Halbleitertechnologie von A bis Z

Metallisierung

Anforderungen an die Metallisierung 3
Anforderungen an die Metallisierung 3
Aluminiumtechnologie 3
Aluminium und Aluminiumlegierungen 3
Siliciumdiffusion 4
Elektromigration 6
Hillockwachstum 7
Kupfertechnologie 8
Kupfertechnologie 8
Damascene-Verfahren 9
Low-k-Technologie 15
Grenzen der Kupferverdrahtung 19
Der Metall-Halbleiter-Kontakt 21
n-Halbleiter-Kontakt 21
p-Halbleiter-Kontakt 24
Kontaktierung von dotierten Halbleitern 25
Bändermodell eines p-n-Übergangs 27
Chemisch-Mechanisches Polieren (CMP) 28
Warum Planarisierung nötig ist 28
Prinzip des CMP-Prozesses 28
STI-CMP 29
ILD-CMP 31
Wolfram-CMP 31
Kupfer-Damascene-CMP 32
Dishing, Erosion & Dummy-Fill 32
Endpoint-Detection & Prozesskontrolle 32
Mehrlagenverdrahtung 33
Mehrlagenverdrahtung 33
BPSG-Reflow 34
Reflowrückätzen 34
Chemisch mechanisches Polieren 35
Kontaktierung der Metallisierungsebenen 36

Anforderungen an die Metallisierung

Anforderungen an die Metallisierung

Bei der Metallisierung werden Kontakte zu den dotierten Gebieten von Halbleiterbauelementen hergestellt, diese werden mit Leiterbahnen verbunden. Von dort werden die Anschlüsse zum Rand der einzelnen Chips geführt um die Verbindung zum Gehäuse herzustellen oder zur Kontrolle der Schaltung mit Messsonden während der Fertigung.

Welche Voraussetzungen müssen die Metallisierungsebenen erfüllen um in der Halbleitertechnik eingesetzt werden zu können?

- gute Haftung auf Siliciumdioxid
- hohe Strombelastbarkeit, geringer elektrischer Widerstand
- geringer Kontaktwiderstand zwischen Metall und Halbleiter
- möglichst einfacher Prozess zum Aufbringen der leitfähigen Schicht
- geringe Korrosionsanfälligkeit für lange Lebensdauer der Chips
- gute Kontaktierbarkeit
- Möglichkeit der Mehrlagenverdrahtung zur Einsparung von Chipfläche
- hohe Reinheit des Materials
- Beständigkeit gegen Elektromigration bei hohen Stromdichten
- Verträglichkeit mit dem chemisch mechanischen Polieren, mit dem jede Ebene eingeebnet wird
- keine Diffusion in die umgebenden Isolationsschichten oder in das Silicium

Die letzten drei Punkte sind erst mit den kleinen Strukturen wichtig geworden, und gerade sie haben zum Materialwechsel geführt. Aluminium erfüllt die klassischen Anforderungen gut und war über Jahrzehnte das Standardmaterial. Mit sinkenden Leiterbahnquerschnitten steigen jedoch die Stromdichten, und Aluminium wandert unter dieser Belastung. Seit Ende der 1990er Jahre ist deshalb Kupfer das Material der Verdrahtung in leistungsfähigen Schaltungen. Aluminium ist damit nicht verschwunden: Bondpads, Leistungs- und Analogbauelemente sowie unkritische Ebenen werden weiterhin damit ausgeführt.

Aluminiumtechnologie

Aluminium und Aluminiumlegierungen

Aluminium und seine Legierungen waren über Jahrzehnte das Standardmaterial

für die Verdrahtung auf dem Chip und werden weiterhin dort eingesetzt, wo es nicht auf kleinste Strukturen ankommt: für Bondpads, in der Leistungs- und Analogtechnik und auf unkritischen oberen Ebenen. Aluminium bietet:

- eine gute Haftung auf SiO_2 und Zwischenoxiden wie BPSG und PSG,
- eine gute Kontaktierbarkeit bei der Verdrahtung zum Gehäuse (z.B. mit Gold- und Aludrähten),
- einen niedrigen spezifischen Widerstand von 2,7 µΩ·cm für reines Aluminium, bei den üblichen Legierungen mit Silicium und Kupfer etwa 3 µΩ·cm,
- und ist sehr gut in Trockenätzverfahren strukturierbar.

Der letzte Punkt war lange ein entscheidender Vorteil: Aluminium bildet mit Chlor flüchtige Verbindungen und lässt sich deshalb wie jede andere Schicht ganzflächig abscheiden, mit Lack maskieren und ätzen. Genau dieses einfache Verfahren steht bei Kupfer nicht zur Verfügung.

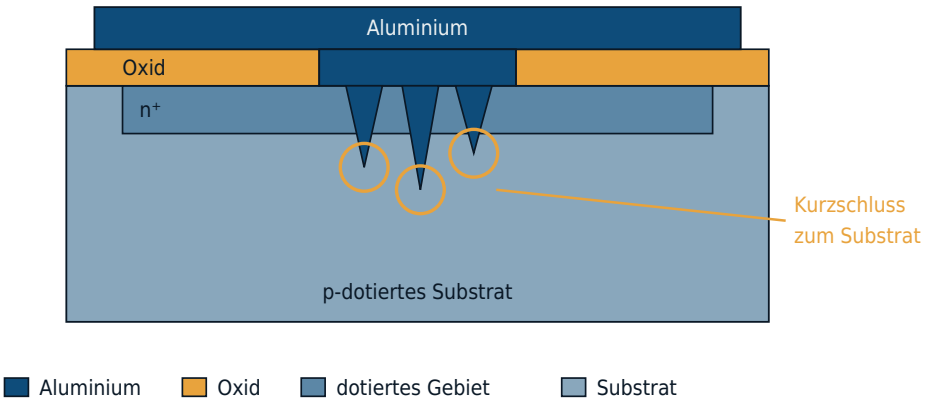
Aluminium erfüllt allerdings die Anforderung an elektrische Belastbarkeit und Korrosionsbeständigkeit nur teilweise. Die folgenden Abschnitte beschreiben die drei Effekte, an denen die Aluminiumtechnik ihre Grenze findet: die Diffusion von Silicium in das Metall, die Elektromigration und das Wachstum von Hügeln. Metalle wie Silber oder Kupfer sind in dieser Hinsicht deutlich besser, lassen sich aber nicht trockenätzen – für sie musste erst ein anderes Strukturierungsverfahren gefunden werden.

Siliciumdiffusion

Bei der Verwendung von reinem Aluminium kann es zu einer Diffusion von Siliciumatomen in das Metall kommen. Silicium löst sich schon weit unterhalb des Schmelzpunkts im festen Aluminium; bei den üblichen Sintertemperaturen von 400–450 °C, mit denen der Kontakt nach dem Aufbringen des Metalls verbessert wird, geschieht das in erheblichem Umfang. Der Halbleiter reagiert dabei mit der Aluminiummetallisierung und hinterlässt durch den Materialschwund, verursacht durch die ausdiffundierten Atome, Gruben an der Kontaktfläche zwischen Silicium und Aluminium. Das Aluminium füllt diese Gruben auf. Dadurch entstehen "Spikes" die unter Umständen zu einer Kurzschlussbildung führen, wenn sie durch die dotierten Gebiete bis in den Siliciumkristall hineinragen.

Spikebildung

Die Aluminiumspitzen durchstoßen das dotierte Gebiet

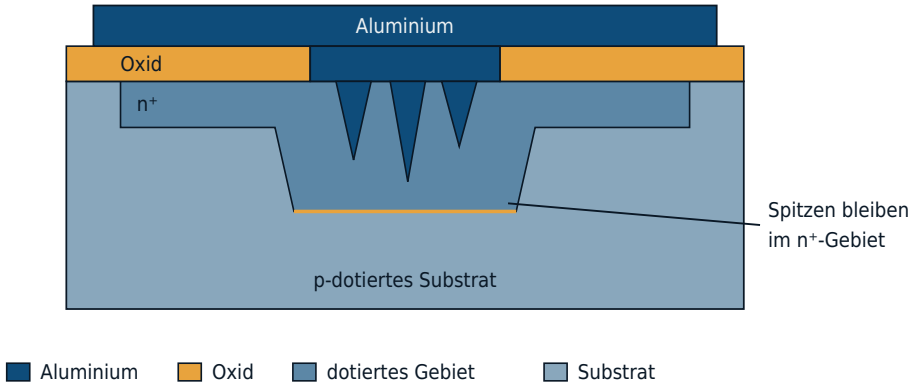


Beim Sintern löst sich Silicium im Aluminium und hinterlässt Gruben, die das Metall auffüllt. Reichen die Spitzen tiefer als das dotierte Gebiet, ist der Übergang kurzgeschlossen.

Die Größe dieser Spikes ist von der Temperatur abhängig mit der das Aluminium auf dem Silicium aufgebracht wird. Um diese Spikes zu verhindern gibt es mehrere Möglichkeiten. An der Stelle des Kontaktlochs kann eine tiefe Ionenimplantation, die Kontaktimplantation, eingebracht werden. Somit ragen die Spikes nicht bis in das Substrat hinein.

Kontaktimplantation

Das tiefer dotierte Gebiet fängt die Spitzen ab



Die tiefe Implantation unter dem Kontaktloch verlegt den Übergang nach unten. Sie kostet einen zusätzlichen Prozessschritt und ändert die elektrischen Eigenschaften des Gebiets.

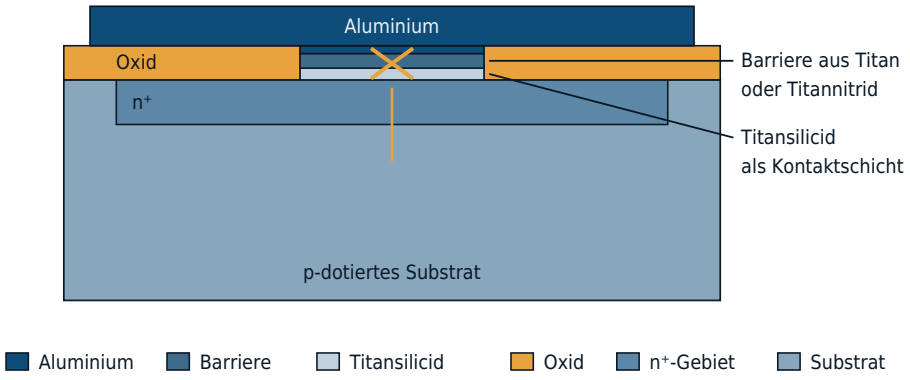
Der Nachteil dabei ist jedoch, dass ein weiterer Prozessschritt eingeführt werden muss, und sich die elektrischen Eigenschaften durch die Vergrößerung des dotierten Gebietes ändern.

Anstelle des reinen Aluminiums kann auch eine Aluminium-Silicium-Legierung verwendet werden die ca. 1–2 % Siliciumanteil enthält. Das Aluminium ist nun bereits mit Silicium versetzt und es diffundieren keine Siliciumatome mehr aus dem Wafer in das Aluminium. Bei sehr kleinen Kontaktlöchern kann jedoch Silicium an der Kontaktfläche ausfallen, was in einem vergrößerten Kontaktwiderstand resultiert.

Für hochwertige Kontakte ist eine Trennung zwischen Aluminium und Silicium erforderlich. Dazu bringt man auf dem Silicium eine Barriere aus verschiedenen Stoffen, wie z.B. Titan, Titannitrid oder Wolfram auf. Damit eine Erhöhung des Kontaktwiderstands an der Grenzschicht zwischen Titan und Silicium unterbunden wird, muss hier noch eine Kontaktschicht aus Titansilicid aufgebracht werden.

Barriereschicht zwischen Aluminium und Substrat

Aluminium und Silicium berühren sich nicht mehr



Der gesperrte Pfeil steht für die unterbundene Diffusion: Die Barriere hält das Silicium zurück, das Silicid hält den Kontaktwiderstand niedrig. Ohne diese Zwischenlage wäre der Übergang von Titan auf Silicium zu hochohmig.

Diese Barriere ist der Vorläufer dessen, was in der Kupfertechnik zur Regel wurde: Dort umschließt eine Diffusionsbarriere jede einzelne Leitbahn. Die Spikebildung selbst ist damit ein historisches Problem – der Kontakt zum Silicium wird heute über ein Silicid und einen Wolfram-, Cobalt- oder Rutheniumpfropfen hergestellt, Aluminium berührt das Silicium nicht mehr.

Elektromigration

Bei hohen Stromdichten (Stromfluss pro Fläche) übertragen die strömenden Elektronen bei jedem Stoß etwas Impuls auf die Atomrümpfe. Einzeln ist dieser Anstoß bedeutungslos, in der Summe schiebt dieser "Elektronenwind" die Atome jedoch langsam in Stromrichtung von ihren Plätzen weg. Besonders an Engstellen – etwa dort, wo Leiterbahnen über eine Kante laufen – ist die

Stromdichte erhöht; der Querschnitt nimmt weiter ab, die Stromdichte steigt, und im Extremfall reißt die Aluminiumbahn ab.

Schäden auf Leiterbahnen durch Elektromigration

(Quelle: Max-Planck-Institut für Metallforschung Stuttgart)

Die Elektromigration begrenzt damit, wie viel Strom eine Leitbahn dauerhaft führen darf. Da mit jeder Verkleinerung der Querschnitt sinkt, der zu schaltende Strom aber nicht im gleichen Maß, ist sie über die Jahre eher wichtiger geworden. Für Aluminium war sie einer der Gründe, zur Kupfertechnologie zu wechseln.

Zur quantitativen Beschreibung – Black-Gleichung, zulässige Stromdichten, Aktivierungsenergien – siehe Elektromigration, TDDB und NBTI/PBTI →

Hillockwachstum

Durch die Elektromigration wird Material verschoben und an Stellen geringer Stromdichte angehäuft. Diese so genannten Hillocks (Hügel) können darüber liegende Schichten durchbrechen und so zu einem Kurzschluss mit einer anderen Metallisierungsebene führen, außerdem kann Feuchtigkeit durch die Risse eindringen und zu Korrosion führen. Hillocks entstehen aber auch auf Grund unterschiedlicher Ausdehnungskoeffizienten der Materialien. Die Stoffe dehnen sich bei Temperaturänderungen unterschiedlich aus, und es entstehen Spannungen zwischen den Schichten. Mit Ausgleichsschichten, die einen "mittleren" Ausdehnungskoeffizienten haben kann dieses Problem gelöst werden (z.B. Titan, Titannitrid).

Weitere negative Effekte die bei der Metallisierung auftreten können:

- Streubelichtung: durch Unebenheiten können bei der Belichtung einfallende Lichtstrahlen in der Art reflektieren werden, so dass Bereiche unkontrolliert belichtet werden. Mit Hilfe einer Antireflexschicht (engl. anti-reflective coating, ARC) wird die Reflexion verhindert; auf Metall dient dazu meist Titannitrid, auf anderen Schichten Siliciumoxinitrid oder ein organischer Lack
- Schlechte Kantenabdeckung: an Kanten kann es zu vermehrtem, in Ecken zu vermindertem Aufwachsen der Schicht kommen. Um dem entgegenzuwirken, können Kanten vor der Metallabscheidung verrundet werden:

Das Design der Leiterbahnen muss also exakt geplant werden um diesen Fällen vorzubeugen. Durch einen geringen Zusatz an Kupfer kann die Lebensdauer der Aluminiumleiterbahnen stark erhöht werden, jedoch wird die Strukturierbarkeit

von mit Kupfer versetztem Aluminium wesentlich erschwert. Zum Schutz vor Korrosion werden die Oberflächen mit Siliciumdioxid oder Siliciumnitrid passiviert. Die Passivierung ist dabei der eigentliche Schutz: Der ganz überwiegende Teil der Bausteine wird in Kunststoff vergossen. Keramikgehäuse bleiben Anwendungen vorbehalten, die dicht abgeschlossen sein müssen, etwa in der Luft- und Raumfahrt. Verfahren wie Metallisierungsschichten auf dem Wafer aufgebracht werden, sind im Kapitel Abscheidung näher beschrieben.

Kupfertechnologie

Kupfertechnologie

Ab einer Strukturgröße von weniger als 250 nm erfüllt Aluminium auch mit Kupferanteilen kaum noch die benötigten Anforderungen zur Verwendung in integrierten Schaltkreisen. Der spezifische Widerstand von Kupfer liegt mit 1,7 µΩ·cm rund ein Drittel unter dem von Aluminium mit 2,7 µΩ·cm. Bei gleichem Querschnitt fällt an einer Kupferleitbahn also weniger Spannung ab, sie erwärmt sich weniger und die Signale laufen schneller. Auch die Stress- und Elektromigration sind in Aluminium wesentlich stärker ausgeprägt: Kupfer hat den höheren Schmelzpunkt und die stärker gebundenen Atome, es verträgt damit deutlich höhere Stromdichten. Ein Umstieg auf Kupfer war bei der stetigen Miniaturisierung somit unausweichlich.

Kupfer hat jedoch die negative Eigenschaft, dass es nahezu alles mit dem es in Kontakt kommt kontaminiert. Im Silicium ist es außerordentlich beweglich und erzeugt dort Störstellen, die Ladungsträger einfangen und Bauelemente unbrauchbar machen. Die Bereiche und Anlagen in der Fertigung, in denen mit Kupfer gearbeitet wird, müssen deshalb von den anderen strikt abgetrennt werden, und jede Kupferleitbahn wird von einer Diffusionsbarriere umschlossen. Zudem korrodiert Kupfer sehr leicht und muss deshalb, wie auch Aluminium, mit einer Passivierungsschicht versiegelt werden.

Während bei Aluminium eine Durchkontaktierung der Ebenen mit Wolfram geschieht, wird bei Kupfer das Metall selbst dazu verwendet (lediglich die erste Schicht beim Kontakt der dotierten Siliciumgebiete erfordert eine Trennung mit Wolfram). So entfallen nicht nur die negativen thermoelektrischen Effekte die an den Übergängen zwischen verschiedenen Metallschichten entstehen, sondern auch die zusätzlichen Arbeitsschritte zum Aufbringen mehrerer Schichten. Im Gegensatz zu Aluminium lässt sich Kupfer jedoch nicht in Trockenätzverfahren strukturieren, weil seine Halogenverbindungen bei Prozesstemperatur nicht flüchtig sind.

Das "gewöhnliche", subtraktive Verfahren zur Strukturierung einer Schicht, wie es auch bei Aluminium angewandt wird, verläuft im Wesentlichen in folgenden Schritten:

- zu strukturierende Schicht aufbringen
- Fotolack aufbringen, belichten und entwickeln
- Lackmaske in einem Ätzschritt in die darunterliegende Schicht übertragen
- Lack entfernen
- Passivierungsschicht aufbringen

Bei Kupfer bedient man sich eines additiven Verfahrens, dem so genannten Damascene-Prozess.

Damascene-Verfahren

Beim Damascene-Verfahren werden in bereits bestehende Zwischenschichten, die zur Isolierung bzw. Passivierung dienen, die Kontaktlöcher (VIA = Vertical Interconnect Access) der einzelnen Metallisierungsebenen geätzt, sowie die Gräben (Trenches), in denen später die Kupferleiterbahnen verlaufen sollen, strukturiert. In die Öffnungen kann dann mittels elektrochemischer Verfahren (Galvanotechnik) Kupfer abgeschieden werden. Ebenso sind auch CVD- oder PVD-Abscheidungen mit Reflowtechniken möglich. Anschließend wird das Kupfer in einem CMP-Prozess planarisiert, bis die Oberfläche eingeebnet ist.

Man unterscheidet zwischen Single- und Dual-Damascene-Prozessen und bei letzteren zwischen VFTL (VIA First Trench Last: VIA zuerst, Graben zuletzt) und TFVL (Trench First VIA Last: Trench zuerst, VIA zuletzt). Im Folgenden werden die zwei Dual-Damascene-Verfahren näher erläutert.

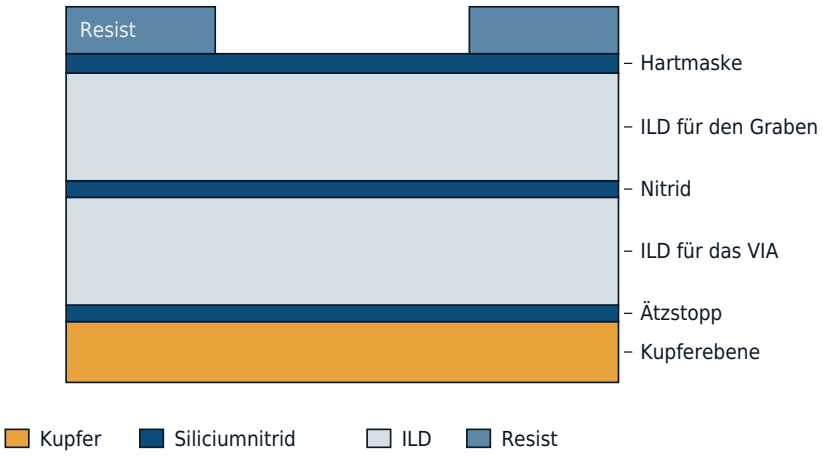
Trench First VIA Last

Auf dem Wafer (hier auf einer bereits bestehenden Kupferebene) werden verschiedene Materialien aufgebracht, die als Isolations-, Passivierungs- oder Schutzschicht dienen. Als Ätzstopp und zum Schutz vor Ätzgasen, kann Siliciumnitrid SiN zum Einsatz kommen. Als dielektrische Zwischenschicht (interlayer dieletric, ILD) kommen Materialien zum Einsatz, die einen geringen k-Wert haben, wie Siliciumdioxid SiO_2. Darüber wird eine Lackmaske strukturiert.

1. Der Wafer wird mit Resist beschichtet, der in einem Lithografieprozess strukturiert wird.

1. Lackmaske für den Graben

Auf dem Schichtstapel wird der Resist strukturiert

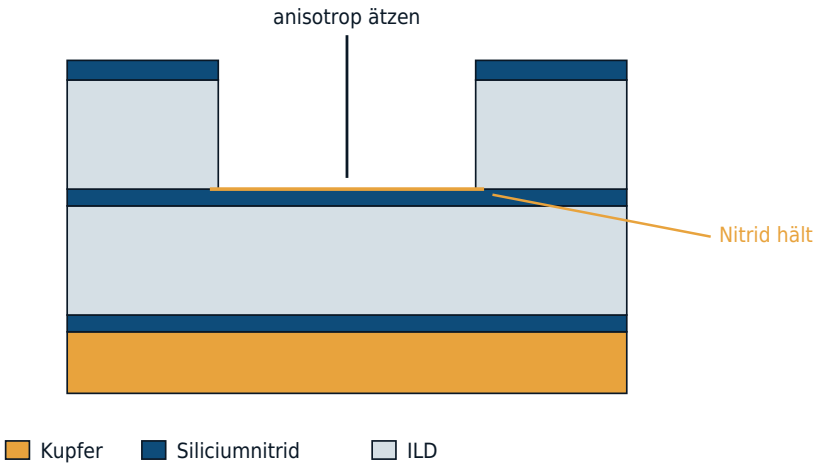


Es liegen zwei etwa gleich dicke ILD-Ebenen übereinander, getrennt durch eine Nitridschicht: In der unteren sitzt später das VIA, in der oberen der Graben.
Die Hartmaske schützt das ILD beim Lackentfernen und dient als CMP-Stopp.

2. In einem anisotropen Ätzprozess wird die Hartmaske (SiN) und die ILD-Schicht bis zum ersten Ätzstopp (ebenfalls SiN) geöffnet.

2. Graben ätzen

Hartmaske und oberes ILD werden bis zum trennenden Nitrid geöffnet



Die Ätzung läuft durch Hartmaske und oberes ILD und kommt auf dem trennenden Nitrid zum Stehen. Der Graben ist damit genau eine ILD-Ebene tief.

Der Fotoresist wird entfernt, und der Graben für die Leiterbahnen ist fertig strukturiert.

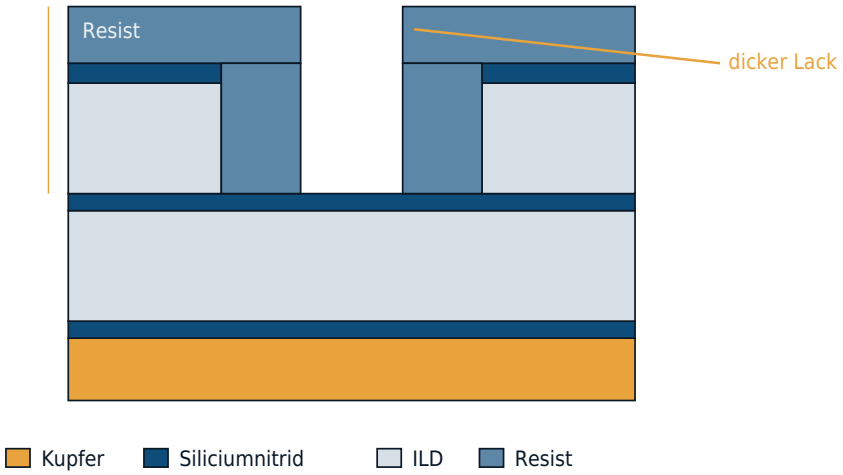
Die Hartmaske an der Oberfläche wird benötigt, um die ILD-Schicht während des Lackentfernens vor dem Plasma zu schützen. Dies ist notwendig, da das ILD chemisch ähnlich aufgebaut ist wie der Fotolack, und so durch die

gleichen Prozessgase angegriffen werden kann. Zusätzlich dient die Hartmaske als CMP-Stopp beim abschließenden Einebnen des Kupfers.

3. Als nächstes wird erneut Lack aufgebracht und strukturiert.

3. Lackmaske für das VIA

Der Lack füllt den Graben und muss deshalb dick aufgetragen werden

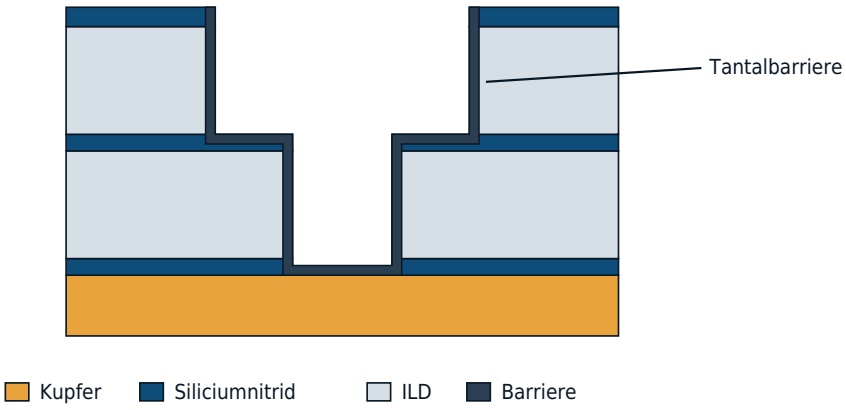


Weil der Lack den Graben mit auffüllt, ist er hier deutlich dicker als beim ersten Schritt. Ein kleines Kontaktloch darin zu belichten, ist der eigentliche Nachteil dieses Verfahrens.

4. Anschließend werden in einem anisotropen Ätzprozess die Kontaktlöcher (VIA) strukturiert.

4. VIA ätzen und Barriere abscheiden

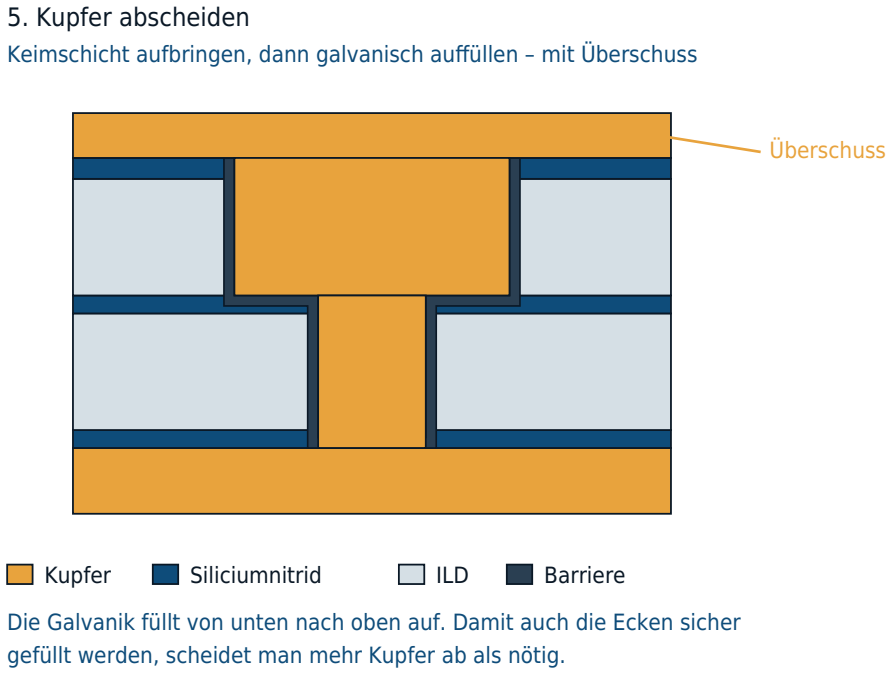
Durch Nitrid, unteres ILD und Ätzstopp bis auf das Kupfer



Das VIA durchläuft das trennende Nitrid, das untere ILD und den Ätzstopp. Der wird zuletzt und mit geringer Energie geöffnet, damit kein Kupfer herausgeschlagen wird. Die Barriere kleidet beide Öffnungen aus.

In einem niederenergetischen Ätzprozess wird dann die Ätzstoppschicht geöffnet, um kein darunterliegendes Kupfer herauszuschlagen, welches sonst in die ILD-Schicht eindiffundieren kann. Danach wird der Fotolack entfernt und eine dünne Barriereschicht aus Tantal abgeschieden die verhindert, dass das anschließend aufgebrachte Kupfer in das ILD eindringt.

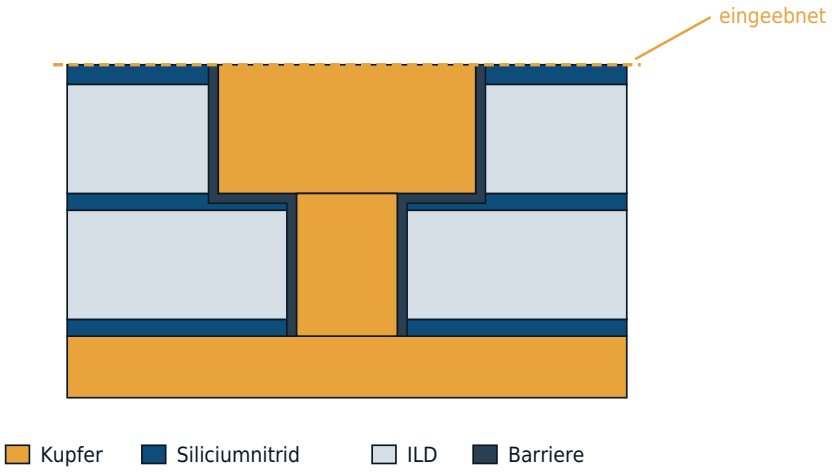
5. Eine dünne Kupferkeimschicht wird abgeschieden und die Strukturen in einem galvanischen Verfahren mit Kupfer aufgefüllt.



6. Das Kupfer wird abschließend durch chemisch mechanisches Polieren planarisiert.

6. Chemisch mechanisch polieren

Der Überschuss wird abgetragen, die Hartmaske dient als Stopp



Leiterbahn und Durchkontaktierung sind in einem Zug entstanden – das ist der Kern des Dual-Damascene-Verfahrens. Die nächste Ebene beginnt wieder bei Schritt 1.

Der größte Nachteil beim TFVL-Verfahren ist die dicke Lackschicht die nach dem Ätzen der Gräben aufgebracht werden muss (3.). Die winzigen Kontaktlöcher können in einer so dicken Lackschicht nur sehr schwer hergestellt werden. Das TFVL-Verfahren wird daher nur für die obersten Metallisierungsebenen verwendet, bei denen die Abmessungen der Strukturen nicht so kritisch sind, wie in den untersten Lagen.

VIA First Trench Last

Der VFTL-Prozess ähnelt dem TFVL-Prozess mit dem Unterschied, dass zuerst die Kontaktlöcher und danach die Gräben strukturiert werden.

1. Zunächst wird eine Lackmaske für die VIAs strukturiert und die Kontaktlöcher in einem anisotropen Ätzschritt bis zum untersten Ätzstopp geöffnet. Wichtig ist, dass der Ätzstopp nicht durchgeätzt wird, damit das darunter befindliche Kupfer nicht herausgeschlagen wird und in das ILD eindiffundiert.

1. VIA zuerst ätzen

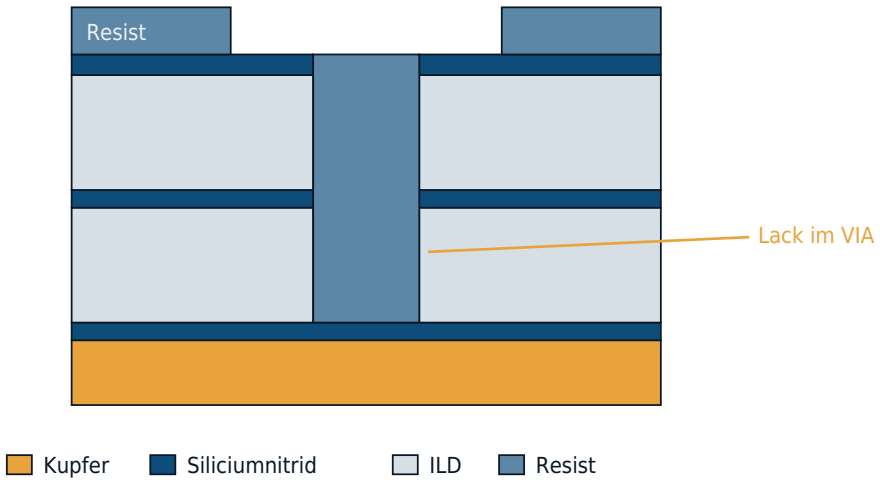
Eine schmale Lackmaske, das Loch reicht durch beide ILD-Ebenen

Das Loch geht gleich durch beide ILD-Ebenen. Sein unterer Teil bleibt später das VIA, der obere wird im nächsten Schritt zum Graben aufgeweitet.

2. Anschließend wird der Fotolack entfernt, und für die Gräben eine neue Lackmaske strukturiert; hierbei werden auch die bereits geöffneten VIAs mit Lack gefüllt.

2. Breite Lackmaske für den Graben

Der Lack füllt zugleich das VIA und schützt es

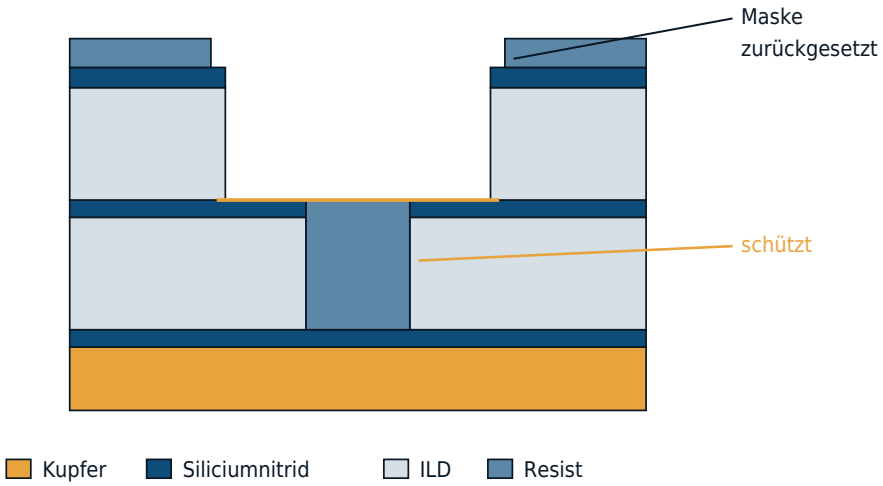


Der Lackpfropfen im Loch ist der Kern des Verfahrens: Ohne ihn würde die folgende Ätzung das VIA weiter vertiefen und den Ätzstopp durchschlagen.

3. Der unterste Ätzstopp wird durch den aufgebrachten Lack im VIA bei der anschließenden Grabenätzung vor den Ätzgasen geschützt wird.

3. Graben ätzen

Der Lack im VIA hält die Ätzgase von den Schichten darunter fern



Danach folgen wie beim anderen Verfahren das Öffnen des Ätzstopps, die Barriere, die Kupferfüllung und das Polieren.

Danach wird analog zum TFVL-Prozess der unterste Ätzstopp geöffnet, eine Tantalbarriere und die Kupferkeimschicht aufgebracht.

4. Als letztes folgt die Kupferabscheidung und die Planarisierung mittels CMP.

Beim Single-Damascene-Verfahren werden die Ebenen für VIAs und Gräben einzeln abgeschieden und strukturiert. Dadurch sind zwei Kupferprozesse (jeweils mit Abscheidung, Strukturierung und Planarisierung) notwendig.

Barriere und Liner

Die oben genannte Tantalbarriere ist in Wirklichkeit ein Schichtpaket. Auf das Dielektrikum kommt zunächst Tantalnitrid, das die Kupferdiffusion aufhält, darüber metallisches Tantal, auf dem die Kupferkeimschicht gut haftet und gleichmäßig aufwächst. Beide leiten deutlich schlechter als Kupfer und tragen zum Stromfluss praktisch nichts bei.

Solange die Leiterbahnen breit sind, fällt das nicht ins Gewicht. Bei den untersten Ebenen heutiger Prozesse ist die Barriere jedoch nicht mehr beliebig dünn zu machen, ohne ihre Sperrwirkung zu verlieren – sie beansprucht dann einen erheblichen Teil des Querschnitts. Die Antwort darauf ist ein Cobaltliner anstelle des Tantals, auf dem sich das Kupfer besser abscheiden lässt, und bei den feinsten Ebenen der Verzicht auf Kupfer zugunsten von Metallen, die ohne Barriere auskommen. Mehr dazu im Abschnitt Grenzen der Kupferverdrahtung.

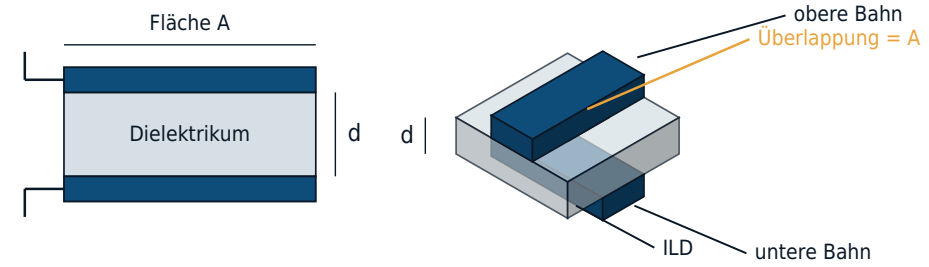
Low-k-Technologie

Durch die stetige Verkleinerung von Strukturen auf Mikrochips, um einerseits die Leistungsaufnahme zu verringern und andererseits maximale Schaltgeschwindigkeiten zu erzielen, rücken die Leiterbahnen zur Verdrahtung der einzelnen Bauelemente sowohl in vertikaler als auch lateraler Richtung immer näher zusammen. Um die Leiterbahnen von einander zu isolieren, müssen Schichten wie bspw. Siliciumdioxid SiO_2 als ILD aufgebracht werden.

Dort, wo Leiterbahnen parallel verlaufen oder sich auf übereinanderliegenden Metallisierungsebenen kreuzen, entstehen parasitäre Kapazitäten (Kondensatoren). Die Leiterbahnen bilden die leitenden Elektroden, das dazwischen liegende SiO_2 das Dielektrikum.

Plattenkondensator und Leiterbahnkreuzung

Zwei Leiterbahnen, die sich kreuzen, bilden einen Kondensator



Plattenkondensator

zwei Platten mit Isolator

Leiterbahnkreuzung

isometrisch, ILD halbtransparent

Leiterbahn bzw. Elektrode | Dielektrikum | wirksame Fläche

Beide Anordnungen gehorchen derselben Formel: Die Kapazität wächst mit der überlappenden Fläche und sinkt mit dem Abstand. Weil beide mit jeder Generation kleiner werden, hilft nur ein kleineres ε_r – ein Low-k-Material.

Die Kapazität C eines Kondensators berechnet sich nach:

$$C=\frac{\epsilon_r\epsilon A}{d}$$

Dabei steht d für den Abstand und A für die Fläche der Elektroden, also der sich überschneidenden Leiterbahnen. ε_0 beschreibt die absolute Dielektrizitätskonstante des Vakuums und ε_r (im Englischen häufig κ (Kappa) bzw. vereinfacht k) die Dielektrizitätszahl des Isolators (hier SiO_2).

Die Größe der parasitären Kapazität beeinflusst nun die elektrischen Eigenschaften wie die maximale Schaltgeschwindigkeit oder den Stromverbrauch des Chips, weshalb versucht wird C möglichst klein zu halten.

Dies ist theoretisch möglich durch eine Verringerung von ε_0, ε_r und A oder durch eine Erhöhung von d. Da d wie zu Beginn erläutert immer kleiner wird, A durch die elektrischen Anforderungen vorgegeben und ε_0 eine physikalische Konstante ist, ergibt sich, dass die Kapazität eines Kondensators im Wesentlichen nur durch eine Verringerung von ε_r gesenkt werden kann.

Man benötigt also Dielektrika mit *niedrigem* ε_r: low-k.

Das klassische Dielektrikum, SiO_2, hat eine Dielektrizitätszahl von ca. 4. Low-k beschreibt nun Materialien mit einem Wert $\varepsilon_r < 4$; Ultra-low-k-Materialien (ULK) mit $\varepsilon_r < 2{,}4$ sind in der Fertigung seit Jahren Stand der Technik. Die Dielektrizitätszahl gibt die Polarisation (Verschiebung von Ladungen innerhalb eines Isolators) im Dielektrikum an, und ist der Faktor, um den die Ladung einer Kapazität im Vergleich zu leerem Raum ansteigt oder um den das elektrische Feld im Kondensator abgeschwächt wird.

Um die Dielektrizitätszahl zu verringern, gibt es im Grunde zwei Ansätze:

- die Polarisierbarkeit von Bindungen im Dielektrikum verringern
- die Anzahl an Bindungen reduzieren

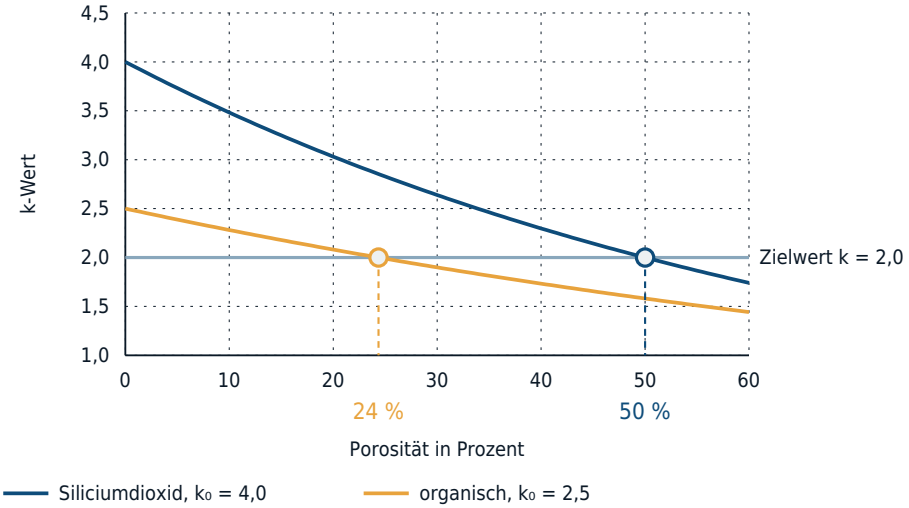
Die Polarisierbarkeit kann durch Stoffe mit wenig polaren Gruppen gesenkt werden, möglich sind fluorierte (FSG, ε_r ca. 3,6) oder organische (OSG) Siliciumoxide. Dies allein ist bei den immer kleiner werdenden Strukturgrößen jedoch nicht mehr ausreichend, weshalb der Trend zu porösen Schichten geht. Durch die Porosität befindet sich dann innerhalb des ILD "leerer Raum", der im Falle von Luft eine Dielektrizitätszahl von ca. 1 aufweist. Dadurch verringert sich ε_r für die gesamte Schicht. Die Poren können erzeugt werden, indem das ILD-Material mit Polymeren versetzt wird, die in einem Temperschritt aus der Schicht getrieben werden.

Jedoch ergeben sich einige Probleme die bewältigt werden müssen, um diese neuen Materialien in der Fertigung von Halbleitern einsetzen zu können.

Durch Poren im Material verringert sich dessen Dichte, was in geringerer mechanischer Stabilität resultiert. Im Falle von SiO_2 müssten ca. 50 % Poren im Material eingebracht werden, um einen k-Wert von 2,0 zu erreichen. Geht man von einem organischen Material aus, dessen k-Wert ohne Poren 2,5 beträgt, so genügt eine Porosität von ca. 22 %.

k-Wert in Abhängigkeit von der Porosität

Je niedriger der Ausgangswert, desto weniger Poren sind nötig



Berechnet nach der logarithmischen Mischungsregel k = k₀^(1−p):
Ein Material mit niedrigem Ausgangswert braucht deutlich weniger Poren.
Poren machen die Schicht jedoch mechanisch weicher und schwerer zu bearbeiten.

(Quelle: Semiconductor International)

Ebenso können Prozessgase oder Kupfer aus den Leiterbahnen leichter in die poröse Schicht eindringen und diese schädigen, wodurch die Dielektrizitätszahl wieder ansteigt. Um dem entgegenzuwirken müssen die Poren möglichst gleichmäßig verteilt sein und dürfen nicht miteinander verbunden sein. Um eine Diffusion von Kupfer in das ILD zu verhindern, muss in einem zusätzlichen Prozessschritt eine dünne Diffusionsbarriere aufgebracht werden, jedoch muss darauf geachtet werden, dass dieses Material die Dielektrizitätszahl nicht erhöht.

Ebenso wie der in der Halbleiterfertigung eingesetzte Fotolack, bestehen auch die organischen Siliciumoxide aus Kohlenwasserstoffgruppen (CH). Wird nun der Lack nach der Strukturierung des Dielektrikums in einem Sauerstoffplasma entfernt, greift das Plasma auch das Dielektrikum an. Auch hier müssen zusätzliche Schutzschichten (SiN im Abschnitt Damascene-Verfahren) aufgebracht werden die verhindern, dass die Isolationsschicht geschädigt wird.

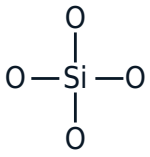
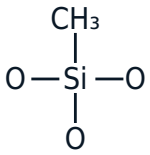
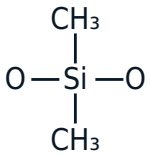
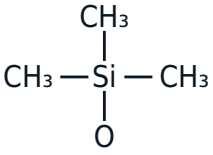
Wenn kein Material mehr hilft: Luftspalte

Am Ende dieser Entwicklung steht der Verzicht auf das Dielektrikum. Wird das Material zwischen zwei eng benachbarten Leitbahnen gezielt entfernt und der entstehende Spalt oben durch eine schlecht deckende Abscheidung überbrückt, bleibt ein Hohlraum zurück – mit einer Dielektrizitätszahl von 1 der

bestmögliche Isolator. Solche Luftspalte werden nur an ausgewählten Stellen eingesetzt, an denen die Kapazität besonders stört, denn sie schwächen den Aufbau mechanisch weiter und erschweren die Wärmeabfuhr.

Genau darin liegt die eigentliche Schwierigkeit der ganzen Low-k-Entwicklung: Jeder Schritt zu einem kleineren k-Wert macht die Schicht poröser und damit weicher. Sie muss aber das chemisch mechanische Polieren aushalten und später die Kräfte ertragen, die beim Aufbau des Gehäuses und bei jedem Temperaturwechsel im Betrieb auf den Chip wirken.

Übersicht verschiedener organischer Siliciumoxide

Chemische Formel	Strukturformel	k-Wert
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

Grenzen der Kupferverdrahtung

Der Umstieg auf Kupfer hat der Verdrahtung rund zwei Jahrzehnte verschafft. Inzwischen ist die Verdrahtung jedoch wieder zum Engpass geworden – und zwar aus einem Grund, den man dem Material nicht ansieht: Kupfer wird schlechter, je schmaler die Leitbahn ist.

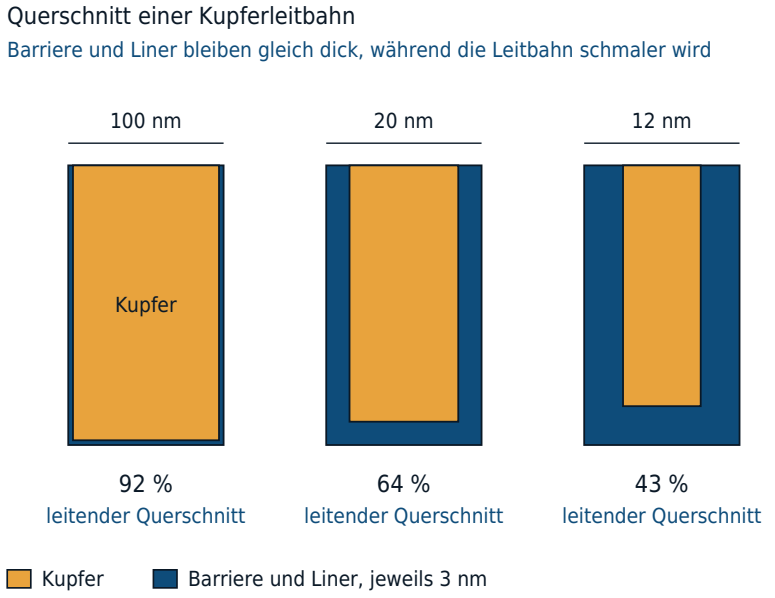
Der spezifische Widerstand ist keine Konstante mehr

Der Wert von 1,7 µΩ·cm gilt für ein ausgedehntes Stück Kupfer. Er beruht darauf, dass die Elektronen zwischen zwei Stößen eine gewisse mittlere freie Weglänge zurücklegen; in Kupfer sind das bei Raumtemperatur etwa 40 nm. Ist die Leitbahn schmaler als diese Weglänge, stoßen die Elektronen nicht mehr überwiegend im Kristall, sondern an den Seitenwänden und an den Korngrenzen. Beide Stoßarten kommen zum normalen Widerstand hinzu, und der Effekt wächst, je enger es wird. Bei Leitbahnbreiten um 20 nm liegt der wirksame spezifische Widerstand bereits um ein Mehrfaches über dem Literaturwert.

Die Barriere frisst den Querschnitt

Hinzu kommt ein rein geometrisches Problem. Jede Kupferleitbahn muss von einer Diffusionsbarriere umschlossen sein, und deren Dicke lässt sich nicht beliebig verringern, ohne dass sie ihre Sperrwirkung verliert. Während die Leitbahn schmaler wird, bleibt die Barriere also nahezu gleich dick und beansprucht einen immer größeren Anteil des Querschnitts – einen Anteil, der nichts zum Stromfluss beiträgt.

Querschnitt einer Kupferleitbahn bei drei Breiten



Beide Effekte wirken in dieselbe Richtung und verstärken einander: Der Kern wird kleiner, und das Kupfer in ihm leitet zugleich schlechter. Der Widerstand

einer Leitbahn steigt dadurch weit stärker an, als es die reine Querschnittsverkleinerung erwarten ließe.

Wege aus dem Problem

- Andere Metalle: Cobalt und Ruthenium haben einen höheren spezifischen Widerstand als Kupfer, aber eine kürzere freie Weglänge der Elektronen – sie verschlechtern sich bei kleinen Abmessungen also weniger stark. Vor allem kommen sie mit einer sehr dünnen oder ganz ohne Barriere aus. Unterhalb einer bestimmten Breite leitet eine Leitbahn aus diesen Metallen deshalb insgesamt besser als eine aus Kupfer, obwohl das Material für sich genommen schlechter ist. In den untersten Ebenen und in den Kontakten werden sie bereits eingesetzt.
- Weniger Kapazität statt weniger Widerstand: Da die Signallaufzeit vom Produkt aus Widerstand und Kapazität abhängt, hilft es auch, die Kapazität zu senken – über poröse Dielektrika und Luftspalte.
- Die Stromversorgung auf die Rückseite: Ein erheblicher Teil der Verdrahtung dient nicht der Signalübertragung, sondern der Versorgung mit Strom. Verlegt man diese Leitungen auf die Rückseite der Scheibe, die bisher nur Träger war, und führt sie durch das Substrat hindurch nach oben, so entlastet das die feinen Ebenen auf der Vorderseite gleich doppelt: Die Signalleitungen bekommen mehr Platz, und die Versorgungsleitungen dürfen auf der Rückseite deutlich breiter und damit widerstandsärmer ausfallen.

Keiner dieser Ansätze löst das Grundproblem, sie verschieben es. Anders als bei den Transistoren, wo die Verkleinerung die Bauelemente schneller macht, verschlechtert sie die Verdrahtung – und deren Anteil an Verzögerung und Verlustleistung wächst mit jeder Generation.

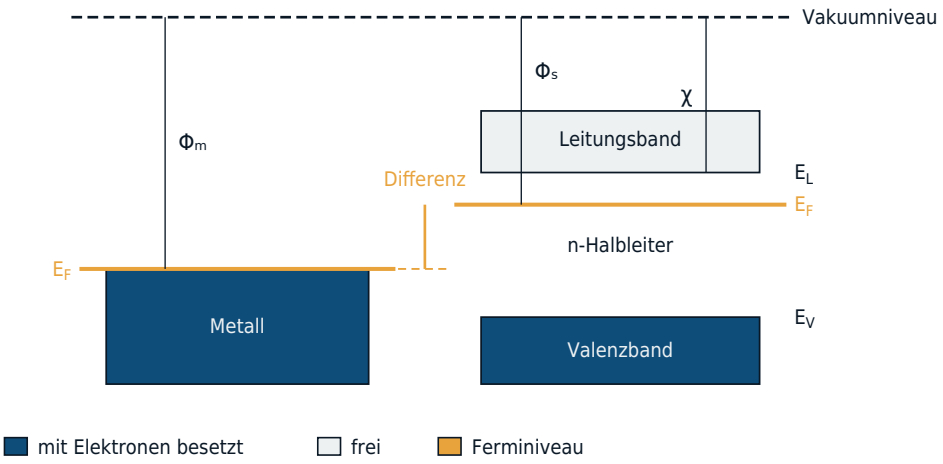
Der Metall-Halbleiter-Kontakt

n-Halbleiter-Kontakt

Ob und in welche Richtung beim Kontakt Elektronen fließen, bestimmt die Austrittsarbeit der beiden Materialien – die Energie, die nötig ist, um ein Elektron vom Ferminiveau aus dem Material zu lösen. Ist die Austrittsarbeit des Metalls größer als die des n-Halbleiters (wie z.B. bei Aluminium auf Silicium), liegt das Ferminiveau des Metalls energetisch tiefer: Bei der Kontaktierung fließen Elektronen aus dem Silicium in das Metall, da sie dort energetisch günstigere Zustände besetzen können. Im umgekehrten Fall – Austrittsarbeit des Metalls kleiner als die des n-Halbleiters – entsteht keine Barriere, sondern von vornherein ein ohmscher Kontakt.

Bändermodell vor dem Kontakt

Das Metall hat die größere Austrittsarbeit, sein Ferminiveau liegt tiefer



Φ ist die Austrittsarbeit, χ die Elektronenaffinität. Weil Φ_m größer ist als Φ_s, liegt das Ferminiveau des Metalls tiefer. Beim Kontakt fließen deshalb Elektronen aus dem Halbleiter in das Metall, bis beide Niveaus gleich sind.

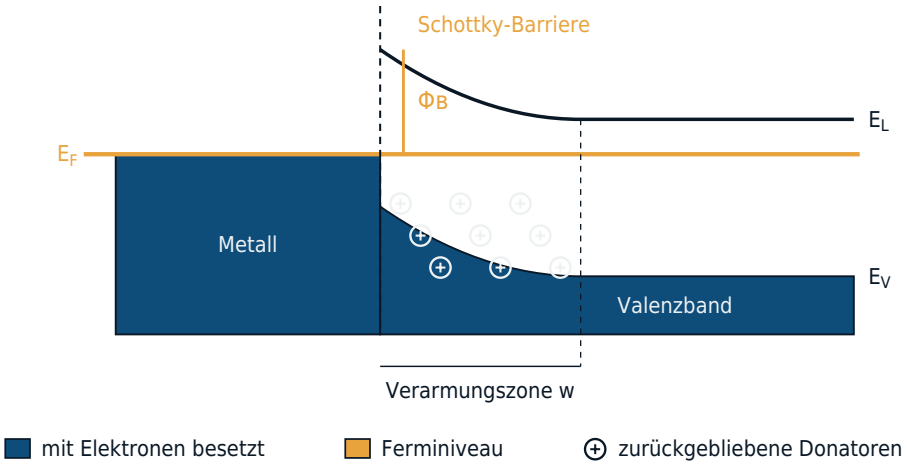
Es verringert sich somit die Aufenthaltswahrscheinlichkeit von Elektronen im Leitungsband des Halbleiters: Der Abstand zwischen Leitungsbandkante und Ferminiveau – welches den höchsten noch mit Elektronen besetzten Energiezustand beschreibt – wird an der Grenzfläche größer.

Durch die abgeflossenen negativen Ladungsträger bleiben positive Donatoren (z.B. Phosphorionen) zurück und es entsteht eine Raumladungszone. Die Verbiegung des Leitungsbandes veranschaulicht die Spannungsbarriere (Schottky-Barriere), welche die verbliebenen Elektronen im n-Leiter überwinden müssen, um in das Metall zu fließen.

Beim Kontakt von Metall und Halbleiter gleichen sich die Ferminiveaus also durch Diffusionsprozesse an, im Bereich der Grenzfläche ist das Ferminiveau konstant.

Bändermodell nach dem Kontakt

Die Ferminiveaus gleichen sich an, die Bänder verbiegen sich



Die abgeflossenen Elektronen hinterlassen positive Donatorrümpfe. Deren Raumladung verbiegt die Bänder – parabolisch, wie es die Poisson-Gleichung verlangt. Je stärker dotiert, desto schmaler wird w.

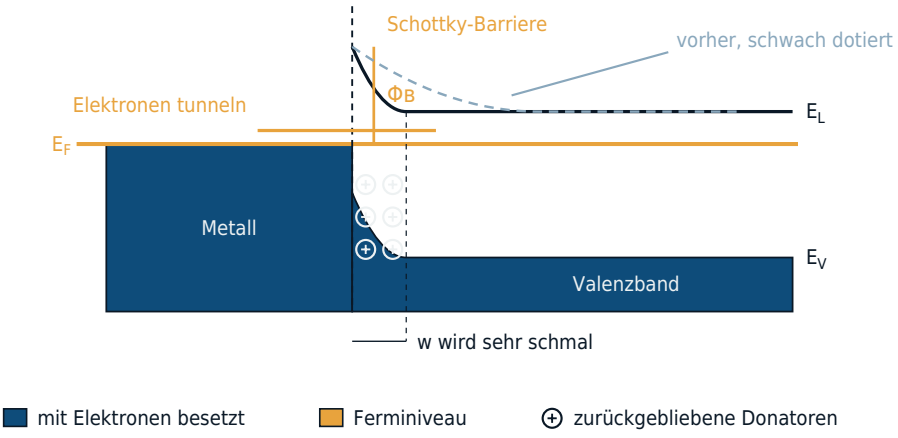
Die Weite *w* der Verarmungszone hängt von der Stärke der Dotierung ab. Die aus dem Halbleiter abgewanderten Elektronen erzeugen im Metall eine negative Raumladung, welche auf den Oberflächenbereich begrenzt ist.

Dieser Metall-Halbleiter-Kontakt weist eine nichtlineare Strom-Spannungscharakteristik auf, eine so genannte Schottky-Diode. Diese Barriere können die Elektronen durch Wärmeenergie von außen überwinden oder durch ein anliegendes elektrisches Feld "untertunneln" (nach der Quantentheorie kann ein Teilchen einen Bereich, in dem es aus energetischen Gründen nicht sein kann, überwinden, indem es sich, stark vereinfacht gesprochen, kurzzeitig Energie ausleiht, um die Barriere zu überwinden und die Energie dann wieder zurückgibt: der Tunneleffekt). Auch bei Aluminium kann dieser Effekt beobachtet werden. Da Aluminium an der Oberfläche immer oxidiert, hätten zwei aneinander liegende Aluminiumflächen eine isolierende Wirkung. Es ist jedoch ein Stromfluss zu verzeichnen, der auf dem Tunneleffekt beruht.

Je nach Anwendung will man diesen Dioden-Effekt herstellen oder aber verhindern. Um einen ohmschen Kontakt, also einen Kontakt ohne diese Potentialbarriere zu erzeugen, kann die Kontaktfläche stark dotiert werden (n^+-Dotierung), so dass die Verarmungszone sehr dünn wird und der Metall-Halbleiter-Kontakt in Folge des Tunneleffekts ein lineares Strom-Spannungsverhältnis aufweist.

Bändermodell nach n^+-Dotierung

Die Barriere bleibt gleich hoch, wird aber so dünn, dass Elektronen durchtunneln



Die starke Dotierung drängt die Raumladung auf wenige Nanometer zusammen. Die Barriere ist dadurch zwar unverändert hoch, aber so dünn, dass die Elektronen sie durchtunneln – der Kontakt wird ohmsch.

Da Aluminium im Silicium als Elektronenakzeptor eingebaut wird (es nimmt Elektronen auf) und sich so eine p-Dotierung an der Grenzfläche bildet, entsteht bei p-dotiertem Silicium ein ohmscher Kontakt. Bei einem n-dotierten Gebiet verursacht das Aluminium eine Dotierungsumkehr, so dass hier ein p-n-Übergang entsteht: eine Diode. Um diese zu vermeiden gibt es zwei Möglichkeiten:

- das n-dotierte Gebiet wird so stark dotiert, dass das Aluminium dieses nur abschwächt, nicht aber umkehrt
- eine Zwischenschicht aus Titan, Chrom oder Palladium verhindert die Umdotierung der n-dotierten Gebiete

Zur Verbesserung der Kontakte können auch Metallsilicide (Metalle in Verbindung mit Silicium) an der Kontaktfläche aufgebracht werden.

Im Gegensatz zur Diode beim p-n-Übergang, bei der die Schaltgeschwindigkeit auf der Diffusion von Elektronen beruht, haben Schottky-Dioden sehr kurze Schaltzeiten. Sie eignen sich daher als Schutzdioden, um Spannungsspitzen abzufangen.

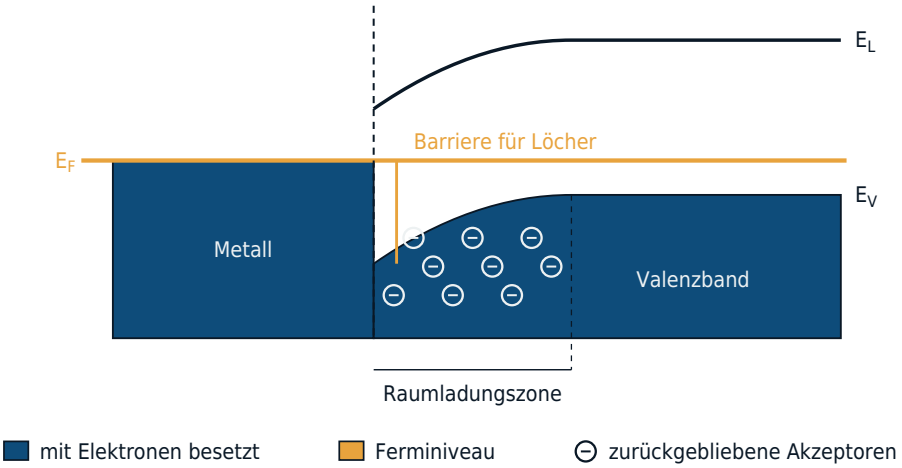
p-Halbleiter-Kontakt

Bei Metall-p-Halbleiter-Kontakten ergibt sich in Folge des Ladungsträgeraustauschs zwischen Metall und Halbleiter eine Bandverbiegung nach unten, Löcher im Halbleiter rekombinieren mit Elektronen aus dem Metall. Durch die Verringerung der Löcherkonzentration ergibt sich eine negative Raumladungszone im Halbleiterkristall, der Abstand zwischen Valenzbandkante

und Ferminiveau – repräsentativ für die höchsten Besetzungszustände durch Elektronen – wird an der Grenzfläche größer, und die so entstandene Potentialbarriere an der Valenzbandkante verhindert eine weitere Bewegung der Löcher, welche – komplementär zu Elektronen – die energetisch höchsten Zustände einnehmen wollen.

Bändermodell nach Metall-p-Halbleiter-Kontakt

Die Bänder verbiegen sich nach unten, die Barriere liegt am Valenzband



Löcher aus dem Halbleiter rekombinieren mit Elektronen aus dem Metall. Zurück bleiben negativ geladene Akzeptorrümpfe, deren Raumladung die Bänder nach unten zieht. Der Abstand vom Valenzband zum Ferminiveau wächst dadurch.

Ohne eine äußere Spannung kommen die Diffusionsprozesse zum Erliegen. Auch hier gleichen sich die Ferminiveaus im thermodynamischen Gleichgewicht einander an.

Kontaktierung von dotierten Halbleitern

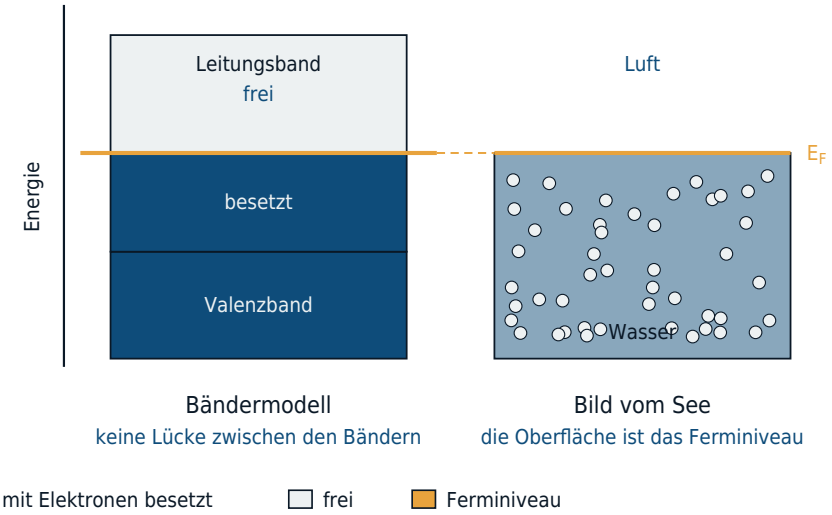
Nach der Herstellung der Transistoren im Siliciumsubstrat müssen diese mittels elektrischer Kontakte miteinander verbunden werden. Dabei wird zum einen die Gateelektrode zur Steuerung des Transistors kontaktiert, zum anderen müssen die dotierten Source- und Draingebiete, über die der Stromfluss erfolgt, angesteuert werden. Hier ergeben sich Probleme an den Kontaktflächen beim Übergang von Metallisierung zu Silicium, da je nach Dotierungstyp von Source und Drain Elektronenmangel (p-dotiert) oder Elektronenüberschuss (n-dotiert) vorherrscht.

Dabei spielt das Ferminiveau eine wichtige Rolle. Das Ferminiveau ist das Energieniveau, bis zu dem sich am absoluten Temperaturnullpunkt (-273,15 °C) noch Elektronen befinden. In Leitern befinden sich Elektronen im Valenzband

und im energetisch höheren Leitungsband, folglich ist das Ferminiveau auf Höhe des Leitungsbandes. Zur Veranschaulichung kann die Wasseroberfläche eines Sees betrachtet werden. Die Wassermoleküle darunter stellen die Elektronen dar, welche bis an die Oberfläche – das Ferminiveau – reichen.

Ferminiveau in Metallen

Beide Bänder grenzen ohne Lücke aneinander, besetzt ist alles bis zum Ferminiveau

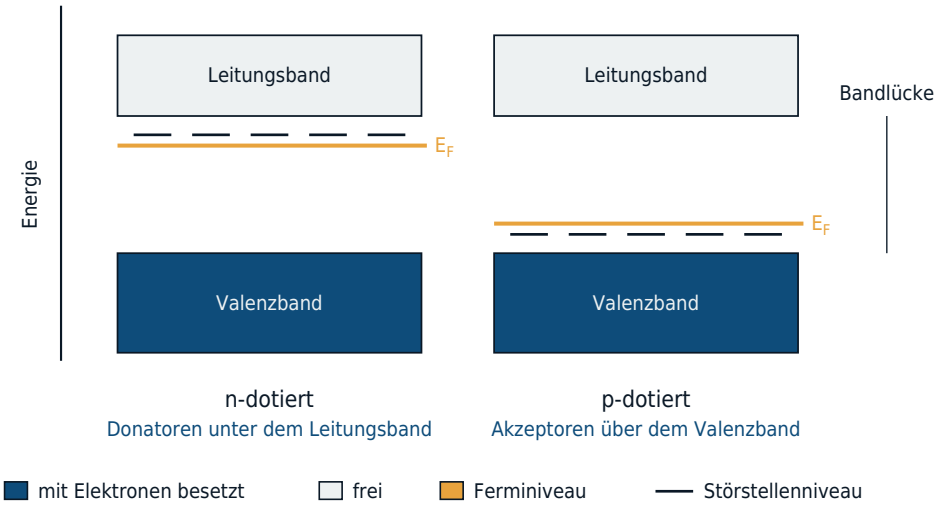


Am absoluten Nullpunkt sind alle Zustände bis zum Ferminiveau besetzt und alle darüber frei. Weil es beim Metall mitten im Leitungsband liegt, genügt eine winzige Energie, um Elektronen in freie Zustände zu heben – das Metall leitet.

In dotierten Halbleitern befinden sich Fremdatome als Donatoren oder Akzeptoren im Kristallgitter. In n-dotierten Halbleitern befindet sich das Ferminiveau in der Nähe der Leitungsbandkante, da die Donatoratome schon bei geringer Energiezufuhr freie Elektronen zur Verfügung stellen können. Dementsprechend befindet sich das Ferminiveau in einem p-dotierten Halbleiter in der Nähe der Valenzbandkante, da Elektronen aus dem Valenzband des Siliciumkristalls leicht vom Akzeptoratom aufgenommen werden können (vgl. Kapitel Dotieren).

Ferminiveau in dotierten Halbleitern

Anders als im Metall liegt zwischen den Bändern eine Lücke



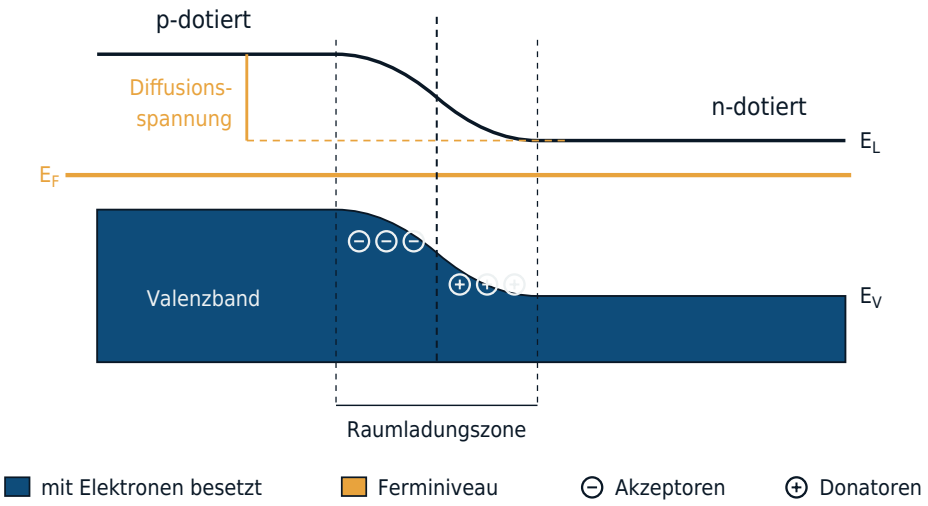
Das Ferminiveau rückt dorthin, wo die Ladungsträger leicht verfügbar sind: bei n-Dotierung nach oben zur Leitungsbandkante, bei p-Dotierung nach unten zur Valenzbandkante. Diese Lage bestimmt später den Kontaktwiderstand.

Bändermodell eines p-n-Übergangs

Aus der Tatsache, dass das Ferminiveau konstant sein muss (andernfalls würden Elektronen an Orte mit niedrigerem Ferminiveau fließen, dort freie Zustände besetzen und damit das Ferminiveau wieder anheben), ergibt sich auch beim p-n-Übergang eine Bänderverbiegung. Diese veranschaulicht die Raumladungszone, welche sich in Folge der abgewanderten Majoritätsladungsträger und der verbleibenden festen Dotieratome einstellt; also die Potentialschwelle, welche im Gleichgewichtszustand (ohne äußere Spannung) eine weitere Diffusion von Elektronen und Löchern in den jeweils anderen Kristall verhindert. Bei Silicium beträgt die Diffusionsspannung zum Überwinden des Potentialgefälles ca. 0,7 V.

Bändermodell am p-n-Übergang

Das Ferminiveau ist konstant, deshalb verbiegen sich die Bänder



Elektronen und Löcher wandern über den Übergang und rekombinieren. Zurück bleiben ortsfeste Dotieratome, deren Raumladung die Bänder verbiegt. Die entstehende Potentialschwelle bringt die Diffusion zum Erliegen; bei Silicium sind dafür etwa 0,7 V zu überwinden.

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,

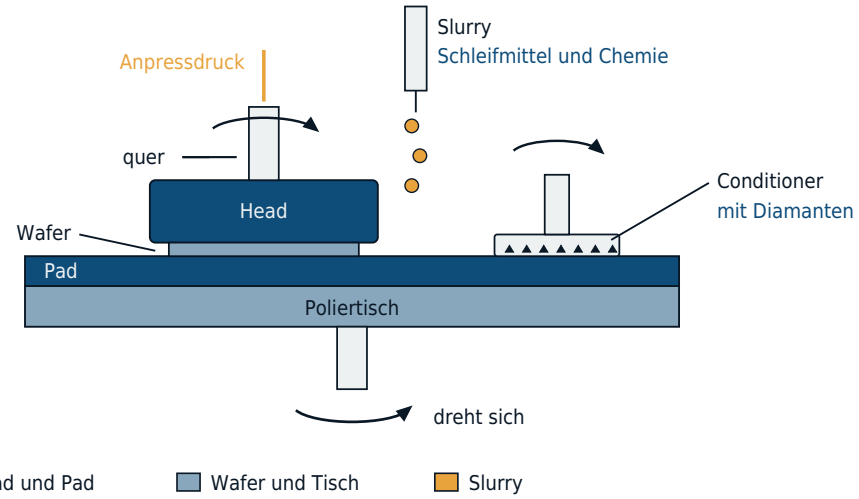
nicht nur eine lokale Glättung.

Der Wafer wird mit der aktiven Seite nach unten in einem Chuck mit Vakuumansaugung (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.

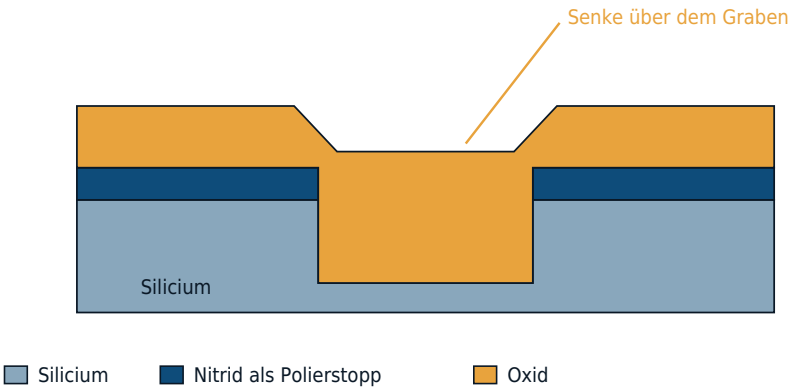
STI-CMP

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

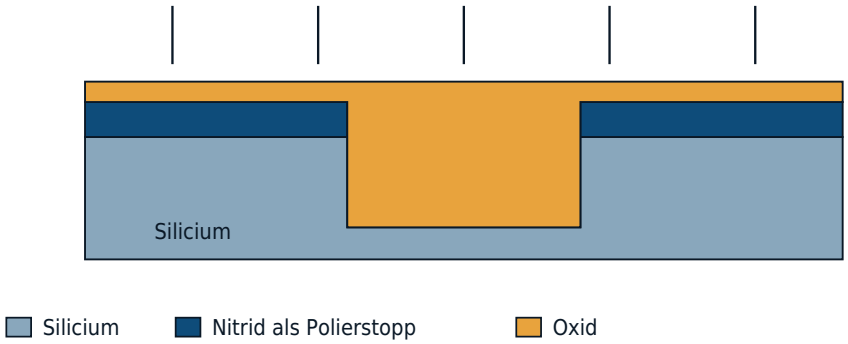
Das Oxid folgt der Topografie und liegt über dem Nitrid am höchsten



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

2. Erster Polierschritt

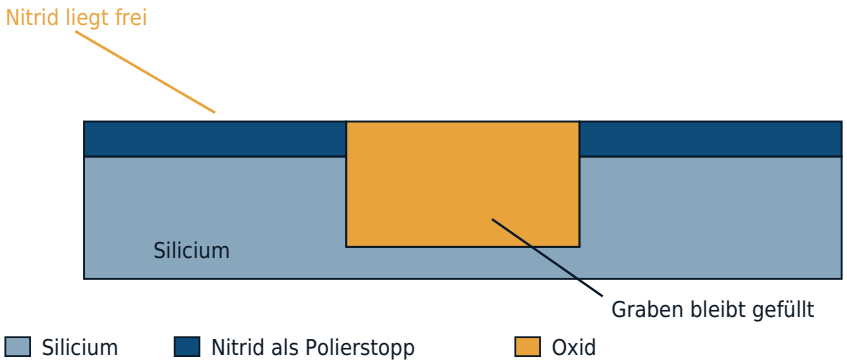
Die Oberfläche ist eben, über dem Nitrid liegt noch eine Restschicht



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 Barriereschicht 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 Barriereschicht (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 Metalldichte 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 Metalldicke 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.

Mehrlagenverdrahtung

Mehrlagenverdrahtung

Die Verdrahtung kann in einer integrierten Schaltung über 80 % der Chipfläche einnehmen, darum wurden Techniken entwickelt, mit denen man die Verdrahtung in mehrere Ebenen übereinander legt. So lässt sich die Summe der Leiterbahnen bei nur einer zusätzlichen Ebene um bis zu 30 % verringern. In einem heutigen Prozessor summieren sich die Leiterbahnen auf mehrere zehn Kilometer Gesamtlänge.

Zwischen den Verdrahtungsebenen sind Isolationsschichten aufgebracht, durch Kontaktöffnungen (VIA, vertical interconnect access) werden die einzelnen Ebenen miteinander verbunden. Gebräuchlich sind heute etwa zehn bis zwanzig Verdrahtungsebenen.

Diese Ebenen sind nicht gleich aufgebaut, sondern nach ihrer Aufgabe gestaffelt:

- Lokale Verdrahtung: Die untersten Ebenen verbinden benachbarte Transistoren über kurze Wege. Sie haben die feinsten Abmessungen, den größten Widerstand je Länge und werden mit den aufwendigsten Lithografieverfahren hergestellt.
- Mittlere Ebenen: Sie verbinden Funktionsblöcke innerhalb des Chips; Breite und Dicke der Leiterbahnen nehmen mit jeder Ebene zu.
- Globale Verdrahtung: Die obersten Ebenen führen Takt und Versorgungsspannung über den ganzen Chip. Sie sind um ein Vielfaches breiter und dicker als die untersten, weil sie große Ströme führen und dabei möglichst wenig Spannung abfallen darf.

Steile Kanten und Stufen müssen entschärft werden, da die Konformität der aufgebrachten Metallisierungen gering ist und somit Engstellen entstehen, die wiederum durch sehr hohe Stromdichten belastet werden. Folge: die Leiterbahnen altern frühzeitig oder reißen ab. Hinzu kommt, dass die Belichtung feiner Strukturen nur eine geringe Schärfentiefe hat und deshalb eine ebene Oberfläche voraussetzt. Um die Kanten bzw. Stufen zu entfernen gibt es mehrere Möglichkeiten zur Planarisierung, die im folgenden erläutert werden.

BPSG-Reflow

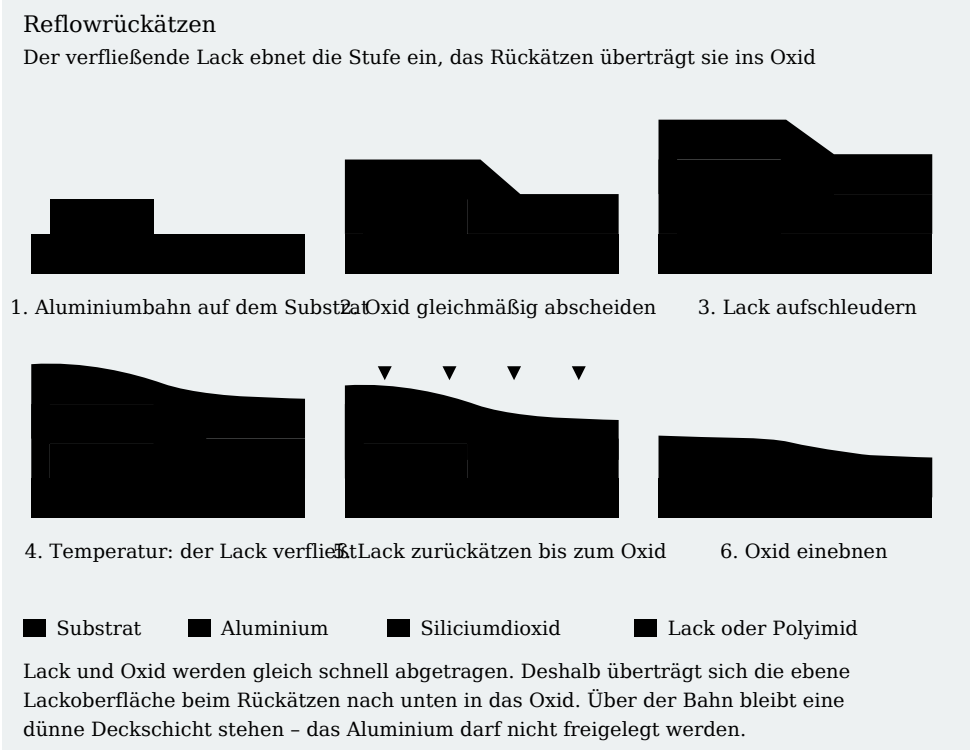
Bei der Reflowtechnik werden Schichten aus dotierten Gläsern auf dem Wafer aufgebracht. Weit verbreitet sind Phosphorsilicatglas (PSG) und Borphosphorsilicatglas (BPSG). In einem Hochtemperaturschritt verfließen die Gläser und bilden eine ebene Oberfläche: bei PSG und BPSG geschieht dies bei ca. 900 °C. Zur Planarisierung auf einer Verdrahtungsebene ist diese Technik jedoch nicht geeignet, da das Aluminium unter den hohen Temperaturen aufschmelzen würde.

Damit ist der Einsatzbereich eng: Das Verfahren kommt nur vor der ersten Metallebene in Frage, solange noch kein Metall auf der Scheibe liegt. Auch dort ist es weitgehend abgelöst worden, weil die Temperatur die bereits eingebrachten Dotierprofile verändert und weil das Verfließen nur örtlich einebnet – über die ganze Scheibe hinweg bleibt die Oberfläche uneben. Beides leistet das chemisch mechanische Polieren besser.

Reflowrückätzen

Auf dem Wafer wird eine Siliciumdioxidschicht aufgebracht, die mindestens so dick ist wie die höchste Stufe auf der Scheibe. Auf der Oxidschicht wird eine Lack- oder Polyimidschicht aufgeschleudert und zur Stabilisierung thermisch behandelt (siehe Fototechnik); durch die Temperatur verfließt die Schicht.

Im Trockenätzverfahren werden der Lack bzw. das Polyimid und das Siliciumdioxid mit gleichen Ätzraten abgetragen, so dass eine eingeebnete Oxidoberfläche zurückbleibt.



Neben der Technik mit Lack oder Polyimid kann auch so genanntes Spin On Glas (SOG) auf dem Wafer aufgebracht werden. Ebenfalls im Schleuderverfahren entsteht so direkt eine planarisierte Schicht, die durch eine thermische Behandlung stabilisiert wird. Die vorhergehende Siliciumdioxidschicht ist hierbei nicht erforderlich. Diese Techniken bieten jedoch keine Homogenität über den gesamten Wafer, sondern gleichen nur lokal Stufen aus.

Genau daran sind sie gescheitert. Sie glätten die Umgebung einer einzelnen Stufe, lassen aber die großräumigen Höhenunterschiede über die Scheibe hinweg bestehen – und gerade die stören die Belichtung feiner Strukturen. Beide Verfahren sind deshalb durch das chemisch mechanische Polieren abgelöst worden und heute nur noch von historischem Interesse.

Chemisch mechanisches Polieren

Nach dem Damascene-Verfahren wird die überschüssige Kupferschicht und die Barriere per chemisch-mechanischem Polieren (CMP) bis auf das Zwischenoxid abgetragen, sodass isolierte Leiterbahnen zurückbleiben. Da dieser Schritt auch bei jeder anderen Ebene des Wafers zum Einsatz kommt, wird er im eigenen Kapitel ausführlich behandelt.

Mehr im Kapitel Chemisch-Mechanisches Polieren (CMP) →

Kontaktierung der Metallisierungsebenen

Um die Metallebenen zu verbinden, werden in die Isolationsschichten Kontaktlöcher mit sehr hoher Anisotropie geätzt, so werden Kanten an den Kontaktlöchern vermieden. Die Kontaktlöcher müssen so aufgefüllt werden, dass einerseits eine optimale Kontaktierung gewährleistet wird und zugleich die Oberfläche planar bleibt.

Zur Auffüllung der Kontaktlöcher hat sich Wolfram als geeignet erwiesen. Unter Zugabe von Silan scheidet sich in einem CVD-Prozess aus Wolframhexafluorid eine dünne Schicht Wolfram als Nukleationskeim ab, Siliciumtetrafluorid und Fluorwasserstoff als Nebenprodukte werden abgesaugt:

$4\ WF_6 + 3SiH_4 \rightarrow 4W + 3SiF_4 + 12HF$

Unter Zugabe von Wasserstoff zu Wolframhexafluorid werden die Kontaktlöcher dann aufgefüllt:

$WF_6 + 3H_2 \rightarrow W + 6HF$

Der Vorteil des Verfahrens liegt darin, dass die Abscheidung aus der Gasphase auch enge, tiefe Löcher gleichmäßig von innen heraus füllt, während ein aufgedampftes oder gesputtertes Metall die Öffnung oben zuwachsen ließe und einen Hohlraum einschlösse. Wolfram wird deshalb ganzflächig abgeschieden und anschließend per CMP so weit zurückpoliert, dass es nur noch in den Löchern steht.

Darüber kann die nächste Metallebene aufgebracht, strukturiert und planarisiert werden. Bei Kupfer als Metallisierung wird Wolfram nur für den ersten Kontakt zum Siliciumsubstrat benötigt. Die Verbindung der einzelnen Kupferebenen geschieht mit dem Kupfer selbst.

Der Kontakt als Engstelle

Mit kleiner werdenden Abmessungen verschiebt sich das Problem. Der Widerstand eines Kontakts setzt sich aus dem Widerstand des Füllmetalls, dem der Barriere und dem Übergangswiderstand zum Silicium zusammen. Die ersten beiden Anteile wachsen mit schrumpfendem Querschnitt schnell an, weil Wolfram eine vergleichsweise dicke Haftschicht aus Titan und Titannitrid benötigt, die selbst kaum leitet. In den untersten Ebenen moderner Prozesse wird Wolfram deshalb durch Cobalt oder Ruthenium ersetzt: Beide haben zwar einen höheren spezifischen Widerstand als Wolfram, kommen aber mit einer sehr dünnen oder ganz ohne Haftschicht aus. Bei den kleinsten Kontakten leitet die Füllung dadurch insgesamt besser, obwohl das Material für sich genommen schlechter leitet.