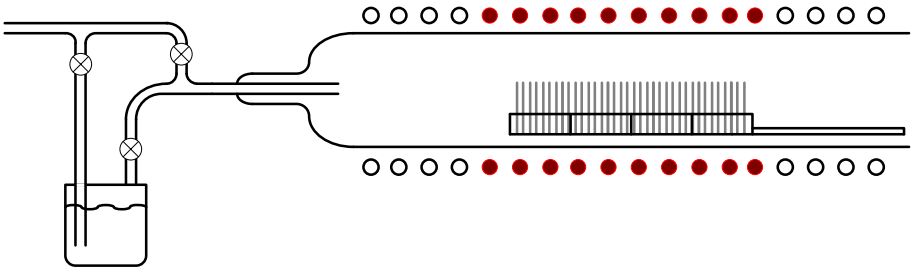
Halbleitertechnologie

von A bis Z



Oxidation

www.halbleiter.org

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Oxidation

1.1 Industrielle Verwendung

1.1.1 Anwendung und Eigenschaften

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)

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 Ätzen. Der Ablöseprozess imt HF läuft nach folgender Reaktion ab:

$$\mathrm{SiO_2} + 6\,\mathrm{HF} \longrightarrow \mathrm{H_2SiF_6} + 2\,\mathrm{H_2O}$$

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

1.2 Erzeugung von Oxidschichten

1.2.1 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 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. $1m$ auf $\pm$ 0,5 °C exakt regelbar.

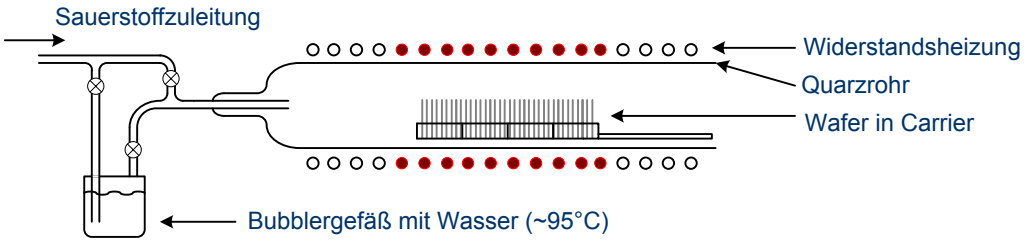


Abb. 1.1: Darstellung eines Oxidationsofens

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 stattfinden). 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 Oxid zu Siliciumdioxid:

$$\mathrm{Si} + \mathrm{O_2} \longrightarrow \mathrm{SiO_2}$$

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:

$$\mathrm{Si} + 2\,\mathrm{H_2O} \longrightarrow \mathrm{SiO_2} + 2\,\mathrm{H_2}$$

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.

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

Tab. 1.1: Vergleich der Aufwachsraten von trockener und nasser Oxidation

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 der Sauerstoff zunächst diffundieren muss, 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 von der Diffusionsgeschwindigkeit des Oxids durch das Siliciumdioxid bestimmt. Mit zunehmender Oxiddicke verlangsamt sich also das Wachstum. 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. Das kann beispielsweise mit höherer Oxidationstemperatur erreicht werden, oder durch Verwendung der nassen Oxidation, die ebenfalls nur eine sehr geringe Ladung verursacht.

Segregation:
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.

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:

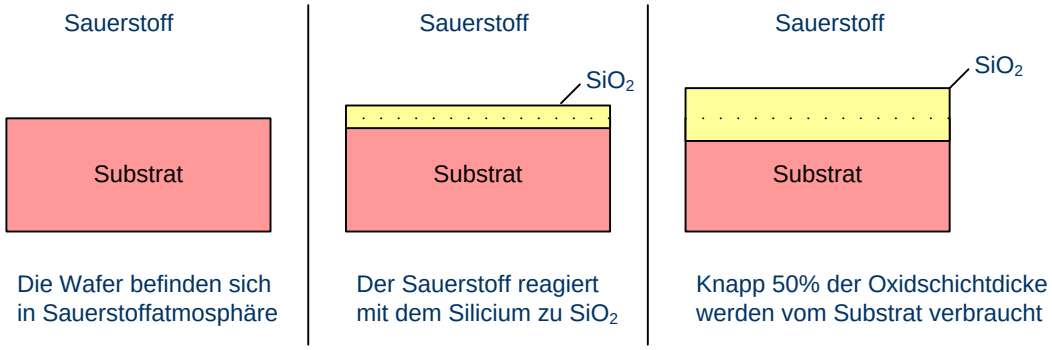


Abb. 1.2: Aufwachsverhalten von Oxid auf Silicium

$$\mathrm{k} = \frac{\text{Löslichkeit des Dotierstoffs in Silicium}}{\text{Löslichkeit des Dotierstoffs in } \mathrm{SiO_2}}$$

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

1.2.2 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 später 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 reagiert das Silan und der Sauerstoff zu Siliciumdioxid und Wasserstoff:

$$\mathrm{SiH_4 + O_2 \longrightarrow SiO_2 + 2\,H_2}$$

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.

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

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.

1.3 Die LOCOS-Technik

1.3.1 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. Dies ist beispielsweise mit der LOCOS-Technik realisierbar: LOCal Oxidation of Silicon (Lokale Oxidation von Silicium).

1.3.2 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.

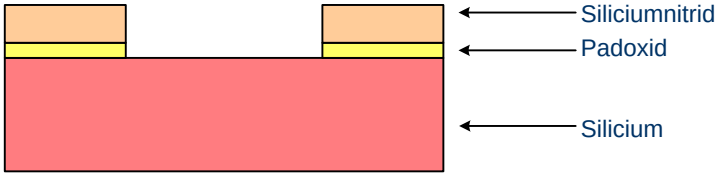


Abb. 1.3: Schichtaufbau vor dem LOCOS-Prozess

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.

Neben diesem Effekt, der bis zu $1\mu m$ 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. Durch Ätzen einer Oxidschicht wären so geringe Abmessungen für die Integration von Transistoren und anderen Bauteilen nicht erreichbar. Des Weiteren ist wegen der geringeren Unebenheit, ohne die Kanten- und Stufenbildung, die Auflösung in der Fototechnik verbessert. Das FOX lässt sich noch etwas zurückätzen, dadurch wird zwar das aufgewachsene Oxid leicht reduziert, die

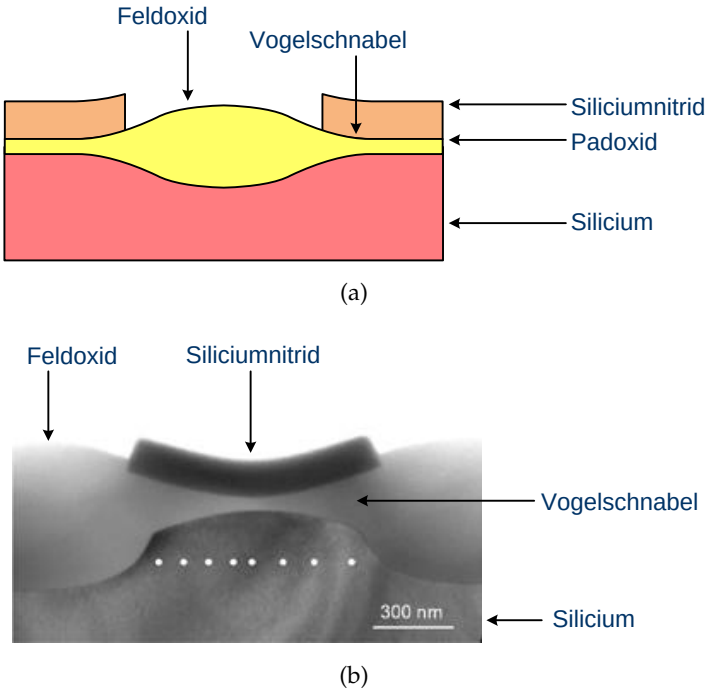


Abb. 1.4: LOCOS-Struktur nach der Oxidation

Länge des Vogelschnabels nimmt jedoch ab und die Oberfläche wird wiederum etwas mehr eingeebnet. Dies wird als fully recessed LOCOS bezeichnet.

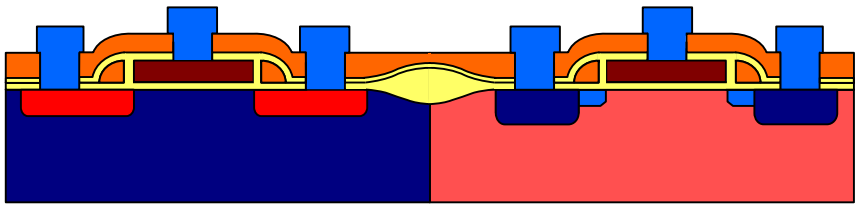


Abb. 1.5: Anwendungsbeispiel der LOCOS-Technik zur seitlichen Isolation zweier Transistoren

1.3.3 Alternative Prozesse

In modernen Prozessen werden anstelle der lokalen Oxidation Gräben in das Substrat geätzt und mit Oxid aufgefüllt, um so benachbarte Bauteile voneinander zu isolieren. Dies beansprucht wesentlich weniger Fläche, erfordert jedoch mehr Prozessschritte als das LOCOS-Verfahren. Diese Grabenisolation (engl. shallow trench isolation, STI) wird

hauptsächlich bei der Integration von Transistoren angewandt und ist einer der ersten Prozesse in der Halbleiterfertigung. Daneben gibt es auch eine tiefe Grabenisolation (engl. deep trench isolation) welche in der Analogtechnik eingesetzt wird.

In der 45-nm-Technologie sind die Gräben bei einem STI-Prozess ca. 100 nm (SOI) bzw. $300nm$ (Bulk) tief.

1.4 Schichtdickenmessung

1.4.1 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.

1.4.2 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.

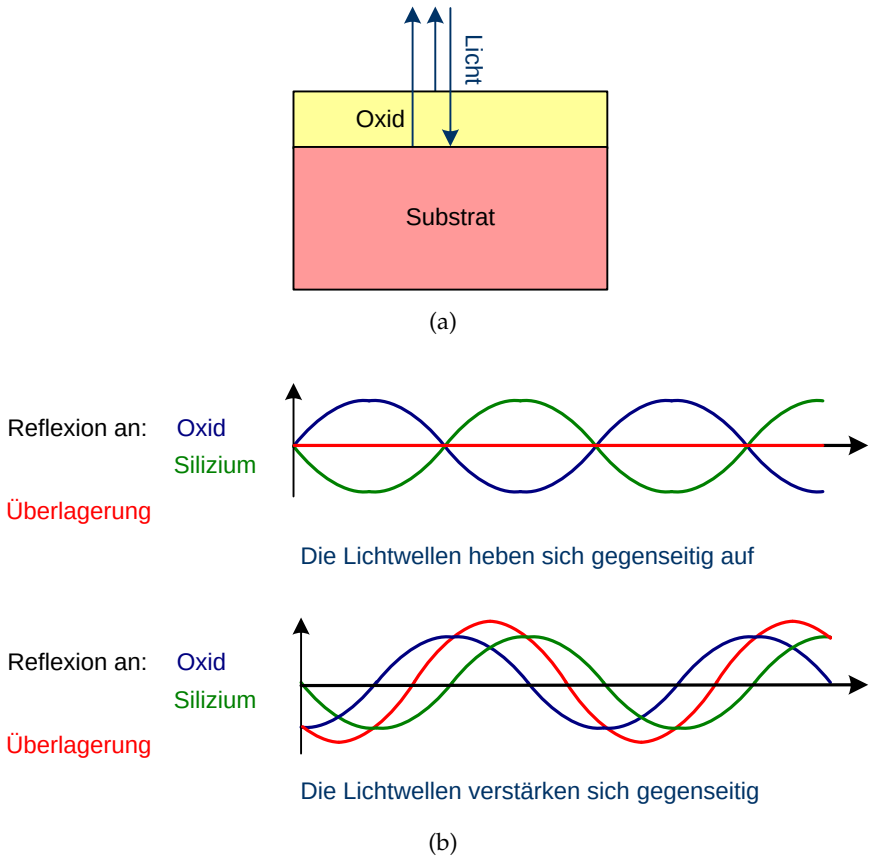


Abb. 1.6: (a) Prinzip der Interferenz, (b) destruktive und konstruktive Überlagerung von Lichtwellen

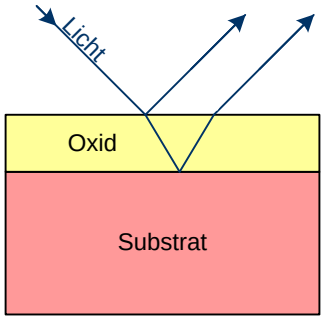
Es wird dabei ein Spektrum an unterschiedlichen Wellenlängen auf den Wafer gestrahlt. Je nach Dicke der durchstrahlten Schicht überlagern sich die reflektierten Strahlen unterschiedlich und ergeben eine für jede Schichtdicke und jedes Material charakteristische Überlagerung. Mit Hilfe eines Photometers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

1.4.3 Ellipsometrie

Ellipsometrie ist die Bestimmung von optischen Eigenschaften durch Änderung der Polarisation von Licht. Linear polarisiertes Licht (die Lichtwellen haben eine bestimm-

te Schwingung) wird dabei unter einem festen Winkel auf den Wafer gestrahlt.



Schräg eingestrahltes Licht wird gebrochen und reflektiert;
gleichzeitig ändert sich die Polarisation

Abb. 1.7: Ellipsometrie

Bei der Reflexion an 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.

1.4.4 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.

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,

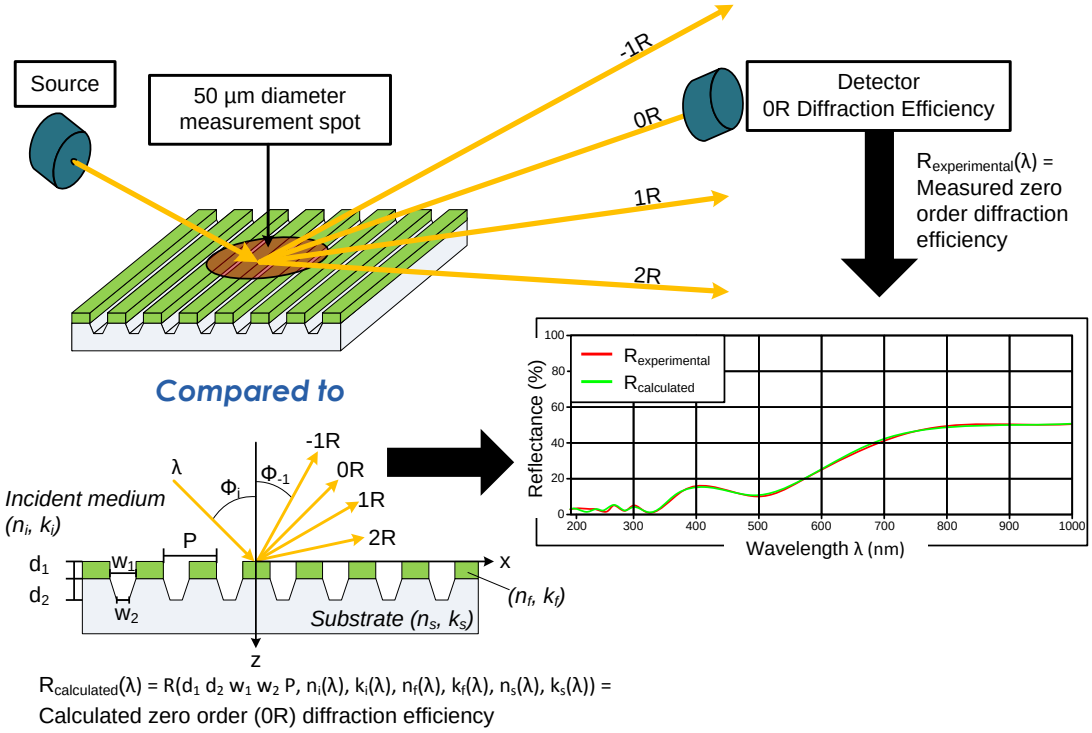


Abb. 1.8: Vergleich von Messung und Simulation

wie gut die Simulation mit der Messung übereinstimmt (0-100 %).