — two poly lines with insufficient spacing — compared to the corrected, rule-compliant geometry.



Layout versus Schematic (LVS)

Introduction & Motivation

Why DRC Alone Is Not Enough

A layout that fully passes the Design Rule Check is geometrically sound: every width, spacing, enclosure, and density falls within the allowed limits. That, however, says nothing about whether the drawn polygons actually implement the circuit designed in the schematic. A missing contact, two swapped terminals, or an accidentally duplicated transistor violate no design rule at all — the layout stays "DRC clean" even though the circuit no longer matches the design electrically.

Such errors happen easily: during manual routing of a complex block, while copying and adapting existing layout cells, or simply through a forgotten connection step. The larger the chip, the more unlikely it becomes to spot a

single missing contact by eye among millions of structures.

The Idea Behind LVS

The Layout-versus-Schematic comparison (LVS) solves this problem by checking not the geometry itself but the circuit topology that results from it. A netlist is automatically extracted from the physical layout — a list of all devices (transistors, resistors, capacitors) together with their terminals and the nets connecting them. This extracted netlist is then compared against a reference netlist derived directly from the schematic.

If the two netlists match — the same devices, the same connections, the same device parameters — the layout is considered "LVS clean." Any deviation is reported as a mismatch and must be resolved before the next step in the design flow. LVS is therefore the necessary complement to DRC: DRC ensures the layout is manufacturable; LVS ensures that what gets manufactured is actually the intended circuit.

Netlist Extraction from the Layout

From Polygons to Devices

The first step of LVS extraction identifies devices purely from layer geometry — much like a DRC tool, just with a different goal. Wherever a polysilicon structure overlaps a diffusion area, the extractor recognizes a transistor: the width of the polysilicon at the overlap yields the gate width W, and the extent of the diffusion perpendicular to it yields the gate length L. Whether the device is NMOS or PMOS follows from the diffusion type (N^+ or P^+), or more precisely, from whether the diffusion sits inside an N-well.

Other devices are recognized analogously: a metal-insulator-metal structure yields a capacitor; a narrow, high-resistance polysilicon or diffusion strip with no gate function yields a resistor. Every device found is added to the netlist together with its extracted geometric parameters.

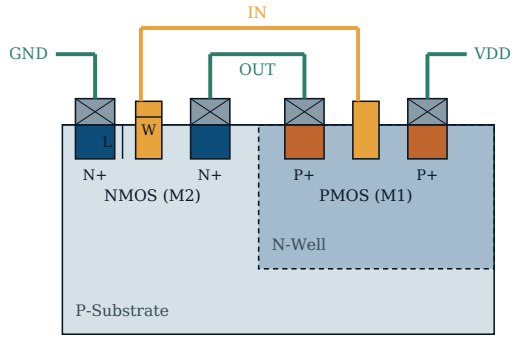
From Contacts to Nets

In parallel, the extractor determines electrical connectivity: all polygons on the same layer that directly touch or overlap, as well as all layers connected to each other through contacts or vias, are merged into a common net — regardless of how many metal layers and vias that net spans. Technically, this comes down at its core to a union-find operation over all conductive polygons: two polygons belong to the same net exactly when a conductive path exists between them.

The result is a complete netlist: a list of all devices with their terminals, where each terminal is assigned to one of the previously determined nets. At this stage, the netlist still has no names like "VDD" or "OUT" — it references nets only through internal identifiers. Only the comparison against the schematic (Section 3) assigns them meaningful names again.

The diagram below shows a simple inverter in layout and the simplified netlist extracted from it.

Netlist Extraction: Inverter Layout




```
* extracted netlist

M1 OUT IN VDD VDD
   PMOS W=0.60u L=0.35u

M2 OUT IN GND GND
   NMOS W=0.40u L=0.35u
```

SPICE-like notation

The Reference Netlist from the Schematic

From Schematic to Netlist

While the layout netlist is derived from geometry, the reference netlist is obtained directly: a schematic capture tool already knows, for every placed symbol — transistor, resistor, capacitor — its terminals (pins) and their meaning. When the designer connects two pins with a drawn wire or with identical net labels, they belong to the same net. This extraction is trivial compared to the layout side, since no geometry needs to be interpreted — the connectivity is already explicit.

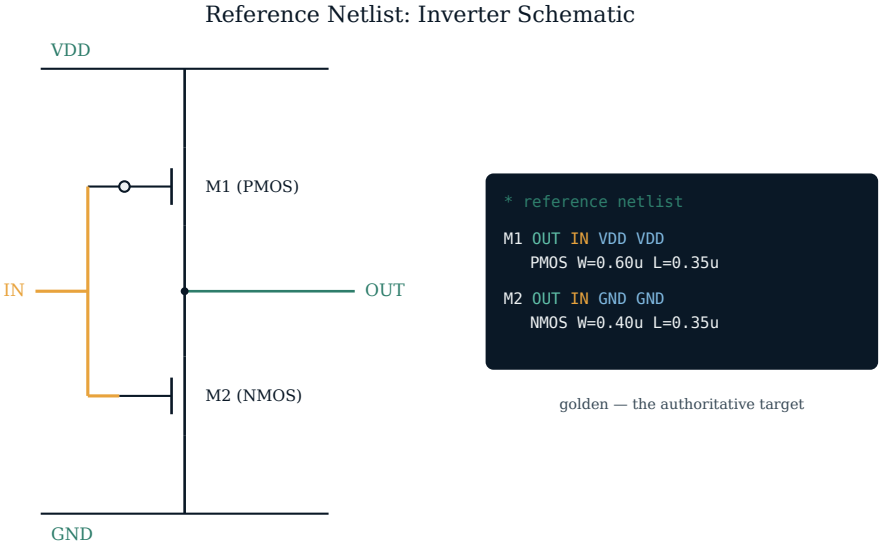
Net names such as VDD, GND, IN, or OUT are usually assigned by the designer themselves, either by labeling individual wires or through special global net symbols for supply rails that are treated as "the same everywhere" across the tool — a VDD symbol placed at several points in the schematic connects all of those points into a single net, even without a continuous wire between them.

Hierarchy and Flattening

Real schematics are almost always built hierarchically: an inverter is a cell reused inside a gate, which is itself part of a multiplexer instantiated inside a register. The netlist compiler must resolve this hierarchy — either fully flat (every instance of a cell is expanded into its own independent devices) or hierarchically, where identically structured instances are checked only once against a reference cell, after which merely their interconnection with one another is compared.

Hierarchical LVS is the only practical approach for memory blocks with thousands of repeated, identical cells (e.g. SRAM arrays) — a fully flattened comparison would recompute the entire cell structure for every single instance, even though only its position in the layout differs.

The reference netlist obtained this way is considered "golden" — the authoritative target state against which the netlist extracted from the layout is compared in the next step. The diagram below shows the schematic of the same inverter from Section 2 and the reference netlist derived from it.



The Comparison Algorithm

The Netlist as a Graph

To compare two netlists algorithmically, each is first converted into a graph: devices and nets become nodes, and every terminal connection of a device to a net becomes an edge between them, labeled with the terminal role (drain, gate, source, bulk). The result is a so-called bipartite graph — device nodes connect

only to net nodes, never directly to one another.

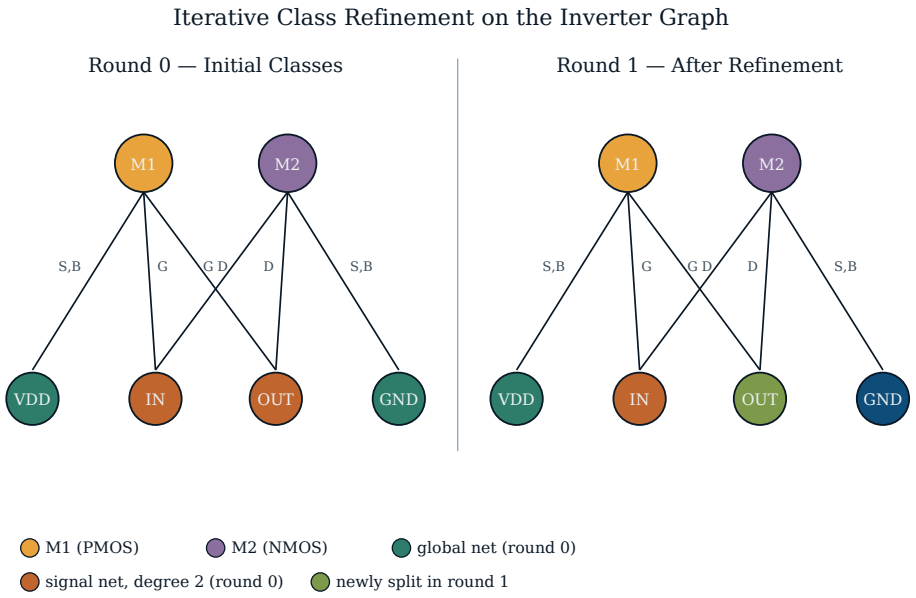
Two netlists are electrically identical exactly when an isomorphism exists between their two graphs: a unique mapping that sends every node of one graph to exactly one node of the other such that edges, edge labels, and node attributes (device type, W/L) are preserved. Graph isomorphism is computationally expensive in the general case — for circuits with millions of transistors, a naive brute-force comparison of all possible mappings would be hopeless.

Iterative Refinement Instead of Brute Force

LVS tools sidestep this problem through iterative class refinement. Every node first receives a coarse initial class based on easily determined features — device type and parameters for device nodes, number of connected terminals for net nodes. The algorithm then repeatedly refines: each node is assigned a new, finer class derived from the multiset of its neighbors' classes (together with the edge labels). Two nodes initially placed in the same class split apart as soon as their neighborhoods differ.

This process repeats until no class can be split any further. In well-designed circuits with little symmetry, what typically remains are almost exclusively singleton classes — each node is then uniquely matched to a node in the other netlist. Only for the few remaining, genuinely symmetric groups (for example, parallel identical dummy fingers of a transistor) does the algorithm still need to try out all remaining assignment possibilities within that small group — an effort that stays practical even with millions of devices, because these residual groups are, in practice, tiny.

The diagram below shows this refinement step using the inverter from the previous sections: initial classes on the left, the result after one refinement round on the right.



Typical Mismatches

Four Recurring Failure Patterns

Although LVS tools produce very different error messages in detail, the vast majority of mismatches trace back to a handful of recurring causes. Knowing these patterns usually finds the root cause in the layout far faster than working blindly through the error list.

An open net results from a missing or misplaced contact: two structures that should be connected according to the schematic remain electrically separate in the layout. The netlist then shows two nets instead of one — the LVS typically reports a "net not found" or a net-count mismatch, because a net expected from the schematic falls apart into two pieces in the layout.

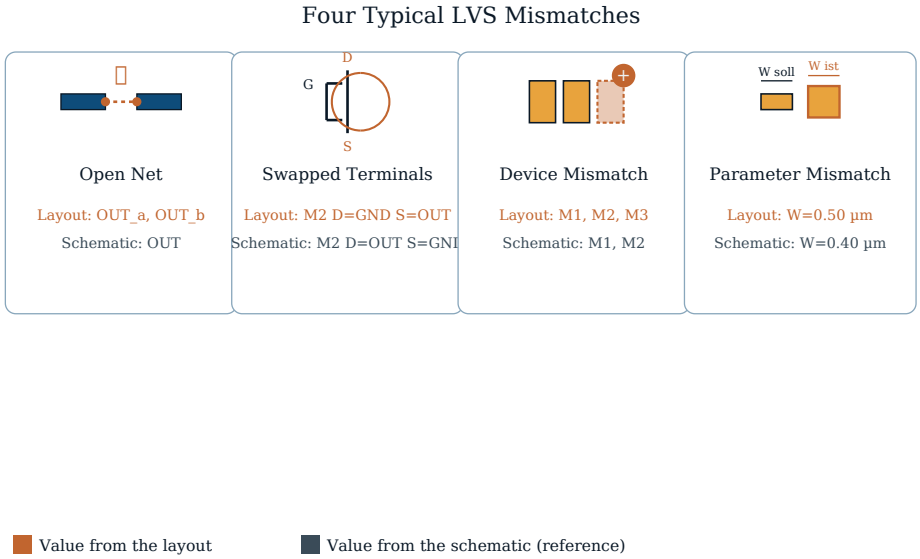
Swapped terminals happen easily when copying or manually wiring a transistor: source and drain are connected to the wrong nets in the layout. Since source and drain are geometrically symmetric — and therefore interchangeable — for most MOSFETs, the layout stays DRC clean, but the circuit then behaves electrically differently than intended, or not at all.

A missing or extra device usually results from a copy error — a cell was accidentally placed twice, or removing a test structure also deleted a structure that was actually needed. The two netlists then already differ in the sheer device count, often the simplest of all mismatches to find.

A parameter mismatch usually concerns transistor width W or length L: if the layout was drawn with a different gate width than the schematic specifies,

connectivity remains fully correct — topology and device type match, only the numeric value deviates. Many LVS tools do not flag this as a hard error but as a warning with an adjustable tolerance, since small deviations from rounding to the manufacturing grid are often unavoidable.

The diagram below shows all four failure patterns side by side, layout versus schematic.



LVS in the Design Flow

Position Between DRC and Parasitic Extraction

In the design flow, LVS fundamentally runs after the Design Rule Check: a layout must first be DRC clean before an LVS run can produce meaningful results — on a geometrically flawed layout (for instance with structures overlapping that should actually be separate), no reliable netlist could be extracted in the first place. Together, both checks form what is referred to in practice as "signoff": a layout that is neither DRC clean nor LVS clean does not proceed to fabrication.

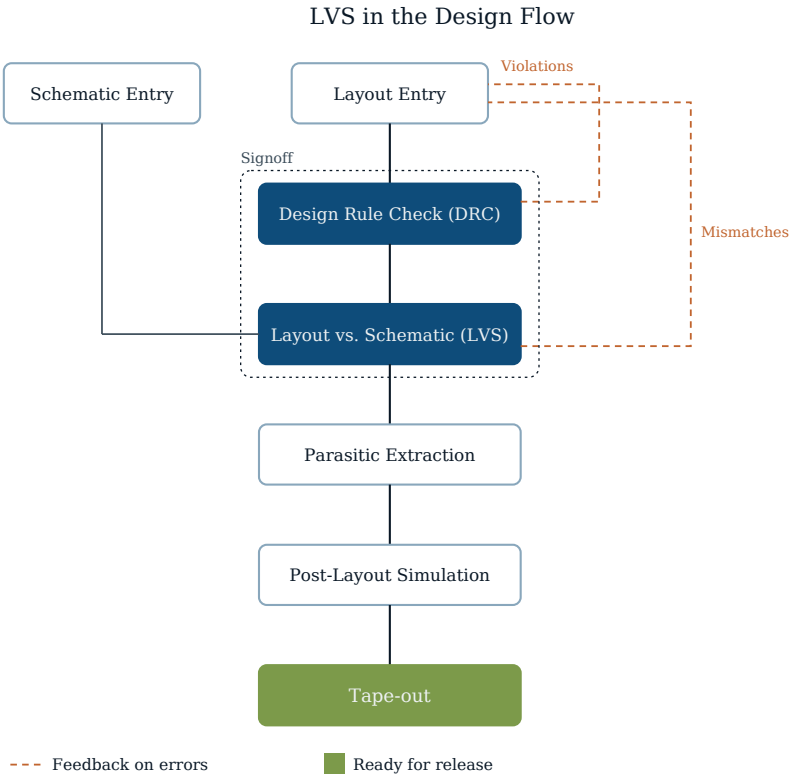
As with DRC, an LVS run is not a one-time event at the end but part of an iterative cycle: every reported deviation — open net, swapped terminal, missing device — must be fixed in the layout, after which extraction and comparison run again. For large blocks, incremental LVS runs that, after a small layout change, only re-compare the affected neighborhood considerably speed up this cycle compared to a full re-comparison of the entire block.

What Follows Next

Once a block is both DRC clean and LVS clean, parasitic extraction usually follows: from the now-verified layout geometry, the parasitic resistances and capacitances of the wiring are additionally computed and added to the netlist — a task LVS itself does not perform, since it checks only connectivity and device parameters, not the electrical fine details of the wiring. The resulting parasitics-annotated netlist then feeds post-layout simulation, which shows whether the circuit still behaves as intended once the real, layout-induced delays and voltage drops are taken into account.

Only once this final step also confirms the specification is a block ready for tape-out — the release of the mask data to the foundry.

The diagram below shows this flow at a glance, from schematic and layout through to tape-out.



Circuit Layouts

From Schematic to Layout

The schematic from the previous chapter shows *how* the six transistors are

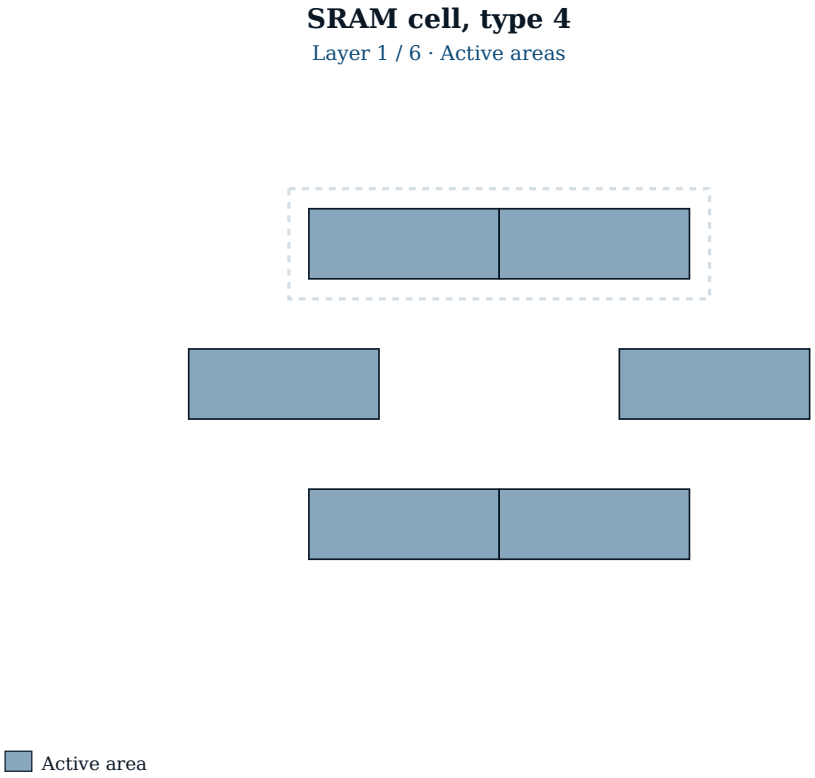
electrically connected. But it says nothing about *where* in the chip these transistors actually sit, or how the connections between them are manufactured. That's exactly what the layout shows: the sequence of masks used to expose and pattern each fabrication layer — diffusion, polysilicon, contacts, and no fewer than three metal layers.

This cell follows a common style often called "type 4": the two storage nodes Data and Data_b run as continuous columns from the top (PMOS) through the middle (access transistors) to the bottom (NMOS), spanning the full cell height. VDD sits centrally between the two columns, BL and BL̄ on the outside, VSS at the top and bottom, and the word line (WL) runs across the middle.

Symbol	Meaning
VDD	supply voltage (central, vertical)
VSS	ground (top and bottom, horizontal)
BL / BL̄	bit line / inverted bit line (outside, vertical)
WL	word line (middle, horizontal)
Data / Data_b	the two cross-coupled storage nodes
PU	pull-up transistor (PMOS, pulls toward VDD)
PD	pull-down transistor (NMOS, pulls toward VSS)

Layer 1: Active Areas

Six rectangles in three rows: the two PMOS (pull-up, in the N-well) on top, the two access transistors in the middle, and the two NMOS (pull-down) at the bottom. Each column — "Data" on the left, "Data_b" on the right — runs as its own active area through all three rows. That's the core idea of the type-4 layout: an inverter isn't made of two transistors side by side, but of a PMOS on top and an NMOS at the bottom *in the same column*, with the access transistor in between.



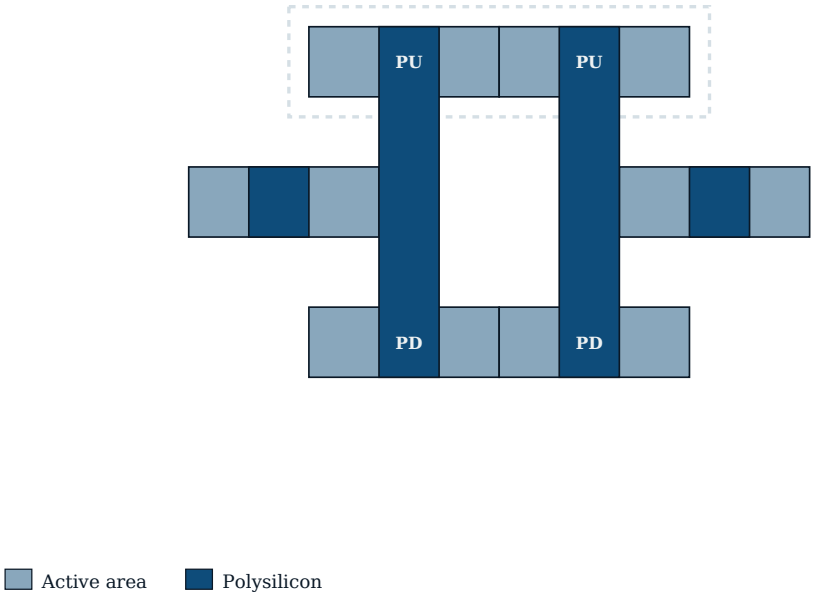
Layer 2: Polysilicon

Two continuous vertical gates are added — gate A (which later drives the Data node, though note: its *own* gate signal is Data_b, see below) and gate B, each running from the PMOS row down to the NMOS row. The two access transistors, in contrast, only get short, local gate stubs — no continuous polysilicon word line. The reason: a continuous poly WL would run straight through gate A and gate B and short them together. WL is instead added in layer 6 as a metal line, sitting one layer higher and therefore crossing over the gates without conflict.

Important for the cross-coupling: the pull-up/pull-down pair of one column is always driven *by the other column* — Data-side transistors are gated by Data_b, and Data_b-side transistors are gated by Data. Otherwise an inverter would be driving its own output, and the cell couldn't hold a bit.

SRAM cell, type 4

Layer 2 / 6 · + polysilicon

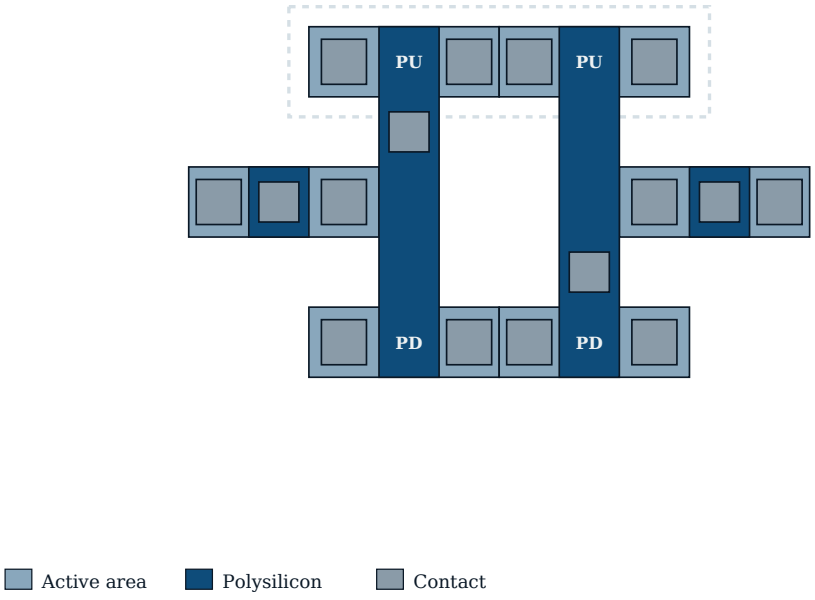


Layer 3: Contacts

Contact squares appear everywhere the next layer (metal-1) needs electrical access to diffusion or polysilicon: at both ends of every Data/Data_b column (PMOS drain, access-transistor drain, NMOS drain), at the source connections (VDD, VSS, BL, BL), and one each on gate A and gate B for the feedback path.

SRAM cell, type 4

Layer 3 / 6 · + contacts



Layer 4: Metal-1

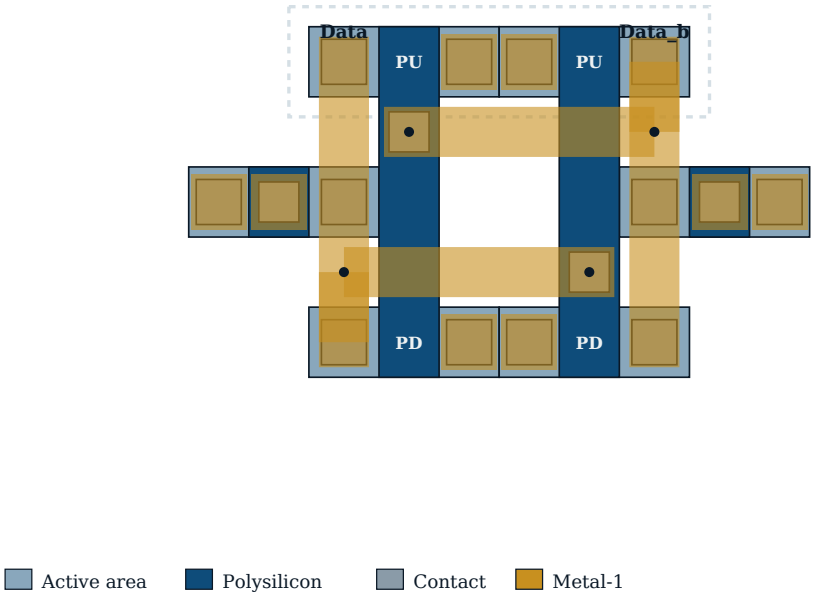
This is where the actual wiring inside the cell happens. Two things at once:

First, the local connection: each Data/Data_b column gets a continuous metal-1 trace that ties the PMOS drain, access-transistor terminal, and NMOS drain together into a single electrical node.

Second, the feedback across the full cell height: Data_b has to reach gate A, and Data has to reach gate B — both paths necessarily cross, but must never touch. This is solved by routing the two paths through different gaps (one above, one below the word-line row), so that in the picture they form an "O"-shaped loop around the two gates without ever crossing each other.

SRAM cell, type 4

Layer 4 / 6 · + metal-1

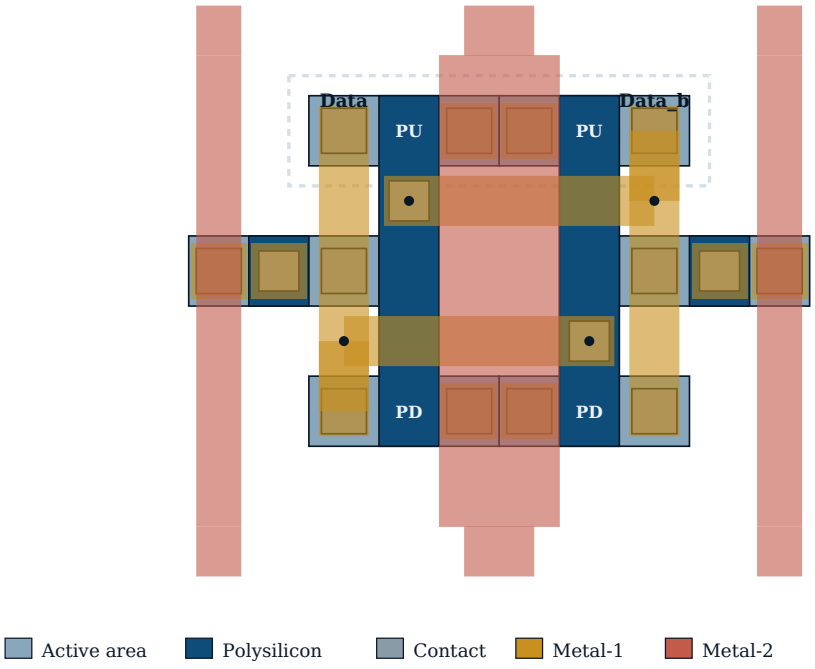


Layer 5: Metal-2

VDD, BL, and BŁ are added as vertical lines — VDD centrally between the two columns, BL and BŁ on the outside. All three run the full height of the cell and are ready at the top and bottom edges to connect to the next cell in the same column of the array.

SRAM cell, type 4

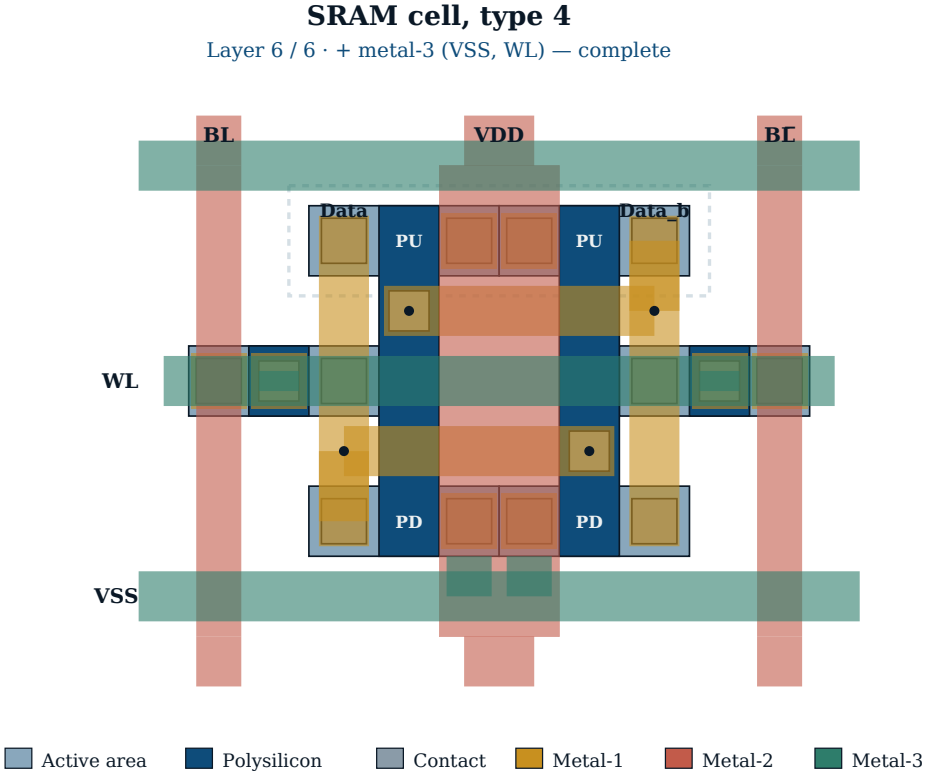
Layer 5 / 6 · + metal-2 (VDD, BL, BL̄)



Layer 6: Metal-3 — Complete

Finally, VSS (top and bottom) and WL (middle), both horizontal. That WL only appears here, instead of in polysilicon, is no accident: on this layer it can run freely over VDD, gate A, and gate B, because every layer beneath it is insulated underneath — crossings between *different* layers are unproblematic; only the *same* layer must never touch itself.

With that, the cell is fully wired. All four connections (VDD, VSS, BL, BL̄, WL) sit ready at the edges to connect seamlessly to neighbouring cells in the array — VDD/BL/BL̄ upward and downward (same column), VSS/WL left and right (same row).



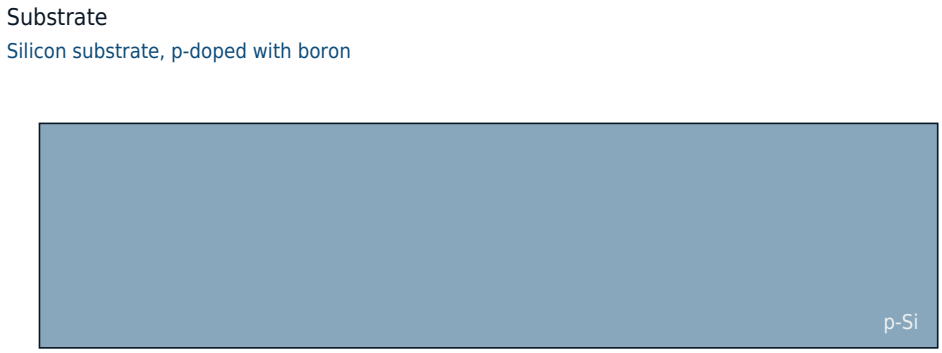
A note on the rendering: from metal-1 onward, every layer is drawn semi-transparent rather than opaque — just like in real layout editors (e.g. Virtuoso, KLayout), where transparent layers let you see what lies underneath.

Field-effect transistors

Fabrication of an n-Channel FET

1. Substrate

The basis for an n-channel field-effect transistor is a p-doped silicon substrate, using boron as the dopant.

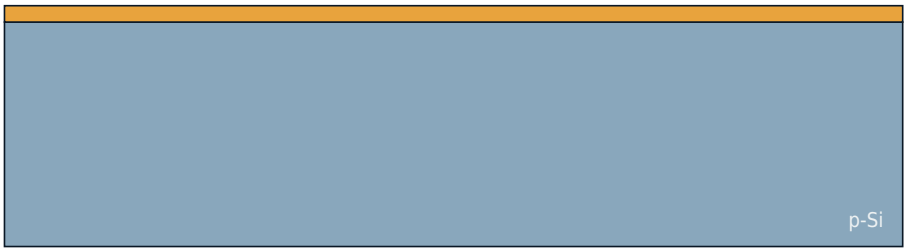


2. Oxidation

Silicon dioxide SiO_2 (the gate oxide, GOX for short) is grown on the substrate by dry oxidation. It insulates the gate, deposited later, from the substrate.

Gate Oxide (GOX)

Thermal oxidation forms a thin oxide layer



Silicon (p-Si) Gate oxide

The thin, thermally grown gate oxide will later insulate the gate
from the channel and must therefore be especially uniform and defect-free.

3. Deposition

Nitride is deposited in an LPCVD process; it later serves as a mask during field oxidation.

Nitride as a Mask

LPCVD deposition of silicon nitride

Silicon (p-Si) Gate oxide Nitride

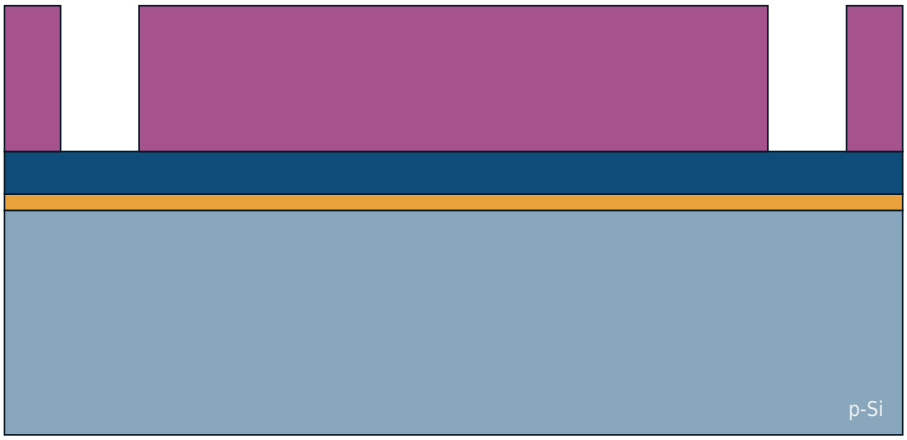
The nitride serves as a mask for the field oxidation in the following steps
– it can later be removed selectively against the oxide.

4. Photolithography

A photoresist is applied on top of the nitride, exposed, and developed, producing a structured resist layer that serves as an etch mask.

Photoresist as an Etch Mask

Resist patterning for the nitride etch



Silicon (p-Si) Gate oxide Nitride Photoresist

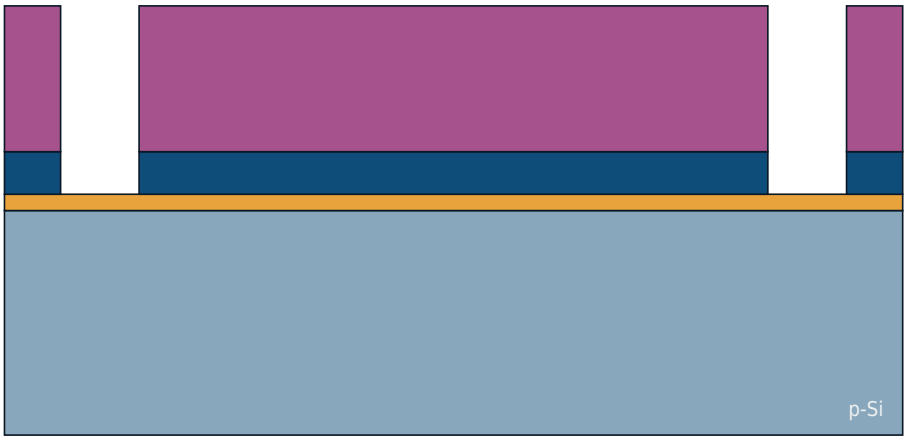
The resist openings define where the nitride will be removed in the next step – the field oxide will later form there.

5. Etching

The resist masks the nitride; the exposed areas are removed by reactive ion etching.

Etching the Nitride

Reactive ion etching transfers the resist pattern into the nitride



Silicon (p-Si) Gate oxide Nitride Photoresist

Through the resist openings, the nitride is etched away down to the gate oxide.

6. Resist Removal

After the pattern has been transferred into the nitride, the resist is removed wet-chemically using a developer solution.

Removing the Resist

The photoresist is removed with developer solution

Silicon (p-Si) Gate oxide Nitride

What remains is the patterned nitride mask, which will locally prevent the field oxidation in the next step.

7. Oxidation

The nitride acts as a mask; thermal wet oxidation occurs only where the nitride has been removed. The field oxide (FOX, e.g. 700 nm) provides lateral isolation from neighbouring devices.

<?xml version="1.0" encoding="UTF-8"?>

Field Oxide (FOX)

Thermal wet oxidation grows only where no nitride masks the surface

Silicon (p-Si) Gate oxide Nitride

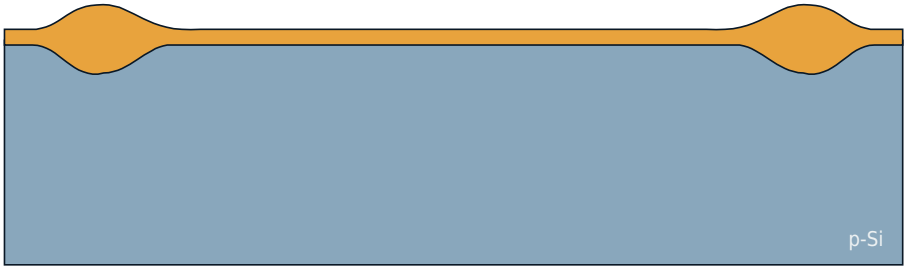
Where the nitride covers the surface, the substrate stays unoxidized. In the exposed areas, the thicker field oxide grows – the characteristic bird's beak forms at the transition to the nitride.

8. Etching

After oxidation, the nitride is removed in a wet-chemical etching process.

Removing the Nitride

Wet etching removes the nitride mask selectively



Silicon (p-Si) Gate oxide

What remains is the finished field oxide structure: thick oxide over the isolation areas, thin gate oxide over the active area.

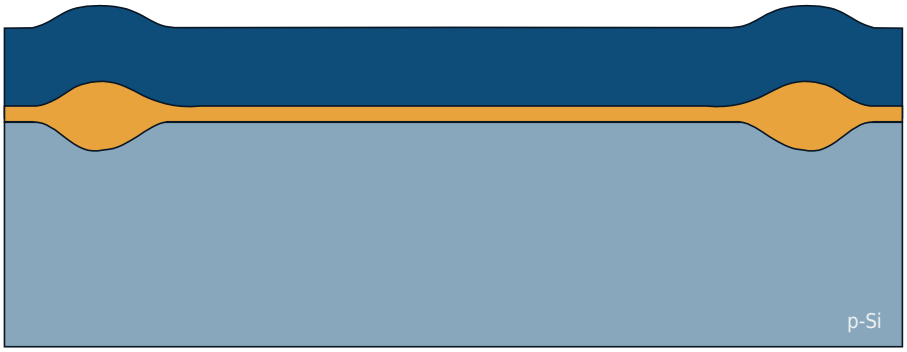
9. Deposition

Polycrystalline silicon is deposited in an LPCVD process.

<?xml version="1.0" encoding="UTF-8"?>

Polysilicon, Gate

LPCVD deposition of the future gate electrode



Silicon (p-Si) Gate oxide Polysilicon

The polysilicon is deposited over the entire surface and patterned into the gate electrode in the next step.

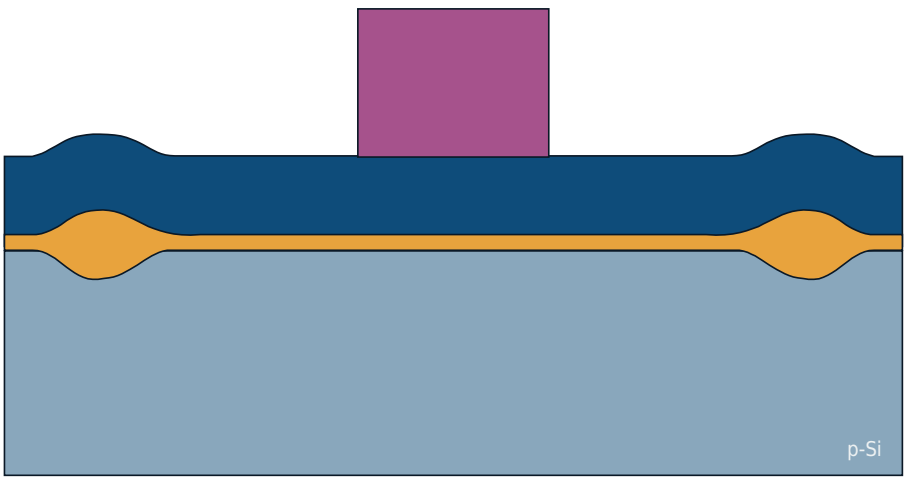
10. Photolithography

A resist layer is patterned on top of the polysilicon to serve as an etch mask.

<?xml version="1.0" encoding="UTF-8"?>

Photoresist as an Etch Mask

Resist patterning for the gate patterning



Silicon (p-Si) Gate oxide Polysilicon Photoresist

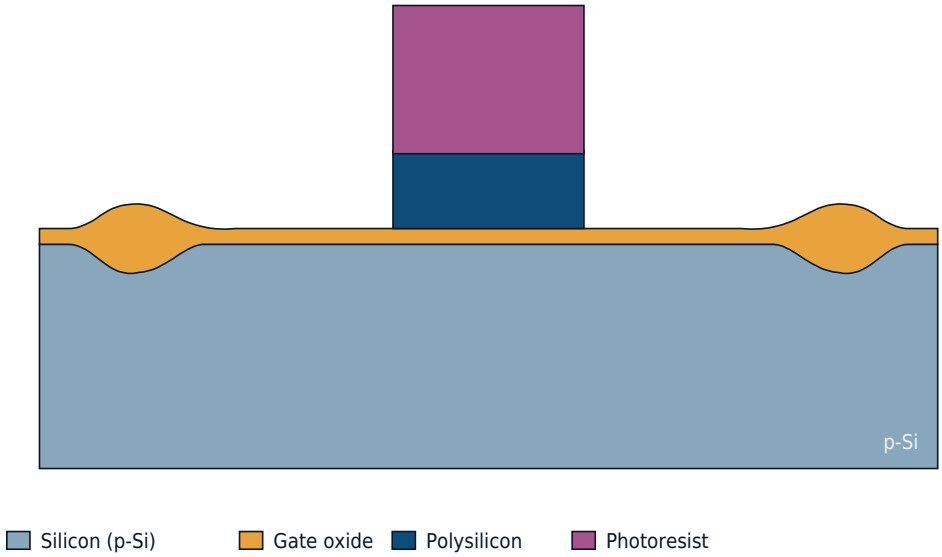
The resist island marks exactly the later position and width of the gate electrode.

11. Etching

The photoresist again acts as an etch mask, and the silicon is patterned in a reactive ion etch step, forming the gate electrode that controls the transistor.

Etching the Polysilicon

Reactive ion etching forms the gate electrode



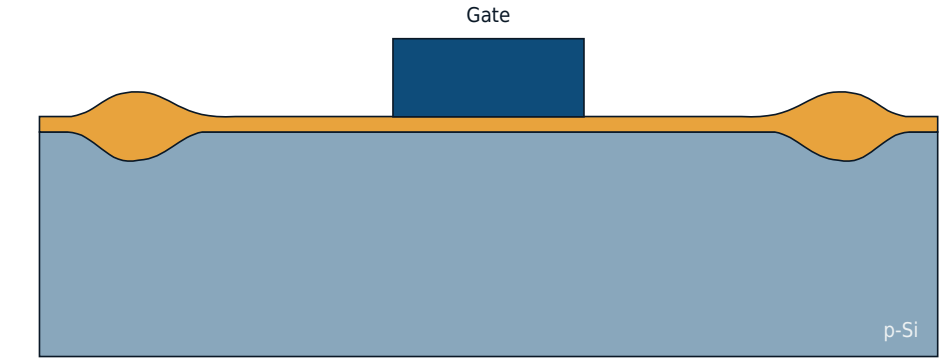
Polysilicon remains only under the resist – this forms the gate electrode with a precisely defined channel length.

12. Resist Removal

After etching, the resist is again removed wet-chemically.

Removing the Resist

The photoresist is removed, the gate is exposed



Silicon (p-Si) Gate oxide Polysilicon

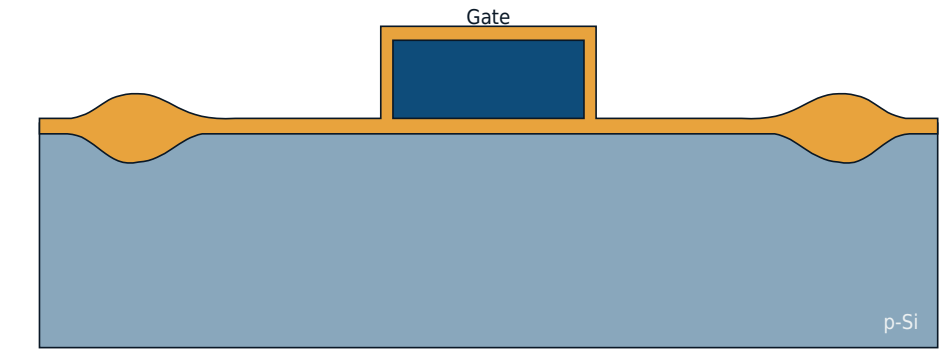
The finished, patterned gate now stands exposed on the thin gate oxide.

13. Oxidation

A thin oxide, the post-oxide, is deposited next. It protects the gate electrode and also acts as a spacer for the source/drain implantation.

Post-Oxidation

Thermal oxidation protects the gate and forms the spacer



Silicon (p-Si) Gate oxide / post-oxide Polysilicon

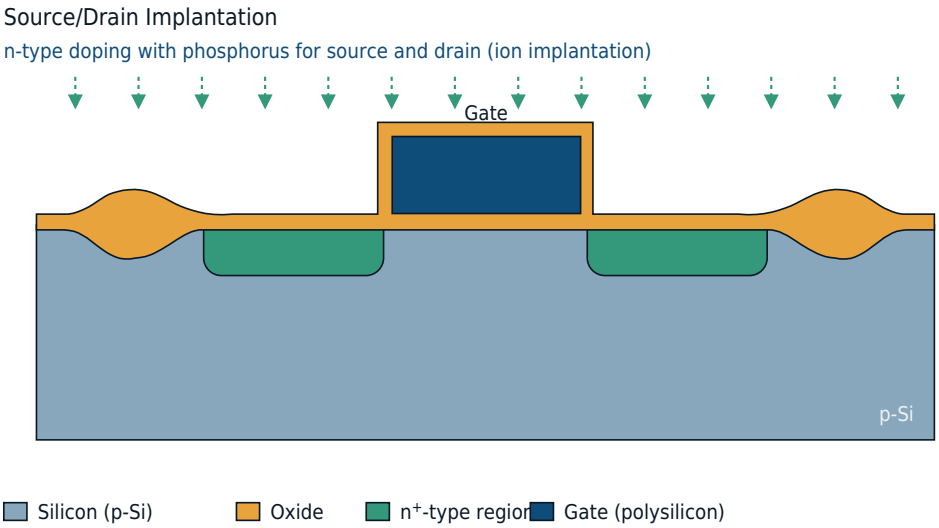
The post-oxide protects the gate edges and serves as a spacer during the subsequent source/drain implantation.

14. Ion Implantation

In an implantation step using phosphorus ions, the source and drain regions are n-doped. Since the gate electrode acts as an implantation mask, defining the width of the n-channel between source and drain, this is referred to as self-

alignment. By adjusting the width of the spacers, the distance between the implanted dopant atoms and the gate can be precisely controlled according to the electrical requirements.

<?xml version="1.0" encoding="UTF-8"?>



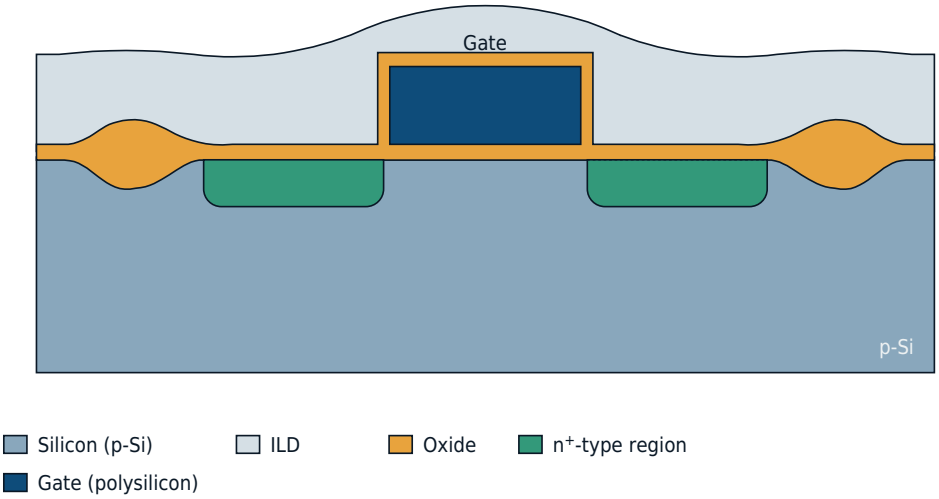
15. Oxidation

An oxide (the interlayer oxide, ILD for short, e.g. 700 nm) is deposited to insulate against the metallization layers above. This is done in an LPCVD process using TEOS, which provides good step coverage.

<?xml version="1.0" encoding="UTF-8"?>

Interlayer Oxide (ILD)

TEOS deposition as insulation for the later interconnects



The interlayer oxide insulates the later aluminum wiring from the transistor underneath and follows the existing topography.

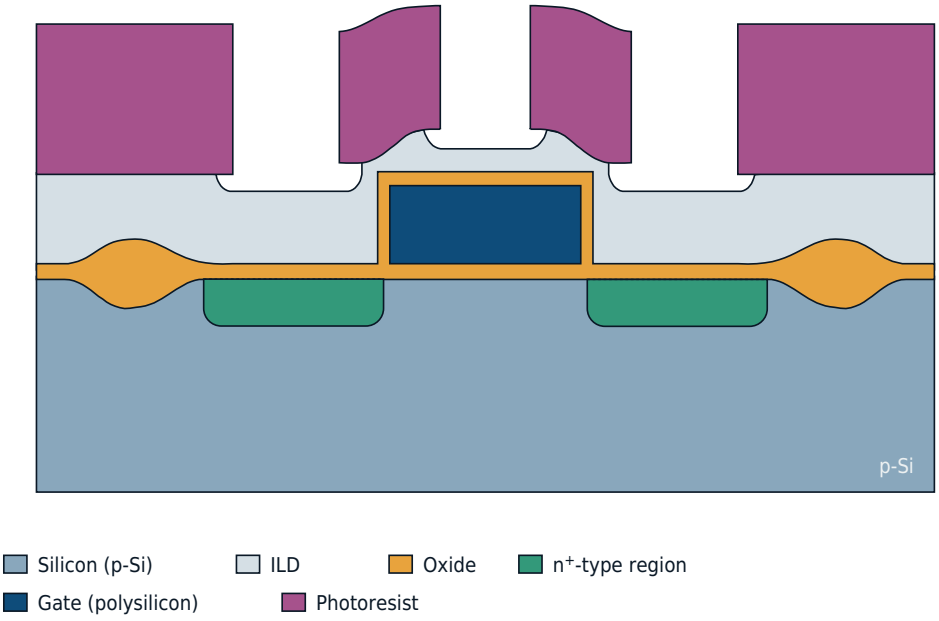
16. Photolithography and Etching

A further resist layer is patterned, and the edges of the contact holes are rounded off in an isotropic etch process.

<?xml version="1.0" encoding="UTF-8"?>

Photoresist as an Etch Mask

Resist patterning for the subsequent contact hole etch



The resist openings define where the contact holes to source, drain, and gate will be etched.

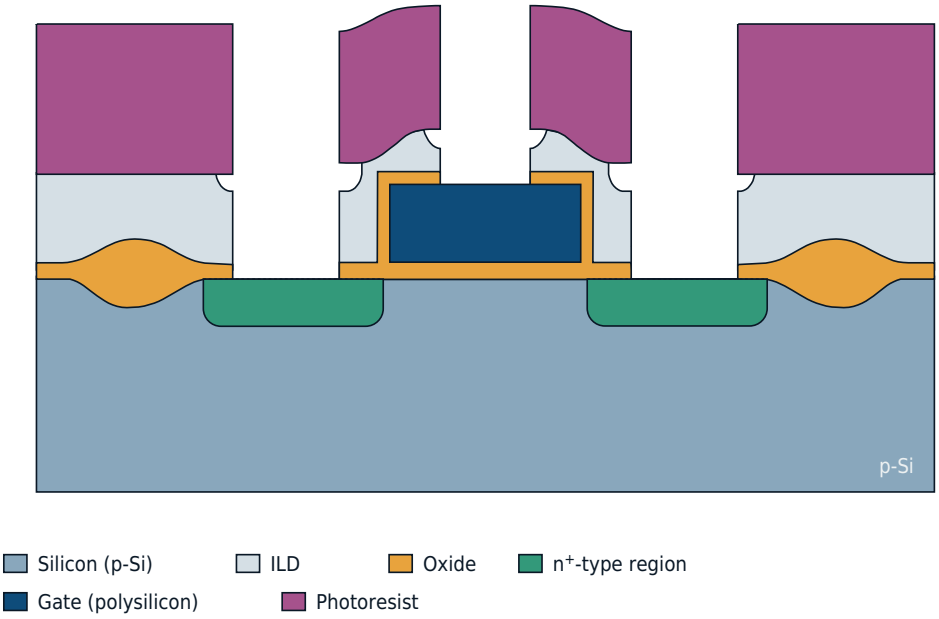
17. Etching

The contact holes are then opened down to the n-doped regions and to the gate in an anisotropic etch process.

<?xml version="1.0" encoding="UTF-8"?>

Contact Hole Etch

Reactive ion etching opens the interlayer oxide down to source, drain, and gate



The resist protects the remaining interlayer oxide while the holes are etched down to the three terminal areas.

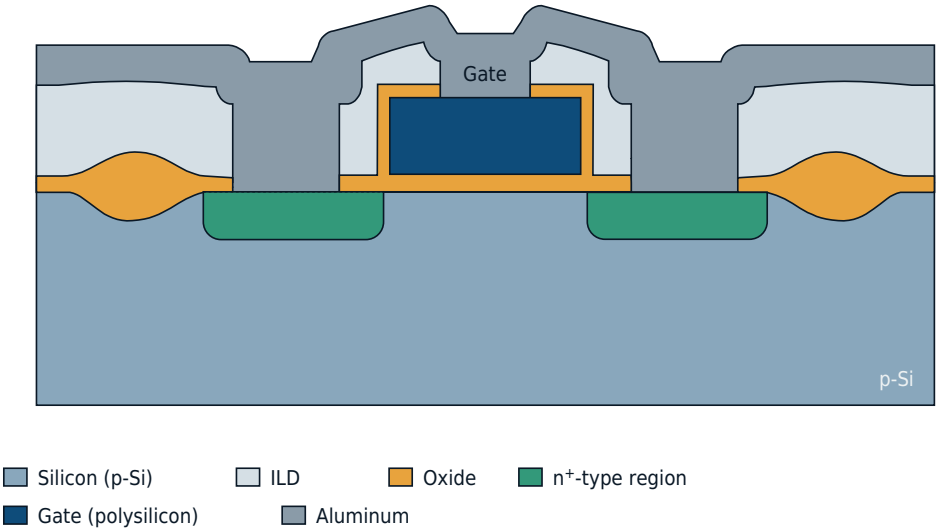
18. Metallization

The contact holes are filled with aluminium in a sputtering process.

<?xml version="1.0" encoding="UTF-8"?>

Aluminum Deposition

Sputtering fills the contact holes and covers the surface



The aluminum completely fills the contact holes and forms a continuous layer over the whole structure.

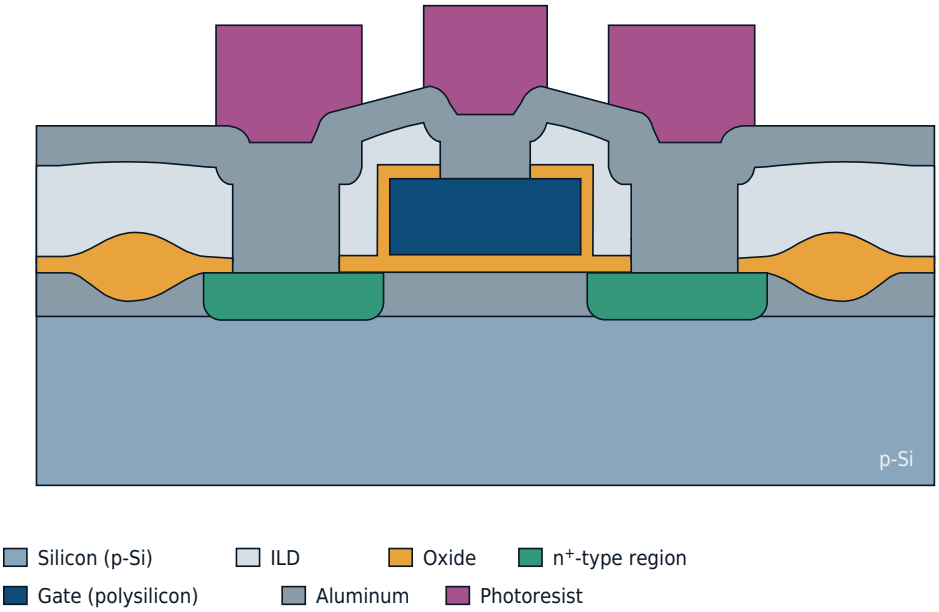
19. Photolithography

In a final lithography step, another resist layer is patterned.

<?xml version="1.0" encoding="UTF-8"?>

Photoresist as an Etch Mask

Resist patterning for the aluminum patterning



The resist defines where the aluminum will remain as an interconnect and where it will be removed in the next step.

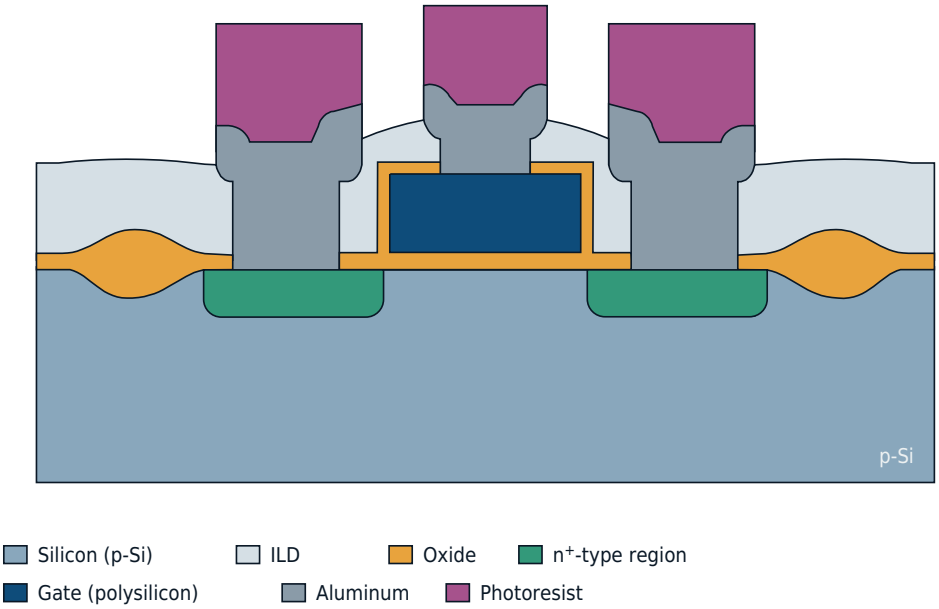
20. Etching

The resist pattern is transferred into the underlying metallization layer in an anisotropic dry etch step.

<?xml version="1.0" encoding="UTF-8"?>

Etching the Aluminum

Anisotropic dry etching patterns the three terminals



The aluminum patterning separates the three terminals for source, gate, and drain from each other. In the last step, the resist is removed.

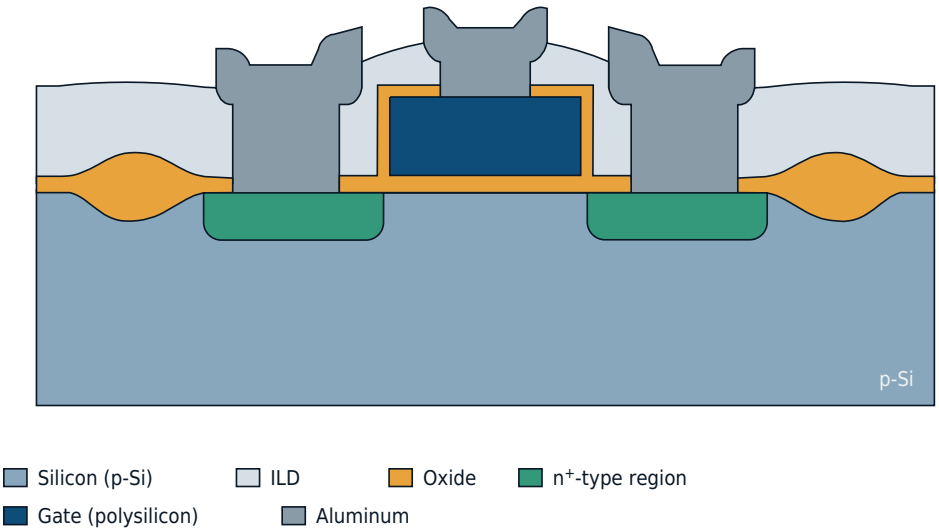
21. Resist Removal

Finally, the resist is removed, leaving behind aluminium interconnects that drive the transistor.

<?xml version="1.0" encoding="UTF-8"?>

Finished n-Channel Transistor

Cross-section after the last step: the resist is removed, the three contacts are exposed



The aluminum patterning separates the three terminals for source, gate, and drain from each other. In the last step, the resist is removed.

The actual structure of a transistor is considerably more complex. Additional planarization layers may be used to support the lithographic processes, and several drain/source implantations may be carried out to precisely set the threshold voltage. Additional spacers at the gate are likewise possible, in order to fine-tune the channel length or influence the doping profile.

Fundamentals

The basic structure and operating principle of this device are covered in the Fundamentals chapter Transistors and Memory. The following sections describe fabrication in detail.

Bipolar transistors

Fabrication of an NPN Bipolar Transistor

1. Substrate

The basis for an NPN bipolar transistor is a p-doped silicon substrate, using boron as the dopant. A thick oxide layer (e.g. 600 nm) is deposited on top of it.

Starting Material

Lightly p-doped silicon wafer as substrate

Silicon (p-Si) Oxide

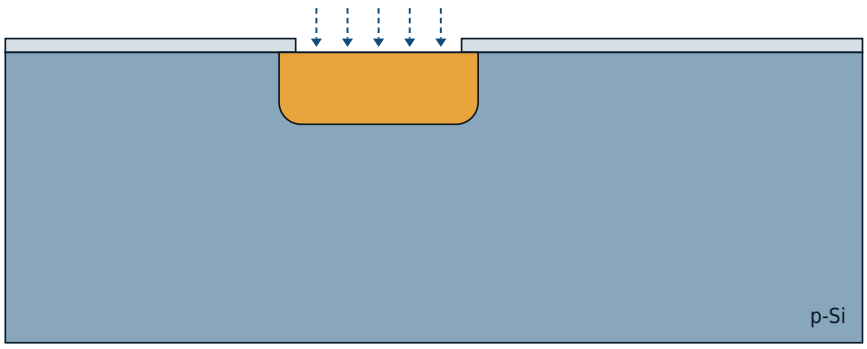
The starting point is a lightly p-doped silicon wafer, on whose surface a thin oxide layer grows as a protective and masking layer.

2. Buried Layer Implantation

The oxide serves as an implantation mask. Antimony (Sb) is used as the dopant, since it diffuses less than phosphorus during subsequent diffusion processes. The heavily n^+-doped buried collector serves as a low-resistance contact area for the collector terminal.

Opening a window and implanting

Lithography defines the window for the buried layer



Silicon (p-Si) Oxide n^+-doped

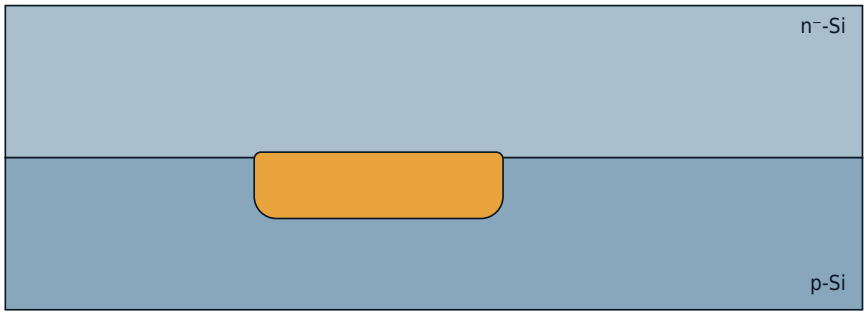
In the exposed area, ion implantation creates a heavily n-doped region. This buried layer later lies beneath the finished transistor.

3. Homoepitaxy

In an epitaxial process, a high-resistance, lightly n^--doped collector layer is deposited (typically 10 µm).

Epitaxy

A new, lightly n-doped silicon layer grows



Silicon (p-Si) n⁻-Si (epitaxial layer) n⁺-doped

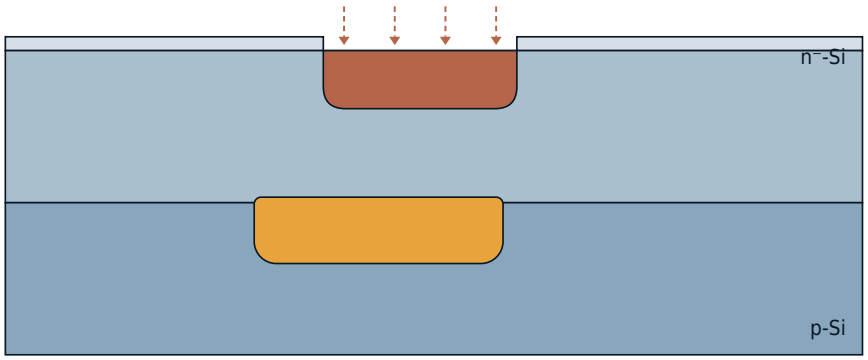
The new epitaxial layer forms the later active region of the transistor.
The buried n^+ layer remains at the interface with the substrate.

4. Base Implantation

The p-doped base is created using boron ions; a subsequent diffusion step enlarges the region.

Base Implantation

A new window defines the position of the base



Silicon (p-Si) n⁻-Si (epitaxial layer) Oxide n⁺-doped p-doped

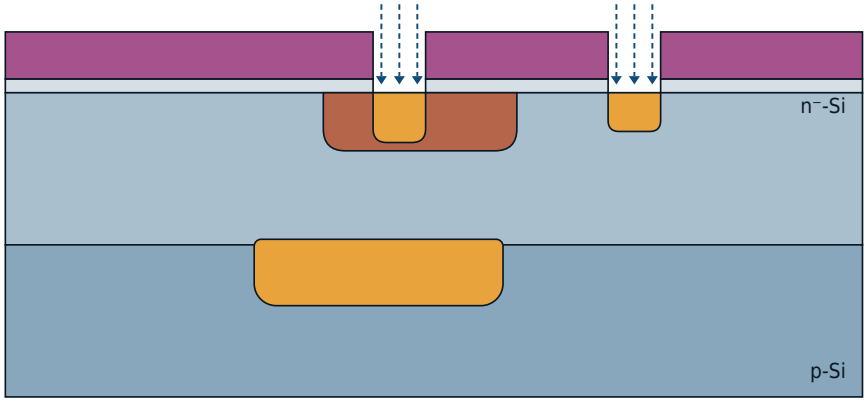
Through the new window, the base is implanted: a lightly p-doped region directly above the buried n⁺ layer.

5. Emitter and Collector Implantation

Using phosphorus ions, the two heavily n^+-doped emitter and collector terminals are introduced.

Emitter and Collector Implantation

A resist mask defines two new windows



Silicon (p-Si) n⁻-Si (epitaxial layer) Oxide Photoresist n⁺-doped p-doped

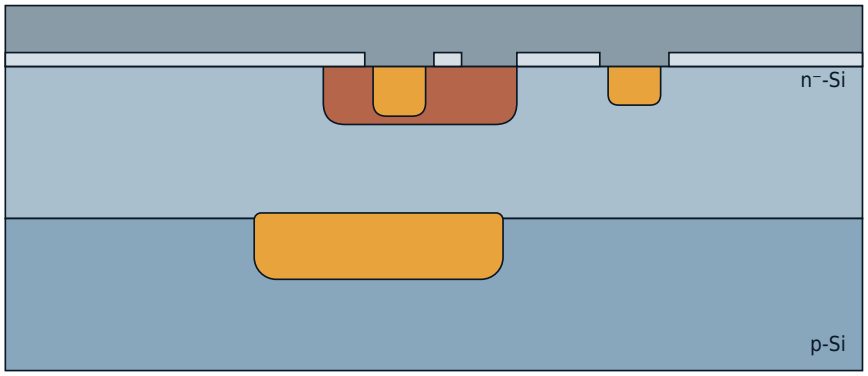
Through the resist mask, two n⁺ regions are created: the emitter within the base, and a low-resistance collector terminal beside it.

6. Metallization and Photolithography

Aluminium is deposited in a sputtering process to contact the three terminals, and a resist layer is patterned on top of it.

Contact Holes and Aluminium

After removing the resist, aluminium is deposited



Silicon (p-Si) n⁻-Si (epitaxial layer) Oxide n⁺-doped p-doped Aluminium

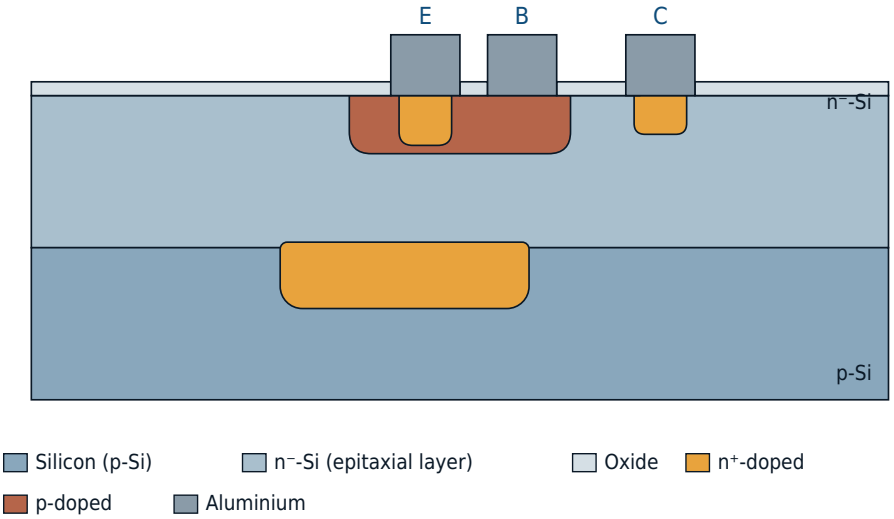
The resist is removed, the oxide is opened down to the silicon at three points and a continuous aluminium layer is then deposited.

7. Etching

Finally, the terminals for emitter, base, and collector are patterned in an anisotropic dry etch step.

Aluminium Patterning

Etching separates the three terminal contacts E, B, and C



The aluminium layer is patterned into three separate contacts for emitter (E), base (B), and collector (C).

Due to various optimizations, bipolar transistors today are built from more than three layers (npn or pnp). The collector region always consists of at least two zones with different doping levels. The terms npn and pnp refer only to the active inner region, not to the actual layer stack.

Fundamentals

The basic structure and operating principle of this device are covered in the Fundamentals chapter Transistors and Memory. The following sections describe fabrication in detail.

Construction of a FinFET

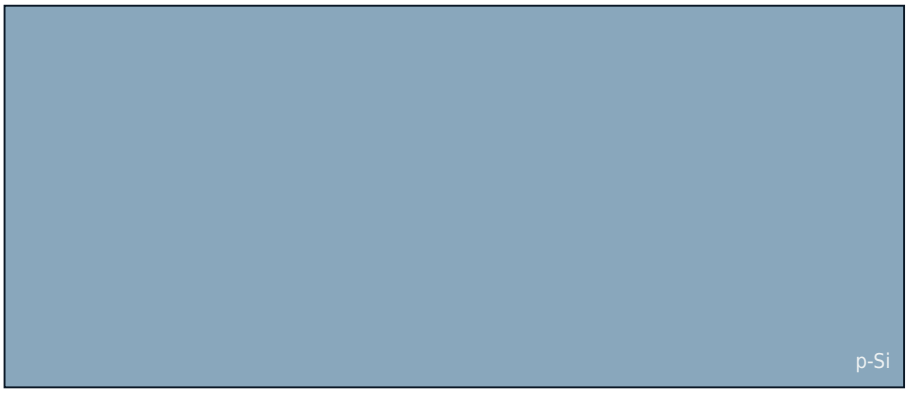
Construction of a FinFET in a bulk process

1. Substrate

As with any silicon transistor, the basis for a FinFET is a lightly p-doped silicon substrate.

Substrate

Silicon substrate, p-doped with boron



Silicon (p-Si)

The basis for the entire fabrication process is a lightly p-doped silicon substrate as the wafer.

2. Hard Mask

A hard mask is deposited on the substrate, usually a stack of oxide and silicon nitride. It later transfers the lithographic pattern into the silicon with markedly less loss than a resist-only process.

Hard Mask

Oxide/nitride stack as an etch mask for fin patterning

Silicon (p-Si) Hard mask (oxide/nitride)

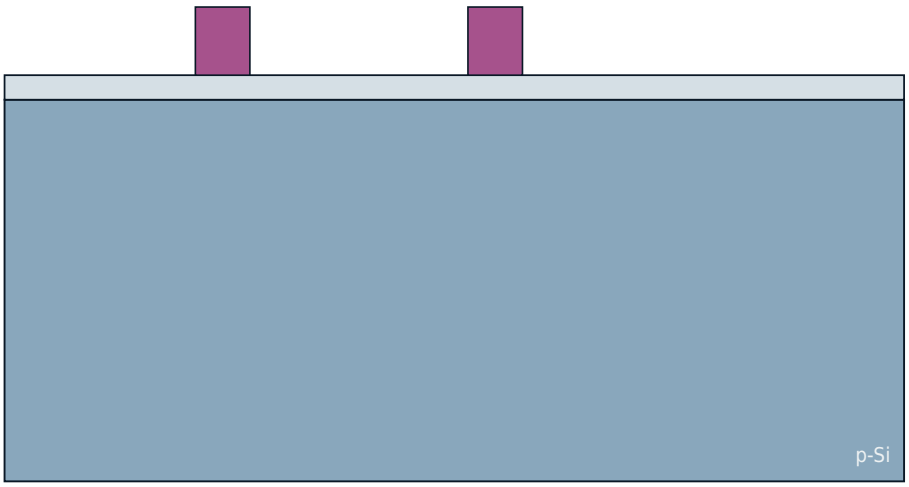
A hard mask of oxide and nitride is deposited – it will later transfer the lithographic pattern into the silicon with little loss.

3. Fin Lithography

Photoresist is patterned into narrow strips that set the position and width of the later fins. In the first volume-production processes, the fin width was 10 to 15 nm and has since dropped to just a few nanometers; the height should ideally be twice that or more. Such narrow fins can no longer be exposed directly; they are instead created through multipatterning: the fin width results from the thickness of a deposited layer, not from a mask.

Fin Lithography

Resist patterning sets the position and width of the fins



Silicon (p-Si) Hard mask Photoresist

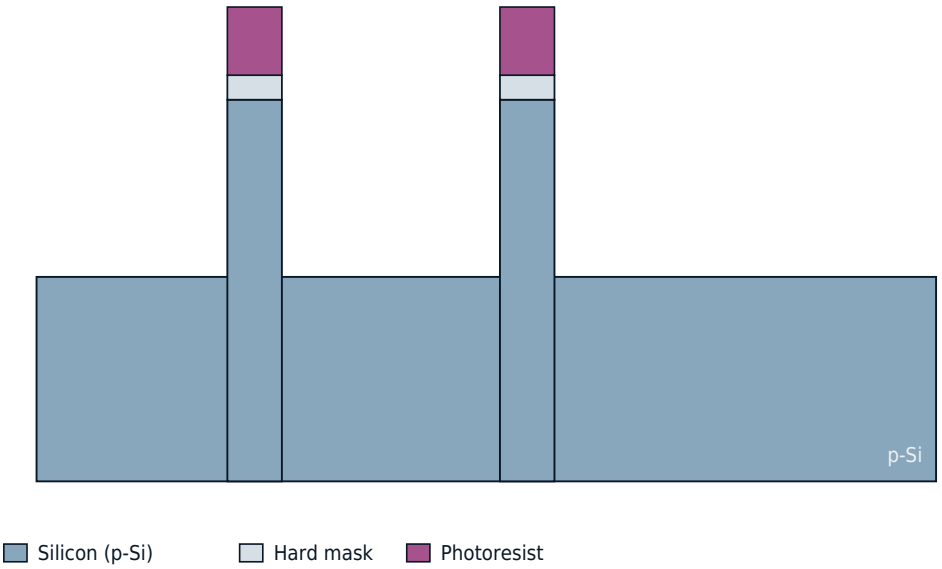
The narrow resist strips define the later fin width – typically only a few tens of nanometers.

4. Fin Etch

In a highly anisotropic dry etch step, the free-standing fins are etched into the silicon substrate. In the bulk process, the etch has to be time-based, since unlike with an SOI substrate there is no stop layer within the silicon to indicate the depth reached.

Fin Etch

Anisotropic etching transfers the pattern into the hard mask and silicon



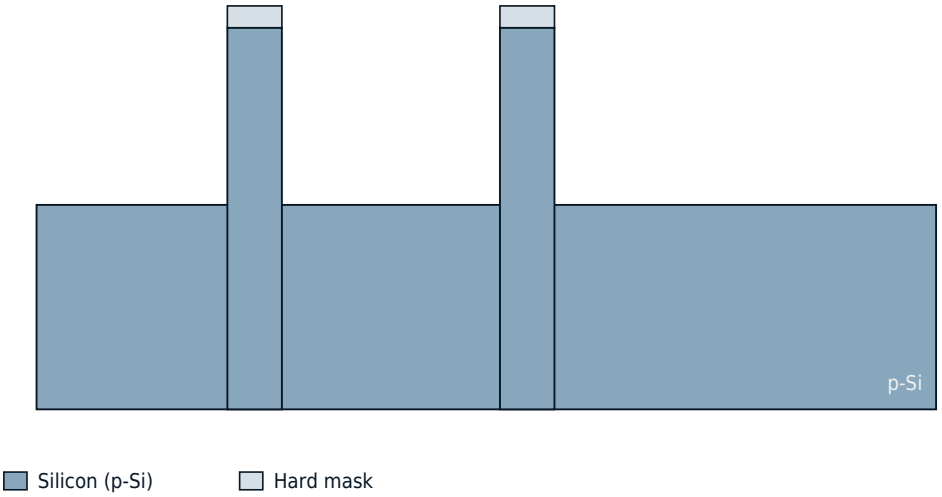
Reactive ion etching removes silicon between the fins – the etch depth later determines how far the fins protrude from the STI.

5. Removing the Hard Mask

Resist and hard mask are removed. The bare etched fins stand ready for isolation.

Removing the Resist

Only the resist is removed – the hard mask remains as a cap on the fins



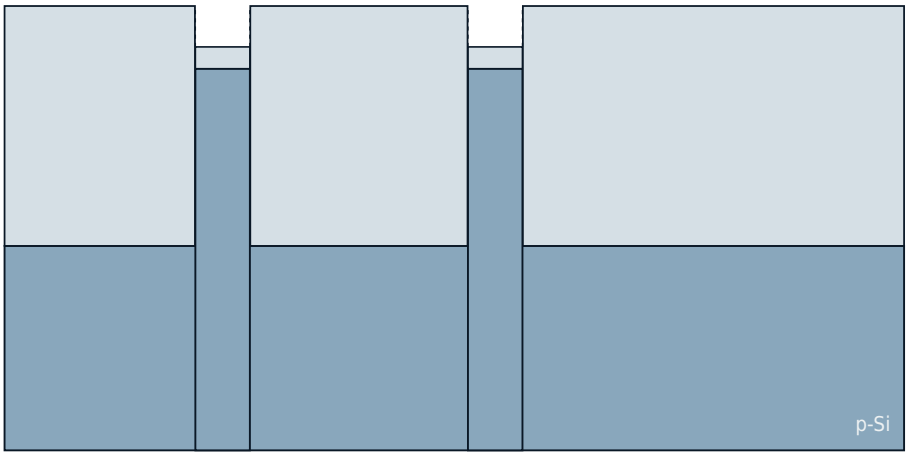
Only the photoresist is removed. The hard mask deliberately remains on the fins – it serves as a polish stop layer in the following STI process and is removed again only after planarization (step 8).

6. STI Oxide

For isolation, a Shallow Trench Isolation (STI) oxide is deposited that must provide good fill behavior for narrow, deep trenches – it completely overfills the trenches between the fins in the process.

STI Oxide

Oxide deposition fills the trenches between the fins



Silicon (p-Si) STI oxide / hard mask

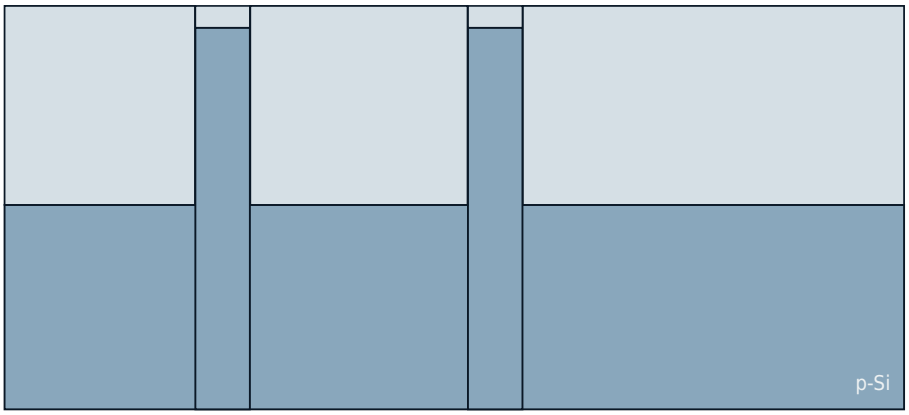
The isolation oxide (Shallow Trench Isolation, STI) is deposited over the whole wafer, overfilling the trenches and the hard mask caps.

7. STI Planarization

The oxide is planarized by chemical mechanical polishing. The hard mask serves as the stop layer.

STI Planarization

CMP polishes the oxide back – the hard mask stops the removal



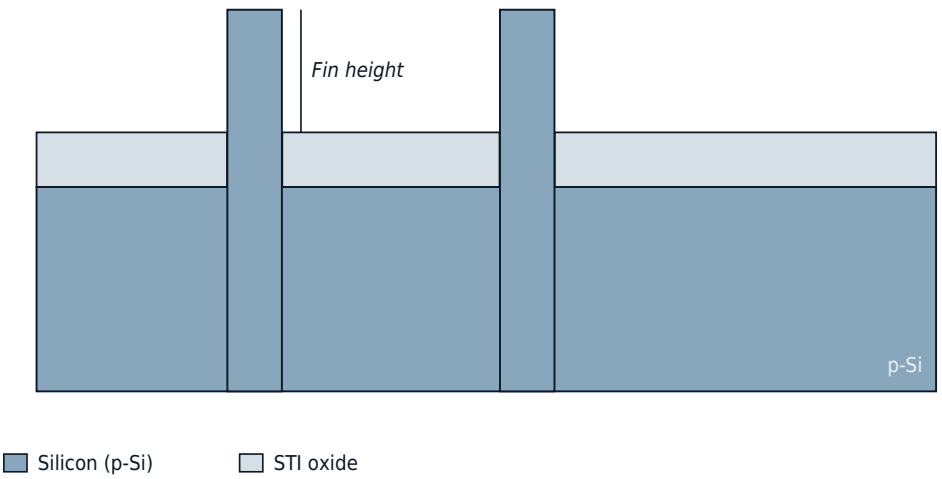
Chemical mechanical polishing removes the excess oxide. The hard mask is considerably harder than the oxide and reliably stops the removal at the same height – without it, the fins themselves would be polished away.

8. STI Recess

In a further, time-based etch step, the oxide is selectively etched back until the fins protrude from the oxide at the desired height – the three-dimensional structure characteristic of the FinFET emerges. What remains is a lateral isolation between the fins. Since the fins are still connected to one another through the substrate, a doping step is required to electrically isolate the silicon fins from each other (not shown). Alternatively, the oxide can be grown into the base region of the fins in a high-temperature step, isolating the channels from the silicon substrate. In volume production, the first approach has prevailed: a region beneath the fin is deliberately heavily doped so that no conductive path can form there.

STI Recess

Hard mask and oxide are etched back – the fin flanks are exposed



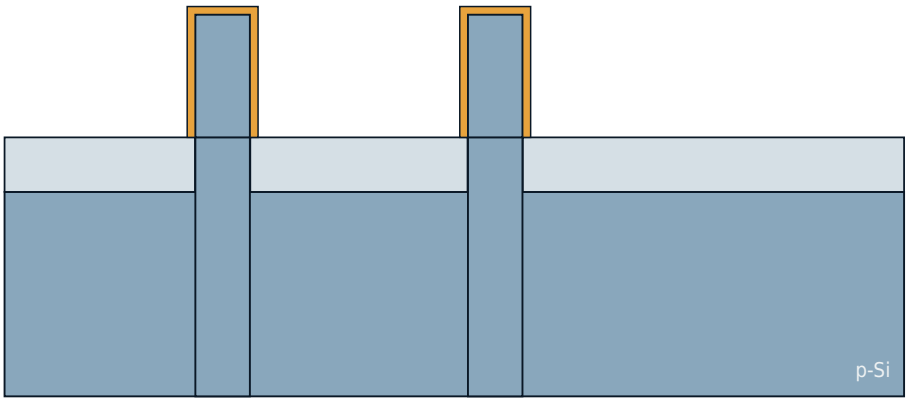
First, the hard mask is selectively removed, then the STI oxide is etched back until the fins protrude at the desired height. This silicon flank forms the channel – the gate will wrap around it on three sides.

9. Gate Dielectric

The gate dielectric is formed on the exposed fin flanks to isolate the channel region from the gate electrode. In volume production, this is no longer a thermally grown oxide, but a layer of hafnium dioxide deposited by atomic layer deposition on top of a thin interfacial layer of silicon dioxide. Only atomic layer deposition coats a tall, narrow fin evenly on all three exposed sides.

Gate Dielectric

A high-k layer grows conformally on the exposed fin flank



Silicon (p-Si) STI oxide Gate dielectric

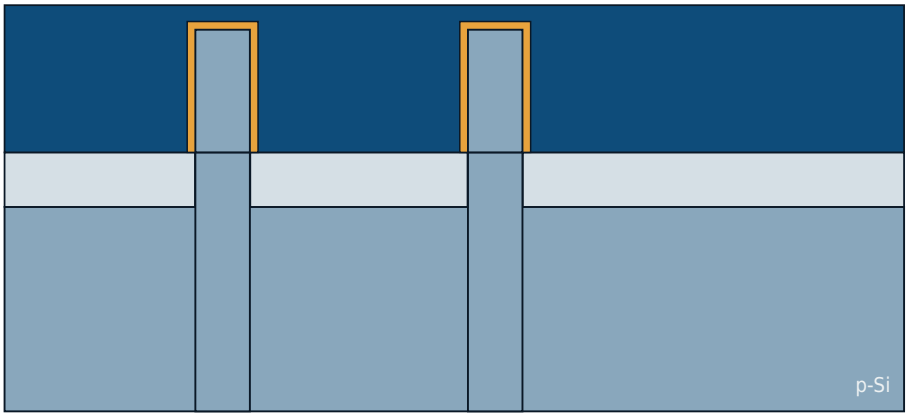
A thin high-k layer (e.g. hafnium oxide) wraps around the exposed fin flank on three sides – the later channel surface.

10. Dummy Gate

The actual gate material used later is a metal, whose work function is chosen so as to yield the desired threshold voltage – a different metal for n-channel and p-channel transistors, respectively. Because this metal would not survive the later high-temperature steps, it is introduced only at the very end. First, a placeholder gate of polysilicon is used, deposited here conformally so that it already wraps around the fins completely. All the high-temperature process steps are carried out with this placeholder, which is removed again afterward (step 18). This approach is called the gate-last, or replacement-gate, process.

Dummy Gate (Polysilicon)

Polysilicon is deposited over the whole wafer as a placeholder gate



Silicon (p-Si) STI oxide Gate dielectric

Dummy gate (polysilicon)

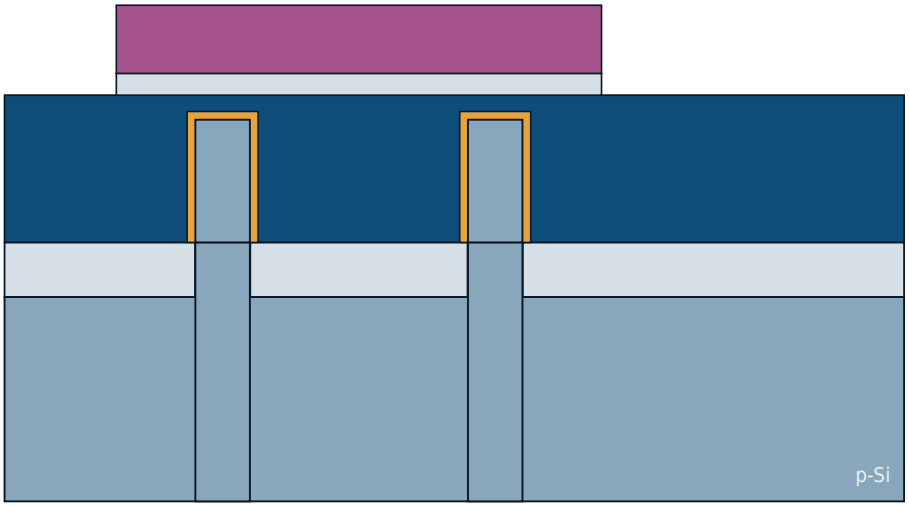
In the gate-last process, polysilicon initially serves only as a placeholder – it will be removed again after source/drain formation.

11. Gate Lithography

A resist strip running across the fins, together with a hard mask, defines the position and length of the gate – perpendicular to the cross-section shown here.

Gate Lithography

Resist and hard mask define the gate length



Silicon (p-Si) STI oxide / hard mask Gate dielectric

Dummy gate Photoresist

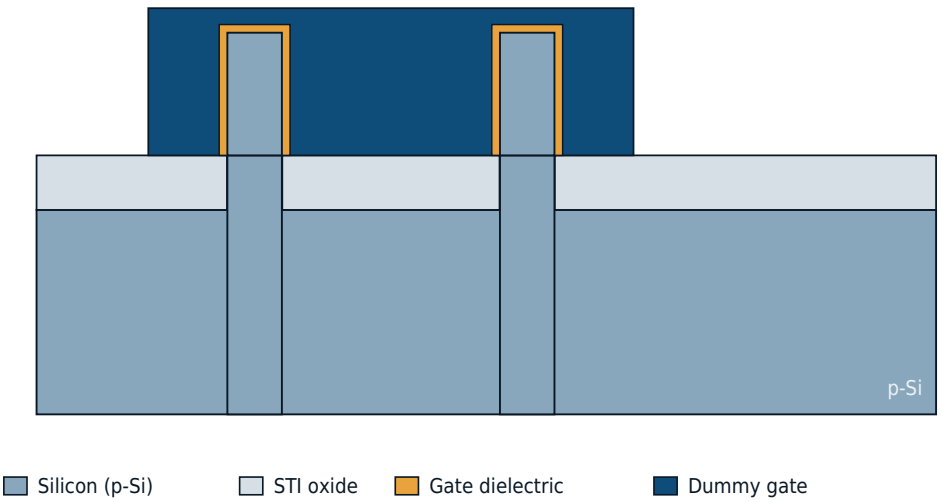
The resist strip running across the fins defines the later gate length – perpendicular to the cross-section shown here.

12. Dummy Gate Etch

Excess polysilicon is removed by reactive ion etching. What remains is the dummy gate stack, which still wraps around the fins on three sides.

Dummy Gate Etch

Reactive ion etching forms the gate stack



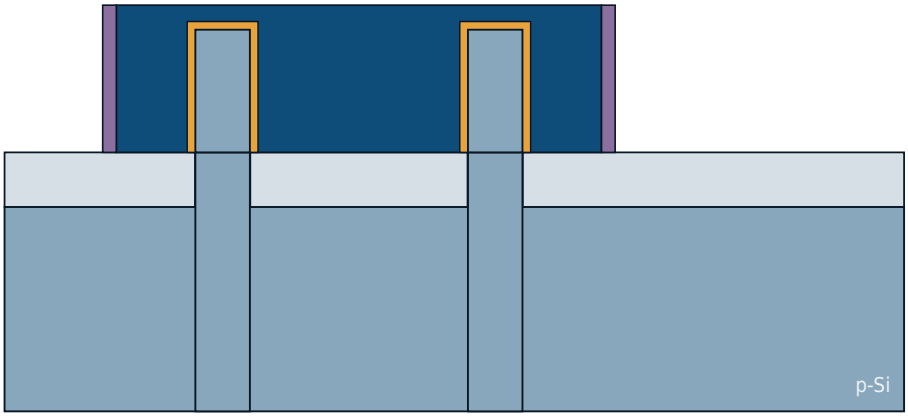
Excess polysilicon is removed – what remains is the dummy gate stack, which wraps around the fins on three sides.

13. Spacer

A nitride spacer is deposited conformally and etched back anisotropically, remaining as a sidewall spacer on the gate flanks. It protects the flanks during the following steps and later sets the distance between the gate and the source/drain contacts. A similar nitride layer between the fin and the gate can also be used deliberately to suppress the influence of the top gate segment on the channel, if only lateral control from the sides is desired.

Spacer

Nitride deposition and etch-back form the sidewall spacer at the gate



Silicon (p-Si) STI oxide Gate dielectric Dummy gate Spacer (nitride)

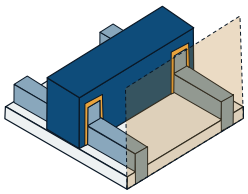
The nitride spacer protects the gate flanks and later sets the distance between the gate and the source/drain contacts.

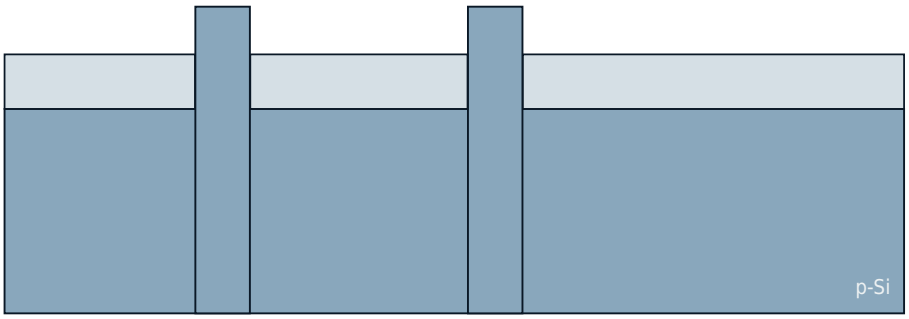
14. Source/Drain Recess

Outside the gate – in the cross-section through the later source/drain region – the fin is deliberately recessed to make room for the subsequent epitaxy. Source and drain sit along the same fin: one before and one behind the gate, along the direction of current flow. Each individual fin therefore forms a complete channel on its own, with its own source and drain region at its two ends – with several parallel fins under a shared gate, every single fin contributes to both source and drain, rather than one fin belonging to source and another to drain.

Source/Drain Recess

The fin is recessed outside the gate – cross-section in the S/D region





Silicon (p-Si) STI oxide

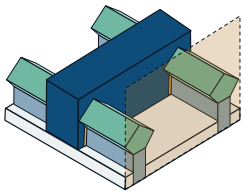
Outside the gate – in the cross-section through the later source/drain region – the fin is deliberately recessed to make room for the subsequent epitaxy.

15. Source/Drain Epitaxy

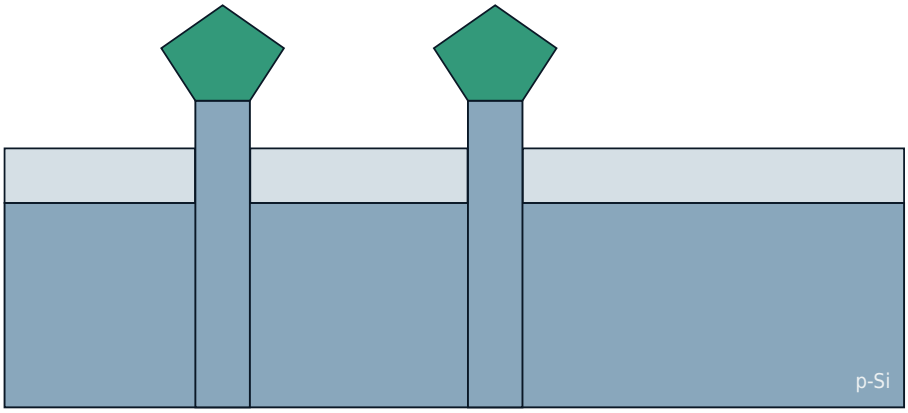
Selective epitaxy grows doped silicon (for nFETs typically Si:P, for pFETs SiGe:B) in a diamond shape on the fin stub. This characteristic "raised source/drain" geometry further increases the effective channel cross-section and lowers the access resistance. For p-channel transistors, a further effect comes into play: SiGe has a larger lattice constant than silicon and is laterally compressed as it grows epitaxially on the fin, matching the fin's crystal lattice. The resulting compressive strain is transmitted directly into the adjacent channel, where it acts uniaxially along the current direction and specifically increases hole mobility, and with it the drive current.

Source/Drain Epitaxy

Raised source/drain grows selectively on the recessed fin



Cut: source/drain



Silicon (p-Si) STI oxide Epitaxial Si (n/p-doped)

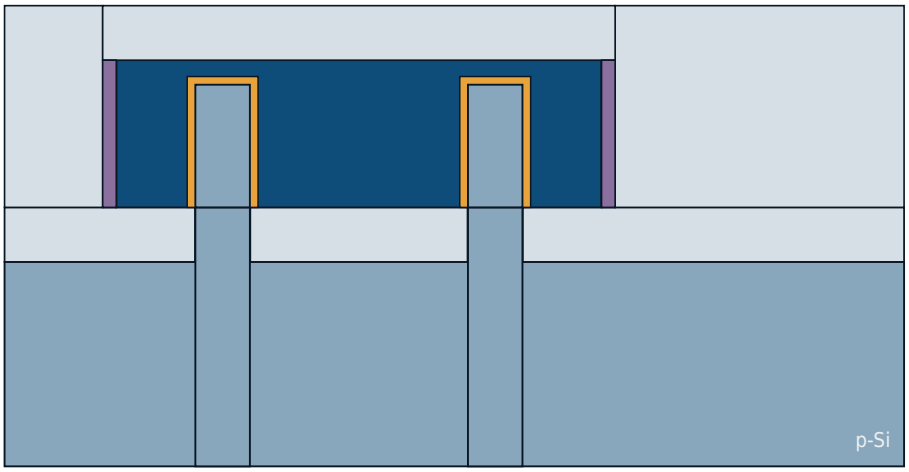
Selective epitaxy grows doped silicon (for nFET typically Si:P)
in a diamond shape on the fin stub – the characteristic
'raised source/drain' geometry also increases the channel cross-section.

16. Interlayer Dielectric

The interlayer dielectric (ILD) is deposited over the whole wafer and initially covers the dummy gate stack completely as well.

Interlayer Dielectric (ILD)

Oxide deposition insulates the gate stack – cross-section through the gate



Silicon (p-Si) STI oxide / interlayer dielectric Gate dielectric
Dummy gate Spacer

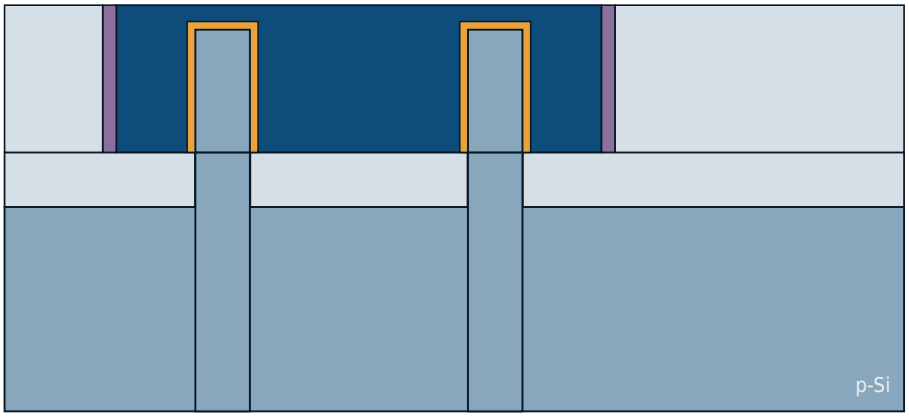
The interlayer dielectric (ILD) is deposited over the whole wafer and initially covers the dummy gate stack completely as well.

17. CMP

Chemical mechanical polishing removes the excess interlayer dielectric until the top of the dummy gate is exposed.

CMP

Planarization exposes the top of the dummy gate



Silicon (p-Si) STI oxide / interlayer dielectric Gate dielectric
Dummy gate Spacer

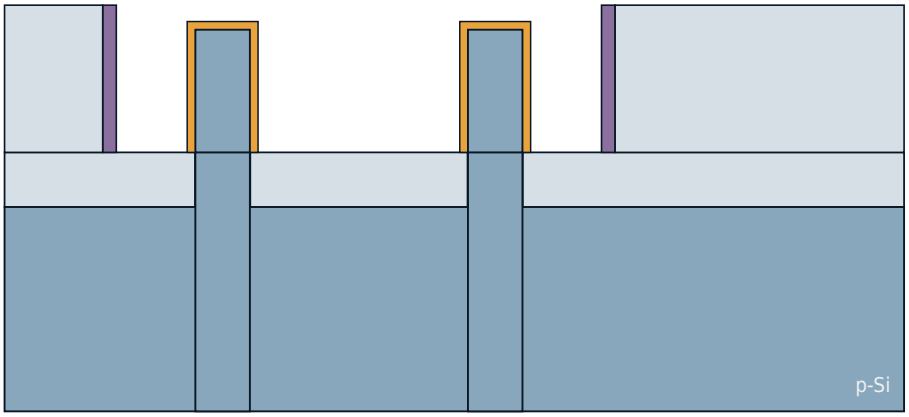
Chemical mechanical polishing removes the excess interlayer dielectric until the top of the dummy gate is exposed.

18. Removing the Dummy Gate

The polysilicon of the placeholder gate is selectively etched out – the spacer forms the walls of the open gate trench, ready for the actual metal gate.

Removing the Dummy Gate

Selective etching opens the gate trench



Silicon (p-Si) STI oxide / interlayer dielectric Gate dielectric Spacer

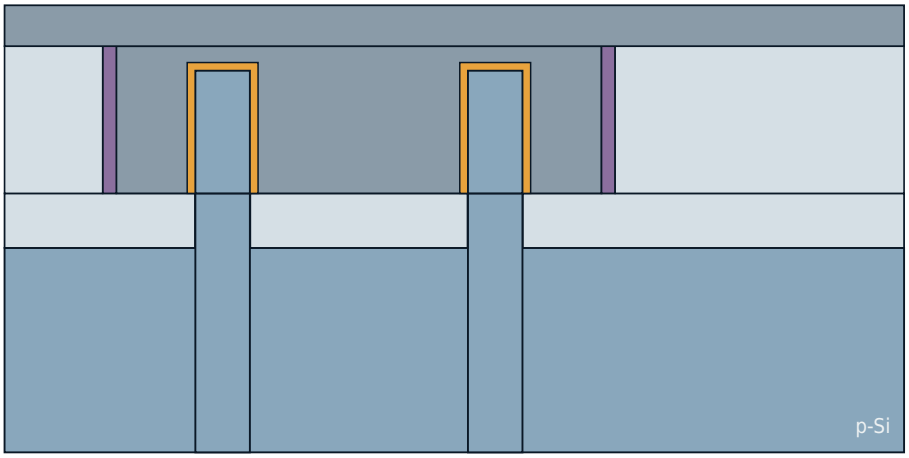
The polysilicon is selectively etched out – the spacer forms the walls of the open gate trench, ready for the metal gate.

19. High-k/Metal Gate

Work-function metal and fill metal are deposited conformally into the trench. Because the metal layers to the left and right of the channel and above it can in principle differ, this can, if needed, produce several separately tunable gate electrodes for a single transistor. The final gate wraps around the channel on three sides.

High-k/Metal Gate

The final metal gate is deposited into the trench



Silicon (p-Si) STI oxide / interlayer dielectric Gate dielectric

Metal gate Spacer

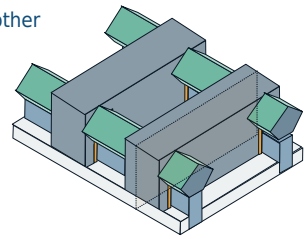
Work-function metal and fill metal are deposited conformally and fill not only the trench but initially overfill the whole surface as well – the following CMP step removes the excess.

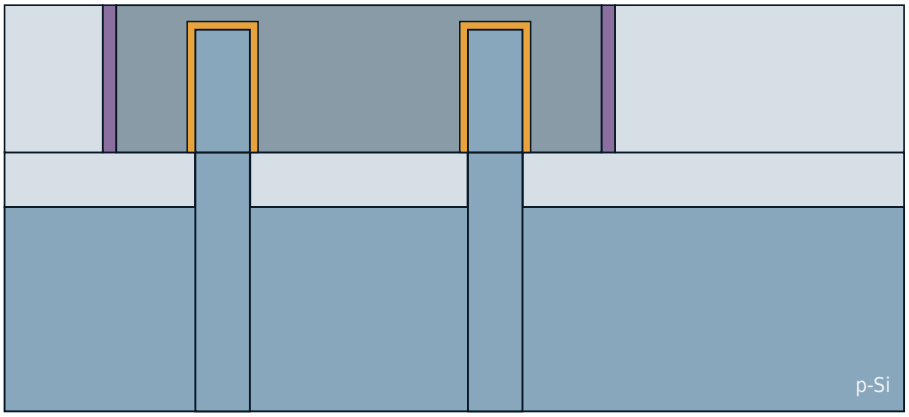
20. Gate CMP

A final polishing step removes excess gate metal and electrically separates neighboring gates.

Gate CMP

Excess metal is polished back – the gates are isolated from each other





Silicon (p-Si) STI oxide / interlayer dielectric Gate dielectric

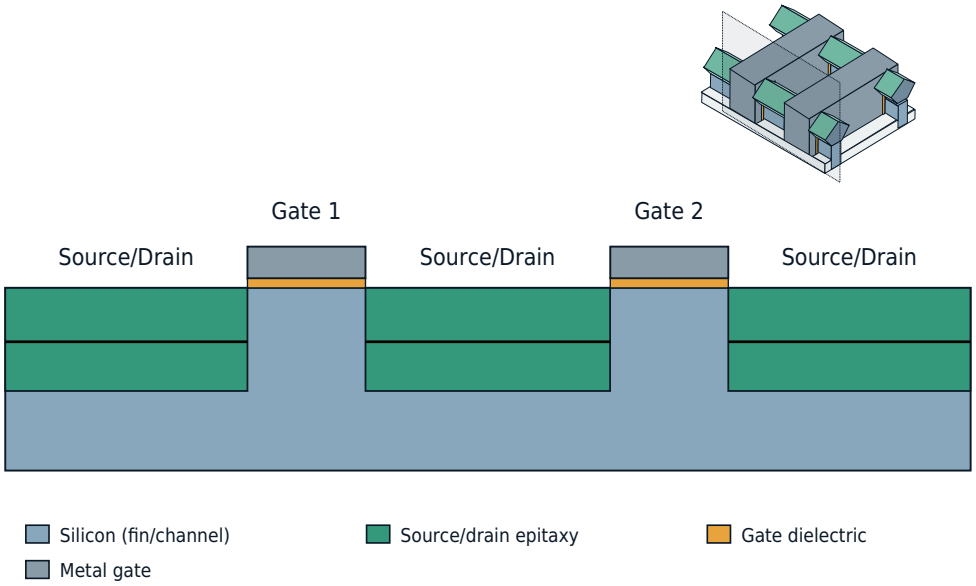
Metal gate Spacer

A final polishing step removes excess gate metal and electrically separates neighboring gates.

Which of a fin's two source/drain regions actually acts as source and which as drain is not a fixed property of the structure, but depends on the applied potential: for an n-channel transistor, the region at the lower potential takes on the role of the source, the one at the higher potential that of the drain (the reverse for p-channel transistors) – in every case, the channel is switched into conduction exclusively by the gate above it. This is why several gates can be placed independently along the same fin, as shown in the figure below: each gate controls only the channel segment directly beneath itself, regardless of whether a neighboring gate happens to be conducting or blocking at that moment. The shared source/drain region between two gates is, electrically, simply a node that acts as the drain of one transistor and the source of the other, depending on the switching state.

FinFET in longitudinal section

Cross-section along a fin – two gates with source/drain regions in between

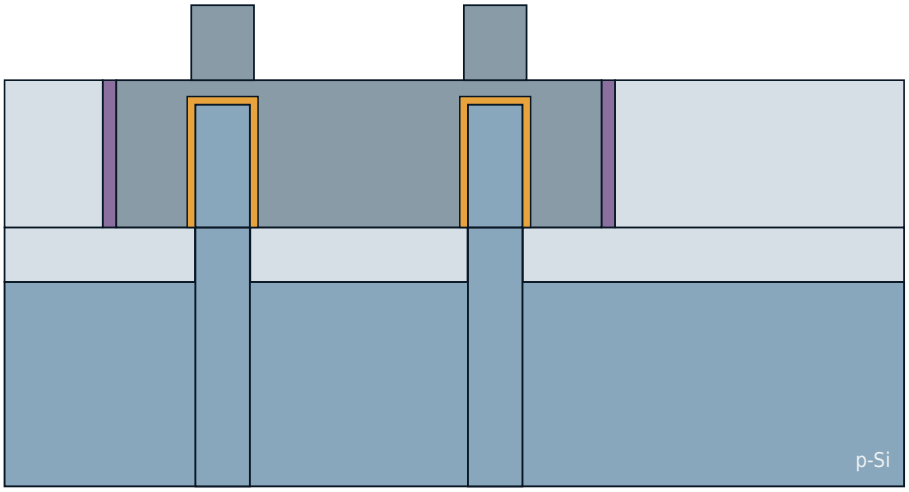


21. Contacts

Contact holes are etched down to the gate as well as source and drain and filled with metal – the FinFET is ready for electrical connection. The source/drain contacts lie outside the cross-sectional plane shown here.

Contacts

Finished FinFET structure



Silicon (p-Si) Oxide Gate dielectric Metal gate / contacts Spacer

Contact holes are etched down to gate, source, and drain and filled with metal – the FinFET is ready for electrical connection. (The source/drain contacts lie outside this cross-sectional plane.)

Since, in an SOI-based process, a full-area oxide layer is already present on the wafer, the electrical isolation of the channels from one another is automatically ensured, without the additional doping step described in step 8. In addition, the fin etch process is less problematic, since it can simply be stopped on the buried oxide instead of having to rely on a time-based etch.

General structure and mode of operation

The basic structure and mode of operation of a FinFET do not differ from those of a conventional MOS field-effect transistor. Here too there are source and drain terminals, through which the current flows. A gate electrode controls the transistor. In contrast to the classic, planar field-effect transistor, however, the channel between source and drain is formed as a three-dimensional structure on the silicon substrate, so that the gate electrode can enclose it from several sides. This enables considerably improved electrical behavior: leakage currents can be reduced and the drive current controlled more effectively.

The three-dimensional structure, however, also creates new parasitic capacitances and critical dimensions that have to be optimized. In a FinFET, the gate length is measured parallel to the channel, while the gate width

corresponds to twice the fin height plus the fin width. The height of the channel limits the drive current and the gate capacitance, while the width influences the breakdown voltage and short-channel effects, and the quantities that result from them, such as power consumption.

Context

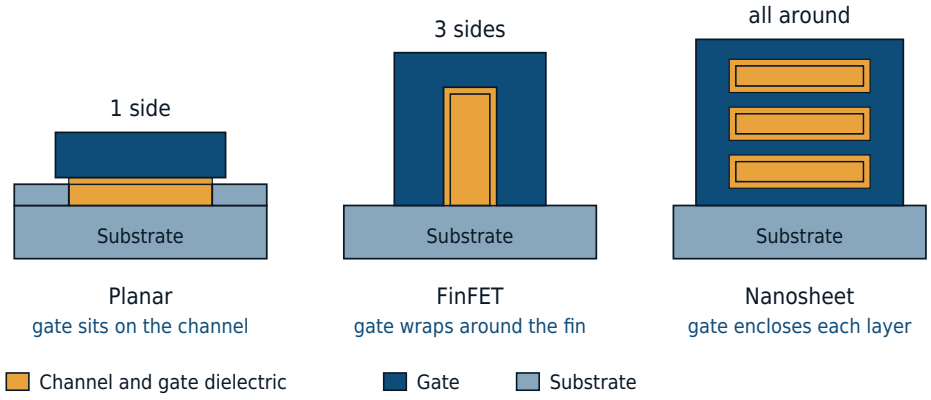
The FinFET is no longer a future technology; rather, it has defined an era that has now come to a close. After being researched for decades, it entered volume production for the first time in 2011 and remained the standard device architecture for all high-performance processors for more than a decade. At the smallest feature sizes, it has since been superseded itself: as the fin becomes ever narrower and taller, it becomes increasingly difficult to control mechanically, and the gate width can only be adjusted in discrete steps – it results from the number of fins and cannot be chosen continuously. Its successor solves both problems by laying the fin down (see From FinFET to nanosheet).

Below, the structure of a multigate transistor with three gates (tri-gate) is described for the bulk process.

From FinFET to nanosheet

The evolution from the planar transistor through the FinFET to the nanosheet follows a single idea: the gate should enclose the channel from as many sides as possible. The more completely it surrounds it, the more reliably the transistor switches off – and the shorter the channel can be made without current flowing through it uncontrolled.

Gate enclosure compared



Why the fin reached its limit

The FinFET solved the problem of the planar transistor, but brought two problems of its own.

First, a fin that has to become ever narrower while also growing taller becomes mechanically unstable. It stands freely on the substrate and has to survive several process steps without toppling over or breaking.

Second, the gate width can now only be adjusted in discrete jumps. In the planar transistor it was a freely chosen quantity in the design; in the FinFET it results from the number of fins. A transistor can therefore have two or three fins, but nothing in between. Anyone needing a somewhat stronger driver immediately gets a substantially stronger one – along with the corresponding area and power consumption.

Laying the fin down

The successor rotates the fin by 90 degrees: instead of a single vertical fin, several horizontal silicon layers are stacked on top of one another, each fully enclosed by the gate. Because the gate now surrounds the channel on all sides, it is called a gate-all-around transistor; based on the shape of the layers, it is also called a nanosheet transistor.

Both disadvantages of the fin disappear as a result. The layers are held by the gate and are not free-standing. And the gate width is once again continuously adjustable, since it results from the width of the layers – which can be freely chosen in the design, rather than counted out in fins.

How the layers are formed

The decisive trick lies in the fabrication. A stack of alternating silicon and silicon-germanium layers is grown epitaxially on the substrate. A fin is first etched out of this stack, just as with the FinFET. The silicon-germanium is then removed selectively – it dissolves in a chemistry that leaves the silicon untouched. What remains are the free-floating silicon layers, whose gaps are then filled with dielectric and gate metal.

The thickness of the layers and their spacing are therefore not determined by lithography, but by the layer thicknesses used during growth. The gap between two layers is only a few nanometers wide, and dielectric and metal have to reach completely into this gap – a task that cannot be accomplished without atomic layer deposition.

What comes next

The next step is already emerging and takes the idea to its logical conclusion: instead of placing n-channel and p-channel transistors side by side, they are meant to be stacked on top of one another. The resulting complementary FET saves area, since both transistors occupy the same footprint. The effort required is considerable, since two differently doped devices have to be built within a single, shared process flow.

DRAM Fabrication: The Trench Capacitor Process

General Structure and Operating Principle

The DRAM Memory Cell: One Transistor, One Capacitor

A DRAM cell (Dynamic Random Access Memory) stores a single bit as electrical charge on a capacitor. An access transistor connects this capacitor to the bitline on demand — while the transistor is off, the charge stays isolated and the bit remains stored; once switched on via the wordline, the charge can be read out or rewritten. This pairing of one transistor and one capacitor is called a 1T1C cell and is by far the most area-efficient way to store a bit — considerably more compact than the six transistors of an SRAM cell, but with one decisive drawback: the stored charge slowly leaks away and must be periodically refreshed every few milliseconds, hence the name "dynamic."

The challenge in cell design is purely geometric: a capacitor needs area to store enough charge for a reliably readable signal — typically on the order of a few femtofarads. Yet with every new technology node, the available cell area keeps

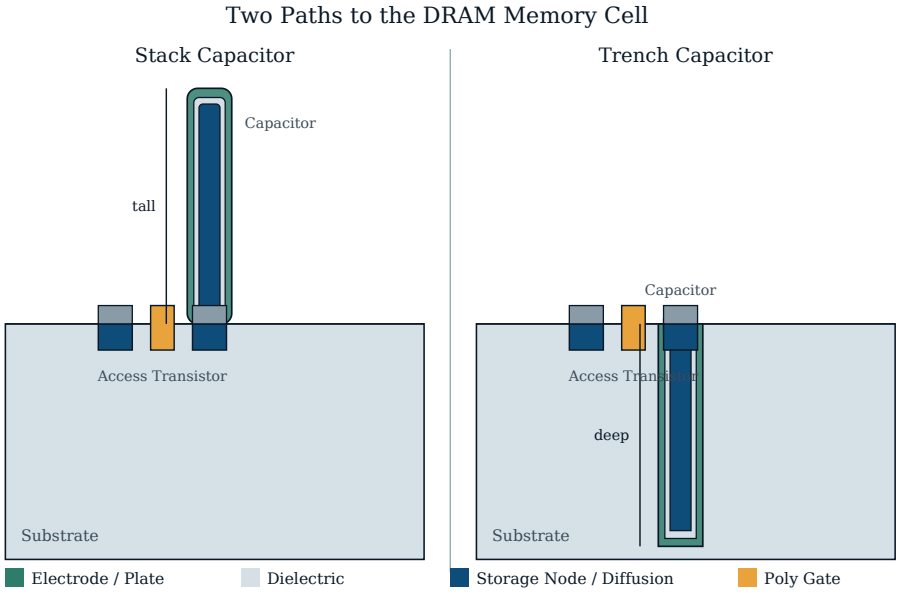
shrinking while the required capacitance stays nearly constant. The manufacturing industry has solved this problem in two fundamentally different ways.

Two Paths to the Same Goal: Stack versus Trench

The stack capacitor (dominant at Samsung, SK Hynix, and Micron) builds the needed area vertically above the access transistor: cylindrical or crown-shaped electrode structures rise above the wafer surface and can, within the limits of lithography and etch technology, grow nearly arbitrarily tall.

The trench capacitor (historically developed by IBM and used at, among others, Infineon/Qimonda) instead pushes the same area downward: a deep, narrow trench is etched several micrometers into the substrate, with the capacitor electrodes forming concentric cylindrical surfaces along its walls. The key advantage of this approach lies in the process sequence: the trench — the most demanding and critical part of fabrication — is formed before the access transistor, so the transistor itself is built on top of an almost planar wafer surface, without having to overcome the tall topography of an already-completed stack capacitor.

The diagram below compares both capacitor types in cross-section, before the following sections build up the complete trench process step by step.



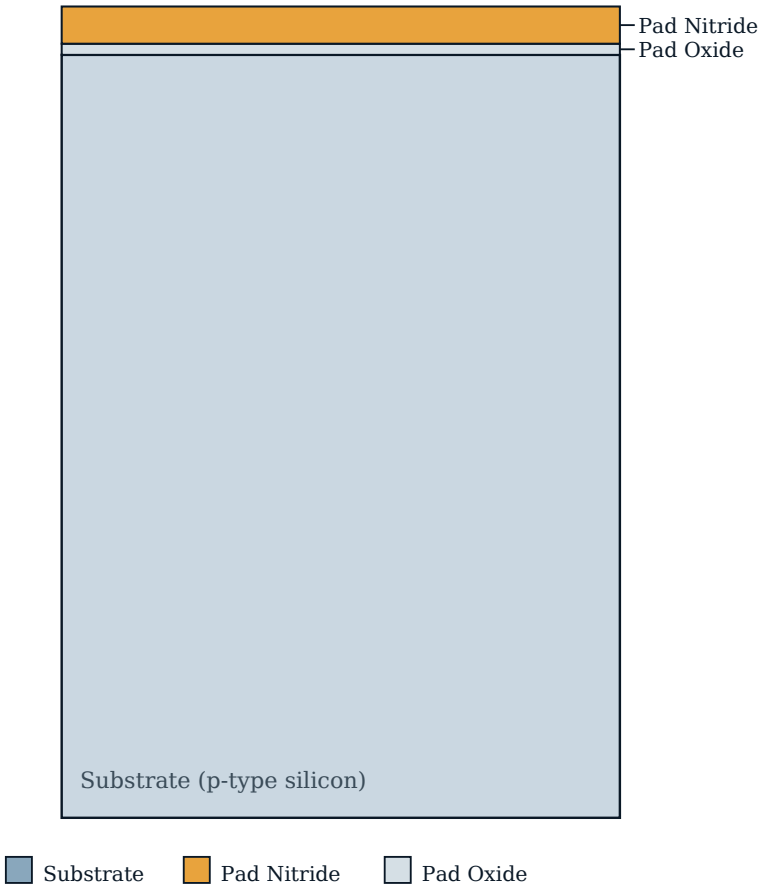
Trench Capacitor Fabrication

From Flat Wafer to Deep Trench

The trench capacitor process begins before a single transistor exists — the trench is etched as the very first structure into the still completely unprocessed wafer. This ordering is no coincidence: the most demanding, most critical part of the entire cell fabrication is meant to take place on a perfectly flat surface, unobstructed by any transistor or interconnect structures that would otherwise already be present.

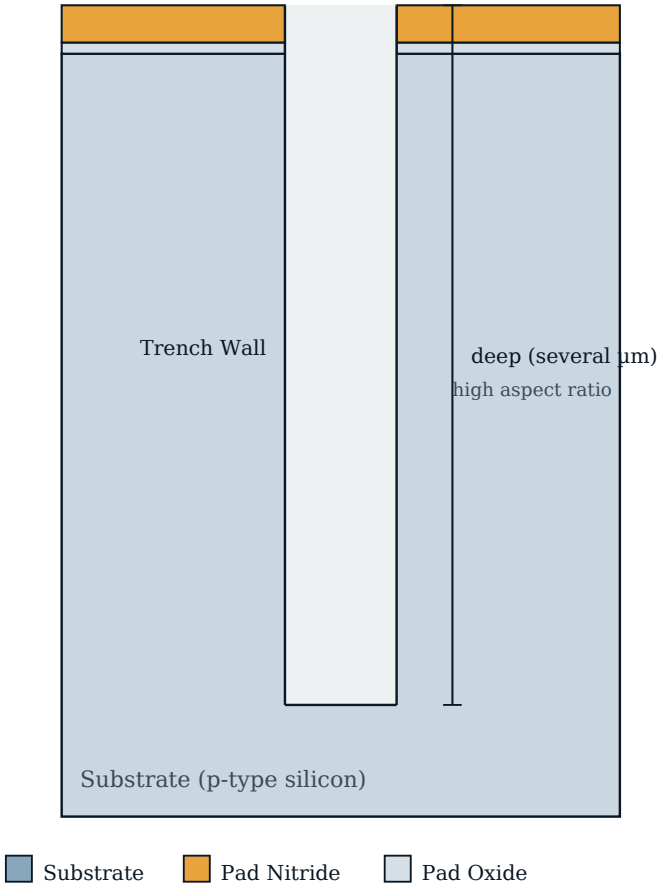
Step 1 – Pad Oxide and Pad Nitride: A thin thermal oxide layer (pad oxide) is first grown on the p-type silicon substrate, acting as a buffer that relieves mechanical stress between the substrate and the nitride layer that follows. Directly above it, a considerably thicker layer of silicon nitride (pad nitride) is deposited. This layer stack serves as the hard mask for the trench etch in the next step, and must be patterned to expose an opening only at the intended trench location.

Step 1: Pad Oxide and Pad Nitride



Step 2 – Deep Trench Etch: Through the opened hard mask, a highly anisotropic RIE process etches a narrow trench several micrometers deep into the substrate. The aspect ratio (depth to width) in modern trench cells is typically 50:1 or higher — one of the most demanding etch tasks in all of semiconductor manufacturing, since the sidewalls must remain uniformly vertical and smooth across the entire depth.

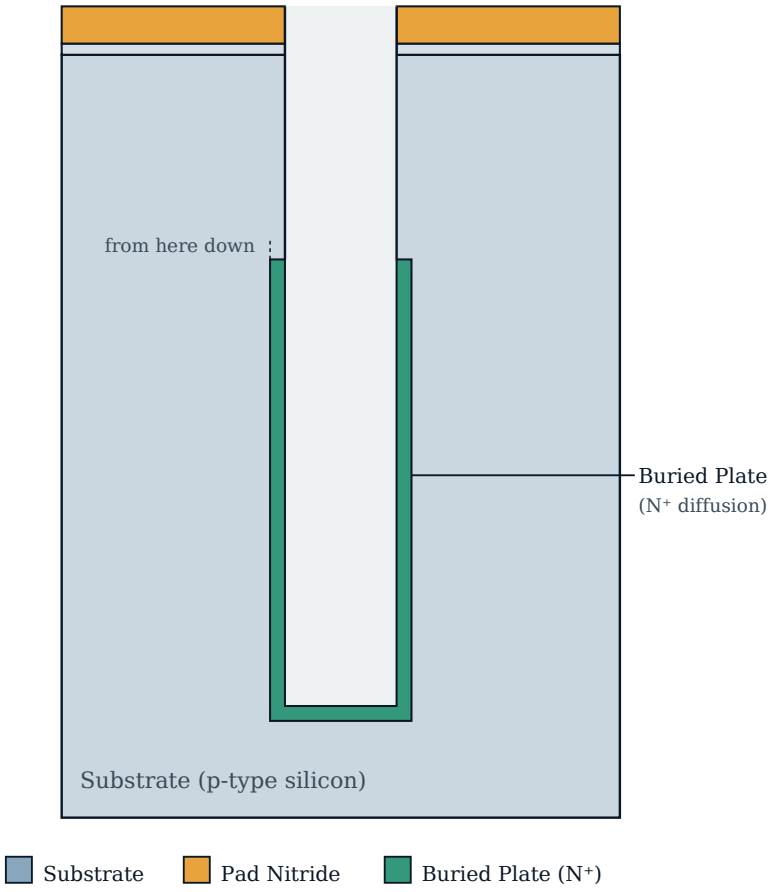
Step 2: Deep Trench Etch



The Capacitor Electrodes Take Shape

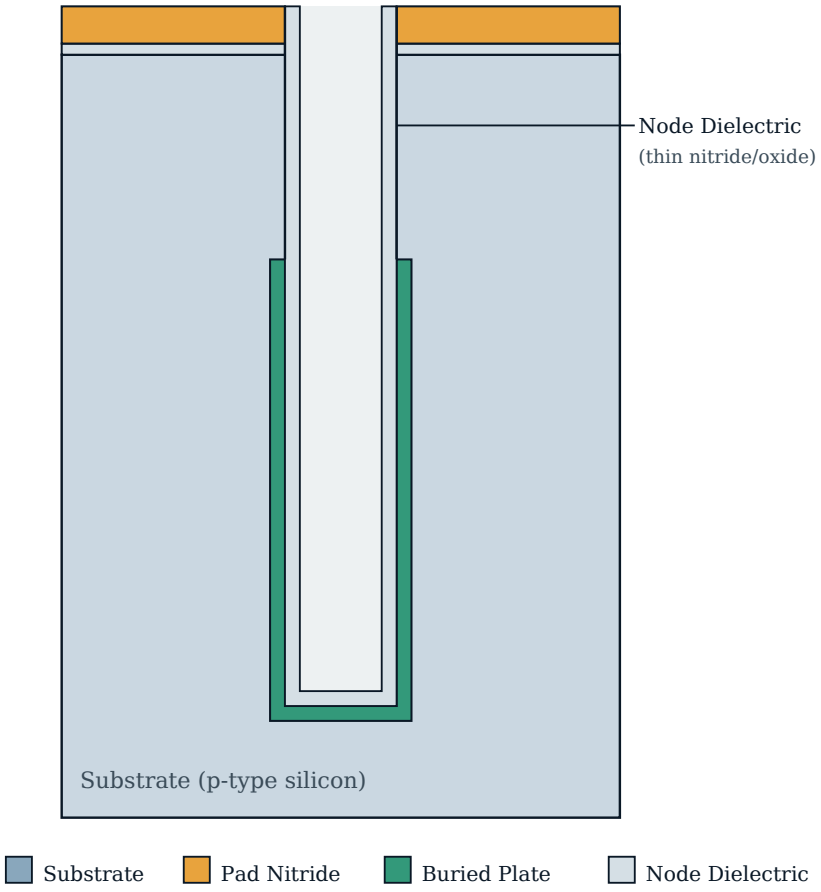
Step 3 – Buried Plate: The lower, deeper region of the trench wall is coated with an arsenic-doped glass layer, from which the dopant drives into the adjacent silicon wall at elevated temperature. This forms the buried plate — a heavily n-doped zone that serves as the outer capacitor electrode, shared in common by every cell in an array (comparable to the "ground plate" of a classic capacitor).

Step 3: Buried Plate



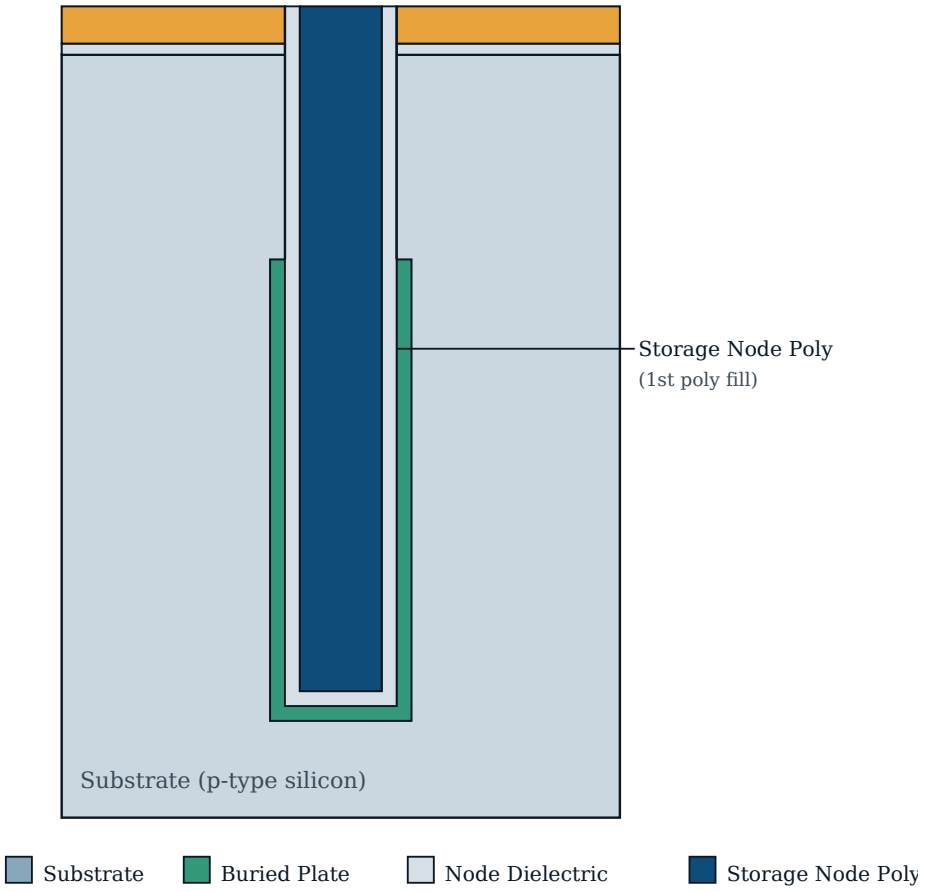
Step 4 – Node Dielectric: A conformal dielectric layer only a few nanometers thick — usually a nitride-oxide stack — is now deposited across the full trench depth. It separates the inner electrode that follows in the next step from the buried plate, and together with the trench wall area, largely determines the achievable capacitance of the cell.

Step 4: Node Dielectric



Step 5 – First Poly Fill: Doped polysilicon completely fills the remaining trench. This "storage node poly," together with the node dielectric and the buried plate, forms the actual capacitor: the charge stored on the poly represents the state of the stored bit.

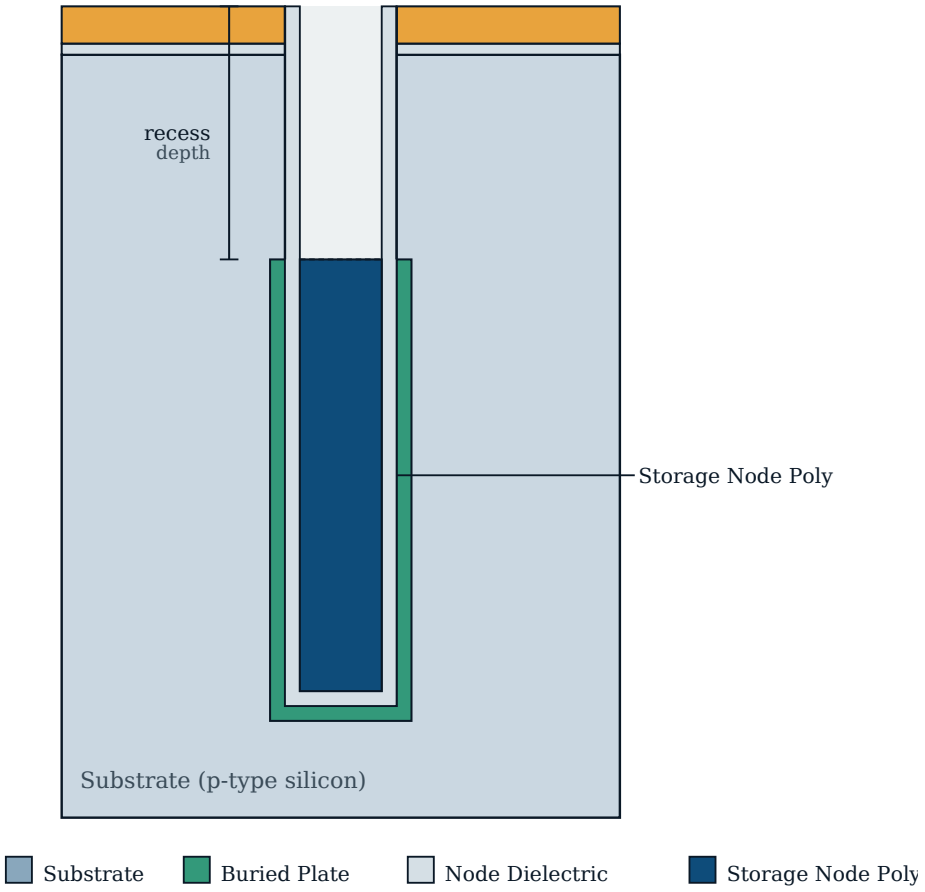
Step 5: First Poly Fill



The Upper Trench Region: Collar and Buried Strap

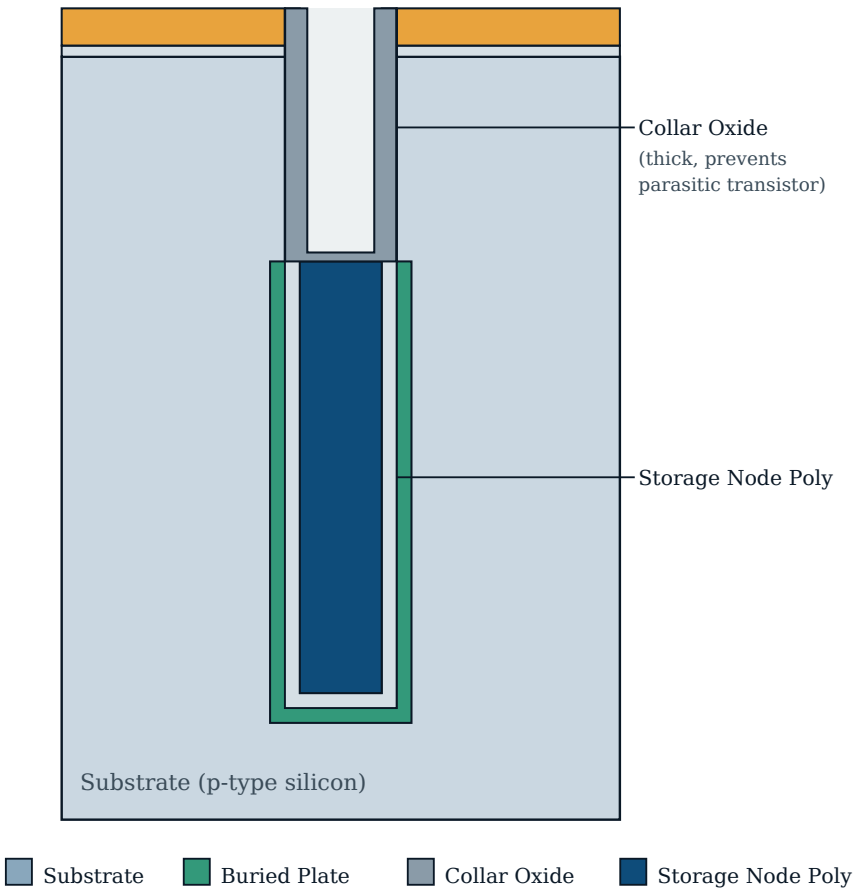
Step 6 – Poly Recess Etch: The storage node poly is selectively etched back so that it fills only the lower trench section surrounded by the buried plate. The upper trench region — where the access transistor will later sit directly beside the trench — is once again left open.

Step 6: Poly Recess Etch



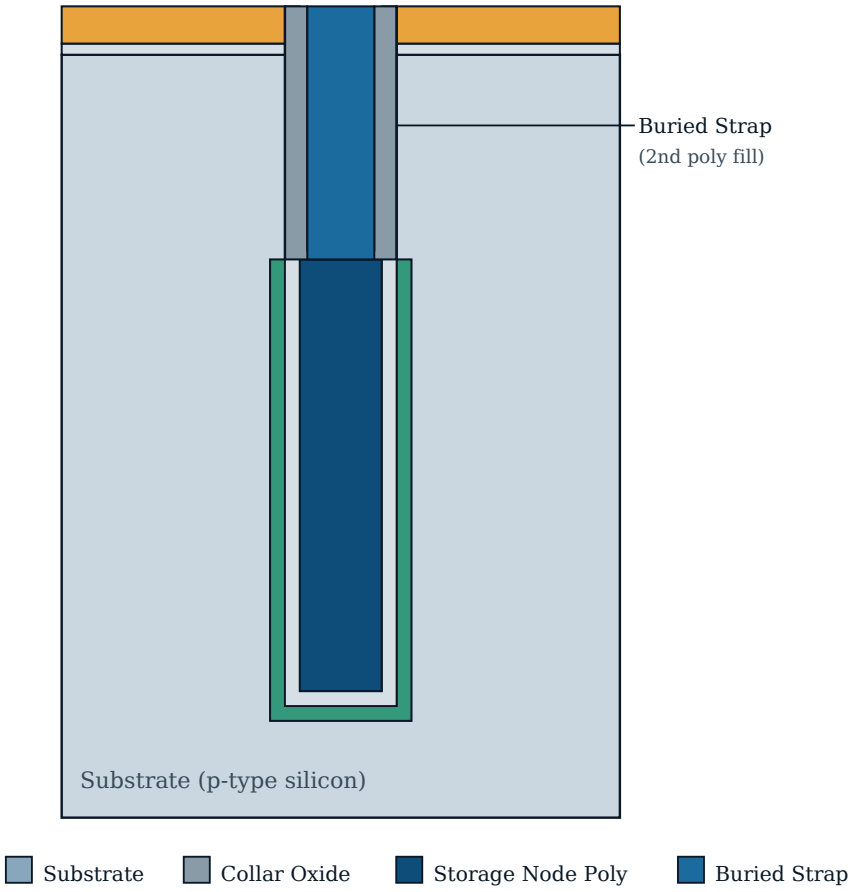
Step 7 – Collar Oxide: In the now-exposed upper trench section, the thin node dielectric is removed and replaced with a considerably thicker oxide layer, the collar. Without this thick oxide, the later wordline voltage could turn on a parasitic, unwanted transistor along the upper trench wall, letting charge leak past the actual access transistor. The collar reliably suppresses this leakage path.

Step 7: Collar Oxide



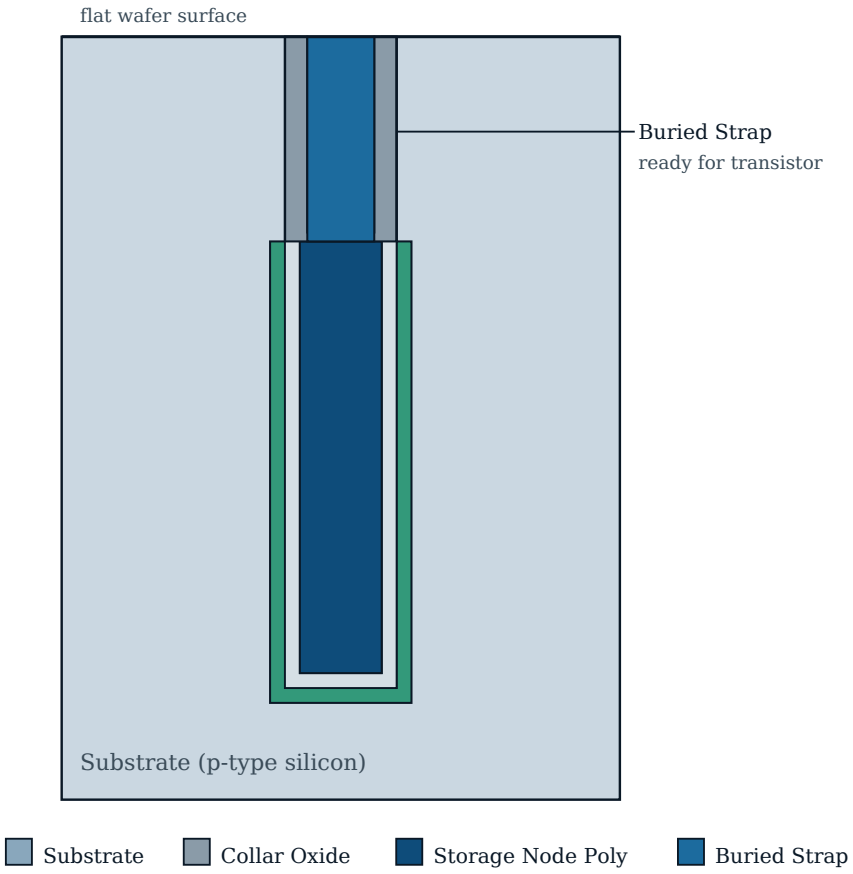
Step 8 – Buried Strap: A second polysilicon deposition completely fills the upper trench section, now narrowed by the collar. This so-called buried strap forms the electrical connection between the deep-lying storage node poly and the access transistor's future source/drain diffusion at the wafer surface — without it, the capacitor would remain electrically isolated from the transistor.

Step 8: Buried Strap



Step 9 – Planarization: A final CMP and etch step removes excess poly along with the entire pad-nitride/pad-oxide hard mask, restoring an almost flat wafer surface. The trench capacitor is now fully complete and, sitting flush with the substrate surface, awaits the access transistor — whose fabrication the next section describes.

Step 9: Planarization



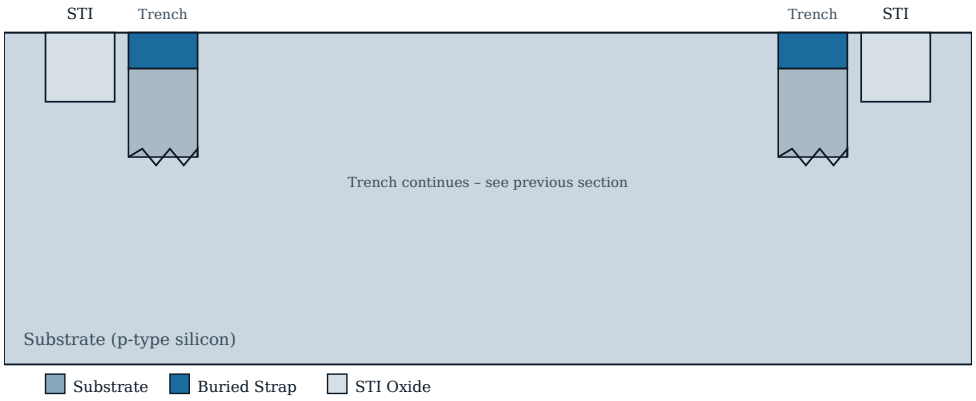
Access Transistor and Interconnect Fabrication

Two Cells, One Shared Bitline Contact

With the finished, planarized trench capacitor from the previous section, fabrication can continue where an ordinary logic process begins: on an almost flat wafer surface. The diagrams below show a typical cell pair — two neighboring trench capacitors whose access transistors share a common bitline contact in the middle. This "folded" arrangement saves one contact per two cells and is standard practice in trench DRAM fabrication.

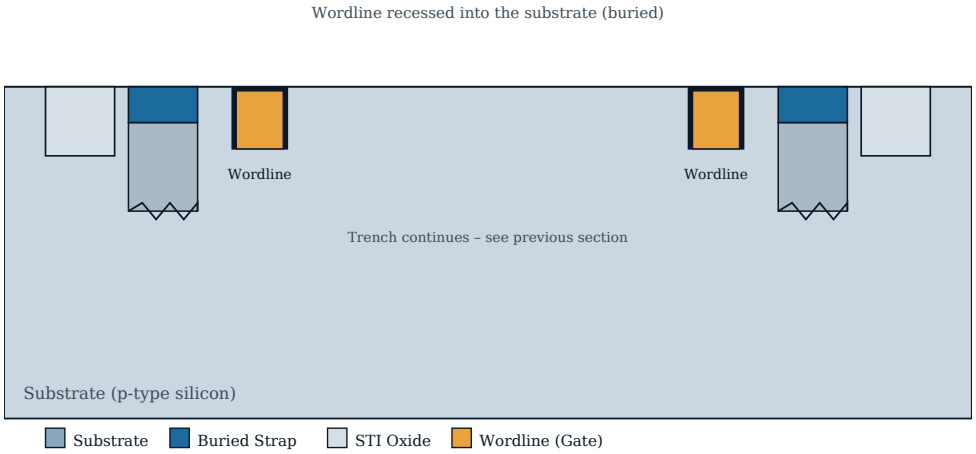
Step 10 – Shallow Trench Isolation (STI): At the outer edge of each cell pair, a shallow trench isolation separates that pair's active area from the next one's. Unlike the deep capacitor trench, the STI is only a few hundred nanometers deep and purely oxide-filled — it has no electrical function beyond separating neighboring devices.

Step 10: Shallow Trench Isolation (STI)



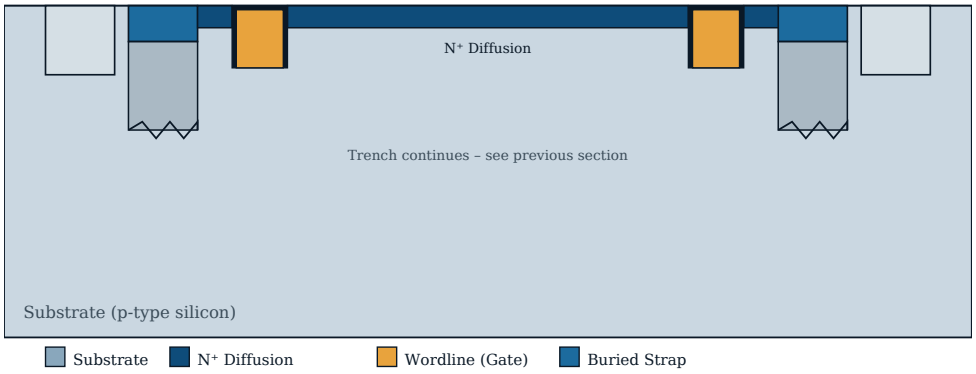
Step 11 – Gate Oxide and Buried Wordline: Between each trench and the future bitline contact, the access transistor is formed. Modern trench cells recess the wordline into a shallow pit in the substrate as a so-called buried wordline, lined with a thin gate oxide. This recessed structure does not shorten the effective channel while allowing a considerably more compact cell than a classic, surface-sitting gate electrode.

Step 11: Gate Oxide and Buried Wordline



Step 12 – Source/Drain Implantation: A shallow N^+ implant connects the active area continuously: from the left trench's buried strap, across the left wordline's channel to the center, onward across the right wordline's channel, to the right trench's buried strap. Each wordline thus switches the current path between "its" trench capacitor and the shared center node.

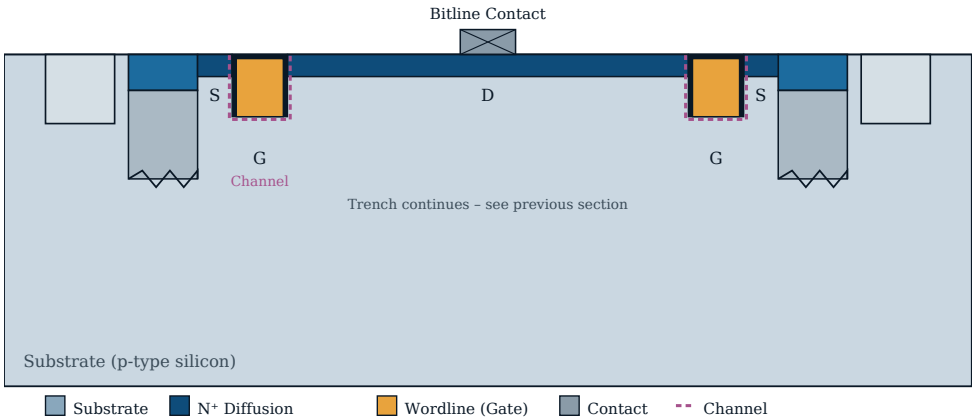
Step 12: Source/Drain Implantation



Connecting to the Interconnect Layers

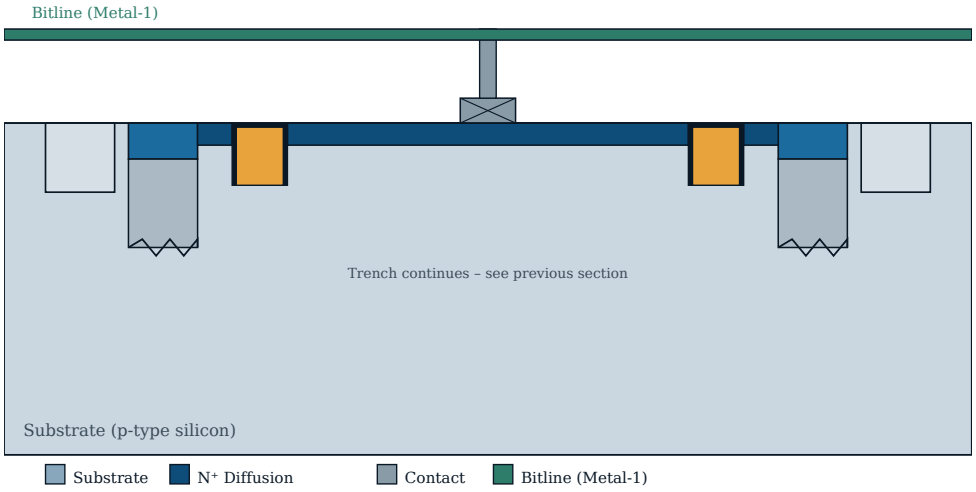
Step 13 – Bitline Contact: A contact hole is opened on the shared diffusion at the cell center and filled with metal. Through this single contact, both access transistors of the pair access the bitline together — hence the "folded" cell arrangement. The diagram uses this point to show, as an example, all three transistor terminals (G, S, D) along with the channel path: with a buried wordline, the channel does not form flat beneath the gate but wraps in a U-shape along the trench walls and floor — effectively lengthening the current path despite the compact cell area. Source and drain are shown here as fixed for clarity; electrically, both diffusion sides are symmetric and swap roles depending on the direction of operation.

Step 13: Bitline Contact



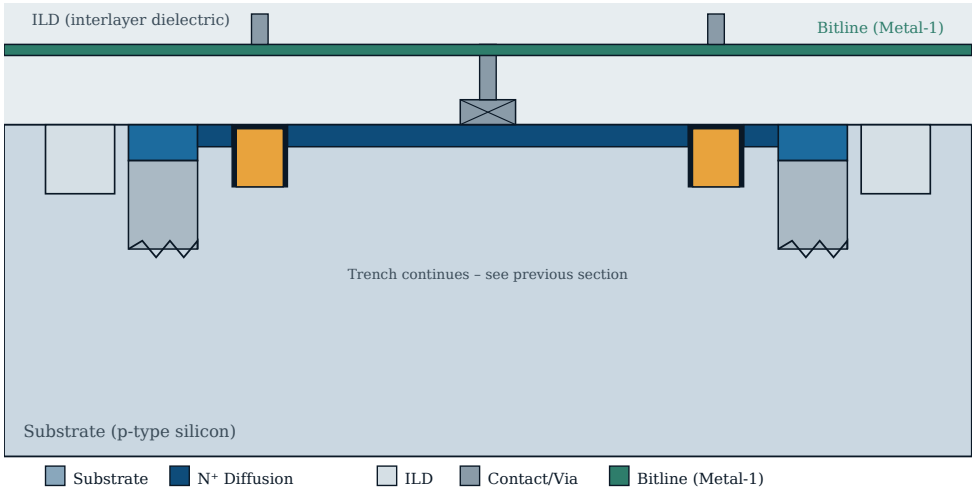
Step 14 – Bitline Metallization: The first metal layer (metal-1) wires the bitline contact into a continuous trace that runs across the entire cell array, connecting thousands of cell pairs in the same column. When reading a cell, a sense amplifier compares this bitline's voltage against a reference voltage to determine whether a "1" or "0" was stored.

Step 14: Bitline Metallization



Step 15 – Interlayer Dielectric and Contacts: An interlayer dielectric (ILD) embeds the bitline wiring while also providing the insulation for additional contacts — for instance the wordline connections outside the actual cell array, where the wordlines terminate in driver circuits. From here on, fabrication follows the same principles as any other logic process: further metal layers, vias, and contacts, as already described in the "Circuit Layouts" chapter. This completes the DRAM cell — from the deep capacitor trench to the connection into the first interconnect layer.

Step 15: Interlayer Dielectric and Contacts



Micromechanics

Introduction to MEMS

What Is a MEMS?

MEMS (Micro-Electro-Mechanical Systems) refers to devices that combine mechanical and electrical functions on the same chip – sensors that convert motion, pressure, or sound into electrical signals, or actuators that translate electrical signals into motion. Unlike classical integrated circuits, which work exclusively with charge carriers, MEMS contain movable mechanical structures: free-standing beams, membranes, comb structures, or springs that deflect in the nanometer-to-micrometer range.

Fabrication uses the same basic tools as classical semiconductor technology – lithography, etching, deposition – but goes beyond their usual purpose: while every layer in a transistor is meant to stay permanently bonded to the next, MEMS structures must be deliberately released so they can move. This requires additional concepts such as sacrificial layers and deep, highly anisotropic etching, which play no role in classical fabrication.

Applications are everywhere: accelerometers in smartphones (screen rotation), pressure sensors in tire-pressure monitoring systems, microphones in practically every mobile device, and micromirrors in projectors.

Fabrication: Mechanical, Not Just Electrical

Fabrication uses the same basic tools as classical semiconductor technology – lithography, etching, deposition – but goes beyond their usual purpose: while every layer in a transistor is meant to stay permanently bonded to the next, MEMS structures must be deliberately released so they can move. This requires additional concepts such as sacrificial layers and deep, highly anisotropic etching, which play no role in classical fabrication.

Application Examples

Applications are everywhere: accelerometers in smartphones (screen rotation), pressure sensors in tire-pressure monitoring systems, microphones in practically every mobile device, and micromirrors in projectors.

Anisotropic Etching of Silicon

Crystal Orientation and Etch Rates

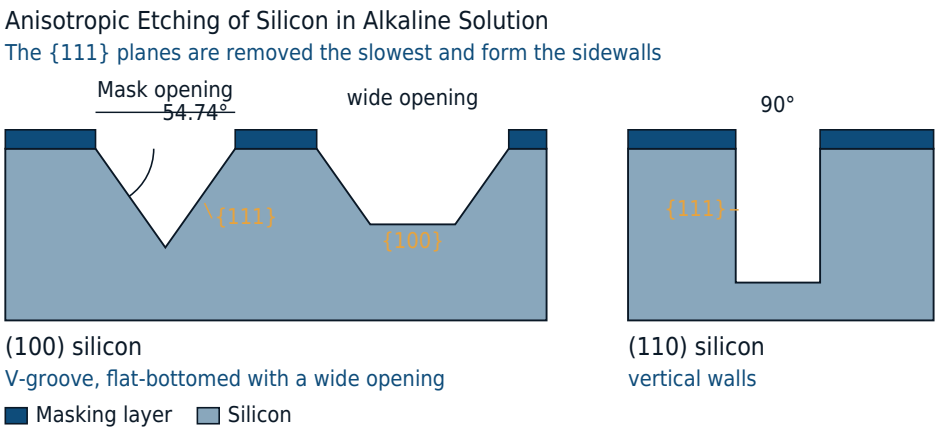
Single-crystal silicon does not etch equally fast in all directions in alkaline solutions such as potassium hydroxide (KOH) or TMAH (tetramethylammonium hydroxide) – this is known as anisotropic etching. The cause lies in the crystal structure: silicon crystallizes in a diamond lattice, and the {111} crystal planes are packed considerably more densely with atoms than the {100} or {110} planes. Because more densely packed planes offer fewer freely accessible bonds to the etchant, they are removed much more slowly – typical etch-rate ratios between {100} and {111} are 100:1 or higher.

The {111} planes therefore form the sidewalls of the etched structure regardless of etch duration: once a {111} plane is exposed, the etch attack there effectively stalls, while {100} or {110} surfaces continue to be removed.

The Characteristic 54.74° Angle

On (100)-oriented wafers – the most common cut in semiconductor manufacturing – the wafer surface lies parallel to the {100} plane. The {111} planes intersect this surface at a fixed angle of 54.74°, dictated by the crystal geometry. Opening the masking layer in a rectangle and etching through it therefore does not produce a vertical trench, but a V-shaped groove with sloped sidewalls inclined at exactly 54.74°.

If the mask opening is wide enough and the etch time long enough, the two {111} sidewalls do not meet at a point; instead, the etch first reaches a flat {100} bottom surface – producing a flat-bottomed trapezoid rather than a pointed V-groove. With a narrow opening or a short etch time, the V-shape persists.



(110) Silicon: Vertical Walls

Using a (110)-oriented wafer instead, whose surface lies parallel to the {110} plane, changes the geometry fundamentally: here the {111} planes stand perpendicular to the wafer surface. A rectangular mask opening therefore produces a trench with exactly vertical (90°) sidewalls instead of the sloped 54.74° flanks seen on a (100) wafer.

This property makes (110) silicon attractive for applications that need high aspect ratios while still relying on simple wet etching – for example, narrow, deep trenches that would require significantly more area on a (100) wafer due to the sloped walls. The drawback: (110) wafers are less common and correspondingly more expensive and harder to source than the standard (100) wafer.