Halbleitertechnologie von A bis Z

Grundlagen

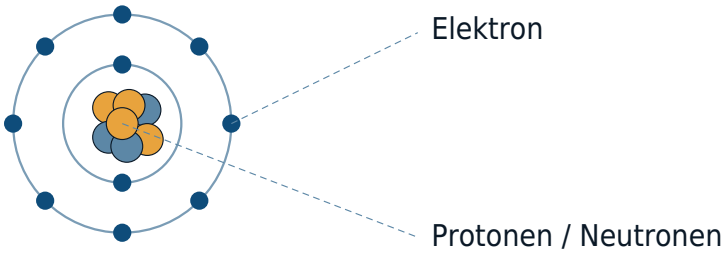
Der Atombau

Das Atommodell

Ein Atom [gr. átomos: unteilbar] ist der kleinste chemisch nicht weiter teilbare Grundbaustein der Materie. Es gibt verschiedene Atome, welche sich aus einer bestimmten Anzahl an Elektronen, Protonen und Neutronen zusammensetzen. Positiv geladene Protonen und ungeladene Neutronen bilden den Atomkern, den negativ geladene Elektronen in bestimmten Abständen umkreisen. Ein in der Natur vorkommendes Atom ist elektrisch neutral, d. h. es befinden sich genauso viele Protonen wie Elektronen im Atom. Die Anzahl der Neutronen kann dabei variieren.

Nach dem Bohrschen Atommodell befinden sich die Elektronen auf so genannten Schalen, welche unterschiedliche Energieniveaus repräsentieren und konzentrisch um den Atomkern angeordnet sind. Es gibt maximal sieben Schalen, die unterschiedlich viele Elektronen aufnehmen können, die Elektronen auf der äußersten Schale werden auch als Valenzelektronen bezeichnet.

Stark vereinfachte Darstellung eines Neonatoms



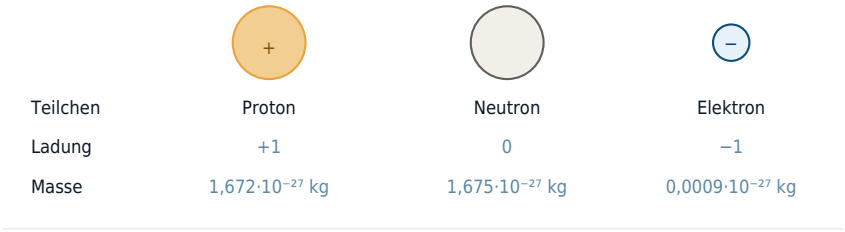
Durch den Austausch von Elektronen mit anderen Atomen, können Atome einen stabileren Zustand erreichen, weshalb es zur Bildung von unterschiedlichen Bindungen von Atomen kommt.

Größenordnungen

Masse
: Die Masse eines Atoms wird hauptsächlich vom Atomkern bestimmt, da die

Massen von Protonen und Neutronen mit $1{,}67 \cdot 10^{-27}$ kg rund 1800 Mal größer sind, als die der Elektronen in der Atomhülle ($9{,}11 \cdot 10^{-31}$ kg).

Atombausteine

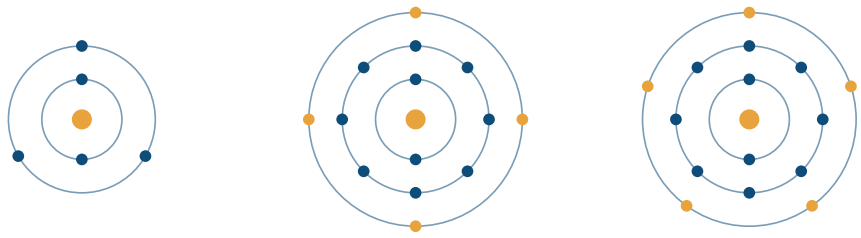


Teilchen	Proton	Neutron	Elektron
Ladung	+1	0	−1
Masse	$1{,}672 \cdot 10^{-27}$ kg	$1{,}675 \cdot 10^{-27}$ kg	$0{,}0009 \cdot 10^{-27}$ kg

Abmessungen
: Der Durchmesser der Atomhülle beträgt 0,1–0,5 nm (0,0000000001 m oder 1 Ångström), der Durchmesser des Atomkerns ist noch einmal um den Faktor 100.000 geringer. Zur Veranschaulichung soll ein Stecknadelkopf in der Mitte eines Fußballfeldes als Atomkern betrachtet werden. Dann entspricht die Entfernung zu den Eckfahnen dem Abstand, mit dem die Elektronen den Kern umkreisen.

Dichte
: Im Kern eines Atoms sind Protonen und Neutronen extrem dicht gepackt. Würde man die Erde auf die gleiche Dichte komprimieren, so würde ihr Radius von ursprünglich 6.700.000 m auf nur noch 100 m reduziert.

Atome, die eine große Bedeutung in der Halbleiterindustrie haben



Bor	Silizium	Phosphor
5p·6n·5e	14p·14n·14e	15p·15n·15e
3 Außenelektronen	4 Außenelektronen	5 Außenelektronen

Das Periodensystem der Elemente

Die Elemente

Ein Element besteht aus mehreren gleichen Atomen (das bedeutet, mit derselben Anzahl an Protonen = Kernladungszahl) und ist ein Stoff, der mit chemischen Mitteln nicht weiter zerlegt werden kann. Die Masse von Elementen wird nur durch die Anzahl von Protonen und Neutronen bestimmt, da die Elektronenmasse vernachlässigbar gering ist. Wasserstoff mit einem Proton und keinem Neutron hat die Massenzahl 1, das nächst schwerere Element, Helium, besitzt die Massenzahl 4 (2 Protonen + 2 Neutronen). Die Massenzahl gibt die Anzahl der Teilchen im Atomkern an. Multipliziert mit der atomaren Masseneinheit u = $1{,}660 \cdot 10^{-27}$ kg erhält man in etwa die Masse eines Atoms, für Helium $6{,}64 \cdot 10^{-27}$ kg.

Die chemischen Elemente werden meist mit den Anfangsbuchstaben ihrer lateinischen oder griechischen Namen benannt (Wasserstoff [H] von lat. hydrogenium, Lithium [Li] von gr. lithos).

Das Periodensystem der Elemente

Das Periodensystem der Elemente (kurz PSE) stellt alle chemischen Elemente mit steigender Protonenanzahl (Kernladung, Ordnungszahl) und entsprechend ihrer chemischen Eigenschaften eingeteilt in Perioden sowie Haupt- und Nebengruppen dar.

Die Periode gibt dabei die Anzahl der Elektronenschalen an, die Hauptgruppe die Anzahl der Elektronen auf der äußersten Schale (1 bis 8 Elektronen). Gruppe 1 und 2 sowie 13-18 bilden die Hauptgruppen, die Gruppen 3-12 die Nebengruppen.

Das erste Element mit einer Schale (Periode 1) und einem Außenelektron (Gruppe 1) ist Wasserstoff H. Das nächste Element, Helium He, besitzt wie Wasserstoff nur eine Elektronenschale und befindet sich somit ebenfalls in Periode 1. Da die erste Schale mit zwei Elektronen bereits vollständig gefüllt ist, steht Helium Nicht in Gruppe 2, sondern in Gruppe 18 (Gruppe der Edelgase).

Um weitere Elektronen aufnehmen zu können, muss eine neue Schale begonnen werden. Somit findet man Lithium in Gruppe 1, Periode 2 (zwei Elektronen auf der ersten Schale, ein Valenzelektron auf der zweiten Schale). Eine Schale kann maximal $2n^2$ Elektronen aufnehmen, wobei n für die Periode steht.

Nachdem in Gruppe 1 und 2 die ersten beiden Valenzelektronen auf der äußersten Schale besetzt wurden, werden ab der vierten Periode zunächst weiter innenliegende Schalen mit Elektronen vervollständigt, bevor die jeweils äußerste Schale der Gruppen 13 bis 18 vollständig mit Elektronen aufgefüllt wird.

Das Periodensystem der Elemente

Tippen Sie auf ein Element für Details.

Bedeutende Elemente in der Halbleitertechnologie

Element	Teilchen	Hinweis
B: Bor	5p, 6n, 5e	3 Außenelektronen: Zur p-Dotierung von Silicium
N: Stickstoff	7p, 7n, 7e	Stabiles N_2-Molekül: Schutz- und Spühlgas, Schutzschichten auf dem Wafer
O: Sauerstoff	8p, 8n, 8e	Sehr reaktionsfreudig: Oxidation von Silicium, Isolationsschichten (SiO_2) u.a.
F: Fluor	9p, 10n, 9e	Reaktionsfreudigstes Element: Wird in Verbindung mit anderen Stoffen zum Ätzen verwendet (z.B. HF, CF_4)
Si: Silicium	14p, 14n, 14e	Grundmaterial in der Halbleitertechnik
P: Phosphor	15p, 16n, 15e	5 Außenelektronen: Zur n-Dotierung von Silicium

Elemente, die sich links im Periodensystem befinden, sind Metalle. Diese haben das Bestreben, Valenzelektronen abzugeben um so eine stabile Elektronenkonfiguration (die Edelgaskonfiguration) zu erreichen. Rechts im Periodensystem stehen die Nichtmetalle, die zum Erreichen der Edelgaskonfiguration zusätzliche Elektronen aufnehmen. Dazwischen befinden sich die Halbmetalle wie Silicium und Germanium.

Chemische Bindungen

Bindungsbestreben

Elektronen, die sich auf der äußersten Schale befinden, können sich vom Atom lösen (z.B. durch Zuführen von Energie in Form von Wärme) und mit anderen Atomen ausgetauscht werden. Verbindungen von mehreren Elementen nennt man Moleküle. Der Grund für das Bindungsbestreben ist das Erreichen der mit acht Elektronen voll besetzten äußersten Schale: die so genannte Edelgaskonfiguration. Stoffe, die die volle Außenschale erreicht haben, gehen keine Verbindungen ein (einige wenige Ausnahmen wie z.B. Xenon-Fluor-Verbindungen sind möglich).

Dabei unterscheidet man hauptsächlich zwischen drei verschiedenen Arten von Bindungen, die im Folgenden näher erläutert werden.

Die Atombindung

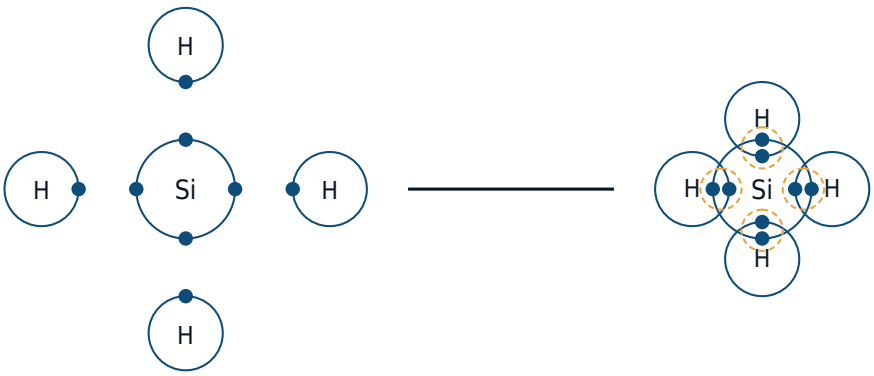
Nichtmetalle gehen diese Verbindung ein um das Elektronenoktett zu

vervollständigen. So können zwei Fluoratome (je sieben Außenelektronen) durch gegenseitigen Austausch eines Elektrons ihr Elektronenoktett auffüllen. Der Abstand zwischen den beiden Atomkernen repräsentiert einen Kompromiss zwischen der Anziehung von Atomkern und Bindungselektronen und der Abstoßung der beiden Atomkerne. Der Grund für Atombindungen ist das Bestreben der Natur den energetisch niedrigsten Zustand herzustellen. Da den Elektronen durch den Zusammenschluss mehrerer Atome zu einem Molekül "mehr Raum" zur Verfügung steht, was einer geringeren Energie entspricht, kommt es überhaupt erst zur Atombindung.

Aus dem Bindungsbestreben zum Erreichen der voll besetzten Außenschale ergibt sich, dass Fluoratome elementar niemals atomar, sondern immer als Fluormolekül auftreten: F_2. Dies gilt auch für Stickstoff (N_2), Sauerstoff (O_2), Chlor (Cl_2), Brom (Br_2) und Jod (I_2). Auf Grund der Elektronenpaare nennt man diese Bindung auch Elektronenpaarbindung oder kovalente Bindung.

In der folgenden Abbildung ist die Atombindung am Beispiel von Silan (SiH_4) dargestellt (beim Siliciumatom ist nur die äußerste Schale abgebildet). Das Siliciumatom erreicht durch die Bindung die volle Achterschale, die Wasserstoffatome die bereits mit zwei Elektronen voll besetzte erste Schale.

Atombindung



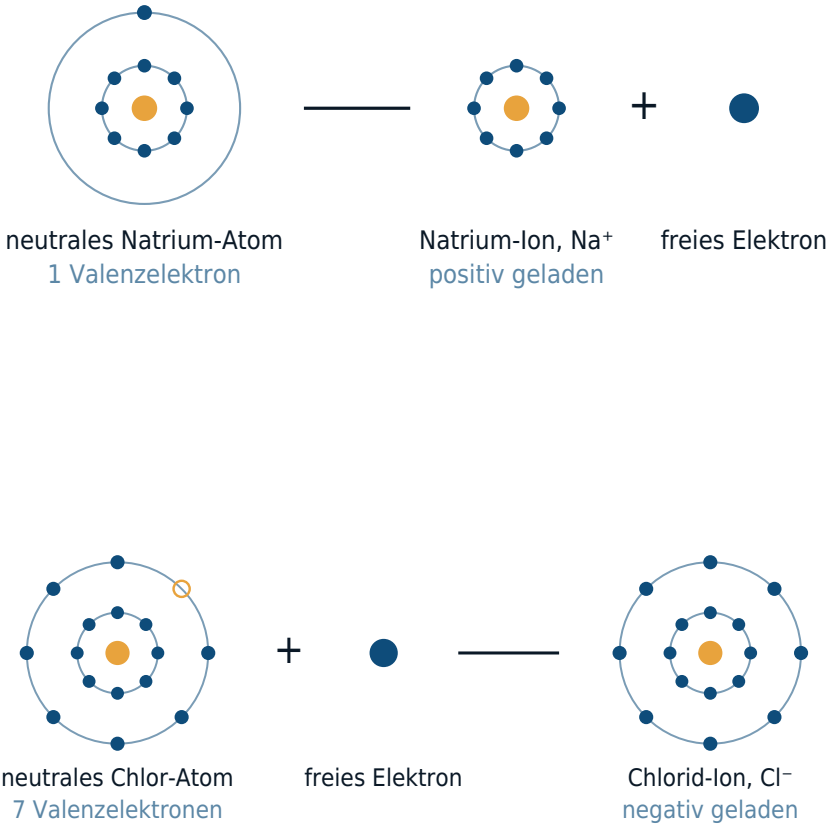
Vier gemeinsame Elektronenpaare vervollständigen die Bindung

Die Ionenbindung

Ionenbindungen entstehen durch den Zusammenschluss von Metallen und Nichtmetallen. Metalle haben das Bestreben, Elektronen abzugeben um eine vollständig gefüllte Außenschale zu erreichen, Nichtmetalle können dagegen zusätzliche Elektronen aufnehmen. Ein Beispiel für eine Ionenbindung ist Natriumchlorid (NaCl, Kochsalz).

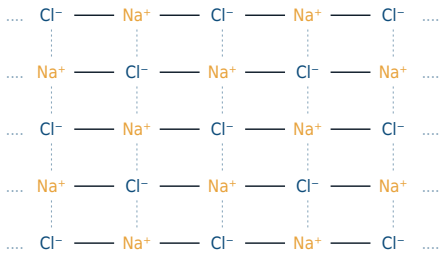
Das Natriumatom gibt sein Valenzelektron ab (damit besitzt es mehr Protonen als Elektronen und ist positiv geladen), Chlor nimmt ein Elektron auf und ist damit einfach negativ geladen. Durch die unterschiedlichen Ladungen ziehen sich die zwei Atome an. Ein geladenes Atom bezeichnet man als Ion, dabei unterscheidet man zwischen Kation (positive Ladung) und Anion (negative Ladung).

Prinzip der Ionenbindung



Da die Atome immer in einer sehr hohen Anzahl auftreten, richten sie sich dabei Dank der Anziehungs- und Abstoßungskräfte zu einem gleichmäßigen Ionengitter aus. Stoffe, die im festen Zustand ein solches Gitter bilden, bezeichnet man als Salze.

Kristallgitter bei der Ionenbindung von Natriumchlorid

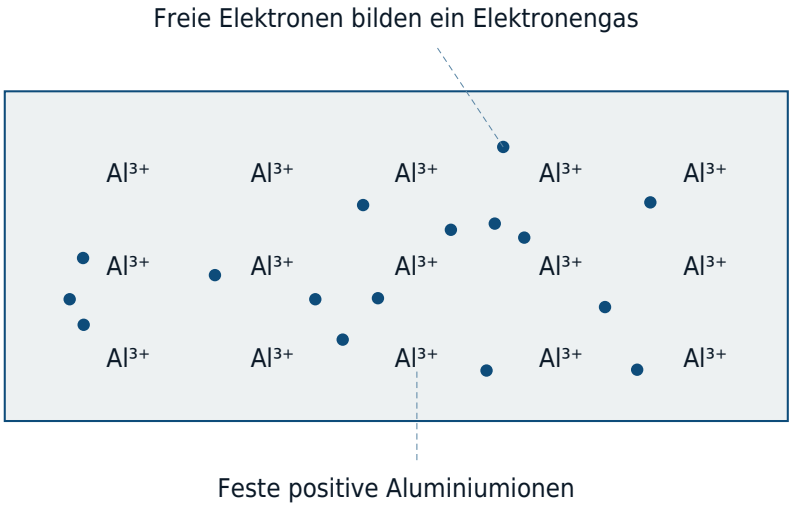


Die Metallbindung

Metalle gehen diese Bindung ein, um die stabile Edelgaskonfiguration zu erreichen. Dazu gibt jedes Metallatom seine Außenelektronen ab: es entstehen positiv geladene Metallionen (Atomrümpfe) und freie Elektronen zwischen denen starke Anziehungskräfte herrschen. Die Metallionen stoßen sich untereinander ebenso ab wie die Elektronen.

Da die Anziehungs- und Abstoßungskräfte in alle Richtungen des Raumes wirken ordnen sich die Atomrümpfe zu einem regelmäßigen Gitter an. In den Zwischenräumen befinden sich die frei beweglichen Elektronen als so genanntes Elektronengas, dieses hält die positiven Metallionen zusammen. Auf Grund der frei beweglichen Elektronen leiten Metalle den elektrischen Strom sehr gut.

Metallbindung



Die physikalischen und chemischen Eigenschaften der Verbindungen sind von der Bindungsart abhängig. So bedeuten stärkere Anziehungskräfte höhere Schmelz- und Siedepunkte, die Anzahl der freien Elektronen beeinflusst die

Leitfähigkeit.

Schwache Bindungen

Die schwachen Bindungen sind keine eigentlichen Bindungen innerhalb von Molekülen, sondern Wechselwirkungen zwischen Molekülen, die in der Regel auf Ladungsverschiebungen innerhalb eines Moleküls zurückzuführen sind. So entstehen positive und negative Ladungsschwerpunkte, durch welche sich Moleküle gegenseitig anziehen.

Die stärkste der schwachen Bindungen ist die Wasserstoffbrückenbindung, sie ergibt sich im Wassermolekül (H–O–H) aus dem Bindungswinkel zwischen Sauerstoff und den Wasserstoffatomen. Dadurch verteilen sich die elektrischen Ladungen im Molekül asymmetrisch (negativ am Sauerstoff, positiv am Wasserstoff), so dass sich benachbarte Moleküle anziehen. Daneben gibt es noch weitere Bindungen, wie die Van-der-Waals-Bindung, welche noch einmal schwächer sind.

Die Bindungsenergie der chemischen Bindungen liegt bei einigen Hundert bis einigen Tausend Kilojoule/Mol (kJ/mol). Bei Wasserstoffbrückenbindungen bei bis zu 100 kJ/mol und bei den Van-der-Waals-Kräften bei 0,5–5 kJ/mol.

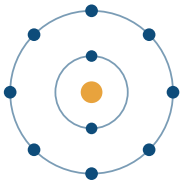
Edelgase

Edelgase und das Elektronenoktett

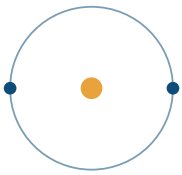
Edelgase sind die Elemente in der achten Hauptgruppe im Periodensystem: Helium, Neon, Argon, Krypton, Xenon, Radon (v.o.n.u.).

Die Besonderheit dieser Elemente ist, dass sie auf der äußersten Schale acht Elektronen besitzen. Dieses Elektronenoktett repräsentiert einen energetisch sehr stabilen Zustand, den alle Elemente einnehmen wollen. Im Kapitel Bindungen wird erläutert, wie die Elemente die volle Achterschale erreichen. Auf Grund des stabilen Zustands der Edelgaskonfiguration gehen diese Elemente so gut wie keine Reaktionen ein (einige wenige Xenon-Fluor-Verbindungen u.a. gibt es).

Edelgase



Neon: 8 Valenzelektronen
auf der zweiten Schale



Helium: 2 Valenzelektronen
auf der ersten Schale

Das Neonatom besitzt 10 Elektronen: zwei auf der ersten und acht Valenzelektronen auf der zweiten/äußersten Schale. Ausnahme: Das Heliumatom, welches nur eine Schale besitzt, erreicht den Edelgaszustand bereits mit zwei Elektronen (Elektronenduett).

Als Inertgase [lat. inert: untätig, träge] finden unter anderem Stickstoff (z.B. als Spülgas bei Ofenprozessen) und Argon (z.B. bei Sputterprozessen) in der Halbleiterherstellung Anwendung.

Leiter – Nichtleiter – Halbleiter

Leiter

Leiter sind generell Stoffe, die die Eigenschaft haben verschiedene Energiearten weiterzuleiten. Im Folgenden steht dabei die Leitfähigkeit des elektrischen Stroms im Vordergrund.

Metalle

Die Leitfähigkeit von Metallen beruht auf den freien Elektronen die bei der Metallbindung als Elektronengas vorliegen. Bereits mit wenig Energie werden genug Elektronen von den Atomen gelöst um eine Leitfähigkeit zu erreichen.

Metallbindung: Feste Atomrümpfe und freie Valenzelektronen (Elektronengas)

Die Leitfähigkeit hängt unter anderem von der Temperatur ab. Steigt diese an, schwingen die Atomrümpfe immer stärker, so dass die Elektronen in ihren Bewegungen behindert werden und infolge dessen der Widerstand ansteigt. Die besten Leiter, Gold und Silber, werden auf Grund der hohen Kosten relativ

selten eingesetzt (Gold u.a. bei der Kontaktierung der fertigen Chips). Die Alternativen in der Halbleitertechnologie zur Verdrahtung der einzelnen Komponenten eines Chips sind Aluminium und Kupfer.

Salze

Neben Metallen können auch Salze elektrischen Strom leiten. Freie Elektronen gibt es hier jedoch nicht. Die Leitfähigkeit beruht auf den Ionen die sich beim Schmelzen oder Lösen von Salzen aus dem Gitterverbund lösen und frei beweglich sind (siehe Thema Bindungen).

Nichtleiter

Nichtleiter besitzen keine freien Ladungsträger in Form von Elektronen oder Ionen. Nichtleiter nennt man auch Isolatoren.

Atombindung

Die Atombindung beruht auf gemeinsamen Elektronenpaaren von Nichtmetallen. Die Elemente mit Nichtmetallcharakter haben alle das Bestreben Elektronen aufzunehmen, somit sind keine freien Elektronen vorhanden die eine Leitfähigkeit bewirken könnten.

Ionenbindung

Im festen Zustand sind Ionen in einem Gitterverbund angeordnet. Durch elektrische Kräfte werden die Teilchen zusammengehalten. Es sind keine freien Ladungsträger für den Stromfluss vorhanden. So können Stoffe, die sich aus Ionen zusammensetzen, sowohl Leiter (im gelösten Zustand) als auch Nichtleiter sein.

Halbleiter

Halbleiter sind Feststoffe, deren Leitfähigkeit zwischen der von Leitern und Nichtleitern liegt. Durch Elektronenaustausch gleichartiger Atome, um das Elektronenoktett zu vervollständigen, ordnen sich diese als Gitterstruktur an. Im Gegensatz zu Metallen nimmt die Leitfähigkeit mit steigender Temperatur - bis zu einem gewissen Maß - zu.

Durch den Temperaturanstieg brechen Bindungen auf und Elektronen werden freigesetzt. An der Stelle an der sich das Elektron befand verbleibt ein so genanntes Defektelektron (auch Loch).

Ausschnitt aus einem Siliciumkristall

Der Elektronenfluss beruht auf der Eigenleitfähigkeit von Halbleitern. Das so genannte Bändermodell veranschaulicht, warum sich Halbleiter so verhalten.

Das Bändermodell

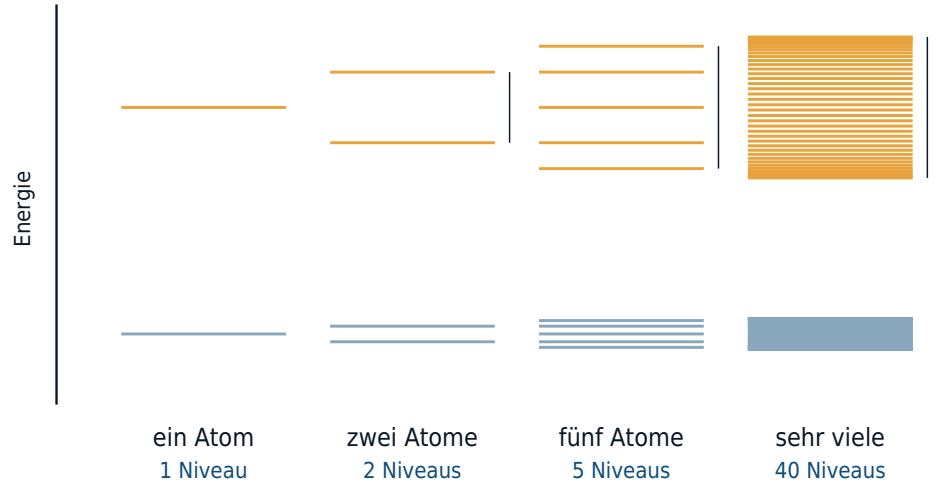
Das Bändermodell ist ein Energieschema, mit Hilfe dessen man die Leitfähigkeit von Leitern, Isolatoren und Halbleitern beschreiben kann. Das Modell besteht aus zwei Energiebändern (Valenz- und Leitungsband) und der Bandlücke. Die Valenzelektronen - die als Ladungsträger dienen - befinden sich im Valenzband; das Leitungsband ist im Grundzustand nicht mit Elektronen besetzt. Zwischen den beiden Energiebändern befindet sich die Bandlücke, ihre Breite beeinflusst u.a. die Leitfähigkeit.

Die Energiebänder

Betrachtet man ein einzelnes Atom, so gibt es nach dem Bohrschen Atommodell scharf voneinander getrennte Energieniveaus, die von Elektronen besetzt werden können. Befinden sich mehrere Atome nebeneinander, so stehen diese miteinander in Wechselwirkung und die diskreten Energieniveaus werden aufgefächert. In einem Siliciumkristall gibt es ca. 5×10^{22} Atome pro Kubikzentimeter, so dass die einzelnen Energieniveaus nicht mehr von einander unterscheidbar sind und breite Energiebereiche bilden.

Aufspaltung von Energieniveaus

Je mehr Atome zusammenwirken, desto dichter liegen die Niveaus

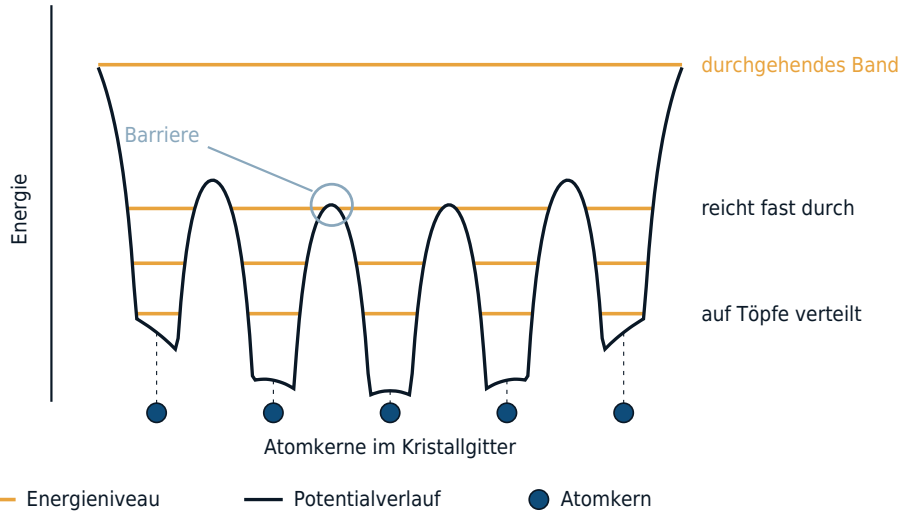


Jedes hinzukommende Atom fügt ein weiteres Niveau hinzu. Die Gesamtbreite wächst dabei kaum noch, die Abstände werden immer kleiner – bei 10^{22} Atomen je Kubikzentimeter sind sie nicht mehr unterscheidbar.

Die Breite der Energiebänder hängt davon ab, wie stark die Elektronen an das Atom gebunden sind. Die Valenzelektronen im höchsten Energieniveau wechselwirken stark mit denen der Nachbaratome und können relativ leicht vom Atom gelöst werden, bei einer sehr großen Anzahl an Atomen lässt sich ein einzelnes Elektron nicht mehr einem bestimmten Atom zuordnen. In Folge dessen verschmelzen die Energiebänder der einzelnen Atome zu einem kontinuierlichen Band, dem Valenzband.

Energiebänder durch in Wechselwirkung stehende Atome

Die Potentialtöpfe benachbarter Atome überlagern sich



Tief gebundene Elektronen bleiben in ihrem Topf gefangen – ihre Niveaus bleiben getrennt. Die Valenzelektronen liegen dagegen über den Barrieren zwischen den Kernen und gehören keinem Atom mehr allein: ein Band entsteht.

Das Bändermodell bei Leitern

Bei Leitern ist das Valenzband entweder nicht voll mit Elektronen besetzt, oder das gefüllte Valenzband überlappt sich mit dem leeren Leitungsband. In der Regel treffen beide Zustände gleichzeitig zu, die Elektronen können sich also im nur teilweise besetzten Valenzband oder in den zwei sich überlappenden Bändern bewegen. Die Bandlücke, die sich zwischen Valenz- und Leitungsband befindet, existiert bei Leitern nicht.

Das Bändermodell bei Nichtleitern

Bei Isolatoren ist das Valenzband durch die Bindungen der Atome voll mit Elektronen besetzt. Sie können sich darin nicht bewegen, da sie zwischen den Atomen "eingesperrt" sind. Um leiten zu können müssten sich die Elektronen aus dem voll besetzten Valenzband in das Leitungsband bewegen. Das verhindert die Bandlücke, die zwischen Valenz- und Leitungsband liegt.

Nur mit sehr großem Energieaufwand (falls überhaupt möglich) kann diese Lücke überwunden werden (in der Bandlücke darf sich nach den Gesetzen der Quantenphysik kein Elektron aufhalten).

Das Bändermodell bei Halbleitern

Auch bei Halbleitern gibt es diese Bandlücke, diese ist im Vergleich zu

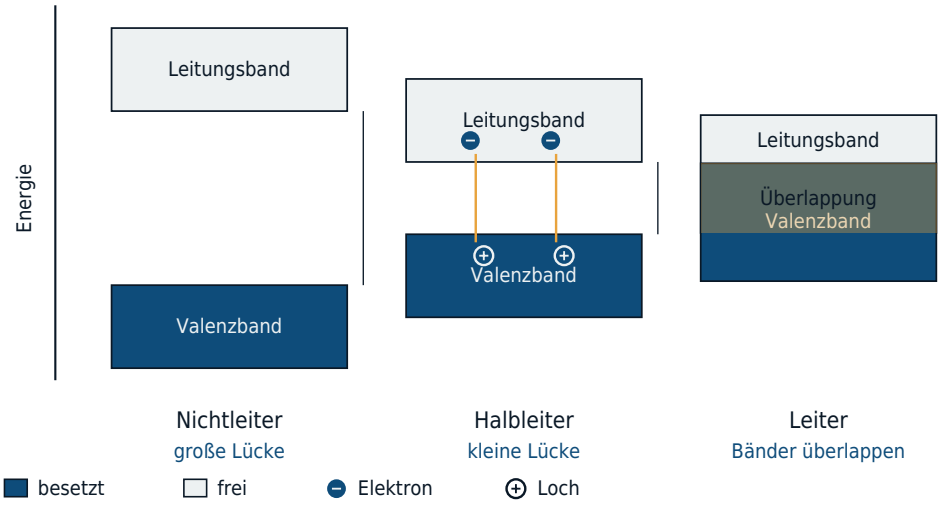
Isolatoren aber so klein, dass bereits bei Raumtemperatur Elektronen aus dem Valenzband in das Leitungsband gelangen. Die Elektronen können sich hier nun frei bewegen und stehen als Ladungsträger zur Verfügung. Jedes Elektron hinterlässt außerdem ein Loch im Valenzband, welches von anderen Elektronen im Valenzband besetzt werden kann. Somit erhält man wandernde Löcher im Valenzband, die als positive Ladungsträger angesehen werden können.

Es treten immer Elektronen-Loch-Paare auf, es gibt also ebenso viele negative wie positive Ladungen, der Halbleiterkristall ist insgesamt neutral. Ein reiner, undotierter Halbleiter wird als intrinsischer Halbleiter bezeichnet, pro Kubikzentimeter gibt es in etwa 10^{10} freie Elektronen und Löcher (bei Raumtemperatur).

Da die Elektronen immer den energetisch günstigsten Zustand annehmen, fallen sie ohne Energiezufuhr wieder in das Valenzband zurück und rekombinieren mit den Löchern. Bei einer bestimmten Temperatur stellt sich ein Gleichgewicht zwischen den ins Leitungsband gehobenen und den zurückfallenden Elektronen ein. Mit zunehmender Temperatur erhöht sich die Anzahl der Elektronen, die die Bandlücke überspringen können. Mit steigender Temperatur nimmt also die Leitfähigkeit von Halbleitern zu.

Das Bändermodell

Die Breite der Bandlücke entscheidet über die Leitfähigkeit



Beim Halbleiter genügt schon die Wärme bei Raumtemperatur, um einzelne Elektronen über die Lücke zu heben. Jedes hinterlässt ein Loch im Valenzband – beide tragen zum Strom bei.

Da die Breite der Bandlücke einer bestimmten Energie und somit einer bestimmten Wellenlänge entspricht, versucht man, die Bandlücke gezielt zu verändern um so bestimmte Farben bei Leuchtdioden zu erhalten. Dies kann u.a. durch Kombination verschiedener Stoffe erreicht werden. Galliumarsenid

(GaAs) hat eine Bandlücke von 1,4 Elektronenvolt (eV, bei Raumtemperatur) und strahlt somit rotes Licht ab.

Die Eigenleitfähigkeit von Silicium ist für die Funktionsweise von Bauelementen uninteressant, da sie sehr stark von der zugeführten Energie abhängt. Sie ändert sich also auch mit der Betriebstemperatur, eine mit Metallen vergleichbare Leitfähigkeit stellt sich zudem erst mit sehr hohen Temperaturen ein (mehrere Hundert Grad Celsius). Um die Leitfähigkeit von Halbleitern gezielt zu beeinflussen, können Fremdatome in das regelmäßige Siliciumgitter eingebaut, und damit die Ladungsträgerkonzentration von Elektronen und Löchern eingestellt werden. Dies nennt man *dotieren*.

Dotieren: n- und p-Halbleiter

Dotieren

Dotieren bedeutet das Einbringen von Fremdatomen in einen Halbleiterkristall zur gezielten Veränderung der Leitfähigkeit. Zwei der wichtigsten Stoffe mit denen Silicium dotiert werden kann sind Bor (3 Valenzelektronen = 3-wertig) und Phosphor (5 Valenzelektronen = 5-wertig). Weitere sind: Aluminium, Indium (3-wertig) und Arsen, Antimon (5-wertig).

Die Dotierelemente werden in die Gitterstruktur des Halbleiterkristalls eingebaut, die Anzahl der Außenelektronen bestimmt die Art der Dotierung. Elemente mit 3 Valenzelektronen werden zur p-Dotierung benutzt, 5-wertige Elemente zur n-Dotierung. Die Leitfähigkeit eines gezielt verunreinigten Siliciumkristalls kann so um den Faktor 10^6 erhöht werden.

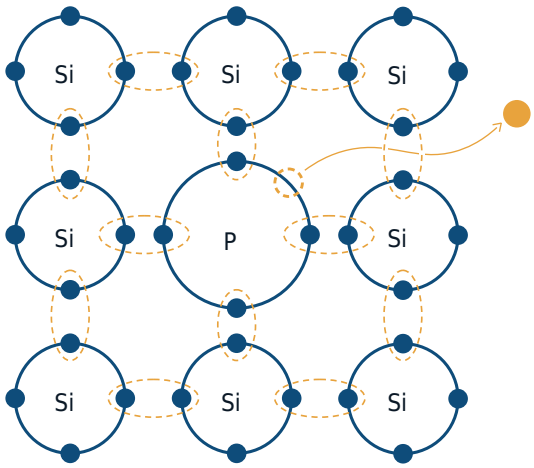
n-Dotierung

Das 5-wertige Dotierelement besitzt ein Außenelektron mehr als die Siliciumatome. Vier Außenelektronen können sich mit je einem Siliciumatom binden, das fünfte ist frei beweglich und dient als Ladungsträger. Dieses ungebundene Elektron benötigt sehr viel weniger Energie um vom Valenzband in das Leitungsband gehoben zu werden, als die Elektronen, die die Eigenleitfähigkeit des Siliciums verursachen. Das Dotierelement, welches ein Elektron abgibt, wird als Elektronendonator (donare, lat. = geben) bezeichnet.

Die Dotierelemente werden durch die Abgabe negativer Ladungsträger positiv geladen und sind fest im Gitter eingebaut, es bewegen sich nur die Elektronen. Dotierte Halbmetalle, deren Leitfähigkeit auf freien (negativen) Elektronen beruht sind n-leitend bzw. n-dotiert. Während Löcher (und ebenso viele Elektronen) im Kristall jederzeit spontan erzeugt werden können, überwiegt nun

die Anzahl freier Elektronen durch die eingebrachten Donatoren, weshalb diese als Majoritätsladungsträger bezeichnet werden. Löcher hingegen als Minoritätsladungsträger.

n-Dotierung mit Phosphor



Vier Bindungselektronen, ein Elektron als freier Ladunsträger

Arsen wird als Alternative zu Phosphor verwendet, da dessen Diffusionskoeffizient geringer ist. Das bedeutet, dass der Dotierstoff bei späteren Prozessschritten weniger stark diffundiert und die Dotierung somit an der Stelle bleibt, wo sie eingebracht wurde.

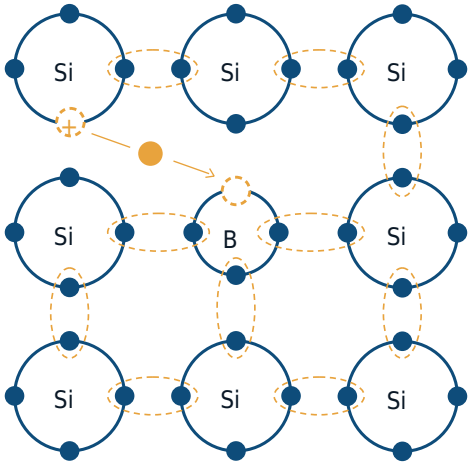
p-Dotierung

Im Gegensatz zum freien Elektron bei der Dotierung mit Phosphor bewirken 3-wertige Dotierelemente genau das Gegenteil. Sie können ein zusätzliches Außenelektron aufnehmen und hinterlassen so ein Loch im Valenzband der Siliciumatome. Dadurch werden die Elektronen im Valenzband beweglich. Die Löcher bewegen sich in entgegengesetzter Richtung zur Elektronenbewegung. Die dazu nötige Energie beträgt bei Indium als Dotierelement nur 1 % der Energie die nötig ist, um die Valenzelektronen der Siliciumatome in das Leitungsband zu heben.

Durch die Aufnahme eines Elektrons wird das Dotierelement einfach negativ geladen; solche Dotieratome nennt man Elektronenakzeptor (acceptare, lat. = aufnehmen). Auch hier ist das Dotierelement fest im Kristallgitter eingebaut, es bewegt sich nur die positive Ladung. P-leitend oder p-dotiert nennt man diese Halbleiter weil die Leitfähigkeit auf positiven Löchern beruht. Analog zu n-dotierten Halbleitern, sind hier die Löcher die Majoritätsladungsträger, freie

Elektronen die Minoritätsladungsträger.

p-Dotierung mit Bor



Vier Bindungselektronen, ein Elektron als freier Ladunsträger

Nach außen sind dotierte Halbleiter elektrisch neutral. Die Bezeichnungen n- bzw. p-Dotierung beziehen sich nur auf die Majoritätsladungsträger.

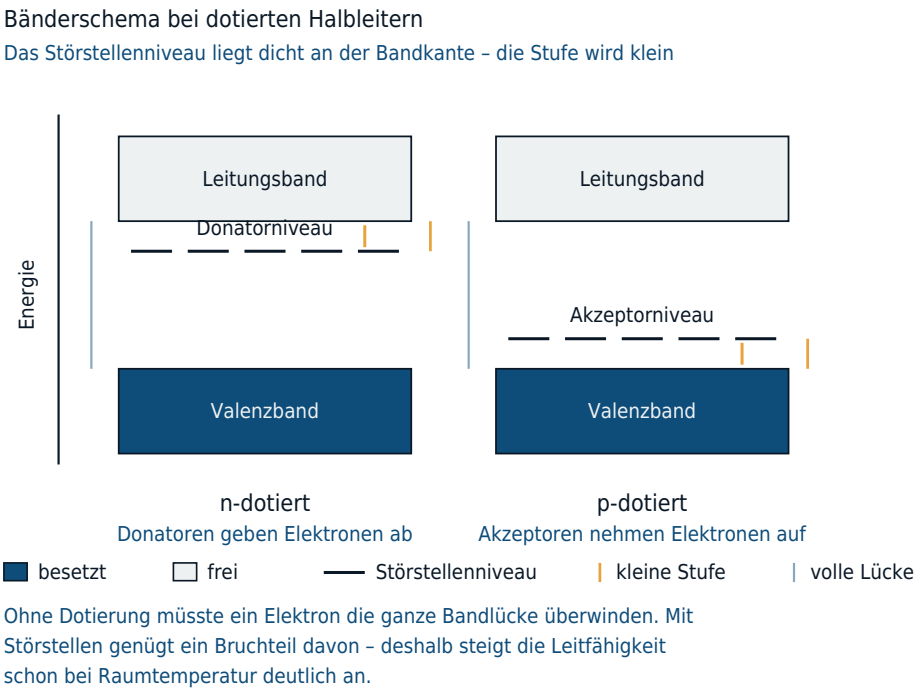
N- und p-dotierte Halbleiter verhalten sich in Bezug auf den Stromfluss annähernd gleich. Mit steigender Anzahl an Dotierelementen nimmt auch die Zahl von Ladungsträgern im Halbleiterkristall zu. Dabei genügt schon eine sehr geringe Menge an Dotierelementen. Schwach dotierte Siliciumkristalle enthalten nur 1 Fremdatom pro 1.000.000.000 Siliciumatomen, bei den höchsten Dotierungen ist das Verhältnis von Fremdatomen zu Siliciumatomen z.B. 1 zu 1000.

Bänderschema bei dotierten Halbleitern

Bei n-dotierten Halbleitern steht durch das Einbringen eines Dotierelements mit fünf Außenelektronen ein Elektron im Kristall zur Verfügung das nicht gebunden ist, und so mit vergleichsweise geringer Energie in das Leitungsband gehoben werden kann. Somit findet man in n-dotierten Halbleitern ein Donatorniveau nahe der Leitungsbandkante, die zu überwindende Bandlücke ist sehr klein.

Durch das Einbringen eines 3-wertigen Dotierelements steht bei p-dotierten Halbleitern eine Leerstelle zur Verfügung, die schon mit geringer Energie von einem Elektron aus dem Valenzband besetzt werden kann. Bei p-dotierten Halbleitern befindet sich somit ein Akzeptorniveau nahe der Valenzbandkante.

Bänderschema bei dotierten Halbleitern



Der p-n-Übergang

p-n-Übergang im Gleichgewicht

Der p-n-Übergang ist der Übergangsbereich aneinander liegender n- und p-dotierter Halbleiterkristalle. In diesem Bereich gibt es keine freien Ladungsträger, da die freien Elektronen des n-Leiters und die freien Löcher des p-dotierten Kristalls in der Nähe der Kontaktfläche der zwei Kristalle miteinander rekombinieren, d.h. die Elektronen besetzen die freien Löcher. Diese Ladungsträgerbewegung (Diffusion) ergibt sich in Folge eines Konzentrationsgefälles: da es im p-Kristall nur wenige Elektronen und im n-Kristall nur wenige Löcher gibt, wandern die Majoritätsladungsträger (Elektronen im n-Gebiet, Löcher im p-Gebiet) in den jeweils andersartig dotierten Halbleiterkristall. Das Kristallgitter an der Grenzfläche darf nicht unterbrochen werden, ein einfaches "Aneinanderpressen" eines p- und eines n-dotierten Siliciumkristalls ermöglicht keinen funktionstüchtigen p-n-Übergang.

Die Gebiete in Nähe der Grenzschicht sind auf Grund der abgewanderten freien Ladungsträger positiv (n-Kristall) bzw. negativ (p-Kristall) geladen. Je mehr Ladungsträger rekombinieren, desto größer wird diese Verarmungs- oder Raumladungszone (RLZ) und damit die Spannungsdifferenz von n- zu p-Kristall.

Bei einer bestimmten Höhe dieses Potentialgefälles kommt die Rekombination der Löcher und Elektronen zum Erliegen, die Ladungsträger können das elektrische Feld nicht mehr überwinden. Bei Silicium liegt diese Grenze bei etwa 0,7 V (vgl. Bändermodell eines p-n-Übergangs).

p-n-Übergang ohne angelegte Spannung

Ein p-n-Übergang entspricht einem elektrischen Bauteil, welches durch Anlegen von Spannung den Strom in der einen Richtung leitet (Durchlassrichtung) und in der anderen sperrt (Sperrrichtung): eine Diode.

p-n-Übergang mit angelegter elektrischer Spannung

Wird am n-Kristall eine positive und am p-Kristall eine negative Spannung angelegt, so zeigen das elektrische Feld im Inneren und das durch die Spannungsquelle erzeugte in die gleiche Richtung. Das Feld am p-n-Übergang wird damit verstärkt. Die jeweils entgegengesetzt geladenen freien Ladungsträger werden von den Polen der Spannungsquelle angezogen, dadurch wird die Sperrschicht vergrößert und es ist kein Stromfluss möglich.

Polt man die angelegte Spannung an den Halbleiterkristallen um, überlagert das durch die Spannungsquelle erzeugte elektrische Feld das innere in entgegengesetzter Richtung und schwächt es ab. Wird das innere Feld vollständig vom äußeren abgebaut, fließen ständig neue Ladungsträger von der Stromquelle zur Sperrschicht und können hier fortlaufend rekombinieren: es fließt Strom.

p-n-Übergang mit angelegter Spannung

Die Diode lässt sich auf Grund dieses Verhaltens als Gleichrichter verwenden: zur Umwandlung von Wechselstrom in Gleichstrom. Bereiche, in denen p- und n-dotierte Halbleiterkristalle in Kontakt stehen, kommen in vielen elektrischen Bauelementen der Halbleitertechnologie vor.

Transistoren und Speicher

MOSFET

Ein Transistor ist ein elektronisches Halbleiterbauelement, das zum Schalten oder Verstärken von Strom dient. Beim Feldeffekttransistor (FET) steuert eine Gateelektrode den Stromfluss zwischen den Anschlüssen Source und Drain, ohne selbst im Stromkreis zu liegen. Die gebräuchlichste Bauform ist der

MOSFET: Liegt am Gate keine Spannung an, trennt das p-dotierte Substrat Source und Drain wie ein gesperrter p-n-Übergang; der Transistor sperrt. Erst eine positive Gatespannung zieht Elektronen an die Oberfläche und bildet dort einen leitenden n-Kanal – derselbe Mechanismus wie beim Dotieren, hier jedoch elektrisch steuerbar. Wie der MOSFET tatsächlich gefertigt wird, zeigt das Kapitel Aufbau eines n-Kanal-FET.

Bipolartransistor

Der zweite grundlegende Transistortyp ist der Bipolartransistor. Er besteht aus zwei gegeneinander geschalteten p-n-Übergängen mit der Schichtfolge n-p-n oder p-n-p; seine Anschlüsse heißen Emitter, Basis und Kollektor. Anders als beim MOSFET tragen hier beide Ladungsträgertypen – Elektronen und Löcher – zum Stromfluss bei, daher „bipolar". Eine kleine Spannung an der dünnen Basisschicht öffnet den Übergang zum Emitter und lässt einen vielfach größeren Strom von Emitter zu Kollektor fließen. Bipolartransistoren sind schneller als MOSFETs, benötigen aber mehr Fläche. Details liefert das Kapitel Aufbau eines npn-Transistors.

FinFET

Mit sinkender Strukturgröße reicht ein flach aufliegendes (planares) Gate irgendwann nicht mehr aus, um den Kanal zuverlässig zu sperren. Beim FinFET wird der Kanal deshalb als dünne, hochstehende Silicium-„Finne" ausgeführt, die das Gate von drei Seiten umschließt – Source, Drain und die grundlegende Schaltfunktion bleiben identisch zum MOSFET. Durch den mehrseitigen Zugriff lassen sich Leckströme reduzieren und die Schwellspannung schärfer definieren. Wie ein FinFET im Detail aufgebaut und gefertigt wird, beschreibt das Kapitel Aufbau eines FinFET.

DRAM-Zelle

Transistoren dienen nicht nur zum Schalten, sondern auch zum Speichern von Daten. Die DRAM-Zelle kombiniert dazu einen einzelnen Transistor mit einem Kondensator (1T1C-Zelle): Der Kondensator speichert ein Bit als elektrische Ladung, der Transistor verbindet ihn bei Bedarf über die Wordline mit der Bitline zum Lesen oder Schreiben. Da die Ladung über Leckströme langsam abfließt, muss sie alle paar Millisekunden aufgefrischt werden – daher „dynamisch". Wie Kondensator und Zugriffstransistor gefertigt werden, zeigt das Kapitel DRAM-Fertigung: der Trench-Kondensator-Prozess.

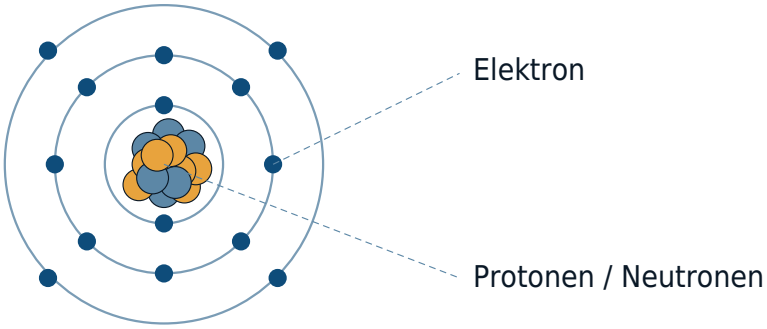
Silicium

Eigenschaften von Silicium

Silicium ist das chemische Element mit der Ordnungszahl 14 im Periodensystem der Elemente, es steht in der 4. Hauptgruppe und der 3. Periode. Silicium ist ein klassischer Halbleiter, seine Leitfähigkeit liegt also zwischen der von Leitern und Nichtleitern. Silicium (von lat. silex / silicis: Kieselstein) kommt in der Natur ausschließlich als Oxid vor: als Siliciumdioxid (SiO_2) in Form von Sand, Quarz oder als Silicat (Verbindungen von Silicium mit Sauerstoff, Metallen u.a.). Silicium gibt es also sprichwörtlich wie Sand am Meer, dementsprechend ist es ein sehr günstiges Ausgangsmaterial, dessen Wert erst mit der Verarbeitung bestimmt wird. Andere Halbleiter wie Germanium oder der Verbindungshalbleiter Galliumarsenid bieten teils wesentlich bessere elektrische Eigenschaften als Silicium: die Ladungsträgerbeweglichkeit und die daraus resultierenden Schaltgeschwindigkeiten sind bei Germanium und GaAs deutlich höher. Doch Silicium hat entscheidende Vorteile gegenüber anderen Halbleitern.

Diese Vorrangstellung gilt für integrierte Schaltungen, nicht für die Halbleitertechnik insgesamt. In der Leistungselektronik haben sich Siliciumcarbid und Galliumnitrid einen festen Platz erobert, weil sie höhere Spannungen, höhere Temperaturen und schnelleres Schalten vertragen. Und für alles, was Licht aussenden soll, ist Silicium ungeeignet: Sein Bandübergang gibt die Energie als Wärme ab statt als Licht, weshalb Leuchtdioden und Laser aus Verbindungshalbleitern bestehen.

Bohrsches Atommodell von Silicium



Auf einem Siliciumkristall lassen sich sehr leicht Oxidschichten erzeugen, das entstandene Siliciumdioxid ist ein hochwertiger Isolator, der sich gezielt auf dem Substrat aufbringen lässt. Bei den angesprochenen Halbleitern Germanium und GaAs ist es dagegen sehr kostenintensiv ähnliche Isolationsschichten zu erzeugen. Auch die Möglichkeit, durch Dotierung ganz gezielt die Leitfähigkeit des reinen Siliciums zu verändern macht das Halbmetall so bedeutsam. Andere Stoffe sind zudem teilweise sehr giftig, und Verbindungen mit diesen Elementen nicht so langlebig und stabil wie bei Silicium. Voraussetzung für den Einsatz von Silicium in der Halbleiterfertigung ist jedoch, dass das Silicium in einer hochreinen Form als Einkristall vorliegt. Das bedeutet, dass die Siliciumatome im Kristallgitter absolut regelmäßig angeordnet sind und sich keine undefinierten Fremdatome im Kristall befinden.

Neben der einkristallinen Form gibt es noch Polysilicium (poly = viel) und amorphes Silicium (a-Si). Während das einkristalline Silicium als Wafer die Grundlage für die Mikroelektronik ist, wird polykristallines Silicium in verschiedenen Bereichen als Schicht auf dem Wafer erzeugt, um bestimmte Aufgaben zu erfüllen (z.B. Maskierschicht, Gate im Transistor u.a.). Es lässt sich einfach erzeugen und leicht strukturieren. Polysilicium setzt sich aus vielen einzelnen, zu einander unregelmäßig angeordneten Einkristallen zusammen. Amorphes Silicium besitzt keinen regelmäßigen Gitteraufbau sondern eine ungeordnete Gitterstruktur. In der Chipfertigung spielt es nur eine untergeordnete Rolle, außerhalb davon jedoch keineswegs: Weil es sich großflächig und bei niedriger Temperatur auf Glas abscheiden lässt, sind die Schaltelemente hinter jedem Flachbildschirm daraus aufgebaut, und auch in Dünnschichtsolarzellen wird es eingesetzt.

Herstellung von Rohsilicium

Ausgangsmaterial Siliciumdioxid

Silicium, das in der Halbleiterfertigung Verwendung findet, wird aus Quarz gewonnen. Dabei muss Sauerstoff, der sich bereits bei Raumtemperatur sehr schnell mit Silicium verbindet, und der auch beim Quarz in Verbindung mit Silicium als Siliciumdioxid vorliegt, entfernt werden. Das geschieht im Lichtbogenofen unter Verwendung von Kohlenstoff, bei Temperaturen, die im heißen Bereich des Ofens bis etwa 2000 °C reichen. Der Sauerstoff löst sich dabei vom Silicium und reagiert mit dem Kohlenstoff (C) zu Kohlenmonoxid (CO):

$SiO_2 + 2C \rightarrow Si + 2CO$

Diese Gleichung fasst zusammen, was im Ofen tatsächlich in mehreren Stufen abläuft. In der kühleren Zone entsteht zunächst Siliciumcarbid, das sich in der heißen Zone mit weiterem Siliciumdioxid zu Silicium umsetzt. Siliciumcarbid ist also kein Störprodukt, sondern ein notwendiges Zwischenprodukt; über das Verhältnis von Kohlenstoff zu Quarz wird eingestellt, dass es am Ende wieder aufgebraucht ist. Das Kohlenmonoxid ist bei diesen Temperaturen gasförmig und kann leicht vom flüssigen Silicium getrennt werden.

Das so gewonnene Rohsilicium (metallurgisches Silicium) ist etwa 98 bis 99 % rein. Der Rest sind Fremdstoffe wie Eisen, Aluminium, Calcium, Phosphor und Bor. Für Bauelemente ist das um viele Größenordnungen zu unrein – diese Stoffe müssen in weiteren Prozessen entfernt werden.

Reinigung des Rohsiliciums

Die Reinigung erfolgt auf dem Umweg über eine gasförmige Verbindung: Man überführt das Silicium in Trichlorsilan, reinigt dieses durch Destillation und scheidet daraus wieder festes Silicium ab. Das Verfahren ist nach seinem Entwickler als Siemens-Prozess bekannt und liefert bis heute den größten Teil des weltweit erzeugten Reinstsiliciums.

Mittels des Trichlorsilanprozesses werden viele Verunreinigungen durch Destillation herausgefiltert. Das Rohsilicium und Chlorwasserstoff (HCl) reagieren bei ca. 300 °C zu gasförmigem Trichlorsilan ($SiHCl_3$) und Wasserstoff (H_2):

$Si + 3HCl \rightarrow SiHCl_3 + H_2$

Die Fremdstoffe, die sich ebenfalls mit dem enthaltenen Chlor verbinden, gehen erst bei höheren Temperaturen in den gasförmigen Zustand über. So lässt sich das Trichlorsilan abtrennen. Lediglich Kohlenstoff, Phosphor und Bor, die ähnliche Kondensationstemperaturen haben, können hier nicht herausgefiltert werden.

Der Trichlorsilanprozess lässt sich umkehren, so dass man anschließend das gereinigte Silicium in polykristalliner Form erhält. Dies geschieht unter Zugabe von Wasserstoff in einem Quarzgefäß, in dem sich beheizte, dünne Siliciumstäbe (Siliciumseelen) befinden, bei ca. 1100 °C:

$SiHCl_3 + H_2 \rightarrow Si + 3HCl$
und
$4SiHCl_3 \rightarrow Si + 3SiCl_4 + 2H_2$

Das Silicium schlägt sich auf den Siliciumseelen nieder, die dadurch über mehrere Tage hinweg auf etwa 150 bis 200 mm dicke Stäbe anwachsen. Diese werden anschließend zerkleinert und als Bruchstücke weiterverarbeitet.

Das Verfahren ist außerordentlich energieaufwendig, denn die Stäbe müssen die ganze Zeit über auf Temperatur gehalten werden, während die Kammerwand gekühlt wird. Als Alternative hat sich die Abscheidung im Wirbelbett etabliert: Dort wachsen kleine Siliciumkörner in einem von Gas durchströmten Bett heran, was deutlich weniger Energie erfordert und fortlaufend statt chargenweise arbeitet. Die Reinheit reicht allerdings nicht an den Siemens-Prozess heran, weshalb dieses Material vor allem in die Solarindustrie geht.

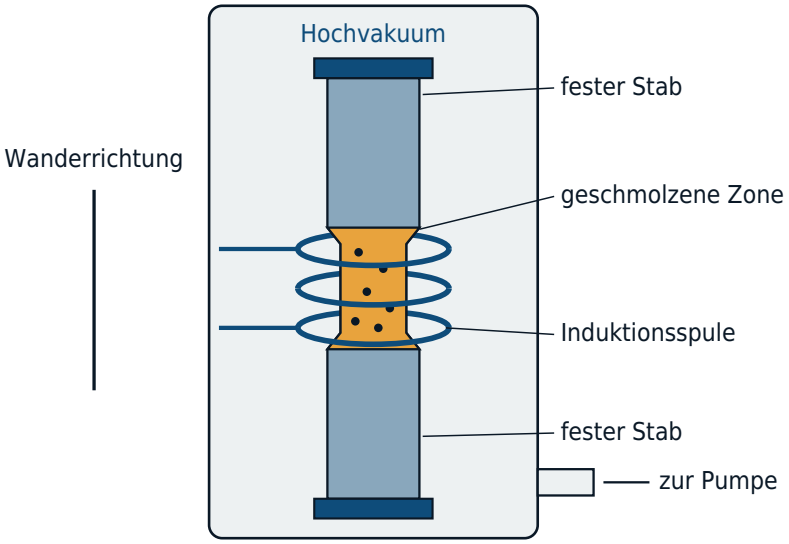
Genau darin liegt eine wichtige Unterscheidung: Solartaugliches Silicium muss etwa sechs Neunen erreichen (99,9999 %), elektroniktaugliches dagegen neun bis elf. Mit dem Czochralski-Verfahren könnte man dieses Polysilicium bereits in einen Einkristall umwandeln, jedoch ist der Reinheitsgrad für die Herstellung von Bauelementen noch immer nicht hoch genug.

Zonenreinigung

Um die Reinheit weiter zu erhöhen wird die Zonenreinigung verwendet. Um die Siliciumstäbe wird eine Spule gelegt, die hochfrequenter Wechselstrom durchfließt. Dadurch schmilzt das Silicium im Inneren der Stäbe und die Verunreinigungen sinken auf Grund der hohen Löslichkeit nach unten (die Oberflächenspannung des Siliciums verhindert, dass die Schmelze herausfließt). Durch mehrfache Wiederholung dieses Verfahrens wird der Gehalt der Fremdstoffe im Silicium so weit gesenkt, dass es zum Einkristall weiter verarbeitet werden kann.

Darstellung der Zonenreinigung

Zonenreinigung eines Siliciumstabs
Die Oberflächenspannung hält die Schmelze zwischen den festen Enden



Verunreinigungen bleiben in der Schmelze und wandern mit der Zone.

Die Prozesse werden alle im Vakuum durchgeführt, um weitere Verunreinigungen zu verhindern.

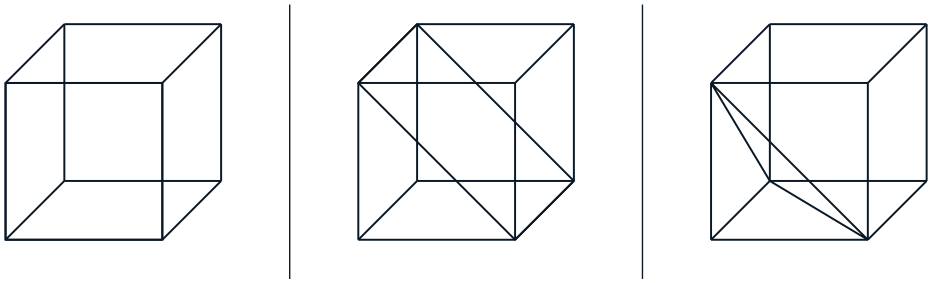
Am Ende dieser Prozesse hat das für die Halbleiterfertigung aufbereitete Silicium eine Reinheit von über 99,9999999 %, dies entspricht weniger als einem Fremdatom pro 1 Mrd. Siliciumatomen.

Herstellung des Einkristalls

Der Einkristall

Ein Einkristall (Monokristall), wie er in der Halbleiterfertigung benötigt wird, ist eine regelmäßige Anordnung von Atomen. Daneben gibt es polykristallines (Zusammensetzung von vielen kleinen Einkristallgittern) und amorphes Silicium (ungeordnete Struktur). Je nachdem, welche Lage das Gitter im Raum hat, besitzen Siliciumscheiben unterschiedliche Oberflächenstrukturen, die Einfluss auf verschiedene Parameter, wie die Ladungsträgerbeweglichkeit, von Bauelementen haben.

Kristallorientierungen



Orientierung 100 | Orientierung 110 | Orientierung 111

In der Chipfertigung wird nahezu ausschließlich (100)-orientiertes Silicium verwendet. Der Grund liegt an der Grenzfläche zum Gateoxid: Dort bleiben nach der Oxidation ungesättigte Bindungen zurück, und ihre Zahl ist bei dieser Orientierung am geringsten. Die (111)-Orientierung hat die dichteste Belegung mit Bindungen und damit die meisten Störstellen; sie oxidiert dafür am schnellsten.

Auch in der Mikromechanik ist die Kristallorientierung von besonderer Bedeutung. So lassen sich auf (110)-Silicium Mikrokanäle mit senkrechten Wänden herstellen, wohingegen sich bei (100)-Orientierung Flankenwinkel von 54,74° ergeben.

Kristallziehverfahren nach Czochralski

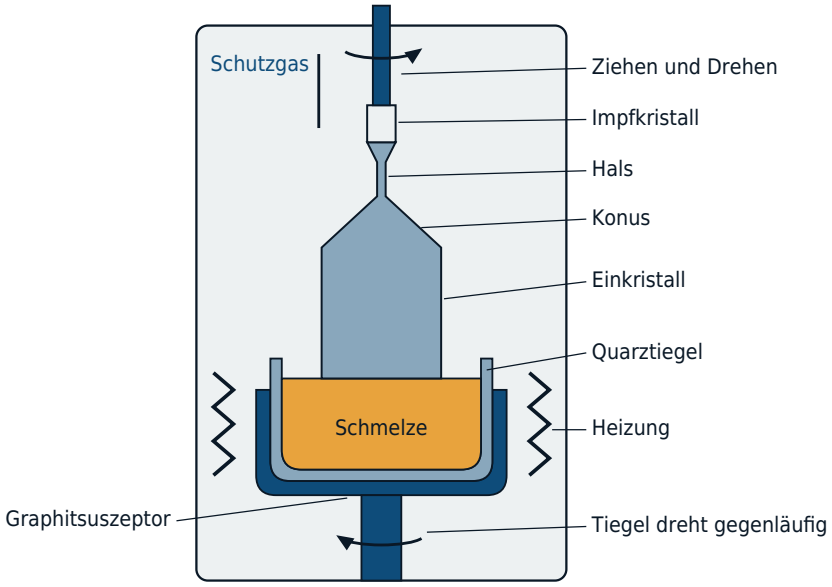
Das polykristalline Silicium, das nach der Zonenreinigung vorliegt, wird in einem Quarztiegel knapp über dem Schmelzpunkt von Silicium aufgeschmolzen. Der Schmelze können jetzt Dotierstoffe (z.B. Bor oder Phosphor) hinzugefügt werden, um entsprechende elektrische Eigenschaften des Einkristalls zu erzielen.

Ein Impfkristall (Einkristall) an einem drehbaren Stab wird an die Oberfläche der Siliciumschmelze gebracht. Dieser Keim gibt die Orientierung des Kristalls vor. Beim Kontakt des Keims mit der Schmelze lagert sich Silicium am Keim an und übernimmt dessen Kristallstruktur. Dadurch, dass die Tiegeltemperatur nur wenig über dem Schmelzpunkt von Silicium liegt, erstarrt das angelagerte Silicium sofort am Keimling und der Kristall wächst.

Der Keim wird unter ständigem Drehen langsam nach oben gezogen, wobei stetiger Kontakt zur Schmelze besteht. Der Tiegel dreht sich dabei entgegengesetzt zum Impfkristall. Eine konstante Temperierung der Schmelze ist unerlässlich, um ein gleichmäßiges Wachstum zu gewährleisten. Der Durchmesser des Einkristalls wird durch die Ziehgeschwindigkeit, die 2–25 cm/h beträgt, bestimmt. Je schneller gezogen wird, desto dünner wird der

Kristall. Die gesamte Apparatur befindet dabei sich in einer Schutzgasatmosphäre, damit es nicht zu einer Oxidation des Siliciums kommt.

Kristallziehen nach Czochralski
Der Impfkristall gibt die Orientierung des Kristalls vor



Durchmesser über Ziehgeschwindigkeit und Schmelztemperatur eingestellt.

Ein wesentlicher Nachteil dieses Verfahrens ist, dass sich die Schmelze während des Vorgangs immer mehr mit Dotierstoffen anreichert, da sich die Dotierstoffe in der Schmelze besser lösen, als im Festkörper. Somit ist die Dotierstoffkonzentration entlang des Siliciumstabes nicht konstant. Zudem können sich Verunreinigungen oder Metalle aus dem Tiegel lösen und im Kristallgitter eingebaut werden.

Die Vorteile dieses Verfahrens sind die geringen Kosten, und die Möglichkeit, größere Wafer herzustellen, als dies beim Zonenziehen der Fall ist (s. unten).

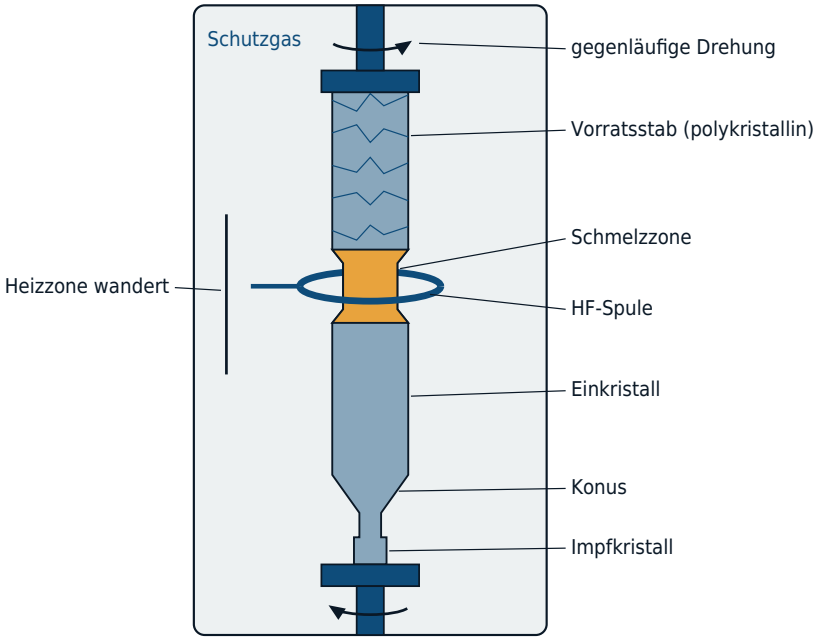
Tiegelfreies Zonenziehen

Im Gegensatz zum Kristallziehverfahren nach Czochralski wird beim tiegelfreien Zonenziehen nicht das gesamte Polysilicium geschmolzen, sondern, wie bei der Zonenreinigung, nur ein kleiner Bereich (wenige Millimeter).

Auch hier dient ein Impfkristall, der an das Ende des polykristallinen Siliciumstabs herangeführt wird, als Vorgabe der Kristallstruktur. Der Polykristall wird aufgeschmolzen und übernimmt die Struktur des Keimlings. Die Heizzone wird langsam am Stab entlang geführt, der polykristalline Siliciumstab wandelt sich langsam in einen Einkristall um.

Da nur ein kleiner Bereich des polykristallinen Siliciums aufgeschmolzen wird, können sich kaum Verunreinigungen anlagern (wie bei der Zonenreinigung verlagern sich die Fremdatome ans Ende des Siliciumstabs). Die Dotierung erfolgt hier durch Zusätze der Dotierstoffe ins Schutzgas (z.B. mit Diboran oder Phosphin), welches die Apparatur umströmt.

Tiegelfreies Zonenziehen
Nur wenige Millimeter des Stabs sind geschmolzen



Ohne Tiegel kein Kontakt zu Fremdmaterial; dotiert wird über das Schutzgas.

Der Wafer

Wafervereinzelung und Oberflächenveredelung

Der Einkristallstab wird zunächst auf den gewünschten Durchmesser abgedreht und bekommt dann, je nach Kristallorientierung und Dotierung, einen oder zwei Flats. Der größere Flat dient dazu, die Wafer in der Fertigung exakt ausrichten zu können. Der zweite Flat dient zur Erkennung des Scheibentyps (Kristallorientierung, p-/n-Dotierung), ist aber nicht immer vorhanden. Bei Wafern ab 200 mm Durchmesser werden an Stelle der Flats sogenannte Notches verwendet. Dabei handelt es sich um eine winzige Einkerbungen am Scheibenrand, die ebenfalls eine Ausrichtung der Wafer ermöglicht, aber sehr viel weniger kostbare Fläche des Wafers beansprucht.

Sägen

Historisch wurde der Einkristallstab mit einer Innenlochsäge zerteilt, deren Schnittkante mit Diamantsplittern besetzt ist. Sie schneidet genau, aber immer nur eine Scheibe auf einmal, und bis zu 20 % des Kristallstabs gehen dabei durch die Dicke des Sägeblatts verloren. Bei den heutigen Stabdurchmessern und Scheibenpreisen ist das nicht mehr vertretbar; Standard ist deshalb das Drahtsägen, bei dem mehrere hundert Wafer in einem Durchgang aus dem Stab geschnitten werden. Dabei wird ein langer Draht, welcher mit einer Suspension aus Siliciumcarbidkörnern und einem Trägermittel wie Glykol oder Öl benetzt wird, über rotierende Walzen geführt. Der Siliciumkristall wird in das Drahtgitter abgelassen und so zu Wafern vereinzelt. Der Draht bewegt sich im Gegenschritt mit ca. 10 m/s und ist typischerweise 0,1–0,2 mm dick. Zunehmend wird statt der aufgeschlämmten Körner ein Draht verwendet, in dessen Oberfläche die Diamantkörner fest eingebunden sind. Er schneidet schneller, erzeugt weniger Abfall und macht die Entsorgung der verbrauchten Suspension überflüssig.

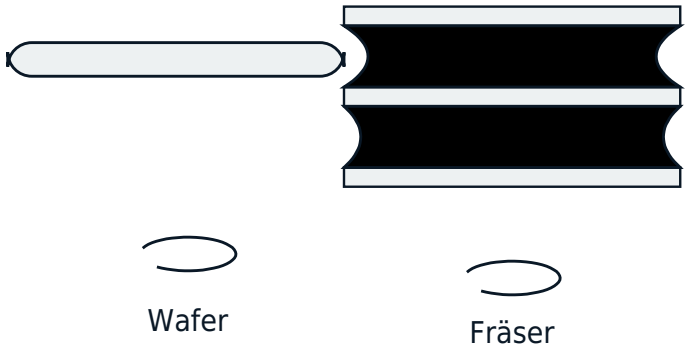
Nach dem Sägen besitzen die Scheiben eine aufgeraute Oberfläche und, auf Grund der mechanischen Belastung, Gitterschäden im Kristall. Zum Veredeln der Oberfläche durchlaufen die Scheiben mehrere Prozessschritte.

Läppen

Mit körnigen Schleifmitteln (z.B. Aluminiumoxid) werden 50 µm (0,05 mm) der Scheibenoberfläche auf einer rotierenden Stahlscheibe abgetragen. Die Körnung wird dabei stufenweise verringert, jedoch wird die Oberfläche auf Grund der mechanischen Behandlung erneut geschädigt. Die Ebenheit nach dem Läppen beträgt ca. 2 µm.

Scheibenrand abrunden

In späteren Prozessen dürfen die Scheiben keine scharfen Kanten besitzen, da aufgebrachte Schichten ansonsten abplatzen können. Dazu wird der Rand der Scheiben mit einem Diamantfräser abgerundet.



Ätzen

In einem Tauchätzschritt, mit einer Mischung aus Fluss-, Essig- und Salpetersäure, werden noch einmal 50 µm abgetragen. Da es sich hierbei um einen chemischen Vorgang handelt, wird die Oberfläche nicht geschädigt. Kristallfehler werden endgültig behoben.

Polieren

Dies ist der letzte Schritt zum fertigen Wafer. Am Ende des Polierschrittes besitzen die Wafer noch eine Unebenheit von weniger als 3 nm (0,000003 mm). Dazu werden die Scheiben mit einem Gemisch aus Natronlauge (NaOH), Wasser und Siliciumdioxidkörnern behandelt. Das Siliciumdioxid entfernt weitere 5 µm von der Scheibenoberfläche, die Natronlauge entfernt Oxid und beseitigt Bearbeitungsspuren der Siliciumdioxidkörner.

Danach folgen eine abschließende Reinigung und eine Vermessung jeder einzelnen Scheibe auf Ebenheit, Dicke und Partikel. Ein erheblicher Teil der Wafer erhält anschließend noch eine epitaktische Schicht: eine dünne,

besonders reine und definiert dotierte Siliciumlage, die auf die polierte Oberfläche aufgewachsen wird. Die eigentlichen Bauelemente entstehen dann in dieser Schicht, während die Scheibe darunter nur noch Träger ist.

Historische Entwicklung der Wafergröße

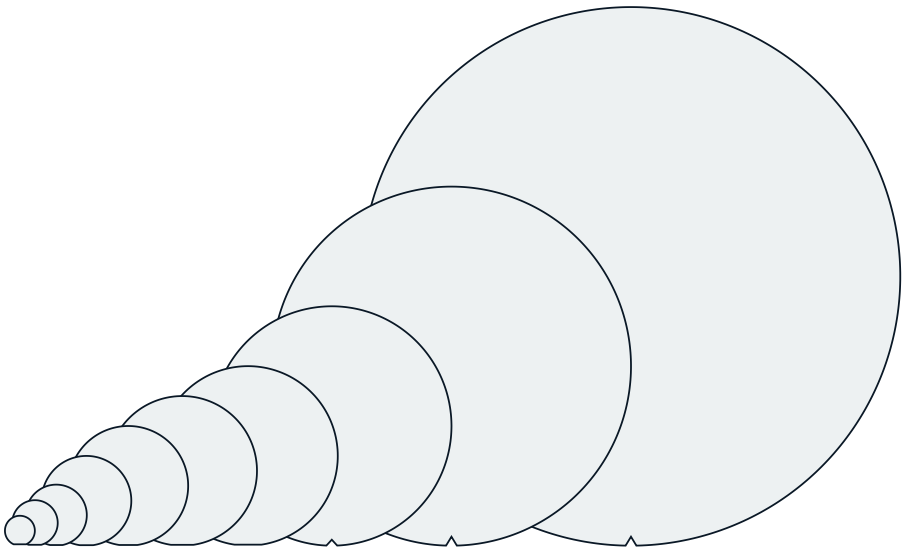
Die Herstellung von integrierten Schaltkreisen auf Siliciumwafern begann Mitte der 1960er Jahre auf Scheiben mit einem Durchmesser von 25 mm. Seither ist der Durchmesser in Stufen gewachsen; seit Ende der 1990er Jahre ist der 300-mm-Wafer das größte Format der Serienfertigung. Seine Fläche ist rund 140-mal so groß wie die der ersten 25-mm-Scheiben.

Mit größeren Wafern steigt der Durchsatz in der Herstellung von Chips erheblich, wodurch die Kosten in der Fertigung entsprechend gesenkt werden können. So können bei identischer Strukturgröße mehr als doppelt so viele Chips auf einem 300-mm-Wafer hergestellt werden, wie es auf einem 200-mm-Wafer der Fall ist.

Typische Daten von Wafern:

Typ [mm]	Durchmesser [mm]	Dicke [µm]	1. Flat [mm]	Durchbiegung [µm]
150	150 ± 0,5	~700	55 - 60	25
200	200 ± 0,5	~800	*Notch*	35
300	300 ± 0,5	~900	*Notch*	45

Wafergrößen in der Übersicht: 25, 38, 51, 75, 100, 125, 150, 200, 300, 450 [mm] (maßstabsgetreu)



Der Schritt auf 450 mm, der nicht kam

Die Reihe hätte sich mit 450 mm fortsetzen sollen. Um 2008 schlossen sich mehrere große Hersteller und Ausrüster zu einem gemeinsamen Programm zusammen, erste Anlagen und Testwafer entstanden, und der Umstieg wurde für den Anfang der 2010er Jahre angekündigt. Er fand nie statt: Das Programm wurde Mitte der 2010er Jahre stillgelegt, und 300 mm ist bis heute das größte Format geblieben.

Die Gründe dafür stehen bereits in den technischen Schwierigkeiten, die man damals zu lösen suchte:

- Die Durchbiegung wächst mit dem Durchmesser. Damit sich gestapelte Scheiben beim Transport nicht berühren, müssten Abstände, Auflagen und Greifer neu ausgelegt werden.
- Schichtspannungen verformen eine große Scheibe leichter. Dicker machen lässt sie sich nicht beliebig, weil Materialkosten und Eigenfrequenz dagegen sprechen.
- Die größere Fläche muss auch bearbeitet werden. Gegenüber 300 mm müssten die Kosten je Quadratzentimeter um den Faktor 2,25 sinken, damit sich am Ende überhaupt etwas rechnet.
- Ein 450-mm-Kristall wächst mehr als doppelt so lange. In dieser Zeit steigt auch die Wahrscheinlichkeit, dass sich Fehler in das Gitter einbauen.
- Nahezu jede Anlage der Fertigungslinie hätte neu entwickelt werden müssen – bei einem Investitionsbedarf, den einzelne Hersteller nicht mehr allein tragen können.

Der entscheidende Punkt war jedoch ein wirtschaftlicher. Der Nutzen eines größeren Wafers besteht darin, die Fixkosten je Anlage auf mehr Chips zu verteilen. Dieser Vorteil fällt umso geringer aus, je teurer die einzelne Anlage wird – und die Lithografieanlagen wurden in derselben Zeit dramatisch teurer. Der Aufwand für die Umstellung wäre also gestiegen, während der Ertrag daraus sank. Die Industrie hat ihre Investitionen stattdessen in kleinere Strukturen und in die dritte Dimension gelenkt: mehr Bauelemente je Fläche statt mehr Fläche je Scheibe.

Wieso sind Wafer rund?

Oft stellt sich die Frage wieso Wafer rund sind, schließlich sind Mikrochips doch rechteckig. Dadurch ergibt sich auf dem Wafer immer ein Verschnitt, also eine Fläche, auf der keine vollständigen Chips Platz finden, und die am Ende der Halbleiterfertigung verworfen werden muss.

Nach Erläuterung der beiden Herstellungsverfahren - dem Kristallziehverfahren und dem Zonenziehen - kann diese Frage leicht beantwortet werden.

Ein Siliciumwafer für die Mikrochipherstellung muss als Einkristall vorliegen. Dies ist nur mit den genannten Verfahren möglich, und diese liefern prinzipbedingt eine kreisrunde Form.

Auch wenn es technisch möglich wäre, die runden Siliciumkristalle nachträglich in eine eckige Form zu bringen (z. B. mittels sägen), so bietet die runde Form der Siliciumwafer trotz rechteckiger Mikrochips dennoch einige Vorteile.

- Das Begradigen der runden Siliciumstäbe bedeutet zusätzlichen Stress für das Material und würde zwangsläufig zu Kristallfehlern führen, die sich auf die Qualität der Chips auswirken würde.
- Runde Wafer sind wesentlich stabiler. Eckige Wafer könnten kaum ohne Beschädigung transportiert und bearbeitet werden.
- Eine gleichmäßige Bearbeitung während der Chipfertigung mit radialsymmetrischen Prozessen (CMP, Spin-on, Ätzen) ist wesentlich einfacher.
- Ein schmaler Randbereich müsste auch bei rechteckigen Wafern immer verworfen werden, da die Scheiben während der Bearbeitung gehalten werden müssen. Abgeschiedene Schichten würden abplatzen und so zusätzliche Partikel verursachen, wenn diese bis an den äußersten Rand reichen.

Mit zunehmender Wafergröße nimmt der Verschnitt auch immer weiter ab.

Rechteckige Wafer findet man hingegen bei der Fertigung von Solarzellen. Zumeist finden hier polykristalline Wafer Anwendung, die in rechteckigen Formen gegossen werden können. Die Fertigung ist verhältnismäßig einfach, so dass auch eckige Wafer bearbeitet werden können. Meist werden die Ecken zusätzlich abgeschrägt. Bei den heute üblichen einkristallinen Solarzellen ist es genau umgekehrt: Sie entstehen aus rund gezogenen Kristallen, die vor dem Sägen quadratisch beschnitten werden – mit abgerundeten Ecken, weil das Wegschneiden der Restkante mehr Material kosten würde als der Flächengewinn einbringt.

Dotiertechniken

Dotieren

Dotieren bedeutet das Einbringen eines Fremdstoffs in den Halbleiterkristall zur gezielten Änderung der Leitfähigkeit durch Elektronenüberschuss oder Elektronenmangel. Im Gegensatz zum Dotieren bei der Waferherstellung selbst, wo der ganze Wafer dotiert wird, ermöglichen die auf dieser Seite beschriebenen Dotiertechniken das partielle Dotieren der Siliciumscheiben. Das

Einbringen des Fremdstoffs kann mit verschiedenen Methoden erreicht werden: Diffusion, Implantation (und Legierung).

Diffusion

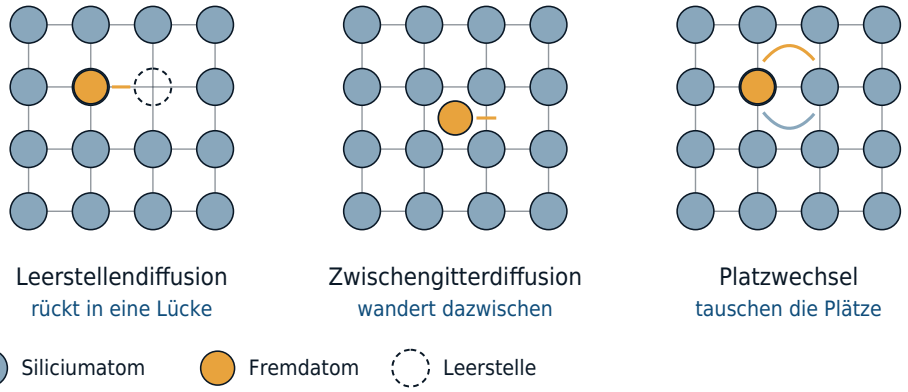
Zur gezielten Dotierung wird die Diffusion heute kaum noch eingesetzt; sie ist durch die Ionenimplantation abgelöst worden. Zu verstehen ist sie trotzdem unverzichtbar, denn sie hört nicht auf zu wirken: Jeder spätere Hochtemperaturschritt lässt bereits eingebrachte Dotierstoffe weiterwandern. Das begrenzt, wie viel Wärme einem Wafer nach der Dotierung noch zugemutet werden darf, und ist der Grund, warum Ofenprozesse zunehmend durch sekundenschnelles Heizen ersetzt wurden.

Diffusion bedeutet das selbständige Ausbreiten eines Stoffes in einem anderen Stoff auf Grund eines Konzentrationsunterschiedes, so verteilt sich z.B. ein Tropfen Tinte in einem Wasserglas nach einer bestimmten Zeit gleichmäßig. Im Siliciumkristall finden wir ein festes Gitter von Atomen vor, durch das sich der Dotierstoff bewegen muss. Das kann auf drei Arten geschehen:

- Leerstellendiffusion: Die Fremdatome besetzen leere Stellen im Kristallgitter die immer auftreten können. Ähnlich einer Löcherleitung entsteht ein Wechselspiel von besetzten Gitterplätzen und Lücken.
- Zwischengitterdiffusion: Die Fremdatome bewegen sich zwischen den Siliciumatomen im Kristallgitter hindurch.
- Platzwechsel: Fremdatome die sich im Kristallgitter befinden tauschen mit Siliciumatomen den Gitterplatz.

Diffusion im Kristallgitter

Drei Wege, auf denen der Dotierstoff durch das Gitter wandert



Alle drei Wege laufen nebeneinander ab. Welcher überwiegt, hängt vom Dotierstoff und von der Temperatur ab – Bor und Phosphor wandern vor allem über Leerstellen, kleinere Atome eher zwischen den Gitterplätzen.

Der Dotierstoff kann sich so lange im Halbleiterkristall bewegen, bis entweder ein Konzentrationsgefälle ausgeglichen ist, oder die Temperatur so weit gesenkt wurde, dass die Atome sich nicht mehr bewegen können.
Die Geschwindigkeit des Diffusionsvorgangs hängt von mehreren Faktoren ab:

- Dotierstoff
- Konzentrationsunterschied
- Temperatur
- Substrat
- Kristallorientierung des Substrats (siehe Herstellung von Einkristallen)

Diffusion mit erschöpflicher Quelle

Diffusion mit einer erschöpflichen Quelle bedeutet, dass der Dotierstoff nur begrenzt zur Verfügung steht. Je länger der Diffusionsprozess andauert, umso geringer wird die Konzentration an der Oberfläche; dafür steigt die Eindringtiefe in das Substrat. Der Diffusionskoeffizient eines Stoffes gibt dabei an, wie schnell er sich im Kristall bewegt. Arsen mit einem geringen Diffusionskoeffizienten dringt langsamer in das Substrat ein, als bspw. Phosphor oder Bor.

Diffusion mit unerschöpflicher Quelle

Der Dotierstoff steht hier unbegrenzt zur Verfügung. Die Konzentration an der Oberfläche bleibt somit während des Vorgangs konstant, da Teilchen, die in das Substrat eingedrungen sind, fortwährend nachgeliefert werden.

Diffusionsverfahren

Bei den nachfolgenden Prozessen befinden sich die Wafer in einem Quarzrohr, das über die gesamte Länge auf eine bestimmte Temperatur aufgeheizt wird.

Diffusion aus der Gasphase

Ein Trägergas (Stickstoff, Argon) wird mit dem gewünschten Dotierstoff (ebenfalls in Gasform, z.B. Phosphin (PH_3) oder Diboran (B_2H_6) angereichert und über die Siliciumscheiben geleitet, wo der Konzentrationsausgleich stattfinden kann.

Feststoffdiffusion

Zwischen den Wafern befinden sich Scheiben, auf denen der Dotierstoff aufgebracht wurde. Steigt die Temperatur im Quarzrohr, diffundiert von den Quellscheiben der Dotierstoff in die Atmosphäre aus. Über ein Trägergas wird

der Dotierstoff dann gleichmäßig im Quarzrohr verteilt und gelangt so an die Oberfläche der Wafer.

Diffusion mit flüssiger Quelle

Als flüssige Quelle dienen Borbromid (BBr_3) oder Phosphoroxychlorid ($POCl_3$). Ein Trägergas wird durch die Flüssigkeit geleitet und transportiert den Dotierstoff so zu den Wafern, wo er dann an die Oberfläche der Siliciumscheiben gelangt.

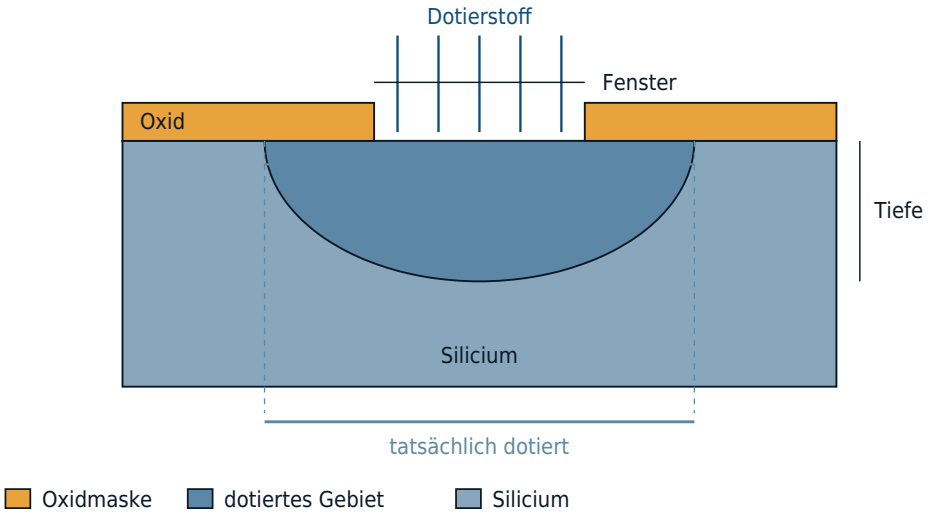
Da nicht der gesamt Wafer dotiert werden soll, werden bestimmte Bereiche mit Siliciumdioxid maskiert. Das Oxid können die Dotierstoffe nicht durchdringen, weshalb an diesen Stellen keine Dotierung stattfindet. Um Spannungen oder gar Brüche der Scheiben zu vermeiden, wird das Rohr schrittweise (10 °C pro Minute) auf ca. 900 °C aufgeheizt, wo der Dotierstoff dann zu den Scheiben geleitet wird. Um den Diffusionsvorgang in Gang zu setzen, wird die Temperatur anschließend auf ca. 1200 °C erhöht.

Charakteristik:

- Da viele Wafer gleichzeitig bearbeitet werden können ist dieses Verfahren recht günstig
- Befinden sich bereits Fremdstoffe früherer Dotierungen im Kristall können diese bei einer erneuten Temperaturbelastung ausdiffundieren
- Im Quarzrohr lagern sich mit der Zeit Dotierstoffe an, die bei nachfolgenden Dotierungen zusätzlich vom Trägergas zu den Wafern transportiert werden
- Dotierstoffe breiten sich im Kristall nicht nur senkrecht, sondern auch seitlich aus, so dass die Dotierfenster immer großflächiger dotiert werden als gewünscht

Diffusion durch ein Fenster in der Maske

Der Dotierstoff wandert auch zur Seite und unterwandert das Oxid



Die seitliche Ausbreitung beträgt etwa acht Zehntel der Eindringtiefe. Das dotierte Gebiet fällt dadurch immer breiter aus als das Fenster – beim Entwurf der Masken muss das einkalkuliert werden.

Ionenimplantation

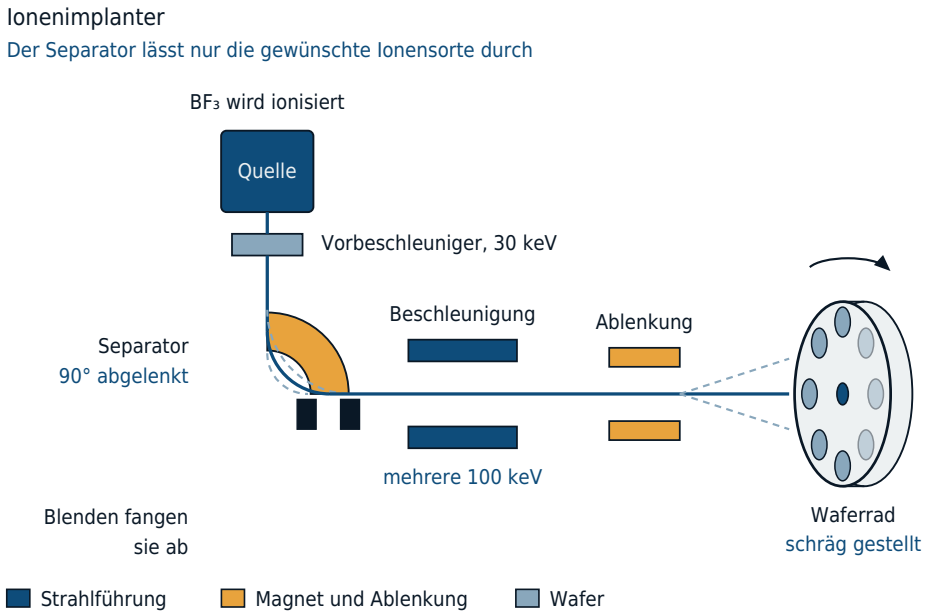
Bei der Ionenimplantation werden geladene Dotierstoffe (Ionen) in einem elektrischen Feld beschleunigt und auf die Wafer gelenkt. Die Eindringtiefe lässt sich sehr genau festlegen, indem die zur Beschleunigung der Ionen benötigte Spannung verringert oder erhöht wird. Da der Prozess bei Raumtemperatur stattfindet, können vorher eingebrachte Dotierungen nicht ausdiffundieren. Wie bei der Diffusion werden Stellen, die nicht dotiert werden sollen, mit einer Maske verdeckt, wobei bei der Implantation eine Maskierung aus Fotolack ausreicht.

Ein Implanter besteht aus folgenden Komponenten:

- Ionenquelle: Das Dotiergas (z.B. Bortrifluorid BF_3) wird ionisiert (Elektronen werden von einer Glühkathode emittiert und stoßen mit den Gasteilchen zusammen. Durch Stoßionisation werden stets positive Ionen und freie Elektronen erzeugt)
- Vorbeschleuniger: Die Ionen werden mit ca. 30 Kiloelektronenvolt aus der Ionenquelle gezogen
- Massenseperator: Die geladenen Teilchen werden durch ein Magnetfeld um 90° abgelenkt. Zu leichte/schwere Teilchen werden mehr/weniger abgelenkt als die gewünschten Ionen und mit Blenden hinter dem Seperator abgefangen

- Beschleunigungsstrecke: Mit mehreren 100 keV werden die Teilchen auf ihre Endenergie beschleunigt (200 keV beschleunigen Borionen auf ca. 2.000.000 m/s)
- Linsen: Über das gesamte System sind Linsen verteilt, die den Ionenstrahl fokussieren
- Ablenkungsvorrichtungen: Kondensatoren lenken die Ionen ab, um die gewünschte Stelle zu bestrahlen
- Waferstation: Die Wafer werden entweder einzeln oder auf großen rotierenden Rädern in den Ionenstrahl gebracht und bestrahlt

Darstellung einer Implantationsanlage



Nur Ionen mit der passenden Masse durchlaufen den Separator auf der richtigen Bahn. Leichtere werden stärker, schwerere schwächer abgelenkt und laufen in die Blenden. So wird die Dotierung sortenrein.

Eindringtiefe von Ionen im Wafer

Im Gegensatz zur Diffusion dringen die Teilchen nicht auf Grund ihrer Eigenbewegung ein, sondern werden mit hoher Geschwindigkeit in das Kristallgitter geschossen. Dabei werden sie durch Zusammenstöße mit den Siliciumatomen abgebremst. Durch den Aufprall werden die Siliciumatome von ihren Gitterplätzen gestoßen, die Dotierionen selbst lagern sich meist auf Zwischengitterplätzen an. Dort sind sie elektrisch nicht aktiv, da keine Bindungen mit anderen Atomen vorliegen, die freie Ladungsträger hervorrufen könnten. Die verschobenen Siliciumatome müssen wieder ins Kristallgitter eingebaut, und die elektrisch nicht aktiven Dotierstoffe aktiviert werden.

Ausheilen des Kristallgitters und Aktivierung der Dotierstoffe

Durch einen Temperaturschritt bei ca. 1000 °C werden die Dotierstoffe auf Gitterplätze bewegt (vorher befinden sich nur ca. 5 % der Dotieratome auf Gitterplätzen). Die Gitterschäden durch die Zusammenstöße werden bereits bei ca. 500 °C ausgeheilt. Da sich die Dotieratome während den hohen Temperaturen im Substrat bewegen, werden diese Schritte nur sehr kurze Zeit durchgeführt.

Genau hier liegt der Zielkonflikt des Verfahrens: Die Temperatur muss hoch genug sein, um die Dotierstoffe auf Gitterplätze zu bringen, und die Zeit kurz genug, damit sie dabei nicht auswandern. Beides gleichzeitig erreicht man nur, indem man die Heizdauer immer weiter verkürzt. Aus dem Ofenprozess über Stunden wurde zunächst das Schnellheizen über Sekunden, dann das Spitzenheizen, bei dem die Zieltemperatur nur durchfahren und nicht gehalten wird, und schließlich das Heizen mit Blitzlampen oder Laser im Millisekundenbereich. Dabei wird nur noch die oberste Schicht der Scheibe erwärmt, während das Substrat darunter kalt bleibt.

Für sehr flache Dotierprofile wird zusätzlich die Ionenmasse erhöht: Statt einzelner Boratome implantiert man Moleküle, die mehrere Boratome enthalten. Sie tragen bei gleicher Beschleunigungsspannung je Boratom weniger Energie und dringen entsprechend weniger tief ein – ein Weg, sehr flache Übergänge zu erzeugen, ohne den Implanter mit unpraktikabel niedrigen Spannungen betreiben zu müssen.

Channeling

Das verwendete Substrat liegt als Einkristall vor, d.h. die Siliciumatome sind regelmäßig angeordnet und bilden Kanäle. Die eingeschossenen Dotierstoffatome verlaufen dann parallel zu diesen Kanälen, werden nur schwach abgebremst und dringen sehr tief in das Substrat ein. Um dies zu verhindern gibt es zwei Möglichkeiten:

- Waferausrichtung: Die Wafer werden um ca. 7° zur Strahlrichtung ausgelenkt. Dadurch werden die Ionen nicht parallel zu den Gitterkanälen eingeschossen und durch Zusammenstöße frühzeitig abgebremst.
- Streuoxid: Auf der Waferoberfläche wird ein dünnes Oxid aufgebracht, das die Ionen ablenkt und ein senkrechtes Eintreffen im Substrat verhindert

Charakteristik

- Die Reproduzierbarkeit der Ionenimplantation ist sehr hoch
- Der Ablauf bei Raumtemperatur verhindert ein Ausdiffundieren anderer Dotierstoffe
- Als Maske dient Fotolack, Oxid wie bei der Diffusion ist nicht nötig
- Ionenimplanter sind teuer, die Kosten pro bearbeiteter Scheibe sind hoch
- Die Dotierstoffe breiten sich nicht seitlich unter der Maskierung aus (nur minimal durch Zusammenstöße)
- Nahezu jedes Element kann in höchster Reinheit implantiert werden
- Ähnlich den Ablagerungen von Dotierstoffen bei der Diffusion im Quarzrohr, können sich an Wänden oder Blenden Ionen ablagern, die bei späteren Implantationen abgelöst werden und auf die Wafer gelangen
- Senkrechte Flächen lassen sich mit einem gerichteten Ionenstrahl schlecht dotieren. Bei schräg gestellter Scheibe gelingt es begrenzt; für Strukturen wie die Finnen eines FinFET wird stattdessen aus einem Plasma heraus dotiert, dessen Ionen aus allen Richtungen auftreffen
- Der Implantationsprozess findet unter Hochvakuum statt, welches mit mehreren Turbomolekular- oder Kryopumpen erzeugt werden muss

Es gibt verschiedene Implantertypen, wobei meistens Mittel- und Hochstromimplanter zum Einsatz kommen. Mittelstromimplanter sind für kleine bis mittlere Dosen an Ionen geeignet ($1 \cdot 10^{11}$–$1 \cdot 10^{15}$ Ionen/cm^2), Hochstromimplanter für Dosen von $1 \cdot 10^{15}$–$1 \cdot 10^{17}$ Ionen/cm^2.
Die Ionenimplantation hat sich auf Grund der Vorteile gegenüber der Diffusion größtenteils durchgesetzt.

Dotieren mittels Legierung

Der Vollständigkeit halber sei noch erwähnt, dass es neben diesen beiden Verfahren noch die Dotierung mittels Legierung gibt. Da dieses Verfahren aber Nachteile, wie z.B. Rissbildung im Substrat mit sich bringt, wird es in der heutigen Halbleitertechnik kaum mehr eingesetzt.

Oxidation

Industrielle Verwendung

Anwendung von Oxidschichten

Oxid findet in der Halbleiterfertigung Anwendung in den verschiedensten Gebieten:

- zur Isolation (z.B. zwischen Metallisierungsschichten)
- als Streuschicht (z.B. Ionenimplantation)
- als Anpassungsschicht (z.B. LOCOS Technik)
- zur Planarisierung (z.B. um Kanten zu entschärfen)
- als Maskierschicht (z.B. Diffusion)
- als Justiermarken (Justierpunkte in der Fototechnik)
- als Schutzschicht (vor mechanischer Beschädigung)

Eigenschaften von Oxidschichten

In Verbindung mit Silicium tritt Oxid als Siliciumdioxid (SiO_2) auf. Es lässt sich auf dem Wafer in sehr dünnen Schichten sehr gleichmäßig herstellen.

SiO_2 ist sehr widerstandsfähig und kann in der Halbleiterfertigung nur durch Flusssäure (HF) nasschemisch geätzt werden. Wasser und andere Säuren greifen das Oxid nicht an, wegen der Kontaminationsgefahr durch Metallionen können Alkalilaugen (KOH, NaOH u.a.) nicht eingesetzt werden. Diese spielen jedoch in der Mikromechanik eine Rolle beim anisotropen Nassätzen – dort greifen sie das Silicium an, während das Oxid nahezu unberührt bleibt und deshalb gerade als Maskierung dient. Der Ablöseprozess mit HF läuft nach folgender Reaktion ab:

$$SiO_2 + 6HF \rightarrow H_2SiF_6 + 2H_2O$$

Des Weiteren bietet sich Oxid zur Funktion von Schaltungen an, da es die unterschiedlichsten Anforderungen erfüllt (z.B. als Gateoxid, Feldoxid oder Zwischenoxid).

Entscheidend ist dabei weniger das Oxid selbst als die Grenzfläche zum Silicium. Sie lässt sich thermisch so störstellenarm erzeugen, dass darüber ein Transistor zuverlässig schaltet. Kein anderer Halbleiter besitzt ein eigenes Oxid mit dieser Eigenschaft – das ist der Hauptgrund, weshalb sich Silicium gegen Materialien durchgesetzt hat, die elektrisch eigentlich günstiger wären.

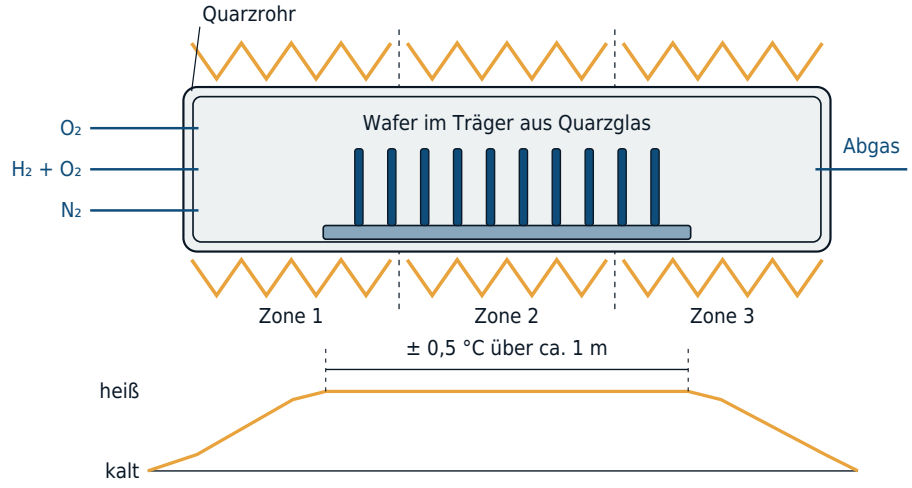
Erzeugung von Oxidschichten

Thermische Oxidation

Bei der thermischen Oxidation werden die Siliciumwafer bei ca. 1000 °C in einem Oxidationsofen oxidiert. Dieser Ofen besteht im Wesentlichen aus einem Quarzrohr in dem sich die Wafer auf einem Träger (Carrier) aus Quarzglas befinden, mehreren getrennt regelbaren Heizwicklungen und verschiedenen Gaszuleitungen. Das Quarzglas besitzt einen sehr hohen Schmelzpunkt (weit über 1500 °C) und ist deshalb sehr gut für Hochtemperaturprozesse geeignet. Damit es nicht zu Scheibenverzug oder Scheibensprüngen kommt, wird das Quarzrohr in sehr kleinen Schritten (max. 10 °C pro Minute) aufgeheizt. Die Temperierung ist mittels der getrennten Heizwicklungen im gesamten Rohr über eine Länge von ca. 1 m auf ± 0,5 °C exakt regelbar.

Oxidationsofen

Quarzrohr mit Waferträger und getrennt geregelten Heizzonen



Prozessgase: O_2 für die trockene, $H_2 + O_2$ für die nasse Oxidation, N_2 zum Spülen.
Prozesstemperatur etwa 800 bis 1200 °C, Aufheizrate höchstens 10 °C pro Minute.

Der Sauerstoff strömt dann als Gas über die Wafer und reagiert an der Oberfläche zu Siliciumdioxid. Es entsteht eine glasartige Schicht mit amorpher Struktur. Je nach Prozessgas finden dann verschiedene Oxidationen statt, eine thermische Oxidation muss naturgemäß auf einer Siliciumoberfläche erfolgen. Die thermische Oxidation unterteilt sich in die trockene und feuchte Oxidation, welche sich wiederum in die nasse Oxidation und die H_2-O_2-Verbrennung gliedern lässt.

Trockene Oxidation

Der Oxidationsprozess findet unter reiner Sauerstoffatmosphäre statt. Dabei reagiert Silicium mit Sauerstoff zu Siliciumdioxid:

$Si + O_2 \rightarrow SiO_2$

Dieser Prozess findet in der Regel bei 1000–1200 °C statt. Zur Erzeugung von sehr stabilen und dünnen Oxiden wird die Oxidation bei ca. 800 °C durchgeführt.

Eigenschaften der Trockenoxidation

- langsames Oxidwachstum
- hohe Dichte
- hohe Durchbruchspannung (für elektrisch stark beanspruchte Oxide, z.B. Gateoxid)

Nasse Oxidation

Bei der nassen Oxidation wird der Sauerstoff durch ein Bubbler-Gefäß mit Wasser (ca. 95 °C) geleitet, so dass sich zusätzlich zum Sauerstoff auch Wasser in Form von Wasserdampf im Quarzrohr befindet. Daraus ergibt sich folgende Reaktionsgleichung:

$Si + 2H_2O \rightarrow SiO_2 + 2H_2 \uparrow$

Dieser Prozess findet bei 900–1000 °C statt. Er weist ein schnelles Oxidwachstum bei niedrigen Temperaturen auf und wird u. a. zur Herstellung von Maskierschichten und Feldoxiden verwendet. Die Qualität der erzeugten Schicht ist geringer als bei der Trockenoxidation.

Vergleich der Aufwachsrate bei trockener und nasser Oxidation

Temperatur	Trockene Oxidation	Nasse Oxidation
900 °C	19 nm/h	100 nm/h
1000 °C	50 nm/h	400 nm/h
1100 °C	120 nm/h	630 nm/h

H_2-O_2-Verbrennung

Bei der H_2-O_2-Verbrennung wird neben hochreinem Sauerstoff auch hochreiner Wasserstoff verwendet. Die beiden Gase werden getrennt in das Quarzrohr geleitet und an der Eintrittsöffnung verbrannt. Damit es nicht zu einer Knallgasreaktion mit dem hochbrennbaren Wasserstoff kommt, muss die Temperatur über 500 °C liegen, die Gase reagieren dann in einer stillen Verbrennung. Dieses Verfahren ermöglicht die Erzeugung von schnell

wachsenden und nur wenig verunreinigten Oxidschichten. Damit lassen sich sowohl dicke Oxide, als auch dünne Schichten bei vergleichsweise geringer Temperatur (900 °C) herstellen. Die niedrige Temperatur erlaubt auch die thermische Belastung von bereits dotierten Wafern (siehe Dotieren mittels Diffusion).

Bei allen thermischen Oxidationen ist das Oxidwachstum auf 111-orientierten Substraten höher als auf 100-orientierten (siehe der Einkristall). Außerdem erhöht ein sehr hoher Anteil an Dotierstoffen im Substrat das Wachstum deutlich.

Ablauf des Oxidationsvorgangs

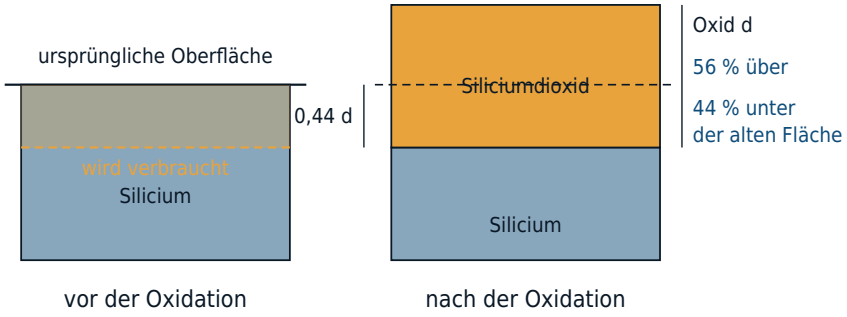
Zu Beginn reagiert der Sauerstoff an der Waferoberfläche zu Siliciumdioxid. Nun befindet sich eine Oxidschicht auf dem Substrat durch die nachfolgende Sauerstoffatome zunächst diffundieren müssen, um mit dem Silicium reagieren zu können. Die Aufwachsrate hängt nur zu Beginn von der Reaktionszeit zwischen Silicium und Oxid ab; ab einer gewissen Dicke wird die Oxidationsgeschwindigkeit davon bestimmt, wie schnell der Sauerstoff durch das bereits aufgewachsene Siliciumdioxid hindurchdiffundiert. Mit zunehmender Oxiddicke verlangsamt sich also das Wachstum. Rechnerisch beschreibt man das mit einem linear-parabolischen Ansatz nach Deal und Grove: Anfangs wächst die Dicke etwa proportional zur Zeit, später nur noch proportional zur Wurzel aus der Zeit. Dicke Oxide sind deshalb unverhältnismäßig zeitaufwendig. Da die entstandene Schicht amorph ist, sind nicht alle Bindungen der Siliciumatome intakt; es gibt teilweise freie Bindungen (freie Elektronen und Löcher) an der Si-SiO_2-Grenzschicht. So ergibt sich an diesem Übergang insgesamt eine leicht positive Ladung. Da sich diese Ladung negativ auf Bauteile auswirken kann wird versucht, sie so gering wie möglich zu halten. Dies kann beispielsweise mit einer höheren Oxidationstemperatur erreicht werden, oder durch die Verwendung der nassen Oxidation, die ebenfalls nur eine sehr geringe Ladung verursacht.

Volumenzuwachs

Bei der thermischen Oxidation wird Silicium durch die Reaktion mit Sauerstoff zu Siliciumdioxid verbraucht. Das Verhältnis der aufgewachsenen Oxidschicht zu verbrauchtem Silicium beträgt 2,27; d.h. das Oxid wächst zu 45 % der Oxiddicke in das Substrat ein.

Aufwachsverhalten von Oxid auf Silicium

Das Oxid nimmt mehr Raum ein als das Silicium, aus dem es entsteht



Aus einer Siliciumschicht der Dicke 1 entsteht Oxid der Dicke 2,27.
Die Grenzfläche wandert dabei nach innen: Das Oxid wächst in das Substrat hinein.

Segregation

Dotierstoffe die sich im Substrat befinden, können im Siliciumkristall oder im Oxid eingebaut werden, dies hängt davon ab, in welchem Material sich der Dotierstoff besser löst. Dieser sogenannte Segregationskoeffizient k berechnet sich nach:

$$k=\frac{Löslichkeit\ des\ Dotierstoffs\ in\ Si}{Löslichkeit\ des\ Dotierstoffs\ in\ SiO_2}$$

Ist k größer 1 werden die Dotierstoffe an der Oberfläche des Substrats eingebaut, bei k kleiner 1 reichern sich die Dotierstoffe im Oxid an.

Wo die thermische Oxidation heute steht

Die klassische Aufgabe des thermischen Oxids, das Gatedielektrikum zu bilden, ist entfallen. Bei Schichtdicken von wenigen Atomlagen tunnelt so viel Strom durch das Oxid, dass es als Isolator nicht mehr taugt. An seine Stelle ist Hafniumdioxid getreten, das per Atomlagenabscheidung aufgebracht wird und bei gleicher Wirkung dicker sein darf. Ganz ohne thermisches Oxid geht es allerdings auch dort nicht: Zwischen Silicium und Hafniumdioxid liegt eine nur wenige Atomlagen dicke Zwischenschicht aus Siliciumdioxid, denn die störstellenarme Grenzfläche zum Silicium lässt sich anders nicht herstellen.

Erhalten geblieben sind der thermischen Oxidation die Aufgaben, für die gerade ihre Eigenart nützlich ist, aus dem Substrat selbst zu wachsen: das Padoxid unter Nitridmasken, Opferoxide zum Schutz der Oberfläche vor Implantationen, die Auskleidung frisch geätzter Gräben bei der Grabenisolation und die Oxide der Leistungselektronik, wo es weiterhin auf Dicke und Durchbruchfestigkeit ankommt.

Geändert hat sich auch die Anlagentechnik. Der oben beschriebene Ofen bearbeitet eine ganze Charge über Stunden. Für dünne Schichten ist das zu ungenau und thermisch zu belastend; dort wird eine einzelne Scheibe mit Lampen innerhalb von Sekunden auf Temperatur gebracht und ebenso schnell wieder abgekühlt. Die eingebrachte Wärmemenge bleibt dadurch so gering, dass sich vorhandene Dotierprofile kaum noch verschieben.

Oxidation durch Abscheidung

Bei der thermischen Oxidation wird Silicium des Wafers zur Oxidbildung verbraucht. Ist die Siliciumoberfläche jedoch durch andere Schichten verdeckt, muss man das Oxid über Abscheideverfahren aufbringen, bei denen neben Sauerstoff auch Silicium selbst hinzugefügt wird. Die zwei wichtigsten Verfahren dabei sind die Silanpyrolyse und die TEOS-Abscheidung. Eine ausführliche Beschreibung der Verfahren folgt im Kapitel Abscheidung.

Silanpyrolyse

Pyrolyse bedeutet, dass chemische Verbindungen durch Wärme gespaltet werden, in diesem Fall das Gas Silan SiH_4 und hochreiner Sauerstoff O_2. Da sich das giftige Silan bei einer Konzentration von 3 % in der Umgebungsluft selbst entzündet muss es mit Stickstoff oder Argon auf 2 % verdünnt werden. Bei ca. 400 °C reagieren Silan und Sauerstoff zu Siliciumdioxid und Wasserstoff:

$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2 \uparrow$

Das Siliciumdioxid ist nur von geringer Qualität. Alternativ kann eine Hochfrequenzanregung über ein Plasma bei ca. 300 °C zur Dioxidabscheidung verwendet werden. So entsteht ein etwas stabileres Oxid.

TEOS-Abscheidung

Das bei dieser Methode verwendete Tetraethylorthosilicat, kurz TEOS ($SiO_4C_8H_{20}$) enthält die beiden benötigten Elemente Silicium und Sauerstoff. Unter Vakuum geht die bei Raumtemperatur flüssige Verbindung bereits in Gasform über. Das Gas wird in ein beheiztes Quarzrohr überführt und dort bei ca. 750 °C gespaltet.

$SiO_4C_8H_{20} \rightarrow SiO_2 + \text{Nebenprodukte}$

Das Siliciumdioxid scheidet sich auf den Wafern ab, die Nebenprodukte (z.B. H_2O - Wasserdampf) werden abgesaugt. Die Gleichmäßigkeit dieses Oxids wird durch den Druck im Quarzrohr und die Prozesstemperatur bestimmt. Es ist elektrisch stabil und sehr rein.

Aus diesen Eigenschaften ergibt sich die Arbeitsteilung: Wo es auf die

Grenzfläche zum Silicium ankommt, wird thermisch oxidiert; wo Oxid auf etwas anderem als blankem Silicium liegen soll oder wo die Temperatur begrenzt ist, wird abgeschieden. Über der ersten Metallebene ist ausschließlich das Abscheiden möglich, denn die Leiterbahnen vertragen die Ofentemperaturen nicht.

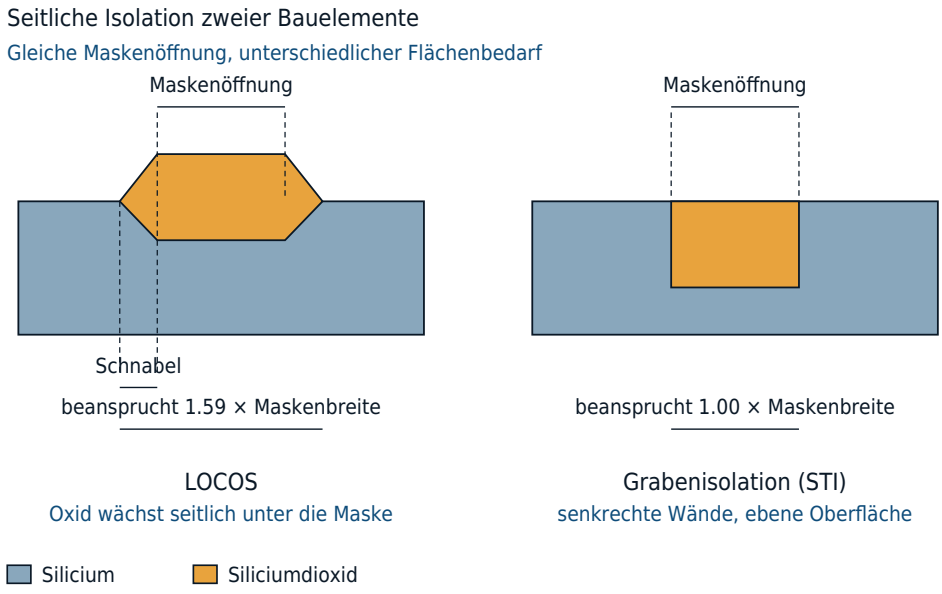
Bauelementisolation

Grabenisolation

Die lokale Oxidation stößt bei kleinen Strukturen an ihre Grenze: Der Vogelschnabel wird nicht kleiner, wenn die Bauelemente es werden. Er beansprucht eine feste Länge, die sich aus der Dicke von Padoxid und Nitrid ergibt, und verbraucht damit einen immer größeren Anteil der verfügbaren Fläche. Unterhalb einer Strukturgröße von etwa einem halben Mikrometer wurde das Verfahren dadurch unbrauchbar.

An seine Stelle ist die Grabenisolation getreten (engl. shallow trench isolation, STI). Statt das Oxid aufwachsen zu lassen, wird ein Graben in das Silicium geätzt und mit Oxid gefüllt. Die Isolation ist damit genau so breit wie die Maskenöffnung, und die Oberfläche bleibt eben.

Vergleich der beiden Verfahren



Ablauf

Die Grabenisolation ist einer der ersten Schritte der gesamten Fertigung; sie wird angelegt, bevor irgendein Bauelement entsteht.

1. Auf dem Wafer werden ein dünnes Padoxid und darüber eine Siliciumnitridschicht aufgebracht. Das Nitrid dient später als Stopp beim Polieren.
2. Über eine Lackmaske werden Nitrid, Padoxid und das darunterliegende Silicium anisotrop geätzt. Die Gräben erhalten dabei bewusst leicht schräge Wände, damit sie sich anschließend besser füllen lassen.
3. Eine kurze thermische Oxidation rundet die scharfen Kanten am Grabenrand ab und beseitigt die Ätzschäden an der Grabenwand. Scharfe Kanten würden das elektrische Feld überhöhen und die Einsatzspannung des benachbarten Transistors verfälschen.
4. Der Graben wird mit Oxid gefüllt. Da er schmal und tief ist, kommen dafür Verfahren zum Einsatz, die keinen Hohlraum hinterlassen.
5. Das überschüssige Oxid wird durch chemisch mechanisches Polieren abgetragen, bis das Nitrid freiliegt.
6. Das Nitrid wird nasschemisch in heißer Phosphorsäure entfernt, das Padoxid in Flusssäure. Zurück bleiben ebene Siliciumflächen, umgeben von Oxid.

Der Preis für die kompaktere Isolation ist die Zahl der Schritte: Wo LOCOS mit Maskieren und Oxidieren auskam, sind hier Ätzen, Füllen, Polieren und zwei Nassprozesse nötig. Ohne das chemisch mechanische Polieren wäre das Verfahren gar nicht möglich – erst damit lässt sich das Füllmaterial sauber auf die Höhe der Siliciumoberfläche zurücknehmen.

Weitere Bauformen

Die Tiefe der Gräben richtet sich danach, was voneinander getrennt werden soll. Zur Trennung benachbarter Transistoren genügen einige hundert Nanometer. Sollen dagegen ganze Schaltungsteile gegeneinander isoliert werden, etwa in der Leistungs- und Analogtechnik oder um empfindliche analoge Bereiche vor Störungen aus dem digitalen Teil zu schützen, werden mehrere Mikrometer tiefe Gräben angelegt (engl. deep trench isolation).

Bei Bauelementen mit senkrecht stehenden Kanälen verliert selbst die Grabenisolation ihre Rolle als flächenbestimmender Faktor: Dort trennt nicht mehr ein Graben zwischen zwei nebeneinanderliegenden Gebieten, sondern die Bauelemente stehen ohnehin frei.

Höchstintegration auf Chips

In der Halbleitertechnik werden Strukturen mittels Belichtungs- und

Ätzverfahren erzeugt. Dabei entstehen Stufen, an denen sich Fotolack ansammeln kann und so das Auflösungsvermögen in der Fototechnik verringert wird. Bei einer isotropen Ätzcharakteristik (der Abtrag erfolgt sowohl in vertikaler als auch horizontaler Richtung) müssen Lackmasken angepasst werden, damit unterätzte Strukturen am Ende des Ätzprozesses die richtigen Abmessungen besitzen.

An diesen Stufen treten auch Probleme bei der Metallisierung auf, da die Leiterbahnen hier verengt werden, so dass Schäden durch Elektromigration die Folge sind.

Um eine hohe Packungsdichte zu erreichen, also möglichst viele Bauelemente auf möglichst geringer Fläche unterzubringen, müssen die Stufen und Unebenheiten vermieden werden. Die erste Antwort darauf war die LOCOS-Technik: LOCal Oxidation of Silicon (Lokale Oxidation von Silicium). Sie prägte den Aufbau integrierter Schaltungen über zwei Jahrzehnte und ist bis heute in der Leistungs- und Analogtechnik anzutreffen.

Für feine Strukturen reicht sie jedoch nicht mehr aus. Warum das so ist, zeigt sich an ihrem charakteristischen Nebeneffekt, der im folgenden Abschnitt beschrieben wird; die heute übliche Grabenisolation vermeidet ihn.

Der Vogelschnabel

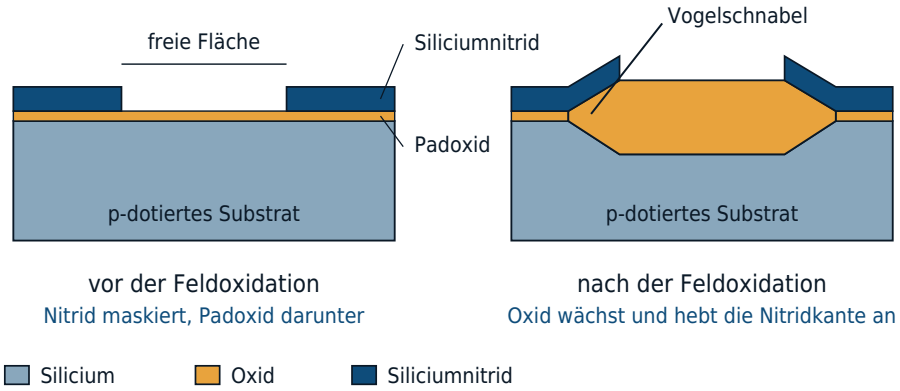
Bei der LOCOS-Technik nutzt man die unterschiedlichen Oxidationsgeschwindigkeiten von Silicium und Siliciumnitrid zur lokalen Maskierung der Scheibenoberfläche aus.

Mit einer Siliciumnitridschicht maskiert man die Stellen an denen kein Oxid aufwachsen soll, es bildet sich nur eine Oxidschicht auf den Nitrid freien Bereichen. Da Silicium und Siliciumnitrid unterschiedliche Ausdehnungskoeffizienten besitzen, wird eine dünne Schicht Oxid, das Padoxid, zwischen Nitridmaske und Substrat aufgebracht um Spannungen durch Temperaturänderungen zu vermeiden.

Zur seitlichen Isolation von Transistoren bringt man nun ein Feldoxid (FOX) auf der freien Siliciumoberfläche auf. Während sich bei der Feldoxidation auf dem Silicium eine Siliciumdioxidschicht bildet, verursacht das Padoxid eine seitliche Sauerstoffdiffusion unter die Nitridmaske und somit ein leichtes Oxidwachstum am Rand der Maskierung. Der Oxidausläufer hat die Form eines Vogelschnabels, dessen Länge vom Oxidationsprozess, sowie von der Dicke des Nitrids und des Padoxids abhängt.

LOCOS: vor und nach der Feldoxidation

Das Oxid wächst nur dort, wo das Nitrid die Oberfläche freilässt



Der Sauerstoff diffundiert seitlich durch das Padoxid unter die Nitridmaske.

(Quelle: Consiglio Nazionale delle Ricerche)

Neben diesem Effekt, der bis zu 1 µm der Fläche für Bauelemente einnehmen kann, tritt bei einer feuchten Oxidation außerdem der so genannte White-Ribbon- oder Kooi-Effekt auf. Dabei reagiert Nitrid aus der Maskierschicht mit Wasserstoff zu Ammoniak NH_3, der zur Siliciumoberfläche diffundiert und dort zu einer Nitridation führt. Vor der Gateoxidation muss dieses Nitrid entfernt werden, da es sonst als Maskierung wirkt.

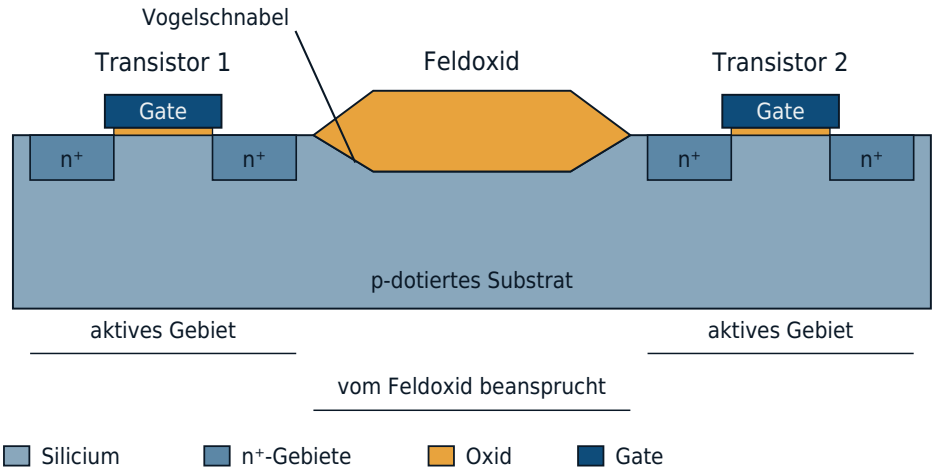
Trotz dieser negativen Effekte ist die LOCOS-Technik ein geeignetes Verfahren um die hohe Packungsdichte zu ermöglichen. Auf Grund der geringeren Unebenheit ohne Kanten- und Stufenbildung wird die Auflösung in der Fototechnik verbessert. Das Feldoxid lässt sich noch etwas zurückätzen, dadurch wird zwar das aufgewachsene Oxid leicht reduziert, die Länge des Vogelschnabels nimmt jedoch ab und die Oberfläche wird wiederum etwas mehr eingeebnet. Dies wird als *fully recessed LOCOS* bezeichnet.

Der Vogelschnabel ist trotzdem der Punkt, an dem das Verfahren scheitert. Seine Länge hängt von der Dicke des Padoxids und des Nitrids ab, nicht von der Größe der Strukturen. Er schrumpft also nicht mit, wenn die Bauelemente kleiner werden. Solange die Isolationsgebiete mehrere Mikrometer breit waren, fiel ein Ausläufer von einigen zehntel Mikrometern kaum ins Gewicht; bei Strukturen unterhalb eines halben Mikrometers verbraucht er den größten Teil der Fläche, die er isolieren soll. Dünneres Padoxid würde ihn verkürzen,

überträgt dann aber die Spannungen der Nitridmaske ungedämpft auf das Silicium und erzeugt Kristallfehler.

LOCOS zur Isolation zweier Transistoren

Das Feldoxid trennt die aktiven Gebiete voneinander



Ohne Feldoxid stünden beide Transistoren über das Substrat in Verbindung. Der Vogelschnabel kostet Fläche, die für Bauelemente ausfällt.

Abscheidung

Plasma, der 4. Aggregatzustand

Plasmazustand

In vielen Prozessen der Halbleiterherstellung wird ein Plasma eingesetzt, sei es beim Sputtern, in Abscheideprozessen oder beim Trockenätzen. Ein wichtiger Punkt dabei ist, dass die Scheibe im Plasma kalt bleibt. Das klingt zunächst widersprüchlich, denn die Elektronen im Plasma sind mit einigen zehntausend Grad außerordentlich heiß. Sie sind jedoch so leicht, dass sie beim Stoß kaum Energie an die schweren Gasteilchen abgeben – das Gas selbst und mit ihm die Scheibe bleiben deshalb nahe der Raumtemperatur. Ein solches Plasma, in dem Elektronen und Gas nicht dieselbe Temperatur haben, heißt Nichtgleichgewichtsplasma. Nur deshalb lassen sich Wafer, die bereits eine Metallisierung tragen, überhaupt im Plasma bearbeiten.

Plasma wird auch als der vierte Aggregatzustand bezeichnet. Ein Aggregatzustand ist ein qualitativer Zustand von Stoffen, der von der Temperatur und vom Druck abhängt. Den drei Aggregatzuständen fest, flüssig und gasförmig begegnet man auch im täglichen Leben. Ist die Temperatur niedrig, so ist jedes Atom in einem Stoff fest an einem Punkt angeordnet. Anziehungskräfte verhindern, dass sie sich bewegen. Am absoluten Nullpunkt (-273,15 °C) gehen Stoffe auch keine Reaktion ein. Bei zunehmender Temperatur fangen die Teilchen an zu schwingen, und die Bindungen der Atome werden instabiler. Ist der Schmelzpunkt erreicht, geht ein Stoff vom ersten in den zweiten Aggregatzustand über: Eis (fest) geht über in Wasser (flüssig).

Die Anziehungskräfte in flüssigen Stoffen sind noch vorhanden, jedoch können sich die Teilchen verschieben und haben keine festen Plätze mehr wie im festen Zustand, die Teilchen passen sich beispielsweise einer vorgegebenen Form an. Wird die Temperatur noch weiter erhöht, brechen die Bindungen komplett auf, die Teilchen bewegen sich unabhängig voneinander. Am Siedepunkt geht ein Stoff vom zweiten in den dritten Aggregatzustand über: Wasser (flüssig) geht über in Wasserdampf (gasförmig).

Während das Volumen bei festen und flüssigen Stoffen konstant ist, nehmen gasförmige Stoffe den vorhandenen Platz vollständig ein, die Teilchen verteilen sich gleichmäßig im gesamten Raum.

Jeder Stoff hat einen ganz bestimmten Schmelz- und Siedepunkt. Silicium schmilzt bei 1414 °C und geht bei 2900 °C in den gasförmigen Zustand über.

Führt man einem Stoff noch mehr Energie zu werden durch Zusammenstöße der Teilchen untereinander Elektronen aus der äußersten Elektronenschale herausgeschlagen. Es befinden sich nun freie Elektronen und positiv geladene Ionen im Raum: der Plasmazustand ist erreicht.

Plasmaerzeugung

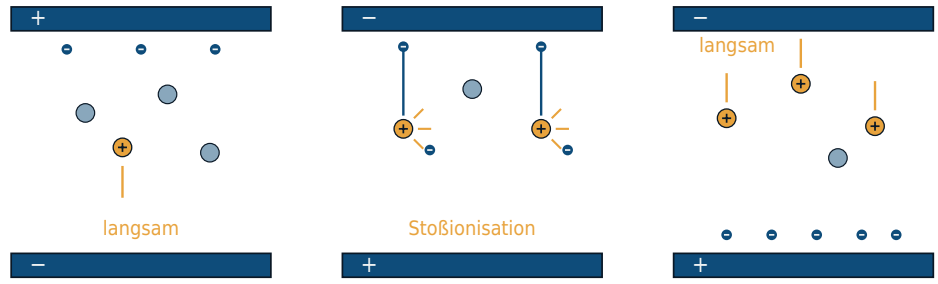
Ein Plasma in der Halbleitertechnologie wird meist durch Hochfrequenzspannung erzeugt, als Gas dient z.B. Argon. Das Gas befindet sich in einem Hochfrequenzfeld zwischen zwei unterschiedlich geladenen Platten (Elektroden) und wird hier ionisiert. Dazu sind Elektronen notwendig, die aus den Argonatomen weitere Elektronen herausschlagen. Diese ersten Elektronen können auf unterschiedliche Weise erzeugt werden:

- Elektronen werden von einer Glühkathode emittiert
- Durch sehr hohe Spannung werden Elektronen aus der negativen Elektrode herausgezogen
- In jedem Gas befinden sich durch Zusammenstöße der Teilchen kurzzeitig freie Elektronen

Da die Elektronen sehr viel leichter als die Ionen sind, werden sie sofort von der positiv geladenen Elektrode angezogen; die schweren Ionen bewegen sich langsam zur negativen Elektrode. Bevor sie diese jedoch erreichen werden die Elektroden umgepolt, die Elektronen werden zur anderen Elektrode beschleunigt und schlagen auf ihrer Flugbahn bei Zusammenstößen weitere Elektronen aus den Atomen heraus. Typische Frequenzen bei der Plasmaerzeugung sind 13,56 Megahertz und 2,45 Gigahertz, die Spannung an den Elektroden wird also 13,56 Millionen bzw. 2,45 Milliarden mal pro Sekunde umgepolt.

Wie das Plasma zündet

Die Elektroden werden millionenfach je Sekunde umgepolt



1. Elektronen wandern
zur positiven Elektrode

2. Umpolung
Elektronen schlagen weitere heraus

3. Mehr Ladungsträger
das Plasma baut sich auf

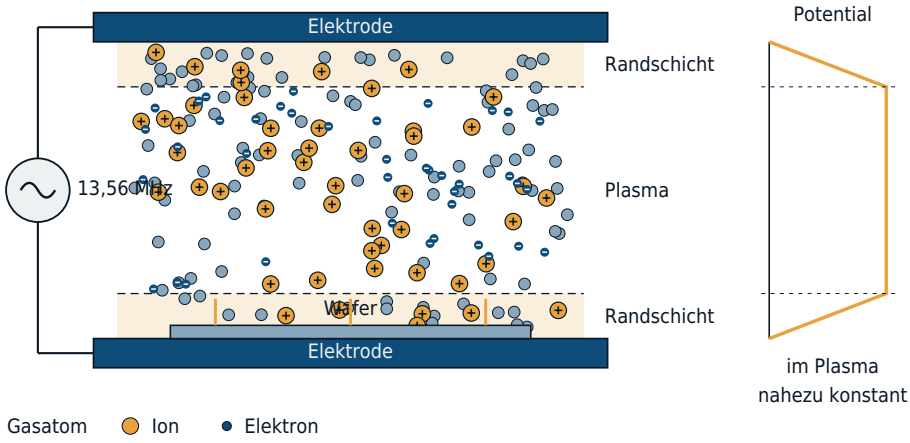
Gasatom Ion Elektron

Die leichten Elektronen folgen dem Wechselfeld, die schweren Ionen kaum.
Mit jedem Umpolen entstehen neue Ladungsträger, bis das Plasma brennt.

Im Inneren des Plasmas halten sich Elektronen und Ionen etwa die Waage, es ist nach außen elektrisch neutral. Anders sieht es unmittelbar vor den Elektroden aus: Dorthin gelangen die leichten, schnellen Elektronen viel häufiger als die trägen Ionen, die dem schnellen Spannungswechsel nicht folgen können. Die Elektrode lädt sich dadurch negativ auf, und vor ihr bildet sich eine elektronenarme Randschicht aus. Über dieser Randschicht fällt fast die gesamte Spannung ab – sie ist es, die die Ionen auf die Scheibe beschleunigt und damit den gerichteten Anteil des Trockenätzens ausmacht.

Plasmaerzeugung zwischen zwei Elektroden

Im Inneren gleichen sich die Ladungen aus, vor den Elektroden nicht



Die Elektronen folgen dem Wechselfeld, die schweren Ionen nicht.
Vor den Elektroden fehlen dadurch Elektronen; über diese Randschicht fällt fast die gesamte Spannung ab und beschleunigt die Ionen auf den Wafer.

Die Plasmaerzeugung findet unter Vakuum statt, das erzeugte Plasma ist dabei nicht erhitzt, was bei vielen Prozessen wichtig ist. Je nach Gas und Anlage in der das Plasma erzeugt wird, kann es in der Abscheidung, beim Sputtern, zum Ätzen oder auch bei der Ionenimplantation verwendet werden. Durch die schnellen Schwingungen der positiven Ionen im Hochfrequenzfeld sind diese sehr energiereich. Es befinden sich dabei nicht nur positive Ionen und freie Elektronen im Plasma, da durch Zusammenstöße auch andere Teilchen erzeugt werden; der Zustand des Plasmas ändert sich laufend. Elektronen werden von den Ionen teilweise wieder eingefangen und erneut herausgeschlagen, diese zusätzlichen Teilchen spielen jedoch bei der weiteren Verwendung des Plasmas keine Rolle. Der Ionisierungsgrad beträgt je nach Teilchendichte im Plasma (10^8–10^{12} Teilchen pro cm^3) 0,001–10 %, die Mehrzahl der Teilchen ist also ungeladen.

Die einfache Anordnung mit zwei Platten hat einen Nachteil: Dieselbe Spannung bestimmt sowohl, wie viele Ionen entstehen, als auch, wie hart sie auf die Scheibe treffen. Beides lässt sich nicht getrennt einstellen. Anlagen für feine Strukturen erzeugen das Plasma deshalb anders – über eine Spule, die von außen ein Wechselfeld einkoppelt, oder über Mikrowellen – und legen die Spannung an der Scheibenelektrode unabhängig davon an. Dichte und Energie der Ionen sind dann zwei getrennte Stellgrößen.

CVD-Verfahren

Silicium-Gasphasenepitaxie

Epitaxie bedeutet "obenauf" oder "zugeordnet", und stellt einen Prozess dar, bei dem eine Schicht auf einer anderen Schicht erzeugt wird und deren Kristallstruktur übernimmt. Ist die abgeschiedene Schicht aus dem gleichen Material wie die Unterlage spricht man von Homoepitaxie, bei zwei unterschiedlichen Materialien von Heteroepitaxie. Der bedeutendste Fall bei der Homoepitaxie ist die Abscheidung von Silicium auf Silicium, bei der Heteroepitaxie wächst ein anderes Material auf, etwa Silicium-Germanium auf Silicium oder Galliumnitrid auf Silicium und Saphir. Voraussetzung ist eine kristalline Unterlage: Auf einer amorphen Schicht wie Siliciumdioxid wächst nichts Einkristallines auf. Silicium-auf-Isolator-Scheiben (Silicon On Insulator, SOI) entstehen deshalb nicht durch Epitaxie, sondern indem zwei Scheiben verbunden und eine davon bis auf eine dünne Schicht abgetragen wird.

Die Homoepitaxie

Je nach Prozess können die Wafer bereits vom Hersteller mit einer Epitaxieschicht geliefert werden (z.B. in der CMOS-Technik), oder der Chiphersteller muss dies selbst durchführen (z.B. in der Bipolartechnik).

Als Gase zur Erzeugung der Schicht benutzt man reinen Wasserstoff in Verbindung mit Silan (SiH_4), Dichlorsilan (SiH_2Cl_2) oder Siliciumtetrachlorid ($SiCl_4$). Bei ca. 1000 °C spalten die Gase das Silicium ab, das sich auf der Waferoberfläche absetzt. Das Silicium übernimmt die Struktur des Substrats und wächst dabei aus energetischen Gründen Ebene für Ebene nacheinander auf. Damit das Silicium nicht polykristallin aufwächst, muss immer ein Mangel an Siliciumatomen vorherrschen, d.h. es steht immer etwas weniger Silicium zur Verfügung, als aufwachsen könnte. Beim Siliciumtetrachlorid verläuft die Reaktion in zwei Schritten:

$SiCl_4 + H_2 \rightarrow SiCl_2 + 2HCl$
$2\ SiCl_2 \rightarrow Si + SiCl_4$

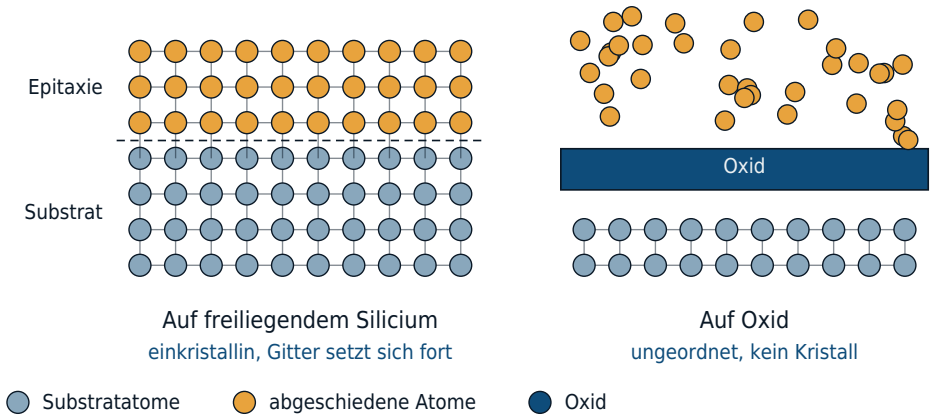
Damit das Silicium auf dem Substrat die Struktur des Kristalls übernehmen kann, muss die Oberfläche absolut rein sein, dabei macht man sich die Gleichgewichtsreaktion zu Nutze. Beide Reaktionen können auch in die andere Richtung verlaufen, je nachdem wie das Verhältnis der eingesetzten Gase ist. Befindet sich nur wenig Wasserstoff in der Atmosphäre wird, wie beim Trichlorsilanprozess zur Reinigung des Siliciums, auf Grund der hohen Chlorkonzentration Silicium von der Waferoberfläche abgetragen. Erst mit zunehmender Konzentration von Wasserstoff wird ein Wachstum erzielt.

Die Aufwachsrate beträgt bei $SiCl_4$ ca. 1–2 µm pro Minute. Da das einkristalline Silicium nur auf der gereinigten Oberfläche aufwächst, kann man mit Oxid bestimmte Bereiche maskieren, auf denen dann polykristallines Silicium aufwächst. Dieses wird jedoch im Vergleich zu einkristallinem Silicium sehr leicht durch die rückwärts ablaufende Reaktion abgeätzt. Werden den Prozessgasen noch Diboran (B_2H_6) oder Phosphin (PH_3) hinzugefügt, kann man dotierte Schichten erzeugen, da sich die Dotiergase bei den hohen Temperaturen zersetzen und die Dotierstoffe im Kristallgitter eingebaut werden.

Der Prozess zur Herstellung von Homoepitaxieschichten wird im Vakuum durchgeführt. Dabei wird die Prozesskammer zunächst auf 1200 °C aufgeheizt, damit sich das natürliche Oxid, das sich immer auf der Siliciumoberfläche bildet, verflüchtigt. Wie oben erwähnt, kommt es bei zu geringer Wasserstoffkonzentration zu einer Rückätzung der Wafer. Dies macht man sich vor dem eigentlichen Prozess zu Nutze, indem man auf diese Weise die Oberfläche der Wafer reinigt. Durch Veränderung der Gaskonzentrationen findet dann die Abscheidung in einer Epitaxieanlage statt.

Epitaxie: die Schicht übernimmt die Kristallstruktur

Sie wächst Ebene für Ebene auf und setzt das Gitter der Unterlage fort



Genau darauf beruht die selektive Epitaxie: Bei passendem Chlorgehalt wird das ungeordnete Material laufend wieder abgetragen, das einkristalline bleibt stehen.

Selektive Epitaxie und verspannte Schichten

Der oben beschriebene Umstand, dass einkristallines Silicium nur auf freiliegendem Silicium aufwächst und polykristallines Material durch die rückwärts laufende Reaktion wieder abgetragen wird, ist heute kein

Nebeneffekt mehr, sondern die Grundlage eines eigenen Verfahrens. Bei passend gewähltem Chlorgehalt wächst die Schicht ausschließlich in den geöffneten Fenstern und gar nicht auf Oxid oder Nitrid – man spricht von selektiver Epitaxie.

Ihre wichtigste Anwendung liegt nicht darin, Silicium zu ergänzen, sondern darin, es zu verspannen. Wird in die Source- und Draingebiete eines Transistors statt Silicium eine Silicium-Germanium-Legierung eingewachsen, so beansprucht diese wegen der größeren Germaniumatome mehr Platz, als das Gitter vorsieht. Sie drückt dadurch auf den dazwischenliegenden Kanal. In einem so gestauchten Gitter bewegen sich Löcher deutlich leichter, der Transistor schaltet schneller. Für den umgekehrten Fall – Zug auf den Kanal, günstig für Elektronen – wird Silicium-Kohlenstoff verwendet. Diese verspannten Schichten gehören seit Mitte der 2000er Jahre zum Standardaufbau eines jeden Hochleistungstransistors.

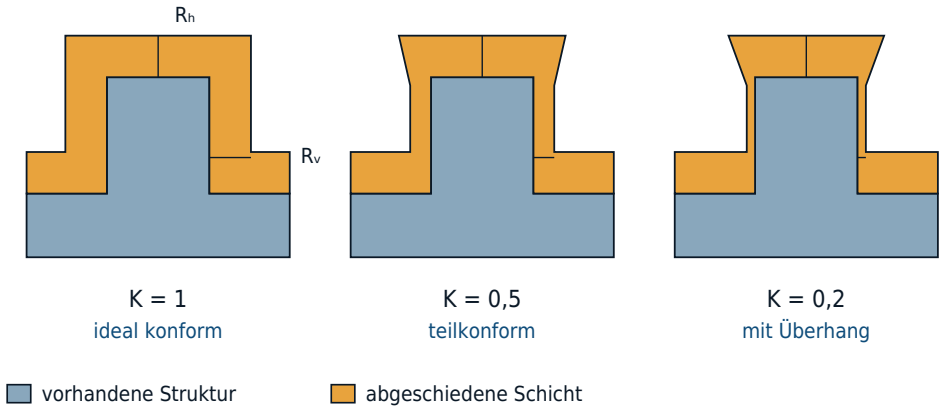
CVD-Verfahren

Oftmals werden in der Halbleitertechnologie Schichten benötigt, die nicht aus dem Siliciumsubstrat erzeugt werden können. Bei Siliciumnitrid und Siliciumoxinitrid muss die Schicht z.B. durch thermische Zersetzung von Gasen erzeugt werden, die alle benötigten Materialien, wie Silicium, enthalten. Darauf basiert die Chemische Dampfabscheidung, kurz CVD (engl.: Chemical Vapor Deposition). Der Wafer dient dabei nur als Grundfläche und reagiert nicht mit den Gasen. Dabei wird das CVD-Verfahren je nach Druck und Temperatur in verschiedene Verfahren unterteilt, deren Schichten sich durch Dichte und Kantenbedeckung unterscheiden. Ist das Schichtwachstum an vertikalen Kanten genau so hoch wie auf horizontalen Flächen ist die Abscheidung konform.

Die Konformität K ist der Quotient aus Schichtwachstum auf vertikalen Flächen R_v und dem Wachstum auf horizontalen Flächen R_h. $K = R_v : R_h$. Ist die Abscheidung nicht ideal konform, ist K deutlich kleiner als 1 (z.B. $R_v : R_h = 1:2 \rightarrow K = 0{,}5$). Hohe Konformitäten lassen sich nur durch hohe Prozesstemperaturen und niedrige Drücke erreichen. Eine Konformität von genau 1 erreicht nur die Atomlagenabscheidung, weil dort nicht die Anlagerung, sondern eine sich selbst begrenzende Reaktion die Schichtdicke bestimmt.

Konformität einer Abscheidung

Verhältnis des Wachstums an senkrechten und waagerechten Flächen



R_h ist das Wachstum auf waagerechten, R_v das an senkrechten Flächen. Je kleiner K, desto eher wächst die Schicht an der Kante zusammen und schließt darunter einen Hohlraum ein.

APCVD: Atmospheric Pressure CVD

APCVD ist eine CVD-Abscheidung bei Normaldruck (atmosphärischer Druck) die zur Herstellung von dotierten und undotierten Oxiden eingesetzt wird. Das erzeugte Oxid hat eine geringe Dichte, die Kantenbedeckung ist auf Grund der niedrigen Temperatur mäßig. Der große Vorteil ist der Durchsatz: Ohne Vakuum lassen sich die Scheiben durchlaufend behandeln, und die Abscheiderate ist hoch. Für dicke, unkritische Oxide ist das Verfahren deshalb weiterhin wirtschaftlich, für dünne oder konforme Schichten kommt es nicht in Frage.

Als Prozessgase dienen Silan SiH_4 (stark verdünnt mit Stickstoff N_2) und Sauerstoff O_2 die bei ca. 400 °C thermisch zersetzt werden und miteinander reagieren.

$SiH_4 + O_2 \rightarrow SiO_2 + 2H_2$ (T = 430 °C, p = 10^5 Pa)

Durch Zugabe von Ozon O_3 kann eine bessere Konformität erzielt werden, da es die Beweglichkeit der sich anlagernden Teilchen erhöht. Das Oxid ist porös und elektrisch instabil und kann durch einen Hochtemperaturschritt verdichtet werden.

Um die Bildung von Kanten, die Probleme bei der Abscheidung weiterer Schichten machen könnten zu verhindern, wird für Schichten, die als Zwischenoxid dienen, so genanntes Phosphorglas (PSG) eingesetzt. Dazu wird den zwei Gasen SiH_4 und O_2 zusätzlich noch Phosphin PH_3 zugesetzt, so dass das

erzeugte Oxid 4–8 % Phosphor enthält. Ein hoher Phosphorgehalt erhöht die Fließeigenschaften deutlich, es kann sich allerdings Phosphorsäure bilden, die Aluminium (Leiterbahnen) korrodieren lässt.

Da Temperschritte aber vorangegangene Prozesse (z.B. Dotierungen) beeinflussen, wird zum Verfließen des PSG nur ein kurzzeitiges Tempern mit starken Argonlampen (mehrere hundert Kilowatt, <10 s, T = 1100 °C) an Stelle von langen Ofenprozessen angewandt.

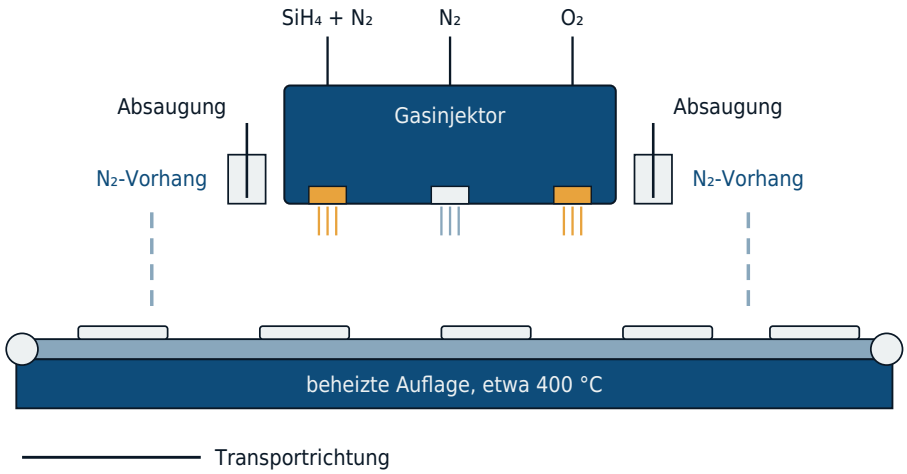
Analog zum Phosphorsilicatglas, kann gleichzeitig auch Bor hinzugefügt werden (Borphosphorsilicatglas, BPSG, 4 % B und 4 % P).

Der Zweck dieser dotierten Gläser war das Verfließen bei hoher Temperatur, mit dem Stufen eingeebnet wurden. Diese Aufgabe hat das chemisch mechanische Polieren übernommen, das ohne Temperaturbelastung auskommt und auch großräumig einebnet. Dotierte Gläser werden heute vor allem noch als Zwischenoxid unterhalb der ersten Metallebene eingesetzt, wo der Phosphor zusätzlich Alkaliionen bindet und sie vom Transistor fernhält.

Darstellung eines Horizontalreaktors zur APCVD-Abscheidung

Horizontalreaktor für die APCVD

Die Scheiben laufen bei Normaldruck unter dem Gasinjektor hindurch



Bei Normaldruck ist keine Schleuse nötig: Die Scheiben laufen fortlaufend hindurch, was den hohen Durchsatz des Verfahrens ausmacht.

Silan und Sauerstoff werden getrennt zugeführt und erst über der Scheibe zusammengebracht, damit sie nicht schon im Injektor reagieren.

LPCVD: Low Pressure CVD

Beim LPCVD-Verfahren findet der Prozess unter Vakuumatmosphäre statt, bei dem sich dünne Schichten wie Siliciumnitrid Si_3N_4, Siliciumoxinitrid SiON, Siliciumdioxid SiO_2 und Wolfram W erzeugen lassen. LPCVD-Verfahren ermöglichen sehr gute Konformitäten von beinahe 1; Grund dafür ist der niedrige Druck von nur 10–100 Pa (normaler Luftdruck ca. 100.000 Pa), durch den die Teilchen keine Bewegung in einer Richtung erfahren, sondern sich selbst durch Zusammenstöße ausbreiten, somit werden die Kanten auf der Waferoberfläche von allen Gasteilchen gleichmäßig bedeckt. Die bis zu 900 °C hohe Prozesstemperatur trägt zur hohen Konformität bei. Im Gegensatz zum APCVD-Verfahren ist die Dichte und Stabilität sehr hoch.

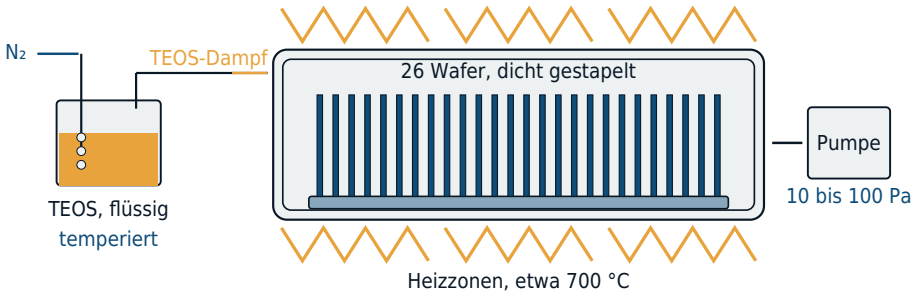
Die Reaktionen für Si_3N_4, SiON, SiO_2 und Wolfram laufen nach den folgenden Reaktionsgleichungen ab:

a) Si_3N_4 (850 °C): $4NH_3 + 3SiH_2Cl_2 \rightarrow Si_3N_4 + 6HCl + 6H_2$
b) SiON (900 °C): $NH_3 + SiH_2Cl_2 + N_2O \rightarrow SiO_xN_y$ + Nebenprodukte
c) SiO_2 (700 °C): $SiO_4C_8H_{20} \rightarrow SiO_2$ + Nebenprodukte
d) Wolfram (400 °C): $WF_6 + 3H_2 \rightarrow W + 6HF$

Im Gegensatz zu den gasförmigen Ausgangsstoffen bei Si_3N_4, SiON und Wolfram, wird zur Erzeugung von SiO_2 flüssiges Tetraethylorthosilicat (TEOS) verwendet; daneben gibt es noch weitere Flüssigquellen wie Ditertiabutylsilan (DTBS, $SiH_2C_8H_{20}$) oder Tetramethylcyclotetrasiloxan ($Si_4O_4C_4H_{16}$).

Eine Wolframschicht lässt sich nur auf Silicium erzeugen. Somit muss, wenn kein Silicium als Grundfläche zur Verfügung steht, Silan hinzugefügt werden.

LPCVD-Anlage für TEOS-Schichten
TEOS ist bei Raumtemperatur flüssig und muss erst verdampft werden



Der niedrige Druck lässt die Gasteilchen weit fliegen und in jede Lücke gelangen. Deshalb dürfen die Scheiben dicht stehen, und die Kanten werden gleichmäßig bedeckt: Die Konformität liegt nahe bei 1.

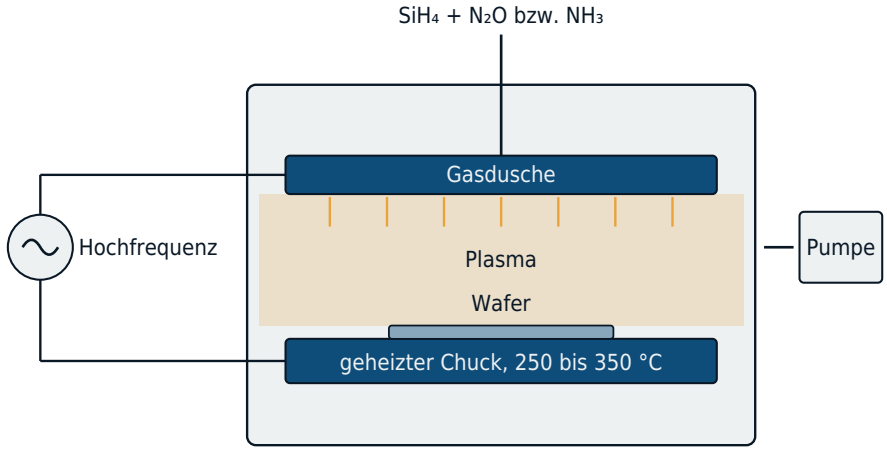
PECVD: Plasma Enhanced CVD

Das Plasma unterstützte CVD-Verfahren findet bei 250–350 °C statt. Da die Temperatur zu niedrig ist um die Gase zu zersetzen, wird es mit Hilfe einer Hochfrequenzspannung in den Plasmazustand versetzt. Im Plasmazustand ist es sehr energiereich und scheidet sich dann auf der Scheibe ab. Besonders Leiterbahnen aus Aluminium können keinen hohen Temperaturen ausgesetzt werden (Aluminium schmilzt bei 660 °C, das thermische Budget fertiger Aluminiumleitbahnen endet jedoch bereits bei etwa 450 °C) weshalb das PECVD-Verfahren vor allem hier zur Abscheidung von SiO_2 und Si_3N_4 genutzt wird. Anstelle von Dichlorsilan SiH_2Cl_2, wie es beim LPCVD-Verfahren verwendet wird, kommt Silan SiH_4 zum Einsatz, da es sich bei den niedrigen Temperaturen leichter zersetzt. Die Konformität ist nicht so ideal wie beim LPCVD-Verfahren (ca. 0,6–0,8), jedoch ist die Abscheiderate mit 500 nm pro Minute deutlich höher.

Aus dieser Kombination – niedrige Temperatur bei hoher Rate – folgt die heutige Bedeutung des Verfahrens. Alles, was nach der ersten Metallebene abgeschieden wird, entsteht per PECVD: die Zwischenoxide zwischen den Verdrahtungsebenen, die Nitridschichten, die als Ätzstopp und als Diffusionssperre gegen Kupfer dienen, die Hartmasken für die Strukturierung und die abschließende Passivierung des fertigen Chips. Auch die kohlenstoffhaltigen Low-k-Dielektrika werden auf diesem Weg erzeugt, indem dem Prozessgas eine siliciumorganische Verbindung zugesetzt wird.

PECVD-Anlage

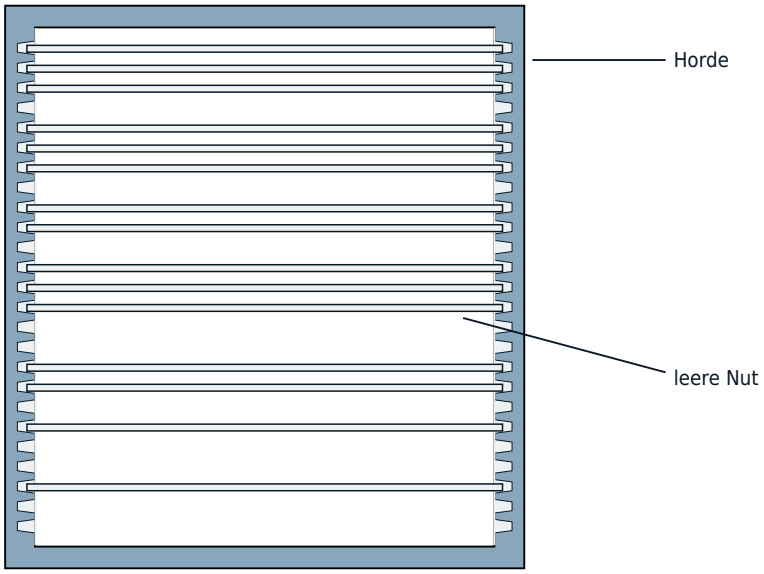
Das Plasma liefert die Energie, die sonst die Temperatur aufbringen müsste



Bei 250 bis 350 °C würden sich die Gase von allein nicht zersetzen. Deshalb lässt sich das Verfahren auch über fertigen Aluminiumleitbahnen einsetzen – dort, wo ein Ofenprozess längst zu heiß wäre.

Horde mit Wafern

Seitenansicht: die Scheiben liegen waagerecht übereinander



Die Horde bietet 25 Plätze; hier sind 15 davon belegt. Die Nut ist höher als die Scheibe dick und weitet sich zur Öffnung hin, damit sie leicht hineinrutscht.

ALD: Atomic Layer deposition

Die Atomlagenabscheidung (Atomic Layer Deposition, kurz ALD) ist ein CVD-Abscheideverfahren zur Erzeugung dünner Schichten. Dabei handelt es sich um ein zyklisches Verfahren, bei dem mehrere Gase abwechselnd in die Prozesskammer eingeleitet werden.

Jedes Gas reagiert dabei in der Art, dass die jeweiligen Oberflächenatome abgesättigt werden, die Reaktion also selbstbegrenzend abläuft. Das jeweils andere Gas kann dann an dieser Oberfläche weiterreagieren. Alternierend zu den reaktiven Gasen wird die Kammer mit Inertgasen, wie Argon oder Stickstoff gespült. Ein einfacher ALD-Zyklus, der meist nur wenige Sekunden beträgt, kann bspw. wie folgt ablaufen:

1. selbst begrenzende Reaktion mit dem ersten Gas, welches die Oberflächenatome absättigt
2. Spülschritt mit einem Inertgas
3. selbst begrenzende Reaktion mit einem zweiten Gas, welches an der neuen Oberfläche reagiert
4. Spülschritt mit einem Inertgas

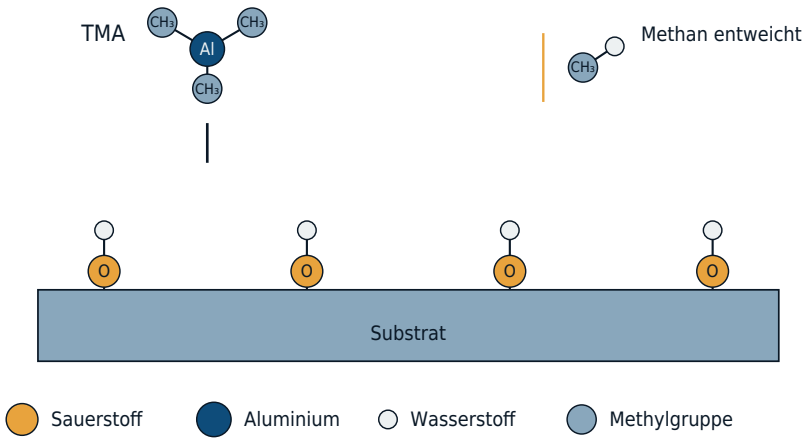
Ein konkretes Beispiel ist die Abscheidung einer Aluminiumoxidschicht Al_2O_3, bei der als reaktive Gase Trimethylaluminium C_3H_9Al (TMA) und Wasser H_2O

verwendet werden.

Dabei entfernt eine Methylgruppe CH_3 der Aluminiumverbindung im ersten Schritt oberflächennahe Wasserstoffatome von OH-Gruppen unter Bildung von Methan CH_4. Die verbleibenden Moleküle werden an den nun ungesättigten Sauerstoffatomen gebunden.

1. Trimethylaluminium trifft auf die Oberfläche

Die Methylgruppen nehmen den Wasserstoff der OH-Gruppen mit

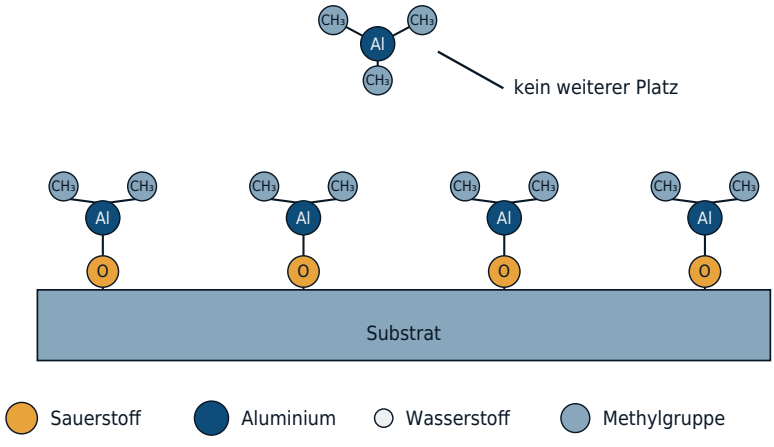


Jede Methylgruppe bindet ein Wasserstoffatom und verlässt die Oberfläche als Methan.

Sind diese Atome abgesättigt, kann keine weitere Anlagerung von TMA erfolgen.

2. Die Oberfläche ist abgesättigt

Weiteres TMA findet keinen Platz mehr – die Reaktion begrenzt sich selbst



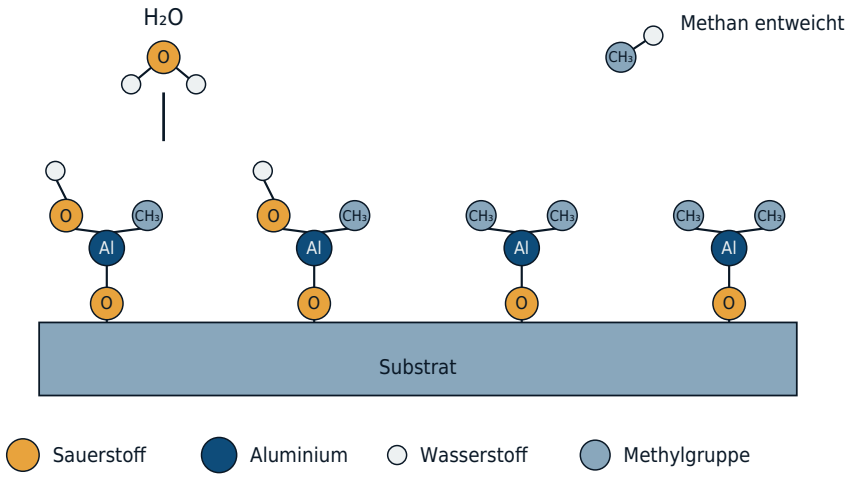
Genau darauf beruht die Genauigkeit des Verfahrens: Die Schicht wächst nur eine Lage.

Die Prozesskammer wird gespült und anschließend Wasserdampf in die Kammer eingeleitet. Durch je ein Wasserstoffatom der H_2O-Moleküle werden die CH_3-

Gruppen der zuvor abgeschiedenen Schicht unter Bildung von Methan entfernt.

3. Nach dem Spülen folgt Wasserdampf

Das Wasser ersetzt die Methylgruppen, wieder entsteht Methan

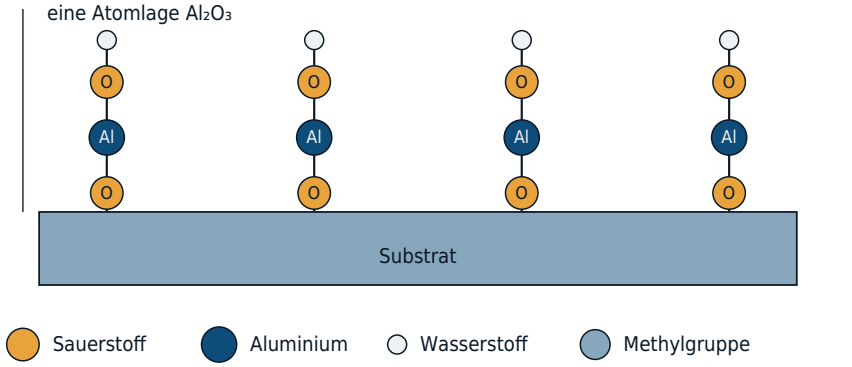


Ein Wasserstoff des Wassers bindet die Methylgruppe, der Sauerstoff bleibt zurück.

Zurück bleiben OH^--Ionen, welche sich nun mit dem Aluminium verbinden, so dass wiederum oberflächennahe Wasserstoffatome zur Verfügung stehen, die mit TMA reagieren können.

4. Eine Lage Aluminiumoxid ist gewachsen

Die neuen OH-Gruppen sind der Ausgangspunkt des nächsten Zyklus



Pro Zyklus wächst etwa ein Zehntel Nanometer – die Dicke folgt aus der Zyklenzahl.

Die Atomlagenabscheidung bietet wesentliche Vorteile gegenüber anderen Abscheideverfahren, weshalb sie ein wichtiger Prozess zur Aufbringung verschiedener Schichten ist. Dabei können auch 3-dimensionale Strukturen (z.B. Gräben) sehr homogen beschichtet werden. Möglich sind sowohl isolierende als

auch leitende Filme auf verschiedenen Substraten (Halbleiter, Polymere u.a.). Durch die Anzahl der Zyklen ist eine exakte Schichtdicke leicht zu realiseren. Da die reaktiven Gase nicht gleichzeitig in die Kammer geleitet werden, kann es nicht schon vor der eigentlichen Abscheidung zu Keimbildung kommen. Die Qualität der Schicht ist damit sehr hoch.

Wofür ALD heute eingesetzt wird

Der Preis für diese Genauigkeit ist die Geschwindigkeit: Pro Zyklus wächst nur etwa ein Zehntel Nanometer auf, und ein Zyklus dauert Sekunden. Für dicke Schichten ist das Verfahren daher ungeeignet. Überall dort, wo es auf wenige Atomlagen ankommt, ist es jedoch konkurrenzlos:

- Gatedielektrikum: Das Siliciumdioxid unter dem Gate ließ sich nicht weiter verdünnen, ohne dass Strom hindurchtunnelt. Ersetzt wurde es durch Hafniumdioxid, das bei gleicher Wirkung dicker sein darf. Abgeschieden wird es per ALD – anders lässt sich eine Schicht von wenigen Atomlagen über die ganze Scheibe hinweg nicht gleichmäßig genug herstellen.
- Abstandsschichten beim Multipatterning: Beim selbstjustierenden Verfahren bestimmt die Dicke einer abgeschiedenen Schicht die spätere Strukturbreite. Sie muss deshalb genauer einstellbar sein, als eine Belichtung es je wäre.
- Barrieren und Keimschichten: in tiefen, engen Kontaktlöchern, die kein anderes Verfahren gleichmäßig auskleidet.
- Speicherkondensatoren: Die Kondensatoren einer DRAM-Zelle sind schmale tiefe Gräben, deren Wände vollständig beschichtet werden müssen.

Muss die Temperatur niedrig bleiben, wird der zweite Reaktionsschritt durch ein Plasma unterstützt (plasma enhanced ALD). Die Radikale aus dem Plasma reagieren bereitwilliger als das neutrale Gas, so dass der Zyklus auch weit unterhalb der sonst nötigen Temperatur abläuft.

Spaltfüllung

Eine Aufgabe stellt sich bei jedem Abscheideverfahren, und keines der bisher beschriebenen löst sie von allein: das Auffüllen enger, tiefer Zwischenräume. Solche Spalte entstehen überall dort, wo zwischen bereits stehenden Strukturen ein Isolator eingebracht werden muss – zwischen den Gräben der Grabenisolation, zwischen dicht benachbarten Leiterbahnen, zwischen den Stegen eines Transistors.

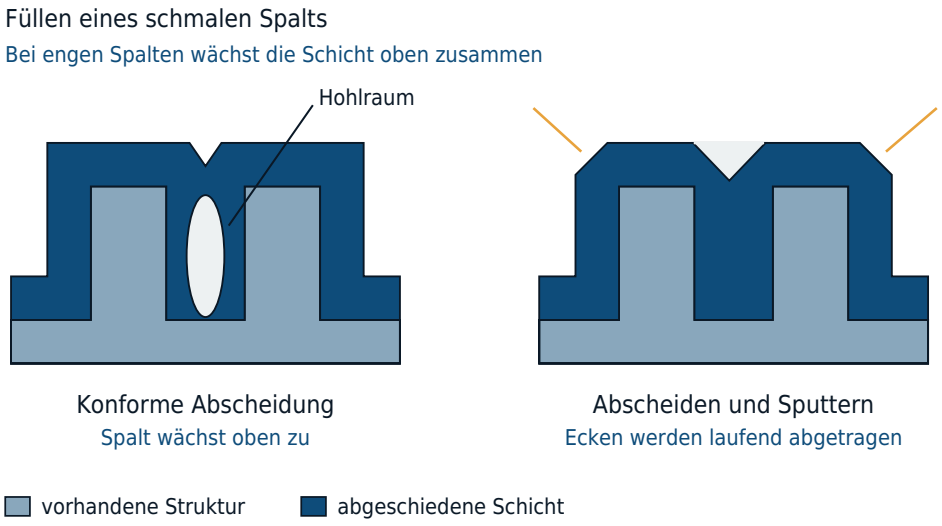
Warum ein Hohlraum entsteht

Eine konforme Schicht wächst an allen Flächen gleich schnell auf, also auch an den beiden gegenüberliegenden Spaltwänden. Ist der Spalt schmaler als die doppelte Schichtdicke, treffen sich die beiden Fronten. Das allein wäre

unproblematisch, wenn sie sich überall gleichzeitig träfen. Tatsächlich wächst die Schicht an der Spaltöffnung schneller, weil die Öffnung von mehr Richtungen aus erreichbar ist als der Grund. Der Spalt schließt sich deshalb zuerst oben, und darunter bleibt ein Hohlraum zurück.

Ein solcher Hohlraum ist kein kosmetisches Problem. Wird er bei einem späteren Schritt angeschnitten, etwa beim Polieren oder beim Ätzen eines Kontaktlochs, füllt sich der Hohlraum mit Ätzlösung oder mit Metall – im ungünstigen Fall entsteht ein Kurzschluss zwischen zwei Leitbahnen.

Abscheiden und Sputtern zugleich



Die Lösung besteht darin, die Ecken an der Spaltöffnung laufend wieder abzutragen, während die Schicht wächst. Dazu wird der Abscheidung ein Sputteranteil überlagert: Argonionen werden aus einem dichten Plasma auf die Scheibe beschleunigt und schlagen dort Material heraus. Es ist derselbe Vorgang, der beim Sputtern zum Beschichten genutzt wird – nur trifft der Ionenstrahl hier nicht auf ein Target, sondern auf die wachsende Schicht selbst.

Weil das Sputtern unter schrägem Einfall am wirksamsten ist, greift es die Kanten an der Spaltöffnung viel stärker an als die waagerechte Fläche am Grund. An der Oberkante entsteht dadurch eine schräge Fläche, die als Facette bezeichnet wird. Die Öffnung bleibt offen, und der Spalt füllt sich von unten nach oben.

Verfahren dieser Art laufen unter der Bezeichnung HDP-CVD (high density plasma CVD). Der entscheidende Prozessparameter ist das Verhältnis von Abscheide- zu Sputterrate, im Englischen als deposition/sputter ratio oder kurz D/S bezeichnet: Zu wenig Sputtern lässt den Hohlraum entstehen, zu viel trägt

die Strukturen selbst ab. Der Sputteranteil liegt typischerweise bei 10 bis 20 Prozent der Nettoabscheiderate.

Wenn auch das nicht mehr reicht

Mit weiter steigenden Tiefen-Breiten-Verhältnissen stößt auch dieses Verfahren an eine Grenze. Der nächste Schritt kehrt zum ältesten Prinzip zurück: Eine Flüssigkeit fließt von selbst in jeden Spalt, ohne Rücksicht auf dessen Form. Aufgebracht wird deshalb ein siliciumhaltiges Vorprodukt, das zunächst flüssig ist, den Spalt vollständig ausfüllt und erst anschließend durch eine Behandlung mit Wasserdampf und Temperatur in festes Siliciumdioxid überführt wird. Die entstehende Schicht ist weniger dicht als ein abgeschiedenes Oxid, füllt dafür aber Spalte, die für jedes gerichtete Verfahren unzugänglich sind.

PVD-Verfahren

MBE: Molekularstrahlepitaxie

Das Aufwachsverhalten bei der Molekularstrahlepitaxie (Molecular beam epitaxy, MBE) entspricht dem der Siliciumgasphasenepitaxie. Jedoch wird in diesem Fall das Silicium auf eine andere Weise auf dem Wafer abgeschieden.

Der Prozess findet unter Ultrahochvakuum statt (UHV, 10^{-8} Pa), der zu prozessierende Wafer wird mit der Oberseite nach unten an der Decke der Prozesskammer gehalten, und auf ca. 600–800 °C erhitzt damit sich das natürliche Oxid verflüchtigt.

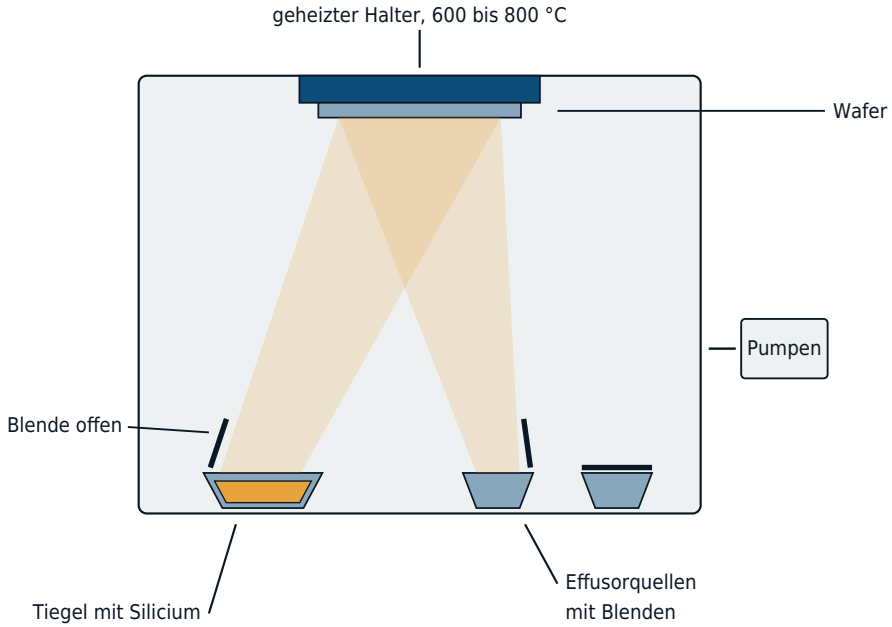
Silicium wird dabei mit einem Elektronenstrahl in einem Tiegel verdampft, und scheidet sich dann auf dem Wafer ab. Dotierstoffe können in einer Effusorquelle verdampft werden und gelangen so mit dem Siliciumdampf auf den Wafer. Durch gezielte Temperaturregelung und mittels Blenden kann der Teilchenstrahl genau dosiert werden. Bei diesem Verfahren ist es außerdem möglich Schichten aus verschiedenen Stoffen übereinander wachsen zu lassen, die sonst auf Grund der unterschiedlichen Atomgrößen nicht möglich wären. So lassen sich Silicium und Germanium als Schicht erzeugen, die bei Hochfrequenzanwendungen in der Bipolartechnik von Nöten ist.

Genau darin liegt die Stärke des Verfahrens: Weil die Atome einzeln und ohne chemische Reaktion ankommen, lässt sich die Zusammensetzung von Atomlage zu Atomlage ändern. Für Schichtfolgen aus Verbindungshalbleitern, wie sie in Laserdioden und Hochfrequenzbauelementen gebraucht werden, ist das unverzichtbar. In der Siliciumfertigung spielt die MBE dagegen keine Rolle – dort ist sie zu langsam.

Jedoch ist der Aufwand bei dieser Epitaxie zur Erzeugung des UHV sehr hoch. Um diesen Druck, der weniger als dem billionsten Teil des normalen Luftdrucks entspricht, zu erreichen, müssen mehrere Pumpen kombiniert werden und die Prozesskammer lange abgepumpt werden. Dabei kann nur ein Wafer auf einmal prozessiert werden, das Wachstum beträgt nur ca. 1 µm pro Stunde.

MBE-Anlage

Die Atome fliegen einzeln durch das Ultrahochvakuum auf den Wafer



Jede Quelle lässt sich einzeln zuschalten; das Wachstum liegt bei etwa einem Mikrometer je Stunde. Nur so lässt sich die Zusammensetzung von Atomlage zu Atomlage ändern. Der Druck liegt bei 10^{-8} Pa.

Aufdampfen

Beim Aufdampfen können metallische Schichten, wie z.B. Aluminium, auf den Wafern erzeugt werden. Das Material befindet sich in einem Tiegel aus hochschmelzbarem Metall wie Tantal und wird dort erhitzt bis es in den gasförmigen Zustand übergeht. Der Metalldampf trifft senkrecht auf die Scheibe, so dass Kanten nicht gut bedeckt werden, die entstandene Schicht ist polykristallin. Alternativ zum Schmelzen im Tiegel kann das Metall auch mit einem Elektronenstrahl verdampft werden. Mit dieser Elektronenstrahlmethode ist es möglich eine, im Vergleich zur thermischen Aufdampfung, sehr genaue Abscheiderate zu erzielen. Da die Kantenabdeckung nicht sehr gut ist, werden diese zwei Verfahren meist zur ganzflächigen Rückseitenbeschichtung für die spätere elektrische Kontaktierung verwendet.

In der Fertigung integrierter Schaltungen ist das Aufdampfen weitgehend durch

das Sputtern verdrängt worden, das dichtere Schichten und eine bessere Kantenbedeckung liefert. Erhalten hat es sich dort, wo gerade die schlechte Kantenbedeckung erwünscht ist: Beim Abhebeverfahren wird zuerst eine Lackmaske strukturiert, dann das Metall aufgedampft und anschließend der Lack samt dem darauf liegenden Metall abgelöst. Weil die Flanken der Lackmaske kaum bedeckt werden, reißt der Film an der Kante sauber ab. So lassen sich Metalle strukturieren, die sich schlecht ätzen lassen.

In der Fertigung integrierter Schaltungen ist das Aufdampfen weitgehend durch das Sputtern verdrängt worden, das dichtere Schichten und eine bessere Kantenbedeckung liefert. Erhalten hat es sich dort, wo gerade die schlechte Kantenbedeckung erwünscht ist: Beim Abhebeverfahren wird zuerst eine Lackmaske strukturiert, dann das Metall aufgedampft und anschließend der Lack samt dem darauf liegenden Metall abgelöst. Weil die Flanken der Lackmaske kaum bedeckt werden, reißt der Film an der Kante sauber ab. So lassen sich Metalle strukturieren, die sich schlecht ätzen lassen.

In der Fertigung integrierter Schaltungen ist das Aufdampfen weitgehend durch das Sputtern verdrängt worden, das dichtere Schichten und eine bessere Kantenbedeckung liefert. Erhalten hat es sich dort, wo gerade die schlechte Kantenbedeckung erwünscht ist: Beim Abhebeverfahren wird zuerst eine Lackmaske strukturiert, dann das Metall aufgedampft und anschließend der Lack samt dem darauf liegenden Metall abgelöst. Weil die Flanken der Lackmaske kaum bedeckt werden, reißt der Film an der Kante sauber ab. So lassen sich Metalle strukturieren, die sich schlecht ätzen lassen.

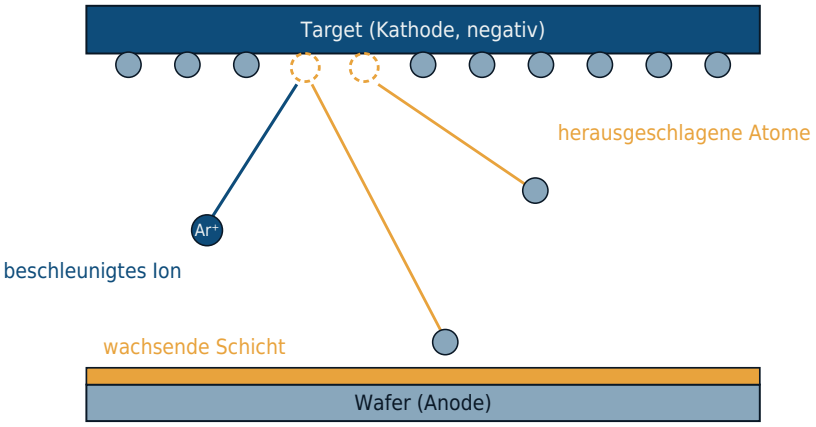
Schematische Darstellung einer Aufdampfanlage

Sputtern: Kathodenzerstäubung

Bei der Kathodenzerstäubung werden Ionen (meist Argon) auf ein Target beschleunigt und schlagen dort Atome oder Moleküle heraus, das Target besteht dabei aus dem Material der aufzubringenden Schicht. Die mittlere Weglänge dieser Teilchen ist wenige Millimeter lang, d.h. sie stoßen oft zusammen, dadurch werden auf der Scheibe auch vertikale Flächen gut bedeckt. Die abgeschiedenen Teilchen bilden eine poröse Schicht, die durch Temperung verdichtet werden kann. Die Kathodenzerstäubung lässt sich in das passive (inerte) und das reaktive Sputtern unterteilen.

Kathodenzerstäubung

Argonionen schlagen Atome aus dem Target, die sich auf dem Wafer niederschlagen



Die Atome fliegen in alle Richtungen und stoßen unterwegs mit Gasteilchen zusammen. Deshalb werden auch senkrechte Flächen gut bedeckt – die Schicht bleibt dabei porös und wird anschließend durch Tempern verdichtet.

Beim passiven Sputtern wird nur das Material des Targets auf den Scheiben abgeschieden, entsprechend dem Material des Targets lassen sich hochreine Schichten erzeugen, da ein genaues Mischungsverhältnis der Stoffe im Target möglich ist. Beim reaktiven Sputtern wird dem Gas in der Prozesskammer ein Reaktionsgas (z.B. Sauerstoff O_2) hinzugefügt, dass sich mit dem zerstäubten Material des Targets verbindet und sich dann auf dem Wafer absetzt. Dadurch lassen sich aus einem Metalltarget (z.B. Aluminium Al) auch Isolierungsschichten herstellen:

$4\ Al + 3\ O_2 \rightarrow 2\ Al_2O_3$

Zur Erzeugung von metallischen Schichten wird das Gleichstrom-Sputtern (DC-Sputtern) verwendet. Dabei werden die Ionen mit bis zu 3 Kilovolt (kV) auf das Target beschleunigt und entladen sich dort. Da diese Ladungen immer abgeführt werden müssen, kann hier nur ein leitendes Material als Target dienen. Für isolierende Schichten muss das reaktive Sputtern eingesetzt werden. Will man eine isolierende Schicht direkt aus dem Target erzeugen, verwendet man das Hochfrequenzsputtern (RF-Sputtern).

Beim RF-Sputtern liegt die Spannung an je einer Elektrode hinter dem Target (Kathode) und dem Wafer (Anode) an. Durch die hochfrequente Spannung werden die Elektronen bei der positiven Halbwelle am Target von diesem angezogen, wodurch sich das Target negativ auflädt. Durch dieses negative Target werden die Ionen angezogen und schlagen dort Teilchen heraus. Beim Magnetronsputtern befinden sich zusätzlich Magnete hinter dem Target, so dass Elektronen auf Kreisbahnen abgelenkt werden und so häufiger Argonatome

ionisieren können, was einen erheblichen Anstieg der Abscheiderate verursacht. Da die Anode mit der Prozesskammer verbunden ist, hat sie auf die Oberfläche bezogen im zeitlichen Mittel eine wesentlich geringere Potentialdifferenz gegenüber dem Plasma als die Kathode, weshalb die Ionen nur zum Target und nicht zum Wafer wandern.

Um die Kantenabdeckung zu erhöhen verwendet man das BIAS-Sputtern, bei dem am Substrat eine negative Spannung angelegt wird. Dadurch werden auch hier, wie am Target, Teilchen der erzeugten Schicht abgetragen und die Oberfläche eingeebnet. Jedoch muss darauf geachtet werden, dass kein Abtrag des Substrats erfolgt. Dieses so genannte Rückätzen ist auch das Prinzip der meisten Plasmaätzanlagen.

Sputtern in enge Strukturen

Die herausgeschlagenen Teilchen sind ungeladen und fliegen in alle Richtungen. Über einem engen, tiefen Kontaktloch bedeutet das: Der größte Teil trifft den Rand der Öffnung, nur wenig gelangt bis auf den Grund. Die Öffnung wächst zu, bevor der Boden bedeckt ist.

Abhilfe schafft es, die abgestäubten Teilchen selbst zu ionisieren. In einem dichten Plasma zwischen Target und Scheibe verlieren sie ein Elektron und lassen sich anschließend durch eine Spannung an der Scheibe senkrecht nach unten beschleunigen. Der Materialstrom wird dadurch gerichtet und erreicht auch den Grund tiefer Löcher. Auf diese Weise werden die Tantalbarriere und die Kupferkeimschicht im Damascene-Verfahren abgeschieden – die Keimschicht muss lückenlos und an den Seitenwänden ebenso vorhanden sein wie am Boden, sonst wächst das Kupfer bei der anschließenden Galvanik nicht durchgehend auf.

Das Sputtern eignet sich somit zur Erzeugung von metallischen Schichten mit guter Konformität und sehr guter Reproduzierbarkeit. Der Aufwand ist gering, der Unterdruck mit 5 Pa recht einfach zu erzeugen.

Metallisierung

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.

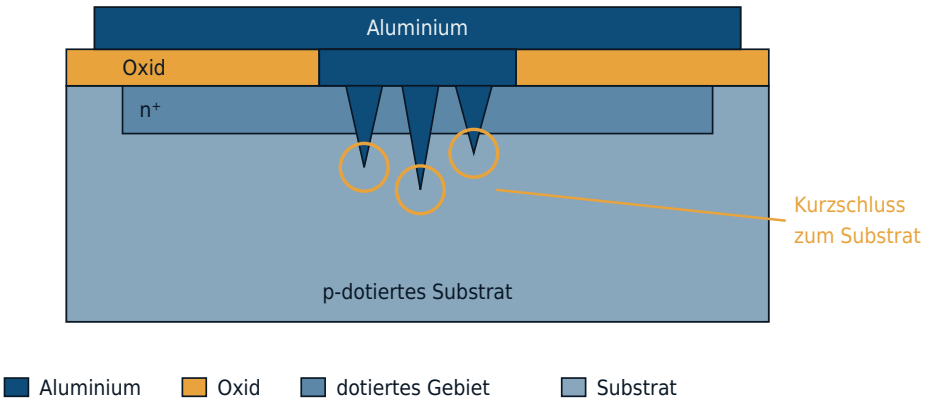
Aluminium erfüllt allerdings die Anforderung an elektrische Belastbarkeit und Korrosionsbeständigkeit nur teilweise. Die folgenden Abschnitte beschreiben die drei Effekte, an denen die Aluminiumtechnik ihre Grenze findet: die Diffusion von Silicium in das Metall, die Elektromigration und das Wachstum 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

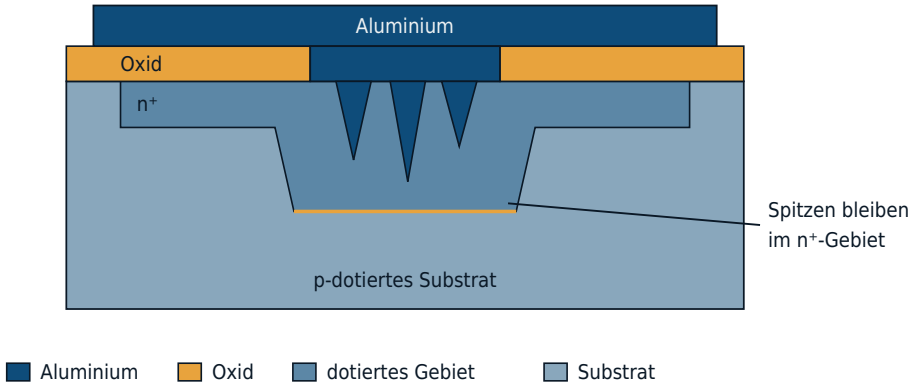


Beim Sintern löst sich Silicium im Aluminium und hinterlässt Gruben, die das Metall auffüllt. Reichen die Spitzen tiefer als das dotierte Gebiet, ist der Übergang kurzgeschlossen.

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



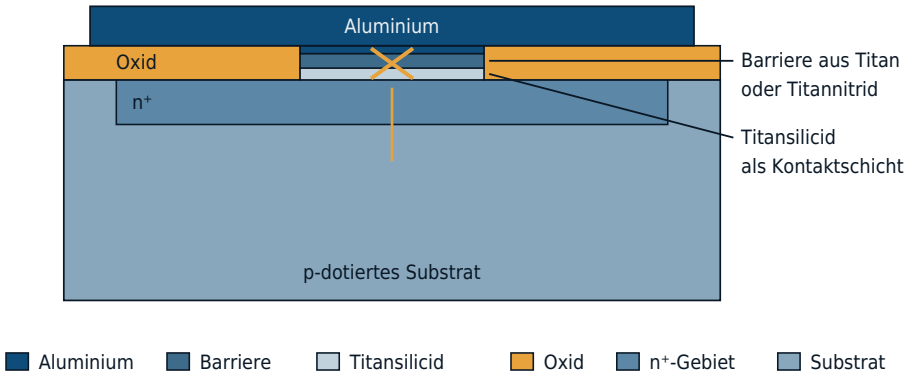
Die tiefe Implantation unter dem Kontaktloch verlegt den Übergang nach unten. Sie kostet einen zusätzlichen Prozessschritt und ändert die elektrischen Eigenschaften des Gebiets.

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 Stellen mit geringem Leiterbahnquerschnitt ist die Stromdichte erhöht, durch die

Verschiebung der Atome nimmt der Querschnitt ab, die Stromdichte steigt weiter an. Insbesondere an Kanten über die die Leiterbahnen verlaufen sind solche Engstellen zu finden. Im Extremfall reißen die Aluminiumleiterbahnen durch den Materialtransport ab.

Schäden auf Leiterbahnen durch Elektromigration

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

Die Elektromigration begrenzt, 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. Kupfer verträgt etwa die hundertfache Lebensdauer von Aluminium bei gleicher Belastung, weil seine Atome fester gebunden sind. Der Materialtransport verlagert sich dabei an die Grenzflächen: Kupferatome wandern bevorzugt an der Oberseite der Leitbahn entlang, dort, wo sie an die Deckschicht grenzt. Eine dünne Cobaltschicht auf der Leitbahn statt eines rein dielektrischen Abschlusses bindet die Oberfläche und verlängert die Lebensdauer deutlich.

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.

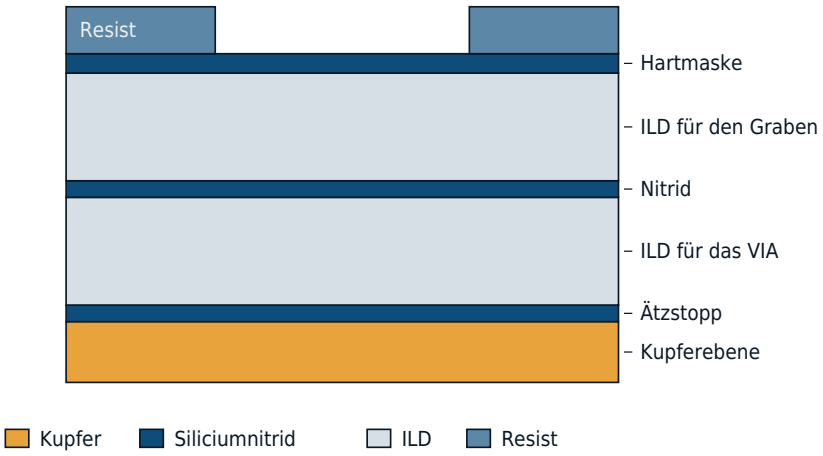
Trench First VIA Last

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

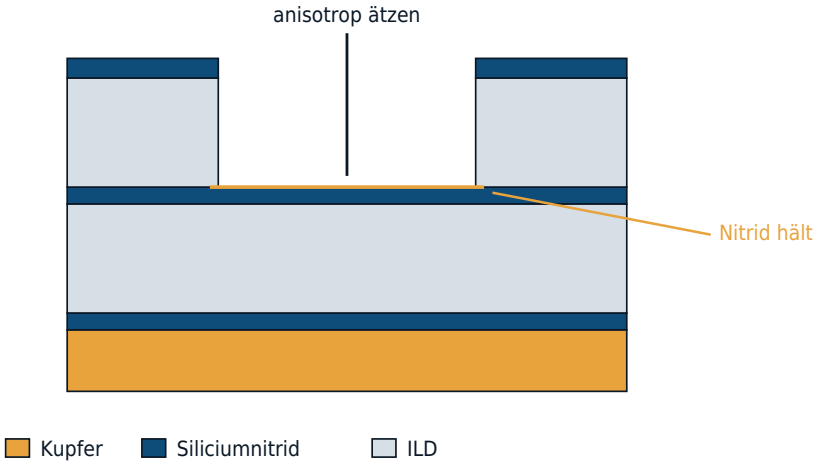


Es liegen zwei etwa gleich dicke ILD-Ebenen übereinander, getrennt durch eine Nitridschicht: In der unteren sitzt später das VIA, in der oberen der Graben. Die Hartmaske schützt das ILD beim Lackentfernen und dient als CMP-Stopp.

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.

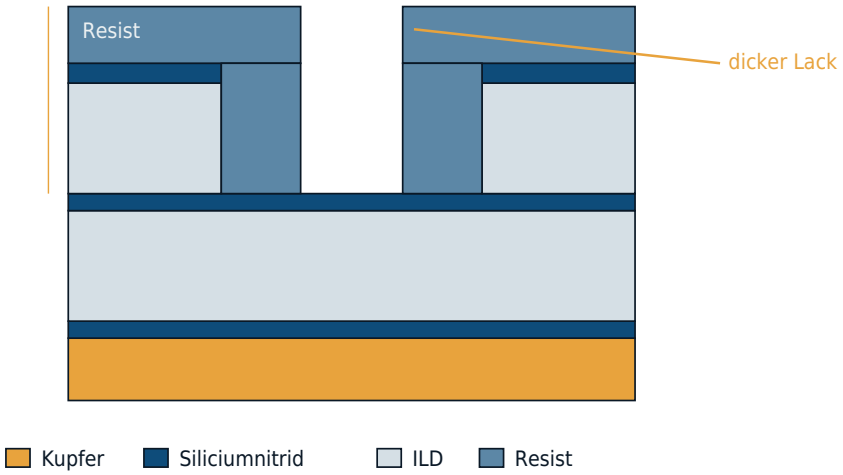
Die Hartmaske an der Oberfläche wird benötigt, um die ILD-Schicht während des Lackentfernens vor dem Plasma zu schützen. Dies ist notwendig, da das ILD chemisch ähnlich aufgebaut ist wie der Fotolack, und so durch die

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

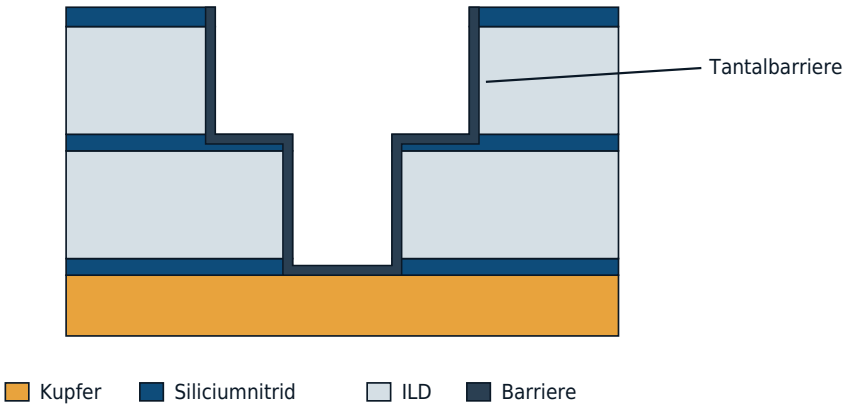


Weil der Lack den Graben mit auffüllt, ist er hier deutlich dicker als beim ersten Schritt. Ein kleines Kontaktloch darin zu belichten, ist der eigentliche Nachteil dieses Verfahrens.

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

4. VIA ätzen und Barriere abscheiden

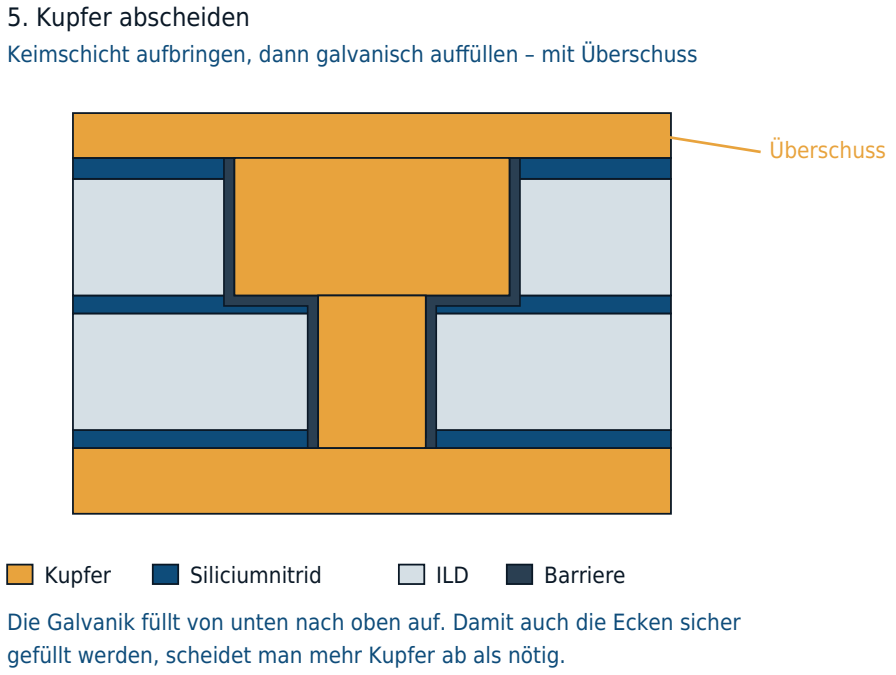
Durch Nitrid, unteres ILD und Ätzstopp bis auf das Kupfer



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.

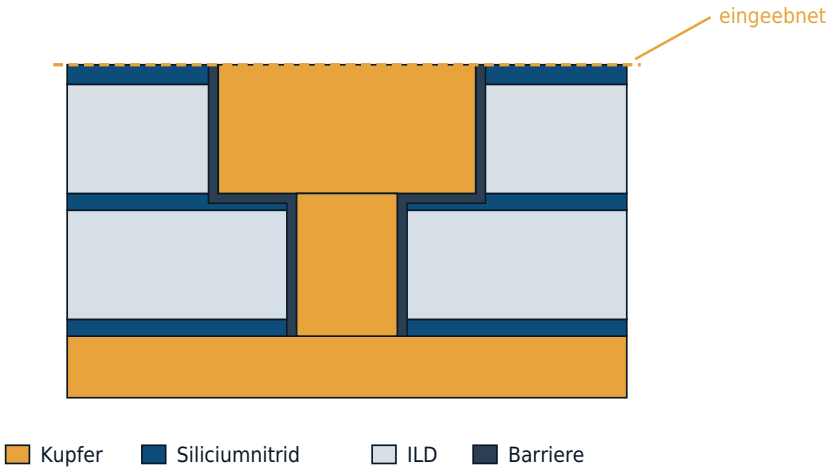
5. Eine dünne Kupferkeimschicht wird abgeschieden und die Strukturen in einem galvanischen Verfahren mit Kupfer aufgefüllt.



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

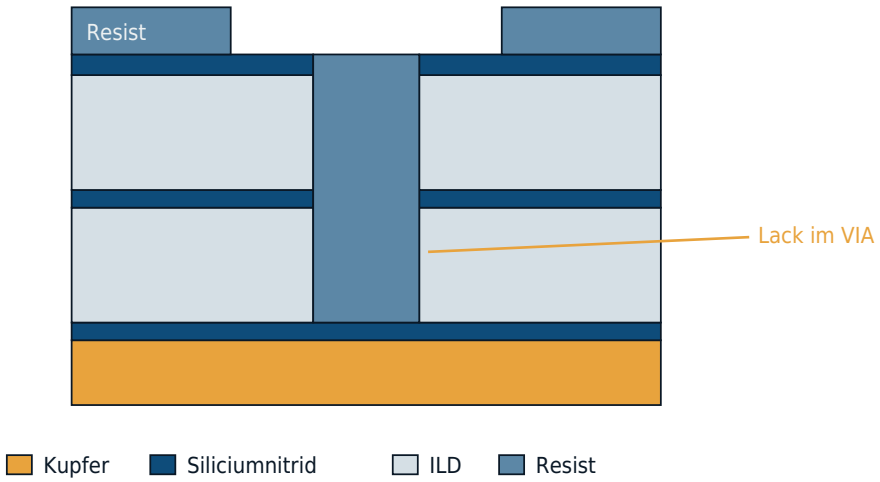
Eine schmale Lackmaske, das Loch reicht durch beide ILD-Ebenen

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

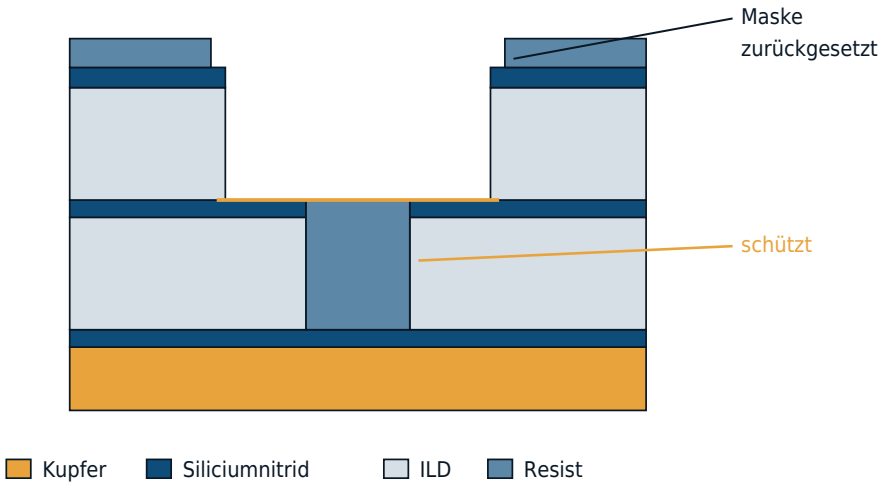


Der Lackpfropfen im Loch ist der Kern des Verfahrens: Ohne ihn würde die folgende Ätzung das VIA weiter vertiefen und den Ätzstopp durchschlagen.

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.

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.

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

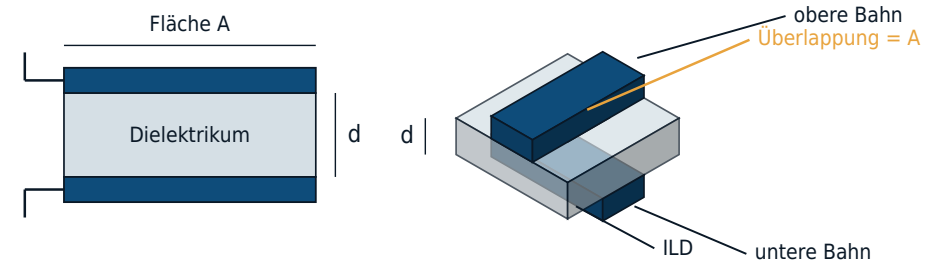
Durch die stetige Verkleinerung von Strukturen auf Mikrochips, um einerseits die Leistungsaufnahme zu verringern und andererseits maximale Schaltgeschwindigkeiten zu erzielen, rücken die Leiterbahnen zur Verdrahtung der einzelnen Bauelemente sowohl in vertikaler als auch lateraler Richtung immer näher 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

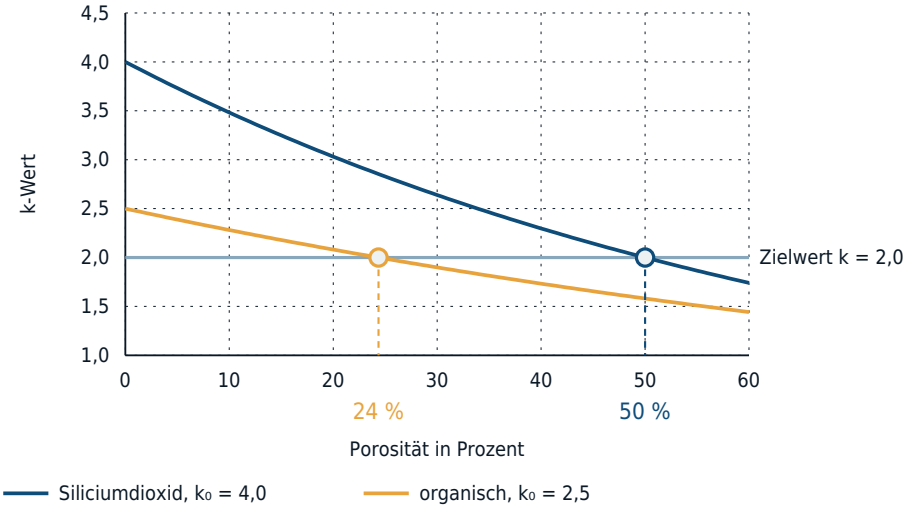
Die Polarisierbarkeit kann durch Stoffe mit wenig polaren Gruppen gesenkt werden, möglich sind fluorierte (FSG, ε_r ca. 3,6) oder organische (OSG) Siliciumoxide. Dies allein ist bei den immer kleiner werdenden Strukturgrößen jedoch nicht mehr ausreichend, weshalb der Trend zu porösen Schichten geht. 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.

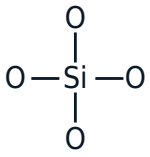
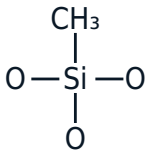
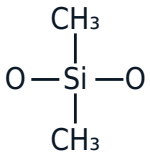
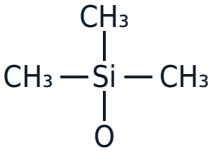
Wenn kein Material mehr hilft: Luftspalte

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.

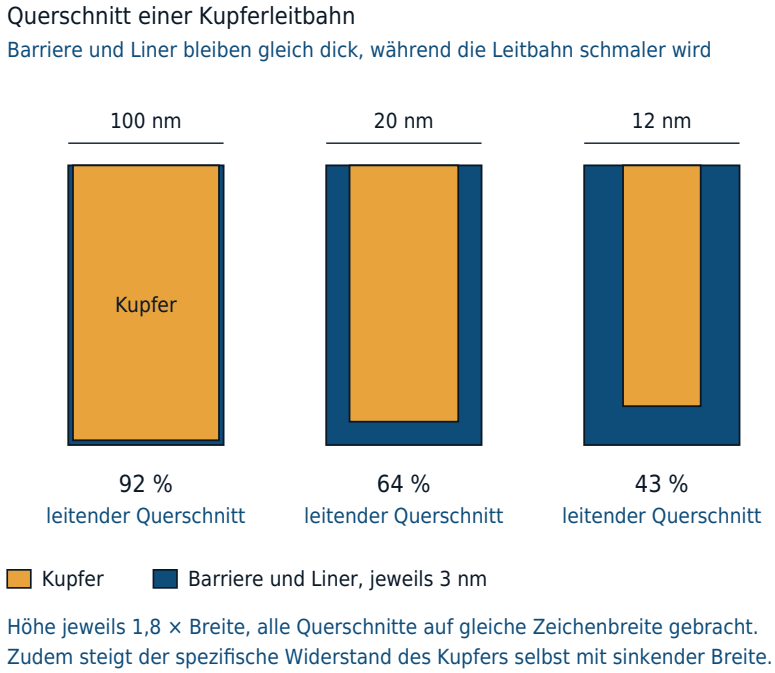
Der spezifische Widerstand ist keine Konstante mehr

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

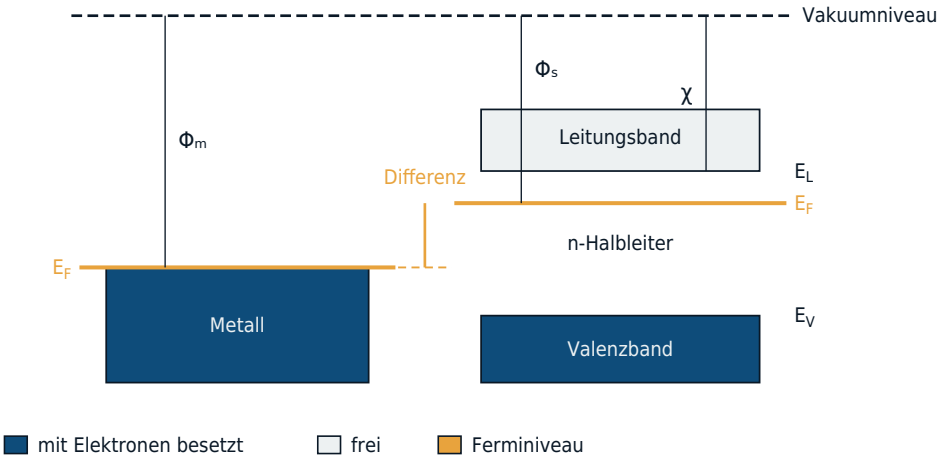
n-Halbleiter-Kontakt

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.

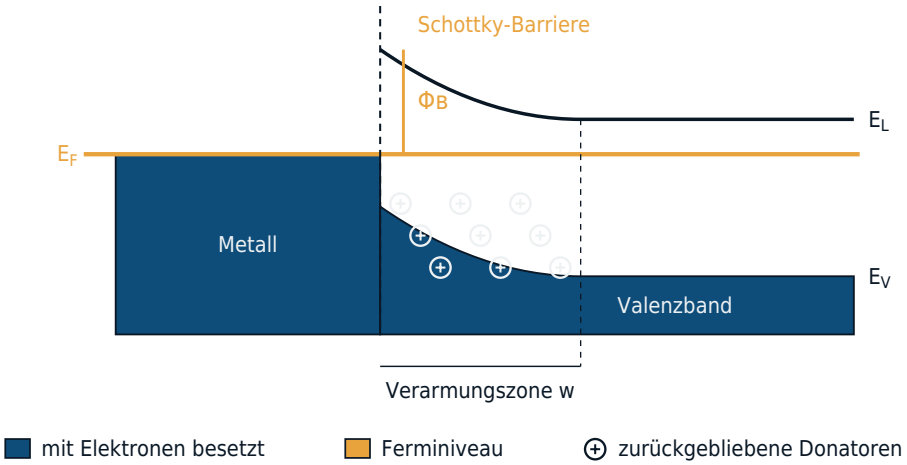
Es verringert sich somit die Aufenthaltswahrscheinlichkeit von Elektronen im Leitungsband des Halbleiters: Der Abstand zwischen Leitungsbandkante und Ferminiveau – welches den höchsten noch mit Elektronen besetzten Energiezustand beschreibt – wird an der Grenzfläche größer.

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.

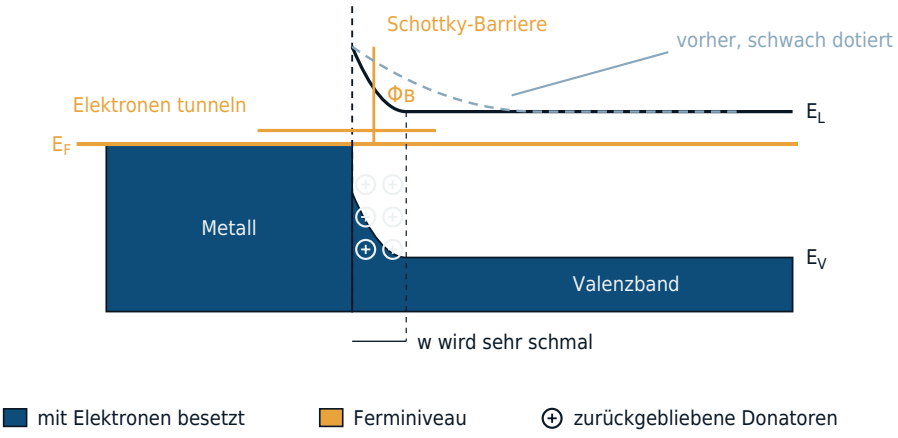
Die Weite *w* der Verarmungszone hängt von der Stärke der Dotierung ab. Die aus dem Halbleiter abgewanderten Elektronen erzeugen im Metall eine negative Raumladung, welche auf den Oberflächenbereich begrenzt ist.

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.

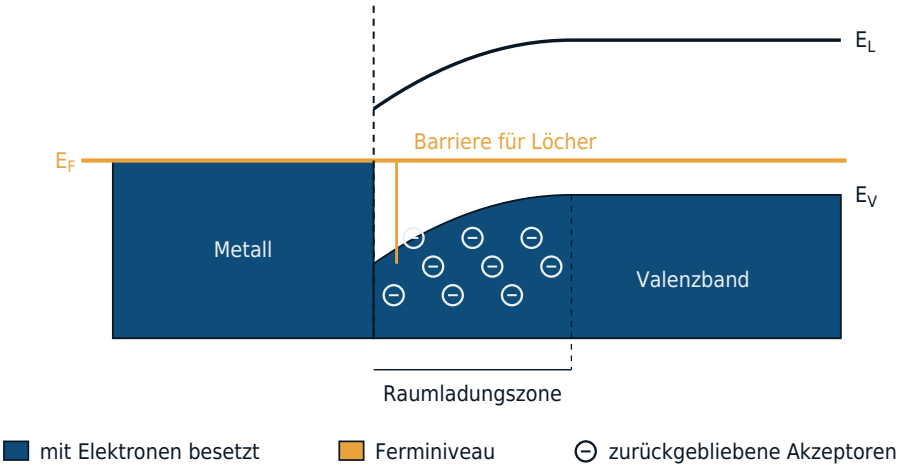
p-Halbleiter-Kontakt

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

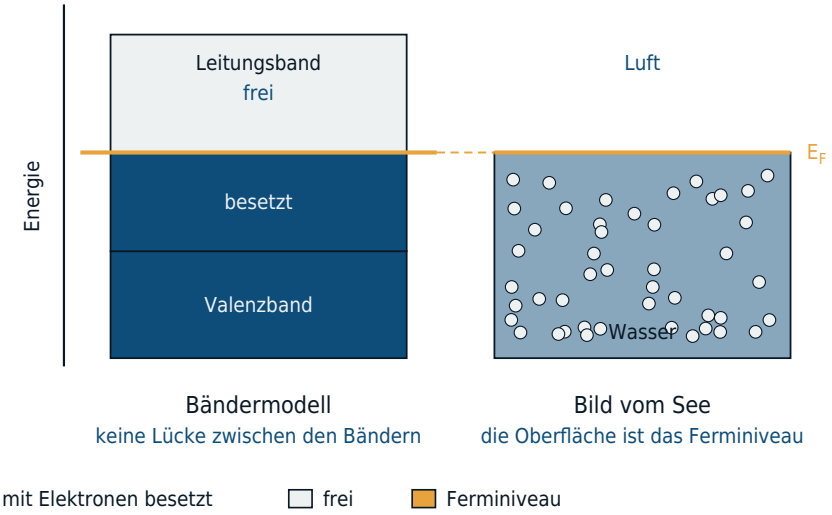
Nach der Herstellung der Transistoren im Siliciumsubstrat müssen diese mittels elektrischer Kontakte miteinander verbunden werden. Dabei wird zum einen die Gateelektrode zur Steuerung des Transistors kontaktiert, zum anderen müssen die dotierten Source- und Draingebiete, über die der Stromfluss erfolgt, 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



Bändermodell

keine Lücke zwischen den Bändern

Bild vom See

die Oberfläche ist das Ferminiveau

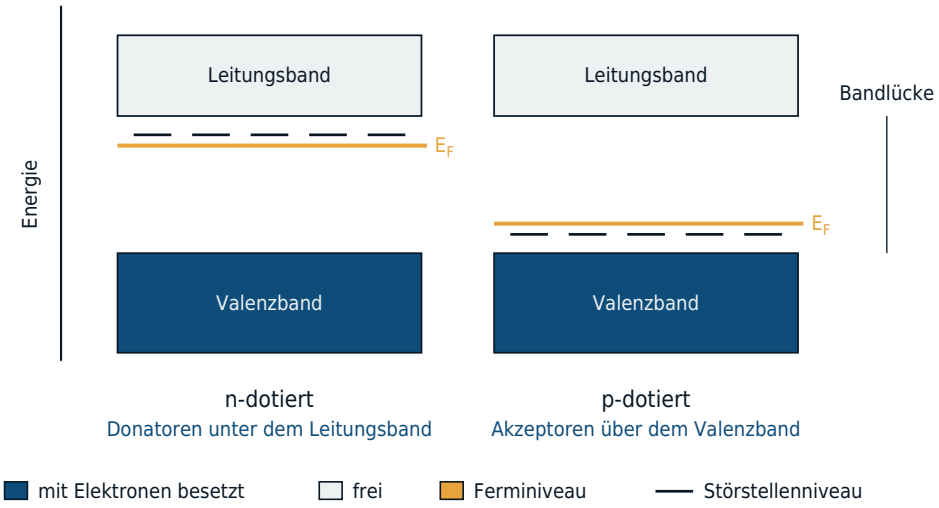
mit Elektronen besetzt frei Ferminiveau

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



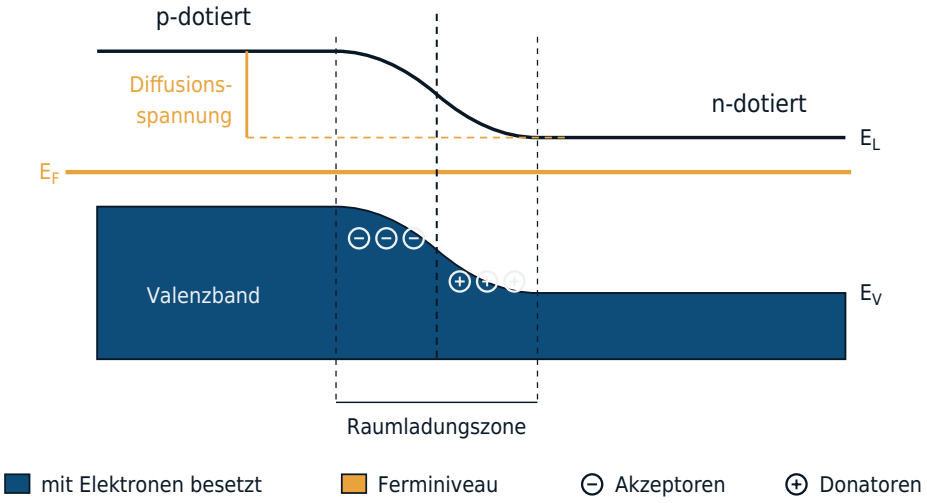
Das Ferminiveau rückt dorthin, wo die Ladungsträger leicht verfügbar sind: bei n-Dotierung nach oben zur Leitungsbandkante, bei p-Dotierung nach unten zur Valenzbandkante. Diese Lage bestimmt später den Kontaktwiderstand.

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.

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.

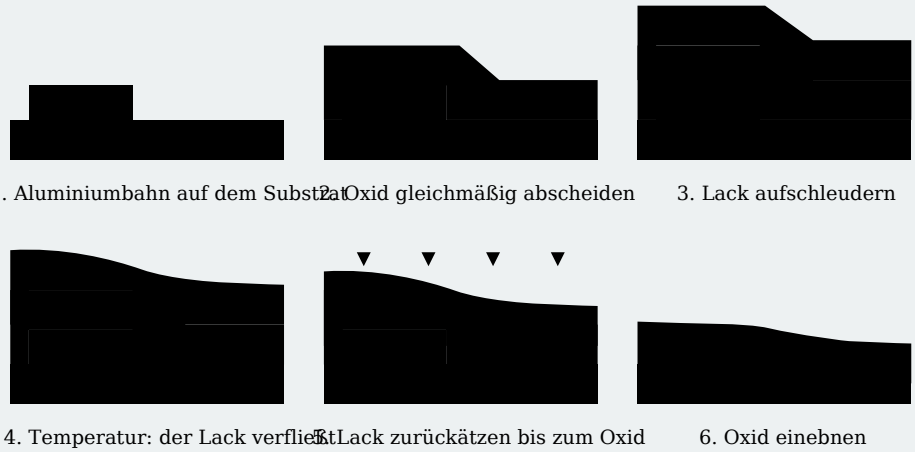
Reflowrückätzen

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.

Reflowrückätzen

Der verfließende Lack ebnet die Stufe ein, das Rückätzen überträgt sie ins Oxid



Lack und Oxid werden gleich schnell abgetragen. Deshalb überträgt sich die ebene Lackoberfläche beim Rückätzen nach unten in das Oxid. Über der Bahn bleibt eine dünne Deckschicht stehen – das Aluminium darf nicht freigelegt werden.

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.

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.

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

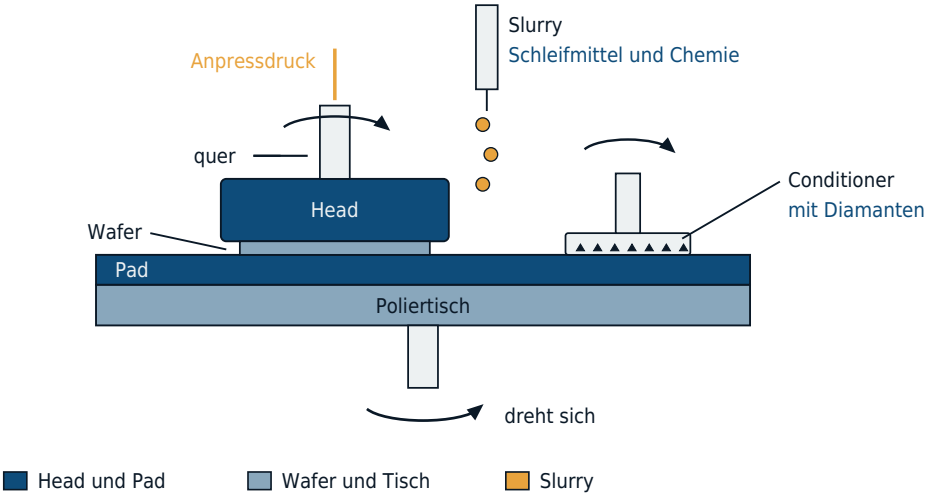
Beim chemisch mechanischen Polieren (auch chemisch mechanisches Planarisieren, kurz CMP) wird im Gegensatz zu den Reflowtechniken eine Homogenität über die gesamte Scheibenoberfläche erzielt. Dies ist vor allem im Hinblick auf lithografische Prozesse wichtig, die zur korrekten Belichtung eine möglichst planare Oberfläche benötigen. Ebenso ist eine Waferoberfläche ohne Topografie für alle nachfolgenden Schichten von Vorteil.

Der Wafer wird dazu in einem Chuck mit Vakuumansaugung (Head) mit der aktiven Seite nach unten gehalten und auf eine Polierfläche (Pad, meist aus Polyurethan) auf dem Poliertisch gepresst. Der Head und der Poliertisch drehen sich, während der Head gleichzeitig horizontale Bewegungen ausführen kann. Als Poliermittel zwischen Tisch und Wafer dient eine Lösung (Slurry) aus Schleifmitteln und chemischen Substanzen, die unter Druck die Oberfläche verändern und so den Polierprozess unterstützen. Zur besseren Verteilung der Slurry und um das Poliertuch zur konditionieren, kann das Pad mit einer mit Diamanten besetzten Stahlscheibe (Dresser/Conditioner) aufgeraut werden. Dies geschieht während (in-situ) oder vor/nach dem Polierschritt (ex-situ).

Der CMP-Prozess erfolgt gewöhnlich in zwei oder drei Stufen auf Pads mit unterschiedlichen Oberflächen und verschiedenen Slurrys. Die Wafer werden dazu nach jedem Polierschritt auf das nächste Pad transferiert. Im Anschluss erfolgt eine Reinigung, um Partikel und Slurryreste zu entfernen.

CMP-Anlage

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



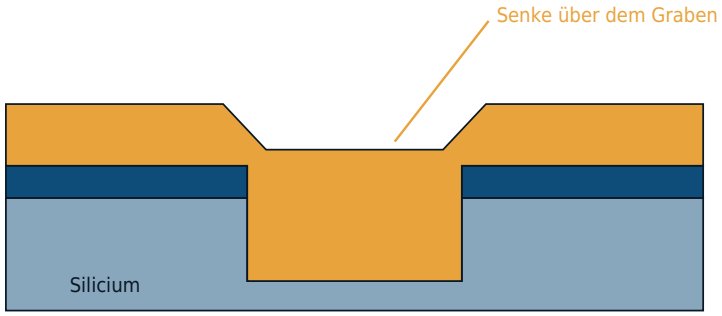
Head und Poliertisch drehen sich gegenläufig, der Head bewegt sich zusätzlich quer über das Pad. Erst diese Überlagerung sorgt dafür, dass jeder Punkt der Scheibe im Mittel gleich stark abgetragen wird.

Typischerweise wird der CMP-Prozess nach der Abscheidung des TEOS zur Shallow-Trench-Isolation eingesetzt um das Oxid soweit abzutragen, dass nur die Isolation zwischen den aktiven Gebieten der Transistoren bestehen bleibt. Ebenso wird das Zwischenoxid zwischen Transistorebene und der ersten Metallebene (First contact) mittels CMP auf die erforderliche Dicke zurückpoliert. In diesem Oxid werden anschließend die Kontakte zu den Source- und Draingebieten mittels Wolfram hergestellt. Auch hier dient das chemisch mechanische Polieren dazu, das auf der Oberfläche befindliche Metall zu entfernen. Wie im Kapitel Damscene-Verfahren beschrieben wurde, werden auch die Verdrahtungsebenen aus Kupfer im CMP-Prozess eingeebnet.

Im Folgenden ist der CMP-Prozess mit zwei Polierschritten im STI-Bereich dargestellt. Nach dem ersten Polierschritt wird das Oxid auf dem aktiven Gebiet und über den Gräben planarisiert. Im zweiten Polierschritt wird das restliche Oxid dann in einem selektiven Prozess bis auf die Passivierungsschicht abgetragen. Hierbei ist wichtig, dass das Oxid auf den Bereichen, auf denen später die Transistoren hergestellt werden, vollständig entfernt wird, da sonst das Nitrid, welches das darunterliegende Silicium während des Polierens schützt, nicht nasschemisch entfernt werden kann

1. Graben mit Oxid überfüllt

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

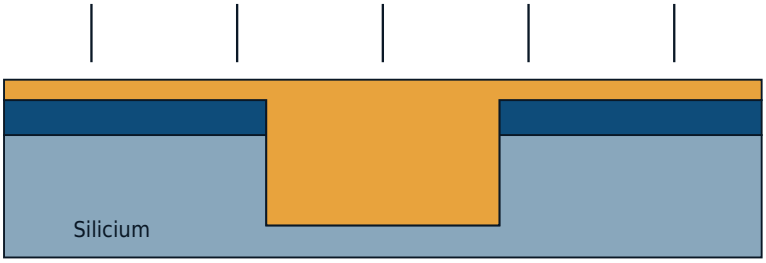


Silicium Nitrid als Polierstopp Oxid

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

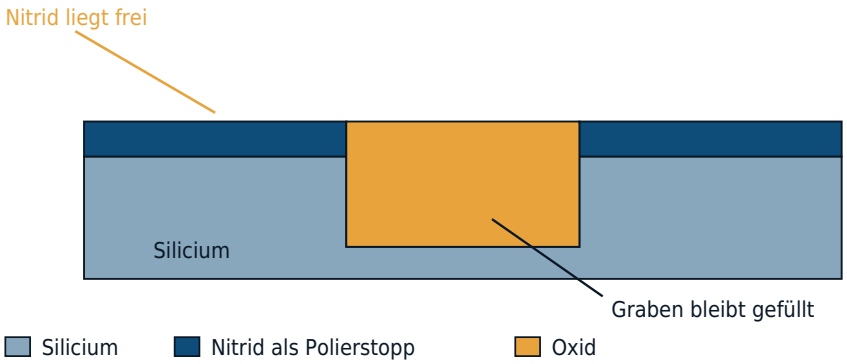


Silicium Nitrid als Polierstopp Oxid

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.

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.

Auch wenn dieses Verfahren recht grob anmutet, lässt sich hierbei doch eine auf wenige Nanometer planare Oberfläche herstellen. Es ist längst kein Sonderschritt mehr: In einem modernen Prozessablauf wird eine Scheibe einige Dutzend Mal poliert.

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.

Belichten und Belacken

Übersicht

In der Halbleiterfertigung werden Strukturen auf Siliciumscheiben mittels lithografischer Verfahren hergestellt. Dabei wird zuerst ein strahlungsempfindlicher Film, meist eine Fotolackschicht, auf dem Wafer aufgebracht, strukturiert und mit Ätzverfahren in die darunter liegende Schicht übertragen.

Die Fototechnik beinhaltet dabei folgende Prozessschritte:

- Aufbringen eines Haftvermittlers und Entfernen von Wasser auf dem Wafer
- Belacken der Wafer
- Stabilisieren der Lackschicht
- Belichten
- Entwickeln des Lackes
- Aushärten des Lackes
- Kontrolle

Bei einigen Prozessen wie der Ionenimplantation, dient der Fotolack als Schutzschicht um bestimmte Bereiche auf dem Wafer von der Implantation auszuschließen. Eine Übertragung der Lackmaske durch einen Ätzprozess findet hier nicht statt.

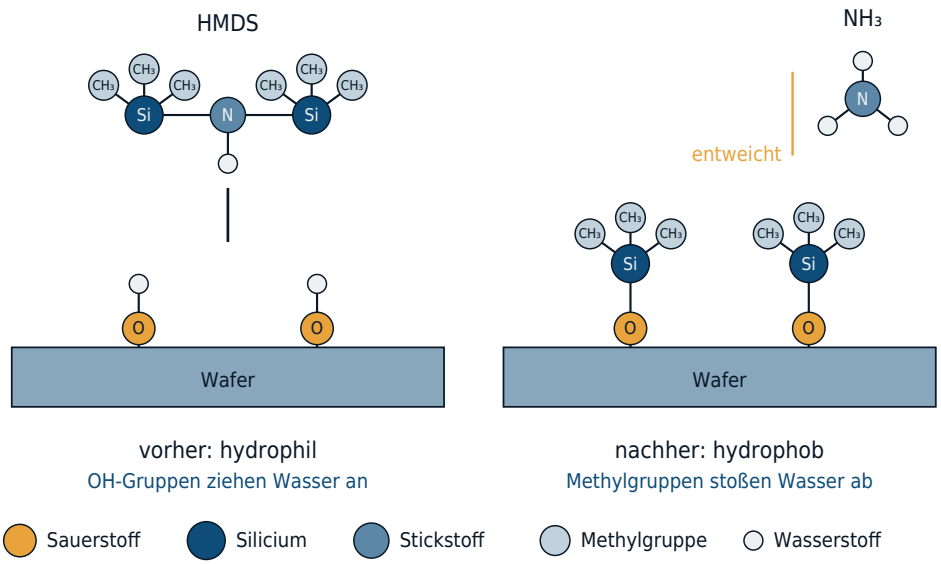
Aufbringen eines Haftvermittlers

Zu Beginn werden die Wafer gereinigt und ausgeheizt (Pre-Bake) um anhaftende Partikel zu beseitigen und angelagertes Wasser zu entfernen. Die Oberfläche der Wafer ist Wasser anziehend (hydrophil) und muss vor dem Aufbringen der Lackschicht hydrophob, also Wasser abstoßend und somit Lack anziehend gemacht werden. Dazu wird auf den Wafern ein Haftvermittler, meist Hexamethyldisilazan (HMDS), aufgebracht. Die Wafer werden dabei dem Dampf dieser Flüssigkeit ausgesetzt, so dass sich die Scheibenoberflächen damit benetzten.

Durch die Feuchtigkeit in der Umgebungsluft befinden sich auch nach dem Ausheizen immer Wasserstoff H oder Hydroxidionen OH^- an der Waferoberfläche. Das HMDS spaltet sich in Trimethylsiliciumgruppen $Si(CH_3)_3$ auf und entfernt den Wasserstoff unter Bildung von Ammoniak NH_3.

Anlagerung von HMDS

Aus der wasseranziehenden Oberfläche wird eine wasserabweisende



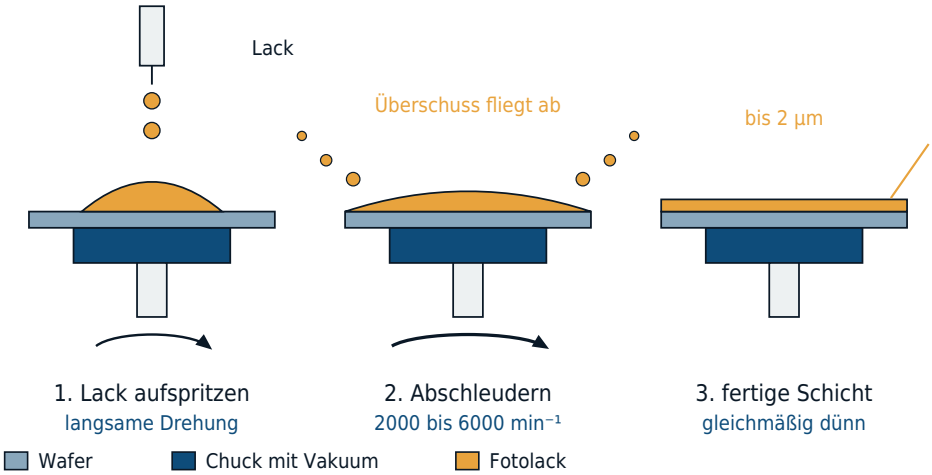
Das HMDS spaltet sich in zwei Trimethylsiliciumgruppen. Jede bindet an ein Sauerstoffatom der Oberfläche und nimmt dessen Wasserstoff mit; zusammen mit dem Stickstoff entsteht daraus Ammoniak.

Belacken

Die Belackung der Wafer erfolgt durch eine Schleuderbeschichtung auf einem drehbaren Teller mit Vakuumansaugung (Chuck). Bei niedriger Drehzahl wird Lack in der Mitte der Scheibe aufgespritzt und dann bei 2000–6000 Umdrehungen pro Minute durch die Zentrifugalkraft zu einer homogenen Lackschicht auseinander gezogen. Deren Dicke beträgt je nach anschließendem Prozess bis zu 2 µm. Die Dicke hängt dabei von der Drehzahl und der Zähigkeit (Viskosität) des Lackes ab.

Belacken durch Aufschleudern

Die Zentrifugalkraft zieht den Lacktropfen zu einer dünnen Schicht auseinander



Die Schichtdicke folgt aus Drehzahl und Zähigkeit des Lacks: Je schneller der Teller dreht und je dünnflüssiger der Lack, desto dünner die Schicht. Der weitaus größte Teil des aufgespritzten Lacks wird dabei abgeschleudert.

Damit der Lack auf der Scheibe gleichmäßig verfließen kann enthält er Wasser und Lösungsmittel, die ihn weich machen. Zur Stabilisierung der Lackschicht wird der Wafer danach bei ca. 100 °C ausgeheizt (Post-Bake bzw. Soft-Bake). Wasser und Lösungsmittel werden teilweise verdampft, eine Restfeuchtigkeit muss für die Belichtung erhalten bleiben.

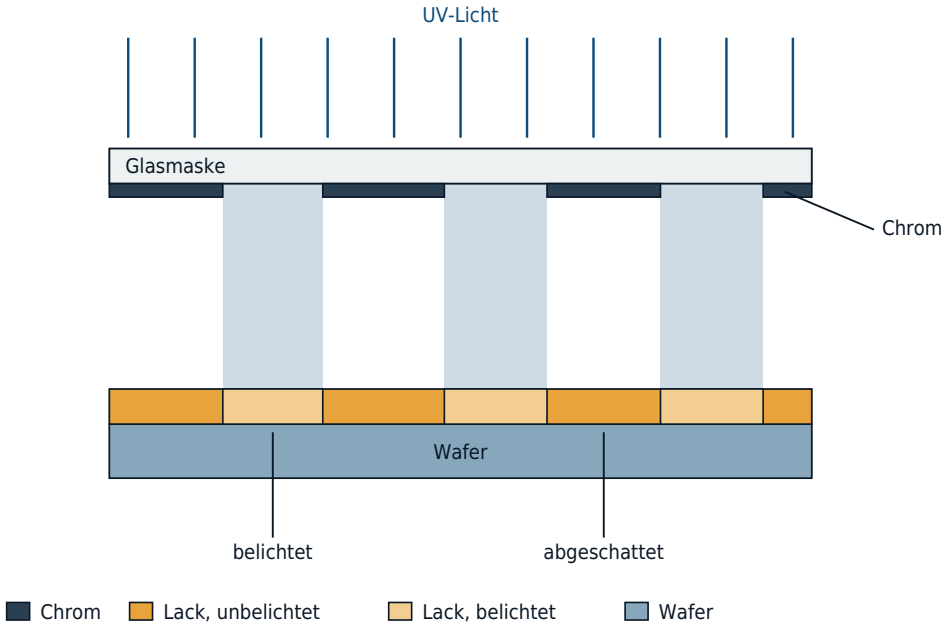
Heutzutage kommen auf Grund der geringen Strukturgrößen oft noch weitere Hilfsschichten zum Einsatz, welche entweder vor oder nach dem Fotolack auf dem Wafer aufgebracht werden. Diese Schichten können u. a. als Antireflexschicht oder zur zusätzlichen Planarisierung dienen.

Belichtung

In einer Belichtungsanlage befindet sich eine Glasmaske die teilweise mit Chrom beschichtet ist, dadurch werden partiell Bereiche des belackten Wafers belichtet und andere abgeschattet.

Belichtung durch eine Maske

Wo Chrom auf der Maske liegt, bleibt der Lack unbelichtet



Die Maske trägt das Muster als Chromschicht auf Glas. Beim Entwickeln wird je nach Lacktyp entweder der belichtete oder der unbelichtete Teil abgelöst – zurück bleibt die strukturierte Lackschicht.

Je nach Art des Lackes werden belichtete Teile löslich oder unlöslich. Mit Hilfe einer Entwicklerlösung werden die löslichen Bereiche entfernt, so dass eine strukturierte Lackschicht erhalten bleibt. Bei klassischem Positivlack der i-Linie spaltet sich ein Stickstoffmolekül N_2 durch das energiereiche UV-Licht ab. Zurück bleibt ein Keto-Karben, das sich aus energetischen Gründen zu Keten ($CH_2{=}C{=}O$) umwandelt. Unter Aufnahme von Feuchtigkeit aus der Umgebungsluft bildet sich aus dem Keten eine Carbonsäure. Bei 248 nm und darunter wird dieser Mechanismus durch chemisch verstärkte Lacke ersetzt, die eine Photosäure freisetzen (siehe Fotolack).

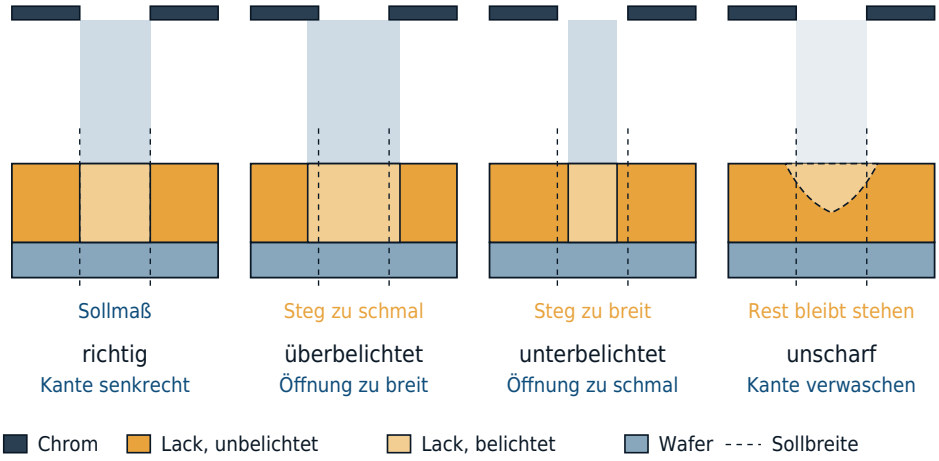
Die Belichtungszeit ist sehr wichtig damit die Strukturen die exakte Größe erhalten. Je länger die Scheibe den Strahlen der Belichtungsanlage ausgesetzt wird, desto größer werden dabei die belichteten Bereiche. Eine exakte Bestimmung der korrekten Belichtungsdauer zum Erreichen der vorgegebenen Strukturbreiten mit einem oder mehreren Testwafern (Vorläufern) ist notwendig, da sich der Lack je nach Umgebungstemperatur auch unterschiedlich verhalten kann.

Bei einer Überbelichtung werden Lackstege und damit die darunter liegenden Strukturen zu klein, Kontaktlöcher werden zu groß. Bei einer zu kurzen Belichtungszeit sind die Kontaktlöcher nicht geöffnet, Leiterbahnen sind zu breit und stehen unter Umständen miteinander in Kontakt. Zudem führt eine

schlechte Fokussierung zu unbelichteten Bereichen, so dass Kontaktlöcher bei der Entwicklung später nicht freigelegt werden und zwischen Leiterbahnen Verbindungen bestehen bleiben, die zu Kurzschlüssen führen.

Belichtungsprofile

Belichtungszeit und Fokus entscheiden über die Breite der Struktur



Je länger belichtet wird, desto breiter der belichtete Bereich: Lackstege werden schmaler, Kontaktlöcher größer. Bei zu kurzer Belichtung oder schlechtem Fokus bleibt am Boden Lack stehen und das Kontaktloch öffnet nicht.

Je nach nachfolgendem Prozess muss das Lackmaß, also die Breite der Lackstege oder die Größe der Kontaktlöcher angepasst werden. Bei isotropen Ätzungen (die Ätzung geschieht sowohl in vertikaler als auch horizontaler Richtung) wird die Maskierung nicht 1:1 in die zu strukturierende Schicht übertragen.

Eingesetzte Belichtungsverfahren

Zur Belichtung werden je nach Anforderung energiereiche Strahlen wie UV-Licht, Röntgenstrahlung, Elektronenstrahlen und Ionenstrahlen eingesetzt. Dabei gilt: Je kürzer die Wellenlänge der Strahlung, desto kleiner sind auch die erreichbaren Strukturbreiten. Den Zusammenhang beschreibt das Rayleigh-Kriterium:

$$CD=k_1\cdot\frac{\lambda}{NA}$$

Dabei ist λ die Belichtungswellenlänge, NA die numerische Apertur des Objektivs und k_1 ein Prozessfaktor, der die Beleuchtungsart, die Maskentechnik und den Fotolack zusammenfasst. Die Schärfentiefe nimmt dabei mit dem Quadrat der Apertur ab:

$$DOF=k_2\cdot\frac{\lambda}{NA^2}$$

Jede Verbesserung der Auflösung wird also mit einem kleineren Fokusfenster bezahlt – deshalb sind planare Waferoberflächen (siehe CMP) eine Voraussetzung für feine Strukturen.

Die Wellenlängen im Überblick

Wellenlänge	Quelle	Einsatz
436 nm (g-Linie) 365 nm (i-Linie)	Quecksilberdampflampe	historisch; die i-Linie wird für unkritische Ebenen, Leistungshalbleiter und MEMS weiterhin genutzt
248 nm	Kryptonfluorid-Excimerlaser	gröbere Ebenen, Analog- und Leistungstechnik
193 nm	Argonfluorid-Excimerlaser	Arbeitspferd der Fertigung, trocken und als Immersionsbelichtung
13,5 nm (EUV)	Zinn-Plasma, mit Laser gezündet	kritische Ebenen aller führenden Logik- und Speicherprozesse

Ursprünglich war als Zwischenschritt auch der Fluorlaser mit 157 nm geplant. Er erwies sich als nicht praktikabel, da die dafür nötigen Calciumfluoridlinsen Feuchtigkeit aus der Luft aufnehmen; zudem wäre eine Immersionsbelichtung mit Wasser nicht möglich, weil Wasser Licht dieser Wellenlänge absorbiert.

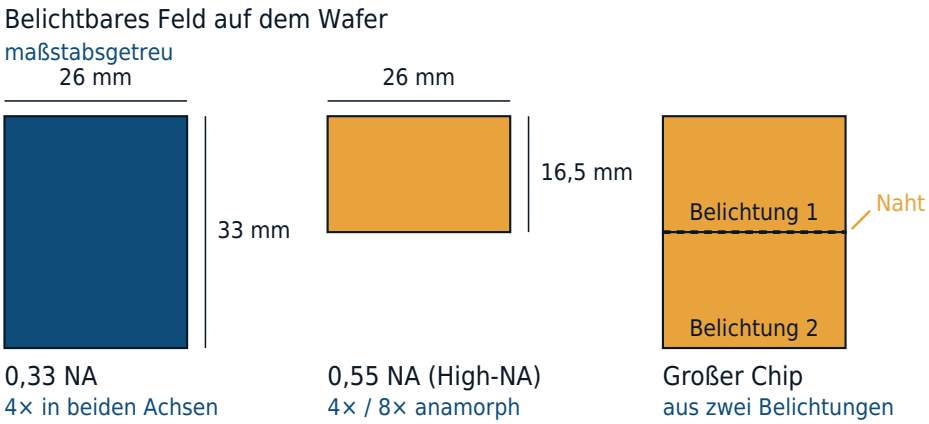
Immersionsbelichtung

Statt die Wellenlänge weiter zu verkürzen, lässt sich die numerische Apertur erhöhen. Dazu wird der Spalt zwischen letzter Linse und Wafer mit hochreinem Wasser gefüllt, dessen Brechungsindex bei 193 nm etwa 1,44 beträgt. Damit sind Aperturen bis 1,35 möglich, gegenüber rund 0,93 in Luft. Eine einzelne Immersionsbelichtung löst dabei Strukturen bis etwa 38 nm Halbperiode auf – das ist die physikalische Grenze dieser Technik. Feinere Strukturen entstehen mit 193 nm nur noch, indem man ein Muster auf mehrere Belichtungen aufteilt (siehe Multipatterning).

EUV-Lithografie

Der Sprung auf extremes Ultraviolett mit 13,5 nm ist seit 2019 in der Serienfertigung. Die Strahlung entsteht, indem Zinntröpfchen von einem Hochleistungslaser zu einem Plasma verdampft werden; da EUV von Luft und von jedem Glas absorbiert wird, arbeitet die gesamte Anlage im Vakuum und bildet ausschließlich über Spiegel ab. Die heutige Anlagengeneration mit einer Apertur von 0,33 erreicht in einer Belichtung etwa 13 nm Halbperiode und ersetzt damit an vielen Ebenen zwei oder mehr Immersionsschritte.

Belichtbares Feld bei 0,33 NA und 0,55 NA



Seit 2024 kommt die High-NA-Generation mit einer Apertur von 0,55 hinzu, die etwa 8 nm Halbperiode auflöst. Der Preis dafür ist eine anamorphe Optik: Das Bild wird in einer Richtung um den Faktor 4, in der anderen um den Faktor 8 verkleinert, wodurch sich das belichtbare Feld auf 26 × 16,5 mm halbiert. Große Chips müssen deshalb aus zwei aneinandergesetzten Belichtungen entstehen. Im Juli 2026 wurden erstmals High-NA-belichtete Ebenen in einem Serienprodukt qualifiziert; der breite Einsatz wird für 2027 und 2028 erwartet.

Weitere Verfahren

Röntgen- und Ionenstrahlen haben die Fertigung nie erreicht und sind der Forschung vorbehalten. Elektronenstrahlschreiber sind dagegen unverzichtbar – nicht zur Belichtung der Wafer, sondern zur Herstellung der Fotomasken.

Multipatterning

Eine einzelne Immersionsbelichtung mit 193 nm löst Strukturen bis etwa 38 nm Halbperiode auf. Als die Bauelemente diese Grenze unterschritten, bevor die EUV-Lithografie verfügbar war, blieb nur ein Weg: Das Muster wird auf mehrere Belichtungs- und Ätzschritte verteilt, die jeder einzeln innerhalb der Auflösungsgrenze liegen. Zusammen ergeben sie ein feineres Muster, als eine Belichtung erzeugen könnte.

Mehrfachbelichtung mit zwei Masken (LELE)

Beim Litho-Etch-Litho-Etch-Verfahren wird das Zielmuster in zwei Teilmuster zerlegt, die jeweils den doppelten Strukturabstand haben. Das erste wird belichtet und geätzt, danach folgt ein zweiter kompletter Durchlauf, dessen Strukturen genau in die Lücken des ersten fallen.

Der Nachteil liegt in der Justierung: Die Lage der zweiten Ebene zur ersten geht

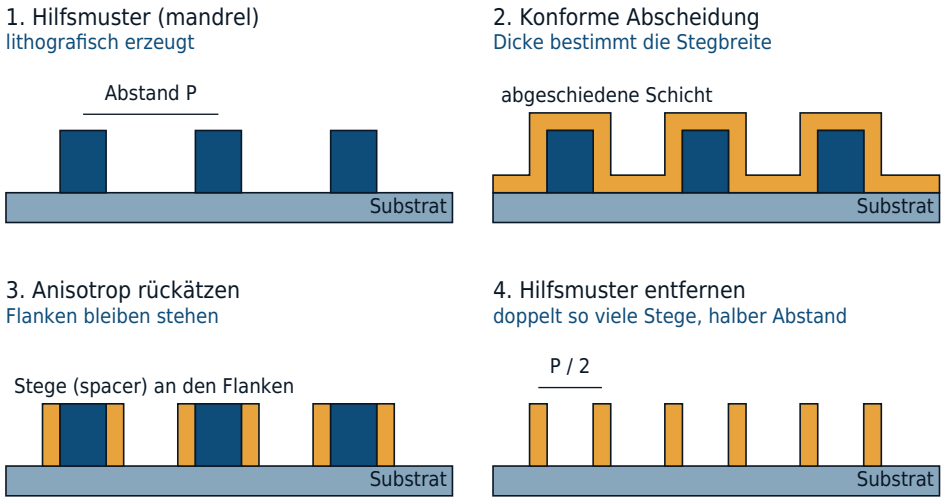
direkt in die Strukturbreite ein. Jeder Justierfehler erscheint als abwechselnd zu breiter und zu schmaler Abstand – ein systematischer Fehler, der die Bauelemente ungleichmäßig macht. Mit drei oder vier Durchläufen (LELELE und weiter) wächst dieses Problem, ebenso die Zahl der Prozessschritte und damit die Kosten.

Selbstjustierende Verdopplung (SADP)

Das selbstjustierende Verfahren umgeht das Justierproblem, indem die feinen Strukturen nicht belichtet, sondern abgeschieden werden:

- Ein Hilfsmuster (mandrel) wird mit normaler Auflösung lithografisch erzeugt.
- Darüber wird konform eine dünne Schicht abgeschieden, üblicherweise per Atomlagenabscheidung. Ihre Dicke bestimmt die späteren Strukturbreiten – und sie ist sehr viel genauer einstellbar als eine Belichtung.
- Ein anisotroper Ätzschritt entfernt die Schicht auf den waagerechten Flächen und lässt an den Flanken des Hilfsmusters Stege stehen (spacer).
- Das Hilfsmuster wird selektiv entfernt. Zurück bleiben doppelt so viele Stege, wie das Hilfsmuster Strukturen hatte.

Ablauf der selbstjustierenden Verdopplung



Da die Stege durch das Hilfsmuster selbst positioniert werden, gibt es zwischen ihnen keinen Justierfehler. Wiederholt man das Verfahren, entstehen vier Strukturen pro ursprünglicher (SAQP). Der Preis: Es entstehen immer geschlossene Ringe um das Hilfsmuster, und die Struktur ist zwangsläufig regelmäßig. Beides muss mit zusätzlichen Masken korrigiert werden, die unerwünschte Abschnitte gezielt durchtrennen (cut mask) oder Enden definieren. Das Verfahren eignet sich deshalb für dichte, regelmäßige Muster –

Leitbahnen, Gate-Reihen, Speicherzellenfelder – und nicht für beliebige Layouts.

Bedeutung heute

Multipatterning ist keine Übergangslösung geblieben. Auch mit EUV wird es weiter eingesetzt, nur an anderer Stelle: Wo EUV die kritischen Ebenen in einer Belichtung schafft, entfällt die Aufteilung; wo Strukturen unterhalb der EUV-Auflösung liegen, wird sie erneut nötig. Zugleich hat sich der Entwurf dauerhaft verändert. Layouts folgen heute strengen Regeln mit einheitlichen Rastern und einheitlichen Vorzugsrichtungen, weil nur solche Muster überhaupt zerlegbar sind – die Lithografie schreibt dem Schaltungsentwurf ihre Beschränkungen vor.

Belichtungsverfahren

Übersicht

Die Fototechnik kann, je nach Art der Bestrahlung, in verschiedene Verfahren unterteilt werden: optische Lithografie (Fotolithografie), Elektronenstrahllithografie, Röntgenstrahllithografie und Ionenstrahllithografie.

Bei der optischen Lithografie werden strukturierte Fotomasken (Reticle) verwendet. Die Belichtung mit UV-Licht oder Gaslasern erfolgt entweder im Maßstab 1:1 oder reduzierend, beispielsweise 4:1 oder 10:1.

Die optische Lithografie zerfällt heute in zwei technisch völlig verschiedene Zweige:

- DUV (deep ultraviolet, 248 und 193 nm): Die Maske ist transparent, abgebildet wird durch ein Linsensystem, die Anlage arbeitet an Luft oder mit einem Wasserfilm zwischen Objektiv und Wafer.
- EUV (extremes Ultraviolett, 13,5 nm): Da diese Strahlung von jedem Material absorbiert wird, arbeitet die Anlage im Vakuum, bildet über Spiegel ab und die Maske ist kein Transparent, sondern ein Spiegel.

Die nachfolgend beschriebenen Belichtungsverfahren sind nach steigendem Auflösungsvermögen geordnet. Kontakt- und Abstandsbelichtung finden sich heute nur noch in der Mikromechanik, in der Forschung und in der Ausbildung; die Chipfertigung nutzt ausschließlich die reduzierende Projektionsbelichtung.

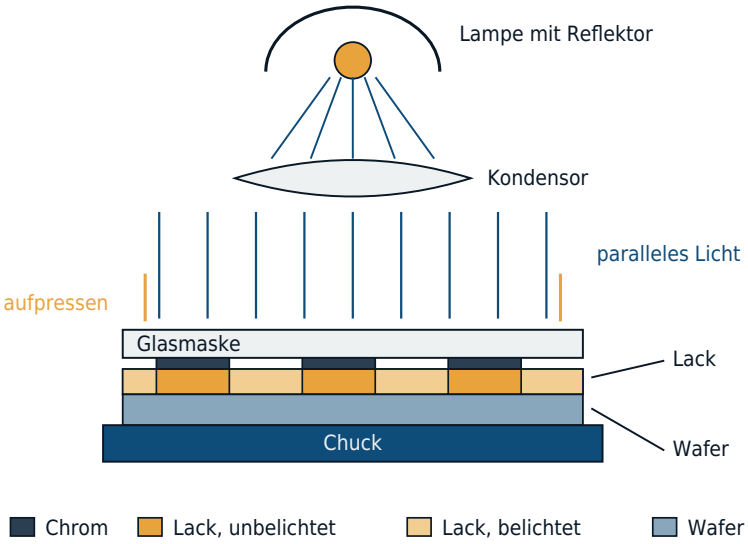
Kontaktbelichtung

Das Kontaktbelichtungsverfahren ist das älteste angewandte Verfahren. Dabei wird die Maske direkt auf die Lackschicht gepresst, die Strukturen im Maßstab 1:1 übertragen. Auflösungsbegrenzende Streu- bzw. Beugungseffekte des Lichts

treten nur an den Strukturkanten auf. Die erzielten Strukturweiten sind bei diesem Verfahren jedoch gering. Da alle Chips auf einmal belichtet werden ist der Scheibendurchsatz bei dieser Technik sehr hoch, der Aufbau der Belichtungsanlagen ist relativ einfach.

Kontaktbelichtung

Die Maske liegt unmittelbar auf dem Lack, die Übertragung erfolgt 1:1



Weil Maske und Lack sich berühren, entstehen Beugungseffekte nur an den Strukturkanten. Der Aufbau ist einfach und belichtet alle Chips auf einmal. Dafür verschmutzt und verkratzt die Maske, weil sie auf den Lack gepresst wird.

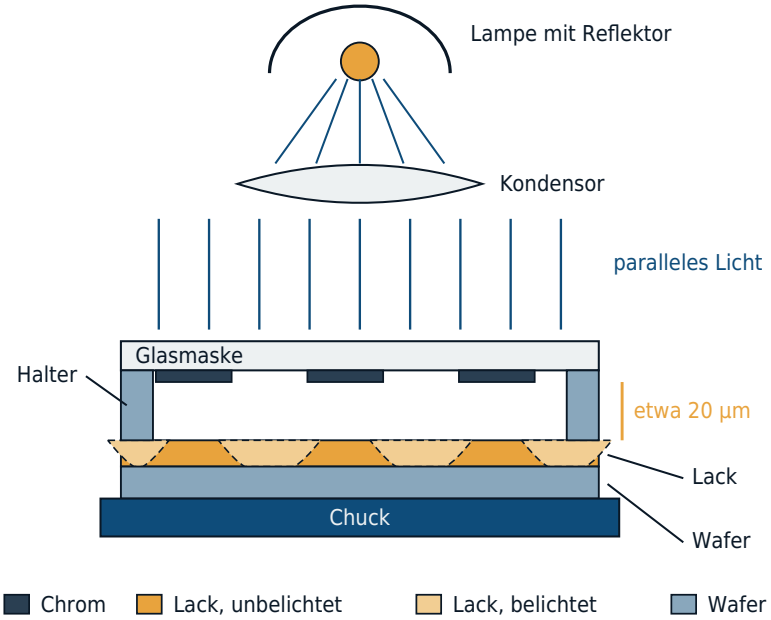
Die Nachteile liegen jedoch auf der Hand: durch die angepresste Maske auf den Lack verschmutzt diese schnell, oder kann, wie auch der Lack, dabei verkratzt werden. Befinden sich Partikel zwischen Scheibe und Maske wird die Abbildung durch den Abstand von Maske und Wafer verschlechtert.

Abstandsbelichtung

Bei der Abstands- oder Proximitybelichtung wird der Kontakt von Maske und Scheibe durch einen Abstandshalter (ca. 20 µm) vermieden. Dabei wird jedoch nur ein Schattenbild der Maske auf dem Wafer abgebildet, welches eine deutlich schlechtere Auflösung der Strukturen bietet.

Abstandsbelichtung

Abstandshalter halten die Maske über dem Lack – es entsteht ein Schattenbild



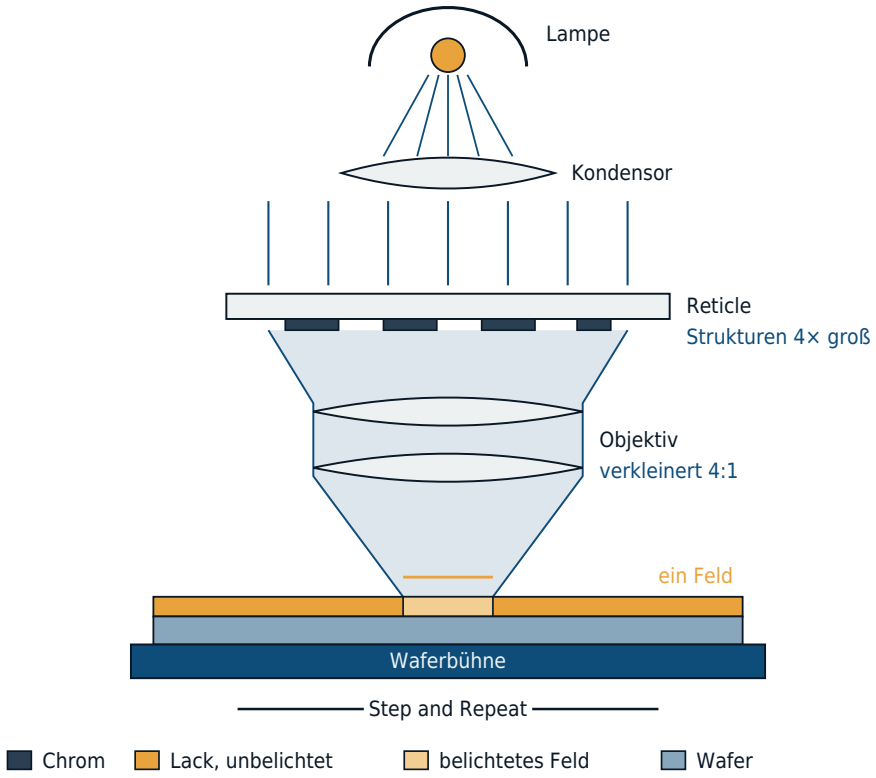
Der Spalt schützt Maske und Lack vor Kratzern und vor Partikeln dazwischen. Dafür fällt nur ein Schattenbild auf die Scheibe: Das Licht wird an den Chromkanten gebeugt, die Strukturen werden breiter und ihre Ränder unscharf.

Reduzierende Projektionsbelichtung

Dies ist das einzige Belichtungsverfahren, das in der Chipfertigung noch eingesetzt wird. Bei diesem Verfahren ist die Step-and-Repeat-Technik gebräuchlich. Dabei wird ein einzelner Chip - bei geringer Größe auch mehrere - über ein Reticle auf dem Wafer abgebildet. So wird die gesamte Fläche der Scheibe Chip für Chip belichtet.

Reduzierende Projektionsbelichtung

Das Objektiv bildet das Reticle verkleinert auf den Wafer ab



Weil die Strukturen auf dem Reticle vier- bis zehnfach größer sind, werden auch Partikel und Maskenfehler mitverkleinert und fallen oft unter die Auflösungsgrenze. Belichtet wird Feld für Feld, dazwischen verfährt die Waferbühne.

Bei hoher Auflösung wird das Feld nicht mehr als Ganzes belichtet, sondern Maske und Wafer werden gegenläufig synchron unter einem schmalen Lichtspalt hindurchgeführt (step and scan). Nur so lässt sich das Objektiv über das ganze Feld korrigiert halten. Der Vorteil bei diesem Verfahren ist, dass die auf dem Reticle abgebildeten Strukturen um den Faktor 4 vergrößert vorliegen (bei den High-NA-EUV-Anlagen in einer Achse um den Faktor 8). Bei der verkleinerten Abbildung auf den Wafer werden neben den Strukturen auch alle Fehler, wie Partikel, verkleinert dargestellt oder fallen sogar unter die Auflösungsgrenze; die Auflösung selbst wird bei diesem Verfahren verbessert.

Da für eine Ebene nicht die komplette Maske genutzt wird, können mehrere Ebenen auf einer Glasplatte aufgebracht werden, so dass die Maskenkosten gesenkt werden.

Zusätzlich kann ein Pellicle, eine dünne Folie mit Alurahmen, auf die Maske geklebt werden, wodurch Partikel von der Maske ferngehalten werden. Diese befinden sich nun außerhalb des Schärfebereichs und werden nicht abgebildet.

Daneben gibt es noch die Spiegelprojektionsbelichtung (1:1), bei der der Wafer

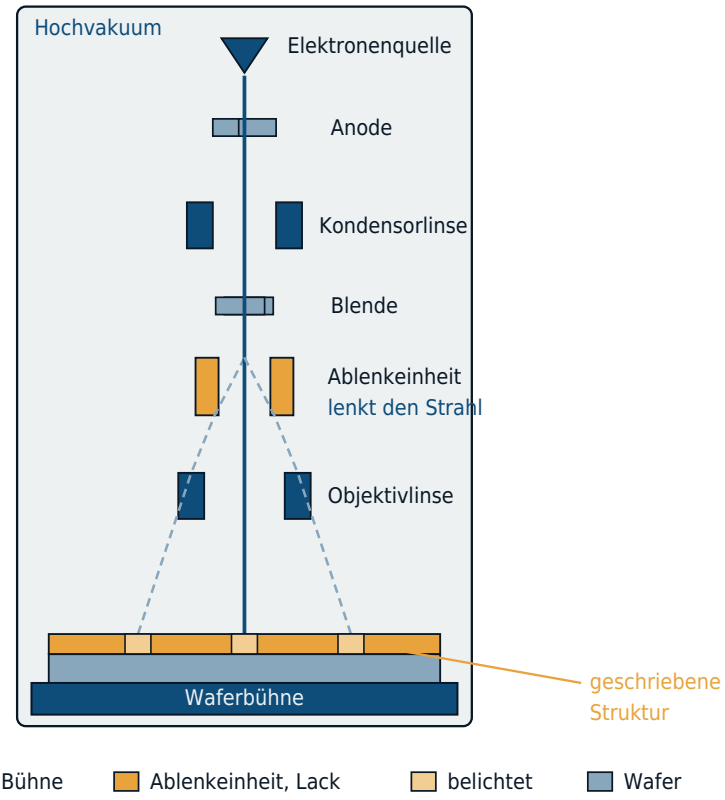
über ein komplexes System aus Spiegeln belichtet wird. Durch die Verwendung von Spiegeln entstehen keine Farbfehler, wie sie bei Linsensystem auftreten, dazu können Ausdehnungen der Scheibe durch Temperatureinfluss in Prozessen ausgeglichen werden. Jedoch werden über Spiegelsysteme Bilder verzehrt/gekrümmt dargestellt. Durch die 1:1-Abbildung ist die Auflösung außerdem stark begrenzt.

Elektronenstrahllithografie

Wie bei der Maskenherstellung wird ein fokussierter Elektronenstrahl über den belackten Wafer gescannt. Dabei kann das Scannen zeilenweise im Raster-Scan-Verfahren oder im Vektorscanverfahren durchgeführt werden. Jedoch muss auch hier jede Struktur einzeln geschrieben werden, was sehr zeitintensiv ist. Der Vorteil ist, dass keine Masken benötigt werden, was Kosten spart. Die gesamte Apparatur befindet sich dabei im Hochvakuum.

Elektronenstrahllithografie

Ein fokussierter Strahl schreibt die Struktur direkt – ohne Maske



Der Strahl fährt die Struktur Punkt für Punkt ab – zeilenweise im Rasterscan oder gezielt im Vektorscan. Das kostet Zeit, spart aber die Maske und ist deshalb vor allem für Prototypen und für die Maskenherstellung selbst üblich.

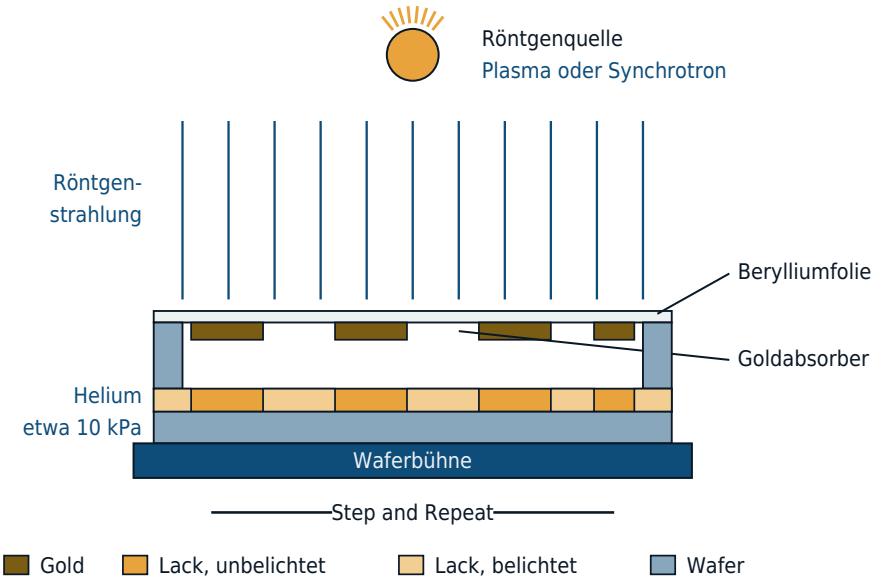
Röntgenstrahllithografie

Die Auflösungsgrenze der Röntgenstrahllithografie liegt bei etwa 40 nm. Die Abbildung erfolgt im 1:1 Step-and-Repeat-Verfahren durch Schattenwurf, und wird unter Atmosphärendruck in Luft oder bei leichtem Unterdruck in Heliumatmosphäre (ca. 10.000 Pa) durchgeführt. Die Röntgenquelle kann dabei eine Plasmaquelle oder Synchrotonstrahlung sein.

Anstelle der chrombeschichteten Glasmasken werden dünne, mechanisch stabile Folien aus Beryllium, teils auch aus Silicium verwendet. Um die Röntgenstrahlung zu absorbieren werden die Folien mit schweren Elementen wie Gold beschichtet. Die Anlagen sowie die Masken sind sehr teuer.

Röntgenstrahllithografie

Schattenwurf im Maßstab 1:1 – für Röntgenstrahlen gibt es keine Linsen



Röntgenstrahlen lassen sich nicht mit Linsen bündeln, deshalb wird das Muster direkt als Schatten übertragen – eins zu eins, Feld für Feld. Die Auflösung reicht bis etwa 40 nm, doch Anlagen und Masken sind sehr teuer.

Weitere Verfahren

Eine weitere Möglichkeit der Lithografie ist die Bestrahlung der Wafer mit Ionen. Mit den Ionen kann der Wafer sowohl über eine Maske strukturiert, als auch direkt wie bei der Elektronenstrahlmethode beschrieben werden. Im Falle von Wasserstoffionen beträgt die Wellenlänge 0,0001 nm. Mit anderen Elementen ist auch eine direkte Dotierung ohne Maskierung denkbar.

Der Fotolack

Lacktechnik

Es gibt Positiv- und Negativlacke, die bei unterschiedlichen Anwendungen eingesetzt werden. Während beim Positivlack die belichteten Stellen löslich werden, werden bestrahlte Bereiche beim Negativlack für die Entwicklerlösung unlöslich.

Charakteristik der Positivlacke:

sehr gute Auflösung
stabil in der Entwicklerlösung
in Laugenlösungen entwickelbar
nur bedingt resistent gegen Ätzung und Implantationsprozesse
schlechte Haftung auf dem Wafer

Charakteristik der Negativlacke:

sehr empfindlich (exaktere Strukturierung bei der Belichtung möglich als Positivlack)
gute Haftung
resistent gegen Ätzverfahren und Implantation
billiger als Positivlack
geringe Auflösung
Xylol-Entwickler (giftig)

Zur Strukturierung der Wafer in der Fertigung werden fast ausschließlich Positivlacke verwendet. Negativlacke finden meist als Passivierungsschicht (Fotoimidschicht) Anwendung, die mittels UV-Licht ausgehärtet werden können. Sofern im Text keine Angaben zum Lack gemacht werden, handelt es sich um Positivlack; bei Negativlack wird dies ausdrücklich erwähnt.

Chemische Zusammensetzung

Fotolacke (auch als Fotoresist bezeichnet; resist = widerstehen) setzen sich aus einem Bindemittel, einem Sensibilisator und einem Lösemittel (Verdünner) zusammen.

Bindemittel (Anteil ~20%)
: Als Bindemittel werden häufig Novolake eingesetzt. Dabei handelt es sich um Phenolharze (Kunstharz, Kunststoff), welche in erster Linie die thermischen Eigenschaften des Lackes bestimmen.

Sensibilisator (Anteil ~10%)
: Der Sensibilisator bestimmt die Lichtempfindlichkeit des Lackes. Sensibilisatoren setzen sich aus Molekülen zusammen, die bei einer

Belichtung mit energiereicher Strahlung die Löslichkeit des Lackes verändern (im Positivlack entsteht aus dem Sensibilisator durch die Belichtung Carbonsäure. Mehr dazu im Kapitel Entwickeln). Damit der Lack nicht durch das Licht in den Fertigungshallen belichtet wird, finden die Fototechnikprozesse unter Gelblicht statt, gegen das der Lack unempfindlich ist.

Lösemittel (Anteil ~70%)
: Die Lösemittel bestimmen die Viskosität des Lackes. Durch das Verdampfen der Lösemittel auf einer Heizplatte wird der Lack stabilisiert und resistent für nachfolgende Prozesse.

Ein vom Hersteller gelieferter Lack hat eine definierte Oberflächenspannung und Dichte, einen bestimmten Feststoffgehalt und eine bestimmte Viskosität. Somit ist bei der Belackung in der Chipfertigung die erzielte Fotolackdicke ausschließlich von der Drehzahl der Belackungsanlage abhängig.

Chemisch verstärkte Lacke

Der oben beschriebene Aufbau gilt für die klassischen Lacke der i-Linie. Bei 248 nm und darunter reicht die Zahl der eintreffenden Photonen nicht mehr aus, um genügend Sensibilisatormoleküle direkt umzusetzen. Eingesetzt werden deshalb chemisch verstärkte Lacke (chemically amplified resist, CAR): Statt eines Sensibilisators enthalten sie einen Photosäurebildner (photo acid generator, PAG), der bei Belichtung eine starke Säure freisetzt. Diese spaltet in einem anschließenden Temperschritt (post exposure bake) Schutzgruppen vom Bindemittel ab und wird dabei nicht verbraucht, sondern wirkt katalytisch weiter. Ein einzelnes Photon löst so hunderte bis tausende Reaktionen aus.

Der Verstärkungsmechanismus hat einen Preis: Die Säure diffundiert während des Temperschritts, was die Kante der späteren Lackstruktur verschmiert. Diffusionslänge, Bake-Temperatur und -Zeit sind damit auflösungsbestimmende Prozessparameter.

Lacke für die EUV-Belichtung

Ein EUV-Photon trägt mit 92 eV rund vierzehnmal mehr Energie als ein Photon bei 193 nm. Für dieselbe eingebrachte Energie treffen also entsprechend weniger Photonen auf die Fläche, und deren Ankunft ist ein statistischer Prozess. Bei Strukturen von wenigen Nanometern schwankt die Photonenzahl pro Struktur merklich – die Folge sind zufällig fehlende oder zusätzliche Strukturen sowie eine rauere Kante. Diese stochastischen Defekte, nicht die Optik, sind heute die praktische Auflösungsgrenze der EUV-Lithografie.

Gegenmaßnahmen sind höhere Belichtungsdosen (die den Durchsatz kosten) und Lacke mit höherer Absorption. Neben weiterentwickelten CAR kommen

dafür Metalloxidlacke auf Zinnbasis in Betracht, die EUV deutlich stärker absorbieren als organische Systeme. Ebenfalls in Entwicklung sind trocken aufgebrachte Lacke, die aus der Gasphase abgeschieden statt aufgeschleudert werden.

Die Schichtdicken sind dabei deutlich geringer als in der klassischen Fototechnik: Bei EUV liegen sie im Bereich einiger zehn Nanometer, da hohe schmale Lackstege sonst durch die Oberflächenspannung der Entwicklerlösung umkippen (pattern collapse). Die mechanische Stabilität und die Ätzresistenz übernehmen deshalb zusätzliche Hilfsschichten unter dem Lack.

Entwickeln und Kontrolle

Entwickeln

Die belichteten Scheiben werden in Tauch- oder Sprühentwicklungen prozessiert. Während beim Tauchätzschritt eine komplette Horde in einer Laugenlösung entwickelt und anschließend gespült werden kann, wird bei der Sprühentwicklung jeder Wafer einzeln prozessiert. Wie beim Belacken befindet sich der Wafer auf einem Chuck und wird bei langsamer Umdrehung stetig mit Entwicklerlösung besprüht. Nach dem vollständigen Entwickeln des Lackes wird der Wafer mit aufgespritztem Wasser gespült um den Entwicklungsprozess zu stoppen. Einige Vorteile der Sprühentwicklung gegenüber dem Tauchverfahren sind folgende:

- kleinste Strukturen können freigelegt werden
- Die Entwicklerlösung wird stetig erneuert: Verunreinigungen werden verhindert
- Die Menge der eingesetzten Entwicklerlösung ist wesentlich geringer

Durch das Entwickeln lösen sich, entsprechend der Lackart, bestimmte Bereiche des Lackes, so dass am Ende eine strukturierte Scheibe bestehen bleibt. Durch die Belichtung wird im Lack eine Reaktion ausgelöst, bei der der Sensibilisator in Säure umgewandelt wird. Diese Carbonsäure wird beim Entwickeln nach folgender Gleichung mit Natronlauge (NaOH) zu wasserlöslichem Salz umgewandelt:

$R\text{-}COOH + NaOH \rightarrow (R\text{-}COO)^- + Na^+ + H_2O$ (R = weitere unbeteiligte Stoffe)

Darstellung der belichteten Lackarten nach dem Entwickeln

Da bei Kali- oder Natronlösungen Rückstände des Entwicklers auf dem Wafer

zurückbleiben, werden auch Metallionen freie Entwickler wie TMAH (Tetramethylammoniumhydroxid) eingesetzt. Durch einen erneuten Aushärteschritt (Hard-Bake) wird der Lack für nachfolgende Prozesse, wie Ätzen oder für eine Ionenimplantation, beständig gemacht.

Lackkontrolle

Nun wird die Lackstruktur kontrolliert. Unter schrägem Lichteinfall lassen sich im Mikroskop die Gleichmäßigkeit der Lackschicht, sowie schlechte Fokussierungen und Lackanhäufungen erkennen. Sind die Lackstege zu dick oder dünn muss der Lack entfernt und der Prozess wiederholt werden. Ebenso muss die Lackstruktur zu der sich darunter befindlichen Ebene exakt justiert sein, andernfalls ist ebenso eine Wiederholung der Belackungs- und Belichtungsprozesse notwendig. Dazu gibt es unterschiedliche Justiermarken für die Justiergenauigkeit und die Linienweitenkontrolle:

Darstellung von Justiermarken

Die Breite der Lackstege wird mit einem Mikroskop kontrolliert: Lichtstrahlen treffen senkrecht auf den Wafer, und werden an Kanten nicht zum Objektiv zurückgestreut. So sind schwarze Linien als Strukturbegrenzungen sichtbar, mit deren Abständen zueinander man unter zu Hilfenahme der Mikroskopvergrößerung die Breite der Stege berechnen kann.

Lackentfernen

Nachdem die Struktur unter dem Lack in Ätzverfahren abgetragen wurde, oder der Lack beim Implantationsprozess als Maskierungsschicht gedient hat, muss der Lack entfernt werden. Dies geschieht mit starken Ätzlösungen (Remover), in einem Trockenätzschritt oder mit Lösungsmitteln. Als Lösungsmittel zum Entfernen der Lackschicht eignet sich Aceton, da es den Wafer und andere Schichten nicht angreift. Durch eine Ionenimplantation oder einen Trockenätzprozess kann die Lackschicht jedoch weiter ausgehärtet sein, so dass Lösungsmittel den Lack nicht mehr angreifen und entfernen können.

In diesem Fall kann der Lack mit einem Remover bei ca. 80 °C in einem Tauchverfahren entfernt werden. Hat sich der Lack während der Bearbeitung auf über 200 °C erwärmt kann er auch vom Remover nicht mehr abgetragen werden. Dann muss der Lack mittels Lackveraschung oder Lackverbrennung entfernt werden.

Unter Anwesenheit von Sauerstoff wird dabei durch Hochfrequenzanregung

eine Gasentladung gezündet, so dass angeregte Sauerstoffatome entstehen. Diese verbrennen den Lack rückstandsfrei. Die geladenen Teilchen werden jedoch durch das elektrische Feld stark beschleunigt und können so einen leichten Abtrag der Scheibenoberfläche oder eine geringe Schädigung der Scheibe verursachen. Das Lackveraschen (Strippen) kann direkt nach dem Ätzen noch in derselben Prozesskammer (in-situ) oder in einer anderen Kammer (ex-situ) erfolgen.

Maskentechnik

Maskentechnik

Die in der Fototechnik eingesetzten Masken enthalten ein Muster mit dem die jeweilige Schicht auf dem Wafer strukturiert wird. Ausgangsmaterial für die Masken sind Glasplatten, die ganzflächig mit Chrom und Lack beschichtet sind (Blanks). Über den für Elektronenstrahlen empfindlichen Lack wird die Chromschicht strukturiert, welche dann die lichtundurchlässigen Bereiche auf der Glasmaske darstellt.

Mit einem Elektronenstrahl werden die Masken direkt beschrieben. Die gesamte Apparatur – die Elektronenstrahlquelle, Fokussier- und Ablenkeinheit und der Blank – befindet sich im Hochvakuum (0,01–100 Pa; normaler Luftdruck ca. 100.000 Pa). Der Elektronenstrahl wird computergesteuert über die Maske gelenkt und belichtet den Lack. Bei dieser Methode lassen sich Strukturen bis weit unter 100 nm auflösen.

Ein einzelner Strahl schreibt eine hochaufgelöste Maske allerdings nicht in vertretbarer Zeit. Heutige Maskenschreiber arbeiten deshalb mit mehreren hunderttausend parallel geführten Einzelstrahlen (multi beam mask writer), die gemeinsam über den Blank geführt und einzeln ein- und ausgeschaltet werden. Erst damit sind Schreibzeiten von einigen Stunden pro Maske erreichbar – und erst damit lassen sich beliebig geformte, gekrümmte Strukturen genauso schnell schreiben wie rechtwinklige.

Korrektur der Abbildungsfehler

Auf Grund wellenoptischer Effekte (z.B. Beugung) kann es bei der Belichtung der Wafer zu Abbildungsfehlern kommen, welche durch die sogenannte *optical proximity correction* (OPC, etwa: optische Nahbereichskorrektur) korrigiert bzw. verringert werden.

Dazu können die eigentlichen Strukturen so abgeändert werden, dass die Abbildung auf dem Wafer dem gewünschten Muster entspricht. Außerdem

können zusätzliche Hilfsstrukturen auf die Maske geschrieben werden welche nur dazu dienen, Abbildungsfehler zu minimieren, jedoch keine Funktion für die Schaltung haben.

Bei den kleinsten Strukturen genügt es nicht mehr, die Maske allein zu korrigieren. Optimiert werden Maske und Beleuchtung gemeinsam (source mask optimization, SMO): Auch die Winkelverteilung des einfallenden Lichts wird auf das jeweilige Muster zugeschnitten. Die weitestgehende Variante ist die inverse Lithografie (inverse lithography technology, ILT). Dabei wird nicht mehr eine gezeichnete Maske korrigiert, sondern aus dem gewünschten Ergebnis auf dem Wafer rückwärts diejenige Maskenstruktur berechnet, die dieses Ergebnis erzeugt. Heraus kommen frei geformte, oft gekrümmte Konturen, die mit einer gezeichneten Vorlage keine Ähnlichkeit mehr haben. Die Rechenzeit dafür übersteigt inzwischen die Schreibzeit der Maske deutlich und wird auf großen Rechnerclustern erbracht.

Weil eine einzige Maske viele tausend Wafer belichtet, überträgt sich jeder Fehler auf ihr auf jeden Chip. Masken werden deshalb nach dem Schreiben mit hochauflösenden Verfahren geprüft und einzelne Defekte gezielt repariert – etwa durch Abtragen überschüssigen Absorbermaterials mit einem fokussierten Elektronen- oder Ionenstrahl.

Maskenherstellung

Grundsätzlich werden die Strukturen auf den Masken auf die gleiche Weise hergestellt, wie auf den Wafern. Im Gegensatz zum Belichtungsprozess in der Waferfertigung, bei dem die Strukturen der Maske durch Schattenwurf auf dem Lack abgebildet werden, werden die Masken aber mit einem Elektronenstrahl direkt beschrieben.

1. Belichten eines fotoempfindlichen Lacks zur Strukturierung einer Chromschicht auf dem Glassubstrat mittels Laser oder Elektronenstrahl

2. Entwickeln der Lackschicht

3. Ätzen der Chromschicht

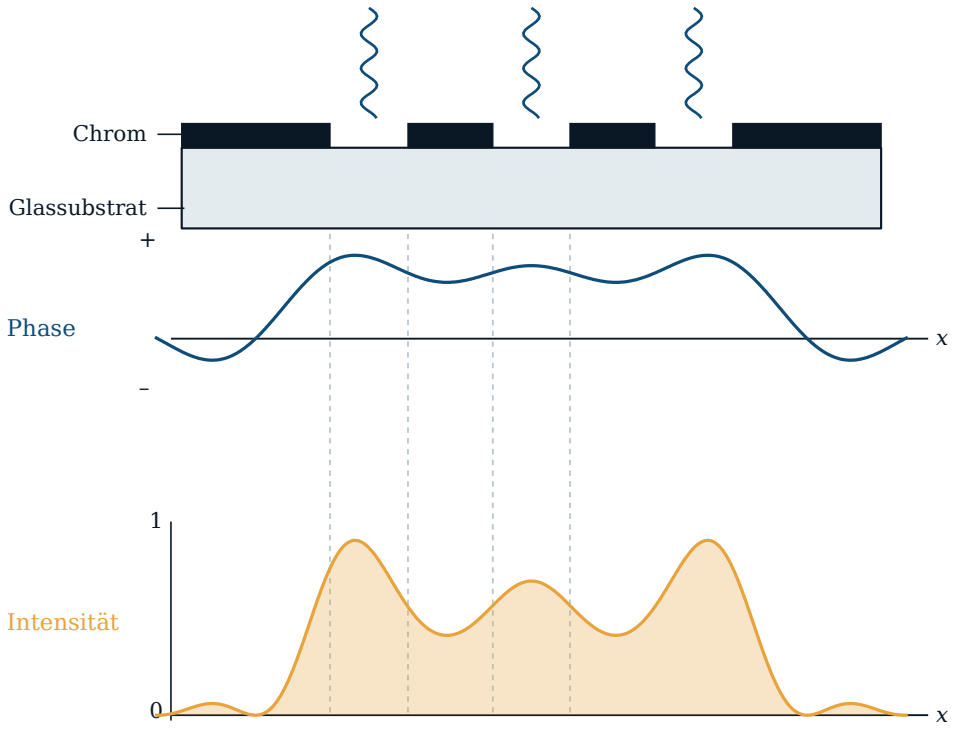
4. Lackentfernen

5. Anbringen des Pellicles

Maskentypen

Neben der klassischen Chrommaske (COG, Chrome On Glass) gibt es noch weitere Maskentypen, die eine verbesserte Strukturauflösung ermöglichen. Hauptproblem der COG-Maske ist die Beugung, die das Licht an den Strukturkanten erfährt. Dadurch fällt das Licht nicht nur senkrecht auf den Wafer, sondern wird auch in den Schattenbereich abgelenkt, wo der Fotolack nicht belichtet werden soll.

Intensitätsprofil einer Chrom On Glass-Mask

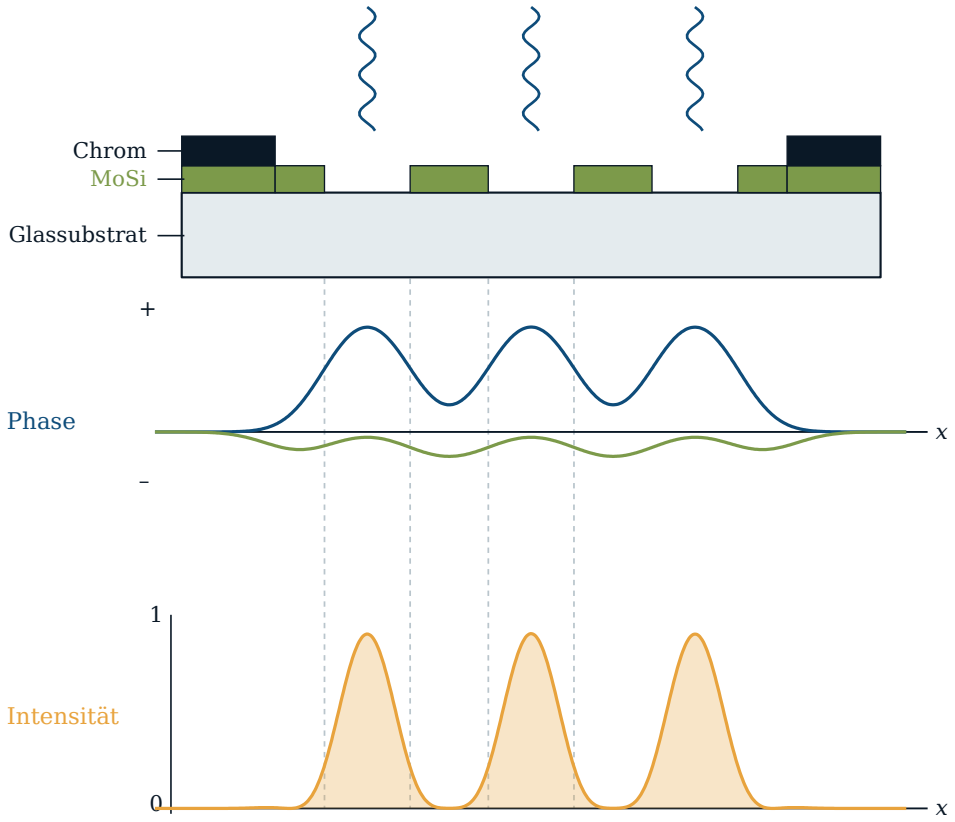


Mit verschiedenen Maßnahmen versucht man nun, die Intensität des gebeugten Lichts zu vermindern. Diese werden im Folgenden anhand der unterschiedlichen Maskentypen näher beschrieben.

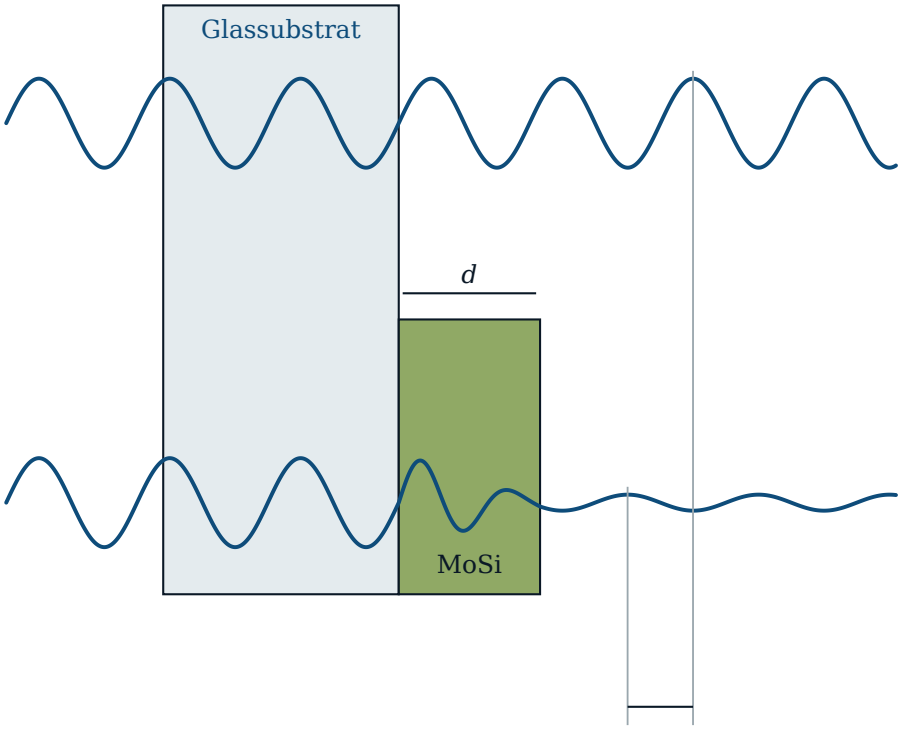
Attenuated Phase Shift Mask (AttPSM)

Bei der so genannten Halbtonmaske oder weichen Phasenmaske bildet eine Schicht aus Molybdänsilicid (MoSi) den strukturgebenden Teil, eine Chromschicht gibt es hier nicht. Die Dicke der MoSi-Schicht ist so gewählt, dass das Licht beim Durchgang eine Phasenverschiebung um 180° erfährt gegenüber dem Licht, das lediglich Glas durchläuft. Dies geschieht durch die unterschiedliche Lichtgeschwindigkeit in Luft und in MoSi. Gleichzeitig ist die Schicht je nach Molybdänanteil im Silicium zu 6 oder 18 % lichtdurchlässig (bei einer Belichtungswellenlänge von 193 nm), das Licht wird also abgeschwächt (attenuate, engl.: vermindern). Die gegenläufigen Lichtwellen löschen sich so unterhalb der MoSi-Strukturen nahezu aus, somit wird der Kontrast zwischen Hell und Dunkel erhöht. Zusätzlich kann in Bereichen, die nicht zur Belichtung benötigt werden, Chrom aufgebracht werden, um Licht vollständig auszublenden. Diese Masken werden als Tritonemaske bezeichnet.

Intensitätsprofil einer Attenuated Phase Shift Mask



Prinzip der Phasenschiebung mittels Molybdänsilicid



Chromfreie Phasenschiebermaske

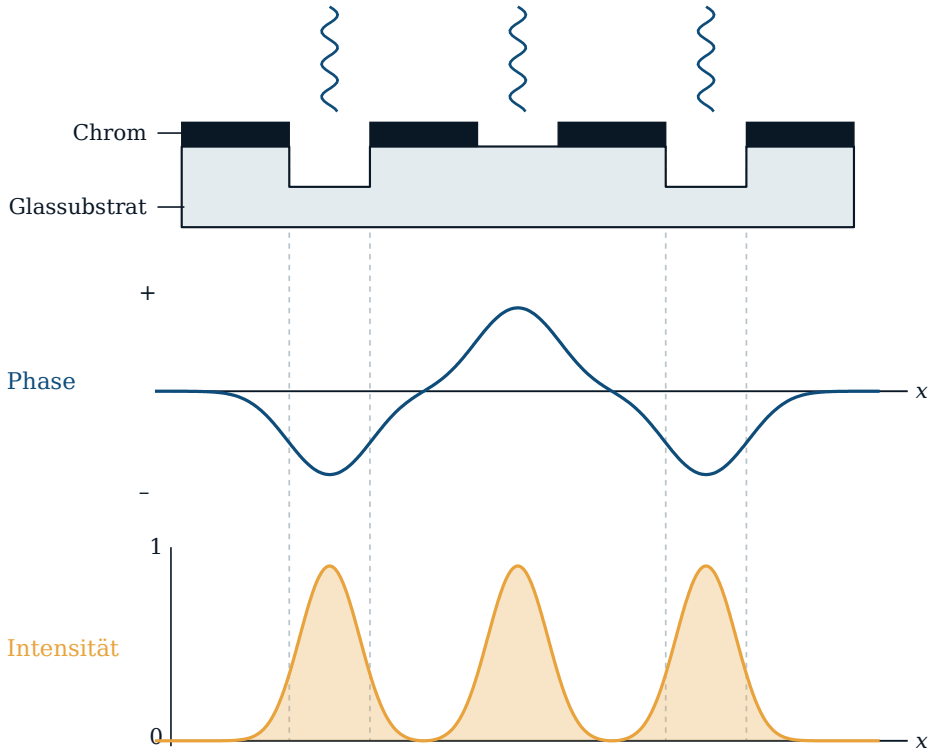
Die chromfreien Masken besitzen keine strukturgebende Beschichtung. Die Phasenverschiebung wird durch Gräben erzeugt, die direkt in die Glasplatte geätzt sind. Die Herstellung dieser Masken gestaltet sich schwierig, da der Ätzvorgang mitten im Glas gestoppt werden muss. Im Gegensatz zu Ätzprozessen, bei denen eine Schicht vollständig durchgeätzt wird und Änderungen im Plasma Auskunft darüber geben, wann die darunterliegende Schicht freigelegt ist, erhält man hier keine Information, wann die benötigte Tiefe erreicht ist.

Zusätzlich ergeben sich Probleme bei der Herstellung großer Strukturen. Dadurch, dass es keine abschattenden Bereiche gibt und das Licht überall mit derselben Intensität auf die Maske trifft, erfolgt nur an den Strukturkanten die gewünschte Interferenz und die daraus resultierende Abschwächung des Lichts. In der Mitte der Bereiche findet jedoch keine oder nur noch eine geringe Abschwächung statt. Um eine Belichtung dieser Bereiche zu verhindern, muss die Lichtintensität also von vorn herein reduziert werden, wodurch die Gefahr einer Unterbelichtung aller Strukturen entsteht.

Alternating Phase Shift Mask (AltPSM)

Bei der alternierenden Phasenmaske werden wie bei der chromfreien Maske Gräben direkt in das Glassubstrat geätzt, jedoch abwechselnd (alternierend) mit ungeätzten Bereichen. Zusätzlich werden Stellen mit Chrom beschichtet um die Lichtintensität an entsprechenden Stellen herabzusetzen.

Intensitätsprofil einer Alternating Phase Shift Mask



Dadurch ergeben sich jedoch Bereiche mit undefinierter Phasenverschiebung, so dass bei diesem Maskentyp, der eine sehr hohe Auflösung ermöglicht, grundsätzlich zweimal belichtet werden muss. Die erste Maske enthält dabei die Strukturen die in x-Richtung verlaufen, die zweite Maske die Struktur in y-Richtung.

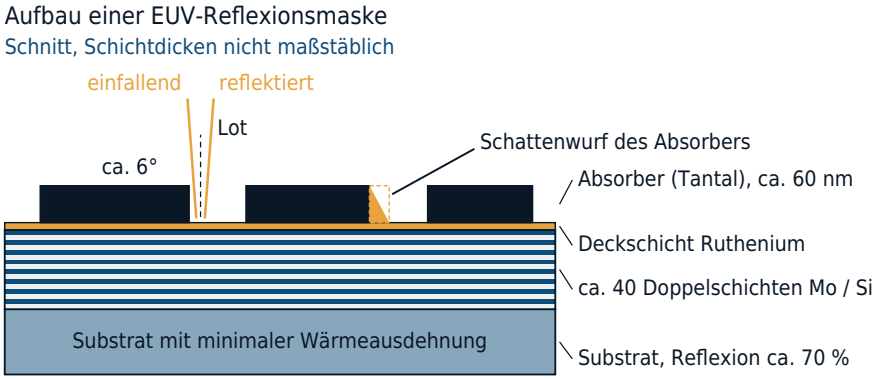
Zielstruktur auf dem Wafer und entsprechende Maskenstruktur

Reflexionsmasken für die EUV-Belichtung

Alle bisher beschriebenen Maskentypen sind Transparente: Licht tritt durch das Glassubstrat hindurch. Für EUV mit 13,5 nm funktioniert das nicht, weil jedes Material diese Strahlung absorbiert. Die EUV-Maske ist deshalb ein Spiegel.

Als Substrat dient ein Glas mit nahezu verschwindender Wärmeausdehnung, da sich die Maske unter der Bestrahlung sonst verzieht. Darauf liegen etwa vierzig Doppelschichten aus Molybdän und Silicium, deren Dicken so gewählt sind, dass sich die an den einzelnen Grenzflächen reflektierten Anteile konstruktiv überlagern. Selbst dieser Vielschichtspiegel reflektiert nur rund 70 % der auftreffenden Strahlung; da im Strahlengang mehrere solche Spiegel liegen, erreicht nur ein kleiner Teil der erzeugten Leistung den Wafer. Eine dünne Deckschicht aus Ruthenium schützt den Stapel, darüber bildet ein Absorber, meist eine Tantalverbindung, das eigentliche Muster.

Aufbau einer EUV-Reflexionsmaske



Aus dem Spiegelprinzip folgt eine Eigenheit, die es bei Transparentmasken nicht gibt: Der Strahl muss schräg einfallen, damit einfallendes und reflektiertes Licht getrennt werden können. Bei diesem Einfall von einigen Grad wirft der etwa 60 nm dicke Absorber einen Schatten, und zwar je nach Orientierung der Struktur einen unterschiedlichen. Diese Maskeneffekte müssen bei der Korrektur der Maske mitgerechnet werden.

Auch der Schutz vor Partikeln ist schwieriger. Ein Pellicle muss hier von der Strahlung zweimal durchlaufen werden und darf sie dabei kaum absorbieren; gleichzeitig muss die Membran der Wärmelast einer Hochleistungsquelle standhalten. Verwendet werden dafür extrem dünne Membranen, unter anderem auf Basis von Kohlenstoffnanoröhren.

Phasenschiebende EUV-Masken, die den Kontrast wie bei 193 nm zusätzlich erhöhen würden, sind Gegenstand der Entwicklung, in der Fertigung aber noch

nicht etabliert. Die Auflösung wird bei EUV daher vor allem über die numerische Apertur, die Beleuchtung und den Lack gewonnen, nicht über die Maskentechnik.

Belichtungsverfahren der nächsten Generation

Der Wechsel, den dieses Kapitel jahrelang als bevorstehend beschrieben hat, ist vollzogen: Die EUV-Lithografie mit 13,5 nm ist seit 2019 in der Serienfertigung und belichtet die kritischen Ebenen aller führenden Logik- und Speicherprozesse. Sie arbeitet, wie vorhergesagt, im Vakuum, bildet über Spiegel ab und nutzt Masken mit spiegelnder statt transparenter Oberfläche.

Gleichzeitig ist die 193-nm-Immersionsbelichtung nicht verschwunden. Sie belichtet weiterhin die Mehrzahl der Ebenen eines modernen Bauelements, weil sie schneller und je Belichtung deutlich günstiger ist. EUV wird nur dort eingesetzt, wo es sich rechnet: Wo eine EUV-Belichtung zwei oder mehr Immersionsschritte samt Zwischenprozessen ersetzt, ist sie im Vorteil.

High-NA-EUV

Die aktuelle Ausbaustufe erhöht die numerische Apertur von 0,33 auf 0,55 und erreicht damit rund 8 nm Halbperiode in einer Belichtung. Erkauft wird das mit einer anamorphen Optik, die in den beiden Achsen unterschiedlich stark verkleinert und damit nur noch das halbe Feld belichtet, sowie mit erheblichem Aufwand: Eine Anlage kostet etwa das Doppelte einer herkömmlichen EUV-Anlage und braucht am Standort Monate bis zur Qualifikation. Erste Ebenen eines Serienprodukts wurden im Juli 2026 freigegeben, der breite Einsatz ist für 2027 und 2028 zu erwarten. Nicht alle Hersteller gehen diesen Weg – teils wird die 0,33er-Generation mit Mehrfachbelichtung als wirtschaftlicher eingeschätzt.

Wo die Grenze inzwischen liegt

Die optische Auflösung ist nicht mehr der eigentliche Engpass. Begrenzend wirken:

- Statistik: die stochastischen Defekte durch die geringe Photonenzahl pro Struktur (siehe Fotolack)
- Überlagerung: der zulässige Fehler bei der Justierung mehrerer Ebenen zueinander liegt im Bereich weniger Nanometer und wird bei geteilten Feldern und Mehrfachbelichtung weiter aufgezehrt
- Lichtleistung: der Durchsatz einer EUV-Anlage hängt direkt an der Leistung der Zinn-Plasmaquelle und an der nötigen Belichtungsdosis
- Kosten: Anlagen und Masken sind so teuer, dass sich feinste Strukturen nur bei sehr hohen Stückzahlen rechnen

Alternative Ansätze

Neben der weiteren Steigerung der Apertur werden Verfahren verfolgt, die ohne abbildende Optik arbeiten:

Nanoimprint
: Eine Vorlage mit dem Negativ der Struktur wird in eine flüssige Lackschicht gepresst, die anschließend aushärtet. Ohne Beugungsgrenze und ohne teure Optik, dafür mit den Problemen jedes Kontaktverfahrens: Defekte und Verschleiß der Vorlage. Im Einsatz für Speicherbauelemente und optische Komponenten.

Direktschreiben mit Elektronen
: Vielstrahlsysteme schreiben die Struktur maskenlos direkt auf den Wafer. Für Prototypen, Kleinserien und kundenspezifische Bausteine attraktiv, für die Massenfertigung nach wie vor zu langsam.

Gerichtete Selbstorganisation
: Blockcopolymere ordnen sich innerhalb einer lithografisch grob vorgegebenen Führungsstruktur selbst zu regelmäßigen feinen Mustern. Geeignet für streng periodische Strukturen wie Kontaktlochraster, nicht für beliebige Layouts.

Keiner dieser Ansätze ersetzt die optische Projektion in der Logikfertigung. Die Entwicklung der vergangenen zwanzig Jahre hat vielmehr gezeigt, dass ein etabliertes Belichtungsverfahren durch Maskentechnik, Beleuchtung, Rechenkorrektur und Mehrfachbelichtung sehr viel weiter getragen werden kann, als es die Wellenlänge zunächst erwarten lässt.

Optical Proximity Correction (OPC)

Einführung & Motivation

Wenn die Maske nicht mehr das druckt, was sie zeigt

Eine Fotomaske enthält im Idealfall exakt die Geometrie, die später auf dem Wafer erscheinen soll: ein Rechteck auf der Maske ergibt ein Rechteck im Resist. Dieses Ideal gilt jedoch nur, solange die kleinsten Strukturen auf der Maske deutlich größer sind als die Wellenlänge des verwendeten Lichts. Sobald Strukturgröße und Wellenlänge in dieselbe Größenordnung geraten, dominieren Beugungseffekte das Abbildungsverhalten – Kanten werden nicht mehr scharf abgebildet, sondern durch das optische System systematisch verzerrt.

Dieses Verhältnis wird durch den sogenannten k1-Faktor beschrieben, der die erreichbare Auflösung ins Verhältnis zu Wellenlänge und Apertur des

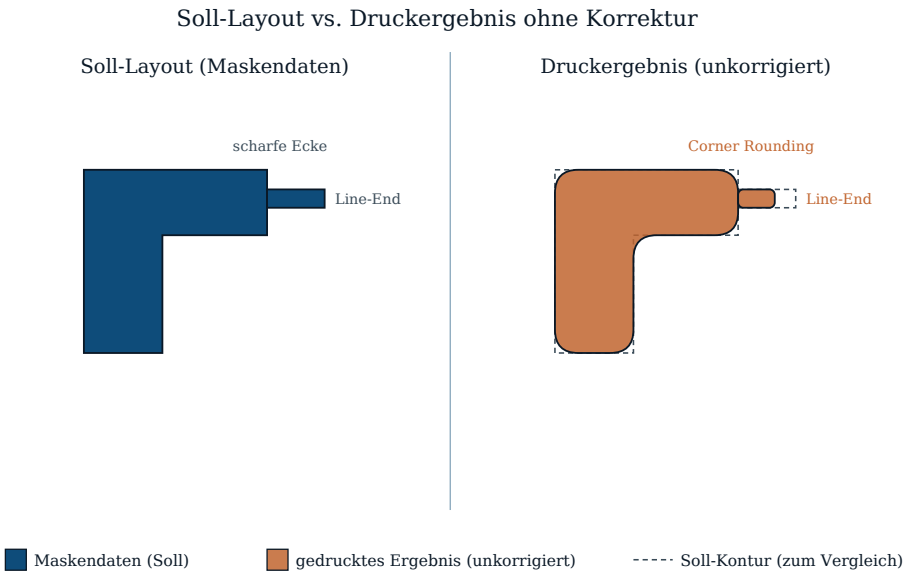
Belichtungssystems setzt. Moderne Prozesse arbeiten mit k1-Werten deutlich unterhalb dessen, was klassische, unkorrigierte Optik noch scharf abbilden kann – die Wellenlänge der Belichtung (in vielen Nodes weiterhin 193 nm) ist damit größer als die zu druckende Strukturbreite selbst. Bei einem derart „unterauflösenden" Prozess reicht eine reine 1:1-Übertragung der gewünschten Geometrie nicht mehr aus.

Optical Proximity Correction: die Maske vorverzerren

Optical Proximity Correction (OPC) löst dieses Problem, indem sie das Problem umkehrt: Statt zu versuchen, die Optik perfekter zu machen, wird die Maskengeometrie so vorverzerrt, dass die bekannten, vorhersagbaren Abbildungsfehler des optischen Systems das gewünschte Ergebnis auf dem Wafer erzeugen. Eine Ecke, die ohne Korrektur abgerundet würde, wird auf der Maske bereits überzeichnet, damit sie nach der Abbildung wieder scharf erscheint.

OPC ist damit kein nachträglicher Reparaturschritt, sondern fester Bestandteil des Maskendatenflusses: Jedes moderne Layout durchläuft vor der eigentlichen Maskenherstellung eine automatisierte OPC-Software, die auf Basis eines kalibrierten Modells des Belichtungsprozesses die Geometrie chipweit anpasst. Ohne diesen Schritt wären die meisten heutigen Strukturgrößen mit der verfügbaren Belichtungswellenlänge schlicht nicht fertigbar.

Das folgende Diagramm zeigt das Grundproblem an einem einfachen Beispiel: links das gewünschte Layout, rechts das Ergebnis, das ohne Korrektur tatsächlich gedruckt würde.



Typische Druckfehler ohne Korrektur

Corner Rounding und Line-End Shortening

Die im vorigen Abschnitt gezeigten Effekte sind die unmittelbarste Folge der Beugung: Scharfe, rechtwinklige Ecken einer Maskenstruktur werden im Resist zu abgerundeten Konturen, weil das optische System hochfrequente Ortsfrequenzanteile – die für scharfe Kanten nötig wären – nicht mehr überträgt. Bei konvexen Ecken „frisst" sich die Rundung in die Struktur hinein, bei konkaven Ecken (etwa in einer inneren Ecke eines L-förmigen Layouts) füllt sie die Kerbe stattdessen auf.

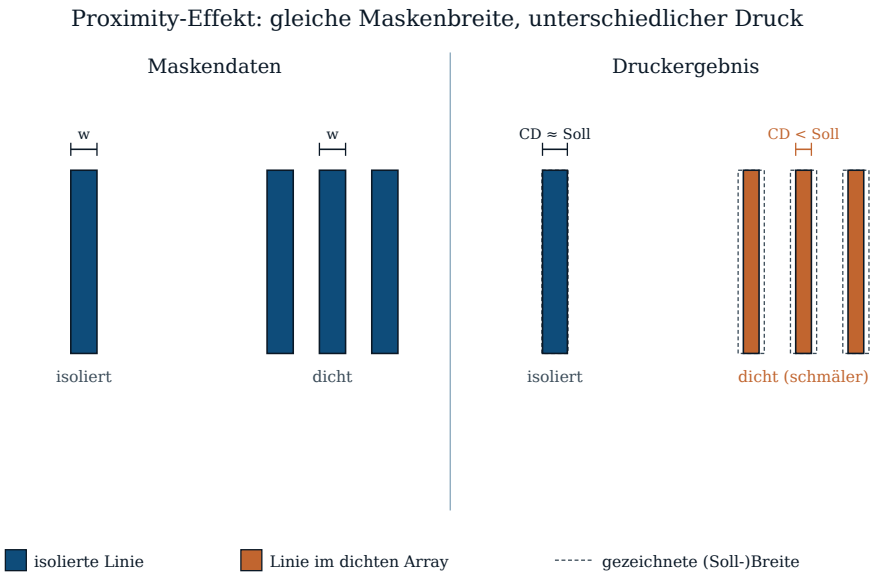
Am Ende einer Leiterbahn tritt ein verwandtes Phänomen auf: das Line-End Shortening. Weil das optische System das Ende einer Linie nicht als scharfe Kante, sondern als allmählich abfallende Intensität abbildet, zieht sich der tatsächlich entwickelte Resist am Leitungsende gegenüber der Maskenvorgabe zurück – eine Leitung, die exakt bis zu einem bestimmten Kontakt reichen sollte, endet im Druck möglicherweise einige Nanometer davor.

Der Proximity-Effekt: gleiche Breite, unterschiedliches Ergebnis

Der namensgebende Effekt von Optical Proximity Correction ist jedoch ein dritter, subtilerer: Zwei Linien mit exakt derselben auf der Maske gezeichneten Breite drucken unterschiedlich breit, je nachdem, wie viele weitere Strukturen sich in ihrer unmittelbaren Nachbarschaft befinden. Eine isoliert stehende Linie erhält ihr Beugungsbild ausschließlich von den eigenen Kanten. Eine Linie inmitten eines dichten Linienarrays dagegen überlagert sich mit den Beugungsmustern ihrer Nachbarn – das resultierende Intensitätsprofil im Resist unterscheidet sich systematisch von dem der isolierten Linie, obwohl die Maskengeometrie identisch ist.

Dieser sogenannte Iso-Dense-Bias ist über weite Bereiche des Pitch-Spektrums (Abstand von Strukturmitte zu Strukturmitte) hinweg unterschiedlich stark ausgeprägt und verläuft nicht linear – er muss für jeden Prozess separat charakterisiert werden, üblicherweise durch systematisches Belichten und Vermessen von Teststrukturen mit variierendem Pitch. Das Ergebnis dieser Charakterisierung fließt direkt in das Korrekturmodell ein, das im weiteren Verlauf dieses Kapitels vorgestellt wird.

Das folgende Diagramm zeigt den Proximity-Effekt an einer isolierten Linie im Vergleich zu einer Linie inmitten eines dichten Arrays – beide mit identischer Maskenbreite, aber unterschiedlichem Druckergebnis.



Mask Bias & Serifs

Mask Bias: die Maßkorrektur

Die einfachste Form der Korrektur setzt direkt am Proximity-Effekt aus dem vorigen Abschnitt an: Ist bekannt, dass eine Linie in einem dichten Array systematisch schmaler druckt als vorgesehen, wird sie auf der Maske einfach von vornherein etwas breiter gezeichnet – der sogenannte Mask Bias. Da der Betrag dieser Abweichung vom lokalen Pitch abhängt, ist der Bias keine feste Konstante, sondern wird anhand der zuvor charakterisierten Iso-Dense-Kurve für jede Umgebungssituation im Layout separat bestimmt: dicht gepackte Bereiche erhalten einen anderen Bias-Wert als isolierte Linien oder mittlere Pitch-Bereiche.

Mask Bias korrigiert damit ausschließlich die Linienbreite als Ganzes – er verändert nicht die Form einer Ecke oder eines Leitungsendes. Für diese lokalen, geometrisch begrenzten Fehler braucht es einen zweiten, gezielteren Korrekturmechanismus.

Serifs und Hammerheads: gezielte Ecken-Verstärkung

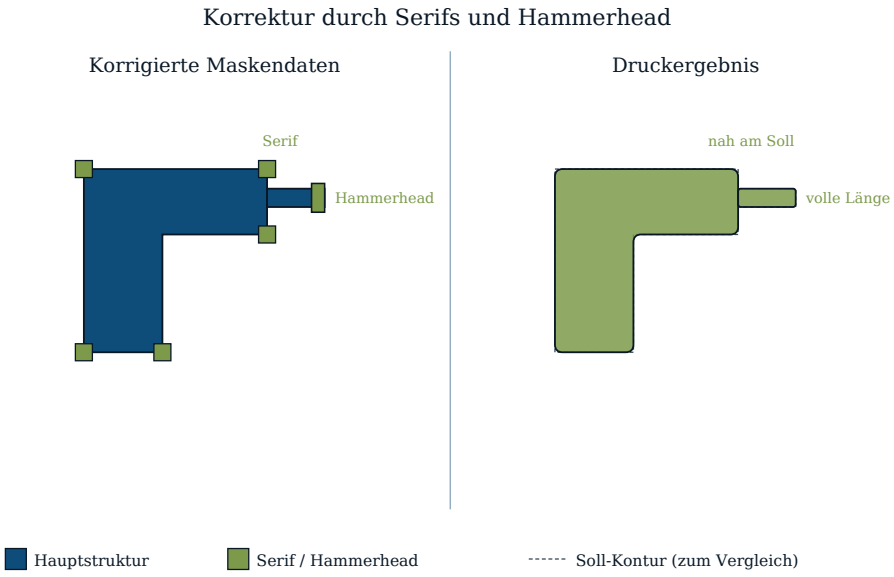
Um Corner Rounding entgegenzuwirken, wird an besonders empfindlichen Ecken – vor allem an konvexen Außenecken – eine kleine zusätzliche quadratische Struktur direkt auf der Maske ergänzt, ein sogenannter Serif. Der Serif selbst wird ebenfalls unterhalb der Auflösungsgrenze bleiben und nicht als eigenständige Struktur gedruckt; er verändert jedoch lokal die Lichtintensität am Rand der Hauptstruktur so, dass die Ecke nach der Abbildung schärfer

erscheint, als sie es ohne diese Verstärkung täte.

Am Ende einer Leitung übernimmt eine vergrößerte Variante desselben Prinzips die Korrektur des Line-End Shortening: ein sogenannter Hammerhead – eine lokale Verbreiterung des Leitungsendes senkrecht zur Leitungsrichtung – kompensiert den systematischen Rückzug des Resists am Leitungsende, sodass die Leitung auch nach der optischen Abbildung noch bis zur vorgesehenen Position reicht.

Beide Techniken – Serifs und Hammerheads – lassen sich als einfache geometrische Regeln formulieren: „Füge an jeder konvexen Ecke mit Innenwinkel kleiner als X ein Quadrat der Größe Y hinzu" oder „Verbreitere jedes Leitungsende um Z nm". Diese regelbasierte Herangehensweise ist die einfachste Form von OPC und wird im übernächsten Abschnitt der modellbasierten Variante gegenübergestellt.

Das folgende Diagramm zeigt die Layoutgeometrie aus Abschnitt 1 nach Ergänzung von Serifs und Hammerhead, sowie das dadurch deutlich verbesserte Druckergebnis.



Sub-Resolution Assist Features

Das Problem: isolierte Linien verhalten sich anders

Serifs und Mask Bias korrigieren die Form und Breite einer Struktur direkt an ihrer eigenen Kontur. Für ein grundlegenderes Problem reicht das jedoch nicht aus: Eine isolierte Linie erhält ihr Beugungsbild – wie in Abschnitt 2 beschrieben – ausschließlich von den eigenen Kanten, während eine Linie in einem dichten Array von den Beugungsmustern ihrer Nachbarn profitiert. Diese

unterschiedliche Beugungssituation wirkt sich nicht nur auf die gedruckte Breite aus, sondern auch darauf, wie tolerant die jeweilige Struktur gegenüber Fokus- und Dosisschwankungen im Belichtungsprozess ist – ihr sogenanntes Prozessfenster. Isolierte Strukturen haben typischerweise ein deutlich kleineres Prozessfenster als dicht gepackte, selbst wenn beide nach Korrektur exakt dieselbe Breite drucken.

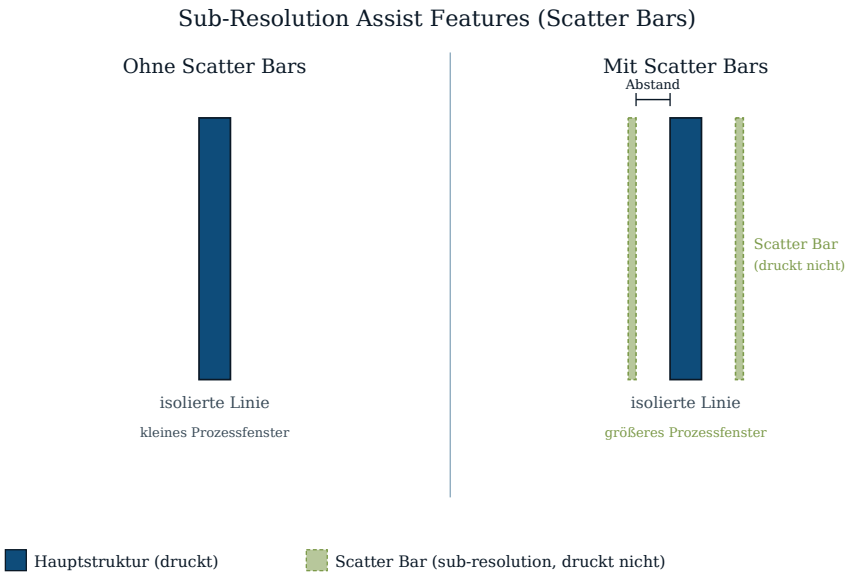
Serif und Bias allein können dieses Defizit nicht beheben, da sie nur die Struktur selbst verändern, nicht aber deren optische Umgebung.

Scatter Bars: Nachbarn, die nicht gedruckt werden

Sub-Resolution Assist Features (SRAFs), umgangssprachlich auch Scatter Bars genannt, lösen dieses Problem, indem sie der isolierten Linie künstlich Nachbarschaft verschaffen: Parallel zur eigentlichen Struktur werden zusätzliche, sehr schmale Linien auf die Maske gezeichnet – schmal genug, dass sie selbst unterhalb der Auflösungsgrenze des Belichtungssystems bleiben und daher nicht als eigenständige Struktur im Resist erscheinen. Sie verändern jedoch das Beugungsbild am Ort der Hauptstruktur so, dass es dem einer dicht gepackten Umgebung ähnlicher wird – die isolierte Linie „sieht" optisch nun fast so aus wie eine Linie im Array, mit einem entsprechend größeren, stabileren Prozessfenster.

Die Platzierung von SRAFs unterliegt strengen geometrischen Regeln: Der Abstand zur Hauptstruktur muss groß genug sein, um Verschmelzung zu vermeiden, aber klein genug, um noch einen Effekt zu haben; die Breite der Scatter Bars selbst muss sicher unterhalb der Druckschwelle bleiben, auch unter ungünstigen Prozessbedingungen. Softwaregestützte SRAF-Platzierung generiert diese Hilfsstrukturen heute automatisch anhand des charakterisierten Beugungsverhaltens des jeweiligen Prozesses, typischerweise als letzter Schritt vor der eigentlichen Serif- und Bias-Korrektur.

Das folgende Diagramm zeigt eine isolierte Linie ohne und mit Scatter Bars – die Hilfsstrukturen selbst erscheinen nicht im gedruckten Resist.



Rule-based vs. Model-based OPC

Regelbasierte OPC: schnell, aber grobmaschig

Die in den vorigen Abschnitten vorgestellten Korrekturen – Mask Bias, Serifs, Scatter Bars – lassen sich als Nachschlagetabelle organisieren: Für eine gegebene Linienbreite und einen gegebenen Abstand zur nächsten Nachbarstruktur liefert die Tabelle einen vorab charakterisierten Korrekturwert. Diese regelbasierte OPC ist rechnerisch günstig, da für jede Kante nur ein einziger Tabellen-Lookup nötig ist, keine aufwendige Simulation.

Die Grenzen dieses Ansatzes zeigen sich bei komplexer, zweidimensionaler Geometrie: In der Nähe einer Ecke, eines Line-Ends oder einer ungewöhnlichen Nachbarschaftskonstellation überlagern sich mehrere Proximity-Effekte gleichzeitig, für die keine einzelne Tabellenzeile mehr eine passende Antwort liefert. Regelbasierte OPC behandelt eine Kante dabei stets als Ganzes: Der gesamte gerade Kantenabschnitt wird um denselben Betrag verschoben, unabhängig davon, ob sich in der Nähe des einen Endes eine zusätzliche Struktur befindet, die dort einen anderen Korrekturbedarf erzeugt als am anderen Ende.

Modellbasierte OPC: Segment für Segment simuliert

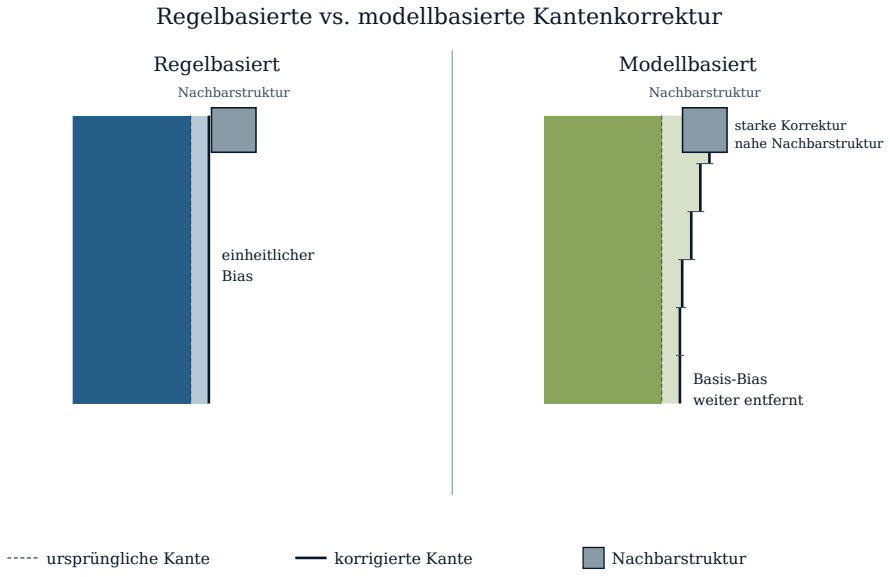
Modellbasierte OPC geht einen grundlegend anderen Weg. Statt Tabellenwerte nachzuschlagen, nutzt sie ein kalibriertes numerisches Modell des gesamten Abbildungsprozesses – optisches Modell und Resist-Modell kombiniert – um für eine gegebene Maskengeometrie den tatsächlich zu erwartenden gedruckten

Konturverlauf vorherzusagen. Jede Kante eines Polygons wird dafür in viele kurze Segmente unterteilt, die sich anschließend unabhängig voneinander verschieben lassen.

Der Korrekturalgorithmus arbeitet iterativ: Er simuliert den Druck der aktuellen Maskengeometrie, vergleicht das Ergebnis mit der Ziel-Kontur, verschiebt jedes Segment einzeln in Richtung einer besseren Übereinstimmung, und wiederholt diesen Zyklus, bis die simulierte Kontur innerhalb einer festgelegten Toleranz mit dem Soll übereinstimmt oder eine maximale Iterationszahl erreicht ist. Weil jedes Segment einzeln reagiert, kann die Korrektur in der Nähe einer komplexen 2D-Situation – etwa nahe einer Ecke – deutlich stärker ausfallen als weiter entfernt auf derselben Kante, wo einfache regelbasierte Korrektur nur einen einzigen, gemittelten Wert kennen würde.

Dieser Detailgrad hat seinen Preis: Modellbasierte OPC erfordert für jede Iteration eine Lithografiesimulation an tausenden Segmenten gleichzeitig, chipweit – ein Rechenaufwand, der um Größenordnungen über dem der regelbasierten Variante liegt. In der Praxis kombinieren moderne OPC-Flows daher beide Ansätze: Regelbasierte Korrektur für einfache, gut charakterisierte Situationen, modellbasierte Verfeinerung dort, wo die Geometrie es erfordert.

Das folgende Diagramm vergleicht beide Ansätze an derselben Kante in der Nähe einer benachbarten Struktur: links die einheitliche Verschiebung der regelbasierten Korrektur, rechts die segmentweise, an die lokale Geometrie angepasste Verschiebung der modellbasierten Korrektur.



Verifikation: Lithography Simulation & Process Window

Warum Korrektur allein nicht genügt

Eine angewendete OPC-Korrektur ist zunächst nur eine Behauptung: Das Korrekturmodell sagt voraus, dass die angepasste Maskengeometrie unter den angenommenen Prozessbedingungen das gewünschte Ergebnis druckt. Ob diese Vorhersage über das gesamte, oft milliardenfach wiederholte Chipdesign hinweg tatsächlich zutrifft – und ob sie auch unter realistischen Schwankungen im Fertigungsprozess Bestand hat –, muss anschließend gesondert geprüft werden. Diese OPC-Verifikation läuft technisch ähnlich wie die eigentliche Korrektur: eine vollflächige Lithografiesimulation, diesmal jedoch nicht zur Anpassung der Geometrie, sondern zur Kontrolle des bereits korrigierten Layouts.

Das Prozessfenster: mehr als nur der Idealfall

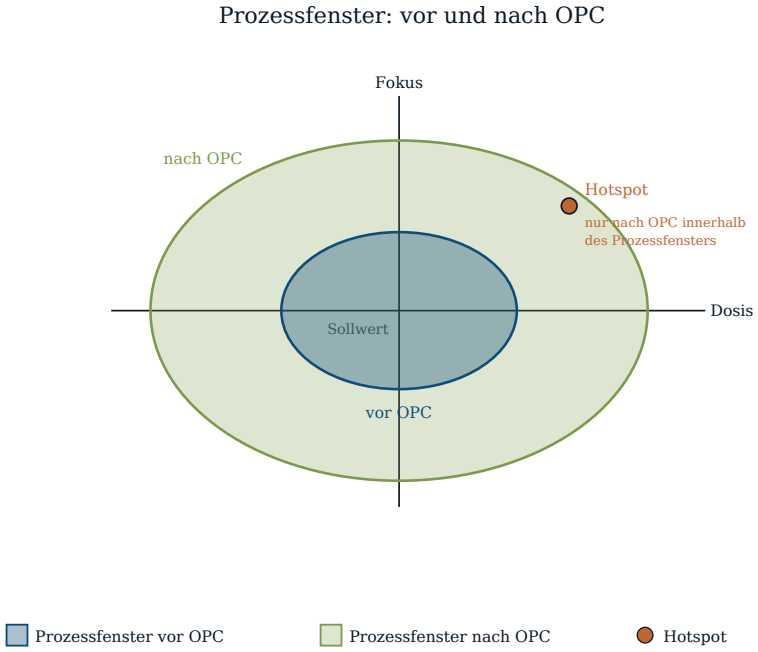
Belichtungsanlagen halten Fokus und Dosis nicht perfekt konstant – von Wafer zu Wafer, über die Waferfläche hinweg und selbst innerhalb eines einzelnen Belichtungsfeldes schwankt beides in einem gewissen Rahmen. Eine Korrektur, die nur beim exakten Sollfokus und der exakten Solldosis funktioniert, wäre in der Praxis wertlos. Die OPC-Verifikation simuliert deshalb nicht nur den nominalen Arbeitspunkt, sondern ein ganzes Prozessfenster aus Fokus- und Dosis-Kombinationen – typischerweise die vier Extremecken plus mehrere Zwischenpunkte – und prüft, ob die Struktur an jedem einzelnen dieser Punkte noch innerhalb der zulässigen Toleranz bleibt.

Gut ausgeführte OPC vergrößert dieses Prozessfenster spürbar gegenüber dem unkorrigierten Layout: Strukturen, die vorher nur in einem engen Fokus-Dosis-Bereich sauber druckten, bleiben nach der Korrektur auch bei größeren Abweichungen noch innerhalb der Spezifikation. Das größere Fenster ist damit ein direktes, messbares Qualitätsmaß für die Korrektur selbst – zusätzlich zur reinen Übereinstimmung am nominalen Arbeitspunkt.

Hotspots: wo die Korrektur nicht ausreicht

Stellen, an denen die simulierte Kontur an einem oder mehreren Punkten des Prozessfensters eine Design Rule verletzt – etwa weil sich zwei benachbarte Leitungen bei ungünstigem Fokus fast berühren (Bridging) oder eine Leitung sich so weit verschmälert, dass sie beinahe unterbrochen wird (Necking) –, werden als Hotspots markiert. Jeder gefundene Hotspot muss vor dem Tape-out behoben werden: durch eine gezielte lokale Nachkorrektur, notfalls durch eine manuelle Layoutänderung an dieser Stelle. Anschließend läuft die Verifikation an der betroffenen Stelle erneut, bis auch der ungünstigste Punkt im Prozessfenster innerhalb der Toleranz bleibt.

Das folgende Diagramm zeigt das Prinzip an einem Prozessfenster-Diagramm: Auf den Achsen liegen Fokus- und Dosisabweichung vom Sollwert, die Fläche markiert die Kombinationen, bei denen die Struktur noch korrekt druckt – vor und nach OPC im Vergleich, mit einem Hotspot, der erst durch die Korrektur innerhalb des Fensters liegt.



Grenzen und Ausblick

Der Rechenaufwand von Full-Chip-OPC

Ein moderner Chip enthält mehrere Milliarden Polygone. Wird jede Kante davon für die modellbasierte Korrektur in Dutzende Segmente unterteilt und jedes Segment über mehrere Iterationen hinweg simuliert, ergibt sich ein Rechenaufwand, der selbst für große Rechenzentren eine ernsthafte Herausforderung darstellt. Full-Chip-OPC-Läufe verteilen sich heute routinemäßig über tausende CPU-Kerne parallel und benötigen trotzdem oft viele Stunden bis einzelne Tage Rechenzeit – pro Maskenebene, und ein moderner Prozess hat davon mehrere Dutzend.

Diese Datenmenge hat auch das Maskendatenformat selbst verändert: Das klassische GDSII-Format, ursprünglich für vergleichsweise einfache, unkorrigierte Layouts entworfen, stößt bei den stark fragmentierten, serifenreichen Geometrien nach OPC an praktische Grenzen. Das OASIS-Format wurde gezielt entwickelt, um die nach der Korrektur stark vergrößerten Datenmengen kompakter zu speichern, unter anderem durch effizientere Kompression sich wiederholender Geometriemuster.

EUV: neue Wellenlänge, neue Effekte

Mit dem Übergang zur EUV-Lithografie (Extreme Ultraviolet, 13,5 nm Wellenlänge) verschiebt sich die Ausgangslage grundlegend: Da die Wellenlänge nun weit unterhalb der zu druckenden Strukturgröße liegt, ist der in Abschnitt 1 beschriebene k1-Faktor für viele Strukturen deutlich entspannter als bei der 193-nm-Immersionslithografie – die klassischen Proximity-Effekte treten in vielen Fällen schwächer auf, und der Korrekturbedarf durch Mask Bias und Serifs sinkt entsprechend.

Dafür bringt EUV eigene, neuartige Effekte mit, die eine eigene Korrekturbehandlung benötigen. Da bei dieser Wellenlänge keine transmittierenden Linsen mehr existieren, arbeitet EUV-Lithografie ausschließlich mit reflektiven Spiegeloptiken, und das Licht trifft die Maske nicht senkrecht, sondern in einem Winkel – dadurch entstehen sogenannte Mask-3D-Effekte, bei denen die endliche Dicke der Maskenabsorberschicht selbst Schatten wirft und die Abbildung asymmetrisch verzerrt, je nach Orientierung einer Struktur zum Einfallswinkel. Hinzu kommt eine geringere Photonenzahl pro Belichtung, die zu statistischen Schwankungen führt – sogenannten stochastischen Effekten –, welche mit klassischer, deterministischer OPC allein nicht vollständig beherrschbar sind.

OPC verschwindet mit EUV also nicht, sondern wandelt sich: Die in diesem Kapitel vorgestellten Grundprinzipien – Mask Bias, Serifs, SRAFs, modellbasierte Segmentkorrektur, Prozessfenster-Verifikation – bleiben gültig, werden jedoch um zusätzliche Modellkomponenten für Mask-3D- und stochastische Effekte erweitert. Bei Strukturen, die selbst mit EUV die Auflösungsgrenze einer einzelnen Belichtung unterschreiten, kommt zudem Multi-Patterning hinzu: Ein Layout wird auf mehrere Masken aufgeteilt, deren OPC-Korrekturen aufeinander abgestimmt werden müssen – ein Thema, das den Rahmen dieses Kapitels sprengt, aber auf denselben hier vorgestellten Grundlagen aufbaut.

Nasschemie

Ätztechnik allgemein

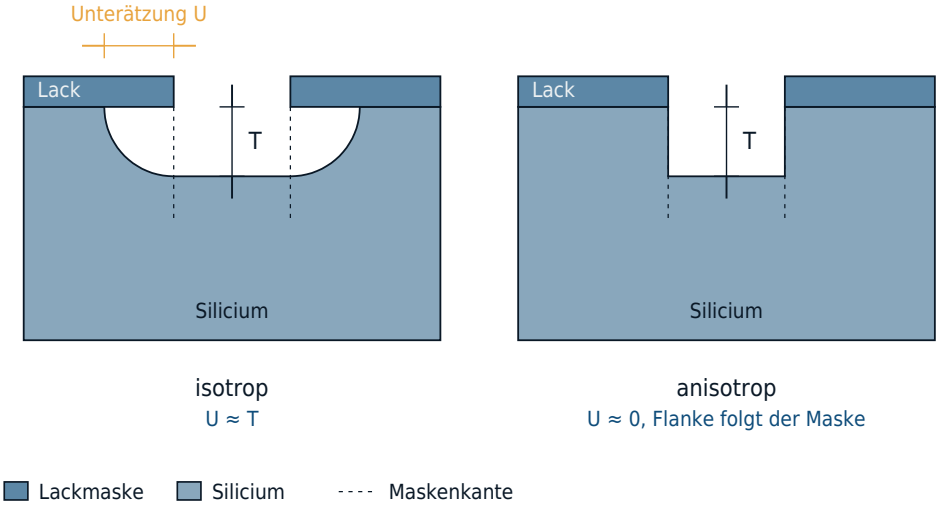
Übersicht

In der Halbleiterfertigung müssen verschiedenste Schichten geätzt werden. Sei es zur ganzflächigen Entfernung oder zum Übertragen eines strukturierten Fotolackfilms in eine darunter liegende Schicht. Dabei lässt sich die Ätztechnik in das nasschemische Ätzen und das Trockenätzen unterteilen. Darüber hinaus unterscheidet man zwischen isotropen und anisotropen Prozessen, sowie dem chemischen und dem physikalischen Charakter des Ätzprozesses.

Bei einem isotropen Ätzprozess geschieht die Ätzung in alle Richtungen. So werden Schichten nicht nur in ihrer Dicke abgetragen, sondern auch in ihrer lateralen Ausdehnung. Bei anisotropen Ätzungen wird die Schicht nur in einer Richtung abgetragen. Je nach Prozess kann ein isotroper oder ein anisotroper Ätzvorgang erwünscht sein.

Isotrope und anisotrope Ätzung

Bei idealer Isotropie ist die Unterätzung etwa so groß wie die Ätztiefe



Nasschemisches Ätzen greift meist ungerichtet an: Die Lösung erreicht die Fläche unter der Maske ebenso wie den Grabenboden – die Struktur wird dadurch breiter als die Maskenöffnung. Anisotrope Verfahren wie RIE ätzen dagegen gerichtet und übertragen die Maskengeometrie nahezu unverändert.

Ein wichtiger Parameter der Ätzprozesse ist die Selektivität. Diese gibt das

Abtragsverhältnis zweier Schichten an. Ist die Selektivität 2:1, so wird die eine Schicht doppelt so schnell durch den Ätzprozess entfernt wie die andere.

Die im Folgenden beschriebene Nasschemie wird jedoch nicht allein zum Ätzen von Schichten verwendet, sondern kommt auch bei anderen Prozessen zum Einsatz:

- Nassätzen: zum ganzflächigen Entfernen von dotierten und undotierten Oxidschichten
- Scheibenreinigung
- Lackentfernen
- Rückseitenbehandlung: Entfernen von Schichten die bei Ofenprozessen auf den Scheibenrückseiten entstanden sind
- Polymerentfernung: Entfernung von Nebenprodukten die beim Plasmaätzen entstehen und sich auf den Scheiben festsetzen
- Kantenentschichtung: gezieltes Entfernen von Schichten am äußersten Rand der Scheibe, wo sie später abplatzen und Partikel erzeugen würden

Auf Grund des meist isotropen Ätzprofils wird das Nassätzen kaum zur Strukturierung verwendet. Eine Ausnahme ist hier die Mikromechanik. Auf Grund der Gitterstruktur von Siliciumeinkristallen können mit nasschemischen Ätzlösungen definierte Strukturen erzeugt werden, welche bspw. Flankenwinkel von 90° oder 54,74° besitzen.

Aus dem Verlust der Strukturierungsaufgabe folgt allerdings nicht, dass die Nasschemie an Bedeutung verloren hätte – im Gegenteil. Je kleiner die Strukturen wurden, desto mehr Prozessschritte kamen hinzu und desto empfindlicher reagiert das Bauelement auf Verunreinigungen. Ein moderner Fertigungsablauf enthält mehr nasschemische Schritte als je zuvor, nur sind es überwiegend Reinigungs-, Abtrags- und Vorbehandlungsschritte statt Strukturierungen. Gleichzeitig hat sich die Anlagentechnik von der Charge zur Einzelscheibenbearbeitung verschoben.

Nassätzen

Prinzip

Das Prinzip beim Nassätzen ist die Umwandlung des festen Materials der Schicht in flüssige Verbindungen unter zu Hilfenahme einer chemischen Lösung. Die Selektivität ist sehr hoch, da die eingesetzten Chemikalien genau auf die vorhandenen Schichten abgestimmt werden können; sie beträgt bei den meisten Lösungen mehr als 100:1.

Der Vorgang läuft in drei Teilschritten ab, die nacheinander durchlaufen werden

müssen:

1. Die reaktiven Teilchen gelangen aus der Lösung an die Oberfläche der zu ätzenden Schicht.
2. An der Oberfläche findet die chemische Reaktion statt; dabei entsteht eine lösliche Verbindung.
3. Das Reaktionsprodukt löst sich von der Oberfläche und wird in die Lösung abtransportiert.

Welcher dieser Schritte der langsamste ist, bestimmt das Verhalten des gesamten Prozesses. Ist die eigentliche Reaktion der Engpass, spricht man von einem reaktionsbegrenzten Prozess: Die Ätzrate hängt dann stark von der Temperatur ab und steigt näherungsweise exponentiell mit ihr an, dafür ätzt der Prozess über die ganze Scheibe gleichmäßig. Ist dagegen der An- und Abtransport der Engpass, ist der Prozess diffusionsbegrenzt. Die Temperaturabhängigkeit ist dann schwächer, dafür wirkt sich jede Ungleichmäßigkeit der Strömung unmittelbar auf den Abtrag aus – an den Rändern der Scheibe wird schneller geätzt als in der Mitte.

Für die Praxis folgt daraus: Reaktionsbegrenzte Prozesse werden über eine sehr genaue Temperierung geführt, diffusionsbegrenzte über die Bewegung der Lösung, also über Umwälzung, Rotation der Scheibe oder Ultraschall. In engen Gräben und Kontaktlöchern wird jeder Prozess diffusionsbegrenzt, weil die Lösung dort kaum ausgetauscht wird. Das ist einer der Gründe, weshalb sich die Nassätzung für feine Strukturen nicht eignet.

Anforderung

Folgende Anforderungen müssen von den chemischen Lösungen erfüllt werden:

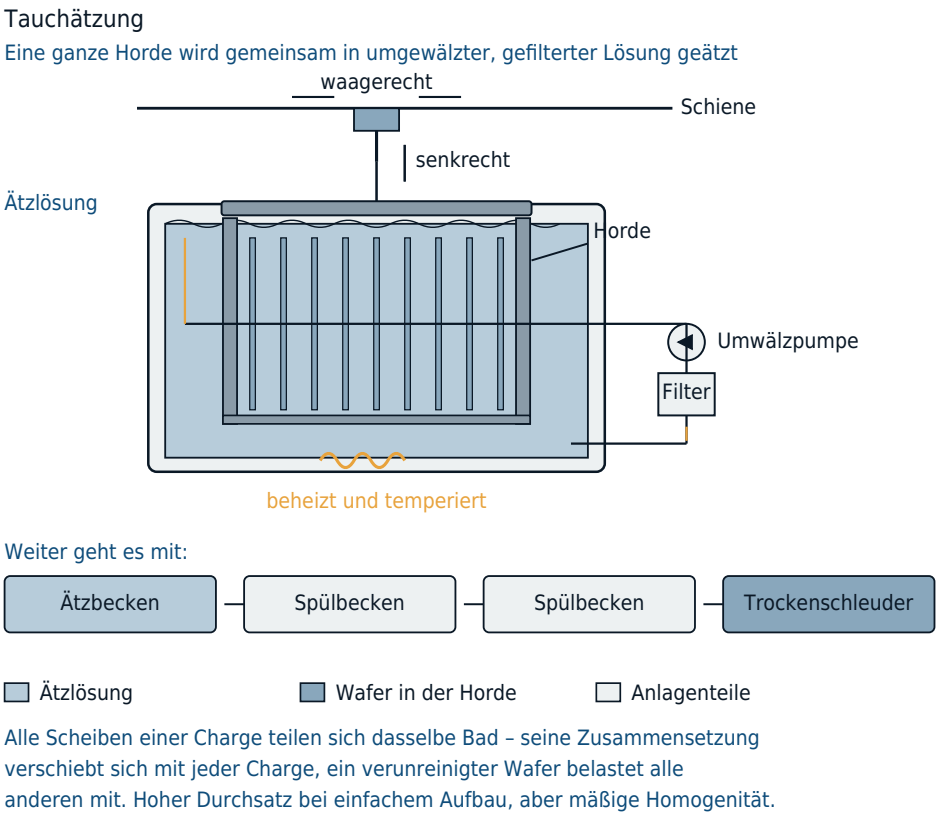
- die Maskierschicht darf nicht angegriffen werden
- die Selektivität muss hoch sein, sowohl gegenüber der Maskierung als auch gegenüber der darunter liegenden Schicht
- die Ätzung muss durch Verdünnung mit Wasser gestoppt werden können
- es dürfen sich keine gasförmigen Reaktionsprodukte bilden, da die Bläschen vereinzelt Bereiche abschatten können
- konstante Ätzrate über lange Zeit
- die Reaktionsprodukte müssen direkt gelöst werden, damit keine Partikel die Ätzlösung verunreinigen
- die Lösung muss die Oberfläche vollständig benetzen; wasserabweisende Flächen und enge Gräben werden sonst nicht erreicht
- keine Einträge von Alkali- oder Schwermetallionen, die als bewegliche Ladungen in das Oxid gelangen und Bauelemente unbrauchbar machen würden
- gute Umweltverträglichkeit und leichte Entsorgung

Die letzte Anforderung ist in der Praxis die schwierigste. Die wirksamsten Ätzlösungen sind zugleich die gefährlichsten, und ein Teil der eingesetzten Chemikalien lässt sich nicht durch harmlosere ersetzen. Aufwand für Abluft, Abwasser und Arbeitsschutz macht daher einen erheblichen Anteil der Betriebskosten einer nasschemischen Anlage aus.

Tauchätzung

Beim Tauchätzverfahren wird eine ganze Horde mit Wafern auf einmal in einem Becken mit der Ätzlösung prozessiert. Durch Filter und Umwälzpumpen werden Partikel von den Wafern ferngehalten. Da die Konzentration der Ätzlösung mit zunehmender Menge an bearbeiteten Scheiben abnimmt, muss sie oft erneuert werden.

Die Ätzrate, also der Materialabtrag pro Zeit, muss genau bekannt sein, um reproduzierbare Ätzungen durchführen zu können. Eine exakte Temperierung der Ätzlösung ist erforderlich, da die Ätzrate der meisten chemischen Lösungen mit der Temperatur zunimmt.



Die Hebevorrichtung kann die Horde waagerecht und senkrecht transportieren. Nachdem die Wafer geätzt wurden wird die Ätzung auf den Wafern durch Spülen in mehreren Tauchbecken gestoppt; anschließend kommen die Wafer in

eine Trockenschleuder.

Die Vorteile der Tauchätzung sind die hohe Durchsatzrate an Wafern und der relativ einfache Aufbau der Ätzanlagen. Die Homogenität des Schichtabtrags ist jedoch gering.

Ein grundsätzliches Problem des Verfahrens ist das Bad selbst: Alle Scheiben einer Charge sehen dieselbe, im Lauf der Zeit veränderliche Lösung. Ihre Zusammensetzung verschiebt sich mit jeder bearbeiteten Charge, gelöstes Material reichert sich an, und ein einzelner verunreinigter Wafer kann die ganze Charge und alle nachfolgenden beeinträchtigen. Die Konzentration wird deshalb laufend gemessen und nachdosiert, und die Standzeit eines Bades ist begrenzt.

Für kritische Schritte ist die Tauchätzung daher weitgehend durch die Einzelscheibenbearbeitung ersetzt worden. Sie hat sich aber dort gehalten, wo lange Ätzzeiten bei hohen Temperaturen anfallen und die Homogenitätsanforderung moderat ist – das klassische Beispiel ist das Entfernen von Siliciumnitrid in heißer Phosphorsäure, das je nach Schichtdicke eine Stunde und länger dauert und einzeln pro Scheibe unwirtschaftlich wäre.

Sprühätzung

Die Sprühtechnik ist vergleichbar mit der Sprühentwicklung in der Fototechnik. Durch die Rotation auf dem Chuck unter stetiger Zugabe frischer Ätzlösung ist die Homogenität sehr gut. Bläschen können sich hier auf Grund der schnellen Drehung nicht bilden, jedoch ist auch hier der Nachteil, dass jeder Wafer einzeln prozessiert werden muss.

Alternativ zu der sequentiellen Prozessierung gibt es auch die Sprühätzung mehrerer Horden gleichzeitig in einer Trommel, bei der die Wafer schnell um die Sprühvorrichtung im Zentrum kreisen. Direkt im Anschluss werden die Scheiben unter Rotation in heißer Stickstoffatmosphäre getrocknet.

Was als Nachteil begann, ist heute der Regelfall: Die Einzelscheibenbearbeitung hat sich für nahezu alle kritischen nasschemischen Schritte durchgesetzt. Der Grund liegt in der Kontrolle. Jede Scheibe sieht ausschließlich frische Chemie, die Prozesszeit ist auf die Sekunde einstellbar, und ein Fehler betrifft immer nur eine Scheibe statt einer ganzen Charge.

Eine moderne Einzelscheibenanlage verfügt über mehrere schwenkbare Düsen, die nacheinander verschiedene Medien auf die rotierende Scheibe geben – Ätzlösung, Spülwasser, Trocknungsmittel – ohne dass die Scheibe den Prozessraum verlässt. Der Verbrauch je Scheibe liegt dabei um Größenordnungen unter dem eines Bades, was die Kosten für Chemie und

Entsorgung erheblich senkt.

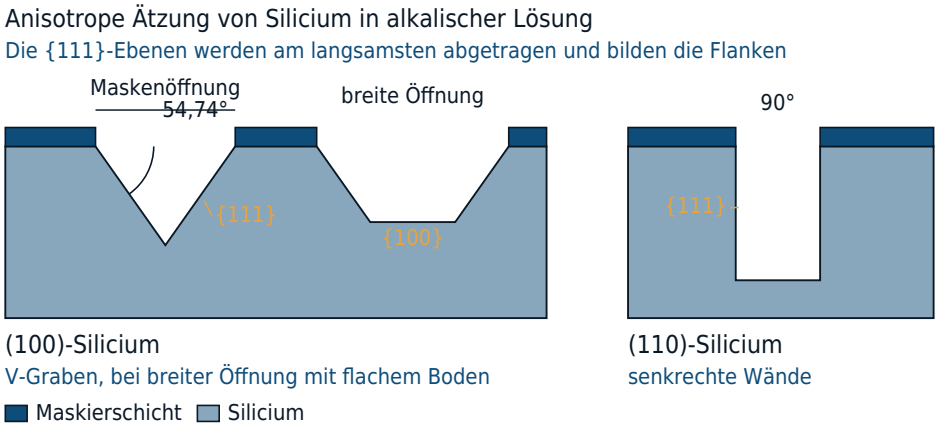
Möglich wird damit auch, was in einem Becken grundsätzlich nicht geht: die räumlich begrenzte Behandlung einer Scheibe. Über eine Düse am Rand lässt sich ausschließlich die äußerste Kante entschichten, über eine Zuführung von unten ausschließlich die Rückseite behandeln, während die Vorderseite durch einen Stickstoffstrom geschützt bleibt.

Anisotrope Siliciumätzung

Obwohl in einer Ätzlösung die Teilchen in der Flüssigkeit die Schicht auf dem Wafer in jeder Richtung abätzen können gibt es auch Verfahren, mit denen man bei der Nassätzung ein nahezu anisotropes Ätzprofil erhält. Dabei nutzt man die unterschiedliche Ätzrate auf den verschiedenen Kristallstrukturen aus.

Ursache ist die Zahl der Bindungen, mit denen ein Siliciumatom in der jeweiligen Ebene im Kristall gehalten wird. In der {111}-Ebene ist die Belegung mit Bindungen am dichtesten, entsprechend langsam wird sie abgetragen. Die {100}- und {110}-Ebenen werden deutlich schneller geätzt – das Verhältnis erreicht in Kalilauge je nach Konzentration und Temperatur mehr als 100:1. Die Ätzung läuft dadurch so lange in die Tiefe, bis sie auf {111}-Ebenen trifft; diese bleiben stehen und bilden die Flanken der Struktur. Die Form des Ergebnisses ist damit nicht durch die Ätzzeit bestimmt, sondern durch die Kristallorientierung der Scheibe.

Ätzprofile in Abhängigkeit der Kristallorientierung



Bei (100)-orientierten Scheiben schneiden die {111}-Ebenen die Oberfläche unter 54,74°, es entstehen V-förmige Gräben und Gruben mit pyramidenförmigem Querschnitt. Ist die Maskenöffnung breit oder die Ätzzeit kurz, bleibt ein flacher Boden stehen und die Struktur ist ein Trapez. Bei (110)-

orientierten Scheiben stehen die {111}-Ebenen senkrecht auf der Oberfläche, so dass sich Gräben mit senkrechten Wänden und sehr großem Tiefen-Breiten-Verhältnis ätzen lassen.

Ätzlösungen

Die Ätzung erfolgt mit Kali-, Natron- oder Lithiumlauge (KOH, NaOH, LiOH), mit Tetramethylammoniumhydroxid (TMAH) oder mit einer EDP-Lösung (ein Gemisch aus Wasser, Pyrazin, Brenzkatechin und Ethylendiamin). Verantwortlich für die Reaktion ist in jedem Fall die OH^--Gruppe (Hydroxidgruppe) der Stoffe:

$$Si + 2OH^- + 2H_2O \rightarrow SiO_2(OH)_2^{2-} + 2H_2$$

Die Wahl der Lösung ist weniger eine Frage der Ätzrate als eine Frage der Verträglichkeit mit der übrigen Fertigung. Kalilauge ätzt am stärksten anisotrop, bringt aber Kaliumionen ein: Alkaliionen sind im Gateoxid beweglich, verschieben die Einsatzspannung von Transistoren und machen ein Bauelement unbrauchbar. In einer Fertigungslinie, in der auch integrierte Schaltungen entstehen, ist Kalilauge deshalb ausgeschlossen. TMAH enthält keine Metallionen und ist mit der CMOS-Fertigung verträglich; die Anisotropie fällt geringer aus, dafür wird Siliciumdioxid kaum angegriffen. EDP wird wegen seiner Giftigkeit heute weitgehend gemieden. TMAH ist als Lösung für die Bearbeitung ebenfalls keineswegs harmlos, sondern über die Haut hochgiftig – in deutlich verdünnter Form, nämlich als 2,38-prozentige Lösung, ist derselbe Stoff der Standardentwickler der Fototechnik.

Begrenzen der Ätztiefe

Da die Ätzung von selbst nur an {111}-Ebenen stoppt, muss die Tiefe auf andere Weise festgelegt werden, wenn eine Membran definierter Dicke entstehen soll. Üblich sind zwei Wege: eine sehr hoch mit Bor dotierte Schicht, die von der Lauge kaum noch angegriffen wird, oder eine elektrochemische Führung, bei der ein p-n-Übergang unter Spannung gesetzt wird und die Ätzung an ihm zum Stehen kommt.

Anwendungen

Für die Strukturierung integrierter Schaltungen spielt die anisotrope Nassätzung keine Rolle. Außerhalb davon ist sie ein Standardverfahren:

- Mikromechanik: Membranen für Drucksensoren, Massen für Beschleunigungssensoren, Düsen für Tintenstrahldruckköpfe
- Solarzellen: eine kurze Behandlung in verdünnter Lauge erzeugt auf (100)-Silicium eine dichte Folge kleiner Pyramiden. Diese Textur lässt einfallendes Licht mehrfach auf die Oberfläche treffen und senkt den Reflexionsgrad

erheblich.
- Optik und Aufbautechnik: V-Gräben zur Aufnahme und Ausrichtung von Glasfasern, Justierstrukturen für die Montage

Ätzlösungen für isotrope Ätzungen

Für die jeweiligen Materialien stehen unterschiedliche Ätzlösungen zur Verfügung. Oxidschichten werden in der Halbleiterindustrie mit Flusssäure HF geätzt:

$SiO_2 + 6HF \rightarrow H_2SiF_6 + 2H_2O$

Die Lösung wird dabei mit NH_4F gepuffert um die HF-Konzentration konstant zu halten. Bei einer Mischung aus einer 40-prozentigen NH_4F-Lösung und 49-prozentigen Flusssäure (Verhältnis 10:1) beträgt die Ätzrate bei thermischem Oxid 50 nm/min. TEOS-Oxide werden mit ca. 150 nm/min und PECVD-Oxide mit ca. 350 nm/min deutlich schneller geätzt. Die Selektivität gegenüber kristallinem Silicium, Siliciumnitrid und Polysilicium ist wesentlich größer als 100:1.

Dass die Ätzrate so stark von der Herkunft des Oxids abhängt, ist dabei kein Nebeneffekt, sondern ein nützliches Werkzeug: Über die Ätzrate in gepufferter Flusssäure lässt sich umgekehrt die Dichte und damit die Qualität einer abgeschiedenen Oxidschicht beurteilen.

Eine kurze Behandlung in stark verdünnter Flusssäure, der HF-Dip, entfernt das natürliche Oxid, das sich auf jeder Siliciumoberfläche an Luft innerhalb kurzer Zeit bildet. Er steht deshalb vor vielen Prozessen, bei denen ein sauberer Kontakt zum Silicium nötig ist, etwa vor der Epitaxie, vor der Gateoxidation oder vor dem Aufbringen von Metall. Die Oberfläche ist danach mit Wasserstoff abgesättigt und damit wasserabweisend – sie perlt sichtbar ab. Da sich sofort wieder neues Oxid bildet, muss der nachfolgende Schritt zügig folgen.

Siliciumnitrid wird mit heißer Phosphorsäure H_3PO_4 bei etwa 160 °C geätzt. Dabei ist die Selektivität zu SiO_2 jedoch mit 10:1 sehr gering. Bei Polysilicium wird die Selektivität zum Nitrid im Wesentlichen vom Gehalt der Phosphorsäure bestimmt. Um die Selektivität gegenüber Oxid zu erhöhen, wird der Lösung gelöstes Silicium zugesetzt, so dass sie bezüglich Siliciumdioxid abgesättigt ist und dieses kaum noch angreift.

Kristallines und polykristallines Silicium, werden zunächst mit Salpetersäure (HNO_3) oxidiert, das Siliciumdioxid wird dann mit Flusssäure geätzt:

1. $3Si + 4HNO_3 \rightarrow 3SiO_2 + 4NO + 2H_2O$
2. $SiO_2 + 6HF \rightarrow H_2SiF_6 + 2H_2O$

Aluminium wird bei ca. 50 °C mit einer Mischung aus Phosphor-, Salpeter- und

Essigsäure geätzt. Die Salpetersäure oxidiert das Aluminium, die Phosphorsäure löst das entstandene Oxid, und die Essigsäure setzt die Oberflächenspannung herab, damit die Lösung die Struktur vollständig benetzt und der entstehende Wasserstoff nicht als Bläschen haften bleibt. Titan wird mit einer Lösung aus Ammoniakwasser NH_4OH, Wasserstoffperoxid H_2O_2 und Wasser (Verhältnis 1:3:5) geätzt. Da die Lösung auch Silicium angreift wenn das Peroxid verbraucht ist, ist die Standzeit der Lösung gering. Dieselbe Mischung dient in stärkerer Verdünnung als Reinigungslösung für die Scheibenreinigung.

Umgang mit Flusssäure

Flusssäure nimmt unter den Prozesschemikalien eine Sonderstellung ein und verdient eine ausdrückliche Warnung. Sie ist nur eine schwache Säure und verursacht auf der Haut zunächst kaum Schmerz, dringt aber tief in das Gewebe ein. Dort bindet das Fluorid das Calcium des Körpers, was zu Verätzungen bis in den Knochen und zu Störungen des Herzrhythmus führen kann. Beschwerden treten oft erst Stunden nach dem Kontakt auf, wenn der Schaden bereits entstanden ist. Der Umgang setzt daher entsprechende Schutzausrüstung, ein bereitliegendes Calciumgluconat-Gel und eine Unterweisung voraus – in der Fertigung wird Flusssäure ausschließlich in geschlossenen Anlagen gehandhabt.

Einordnung

Generell eignet sich das nasschemische Ätzen für das Abtragen kompletter Schichten auf einem Wafer. Die Selektivität von der zu ätzenden Schicht zu der darunter liegenden ist meist sehr gut, so dass keine Gefahr besteht falsche Schichten abzutragen. Zudem ist der Abtrag pro Zeiteinheit sehr hoch, in Tauchätzverfahren lassen sich viele Scheiben gleichzeitig ätzen. Bei kleinen Strukturen kann die Nassätzung jedoch nicht eingesetzt werden, da das isotrope Ätzprofil dafür nicht geeignet ist. In diesem Fall werden die Schichten mit Trockenätzverfahren anisotrop abgetragen.

Bei sehr feinen Strukturen kommt ein zweiter Grund hinzu, der nichts mehr mit dem Ätzprofil zu tun hat: Die Flüssigkeit muss am Ende wieder von der Scheibe herunter, und genau dabei können die freigelegten Strukturen zerstört werden.

Trocknen der Scheiben

Am Ende jedes nasschemischen Schrittes steht das Spülen mit hochreinem Wasser und das Trocknen. Was nach einem Nebenschritt klingt, ist bei feinen Strukturen der kritischste Teil des gesamten Prozesses.

Warum Trocknen ein Problem ist

Solange eine Flüssigkeit zwischen zwei benachbarten Stegen steht, zieht ihre

Oberflächenspannung die Stege zusammen. Die Kraft wächst, je enger die Strukturen stehen und je höher sie sind. Berühren sich zwei Stege einmal, halten sie über die Bindungen ihrer Oberflächen aneinander fest und richten sich nicht wieder auf. Dieser Strukturkollaps – im Englischen als *stiction* bezeichnet, aus *static friction* – ist bei schlanken Stegen der häufigste Ausfallgrund am Ende eines Nassprozesses. Betroffen sind vor allem freistehende Elemente der Mikromechanik und hohe schmale Lackstege der Fototechnik.

Ein Trocknen durch bloßes Abschleudern oder Verdampfen verschärft das Problem, weil die zurückweichende Flüssigkeitsfront dabei genau durch die Struktur hindurchläuft. Die Verfahren zielen deshalb alle darauf, den Übergang von flüssig zu trocken entweder zu entschärfen oder ganz zu vermeiden.

Verfahren

Schleudertrocknen
: Die Scheibe rotiert schnell, die Flüssigkeit wird nach außen abgeschleudert, ein Stickstoffstrom trocknet nach. Einfach und schnell, aber für feine Strukturen ungeeignet und anfällig für zurückbleibende Trockenflecken.

Isopropanol-Trocknung
: Das Spülwasser wird durch Isopropanol verdrängt, dessen Oberflächenspannung nur etwa ein Drittel der von Wasser beträgt. Entsprechend geringer fallen die Kapillarkräfte aus.

Marangoni-Trocknung
: Die Scheibe wird langsam aus dem Wasser gezogen, während an der Wasseroberfläche gezielt Isopropanoldampf zugeführt wird. Der dadurch entstehende Unterschied der Oberflächenspannung entlang der Grenzfläche zieht den Wasserfilm von der Scheibe ab, so dass diese trocken aus dem Bad kommt und keine Flüssigkeit auf ihr verdampfen muss.

Trocknen mit überkritischem Kohlendioxid
: Die Flüssigkeit wird durch Kohlendioxid ersetzt, das anschließend über Druck und Temperatur in den überkritischen Zustand gebracht wird. In diesem Zustand gibt es keine Grenzfläche zwischen flüssig und gasförmig mehr und damit auch keine Oberflächenspannung; das Kohlendioxid kann rückstandsfrei abgelassen werden. Das aufwendigste, aber auch schonendste Verfahren – es kommt dort zum Einsatz, wo alle anderen an ihre Grenze kommen.

Der gleiche Gedanke steht hinter der Entwicklung trockener Alternativen zu einzelnen Nassschritten: Wo eine Schicht mit einem Gas statt mit einer Flüssigkeit entfernt werden kann, entfällt das Problem von vornherein.

Scheibenreinigung

Der Reinraum

Die Halbleiterfertigung findet in Reinräumen statt um die hochkomplexen Schaltungen der Halbleiterbauteile vor Verunreinigungen zu schützen, die die Funktionsfähigkeit der Bauteile beeinflussen können. Die Reinräume werden nach der Größe der Partikel und deren Anzahl pro Kubikfuß (= cbf; 1 Fuß = 30,48 cm) bzw. nach DIN/ISO pro 1 m^3 klassifiziert:

Zulässige Anzahl an Partikeln pro 1m^3

Klasse ↓	≥ 0,1 µm	≥ 0,2 µm	≥ 0,3 µm	≥ 0,5 µm	≥ 1,0 µm	≥ 5,0 µm
ISO 1	10	2				
ISO 2	100	24	10	4		
ISO 3	1.000	237	102	35	8	
ISO 4	10.000	2.370	1.020	352	83	
ISO 5	100.000	23.700	10.200	3.520	832	29
ISO 6	1.000.000	237.000	102.000	35.200	8.320	293
ISO 7	10.000.000	2.370.000	1.020.000	352.000	83.200	2.930

So dürfen sich in einem Klasse-3-Reinraum maximal 1.000 Partikel mit einem Durchmesser ≥0,1 µm befinden, 237 Partikel ≥0,2 µm, 102 Partikel ≥0,3 µm, 35 Partikel ≥0,5 µm und 8 Partikel ≥1 µm. In einem Operationssaal beträgt die Reinraumklasse 2 oder 3, in einem Volumen von 1 m^3 Stadtluft sind 400 Millionen Partikel der Größe 5 µm enthalten.

Die Luft in Reinräumen für die Mikroelektronik wird über Feinstfilter gereinigt und dann durch die Decke eingeleitet. Durch Löcher im Boden wird die Luft abgesaugt, so dass dieser Laminarstrom Partikel von oben nach unten durch den Boden abtransportiert. Damit keine Verunreinigungen von außen in den Reinraum eindringen können herrscht dort in der Regel ein leichter Überdruck. Um die Wafer bestmöglich gegen Partikel zu schützen, werden diese in Transportboxen von einer Produktionsanlage zur nächsten transportiert. In modernen Reinräumen handelt es sich dabei um so genannte FOUPs (Front Opening Unified Pod), welche direkt an die Ladestationen der Maschinen andocken und so geöffnet werden, dass die Wafer vom FOUP in die Anlage gelangen, ohne der Reinraumluft ausgesetzt zu werden.

Da die Wafer in den FOUPs vollständig von der Umgebungsluft im Reinraum abgeschottet sind, kann in den Transportboxen eine andere Reinraumklasse herrschen, als im Reinraum selbst. So ist es möglich, die großen Fertigungsbereiche mit einer relativ hohen Reinraumklasse von ISO 5 zu betreiben, während in den FOUPs eine Reinraumklasse von ISO 1 oder ISO 2

herrscht. Dadurch können enorme Kosten gespart werden, da nur ein kleines Volumen der höchsten Reinheitsklasse genügen muss.

Darstellung eines Klasse 1 Reinraums

(Anklicken für Großansicht, 700 KB · Quelle: unbekannt)

Die Fertigung im Reinraum ist in der Abbildung nur der Bereich zwischen rosafarbener Lüftungsanlage unter dem Dach und dem gelben Bodenbereich. Darunter befindet sich das Basement mit den Versorgungsanlagen (Pumpen, Chemietanks etc.). Manche Reinräume sind von einem so genannten Grauraum umgeben in dem sich die Anlagen befinden, so dass über eine Wand getrennt nur die Bedienfelder und die Schleusen zum Einbringen der Wafer in die Anlagen vom Reinraum aus zugänglich sind.

Das Personal trägt spezielle Reinraumanzüge, die keine Partikel absondern. Dabei bedeckt der Anzug je nach Bedarf den kompletten Körper. Der Kopf wird entweder mit einem simplen Haarnetz oder einer Haube und einem Mundschutz oder mit einer kompletten Maske bedeckt. Zusätzlich gibt es besonders partikelarme Schuhe, Unterbekleidung und Handschuhe. Im Eingangsbereich des Reinraums befinden sich oft Luftduschen, die vor Betreten noch einmal Partikel vom Anzug abblasen.

Arten der Verunreinigung

Trotz des Reinraums gibt es verschiedene Arten von Verunreinigungen, die hauptsächlich vom Personal in der Fertigungslinie, der Umgebungsluft, von Chemikalien (Gasen, Lösungen) und den Anlagen verursacht werden:

- mikroskopische Verunreinigungen: z.B. Partikel aus der Umgebungsluft oder Gasen
- molekulare Verunreinigungen: z.B. Kohlenwasserstoff aus Öl in den Pumpsystemen
- ionische Verunreinigung: z.B. Handschweiß
- atomare Verunreinigung: z.B. Schwermetalle aus Lösungen, Abrieb von Festkörpern

Mikroskopische Verunreinigungen

Mikroskopische Verunreinigungen sind Partikel, die sich an der Waferoberfläche anlagern. Quellen dieser Verunreinigungen sind die Umgebungsluft, Kleidung des Personals, Abriebe an beweglichen Teilen in Prozessanlagen, unzureichend gefilterte Flüssigkeiten, wie Ätz- und Reinigungslösungen oder Entwickler, oder

Ätzrückstände nach Prozessen beim Trockenätzen.

Mikroskopische Verunreinigungen verursachen Abschattungseffekte, wenn sie an der Scheibenoberfläche die Belichtung des Fotolacks verhindern; bei der Kontaktbelichtung kommt es zu schlechten Auflösungen, wenn sich relativ große Partikel zwischen Maske und Scheibe befinden. Des Weiteren können sie bei Implantationsprozessen oder beim Trockenätzen beschleunigte Ionen von der Scheibenoberfläche fernhalten.

Partikel können auch von Schichten eingeschlossen werden, so dass Unebenheiten entstehen. Nachfolgende Schichten können an diesen Stellen aufplatzen oder es sammelt sich Fotolack an, in Folge dessen es durch die zu dicke Lackschicht zu nicht vollständig belichteten Bereichen kommt.

Molekulare Verunreinigungen

Molekulare Verunreinigungen resultieren aus Lack- und Lösemittelresten auf den Scheiben oder durch Ölnebelablagerungen aus Vakuumpumpen. Diese Verunreinigungen können sich sowohl auf der Scheibenoberfläche befinden, als auch in die Schichten eindiffundieren. Die oberflächlich anhaftenden Verunreinigungen behindern teilweise die Haftung nachfolgender Schichten, im Falle von Metallisierungen (dünne Leiterbahnen) erheblich. Durch die Diffusion in Oxide wird die elektrische Belastbarkeit dieser Schichten verschlechtert.

Alkalische und metallische Verunreinigungen

Die Hauptquelle dieser Verunreinigungen ist der Mensch, der ständig über die Haut, aber auch über die Atemluft Salze aussondert. Aber auch durch unzureichend deionisiertes Wasser gelangen (Alkali-)Ionen von Natrium oder Kalium auf die Scheiben. Schwermetalle die in Ätzlösungen vorhanden sind können zur Kontamination führen. Durch Strahlen in Anlagen (Implanter, Trockenätzen) kann Material von Wänden abgesputtert werden, das sich dann auf den Scheiben niederschlägt.

Ionische Verunreinigungen beeinflussen beispielsweise die elektrischen Eigenschaften von MOS-Transistoren, da ihre Ladung die Schwellspannung – also die Spannung ab der der Transistor leitend wird – verändert. Schwermetalle wie Eisen oder Kupfer liefern freie Elektronen, so dass die Leistungsaufnahme in Dioden steigt. Metalle können aber auch Rekombinationszentren für freie Ladungsträger bilden, so dass in der Schaltung

nicht mehr genügend freie Ladungsträger zur einwandfreien Funktion zur Verfügung stehen.

Reinigungstechniken

Die Scheiben werden nach jedem nasschemischen Prozessschritt in Reinstwasser gespült, aber auch nach anderen Prozessen werden die Scheiben von Verunreinigungen befreit. Dabei gibt es verschiedene Reinigungstechniken, die die unterschiedlichen Verschmutzungen beseitigen. Der Verbrauch an Reinstwasser in einer Halbleiterfertigung ist dabei extrem hoch und beträgt mehrere Millionen Liter im Jahr. Dabei befinden sich im Reinstwasser fast keine Verunreinigungen mehr, 1–2 ppm (parts per million = Anzahl an Verunreinigungen pro Millionen Wasserteilchen) sind erlaubt (Beispiel, Stand 2002). Das Wasser wird meist direkt vor Ort in Aufbereitungsanlagen gereinigt.

Eine Möglichkeit der Scheibenreinigung ist das Ultraschallbad, bei dem die Wafer in eine Lösung aus Wasser und Ultraschallreinigungs- und Netzmitteln gegeben werden. Durch die Ultraschallanregung lösen sich Partikel von der Oberfläche, Metalle und molekulare Verunreinigungen werden vom Reinigungsmittel teilweise gebunden. Nicht stark anhaftende Partikel können auch mit Stickstoff abgeblasen werden.

Zum Entfernen organischer Verunreinigungen wie Fett, Öl oder Hautschuppen eignen sich Lösungsmittel wie Aceton oder Ethanol. Diese können jedoch Kohlenstoffrückstände hinterlassen.

Ionische Verunreinigungen (Ionen von Kalium, Natrium etc.) werden durch Spülen mit deionisiertem Wasser entfernt. Auch die Reinigung mit rotierenden Bürsten und einer Reinigungsflüssigkeit ist möglich. Jedoch werden bei strukturierten Scheiben die Partikel an den Kanten angelagert, die Bürsten können die Scheibenoberfläche beschädigen. Kontaktlöcher oder andere Vertiefungen können mit einer Hochdruckreinigung bei ca. 50 bar ausgespült werden. Hierbei werden jedoch keine ionischen oder metallischen Verunreinigungen entfernt.

Die Reinigung mit Wasser und verschiedenen Reinigungsmitteln reicht jedoch oft nicht aus. Oftmals werden die Verunreinigungen mit aggressiven Ätzlösungen entfernt, oder Oberflächen gezielt minimal abgetragen. Gemische aus Wasserstoffperoxid und Schwefelsäure oder Ammoniak oder die Carosche Säure/Piranha-Lösung (Schwefelsäure mit Wasserstoffperoxid) lösen organische Verunreinigungen bei ca. 90 °C durch Oxidation ab.

Ein Gemisch aus Salzsäure und Wasserstoffperoxid bildet Alkalimetalle zu leicht löslichen Chloriden (Salzen) aus, Schwermetalle bilden Komplexe und gehen in Lösung. Mit Flusssäure kann natürliches Oxid entfernt werden, mit einer gezielt

aufgebrachten Schicht Oxid mit Wasserstoffperoxid lässt sich die Oberfläche dagegen auch schützen.

Neben der Spülung der Scheiben nach jedem nasschemischen Prozessschritt, durchlaufen die Wafer bei einer vollständigen Reinigung eine Reihe von Reinigungsschritten nacheinander. Dabei ist die Reihenfolge der Reinigungsverfahren wichtig, da sich diese teilweise gegenseitig in der Reinigungswirkung behindern. Eine Reinigungssequenz kann beispielsweise wie folgt ablaufen:

- Abblasen von Partikeln mit Stickstoff
- Reinigung im Ultraschallbad
- Entfernen von organischen Verunreinigungen mit Caroscher Säure (H_2SO_4–H_2O_2)
- Entfernen feinerer organischer Verunreinigungen mit einer NH_4–H_2O_2-Lösung
- Entfernen von metallischen Verunreinigungen mit Salzsäure und Wasserstoffperoxid
- Trocknen der Scheiben in der Trockenschleuder unter heißer Stickstoffatmosphäre

Nach jedem Reinigungsschritt werden die Scheiben in Reinstwasser gespült, gegebenenfalls natürliches Oxid mit Flusssäure entfernt. Je nach aktueller Scheibenoberfläche sehen die Reinigungssequenzen unterschiedlich aus, da die Lösungen manche Schichten angreifen. Bei den immer kleiner werdenden Strukturen wird auch die Reinigung mehr und mehr erschwert. Nicht nur, dass sich kleine Öffnungen nicht so leicht erreichen lassen. Die Oberflächenspannung und Kapillarwirkungen können Strukturen umkippen lassen und so die Schaltung zerstören.

Trockenätzen

Übersicht

Allgemein

Bei der nasschemischen Ätzung werden Schichten in der Regel isotrop abgetragen, doch gerade bei kleinen Strukturen ist ein anisotropes Ätzprofil wichtig. Dafür eignen sich die Trockenätzverfahren, die eine ausreichende Selektivität bieten. Die Verfahren ermöglichen reproduzierbare homogene Ätzungen der meisten Schichten, die in der Halbleiterfertigung zum Einsatz kommen. Dabei können neben anisotropen Ätzprofilen auch isotrope realisiert werden. Trotz der hohen Anlagenkosten und der sequentiellen Einzelscheibenbearbeitung hat sich das Trockenätzen gegen die nasschemische Ätzung durchgesetzt.

Wichtige Größen beim Trockenätzen

Trockenätzprozesse werden über einige wenige Kenngrößen beschrieben, die zugleich die Zielgrößen jeder Prozessentwicklung sind.

Ätzrate r
: Die Ätzrate beschreibt den Ätzabtrag pro Zeit und wird z.B. in Nanometer pro Minute oder Ångström pro Sekunde angegeben.

$$r=\frac{Ätzabtrag \Delta z}{Ätzzeit \Delta t}$$

Anisotropiefaktor f
: Der Anisotropiefaktor beschreibt das Verhältnis von horizontaler Ätzrate r_h zu vertikaler Ätzrate r_v.

$$f=1-\frac{r_h}{r_v}$$

Zur Strukturübertragung sind stark anisotrope Prozesse erwünscht, also Ätzungen nur in vertikaler Richtung, so dass die Lackmaske nicht unterätzt wird. Für anisotrope Ätzungen ergibt sich $f \to 1$, und dementsprechend für isotrope Prozesse $f \to 0$.

Selektivität S_{jk}
: Die Selektivität S_{jk} zwischen Material j und Material k beschreibt das Verhältnis der Ätzraten zweier Materialien, z.B. von zu strukturierender

Schicht (j) und Lackmaske (k).

$$S_{jk}=\frac{r_j}{r_k}$$

Ob eine hohe oder geringe Selektivität gewünscht ist, hängt vom jeweiligen Prozess ab. Bei der Strukturierung von Schichten sollte der Wert möglichst groß sein, d.h. die zu strukturierende Schicht wird schneller abgetragen als die Lackmaske. Beim Reflowrückätzen muss die Selektivität 1 betragen um Topografien gleichmäßig einzuebnen.

Uniformität *U*
: Die Uniformität beschreibt, wie gleichmäßig der Abtrag ausfällt – über eine Scheibe hinweg, von Scheibe zu Scheibe und von Charge zu Charge. Sie wird als Streuung der Ätzrate bezogen auf deren Mittelwert angegeben.

$$U=\frac{r_{max}-r_{min}}{r_{max}+r_{min}}\cdot 100\,\%$$

Aspektverhältnis *A*
: Das Aspektverhältnis ist das Verhältnis von Ätztiefe *z* zu Strukturbreite *b*. Es beschreibt, wie schlank ein Graben oder Loch ist.

$$A=\frac{z}{b}$$

Mit steigendem Aspektverhältnis sinkt die Ätzrate, weil die Ätzteilchen schlechter an den Strukturboden gelangen und die Reaktionsprodukte schlechter abgeführt werden. Dieser Effekt wird als aspektverhältnisabhängiges Ätzen (aspect ratio dependent etching, ARDE) oder RIE-Lag bezeichnet. Er ist heute die maßgebliche Grenze beim Ätzen tiefer Strukturen, etwa der Speicherlöcher im 3D-NAND-Flash mit Aspektverhältnissen von mehr als 50:1. Typische Profilfehler solcher Ätzungen sind das Bowing (bauchige Aufweitung im oberen Bereich), das Twisting (Verkippen tiefer Löcher gegeneinander) und das Microtrenching (überhöhter Abtrag in den Bodenecken).

Loading-Effekt
: Die Ätzrate hängt von der insgesamt freiliegenden Fläche ab: Eine große zu ätzende Fläche verbraucht mehr Reaktivteilchen, die Ätzrate sinkt. Makroloading betrifft die gesamte Scheibe bzw. Charge, Mikroloading die unmittelbare Umgebung einer einzelnen Struktur.

Da die Ätzrate von Anlagenzustand, Gasfluss und Belegung abhängt, wird eine Ätzung in der Fertigung nicht über eine feste Zeit gesteuert, sondern über eine Endpunkterkennung. Dazu verfolgt ein Spektrometer die optische Emission des Plasmas (optical emission spectroscopy, OES) – beim Durchätzen einer Schicht ändert sich die Intensität charakteristischer Linien, weil andere Reaktionsprodukte entstehen – oder ein Interferometer misst die verbleibende

Schichtdicke. Nach dem erkannten Endpunkt folgt eine kurze, definierte Überätzzeit, um Reste über die gesamte Scheibe hinweg sicher zu entfernen.

Trockenätzverfahren

Bei den Trockenätzverfahren werden Gase durch hochfrequente Wechselfelder angeregt, typisch sind 13,56 MHz und 2,45 GHz. Bei einem Druck zwischen 0,1 Pa und 100 Pa liegt die freie Weglänge der Teilchen bei wenigen Millimetern bis Zentimetern.

Moderne Anlagen trennen die Erzeugung des Plasmas von der Beschleunigung der Ionen: Ein Generator hoher Frequenz (z.B. 13,56 MHz oder 60 MHz, induktiv oder per Mikrowelle eingekoppelt) stellt die Ladungsträgerdichte und damit die Ätzrate ein, ein zweiter Generator niedriger Frequenz (z.B. 400 kHz bis 2 MHz) an der Waferelektrode die Ionenenergie und damit den physikalischen Anteil der Ätzung. Beide Größen sind so weitgehend unabhängig voneinander einstellbar. Zusätzlich wird das Plasma häufig gepulst, um Ladungsansammlungen und Profilfehler in tiefen Strukturen zu verringern.

Es gibt generell drei verschiedene Arten des Trockenätzens, welche auf der folgenden Seite näher beschrieben werden:

- Physikalisches Trockenätzen: physikalischer Abtrag der Scheibenoberfläche durch beschleunigte Teilchen
- Chemisches Trockenätzen: Reaktion zwischen Gas und Scheibenoberfläche
- Chemisch physikalisches Trockenätzen: physikalisches Ätzen mit chemischen Anteil

Als jüngere Entwicklung kommt das atomlagengenaue Ätzen (atomic layer etching, ALE) hinzu. Dabei werden die chemische Aktivierung der Oberfläche und der eigentliche Abtrag zeitlich getrennt und zyklisch wiederholt, so dass pro Zyklus nur eine oder wenige Atomlagen entfernt werden. Der Abtrag wird damit nicht über die Ätzzeit, sondern über die Anzahl der Zyklen eingestellt – eine Voraussetzung für Bauelemente, deren Schichtdicken nur noch wenige Nanometer betragen, etwa Gate-all-around-Transistoren.

Trockenätzverfahren

Ionenstrahlätzen

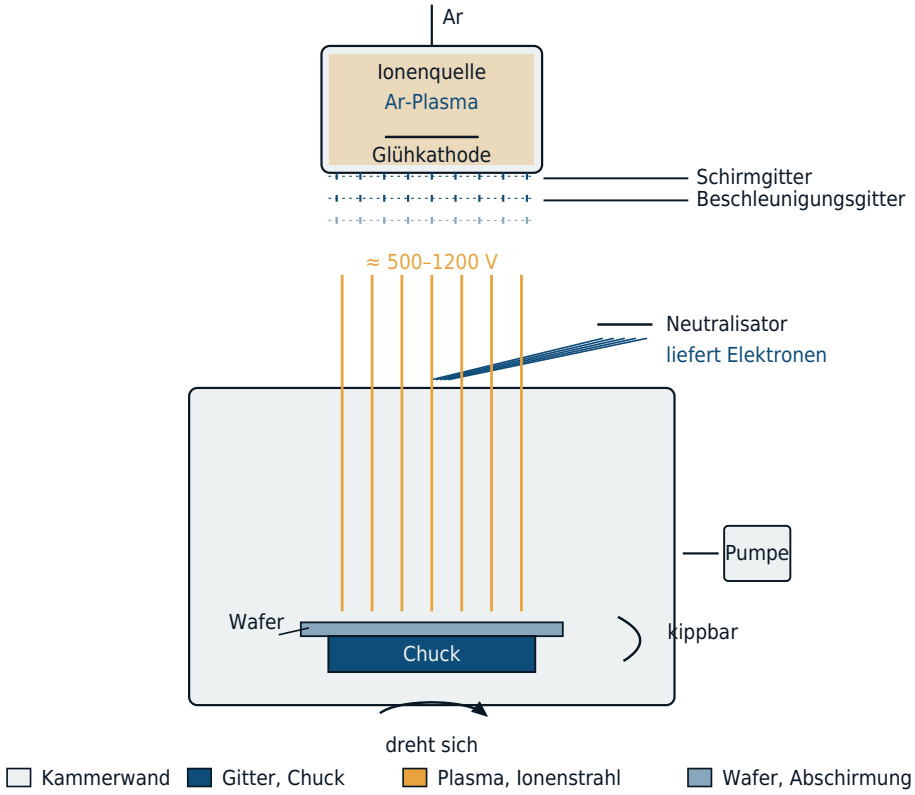
Das Ionenstrahlätzen (ion beam etching, IBE) ist ein rein physikalisches Trockenätzverfahren. Dabei werden Argonionen als gerichteter Ionenstrahl mit 1–3 keV auf die Scheibe gelenkt. Durch die Energie der Teilchen schlagen diese

Material aus der Oberfläche heraus. Die Ionen werden in einer vom Prozessraum getrennten Quelle erzeugt und über ein Gittersystem extrahiert; der Arbeitsdruck liegt unterhalb von 0,1 Pa, damit die Ionen ihre Richtung bis zur Scheibe behalten. Der Wafer wird senkrecht oder gekippt zum Ionenstrahl in der Prozesskammer gehalten, die Ätzung erfolgt völlig anisotrop. Die Selektivität ist nur gering, da keine Unterscheidung der Materialien von den beschleunigten Ionen ausgeht. Das Gas und herausgeschlagenes Material werden vom Pumpensystem abgesaugt, jedoch setzen sich auch Partikel an den Kammerwänden und an vertikalen Kanten der Scheibe selbst ab, da das abgetragene Material nicht in den gasförmigen Zustand übergeht.

Um die Partikelablagerungen zu verhindern wird neben Argon ein reaktives Gas in die Kammer eingeleitet. Das Gas reagiert dabei mit den Argonionen und führt zu einem physikalischen Ätzprozess mit chemischem Charakter. Das Gas reagiert nun teilweise mit der Oberfläche, aber auch mit dem durch den physikalischen Anteil herausgeschlagenen Material, zu einem gasförmigen Reaktionsprodukt. Wird das reaktive Gas bereits in der Ionenquelle zugegeben, spricht man von reaktivem Ionenstrahlätzen (reactive ion beam etching, RIBE), wird es erst vor der Scheibe eingeleitet, von chemisch unterstütztem Ionenstrahlätzen (chemically assisted ion beam etching, CAIBE).

Ionenstrahl-Ätzreaktor

Ein kollimierter Ionenstrahl trägt das Material rein physikalisch ab



Das Gittersystem beschleunigt die Argon-Ionen zu einem gebündelten Strahl; der Neutralisator liefert Elektronen, damit sich der Wafer nicht auflädt. Weil kein reaktives Gas beteiligt ist, wird nahezu jedes Material rein mechanisch abgetragen – dafür fehlt die chemische Selektivität des reaktiven Ionenätzens.

Mit diesem Verfahren können nahezu alle Materialien geätzt werden. Durch die senkrechte Bestrahlung ist der Abtrag an senkrechten Kanten sehr gering (hohe Anisotropie).

Auf Grund der geringen Selektivität und der geringen Ätzrate (geringe Ionendichte im Strahl) spielt das Ionenstrahlätzen bei der Strukturierung von Silicium und Siliciumverbindungen keine Rolle mehr. Unverzichtbar ist es dagegen bei Materialien, die keine flüchtigen Reaktionsprodukte bilden und daher chemisch kaum zu ätzen sind:

- magnetische Schichtstapel, etwa die Tunnelelemente von MRAM-Speicherzellen aus CoFeB und MgO, sowie Leseköpfe und Magnetfeldsensoren
- Edelmetalle wie Platin, Iridium und Ruthenium
- piezoelektrische und ferroelektrische Schichten wie PZT
- Lithiumniobat und ähnliche Kristalle in der integrierten Photonik

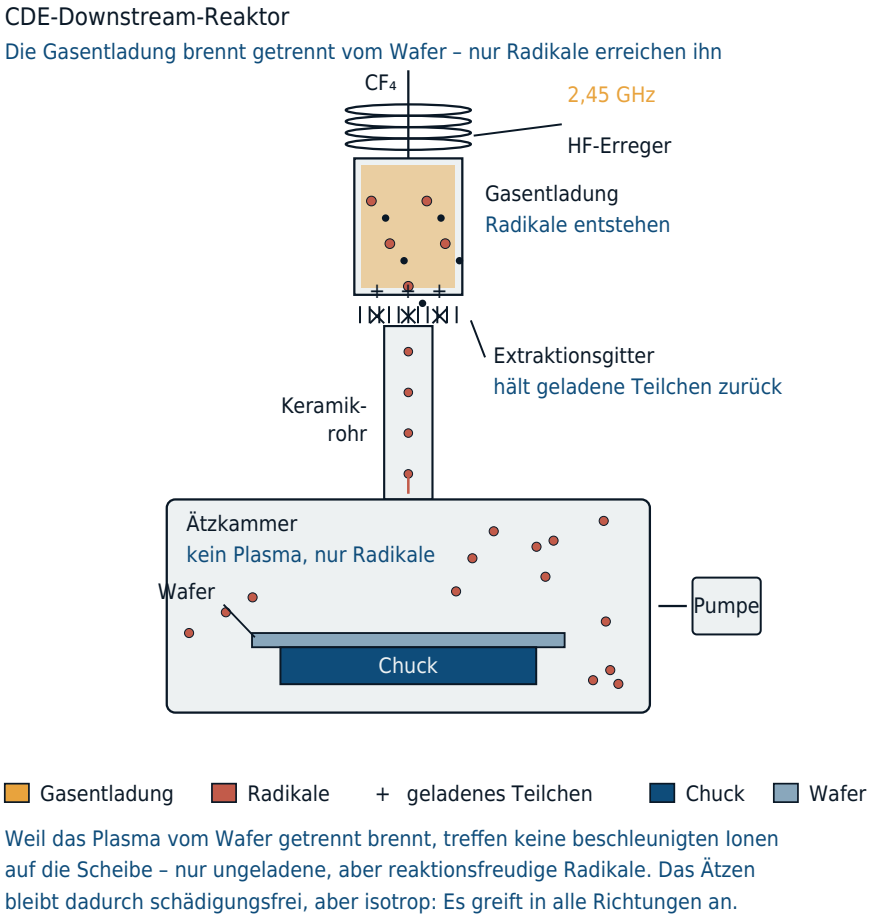
Der kippbare und rotierende Substrathalter ist dabei ein wesentlicher Vorteil: Über den Einfallswinkel lassen sich der Kantenwinkel der Struktur einstellen

und wieder angelagertes Material von den Seitenwänden entfernen – bei Schichtstapeln aus vielen unterschiedlichen Materialien ein Ergebnis, das mit plasmachemischen Verfahren nicht erreichbar ist. Da isolierende Substrate sich unter Ionenbeschuss aufladen würden, arbeitet die Quelle mit einem Neutralisator, der dem Strahl Elektronen zusetzt.

Plasmaätzen

Das Plasmaätzen ist ein rein chemisches Ätzen (engl. chemical dry etching, CDE). Der Vorteil dabei ist, dass die Scheibenoberfläche nicht durch beschleunigte Ionen geschädigt wird. Das Ätzprofil ist auf Grund der frei beweglichen Gasteilchen isotrop, weshalb das Plasmaätzen meist zum Abtragen kompletter Schichten bei hoher Ätzrate eingesetzt wird.

Eine Anlagenart beim Plasmaätzen ist der Down-Stream-Reaktor. Dabei wird durch Anlegen einer hochfrequenten Spannung (z.B. 2,45 GHz) ein Plasma durch Stoßionisation erzeugt. Der Raum der Gasentladung ist dabei vom Wafer getrennt, das Gas gelangt über ein Keramikrohr zum Wafer.



Weil das Plasma vom Wafer getrennt brennt, treffen keine beschleunigten Ionen auf die Scheibe – nur ungeladene, aber reaktionsfreudige Radikale. Das Ätzen bleibt dadurch schädigungsfrei, aber isotrop: Es greift in alle Richtungen an.

In der Zone der Gasentladung entstehen durch die Stöße der Gasmoleküle

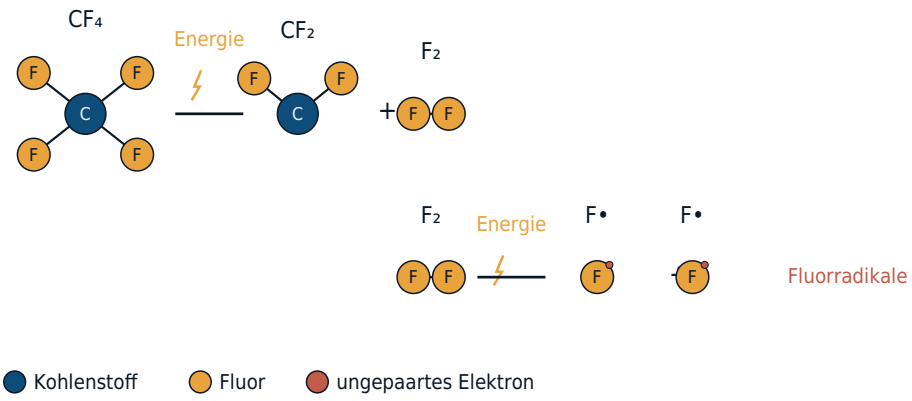
untereinander verschiedene Teilchen, unter anderem Radikale. Radikale sind Moleküle mit aufgespaltener Bindung, die durch ein ungepaartes Außenelektron sehr reaktionsfreudig sind. Als neutrales Gas wird beispielsweise Tetrafluormethan CF_4 (auch Kohlenstofftetrafluorid) eingeleitet und durch die Wechselspannung in CF_2 und ein Fluormolekül F_2 zerlegt. Ebenso kann durch Zugabe von Sauerstoff Fluor von CF_4 abgespalten werden:

$2CF_4 + O_2 \rightarrow 2COF_2 + 2F_2$

Das Fluormolekül kann nun durch die Energie in der Gasentladungszone in zwei einzelne Fluoratome aufgespaltet werden: Fluorradikale (Radikale sind ungeladene Teilchen, ein einzelnes Fluoratom besitzt neun Protonen im Kern und neun Elektronen in der Atomhülle).

Bildung von Fluorradikalen

Energie in der Gasentladung spaltet CF_4 in zwei Schritten auf



Ein Radikal ist ein ungeladenes Teilchen mit einem ungepaarten Außenelektron – das macht es sehr reaktionsfreudig. Zusätzlich lässt sich Fluor auch mit zugesetztem Sauerstoff abspalten: $2\ CF_4 + O_2 \rightarrow 2\ COF_2 + 2\ F_2$.

Neben den neutralen Radikalen entstehen weitere, teils geladene Teilchen ($CF4^+$, $CF3^+$, $CF2^+$, ...), welche gemeinsam über das Gasrohr zur Ätzkammer geleitet werden. Geladene Teilchen werden entweder durch ein Extraktionsgitter abgefangen oder rekombinieren auf dem Weg wieder zu ungeladenen Molekülen. Auch die Fluorradikale werden teilweise wieder gebunden. Es gelangen jedoch genügend Radikale in die Ätzkammer, wo sie mit der Scheibenoberfläche reagieren. Neutrale Teilchen nehmen nicht an der Reaktion Teil und werden, so wie auch die entstandenen Reaktionsprodukte, abgesaugt.

Beispiele für Schichten die im CDE-Verfahren geätzt werden:

Silicium: $Si + 4F \rightarrow SiF_4$
Siliciumdioxid: $SiO_2 + 4F \rightarrow SiF_4 + O_2$
Siliciumnitrid: $Si_3N_4 + 12F \rightarrow 3SiF_4 + 2N_2$

Heutige Bedeutung

Zur Strukturierung von Schichten ist das rein chemische Trockenätzen ungeeignet, weil das isotrope Profil die Maske unterätzt. Als Remote- oder Downstream-Prozess – das Plasma brennt getrennt von der Scheibe, nur langlebige neutrale Radikale erreichen sie – erfüllt es in der Fertigung aber klar umgrenzte Aufgaben:

- Lackentfernung (Ashing): Sauerstoffradikale verbrennen den Fotolack nach der Ätzung oder Implantation zu CO_2 und H_2O.
- Kammerreinigung: Stickstofftrifluorid NF_3 wird in einer vorgeschalteten Plasmaquelle zerlegt; die Fluorradikale entfernen Abscheidungen aus Prozesskammern, ohne dass diese geöffnet werden müssen.
- Selektives Rückätzen: Radikalprozesse aus NH_3 und NF_3 entfernen natives Oxid und dünne Nitridschichten schädigungsfrei – auch in Kontaktlöchern, in die eine gerichtete Ätzung nur schwer hineinreicht.
- Freilegen von Kanälen: Beim Nanosheet- bzw. Gate-all-around-Transistor wird Silicium-Germanium selektiv gegen Silicium isotrop entfernt, um die Kanalschichten freizustellen. Hier wird die Isotropie, die bei der Strukturierung stört, gezielt ausgenutzt.

Reaktives Ionenätzen

Die Ätzcharakteristik - Selektivität, Ätzprofil, Ätzrate, Homogenität, Reproduzierbarkeit - ist beim reaktiven Ionenätzen (Reactive Ion Etching, kurz RIE) durch die eingesetzten Gase und die Prozessparameter (Generatorleistung, Druck, Plattenabstand, Gasfluss) exakt einstellbar. Dabei ist sowohl ein isotropes als auch ein anisotropes Ätzprofil möglich. Damit ist das RIE-Ätzen, ein chemisch-physikalisches Ätzverfahren, das wichtigste Ätzverfahren zur Strukturierung diverser Schichten in der Halbleiterfertigung.

In der Prozesskammer befinden sich die Wafer auf einer mit Wechselspannung gespeisten Elektrode (HF-Elektrode). Durch Stoßionisation wird ein Plasma erzeugt, das freie Elektronen und positiv geladene Ionen enthält. Bei positiver Spannung an der HF-Elektrode lagern sich die Elektronen an dieser an, und können sie auf Grund der Austrittsarbeit bei der positiven Halbwelle nicht verlassen, die Elektrode lädt sich somit auf bis zu 1000 V negativ auf (BIAS-Spannung). Die langsamen Ionen, die der schnellen Wechselspannung nicht folgen konnten, bewegen sich nun in Richtung der negativ aufgeladenen Elektrode mit den Wafern.

Ist die freie Weglänge der Ionen groß, treffen die Teilchen auf Grund ihrer Geschwindigkeit nahezu senkrecht auf die Scheiben. Dabei wird Material durch die beschleunigten Ionen von der Oberfläche herausgelöst (physikalisches Ätzen), teilweise reagieren die Teilchen aber auch chemisch mit dem Substrat.

Vertikale Kanten werden dabei nicht getroffen, so dass hier kein Abtrag erfolgt und das Ätzprofil anisotrop ist. Die Selektivität ist durch den physikalischen Abtrag nicht sehr hoch, zudem wird die Scheibenoberfläche durch die beschleunigten Ionen geschädigt. Diese muss später durch thermische Behandlung ausgeheilt werden.

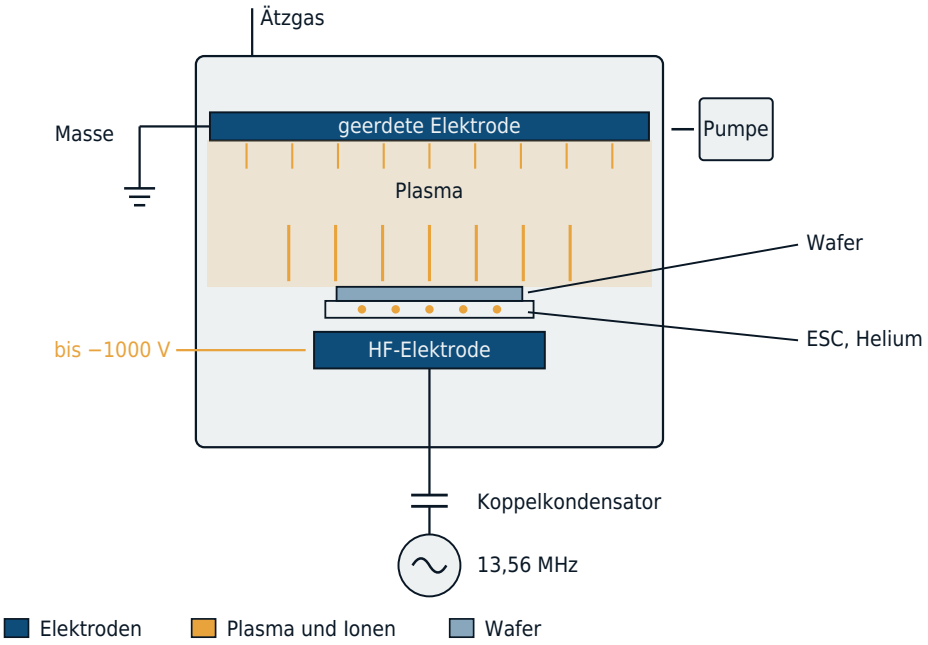
Der chemische Anteil der Ätzung geschieht durch die Reaktion von freien Radikalen auf der Scheibenoberfläche und mit dem physikalisch abgetragenen Material, so dass sich dieses nicht wie beim Ionenstrahlätzen an den Kammerwänden oder auf den Scheiben anlagern kann. Durch Druckerhöhung nimmt die freie Weglänge der Teilchen ab, es passieren auf dem Weg zur Scheibe also viele Stöße zwischen den Teilchen, die so ihre Richtung fortlaufend ändern. So geschieht keine gerichtete Ätzung mehr auf der Scheibenoberfläche, der Ätzprozess nimmt einen stärker chemischen Charakter an, das Ätzprofil ist isotrop, die Selektivität erhöht sich.

Die Hexodenbauform ätzte mehrere Scheiben gleichzeitig auf einem sechsseitigen Träger. Solche Chargenanlagen sind heute nur noch von historischem Interesse, weil sich Homogenität und Endpunkt über viele Scheiben hinweg nicht ausreichend genau kontrollieren lassen.

Heute wird ausschließlich einzelscheibenweise in Parallelplattenbauweise geätzt. Der Wafer liegt auf einem elektrostatischen Chuck (electrostatic chuck, ESC), der ihn plan festhält und über Helium auf der Scheibenrückseite thermisch an die temperierte Elektrode koppelt. Damit sind Wafertemperaturen von etwa −100 °C bis über 100 °C einstellbar – die Temperatur ist eine der wirksamsten Stellgrößen für Selektivität und Profil.

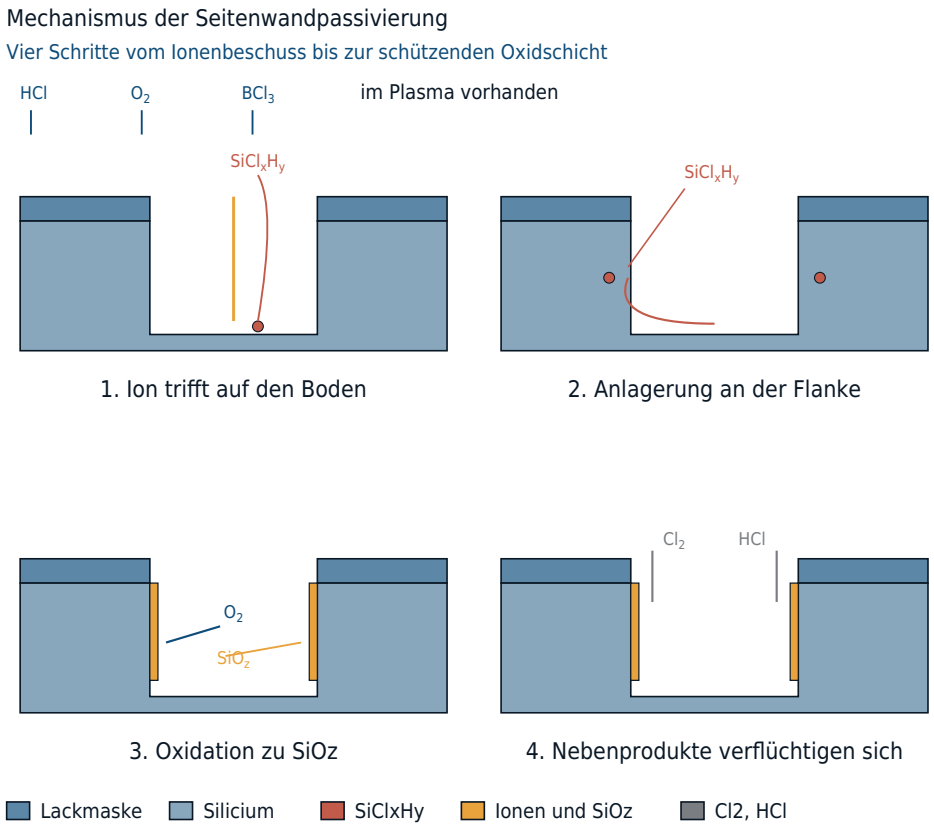
Reaktives Ionenätzen

Parallelplattenbauweise: Der Wafer liegt auf der Hochfrequenzelektrode



Weil die untere Elektrode kleiner ist als die obere, lädt sie sich gegenüber dem Plasma negativ auf. Die positiven Ionen werden dadurch senkrecht beschleunigt und tragen gerichtet ab – das Ätzprofil wird anisotrop.

Durch eine Passivierung der Seitenwände bei einer Siliciumätzung wird ein anisotropes Ätzprofil erzielt. Dabei reagiert Sauerstoff in der Prozesskammer mit herausgelöstem Silicium der Waferoberfläche zu Siliciumdioxid, das an den vertikalen Kanten aufwächst. Eine Oxidschicht auf horizontalen Flächen wird durch den Ionenbeschuss entfernt, so dass eine Ätzung in die Tiefe fortschreiten kann.



Die Ätzrate hängt in jedem Fall vom Druck, der Leistung des HF-Generators, den Prozessgasen, dem Gasdurchfluss und der Wafer- bzw. Elektrodentemperatur ab.

Die Anisotropie nimmt bei steigender HF-Leistung, sinkendem Druck und abnehmender Temperatur zu. Die Homogenität der Ätzung wird von den Gasen, dem Elektrodenabstand und dem Elektrodenmaterial bestimmt. Bei zu geringem Elektrodenabstand ist das Plasma nicht gleichmäßig in der Kammer verteilt und führt so zur Inhomogenität. Bei größer werdendem Abstand nimmt die Ätzrate ab, da das Plasma auf ein größeres Volumen verteilt ist. Als Elektrodenmaterial wurde früher Kohlenstoff eingesetzt: Da die Fluor- und Chlorgase auch Kohlenstoff abtragen, bewirkt die Elektrode eine gleichmäßige Belastung des Plasmas, Scheibenränder werden so nicht stärker belastet als die Scheibenmitte. Heute sind Elektroden aus Silicium, Siliciumcarbid oder Quarz üblich; die dem Plasma zugewandten Kammerteile werden mit Yttrium- oder Aluminiumoxid beschichtet, um Partikel und metallische Verunreinigungen zu vermeiden.

Weiterentwicklungen des reaktiven Ionenätzens

Der klassische RIE-Reaktor koppelt Plasmadichte und Ionenenergie über dieselbe Elektrode: Mehr Leistung bedeutet gleichzeitig mehr Ätzrate und mehr Schädigung. Die heutigen Anlagengenerationen lösen diese Kopplung auf und erweitern das Verfahren in mehrere Richtungen.

Hochdichte Plasmaquellen
: Bei induktiv gekoppelten Anlagen (inductively coupled plasma, ICP, auch TCP) erzeugt eine Spule über einem dielektrischen Fenster das Plasma, bei ECR-Anlagen eine Mikrowelle im Magnetfeld. Die Ionenenergie wird getrennt über einen Bias-Generator an der Waferelektrode eingestellt. Ergebnis sind hohe Ätzraten bei geringer Ionenenergie, also weniger Schädigung, bei Drücken von deutlich unter 1 Pa.

Mehrfrequenz-Anlagen
: Beim Ätzen von Dielektrika arbeiten kapazitiv gekoppelte Reaktoren mit zwei oder drei Generatoren unterschiedlicher Frequenz auf derselben Elektrode, um Ionendichte, Ionenenergie und Energieverteilung getrennt einzustellen. Für tiefe Strukturen werden Bias-Spannungen von mehreren Kilovolt eingesetzt.

Gepulstes Plasma
: Wird die eingespeiste Leistung im Kilohertzbereich getaktet, klingt der Ladungsaufbau in tiefen Strukturen zwischen den Pulsen ab. Das verringert Profilfehler wie Bowing und Notching und verbessert die Selektivität.

Tiefenätzen (DRIE, Bosch-Prozess)
: Beim tiefen reaktiven Ionenätzen wechseln sich sehr kurze Ätzschritte mit SF_6 und Passivierschritte mit C_4F_8 zyklisch ab. So entstehen Gräben mit Aspektverhältnissen über 30:1 und Tiefen von einigen hundert Mikrometern. Die zyklische Führung hinterlässt eine wellige Seitenwand (Scalloping). Hauptanwendungen sind mikromechanische Bauelemente (MEMS) und Durchkontaktierungen durch den Wafer (through-silicon vias).

Kryoätzen
: Bei Wafertemperaturen um −100 °C kondensieren Reaktionsprodukte an den kalten Seitenwänden und passivieren sie, während der Ionenbeschuss den Boden freihält. Das ergibt glatte, senkrechte Seitenwände ohne Scalloping. Seit einigen Jahren wird das Verfahren auch beim Ätzen der Speicherlöcher im 3D-NAND-Flash in der Serienfertigung eingesetzt: Bei Tiefen um 10 µm und Aspektverhältnissen über 50:1 erlauben die tiefen Temperaturen Ätzchemien, die bei Raumtemperatur nicht funktionieren, und etwa die doppelte Ätzrate herkömmlicher Dielektrikaprozesse.

Atomlagengenaues Ätzen (ALE)
: Aktivierung der Oberfläche und Abtrag werden in getrennte, selbstbegrenzende Halbschritte zerlegt und zyklisch wiederholt. Der Abtrag pro Zyklus liegt im Bereich einer Atomlage, gesteuert wird über die Zyklenzahl. Eingesetzt wird ALE dort, wo einzelne Nanometer über die Funktion entscheiden, etwa beim Freilegen von Kanälen und beim Ätzen

von Gate-Strukturen.

Die Selektivität und Ätzrate lassen sich sehr stark durch die eingesetzten Ätzgase beeinflussen. Bei Silicium und Siliciumverbindungen kommen vor allem Chlor und Fluor zum Einsatz.

Eingesetzte Ätzgase beim Trockenätzen

Die Prozesse müssen nicht nur mit einer Gasmischung und gleichen Ätzparametern ablaufen. Natürliches Oxid auf Polysilicium kann beispielsweise erst mit einer hohen Ätzrate und geringer Selektivität abgetragen werden. Das Polysilicium wird dann mit hoher Selektivität zur darunter liegenden Schicht geätzt. Reale Rezepte bestehen deshalb aus mehreren Schritten mit unterschiedlichen Gasmischungen, Leistungen und Drücken.

Die folgende Tabelle gibt einen Überblick über Gase, die beim Trockenätzen verwendet werden.

Schicht	Prozessgase	Bemerkung
SiO_2, Si_3N_4	CF_4, O_2	F ätzt Si, O_2 entfernt Kohlenstoff (C)
	CHF_3, O_2	CHF_3 wirkt als Polymer, erhöhte Selektivität gegen Si
	CHF_3, CF_4	
	CH_3F, CH_2F_2	verbesserte Selektivität von Si_3N_4 gegenüber SiO_2
	C_2F_6 / SF_6	
	C_3F_8	erhöhte Ätzrate gegenüber CF_4
	C_4F_6, C_4F_8 + O_2, Ar	Standardchemie für tiefe Strukturen (Kontaktlöcher, Speicherlöcher im 3D-NAND); die kohlenstoffreichen Moleküle bilden eine Polymerschicht auf Maske und Seitenwand, O_2 steuert deren Dicke
	BCl_3, Cl_2	keine Kontamination durch Kohlenstoff (C)
	$SiCl_4$, Cl_2 / HCl, O_2 / $SiCl_4$, HCl	
Poly-Si	HBr / Cl_2 / O_2	verbesserte Selektivität gegenüber Fotolack und SiO_2; Standardchemie für Gate-Strukturen
	SF_6	hohe Ätzrate, gute Selektivität gegen SiO_2
	NF_3	hohe Ätzrate, isotrop
	HBr, Cl_2	

monokristallines Si	HBr, NF_3, O_2	höhere Selektivität gegen SiO_2
	BCl_3, Cl_2 / HBr, NF_3	
	SF_6 / C_4F_8 SF_6, O_2	Tiefenätzen: zyklisch im Bosch-Prozess bzw. bei tiefen Temperaturen im Kryoprozess
	Cl_2	isotrope Ätzung
	BCl_3	nur geringe Ätzrate, bindet zusätzlich Restfeuchte und natives Al_2O_3
Al-Legierungen	BCl_3 / Cl_2 / CF_4	anisotrope Ätzung
	BCl_3 / Cl_2 / CHF_3	verbesserte Seitenwandpassivierung
	BCl_3 / Cl_2 / N_2	höhere Ätzrate, keine Kohlenstoffkontamination
W, TiN, Ti	SF_6 / NF_3	Rückätzen von Wolfram in Kontaktlöchern
	Cl_2, BCl_3	Barrieren- und Metal-Gate-Schichten

Historisch wurden für monokristallines Silicium auch bromhaltige Fluorchlorkohlenwasserstoffe wie CF_3Br verwendet. Sie sind als ozonschichtschädigende Stoffe nach dem Montreal-Protokoll verboten und in der Fertigung durch HBr-Mischungen ersetzt worden.

Metallisierung: warum Kupfer nicht geätzt wird

Die Ätzung von Aluminiumlegierungen hat mit dem Übergang auf Kupferleitbahnen an Bedeutung verloren. Kupfer bildet mit Chlor und Fluor keine bei Prozesstemperatur flüchtigen Verbindungen und lässt sich deshalb praktisch nicht trockenätzen. Stattdessen wird im Damascene-Verfahren zunächst der Graben im Dielektrikum geätzt, anschließend das Kupfer abgeschieden und der Überschuss durch chemisch-mechanisches Polieren entfernt. Aluminium wird weiterhin für Bondpads sowie in der Leistungs- und Analogtechnik strukturiert.

Klimawirkung der Prozessgase

Die perfluorierten Ätz- und Reinigungsgase sind chemisch außerordentlich stabil – genau die Eigenschaft, die sie im Prozess brauchbar macht. In der Atmosphäre werden sie deshalb kaum abgebaut und wirken als sehr starke Treibhausgase. Die Treibhauspotenziale über 100 Jahre liegen nach dem sechsten IPCC-Bericht bei rund 7400 für CF_4, rund 12 400 für C_2F_6, rund 14 600 für CHF_3, rund 17 400 für NF_3 und rund 25 000 für SF_6. CF_4 verbleibt dabei nach heutiger Abschätzung einige zehntausend Jahre in der Atmosphäre.

Fertigungsseitig wird das Abgas jeder Anlage einzeln nachbehandelt (point-of-use abatement), thermisch, katalytisch oder in einem Plasmabrenner. Reaktive Gase wie NF_3 werden dabei zu über 99 % zerlegt, CF_4 jedoch nur teilweise; es

entsteht überdies als Zerlegungsprodukt anderer Fluorverbindungen. Reduziert wird der Beitrag daher vor allem am Prozess selbst: durch Ersatz von C_2F_6 durch NF_3 bei der Kammerreinigung, geringere Gasflüsse, kürzere Ätzzeiten und Verfahren wie das Kryoätzen mit höherer Ätzrate. Regulatorisch sind die Stoffe von der EU-Verordnung über fluorierte Treibhausgase erfasst; zusätzlich wird im Rahmen von REACH eine Beschränkung von PFAS diskutiert, für die Halbleiterfertigung bislang mit Ausnahmen und langen Übergangsfristen.

Der Wafertest

Grundlagen des Wafertests

Nach Abschluss aller Front-End-Prozessschritte liegt auf dem Wafer eine Vielzahl fertiger, aber noch unvereinzelter Dies vor. Bevor der Wafer zersägt wird, wird jeder einzelne Die noch auf dem Wafer elektrisch getestet – man spricht vom Wafertest oder Probe Test.

Der Test dient zwei Zwecken: Zum einen werden defekte Dies identifiziert, damit sie im nachfolgenden Montageprozess nicht unnötig weiterverarbeitet werden. Zum anderen liefert der Test bereits erste Daten zur elektrischen Charakterisierung der Schaltung, etwa Schwellspannungen oder Leckströme, die zur Prozesskontrolle genutzt werden können.

Ein Testsystem besteht im Wesentlichen aus drei Komponenten:

- Wafer-Prober: ein Präzisionsgerät, das den Wafer exakt positioniert und Die für Die unter die Prüfnadeln bewegt
- Probe Card: die eigentliche Kontaktierungseinheit mit feinen Nadeln, die auf die Kontaktpads (Pads) jedes Dies aufsetzen
- Tester: das elektronische Prüfgerät, das Testmuster erzeugt und die Antwort der Schaltung auswertet

Da moderne Chips mehrere hundert bis tausend Kontaktpads besitzen können, muss die Probe Card entsprechend viele, sehr fein positionierte Nadeln besitzen, die zuverlässig und ohne die empfindliche Chipoberfläche zu beschädigen, Kontakt herstellen.

Die Prüfnadelkarte (Probe Card)

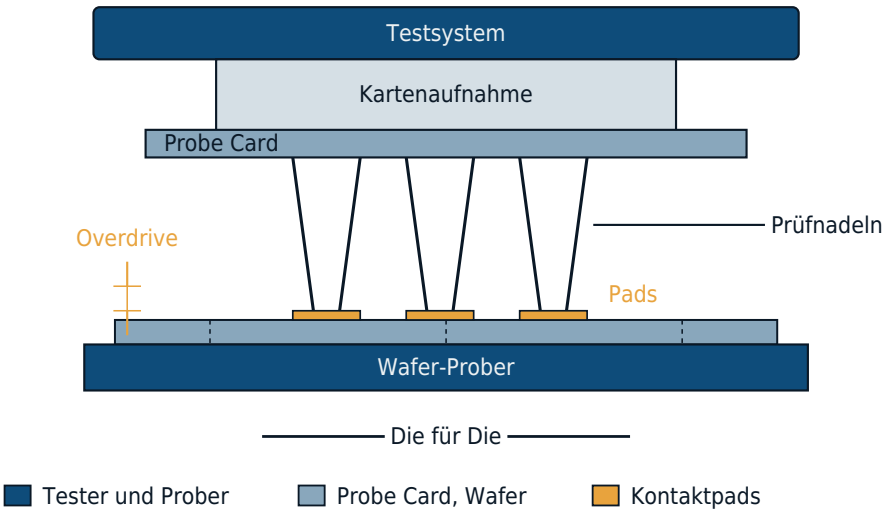
Die Probe Card bildet die mechanische und elektrische Schnittstelle zwischen Tester und Wafer. Ältere Ausführungen nutzen freistehende Nadeln aus Wolfram oder einer Berylliumkupfer-Legierung, die radial auf einen Ring montiert sind. Bei sehr feinen Pad-Abständen (Pitch) kommen heute vermehrt MEMS-basierte Probe Cards zum Einsatz, deren Kontaktspitzen mittels lithografischer Verfahren aus Silicium oder Metall hergestellt werden – ein Verfahren, das eine deutlich höhere Positioniergenauigkeit erlaubt.

Beim Aufsetzen der Nadeln auf die Pads entsteht eine kleine, aber gewollte Kraft (Overdrive), die die native Oxidschicht auf der Padoberfläche durchbricht

und so einen niederohmigen elektrischen Kontakt sicherstellt. Zu hoher Overdrive beschädigt jedoch die Pads und kann darunterliegende Schichten schädigen – die Justierung dieser Kraft ist daher ein kritischer Prozessparameter.

Aufbau des Wafer-Tests

Die Nadeln setzen auf den Pads auf, bevor der Wafer zersägt wird



Die Nadeln werden mit einer kleinen Kraft aufgesetzt, dem Overdrive. Er durchbricht die native Oxidschicht auf den Pads und sichert den Kontakt. Zu viel Overdrive beschädigt die Pads und die Schichten darunter.

Um den Durchsatz zu erhöhen, werden bei modernen Testsystemen häufig mehrere Dies gleichzeitig kontaktiert (Multi-Site-Testing). Damit lassen sich mehrere Chips parallel prüfen, was die Testzeit pro Wafer erheblich verkürzt.

Yield-Mapping und Redundanz

Das Ergebnis jedes Einzeltests wird positionsgenau in einer sogenannten Wafer-Map gespeichert – einer digitalen Karte, die für jeden Die auf dem Wafer vermerkt, ob dieser die Prüfung bestanden hat (Pass) oder nicht (Fail). Diese Karte wird als Yield-Map bezeichnet und begleitet den Wafer durch die gesamte weitere Fertigung.

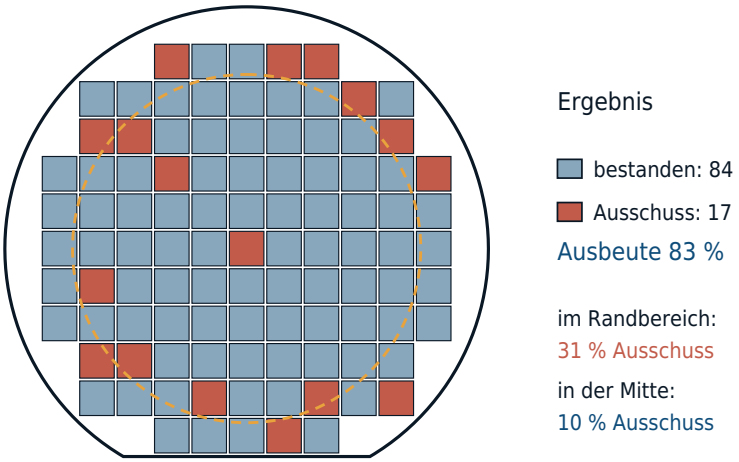
Die Yield-Map hat mehrere wichtige Funktionen:

- Im nachfolgenden Dicing- und Montageprozess werden anhand der Karte nur funktionsfähige Dies weiterverarbeitet, fehlerhafte werden aussortiert
- Räumliche Muster in der Yield-Map (etwa gehäufte Ausfälle am Waferrand) liefern Hinweise auf systematische Prozessprobleme
- Bei Speicherbausteinen (DRAM, NAND-Flash) ermöglicht die Karte den

gezielten Einsatz redundanter Schaltungsblöcke: enthält ein Die einzelne defekte Speicherzellen, können diese durch integrierte Reserveschaltungen (Redundanz) elektrisch ersetzt werden, wodurch der Die insgesamt doch noch nutzbar bleibt

Yield-Map eines Wafers

Jeder Die trägt sein Testergebnis – Ausfälle häufen sich am Rand



Die Karte begleitet den Wafer durch die weitere Fertigung: Beim Dicing werden nur die bestandenen Dies weiterverarbeitet. Räumliche Muster wie dieser Randabfall weisen auf systematische Prozessprobleme hin.

Die durchschnittliche Ausbeute (Yield) eines Wafers – also der Anteil funktionsfähiger Dies an der Gesamtzahl – ist eine der wichtigsten Kennzahlen der Halbleiterfertigung, da sie unmittelbar die Herstellungskosten je Chip bestimmt.

Finaler Test und Binning

Der finale Funktionstest

Nachdem der Die durch Montage und Verkapselung sein fertiges Gehäuse erhalten hat, wird er einem letzten, umfassenden Funktionstest unterzogen – dem Final Test. Im Gegensatz zum Wafertest, der auf Grund der begrenzten Nadelanzahl und Kontaktqualität meist nur eine eingeschränkte Testabdeckung bietet, kann der fertige, gehäuste Chip über seine regulären Anschlüsse deutlich umfassender und unter realistischeren elektrischen Bedingungen geprüft

werden.

Getestet werden unter anderem:

- Vollständige Funktionsprüfung aller logischen und analogen Schaltungsblöcke
- Elektrische Kennwerte wie Betriebsspannung, Stromaufnahme und Signallaufzeiten
- Verhalten unter verschiedenen Temperatur- und Spannungsbedingungen
- Bei Speicherbausteinen: vollständiger Zellentest über den gesamten Adressraum

Erst nach erfolgreichem Final Test gilt ein Chip als verkaufsfähig. Fehlerhafte Bauteile werden aussortiert und in der Regel vernichtet, da eine Reparatur auf dieser Fertigungsstufe wirtschaftlich nicht mehr sinnvoll ist.

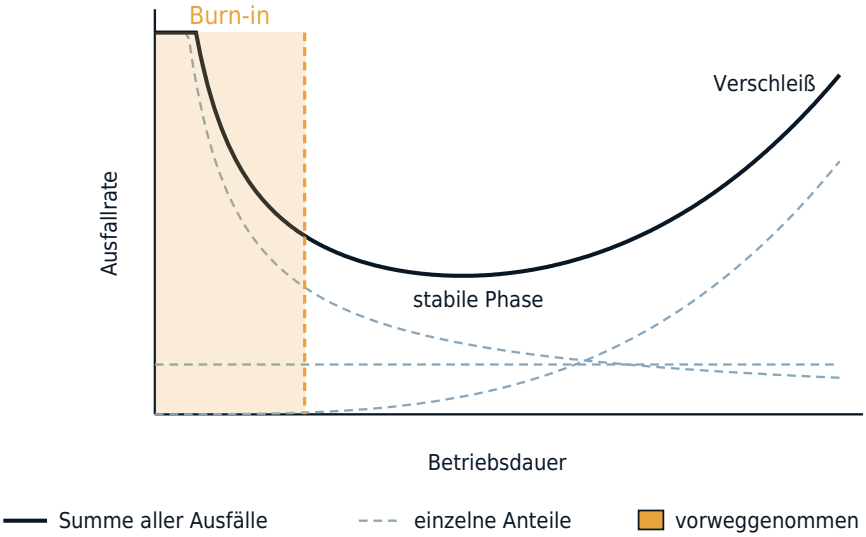
Burn-in

Bei besonders zuverlässigkeitskritischen Anwendungen (etwa in der Automobil- oder Luftfahrtindustrie) wird vor dem eigentlichen Final Test häufig ein sogenannter Burn-in-Prozess durchgeführt. Dabei werden die Chips für einen definierten Zeitraum – von wenigen Stunden bis zu mehreren Tagen – unter erhöhter Temperatur und teilweise erhöhter Betriebsspannung belastet.

Der Hintergrund ist die sogenannte Badewannenkurve der Ausfallrate elektronischer Bauteile: Ein ungewöhnlich hoher Anteil an Frühausfällen tritt typischerweise innerhalb der ersten Betriebsstunden eines Bauteils auf, meist verursacht durch versteckte Fertigungsfehler wie Materialdefekte oder Verunreinigungen, die dem Final Test allein entgangen wären. Nach dieser Frühausfallphase sinkt die Ausfallrate auf ein niedriges, stabiles Niveau ab, bevor sie gegen Ende der Lebensdauer des Bauteils durch Verschleißeffekte wieder ansteigt.

Badewannenkurve der Ausfallrate

Der Burn-in nimmt die Frühausfälle vorweg



Versteckte Fertigungsfehler zeigen sich in den ersten Betriebsstunden. Unter erhöhter Temperatur und Spannung fallen diese Bauteile schon beim Hersteller aus – ausgeliefert wird nur, was den steilen Anfangsbereich überstanden hat.

Durch den Burn-in-Prozess werden diese Frühausfälle vorweggenommen: Bauteile mit versteckten Defekten fallen bereits während des Burn-in aus und werden aussortiert, bevor sie beim Kunden zum Einsatz kommen. Der verbleibende, überlebende Anteil der Chips weist dadurch eine deutlich höhere Zuverlässigkeit im späteren Betrieb auf.

Binning

Selbst innerhalb eines einzelnen Wafers, ja sogar innerhalb desselben Dies-Designs, treten auf Grund unvermeidlicher Prozessschwankungen leichte Unterschiede in den elektrischen Eigenschaften auf – etwa in der maximal erreichbaren Taktfrequenz eines Prozessors oder im Stromverbrauch. Anstatt Chips mit geringfügig schlechteren Kennwerten zu verwerfen, werden sie stattdessen nach ihrer tatsächlichen Leistungsfähigkeit sortiert und als unterschiedliche Produktvarianten vermarktet. Dieses Verfahren wird als Binning bezeichnet.

Ein bekanntes Beispiel sind Prozessorfamilien, bei denen technisch identische Chips je nach erreichter maximaler Taktfrequenz unter verschiedenen Modellbezeichnungen und zu unterschiedlichen Preisen verkauft werden. Chips, die die höchsten Anforderungen nicht erfüllen, werden in ein niedrigeres Leistungssegment (Bin) eingestuft, anstatt als Ausschuss verworfen zu werden.

Binning erlaubt es Herstellern, die natürliche Streuung der Fertigung

wirtschaftlich zu nutzen, statt sie als reinen Ausbeuteverlust zu behandeln. Nach dem Binning erfolgt die endgültige Kennzeichnung der Chips (Marking), üblicherweise durch Laserbeschriftung des Gehäuses, sowie die abschließende Verpackung für den Versand an Gerätehersteller.

Dicing

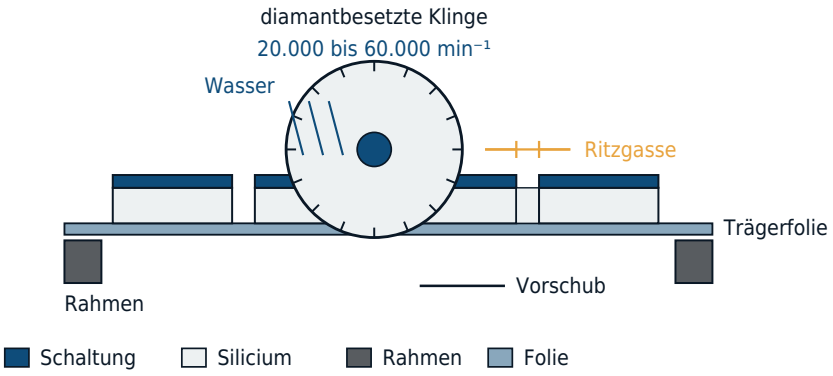
Vom Wafer zum Einzelchip

Nach dem Wafertest liegt der Wafer weiterhin als zusammenhängende Scheibe vor, auf der sich hunderte bis mehrere tausend einzelne Dies befinden, getrennt durch schmale, unbelegte Streifen – die sogenannten Ritzgassen (Scribe Lines oder Dicing Lanes). Diese Streifen enthalten keine funktionalen Schaltungsstrukturen und sind absichtlich für den nachfolgenden Trennschritt vorgesehen.

Beim Dicing (auch Wafersägen oder Die Separation genannt) wird der Wafer entlang dieser Ritzgassen in die einzelnen Dies zerteilt. Dieser Schritt markiert den Übergang vom Front-End der Fertigung, in dem alle Dies noch gemeinsam auf dem Wafer bearbeitet wurden, zum Back-End, in dem jeder Die einzeln weiterverarbeitet wird.

Dicing – das Zersägen des Wafers

Die Säge trennt entlang der Ritzgassen, die Folie hält die Dies



Die Ritzgassen tragen keine Schaltung – ihre Breite begrenzt, wie dick die Klinge sein darf. Weil sie mit der Miniaturisierung schmaler werden, gewinnen Laserverfahren an Bedeutung. Die Folie hält die Dies nach dem Schnitt in Lage.

Vor dem eigentlichen Trennvorgang wird der Wafer üblicherweise auf eine dünne, elastische Trägerfolie (Dicing Tape) aufgeklebt, die auf einem Metallrahmen gespannt ist. Diese Folie hält die einzelnen Dies nach dem Zersägen weiterhin an ihrer ursprünglichen Position, sodass die räumliche

Zuordnung zur zuvor erstellten Yield-Map erhalten bleibt.

Sägeverfahren

Das am weitesten verbreitete Verfahren ist das mechanische Sägen mit einer dünnen, diamantbesetzten Kreissäge (Blade Dicing). Die Säge rotiert mit sehr hoher Drehzahl (typischerweise 20.000–60.000 Umdrehungen pro Minute) und wird während des Schnitts kontinuierlich mit deionisiertem Wasser gekühlt und gespült, um die entstehende Reibungswärme abzuführen und den Sägestaub aus der Schnittfuge zu entfernen.

Die Breite der Ritzgassen begrenzt dabei, wie schmal die Sägeklinge sein darf und wie präzise der Schnitt geführt werden muss. Da mit fortschreitender Miniaturisierung die verfügbare Ritzgassenbreite immer weiter abnimmt, gewinnen alternative Trennverfahren zunehmend an Bedeutung:

- Laserschneiden (Laser Dicing): Ein fokussierter Laserstrahl trennt den Wafer entweder durch direktes Abtragen des Materials oder durch gezieltes Einbringen von Mikrorissen im Waferinneren (Stealth Dicing), entlang derer der Wafer anschließend mechanisch auseinandergebrochen wird
- Plasmaätzen (Plasma Dicing): Der Wafer wird durch ein anisotropes Trockenätzverfahren entlang der Ritzgassen komplett durchgeätzt; dieses Verfahren erzeugt besonders glatte, beschädigungsfreie Kanten und eignet sich gut für sehr dünne Wafer

Jedes Verfahren hat spezifische Vor- und Nachteile hinsichtlich Schnittgeschwindigkeit, mechanischer Belastung des Materials und erreichbarer Kantenqualität, weshalb die Wahl des Trennverfahrens vom jeweiligen Produkt und Waferaufbau abhängt.

Herausforderungen und Qualitätssicherung

Beim mechanischen Sägen können am Rand der Schnittkante feine Ausbrüche (Chipping) entstehen, kleine Absplitterungen des spröden Siliciummaterials. Diese Mikrorisse können sich unter mechanischer oder thermischer Belastung im späteren Betrieb des Bauteils weiter ausbreiten und die Zuverlässigkeit gefährden. Die Sägeparameter – insbesondere Vorschubgeschwindigkeit, Drehzahl und Kühlung – müssen daher sorgfältig auf das jeweilige Schichtsystem des Wafers abgestimmt werden.

Besondere Herausforderungen ergeben sich bei Wafern mit empfindlichen Schichtstapeln, etwa bei Bauteilen mit Low-k-Dielektrika in der Verdrahtungsebene. Diese porösen Materialien sind mechanisch deutlich empfindlicher als klassisches Siliciumdioxid und neigen beim Sägen verstärkt zu Delamination, dem Ablösen einzelner Schichten.

Nach dem Trennvorgang wird die Qualität der Schnittkanten häufig optisch überprüft. Zusätzlich bleibt die zuvor erstellte Yield-Map weiterhin gültig: Da die Dies durch die Trägerfolie an ihrer Position verbleiben, lässt sich anhand der Karte eindeutig bestimmen, welche der nun vereinzelten Dies bereits im Wafertest als funktionsfähig identifiziert wurden und somit in den nächsten Montageschritt, das Die Attach, übernommen werden.

Die Attach

Mechanische Befestigung des Dies

Nach dem Dicing liegen die einzelnen, bereits identifizierten funktionsfähigen Dies weiterhin auf der Trägerfolie vor. Im ersten Schritt der eigentlichen Montage wird jeder Die von der Folie abgehoben (Die Pick) und mechanisch auf seinem endgültigen Trägermaterial befestigt – diesen Vorgang bezeichnet man als Die Attach oder Die Bonding.

Als Trägermaterial (Substrat) kommen je nach Bauteiltyp unterschiedliche Ausführungen zum Einsatz:

- Leadframe: ein gestanztes oder geätztes Metallgitter, meist aus Kupfer oder einer Kupferlegierung, das später die äußeren Anschlussbeinchen des Gehäuses bildet
- Laminat- oder Keramiksubstrat: mehrlagige Trägerplatten mit integrierten Leiterbahnen, wie sie etwa bei Ball-Grid-Array-Gehäusen (BGA) verwendet werden
- Direkte Chip-Montage: bei manchen Anwendungen wird der Die unmittelbar auf eine Leiterplatte oder ein anderes System aufgebracht, ohne zunächst ein eigenes Gehäuse zu erhalten

Der Die wird dabei von einem Greifwerkzeug (Die Bonder) präzise aufgenommen, ausgerichtet und mit definiertem Anpressdruck auf das Substrat aufgesetzt. Die exakte Positionierung ist wichtig, da im nachfolgenden Bonding-Schritt die Kontaktpads des Dies möglichst genau mit den entsprechenden Anschlussstellen auf dem Substrat übereinstimmen müssen.

Verbindungsmaterialien

Zur eigentlichen Befestigung des Dies auf dem Substrat kommen verschiedene Verbindungsmaterialien zum Einsatz, deren Auswahl von den elektrischen, thermischen und mechanischen Anforderungen des jeweiligen Bauteils abhängt:

- Klebstoffe (Epoxidharze): das mit Abstand am häufigsten eingesetzte Verfahren. Der Klebstoff wird entweder als Paste dosiert oder in

vorgefertigter Filmform (Die Attach Film) zwischen Die und Substrat aufgebracht und anschließend thermisch ausgehärtet. Elektrisch leitfähige Klebstoffe enthalten zusätzlich feine Silberpartikel, um neben der mechanischen Verbindung auch eine elektrische Kontaktierung der Rückseite des Dies zum Substrat herzustellen
- Lötverbindungen: insbesondere bei Bauteilen mit hohen Anforderungen an die Wärmeableitung, etwa Leistungshalbleitern, wird der Die mittels eines Lotmaterials (häufig auf Gold- oder Zinnbasis) direkt mit dem Substrat verbunden. Lötverbindungen bieten eine deutlich bessere thermische Leitfähigkeit als Klebstoffe
- Eutektisches Bonden: Der Die wird unter Wärmeeinwirkung direkt mit einer dünnen, eutektischen Metallschicht (häufig einer Gold-Silicium-Legierung) auf dem Substrat verbunden, ohne dass ein zusätzliches Verbindungsmaterial in Pastenform notwendig ist

Die Wahl des Verbindungsmaterials beeinflusst maßgeblich, wie effizient die im Betrieb entstehende Verlustwärme des Chips über das Substrat abgeführt werden kann – ein besonders kritischer Faktor bei leistungsstarken Prozessoren und Leistungshalbleitern.

Mechanische Spannungen und Qualitätssicherung

Ein wesentliches Problem beim Die Attach sind mechanische Spannungen, die durch unterschiedliche Wärmeausdehnungskoeffizienten von Die und Substrat entstehen. Silicium und die üblichen Substratmaterialien wie Kupfer oder Keramik dehnen sich bei Temperaturänderungen unterschiedlich stark aus. Während des Aushärtens des Verbindungsmaterials, aber auch später im Betrieb des fertigen Bauteils, können dadurch Spannungen im Die entstehen, die im ungünstigsten Fall zu Rissen führen.

Um diese Spannungen zu begrenzen, werden Verbindungsmaterialien mit einer gewissen Elastizität eingesetzt, die kleine Ausdehnungsunterschiede kompensieren können, ohne dass die Verbindung selbst oder der Die beschädigt wird. Zudem wird die Schichtdicke des Verbindungsmaterials sorgfältig kontrolliert: Eine zu dünne Schicht bietet zu wenig Ausgleichsvermögen, eine zu dicke Schicht kann wiederum die thermische Anbindung verschlechtern.

Nach dem Aushärten des Verbindungsmaterials wird die Qualität der Die-Attach-Verbindung überprüft, unter anderem durch optische Inspektion auf Lunker (Hohlräume im Verbindungsmaterial) sowie durch Scherfestigkeitstests, bei denen die mechanische Haftung des Dies stichprobenartig geprüft wird. Erst nach erfolgreicher Prüfung wird der Die dem nächsten Prozessschritt, dem Bonding, zugeführt.

Bonding

Elektrische Kontaktierung des Dies

Nachdem der Die durch das Die Attach mechanisch auf dem Substrat befestigt wurde, muss er noch elektrisch mit den Anschlussstrukturen des Substrats verbunden werden. Dieser Vorgang wird als Bonding bezeichnet. Die elektrische Kontaktierung geschieht dabei zu etwa 90 % mittels Draht-Bonden (Wire Bonding), während die mechanische Befestigung des Chips – wie im vorherigen Kapitel beschrieben – hauptsächlich durch Kleben mit Epoxidharzen erfolgt.

Beim Draht-Bonden werden dünne Drähte aus Gold oder Aluminium verwendet, die mit den Anschlusskontakten (Pads) von Chip und Substrat kalt verschweißt werden. Man unterscheidet dabei drei gängige Verfahren:

- Thermosonic-Ball-Wedge-Bonden
- Ultraschall-Wedge-Wedge-Bonden
- Thermokompressions-Ball-Wedge-Bonden

Bei allen drei Verfahren entsteht am Ende des Drahtes zunächst eine kleine Kugel (Ball), die durch lokales Aufschmelzen der Drahtspitze erzeugt wird. Aufgrund der starken Oberflächenkräfte des flüssigen Goldes erstarrt die Schmelze zu einer Kugel, deren größerer Durchmesser gegenüber dem Bonddraht eine zuverlässigere mechanische und elektrische Verbindung gewährleistet.

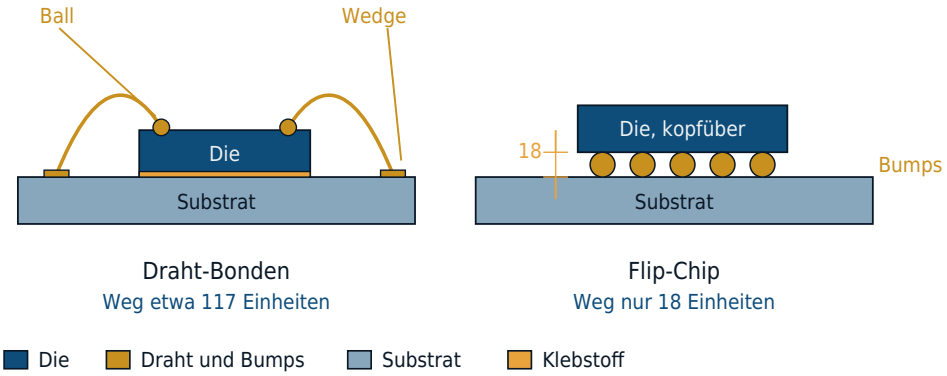
Flip-Chip-Technik

Neben dem klassischen Draht-Bonden existiert mit der Flip-Chip-Technik ein alternatives Verfahren, das ganz ohne lange Bonddrähte auskommt und die kompakteste Form der Verbindung zwischen Chip und Substrat darstellt.

Um eine Flip-Chip-Verbindung herstellen zu können, müssen auf dem Chip und/oder auf dem Substrat kleine Lothöcker, sogenannte Bumps, vorhanden sein. Zur Herstellung der Verbindung wird der Chip kopfüber, aber sehr passgenau auf das Substrat aufgesetzt. Während eines anschließenden Reflow-Prozesses schmelzen die Bumps auf und verbinden Chip und Substrat gleichzeitig mechanisch und elektrisch.

Draht-Bonden und Flip-Chip

Die Länge des Verbindungswegs entscheidet über Induktivität und Wärme



Der Bonddraht ist rund 6-mal so lang wie der Weg über einen Bump. Kurze Wege bedeuten weniger parasitäre Induktivität und eine bessere Wärmeableitung – beides spricht für die Flip-Chip-Technik.

Der wesentliche Vorteil der Flip-Chip-Technik liegt in den sehr kurzen Verbindungswegen vom Chip zum Substrat über die Lötkügelchen – im Gegensatz zu den vergleichsweise langen Wire Bonds beim klassischen Draht-Bonden. Dies reduziert parasitäre Induktivitäten und Kapazitäten in der Verbindung, was bei sehr hohen Signalfrequenzen elektrisch von Vorteil ist.

Da beim Flip-Chip-Verfahren zwischen Chip und Substrat üblicherweise ein Luftspalt verbleibt, ist es meist erforderlich, zusätzlich einen sogenannten Underfiller einzubringen. Dieser füllt den verbleibenden Spalt zwischen Chip und Substrat und gleicht mechanische Spannungen aus, die durch unterschiedliche Wärmeausdehnungskoeffizienten der beiden Materialien entstehen.

Anforderungen und Wärmeableitung

Unabhängig vom gewählten Bonding-Verfahren müssen die elektrischen Verbindungen zwischen Chip und Substrat eine Reihe grundlegender Anforderungen erfüllen: geringer elektrischer Widerstand, hohe Zuverlässigkeit über die gesamte Lebensdauer des Bauteils sowie eine gute mechanische Haftung auch unter thermischer Wechselbelastung.

Ein zunehmend wichtiger Aspekt beim Bonding ist die Wärmeableitung. Mit steigender Verlustleistung moderner Chips wird die Abfuhr der entstehenden Wärme zu einem sehr wichtigen Punkt der gesamten Aufbau- und

Verbindungstechnik. Kurze Leitungswege und Materialien mit guter Wärmeleitfähigkeit sind hierfür unabdingbar. Genau in diesem Punkt bietet die Flip-Chip-Technik gegenüber dem klassischen Draht-Bonden Vorteile, da die kurzen Verbindungswege über die Bumps auch für die Wärmeableitung günstiger sind als lange Bonddrähte.

Nach Abschluss des Bonding-Vorgangs ist der Chip elektrisch vollständig mit seinem Substrat verbunden. Der nächste Schritt in der Montage besteht darin, diese empfindliche Verbindung sowie den Die selbst durch eine geeignete Verkapselung vor äußeren Einflüssen zu schützen.

Verkapselung

Aufgaben der Verkapselung

Nach dem Bonding ist der Die elektrisch vollständig mit seinem Substrat verbunden, liegt jedoch noch offen und ungeschützt vor. Im letzten wesentlichen Schritt der Montage wird der Die deshalb in eine schützende Vergussmasse eingebettet – diesen Vorgang bezeichnet man als Verkapselung oder Molding.

Die Verkapselung als Teil des gesamten Packaging-Prozesses erfüllt dabei mehrere grundlegende Aufgaben:

- Schutz der Schaltung vor äußeren Einflüssen wie Feuchtigkeit, Staub und mechanischer Beschädigung
- Führung der empfindlichen Kontakte auf dem Chip in robuste, handhabbare Anschlüsse nach außen
- Versorgung der Schaltung mit elektrischer Energie über die Gehäuseanschlüsse
- Abfuhr der im Betrieb entstehenden Verlustleistung
- Verteilung von Signalen und Daten zwischen Chip und umgebender Schaltung

Der Trend in der Gehäusetechnik geht dabei zu immer kleineren und dünneren Gehäusen (Chip Size Package, CSP), bei denen das fertige Gehäuse in Länge und Breite nur noch unwesentlich größer ist als der eigentliche Chip, typischerweise um einen Faktor von etwa 1,2.

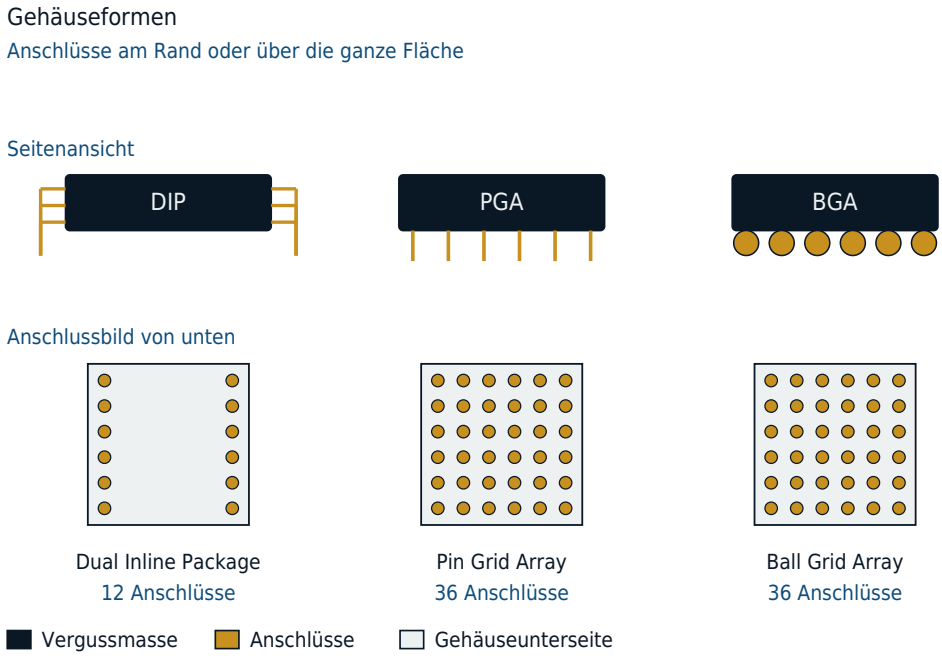
Materialien und Verfahren

Am weitesten verbreitet ist das sogenannte Molding, bei dem der Die zusammen mit den Bonddrähten oder Bumps in eine duroplastische Kunststoffmasse eingespritzt wird. Diese Vergussmasse härtet unter Hitzeeinwirkung aus und

bildet danach das eigentliche, feste Gehäuse des Bauteils.

Bekannte Gehäuseformen, die auf diesem Weg hergestellt werden, sind beispielsweise:

- DIP – Dual Inline Package
- PGA – Pin Grid Array, vor allem bei Prozessoren gebräuchlich
- BGA – Ball Grid Array



Bei Bauteilen mit besonders hohen Anforderungen an die Wärmeableitung, etwa bei Leistungshalbleitern, kommen statt Kunststoffen auch Gehäuse aus gut wärmeleitenden Materialien wie Metall oder Keramik zum Einsatz. Diese Materialien sind jedoch deutlich teurer als Kunststoff, weshalb in der Massenfertigung überwiegend auf Kunststoffgehäuse zurückgegriffen wird.

Bei der zuvor beschriebenen Flip-Chip-Technik entfällt ein klassisches Vollgehäuse häufig vollständig: Hier übernimmt der bereits im Bonding-Schritt eingebrachte Underfiller einen Teil der Schutzfunktion, während der Chip selbst offen auf dem Substrat verbleibt.

Wärmeableitung und abschließende Qualitätsprüfung

Da die Verlustleistung moderner Chips stetig zunimmt, ist die Wärmeableitung einer der wichtigsten Aspekte bei der Wahl von Gehäuseform und

Vergussmaterial. Kurze Leitungswege sowie Materialien mit guter Wärmeleitfähigkeit sind hierfür unabdingbar. Bei besonders leistungsstarken Bauteilen werden deshalb häufig zusätzliche Maßnahmen zur Wärmeableitung in das Gehäusedesign integriert, etwa freiliegende Metallflächen an der Gehäuseunterseite (Exposed Die Pad) oder spezielle Wärmeleitpfade innerhalb des Substrats.

Nach dem Aushärten der Vergussmasse wird das Gehäuse mit einer eindeutigen Kennzeichnung versehen (Marking), üblicherweise durch Laserbeschriftung, die unter anderem Herstellerangaben, Typenbezeichnung und gegebenenfalls das Ergebnis des vorangegangenen Binnings trägt.

Damit ist die eigentliche Montage abgeschlossen. Der fertig verkapselte Chip durchläuft im Anschluss noch den bereits beschriebenen finalen Funktionstest, bevor er für den Versand an Gerätehersteller verpackt wird.

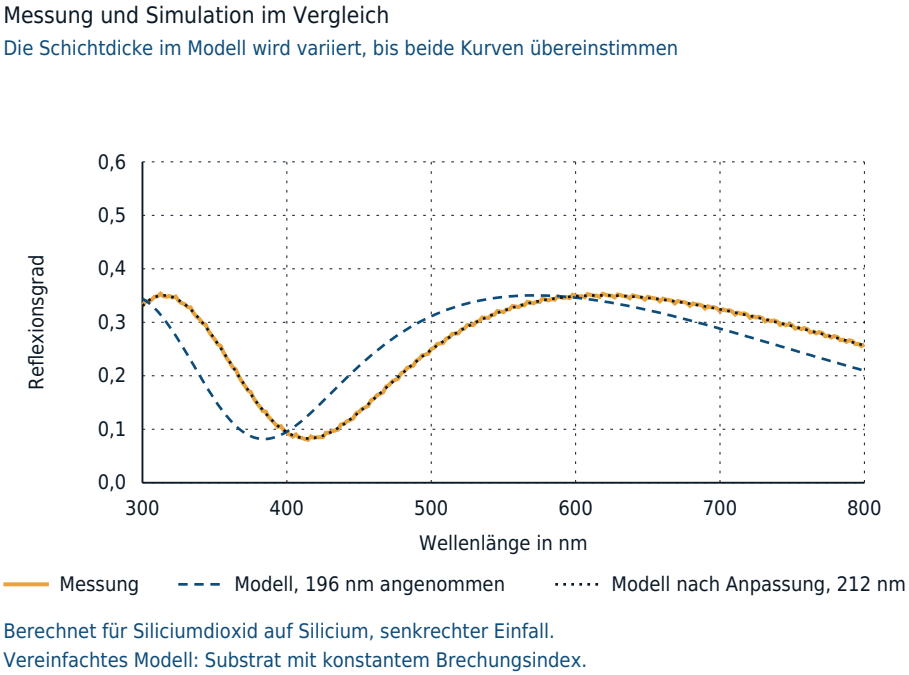
Schichtdickenmessung

Beurteilung der Messung

Bei diesen optischen Messverfahren erfolgt die Schichtdickenbestimmung nur indirekt, die optischen Parameter der zu messenden Schichten müssen dazu hinreichend bekannt sein. Mit Hilfe dieser Werte wird dann ein Modell der Schichtabfolge auf dem Wafer erstellt und eine Messung simuliert. Das Ergebnis wird anschließend mit der tatsächlichen Messung verglichen und die Schichtdicken (aber auch andere optische Indizes) der im Modell hinterlegten Materialien variiert, bis Simulation und Messung bestmöglich übereinstimmen.

Vergleich von Simulation und Messung einer optischen Streulichtmessung

(Quelle: n&k Technology)



Je mehr Parameter im Modell variiert werden, desto leichter kann eine

Übereinstimmung mit der Messung gefunden werden, aber desto unsicherer ist auch das Ergebnis. Meist gibt ein Parameter (goodness of fit, GOF; Anpassungsgüte) darüber Auskunft, wie gut die Simulation mit der Messung übereinstimmt (0–100 %).

Messtechnik

Oxidschichten sind lichtdurchlässige Schichten. Wird Licht auf den Wafer gestrahlt und reflektiert, ändern sich bestimmte Eigenschaften der Lichtwellen, die mit Messgeräten erfasst und ausgewertet werden können. Sollen mehrere übereinander liegende Schichten gemessen werden, müssen diese unterschiedliche optische Indizes haben, damit eine Unterscheidung der Materialien möglich ist.

Damit die Schichtdicke auf dem gesamten Wafer kontrolliert werden kann, werden mehrere Messpunkte (z.B. 5 Punkte bei 150-mm, 9 Punkte bei 200-mm, 13–21 Punkte bei 300-mm-Wafern) gemessen. Dabei dürfen nicht nur die Werte der einzelnen Punkte mit dem vorgegebenen Sollwert verglichen werden, sondern auch die Dicke der Punkte untereinander, da auch die Homogenität wichtig für die weiteren Prozesse ist. Ist die aufgebrachte Schicht zu dick oder zu dünn, muss Material entfernt (z.B. durch chemisch mechanisches Polieren) oder zusätzliches (Wiederholung des entsprechenden Prozesses) aufgebracht werden.

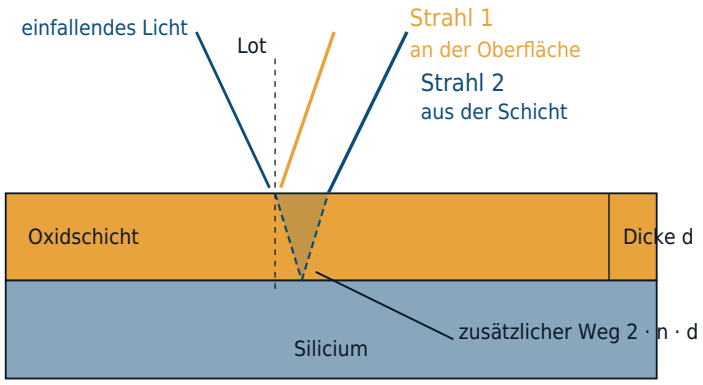
Interferometrie

Bei der Überlagerung von Lichtwellen können sich diese verstärken, abschwächen oder auch auslöschen. Dieses Phänomen nutzt man in der Halbleiterfertigung zur Messung von transparenten Schichten aus.

Auf den Wafer einfallende Lichtstrahlen werden an einer transparenten Schicht teilweise reflektiert, ein Teil der Strahlen durchdringt das Material aber auch und wird an der darunter befindlichen Schicht reflektiert.

Lichteinfall bei der Schichtdickenmessung

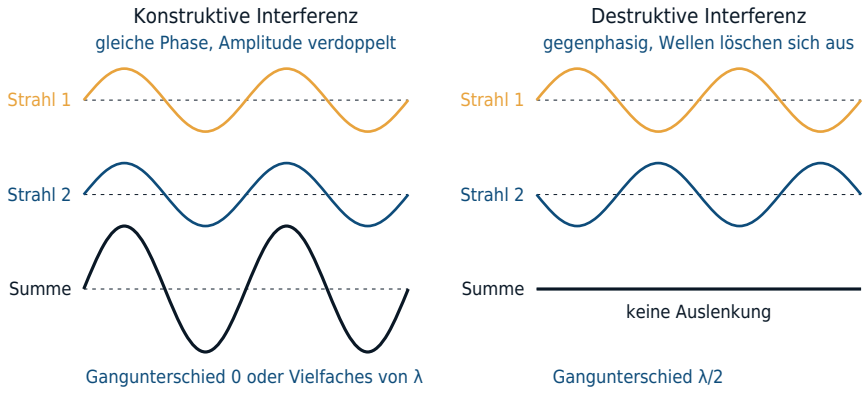
Ein Teil wird oben reflektiert, ein Teil erst an der Grenzfläche darunter



Beide Strahlen verlassen die Schicht parallel und überlagern sich im Messgerät. Strahl 2 hat dabei den längeren Weg zurückgelegt: zweimal die Schichtdicke, multipliziert mit dem Brechungsindex n der Schicht.

Überlagerung zweier Lichtwellen

Der Gangunterschied entscheidet, ob sie sich verstärken oder auslöschen



In der Messung liegt zwischen beiden Extremen ein stetiger Übergang: Aus der Stärke des reflektierten Lichts folgt die Dicke der Schicht.

Es wird dabei ein Spektrum an unterschiedlichen Wellenlängen auf den Wafer gestrahlt. Je nach Dicke der durchstrahlten Schicht überlagern sich die reflektierten Strahlen unterschiedlich und ergeben eine für jede Schichtdicke und jedes Material charakteristische Überlagerung. Mit Hilfe eines Photometers kann aus dem reflektierten Licht die Schichtdicke ermittelt werden.

Interferenzmessungen sind bei Schichten möglich, deren Dicke in etwa einem Viertel der eingestrahlten Lichtwellenlänge oder mehr entspricht.

Ellipsometrie

Ellipsometrie ist die Bestimmung von optischen Eigenschaften durch Änderung der Polarisation von Licht. Linear polarisiertes Licht (die Lichtwellen haben eine bestimmte Schwingung) wird dabei unter einem festen Winkel auf den Wafer gestrahlt.

Veranschaulichung von möglichen Polarisierungen von Licht (links linear, rechts zirkular):

(Quelle: Optics Group, University of Glasgow)

Bei der Reflexion an der Waferoberfläche bzw. an der Grenzschicht zwischen zwei Schichten wird das Licht umpolarisiert. Diese Änderung kann dann mit einem Analysator gemessen werden. Aus den bekannten optischen Eigenschaften der Schichten (z.B. Brechungswinkeln und Absorptionskoeffizient), der eingestrahlten Wellenlänge und der Polarisierung kann die Schichtdicke bestimmt werden.

Im Gegensatz zur Messung mittels Interferenz, ist die Ellipsometrie für dünne Schichten geeignet, deren Dicke weniger als ein Viertel der eingestrahlten Lichtwellenlänge beträgt.

Partikelmessung

Messtechnik

Partikel gehören zu den kritischsten Defektquellen in der Halbleiterfertigung. Bereits ein einzelnes Partikel im Sub-Mikrometerbereich kann, sofern es auf einer aktiven Struktur liegt, zu einem Kurzschluss, einer offenen Leitung oder einem sonstigen elektrischen Defekt führen und damit einen kompletten Chip funktionsunfähig machen. Da moderne Strukturbreiten mittlerweile im einstelligen Nanometerbereich liegen, muss die Kontamination des Wafers mit Partikeln kontinuierlich überwacht werden.

Partikel können aus unterschiedlichsten Quellen stammen: aus der Umgebungsluft des Reinraums, von Anlagenkomponenten (z.B. abgeplatzte Beschichtungen, Dichtungsabrieb), aus Prozessgasen und -chemikalien oder auch vom Wafer selbst (z.B. Kristalldefekte, die an der Oberfläche zutage treten). Um die Ursache eingrenzen zu können und die Reinraumqualität sowie die Anlagenperformance sicherzustellen, werden Wafer nach kritischen

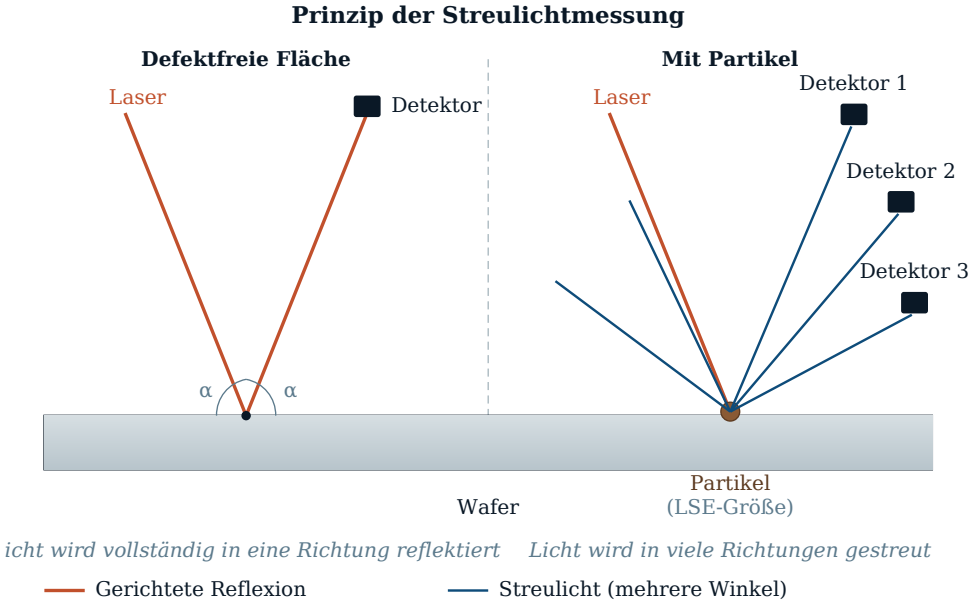
Prozessschritten auf ihre Partikelbelastung hin überprüft.

Die gängigste Methode zur Partikeldetektion nutzt, ähnlich wie einige Schichtdickenmessverfahren, die Wechselwirkung von Licht mit der Waferoberfläche. Da Partikel jedoch keine großflächigen, homogenen Schichten bilden, sondern punktuelle Störstellen darstellen, kommt hier nicht die Interferenz oder Polarisation reflektierten Lichts zum Einsatz, sondern die Streuung von Licht an kleinen Objekten.

Laser-Streulichtmessung

Bei der Laser-Streulichtmessung (Laser Scattering) wird die Waferoberfläche mit einem fokussierten Laserstrahl systematisch abgerastert. Trifft der Strahl auf eine ebene, defektfreie Fläche, wird das Licht weitgehend gerichtet reflektiert. Befindet sich an der betreffenden Stelle jedoch ein Partikel, ein Kratzer oder eine andere topografische Unregelmäßigkeit, wird das Licht in nahezu alle Richtungen gestreut.

Prinzip der Streulichtmessung an einem Partikel



Ein oder mehrere Detektoren, die außerhalb des direkten Reflexionswegs positioniert sind, erfassen dieses Streulicht. Aus der Intensität des gestreuten Lichts lässt sich näherungsweise die Partikelgröße berechnen – typischerweise wird diese jedoch nicht als tatsächliche physikalische Ausdehnung angegeben, sondern als sogenannte Latex Sphere Equivalent (LSE)-Größe, also als die Größe einer Referenzkugel (meist Polystyrol-Latex), die dieselbe Streulichtintensität erzeugen würde.

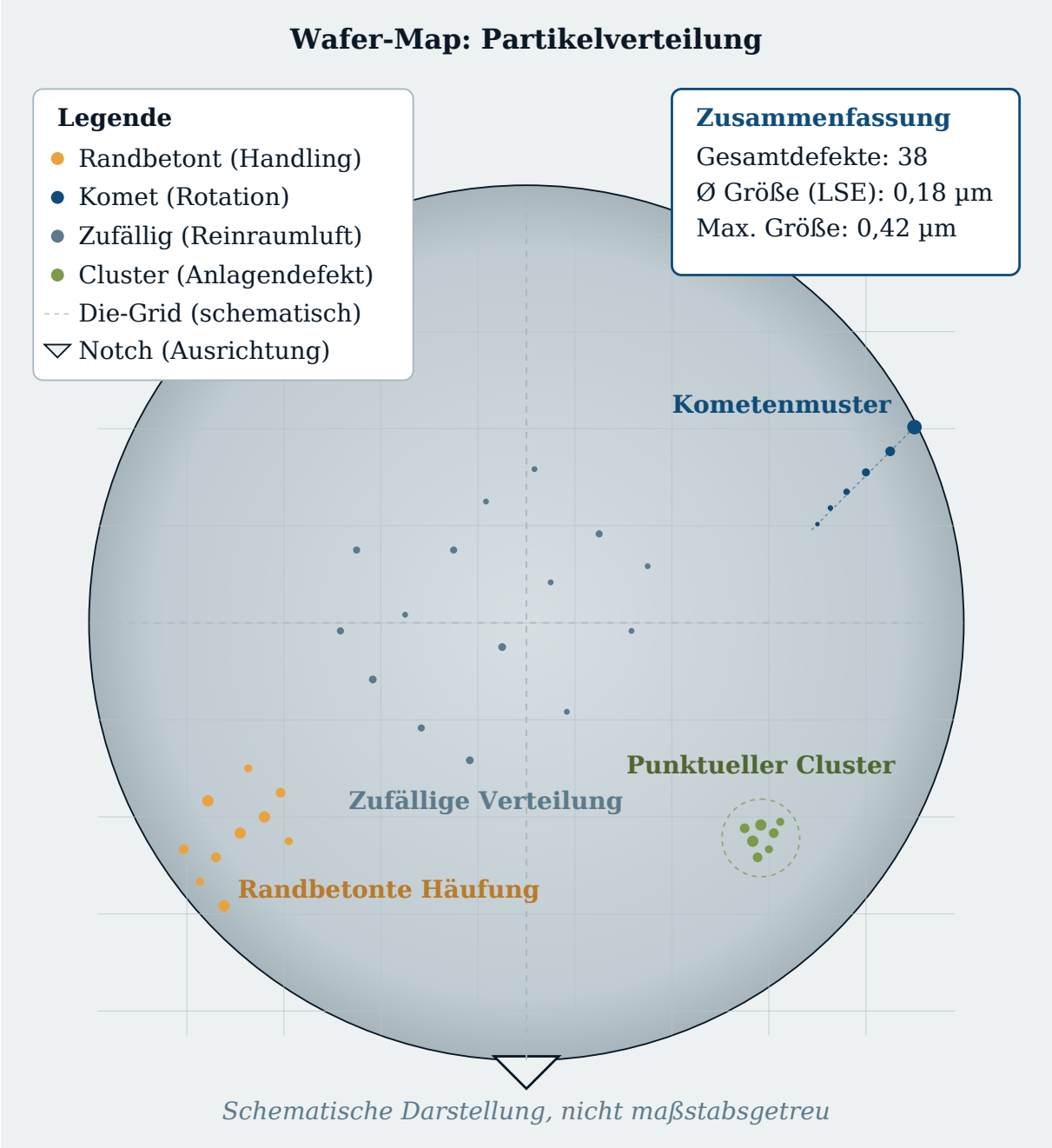
Da unterschiedliche Materialien (Metall, Photoresist, Kristallsplitter etc.) das Licht verschieden stark streuen, weicht die reale Partikelgröße oft von der gemessenen LSE-Größe ab. Moderne Systeme nutzen daher häufig mehrere Wellenlängen und Detektionswinkel gleichzeitig (Multi-Channel- bzw. Multi-Angle-Scattering), um neben der Größe auch erste Rückschlüsse auf die Partikelart zu ermöglichen.

Die erreichbare Nachweisgrenze liegt bei modernen Inspektionssystemen im Bereich von etwa 20–30 nm (LSE), wobei die tatsächliche Empfindlichkeit stark von der Waferoberfläche (Filmstruktur, Rauheit, Reflektivität) und der verwendeten Wellenlänge abhängt. Strukturierte Wafer mit vorhandenen Bauteilen (Pattern-Wafer) sind grundsätzlich schwieriger zu inspizieren als unstrukturierte Wafer (Blanket-Wafer), da Strukturkanten selbst Licht streuen und dadurch fälschlich als Partikel erkannt werden können.

Wafer-Map und Klassifizierung

Das Ergebnis einer Partikelmessung wird üblicherweise als sogenannte Wafer-Map dargestellt: eine grafische Draufsicht auf den Wafer, auf der jeder erkannte Defekt an seiner exakten x/y-Koordinate als Punkt markiert ist. Zusätzlich zur Position werden für jeden Defekt die LSE-Größe sowie ein vorläufiger Klassifizierungscode gespeichert.

Beispiel einer Wafer-Map mit Partikelverteilung



Die räumliche Verteilung der Defekte liefert bereits wichtige Hinweise auf die Fehlerursache. Typische Muster sind zum Beispiel:

- Randbetonte Häufung: deutet häufig auf Probleme beim Handling oder auf ungleichmäßige Prozessbedingungen am Waferrand hin (z.B. bei Spin-Coating- oder CMP-Prozessen).
- Kometenförmige Muster: entstehen oft durch ein einzelnes, größeres Partikel, das während eines Rotationsprozesses (z.B. Lackschleudern) über den Wafer geschleift wird und dabei eine Spur weiterer, kleinerer Defekte hinterlässt.
- Zufällige, gleichmäßige Verteilung: spricht meist für Kontamination aus der Umgebungsluft des Reinraums.

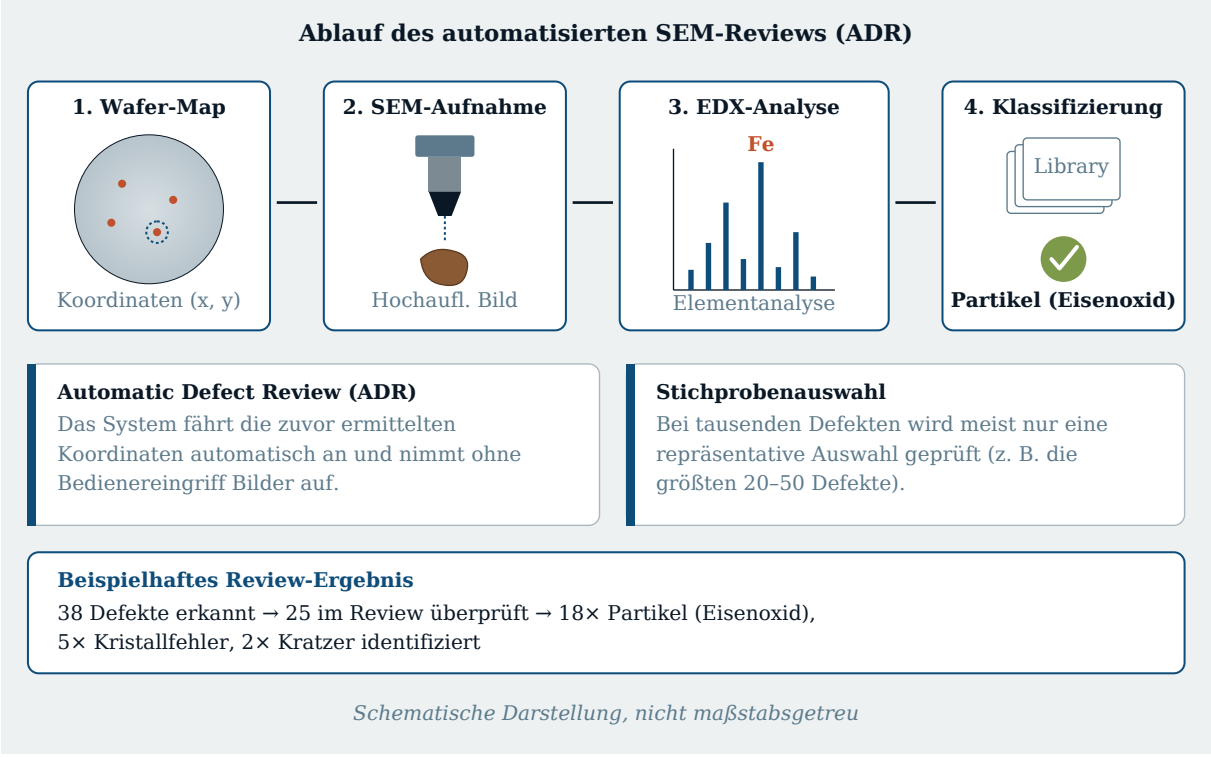
- Punktuelle Cluster an bestimmten Koordinaten: können auf eine defekte Anlagenkomponente hindeuten, die den Wafer immer an derselben Stelle berührt (z.B. ein Wafer-Greifer oder eine Aufnahmevorrichtung).

Die Inspektionssoftware ordnet jedem Defekt anhand von Streulichtsignatur, Größe und Form bereits automatisch eine vorläufige Klasse zu (sogenanntes Automatic Defect Classification, ADC). Diese automatische Vorklassifizierung ist jedoch nicht immer zuverlässig, insbesondere bei neuen oder unbekannten Defekttypen, weshalb kritische oder auffällige Defekte in einem weiteren Schritt genauer untersucht werden müssen.

Review mit dem Rasterelektronenmikroskop (SEM)

Da die optische Streulichtmessung nur Position und eine näherungsweise Größe liefert, jedoch keine verlässliche Aussage über Form und Material des Defekts erlaubt, werden ausgewählte Partikel anschließend im sogenannten Defect Review genauer untersucht. Hierzu wird die Wafer-Map in ein Rasterelektronenmikroskop (SEM) importiert, das die zuvor ermittelten Koordinaten automatisch anfährt (Automatic Defect Review, ADR).

Ablauf des automatisierten SEM-Reviews



Am Zielort angekommen, nimmt das SEM eine hochaufgelöste Aufnahme des Defekts auf. Da bei modernen Inspektionssystemen mehrere tausend Defekte pro Wafer erkannt werden können, ist es in der Praxis nicht möglich, jeden

einzelnen Defekt zu überprüfen. Stattdessen wird meist eine repräsentative Stichprobe (z.B. die größten 20–50 Defekte oder eine zufällige Auswahl je Klasse) automatisiert abgefahren.

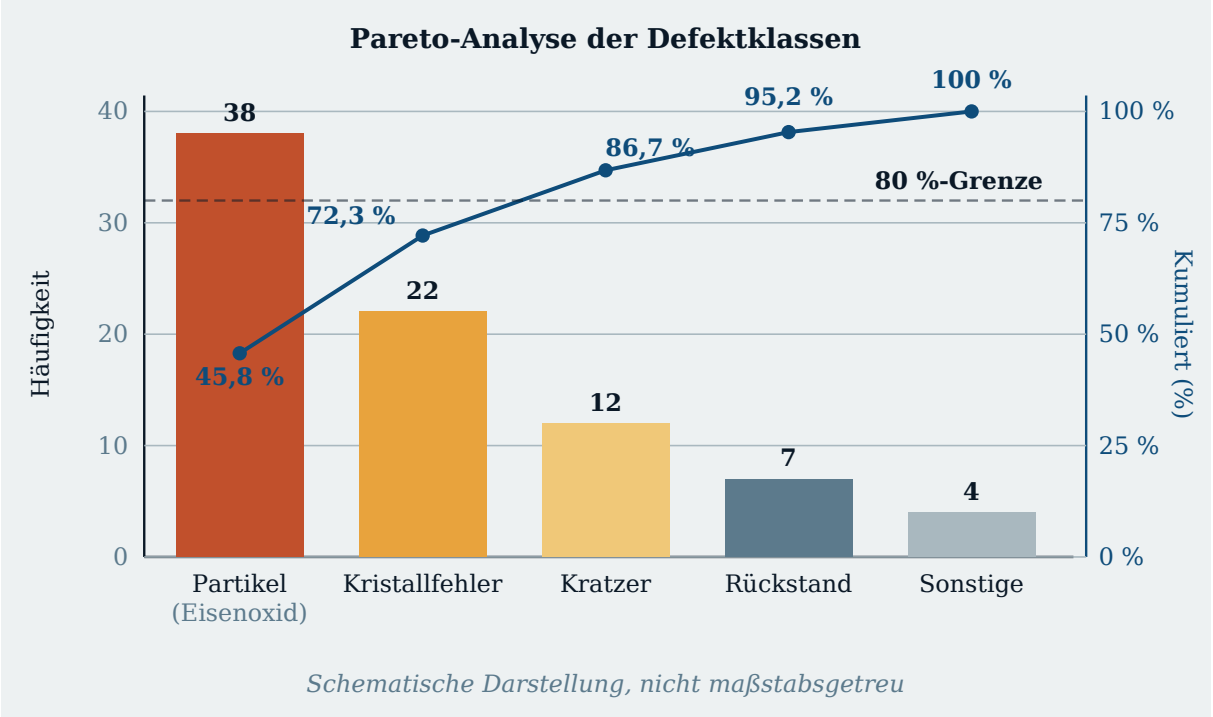
Zusätzlich zur reinen Bildaufnahme verfügen viele SEM-Systeme über einen energiedispersiven Röntgendetektor (Energy Dispersive X-ray Spectroscopy, EDX). Damit kann direkt am Ort des Defekts die chemische Zusammensetzung bestimmt werden, was Rückschlüsse auf die Herkunft zulässt: Wird beispielsweise ein Metall nachgewiesen, das in keinem der aktuellen Prozessschritte verwendet wird, deutet dies auf eine Kontamination durch eine Anlagenkomponente (z.B. abgenutzte Dichtungen, Kammerbeschichtung) hin.

Anhand von Form, Größe und Materialzusammensetzung erfolgt anschließend die endgültige Klassifizierung des Defekts (z.B. Partikel, Kratzer, Kristalldefekt, Rückstand). Diese Klassifizierung wird in der Regel mit historischen Referenzbildern (einer sogenannten Defect Library) abgeglichen, entweder manuell durch einen Techniker oder automatisiert mittels Bilderkennung.

Beurteilung der Messung

Die Ergebnisse der Partikelmessung werden nicht nur für den einzelnen Wafer betrachtet, sondern statistisch über mehrere Wafer, Lose (Lots) und einen längeren Zeitraum ausgewertet. Eine zentrale Kennzahl ist dabei die Gesamtdefektzahl je Wafer (Total Defect Count) oberhalb einer festgelegten Größenschwelle, die mit Kontrollgrenzen (Upper Control Limit, UCL) verglichen wird. Wird diese Grenze überschritten, gilt der Prozess als „out of control“ und muss untersucht werden.

Pareto-Auswertung der Defektklassen über mehrere Wafer



Schematische Darstellung, nicht maßstabsgetreu

Eine gängige Methode zur Priorisierung ist die Pareto-Analyse: Die verschiedenen Defektklassen werden nach ihrer Häufigkeit sortiert dargestellt, sodass sich die Klasse mit dem größten Anteil sofort erkennen lässt. Da erfahrungsgemäß ein Großteil der Defekte auf wenige Hauptursachen zurückgeht, kann die Fehlersuche gezielt auf die häufigsten Klassen konzentriert werden, anstatt jede Einzelursache separat zu verfolgen.

Von besonderer Bedeutung ist zudem die Korrelation zwischen der Defektdichte und der späteren elektrischen Ausbeute (Yield) der Chips. Nicht jeder Defekt führt tatsächlich zu einem Bauteilausfall – ein Partikel, das auf einer unkritischen Fläche liegt (z.B. im Scribe-Bereich zwischen den Chips), hat keinen Einfluss auf die Funktion. Durch den Abgleich von Defektpositionen mit den späteren Testergebnissen der einzelnen Chips (Yield-Map) lässt sich statistisch ermitteln, welche Defektklassen und -größen tatsächlich yieldrelevant sind. Diese sogenannte Defect-to-Yield-Korrelation ist eine der wichtigsten Aufgaben der Fehleranalyse in der Halbleiterfertigung, da sie es erlaubt, Ressourcen gezielt auf die tatsächlich kritischen Defektquellen zu konzentrieren, statt jede gefundene Partikelklasse gleich intensiv zu verfolgen.

Langfristig fließen die gewonnenen Erkenntnisse in ein kontinuierliches Verbesserungsprogramm ein: Anlagen mit erhöhter Partikelbelastung werden gezielt gewartet oder requalifiziert, Prozessparameter angepasst und die Reinraumbedingungen bei Bedarf verbessert, um die Defektdichte und damit letztlich die Fertigungskosten pro funktionsfähigem Chip zu senken.

CD-Messung

Messtechnik

Die Critical Dimension (CD) bezeichnet die kleinste, für die Funktion eines Bauteils entscheidende Strukturbreite auf dem Wafer – klassischerweise die Gate-Länge eines Transistors, aber auch Leiterbahnbreiten, Spacer-Dicken oder Kontaktloch-Durchmesser. Da die CD direkt in elektrische Kenngrößen wie Schaltgeschwindigkeit, Leckstrom und Schwellspannung eingeht, muss sie mit einer Genauigkeit im niedrigen Nanometer- bzw. sogar Sub-Nanometerbereich kontrolliert werden.

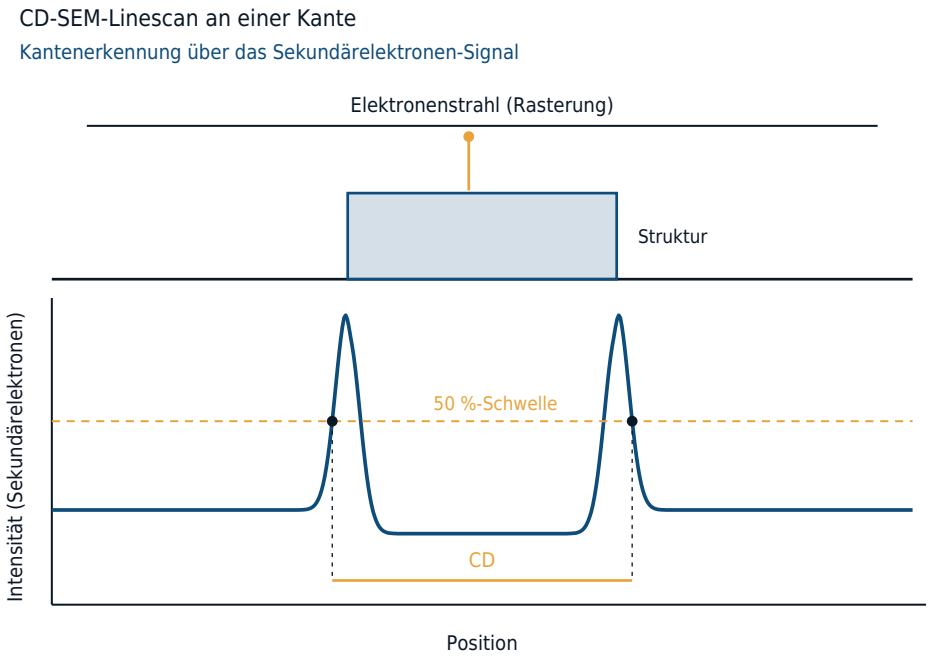
Die CD-Kontrolle erfolgt an mehreren Stellen der Prozesskette: direkt nach der Lithografie (Photoresist-CD, um Belichtungsdosis und Fokus zu überwachen), nach dem Ätzen (Final-CD, da sich die Struktur beim Ätzen weiter verändert) sowie stichprobenartig nach weiteren kritischen Schritten. Man unterscheidet dabei zwischen der Bias, also der systematischen Abweichung zwischen Maskenmaß und tatsächlicher Struktur, und der CD-Uniformity, also der Streuung der Messwerte über den Wafer, zwischen Wafern und zwischen Losen.

Für die eigentliche Messung stehen mehrere Verfahren zur Verfügung, die sich in Auflösung, Geschwindigkeit und Zerstörungsfreiheit unterscheiden. In der Fertigung dominieren zwei Ansätze: die direkte, bildgebende Messung mittels Rasterelektronenmikroskop (CD-SEM) und die indirekte, modellbasierte optische Messung (OCD/Scatterometrie).

CD-SEM – Messprinzip

Ein CD-SEM ist ein auf Durchsatz und Reproduzierbarkeit optimiertes Rasterelektronenmikroskop, das speziell für die automatisierte Vermessung von Strukturbreiten in der Halbleiterfertigung ausgelegt ist. Ein fein fokussierter Elektronenstrahl rastert die Zielstruktur zeilenweise ab; an den Kanten der Struktur ändert sich die Ausbeute an Sekundärelektronen sprunghaft (Kanteneffekt), da dort zusätzliche Elektronen aus den Seitenwänden austreten können.

Sekundärelektronen-Signal beim Abrastern einer Kante



Aus dem resultierenden Intensitätsprofil (Linescan) wird die Kantenposition über einen Schwellwert-Algorithmus bestimmt, beispielsweise als der Punkt, an dem das Signal 50 % zwischen Grundlinie und Peak erreicht. Da Form und Höhe des Sekundärelektronen-Peaks jedoch vom Kantenprofil (senkrecht, geneigt, abgerundet) und vom Material abhängen, liefern unterschiedliche Auswertealgorithmen für dieselbe physikalische Kante leicht unterschiedliche CD-Werte – die Wahl und Konsistenz des Algorithmus ist daher Teil der Messrezept-Qualifizierung.

Um Aufladungseffekte und Strahlschäden an empfindlichen Materialien (insbesondere Photoresist) zu vermeiden, wird mit niedriger Beschleunigungsspannung und möglichst geringer Dosis gearbeitet. Moderne CD-SEM-Systeme erreichen dabei eine Wiederholpräzision im Bereich von unter einem Nanometer.

Kalibrierung und Referenzmetrologie

Da der CD-SEM-Messwert vom verwendeten Auswertealgorithmus und von Gerät zu Gerät leicht abhängt, muss er regelmäßig gegen rückführbare Referenzstandards kalibriert werden. Solche Standards (Reference Materials) besitzen eine zertifizierte, auf nationale Normale zurückgeführte CD und erlauben es, systematische Abweichungen (Bias) zwischen mehreren CD-SEM-Geräten in einer Fertigung zu erfassen und zu korrigieren (Tool-Matching).

Für die absolute Validierung, insbesondere bei neuen Prozessen oder ungewöhnlichen Strukturprofilen, wird zusätzlich destruktive

Referenzmetrologie eingesetzt: Aus dem Wafer wird eine dünne Lamelle präpariert (z.B. mittels Focused Ion Beam) und im Transmissionselektronenmikroskop (TEM) im Querschnitt vermessen. Der TEM-Querschnitt zeigt die tatsächliche Geometrie inklusive Seitenwandwinkel und Profilform und dient damit als Referenz, gegen die das schnellere, zerstörungsfreie CD-SEM-Messrezept abgeglichen wird.

Da eine TEM-Querschnittsmessung aufwändig, langsam und zerstörend ist, kommt sie nur stichprobenartig bei der Rezept-Entwicklung oder bei Abweichungen zum Einsatz – die laufende Fertigungsüberwachung erfolgt praktisch ausschließlich mit CD-SEM und/oder OCD.

OCD / Scatterometrie als Alternative

Neben dem bildgebenden CD-SEM hat sich die Optical Critical Dimension (OCD) Metrologie, auch Scatterometrie genannt, als schnelle und zerstörungsfreie Ergänzung etabliert. Dabei wird eine periodische Teststruktur (z.B. ein Linien-Gitter) mit polarisiertem Licht unter definiertem Winkel beleuchtet; die Struktur wirkt wie ein optisches Beugungsgitter und erzeugt ein charakteristisches Reflexionsspektrum.

Anders als bei der Interferometrie zur Schichtdickenmessung wird dieses Spektrum nicht direkt in eine Dicke umgerechnet, sondern gegen ein rechnerisches Modell der Struktur gefittet: Ausgehend von einem parametrischen Modell (CD, Höhe, Seitenwandwinkel, Schichtdicken) wird das erwartete Spektrum simuliert und iterativ so lange angepasst, bis es mit dem gemessenen Spektrum übereinstimmt. Die Modellparameter am besten passenden Fits ergeben die gesuchten geometrischen Größen.

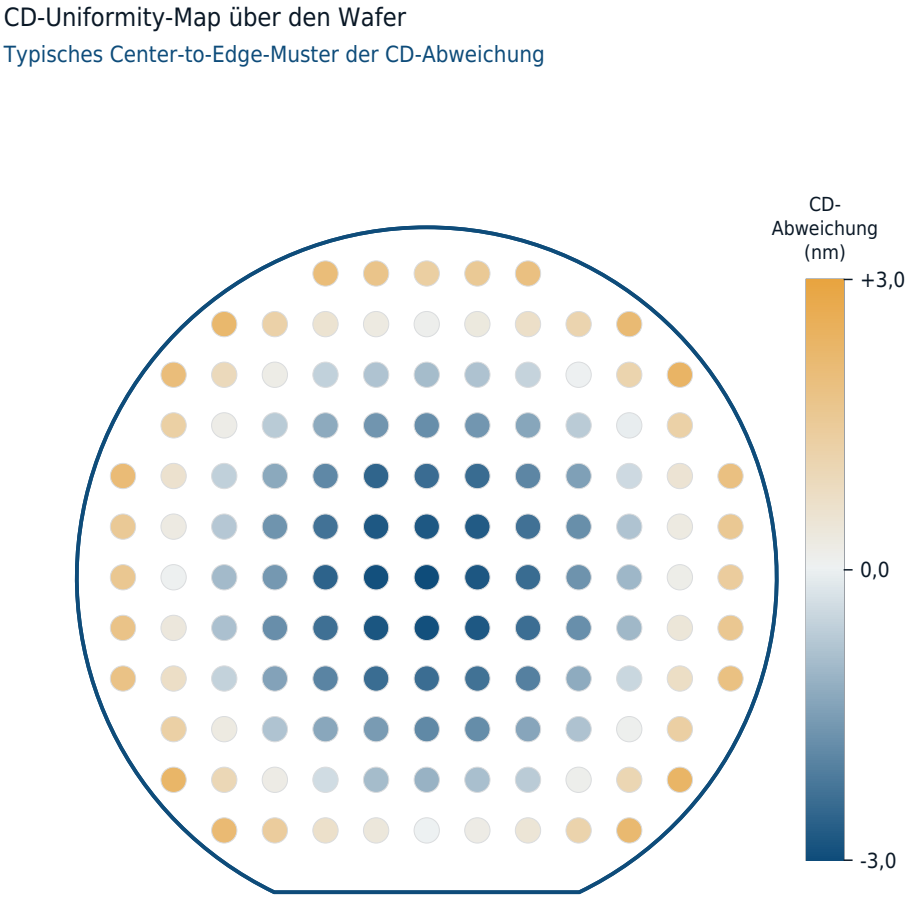
OCD ist deutlich schneller als CD-SEM und liefert zusätzlich zur reinen Linienbreite auch Seitenwandwinkel und Höhe in einer einzigen Messung, benötigt dafür aber eine ausgedehnte periodische Teststruktur (typischerweise in der Scribe Line) und ein präzises optisches Modell – bei unerwarteten Profiländerungen kann ein falsches Modell zu einer scheinbar präzisen, aber falschen CD führen. In der Praxis werden CD-SEM und OCD daher komplementär eingesetzt: OCD für hohen Durchsatz in der Serienüberwachung, CD-SEM zur Modellvalidierung und für Strukturen außerhalb periodischer Testfelder.

CD-Uniformity und Prozesskontrolle

Die einzelnen CD-Messwerte fließen in die statistische Prozessregelung (SPC) ein. Dabei wird zwischen mehreren Uniformity-Ebenen unterschieden: der Within-Die-Uniformity (Streuung innerhalb eines einzelnen Chips), der Within-

Wafer-Uniformity (Streuung über den gesamten Wafer, oft mit einem charakteristischen Rand-Zentrum-Muster) sowie der Wafer-to-Wafer- und Lot-to-Lot-Uniformity.

Typisches CD-Uniformity-Muster über den Wafer



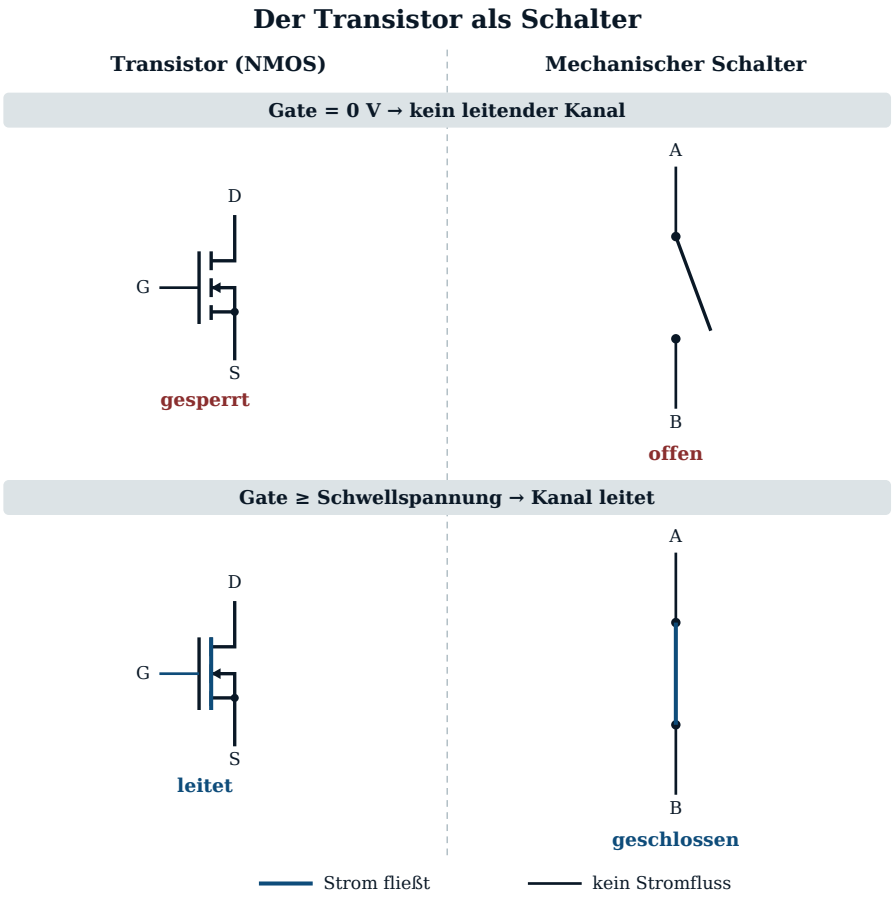
Überschreitet die gemessene CD oder ihre Streuung die festgelegten Regelgrenzen, wird dies als Prozessabweichung gewertet und löst eine Rückkopplung (Feedback Loop) zu den vorgelagerten Prozessschritten aus: Bei einer Abweichung direkt nach der Lithografie werden Belichtungsdosis oder Fokus der Scanner-Anlage nachjustiert (Advanced Process Control, APC), bei einer Abweichung nach dem Ätzen die Ätzzeit oder Gaszusammensetzung. Durch diese automatisierte Rückkopplung wird die CD über viele aufeinanderfolgende Lose stabil im spezifizierten Prozessfenster gehalten, ohne dass für jede Anpassung ein manueller Eingriff nötig ist.

Vom Transistor zum Gatter

Der Transistor als Schalter

Ein MOSFET lässt sich auf seine einfachste Funktion reduzieren: einen spannungsgesteuerten Schalter. Zwischen Source und Drain fließt Strom oder nicht, je nachdem, welche Spannung am Gate anliegt. Damit unterscheidet er sich grundlegend von einem mechanischen Schalter — es bewegt sich nichts, es gibt keinen Verschleiß, und der Schaltvorgang dauert nur einen Bruchteil einer Nanosekunde.

Bei einem selbstsperrenden NMOS-Transistor liegt ohne Gate-Spannung kein leitender Kanal zwischen Source und Drain vor. Erst wenn die Gate-Spannung die Schwellspannung überschreitet, bildet sich der Kanal, und der Transistor leitet. Man kann sich das Gate als Steuereingang eines Schalters vorstellen: 0 V am Gate → Schalter offen, ausreichend positive Spannung am Gate → Schalter geschlossen.



NMOS und PMOS im Vergleich

Neben dem NMOS-Transistor gibt es sein Gegenstück, den PMOS-Transistor. Er verhält sich spiegelverkehrt: Während der NMOS bei High-Pegel am Gate leitet, sperrt der PMOS genau dann — und leitet stattdessen bei Low-Pegel. Diese Komplementarität ist keine Randnotiz, sondern das Grundprinzip, auf dem die gesamte CMOS-Technik ("Complementary MOS") beruht.

Der NMOS wird typischerweise als Verbindung zur Masse (0 V) eingesetzt, der PMOS als Verbindung zur Versorgungsspannung. Genau dieses Zusammenspiel ermöglicht es, mit beiden Transistortypen gemeinsam einen Ausgang sauber zwischen High und Low umzuschalten, ohne dass jemals ein direkter Kurzschluss zwischen Versorgung und Masse entsteht.

NMOS und PMOS im Vergleich

Gate-Spannung	NMOS	PMOS
Low (0 V)	sperrt	leitet
High (VDD)	leitet	sperrt

Der Inverter — das erste Gatter

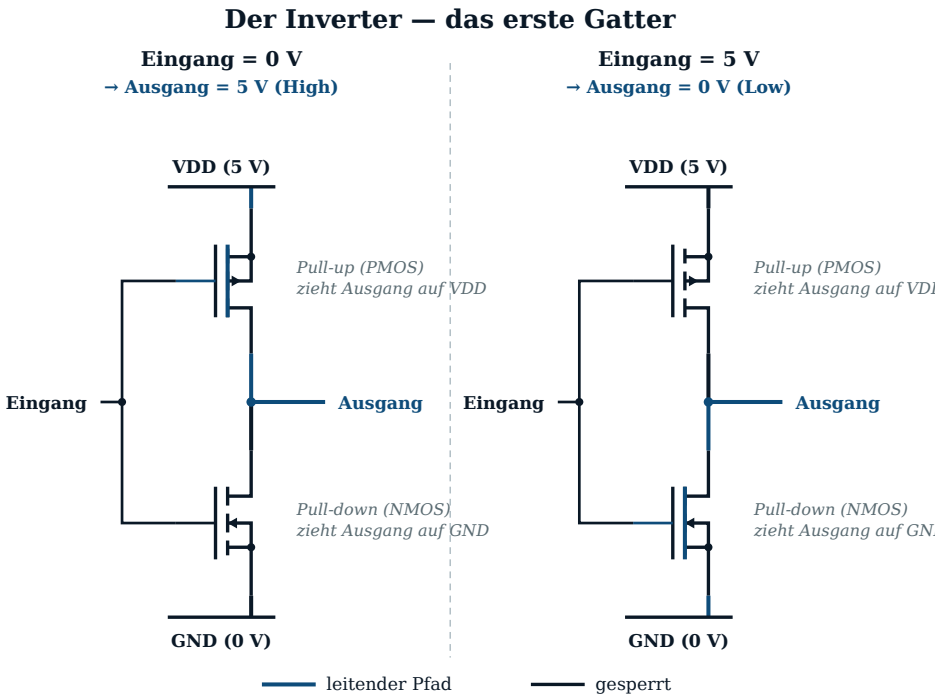
Verschaltet man einen PMOS und einen NMOS wie folgt, entsteht das einfachste aller logischen Gatter: der Inverter (NOT-Gatter).

- Der PMOS liegt zwischen Versorgungsspannung (VDD) und dem Ausgang.
- Der NMOS liegt zwischen dem Ausgang und Masse (GND).
- Beide Gates sind miteinander verbunden und bilden den Eingang.

Legt man am Eingang Low (0 V) an, sperrt der NMOS, aber der PMOS leitet — der Ausgang wird über den PMOS mit VDD verbunden und liegt auf High. Legt man High an, kehrt sich das Bild um: Der PMOS sperrt, der NMOS leitet, der Ausgang wird auf Masse gezogen und liegt auf Low. Der Ausgang ist also stets das Gegenteil des Eingangs — daher der Name Inverter.

Bemerkenswert ist, dass in keinem der beiden stabilen Zustände Strom von VDD nach GND fließt, da immer genau einer der beiden Transistoren sperrt. Das ist der Grund, warum CMOS-Schaltungen im Ruhezustand nur sehr wenig Leistung verbrauchen.

Damit übernehmen die beiden Transistoren im Inverter zugleich die klassischen Rollen von Pull-up und Pull-down: Der PMOS zieht den Ausgang aktiv auf VDD (Pull-up), der NMOS zieht ihn aktiv auf GND (Pull-down). Nie sind beide gleichzeitig aktiv — genau das verhindert den Kurzschluss und sorgt gleichzeitig dafür, dass der Ausgang in jedem Zustand einen klar definierten Pegel hat, statt hochohmig ("floating") in der Luft zu hängen.

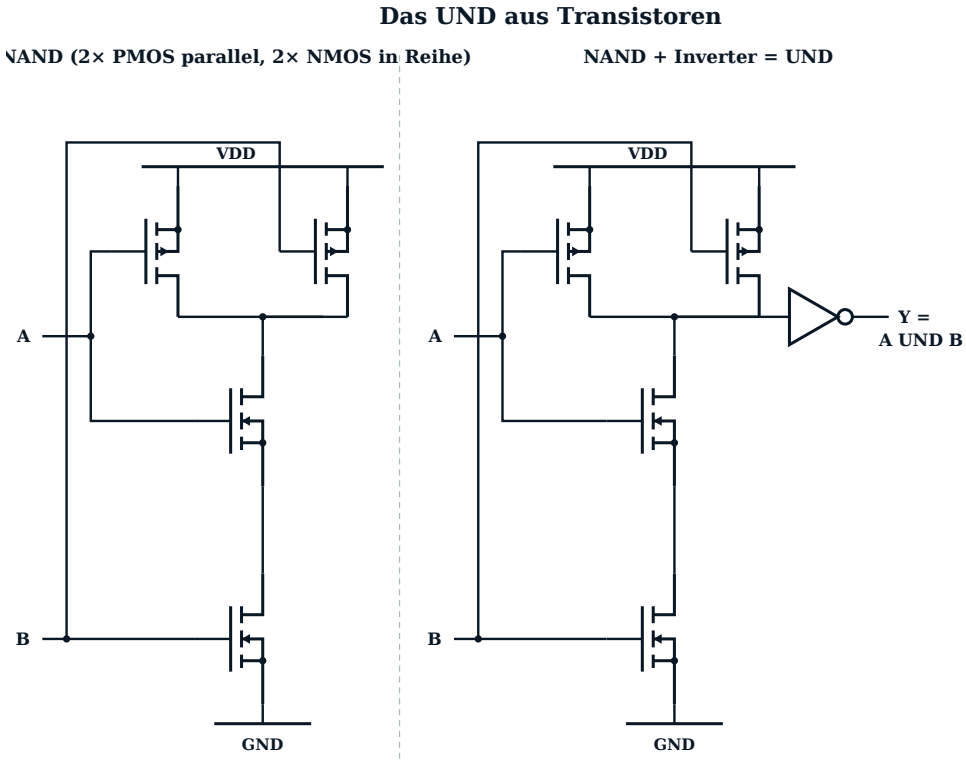


Das UND aus Transistoren

Aus zwei einzelnen Schaltern lässt sich mit wenigen Handgriffen eine UND-Verknüpfung bauen. Das Prinzip: Zwei NMOS-Transistoren werden in Reihe zwischen Ausgang und Masse geschaltet. Nur wenn *beide* Eingänge High sind, leiten *beide* Transistoren, und der Weg zur Masse ist frei — der Ausgang wird auf Low gezogen. Ist auch nur einer der Eingänge Low, ist die Reihe unterbrochen.

Parallel dazu liegen zwei PMOS-Transistoren zwischen Ausgang und Versorgungsspannung. Sind beide Eingänge High, sperren beide PMOS, keiner zieht den Ausgang nach oben. Ist mindestens ein Eingang Low, leitet der zugehörige PMOS und zieht den Ausgang auf High.

Diese Schaltung ist zunächst ein NAND-Gatter (Ausgang Low nur, wenn beide Eingänge High sind — also die Umkehrung von UND). Erst ein nachgeschalteter Inverter macht daraus ein echtes UND-Gatter. Das ist kein Zufall: In der CMOS-Technik sind NAND- und NOR-Gatter die „natürlichen“ Grundbausteine, weil sie sich direkt aus Reihen- und Parallelschaltungen von Transistoren ergeben — UND und ODER entstehen erst durch einen zusätzlichen Inverter am Ausgang.



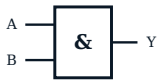
Logische Gatter im Überblick

Aus den Grundschaltungen der letzten Kapitel lassen sich alle gängigen logischen Gatter ableiten. Die folgende Übersicht zeigt Schaltsymbol und Wahrheitstabelle der wichtigsten Gatter:

Gatter	Ausgang = 1, wenn …
UND (AND)	beide Eingänge 1 sind
ODER (OR)	mindestens ein Eingang 1 ist
NICHT (NOT)	Eingang 0 ist
NAND	nicht beide Eingänge 1 sind
NOR	kein Eingang 1 ist
XOR	die Eingänge unterschiedlich sind

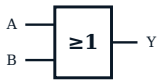
Logische Gatter im Überblick

Symbole nach IEC 60617



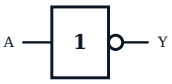
UND

A	B	Y
0	0	0
0	1	0
1	0	0
1	1	1



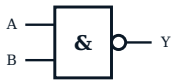
ODER

A	B	Y
0	0	0
0	1	1
1	0	1
1	1	1



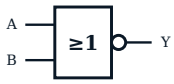
NICHT

A	Y
0	1
1	0



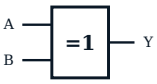
NAND

A	B	Y
0	0	1
0	1	1
1	0	1
1	1	0



NOR

A	B	Y
0	0	1
0	1	0
1	0	0
1	1	0



XOR

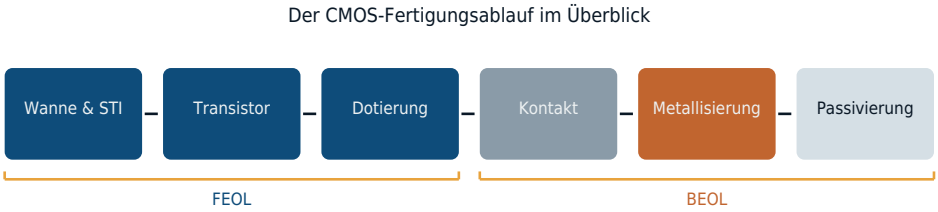
A	B	Y
0	0	0
0	1	1
1	0	1
1	1	0

Der CMOS-Fertigungsablauf

Front-End und Back-End of Line

Ein fertiger Chip ist das Ergebnis von oft mehreren hundert einzelnen Prozessschritten, die sich grob in zwei große Phasen einteilen lassen: das Front-End of Line (FEOL) und das Back-End of Line (BEOL). Im FEOL entstehen die eigentlichen Transistoren – von der Wanne über das Gate bis zur Dotierung. Im BEOL werden diese Transistoren anschließend über mehrere Metalllagen zu einer funktionierenden Schaltung verdrahtet.

Die einzelnen Detailkapitel dieser Seite – etwa zur FET-Fertigung oder zur FinFET-Technologie – zeigen jeweils einen Ausschnitt dieses Ablaufs in großer Tiefe. Dieses Kapitel verfolgt einen anderen Zweck: Es ordnet die Einzelschritte in den Gesamtprozess ein und zeigt, wie FEOL und BEOL ineinandergreifen.



Wafer und Wannen

Ausgangspunkt jeder CMOS-Fertigung ist der Silizium-Wafer – eine dünne, hochreine Scheibe aus einkristallinem Silizium. Auf ihm werden zunächst die Wannen (Wells) angelegt: Bereiche mit gezielt eingebrachter n- oder p-Dotierung, in denen später die komplementären Transistortypen entstehen. In einer n-Wanne wachsen PMOS-Transistoren, im umgebenden p-dotierten Substrat (oder einer p-Wanne) die NMOS-Transistoren.

Damit sich benachbarte Transistoren nicht gegenseitig beeinflussen, werden sie durch Shallow Trench Isolation (STI) voneinander getrennt: schmale, mit Oxid gefüllte Gräben, die elektrisch isolierend wirken und gleichzeitig die aktiven Gebiete klar definieren, in denen später die Transistoren entstehen.

Transistorfertigung im FEOL

Auf den vorbereiteten aktiven Gebieten entsteht anschließend der eigentliche Transistor: Gate-Oxid, Gate-Elektrode, Spacer und die charakteristische Struktur aus Source, Drain und Kanal. Dieser Ablauf ist im Detail bereits in der FET-Fertigungsserie Schritt für Schritt beschrieben und wird hier nicht erneut aufgerollt.

Wichtig für die Gesamtübersicht ist vor allem der zeitliche Rahmen: Alle Schritte bis einschließlich der Gate-Strukturierung gehören zum FEOL. Erst danach beginnt der Übergang zum BEOL.

Dotierung und Silizidierung

Nach der Gate-Strukturierung folgt die Dotierung von Source und Drain – meist durch Ionenimplantation, bei der Dotieratome mit hoher Energie gezielt in den Wafer geschossen werden. Seitlich am Gate angebrachte Spacer sorgen dafür, dass diese Implantation exakt versetzt zum Kanal erfolgt und ein sauber definierter Übergang entsteht.

Im Anschluss wird an den freiliegenden Silizium-Oberflächen – Source, Drain

und Gate – eine dünne Metallschicht (häufig Nickel oder Cobalt) aufgebracht und getempert. Es entsteht ein niederohmiges Metallsilizid, das sogenannte Salicid (self-aligned silicide). Es reduziert den Kontaktwiderstand erheblich und markiert zugleich das Ende des FEOL.

Kontaktierung

Mit dem fertigen Transistor beginnt der Übergang ins BEOL. Zunächst wird eine isolierende Schicht (meist ein Oxid) über der gesamten Struktur abgeschieden und eingeebnet. In diese Schicht werden anschließend winzige Kontaktlöcher geätzt, die exakt auf Source, Drain und Gate treffen und anschließend mit Wolfram gefüllt werden.

Diese Kontakte bilden die erste elektrische Verbindung zwischen dem Transistor und der darüberliegenden Verdrahtungsebene – dem eigentlichen BEOL.

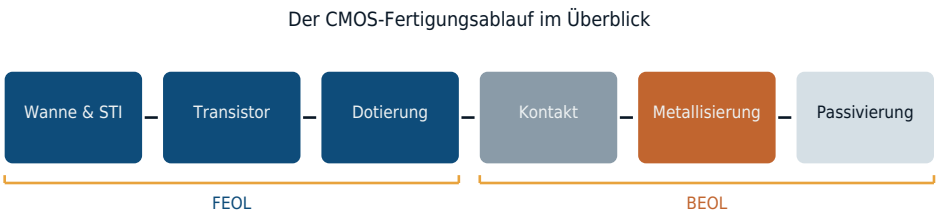
Metallisierung im BEOL

Ab hier wiederholt sich ein Muster über mehrere Lagen hinweg: In eine Oxidschicht werden Gräben und Vias geätzt, die anschließend im sogenannten Damascene-Verfahren mit Kupfer gefüllt werden. Überschüssiges Kupfer wird per chemisch-mechanischem Polieren (CMP) wieder entfernt, sodass eine plane Oberfläche für die nächste Lage entsteht.

Moderne Prozesse verwenden auf diese Weise acht bis fünfzehn Metalllagen – von sehr feinen unteren Lagen für die lokale Verdrahtung bis zu breiten oberen Lagen für die Stromversorgung. Den Abschluss bildet eine Passivierungsschicht, die den Chip vor Feuchtigkeit und mechanischer Beschädigung schützt.

Der Ablauf im Überblick

Die folgende Übersicht fasst den gesamten Ablauf von der Wanne bis zur fertigen Verdrahtung noch einmal zusammen und ordnet die einzelnen Schritte den beiden Hauptphasen zu:



Wer einzelne Schritte vertiefen möchte, findet sie in den entsprechenden Detailkapiteln: die Transistorfertigung in der FET-Serie, die FinFET-Variante in der FinFET-Fertigung, sowie Layout- und Verifikationsthemen in den Kapiteln zu Design Rules und Layout versus Schematic.

Die SRAM-Zelle

Vom Schaltbild zum Chip: EDA

Alle Schaltbilder dieses Bereichs — vom einzelnen Transistor bis zur SRAM-Zelle — sind mit den Mitteln der Electronic Design Automation (EDA) entstanden: der Software-Werkzeugkette, mit der integrierte Schaltungen heute entworfen werden. Ein EDA-Schaltplan wie der hier gezeigte ist dabei mehr als eine Illustration — er ist die Eingabe für die nächsten Entwurfsschritte: aus ihm entsteht per Simulation die Verifikation des elektrischen Verhaltens, per Place & Route das physische Layout auf dem Wafer, und per Verifikation der Abgleich zwischen beiden. Eine SRAM-Zelle ist dafür ein besonders anschauliches Beispiel, weil sie in modernen Chips millionenfach identisch wiederholt wird — genau die Art von regelmäßiger, hochoptimierter Struktur, für die EDA-Werkzeuge unverzichtbar sind.

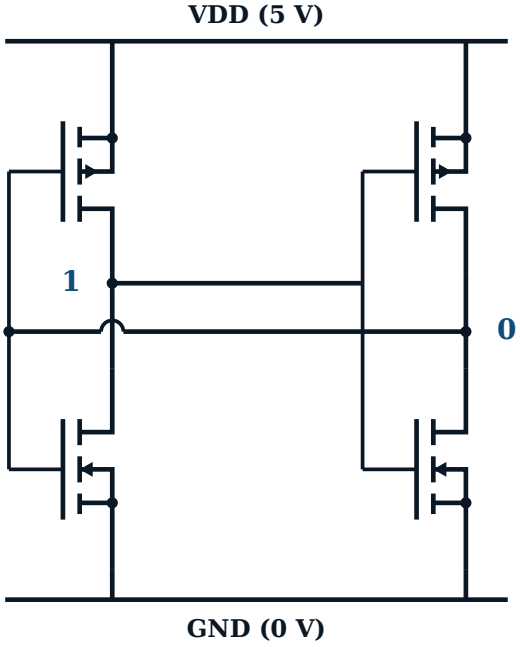
Ein einzelnes Bit zu speichern klingt zunächst nach einer trivialen Aufgabe — tatsächlich steckt genau darin einer der am häufigsten verbauten Schaltungsblöcke überhaupt. Jeder Prozessor-Cache, jedes Register-File besteht aus Millionen bis Milliarden solcher Zellen. Die SRAM-Zelle (Static Random Access Memory) speichert ihr Bit nicht als Ladung auf einem Kondensator wie DRAM, sondern als stabilen Schaltzustand zweier rückgekoppelter Inverter. Solange die Betriebsspannung anliegt, bleibt der gespeicherte Wert erhalten — ohne Refresh, aber auch ohne die hohe Packungsdichte von DRAM, da für ein einziges Bit sechs statt einem Transistor benötigt werden.

Zwei rückgekoppelte Inverter als Speicherkern

Das Herzstück der Zelle sind zwei CMOS-Inverter (aus Kapitel 3 bekannt: je ein PMOS als Pull-up, ein NMOS als Pull-down), deren Ein- und Ausgänge über Kreuz verbunden sind — der Ausgang des einen Inverters treibt den Eingang des anderen. Diese Rückkopplung erzeugt zwei stabile Zustände: Liegt Knoten 1 auf High, zwingt das den zweiten Inverter dazu, Knoten 0 auf Low zu halten, was wiederum den ersten Inverter bestätigt, Knoten 1 auf High zu halten. Der Zustand stabilisiert sich selbst, ganz ohne Takt oder Refresh — daher "statisch". Ein Umkippen ist nur möglich, wenn von außen mit ausreichend Kraft gegen

diese Rückkopplung geschrieben wird.

Speicherkern: zwei rückgekoppelte Inverter



Die Zugriffstransistoren

Die beiden gespeicherten Knoten 1 und 0 liegen nicht direkt an den Ein-/Ausgabeleitungen der Zelle, sondern über je einen NMOS-Zugriffstransistor an den Bitleitungen BL und B$\overline{\text{L}}$. Deren Gates sind gemeinsam mit der Wortleitung (Word Line, WL) verbunden. Ist WL auf Low, sperren beide Zugriffstransistoren vollständig — die Zelle ist von den Bitleitungen getrennt und hält ihren Zustand isoliert. Erst wenn WL auf High gezogen wird, öffnen beide Transistoren gleichzeitig und verbinden Knoten 1 mit BL sowie Knoten 0 mit B$\overline{\text{L}}$. Über diese eine gemeinsame Wortleitung lässt sich später eine ganze Zeile von Zellen im Speicherarray gleichzeitig adressieren.

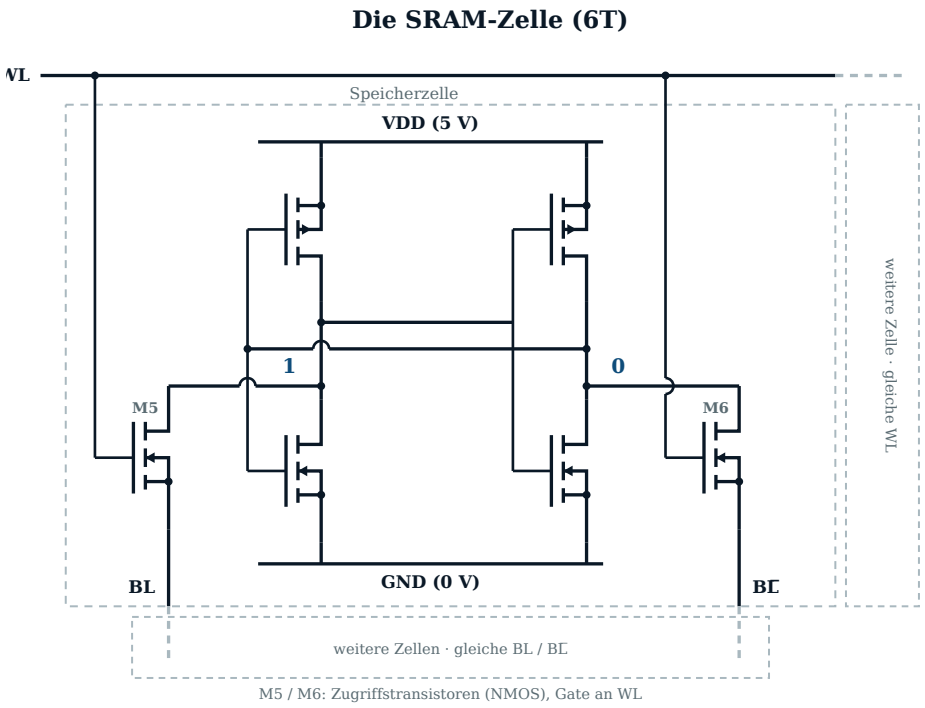
Schreiben

Zum Schreiben werden BL und B$\overline{\text{L}}$ von außen mit invertiertem Wertepaar sehr niederohmig auf die gewünschten Pegel getrieben (z. B. BL = 5 V, B$\overline{\text{L}}$ = 0 V für eine "1"), während WL aktiviert wird. Die externen Treiber sind bewusst stärker ausgelegt als die Rückkopplungsschleife der Zelle — sie überstimmen den bisherigen Zustand über die Zugriffstransistoren und zwingen die Inverter in den neuen, gewünschten Zustand. Sobald WL wieder auf Low fällt, halten die

beiden Inverter den frisch geschriebenen Wert eigenständig.

Vom einzelnen Bit zum Array

Die Wortleitung (WL) einer Zelle wird nicht individuell verdrahtet, sondern durchläuft eine ganze Zeile benachbarter Zellen — jede Zelle dieser Zeile teilt sich dieselbe WL. Senkrecht dazu läuft jedes Bitleitungspaar (BL/BL̄) durch alle Zellen einer Spalte. So entsteht ein Raster: Nur die Zelle, an deren Kreuzungspunkt sowohl WL aktiv ist als auch die passenden Bitleitungen anliegen, wird beim Schreiben oder Lesen tatsächlich angesprochen — alle anderen Zellen derselben Spalte bleiben durch ihre eigene, inaktive Wortleitung isoliert, obwohl sie an denselben Bitleitungen hängen. Über diese Adressierung per Zeile/Spalte lässt sich mit vergleichsweise wenigen Leitungen ein Array aus Millionen Zellen gezielt einzeln ansprechen.



Lesen

Vor jedem Lesezugriff werden beide Bitleitungen zunächst auf ein gemeinsames Mittelpotenzial vorgeladen (Precharge). Aktiviert man dann WL, zieht die Zelle über ihre Zugriffstransistoren eine der beiden Leitungen leicht in Richtung des gespeicherten Zustands, während die andere nahezu auf dem Precharge-Pegel

verbleibt — es entsteht ein kleines Spannungsdifferential zwischen BL und B̄L̄. Ein nachgeschalteter Leseverstärker (Sense Amplifier) erkennt bereits diesen geringen Unterschied und verstärkt ihn zu einem sauberen Digitalpegel. Wichtig dabei: Der Lesevorgang selbst greift den gespeicherten Zustand nur schwach an — die Zelle bleibt dabei im Gegensatz zu DRAM unverändert erhalten.

Die DRAM-Zelle

Ein Kondensator statt sechs Transistoren

Wo die SRAM-Zelle ihr Bit als stabilen Schaltzustand zweier rückgekoppelter Inverter hält, geht die DRAM-Zelle (Dynamic Random Access Memory) einen fundamental anderen Weg: Sie speichert die Information als elektrische Ladung auf einem einzelnen Kondensator. Ein einzelner NMOS-Transistor genügt, um diesen Kondensator gezielt mit der Bitleitung zu verbinden oder von ihr zu trennen – daher die Bezeichnung 1T1C-Zelle (ein Transistor, ein Kondensator).

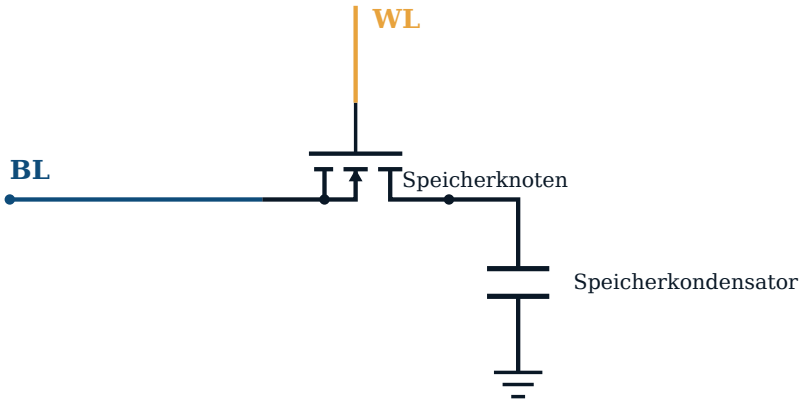
Diese radikale Reduktion auf zwei Bauelemente statt sechs macht DRAM erheblich dichter und günstiger pro Bit als SRAM – Hauptspeicher wird deshalb praktisch ausschließlich in DRAM ausgeführt. Der Preis dafür: Die gespeicherte Ladung ist nicht stabil. Sie fließt über parasitäre Leckpfade langsam ab, weshalb jede Zelle regelmäßig aufgefrischt werden muss, um den Inhalt nicht zu verlieren – daher „dynamisch“.

Zugriffstransistor, Wortleitung und Bitleitung

Wie bei der SRAM-Zelle übernehmen Wortleitung (WL) und Bitleitung (BL) die Ansteuerung – nur mit einem einzigen Zugriffstransistor statt zweien. Sein Gate liegt an der Wortleitung, sein Source-Drain-Pfad verbindet den Speicherkondensator mit der Bitleitung. Ist WL auf Low, sperrt der Transistor vollständig, und der Kondensator bleibt elektrisch isoliert – die gespeicherte Ladung kann (abgesehen von Leckströmen) nirgendwo hinfließen. Erst wenn WL aktiviert wird, öffnet der Zugriffstransistor und stellt die Verbindung zur Bitleitung her.

Zum Schreiben wird die Bitleitung von außen niederohmig auf den gewünschten Pegel getrieben, während WL aktiv ist – der Kondensator lädt oder entlädt sich entsprechend über den geöffneten Zugriffstransistor. Sobald WL wieder auf Low fällt, ist der Kondensator erneut isoliert und hält seine Ladung – vorerst.

Die DRAM-Zelle (1T1C)



Leckströme und der Ladungszerfall

Der Speicherkondensator einer DRAM-Zelle ist winzig – typischerweise nur wenige Femtofarad. Selbst kleinste Leckströme über den gesperrten Zugriffstransistor, über das Dielektrikum des Kondensators oder über den pn-Übergang zum Substrat genügen, um die gespeicherte Ladung innerhalb weniger Millisekunden signifikant abfließen zu lassen. Ohne Gegenmaßnahme würde der gespeicherte Wert schlicht verschwinden – ein Zustand, den keine noch so ausgeklügelte Schaltung tolerieren kann.

Genau dieses Ladungsproblem unterscheidet DRAM fundamental von SRAM: Während die rückgekoppelten Inverter einer SRAM-Zelle ihren Zustand aktiv gegen Störungen verteidigen, ist eine DRAM-Zelle rein passiv – sie kann ihren Inhalt nicht selbst erhalten.

Der Refresh-Zyklus

Um den Ladungsverlust auszugleichen, liest die Speichersteuerung jede Zelle in regelmäßigen Abständen aus und schreibt den erkannten Wert unmittelbar wieder zurück – ein Vorgang, der als Refresh bezeichnet wird. Typische DRAM-Bausteine müssen jede Zelle etwa alle 64 Millisekunden auffrischen, moderne dichte Zellen teils noch häufiger. Da ein einzelner Lesezugriff bereits eine ganze Zeile des Arrays gleichzeitig aktiviert, lässt sich der Refresh effizient zeilenweise durchführen, nicht Zelle für Zelle.

Dieser Refresh kostet Zeit und Energie: Während eine Zeile aufgefrischt wird,

steht sie für reguläre Lese- oder Schreibzugriffe nicht zur Verfügung. Bei sehr großen Speicherbausteinen mit Milliarden Zellen macht der Refresh-Overhead einen spürbaren Anteil der Gesamt-Bandbreite und -Leistungsaufnahme aus.

Der Sense Amplifier

Beim Lesen entlädt sich der winzige Speicherkondensator kurzzeitig über die Bitleitung, die zuvor auf ein Precharge-Mittelpotenzial vorgeladen wurde – die resultierende Spannungsänderung ist minimal, oft nur wenige zehn Millivolt. Ein Sense Amplifier am Ende jeder Bitleitung erkennt dieses winzige Signal und verstärkt es zu einem sauberen digitalen Pegel.

Diese Verstärkung erfüllt gleich zwei Aufgaben zugleich: Sie liefert nicht nur den gelesenen Wert nach außen, sondern treibt über die aktivierte Wortleitung auch den ursprünglichen Kondensator wieder auf seinen vollen Pegel zurück. Jeder Lesezugriff ist damit automatisch auch ein Refresh der gerade aktiven Zeile – anders als bei SRAM, wo Lesen den Zustand praktisch nicht beeinflusst, ist das Auslesen einer DRAM-Zelle destruktiv und macht dieses „Rewrite“ zwingend erforderlich.

DRAM und SRAM im Vergleich

Die grundlegend unterschiedlichen Speicherprinzipien führen zu sehr unterschiedlichen Eigenschaften, die jeweils für unterschiedliche Einsatzgebiete prädestinieren:

	SRAM	**DRAM**
Transistoren pro Bit	6	1 (+ 1 Kondensator)
Zellgröße	groß	sehr klein
Refresh nötig	Nein	Ja (~alle 64 ms)
Zugriffsgeschwindigkeit	sehr schnell	langsamer
Kosten pro Bit	hoch	niedrig
Typischer Einsatz	Cache, Register	Hauptspeicher (RAM)

In der Praxis ergänzen sich beide daher: schnelle, aber teure SRAM-Zellen in den Cache-Ebenen nahe am Prozessorkern, dichte und günstige DRAM-Zellen für den großvolumigen Hauptspeicher.

Die Flash-Speicherzelle

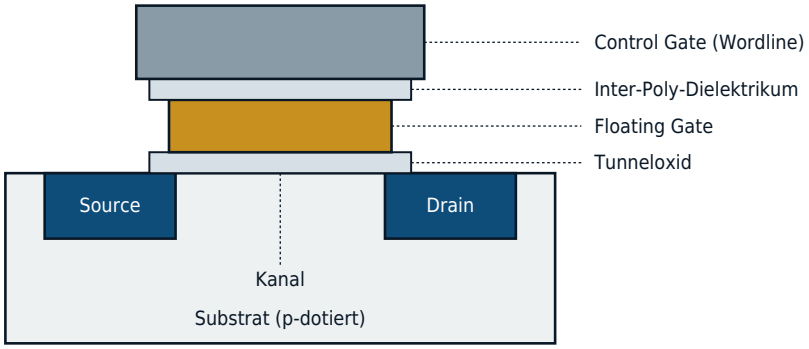
Aufbau der Floating-Gate-Zelle

Eine Flash-Speicherzelle ist im Kern ein MOSFET mit einer zusätzlichen,

elektrisch isolierten Gate-Elektrode – dem Floating Gate. Es liegt zwischen dem regulären Control Gate und dem Kanal, vollständig umschlossen von Siliciumdioxid: nach unten durch das dünne Tunneloxid (wenige Nanometer) zum Kanal getrennt, nach oben durch ein dickeres Zwischenoxid vom Control Gate isoliert.

Weil das Floating Gate keinen elektrischen Kontakt nach außen hat, bleibt Ladung, die einmal darauf gebracht wurde, über Jahre gespeichert – das ist das Grundprinzip der nichtflüchtigen Speicherung.

Die Floating-Gate-Zelle
Aufbau im Querschnitt



Das Floating Gate ist vollstaendig von Oxid umschlossen
und dadurch elektrisch isoliert.
Die gespeicherte Ladung bleibt ueber Jahre erhalten.

Programmieren und Löschen

Um eine Zelle zu programmieren (Ladung auf das Floating Gate zu bringen), nutzt man je nach Zelltyp Hot-Electron-Injection oder Fowler-Nordheim-Tunneling: Bei Hot-Electron-Injection beschleunigt ein hohes Drain-Source-Feld Elektronen so stark, dass einige die Energie besitzen, die Oxidbarriere zu überwinden und auf dem Floating Gate eingefangen zu werden.

Fowler-Nordheim-Tunneling nutzt stattdessen ein sehr hohes elektrisches Feld über dem Tunneloxid, das Elektronen quantenmechanisch durch die Barriere tunneln lässt – dieser Mechanismus wird typischerweise für den Löschvorgang verwendet, bei dem die Ladung wieder vom Floating Gate entfernt wird. Die gespeicherte Ladung verschiebt die Schwellspannung der Zelle messbar, wodurch beim Auslesen zwischen „0“ und „1“ unterschieden werden kann.

NOR- vs. NAND-Architektur

Die Art, wie einzelne Zellen zu einem Array verschaltet werden, unterscheidet zwei Grundarchitekturen. Bei NOR-Flash ist jede Zelle einzeln parallel an Bitline und Wordline angeschlossen, ähnlich einem NOR-Gatter – das ermöglicht schnellen, wahlfreien Zugriff auf einzelne Bytes, kostet aber Fläche.

Bei NAND-Flash sind mehrere Zellen (typisch 32–128) in Reihe zwischen Bitline und Source-Leitung verschaltet, wie bei einem NAND-Gatter; das spart erheblich Fläche und ermöglicht die heute üblichen hohen Speicherdichten, erlaubt aber nur seitenweisen statt wahlfreien Zugriff. Deshalb dominiert NAND-Flash bei USB-Sticks, SSDs und Speicherkarten, während NOR-Flash dort eingesetzt wird, wo schneller Direktzugriff auf einzelne Adressen gebraucht wird, etwa bei Firmware-Speichern.

Multi-Level-Cell (MLC/TLC)

Statt nur zwischen zwei Ladungszuständen (0/1) zu unterscheiden, lässt sich die Ladungsmenge auf dem Floating Gate feiner abstufen. Eine SLC-Zelle (Single-Level-Cell) speichert 1 Bit, eine MLC-Zelle 2 Bit über vier unterscheidbare Ladungsniveaus, eine TLC-Zelle 3 Bit über acht Niveaus.

Das vervielfacht die Speicherdichte pro Zelle, verkleinert aber den Störabstand zwischen den Niveaus – die Zellen werden empfindlicher gegenüber Ladungsverlust und benötigen aufwendigere Fehlerkorrektur, zudem sinkt die Anzahl möglicher Schreib-/Löschzyklen.

Vergleich: DRAM, SRAM und Flash

Die drei Speichertypen unterscheiden sich grundlegend in Flüchtigkeit und Einsatzzweck. DRAM speichert Ladung in einem Kondensator und muss ständig aufgefrischt werden, verliert also bei Stromausfall sofort seinen Inhalt – dafür ist es schnell und dient als Arbeitsspeicher.

SRAM speichert den Zustand in einer bistabilen Schaltung aus sechs Transistoren, ist noch schneller, aber flächenintensiver, und wird etwa als Cache-Speicher eingesetzt. Flash dagegen behält seinen Inhalt auch ohne Stromversorgung, ist jedoch beim Schreiben deutlich langsamer als DRAM und SRAM und daher für dauerhafte Datenspeicherung statt für Arbeitsspeicher gedacht.

Design Rules

Einführung & Motivation

Warum Design Rules?

Jeder integrierte Schaltkreis entsteht als geometrisches Layout: eine Abfolge übereinanderliegender Ebenen aus Diffusion, Polysilizium, Kontakten und mehreren Metalllagen, die gemeinsam Transistoren und deren Verdrahtung bilden. Damit dieses Layout auf dem Wafer korrekt und zuverlässig gefertigt werden kann, muss es eine Reihe geometrischer Randbedingungen einhalten – die sogenannten Design Rules. Sie übersetzen die physikalischen Grenzen eines Fertigungsprozesses in konkrete Vorgaben für den Chip-Entwurf.

Diese Grenzen ergeben sich aus mehreren Quellen zugleich. Die Lithografie kann Strukturen nur bis zu einer bestimmten Auflösung sauber abbilden; darunter verschwimmen Kanten oder Strukturen reißen ab. Masken lassen sich beim Belichten nicht beliebig genau zueinander justieren, sodass zwischen aufeinanderfolgenden Ebenen ein Sicherheitsabstand für Justierfehler eingeplant werden muss. Ätz- und Abscheideprozesse streuen von Wafer zu Wafer und über die Waferfläche, sodass Mindestabstände auch elektrische Kurzschlüsse durch Prozessschwankungen verhindern müssen. Hinzu kommen Effekte wie Latchup in CMOS-Wannenstrukturen oder Ladungsschäden durch das sogenannte Antenna-Phänomen, die zusätzliche, elektrisch motivierte Regeln erfordern.

Design Rules sind damit immer ein Kompromiss: Großzügige Abstände erhöhen die Fertigungsausbeute und Zuverlässigkeit, kosten aber Chipfläche und damit Kosten. Enge, aggressive Regeln erlauben höhere Packungsdichte, verringern aber die Ausbeute und erhöhen das Risiko von Fertigungsfehlern. Jede Foundry veröffentlicht für ihren jeweiligen Prozess einen vollständigen Regelsatz, der von automatisierten Werkzeugen (Design Rule Check, DRC) gegen das fertige Layout geprüft wird, bevor eine Maske belichtet werden darf.

Grundbegriffe

Grundlegende Regeltypen

Auch wenn sich einzelne Regelsätze zwischen Foundries und Technologieknoten stark unterscheiden, lassen sich fast alle Design Rules auf eine kleine Zahl geometrischer Grundbegriffe zurückführen. Diese werden meist an einem einzigen Layer (Ein-Layer-Regel) oder zwischen zwei Layern (Zwei-Layer-Regel) definiert:

Width (Breite) gibt die minimale Breite an, die eine Struktur auf einer Ebene haben darf – etwa eine Polysilizium-Gateleitung oder eine Metallbahn. Unterhalb dieser Breite ist die Lithografie- und Ätztreue nicht mehr gewährleistet, die Struktur könnte unterbrochen sein oder elektrisch zu hochohmig werden.

Spacing (Abstand) beschreibt den minimalen Zwischenraum zwischen zwei Strukturen derselben Ebene. Er verhindert, dass benachbarte Leiterbahnen durch Prozessstreuung verschmelzen oder kurzschließen, und berücksichtigt gleichzeitig die optische Auflösungsgrenze der Belichtung.

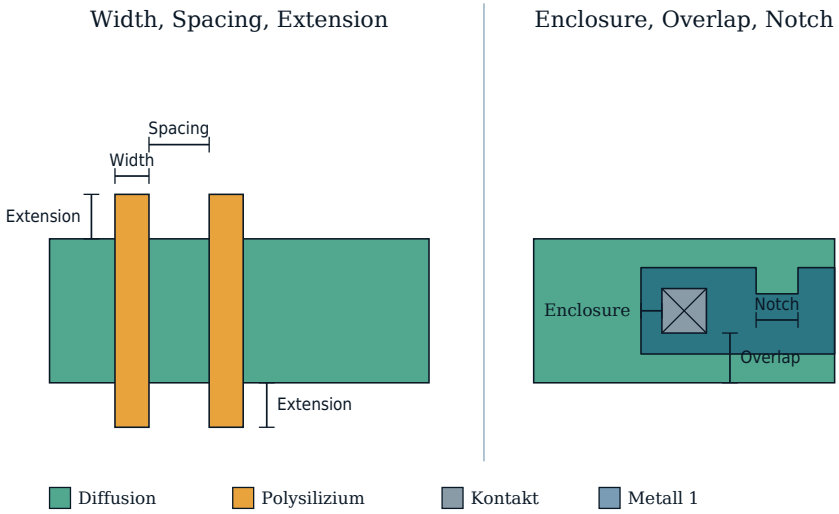
Enclosure (Umschließung) fordert, dass eine Ebene eine andere um einen Mindestbetrag umschließt – klassisch beim Kontakt: Das Metall muss den darunterliegenden Kontakt an allen Seiten um einen festen Betrag überragen, damit Justierfehler nicht zu einem freiliegenden, unkontaktierten Rand führen.

Overlap (Überlappung) ist eng mit Enclosure verwandt, bezieht sich aber auf die Perspektive der unteren Ebene: Diffusion muss den Kontakt überlappen, damit dieser vollständig auf leitfähigem Material aufsetzt.

Extension (Überstand) verlangt, dass eine Struktur über den Rand einer anderen hinausragt – etwa das Polysilizium-Gate, das beidseitig über die Diffusion hinausragen muss, damit Source und Drain beim Ionenimplantieren sicher getrennt bleiben, selbst bei leichtem Justierversatz.

Notch (Kerbe) begrenzt die minimale Abmessung eines einspringenden Winkels oder einer schmalen Aussparung in einer Struktur. Sehr spitze oder schmale Kerben lassen sich lithografisch nicht scharf abbilden und runden sich im realen Prozess ab.

Das folgende Diagramm zeigt diese sechs Grundbegriffe an einem einzelnen, vereinfachten Layout-Ausschnitt: einem Polysilizium-Gate über einer Diffusionsfläche mit einem Kontakt und der darüberliegenden ersten Metalllage.



Lambda-basierte vs. absolute Regeln

Skalierbare Regeln: das Lambda-Konzept

In den frühen 1980er-Jahren, als integrierte Schaltungen erstmals in großem Umfang außerhalb einzelner Halbleiterhersteller entworfen wurden, entstand mit dem Mead-Conway-Ansatz ein einflussreiches Konzept: Statt jede Design Rule in absoluten Maßeinheiten wie Mikrometern anzugeben, wurden alle Abstände und Breiten als Vielfache einer einzigen Bezugsgröße ausgedrückt, genannt Lambda (λ). Eine typische Regel lautete dann beispielsweise "minimale Poly-Breite = 2λ" oder "minimaler Diffusion-Kontakt-Abstand = 3λ".

Der Vorteil dieses Ansatzes lag in der Portabilität: Ein Layout, das vollständig in Lambda-Einheiten entworfen wurde, ließ sich prinzipiell auf einen anderen Fertigungsprozess mit feinerer oder gröberer Auflösung übertragen, indem man einfach den Zahlenwert von λ neu definierte – ohne jede Geometrie im Layout einzeln anzupassen. Für Lehre und akademische Entwürfe, bei denen Schaltungen oft bei mehreren Foundries gefertigt wurden (etwa im Rahmen von Multi-Project-Wafer-Programmen), war das ein erheblicher praktischer Gewinn.

Warum moderne Prozesse absolute Regeln verwenden

Mit fortschreitender Skalierung der Technologieknoten stieß das Lambda-Modell jedoch an seine Grenzen. Nicht alle Strukturen eines modernen Prozesses skalieren gleichmäßig: Transistor-Gates schrumpfen typischerweise schneller als Kontaktabstände, und Metallschichten unterschiedlicher Ebenen

haben oft völlig unterschiedliche minimale Breiten und Abstände, je nach ihrer Rolle in der Verdrahtungshierarchie. Ein einziger Skalierungsfaktor kann diese Asymmetrien nicht mehr abbilden.

Hinzu kommen physikalische Effekte, die sich nicht linear mit der Strukturgröße skalieren – etwa Antenna-Verhältnisse, Latchup-Abstände oder Dichteanforderungen für das chemisch-mechanische Planarisieren (CMP). Aus diesen Gründen veröffentlichen moderne Foundries ihre Design Rules praktisch ausnahmslos in absoluten Einheiten (Mikrometer oder Nanometer), mit jeweils eigenen, oft mehreren hundert Einzelregeln umfassenden Regelsätzen pro Prozessvariante. Das Lambda-Konzept bleibt didaktisch wertvoll, um das Prinzip von Design Rules zu vermitteln, spielt in der kommerziellen Chipentwicklung heute aber praktisch keine Rolle mehr.

Beispiel-Regelsatz

Ein typischer, vereinfachter Regelsatz

Um die in Abschnitt 2 vorgestellten Grundbegriffe an konkreten Zahlen zu verdeutlichen, zeigt die folgende Tabelle einen vereinfachten, illustrativen Regelsatz, wie er für einen generischen Single-Poly-/Multi-Metal-CMOS-Prozess im Sub-Mikrometer-Bereich typisch sein könnte. Die Werte orientieren sich an gängigen Lehr-Designkits und sind nicht die Regeln eines bestimmten realen Foundry-Prozesses – reale Regelsätze umfassen je nach Technologieknoten mehrere hundert Einzelregeln und unterscheiden sich naturgemäß von Hersteller zu Hersteller.

Layer / Regel	Bedeutung	Minimalwert
Poly Width	min. Breite einer Poly-Gateleitung	0,35 µm
Poly Spacing	min. Abstand zweier Poly-Leitungen	0,35 µm
Poly Extension	Gate-Überstand über die Diffusion	0,25 µm
Diffusion Width	min. Breite eines Diffusionsgebiets	0,4 µm
Diffusion Spacing	min. Abstand zweier Diffusionsgebiete gleichen Typs	0,5 µm
Contact Size	feste Kontaktlochgröße	0,4 × 0,4 µm
Contact Spacing	min. Abstand zweier Kontaktlöcher	0,4 µm
Diffusion Overlap of Contact	Diffusion-Umschließung um Kontakt	0,15 µm
Metal-1 Width	min. Breite Metall-1	0,5 µm
Metal-1 Spacing	min. Abstand zweier Metall-1-Leiterbahnen	0,5 µm
Metal-1 Enclosure of Contact	Metall-1-Umschließung um Kontakt	0,15 µm
Via-1 Size	feste Via-Größe (Metall-1 zu Metall-2)	0,4 × 0,4 µm

Layer / Regel	Bedeutung	Minimalwert
Metal-2 Width / Spacing	min. Breite / Abstand Metall-2	0,5 / 0,5 µm

Bereits an diesem stark vereinfachten Auszug lässt sich ein typisches Muster erkennen: Kontakte und Vias sind meist quadratisch und in ihrer Größe fest vorgegeben (nicht frei skalierbar), während umschließende Layer wie Diffusion und Metall einen zusätzlichen Sicherheitsrand einhalten müssen. Die Enclosure-Werte sind bewusst kleiner als die Width- und Spacing-Werte der jeweiligen Ebene – sie kompensieren in erster Linie den Justierfehler zwischen den beteiligten Masken, nicht die Auflösungsgrenze der Lithografie selbst.

In der Praxis wird ein solcher Regelsatz nicht von Hand nachgeschlagen, sondern in eine maschinenlesbare Technologiedatei (z. B. im LEF/DEF- oder herstellerspezifischen Format) überführt, gegen die anschließend automatisierte Werkzeuge wie Layout-Editoren und der Design Rule Checker (siehe Abschnitt 8) das Layout kontinuierlich prüfen.

Latchup- und Well-Regeln

Der Latchup-Mechanismus

In einem CMOS-Prozess liegen NMOS- und PMOS-Transistoren nicht elektrisch isoliert nebeneinander, sondern teilen sich dasselbe Substrat: Die PMOS-Transistoren sitzen in einer N-Wanne, die NMOS-Transistoren direkt im P-Substrat. Diese Anordnung erzeugt unweigerlich zwei parasitäre Bipolartransistoren: einen vertikalen PNP-Transistor (Emitter = P^+-Source/Drain des PMOS, Basis = N-Wanne, Kollektor = P-Substrat) und einen lateralen NPN-Transistor (Emitter = N^+-Source/Drain des NMOS, Basis = P-Substrat, Kollektor = N-Wanne).

Der Kollektor des einen parasitären Transistors ist zugleich die Basis des anderen – zusammen mit den lateralen Widerständen der Wanne (R_{well}) und des Substrats (R_{sub}) bilden PNP und NPN eine rückgekoppelte Vierschichtstruktur, elektrisch identisch mit einem Thyristor (SCR). Wird durch einen Spannungsstoß, einen Substratstrom oder ein Übersprechen einer der beiden Basis-Emitter-Übergänge kurzzeitig leitend, kann sich diese Rückkopplung selbst verstärken: Der resultierende niederohmige Pfad zwischen Versorgung und Masse – der Latchup – bleibt bestehen, bis die Versorgungsspannung abgeschaltet wird, und kann den Chip thermisch zerstören.

Well- und Substrat-Kontakte als Gegenmaßnahme

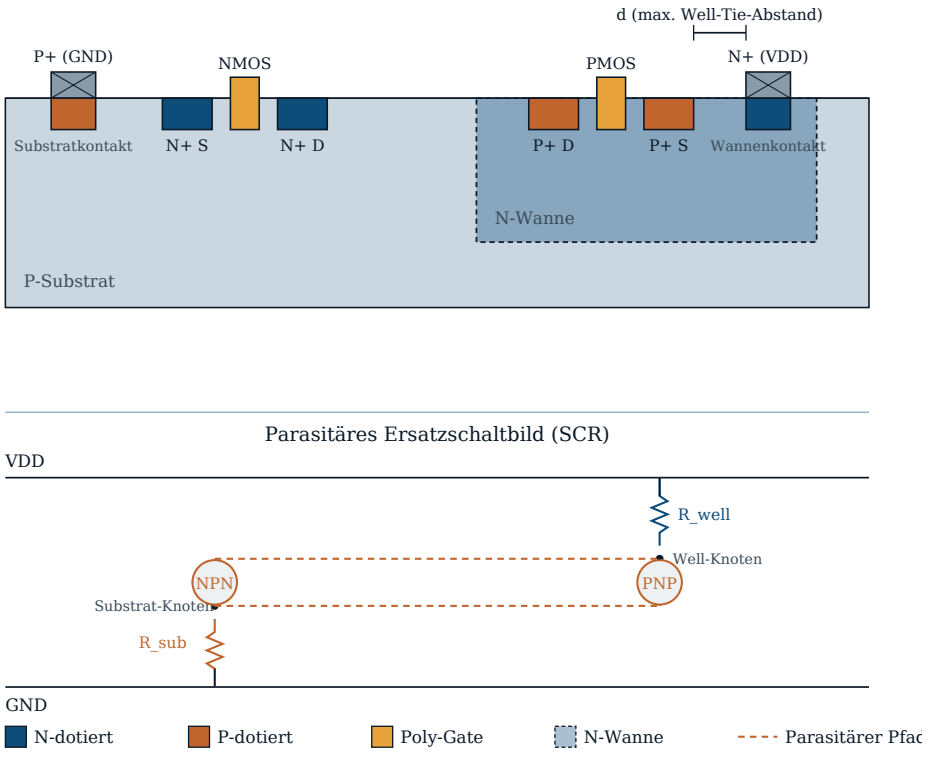
Die Stromverstärkung der parasitären Transistoren lässt sich durch das Layout

wirksam begrenzen: Je niedriger die lateralen Widerstände R_{well} und R_{sub} sind, desto kleiner ist der Spannungsabfall, der zum Einschalten des jeweils anderen Transistors nötig wäre. Design Rules schreiben deshalb einen maximalen Abstand zwischen jeder Source/Drain-Diffusion und dem nächsten Wannen- bzw. Substratkontakt vor – üblicherweise deutlich enger bemessen als die reinen Fertigungstoleranzen es erfordern würden.

In der Praxis werden diese Kontakte oft als durchgehender Guard Ring ausgeführt: ein N^+-Ring um besonders latchup-empfindliche PMOS-Strukturen (an VDD angeschlossen) bzw. ein P^+-Ring um NMOS-Strukturen (an GND angeschlossen), der den Basisstrom auf kürzestem Weg ableitet, bevor er den parasitären Transistor durchsteuern kann. Besonders wichtig sind solche Ringe rund um Ein-/Ausgangstreiber und andere Schaltungsteile, die hohe oder schnell wechselnde Ströme führen.

Das folgende Diagramm zeigt den Querschnitt eines einfachen Inverters mit Substrat- und Wannenkontakt sowie den überlagerten parasitären Thyristorpfad.

Latchup: Querschnitt und parasitäres Ersatzschaltbild



Antenna Rules

Der Plasma-Charging-Effekt

Viele Ätz- und Abscheideschritte in der Halbleiterfertigung – insbesondere das reaktive Ionenätzen (RIE) von Polysilizium und Metall – arbeiten mit einem Plasma. Dabei laden sich leitfähige, elektrisch isolierte Strukturen auf der Waferoberfläche während des Prozessschritts elektrisch auf, weil Elektronen- und Ionenstrom aus dem Plasma nicht symmetrisch auf alle Flächen treffen. Eine lange, noch unvollständig verdrahtete Polysilizium- oder Metallleitung wirkt dabei wie eine Antenne: Sie sammelt während des Ätzens Ladung aus dem Plasma, die sich – solange die Leitung nur an einem Transistor-Gate endet und noch nicht mit Source/Drain oder einer stabilisierenden unteren Ebene verbunden ist – über das dünne Gate-Oxid entladen muss.

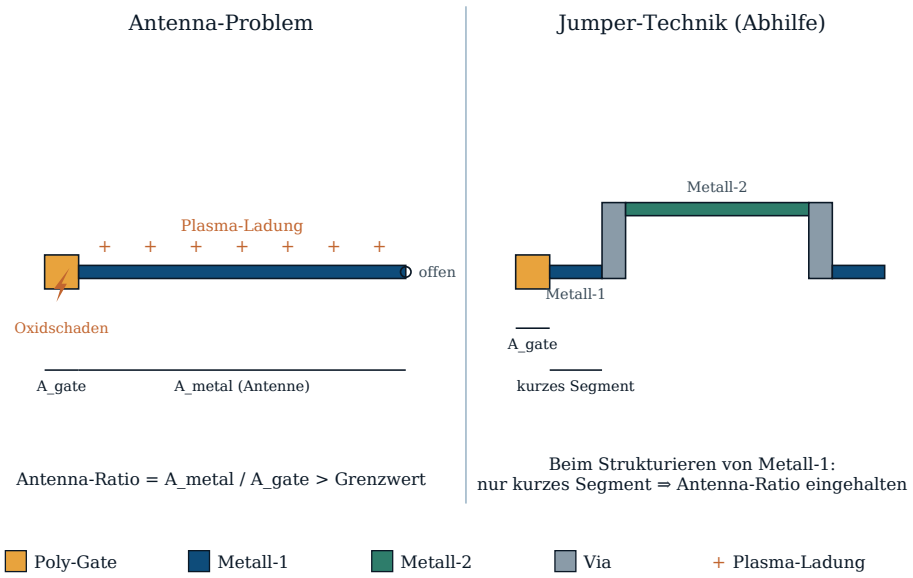
Das Gate-Oxid ist mit wenigen Nanometern Dicke die empfindlichste Struktur im gesamten Bauelement. Reicht die angesammelte Ladung aus, um ein ausreichend hohes elektrisches Feld über dem Oxid aufzubauen, fließt ein Fowler-Nordheim-Tunnelstrom, der das Oxid schädigt oder – im Extremfall – durchschlägt. Solche Schäden sind oft nicht sofort als Ausfall erkennbar, sondern äußern sich erst später als vorzeitige Bauteilalterung oder erhöhte Leckströme, was die Fehlersuche zusätzlich erschwert.

Antenna-Ratio als Design Rule

Um dieses Risiko zu begrenzen, definieren Foundries für jede leitfähige Ebene ein maximal zulässiges Antenna-Ratio: das Verhältnis der Fläche aller mit einem Gate elektrisch verbundenen Leiterbahnabschnitte (Poly, Metall-1, Metall-2 usw., jeweils bis zu dem Fertigungsschritt, in dem sie strukturiert werden) zur Fläche des angeschlossenen Gate-Oxids selbst. Überschreitet eine Netzverbindung diesen Grenzwert, meldet der DRC eine Antenna-Verletzung.

In der Praxis existieren zwei gängige Abhilfen. Die Antenna-Diode fügt an der gefährdeten Leitung eine zusätzliche, in Sperrrichtung gepolte Diode zum Substrat ein, die die angesammelte Ladung kontrolliert ableitet, ohne die Schaltungsfunktion zu beeinträchtigen. Die Jumper-Technik unterbricht stattdessen die lange Leitung auf ihrer ursprünglichen Ebene und führt sie über eine höhere Metalllage weiter: Da diese höhere Ebene erst in einem späteren, zeitlich getrennten Fertigungsschritt strukturiert wird, ist die zum jeweiligen Zeitpunkt mit dem Gate verbundene Fläche deutlich kleiner – das Antenna-Ratio bleibt für jeden einzelnen Fertigungsschritt innerhalb des zulässigen Grenzwerts.

Das folgende Diagramm zeigt links das Antenna-Problem an einer langen, direkt mit dem Gate verbundenen Leitung, rechts die Entschärfung durch einen Metall-Jumper.



Dichte-Regeln (CMP)

Warum Strukturdichte eine Design Rule ist

Moderne Prozesse mit mehreren Metalllagen benötigen zwischen den einzelnen Ebenen eine möglichst plane Oberfläche – nur so lässt sich die nächste Lage scharf belichten, ohne dass Fokusabweichungen durch Höhenunterschiede die Lithografie beeinträchtigen. Diese Einebnung übernimmt das chemisch-mechanische Planarisieren (Chemical-Mechanical Planarization, CMP), bei der eine rotierende, mit Schleifmittel benetzte Polierscheibe die Oxid- oder Metalloberfläche des Wafers mechanisch und chemisch abträgt.

CMP poliert jedoch nicht überall gleich schnell: Die Abtragsrate hängt von der lokalen Strukturdichte der darunterliegenden Ebene ab. In großen, weitgehend leeren Bereichen – etwa über einer breiten, wenig strukturierten Leiterbahn – poliert das Pad tiefer ein, als beabsichtigt, ein Effekt, der als Dishing bezeichnet wird. In dicht gepackten Bereichen mit vielen schmalen, eng benachbarten Strukturen tritt umgekehrt Erosion auf, bei der auch das umgebende Dielektrikum stärker als vorgesehen abgetragen wird. Beide Effekte erzeugen Höhenunterschiede über der Waferfläche, die sich auf nachfolgende Fertigungsschritte fortpflanzen und dort Fokus- und Ausbeuteprobleme verursachen können.

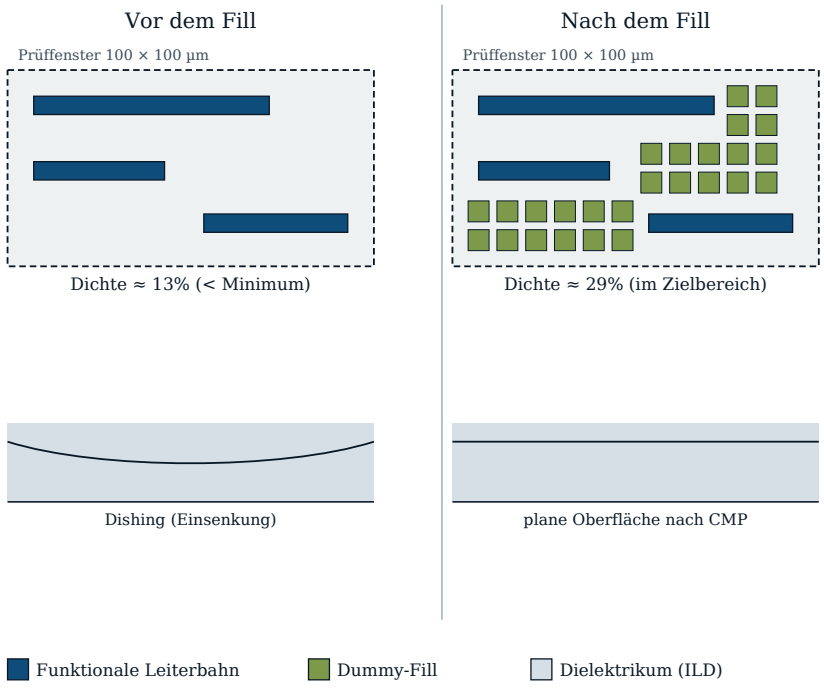
Fill-Strukturen als Gegenmaßnahme

Um dies zu vermeiden, schreiben Design Rules für jede Ebene eine minimale und maximale lokale Strukturdichte vor, üblicherweise gemessen über ein

gleitendes Prüffenster fester Größe (z. B. 100 µm × 100 µm), das über das gesamte Layout geschoben wird. Liegt die Dichte in einem Fenster unterhalb des Minimalwerts, fügen automatisierte Fill-Generatoren zusätzliche, elektrisch funktionslose Dummy-Strukturen in die Lücken ein – kleine, gleichmäßig verteilte Metall- oder Polysilizium-Flächen, die nicht mit dem Schaltungsnetz verbunden sind (floating fill) und ausschließlich dazu dienen, die Dichtevorgabe zu erfüllen.

Diese Fill-Strukturen unterliegen selbst wieder Design Rules – etwa Mindestabstand zu funktionalen Leiterbahnen, um parasitäre Kapazitäten oder ungewollte Kopplungen zu begrenzen. Die Erzeugung erfolgt heute praktisch immer automatisiert als letzter Schritt vor dem Tape-out, nachdem das eigentliche Schaltungslayout bereits fertiggestellt ist.

Das folgende Diagramm zeigt ein Prüffenster vor und nach dem Einfügen von Fill-Strukturen sowie schematisch die daraus resultierende Oberflächentopografie nach dem CMP-Schritt.



Design Rule Check (DRC)

Automatisierte Regelprüfung

Ein vollständiger Regelsatz nützt wenig, wenn er nicht konsequent gegen jedes Layout geprüft wird – bei modernen Chips mit Milliarden von Einzelstrukturen ist eine manuelle Kontrolle von vornherein ausgeschlossen. Diese Aufgabe

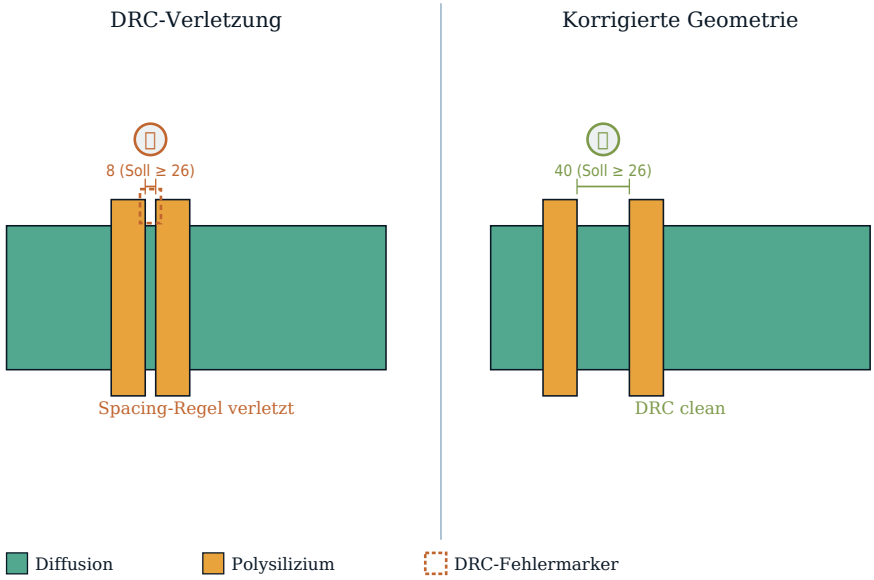
übernimmt der Design Rule Check (DRC): ein Software-Werkzeug, das das Layout gegen die in einer Technologiedatei (dem sogenannten DRC-Deck) hinterlegten Regeln prüft und jede gefundene Verletzung als markierten Fehler (Violation) ausgibt.

Technisch arbeitet ein DRC-Tool im Kern mit booleschen und geometrischen Operationen auf den einzelnen Layern: Es bildet Schnittmengen, Vereinigungen und Abstandsmessungen zwischen Polygonen verschiedener oder gleicher Ebenen und vergleicht die Ergebnisse mit den im Regelsatz hinterlegten Grenzwerten. Eine Width-Regel etwa wird geprüft, indem das Werkzeug jede Struktur der betreffenden Ebene an ihrer schmalsten Stelle vermisst; eine Spacing-Regel, indem es den minimalen Abstand zwischen benachbarten Polygonen bestimmt.

Ablauf im Entwurfsprozess

In der Praxis läuft der DRC nicht nur einmal am Ende, sondern wiederholt während des gesamten Layout-Prozesses: Viele Layout-Editoren bieten eine Echtzeitprüfung an, die Verletzungen bereits während des Zeichnens farblich markiert, ergänzt durch vollständige Batch-Läufe über das gesamte Design vor größeren Meilensteinen. Jede gefundene Verletzung muss vor dem nächsten Fertigungsschritt behoben werden – ein Layout gilt erst als „DRC clean", wenn der Lauf über das komplette Design keine offenen Verletzungen mehr meldet. Erst danach folgen in der Regel weitere Prüfungen wie der Layout-versus-Schematic-Abgleich (LVS), der kontrolliert, ob das gezeichnete Layout tatsächlich dieselbe Schaltung realisiert wie der Schaltplan.

Das folgende Diagramm zeigt eine typische Spacing-Verletzung – zwei Poly-Leitungen mit zu geringem Abstand – im Vergleich zur korrigierten, regelkonformen Geometrie.



Layout versus Schematic (LVS)

Einführung & Motivation

Warum DRC allein nicht reicht

Ein Layout, das den Design Rule Check vollständig besteht, ist geometrisch einwandfrei: Alle Breiten, Abstände, Umschließungen und Dichten liegen innerhalb der zulässigen Grenzen. Das sagt jedoch nichts darüber aus, ob die gezeichneten Polygone tatsächlich die Schaltung realisieren, die im Schaltplan entworfen wurde. Ein fehlender Kontakt, zwei vertauschte Anschlüsse oder ein versehentlich doppelt gezeichneter Transistor verletzen keine einzige Design Rule – das Layout bleibt „DRC clean", obwohl die Schaltung elektrisch nicht mehr dem Entwurf entspricht.

Solche Fehler entstehen leicht: Beim manuellen Route eines komplexen Blocks, beim Kopieren und Anpassen bestehender Layoutzellen oder schlicht durch einen vergessenen Verbindungsschritt. Je größer der Chip, desto unwahrscheinlicher wird es, einen einzelnen fehlenden Kontakt unter Millionen von Strukturen visuell zu entdecken.

Die Idee hinter LVS

Der Layout-versus-Schematic-Vergleich (LVS) löst dieses Problem, indem er nicht die Geometrie selbst prüft, sondern die daraus resultierende

Schaltungstopologie. Aus dem physischen Layout wird automatisiert eine Netzliste extrahiert – eine Liste aller Bauelemente (Transistoren, Widerstände, Kapazitäten) mit ihren Anschlüssen und den Netzen, die sie verbinden. Diese extrahierte Netzliste wird anschließend mit einer Referenz-Netzliste verglichen, die direkt aus dem Schaltplan stammt.

Stimmen beide Netzlisten überein – dieselben Bauelemente, dieselben Verbindungen, dieselben Bauelement-Parameter – gilt das Layout als „LVS clean". Jede Abweichung wird als Mismatch gemeldet und muss vor dem nächsten Schritt im Entwurfsablauf behoben werden. LVS ist damit die notwendige Ergänzung zum DRC: Der DRC stellt sicher, dass das Layout fertigbar ist, der LVS stellt sicher, dass das Gefertigte auch die richtige Schaltung ist.

Netzlisten-Extraktion aus dem Layout

Von Polygonen zu Bauelementen

Der erste Schritt der LVS-Extraktion identifiziert Bauelemente rein aus der Layer-Geometrie – ganz ähnlich wie ein DRC-Werkzeug, nur mit anderem Ziel. Überlappt eine Polysilizium-Struktur eine Diffusionsfläche, erkennt der Extraktor dort einen Transistor: Die Breite des Polysiliziums an der Überlappung liefert die Gate-Weite W, die Ausdehnung der Diffusion senkrecht dazu die Gate-Länge L. Ob es sich um einen NMOS oder PMOS handelt, ergibt sich aus dem Diffusionstyp (N^+ oder P^+) beziehungsweise daraus, ob die Diffusion in einer N-Wanne liegt.

Analog werden weitere Bauelemente erkannt: Eine Metall-Isolator-Metall-Struktur liefert einen Kondensator, ein schmaler, hochohmiger Polysilizium- oder Diffusionsstreifen ohne Gate-Funktion einen Widerstand. Jedes gefundene Bauelement wird mit seinen extrahierten geometrischen Parametern in die Netzliste aufgenommen.

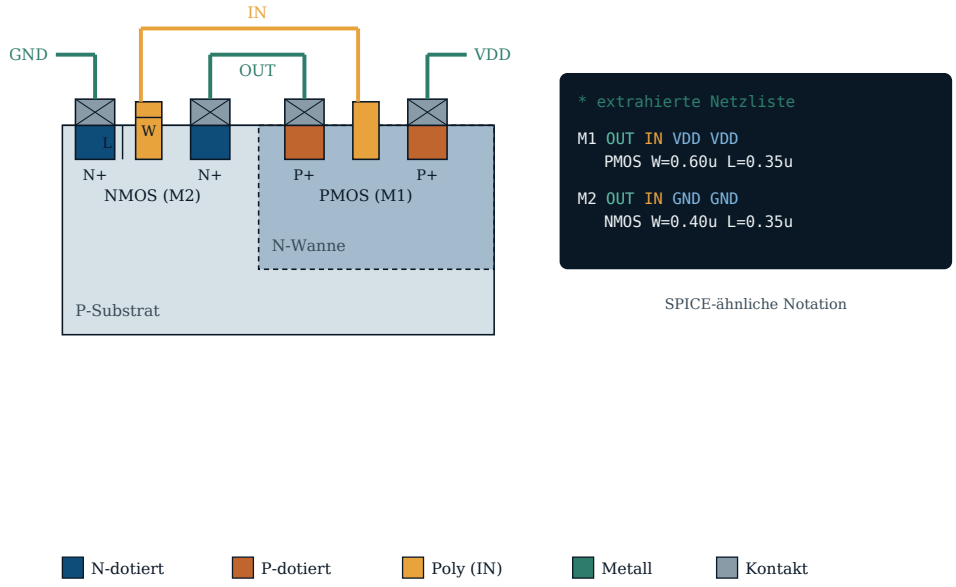
Von Kontakten zu Netzen

Parallel dazu ermittelt der Extraktor die elektrische Konnektivität: Alle Polygone, die auf derselben Ebene direkt aneinandergrenzen oder sich überlappen, sowie alle Ebenen, die über Kontakte oder Vias miteinander verbunden sind, werden zu einem gemeinsamen Netz zusammengefasst – unabhängig davon, über wie viele Metalllagen und Vias sich dieses Netz erstreckt. Technisch läuft das im Kern auf eine Union-Find-Operation über alle leitfähigen Polygone hinaus: Zwei Polygone gehören genau dann zum selben Netz, wenn ein leitfähiger Pfad zwischen ihnen existiert.

Das Ergebnis ist eine vollständige Netzliste: eine Liste aller Bauelemente mit ihren Anschlüssen, wobei jeder Anschluss einem der zuvor ermittelten Netze zugeordnet ist. Diese Netzliste enthält zu diesem Zeitpunkt noch keine Namen wie „VDD" oder „OUT" – sie referenziert Netze zunächst nur über interne Kennungen. Erst der Abgleich mit dem Schaltplan (Abschnitt 3) ordnet ihnen wieder sprechende Namen zu.

Das folgende Diagramm zeigt einen einfachen Inverter im Layout und die daraus extrahierte, vereinfachte Netzliste.

Netzlisten-Extraktion: Inverter-Layout



Die Referenz-Netzliste aus dem Schaltplan

Vom Schaltplan zur Netzliste

Während die Layout-Netzliste aus Geometrie gewonnen wird, entsteht die Referenz-Netzliste auf direktem Weg: Ein Schaltplan-Eingabewerkzeug (Schematic Capture) kennt für jedes platzierte Symbol – Transistor, Widerstand, Kondensator – bereits dessen Anschlüsse (Pins) und deren Bedeutung. Verbindet der Entwerfer zwei Pins durch eine gezeichnete Leitung oder durch identische Netzbezeichner, gehören sie zum selben Netz. Diese Extraktion ist trivial im Vergleich zur Layout-Seite, da keine Geometrie interpretiert werden muss – die Konnektivität liegt bereits explizit vor.

Netznamen wie VDD, GND, IN oder OUT vergibt meist der Entwerfer selbst, entweder durch Beschriftung einzelner Leitungen oder durch spezielle globale Netz-Symbole für Versorgungsspannungen, die werkzeugübergreifend als

„überall gleich" gelten – ein an mehreren Stellen im Schaltplan platziertes VDD-Symbol verbindet all diese Punkte zu einem einzigen Netz, auch ohne durchgezogene Leitung dazwischen.

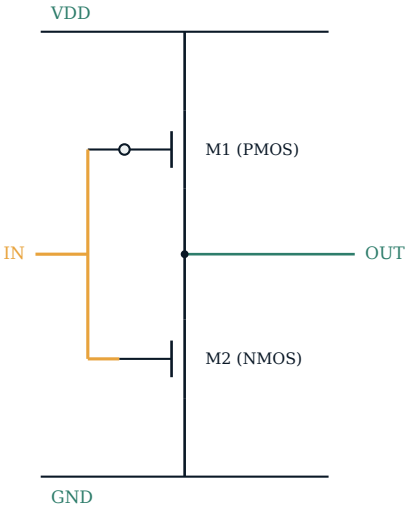
Hierarchie und Flattening

Reale Schaltpläne sind praktisch immer hierarchisch aufgebaut: Ein Inverter ist eine Zelle, die in einem Gatter wiederverwendet wird, das wiederum Teil eines Multiplexers ist, der in einem Register instanziiert wird. Der Netzlisten-Compiler muss diese Hierarchie auflösen – entweder vollständig flach (jede Instanz einer Zelle wird zu eigenständigen Bauelementen expandiert) oder hierarchisch, wobei gleich aufgebaute Instanzen nur einmal als Referenzzelle geprüft und danach lediglich ihre Verschaltung untereinander verglichen wird.

Hierarchisches LVS ist bei Speicherblöcken mit tausendfach wiederholten, identischen Zellen (z. B. SRAM-Arrays) der einzig praktikable Ansatz – ein vollständig geflachter Vergleich würde bei jeder Instanz erneut die komplette Zellstruktur durchrechnen, obwohl sich nur die Position im Layout unterscheidet.

Die so gewonnene Referenz-Netzliste gilt als „golden" – als der verbindliche Sollzustand, gegen den die aus dem Layout extrahierte Netzliste im nächsten Schritt verglichen wird. Das folgende Diagramm zeigt den Schaltplan desselben Inverters aus Abschnitt 2 und die daraus gewonnene Referenz-Netzliste.

Referenz-Netzliste: Inverter-Schaltplan




```
* Referenz-Netzliste
M1 OUT IN VDD VDD
   PMOS W=0.60u L=0.35u
M2 OUT IN GND GND
   NMOS W=0.40u L=0.35u
```

golden — verbindlicher Sollzustand

Der Vergleichsalgorithmus

Die Netzliste als Graph

Um zwei Netzlisten algorithmisch zu vergleichen, wird jede von ihnen zunächst in einen Graphen überführt: Bauelemente und Netze werden zu Knoten, jeder Anschluss eines Bauelements an ein Netz wird zu einer Kante zwischen ihnen, beschriftet mit der Anschlussrolle (Drain, Gate, Source, Bulk). Das Ergebnis ist ein sogenannter bipartiter Graph – Bauelement-Knoten sind ausschließlich mit Netz-Knoten verbunden, nie direkt untereinander.

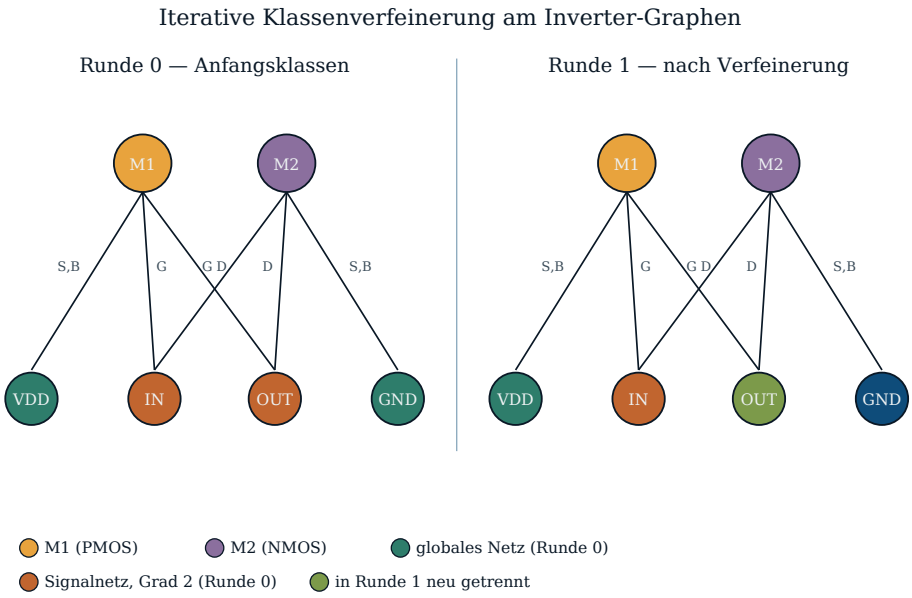
Zwei Netzlisten sind genau dann elektrisch gleich, wenn zwischen ihren beiden Graphen eine Isomorphie existiert: eine eindeutige Zuordnung, die jeden Knoten des einen Graphen auf genau einen Knoten des anderen abbildet, sodass Kanten, Kantenbeschriftungen und Knotenattribute (Bauelement-Typ, W/L) erhalten bleiben. Graphisomorphie ist im allgemeinen Fall rechnerisch aufwendig – für Schaltungen mit Millionen von Transistoren wäre ein naiver Brute-Force-Vergleich aller möglichen Zuordnungen aussichtslos.

Iterative Verfeinerung statt Brute Force

LVS-Werkzeuge umgehen dieses Problem durch iterative Klassenverfeinerung. Zunächst erhält jeder Knoten eine grobe Anfangsklasse aus leicht bestimmbaren Merkmalen – Bauelement-Typ und Parameter bei Bauelement-Knoten, Anzahl angeschlossener Terminals bei Netz-Knoten. Anschließend wird wiederholt verfeinert: Jeder Knoten bekommt eine neue, feinere Klasse zugewiesen, die sich aus der Multimenge der Klassen seiner Nachbarn (samt Kantenbeschriftung) ergibt. Zwei zunächst gleich klassifizierte Knoten trennen sich, sobald sich ihre Nachbarschaften unterscheiden.

Dieser Vorgang wiederholt sich, bis sich keine Klasse mehr weiter aufteilen lässt. In gut entworfenen, wenig symmetrischen Schaltungen bleiben danach fast nur noch einelementige Klassen übrig – jeder Knoten ist damit eindeutig einem Knoten der jeweils anderen Netzliste zugeordnet. Nur für die wenigen verbleibenden, tatsächlich symmetrischen Gruppen (etwa parallel geschaltete, identische Dummy-Finger eines Transistors) muss der Algorithmus noch alle verbleibenden Zuordnungsmöglichkeiten innerhalb dieser kleinen Gruppe durchprobieren – ein Aufwand, der selbst bei Millionen Bauelementen praktikabel bleibt, weil diese Restgruppen erfahrungsgemäß winzig sind.

Das folgende Diagramm zeigt diesen Verfeinerungsschritt am Beispiel des Inverters aus den vorigen Abschnitten: links die Anfangsklassen, rechts das Ergebnis nach einer Verfeinerungsrunde.



Typische Mismatches

Vier wiederkehrende Fehlerbilder

Obwohl LVS-Werkzeuge im Detail sehr unterschiedliche Fehlermeldungen ausgeben, lassen sich die allermeisten Mismatches auf eine Handvoll wiederkehrender Ursachen zurückführen. Wer diese Muster kennt, findet die Fehlerursache im Layout meist deutlich schneller, als sich blind durch die Fehlerliste zu arbeiten.

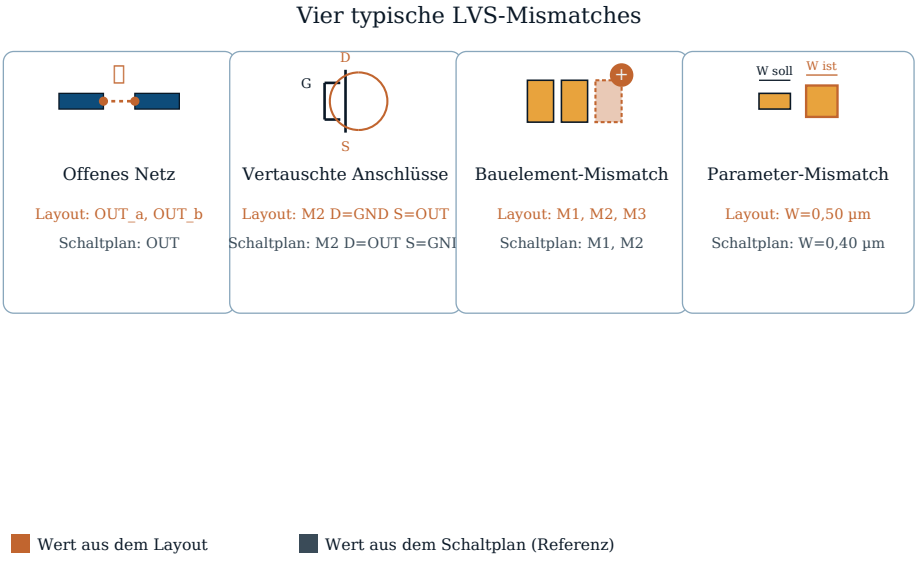
Offenes Netz entsteht durch einen fehlenden oder falsch platzierten Kontakt: Zwei Strukturen, die laut Schaltplan verbunden sein sollen, bleiben im Layout elektrisch getrennt. Die Netzliste zeigt dann statt eines Netzes zwei – der LVS meldet meist ein „net not found" oder eine Netzanzahl-Abweichung, weil ein im Schaltplan erwartetes Netz im Layout in zwei Teile zerfällt.

Vertauschte Anschlüsse passieren leicht beim Kopieren oder manuellen Verdrahten eines Transistors: Source und Drain sind im Layout an die jeweils falschen Netze angeschlossen. Da Source und Drain bei den meisten MOSFETs geometrisch symmetrisch und damit austauschbar sind, bleibt das Layout DRC-clean – die Schaltung funktioniert danach aber elektrisch anders als beabsichtigt, oder gar nicht.

Fehlendes oder zusätzliches Bauelement entsteht meist durch einen Kopier-Fehler – eine Zelle wurde versehentlich doppelt platziert, oder beim Löschen eines Testelements wurde eine benötigte Struktur mit entfernt. Die beiden Netzlisten unterscheiden sich dann bereits in der schieren Bauelement-Anzahl, oft schon der einfachste aller Mismatches zu finden.

Parameter-Mismatch betrifft meist die Transistorweite W oder -länge L: Wurde im Layout eine andere Gate-Breite gezeichnet als im Schaltplan vorgesehen, bleibt die Konnektivität vollständig korrekt – Topologie und Bauelement-Typ stimmen überein, nur der Zahlenwert weicht ab. Viele LVS-Werkzeuge melden das nicht als harten Fehler, sondern als Warnung mit einer einstellbaren Toleranzgrenze, da geringfügige Abweichungen durch Rundung auf das Grid oft unvermeidlich sind.

Das folgende Diagramm zeigt alle vier Fehlerbilder im direkten Vergleich Layout vs. Schaltplan.



LVS im Entwurfsablauf

Position zwischen DRC und Parasitic Extraction

LVS läuft im Entwurfsablauf grundsätzlich nach dem Design Rule Check: Ein Layout muss zunächst DRC-clean sein, bevor ein LVS-Lauf sinnvolle Ergebnisse liefert – auf einem geometrisch fehlerhaften Layout (etwa mit sich überlappenden, eigentlich getrennten Strukturen) ließe sich ohnehin keine verlässliche Netzliste extrahieren. Beide Prüfungen zusammen bilden das, was in der Praxis als „Signoff" bezeichnet wird: Ein Layout, das weder DRC- noch LVS-clean ist, geht nicht zur Fertigung.

Wie beim DRC ist auch der LVS-Lauf kein einmaliges Ereignis am Ende, sondern Teil eines iterativen Zyklus: Jede gemeldete Abweichung – offenes Netz, vertauschter Anschluss, fehlendes Bauelement – muss im Layout behoben werden, wonach die Extraktion und der Vergleich erneut laufen. Bei umfangreichen Blöcken beschleunigen inkrementelle LVS-Läufe, die nach einer

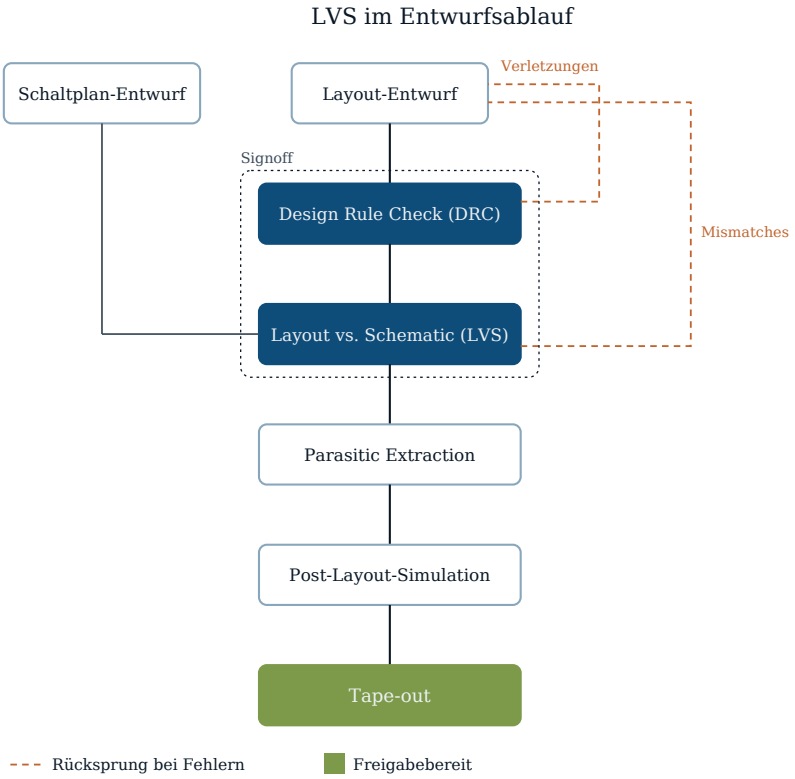
kleinen Layout-Änderung nur die betroffene Umgebung neu vergleichen, diesen Zyklus erheblich gegenüber einem vollständigen Neuvergleich des gesamten Blocks.

Was danach folgt

Ist ein Block sowohl DRC- als auch LVS-clean, schließt sich üblicherweise die Parasitic Extraction an: Aus der bereits verifizierten Layout-Geometrie werden nun zusätzlich die parasitären Widerstände und Kapazitäten der Verdrahtung berechnet und der Netzliste hinzugefügt – eine Aufgabe, die LVS selbst nicht übernimmt, da sie ausschließlich auf Konnektivität und Bauelement-Parameter prüft, nicht auf elektrische Feinheiten der Verdrahtung. Die so parasitär-annotierte Netzliste dient anschließend der Post-Layout-Simulation, die zeigt, ob die Schaltung auch mit den realen, layoutbedingten Verzögerungen und Spannungsabfällen noch wie vorgesehen funktioniert.

Erst wenn auch dieser letzte Schritt die Spezifikation bestätigt, ist ein Block bereit für den Tape-out – die Freigabe der Maskendaten an die Foundry.

Das folgende Diagramm zeigt diesen Ablauf im Überblick, von Schaltplan und Layout bis zum Tape-out.



Schaltungslayouts

Vom Schaltbild zum Layout

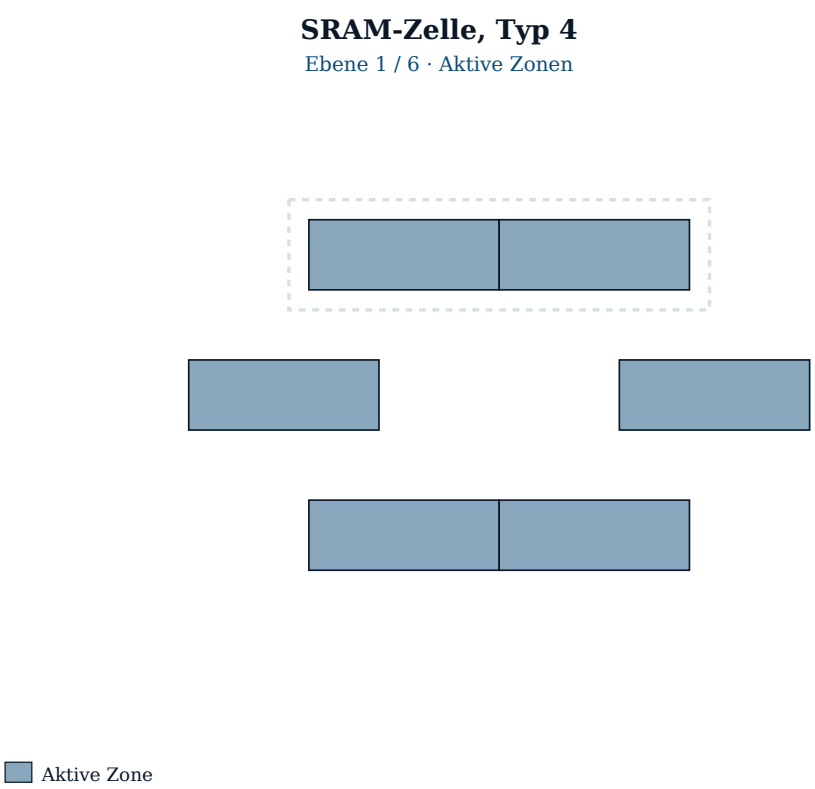
Das Schaltbild aus dem vorigen Kapitel zeigt, *wie* die sechs Transistoren elektrisch verschaltet sind. Es sagt aber nichts darüber aus, *wo* im Chip diese Transistoren tatsächlich liegen und wie die Verbindungen zwischen ihnen gefertigt werden. Genau das zeigt das Layout: die Abfolge von Masken, mit denen jede Fertigungsebene — Diffusion, Polysilizium, Kontakte, gleich drei Metalllagen — belichtet und strukturiert wird.

Diese Zelle folgt einem verbreiteten Aufbau, oft "Typ 4" genannt: Die beiden Speicherknoten Data und Data_b laufen als durchgehende Spalten von oben (PMOS) über die Mitte (Zugriffstransistoren) bis unten (NMOS) durch die ganze Zellhöhe. VDD liegt zentral zwischen den beiden Spalten, BL und B̄L̄ außen, VSS oben und unten, Wortleitung (WL) quer in der Mitte.

Kürzel	Bedeutung
VDD	Versorgungsspannung (zentral, vertikal)
VSS	Masse (oben und unten, horizontal)
BL / B̄L̄	Bitleitung / invertierte Bitleitung (außen, vertikal)
WL	Wortleitung (mittig, horizontal)
Data / Data_b	die beiden rückgekoppelten Speicherknoten
PU	Pull-up-Transistor (PMOS, zieht Richtung VDD)
PD	Pull-down-Transistor (NMOS, zieht Richtung VSS)

Ebene 1: Aktive Zonen

Sechs Rechtecke in drei Zeilen: oben die beiden PMOS (Pull-up, in der N-Wanne), in der Mitte die beiden Zugriffstransistoren, unten die beiden NMOS (Pull-down). Jede Spalte — links "Data", rechts "Data_b" — zieht sich als eigene aktive Fläche durch alle drei Zeilen. Das ist der Kern des Typ-4-Aufbaus: Ein Inverter besteht nicht aus zwei nebeneinanderliegenden Transistoren, sondern aus einem PMOS oben und einem NMOS unten *in derselben Spalte*, mit dem Zugriffstransistor dazwischen.



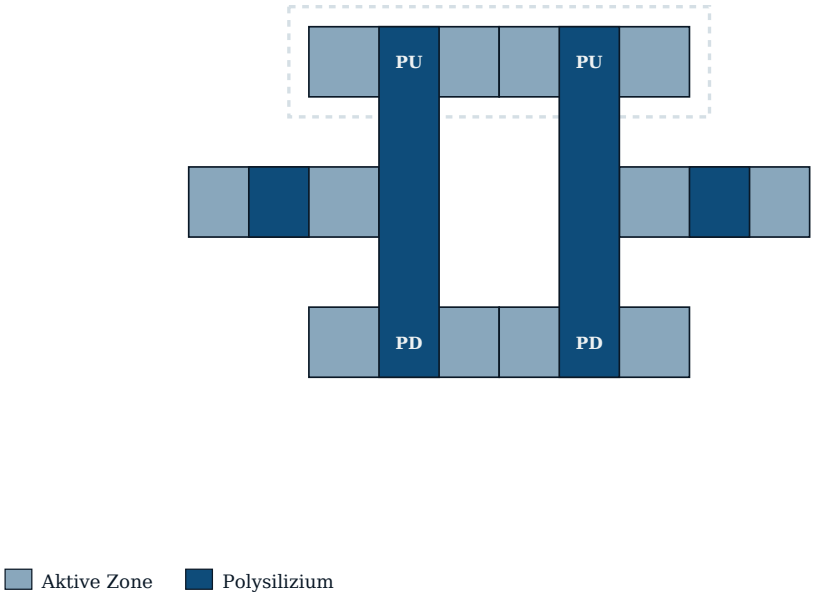
Ebene 2: Polysilizium

Zwei durchgehende, vertikale Gates kommen hinzu — Gate A (treibt später den Data-Knoten, aber Achtung: sein *eigenes* Gate-Signal ist Data_b, siehe unten) und Gate B, jeweils von der PMOS- bis zur NMOS-Zeile. Die beiden Zugriffstransistoren bekommen dagegen nur kurze, lokale Gate-Stücke, keine durchgehende Wortleitung aus Poly. Der Grund: Eine durchgehende Poly-WL würde exakt durch Gate A und Gate B hindurchlaufen und sie kurzschließen. WL wird stattdessen erst in Ebene 6 als Metall-Leitung ergänzt, die eine Ebene höher liegt und deshalb kreuzungsfrei über die Gates hinweggeht.

Wichtig für die Kreuzkopplung: Der Pull-up/Pull-down einer Spalte wird immer *von der jeweils anderen Spalte* gesteuert — Data-Transistoren von Data_b-Gates, Data_b-Transistoren von Data-Gates. Sonst würde ein Inverter sein eigenes Ausgangssignal treiben und die Zelle könnte kein Bit halten.

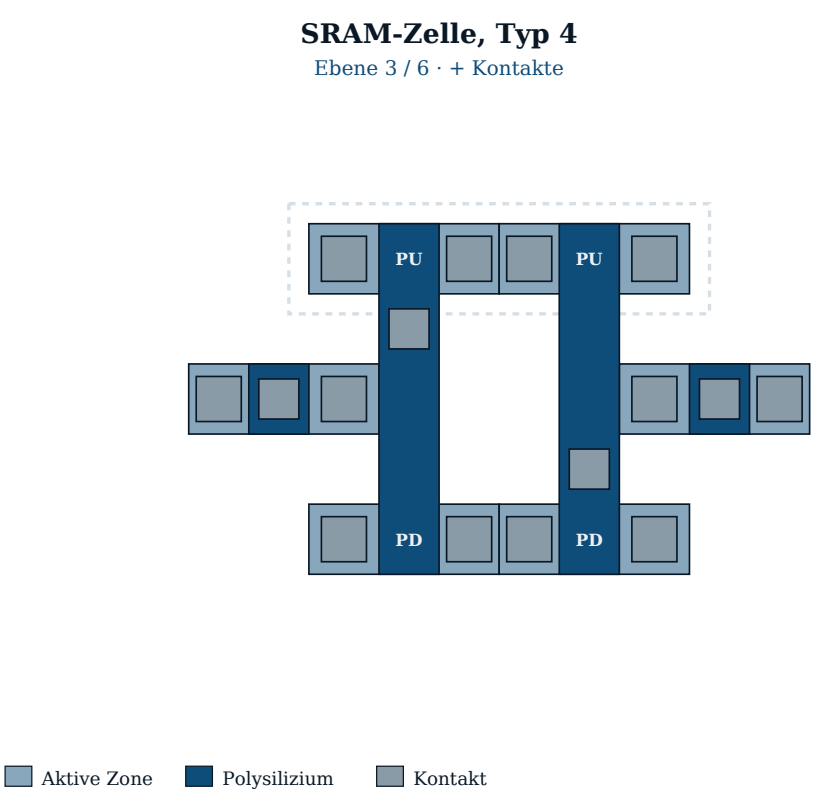
SRAM-Zelle, Typ 4

Ebene 2 / 6 · + Polysilizium



Ebene 3: Kontakte

Kontaktquadrate an jedem Punkt, an dem die nächste Ebene (Metall-1) elektrischen Zugriff auf Diffusion oder Polysilizium braucht: an beiden Enden jeder Data-/Data_b-Säule (PMOS-Drain, Zugriffstransistor-Drain, NMOS-Drain), an den Source-Anschlüssen (VDD, VSS, BL, B$\overline{\text{L}}$) und je einer an Gate A und Gate B für die Rückkopplung.



Ebene 4: Metall-1

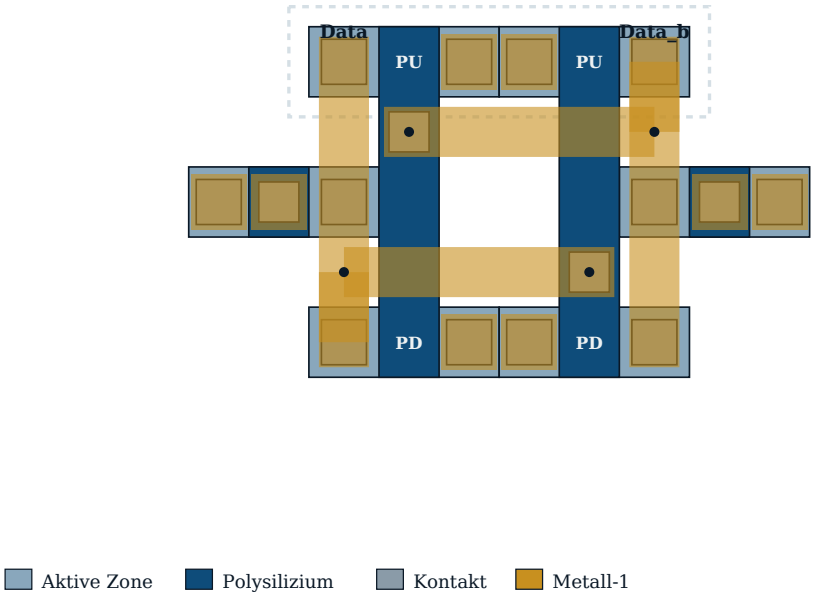
Hier passiert die eigentliche Verdrahtung innerhalb der Zelle. Zwei Dinge gleichzeitig:

Erstens die lokale Verbindung: Jede Data-/Data_b-Säule bekommt eine durchgehende Metall-1-Bahn, die PMOS-Drain, Zugriffstransistor und NMOS-Drain zu einem einzigen elektrischen Knoten zusammenfasst.

Zweitens die Rückkopplung über die volle Zellhöhe: Data_b muss auf Gate A treffen, Data auf Gate B — beide Wege müssen sich dabei kreuzen, dürfen sich aber nicht berühren. Gelöst wird das, indem die beiden Wege durch unterschiedliche Lücken geführt werden (die eine oberhalb, die andere unterhalb der Wortleitungs-Zeile), sodass sie sich im Bild wie eine "O"-Schleife um die beiden Gates legen, ohne einander je zu kreuzen.

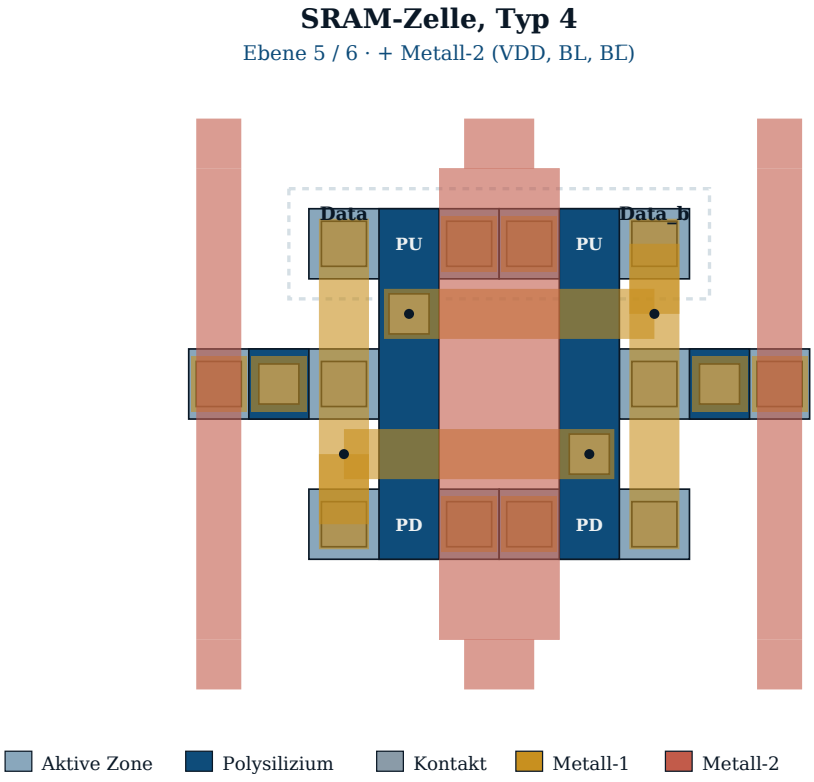
SRAM-Zelle, Typ 4

Ebene 4 / 6 · + Metall-1



Ebene 5: Metall-2

VDD, BL und BL̄ kommen als vertikale Leitungen hinzu — VDD zentral zwischen den beiden Spalten, BL und BL̄ außen. Alle drei laufen über die gesamte Zellhöhe durch und stehen am oberen und unteren Rand bereit, sich mit der nächsten Zelle in derselben Spalte des Arrays zu verbinden.



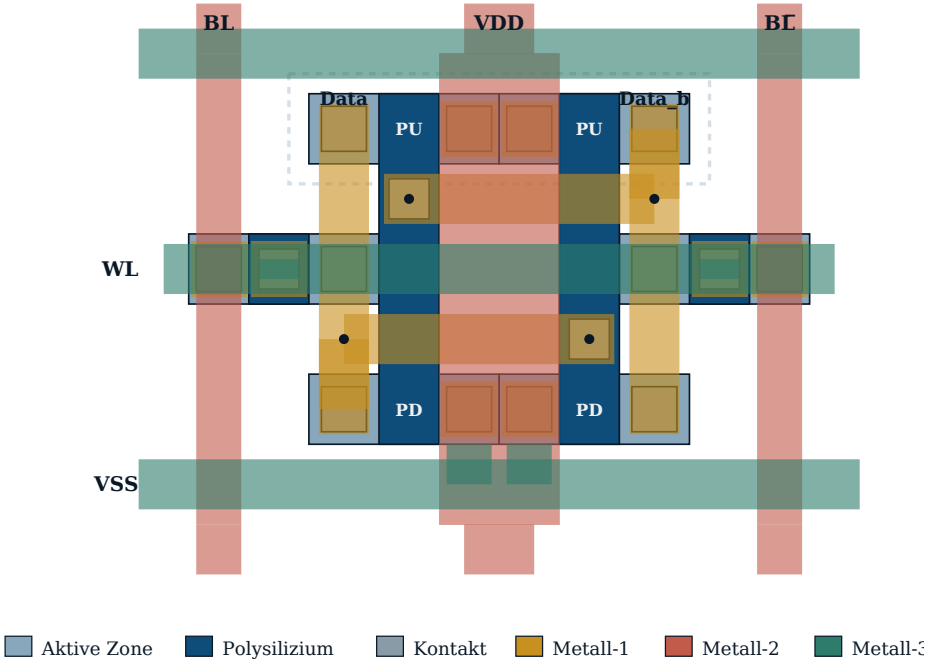
Ebene 6: Metall-3 — fertig

Zuletzt VSS (oben und unten) und WL (mittig), beide horizontal. Dass WL erst hier erscheint statt in Poly, ist kein Zufall: Auf dieser Ebene kann sie ungehindert über VDD, Gate A und Gate B hinweglaufen, weil jede tiefere Ebene isoliert darunter liegt — Kreuzungen zwischen *verschiedenen* Ebenen sind unproblematisch, nur *dieselbe* Ebene darf sich nicht berühren.

Damit ist die Zelle komplett verdrahtet. Alle vier Anschlüsse (VDD, VSS, BL, BL̄, WL) liegen am Rand bereit, um sich nahtlos mit den Nachbarzellen im Array zu verbinden — VDD/BL/BL̄ nach oben und unten (gleiche Spalte), VSS/WL nach links und rechts (gleiche Zeile).

SRAM-Zelle, Typ 4

Ebene 6 / 6 · + Metall-3 (VSS, WL) — fertig



Hinweis zur Darstellung: Ab Metall-1 sind alle Ebenen halbtransparent gezeichnet, nicht deckend — genau wie in echten Layout-Editoren (z. B. Virtuoso, KLayout), wo transparente Ebenen den Blick auf das freigeben, was darunterliegt.

Aufbau eines n-Kanal-FET

Allgemeiner Aufbau

Ein Transistor ist ein elektronisches Halbleiterbauelement das zum Schalten oder Verstärken von Strom verwendet werden kann. Der Stromfluss erfolgt über zwei Anschlüsse (Drain, Source), während der dritte (Gate) zur Steuerung dient. Neben dem Feldeffekttransistor (FET) gibt es noch einen weiteren grundlegenden Transistortyp, den Bipolartransistor. Bei ihm heißen die Anschlüsse Emitter (Source beim FET), Basis (Gate) und Kollektor (Drain). Die Funktionsweise des Bipolartransistors beruht auf Ladungsträgern beider Polaritäten (daher bipolar), Löchern und Elektronen. Beim Feldeffekttransistor, auch als unipolarer Transistor bezeichnet, sind abhängig von der Bauart entweder Elektronen *oder* Löcher am Stromtransport beteiligt.

Bei dem nachfolgend beschriebenen Transistor handelt es sich um einen sogenannten MOSFET (engl. metal oxide semiconductor field-effect transistor, Metall-Oxid-Halbleiter-FET). Obwohl heute meist hochdotiertes Polysilicium als Gatematerial Verwendung findet und kein Aluminium mehr zum Einsatz kommt, wird auch bei diesem Transistortyp nach wie vor die Bezeichnung MOSFET benutzt. Besser wäre in diesem Fall die Bezeichnung IGFET (engl. insulated gate FET, FET mit isoliertem Gate). Bei neuartigen Transistoren mit High-k-Metal-Gate-Technologie ist die Bezeichnung MOSFET dagegen wieder korrekt, sofern als Isolator weiterhin ein Oxid verwendet wird.

Der Transistor ist das grundlegende Bauelement in der Halbleiterfertigung, in modernen Mikrochips finden sich mehrere hundert Milliarden Transistoren. Durch die Kombination mehrerer Transistoren können sämtliche logische Gatter realisiert werden, um aus Eingangssignalen entsprechende logische Ausgangssignale zu erhalten. Dadurch bilden Transistoren das Herzstück eines jeden Mikroprozessors, Speicherchips usw. Der Transistor ist die von der Menschheit in der höchsten Gesamtstückzahl produzierte technische Funktionseinheit, und aus dem heutigen Leben nicht mehr wegzudenken.

Der Transistor wird in der Produktion Schicht für Schicht aufgebaut. Dabei steht hier der grundlegende Aufbau eines einfachen MOSFETs im Vordergrund, die verschiedenen Möglichkeiten zur Realisierung dieser Schichten folgen in den späteren Kapiteln.

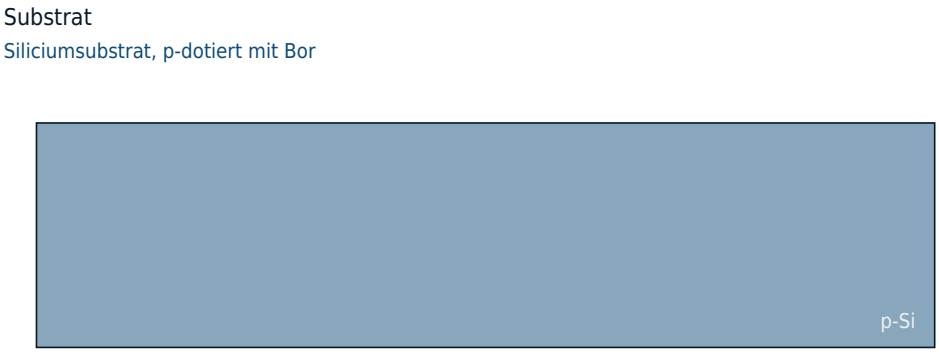
Der hier beschriebene Aufbau ist der planare Transistor: Das Gate liegt flach auf

der Scheibenoberfläche und steuert den Kanal von oben. Diese Bauform bestimmte die Halbleitertechnik über vier Jahrzehnte und ist für das Verständnis der Funktionsweise nach wie vor die richtige Ausgangsform. In den Prozessen für kleinste Strukturen ist sie inzwischen abgelöst worden, weil ein Gate, das nur von einer Seite wirkt, den immer kürzeren Kanal nicht mehr zuverlässig sperren kann. Wie die Nachfolger aussehen, beschreibt das Kapitel Aufbau eines FinFET.

Herstellung eines n-Kanal-FET

1. Substrat

Grundlage für einen n-Kanal-Feldeffekttransistor ist ein p-dotiertes Siliciumsubstrat, als Dotierstoff dient Bor.

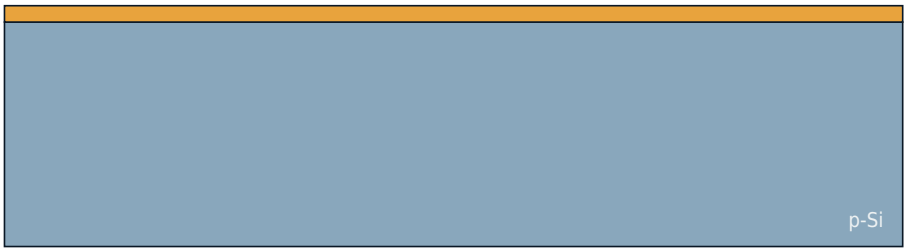


2. Oxidation

Auf dem Substrat wird Siliciumdioxid SiO_2 (das Gateoxid, kurz GOX) in einer Trockenoxidation erzeugt. Es dient zur Isolation zwischen dem später abgeschiedenen Gate und dem Substrat.

Gateoxid (GOX)

Thermische Oxidation erzeugt eine dünne Oxidschicht



Silicium (p-Si) Gateoxid

Das dünne, thermisch gewachsene Gateoxid isoliert später das Gate vom Kanal und muss deshalb besonders gleichmäßig und defektfrei sein.

3. Abscheidung

In einem LPCVD-Prozess wird Nitrid abgeschieden, es dient später bei der Feldoxidation als Maskierung.

Nitrid als Maskierung

LPCVD-Abscheidung von Siliciumnitrid

Silicium (p-Si) Gateoxid Nitrid

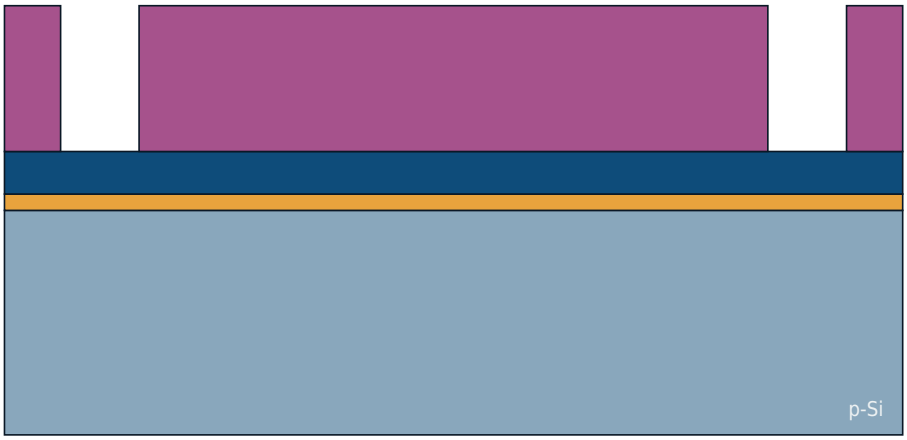
Das Nitrid dient in den folgenden Schritten als Maskierung für die Feldoxidation – es lässt sich später selektiv zum Oxid entfernen.

4. Fototechnik

Auf dem Nitrid wird ein Fotolack aufgebracht, belichtet und entwickelt. So erhält man eine strukturierte Lackschicht, die als Ätzmaske dient.

Fotoresist als Ätzmaske

Lackstrukturierung für die Nitridätzung



Silicium (p-Si) Gateoxid Nitrid Fotoresist

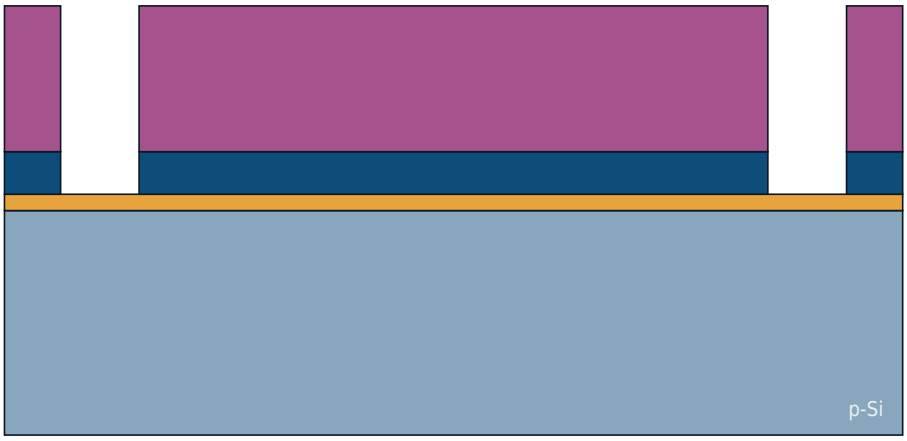
Die Lacköffnungen legen fest, wo das Nitrid im nächsten Schritt entfernt wird – dort entsteht später das Feldoxid.

5. Ätzen

Der Lack maskiert das Nitrid, lackfreie Bereiche werden mittels reaktivem Ionenätzen entfernt.

Nitrid ätzen

Reaktives Ionenätzen überträgt das Lackmuster ins Nitrid



Silicium (p-Si) Gateoxid Nitrid Fotoresist

Durch die Lacköffnungen hindurch wird das Nitrid bis auf das Gateoxid abgetragen.

6. Lackentfernen

Nach der Strukturübertragung der Lackmaske in das Nitrid, wird der Resist mit einer Entwicklerlösung nasschemisch entfernt.

Resist entfernen

Der Fotolack wird mit Entwicklerlösung entfernt

Silicium (p-Si) Gateoxid Nitrid

Übrig bleibt die strukturierte Nitridmaske, die im nächsten Schritt die Feldoxidation lokal verhindert.

7. Oxidation

Das Nitrid dient als Maske, nur auf den Nitrid freien Bereichen findet eine thermische Nassoxidation statt. Das Feldoxid (FOX, z.B. 700 nm) dient zur seitlichen Isolation zu benachbarten Bauteilen.

<?xml version="1.0" encoding="UTF-8"?>

Feldoxid (FOX)

Thermische Nassoxidation wächst nur dort, wo kein Nitrid maskiert

Silicium (p-Si) Gateoxid Nitrid

Wo das Nitrid die Oberfläche abdeckt, bleibt das Substrat unoxidiert. In den freien Bereichen wächst das dickere Feldoxid – der charakteristische Vogelschnabel entsteht am Übergang zum Nitrid.

8. Ätzen

Nach der Oxidation wird das Nitrid in einem nasschemischen Ätzprozess entfernt.

Nitrid entfernen

Nassätzen entfernt die Nitridmaske selektiv

Silicium (p-Si) Gateoxid

Zurück bleibt die fertige Feldoxid-Struktur: dickes Oxid über den Isolationsbereichen, dünnes Gateoxid über dem aktiven Gebiet.

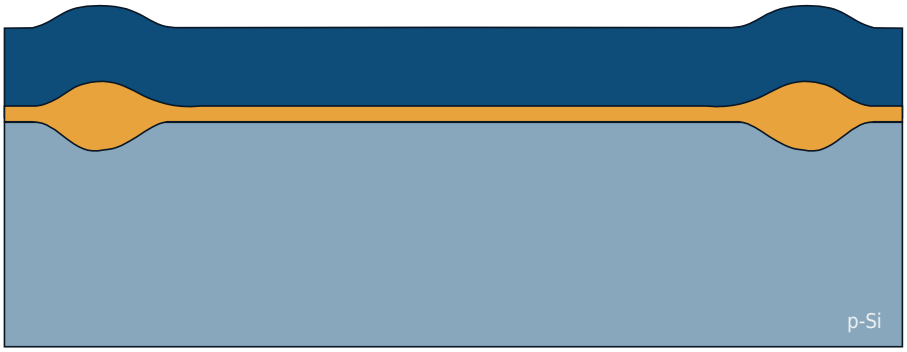
9. Abscheidung

In einem LPCVD-Prozess wird polykristallines Silicium abgeschieden.

<?xml version="1.0" encoding="UTF-8"?>

Polysilicium, Gate

LPCVD-Abscheidung der späteren Gate-Elektrode



Silicium (p-Si) Gateoxid Polysilicium

Das Polysilicium wird ganzflächig abgeschieden und im nächsten Schritt zur Gate-Elektrode strukturiert.

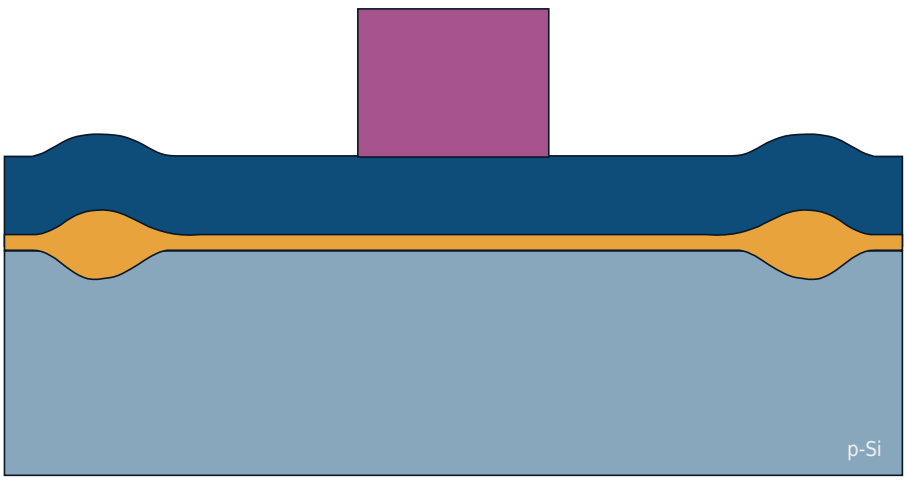
10. Fototechnik

Auf dem Polysilicium wird eine Lackschicht als Ätzmaske strukturiert.

<?xml version="1.0" encoding="UTF-8"?>

Fotoresist als Ätzmaske

Lackstrukturierung für die Gate-Strukturierung



Silicium (p-Si) Gateoxid Polysilicium Fotoresist

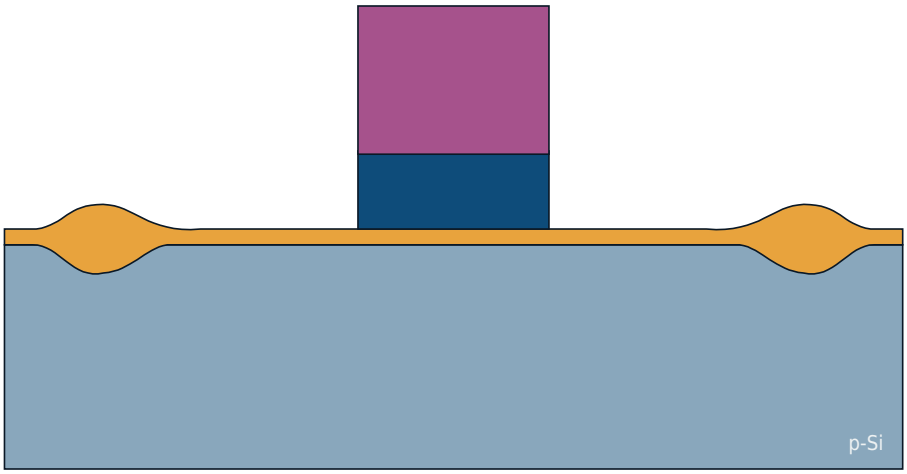
Die Lackinsel markiert genau die spätere Position und Breite der Gate-Elektrode.

11. Ätzen

Der Fotolack dient wiederum als Lackmaske, in einem reaktiven Ionenätzschritt wird das Silicium strukturiert. Es dient als Gateelektrode zur Steuerung des Transistors.

Ätzen von Polysilicium

Reaktives Ionenätzen formt die Gate-Elektrode



Silicium (p-Si) Gateoxid Polysilicium Fotoresist

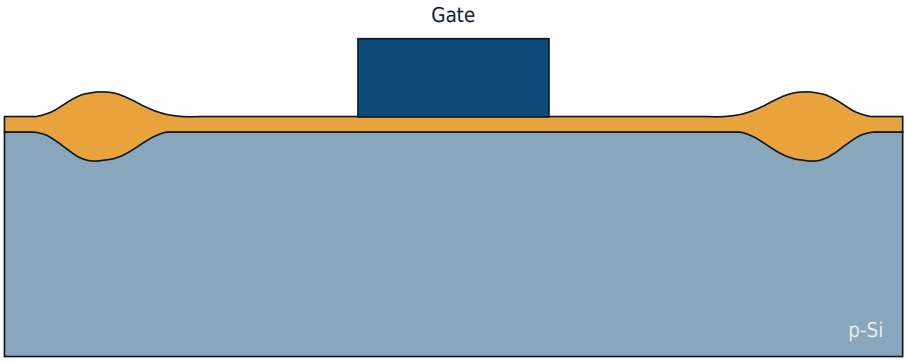
Nur unter dem Lack bleibt Polysilicium stehen – so entsteht die Gate-Elektrode mit exakt definierter Kanallänge.

12. Lackentfernen

Der Lack wird nach dem Ätzschritt wieder nasschemisch entfernt.

Resist entfernen

Der Fotolack wird entfernt, das Gate steht frei



Silicium (p-Si) Gateoxid Polysilicium

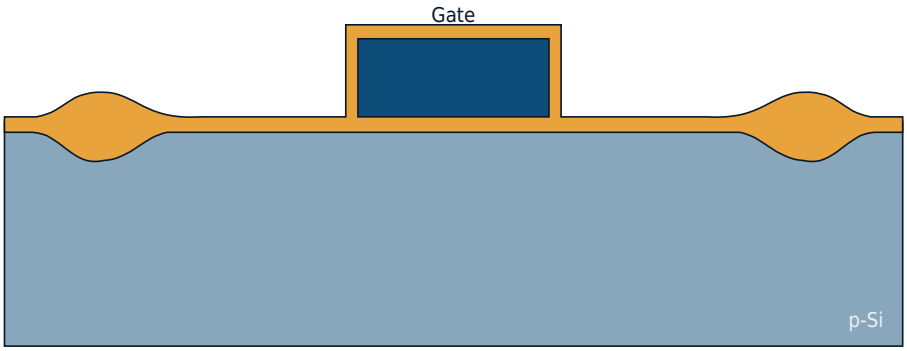
Das fertig strukturierte Gate steht nun frei auf dem dünnen Gateoxid.

13. Oxidation

Nachfolgend wird ein dünnes Oxid, das Postoxid abgeschieden. Es dient zum einen als Schutz der Gateelektrode, zum anderen als Spacer für die Source-/Drainimplantation.

Postoxidation

Thermische Oxidation schützt das Gate und bildet den Spacer



Silicium (p-Si) Gateoxid / Postoxid Polysilicium

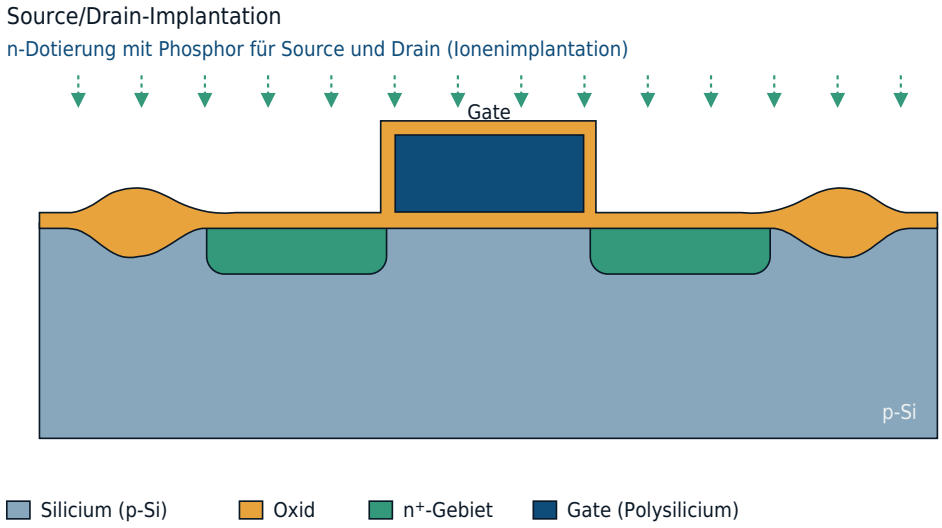
Das Postoxid schützt die Gate-Kanten und dient bei der folgenden Source/Drain-Implantation als Abstandshalter (Spacer).

14. Ionenimplantation

In einem Implantationsschritt mit Phosphorionen werden das Source- und Draingebiet n-dotiert. Da die Gateelektrode als Implantationsmaske dient und so

die Weite des n-Kanals zwischen Source und Drain vorgegeben ist, bezeichnet man dies als Selbstjustierung. Über die Breite der Spacer kann der Abstand der eingebrachten Fremdatome zum Gate je nach elektrischen Anforderungen exakt eingestellt werden.

<?xml version="1.0" encoding="UTF-8"?>



Phosphorionen dringen überall dort ein, wo weder Gate noch Feldoxid im Weg stehen – die Dotierung ist damit selbstjustierend zum Gate. Auch andere Bereiche werden implantiert, aber nur im Substrat werden die Atome elektrisch relevant

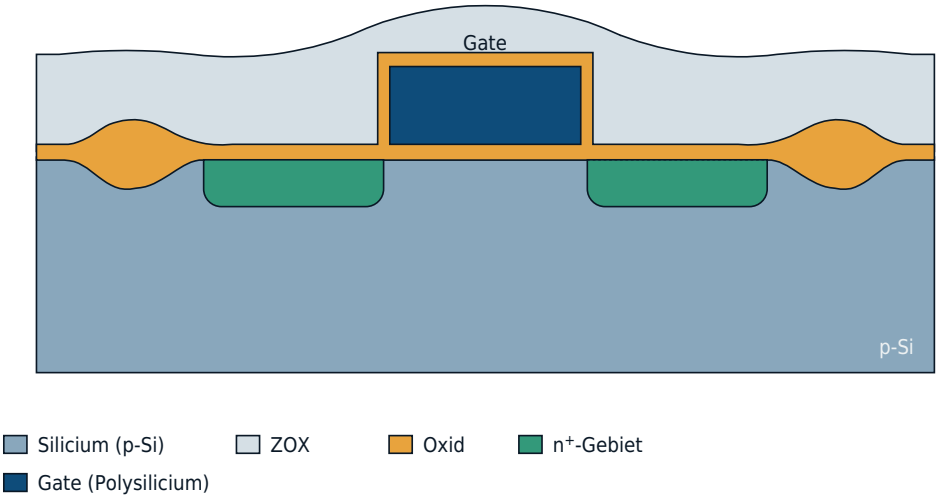
15. Oxidation

Als Isolation zu darüberliegenden Metallisierungsschichten, wird ein Oxid (das Zwischenoxid, kurz ZOX, z.B. 700 nm) abgeschieden. Dies geschieht in einem LPCVD-Prozess mit TEOS, welches eine gute Kantenbedeckung bietet.

<?xml version="1.0" encoding="UTF-8"?>

Zwischenoxid (ZOX)

TEOS-Abscheidung als Isolation zu den späteren Leiterbahnen



Das Zwischenoxid isoliert die spätere Aluminiumverdrahtung vom darunterliegenden Transistor und folgt dabei der vorhandenen Topografie.

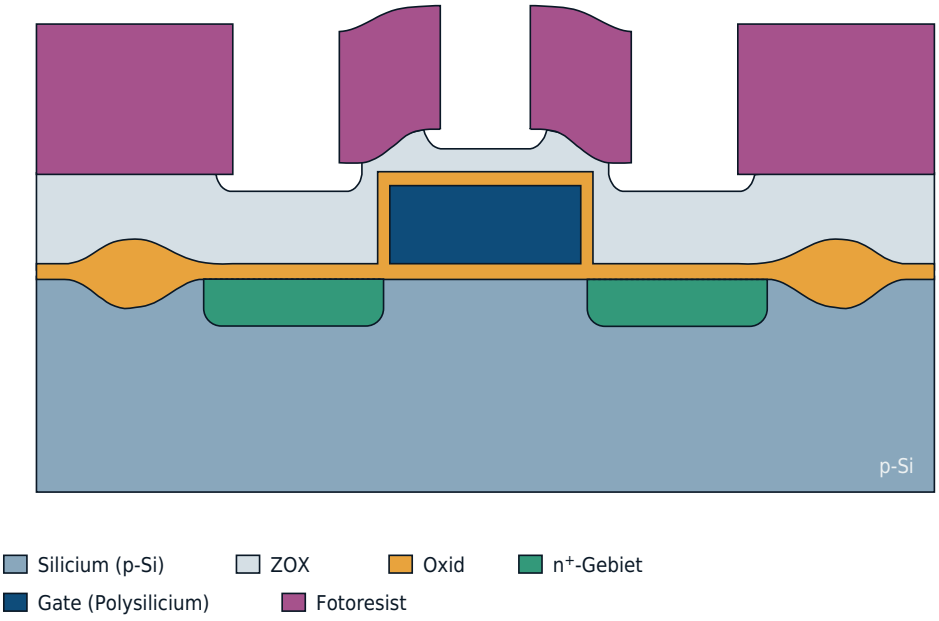
16. Fototechnik und Ätzen

Darüber wird eine weitere Lackschicht strukturiert und in einem isotropen Ätzprozess die Kanten der Kontaktlöcher verrundet.

<?xml version="1.0" encoding="UTF-8"?>

Fotoresist als Ätzmaske

Lackstrukturierung für die spätere Kontaktlochätzung



Die Lacköffnungen legen fest, wo später die Kontaktlöcher zu Source, Drain und Gate geätzt werden.

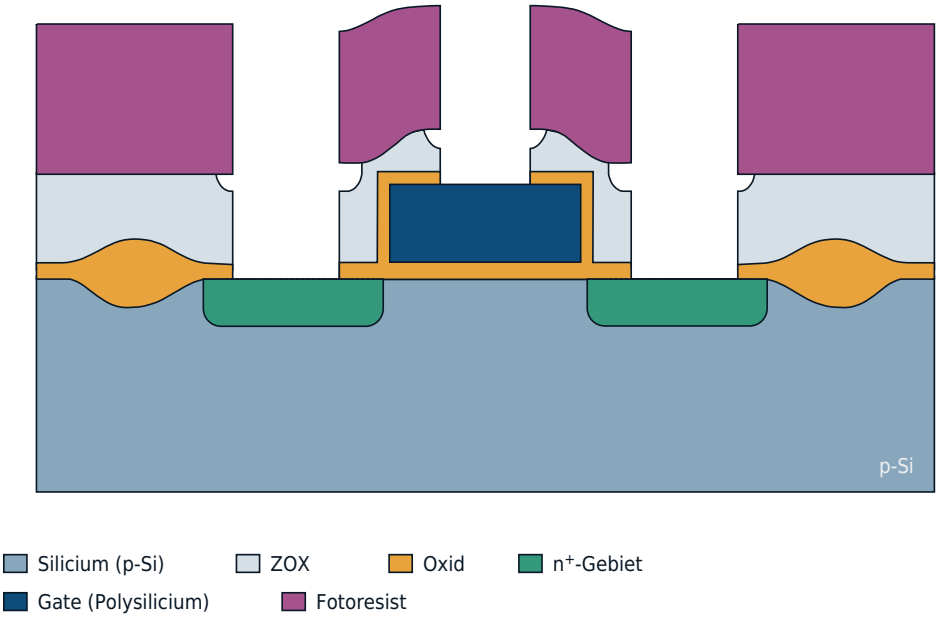
17. Ätzen

Anschließend werden die Kontaktlöcher in einem anisotropen Ätzprozess bis zu den n-dotierten Gebieten und zum Gate freigelegt.

<?xml version="1.0" encoding="UTF-8"?>

Kontaktlochätzen

Reaktives Ionenätzen öffnet das Zwischenoxid bis auf Source, Drain und Gate



Der Lack schützt das übrige Zwischenoxid, während die Löcher bis hinunter zu den drei Anschlussgebieten geätzt werden.

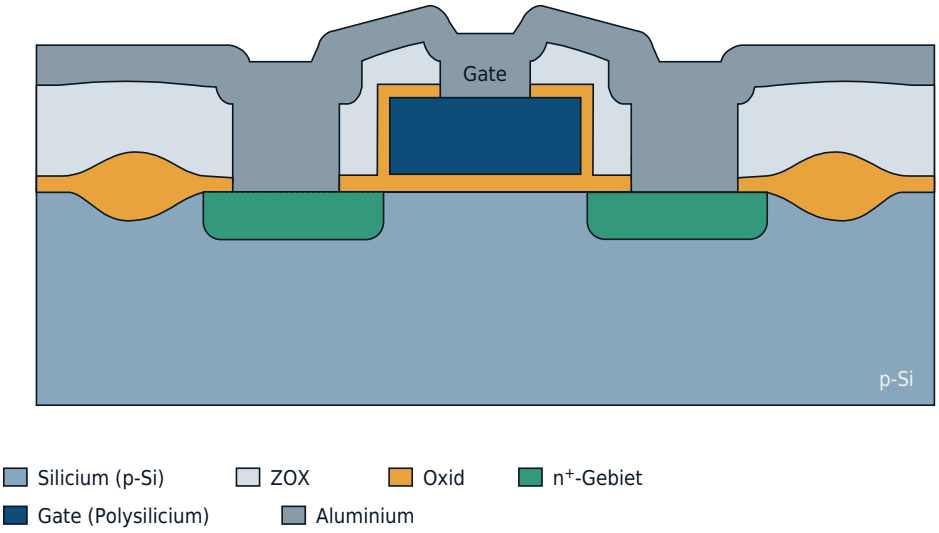
18. Metallisierung

Die Kontaktlöcher werden in einem Sputterprozess mit Aluminium aufgefüllt.

<?xml version="1.0" encoding="UTF-8"?>

Aluminiumabscheidung

Kathodenzerstäubung füllt die Kontaktlöcher und bedeckt die Oberfläche



Das Aluminium füllt die Kontaktlöcher vollständig und bildet eine durchgehende Schicht über der gesamten Struktur.

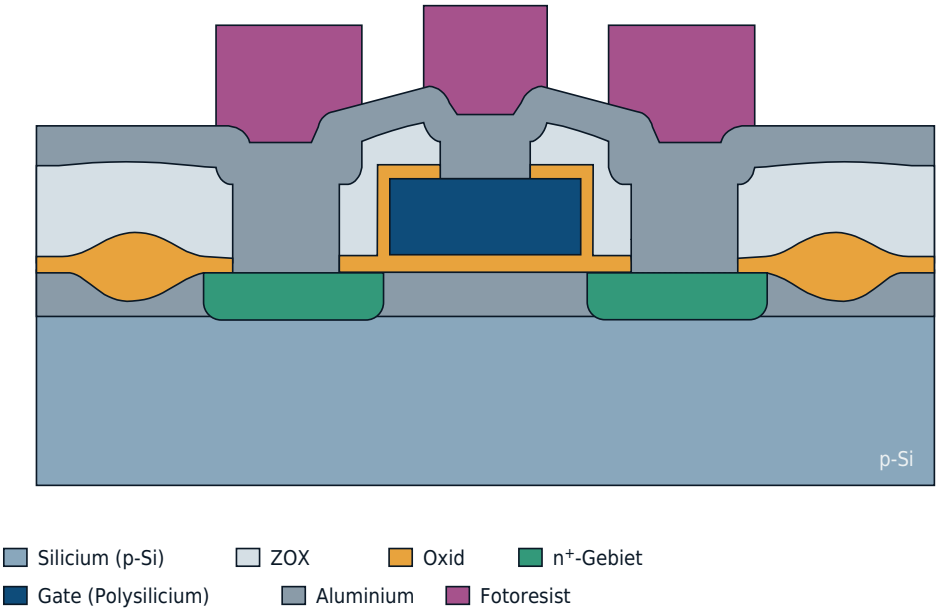
19. Fototechnik

In einem letzten Lithografieschritt wird eine weitere Lackschicht strukturiert.

<?xml version="1.0" encoding="UTF-8"?>

Fotoresist als Ätzmaske

Lackstrukturierung für die Aluminiumstrukturierung



Der Lack legt fest, wo das Aluminium als Leiterbahn stehen bleibt und wo es im nächsten Schritt entfernt wird.

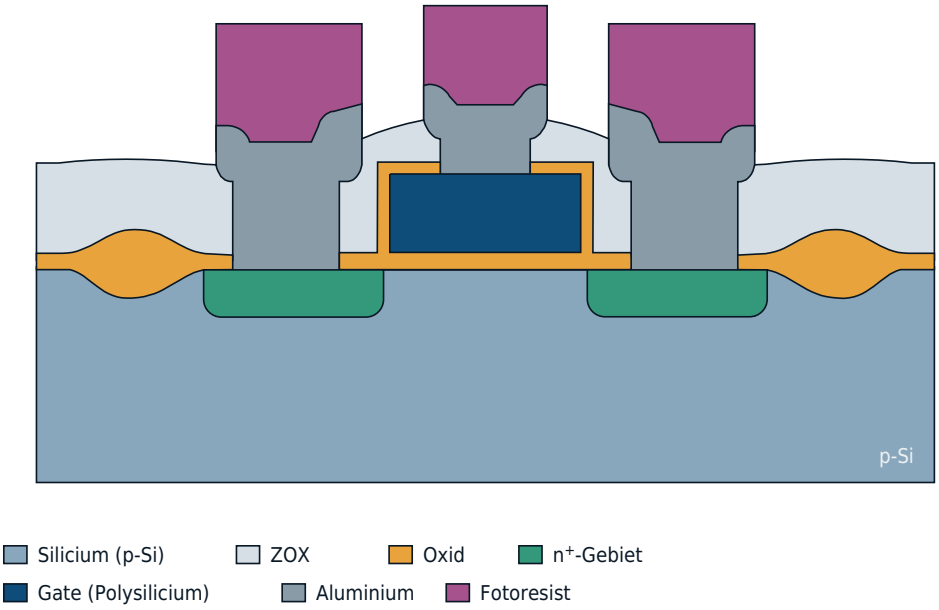
20. Ätzen

Die Lackstrukturen werden in einem anisotropen Trockenätzschritt in die darunterliegende Metallisierungsebene übertragen.

<?xml version="1.0" encoding="UTF-8"?>

Aluminiumätzen

Anisotrope Trockenätzung strukturiert die drei Anschlüsse



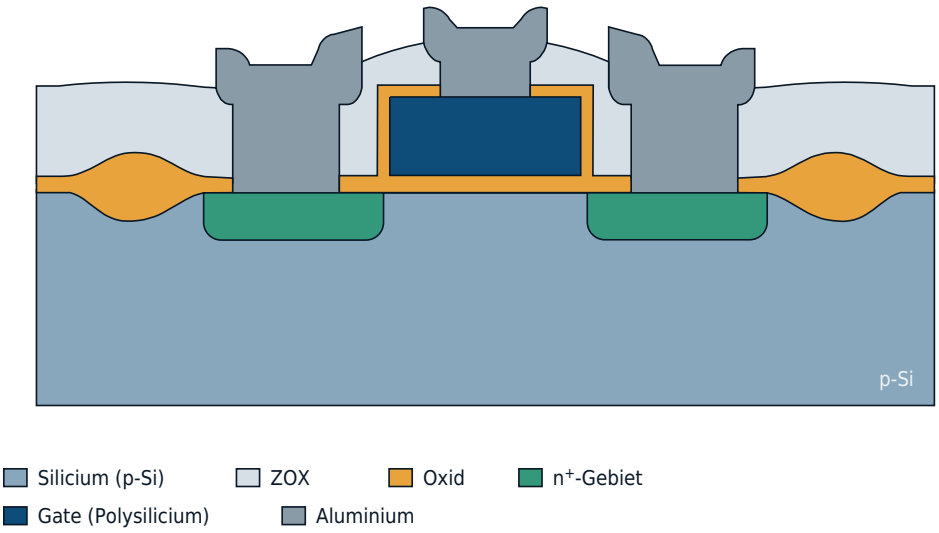
Die Aluminiumstrukturierung trennt die drei Anschlüsse für Source, Gate und Drain voneinander. Im letzten Schritt wird der Lack entfernt.

21. Lackentfernen

Abschließend wird der Resist entfernt, zurück bleiben Aluminiumleiterbahnen zur Ansteuerung des Transistors.

<?xml version="1.0" encoding="UTF-8"?>

Fertiger n-Kanal-Transistor
Querschnitt nach dem letzten Schritt: der Lack ist entfernt, die drei Kontakte liegen frei



Die Aluminiumstrukturierung trennt die drei Anschlüsse für Source,
Gate und Drain voneinander. Im letzten Schritt wird der Lack entfernt.

Der tatsächliche Aufbau eines Transistors ist wesentlich komplexer. So können zusätzliche Planarisierungsschichten zur Unterstützung der lithografischen Prozesse zum Einsatz kommen und auch mehrere Drain-/Sourceimplantationen erfolgen um die Einsatzspannung exakt einzustellen. Ebenso sind weitere Spacer ("Abstandshalter") am Gate möglich, um die Kanallänge exakt zu justieren bzw. das Dotierprofil zu beeinflussen.

Funktionsweise

Anreicherungs-FET: Ohne eine positive Spannung am Gate stehen zwischen Source und Drain keine freien Ladungsträger in Form von Elektronen zur Verfügung, da das Substrat p-dotiert ist. Im stationären Zustand sind hier Löcher die Majoritätsladungsträger und Elektronen Minoritätsladungsträger.

Durch eine positive Spannung am Gate werden Elektronen durch das elektrische Feld aus dem Substrat angezogen (Löcher dementsprechend verdrängt) und bilden so einen leitenden n-Kanal zwischen Source und Drain. Die isolierende Siliciumdioxidschicht verhindert einen Stromfluss zwischen Substrat und Gate.

Da der Transistor den Stromfluss ohne angelegte Spannung sperrt nennt man den Transistor auch selbstsperrend. Der Kanal bildet sich dabei nicht allmählich, sondern erst oberhalb einer bestimmten Gatespannung, der Schwellspannung. Sie ist die wichtigste Kenngröße des Transistors: Unterhalb davon sperrt er, oberhalb leitet er. Über die Dotierung des Substrats und die Wahl des Gatematerials lässt sie sich gezielt einstellen.

Verarmungs-FET: Durch eine leichte n-Dotierung zwischen Source und Gate erreicht man, dass ein Transistor auch ohne Gatespannung leitend ist (eine Spannung zwischen Source und Drain reicht aus). So genannte Verarmungs-FETs, oder auch selbstleitende Transistoren, sperren nur dann, wenn am Gate eine negativere Spannung als am Source-Anschluss anliegt. Dadurch werden die Elektronen, die sich unter dem Gate befinden, verdrängt - die leitende Elektronenbrücke geht verloren.

Aufbau eines npn-Transistors

Allgemeiner Aufbau

Der zweite wichtige Transistortyp neben dem Feldeffekttransistor ist der Bipolartransistor. Seine Funktionsweise beruht auf beiden Ladungsträgern (bipolar), Elektronen und Löchern. Bipolartransistoren sind schneller als Feldeffekttransistoren, beanspruchen jedoch mehr Platz und können somit nicht so kostengünstig gefertigt werden.

Bipolartransistoren bestehen im Wesentlichen aus zwei gegeneinander geschalteten p-n-Übergängen mit der Schichtfolge n-p-n oder p-n-p. Die Anschlüsse des Bipolartransistors werden als Emitter (E), Basis (B) und Kollektor (C) bezeichnet, Emitter und Kollektor besitzen jeweils die gleiche Dotierungsart. Zwischen den beiden Anschlüssen befindet sich die sehr dünne Basisschicht, die dementsprechend jeweils anders dotiert ist.

Beschrieben wird hier ein npn-Transistor in Standard Buried Collector-Bauweise (SBC, *vergrabener Kollektor*). Die Funktionsweise des pnp-Transistors ist analog dazu, die Vorzeichen der angelegten Spannung müssen lediglich vertauscht werden.

Herstellung eines npn-Bipolartransistors

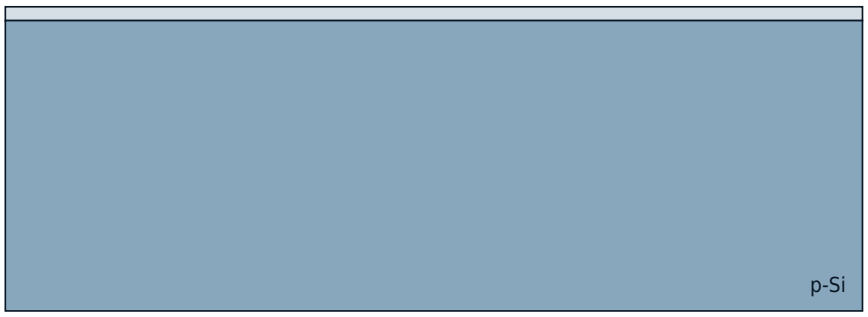
1. Substrat

Grundlage für einen npn-Bipolartransistor ist ein p-dotiertes Siliciumsubstrat,

als Dotierstoff dient Bor. Darauf wird ein dickes Oxid (z.B. 600 nm) abgeschieden.

Ausgangsmaterial

Schwach p-dotierter Silicium-Wafer als Substrat



Silicium (p-Si) Oxid

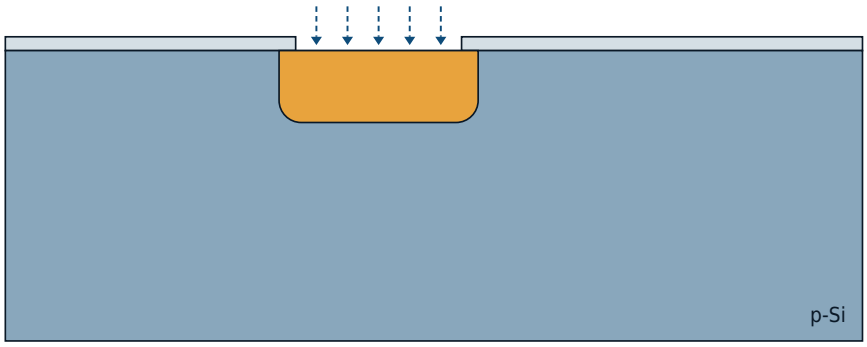
Ausgangspunkt ist ein schwach p-dotierter Silicium-Wafer, auf dessen Oberflaeche eine duenne Oxidschicht als Schutz- und Maskierschicht waechst.

2. Buried-Layer-Implantation

Das Oxid dient als Implantationsmaske, als Dotierstoff wird Antimon Sb eingesetzt, da dies im Vergleich zu Phosphor bei späteren Diffusionsprozessen weniger stark diffundiert. Der stark n^+-dotierte, vergrabene Kollektor dient als niederohmige Kontaktfläche für den Kollektoranschluss.

Fenster oeffnen und implantieren

Lithografie legt das Fenster fuer die vergrabene Schicht fest



Silicium (p-Si) Oxid n^+-dotiert

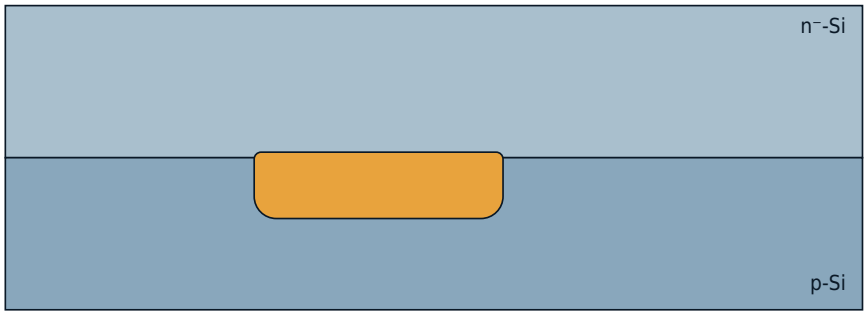
Im freigelegten Bereich erzeugt eine Ionenimplantation ein stark n-dotiertes Gebiet. Diese vergrabene Schicht liegt spaeter unter dem fertigen Transistor.

3. Homoepitaxie

In einem Epitaxieprozess wird eine hochohmige, schwach n^--dotierte Kollektorschicht abgeschieden (typisch 10 µm).

Epitaxie

Eine neue, schwach n-dotierte Siliciumschicht waechst auf



Silicium (p-Si) n^--Si (Epitaxieschicht) n^+-dotiert

Die neue Epitaxieschicht bildet das spaetere aktive Gebiet des Transistors. Die vergrabene n^+-Schicht bleibt an der Grenzflaeche zum Substrat erhalten.

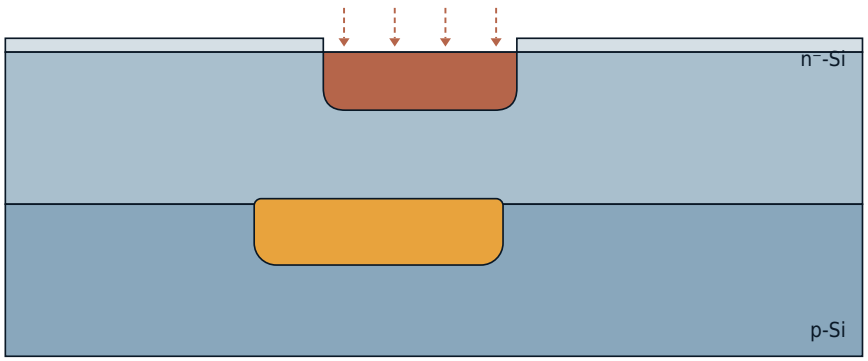
4. Basis-Implantation

Mit Borionen wird die p-dotierte Basis erzeugt, bei einem anschließenden

Diffusionsschritt wird das Gebiet vergrößert.

Basis-Implantation

Ein neues Fenster legt die Position der Basis fest



Silicium (p-Si) n⁻-Si (Epitaxieschicht) Oxid n⁺-dotiert p-dotiert

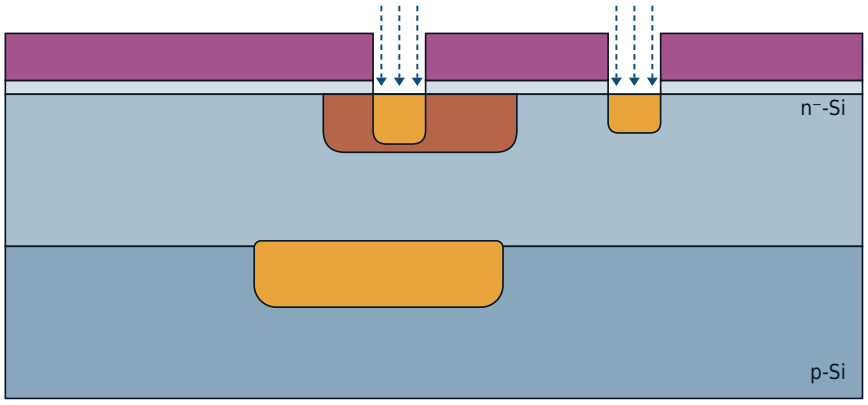
Durch das neue Fenster wird die Basis implantiert: ein schwach p-dotiertes Gebiet direkt oberhalb der vergrabenen n⁺-Schicht.

5. Emitter- und Kollektor-Implantation

Mit Phosphorionen werden die beiden stark n^+-dotierten Emitter- und Kollektoranschlüsse eingebracht.

Emitter- und Kollektor-Implantation

Eine Lackmaske legt zwei neue Fenster fest



Silicium (p-Si) n⁻-Si (Epitaxieschicht) Oxid Fotoresist n⁺-dotiert p-dotiert

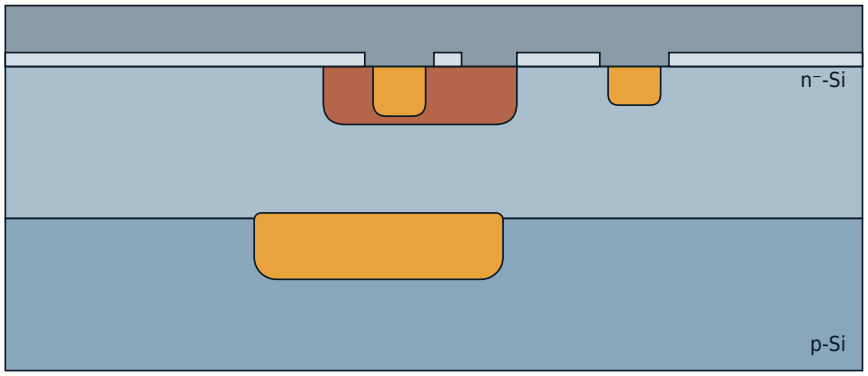
Durch die Lackmaske hindurch entstehen zwei n⁺-Gebiete: der Emitter innerhalb der Basis und ein niederohmiger Kollektor-Anschluss daneben.

6. Metallisierung und Fototechnik

In einem Sputterprozess wird zur Kontaktierung der drei Anschlüsse Aluminium abgeschieden und darüber eine Lackschicht strukturiert.

Kontaktloecher und Aluminium

Nach dem Entfernen des Lacks wird Aluminium aufgebracht



Silicium (p-Si) n⁻-Si (Epitaxieschicht) Oxid n⁺-dotiert
p-dotiert Aluminium

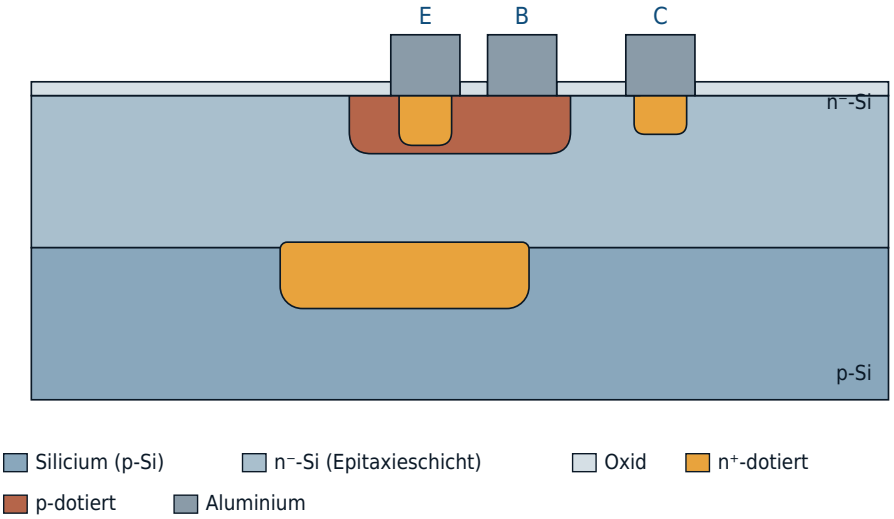
Der Lack wird entfernt, das Oxid an drei Stellen bis auf das Silicium geoeffnet und anschliessend eine durchgehende Aluminiumschicht aufgebracht.

7. Ätztechnik

Abschließend werden die Anschlüsse für Emitter, Basis und Kollektor in einem anisotropen Trockenätzschritt strukturiert.

Aluminium-Strukturierung

Aetzen trennt die drei Anschlusskontakte E, B und C



Die Aluminiumschicht wird zu drei voneinander getrennten Kontakten strukturiert: Emitter (E), Basis (B) und Kollektor (C).

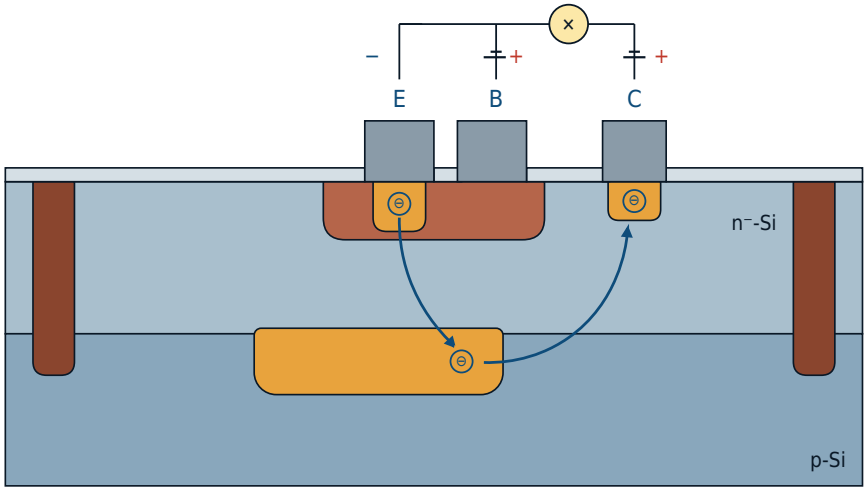
Aufgrund von Optimierungen sind Bipolartransistoren heutzutage aus mehr als drei Schichten (npn bzw. pnp) aufgebaut. Die Kollektorzone besteht hierbei immer aus mindestens zwei unterschiedlich stark dotierten Zonen. Die Bezeichnungen npn und pnp beziehen sich nur auf den aktiven inneren Bereich, jedoch nicht auf den tatsächlichen Aufbau.

Funktionsweise

Die beiden p-n-Übergänge werden im Folgenden als EB (Emitter-Basis) bzw. CB (Kollektor-Basis) bezeichnet. Ohne äußere Spannung bilden sich an EB und CB Raumladungszonen aus (siehe Der p-n-Übergang). Mit einer negativen Spannung am Emitter und einer positiven Spannung am Kollektor wird die Raumladungszone an EB abgebaut, an CB jedoch vergrößert. Wird an der Basis nun eine positive Spannung angelegt, so wird EB leitend - Elektronen gelangen in die Basisschicht. Da diese sehr dünn ist, können die Ladungsträger in den Kollektor injiziert werden, wo sie auf Grund der angelegten positiven Spannung abgesaugt werden. Somit fließt ein Strom von Emitter zu Kollektor. Nahezu alle Elektronen gelangen so bereits bei einer geringen Spannung an der Basis zum Kollektor (>95 %), was bedeutet, dass mit einem relativ kleinen Basisstrom (Emitter zu Basis) ein sehr großer Kollektorstrom (Emitter zu Kollektor) ermöglicht wird.

Fertiger Bipolartransistor

Elektronenfluss vom Emitter ueber die Basis zum Kollektor



Silicium (p-Si) | n⁻-Si (Epitaxieschicht) | Oxid | n⁺-dotiert | p-dotiert | p⁺-dotiert | Aluminium

Am Emitter-Basis-Uebergang in Durchlassrichtung injizierte Elektronen durchqueren die duenne Basis und werden vom Kollektorfeld abgesaugt. Seitliche p⁺-Diffusionen isolieren den Transistor von Nachbarbauteilen.

Die beiden tiefen p^+-dotierten Gebiete dienen zur seitlichen Isolation gegen andere Bauteile. Neben dem Transistor ist zudem ein Widerstand notwendig (nicht in Grafik), da Bipolartransistoren nicht stromlos angesteuert werden können.

Aufbau eines FinFET

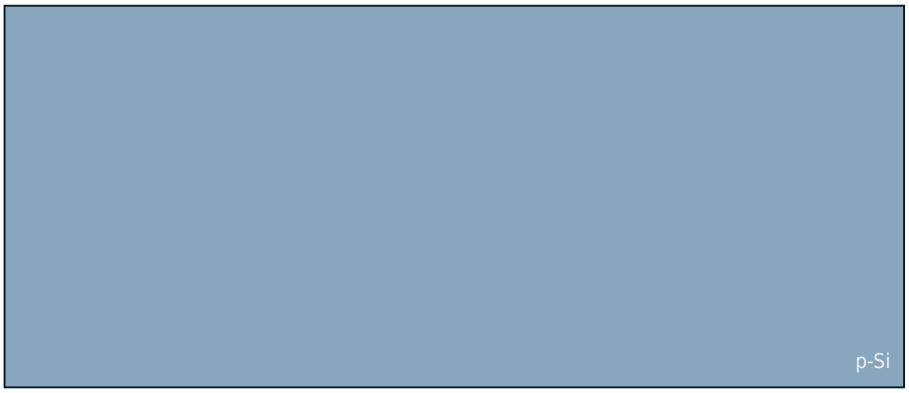
Aufbau eines FinFET im Bulkprozess

1. Substrat

Grundlage für einen FinFET ist, wie für jeden Siliciumtransistor, ein schwach p-dotiertes Siliciumsubstrat.

Substrat

Siliciumsubstrat, p-dotiert mit Bor



Silicium (p-Si)

Grundlage für die gesamte Fertigung ist ein schwach p-dotiertes Siliciumsubstrat als Wafer.

2. Hartmaske

Auf das Substrat wird eine Hartmaske aufgebracht, meist ein Stapel aus Oxid und Siliciumnitrid. Sie überträgt später das Lithografiemuster deutlich verlustärmer in das Silicium als ein reiner Lackprozess.

Hartmaske

Oxid/Nitrid-Stapel als Ätzmaske für die Fin-Strukturierung

Silicium (p-Si) Hartmaske (Oxid/Nitrid)

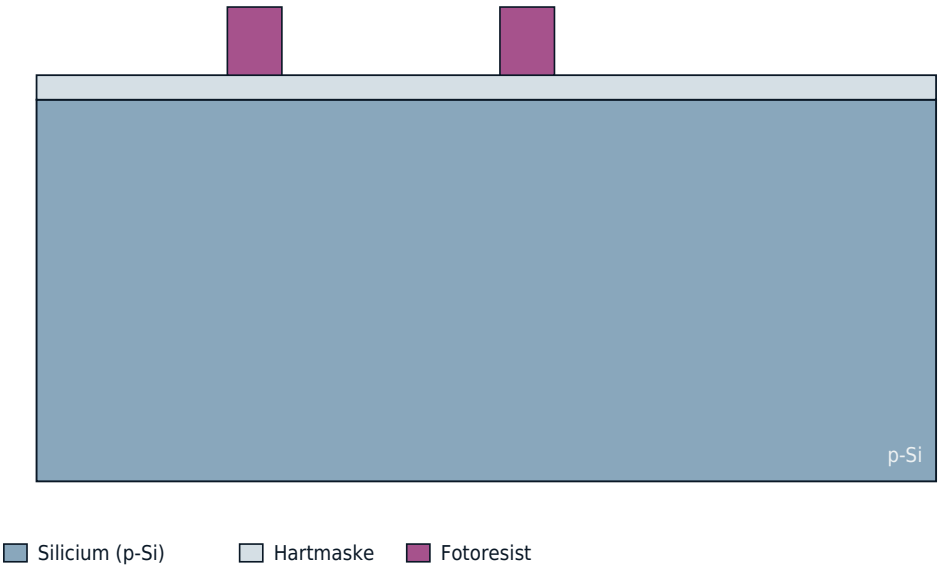
Eine Hartmaske aus Oxid und Nitrid wird abgeschieden – sie überträgt später das Lithografiemuster verlustarm in das Silicium.

3. Fin-Lithografie

Fotoresist wird zu schmalen Streifen strukturiert, welche die Lage und Breite der späteren Finnen festlegen. Die Finnenbreite lag bei den ersten Serienprozessen bei 10 bis 15 nm und ist seither auf wenige Nanometer gesunken; die Höhe sollte idealerweise dem Doppelten oder mehr entsprechen. So schmale Stege lassen sich nicht mehr direkt belichten, sie entstehen über Multipatterning: Die Finnenbreite ergibt sich aus der Dicke einer abgeschiedenen Schicht, nicht aus einer Maske.

Fin-Lithografie

Lackstrukturierung legt Lage und Breite der Fins fest



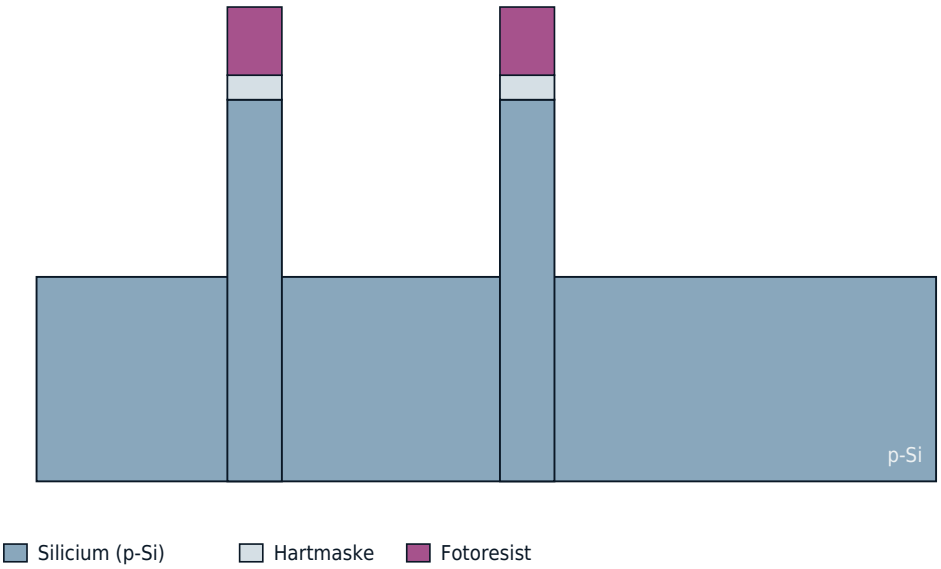
Die schmalen Lackstreifen definieren die spätere Fin-Breite – typischerweise nur wenige Zehn Nanometer.

4. Fin-Ätzung

In einem stark anisotropen Trockenätzschritt werden die freistehenden Finnen in das Siliciumsubstrat geätzt. Beim Bulk-Prozess muss die Ätzung mit Festzeit erfolgen, da anders als bei einem SOI-Substrat keine Stoppschicht im Silicium Aufschluss über die erreichte Tiefe gibt.

Fin-Ätzung

Anisotropes Ätzen überträgt das Muster in Hartmaske und Silicium



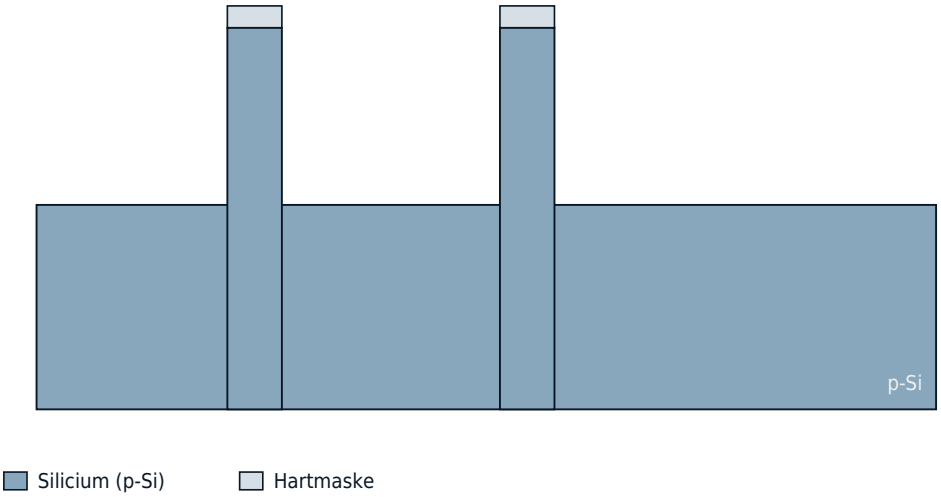
Reaktives Ionenätzen trägt Silicium zwischen den Fins ab – die Ätztiefe bestimmt später, wie weit die Fins aus dem STI herausragen.

5. Hartmaske entfernen

Lack und Hartmaske werden entfernt. Die frei geätzten Finnen stehen bereit für die Isolation.

Resist entfernen

Nur der Lack wird entfernt – die Hartmaske bleibt als Kappe auf den Fins stehen



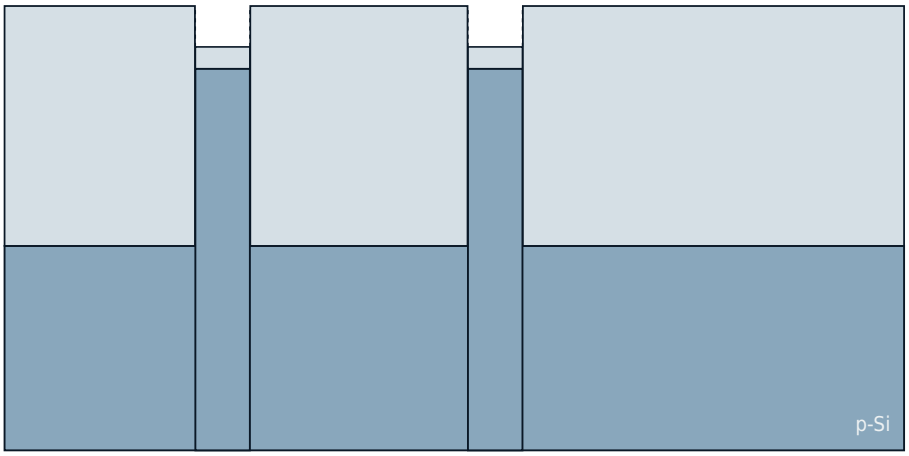
Nur der Fotoresist wird entfernt. Die Hartmaske bleibt bewusst auf den Fins stehen – sie dient beim folgenden STI-Prozess als Poliér-Stoppschicht und wird erst nach der Planarisierung wieder entfernt (Schritt 8).

6. STI-Oxid

Zur Isolation folgt eine Oxidabscheidung (Shallow Trench Isolation, STI), die ein gutes Füllverhalten von schmalen, tiefen Gräben erlauben muss – sie überfüllt die Gräben zwischen den Finnen dabei vollständig.

STI-Oxid

Oxidabscheidung füllt die Gräben zwischen den Fins auf



Silicium (p-Si) STI-Oxid / Hartmaske

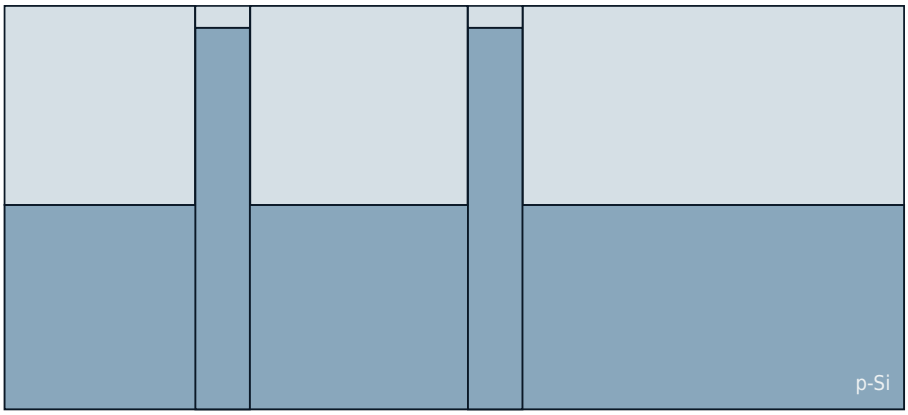
Das Isolationsoxid (Shallow Trench Isolation, STI) wird ganzflächig abgeschieden und überfüllt dabei die Gräben und die Hartmasken-Kappen.

7. STI-Planarisierung

Mittels chemisch-mechanischem Polieren wird das Oxid eingeebnet. Die Hartmaske dient dabei als Stoppschicht.

STI-Planarisierung

CMP poliert das Oxid zurück – die Hartmaske stoppt den Abtrag



Silicium (p-Si) STI-Oxid / Hartmaske

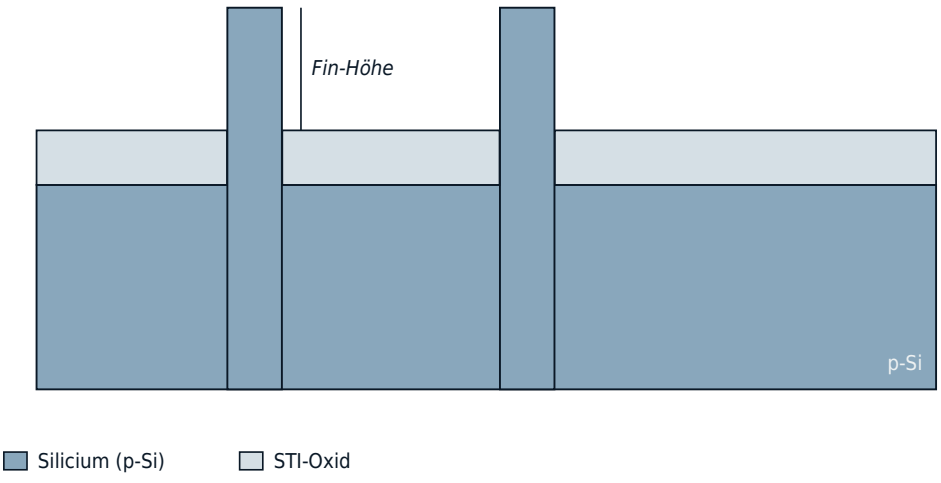
Chemisch-mechanisches Polieren trägt das überschüssige Oxid ab. Die Hartmaske ist deutlich härter als das Oxid und stoppt den Abtrag zuverlässig auf gleicher Höhe – ohne sie würden die Fins angeschliffen.

8. STI-Recess

In einem weiteren Ätzschritt mit Festzeit wird das Oxid gezielt zurückgeätzt, bis die Finnen auf der gewünschten Höhe aus dem Oxid herausragen – die für den FinFET charakteristische dreidimensionale Struktur entsteht. Zurück bleibt eine seitliche Isolation zwischen den Finnen. Da diese über das Substrat weiterhin in Kontakt stehen, muss ein Dotierschritt erfolgen, durch den die Siliciumstege elektrisch voneinander isoliert werden (nicht dargestellt). Alternativ kann das Oxid in einem Hochtemperaturschritt in den Bodenbereich der Finnen hineinwachsen, sodass die Kanäle vom Siliciumsubstrat isoliert werden. In der Serienfertigung hat sich der erste Weg durchgesetzt: Unterhalb der Finne wird gezielt hoch dotiert, sodass sich dort kein leitender Pfad ausbilden kann.

STI-Recess

Hartmaske und Oxid werden zurückgeätzt – die Fin-Flanken liegen frei



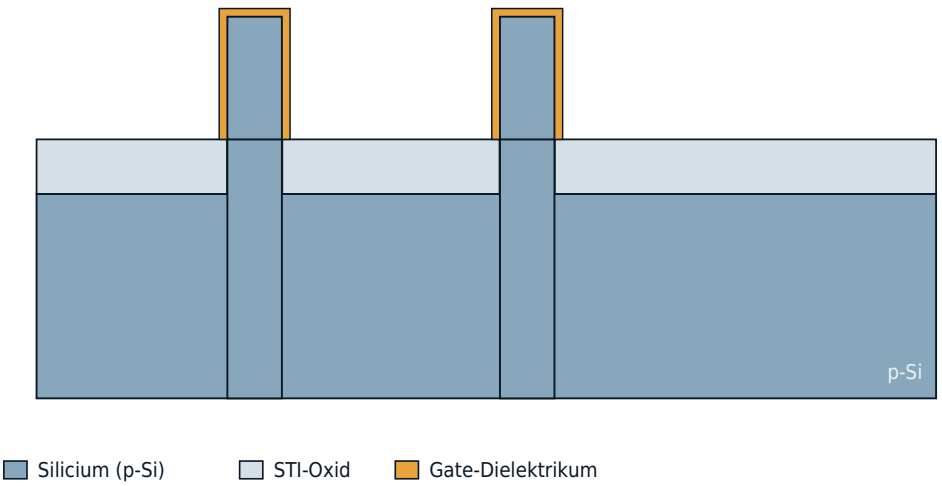
Zunächst wird die Hartmaske selektiv entfernt, danach das STI-Oxid zurückgeätzt, bis die Fins auf der gewünschten Höhe herausragen. Diese Silicium-Flanke bildet den Kanal – das Gate umschließt sie von drei Seiten.

9. Gate-Dielektrikum

Auf den freiliegenden Finnenflanken wird das Gatedielektrikum erzeugt, um den Kanalbereich von der Gateelektrode zu isolieren. In der Serienfertigung ist das kein thermisch gewachsenes Oxid mehr, sondern eine per Atomlagenabscheidung aufgebrachte Schicht aus Hafniumdioxid auf einer dünnen Zwischenlage aus Siliciumdioxid. Nur die Atomlagenabscheidung bedeckt eine hohe, schmale Finne gleichmäßig an allen drei freiliegenden Seiten.

Gate-Dielektrikum

High-k-Schicht wächst konform auf der freiliegenden Fin-Flanke



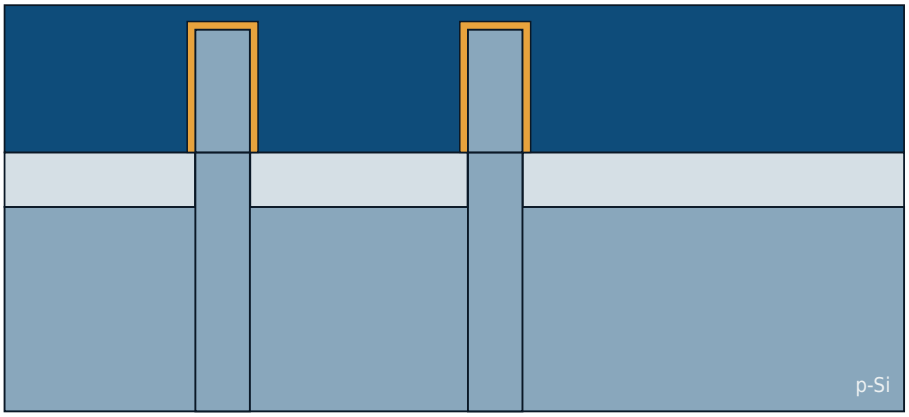
Eine dünne High-k-Schicht (z. B. Hafniumoxid) umschließt die freiliegende Fin-Flanke von drei Seiten – die spätere Kanaloberfläche.

10. Dummy-Gate

Als eigentliches Gatematerial dient später ein Metall, dessen Austrittsarbeit so gewählt wird, dass sich die gewünschte Schwellspannung ergibt – für n- und p-Kanal-Transistoren jeweils ein anderes Metall. Weil dieses Metall die späteren Hochtemperaturschritte nicht überstehen würde, wird es erst ganz zum Schluss eingebaut. Zunächst setzt man ein Platzhaltergate aus Polysilicium, das hier konform abgeschieden wird und die Finnen bereits vollständig umschließt. Alle heißen Prozessschritte laufen mit diesem Platzhalter, der anschließend wieder entfernt wird (Schritt 18). Dieses Vorgehen heißt Gate-Last- oder Replacement-Gate-Verfahren.

Dummy-Gate (Polysilicium)

Polysilicium wird ganzflächig als Platzhalter-Gate abgeschieden



Silicium (p-Si) STI-Oxid Gate-Dielektrikum

Dummy-Gate (Polysilicium)

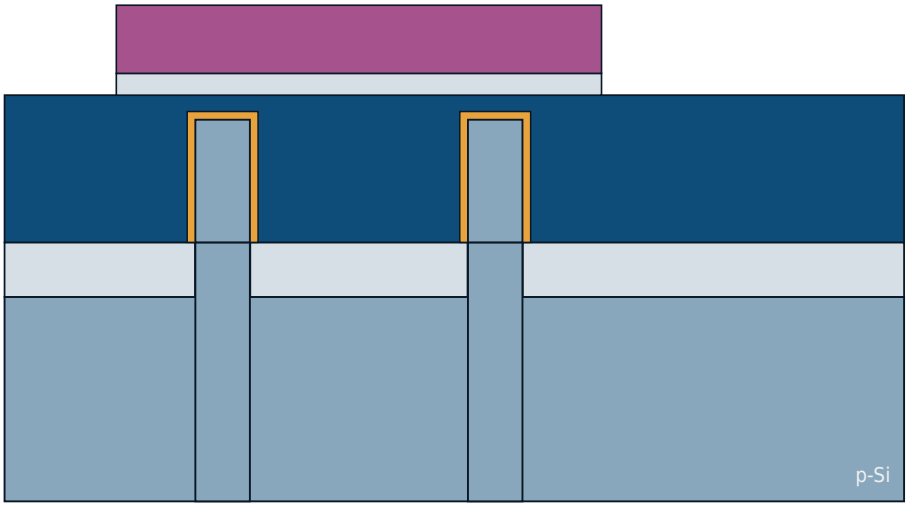
Im Gate-Last-Verfahren dient Polysilicium zunächst nur als Platzhalter – es wird nach der Source/Drain-Bildung wieder entfernt.

11. Gate-Lithografie

Ein Lackstreifen quer über den Finnen legt zusammen mit einer Hartmaske Lage und Länge des Gates fest – senkrecht zur hier gezeigten Schnittebene.

Gate-Lithografie

Lack und Hartmaske legen die Gate-Länge fest



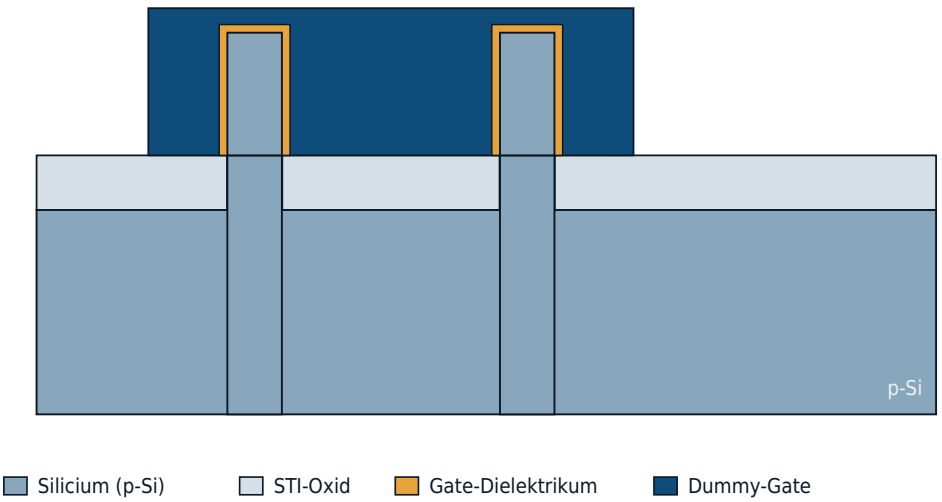
Der Lackstreifen quer über den Fins definiert die spätere Gate-Länge – senkrecht zur hier gezeigten Schnittebene.

12. Dummy-Gate-Ätzung

Überschüssiges Polysilicium wird durch reaktives Ionenätzen entfernt. Übrig bleibt der Dummy-Gate-Stapel, der die Finnen weiterhin von drei Seiten umschließt.

Dummy-Gate-Ätzung

Reaktives Ionenätzen formt den Gate-Stapel



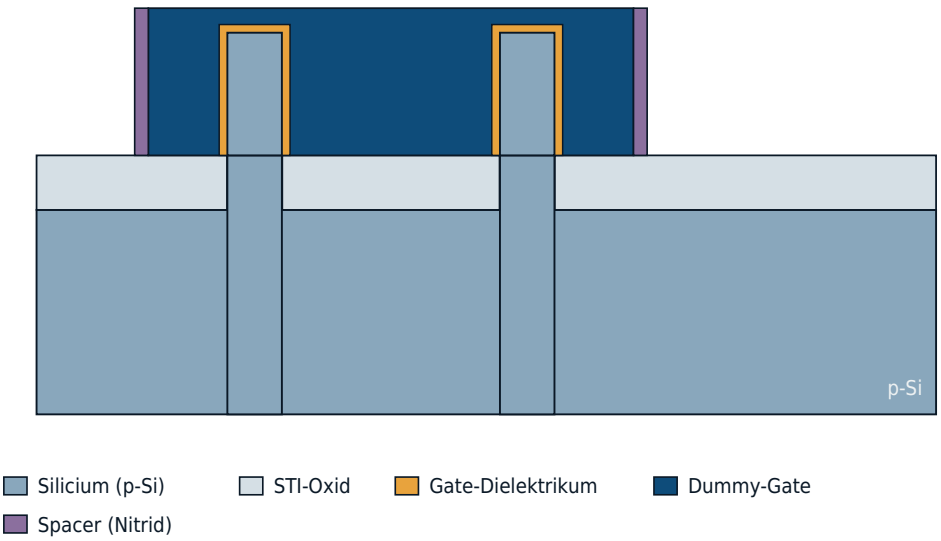
Überschüssiges Polysilicium wird entfernt – übrig bleibt der Dummy-Gate-Stapel, der die Fins von drei Seiten umschließt.

13. Spacer

Ein Nitrid-Spacer wird konform abgeschieden und anisotrop rückgeätzt, sodass er als Abstandshalter an den Gate-Flanken stehen bleibt. Er schützt die Flanken in den folgenden Schritten und legt später den Abstand zwischen Gate und Source/Drain-Kontakten fest. Eine ähnliche Nitridschicht zwischen Finne und Gate kann außerdem gezielt genutzt werden, um den Einfluss des oberen Gate-Segments auf den Kanal zu unterbinden, falls nur die seitliche Steuerung erwünscht ist.

Spacer

Nitridabscheidung und Rückätzen bilden den Abstandshalter am Gate



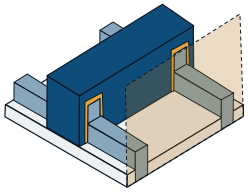
Der Nitrid-Spacer schützt die Gate-Flanken und legt später den Abstand zwischen Gate und Source/Drain-Kontakten fest.

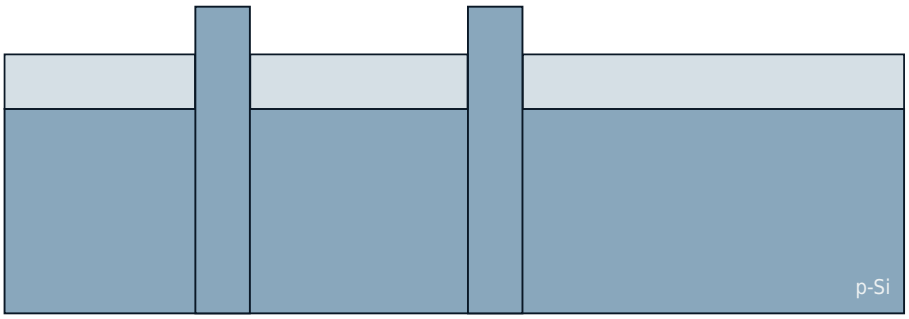
14. Source/Drain-Recess

Außerhalb des Gates – im Schnitt durch den späteren Source/Drain-Bereich – wird die Finne gezielt abgesenkt, um Platz für die nachfolgende Epitaxie zu schaffen. Source und Drain liegen dabei entlang derselben Finne: einmal vor und einmal hinter dem Gate, in Stromflussrichtung. Jede einzelne Finne bildet also für sich einen vollständigen Kanal mit eigenem Source- und Drain-Gebiet an ihren beiden Enden – bei mehreren parallelen Finnen unter einem gemeinsamen Gate trägt demnach jede Finne sowohl zu Source als auch zu Drain bei, nicht etwa die eine Finne nur zu Source und die andere nur zu Drain.

Source/Drain-Recess

Der Fin wird außerhalb des Gates abgesenkt – Schnitt im S/D-Bereich





Silicium (p-Si) STI-Oxid

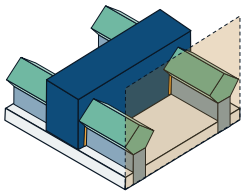
Außerhalb des Gates – im Schnitt durch den späteren Source/Drain-Bereich – wird der Fin gezielt abgesenkt, um Platz für die nachfolgende Epitaxie zu schaffen.

15. Source/Drain-Epitaxie

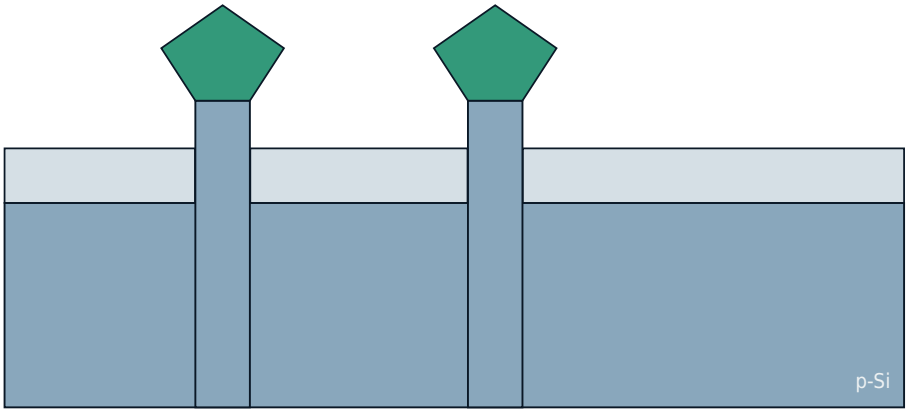
Selektive Epitaxie lässt dotiertes Silicium (bei n-Kanal-Transistoren meist Si:P, bei p-Kanal-Transistoren SiGe:B) rautenförmig auf dem Fin-Stumpf aufwachsen. Diese charakteristische „Raised Source/Drain"-Geometrie vergrößert den wirksamen Kanalquerschnitt zusätzlich und senkt den Zugangswiderstand. Bei p-Kanal-Transistoren kommt ein weiterer Effekt hinzu: SiGe besitzt eine größere Gitterkonstante als Silicium und wird beim epitaktischen Aufwachsen auf der Finne lateral gestaucht, um sich an dessen Kristallgitter anzupassen. Die daraus resultierende kompressive Verspannung überträgt sich unmittelbar in den angrenzenden Kanal und erhöht dort – uniaxial in Stromrichtung wirkend – gezielt die Löcherbeweglichkeit und damit den Antriebsstrom.

Source/Drain-Epitaxie

Raised Source/Drain wächst selektiv auf dem abgesenkten Fin



Schnitt: Source/Drain



Silicium (p-Si) STI-Oxid Epitaktisches Si (n/p-dotiert)

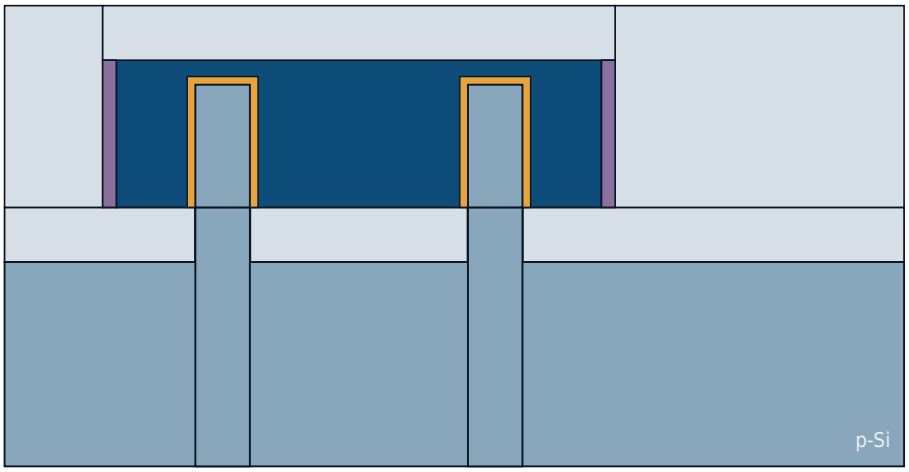
Selektive Epitaxie lässt dotiertes Silicium (bei nFET meist Si:P) rautenförmig auf dem Fin-Stumpf aufwachsen – die charakteristische 'Raised Source/Drain'-Geometrie erhöht den Kanalquerschnitt zusätzlich.

16. Zwischenoxid

Das Zwischenoxid (ILD) wird ganzflächig abgeschieden und bedeckt zunächst auch den Dummy-Gate-Stapel vollständig.

Zwischenoxid (ILD)

Oxidabscheidung isoliert den Gate-Stapel – Schnitt durch das Gate



Silicium (p-Si) STI-Oxid / Zwischenoxid Gate-Dielektrikum
Dummy-Gate Spacer

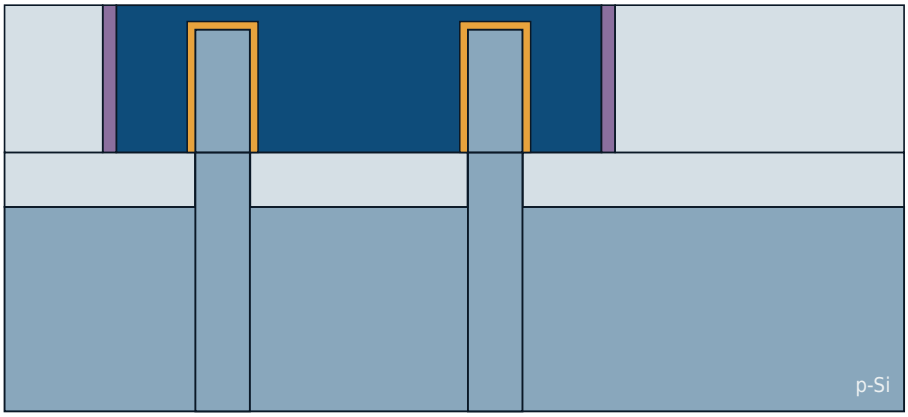
Das Zwischenoxid (ILD) wird ganzflächig abgeschieden und überdeckt zunächst auch den Dummy-Gate-Stapel vollständig.

17. CMP

Chemisch-mechanisches Polieren trägt das überschüssige Zwischenoxid ab, bis die Oberkante des Dummy-Gates freiliegt.

CMP

Planarisierung legt die Oberkante des Dummy-Gates frei



Silicium (p-Si) STI-Oxid / Zwischenoxid Gate-Dielektrikum
Dummy-Gate Spacer

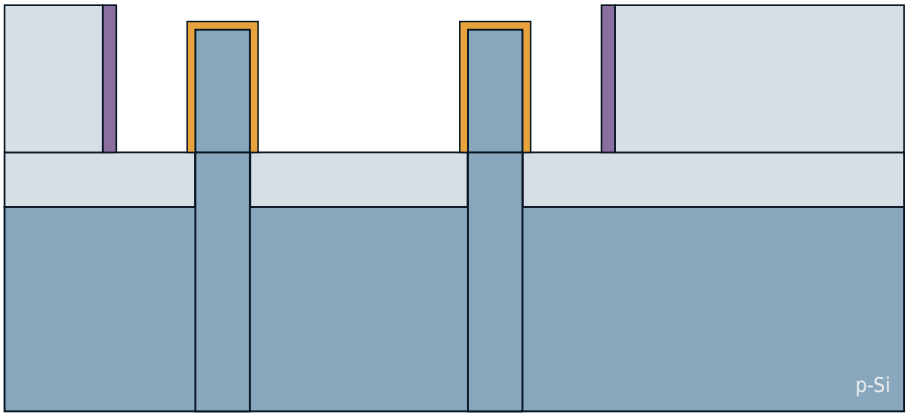
Chemisch-mechanisches Polieren trägt das überschüssige Zwischenoxid ab, bis die Oberkante des Dummy-Gates freiliegt.

18. Dummy-Gate entfernen

Das Polysilicium des Platzhaltergates wird selektiv herausgeätzt – der Spacer bildet die Wände des freien Gate-Grabens, bereit für das eigentliche Metallgate.

Dummy-Gate entfernen

Selektives Ätzen öffnet den Gate-Graben



Silicium (p-Si) STI-Oxid / Zwischenoxid Gate-Dielektrikum Spacer

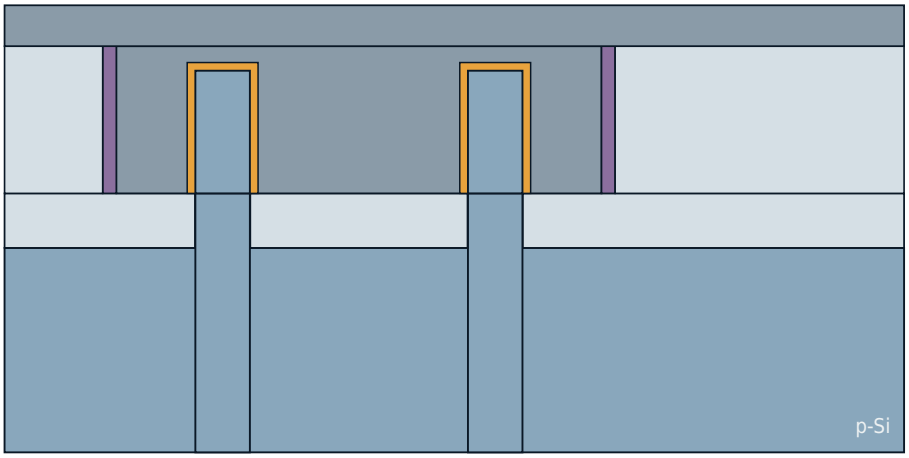
Das Polysilicium wird selektiv herausgeätzt – der Spacer bildet die Wände des freien Gate-Grabens, bereit für das Metallgate.

19. High-k/Metallgate

Austrittsarbeits- und Füllmetall werden konform in den Graben abgeschieden. Weil sich links und rechts vom Kanal sowie darüber grundsätzlich unterschiedliche Metallschichten ergeben können, entstehen so bei Bedarf mehrere, getrennt einstellbare Gateelektroden für einen einzigen Transistor. Das finale Gate umschließt den Kanal von drei Seiten.

High-k/Metallgate

Das finale Metallgate wird in den Graben abgeschieden



Silicium (p-Si) STI-Oxid / Zwischenoxid Gate-Dielektrikum

Metallgate Spacer

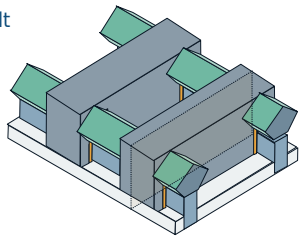
Austrittsarbeits- und Füllmetall werden konform abgeschieden und füllen dabei nicht nur den Graben, sondern überfüllen zunächst die gesamte Oberfläche – erst die folgende CMP trägt den Überschuss ab.

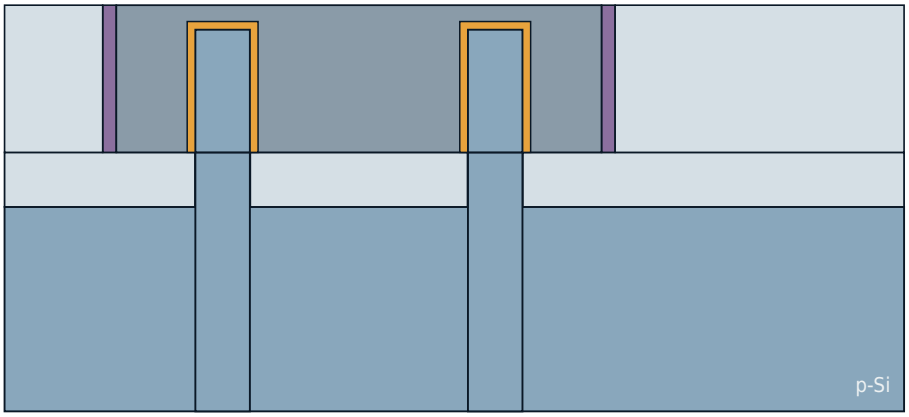
20. Gate-CMP

Ein letzter Polierschritt trägt überschüssiges Gate-Metall ab und trennt benachbarte Gates elektrisch voneinander.

Gate-CMP

Überschüssiges Metall wird zurückpoliert – die Gates sind vereinzelt





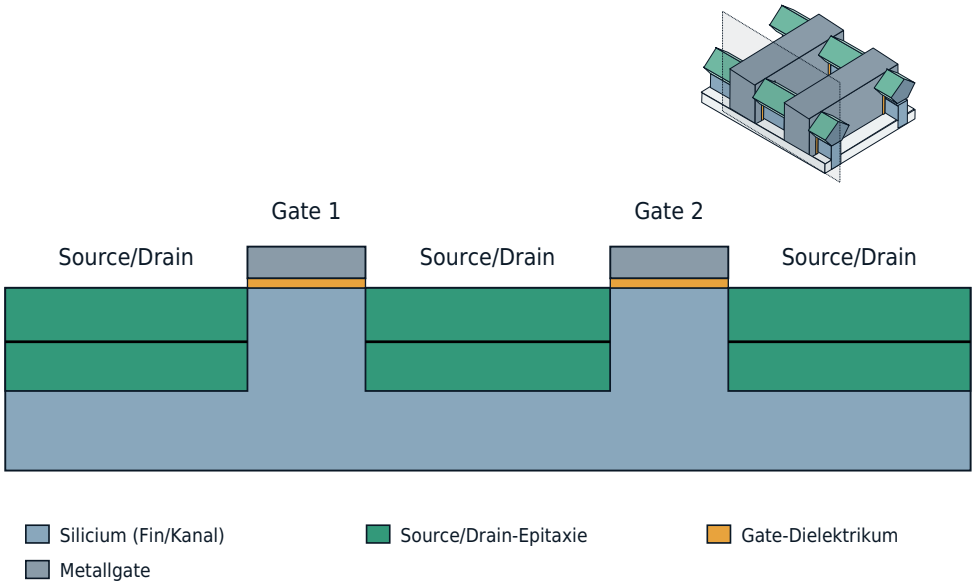
Silicium (p-Si) STI-Oxid / Zwischenoxid Gate-Dielektrikum
Metallgate Spacer

Ein letzter Polierschritt trägt überschüssiges Gate-Metall ab und trennt benachbarte Gates elektrisch voneinander.

Welcher der beiden Source/Drain-Bereiche einer Finne tatsächlich als Source und welcher als Drain wirkt, ist dabei keine feste Eigenschaft der Struktur, sondern hängt vom angelegten Potential ab: Bei einem n-Kanal-Transistor übernimmt der Bereich mit dem niedrigeren Potential die Rolle der Source, der mit dem höheren die der Drain (bei p-Kanal-Transistoren umgekehrt) – leitend geschaltet wird der Kanal in jedem Fall ausschließlich durch das Gate darüber. Deshalb lassen sich mehrere Gates unabhängig voneinander auf derselben Finne platzieren, wie im folgenden Bild gezeigt: Jedes Gate steuert nur den Kanalabschnitt direkt unter sich selbst, unabhängig davon, ob ein benachbartes Gate gerade leitet oder sperrt. Der gemeinsame Source/Drain-Bereich zwischen zwei Gates ist elektrisch schlicht ein Knoten, der je nach Schaltzustand mal als Drain des einen, mal als Source des anderen Transistors wirkt.

FinFET im Längsschnitt

Schnitt entlang einer Finne – zwei Gates mit Source/Drain-Bereichen dazwischen

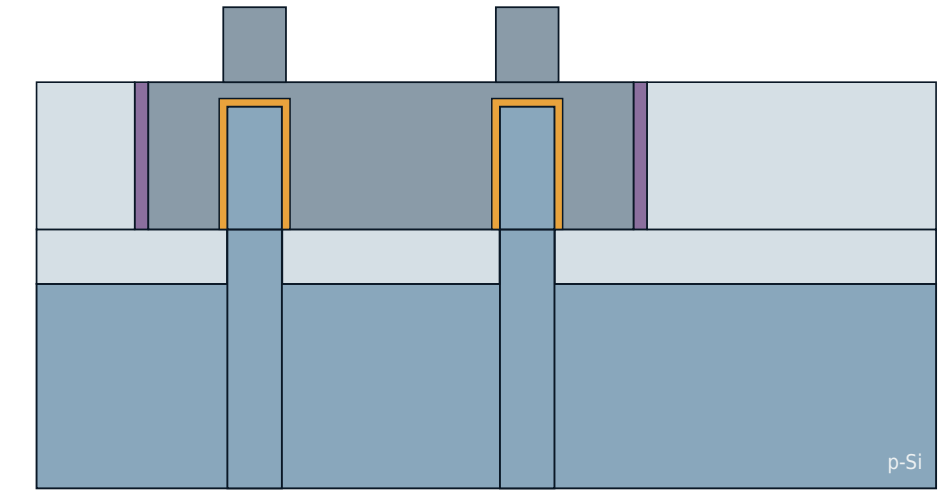


21. Kontakte

Kontaktlöcher werden bis auf Gate sowie Source und Drain geätzt und mit Metall gefüllt – der FinFET ist elektrisch anschließbar. Die Source/Drain-Kontakte liegen dabei außerhalb der hier gezeigten Schnittebene. Die Funktionsweise ist dabei analog zum normalen, planaren FET.

Kontakte

Fertige FinFET-Struktur mit Gate-Kontakten



Kontaktlöcher werden bis auf Gate sowie Source und Drain geätzt
und mit Metall gefüllt – der FinFET ist elektrisch anschließbar.
(Die Source/Drain-Kontakte liegen dabei außerhalb dieser Schnittebene.)

Da bei einem SOI-basierten Prozess bereits eine ganzflächige Oxidschicht auf dem Wafer vorhanden ist, ist die elektrische Isolation der Kanäle voneinander automatisch gegeben, ohne den in Schritt 8 beschriebenen zusätzlichen Dotierschritt. Zusätzlich ist der Ätzprozess der Finnen unproblematischer, da dieser einfach auf dem vergrabenen Oxid gestoppt werden kann, statt mit Festzeit arbeiten zu müssen.

Allgemeiner Aufbau und Funktionsweise

Der grundlegende Aufbau und die Funktionsweise eines FinFET unterscheiden sich nicht von einem herkömmlichen MOS-Feldeffekttransistor. So gibt es auch hier einen Source- und Drainanschluss, über die der Stromfluss erfolgt. Die Steuerung des Transistors regelt eine Gateelektrode. Im Gegensatz zum klassischen, in Planarbauweise hergestellten Feldeffekttransistor wird der Kanal zwischen Source und Drain jedoch als dreidimensionale Struktur auf dem Siliciumsubstrat erzeugt, so dass die Gateelektrode diesen von mehreren Seiten umschließen kann. So wird ein wesentlich besseres elektrisches Verhalten ermöglicht: Leckströme können reduziert und Steuerströme besser kontrolliert werden.

Die dreidimensionale Struktur erzeugt jedoch auch neue parasitäre Kapazitäten und minimale Abmessungen (engl. critical dimension), die optimiert werden müssen. Die Gatelänge wird in einem FinFET parallel zum Kanal gemessen, während die Weite des Gates der doppelten Finnenhöhe plus der Finnenbreite entspricht. Die Höhe des Kanals limitiert den Steuerstrom und die Gatekapazität, die Breite beeinflusst die Durchbruchspannung und Kurzkanaleffekte und die daraus resultierenden Größen wie den Stromverbrauch.

Einordnung

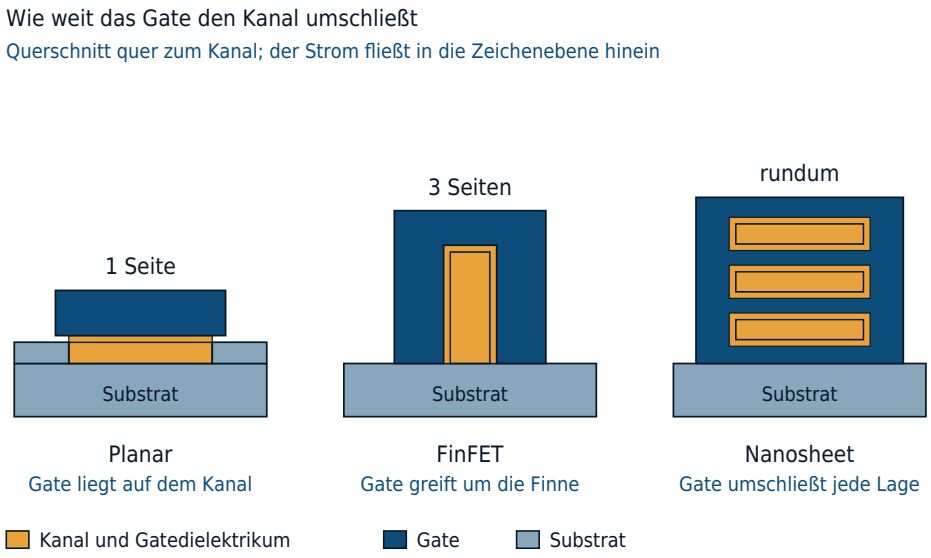
Der FinFET ist keine Zukunftstechnik mehr, sondern hat eine abgeschlossene Epoche geprägt. Nachdem er jahrzehntelang erforscht worden war, kam er 2011 erstmals in die Serienfertigung und war über ein Jahrzehnt die Standardbauform aller leistungsfähigen Prozessoren. An den kleinsten Strukturen ist er inzwischen selbst abgelöst worden: Wird die Finne immer schmaler und höher, lässt sie sich mechanisch kaum noch beherrschen, und die Weite des Gates ist nur in Stufen einstellbar – sie ergibt sich aus der Anzahl der Finnen und lässt sich nicht stufenlos wählen. Der Nachfolger löst beide Probleme, indem er die Finne umlegt (siehe Vom FinFET zum Nanosheet).

Im Folgenden wird der Aufbau eines Multigate-Transistors mit drei Gates (Tri-Gate) im Bulk-Prozess beschrieben.

Vom FinFET zum Nanosheet

Die Entwicklung vom planaren Transistor über den FinFET bis zum Nanosheet folgt einem einzigen Gedanken: Das Gate soll den Kanal von möglichst vielen Seiten umschließen. Je vollständiger es ihn umgibt, desto zuverlässiger sperrt der Transistor – und desto kürzer darf der Kanal werden, ohne dass der Strom unkontrolliert hindurchfließt.

Der Gateumschluss im Vergleich



Warum die Finne an ihre Grenze kam

Der FinFET löste das Problem des planaren Transistors, brachte aber zwei eigene mit.

Zum einen wird eine Finne, die immer schmaler und zugleich höher werden muss, mechanisch instabil. Sie steht frei auf dem Substrat und muss mehrere Prozessschritte überstehen, ohne umzukippen oder zu brechen.

Zum anderen lässt sich die Weite des Gates nur noch in Sprüngen einstellen. Beim planaren Transistor war sie eine frei wählbare Größe im Entwurf; beim FinFET ergibt sie sich aus der Anzahl der Finnen. Ein Transistor kann also zwei oder drei Finnen haben, aber nichts dazwischen. Wer einen etwas stärkeren Treiber braucht, bekommt gleich einen deutlich stärkeren – mit dem entsprechenden Flächen- und Stromverbrauch.

Die Finne umlegen

Der Nachfolger dreht die Finne um 90 Grad: Statt eines senkrechten Stegs liegen mehrere waagerechte Siliciumlagen übereinander, jede vollständig vom Gate umschlossen. Weil das Gate den Kanal nun rundum umgibt, spricht man vom Gate-all-around-Transistor; nach der Form der Lagen heißt er auch Nanosheet-Transistor.

Beide Nachteile der Finne entfallen damit. Die Lagen werden vom Gate gehalten und stehen nicht frei. Und die Gateweite ist wieder stufenlos einstellbar, denn sie ergibt sich aus der Breite der Lagen – die man im Entwurf frei wählen kann, statt Finnen abzuzählen.

Wie die Lagen entstehen

Der entscheidende Kunstgriff liegt in der Herstellung. Auf das Substrat wird ein Stapel aus abwechselnd Silicium und Silicium-Germanium epitaktisch aufgewachsen. Aus diesem Stapel wird zunächst eine Finne geätzt, ganz wie beim FinFET. Anschließend wird das Silicium-Germanium selektiv entfernt – es löst sich in einer Chemie auf, die das Silicium unangetastet lässt. Zurück bleiben die freischwebenden Siliciumlagen, deren Zwischenräume danach mit Dielektrikum und Gatemetall gefüllt werden.

Die Dicke der Lagen und ihr Abstand sind damit nicht durch Lithografie festgelegt, sondern durch die Schichtdicken beim Aufwachsen. Der Abstand zwischen zwei Lagen beträgt nur wenige Nanometer, und in diesen Spalt müssen Dielektrikum und Metall vollständig hineingelangen – eine Aufgabe, die ohne Atomlagenabscheidung nicht zu lösen ist.

Was danach kommt

Der nächste Schritt zeichnet sich bereits ab und führt den Gedanken konsequent weiter: Statt n- und p-Kanal-Transistoren nebeneinander zu setzen, sollen sie übereinander gestapelt werden. Der so entstehende komplementäre FET spart Fläche, weil beide Transistoren dieselbe Grundfläche belegen. Der Aufwand dafür ist erheblich, denn zwei unterschiedlich dotierte Bauelemente müssen in einem gemeinsamen Ablauf entstehen.

DRAM-Fertigung: der Trench-Kondensator-Prozess

Allgemeiner Aufbau und Funktionsweise

Die DRAM-Speicherzelle: 1 Transistor, 1 Kondensator

Eine DRAM-Zelle (Dynamic Random Access Memory) speichert ein einzelnes Bit als elektrische Ladung auf einem Kondensator. Ein Zugriffstransistor verbindet diesen Kondensator bei Bedarf mit der Bitline – ist der Transistor gesperrt, bleibt die Ladung isoliert und das Bit gespeichert; wird er über die Wordline leitend geschaltet, kann die Ladung ausgelesen oder neu geschrieben werden. Diese Zweierkombination aus einem Transistor und einem Kondensator wird als 1T1C-Zelle bezeichnet und ist mit Abstand die flächeneffizienteste Möglichkeit, ein Bit zu speichern – deutlich kompakter als die sechs Transistoren einer SRAM-Zelle, dafür jedoch mit einer entscheidenden Einschränkung: Die gespeicherte Ladung fließt über Leckströme langsam ab und muss periodisch

alle paar Millisekunden aufgefrischt werden, daher der Name „dynamisch".

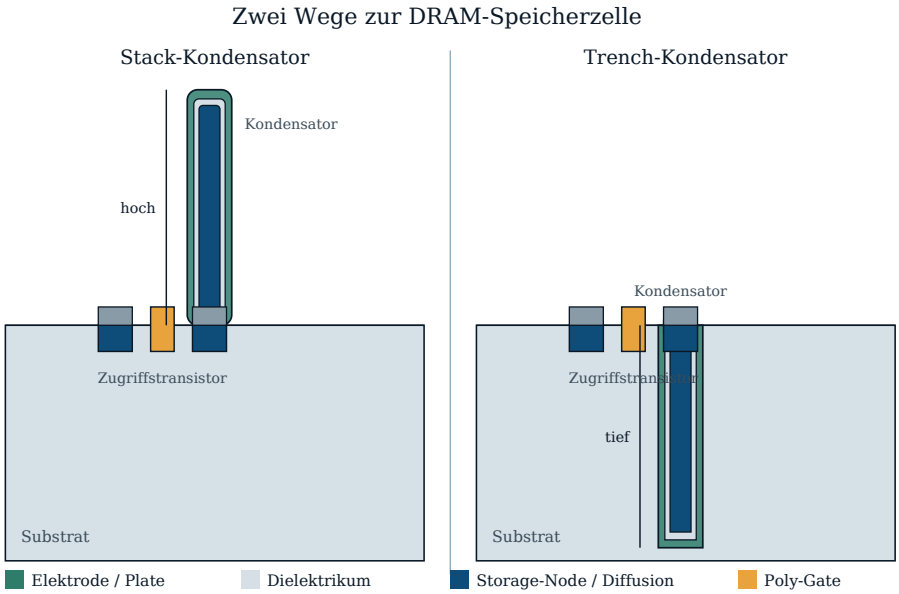
Die Herausforderung beim Zellenentwurf ist rein geometrisch: Ein Kondensator braucht Fläche, um genügend Ladung für ein zuverlässig auslesbares Signal zu speichern – typischerweise im Bereich einiger Femtofarad. Mit jedem neuen Technologieknoten schrumpft die verfügbare Zellfläche jedoch weiter, während die nötige Kapazität nahezu konstant bleibt. Die Fertigungsindustrie hat dieses Problem auf zwei grundlegend verschiedene Weisen gelöst.

Zwei Wege zum selben Ziel: Stack versus Trench

Der Stapel-Kondensator (Stack-Kondensator, dominant bei Samsung, SK Hynix und Micron) baut die benötigte Fläche vertikal oberhalb des Zugriffstransistors auf: Zylindrische oder kronenförmige Elektrodenstrukturen türmen sich über der Waferoberfläche auf und können bei Bedarf – innerhalb der Grenzen der Lithografie und Ätztechnik – nahezu beliebig hoch werden.

Der Trench-Kondensator (Graben-Kondensator, historisch von IBM entwickelt und u. a. bei Infineon/Qimonda eingesetzt) verlagert dieselbe Fläche stattdessen nach unten: Ein tiefer, schmaler Graben wird mehrere Mikrometer in das Substrat geätzt, und die Kondensatorelektroden liegen als konzentrische Zylinderflächen an dessen Wänden. Der große Vorteil dieses Ansatzes liegt in der Prozessreihenfolge: Der Graben – der aufwendigste und kritischste Teil der Fertigung – entsteht bereits vor dem Zugriffstransistor, wodurch der Transistor selbst am Ende einer nahezu ebenen Waferoberfläche gefertigt wird, ohne die hohen Stufen eines fertigen Stapel-Kondensators überwinden zu müssen.

Das folgende Diagramm stellt beide Bauformen im Querschnitt gegenüber, bevor die folgenden Abschnitte den vollständigen Trench-Prozess Schritt für Schritt aufbauen.



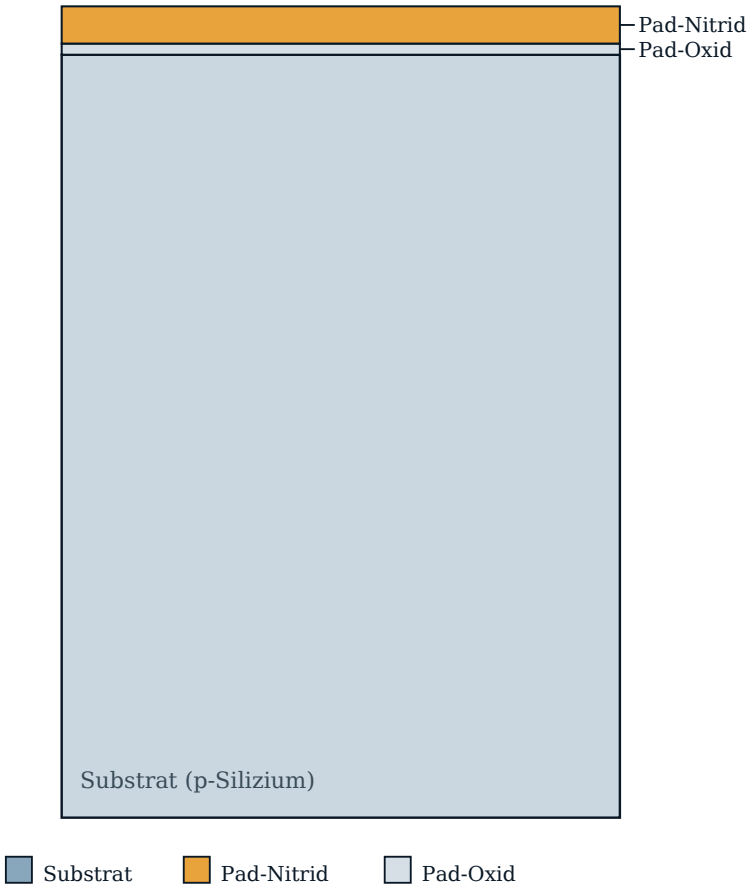
Herstellung des Trench-Kondensators

Vom flachen Wafer zum tiefen Graben

Der Trench-Kondensator-Prozess beginnt, bevor auch nur ein Transistor existiert – der Graben wird als allererste Struktur in den noch völlig unbearbeiteten Wafer geätzt. Diese Reihenfolge ist kein Zufall: Der aufwendigste, kritischste Teil der gesamten Zellfertigung soll auf einer perfekt ebenen Oberfläche stattfinden, ohne durch bereits vorhandene Transistor- oder Verdrahtungsstrukturen behindert zu werden.

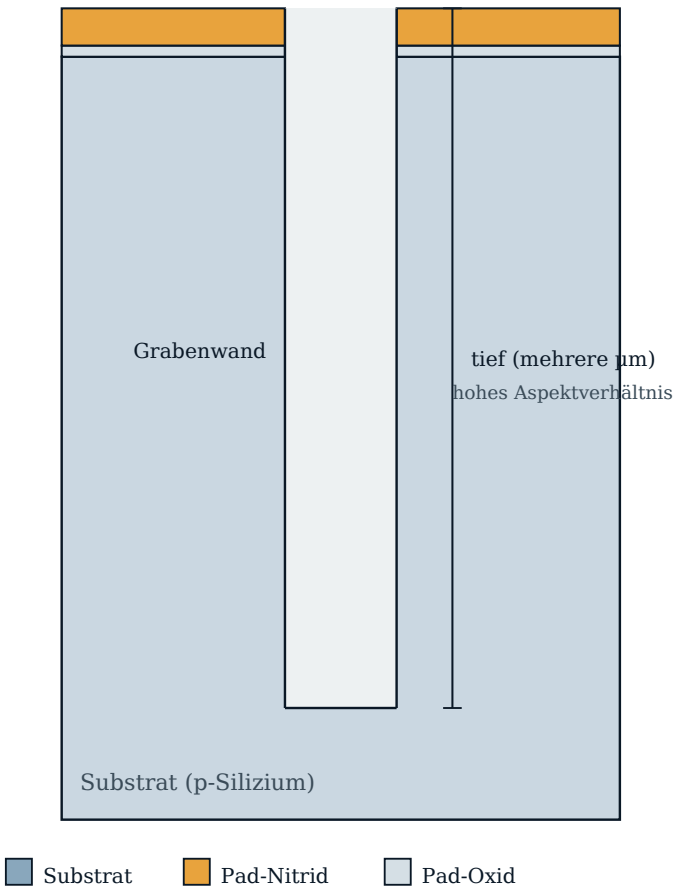
Schritt 1 – Pad-Oxid und Pad-Nitrid: Auf das p-dotierte Silizium-Substrat wird zunächst eine dünne thermische Oxidschicht (Pad-Oxid) aufgewachsen, die als Pufferschicht mechanische Spannungen zwischen Substrat und der folgenden Nitridschicht abbaut. Direkt darüber wird eine deutlich dickere Siliziumnitrid-Schicht (Pad-Nitrid) abgeschieden. Dieser Schichtstapel dient im nächsten Schritt als Hartmaske für die Grabenätzung und muss dafür so strukturiert werden, dass er nur an der vorgesehenen Trench-Position eine Öffnung freigibt.

Schritt 1: Pad-Oxid und Pad-Nitrid



Schritt 2 – Deep-Trench-Ätzung: Durch die geöffnete Hartmaske ätzt ein hochanisotroper RIE-Prozess einen schmalen, mehrere Mikrometer tiefen Graben in das Substrat. Das Aspektverhältnis (Verhältnis von Tiefe zu Breite) liegt bei modernen Trench-Zellen typischerweise bei 50:1 oder höher – eine der anspruchsvollsten Ätzaufgaben der gesamten Halbleiterfertigung, da die Seitenwände über die gesamte Tiefe gleichmäßig senkrecht und glatt bleiben müssen.

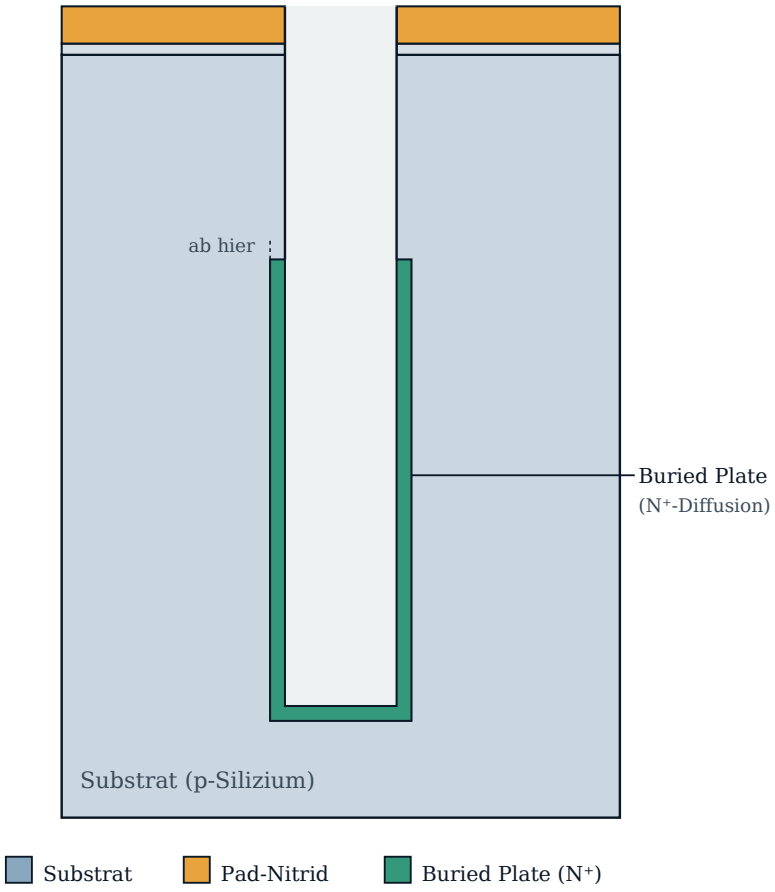
Schritt 2: Deep-Trench-Ätzung



Die Kondensatorelektroden entstehen

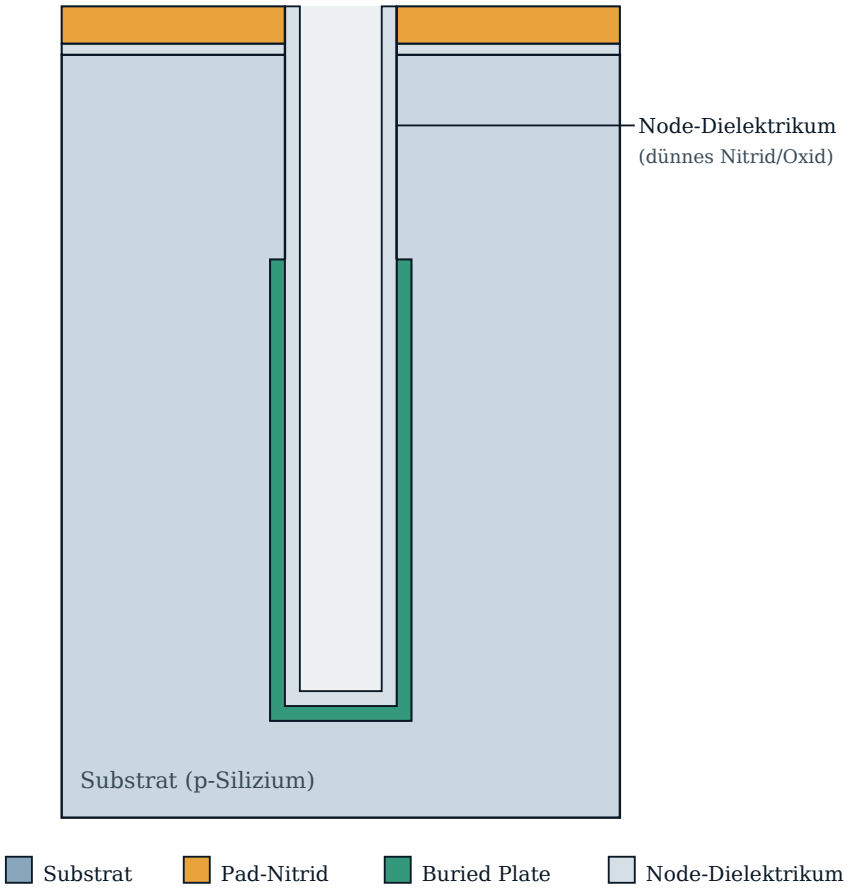
Schritt 3 – Buried Plate: Die untere, tiefere Region der Grabenwand wird mit einer Arsen-dotierten Glasschicht (Arsenglas) beschichtet, aus der der Dotierstoff bei erhöhter Temperatur in die angrenzende Siliziumwand eintreibt. So entsteht die Buried Plate – eine hoch n-dotierte Zone, die als äußere, mit allen Zellen eines Arrays gemeinsam verbundene Kondensatorelektrode dient (vergleichbar mit der „Ground Plate" eines klassischen Kondensators).

Schritt 3: Buried Plate



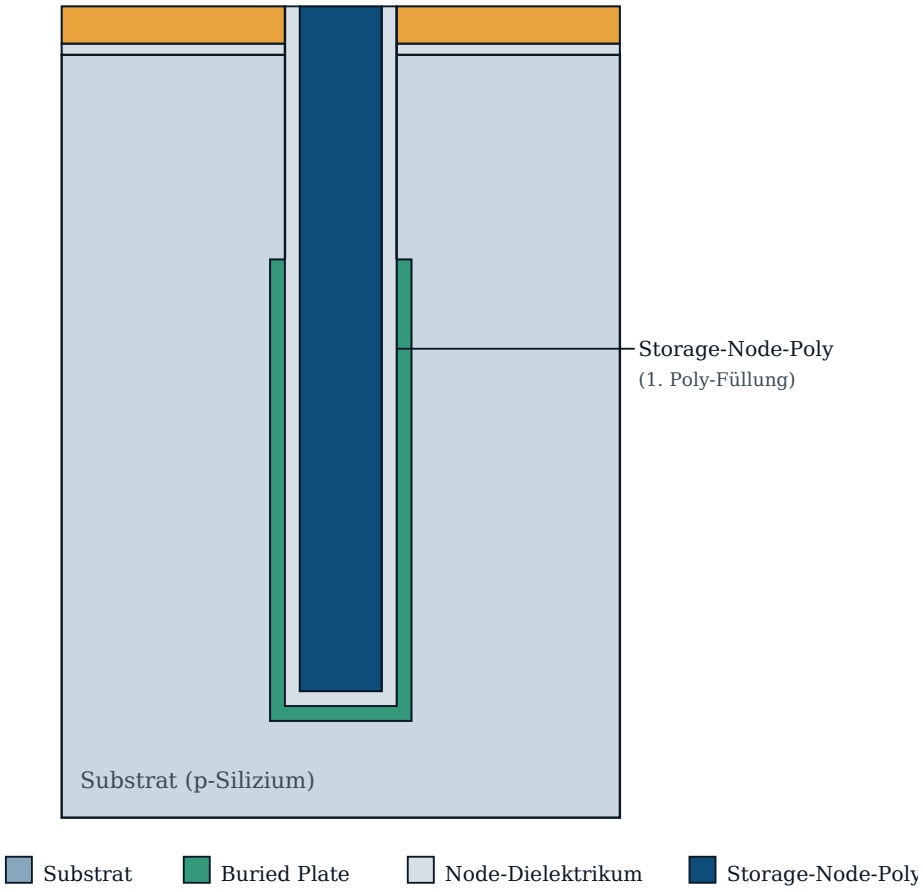
Schritt 4 – Node-Dielektrikum: Über die gesamte Grabentiefe wird nun eine wenige Nanometer dünne, konforme Dielektrikumsschicht abgeschieden – meist ein Nitrid-Oxid-Schichtstapel. Sie trennt die im nächsten Schritt folgende innere Elektrode von der Buried Plate und bestimmt zusammen mit der Grabenwandfläche maßgeblich die erreichbare Kapazität der Zelle.

Schritt 4: Node-Dielektrikum



Schritt 5 – Erste Poly-Füllung: Dotiertes Polysilizium füllt den verbleibenden Graben vollständig auf. Dieses „Storage-Node-Poly" bildet zusammen mit dem Node-Dielektrikum und der Buried Plate den eigentlichen Kondensator: Die auf dem Poly gespeicherte Ladung repräsentiert den Zustand des gespeicherten Bits.

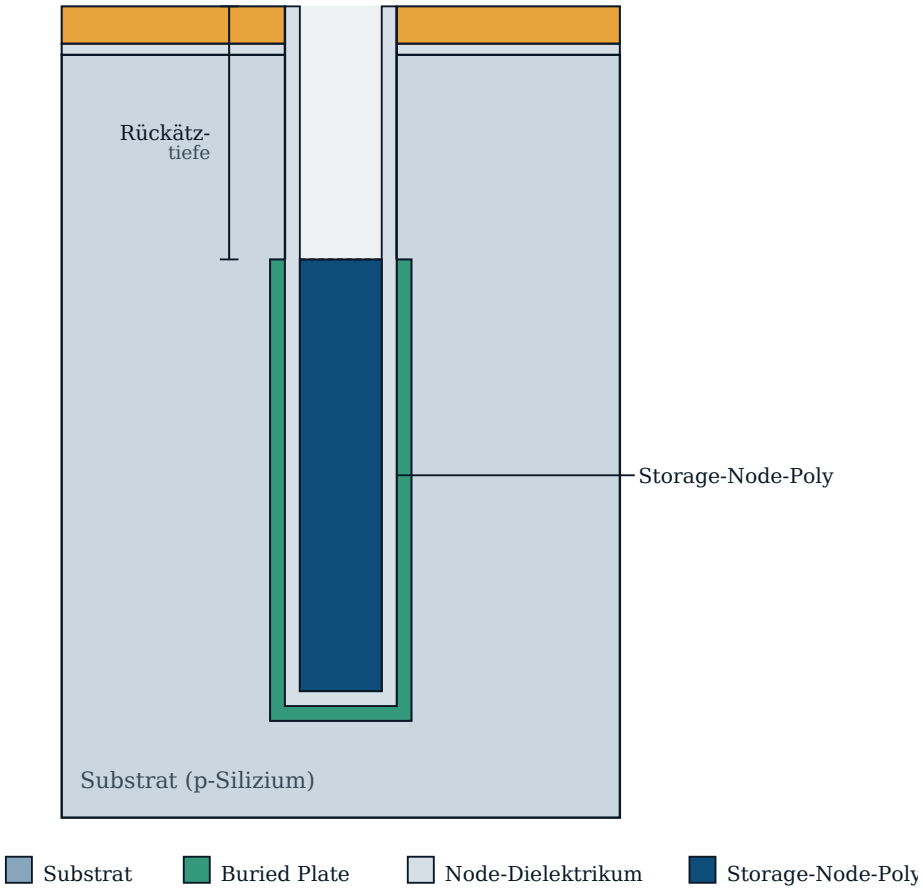
Schritt 5: Erste Poly-Füllung



Der obere Grabenbereich: Collar und Buried Strap

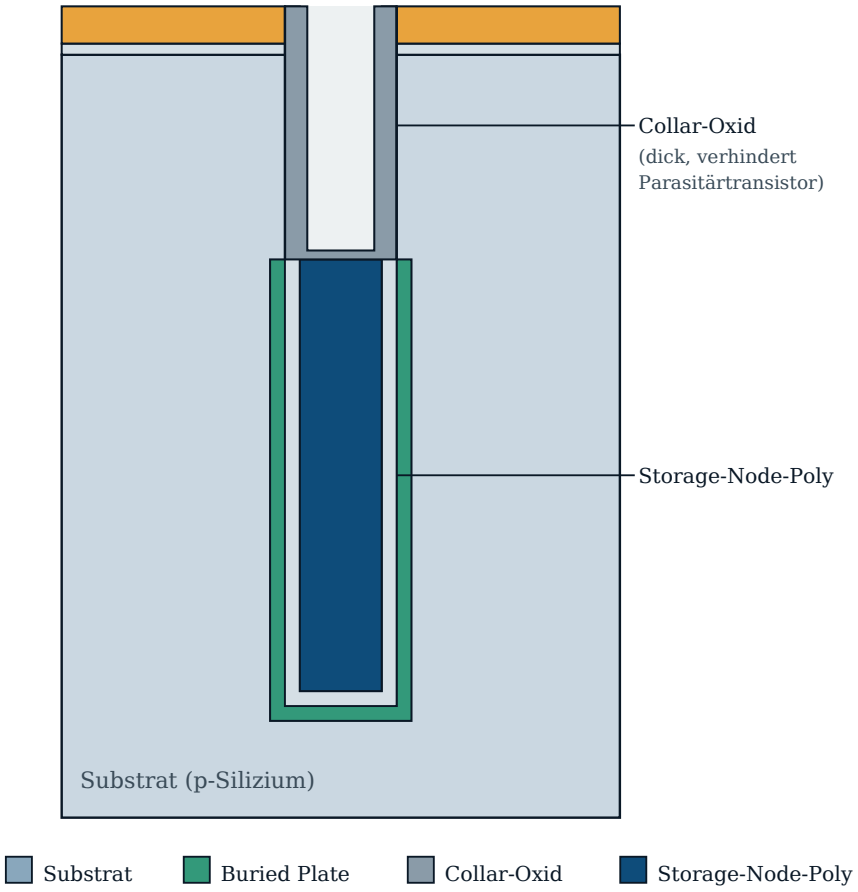
Schritt 6 – Poly-Rückätzung: Das Storage-Node-Poly wird gezielt zurückgeätzt, sodass es nur noch den unteren, von der Buried Plate umgebenen Grabenabschnitt ausfüllt. Der obere Grabenbereich – dort, wo später der Zugriffstransistor direkt neben dem Graben sitzen wird – bleibt zunächst wieder offen.

Schritt 6: Poly-Rückätzung



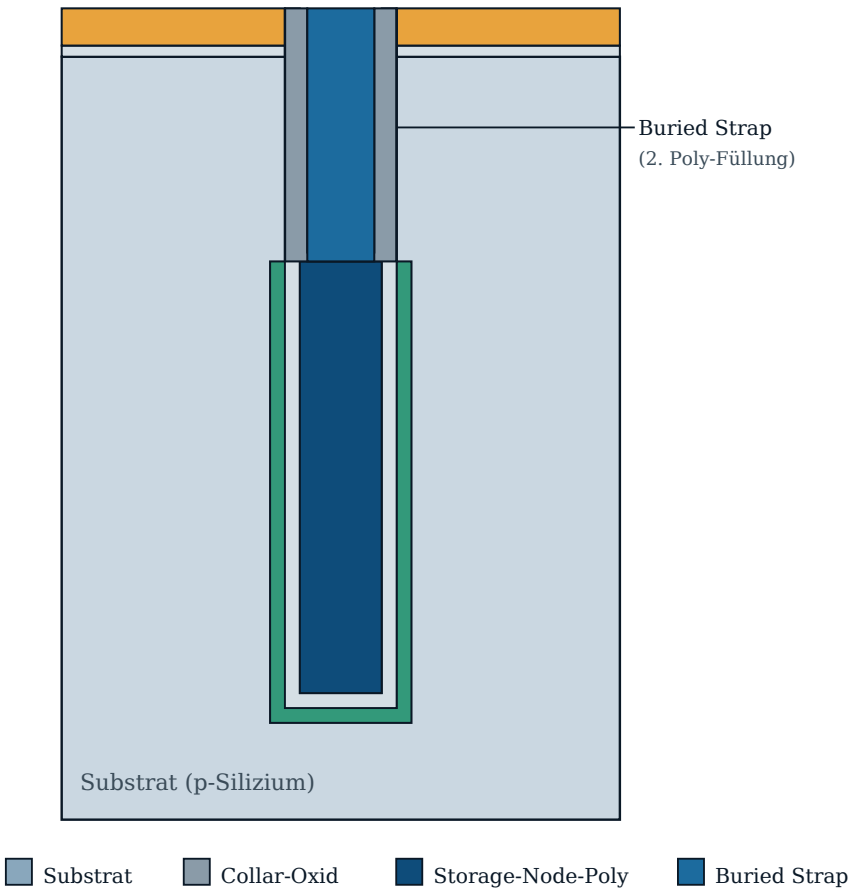
Schritt 7 – Collar-Oxid: Im freigelegten oberen Grabenabschnitt wird das dünne Node-Dielektrikum entfernt und durch eine deutlich dickere Oxidschicht ersetzt, den Collar. Ohne dieses dicke Oxid würde die spätere Wordline-Spannung entlang der oberen Grabenwand einen parasitären, ungewollten Transistor aufsteuern können, der Ladung am eigentlichen Zugriffstransistor vorbei abfließen ließe. Der Collar unterdrückt diesen Leckpfad zuverlässig.

Schritt 7: Collar-Oxid



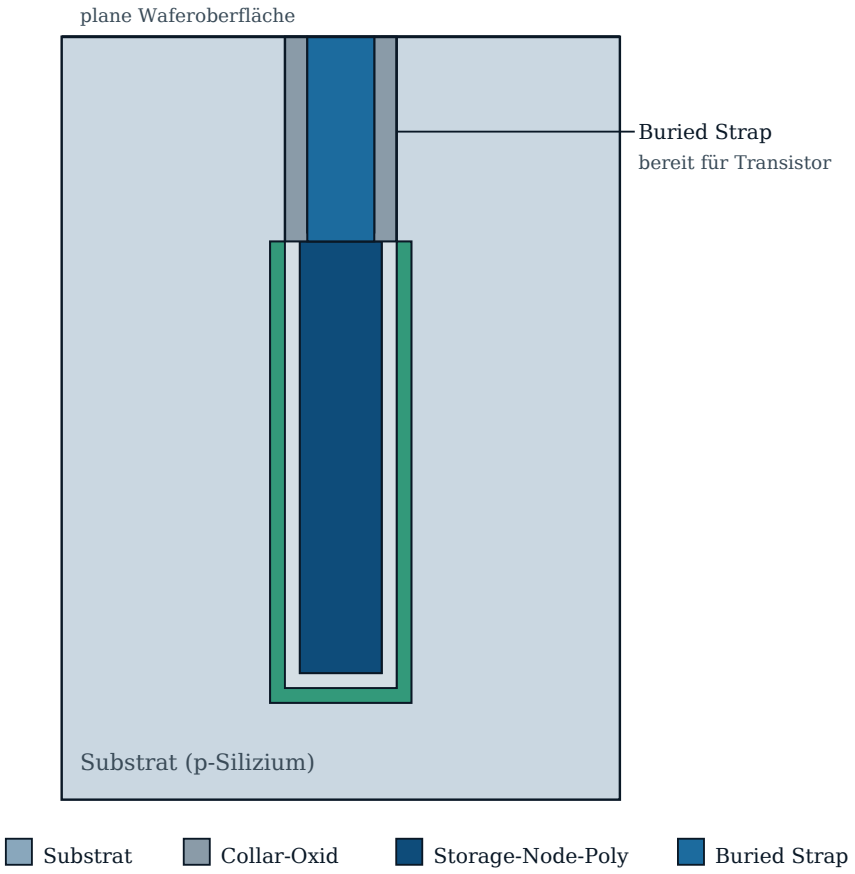
Schritt 8 – Buried Strap: Eine zweite Polysilizium-Abscheidung füllt den durch den Collar verengten oberen Grabenabschnitt vollständig auf. Dieser sogenannte Buried Strap stellt die elektrische Verbindung zwischen dem tief liegenden Storage-Node-Poly und der späteren Source/Drain-Diffusion des Zugriffstransistors an der Waferoberfläche her – ohne ihn bliebe der Kondensator vom Transistor elektrisch isoliert.

Schritt 8: Buried Strap



Schritt 9 – Planarisierung: Ein abschließender CMP- und Ätzschritt trägt überschüssiges Poly sowie die komplette Pad-Nitrid/Pad-Oxid-Hartmaske ab und stellt eine wieder nahezu plane Waferoberfläche her. Der Trench-Kondensator ist damit vollständig fertiggestellt und wartet, bündig mit der Substratoberfläche abschließend, auf den Zugriffstransistor – dessen Herstellung der nächste Abschnitt beschreibt.

Schritt 9: Planarisierung



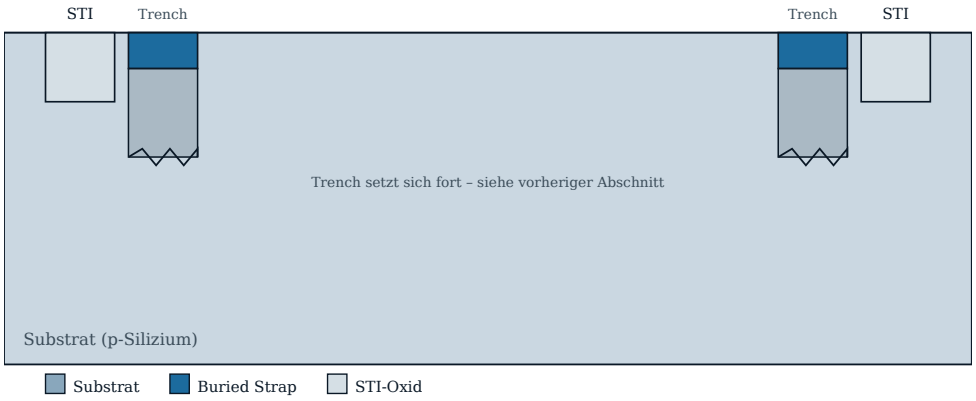
Herstellung des Zugriffstransistors und der Verdrahtung

Zwei Zellen, ein gemeinsamer Bitline-Kontakt

Mit dem fertigen, planarisierten Trench-Kondensator aus dem vorigen Abschnitt kann die Fertigung dort weitermachen, wo ein gewöhnlicher Logikprozess beginnt: an einer nahezu ebenen Waferoberfläche. Die folgenden Diagramme zeigen ein typisches Zellenpaar – zwei benachbarte Trench-Kondensatoren, deren Zugriffstransistoren sich einen gemeinsamen Bitline-Kontakt in der Mitte teilen. Diese „gefaltete" Anordnung spart einen Kontakt pro zwei Zellen und ist in der Trench-DRAM-Fertigung Standard.

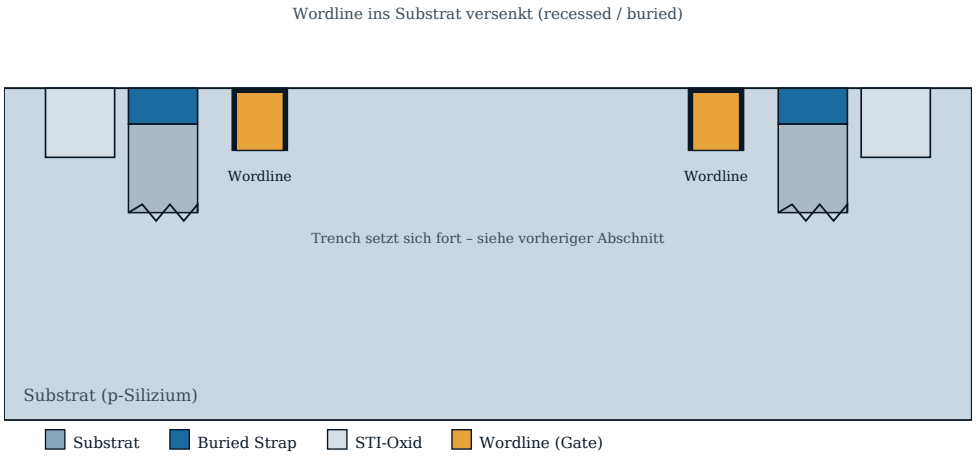
Schritt 10 – Flachgraben-Isolation (STI): Am äußeren Rand jedes Zellenpaars trennt eine flache Grabenisolation (Shallow Trench Isolation) die aktive Fläche dieses Paares von der des nächsten. Anders als der tiefe Kondensator-Graben ist die STI nur wenige hundert Nanometer tief und rein oxidgefüllt – sie hat keine elektrische Funktion außer der Trennung benachbarter Bauelemente.

Schritt 10: Flachgraben-Isolation (STI)



Schritt 11 – Gate-Oxid und Buried Wordline: Zwischen jedem Trench und dem zukünftigen Bitline-Kontakt entsteht der Zugriffstransistor. Moderne Trench-Zellen versenken die Wordline als sogenannte Buried Wordline in eine flache Grube im Substrat, ausgekleidet mit dünnem Gate-Oxid. Der versenkte Aufbau verkürzt den effektiven Kanal nicht und erlaubt gleichzeitig eine deutlich kompaktere Zelle als eine klassische, auf der Oberfläche liegende Gate-Elektrode.

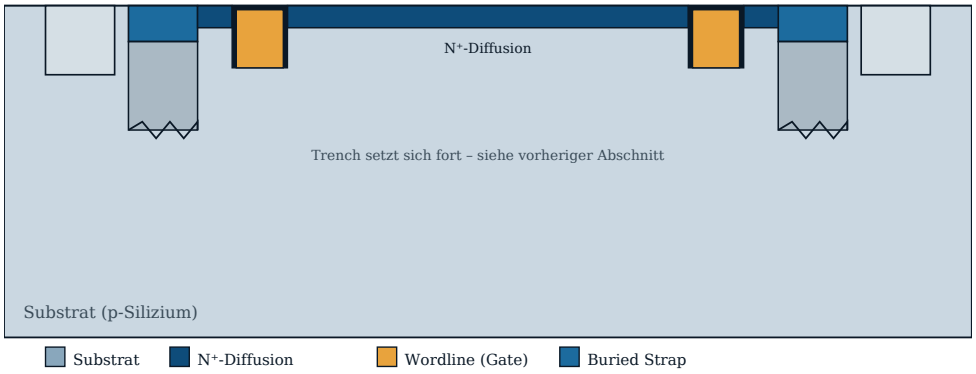
Schritt 11: Gate-Oxid und Buried Wordline



Schritt 12 – Source/Drain-Implantation: Eine flache N^+-Implantation verbindet die aktive Fläche durchgehend: vom Buried Strap des linken Trenches über den Kanal der linken Wordline zur Mitte, weiter über den Kanal der rechten Wordline bis zum Buried Strap des rechten Trenches. Jede Wordline schaltet

damit den Strompfad zwischen „ihrem" Trench-Kondensator und dem gemeinsamen Mittelknoten.

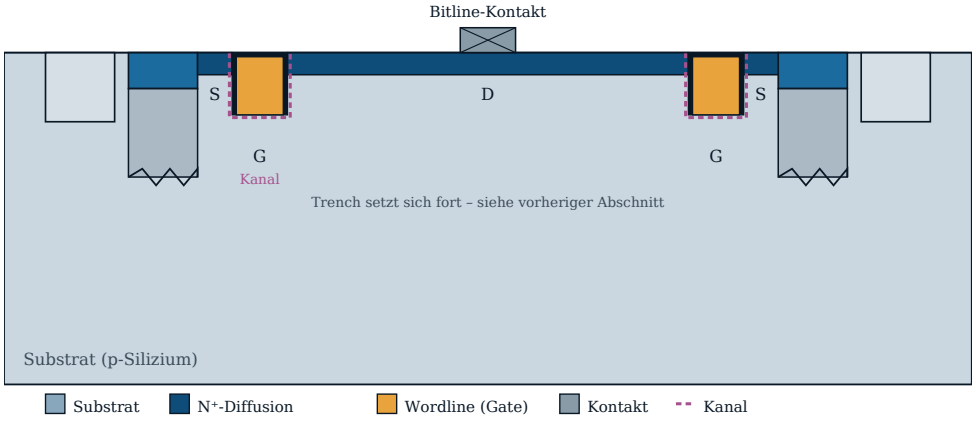
Schritt 12: Source/Drain-Implantation



Der Anschluss an die Verdrahtungsebenen

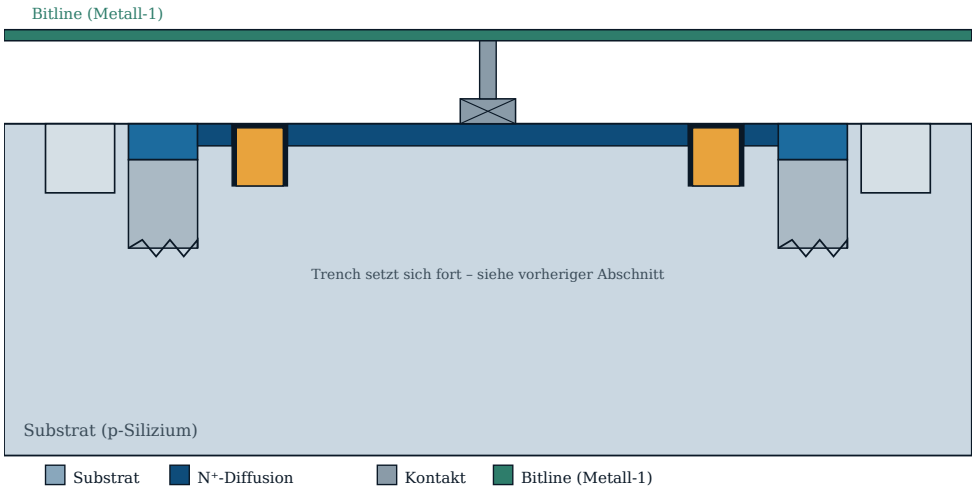
Schritt 13 – Bitline-Kontakt: Auf der gemeinsamen Diffusion in der Zellmitte wird ein Kontaktloch geöffnet und mit Metall gefüllt. Über diesen einen Kontakt greifen beide Zugriffstransistoren des Paares gemeinsam auf die Bitline zu – daher „gefaltete" Zellanordnung. Das Diagramm zeigt an dieser Stelle beispielhaft alle drei Transistoranschlüsse (G, S, D) sowie den Kanalverlauf: Bei einer versenkten Wordline bildet sich der Kanal nicht flach unter dem Gate, sondern U-förmig entlang der Grabenwände und des Grabenbodens – dadurch wird der Strompfad trotz kompakter Zellfläche effektiv länger. Source und Drain sind hier der Übersichtlichkeit halber fest zugeordnet; elektrisch sind beide Diffusionsseiten symmetrisch und tauschen ihre Rolle je nach Betriebsrichtung.

Schritt 13: Bitline-Kontakt



Schritt 14 – Bitline-Metallisierung: Die erste Metalllage (Metall-1) verdrahtet den Bitline-Kontakt zu einer durchgehenden Leiterbahn, die sich über das gesamte Zellenfeld erstreckt und tausende Zellenpaare in derselben Spalte verbindet. Beim Auslesen einer Zelle vergleicht ein Leseverstärker die Spannung dieser Bitline mit einer Referenzspannung, um zu bestimmen, ob eine „1" oder „0" gespeichert war.

Schritt 14: Bitline-Metallisierung



Schritt 15 – Zwischenoxid und Kontakte: Ein Zwischenoxid (ILD) bettet die Bitline-Verdrahtung ein und stellt gleichzeitig die Isolationsschicht für zusätzliche Kontakte bereit – etwa für die Wordline-Anschlüsse außerhalb des eigentlichen Zellenfelds, wo die Wordlines in Treiberschaltungen münden. Ab hier folgt die Fertigung denselben Prinzipien wie jeder andere Logikprozess:

weitere Metalllagen, Vias und Kontakte, wie sie bereits im Kapitel „Schaltungslayouts" beschrieben sind. Damit ist die DRAM-Zelle vollständig – vom tiefen Kondensator-Graben bis zur Anbindung an die erste Verdrahtungsebene.

Schritt 15: Zwischenoxid und Kontakte