Semiconductor Technology from A to Z

Doping Techniques

Diffusion 3
Doping 3
Diffusion 3
Diffusion Methods 4
Fick's Laws and Concentration Profiles 6
Thermal Budget in Process Integration 6
Where Diffusion Is Still Used Today 7
Ion Implantation 7
Ion implantation 8
Penetration Depth and Annealing 9
Channeling 10
Characteristics and Implanter Types 10
Plasma Doping (PLAD) 11
Halo and Well Implants in CMOS Integration 12
Amorphization and Defect Annealing 13
Dose Metrology 14

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

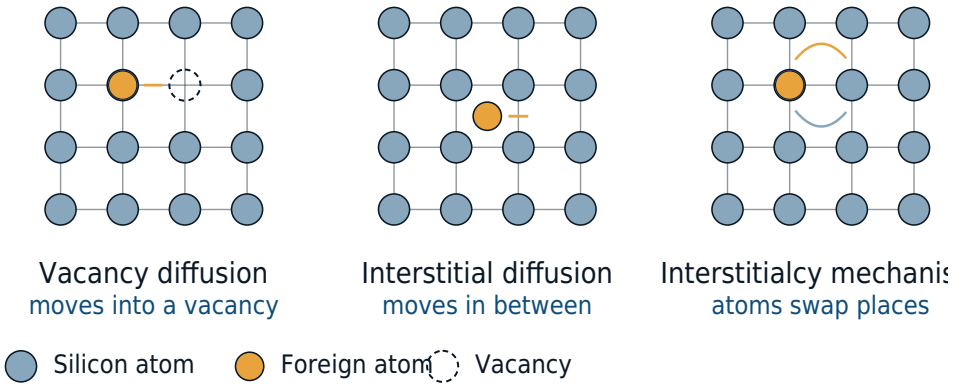
Diffusion is barely used anymore for targeted doping today; it has been superseded by ion implantation. Nevertheless, understanding it remains essential, because it never stops acting: every subsequent high-temperature step causes dopants that have already been introduced to keep migrating. This limits how much heat a wafer may still be subjected to after doping, and is the reason why furnace processes have increasingly been replaced by rapid, second-scale heating.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

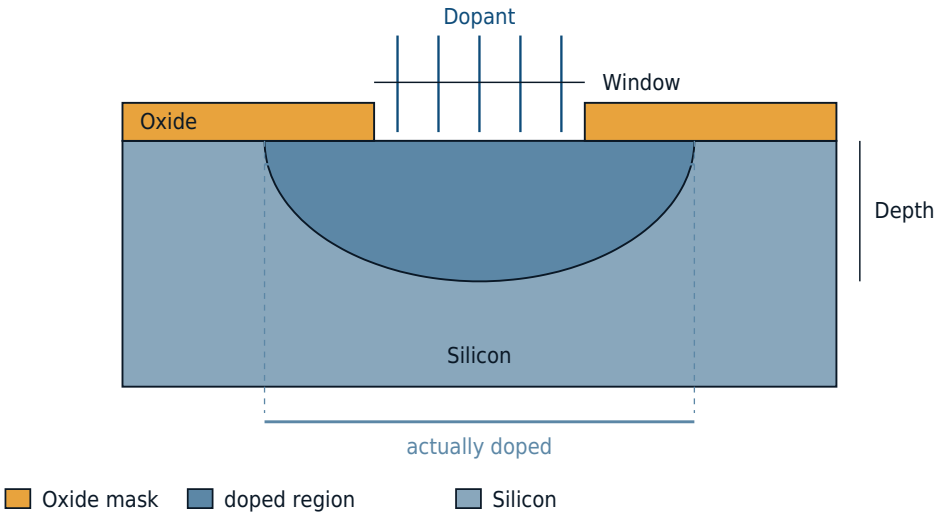
Since the entire wafer is not meant to be doped, certain regions are masked with silicon dioxide. Dopants cannot penetrate this oxide, so no doping takes place at these locations. To avoid stress or even breakage of the wafers, the tube is heated in steps (10 °C per minute) up to approximately 900 °C,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

Ion Implantation

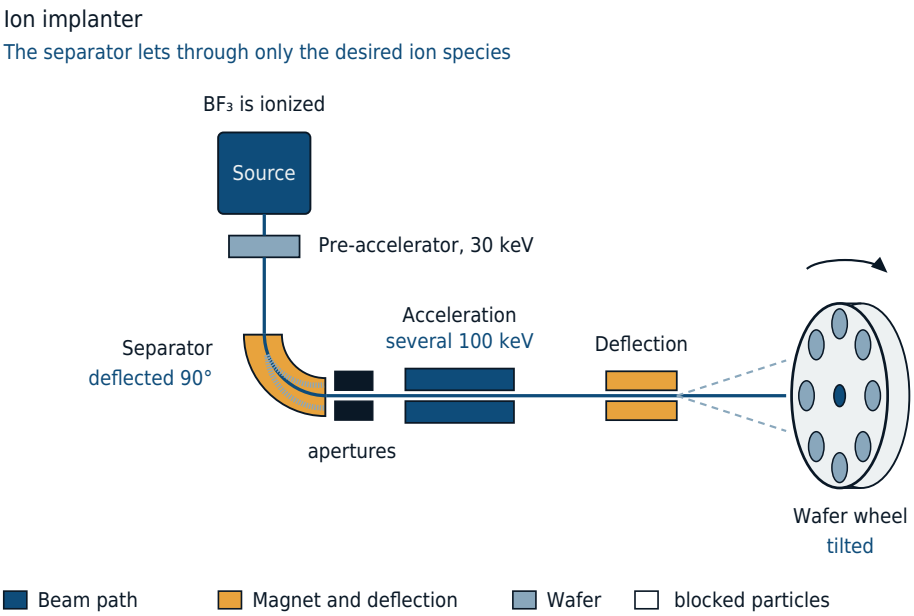
Ion implantation

In ion implantation, charged dopants (ions) are accelerated in an electric field and directed onto the wafer. The penetration depth can be set very precisely by reducing or increasing the voltage used to accelerate the ions. Since the process takes place at room temperature, dopants introduced earlier cannot diffuse out. As with diffusion, regions that should not be doped are covered with a mask, though for implantation a photoresist mask is sufficient.

An implanter consists of the following components:

- Ion source: the dopant gas (e.g. boron trifluoride BF_3) is ionized (electrons emitted from a hot cathode collide with the gas particles; impact ionization continuously generates positive ions and free electrons)
- Pre-accelerator: the ions are drawn out of the ion source at roughly 30 kiloelectronvolts
- Mass separator: the charged particles are deflected by 90° in a magnetic field. Particles that are too light or too heavy are deflected more or less than the desired ions and are caught by screens behind the separator
- Acceleration lane: several hundred keV accelerate the particles to their final energy (200 keV accelerate boron ions to roughly 2,000,000 m/s)
- Lenses: lenses are distributed throughout the system to focus the ion beam
- Beam deflection: electric fields deflect the ions to irradiate the desired location
- Wafer station: the wafers are brought into the ion beam either individually or mounted on large rotating wheels

Illustration of an ion implanter



Only ions with the correct mass pass through the separator on the right path. Lighter ones are deflected more, heavier ones less, and they end up in the apertures. This keeps the doping species-pure.

Penetration Depth and Annealing

Penetration depth of ions in the wafer

Unlike diffusion, the particles do not penetrate the crystal through their own thermal motion but are shot into the crystal lattice at high velocity. Collisions with silicon atoms slow them down along the way. The impacts knock silicon atoms out of their lattice sites, while the dopant ions themselves mostly come to rest on interstitial sites. There, they are not electrically active, since they form no bonds with neighboring atoms that could give rise to free charge carriers. The displaced silicon atoms must be reincorporated into the crystal lattice, and the electrically inactive dopants must be activated.

Annealing the crystal lattice and activating the dopants

A temperature step at around 1000 °C moves the dopants onto lattice sites (beforehand, only about 5 % of the dopant atoms sit on lattice sites). Lattice damage from the collisions is already cured at around 500 °C. Because the dopant atoms keep moving through the substrate at these high temperatures, these steps are carried out for only a very short time.

This is exactly where the trade-off of the process lies: the temperature must be high enough to move the dopants onto lattice sites, and the time short enough that they do not migrate away while doing so. Both can only be achieved

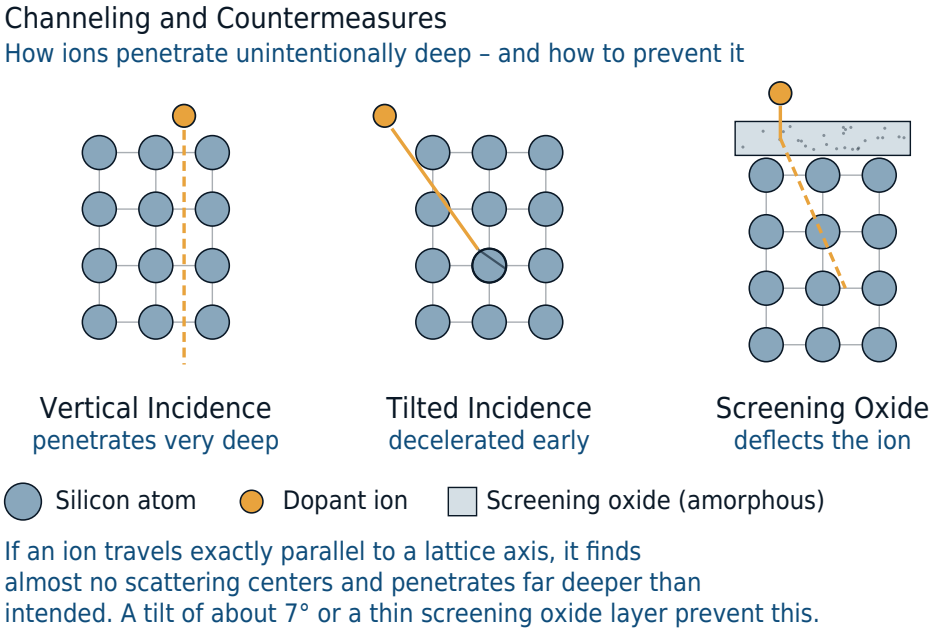
together by continually shortening the heating duration. What began as an hours-long furnace process first became rapid thermal processing over seconds, then spike annealing, where the target temperature is only passed through rather than held, and finally flash-lamp or laser annealing in the millisecond range. Here, only the topmost layer of the wafer is heated while the substrate underneath stays cold.

For very shallow doping profiles, the ion mass is additionally increased: instead of implanting individual boron atoms, molecules containing several boron atoms are implanted. At the same acceleration voltage they carry less energy per boron atom and therefore penetrate less deeply – a way to create very shallow junctions without having to operate the implanter at impractically low voltages.

Channeling

The substrate used is a single crystal, meaning the silicon atoms are arranged periodically and form channels. Implanted dopant atoms that travel parallel to these channels are only weakly decelerated and penetrate very deeply into the substrate. There are two ways to prevent this:

- Wafer tilt: the wafers are tilted by roughly 7° relative to the beam direction. This means the ions are not shot in parallel to the lattice channels and are decelerated early through collisions.
- Screening oxide: a thin oxide is applied to the wafer surface, which deflects the ions and prevents them from arriving perpendicular to the substrate



Characteristics and Implanter Types

Characteristics

- The reproducibility of ion implantation is very high
- The room-temperature process prevents other dopants from diffusing out
- A photoresist mask is sufficient; an oxide mask, as needed for diffusion, is not required
- Ion implanters are expensive, and the cost per processed wafer is high
- Dopants do not spread laterally underneath the mask (only minimally, due to collisions)
- Nearly every element can be implanted at the highest purity
- Similar to dopant deposits in the quartz tube during diffusion, ions can accumulate on walls or screens and later be dislodged and carried onto wafers during subsequent implants
- Vertical surfaces are difficult to dope with a directional ion beam. Tilting the wafer helps to a limited extent; for structures such as FinFET fins, doping is instead performed from a plasma whose ions arrive from all directions
- The implantation process takes place under high vacuum, which must be generated with several turbomolecular or cryopumps

Several types of implanters exist, with medium-current and high-current implanters used most often. Medium-current implanters are suited to small and medium ion doses ($1 \cdot 10^{11}$–$1 \cdot 10^{15}$ ions/cm^2), high-current implanters to doses of $1 \cdot 10^{15}$–$1 \cdot 10^{17}$ ions/cm^2.
Due to its advantages over diffusion, ion implantation has largely become the standard.

Doping by Alloying

For completeness, it is worth mentioning that besides these two methods there is also doping by alloying. Since this method brings disadvantages such as cracking of the substrate, it is barely used in semiconductor technology today.

Plasma Doping (PLAD)

For structures that no longer lie flat on the wafer but rise into the third dimension like FinFET fins or gate-all-around nanosheets, the classic directional ion beam reaches its limits: sidewalls sit transverse to the beam and receive almost no dopant, even with a tilted wafer. Plasma doping (PLAD) solves this by abandoning the beam-line architecture of the classic implanter entirely.

The wafer sits on a pulsed high-voltage platen inside a plasma chamber, in which a process gas of the desired dopant (e.g. diborane B_2H_6 or phosphine PH_3)

is struck. Each high-voltage pulse builds up a plasma sheath around the wafer, across which ions are accelerated toward the surface – not as a narrow beam from one direction, but broadly from the entire plasma. Because the sheath follows the wafer's contour, sidewalls of fins and ridges are doped nearly conformally as well.

The price is giving up the mass separator: since no magnetic field sorts the ions by mass, every species ionized in the plasma reaches the wafer together – alongside the desired dopant ion also hydrogen, molecular fragments, and possible contaminants from the process gas. Dose and species purity are therefore controlled less precisely than in classic mass-separated implantation (see Ion Implantation). Because the entire wafer is exposed at once instead of being scanned point by point, throughput is considerably higher, and the typically very low pulse energies of a few kiloelectronvolts suit extremely shallow doping profiles well.

In production, PLAD is therefore used specifically where classic implantation fails on geometry alone – for example to dope fin sidewalls, or as a replacement for heavily tilted extension implants where shadowing by neighboring structures would otherwise become a problem (see also Characteristics and Implanter Types).

Halo and Well Implants in CMOS Integration

A finished transistor doesn't result from a single implant, but from the interplay of several implantation steps, each separated in timing and energy, each with its own depth and purpose.

Wells

Right at the start of front-end processing, well implants at MeV-range energies create large-area n- or p-doped regions deep in the substrate, in which transistors of the opposite type are later built – an n-well hosts p-channel transistors, a p-well n-channel transistors. These implants extend several hundred nanometers to a few micrometers deep and mostly determine the electrical behavior of the substrate as a whole.

Halo/Pocket Implants

Directly beneath the future source and drain edge sits an oppositely doped, narrow region, the halo or pocket implant. It locally raises the dopant concentration at the edge of the channel and thereby counteracts the short-channel effect: without it, the depletion zones of source and drain would expand

far enough at short channel lengths to touch (punch-through), and the transistor would lose control over the channel. Because this region must sit directly beneath the gate edge, it is implanted at a markedly steeper tilt angle than the other steps.

Four-Fold Rotation

Since transistors on a chip can be oriented in any of four directions relative to each other, a single tilted implant step would only dope symmetrically the gates whose channel direction happens to align with the tilt plane – all others would receive an asymmetric, orientation-dependent profile. The halo step is therefore split into four: the same dose is applied in four sub-steps at the same tilt angle, but rotated 90 ° around the wafer axis each time. In total, every gate receives a symmetric halo dose from all four sides regardless of its orientation on the chip.

LDD and Source/Drain Implants

The shallow, lightly doped extensions known as LDD (Lightly Doped Drain) sit right at the gate edge; they smooth the concentration profile between the channel and the deep, heavily doped source/drain implants, lowering the electric field at the drain edge – too abrupt a transition would otherwise generate hot carriers that damage the gate oxide. Only the final, high-dose source/drain implants provide the actual, low-resistance connection regions of the transistor.

Amorphization and Defect Annealing

At sufficiently high dose, especially with heavy ions, implantation disrupts the crystal structure so thoroughly that the affected layer turns amorphous – the silicon locally loses its regular lattice order entirely. Rather than avoiding this, modern processes exploit it deliberately.

Pre-Amorphization (PAI)

Before the actual dopant implant, a pre-amorphization step (Pre-Amorphization Implant, PAI) shoots electrically inactive ions such as germanium or silicon itself into the surface, deliberately creating a uniformly amorphous layer. The subsequent dopant then no longer hits a crystal lattice with open channels but disordered material – channeling is thereby ruled out from the start, regardless of tilt angle.

Solid-Phase Epitaxy and End-of-Range Defects

During the subsequent anneal, the amorphous layer regrows single-crystalline

starting from the intact crystal structure underneath (solid-phase epitaxy). Right at the original boundary between amorphous and crystalline material, however, a band of defects often remains, the end-of-range (EOR) defects: dislocation loops of excess interstitial atoms that were shot past the actual amorphous zone during implantation and are not fully incorporated during annealing.

Transient Enhanced Diffusion (TED)

These excess interstitial atoms are not just electrically inactive crystal defects; they also accelerate the diffusion of nearby dopants – boron above all – by a large factor for the short duration of their own annealing. This effect is called transient enhanced diffusion (TED) and causes doping profiles to spread significantly further during annealing than equilibrium diffusion at the given temperature would predict. Because TED only acts as long as excess interstitial atoms are present, countermeasures target exactly this window: very short anneal times (see Penetration Depth and Annealing), which give the defects no time to migrate, or a carbon co-implant that specifically traps interstitial atoms before they can drive dopant diffusion.

Dose Metrology

How many ions actually made it into the wafer must be verifiable both during implantation and afterward – the two checks fundamentally measure different quantities.

Faraday Cup: Real-Time Dose Control

Inside the implanter itself, one or more Faraday cups handle dose control: they capture a known fraction of the ion beam and measure its current. Integrated over time, this yields the accumulated charge, and from that, relative to the exposed area and the charge per ion, the actually applied dose. The implanter automatically stops the process once the target dose is reached. Multiple cups at the edge and center of the wafer also allow the beam's uniformity across the wafer to be monitored. This measurement is fast and runs with every wafer, but says nothing about where in the wafer the ions actually ended up or how many are electrically active after annealing.

SIMS: The Depth Profile Afterward

Secondary ion mass spectrometry (SIMS) sputters the wafer surface layer by layer with a primary ion beam and mass-analyzes the secondary ions released in the process. This yields a highly precise concentration profile of the dopant

versus depth – the reference method for verifying range, straggle, and, by integrating the profile, the actually introduced total dose after the fact. SIMS is destructive and too slow to check every single wafer; it is used for process qualification and calibration, not in routine production.

For routine production, four-point-probe sheet resistance measurement is used instead: fast, non-destructive, and sensitive to the electrically active dose after annealing – though without depth resolution. Only the combination of routine Faraday-cup and sheet-resistance measurement with occasional SIMS calibration gives a complete picture of the doping.