Halbleitertechnologie von A bis Z

Oxidation

Industrielle Verwendung 3
Anwendung von Oxidschichten 3
Eigenschaften von Oxidschichten 4
Erzeugung von Oxidschichten 4
Thermische Oxidation 5
Oxidation durch Abscheidung 9
Bauelementisolation 10
Grabenisolation 10
Höchstintegration auf Chips 11
Der Vogelschnabel 12

Industrielle Verwendung

Anwendung von Oxidschichten

Siliziumdioxid (SiO_2) gehört zu den wichtigsten und vielseitigsten Materialien in der Halbleiterfertigung. Aufgrund seiner exzellenten elektrischen Isoliereigenschaften, seiner chemischen Stabilität und der Möglichkeit, es sowohl thermisch als auch durch Abscheidung herzustellen, findet Oxid in nahezu jedem Prozessschritt der Chipherstellung Verwendung. Die folgenden Abschnitte geben einen Überblick über die wichtigsten Einsatzgebiete:

Zur Isolation

Oxidschichten dienen als elektrischer Isolator zwischen verschiedenen Leiterebenen, beispielsweise zwischen Metallisierungsschichten. Dies verhindert unerwünschte Kurzschlüsse und parasitäre Kapazitäten zwischen den Leiterbahnen.

Als Streuschicht

Bei der Ionenimplantation wird eine dünne Oxidschicht genutzt, um die einfallenden Ionen zu streuen und deren Eindringtiefe sowie -verteilung im Substrat gezielt zu beeinflussen. Dies ermöglicht eine präzisere Kontrolle des Dotierungsprofils.

Als Anpassungsschicht

Im Rahmen der LOCOS-Technik (Local Oxidation of Silicon) wird Oxid als Übergangsschicht eingesetzt, um mechanische Spannungen zwischen unterschiedlichen Materialien abzubauen und so die Bauteilzuverlässigkeit zu erhöhen.

Zur Planarisierung

Aufgrund seiner Fließeigenschaften bei hohen Temperaturen (Reflow) kann Oxid genutzt werden, um Kanten zu entschärfen und Stufen in der Topografie des Wafers einzuebnen. Dies ist besonders wichtig für nachfolgende Lithografieschritte, die eine möglichst plane Oberfläche erfordern.

Als Maskierschicht

Oxid wird häufig als Maskierschicht bei Diffusionsprozessen verwendet, da es

für viele Dotierstoffe eine effektive Barriere darstellt. Dadurch lassen sich selektive Dotierungen in genau definierten Bereichen des Wafers realisieren.

Als Justiermarken

In der Fototechnik dienen strukturierte Oxidbereiche als Justierpunkte, die eine präzise Ausrichtung nachfolgender Maskenebenen ermöglichen. Dies ist essentiell für die korrekte Überlagerung mehrerer Prozessebenen.

Als Schutzschicht

Abschließend fungiert Oxid als passivierende Schutzschicht, die den Chip vor mechanischer Beschädigung, Feuchtigkeit und Verunreinigungen während der weiteren Verarbeitung und im späteren Betrieb schützt.

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:

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

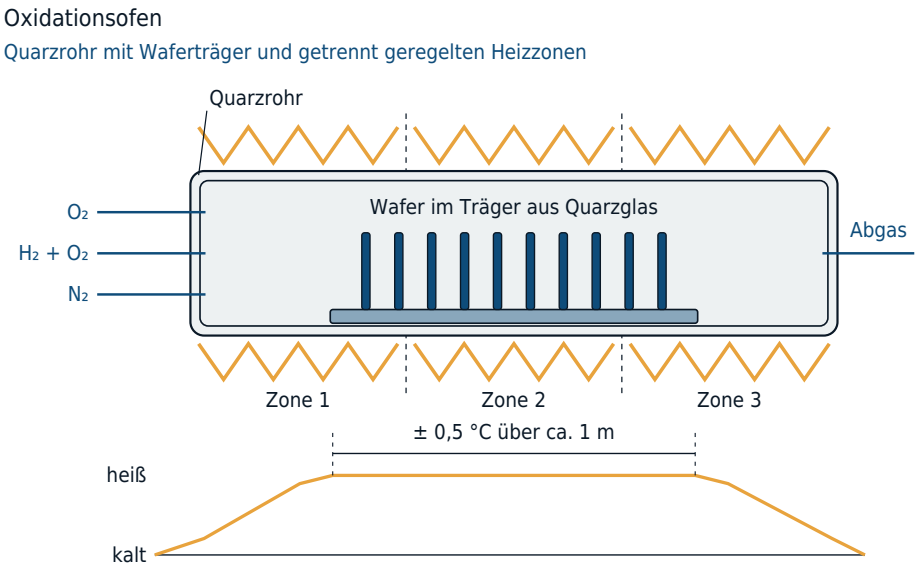
Des Weiteren bietet sich Oxid zur Funktion von Schaltungen an, da es die unterschiedlichsten Anforderungen erfüllt (z.B. als Gateoxid, Feldoxid oder Zwischenoxid).

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.



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

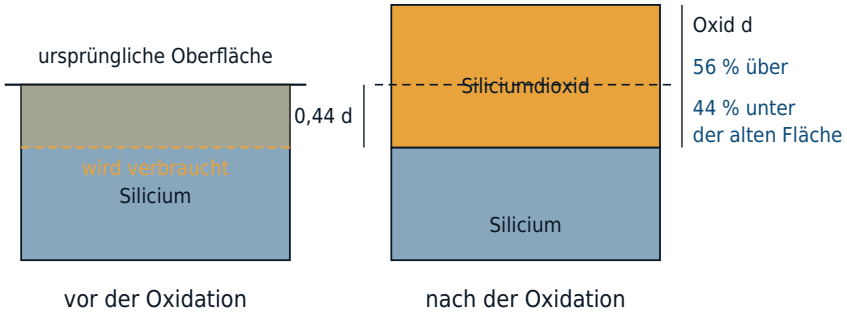
Zu Beginn reagiert der Sauerstoff an der Waferoberfläche zu Siliciumdioxid. Nun befindet sich eine Oxidschicht auf dem Substrat durch die nachfolgende Sauerstoffatome zunächst diffundieren müssen, um mit dem Silicium reagieren zu können. Die Aufwachsrate hängt nur zu Beginn von der Reaktionszeit zwischen 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.

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.

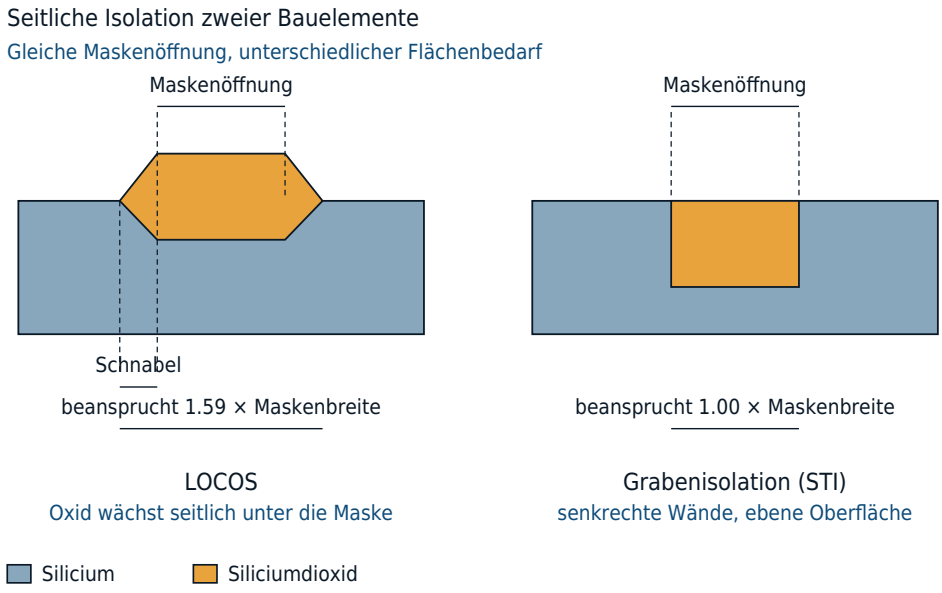
Bauelementisolation

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

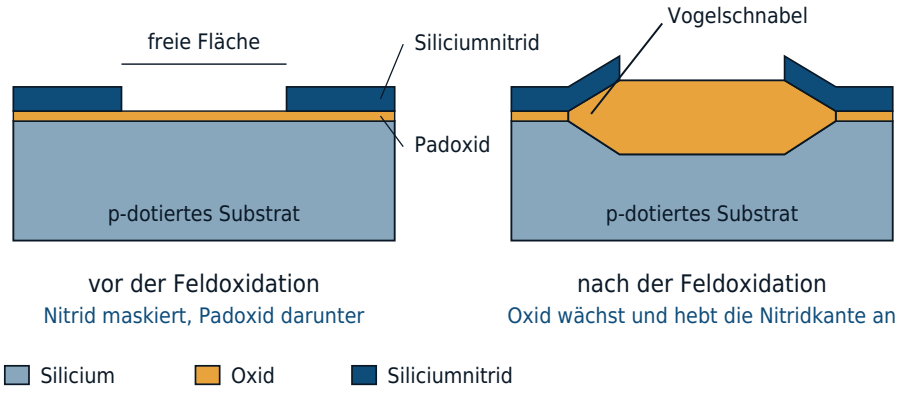
Bei der LOCOS-Technik nutzt man die unterschiedlichen Oxidationsgeschwindigkeiten von Silicium und Siliciumnitrid zur lokalen Maskierung der Scheibenoberfläche aus.

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.

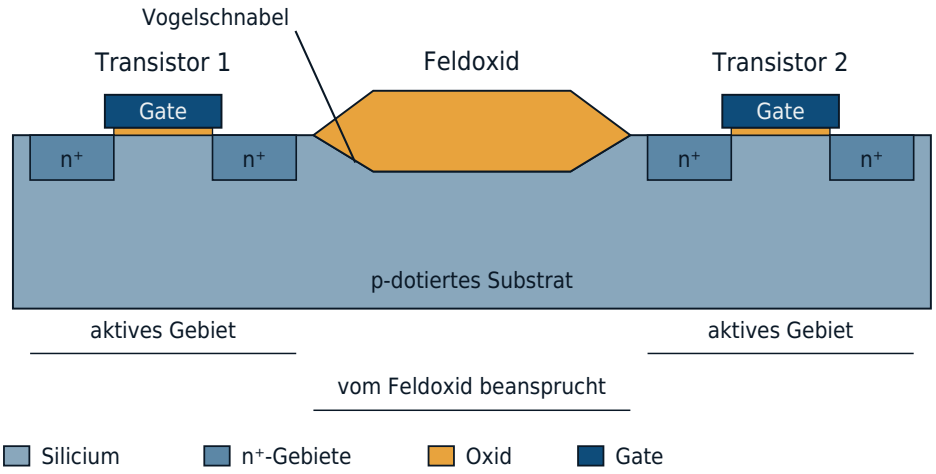
Trotz dieser negativen Effekte ist die LOCOS-Technik ein geeignetes Verfahren um die hohe Packungsdichte zu ermöglichen. Auf Grund der geringeren Unebenheit ohne Kanten- und Stufenbildung wird die Auflösung in der Fototechnik verbessert. Das Feldoxid lässt sich noch etwas zurückätzen, dadurch wird zwar 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.