

Halbleitertechnologie von A bis Z

Metallisierung

Chemisch-Mechanisches Polieren (CMP)	3
<i>Warum Planarisierung nötig ist</i>	3
<i>Prinzip des CMP-Prozesses</i>	3
<i>STI-CMP</i>	4
<i>ILD-CMP</i>	6
<i>Wolfram-CMP</i>	6
<i>Kupfer-Damascene-CMP</i>	7
<i>Dishing, Erosion & Dummy-Fill</i>	7
<i>Endpoint-Detection & Prozesskontrolle</i>	7

Chemisch-Mechanisches Polieren (CMP)

Warum Planarisierung nötig ist

Mit jeder zusätzlichen Schicht – Isolationsoxid, Kontaktebene, Metalllage – wächst die Topografie auf der Waferoberfläche weiter an. Für die Lithografie ist das ein Problem: Die Tiefenschärfe moderner Belichtungssysteme liegt im Bereich weniger hundert Nanometer, sodass Strukturen auf unterschiedlich hohem Untergrund nicht mehr gleichzeitig scharf abgebildet werden können. Auch nachfolgende Abscheide- und Ätzschritte laufen auf einer ebenen Fläche gleichmäßiger ab als über Stufen und Gräben hinweg. Eine global plane Oberfläche ist deshalb ab einer gewissen Integrationsdichte keine Komfortfrage mehr, sondern Voraussetzung für die nächste Prozessebene.

Prinzip des CMP-Prozesses

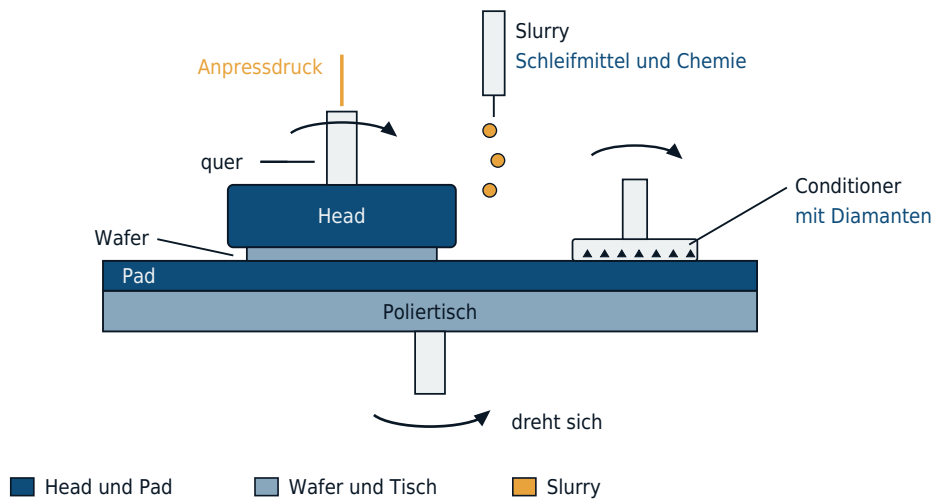
Beim chemisch-mechanischen Polieren (auch chemisch-mechanisches Planarisieren, kurz CMP) wird die Waferoberfläche gleichzeitig chemisch angegriffen und mechanisch abgetragen – anders als bei reinen Reflow-Techniken entsteht dabei eine Homogenität über die gesamte Scheibe hinweg, nicht nur eine lokale Glättung.

Der Wafer wird mit der aktiven Seite nach unten in einem Chuck mit Vakuumsaugung (Head) gehalten und auf eine Polierfläche (Pad, meist aus Polyurethan) auf dem Poliertisch gepresst. Head und Poliertisch drehen sich, während der Head zusätzlich horizontale Bewegungen ausführen kann. Als Poliermittel dient eine Lösung (Slurry) aus Schleifpartikeln und chemisch aktiven Substanzen, die unter Druck die Oberfläche verändert und so den Materialabtrag unterstützt. Zur besseren Verteilung der Slurry und um das Poliertuch zu konditionieren, wird das Pad mit einer diamantbesetzten Stahlscheibe (Dresser/Conditioner) aufgeraut – entweder während des Polierens (in-situ) oder davor/danach (ex-situ).

Der CMP-Prozess läuft meist in zwei bis drei Stufen auf Pads mit unterschiedlicher Oberflächenbeschaffenheit und mit verschiedenen Slurrys ab; der Wafer wird dafür nach jedem Schritt auf das nächste Pad transferiert. Abschließend folgt eine Reinigung, um Partikel- und Slurryreste zu entfernen.

CMP-Anlage

Der Wafer wird mit der aktiven Seite nach unten auf das rotierende Pad gepresst



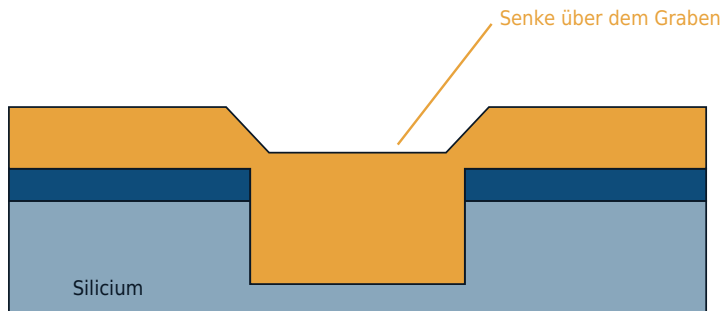
Head und Poliertisch drehen sich gegenläufig, der Head bewegt sich zusätzlich quer über das Pad. Erst diese Überlagerung sorgt dafür, dass jeder Punkt der Scheibe im Mittel gleich stark abgetragen wird.

STI-CMP

Nach der Abscheidung des TEOS-Oxids zur Shallow-Trench-Isolation (STI) wird CMP eingesetzt, um das Oxid so weit abzutragen, dass es nur noch in den Isolationsgräben zwischen den aktiven Gebieten verbleibt. Der Prozess läuft typischerweise zweistufig: Im ersten Schritt wird das Oxid über den aktiven Gebieten und den Gräben planarisiert, im zweiten Schritt wird das restliche Oxid selektiv bis auf die Nitrid-Passivierungsschicht abgetragen. Dabei muss das Oxid auf den späteren Transistorgebieten vollständig entfernt werden – sonst lässt sich das darunterliegende, das Silicium schützende Nitrid im nachfolgenden Schritt nicht nasschemisch entfernen.

1. Graben mit Oxid überfüllt

Das Oxid folgt der Topografie und liegt über dem Nitrid am höchsten

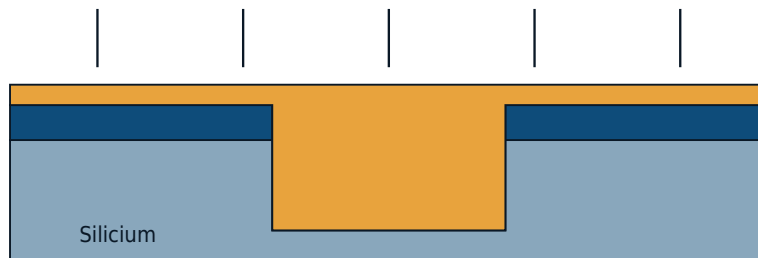


■ Silicium ■ Nitrid als Polierstopp ■ Oxid

Abgeschieden wird deutlich mehr Oxid, als der Graben fasst – sonst bliebe nach dem Polieren eine Mulde zurück.

2. Erster Polierschritt

Die Oberfläche ist eben, über dem Nitrid liegt noch eine Restschicht

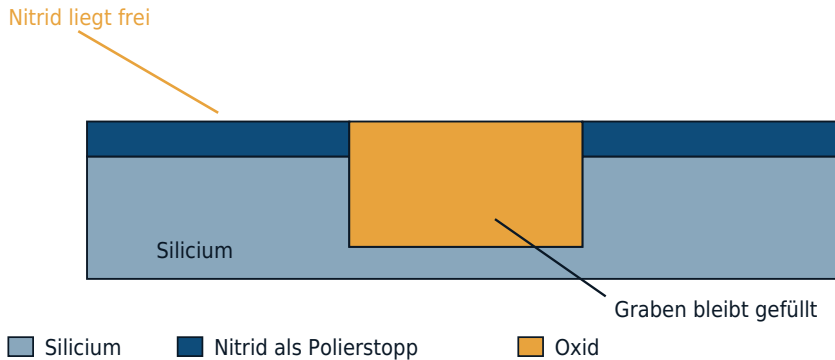


■ Silicium ■ Nitrid als Polierstopp ■ Oxid

Der erste Schritt trägt kräftig ab und ebnet ein. Er hält bewusst vor dem Nitrid an, damit dieses nicht angegriffen wird.

3. Zweiter Polierschritt

Selektiv bis auf das Nitrid – auf den aktiven Gebieten bleibt kein Oxid



Bliebe hier Oxid auf dem Nitrid stehen, ließe sich das Nitrid anschließend nicht nasschemisch entfernen – deshalb muss dieser Schritt vollständig sein.

ILD-CMP

Auch das Zwischenoxid (Interlayer-/Pre-Metal-Dielektrikum, ILD bzw. PMD) zwischen der Transistorebene und der ersten Kontaktebene wird per CMP auf die erforderliche Dicke zurückpoliert. Ziel ist wieder eine plane Oberfläche, unabhängig davon, wie uneben das darunterliegende Gate- und Source-/Drain-Relief ist – erst danach lassen sich die Kontaktlöcher mit gleichbleibender Tiefenschärfe belichten und ätzen.

Wolfram-CMP

In das planarisierte ILD werden die Kontaktlöcher zu Source-, Drain- und Gate-Gebieten geätzt und anschließend mit Wolfram gefüllt. Da sich Wolfram schlecht selektiv in die schmalen, tiefen Löcher abscheiden lässt, wird zunächst eine dünne Titan-/Titannitrid-Barriere als Haftvermittler und Diffusionssperre abgeschieden, gefolgt von einer ganzflächigen, konformen Wolfram-Abscheidung per CVD (aus WF_6), die auch die Kontaktlöcher vollständig auffüllt.

Auf der Oberfläche bleibt dabei eine durchgehende Wolframschicht zurück, die keine elektrische Funktion hat und die einzelnen Kontakte kurzschließen würde. Das CMP entfernt dieses überschüssige Wolfram vollständig von der ILD-Oberfläche und lässt nur die isolierten Plugs in den Kontaktlöchern stehen. Die Slurry-Chemie unterscheidet sich dabei deutlich von der oxidbasierten STI-/ILD-Politur, da Wolfram selektiv gegenüber dem darunterliegenden Oxid und der Barrierschicht abgetragen werden muss.

Kupfer-Damascene-CMP

Auch die Verdrahtungsebenen aus Kupfer werden im CMP-Prozess eingeebnet – hier ist das Polieren allerdings nicht nur ein Planarisierungsschritt, sondern Teil der eigentlichen Strukturierung. Da sich Kupfer nicht trocken ätzen lässt, wird es im Damascene-Verfahren ganzflächig in vorstrukturierte Gräben und Vias abgeschieden; das CMP trägt anschließend sowohl das überschüssige Kupfer als auch die darunterliegende Barrierschicht (meist Ta/TaN) bis auf das umgebende Zwischenoxid ab, sodass isolierte Leiterbahnen zurückbleiben. Details zum Damascene-Verfahren selbst finden sich im Kapitel [Kupfertechnologie](#).

Dishing, Erosion & Dummy-Fill

Dishing und Erosion

Das Verfahren hat eine Eigenheit, die bis in den Schaltungsentwurf hineinreicht: Es trägt weiche Stellen schneller ab als harte. Über einer breiten Kupferfläche wird das Poliertuch in den Graben hineingedrückt und höhlt ihn aus (Dishing); in Bereichen mit vielen dicht stehenden Leiterbahnen wird das gesamte Gebiet gegenüber seiner Umgebung abgesenkt (Erosion). Beides erzeugt genau die Unebenheit, die das Polieren beseitigen sollte, und beides hängt nicht vom Prozess ab, sondern davon, wie das Layout aussieht.

Die Gegenmaßnahme ist ungewöhnlich: In leere Bereiche des Layouts werden Metallstrukturen eingefügt, die elektrisch keine Funktion haben und nur dafür sorgen, dass die Metалldichte über den Chip hinweg gleichmäßig ist. Diese Füllstrukturen werden automatisch erzeugt und machen auf manchen Ebenen einen erheblichen Teil des Musters aus.

Endpoint-Detection & Prozesskontrolle

Da jede Schicht unterschiedlich dick und der Abtrag von Wafer zu Wafer nicht perfekt reproduzierbar ist, muss der Poliervorgang gezielt gestoppt werden, sobald die Zielschicht erreicht ist – zu früh, und Reste der abzutragenden Schicht bleiben stehen; zu spät, und darunterliegende Schichten werden angegriffen. Gängige Verfahren zur Endpunkterkennung sind die optische Interferometrie, bei der sich das reflektierte Lichtsignal mit abnehmender Schichtdicke periodisch ändert, die Messung des Reibungsmoments zwischen Pad und Wafer, das sich beim Übergang auf ein anderes Material sprunghaft ändert, sowie bei Metallschichten die Wirbelstrommessung, die direkt auf die verbleibende Metалldicke reagiert.

Auch wenn das Verfahren insgesamt recht grob anmutet, lässt sich damit eine auf wenige Nanometer plane Oberfläche erzeugen. Es ist längst kein Sonderschritt mehr: In einem modernen Prozessablauf wird eine Scheibe einige Dutzend Mal poliert.