Semiconductor Technology from A to Z

Doping Techniques

Diffusion

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 in this section allow silicon wafers to be doped partially. The introduction of the foreign substance can be achieved using various methods: diffusion, ion implantation (and alloying).

Diffusion

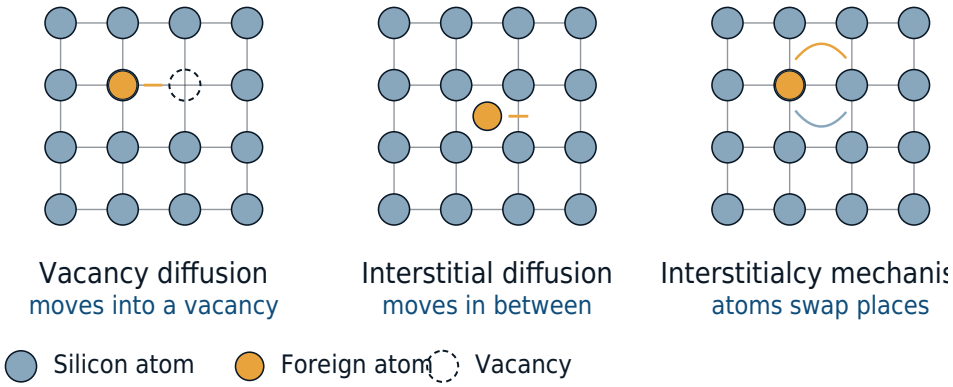
Diffusion is barely used anymore for targeted doping today; it has been superseded by ion implantation. Nevertheless, understanding it remains essential, because it never stops acting: every subsequent high-temperature step causes dopants that have already been introduced to keep migrating. This limits how much heat a wafer may still be subjected to after doping, and is the reason why furnace processes have increasingly been replaced by rapid, second-scale heating.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

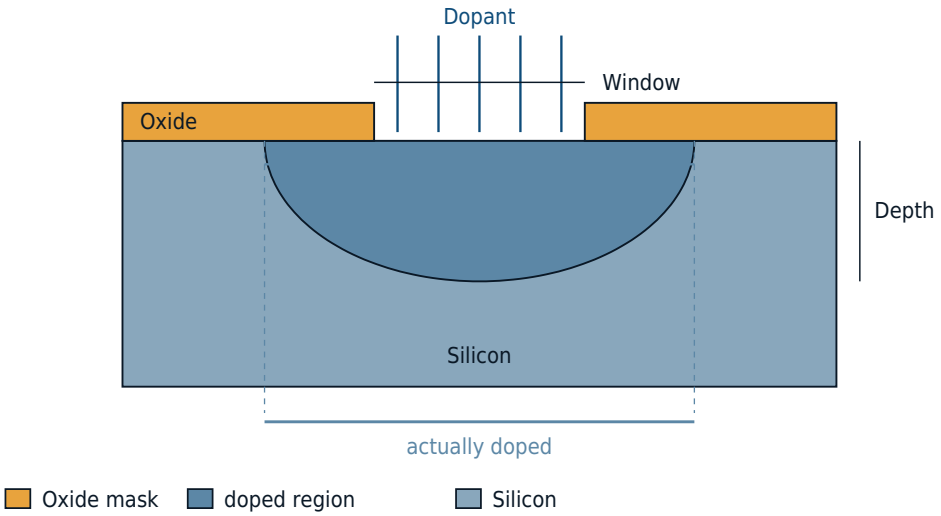
Since the entire wafer is not meant to be doped, certain regions are masked with silicon dioxide. Dopants cannot penetrate this oxide, so no doping takes place at these locations. To avoid stress or even breakage of the wafers, the tube is heated in steps (10 °C per minute) up to approximately 900 °C,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

Fick's Laws and Concentration Profiles

The qualitative description from the previous section can be made more precise: diffusion follows Fick's laws. The first states that the particle flux is proportional to the concentration gradient; the second describes how the concentration profile evolves over time. Depending on the boundary condition, two characteristic profile shapes emerge that are deliberately exploited in practice.

Predeposition holds the surface concentration constant at the dopant's solubility limit in silicon – the source keeps replenishing without limit. This produces a complementary error function profile (erfc profile): steeply falling, with a total dose capped by the solubility limit.

Drive-in, by contrast, works without further supply: the dose already introduced is fixed, and only the existing atoms keep migrating into the substrate at the process temperature. This produces a Gaussian profile that becomes wider and shallower over time while the total dose is conserved.

In practice, both steps are combined: a short predeposition sets a precisely defined total dose, and a subsequent drive-in step without further supply shapes it into the desired depth profile. The characteristic penetration depth grows with the square root of the product of diffusion coefficient and time ($D \cdot t$) – this product is exactly the thermal budget a wafer "spends" during a diffusion step.

Thermal Budget in Process Integration

Because diffusion never fully stops, the product D · t accumulates across every high-temperature step a wafer passes through during fabrication – the thermal budget. This coupling dictates much of the ordering of the entire front-end flow: steps that themselves require a large thermal budget – well drive-in, field oxide growth, gettering – must happen early in the process, while no shallow, sensitive profiles yet exist that could be blurred by them. The shallowest, most tightly toleranced profiles, above all the source/drain extensions, are therefore formed as late as possible and afterward exposed to only minimal further heat – using the short anneal techniques already mentioned, such as spike annealing or flash-lamp annealing.

Besides diffusion itself, segregation at the oxide-silicon interface also reshapes the profile during every subsequent oxidation step (see Thermal Oxidation).

This principle also explains the overall trend of recent decades: wherever a shallow, precise profile is needed, ion implantation with a brief anneal displaces classic diffusion, because it conserves the wafer's overall thermal budget far better.

Where Diffusion Is Still Used Today

Despite being displaced by ion implantation, diffusion is far from obsolete – it remains the first choice wherever deep, less critical profiles are needed or thermal budget isn't a concern.

Deep wells and power devices: For junction depths of several micrometers, as found in power semiconductors or older, less critical technologies, ion implantation would require impractically high acceleration voltages – diffusion accomplishes this with a simple furnace process.

Gettering: Fast-diffusing metal contaminants such as iron, copper, or nickel can be diffused out of the active device region into a sacrificial zone (e.g. mechanically damaged backside, or a heavily doped layer) during a dedicated high-temperature step, where they become harmless. This process uses diffusion itself as the cleaning mechanism.

Solar cells: The emitter layer of crystalline silicon solar cells is still predominantly manufactured industrially via $POCl_3$ tube diffusion – inexpensive, high-throughput, and the requirements on profile depth and sharpness are far below those of logic fabrication.