Semiconductor Technology from A to Z

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.

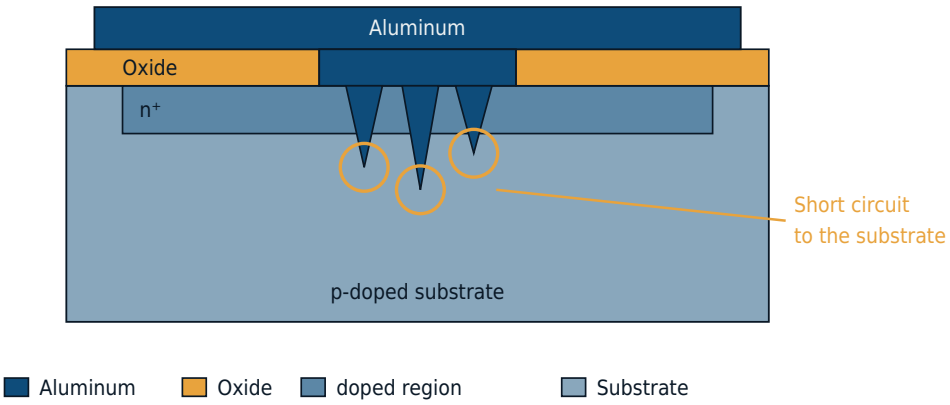
However, aluminum meets the requirements for electrical load capacity and corrosion resistance only partially. The following sections describe the three effects at which aluminum technology reaches its limits: the diffusion of silicon into the metal, electromigration, and the growth of hillocks. Metals 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

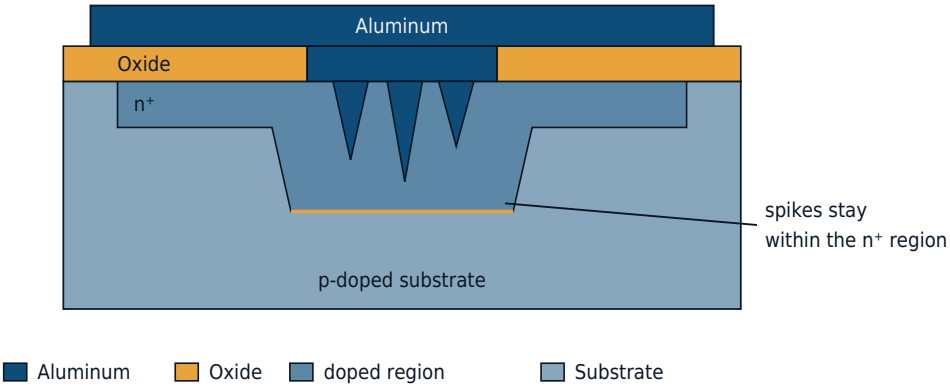


During sintering, silicon dissolves into the aluminum, leaving pits that the metal fills in. If the spikes reach deeper than the doped region, the junction is short-circuited.

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.

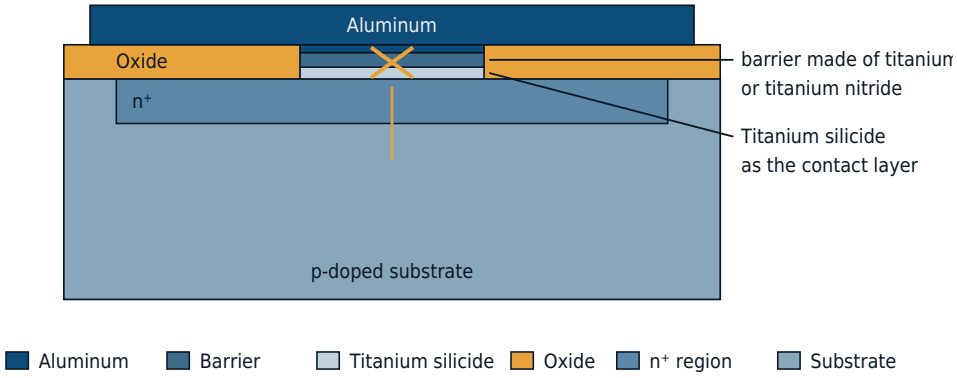
The disadvantage, however, is that an additional process step has to be introduced, and the electrical properties change due to the enlargement of the doped region.

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.

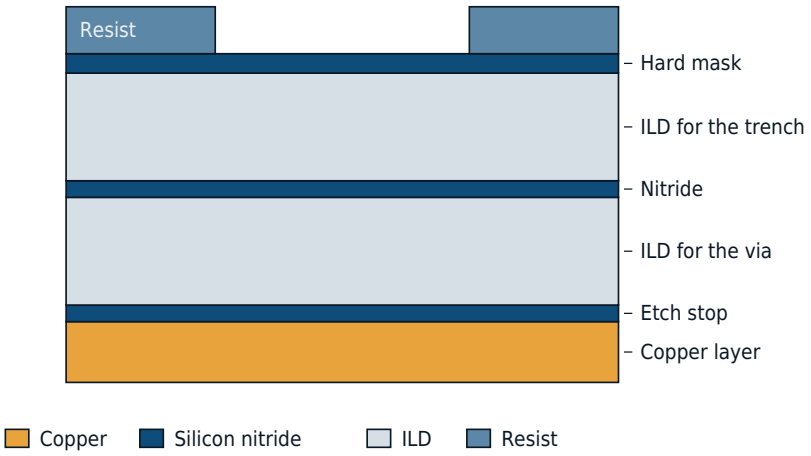
Trench First VIA Last

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

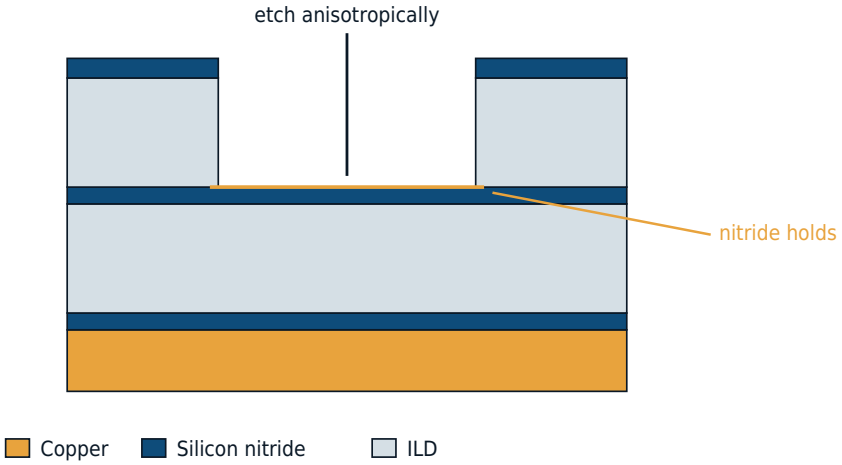


Two roughly equally thick ILD levels are stacked, separated by a nitride layer: the via will later sit in the lower one, the trench in the upper one. The hard mask protects the ILD during resist removal and serves as the CMP stop.

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.

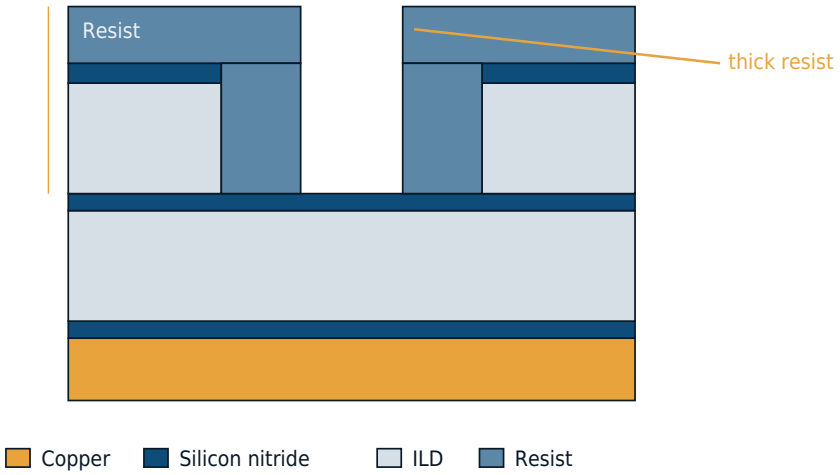
The hard mask at the surface is needed to protect the ILD layer from the plasma during resist removal. This is necessary because the ILD is chemically similar in composition to the photoresist and can therefore be

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

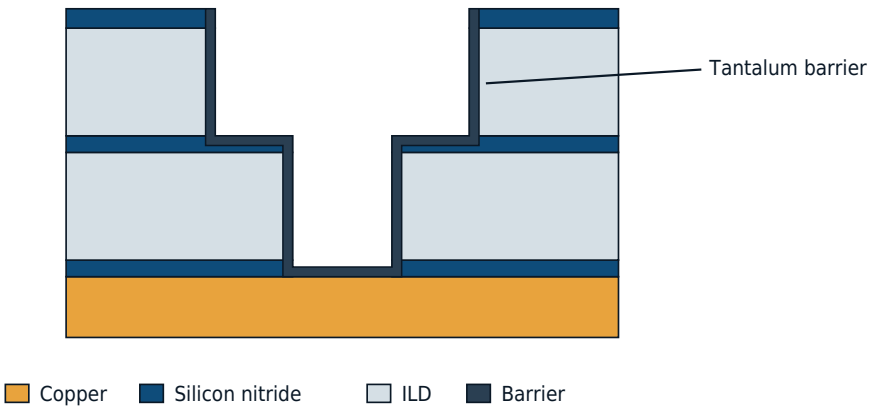


Because the resist also fills the trench, it is noticeably thicker here than in the first step. Exposing a small contact hole through it is the actual drawback of this process.

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

4. Etch the Via and Deposit the Barrier

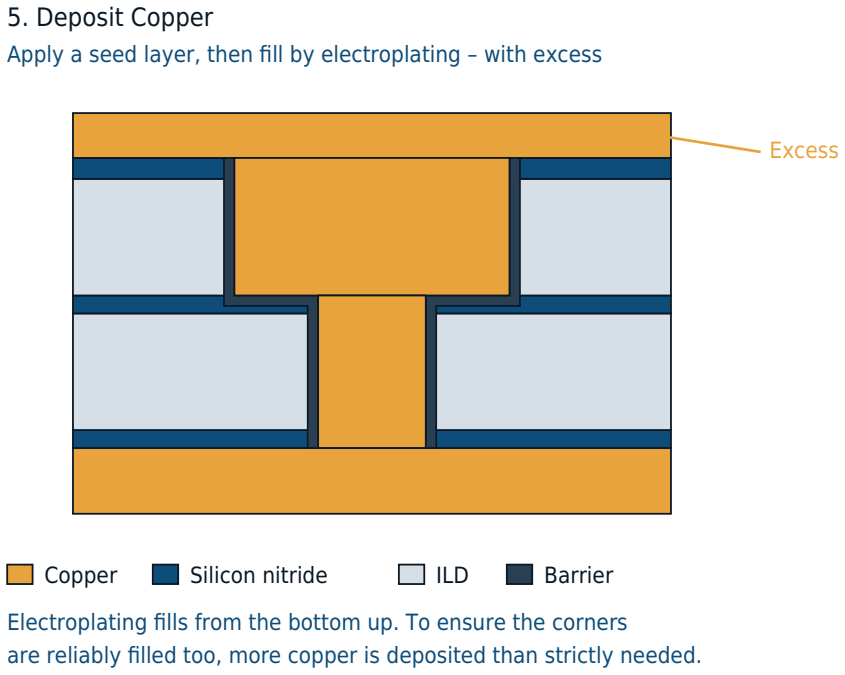
Through nitride, lower ILD, and etch stop, down to the copper



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.

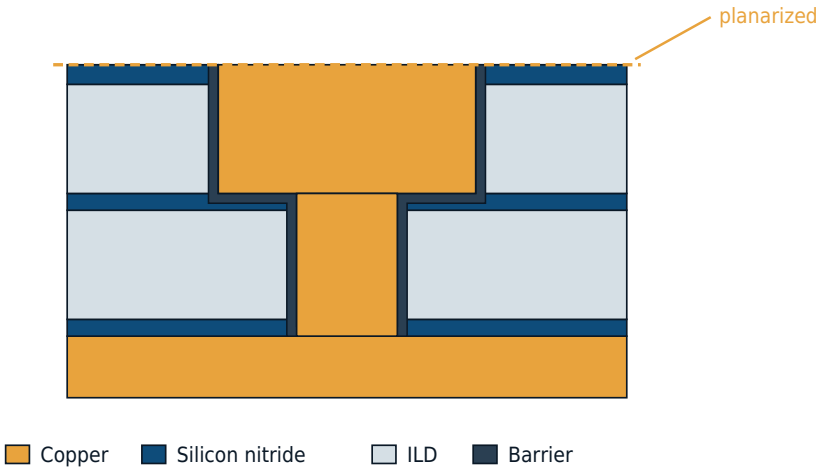
5. A thin copper seed layer is deposited, and the structures are filled with copper in an electroplating process.



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.

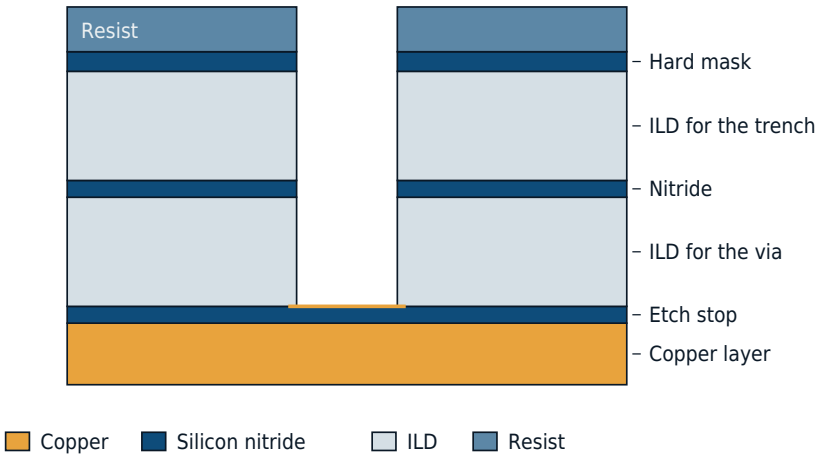
VIA First Trench Last

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

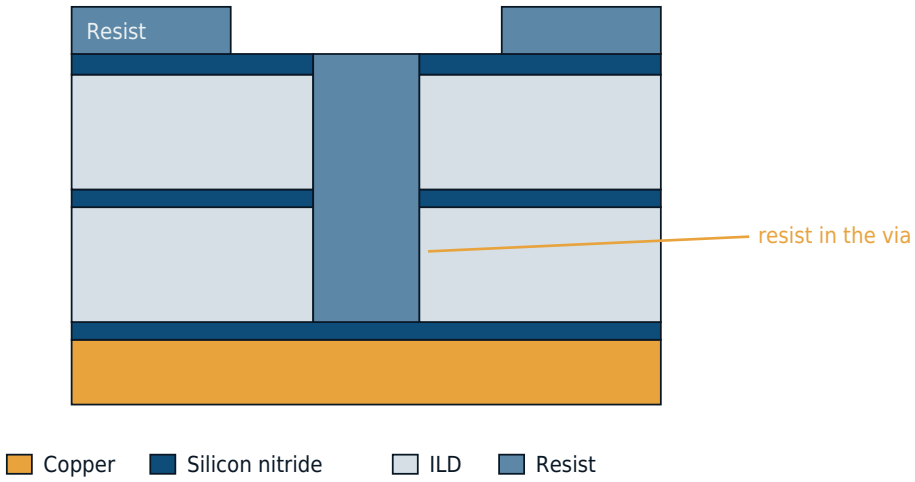


The hole goes straight through both ILD levels. Its lower part later becomes the via, while the upper part is widened into the trench in the next step.

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

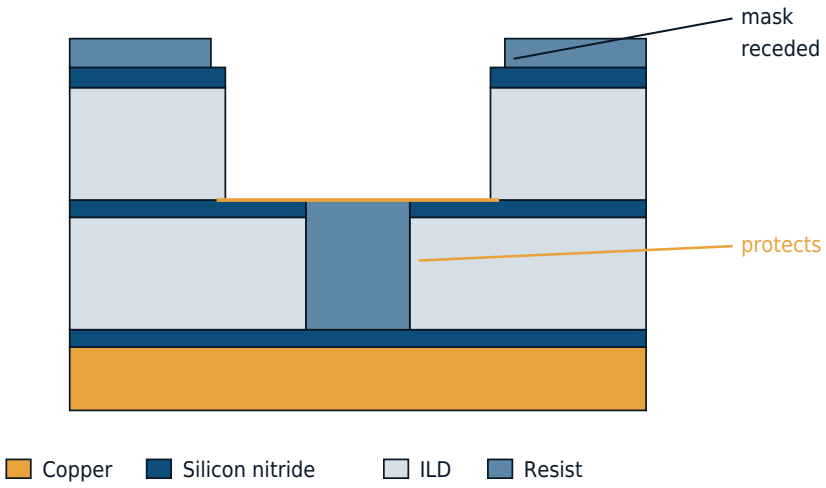


The resist plug in the hole is the key to this process: without it, the subsequent etch would deepen the via further and break through the etch stop.

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

3. Etch the Trench

The resist in the via keeps the etch gases away from the layers beneath



After this, as in the other process, come opening the etch stop, the barrier, the copper fill, and polishing.

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.

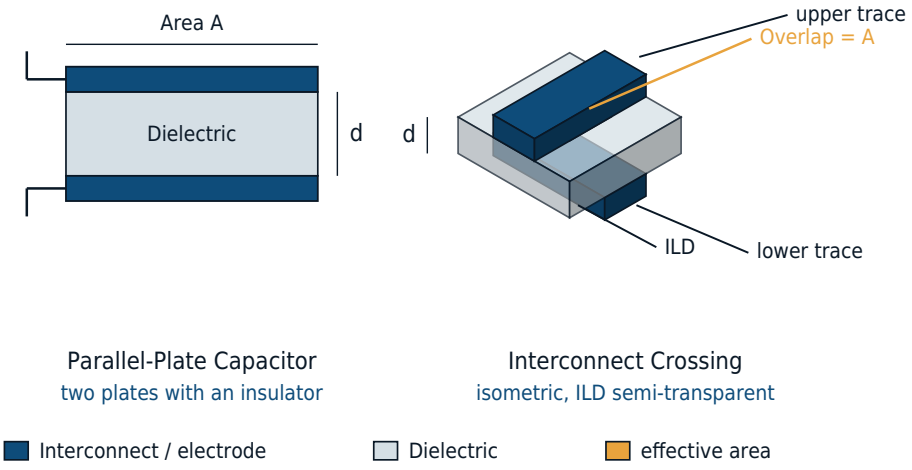
Low-k technology

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.

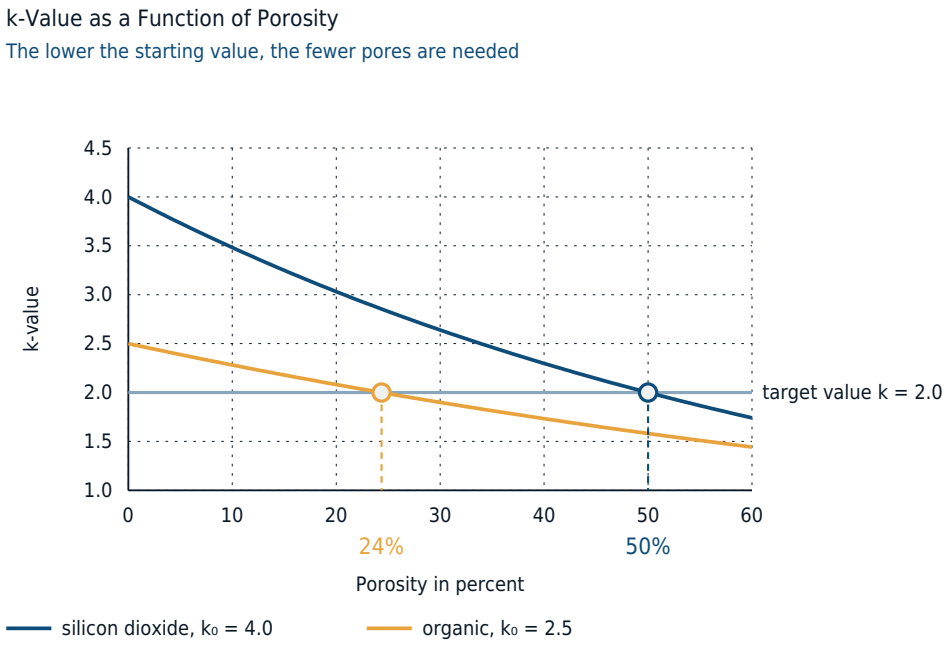
To reduce the permittivity, there are basically two approaches:

- 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.



(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.

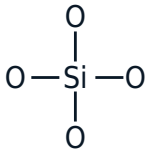
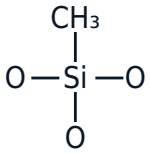
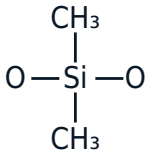
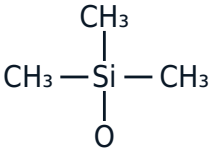
When no material helps anymore: air gaps

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

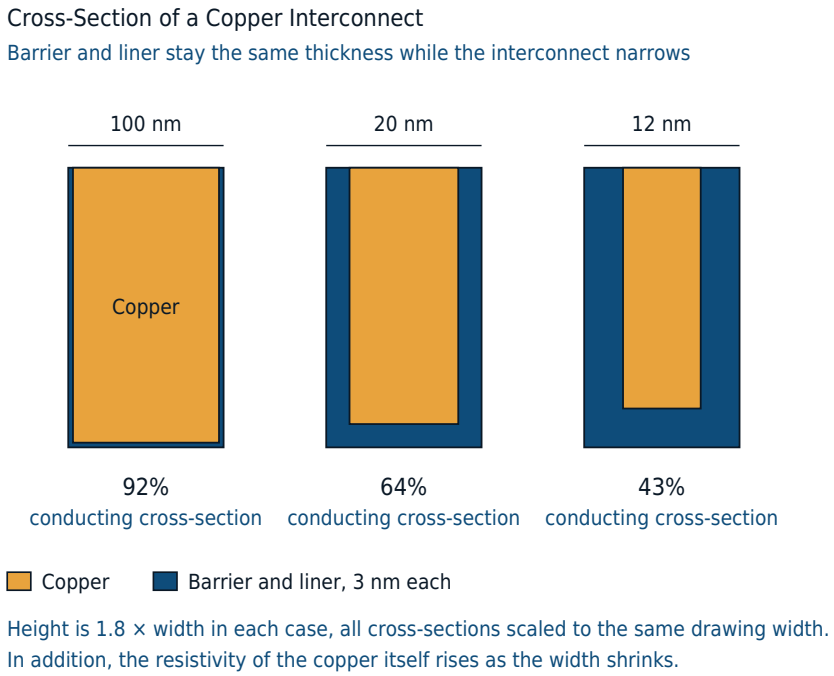
The value of 1.7 µΩ·cm applies to an extended piece of copper. It is based on the fact that electrons travel a certain mean free path between two collisions; in

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.

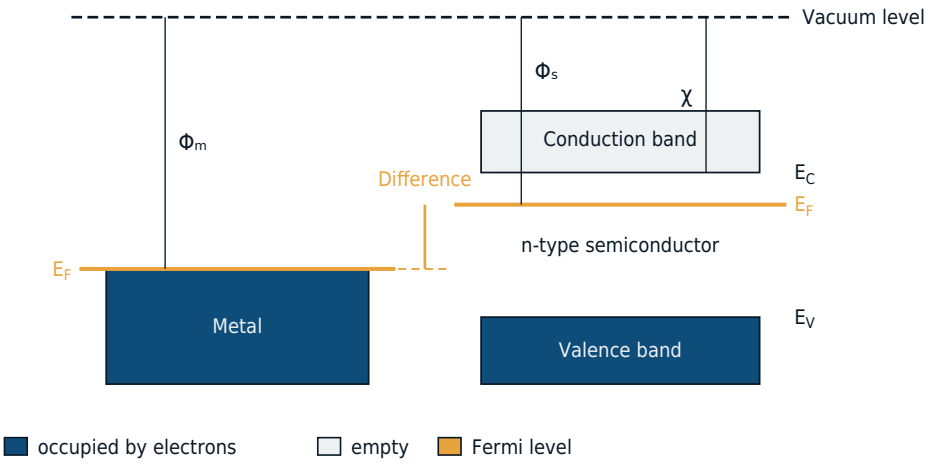
Metal semiconductor junction

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.

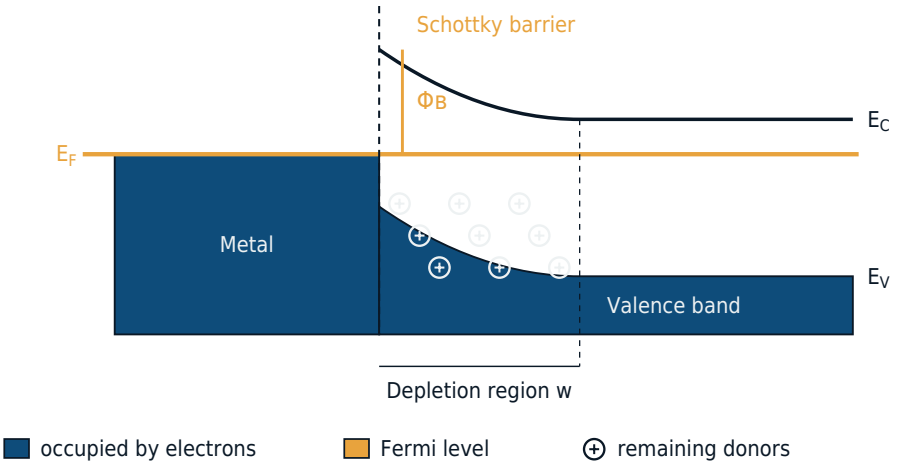
Thus, the probability of electrons being present in the conduction band of the semiconductor decreases: the distance between the conduction band edge and the Fermi level – which describes the highest energy state still occupied by electrons – increases at the interface.

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.

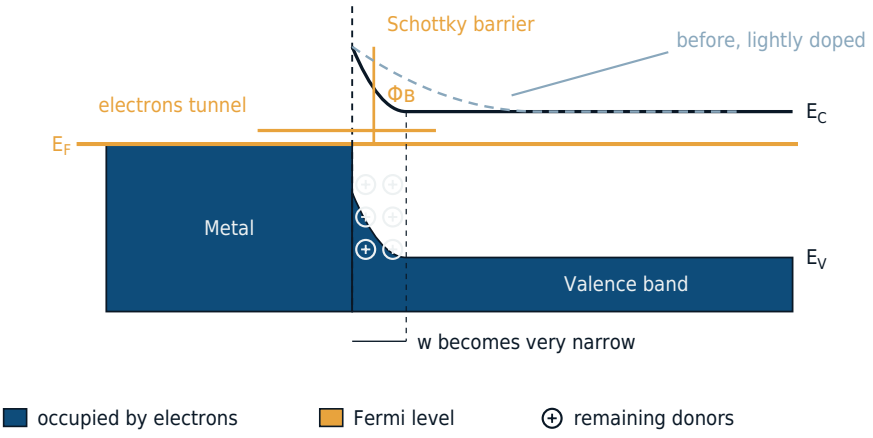
The width *w* of the depletion zone depends on the strength of the doping. The electrons that have migrated out of the semiconductor generate a negative space charge in the metal, which is confined to the surface region.

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.

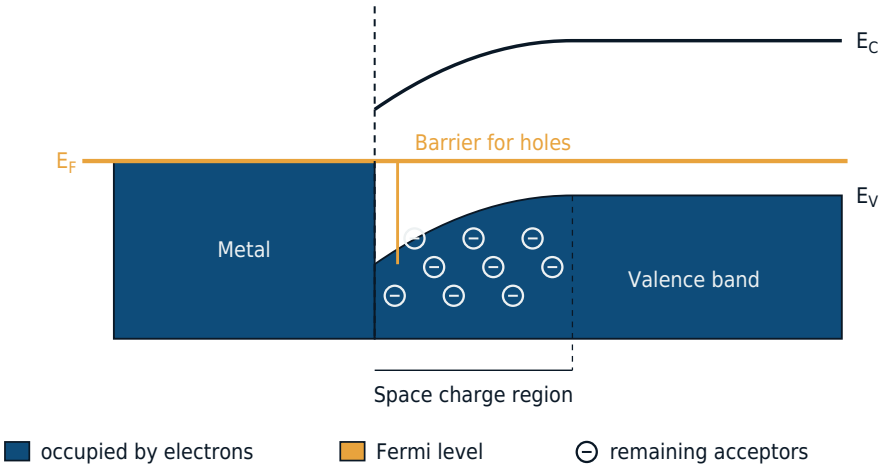
p-type semiconductor contact

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.

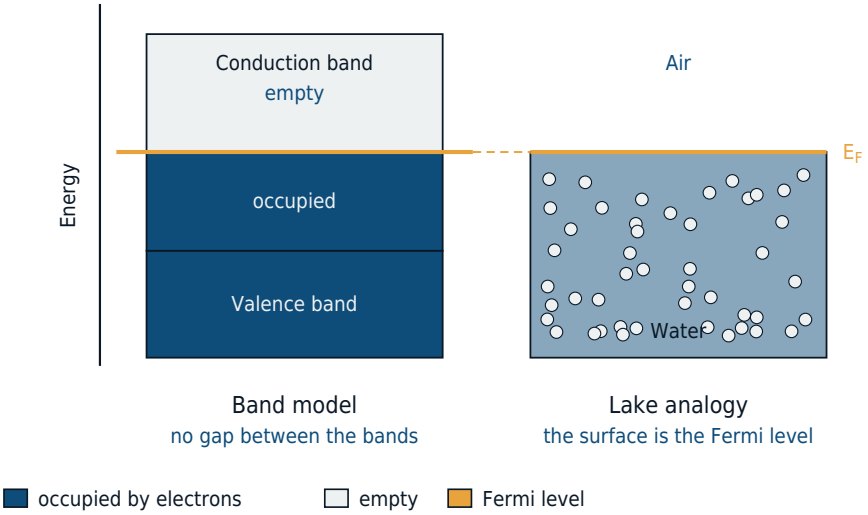
Contacting doped semiconductors

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.

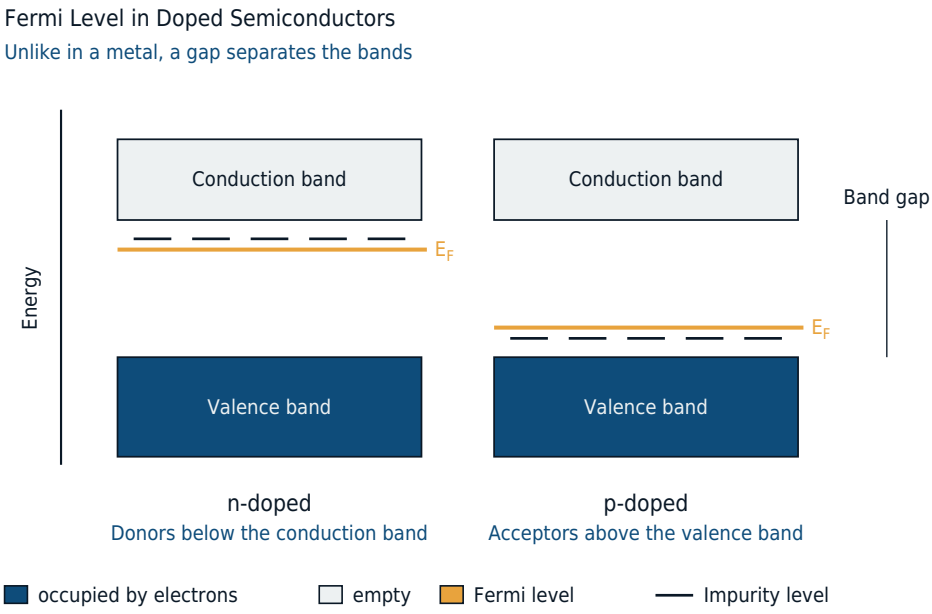
Fermi Level in Metals

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).

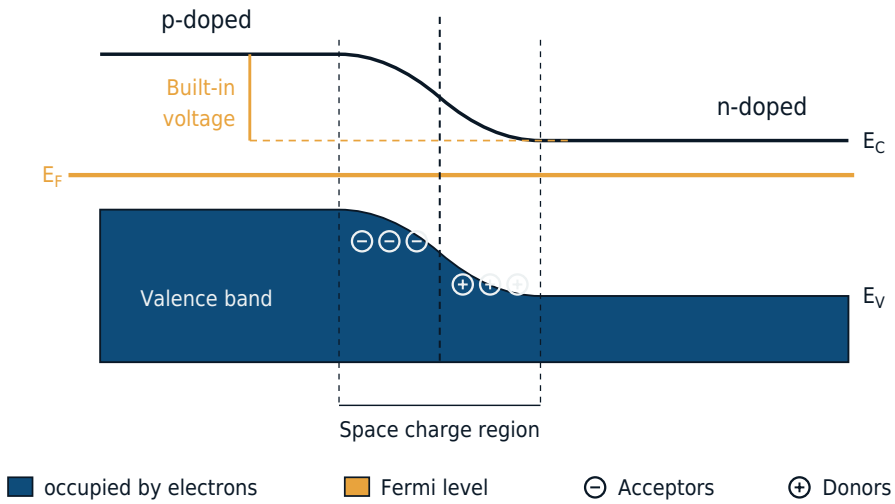


Band model of a p-n junction

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

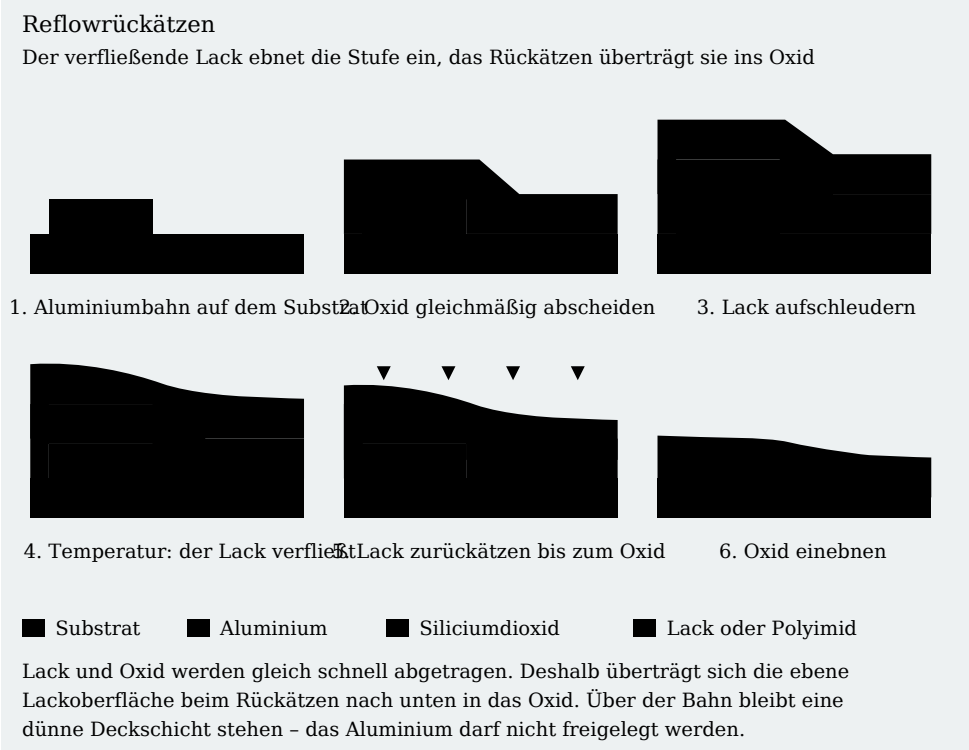
In the reflow technique, layers of doped glasses are deposited onto the wafer. Widely used are phosphosilicate glass (PSG) and borophosphosilicate glass (BPSG). In a high-temperature step, the glasses flow and form a level surface: for PSG and BPSG, this happens at about 900 °C. However, this technique 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.

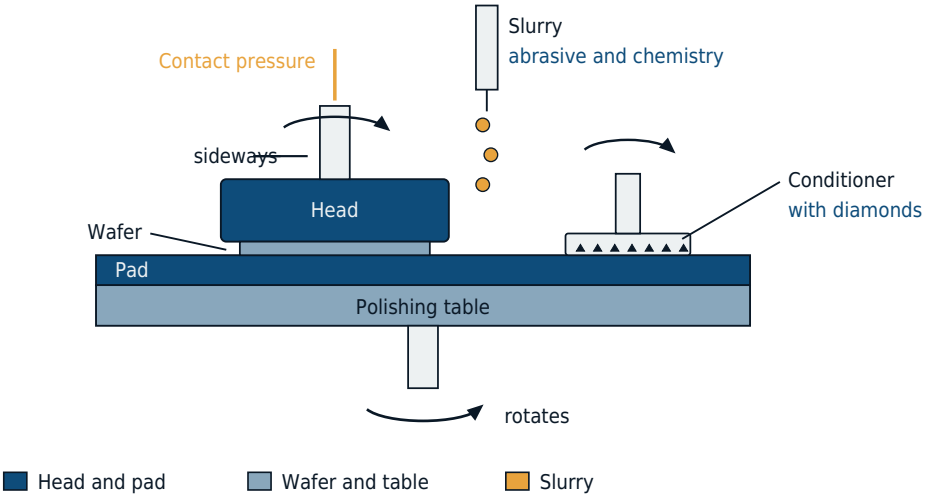
For this purpose, the wafer is held with its active side facing down in a chuck

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.

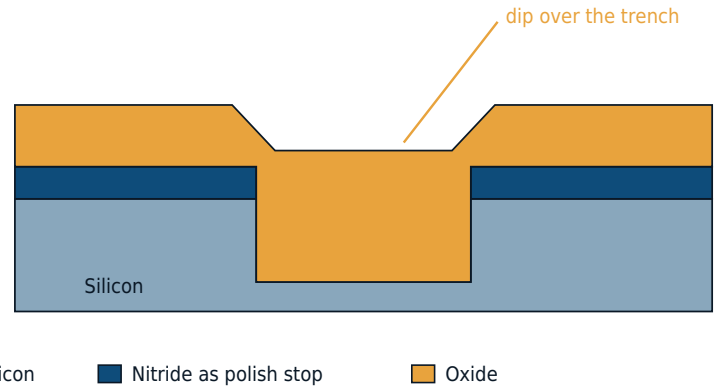
Typically, the CMP process is used after depositing the TEOS for shallow trench isolation, to remove the oxide until only the isolation between the active areas of the transistors remains. Likewise, the interlayer dielectric between the transistor level and the first metal layer (first contact) is polished 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

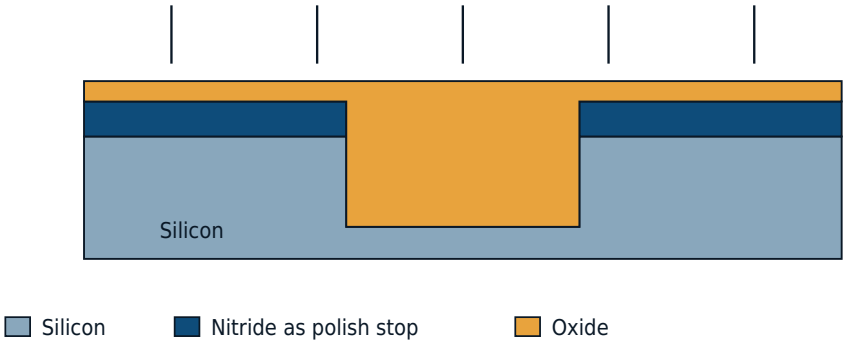
The oxide follows the topography and sits highest over the nitride



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

2. First Polishing Step

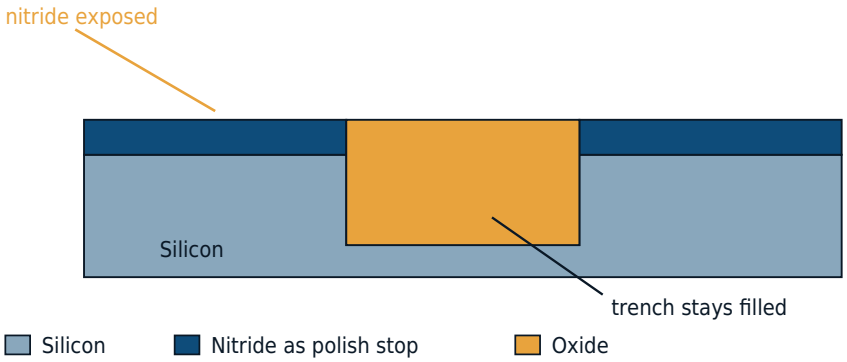
The surface is flat, with a residual layer still over the nitride



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.