Semiconductor Technology from A to Z

Metallization

Chemical Mechanical Polishing (CMP)

Why planarization is necessary

With every additional layer – isolation oxide, contact level, metal layer – the topography on the wafer surface keeps growing. This is a problem for lithography: the depth of focus of modern exposure systems is only a few hundred nanometers, so structures on surfaces of different height can no longer be imaged sharply at the same time. Subsequent deposition and etch steps also proceed more uniformly on a flat surface than across steps and trenches. Beyond a certain level of integration density, a globally planar surface is therefore no longer a matter of convenience but a precondition for the next process level.

Principle of the CMP process

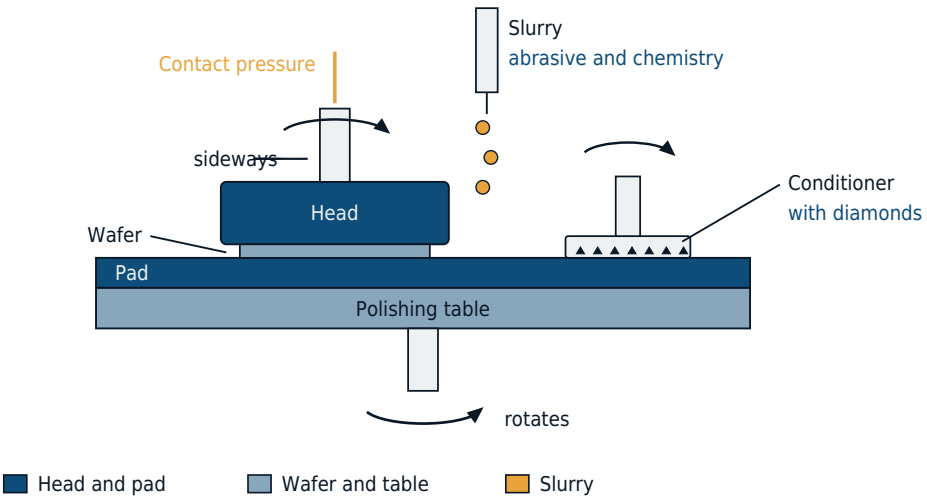
In chemical-mechanical polishing (also chemical-mechanical planarization, CMP for short), the wafer surface is attacked chemically and removed mechanically at the same time – unlike with pure reflow techniques, this produces uniformity across the entire wafer rather than just local smoothing.

The wafer is held face-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 additionally perform horizontal movements. A solution (slurry) of abrasive particles and chemically active substances serves as the polishing medium; under pressure it alters the surface and thereby supports the material removal. To distribute the slurry evenly and to condition the polishing cloth, the pad is roughened with a diamond-studded steel disc (dresser/conditioner) – either during polishing (in-situ) or before/after it (ex-situ).

The CMP process usually runs in two to three stages on pads with different surface properties and different slurries; the wafer is transferred to the next pad after each step. A cleaning step follows to remove particle and slurry residues.

CMP Tool

The wafer is pressed face-down onto the rotating pad



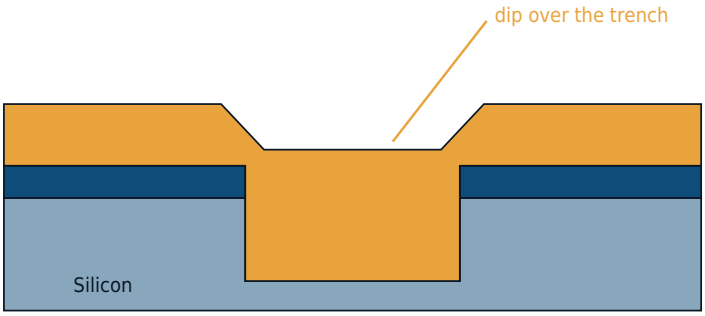
Head and polishing table rotate in opposite directions, and the head also moves sideways across the pad. Only this combined motion ensures that every point on the wafer is removed at the same average rate.

STI-CMP

After depositing the TEOS oxide for shallow trench isolation (STI), CMP is used to remove the oxide until it remains only in the isolation trenches between the active areas. The process typically runs in two stages: the first step planarizes the oxide over the active areas and the trenches, the second step selectively removes the remaining oxide down to the nitride passivation layer. It is important here that the oxide be completely removed from the areas where the transistors will later be fabricated – otherwise the nitride protecting the underlying silicon cannot be removed by wet chemistry in the following step.

1. Trench overfilled with oxide

The oxide follows the topography and sits highest over the nitride

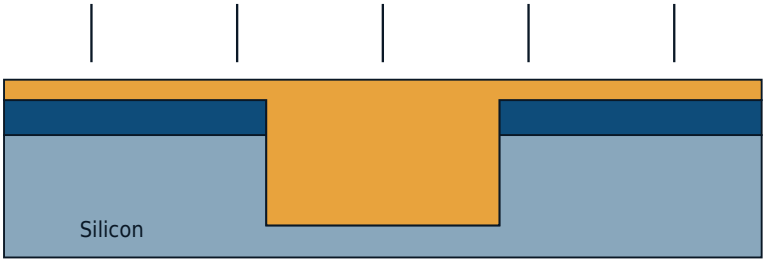


Silicon Nitride as polish stop Oxide

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

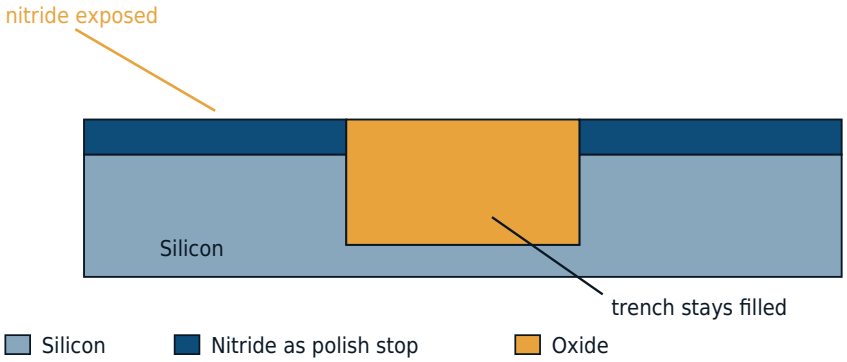


Silicon Nitride as polish stop Oxide

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.

ILD-CMP

The interlayer dielectric (also pre-metal dielectric, ILD or PMD) between the transistor level and the first contact level is likewise polished back to the required thickness using CMP. The goal, again, is a planar surface regardless of how uneven the underlying gate and source/drain relief is – only then can the contact holes be exposed and etched with consistent depth of focus.

Tungsten CMP

Contact holes to the source, drain, and gate regions are etched into the planarized ILD and then filled with tungsten. Since tungsten is difficult to deposit selectively into narrow, deep holes, a thin titanium/titanium nitride barrier is first deposited as an adhesion layer and diffusion barrier, followed by a blanket, conformal tungsten deposition by CVD (from WF_6) that also completely fills the contact holes.

This leaves a continuous tungsten layer on the surface, which has no electrical function and would short-circuit the individual contacts. CMP removes this excess tungsten completely from the ILD surface, leaving only the isolated plugs in the contact holes. The slurry chemistry differs substantially from the oxide-based STI/ILD polish, since tungsten must be removed selectively with respect to the underlying oxide and barrier layer.

Copper Damascene CMP

The copper wiring layers are likewise planarized in a CMP process – here, however, polishing is not merely a planarization step but part of the actual patterning. Since copper cannot be dry-etched, it is deposited across the entire wafer into pre-patterned trenches and vias in the damascene process; CMP then removes both the excess copper and the underlying barrier layer (usually Ta/TaN) down to the surrounding interlayer dielectric, leaving isolated interconnect lines behind. Details on the damascene process itself can be found in the chapter Copper technology.

Dishing, Erosion & Dummy Fill

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.

Endpoint Detection & Process Control

Since every layer has a different thickness and removal is never perfectly reproducible from wafer to wafer, polishing must be stopped deliberately once the target layer is reached – too early, and residues of the layer being removed remain; too late, and underlying layers are attacked. Common endpoint detection methods include optical interferometry, where the reflected light signal changes periodically as the layer thickness decreases, measurement of the friction torque between pad and wafer, which changes abruptly at the transition to a different material, and, for metal layers, eddy current measurement, which responds directly to the remaining metal thickness.

Even though the process as a whole 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.