Semiconductor Technology from A to Z

Deposition

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:

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

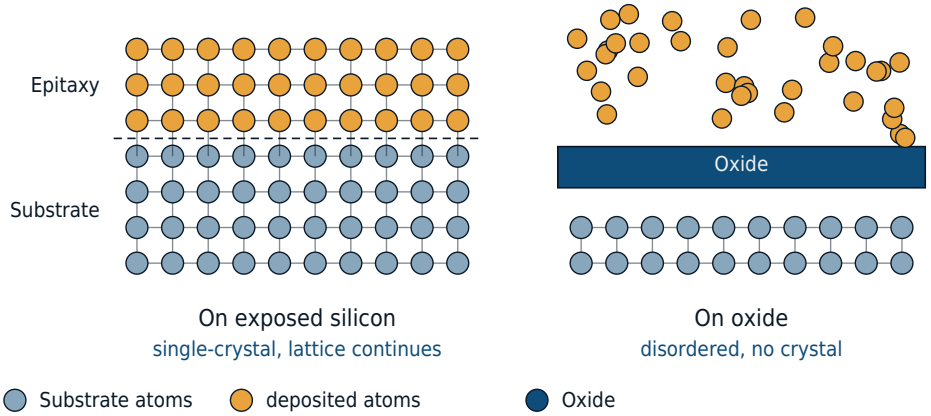
In order for the silicon on the substrate to adopt the crystal structure, the surface must be absolutely clean; this makes use of the equilibrium reaction. Both reactions can also proceed in the opposite direction, depending on the ratio of the gases used. If there is only little hydrogen present in 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.

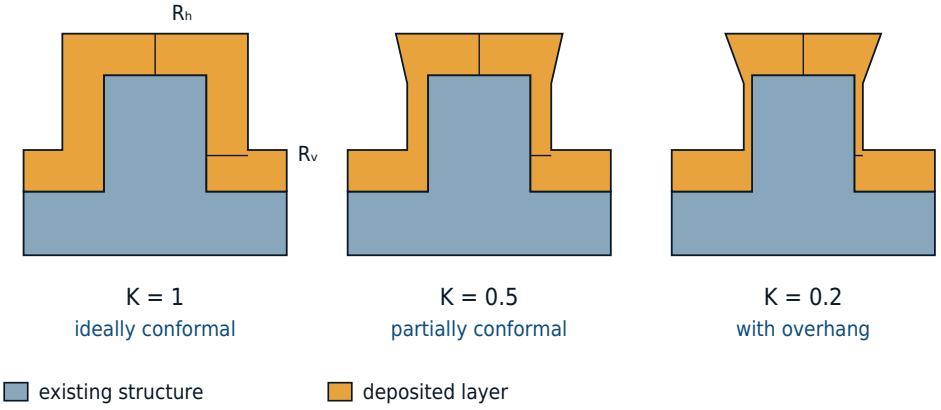
CVD process: Chemical Vapor Deposition

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.

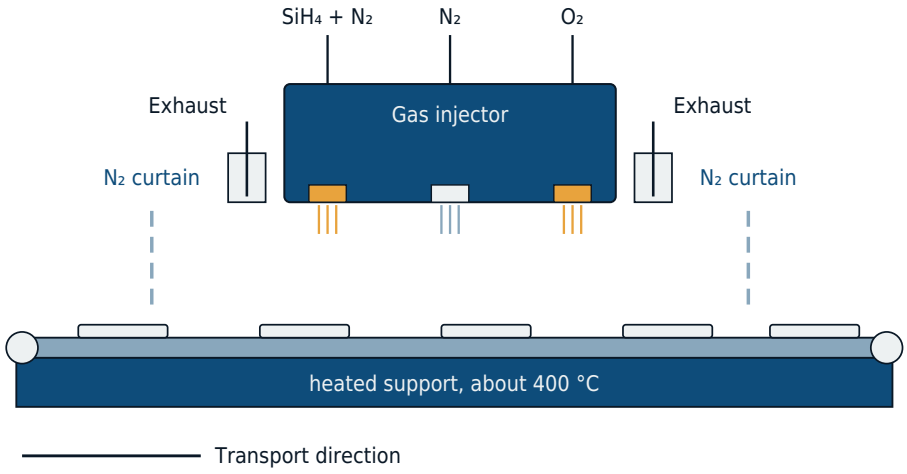
Analogous to phosphosilicate glass, boron can also be added at the same time (borophosphosilicate glass, BPSG, 4 % B and 4 % P).

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:

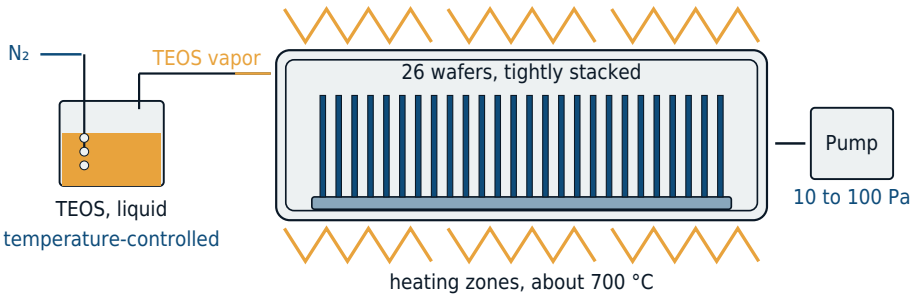
a) Si_3N_4 (850 °C): $4NH_3 + 3SiH_2Cl_2 \rightarrow Si_3N_4 + 6HCl + 6H_2$
b) SiON (900 °C): $NH_3 + SiH_2Cl_2 + N_2O \rightarrow SiO_xN_y$ + byproducts
c) SiO_2 (700 °C): $SiO_4C_8H_{20} \rightarrow SiO_2$ + byproducts
d) Tungsten (400 °C): $WF_6 + 3H_2 \rightarrow W + 6HF$

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

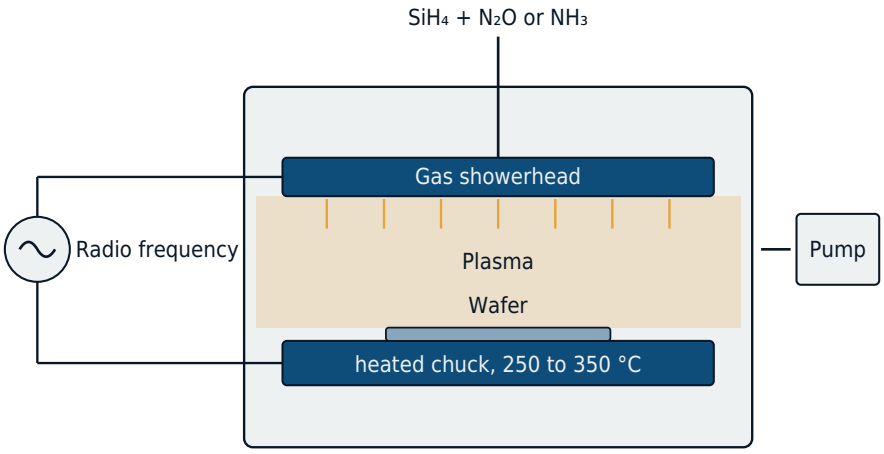
The plasma-enhanced CVD process takes place at 250–350 °C. Since the temperature is too low to decompose the gases, a high-frequency voltage is used to bring the gas into the plasma state. In the plasma state it is very energetic

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

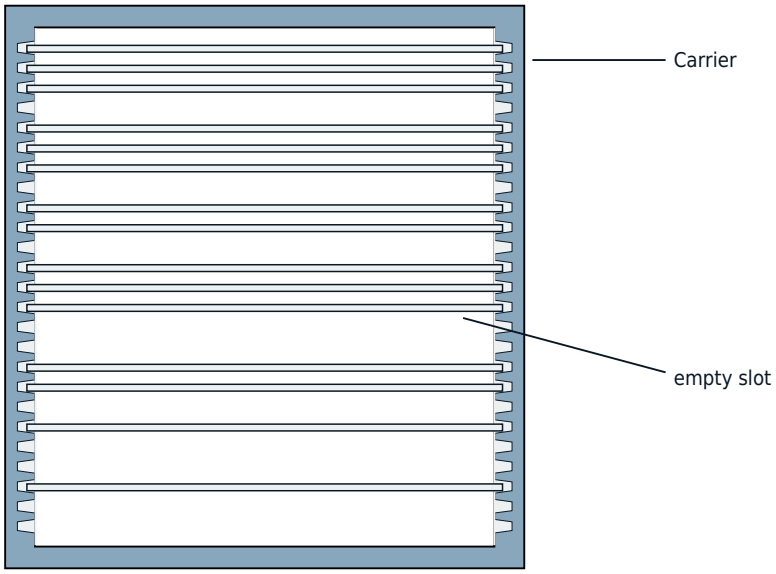
The plasma supplies the energy that would otherwise have to come from temperature



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:

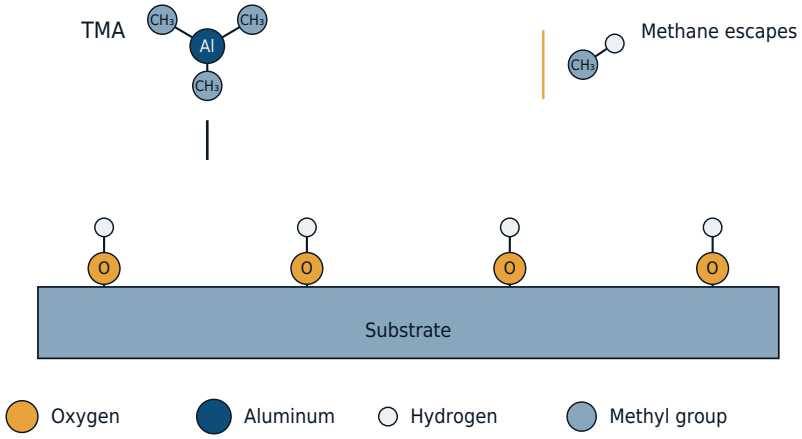
1. self-limiting reaction with the first gas, which saturates the surface atoms
2. purge step with an inert gas
3. self-limiting reaction with a second gas, which reacts at the new surface
4. purge step with an inert gas

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

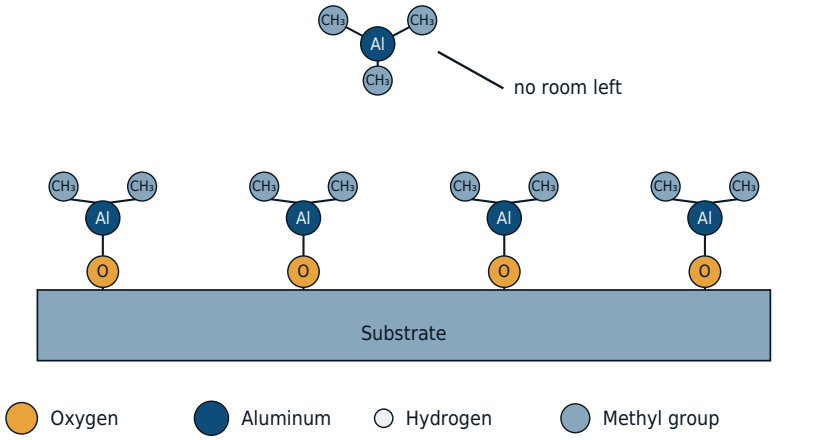


Each methyl group binds one hydrogen atom and leaves the surface as methane.

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

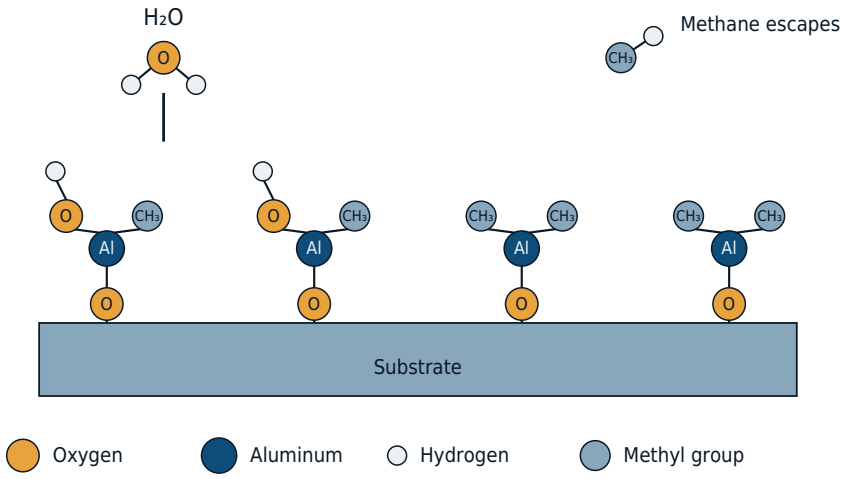


This is exactly what makes the process so precise: the layer grows only one layer at a time.

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

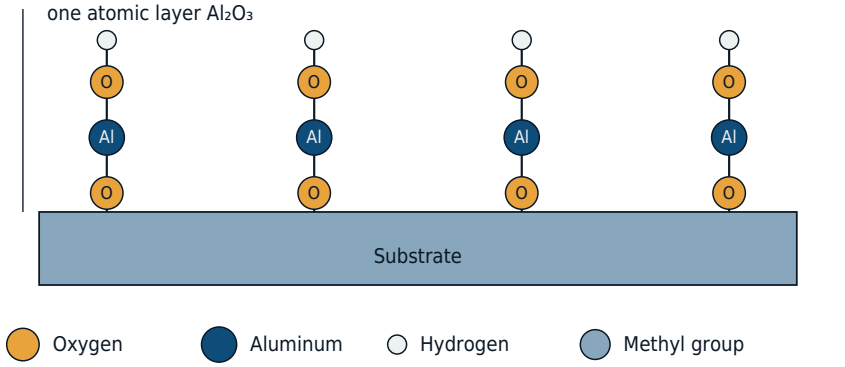


One hydrogen from the water binds the methyl group, leaving the oxygen behind.

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.

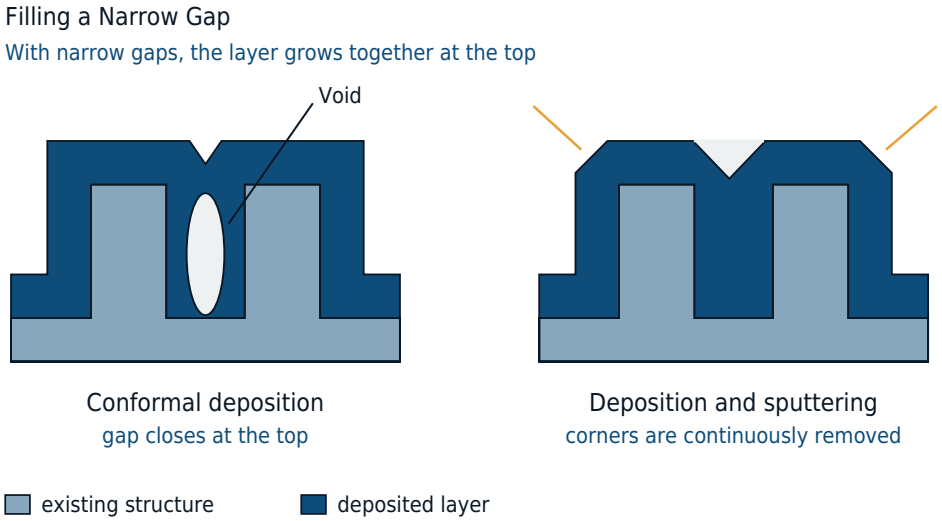
Why a void forms

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.