UNIVERSITÄT DER BUNDESWEHR MÜNCHEN
Fakultät für Elektrotechnik und Informationstechnik

Ladungsspeicherung in Oxid-Nitrid-Oxid (ONO) Strukturen für nichtflüchtige Speicherbauelemente

Jan-Malte Schley
jm.schley@gmx.de

Vorsitzender des Promotionsausschusses:	Prof. Dr.-Ing. H. Baumgärtner
1.Berichterstatter:	Prof. Dr.-Ing. K. Hoffmann
2.Berichterstatter:	Prof. Dr. rer. nat. I. Eisele

Tag der Prüfung: 2. März 2005

Mit der Promotion erlangter akademischer Grad:
Doktor-Ingenieur
(Dr.-Ing.)

Dresden, den 18. März 2005

Vorwort und Danksagung

Die vorliegende Arbeit entstand begleitend zu meiner Tätigkeit als Entwicklungsingenieur der Firma Infineon Technologies im Bereich der Technologieentwicklung.
Ganz herzlich bedanken möchte ich mich bei Herrn Prof. Dr.-Ing. K. Hoffmann, der mir diese externe Arbeit am Institut für Elektronik der Universität der Bundeswehr München ermöglicht hat.
Sehr dankbar bin ich Herrn Prof. Hoffmann auch für seine vielen wertvollen Anregungen, seinen begeisternden Optimismus und die Freiheiten, die er mir bei der Ausgestaltung des Themas ließ.
Zu großem Dank bin ich auch meinem Vorgesetzten bei Infineon Technologies Herrn Dr. C. Ludwig verpflichtet, er hat durch seine stetige Unterstützung wesentlich zum Gelingen dieser Arbeit beigetragen.
Weiterhin möchte ich mich bei allen meinen Kollegen von Infineon Technologies bedanken, die mich bei dieser Arbeit unterstützt haben. Mein besonderer Dank gilt Dr. T. Mikolajick und Dr. T. Kern.
Mein ganz besonderer Dank für so unglaublich Vieles gebührt Inka.

Patentanmeldungen und Veröffentlichungen

veröffentlichte Patentanmeldung:

Method for efficient carrier generation in silicon waveguide systems for switching/modulating purposes using parallel pump and signal waveguides (United States Patent and Trademark Office, application No.: 20040114847)

Als Teil der Technologieentwicklung bei Infineon wurden zu folgenden Veröffentlichungen Beiträge geleistet:

Willer et al., 110nm NROM Technology for Code and Data Flash Products, IEEE VLSI 2004

Mikolajick et al., Optimisation of a Multi-Bit Charge Trapping Memory Cell using Process/Device Simulation, IEEE Non-Volatile Semiconductor Workshop 2004

Hagenbeck et al., Modeling and Simulation of Electron Injection during Programming in $TwinFlash^{TM}$ Devices Based on Energy Transport and Non-Local Lucky Electron Concept, International Workshop on Computational Electronics 2004

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Liste der verwendeten Formelzeichen und Abkürzungen

Physikalische Konstanten

Größe	Bedeutung	Zahlenwert
ϵ_0	Dielektrizitätskonstante des Vakuums	$8,8542 \cdot 10^{-12} \frac{As}{Vm}$
ϵ_{ox}	relative Dielektrizitätskonstante von Siliziumoxid	$3,9$
ϵ_{Si}	relative Dielektrizitätskonstante von Silizium	$11,9$
k	Boltzmann-Konstante	$1,38 \cdot 10^{-23} \frac{J}{K}$
h	Plancksches Wirkungsquantum	$6.626 \cdot 10^{-34} Js$
q	Elementarladung	$1,602 \cdot 10^{-19} C$

Verwendete Bezeichnungen

Größe	Bedeutung	Einheit
A	allgemeine Flächenbezeichnung	m^2
β_1	Anpassungsfaktor für U_{th}	1
β_2	Anpassungsfaktor für U_{th}	1
C'_d	flächenbezogene Kapazität der Verarmungszone	$\frac{F}{m^2}$
C'_{ox}	flächenbezogene Oxidkapazität	$\frac{F}{m^2}$

D	elektrische Flussdichte	$\frac{C}{m^2}$
d_{BD}	Weite der Verarmungszone an der Drain	m
d_{BS}	Weite der Verarmungszone an der Source	m
d_i	Dicke der Inversionsschicht	m
d_j	Tiefe der Source- bzw. Drain-Gebiete	m
d_{ox}	Dicke der Oxidschicht	m
d_1	Korrekturfaktor	1
γ	Substratsteuerfaktor	$\sqrt{V}$
E	elektrische Feldstärke	$\frac{V}{m}$
E_{eff}	effektive elektrische Feldstärke	$\frac{V}{m}$
E_k	kritische elektrische Feldstärke	$\frac{V}{m}$
ϵ_r	relative Dielektrizitätskonstante	1
ν_0	attempt-to-escape Frequenz von SiN	$\frac{1}{s}$
ν_T	thermische Emissionsrate	$\frac{1}{s}$
ν_{PF}	Poole-Frenkel Emissionsrate	$\frac{1}{s}$
I_{DB}	Drain-Bulk-Strom	A
k_A	Propotionalitätsfaktor für die Early-Spannung	$V\sqrt{cm}$
L	Länge	m
L_{Dp}	Debye-Länge im p-Gebiet	m
L_{eff}	effektive Kanallänge	m
l_{km}	Strecke zwischen pinch-off und Drain bei Kanallängenmodulation	m
l_S	typische Länge	m
λ_S	subthreshold swing Anpassungsfaktor	1
μ	Ladungsträgerbeweglichkeit	$\frac{cm^2}{Vs}$
μ_n	Elektronenbeweglichkeit	$\frac{cm^2}{Vs}$
N_A	Akzeptorendichte	$\frac{1}{cm^3}$
N_C	äquivalente Zustandsdichte des Leitungsbands	$\frac{1}{cm^3}$

N_V	äquivalente Zustandsdichte des Valenzbands	$\frac{1}{cm^3}$
n_i	Intrinsicdichte	$\frac{1}{cm^3}$
$n(x)$	Dichte der negativen, beweglichen Ladungsträger	$\frac{1}{cm^3}$
$p(x)$	Dichte der positiven, beweglichen Ladungsträger	$\frac{1}{cm^3}$
ϕ	Potential	V
ϕ_F	Fermispannung	V
ϕ_i	Potential zwischen verschieden dotierten Gebieten	V
ϕ_{ox}	Potentialabfall über SiO_2	V
ϕ_K	Kontaktspannung	V
ϕ_r	Energie einer Fangstelle	eV
ϕ_S	Oberflächenpotential	V
ϕ_t	Temperaturspannung	V
Q	Ladung	C
ρ	Raumladungsdichte	$\frac{1}{cm^3}$
S	subthreshold swing	$\frac{V}{Dekade}$
σ_d	flächenbezogene Ladung der Raumladungszone	$\frac{C}{cm^2}$
σ_n	flächenbezogene Ladung der Inversionsschicht	$\frac{C}{cm^2}$
T	Temperatur	K
U_A	Early-Spannung	V
U_{DB}	Drain-Bulk-Spannung	V
U_{DS}	Drain-Source-Spannung	V
U'_{DS}	Drain-Source-Sättigungsspannung	V
U_{FB}	Flachbandspannung	V
U_G	Gate-Spannung	V
U_{GB}	Gate-Bulk-Spannung	V
U_{GS}	Gate-Source-Spannung	V
U_{SB}	Source-Bulk-Spannung	V
U_{th}	Einsatzspannung	V

$v_{e,max}$	Sättigungsgeschwindigkeit von Elektronen	$\frac{m}{s}$
W	Weite	m
W_L	min. Energie des Leitungsbandes	eV
W_V	max. Energie des Valenzbandes	eV
x_d	Weite der Raumladungszone	m

Verwendete Abkürzungen

Abkürzung	**Bedeutung**
AA	Active Area
BL	Bitleitung (n^+ Gebiet)
CC	Conventional Concept
CHE	Channel Hot Electron
$DIBL$	Drain Induced Barrier Lowering
$INCE$	Inverse Narrow Channel Effect
LDD	Lightly Doped Drain
LI	Local Interconnect
$LOCOS$	Local Oxidation of Silicon
LVZ	Ladungsverlust nach Zykeln
$MNOS$	Metal-Nitride-Oxide-Semiconductor
$NROM$	Nitride Read Only Memory
NVM	Nonvolatile Memory
ONO	Oxide-Nitride-Oxide
$SONOS$	Semiconductor-Oxide-Nitride-Oxide-Semiconductor
STI	Shallow Trench Isolation
WL	Wortleitung (Gate)

Kapitel 1

Einleitung

1.1 Motivation für NROM

In den letzten fünf Jahren hat sich der Markt für nichtflüchtige Speicher rasant entwickelt. Er expandiert seit Jahren mit zweistelligen Zuwachsraten. Zudem gehen Analysten davon aus, dass in Zukunft der Anteil der nichtflüchtigen Speicher am gesamten Speichermarkt stetig wachsen wird. Gartner Dataquest rechnet bis 2007 mit einem Wachstum von durchschnittlich 20% pro Jahr, [11]. Dies ist in Abbildung 1.1 veranschaulicht.

Vor diesem Hintergrund ist der technische Fortschritt bei nichtflüchtigen Speichern von großer kommerzieller Bedeutung. Bisher ist das Floating Gate Konzept die häufigst verwendete Technologie für diese Klasse von Speichern. Durch die Attraktivität des Marktes und durch die stetig wachsenden Anforderungen drängen neuartige Konzepte in Richtung Realisierung und in Richtung Markt.

Eines dieser neuen Konzepte ist NROM. Es handelt sich hierbei um eine Technologie, die auf der lokalisierten Ladungsspeicherung in Nitrid basiert. Hierdurch wird es im Gegensatz zur herkömmlichen Floating Gate Technologie möglich, zwei physikalisch voneinander getrennte Bits in einer Zelle zu speichern. Zudem baut NROM auf einem normalen CMOS Prozess auf und somit aus prozesstechnischer Sicht auf weitgehend bekannten und produktionstauglichen Verfahren.

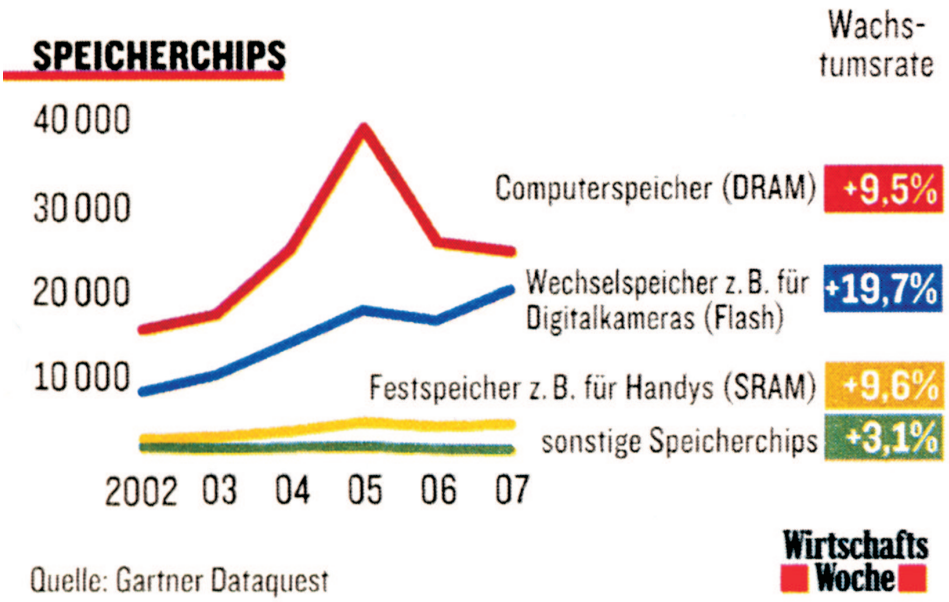


Abbildung 1.1: Wachstumsraten für verschiedene Segmente des Speichermarktes, [11].

Es gibt bereits erste Produkte mit dieser neuen Technologie auf dem Markt.

1.2 Aufbau der Arbeit

Ziel dieser Arbeit ist es, einen Beitrag für das Verständnis, sowie für die Weiterentwicklung von NROM-Speicherzellen zu leisten. So wird neben den Modellbetrachtungen und der Vorstellung eines neuen Multilevel-Betriebs für NROM auch ein neuartiges Zellkonzept, das mit shallow trench isolation (STI) arbeitet, experimentell untersucht.
Für das Verständnis der Funktionsweise von NROM-Zellen ist es zuvor unerlässlich, das Verhalten der MOS-Struktur und des MOS-Transistors zu verstehen. Aus diesem Grund werden in Kapitel 2 die wesentlichen Grundlagen hierfür besprochen. Besonderes Augenmerk wird auf die Auswirkung von sub-μm Effekten gelegt, da diese von großer Bedeutung für die hier behandelten Zellen sind. Zudem wird eine Einführung in die Funktionsweise von NROM-Speicherzellen gegeben.

In Kapitel 3 werden ein bekanntes Konzept und das neuartige STI-Konzept für NROM detailliert vorgestellt und besprochen. Zudem werden Modelle für wesentliche Aspekte der NROM-Zelle erörtert. Kapitel 4 beschäftigt sich mit der experimentellen Untersuchung des STI-Konzepts. Es werden eine Vielzahl von Messungen vorgestellt und deren Ergebnisse besprochen. Es wird gezeigt, dass die shallow trench isolated NROM-Zelle funktionstauglich ist. Zudem werden einige für dieses Konzept charakteristische Eigenschaften herausgearbeitet. Das Kapitel 5 stellt einen neuen Multilevel-Betrieb für NROM vor. Dieser macht NROM nicht nur multilevel tauglich, sondern verbessert auch für zwei Bits pro Zelle die Informationshaltung. Abschließend wird eine Zusammenfassung in Kapitel 6 gegeben.

Kapitel 2

Grundlagen

2.1 MOS-Struktur

Die MIS-Struktur (Metall-Isolator-Halbleiter) bildet die Grundlage für den MOS-Feldeffekttransistor (hier ist der Isolator das natürliche Oxid des Siliziums, SiO_2, damit wird aus MIS $\rightarrow$ MOS) und damit auch für die NROM-Speicherzelle. Aus diesem Grund wird hier auf die grundlegende Funktionsweise der MOS-Stuktur eingegangen.
Abbildung 2.1 zeigt die Verhältnisse einer MOS-Struktur, wenn zwischen Gate und Bulk ($\equiv$ Substrat) eine Spannung $U_{GB} > 0V$ angelegt wird. Der Halbleiter ist homogen p-dotiert mit der Dichte N_A. Das positive Potential des Metalls gegenüber dem Halbleiter führt zu einer negativen Influenzladung an der Halbleiteroberfläche, welche durch eine entsprechende positive Oberflächenladung auf der Metalloberfläche kompensiert wird. Da die Raumladungsdichte stückweise konstant ist, folgt aus

$$div\vec{E} = -\frac{\rho}{\epsilon_0\epsilon_r}, \tag{2.1}$$

dass die resultierende elektrische Feldstärke linear verläuft. Im Oxid, das als Isolator raumladungsfrei ist, ist die Feldstärke konstant, siehe Bild 2.1 b.
Ausgangspunkt für die Berechnung des Feldverlaufs ist die Poissongleichung:

$$\frac{d^2\phi}{dx^2} = -\frac{\rho(x)}{\epsilon_0\epsilon_{Si}} \tag{2.2}$$

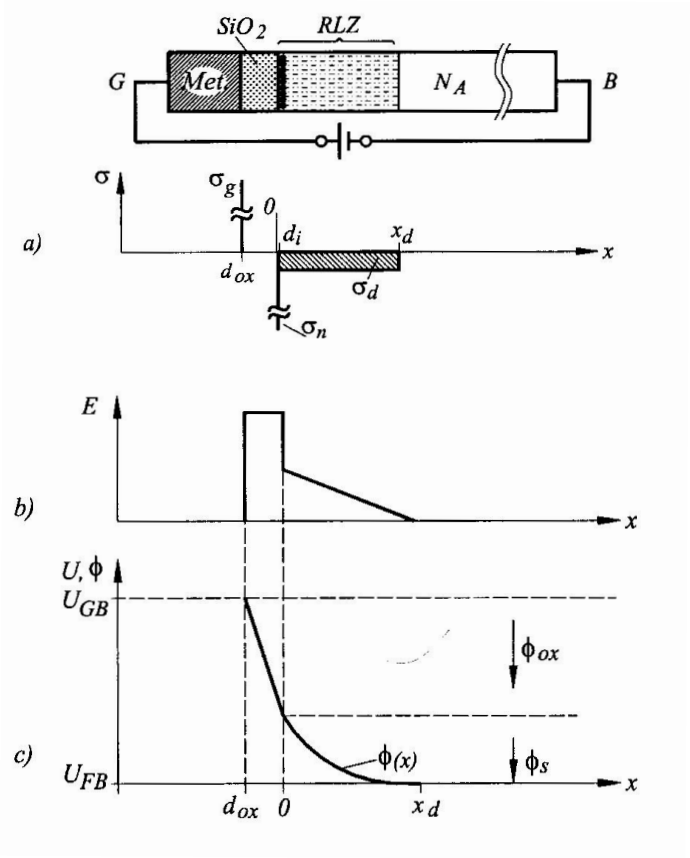


Abbildung 2.1: MOS-Struktur; a) Ladungsträgerverteilung; b) Feldverteilung bei charge-sheet Näherung; c) resultierender Spannungsverlauf; [24].

wobei für die Raumladungsdichte ρ unter Annahme vollständiger Ionisation gilt:

$$\rho(x) = -q[N_A + n(x) - p(x)] \tag{2.3}$$

Mit der Schottky-Näherung, d.h. mit völliger Verdrängung der beweglichen Ladungsträger aus der Raumladungszone im Halbleiter ($p(x) = 0$), ergibt sich für die Feldstärkenberechnung:

$$\int_{E(x)}^{E(x_d)} dE(x) = -\frac{q}{\epsilon_0 \epsilon_{Si}} \int_x^{x_d} (N_A + n(x))dx \qquad \text{für} \qquad 0 \leq x \leq x_d \tag{2.4}$$

Zum Lösen dieser Gleichung ist eine zusätzliche Näherung hilfreich. Die Ausdehnung der Raumladungszone, x_d, ist groß im Vergleich mit der Debye-Länge, L_{Dp}. Weiterhin lässt

sich zeigen, dass die Dicke der Inversionsschicht, d_i, klein ist gegenüber der Debye-Länge, [33]. Es gilt also: $d_i \ll L_{Dp} \ll x_d$. Daher kann mit guter Näherung angenommen werden, dass das gesamte Potential ϕ_S über der Verarmungszone abfällt. Diese Annahme wird als Charge Sheet Näherung bezeichnet, [24]. Mit ihrer Hilfe folgt aus Gleichung (2.4):

$$E_{Si}(x) = \frac{qN_A}{\epsilon_0 \epsilon_{Si}}(x_d - x) \tag{2.5}$$

und mit $E = -grad(\phi)$ für die Spannung:

$$\phi(x) = \frac{qN_A}{2\epsilon_0 \epsilon_{Si}}(x_d - x)^2 \tag{2.6}$$

Damit lässt sich die Oberflächenspannung, ϕ_S, leicht berechnen zu:

$$\phi(x = 0) = \phi_S = \frac{qN_A}{2\epsilon_0 \epsilon_{Si}} x_d^2 \tag{2.7}$$

Unter Berücksichtigung der Schottky-Näherung lässt sich nun die flächenbezogene Ladung der Raumladungszone berechnen:

$$\sigma_d = -qN_A x_d = -\sqrt{2qN_A \epsilon_0 \epsilon_{Si} \phi_S} \tag{2.8}$$

Die zweite Ladung, die sich unter dem Oxid befindet, ist die Ladung des Inversionskanals σ_n. Sie ist von entscheidender Bedeutung für die Funktion des Transistors. Mit Hilfe des Gaußschen Satzes, [52],

$$Q = \oint \vec{D} \cdot d\vec{A} \tag{2.9}$$

gelangen wir zu:

$$\sigma_n + \sigma_d = -D_{ox} \tag{2.10}$$

Zudem gilt:

$$\begin{aligned} D_{ox} &= \epsilon_0 \epsilon_{ox} \frac{d\phi}{dx} \\ &= \epsilon_0 \epsilon_{ox} \frac{\phi_{ox}}{d_{ox}} \\ &= C'_{ox} \phi_{ox} \end{aligned} \tag{2.11}$$

hierbei ist C'_{ox} die flächenbezogene Oxidkapazität und ϕ_{ox} die über der Oxidschicht abfallende Spannung.

Für die weiteren Berechnungen wird an dieser Stelle die Flachbandspannung, U_{FB}, eingeführt. Das ist diejenige Spannung, die man von außen an eine Halbleiterstruktur anlegen muss, damit die Bänder im Bändermodell des Halbleiters waagerecht verlaufen. U_{FB} ist somit die Summe der Kontaktspannungen zwischen Metall und Halbleiter, ϕ_K, und zwischen verschieden dotierten Bereichen im Halbleiter, ϕ_i. Für die Flachbandspannung gilt:

$$\begin{aligned} U_{FB} &= \phi_K + \phi_i \\ &= \phi_K + \phi_t \cdot \ln \frac{N_{A2}}{N_{A1}} \end{aligned} \tag{2.12}$$

N_{A1} und N_{A2} sind zwei unterschiedlich dotierte Gebiete im Silizium. ϕ_t ist die Temperaturspannung:

$$\phi_t = \frac{kT}{q} \tag{2.13}$$

Nach der Einführung der Flachbandspannung, wird wieder Abbildung 2.1 betrachtet, speziell Teil c). Aus diesem lässt sich leicht ablesen, dass gilt:

$$U_{GB} = \phi_{ox} + \phi_S + U_{FB} \tag{2.14}$$

Löst man diese Gleichung nach ϕ_{ox} auf und setzt dies in Gleichung (2.11) ein, so ergibt sich:

$$D_{ox} = C'_{ox} \left(U_{GB} - \phi_S - U_{FB}\right) \tag{2.15}$$

Nun kann die Ladung der Inversionsschicht unter Verwendung der Beziehungen (2.8) und (2.10) berechnet werden zu:

$$\begin{aligned} \sigma_n &= -C'_{ox} \left(U_{GB} - U_{FB} - \phi_S\right) + \sqrt{2qN_A\epsilon_0\epsilon_{Si}\phi_S} \\ &= -C'_{ox} \left(U_{GB} - U_{FB} - \phi_S - \gamma\sqrt{\phi_S}\right) \end{aligned} \tag{2.16}$$

Um die Übersichtlichkeit der Gleichung zu verbessern, ist der Faktor γ eingeführt worden. Er wird als Substratsteuerfaktor bezeichnet und ist definiert als:

$$\gamma = \frac{1}{C'_{ox}} \sqrt{2qN_A\epsilon_0\epsilon_{Si}} \tag{2.17}$$

Dieser Faktor wird im weiteren Verlauf der Arbeit noch häufiger betrachtet werden, da er dazu dient, die Eigenschaften von Transistoren bzw. NROM-Speicherzellen zu beschreiben.

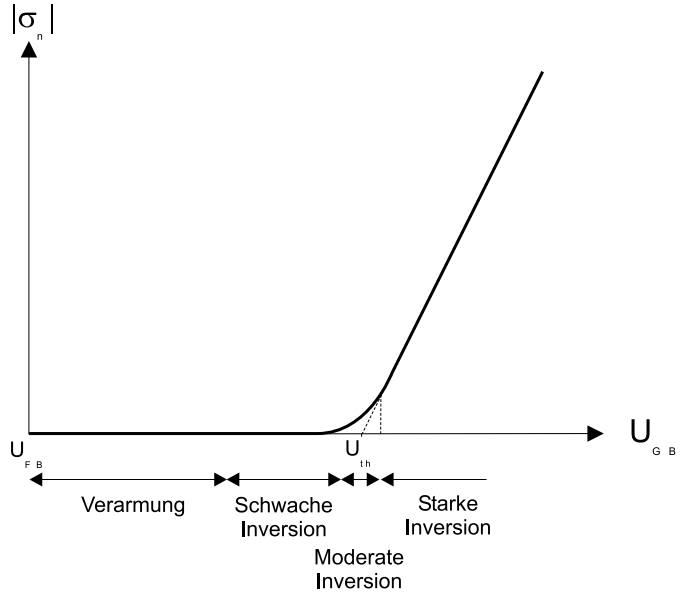


Abbildung 2.2: Betrag der Inversionsschichtladung pro Fläche, aufgetragen über der Gate-Substrat-Spannung.

In Abbildung 2.2 lässt sich erkennen, dass man durch Extrapolation der Kurve in starker Inversion auf $\sigma_n = 0$ die Spannung U_{th} erhält. Diese Spannung wird als Einsatzspannung bezeichnet. Für die Berechnung erkennt man dann aus Abbildung 2.3, dass in starker Inversion große Änderungen von U_{GB} nur sehr kleine Änderungen von ϕ_S zur Folge haben. Daher ist die Näherung üblich, dass in starker Inversion die Oberflächenspannung als konstant betrachtet wird, [69]:

$$\phi_S \approx \phi_0 \tag{2.18}$$

Dies ergibt zusammen mit Gleichung (2.16) für $\sigma_n = 0$ die Einsatzspannung:

$$U_{th} = U_{FB} + \phi_0 + \gamma\sqrt{\phi_0} \tag{2.19}$$

An dieser Stelle wird darauf hingewiesen, dass die Einsatzspannung eine Größe ist, die mit Hilfe einer Gleichung berechnet wird, welche für starke Inversion gilt. Die MOS-Struktur ist bei der Spannung $U_{GB} = U_{th}$ jedoch noch nicht in starker Inversion, was an Bild 2.2 verdeutlicht wird.

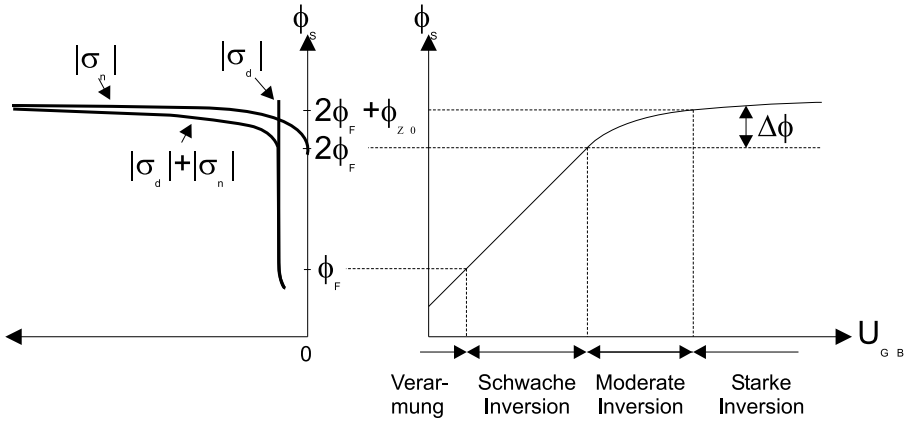


Abbildung 2.3: Oberflächenspannung über Gate-Substrat-Spannung (rechts) und Ladungen über Oberflächenspannung (links)

Um die Anwendbarkeit von Gleichung (2.19) zu erleichtern, wird eine Näherung für den Wert ϕ_0 bei starker Inversion benötigt. Hierzu dient Abbildung 2.3, die in ähnlicher Form z.B. bei Tsividis, [69], zu finden ist. Der Wert ϕ_0 ist schwer exakt zu bestimmen, er ist in der Abbildung 2.3 jedoch annähernd $2\phi_F + \phi_{Z0}$. Hierbei ist ϕ_F die Fermispannung, die durch

$$\phi_F = \phi_t \cdot \ln \frac{N_A}{n_i} \tag{2.20}$$

gegeben ist.

Am häufigsten wird in starker Inversion ϕ_0 genähert mit:

$$\phi_0 = 2\phi_F \qquad \text{siehe z.B.: [24],[69]} \tag{2.21}$$

Unter Betrachtung von Abbildung 2.3 ist diese Näherung jedoch nicht sonderlich exakt. Man sieht, dass die Oberflächenspannung noch weiter ansteigt, wenn U_{GB} erhöht wird.

Genauer wäre:

$$\phi_0 = 2\phi_F + \Delta\phi \tag{2.22}$$

Die Angabe von $\Delta\phi$ ist allerdings schwierig. Für ein homogen dotiertes Substrat beträgt sie einige $\phi_t, (\Delta\phi \approx 6\phi_t)$. Für nicht homogen dotierte Substrate, wie sie in der Realität meist auftreten, kann $\Delta\phi$ jedoch deutlich von diesem Wert abweichen.

2.2 MOS-Transistor

Der MOS-Transistor beruht wesentlich auf der in Abschnitt 2.1 besprochenen Zweipol-MOS-Struktur. Eine schematische Darstellung eines MOS-Transistors ist in Abbildung 2.4 mit dem zugehörigen Ausgangskennlinienfeld abgebildet.

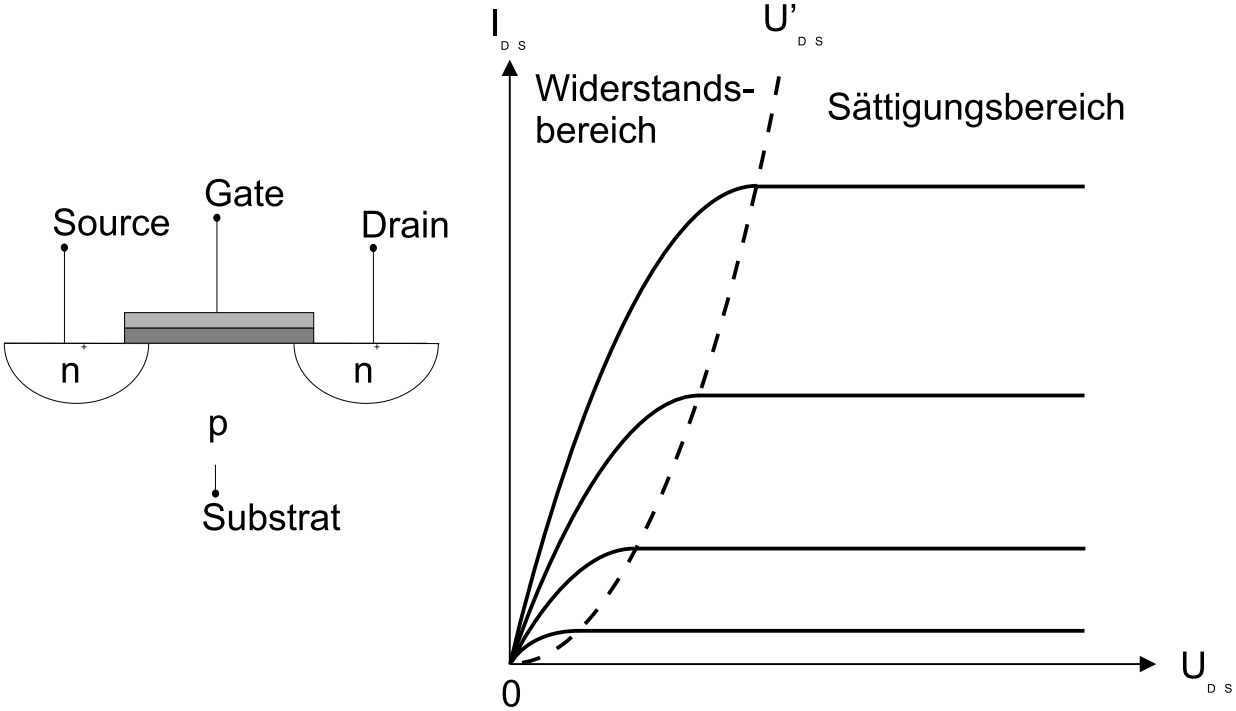


Abbildung 2.4: Schematische Darstellung eines Transistors (links) und dessen Ausgangskennlinienfeldes (rechts)

Im hier behandelten Fall eines n-Kanal Transistors (p-dotiertes Substrat) sind gegenüber

der Zweipol-MOS-Struktur zwei hochdotierte n-Gebiete, Source und Drain, hinzugekommen. Die allgemeine Funktionsweise eines Transistors soll hier nicht Erklärungsgegenstand sein, vielmehr soll auf Aspekte eingegangen werden, die für die untersuchten sub μm - Speicherzellen von großer Bedeutung sind. Grundlegendes zum Transistor ist leicht in der Literatur zu finden, z.B. [24],[29],[33],[69],... .

Eine elementare Größe ist die Einsatzspannung, U_{th}, des Transistors, insbesondere da ihre Verschiebung als Mittel zur Informationsspeicherung in NROM-Zellen verwendet wird. Hier muss eine Anpassung der Gleichung (2.19) aus Abschnitt 2.1 erfolgen. Für die Einsatzspannung eines Transistors ist die Gate-Source-Spannung, U_{GS}, wichtig, die nicht zwangsweise mit der Gate-Substrate-Spannung, U_{GB}, identisch ist. Daher gilt Gleichung (2.19) nur für einen Transistor, wenn die Source-Substrat-Spannung gleich Null ist ($U_{SB} = 0$). Sonst muss Gleichung (2.16) angepasst werden, woraus sich für die Einsatzspannug des Transistors ergibt:

$$U_{th} = U_{FB} + \phi_0 + \gamma\sqrt{\phi_0 + U_{SB}} \tag{2.23}$$

Für ϕ_0 gelten weiterhin die Gleichungen aus Abschnitt 2.1.

Darüber hinaus ist es für das Verständnis des Programmierens einer NROM-Zelle von Bedeutung, den Operationsbereich auf dem Kennlinienfeld zu kennen. Wie später näher erläutert wird, werden NROM-Zellen mit heißen Elektronen programmiert; ob und wo diese auftreten, hängt von den Betriebsbedingungen der Zelle ab. Aus diesem Grund soll hier der Übergang vom Widerstandsbereich in den Sättigungsbereich näher betrachtet werden. Dies ist im Ausgangskennlinienfeld in Abbildung 2.4 veranschaulicht. Die gestrichelte Linie stellt den Übergang zwischen den beiden Bereichen dar. Die Drain-Source-Spannung, bei der dieser Übergang stattfindet, wird Drain-Source-Sättigungsspannung, U'_{DS}, genannt. Sie lässt sich errechnen aus:

$$U'_{DS} = \frac{U_{GS} - U_{th}}{\alpha} \tag{2.24}$$

α ist ein Anpassungsfaktor, der nachfolgend näher betrachtet wird. Zur Spannung U'_{DS}

gehört ein Drain-Sättigungsstrom bei ($U_{DS} = U'_{DS}$) von:

$$I'_{DS} = \frac{W}{L}\mu C'_{ox}\frac{(U_{GS} - U_{th})^2}{2\alpha} \tag{2.25}$$

wobei W die Weite des Transistors und L seine Länge ist. μ ist die Beweglichkeit der Ladungsträger im Kanal. Für den Strom zwischen Drain und Source gilt insgesamt, [69]:

$$I_{DS} = \begin{cases} \frac{W}{L}\mu C'_{ox}\left[(U_{GS} - U_{th})\,U_{DS} - \frac{\alpha}{2}U_{DS}^2\right], & U_{DS} \le U'_{DS} \\ \frac{W}{L}\mu C'_{ox}\frac{(U_{GS}-U_{th})^2}{2\alpha} & U_{DS} > U'_{DS} \end{cases} \tag{2.26}$$

Es steht noch die Wahl für den Wert von α aus. In älteren oder einfachen Transistormodellen findet man häufig für α:

$$\alpha_0 = 1 \tag{2.27}$$

Diese Näherung bedeutet, dass die Tiefe der Verarmungsschicht über den ganzen Kanal als konstant angenommen wird. Die Tiefe wird überall gleich der Tiefe an der Source-Seite gesetzt. Dies hat eine Unterschätzung von $|\sigma_d|$ und eine Überschätzung von $|\sigma_n|$ zur Folge. Aus dieser Tatsache resultiert ein zu großer Wert für I_{DS}. Dieser Fehler wird besonders groß, wenn γ nicht klein ist. Zudem bekommt man für U'_{DS} zu hohe Werte.
In anderen Transistormodellen wird ein Wert angegeben, der aus einer Taylor-Reihenentwicklung abgeleitet wird, [50]:

$$\alpha_1 = 1 + \frac{\gamma}{2\sqrt{\phi_0 + U_{SB}}} \tag{2.28}$$

Diese Gleichung liefert nur für kleine Werte von $U_{DS} = U_{DB} - U_{SB}$ gute Werte. Im Allgemeinen wird $|\sigma_d|$ überschätzt und $|\sigma_n|$ unterschätzt. Dies resultiert in zu niedrigen Werten für I_{DS} und U'_{DS}. Die Fehler gehen also in die entgegengesetzte Richtung im Vergleich mit Gleichung (2.27). Es liegt daher nahe, in Gleichung (2.28) einen Korrekturfaktor einzuführen, damit folgt:

$$\alpha_2 = 1 + d_1\frac{\gamma}{2\sqrt{\phi_0 + U_{SB}}} \tag{2.29}$$

Für diesen Korrekturfaktor, d_1, werden in der Literatur Werte zwischen $0,5$ und $0,8$ verwendet,[34],[50].

2.3 MOS-Transistor im sub μm - Bereich

In diesem Abschnitt werden Effekte behandelt, die bei besonders kleinen Bauelementen auftreten. Die Speicherzellen, die in dieser Arbeit behandelt werden, haben elektrische Dimensionen von Länge zu Weite in der Größenordnung von $150nm$ zu $100nm$, und sind somit tief im sub-Mikrometer-Bereich. Die hier beschriebenen Mechanismen sollen in erster Linie dem besseren Verständnis dienen und weniger, um Formeln zur realitätsnahen Berechnung zu liefern.

2.3.1 Kanallängenmodulation

Die Kanallängenmodulation kann als Kurzkanaleffekt betrachtet werden. Da sie jedoch auch bei z.B. $10\mu m$ langen Transistoren beobachtet werden kann, wird sie hier nicht unter Kurzkanaleffekten behandelt. Die Kanallängenmodulation ist für spätere Betrachtungen wichtig.

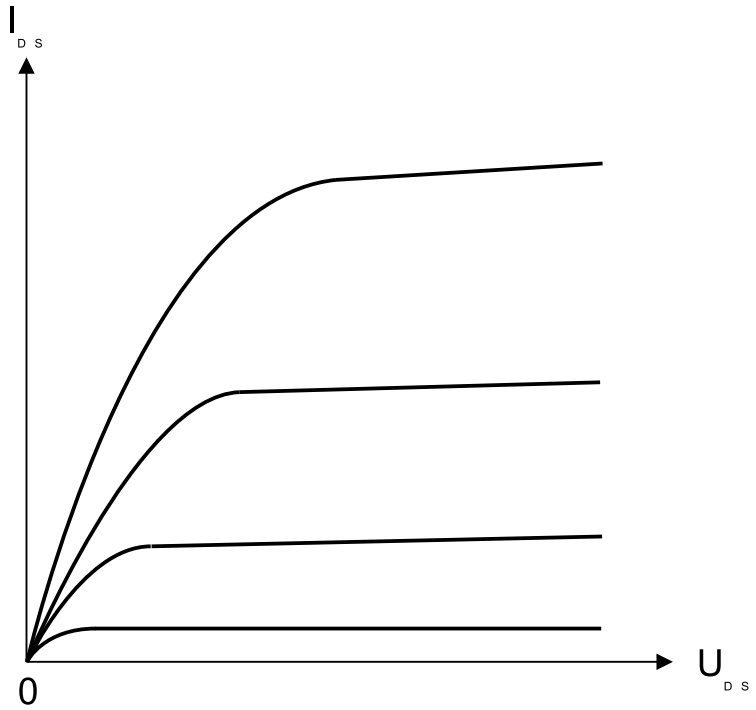


Abbildung 2.5: Kennlinienfeld unter Einfluss von Kanallängenmodulation

In Abbildung 2.5 ist die Auswirkung der Kanallängenmodulation zu sehen. Im Gegensatz, zu dem in Abbildung 2.4 dargestellten Kennlinienfeld, verlaufen die Kurven im Sättigungsbereich nicht mehr horizontal. Der Drain-Source-Strom steigt auch in diesem Bereich bei Erhöhung von U_{DS} weiter leicht an.
Abbildung 2.6 bietet eine anschauliche Erklärung für die Kanallängenmodulation. Dies hat nicht den Anspruch, eine exakte Beschreibung zu sein. So wird in dieser Betrachtung das am Gate anliegende Potential nicht für den Feldverlauf nahe der Drain berücksichtigt.

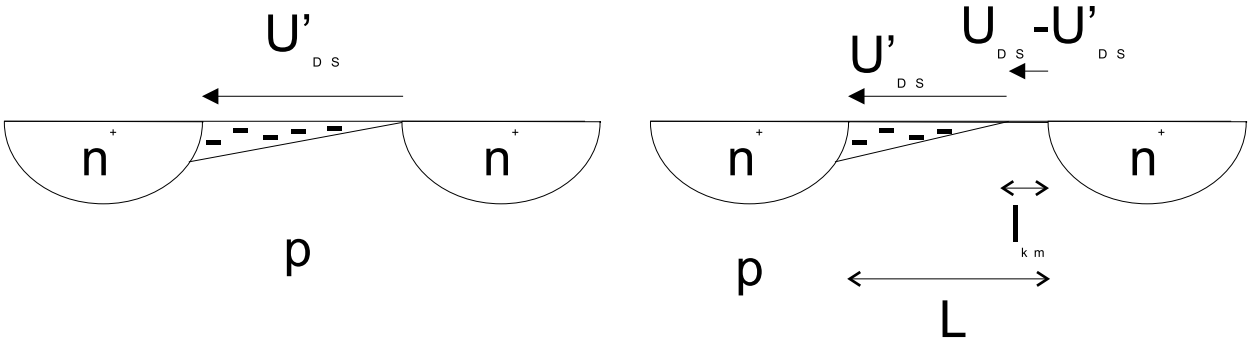


Abbildung 2.6: links: Transistor am pinchoff; rechts: Transistor bei weiter erhöhter Drain-Source-Spannung

Auf der linken Seite der Abbildung ist der Transistor am pinchoff-Punkt ($U_{DS} = U'_{DS}$, vgl. Abbildung 2.4). Für diese Betrachtung nehmen wir an, dass pinchoff auftritt, wenn betragsmäßig die Inversionsschichtladung $|\sigma_n|$ an der Drain sehr klein ist. Dies bedeutet eine sehr viel kleinere Ladung als die der Verarmungszone $|\sigma_d|$. Wenn die Spannung U_{DS} weiter über U'_{DS} hinaus erhöht wird, so nimmt die Ladung der Inversionsschicht am Drain-Ende weiter ab. Der pinchoff-Punkt verschiebt sich weiter nach links, wie es auf der rechten Seite der Abbildung 2.6 dargestellt ist. Dies bedeutet jedoch nicht, dass kein Strom mehr zwischen Source und Drain fließt, was zudem der Kontinuitätsgleichung widersprechen würde. Wie allgemein von *pn*-Übergängen bei Bipolar-Transistoren bekannt ist, können große Ströme durch Verarmungszonen fließen. Die Elektronen müssen sich in diesem Bereich mit sehr hoher Geschwindigkeit bewegen, damit bei sehr kleinen Werten von $|\sigma_n|$ ein großer Strom zustande kommt.

Die Bedingung, dass die Elektronen sich zwischen dem pinchoff-Punkt und der Drain mit sehr hoher Geschwindigkeit bewegen, wird als eine weitere Definition für den pinchoff verwendet. Sie besagt, dass der pinchoff-Punkt derjenige Punkt ist, ab dem die Elektronen sich mit Sättigungsgeschwindigkeit bewegen.

In Abbildung 2.6 ist rechts gezeigt, dass über dem Kanal von der Source bis zum pinchoff-Punkt stets die Spannung U'_{DS} abfällt, der Rest der Spannung $U_{DS}-U'_{DS}$ fällt zwischen dem pinchoff-Punkt und der Drain ab. Diese Strecke wird mit l_{km} bezeichnet. Es ist leicht nachvollziehbar, dass die Strecke l_{km} immer größer wird, je mehr U_{DS} die Sättigungsspannung U'_{DS} überschreitet. Dies ist gleichbedeutend mit einer Verkürzung der Inversionsschicht. Dieser Vorgang wird als Kanallängenmodulation bezeichnet.

Mit der Annahme, dass das Feld in der Verarmungszone horizontal verläuft, kann mit Hilfe der Poisson-Gleichung ein Wert für l_{km} angegeben werden, [69]:

$$l_{km} = \sqrt{\frac{2\epsilon_0\epsilon_{Si}}{qN_A}}\left[\sqrt{\phi_D + (U_{DS} - U'_{DS})} - \sqrt{\phi_D}\right] \tag{2.30}$$

wobei für ϕ_D gilt:

$$\phi_D = \frac{\epsilon_0\epsilon_{Si}E_k^2}{2qN_A} \tag{2.31}$$

E_k ist die kritische Feldstärke oberhalb derer sich die Elektronen mit Sättigungsgeschwindigkeit bewegen. E_k liegt hier in horizontaler Richtung an. Bei der Betrachtung von Elektronen sind für E_k Werte im Bereich $8 \times 10^3 \ldots 3 \times 10^4 V/cm$ einzusetzen. Dies ergibt sich durch

$$E_k = \frac{|v_{e,max}|}{\mu} \tag{2.32}$$

und beruht auf folgenden Annahmen:

$$\text{maximale Elektronengeschwindigkeit:} \qquad |v_{e,max}| = 5 \times 10^6 \ldots 2 \times 10^7 \frac{cm}{s} \tag{2.33}$$

$$\text{Elektronenbeweglichkeit:} \qquad \mu = 650 \frac{cm^2}{Vs} \tag{2.34}$$

Ein zur Gleichung (2.30) sehr ähnlicher Ausdruck wurde von Reddi und Sah vorgestellt,

[56]:

$$l_{km,2} = \sqrt{\frac{2\epsilon_0\epsilon_{Si}}{qN_A}\left(U_{DS} - U'_{DS}\right)} \tag{2.35}$$

Verglichen mit Gleichung (2.30) liefert (2.35) größere Werte. Sie ist für Zellen mit einer Kanallänge von $2 \ldots 4\mu m$ und Dotierungen im Bereich $N_A = 10^{15} \ldots 10^{16} cm^{-3}$ geeignet, [75], [16]. Für Zellen mit einer Kanallänge von nur $\sim 200nm$ wird die Strecke zwischen pinchoff-Punkt und Drain mit der Gleichung (2.35) überschätzt. Die gleiche Tendenz gilt für Gleichung (2.30), dennoch sind beide Gleichungen für das Verständnis hilfreich.
Nun lässt sich der Anstieg des Source-Drain-Stromes im Sättigungsbereich der Kennlinien aus Abbildung 2.5 erklären. Der Sättigungsstrom ist aus Gleichung (2.25) bekannt:

$$\begin{aligned} I'_{DS} &= \frac{W}{L}\mu C'_{ox}\frac{(U_{GS} - U_{th})^2}{2\alpha} \\ &= c_1\frac{1}{L} \end{aligned} \tag{2.36}$$

Für Spannungen die größer als U'_{DS} sind, muss Gleichung (2.26) modifiziert werden zu:

$$I_{DS} = c_1\frac{1}{L - l_{km}} \tag{2.37}$$

Aus den Gleichungen (2.36) und (2.37) ergibt sich:

$$I_{DS} = \frac{I'_{DS}}{1 - \frac{l_{km}}{L}} \tag{2.38}$$

Aus dieser Gleichung lässt sich unmittelbar ablesen, dass ein weiteres Erhöhen von U_{DS} oberhalb von U'_{DS}, welches eine Vergrößerung von l_{km} hervorruft, in einem Anstieg des Source-Drain-Stromes, I_{DS}, resultiert.
Ein vereinfachtes empirisches Modell wurde von Merckel, [48], vorgeschlagen. Der Source-Drain-Strom berechnet sich hierzu:

$$I_{DS} = I'_{DS}\left(1 + \frac{U_{DS} - U'_{DS}}{U_A + U'_{DS}}\right) \tag{2.39}$$

U_A wird als Early-Spannung bezeichnet, für sie gilt:

$$U_A = k_A L\sqrt{N_A} \tag{2.40}$$

wobei k_A ein Proportionalitätsfaktor ist, der im Bereich $1 \times 10^{-3} \ldots 2 \times 10^{-3} V \cdot cm^{1/2}$ liegt. Die Spannung U_A erhält man, indem man die Kennlinie im Sättigungsbereich nimmt und diese bis $I_{DS} = 0$ extrapoliert.

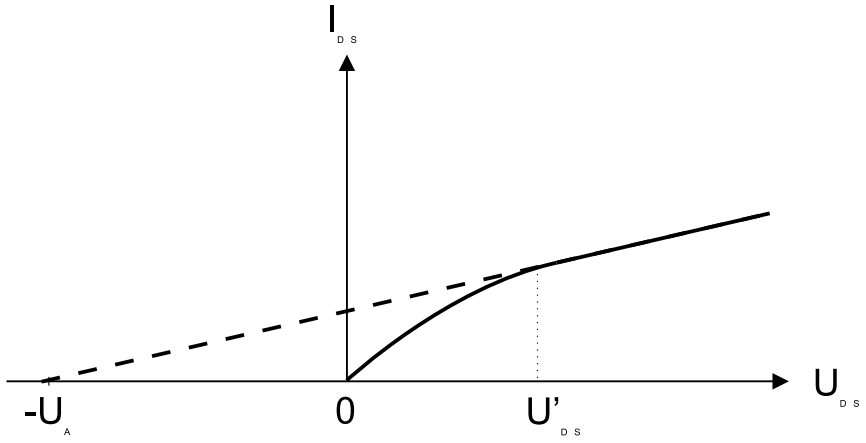


Abbildung 2.7: Vereinfachtes Modell zur Kanallängenmodulation

Dies ist in Abbildung 2.7 dargestellt. Die extrapolierten Kurven für verschiedene Werte von U_{GS} schneiden die x-Achse in einem Punkt, nämlich bei $-U_A$.

2.3.2 Ladungsträgerbeweglichkeit

Die Beweglichkeit der Elektronen im Kanal nimmt bei kleinen Dimensionen ab. Hierzu ist von K. Chen et al.,[8], [9], ein Modell für die Ladungsträgerbeweglichkeit vorgeschlagen worden. Die Abhängigkeit der Beweglichkeit von der effektiven Feldstärke ist in Abbildung 2.8 (es gilt: $T_{ox} = d_{ox}$) dargestellt.

Da für NROM-Speicherzellen immer n-Kanal Transistoren verwendet werden, ist hier in erster Linie die Beweglichkeit der Elektronen von Interesse. Für sie gilt im Rahmen der Achsenbeschriftung $\alpha = 0$.

Die Abhängigkeit der Beweglichkeit kann als Funktion der effektiven, vertikalen elektrischen Feldstärke in der Inversionsschicht ausgedrückt werden, [58]. Diese berechnet sich

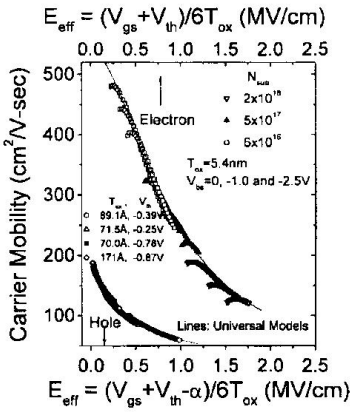


Abbildung 2.8: Modell für die Ladungsträgerbeweglickeit von NMOS Elektronen und PMOS Löchern zusammen mit experimentellen Daten,[8]

für Elektronen zu:

$$\begin{aligned} E_{eff} &= C'_{ox}\frac{\frac{U_{GS}-U_{th}}{2}+U_{th}}{\epsilon_0\epsilon_{Si}} = \frac{\epsilon_0\epsilon_{ox}}{d_{ox}}\frac{U_{GS}+U_{th}}{2\epsilon_0\epsilon_{Si}} \\ &= \frac{U_{GS}+U_{th}}{6d_{ox}} \end{aligned} \tag{2.41}$$

Unter Verwendung dieser Gleichung lässt sich eine Beziehung für die Beweglichkeit der Elektronen angeben, die in ähnlicher Form auch von Liang et al.,[38] , vorgeschlagen wurde:

$$\begin{aligned} \mu_n(U_{GS},U_{th},d_{ox}) &= \frac{540}{1+\left(\frac{E_{eff}}{0{,}9}\right)^{1{,}85}} \\ &= \frac{540}{1+\left(\frac{U_{GS}+U_{th}}{5{,}4d_{ox}}\right)^{1{,}85}} \end{aligned} \tag{2.42}$$

Es gelten folgende Dimensionen für diese Formel: μ_n in $cm^2/(Vs)$, E_{eff} in MV/cm, U_{GS} und U_{th} in MV und d_{ox} in cm.

Bei Anwendung der Gleichung (2.42) für eine Einsatzspannung von $U_{th} = 2V$ und eine Oxiddicke von $d_{ox} = 20nm$ ergibt sich der in Abbildung 2.9 dargestellte Verlauf der Elektronenbeweglichkeit als Funktion der Gate-Source-Spannung.

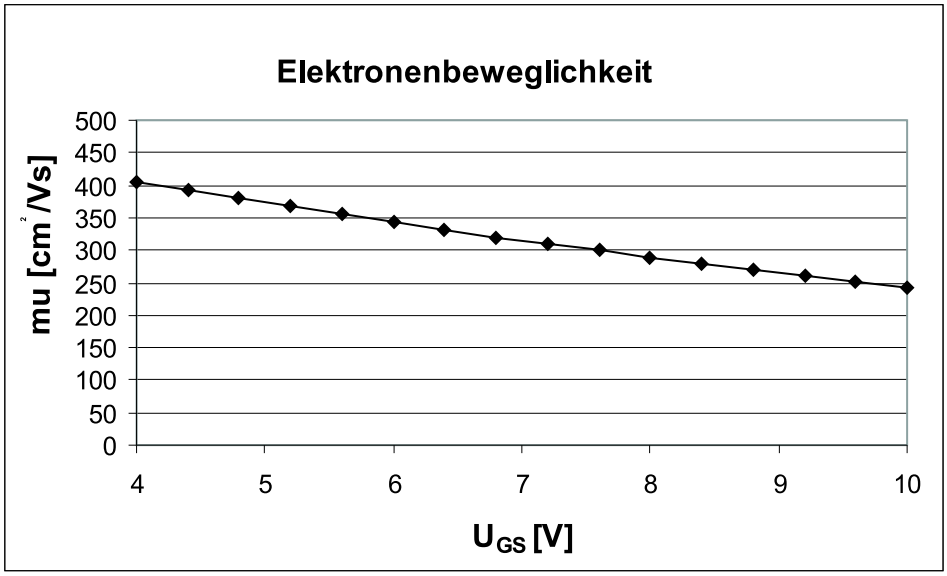


Abbildung 2.9: Elektronenbeweglichkeit als Funktion von U_{GS} für $U_{th} = 2V$ und $d_{ox} = 20nm$

2.3.3 Kurzkanaleffekt

Bei der Herleitung der Einsatzspannung U_{th} (vgl. Gleichungen (2.19) und (2.23)) sind die Auswirkungen der Source- und Drain-Gebiete nicht berücksichtigt worden. Es wurde dasjenige Potential ermittelt, welches benötigt wird, um einen Inversionskanal in einer MOS-Struktur zu erzeugen. Die Situation ist in Abbildung 2.10 dargestellt.

Die Vernachlässigung der Randbereiche ist für Transistoren mit einer großen Kanallänge eine gute Näherung, wie der Vergleich von (a) und (b) verdeutlicht. Wird die Kanallänge jedoch sehr klein, so haben die Raumladungszonen der Source- bzw. Drain-Gebiete eine nicht mehr zu vernachlässigbare Auswirkung. Man sieht, dass Ausschnitt (d) keine gute Näherung für die Situation in Ausschnitt (c) ist. Der Transistor (c) hat eine geringere Einsatzspannung und damit bei gleicher Drain-Source-Spannung einen höheren Strom I_{DS}, als man dies aus einer Betrachtung von (d) ableiten würde.

Die Tatsache, dass die Verarmungszone unter dem Gate nicht mehr nur durch das Gatepotential gesteuert wird, sondern auch durch die Raumladungszonen der pn-Übergänge von

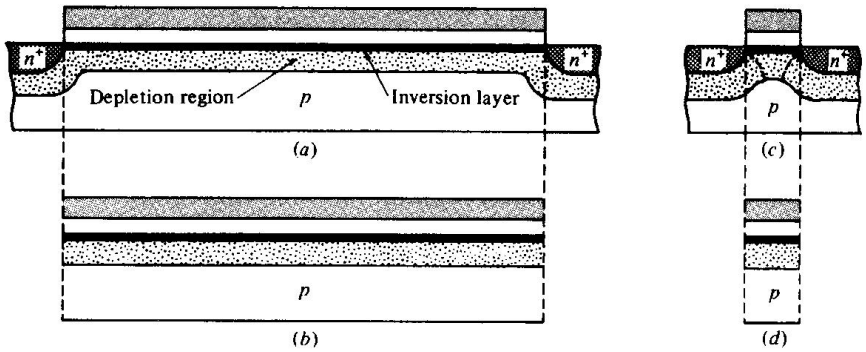


Abbildung 2.10: (a)Langkanal Transistor; (b) Kanal von (a) unter Vernachlässigung der Randeffekte; (c) Kurzkanal Transistor; (d) Kanal von (c) unter Vernachlässigung der Randeffekte;[69].

Source und Drain beeinflusst wird, wird als charge-charring bezeichnet.
Ein einfaches Trapezmodell aus der Literatur ,[71],[76], ist in Abbildung 2.11 dargestellt.

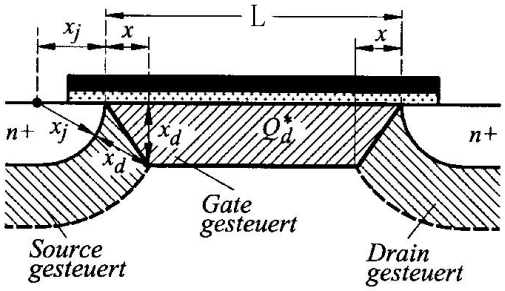


Abbildung 2.11: Trapezmodell zur Beschreibung des charge-charring bei $U_{DS} = 0V$,[24].

Die Einsatzspannung ist aus Gleichung(2.23) als

$$\begin{aligned} U_{th} &= U_{FB} + \phi_0 + \gamma\sqrt{\phi_0 + U_{SB}} \\ &= U_{FB} + \phi_0 - \frac{Q_d}{C_{ox}} \end{aligned} \tag{2.43}$$

bekannt. Mit

$$Q_d = \sigma_d \cdot W \cdot L, \tag{2.44}$$

wird die Ladung der Verarmungszone bezeichnet, die durch das Gate verursacht wird. Das Trapezmodell besagt, dass nur noch die Ladung Q_d^* durch das Gate gesteuert wird. Für diese Ladung ist aus Abbildung 2.11 abzulesen:

$$Q_d^* = \sigma_d \cdot W \cdot (L - W) \tag{2.45}$$

Unter Verwendung von

$$(x_j + x)^2 + x_d^2 = (x_j + x_d)^2 \tag{2.46}$$

lässt sich die Abhängigkeit der Einsatzspannung von der reduzierten Kanallänge L berechnen zu:

$$\begin{aligned} U_{th} &= U_{FB} + \phi_0 + \frac{Q_d^* \cdot Q_d}{Q_d \cdot C_{ox}} \\ &= U_{FB} + \phi_0 + \underbrace{\left[1 - \frac{x_j}{L}\left(\sqrt{1 + 2\frac{x_d}{x_j}} - 1\right)\right]}_{F_S} \gamma\sqrt{\phi_0 + U_{SB}} \\ &= U_{FB} + \phi_0 + F_S \cdot \gamma\sqrt{\phi_0 + U_{SB}} \end{aligned} \tag{2.47}$$

F_S beschreibt die Absenkung der Einsatzspannung. Die Differenz gegenüber der in Gleichung (2.23) berechneten Einsatzspannung beträgt:

$$\Delta U_{th,cc1} = -\left(1 - \frac{Q_d^*}{Q_d}\right)\gamma\sqrt{\phi_0 + U_{SB}} \tag{2.48}$$

Gleichung (2.47) lässt sich weiter vereinfachen, indem man F_S in eine Taylor-Reihe entwickelt und die Glieder höherer Ordnung vernachlässigt,[49]. Zur Kompensation dieser Vernachlässigung wird ein Anpassungsfaktor β_1 eingeführt, der normalerweise gleich eins gesetzt wird.

$$\frac{Q_d^*}{Q_d} = 1 - \beta_1 \frac{x_d}{L} \tag{2.49}$$

Verwendet man diese Gleichung, um die Differenz in der Einsatzspannung zu berechnen, so ergibt sich schließlich:

$$\Delta U_{th,cc2} = -2\beta_1 \frac{\epsilon_{Si} \cdot d_{ox}}{\epsilon_{ox} \cdot L}(\phi_0 + U_{SB}) \tag{2.50}$$

Man erkennt an Gleichung (2.50) leicht, dass eine kürzere Kanallänge durch ein dünneres Gateoxid kompensiert werden kann.

Drain induced barrier lowering (DIBL)

Die bisherigen Betrachtungen im Abschnitt 2.3.3 wurden nur für vernachlässigbar kleine Drain-Source-Spannungen gemacht. Wird jedoch U_{DS} erhöht und somit auch die Drain-Bulk-Spannung U_{DB}, so vergrößert sich die Verarmungszone an der Drain. Die Einsatzspannung sinkt folglich mit steigender Spannung U_{DS} weiter ab.
Zur Berechnung kann ähnlich wie oben vorgegangen werden, jedoch wird das Trapez verzerrt. Es ergibt sich eine Anpassung von Gleichung (2.50) zu:

$$\Delta U_{th,cc2} = -2\beta_1 \frac{\epsilon_{Si} \cdot d_{ox}}{\epsilon_{ox} \cdot L} \left[(\phi_0 + U_{SB}) + \beta_2 U_{DS}\right] \tag{2.51}$$

Aus der Rechnung folgt $\beta_2 = 0,25$. β_2 kann jedoch auch als Anpassungsfaktor verwendet werden.
Gleichung (2.51) wurde mit Hilfe des charge-sharing Modells errechnet. Es sagt aus, dass mit steigender Spannung U_{DS} die Einsatzspannung sinkt. Dies ist gleichbedeutend mit der Aussage, dass die Potentialbarriere für Elektronen zum Eintritt in den Kanal sinkt, [66]. Die Senkung der Barriere ist durch die Spannungsverhältnisse an der Drain verursacht, daher wird sie auch als Drain induced barrier lowering (DIBL) bezeichnet.

2.3.4 Schmalkanaleffekt

Bei schmalen Transistoren gewinnen die Randbereiche an Bedeutung. Die Auswirkung auf die Einsatzspannung hängt in diesem Fall wesentlich von den Art der seitlichen Begrenzung der Zellen ab. Hier werden zwei gebräuchliche Typen betrachtet, LOCOS (local oxidation of silicon) und STI (shallow-trench isolation). Sie sind anschaulich in Abbildung 2.12 dargestellt.

Die Auswirkungen auf die Einsatzspannungen von Transistoren, die auf diese beiden unterschiedlichen Weisen seitlich begrenzt werden, sind in Abbildung 2.13 zu sehen. Wie es zu diesem Verhalten kommt, wird im Folgenden besprochen.

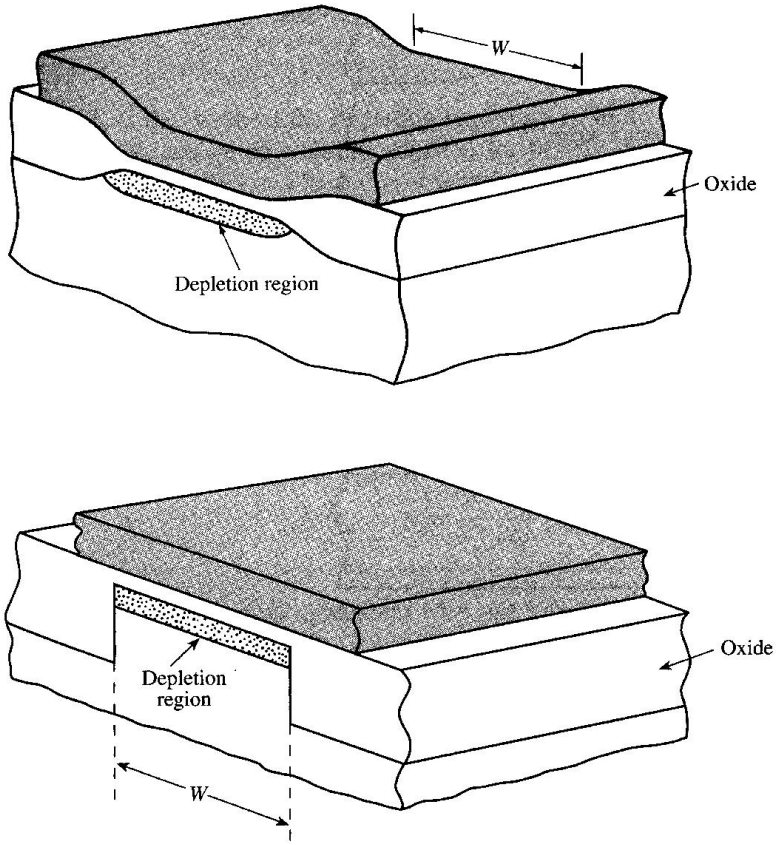


Abbildung 2.12: Querschnitte durch Transistoren; oben: LOCOS isolierter Transistor; unten: shallow-trench-isolated (STI) Transistor; [69].

LOCOS Isolation

Die Ausdehnung der Verarmungszone (bei positiver Gate-Spannung an einem NMOS-Transistor) ist bei dieser Art der Isolation seitlich nicht physikalisch begrenzt. Das Oxid wird zwar dicker zum Rand hin, aber es gibt einen Randbereich, der durch die Randfelder des Gates noch beeinflusst wird (siehe Abb. 2.12 oben). Im Randbereich werden ebenfalls die ionisierten Akzeptorladungen ausgeräumt. Somit wird durch das Gate ein Kanal gesteuert, der effektiv breiter ist, als die angenommene Breite. Bei sehr breiten Transistoren ist dieses Randvolumen der Verarmungszone verschwindend klein, bezogen auf das Gesamtvolumen der Verarmungszone und somit zu vernachlässigen. Ist ein Transistor jedoch

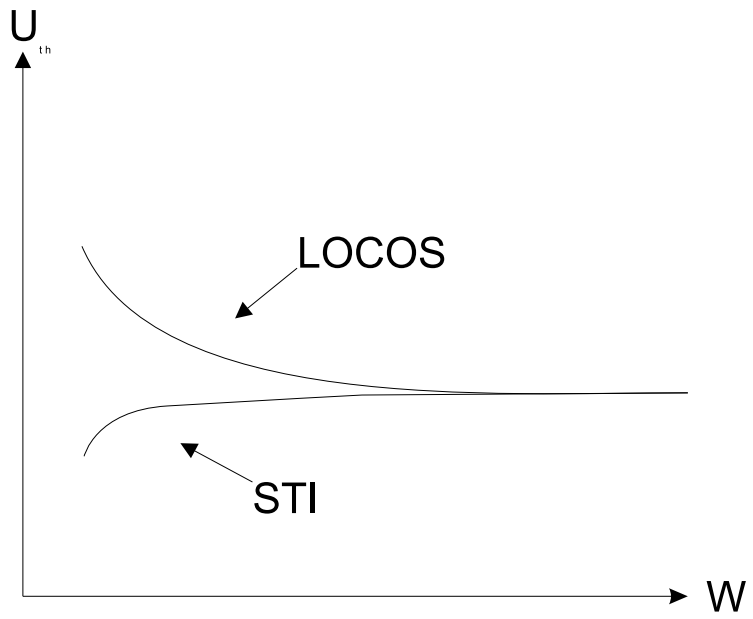


Abbildung 2.13: Effektiver Einsatzspannungsverlauf bei, bis auf die Isolation, identischen MOSFETs

sehr schmal, so erklärt dies, warum die Einsatzspannung steigt.
Dieser Effekt ist bei den NROM-Zellen, die in Kapitel 3.1 behandelt werden, zu beobachten. Sie sind von ihrer elektrischen Ausdehnung her seitlich nicht physikalisch begrenzt, bei ihnen hängt die elektrische Kanalweite maßgeblich von der anliegenden Gate-Spannung ab.

STI Isolation

Im Gegensatz zur LOCOS Isolation kann sich der Kanal bei STI Isolation nicht über die Breite W des Transistors hin ausdehnen. In diesem Fall tragen die Randfelder des Gates sogar zu einer Verringerung der Einsatzspannung bei. Sie unterstützen das Ausräumen der Ladungsträger an den Außenkanten und erleichtern somit das Entstehen einer Inversionsschicht. Hierdurch sinkt die Einsatzspannung.
In der Praxis muss zusätzlich das Verhalten des Dotierstoffes im Kanalbereich berücksichtigt werden. Für eine p-Dotierung ist es üblich, Bor zu verwenden. Die Löslichkeit von Bor

in Oxid ist höher, als die in Silizium. Hierdurch kann es zu einer Segregation des Bors im Randbereich in das STI kommen. Dies hat eine lokal geringere Bor-Konzentration am Rand zur Folge. Es führt ebenfalls zu einer Senkung der Einsatzspannung.
Das Diffusions- und Segregationsverhalten ist z.B. von Jung et al.,[28], behandelt worden. Jung legt zudem die Auswirkung der Defekte, die durch die Source- bzw. Drain-Implantation entstehen, auf das Verhalten des Dotierstoffs an den verschiedenen Grenzflächen dar. So ist in der Mitte unter dem Gate ein Anreicherung von Bor zu beobachten, wohingegen es an der Kante zum STI zu einer Verarmung kommt.
Ist einer oder sind beide dieser Effekte stark ausgeprägt (kommt z.B. sogar eine physikalische Dünnung des Gateoxides hinzu), so kann es zu der Ausbildung eines 'corner-devices' führen. Dies bedeutet, dass am Rand ein Inversionskanal deutlich messbar zu einem früheren Zeitpunkt geöffnet wird als in der Mitte des Transistors. Man kann dies als eine Parallelschaltung von zwei 'Rand-Transistoren' und einem 'Mittel-Transistor' betrachten. Dies lässt sich bei starker Ausprägung durch einen 'hump' in der Kennlinie erkennen.
Das Absinken der Einsatzspannung bei kleinen Weiten (INCE=inverse narrow channel effect) ist jedoch nur bei STI-begrenzten Transistoren mit Oberflächenkanal zu beobachten, bei buried-channel Transistoren steigt, wie bei LOCOS Transistoren, die Einsatzspannung für kleine Weiten,[61],[36].
Die Auswirkungen des INCE können u.a. durch folgende Maßnahmen reduziert werden:

- Implantation in die Seitenwand,[37],[60]
- Abschrägung der Seitenwand,[61]
- Rundung der Kanten,[15]

Jedoch sind bei weiterer Verkleinerung der Strukturen nicht alle diese Möglichkeiten anwendbar, so bleibt z.B. kein Platz mehr für das Abschrägen der Seitenwände.

2.3.5 Heiße Elektronen

Heiße Elektronen (CHE = channel hot electrons) sind in normalen Transistoren unerwünscht, da sie zu Degradation und somit zu Alterung führen. Für NROM-Speicherzellen sind sie jedoch notwendig und folglich erwünscht.
Zuvor kommen wir aber zum Entstehungsprozess und zu den Eigenschaften von heißen Elektronen. Wie bereits diskutiert, steigt die horizontale Komponente der elektrischen Feldstärke von der Source zur Drain hin an. Der Wert ihres Maximums ist dabei sowohl von U_{DS}, wie auch von der Kanallänge, L, abhängig. Für kürzere Kanallängen steigt der Spitzenwert der Feldstärke für konstantes U_{DS} stark an. Überschreitet die horizontale Feldstärke den Wert der kritischen Feldstärke, E_k, (siehe Gl.(2.32)) so bewegen sich die Elektronen in guter Näherung ab diesem Punkt (pinchoff-Punkt) mit Sättigungsgeschwindigkeit bis zur Drain. Die Geschwindigkeit in Feldrichtung nimmt nicht mehr zu, jedoch steigt die kinetische Energie der Elektronen weiter. Diese wird jedoch durch zufällige Kollisionen immer wieder abgegeben. Einige der Elektronen in diesem Bereich gewinnen eine beträchtlich hohe Energie, sie werden als heiße Elektronen bezeichnet. Diese können durch Stoßionisation kovalente Bindungen aufschlagen und somit Elektronen-Loch-Paare erzeugen. Die erzeugten Elektronen werden durch das Feld zur Drain hin beschleunigt, wohingegen die Löcher zum Substrat hin abfließen. Durch den Fluß dieser Löcher kommt es zu einem Strom, der als Drain-Bulk Strom ,I_{DB}, bezeichnet wird.
Einige der erzeugten Elektronen besitzen so viel Energie, dass sie die Potentialbarriere des Gateoxides überwinden können und ins Gate abfließen. Da im Fall der NROM-Zelle das Gateoxid durch einen ONO-Stapel ersetzt wird, werden diese Elektronen dann mit einer hohen Wahrscheinlichkeit im Nitrid eingefangen. Dort verursachen sie dann die gewünschte Verschiebung der Einsatzspannung.
Die heißen Elektronen können jedoch auch, genauso wie bei üblichen Transistoren, negative Folgen haben. Ein kleiner Teil von ihnen schädigt die Silizium-Oxid-Grenzschicht und erhöht die Dichte der Grenzschichtzustände. Ein anderer Teil kann das Oxid selbst schädigen, indem er neue Störstellen erzeugt. Dies ist als Alterung bei herkömmlichen Tran-

sistoren bekannt, [20],[26],[73]. Die Entstehung von heißen Elektronen kann stark reduziert werden, indem ein Teil der Drain nur schwach dotiert wird (LDD = Lightly Doped Drain), [54].

Injektion von sekundären Elektronen

Für das Programmieren einer NROM-Zelle werden, wie erläutert, heiße Elektronen benötigt. Über die erwähnten negativen Folgen, die durch heiße Elektronen verursacht werden können, kommt für NROM noch eine weitere negative Auswirkung hinzu. Die Löcher, die bei der Stoßionisation entstehen, werden zum Substrat hin beschleunigt. Ein sehr geringer Anteil dieser Löcher sammelt so viel Energie auf diesem Weg, dass er seinerseits wieder Stoßionisationen durchführen kann. Dies geschieht deutlich tiefer im Silizium als das erste Aufbrechen von kovalenten Bindungen. Mit einer geringen Wahrscheinlichkeit können die bei dieser zweiten Stoßionisation entstandenen Elektronen (auch 'Sekundärelektronen' genannt) die Potentialbarriere des Oxides überwinden und werden im Nitrid des ONO (Oxid-Nitrid-Oxid Stapels) eingefangen. Somit verändern auch diese Sekundärelektronen die Einsatzspannung der Zelle.
Durch den tieferen Entstehungsort der Sekundärelektronen wird ein grosser Anteil von ihnen weiter zum Kanal hin injiziert, als die erwünschten Primaries. Somit wird die Verteilung der Elektronen breiter. Berücksichtigt man nun, dass die Injektionsposition der Löcher, die zum Löschen verwendet werden, nicht beliebig angepasst werden kann, so stellt sich heraus, dass Elektronen, die weit über dem Kanal (mit großem Abstand zum pn-Übergang) injiziert werden, später nicht mehr durch Löcher im gleichen Gebiet gelöscht werden können.
Um Sekundärelektronen beim späteren Löschen elektrisch zu kompensieren, müssen viele Löcher im Bereich des pn-Übergangs eingeschossen werden. Bei wiederholtem Programmieren und Löschen kann es so zu einer lokalen Anhäufung von Löchern kommen, die die Ladungsspeichereigenschaften der Zelle negativ beeinflusst.
Der Effekt der Sekundärelektronen hängt von einer ganzen Reihe von Parametern ab, so

z.B. der Wannendotierung, der Kanallänge und der Substratvorspannung. Auf diesen Effekt wird im weiteren Verlauf der Arbeit noch näher eingegangen.

2.3.6 Diodendurchbruch und Punchthrough

Diodendurchbruch

Mit Diodendurchbruch ist hier der Durchbruch der Drain- bzw. Source-Substrat Diode gemeint. Wird z.B. an der Drain eine zu hohe Spannung gegenüber dem Substrat angelegt, so bricht die n^+p-Diode durch, und es fließt ein großer Strom direkt von der Drain ins Substrat. Dies begrenzt den Betriebsspannungsbereich der Speicherzelle. Der reine Diodendurchbruch ist von der Länge des Transistors unabhängig.

Punchthrough

Der Punchthrough ist im Gegenteil zum Diodendurchbruch stark von der Kanallänge des Transistors abhängig. Punchtrough bezeichnet den Zustand, in dem sich die Raumladungszonen von Source und Drain berühren und so Ladungen direkt von Source zu Drain fließen können. Diese Betrachtungen werden ohne anliegende Gatespannung durchgeführt. Damit ist sofort klar, dass Punchthrough vor allem ein Problem kurzer Transistoren ist, [3],[25],[62].
Es werden zwei unterschiedliche Fälle von Punchthrough klassifiziert. Diese beiden Fälle sind in Abbildung 2.14 dargestellt.

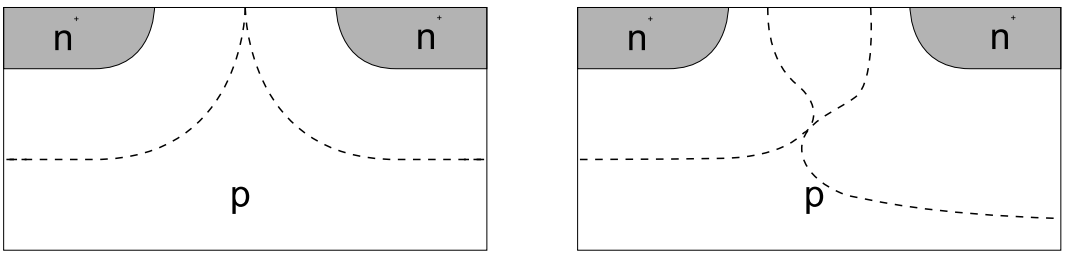


Abbildung 2.14: Veranschaulichung der Verarmungszonen von Source und Drain; links: Oberflächen-Punchthrough; rechts: Punchtrough in der Tiefe

Links ist der Punchthrough an der Oberfläche dargestellt, rechts ist der Punchthrough in der Tiefe zu sehen. Welcher von beiden Fällen zuerst auftritt, hängt maßgeblich von dem Dotierprofil der Wanne ab. Aus der Erklärung mit Hilfe der Raumladungszonen, wie in Abbildung 2.14 zu sehen ist, wird unmittelbar eine Methode zum Verringern des Punchens nahegelegt. Die p-Dotierung der Wanne muss erhöht werden, um die Ausdehnung der Raumladungszonen zu verkleinern. Dann tritt der Punchthrough erst bei höheren Spannungen auf.
Bei den in dieser Arbeit behandelten NROM-Zellen mit Kanallängen in der Größenordnung von nur $100 \ldots 200nm$ ist der Punchtrough ein kritischer Faktor. Beim Programmieren und Löschen wird mit hohen Spannungen an der Drain gearbeitet. Das Punchen kann vor allem die Löscheffizienz stark senken, oder ein Löschen sogar unmöglich machen, daher muss dieser Faktor stets beobachtet werden.

2.3.7 Subthreshold swing

Zur Beurteilung des Transistorverhaltens und seiner Güte wird häufig der subthreshold swing, S, verwendet. Er ist für den Bereich der schwachen Inversion definiert und gibt an, um wie viel die Spannung U_{GS} reduziert werden muss, damit der Drain-Strom, I_D, um eine Dekade gesenkt wird. Als Gleichung formuliert, bedeutet dies:

$$S = \frac{dU_{GS}}{d(\log I_{DS})} \tag{2.52}$$

Es ist klar, dass für einen guten Transistor kleine Werte für S gewünscht sind ($\sim 80mV/Dekade$).
Bei sehr kleinen Bauelementen wird dieser Kennwert durch Punchthrough verschlechtert.
Bei NROM-Speicherzellen ist auf Grund der lokal gespeicherten Ladungen eine Degradation des subthreshold swings zu beobachten, [59].
Eine einfache Formel, bei der der Swing unabhängig von U_{GS} formuliert ist, findet man bei Liu et al., [40]:

$$S = \phi_t \ln 10 \left(1 + \frac{\gamma}{2\sqrt{1,5 \cdot \phi_F + U_{SB}}}\right) \tag{2.53}$$

Darüber hinaus entwickelt Liu,[40], ein Modell, das in Übereinstimmung mit den Ergebnissen von Tsividis,[69], den Swing als Funktion der Gate-Source-Spannung angibt, wobei dieser ein absolutes Minimum besitzt.
Der Swing kann auch in Abhängigkeit der flächenbezogenen Oxidkapazität, C'_{ox}, und der flächenbezogenen Kapazität der Verarmungszone, C'_d, angegeben werden,[19]:

$$S = \left(1 + \frac{C'_d}{C'_{ox}}\right) U_{th} \cdot \ln(10) \tag{2.54}$$

Zur Berücksichtigung von Kurzkanaleffekten haben Godoy et al. diese Formel erweitert:

$$S = \left(\frac{1}{\lambda_S} + \frac{C'_d}{C'_{ox}}\right) U_{th} \cdot \ln(10) \tag{2.55}$$

wobei λ_S durch einen unhandlichen Term gegeben ist, der für kleine Drain-Source-Spannungen vereinfacht werden kann zu:

$$\lambda_S \simeq 1 - \frac{1}{\cosh\left(\frac{L}{2l_S}\right)} \tag{2.56}$$

L ist die Kanallänge und l_S wird als typische Länge bezeichnet, für die gilt:

$$l_S = \sqrt{\frac{\epsilon_{Si} \cdot d_{ox} \cdot x_d}{\epsilon_{ox}}} \tag{2.57}$$

Aus dieser Erweiterung folgt, dass der Swing nicht nur von der Wannendotierung, sondern auch von der Kanallänge abhängt. Für kurze effektive Kanallängen nehmen die Werte für den Swing deutlich zu.
In Abbildung 2.15 ist der Swing über die Wannendotierung aufgetragen. Man sieht, dass er zu sehr hohen Werten der Dotierung hin ansteigt, dies folgt bereits aus Gleichung (2.53), da die Dotierung dort im Zähler über das γ stärker gewichtet ist, als im Nenner über ϕ_F. Geht man zu sehr niedrigen Werten für die Dotierung, so zeigt der Transistor Kurzkanalverhalten. In diesem Bereich steigt der Wert für S auf Grund des Einflusses von λ_S. Es existiert also ein absolutes Minimum für den Swing.

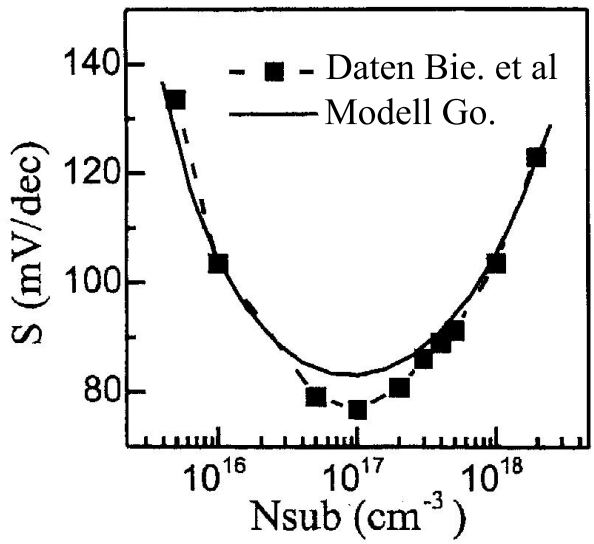


Abbildung 2.15: subthreshold swing vs. Wannendotierung (N_{sub}); Vergleich des Modells von Godoy,[19], mit den Daten von Biesemans et al.,[4], für einen MOS Transistor mit $d_{ox} = 7,5nm$, $L = 250nm$ und $U_{DS} = 0.1V$, [19].

2.4 NROM-Speicherzelle

Die Grundstruktur der NROM-Zelle basiert auf einem n-Kanal MOSFET, wie er zuvor besprochen wurde. Der entscheidende Unterschied ist, dass das Gateoxid durch einen Oxid-Nitrid-Oxid Stapel (ONO) ersetzt wird. Hierdurch wird die Möglichkeit geschaffen, Ladungen zu speichern. Diese werden lokal im Nitrid gespeichert. Vorschläge zu Speicherbausteinen, die auf diesem Prinzip beruhen, sind bereits in den 70er Jahren zu finden. So ist bereits in dem United States Patent 4,173,766 von 1979 ein Bauelement zu finden, das mit dem hier beschriebenen Mechanismus arbeitet, siehe Abbildung 2.16.

Die Beweglichkeit der Ladungsträger im Nitrid ist sehr gering, sie können sich nicht lateral über die Kanallänge verteilen. Hierdurch ergibt sich ein wesentlicher Vorteil zu anderen Technologien im NVM (non volotile memory) Bereich. In einer NROM-Zelle können zwei physikalisch getrennte Bits pro Zelle gespeichert werden, dies ist schematisch in Abbildung 2.17 dargestellt, [14]. Die gespeicherten Ladungen ändern die Einsatzspannungen des

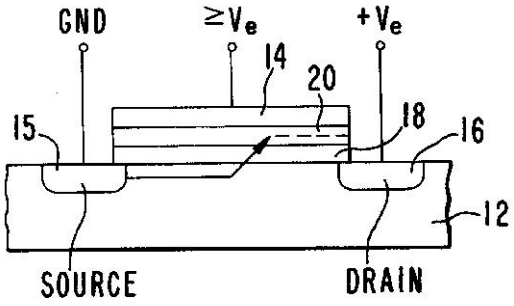


Abbildung 2.16: Frühe ONO-Speicherzelle, die Schichten mit den Zahlen 18, 20 und 14 bilden den ONO-Stapel; die anderen Zahlen sind hier nicht weiter von Bedeutung; [23].

Transistors. Dies wird zur Detektion des Speicherzustands ausgenutzt.
Hier ist eine herkömmliche NROM-Zelle abgebildet, das Wachsen eines Bitline-Oxids und die Biegung des ONO-Stapels am Rand sind nicht zwangsläufig bei einer NROM-Zelle vorhanden.

2.4.1 Schreiben, Lesen und Löschen

Wie aus Abbildung 2.17 leicht ersichtlich, ist die NROM-Zelle symmetrisch. Aus diesem Grund wird nur Bit 1 betrachtet, die Überlegungen gelten analog für Bit 2.

Schreiben

Für das Schreiben eines Bits wird zuerst eine hohe positive Spannung am Gate angelegt (z.B. $U_{GS} = 10V$). Durch diese Spannung wird ein Inversionskanal geöffnet. Nun wird an die Drain eine positive Spannung (z.B. $U_{DS} = 5V$) angelegt. Hierdurch werden die Elektronen auf ihrem Weg zur Drain immer stärker beschleunigt. Auf dem letzten Stück der Strecke fällt der größte Teil der Spannung ab, hier bekommen die Elektronen so viel Energie, dass sie zu heißen Elektronen (CHE = channel hot electrons) werden und durch Stoßionisation Bindungen aufschlagen können. Die erzeugten Elektronen können die Potentialbarriere des Oxids überwinden und im Nitrid gespeichert werden. So wird eine Elektronenverteilung

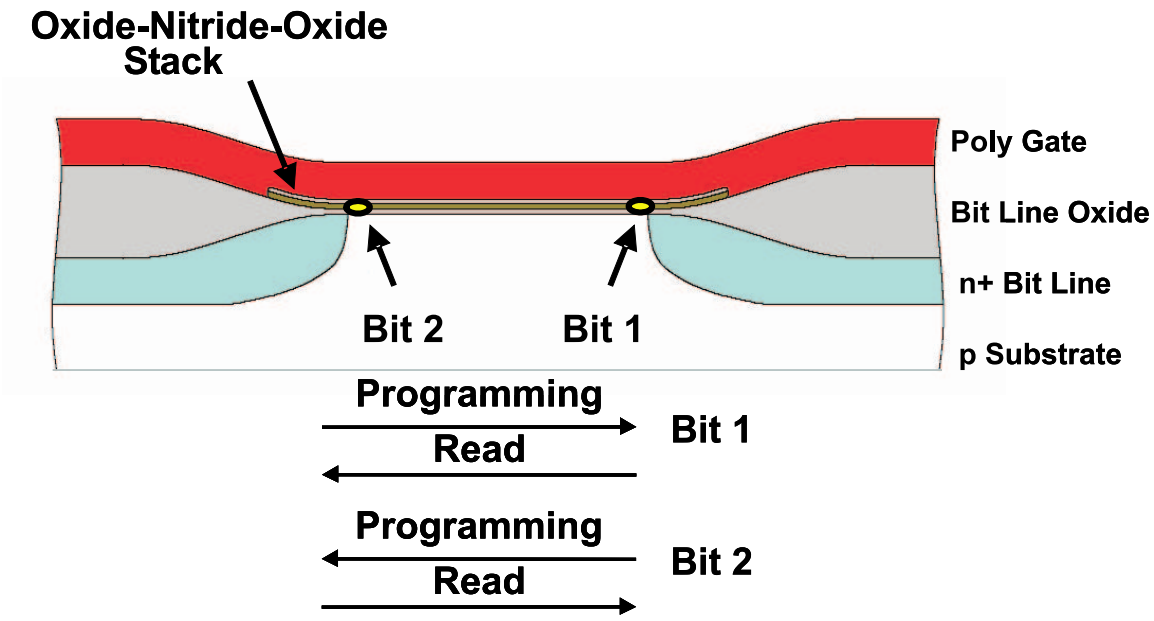


Abbildung 2.17: Grundstruktur einer NROM-Speicherzelle

(siehe 2.17: Bit 1) erzeugt.

Lesen

Für das Lesen des nun geschriebenen Bits 1 wird der Transistor in entgegengesetzter Richtung betrieben (Source und Drain werden vertauscht). Die angelegten Spannungen sind sehr viel geringer (z.B. $U_{GS} = 3 \ldots 4.5V$ und $U_{DS} = 1.6V$). Dies ist aus zwei Gründen sofort klar; zum einen will man die Einsatzspannung messen, und zum anderen will man natürlich beim Lesen von Bit 1 nicht gleichzeitig Bit 2 schreiben. Durch eine derartige Messung wird die Einsatzspannung bestimmt, welche sehr sensitiv auf die eingeschossene Elektronenladung (Bit 1) ist. Da man beide Ladungen getrennt detektieren will, wird - wie oben beschrieben - eine Source-Drain-Spannung von ca. $U_{DS} = 1.6V$ verwendet. Hierdurch wird die Verarmungszone an der neuen Drain vergrößert und bei der Messung der Transferkurve haben die gespeicherten Elektronen auf dieser Seite nur einen geringen Einfluss, [42]. Es ist leicht möglich, durch das Programmieren die Einsatzspannung um $2V$ zu erhöhen.

Löschen

Gelöscht wird eine Elektronenladung durch das Einschiessen von Löchern. Hierzu wird eine negative Spannung am Gate angelegt (z.B. $U_{GS} = -10V$), und zudem wird die Drainspannung erhöht ($U_{DS} = 4 \ldots 6V$). Dies führt drainseitig zu einer sehr starken Bandverbiegung, es werden Löcher generiert und beschleunigt. Gelangen diese in das Nitrid, so kompensieren sie die vorhandene Elekronenladung und somit wird die Einsatzspannung des Transistors wieder auf den Ausgangswert gesenkt. Das Bit ist wieder gelöscht. Ein guter räumlicher Überlapp der Löcher und Elektronen wird zumindest am Beginn des Löschens durch das lokale Feld der Elektronen begünstigt, es liegt eine Selbstjustierung vor.

Kapitel 3

Zellkonzepte und Modellbildung für NROM-Zellen

In diesem Kapitel werden zwei grundlegend unterschiedliche Konzepte für den Aufbau einer NROM-Speicherzelle vorgestellt. Zuerst wird ein „konventionelles“ Konzept vorgestellt. Konventionell bezieht sich in diesem Zusammenhang darauf, dass dieses Konzept bereits für Produkte, die auf dem Markt erhältlich sind, Verwendung findet und in einer Reihe von Publikationen behandelt wird, [5],[13],[27],[45].
Das zweite Konzept ist ein neuartiges Konzept, das von Dr. Josef Willer, [74], angestoßen wurde. Die Zellen werden seitlich durch STI begrenzt. Dies führt zu einem anderen Verhalten der Zellen als beim konventionellen Konzept.
Im Anschluss an die Vorstellung der beiden Zellkonzepte, folgt eine Abhandlung zur Modellbildung der NROM-Zelle, die nicht konzeptgebunden ist. Hier sollen Modelle behandelt werden, welche das Verständnis zu physikalischen Vorgängen in der NROM-Zelle vertiefen, respektive später für die Erklärung von Verlustmechanismen herangezogen werden können. Somit wird die Grundlage für eine besseres Verständnis der experimentellen Resultate im nachfolgenden Kapitel geschaffen.

3.1 Konventionelles Konzept (C-Konzept)

Im Verlaufe dieser Arbeit wird dieses Konzept mit CC (conventional concept) abgekürzt. Zuerst betrachten wir den prinzipiellen Aufbau eines Zellenfeldes nach diesem Konzept. Eine solche Architektur ist in Abbildung 3.1 zu sehen, sie ist aus der Literatur bekannt, [5], [13].

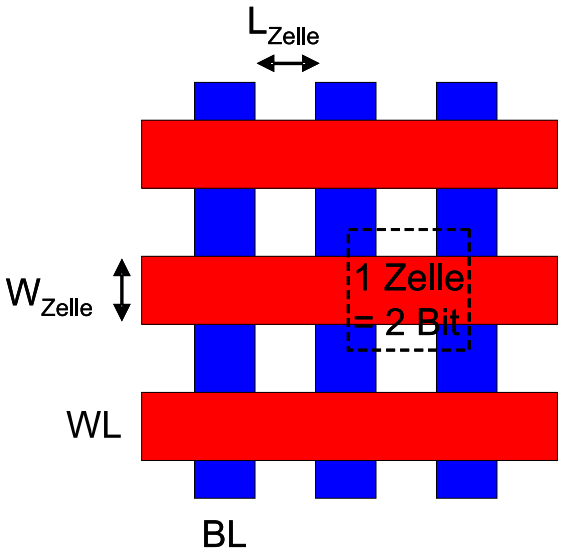


Abbildung 3.1: Struktur der Zellanordnung im C-Konzept; WL = Wortleitung, BL = Bitleitung, W_{Zelle} = gezeichnete Weite einer Zelle, L_{Zelle} = gezeichnete Länge einer Zelle

Die horizontalen, roten Bahnen stellen das Gate dar. Diese Bahnen werden zugleich Wortleitungen (WL) genannt. Senkrecht dazu verlaufen die blauen Bitleitungen (BL). Sie fungieren als Source- bzw. Drain-Gebiete (n^+-dotiert). Der Bereich zwischen den Bitleitungen ist p-dotiert. Diese Bitleitungen sind durchgehend, laufen also unter den Wortleitungen hindurch.

Dies ist nur möglich, weil die Arsen-Implantation der Bitleitungen vor dem Wachsen des Gatestapels durchgeführt wird. Hieraus resultiert sofort ein entscheidender Nachteil dieses Konzepts. Die Arsen-Implatation erfolgt sehr früh im Verlaufe des Herstellungsprozesses. Dies bedeutet, dass danach noch viele thermische Schritte mit hohen Temperaturen folgen.

Folglich sieht das Arsen ein hohes thermisches Budget und diffundiert stark aus. Dies ist ein limitierender Faktor für das Verkleinerungspotential der Zelle.
Die Lage einer NROM-Zelle ist in Abbildung 3.1 durch das gestrichelte Kästchen markiert, sie liegt horizontal. Der Abstand der Bitleitungen definiert die effektive Kanallänge der Zelle, ihre Breite ist durch die Breite der Wortleitung bestimmt. Abhängig davon, ob die rechte/linke BL einer Zelle als Source bzw. Drain verwendet wird, kann mit entsprechenden Spannungsbedingungen das physikalisch rechte/linke Bit programmiert, gelöscht bzw. gelesen werden (siehe Abschnitt 2.4).
Betrachtet man zwei übereinanderliegende Zellen, so fällt auf, dass diese nicht durch ein Oxid voneinander isoliert sind. Um eine gute Trennung bei geringen WL-Abständen zu erzielen und um das Punchen von BL zu BL (beim Programmieren und Löschen) in diesem Bereich zu vermeiden, wird eine Anti-Punch-Implantation zwischen den Zellen geschossen. Sie besteht zumeist aus Bor und/oder Indium. Indium lässt sich nicht sehr tief implantieren, diffundiert jedoch nicht so stark aus wie Bor. Die starke Ausdiffusion von Bor kann die Eigenschaften der Speicherzelle beeinflussen, wenn es von den Seiten her unter den Rand des Kanals diffundiert und somit hier lokal die Kanaldotierung verändert.
Die elektrische Verschaltung der Zellen in einem Zellenfeld, wie es in Abbildung 3.1 zu sehen ist, wird in Abbildung 3.2 veranschaulicht.

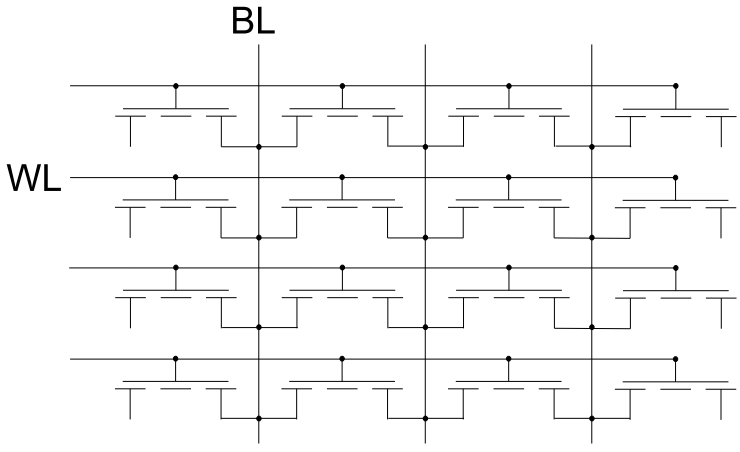


Abbildung 3.2: Elektrische Zellenfeldarchitektur für das C-Konzept

Für die NROM-Speicherzellen ist hier das übliche Transistorsymbol verwendet worden. Man sieht deutlich, dass die Zellen, die an einer Wortleitung angeschlossen sind, alle in Reihe liegen. Zudem werden hierzu durch die Bitleitungen viele Zellen parallel geschaltet. Eine solche Anordnung von Speicherzellen ist z.B. in einem Patent von Allan T. Mitchell und Bert R. Riemenschneider aus dem Jahre 1992 anschaulich dargestellt, siehe Abbildung 3.3.

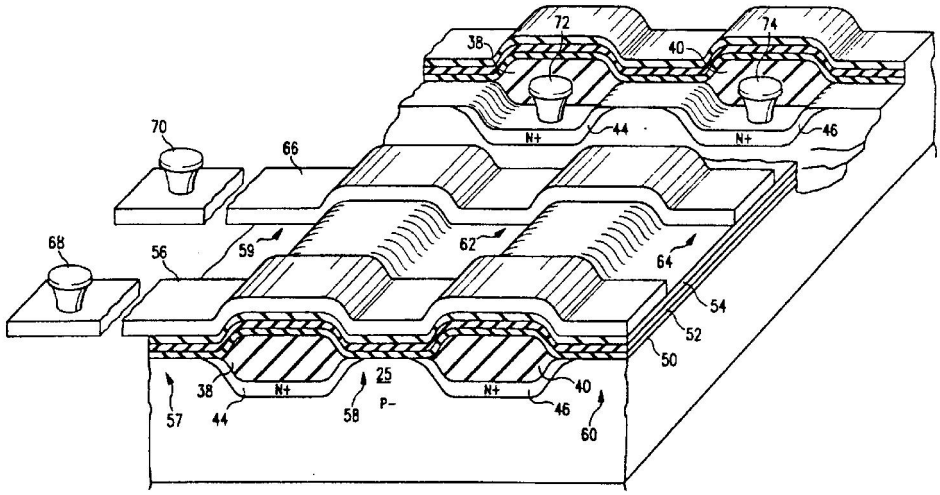


Abbildung 3.3: Patent auf Speicherzellen mit einer C-konzeptartigen Architektur, [70]

Die Abbildung dient nur der Veranschaulichung. Die Speicherzellen, die in diesem Patent behandelt werden, sind um mehr als einen Faktor 10 größer im Vergleich mit den in dieser Arbeit verwendeten Zellen.

Neben der Zellarchitektur lässt sich an der Abbildung 3.3 eine weitere Eigenschaft des C-Konzeptes erkennen. Dies ist das mit Nummer 40 gekennzeichnente Bitleitungsoxid, das zwangsläufig auftritt. Denn in diesem Konzept geschieht die Arsen-Implantation der Bitleitungen, wie oben erläutert, zu einem sehr frühen Zeitpunkt im Prozess. Dies hat nicht nur eine starke Ausdiffusion des Arsens zur Folge, sondern auch das Wachsen einer Oxidlinse (Bitleitungsoxid genannt) in diesem Bereich. Das durch die Arsen-Implantation geschädigte Silizium oxidiert sehr viel stärker, als das Silizium im Zwischenbereich. Eine

reale Situation ist in Abbildung 3.4 zu sehen.

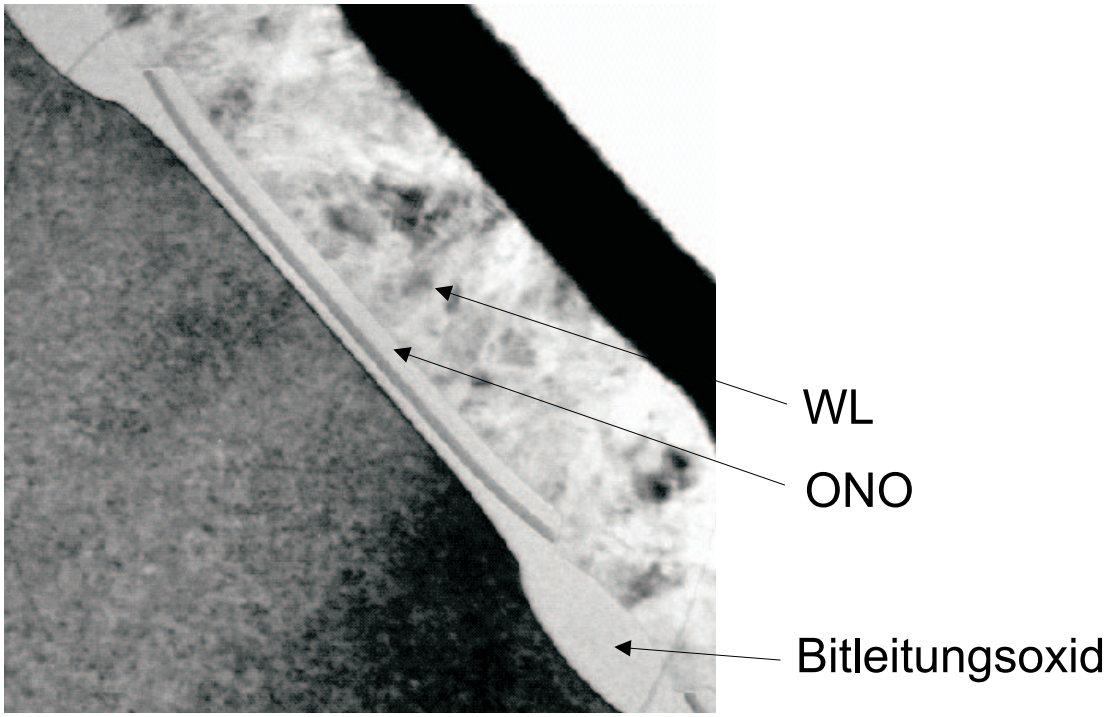


Abbildung 3.4: Längsschnitt (parallel der WL) durch eine NROM-Zelle, $0.17\mu m$ Technologie

Der hier gezeigte Schnitt verläuft in Längsrichtung durch die Zelle. Die Arsengebiete der Bitleitungen sind nicht andekoriert, sie liegen unter den Bitleitungsoxiden. Bei der hier dargestellten Zelle ist deutlich zu erkennen, dass der ONO-Stapel nicht über dem gesamten Bitleitungsoxid liegt. Zudem ist eine deutliche Verbiegung des ONO am Rand der Zelle zu sehen, die durch die Bitleitungslinse hervorgerufen wird.
Ein Schnitt quer durch die NROM-Zelle von Abbildung 3.4 ist in Bild 3.5 zu sehen.

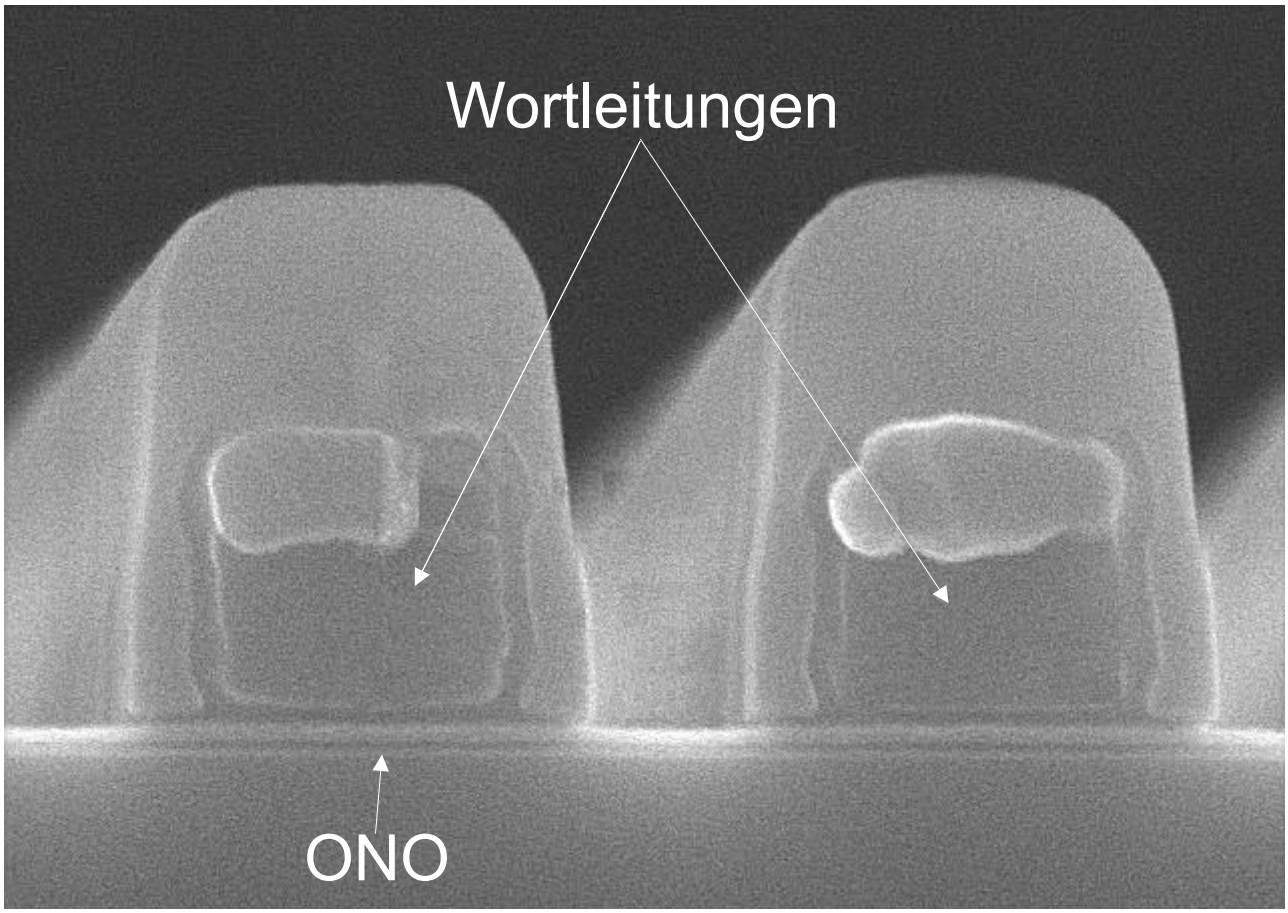


Abbildung 3.5: Querschnitt (senkrecht zur WL) durch eine NROM-Zelle, $0.17\mu m$ Technologie

3.2 STI-Konzept

Im Gegensatz zum C-Konzept, bei dem das gesamte Zellenfeld aus aktivem Gebiet, also aus Silizium, besteht, werden die Zellen im Zellenfeld des STI-Konzeptes seitlich durch STI (Shallow Trench Isolation) begrenzt. Dies ist grundsätzlich neu für die NROM-Technologie. Da die Isolationsgräben das wesentliche Merkmal dieses neuartigen Konzeptes sind, wird es hier als STI-Konzept bezeichnet. Diese grundsätzliche Änderung in der Architektur des Zellenfeldes hat wesentliche Auswirkungen auf die Eigenschaften der einzelnen Speicherzelle und auf den gesamten Prozessablauf.
Die Unterschiede zwischen C- und STI-Konzept werden analysiert und neue Möglichkeiten des STI-Konzepts werden aufgezeigt.
In Abbildung 3.6 ist die Architektur des Zellenfeldes nach STI-Konzept zu sehen.

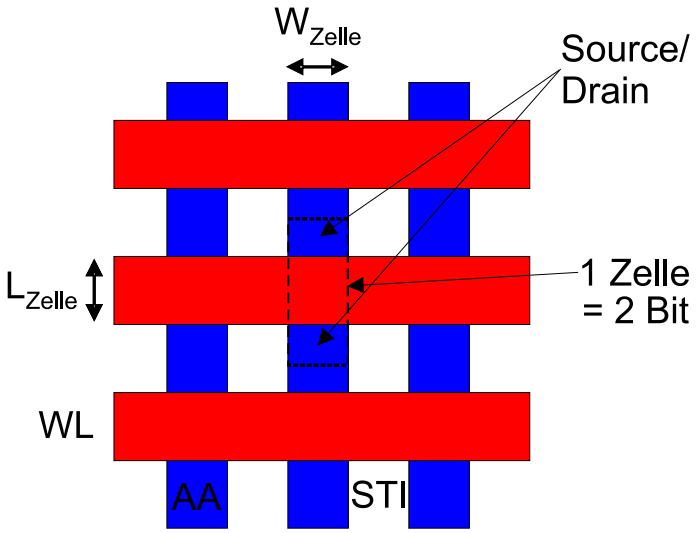


Abbildung 3.6: Struktur der Zellanordnung im STI-Konzept; WL = Wortleitung, AA = Active Area, STI = Shallow Trench Isolation, W_{Zelle} = gezeichnete Weite einer Zelle, L_{Zelle} = gezeichnete Länge einer Zelle

Wie vom C-Konzept bekannt, liegen auch hier die Wortleitungen in horizontaler Richtung. Jedoch besteht nicht mehr das gesamte Gebiet aus aktivem Bereich (Silizium), vielmehr wechseln sich in horizontaler Richtung Streifen mit aktivem Bereich (AA) mit

STI-Isolationnsstreifen (STI) ab. Sowohl STI, als auch AA gehen unter den Wortleitungen hindurch. Aus dieser Konfiguration ergibt sich unmittelbar, dass die Speicherzellen nicht mehr in horizontaler Richtung liegen, sondern in vertikaler Richtung. Somit wird durch die Breite der Wortleitung nicht mehr die gezeichnete Breite, sondern die gezeichnete Länge des Zellkanals bestimmt. Die AA-Breite ist nun die gezeichnete Kanalbreite. Source- und Drain-Gebiete, Zuordnung je nach Betriebsrichtung, sind ebenfalls in Abbildung 3.6 eingezeichnet.

Es fällt unmittelbar auf, dass Source- bzw. Drain-Gebiete isoliert liegen und keine zusammenhängenden Bitleitungen wie im C-Konzept existieren. Um jedoch zu einer realisierbaren Verschaltung zu gelangen, die das Betreiben aller Zellen ermöglicht, muss wiederum eine Verschaltung, wie sie in Abbildung 3.2 zu sehen ist, erzielt werden. Dies wird durch ein für NROM neues Kontaktschema erzielt, wie es in Abbildung 3.7 dargestellt ist.

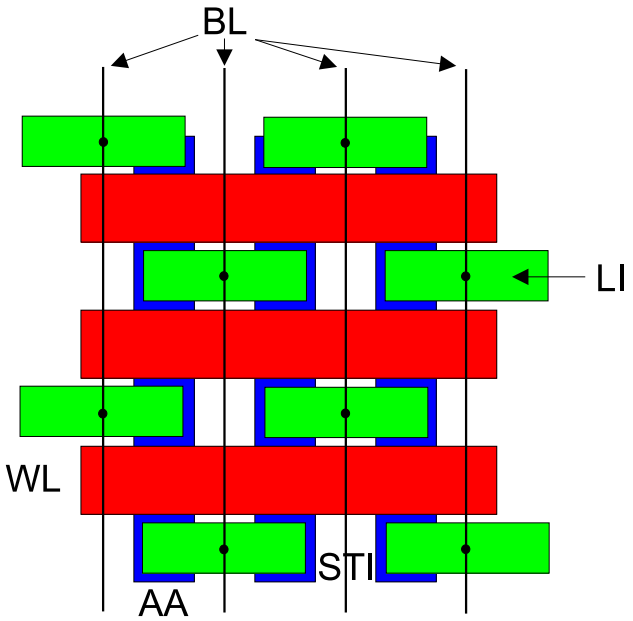


Abbildung 3.7: Struktur der Zellverschaltung im STI-Konzept; WL = Wortleitung, AA = Active Area, STI = Shallow Trench Isolation, BL = Bitleitung, LI = Local Interconnect

Durch lokale Kontakte (LI für Local Interconnect) werden jeweils zwei Source- bzw. Drain-

Gebiete elektrisch miteinander verbunden. Die Anordnung der LIs ist der Abbildung 3.7 zu entnehmen (hier grün dargestellt). Diese LIs werden nun durch Metallbahnen in einer höheren Ebene, wie ebenfalls in der Abbildung 3.7 dargestellt, miteinander verbunden. Trotz der physikalisch anderen Orientierung der Speicherzellen im STI-Konzept, ergibt sich durch dieses Kontaktschema wieder elektrisch die gleiche Verschaltung der Zellen wie im C-Konzept. Prozesstechnisch ist sowohl der geringe Abstand zwischen den LIs als auch die grosse Anzahl der LIs, die ohne Ausfälle funktionieren müssen, eine Herausforderung.
Nachdem der Aufbau des Zellenfeldes behandelt ist, sollen nun die Unterschiede für die einzelne Zelle im STI-Konzept gegenüber dem herkömmlichen Konzept weiter ausgeführt werden. Zur Veranschaulichung und als Ausgangspunkt für die weitere Diskussion dienen zuvor, analog zum C-Konzept, nun für das STI-Konzept zwei Darstellungen zueinander orthogonaler Schnitte durch die Zelle.
Abbildung 3.8 zeigt einen Schnitt entlang des Kanals. Neben dem Gate ist hier besonders deutlich der ONO-Stapel zu sehen. Die Source- bzw. Drain-Gebiete sind hier nicht zu sehen, da sie nicht andekoriert wurden. Sie sind rechts und links neben dem Gate implantiert. Durch thermische Schritte diffundieren sie bis zum Ende der Prozessierung aus. So gelangen sie ein Stück weit unter das Gate und sorgen für einen guten elektrischen Anschluss der Zelle.
Im Schnitt quer durch den Kanal, Abbildung 3.9, sind rechts und links die STI-Gräben zu sehen, durch die die Zelle seitlich begrenzt wird. Zudem ist deutlich ein Höhenunterschied zwischen Silizium und STI erkennbar.

Geringere Ausdiffusion der Source- bzw. Drain-Gebiete

Einen ganz wesentlichen Unterschied zum C-Konzept bildet die Implantation des Arsens für die Source- bzw. Drain-Gebiete. Zum einen wird für diese Implantation keine Maske benötigt, sie geschieht selbstjustiert durch das Vorhandensein der Gatestapel (WL), zum anderen wird sie deutlich später im Prozessverlauf geschossen als beim C-Konzept. Das Arsen wird erst nach der Erzeugung des ONO und des gesamten Gatestapels implantiert. Die

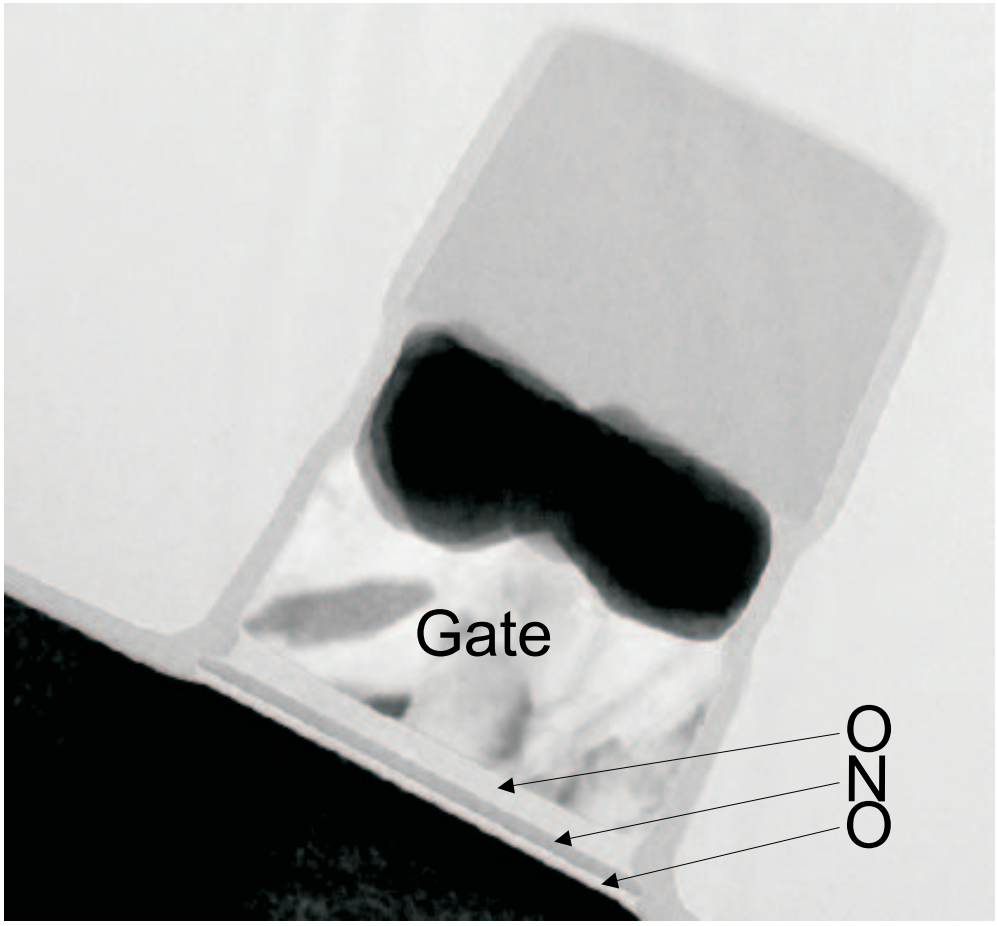


Abbildung 3.8: Längsschnitt (senkrecht zur WL) durch eine STI begrenzte NROM-Zelle

unmittelbare Konsequenz ist, dass dieses Arsen mit einem sehr viel geringeren thermischen Budget beaufschlagt werden kann, als dies bei der frühen Implantation beim herkömmlichen Konzept der Fall ist. Folglich ist die Ausdiffusion der Source- bzw. Drain-Gebiete sehr viel geringer, und somit kann bei gleicher effektiver Kanallänge die physikalische Länge der ganzen Zelle sehr viel keiner realisiert werden. Dies eröffnet erhebliches Verkleinerungspotential.

Die geringere Ausdiffusion der Source-/Drain-Gebiete hat noch einen weiteren Vorteil. Die Tiefe d_j der Arsengebiete ist geringer, hierdurch werden Kurzkanaleffekte reduziert. Nach Brews et al.,[6] , ist die minimale Kanallänge, bei der sich ein Transistor noch wie ein

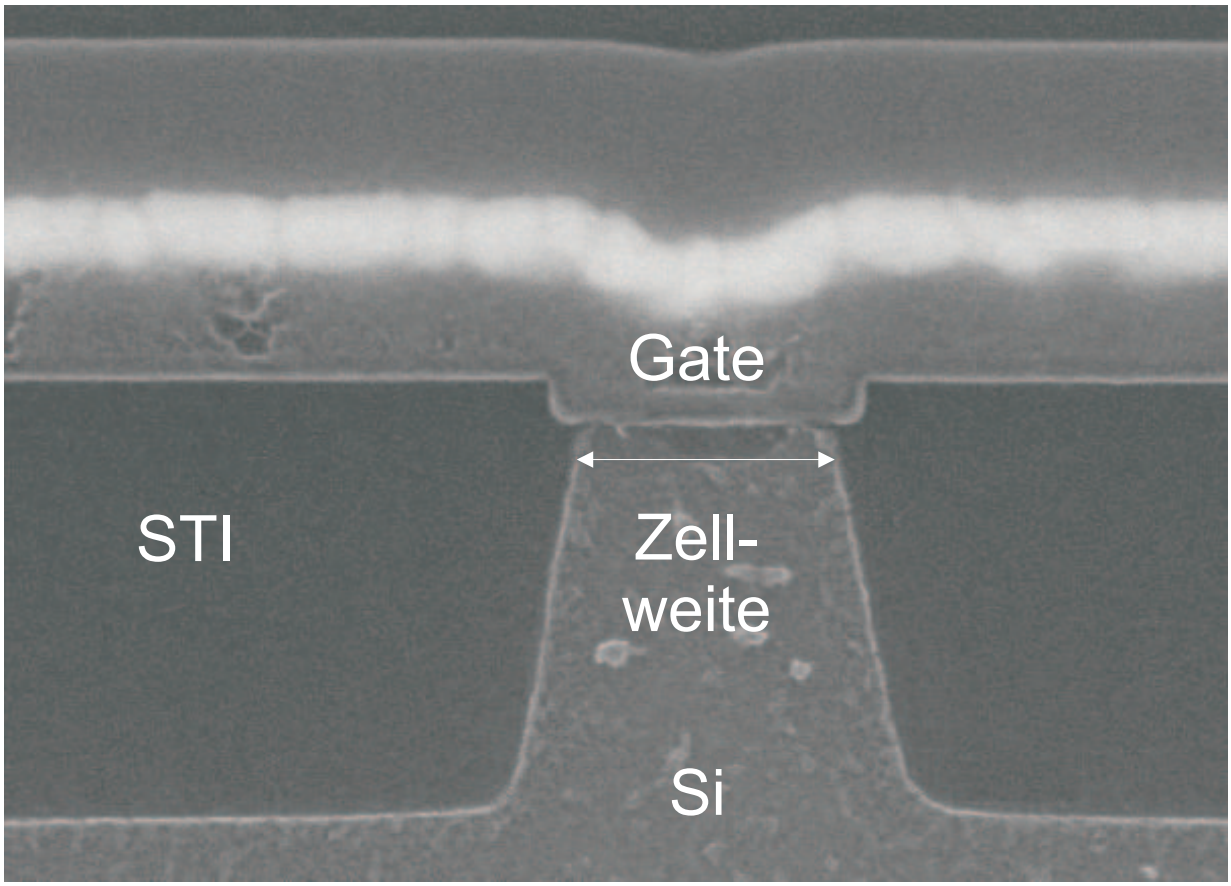


Abbildung 3.9: Querschnitt (parallel der WL) durch eine STI begrenzte NROM-Zelle

Langkanalbauelement verhält:

$$L_{min} \sim \left[d_j d_{ox} \left(d_{BS} + d_{BD}\right)^2\right]^{\frac{1}{3}} \tag{3.1}$$

wobei d_{BS} und d_{BD} die Weiten der Verarmungszonen von Source und Drain sind. Der Proportionalitätsfaktor hat einen empirisch ermittelten Wert von $8,8\mu m^{-1/3}$.

Pocket-Implantation

Pocket bedeutet in diesem Zusammenhang, dass eine lokale Implantation eines dreiwertigen Elements nahe des pn-Übergangs von Source bzw. Drain vorgenommen wird. Diese lokale Änderung der Wannendotierung bietet einen weiteren Freiheitsgrad bei der Optimierung

der Zelle. Beim STI-Konzept wird diese Pocket-Implantation möglich, da die Wortleitungen nicht mehr über die Source- bzw. Drain-Gebiete der Zellen hinweggehen. Dies ist ein wesentlicher Vorteil des STI-Konzeptes gegenüber dem C-Konzept.

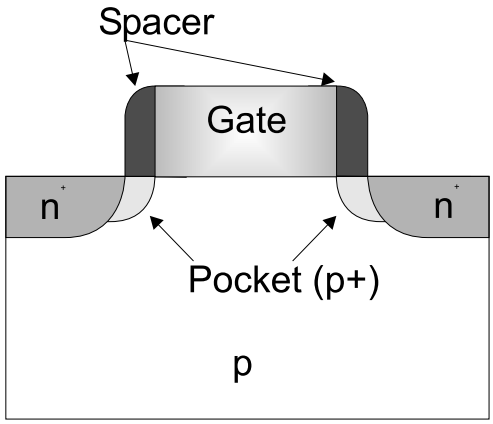


Abbildung 3.10: Schematische Darstellung zum Einsatz von Pocket und Spacer vor Ausdiffusion

Das Zusammenwirken von Pocket und Spacer ist schematisch in Abbildung 3.10 dargestellt.

Der Prozessablauf ist folgendermaßen:

- die p-Wanne wird implantiert (üblicherweise Bor)
- der Gatestapel wird erzeugt und strukturiert
- die Pocket wird implantiert (z.B. Bor, BF2 oder Indium)
- der Spacer wird erzeugt
- die Source- bzw. Drain-Gebiete werden implantiert (üblicherweise Arsen)
- durch thermische Schritte wird ein guter Anschluss der Zelle hergestellt

In Abbildung 3.10 ist die Situation direkt nach der Implantation der Source- bzw. Drain-Gebiete dargestellt, damit die Funktion des Spacers deutlicher wird. Man erkennt, dass in diesem Stadium die Zelle nicht adäquat angeschlossen ist. Unter einem guten Anschluss versteht man, dass die Source- bzw. Drain-Gebiete unter das Gate reichen, damit hierdurch der

gesamte Kanal gesteuert werden kann, und es nicht zu hohen Serienwiderständen kommt. Wie im Prozessablauf beschrieben, muss nach dem abgebildeten Zustand noch durch Ausdiffusion für einen ausreichenden Anschluss gesorgt werden. Hierbei diffundieren sowohl die Source- bzw. Drain-Gebiete, als auch die Pocket unter das Gate, so dass sich an ihrer relativen Lage zueinander nichts ändert.
Es sei noch an dieser Stelle erwähnt, dass auch die Möglichkeit besteht, statt eines dreiwertigen Elements ein fünfwertiges Element für die Pocket-Implantation zu verwenden. Dies würde einer LDD (lightly doped drain) entsprechen und bietet einen weiteren Freiheitsgrad. Damit steht ein Hebel zur Verfügung, um die injizierte Ladungsmenge zu steuern, wie es in Abschnitt 4.5.2 genauer beleuchtet wird.

ONO

Der ONO-Stapel ist das wesentliche Element einer NROM-Zelle. Daher lohnt sich eine kurze Betrachtung, was sich für diese Schichten beim STI-Konzept ändert. Betrachtet man Abbildung 3.8, so sieht man, dass der ONO planar ist. Schaut man zum Vergleich auf Abbildung 3.4, so stellt man fest, dass der ONO-Stapel durch das Bitleitungsoxid an den Seiten verbogen ist. Diese Verbiegung geschieht bei relativ niedrigen Temperaturen. Hierdurch kann es an den Grenzschichten zu Defekten kommen. Das kann die Beweglichkeit der im Nitrid gespeicherten Ladungsträger erhöhen und somit die Eigenschaften der Speicherzelle negativ beeinflussen. Hier bietet der planare ONO des STI-Konzepts einen Vorteil.

3.3 Tabellarischer Vergleich von C- und STI-Konzept

In der Tabelle 3.1 sind noch einmal die wesentlichen Unterschiede zwischen C- und STI-Konzept aufgelistet und bewertet.
Aus dieser Tabelle geht deutlich hervor, dass das STI-Konzept mehrere wesentliche Vorteile bietet. Vor allem die wesentlich günstigere Möglichkeit zur weiteren Miniaturisierung macht das Konzept für die Zukunft besonders interessant. Daher soll diese Arbeit auf Grund experimenteller Resultate die Funktionalität des neuartigen STI-Konzeptes untersuchen.

	C - Konzept	STI - Konzept
Schmalkanaleffekt	Kanalbreite ist stark von der Gatespannung abhängig	STI-begrenzt geringe Abhängigkeit von U_G, Borsegregation
Pocket-implantation	nicht möglich	möglich
S/D - Gebiete	grosse Tiefe, geringer Dotiergradient	geringe Tiefe, grosser Dotiergradient
ONO	im Randbereich verspannt	planares ONO
Verkleinerungs-potential	stark limitiert durch starke Ausdiffusion der S/D-Gebiete	hoch durch späte S/D-Implantation

Tabelle 3.1: Konzeptvergleich

3.4 Modellbildung für NROM-Speicherzellen

Um die komplexen Vorgänge in einer NROM-Zelle besser verstehen zu können, ist es hilfreich, mit anschaulichen Modellen zu arbeiten, die jedoch in der Lage sein müssen, möglichst alle experimentellen Befunde zu erklären.
In Kapitel 2.4 wurden bereits einige grundlegende Erläuterungen zum Funktionsprinzip der NROM-Zelle vorgestellt.
Für eine ideale Zelle kann man mit einem trivialen Modell arbeiten, das folgendermaßen aussieht:

- Programmieren: Es werden Elektronen in die Nitridschicht des ONO injiziert, diese verursachen eine Erhöhung der Einsatzspannung
- Löschen: Löcher werden injiziert, diese kompensieren die Elektronen und versetzen die Zelle wieder in den Ursprungszustand

Betrachtet man die Messungen an einer realen NROM-Zelle, so stellt man Ladungsverlust und andere Störungen fest. Nun gilt es, diese bestmöglich zu erklären. Naturgemäß ist der Ladungsverlust, hiermit ist in diesem Fall das Absinken der Einsatzspannung einer programmierten Zelle gemeint, für eine Speicherzelle der zentrale Punkt. Daher werden hier verschiedene Theorien für die Erklärung des Absinkens der Einsatzspannung vorgestellt und besprochen.
Zuvor wird die Situation beim Programmieren und Löschen noch anschaulich mit Hilfe des Bändermodells dargestellt. Darauf folgend wird zum besseren Verständnis noch ein Überblick über die Trapeigenschaften und die Ladungstransporteigenschaften von Siliziumnitrid gegeben.

3.4.1 Programmieren und Löschen im Bändermodell

Es werden vertikale Schnitte durch die NROM-Zelle nahe an der Drain, wo die Ladungsträgerinjektion stattfindet, dargestellt. Eine Übersicht über die Materialparameter, die für eine Darstellung im Bändermodell notwendig sind, ist in Tabelle 3.2 gegeben.

	Si	SiO_2	Si_3N_4
Bandlücke	$1,12eV$, [51]	$\sim 9eV$, [51]	$5,1eV$, [2]
Lage der Leitungsbandkante bezogen auf selbige von Si	–	$3,1eV$, [47]	$2,05eV$, [47]

Tabelle 3.2: Materialparameter für die Darstellung im Bändermodell

NROM Zellen werden mit heißen Elektronen programmiert. Folglich wird eine hohe Source-Drain-Spannung benötigt, um den Elektronen im Kanal genügend Energie zuzuführen, damit diese heiß werden. Zusätzlich wird eine hohe positive Spannung am Gate angelegt, z.B. 10V. Durch die Gate-Spannung werden einige der heißen Elektronen in vertikaler Richtung beschleunigt und werden so in der Nitridschicht des ONO gefangen. Die Spannung am Gate

ist somit deutlich höher als jene an der Drain. Nahe der Drain, wo die Elektroneninjektion stattfindet, ergibt sich energetisch die in Abbildung 3.11 dargestellte Situation.

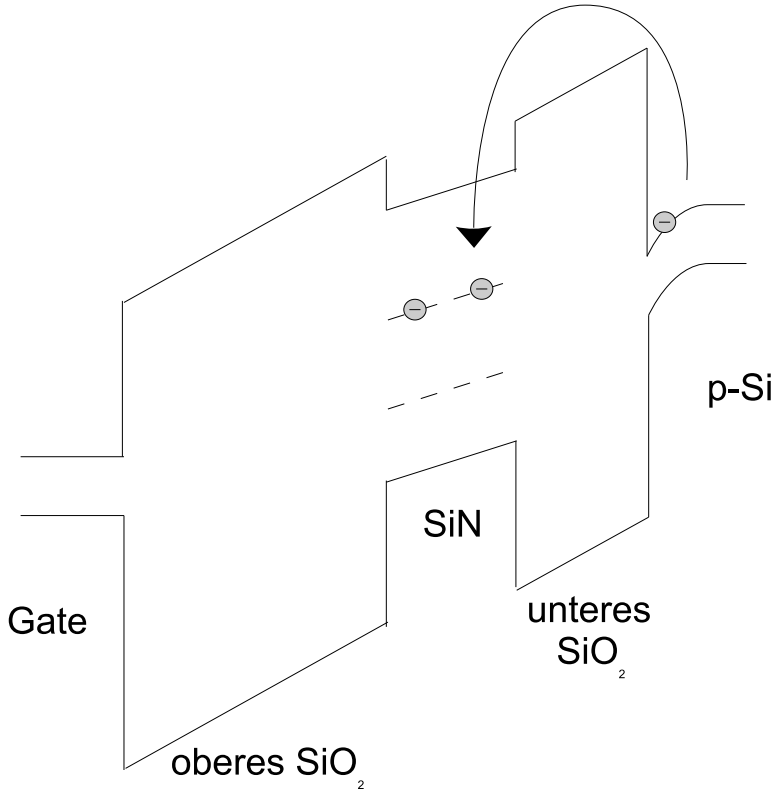


Abbildung 3.11: Schematische Darstellung des Programmierzustands im Bändermodell

Die relativen Dicken der dargestellten ONO-Schichten sind realistisch. Die untere Oxidschicht und die Nitridschicht haben eine vergleichbare Dicke, wohingegen die obere Oxidschicht ungefähr so dick ist, wie die beiden unteren Schichten zusammen.
Mit einer gewissen Wahrscheinlichkeit überwinden die heißen Elektronen die Potential-Barriere der unteren Oxidschicht und werden dann in den Traps der Nitridschicht gefangen. Somit steigt die Einsatzspannung der Zelle.
Beim Löschen werden Löcher in die Nitridschicht des ONO injiziert. Zu diesem Zweck wird eine stark negative Spannung am Gate angelegt ($\sim -10V$), sowie an der Drain eine positive Spannung, die deutlich höher ist im Vergleich mit der Spannung an der Source. So werden durch Band zu Band Generation Löcher erzeugt, die durch das laterale Feld

beschleunigt werden, bis sie genügend Energie besitzen, um mit Hilfe des vertikalen Felds in die Nitridschicht des ONO injiziert zu werden. Ein Schnitt am Ort der Injektion ist in Abbildung 3.12 zu sehen.

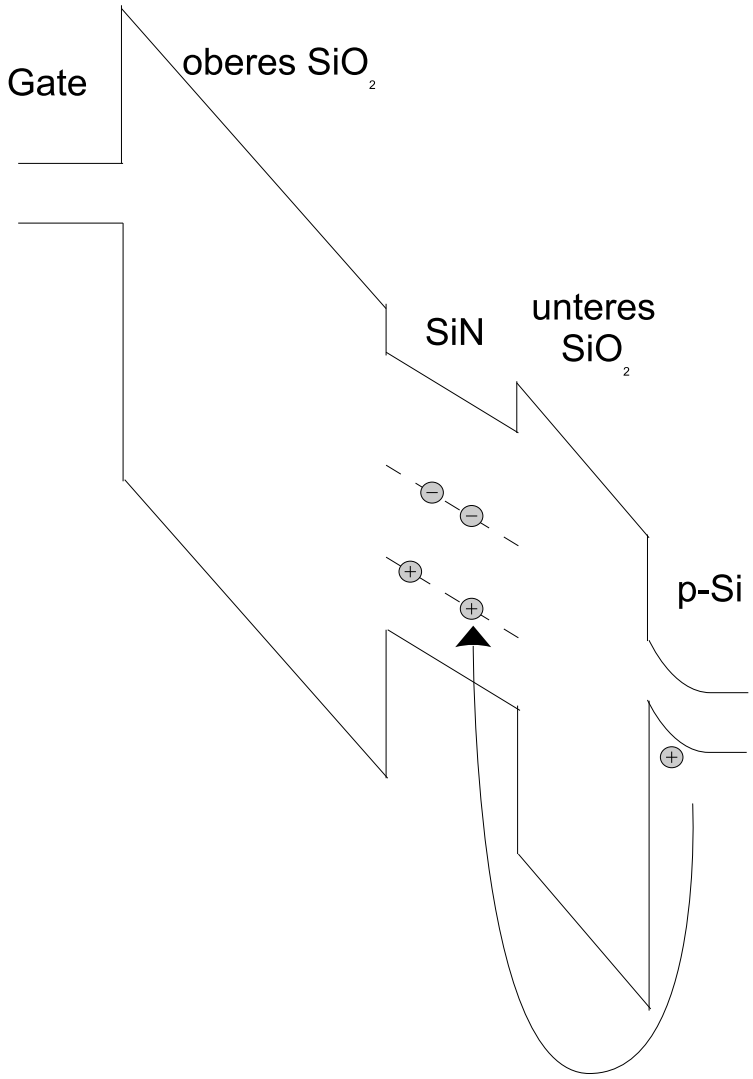


Abbildung 3.12: Schematische Darstellung des Löschzustands im Bändermodell

3.4.2 Traps und Ladungstransport in Siliziumnitrid

Traps und Ladungstransport in Siliziumnitrid werden in der Literatur vor allem im Kontext von SONOS und von MNOS Strukturen behandelt. Hier sollen nur einige, für diese Arbeit

wichtige Aspekte beleuchtet werden.

Trap-Eigenschaften

Kapoor, [32], berichtet von fünf wohldefinierten Trap-levels, die $2,5/2,76/3,03/3,36$ und $3,76eV$ unterhalb der Leitungsbandkante von Siliziumnitrid liegen. Dies ist schematisch in Abbildung 3.13 dargestellt.

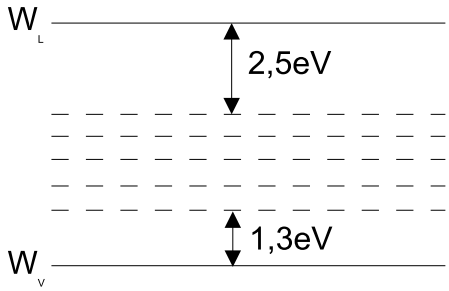


Abbildung 3.13: Fünf wohldefinierte Trap-levels in Siliziumnitrid nach Kapoor, [32]

Die Lokalisierung dieser Traps zeigt, dass es flachere Traps für Löcher als für Elektronen gibt.
Vergleicht man Werte für Dichte und Tiefe der Traps in der Literatur, so stößt man auf größere Diskrepanzen. Eine kurze Übersicht bietet Tabelle 3.3.
Es ist anzumerken, dass sich die letzten beiden Publikationen explizit mit NROM-Speicherzellen beschäftigen. Die Untersuchungen wurden an Zellen, die auf dem C-Konzept basieren, durchgeführt.
Bei Lusky et al. ist eine Auftragung der Traps für Elektronen und Löcher zu finden, sie ist in Abbildung 3.14 gezeigt.
Diese Werte wurden an NROM-Zellen der $0,35\mu m$ Technologie extrahiert. Sie zeigen deutlich, das die Verteilung der Löcher-Traps breiter und flacher ist als jene der Elektronen.
Weitere Abhandlungen, die sich mit Traps in Siliziumnitrid beschäftigen sind: [35], [47], [65], ...
Zudem wird von einer sehr hohen Trapdichte an der $Si_3N_4 - SiO_2$ (thermisches Oxid)

	Trap-Tiefe	Trap-Dichte	Quelle
Elektronen Löcher	$0,8 \ldots 0,9eV$ $0,8 \ldots 0,9eV$	$1 - 5 \times 10^{18} cm^{-3}$ $1 - 4 \times 10^{18} cm^{-3}$	[1]
Elektronen Löcher		$\sim 7 \times 10^{18} cm^{-3}$ $\sim 1,2 \times 10^{20} cm^{-3}$	[31]
Elektronen	median value of $\sim 2,12eV$; only $\sim 10\%$ shallower $1,8eV$		[44]
Löcher	$\sim 1,4eV$ (grosse Verteilungsbreite)		[41]

Tabelle 3.3: Trap-Eigenschaften von Siliziumnitrid

Grenzschicht auf Grund von überschüssigem Silizium berichtet, [21].

Detrappingmechanismen

Es gibt verschiedene Mechanismen, die zu einer Emission von in Trapstellen des ONO-Nitrids gefangenen Ladungsträgern führen können. Im Wesentlichen sind folgende Emissionsarten relevant:

- thermische Emission
- Poole-Frenkel-Effekt
- phononenunterstütztes Tunneln

Die zwei letzten Mechanismen werden üblicherweise bei Anliegen eines äußeren elektrischen Feldes betrachtet. Betrachten wir jedoch die Ladungshaltung in einer NROM-Zelle, so liegt während der Lagerung kein äußeres Feld an. Durch Programmieren und Löschen können sich aber Ladungspakete im ONO ansammeln. Die Coulomb-Kräfte zwischen diesen Paketen wirken wie ein angelegtes Feld.

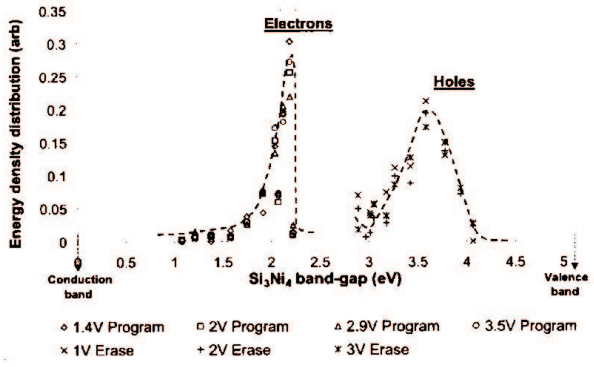


Abbildung 3.14: Trap-Verteilung in der Nitridschicht des ONO, [41]

• Thermische Emission:

Die thermische Emission lässt sich durch eine Emissionsrate der Form

$$\nu_T = \nu_0 \cdot \exp\left(-\frac{q\phi_r}{kT}\right) \tag{3.2}$$

beschreiben, [44], wobei ϕ_r die Energie der Fangstelle ist und ν_0 die 'attempt-to-escape' Frequenz ($\sim 10^{13} s^{-1}$). Lusky et al., [44], zeigen, dass sich mit Hilfe der thermischen Emission der Ladungsverlust einer einmal programmierten NROM-Zelle modellieren lässt. In diesem Fall wurden nur Elektronen injiziert, es können sich noch keine Ladungspakete von Elektronen und Löchern gebildet haben. Die sehr viel größeren Einsatzspannungsverluste bei der Lagerung gezykelter Zellen lassen sich nicht durch thermische Emission erklären, [41].

• Poole-Frenkel-Effekt:

Der Poole-Frenkel-Effekt beschreibt die Erhöhung der thermischen Emissionsrate von Ladungsträgern durch eine Verringerung der Potentialbarriere in einem elektrischen Feld durch ihr Coulombpotential. Der Effekt lässt sich beschreiben durch, [46]:

$$\nu_{PF} = \nu_T \cdot \exp\left(\frac{q}{kT}\sqrt{\frac{qE}{\pi\epsilon_0\epsilon_r}}\right) \tag{3.3}$$

Poole-Frenkel ist bei hohen Temperaturen dominant, $T > 200K$, [53].

Der Poole-Frenkel-Effekt und das anschließend besprochene phononenunterstützte Tun-

neln sind in Abbildung 3.15 grafisch dargestellt. Es liegt ein elektrisches Feld an. Ein solches kann sich im ONO, z.B. durch eine stark inhomogene Verteilung von Elektronen und Löchern ausbilden.

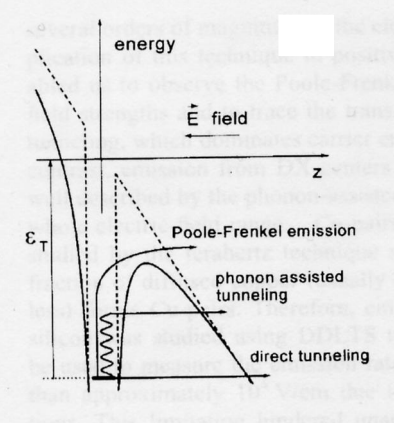


Abbildung 3.15: Potentailbarriere für die Emission von einer geladenen Fangstelle. Die Pfeile zeigen unterschiedliche Ionisationsprozesse, [17].

• Phononenunterstütztes Tunneln:
Phononenunterstütztes Tunneln gewinnt bei nierdrigeren Temperaturen als Poole-Frenkel an Bedeutung. Hier setzen Tunnelmechanismen ein. Dieses Modell besagt, dass Elektronen, die nicht genügend Energie besitzen, um die Poole-Frenkel-Barriere zu überwinden, durch den oberen, schmaleren Bereich tunneln. Somit nimmt dieser Transportmechanismus bei sehr hohen Feldstärken ebenfalls zu.

Betrachtet man die Situation bei NROM, so lassen sich die Feldstärken, die sich im ONO durch Ladungspakete aufbauen, nicht direkt bestimmen. Berücksichtigen wir den Temperaturbereich, in dem Flash-Produkte betrieben werden (ca.$-40°C$ bis $85°C$), so ist es naheliegend, dass der Ladungstransport durch Poole-Frenkel dominiert wird. Dies gilt insbesondere für die langzeitige Ladungshaltung. Während der Injektion von Ladungsträgern kann durchaus auch tunnelunterstützter Ladungstransport stattfinden.

Primär Ladungsträger in Siliziumnitrid

Wie bereits erwähnt, gibt es für SONOS und MNOS Strukturen eine Reihe von Untersuchungen zum Ladungstransport. Hier interessiert für die spätere Modellbildung im Besonderen, ob Elektronen oder Löcher primär für den Ladungstransport im Siliziumnitrid verantwortlich sind.
Die meisten Abhandlungen betrachten eine Situation bei positiver bzw. negativer Gate-Spannung. Es besteht weitgehend Einigkeit, dass bei negativer Gate-Spannung der Löcherstrom überwiegt, [64]. Suzuki et al., [63], sehen dies ebenfalls für positive Gate-Spannungen zutreffen, wohingegen Yau, [77], für beide Gate-Polaritäten die Elektronen als primäre Ladungsträger identifiziert hat.
Für die Informationshaltung in NROM-Speicherzellen sind jedoch vor allem die Ladungsbewegungen ohne ein wesentliches von außen anliegendes Feld von Bedeutung. Eine Zelle soll z.B. zehn Jahre lang eine eingeprägte Information halten, in diesem Zeitraum wird die anliegende Gate-Spannung zumeist Null sein. Es können sich allerdings im ONO Felder zwischen eingeschossenen Elektronen und Löchern aufbauen. Die entscheidende Frage ist nun, welche Spezies von Ladungsträgern sich leichter bewegen kann.
Arnett und Weinberg haben Löcher als die primären Ladungsträger, die zum Stromfluß beitragen, bestimmt, [2]. Dies wird von einer Untersuchung von Liou und Chen untermauert, [39]. Beide haben dies aus einer Leckstrommessung an einem $25nm$ dicken Siliziumnitrid Sample, dessen Ergebnis in Abbildung 3.16 zu sehen ist, hergeleitet.

Aus der Steigung der Kurven wurde die Höhe der Potentialbarriere für beide Gate-Polaritäten zu $1,1eV$ berechnet. Daraus wird gefolgert, dass es sich um das niedrigste Trap-Niveau aus Abbildung 3.13 handelt. Somit kommen sie zu dem Schluss, dass die Löcherleitung in Siliziumnitrid dominiert.
Die in Abschnitt 4.9 vorgestellten Versuche, bestätigen eine deutlich höhere Beweglichkeit der Löcher gegenüber den Elektronen.

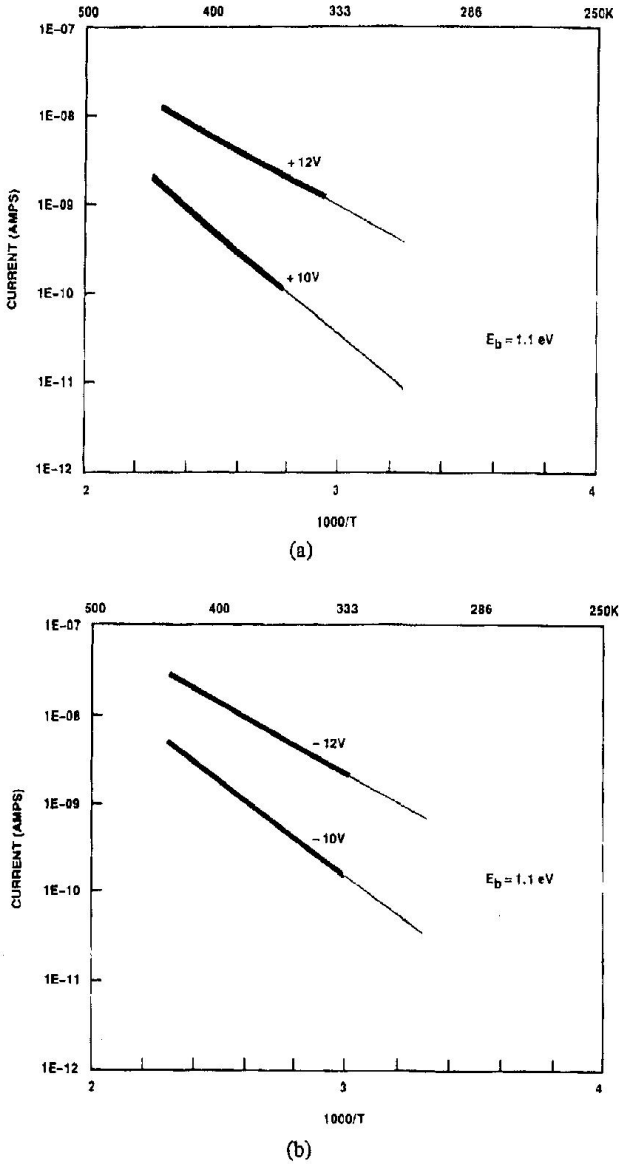


Abbildung 3.16: Temperaturabhängigkeit des Stromes durch eine $25nm$ dicke Siliziumnitrid Schicht: (a) mit positiver Gate-Spannung, (b) mit negativer Gate-Spannung. Aus den Steigungen der Kurven wird für beide Gate-Polaritäten eine Potentialbarrierenhöhe von $1,1eV$ berechnet, [39]

.

3.4.3 Ladungsverlust durch thermische Emission von Ladungsträgern

In einer Veröffentlichung von 2002 haben Lusky et al.,[44], die thermische Emission von Elektronen als primären Verlustmechanismus von NROM-Zellen identifiziert. Hierzu wurde die energetische Tiefe der Traps für Elektronen im Nitrid bestimmt und von da aus die thermische Emission in Abhängigkeit von Zeit und Temperatur bestimmt. Die Besetzung der Traps als Funktion der Energie ist in Abbildung 3.17 dargestellt.

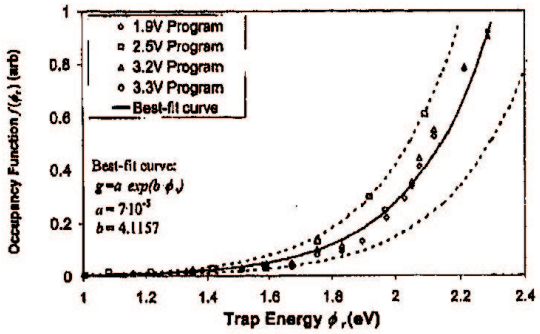


Abbildung 3.17: Die Besetzungsfunktion $f(\phi_r)$ ist für unterschiedliche Programmierfenster aufgetragen (1.9, 2.5, 3.2 und 3.3 V). Die gestrichelten Linien repräsentieren die maximalen Abweichungen der experimentellen Ergebnisse von der errechneten Kurve, [44].

Diese Grafik zeigt, dass die Fangstellen überwiegend sehr tief liegen und somit sehr gut für die Ladungsspeicherung geeignet sind. Darauf beruht auch der geringe Ladungsverlust, von dem in dieser Veröffentlichung berichtet wird.
Diese Ergebnisse beruhen auf Zellen mit einer effektiven Weite von $0,35\mu m$ und einer effektiven Länge von $0,32\mu m$. Lusky verwendet also eine wesentlich ältere Technologie mit sehr viel grösseren Zellabmessungen. Es ist wichtig darauf hinzuweisen, dass diese Messungen nicht an gezykelten Zellen durchgeführt wurden. An gezykelten Zellen werden höhere Verluste gemessen, [41]. Somit wurde nicht gezeigt, dass die thermische Emission auch für gezykelte Zellen den primären Verlustmechanismus darstellt.

Die Erkenntnisse bezüglich der Traptiefen im Nitrid sind jedoch übertragbar, da sich der Aufbau des ONO-Stapels nicht ändert. Es muss folglich noch weitere Mechanismen geben, die zu Ladungsverlust führen.

3.4.4 Ladungsverlust in vertikaler Richtung

Bei der Suche nach Ursachen für den Ladungsverlust in NROM-Zellen liegt die Überlegung nahe, dass Ladungsträger durch das top- bzw. bottom-Oxid entweichen. Dies ist naheliegend, zumal aus einer Vielzahl von Literaturquellen bekannt ist, das Oxide durch heiße Ladungsträger geschädigt werden, [30],[55],[18],[72].
Durch die Schädigung des Oxids braucht ein Ladungsträger, der das Nitrid verlässt, nicht mehr notwendig so viel Energie zu besitzen, dass er die Potentialbarriere des Oxids überwinden kann. Wenn genügend Traps durch die Schädigung der Oxidschicht entstanden sind, so kann es zu 'trap assisted tunneling' kommen, [12], [67]. Dies bedeutet, dass Ladungsträger auf einer Kette von Traps von Trap zu Trap tunneln und so aus der Nitridschicht durch ein geschädigtes Oxid entweichen können. Zudem können auch Elektronen beim Programmieren in neu entstandenen flachen Traps gefangen werden, welche sie später einfacher wieder verlassen können, [10].
In einer Publikation von Tsai et al. wird der Ladungsträgerverlust von Elektronen durch das bottom-Oxid im programmierten Zustand als Hauptverlustmechanismus identifiziert, [68]. Leider sind die Dimensionen der gemessenen Zellen nicht spezifiziert, aus der Marktlage ist jedoch naheliegend, dass es sich um Zellen handelt, die auf einer älteren Technologie beruhen, als die hier behandelten STI-begrenzten Zellen.

3.4.5 Ladungsverlust durch laterale Bewegung von Löchern

Bei diesem Modell wird angenommen, dass die Ladungsträger, die im Nitrid gefangen sind, dieses nicht wieder verlassen können. Es kann folglich nur einen Ladungsträgerverlust durch Rekombination von Elektronen mit Löchern geben. Hier ist es wichtig, dass Ladungsträgerverlust nicht mit Ladungsverlust gleichbedeutend ist. Der Begriff Ladungsverlust, wie er

in diesem gesamten Kapitel verwendet wird, besagt, dass das zu Anfang durch Programmieren erzielte ΔU_{th} kleiner wird. Dies ist nicht gleichbedeutend mit einem physikalischen Verlust von Ladungsträgern aus der Nitridschicht.

Die Ladungsträger können sich jedoch lateral in der Nitridschicht des ONO-Stapels bewegen. Dies kann zu einer Veränderung der Einsatzspannung führen, [57], [43], [41]. Zur weiteren Vereinfachung wird angenommen, dass sich nur die Löcher bewegen. Dies folgt aus den Betrachtungen in Abschnitt 3.4.2. Die Löcher sind einfacher thermisch zu aktivieren als die Elektronen. Ihre Bindungsenergie ist nicht so hoch, wie die der Elektronen, dies führt zu einer höheren Beweglichkeit der Löcher. Somit tragen sie wesentlich stärker zur Ladungsträgerumverteilung in der Nitrid-Schicht des ONO-Stapels bei. Die zur Ladungsträgerbeweglichkeit durchgeführten Experimente aus Abschnitt 4.9 untermauern die Aussage, dass die Beweglichkeit der Löcher wesentlich größer ist, als jene der Elektronen.

Zur genauen Definition des behandelten Sachverhalts ist es wichtig zu unterscheiden, ob der Ladungsverlust einer Zelle nach dem ersten Programmieren, dem zehnten, oder dem zehntausendsten Programmieren betrachtet wird. Der Verlust nimmt mit der Anzahl der Zyklen, mit der eine Zelle beaufschlagt wurde, zu. Daher ist der kritische Fall nach einer großen Anzahl von Zyklen der entscheidende. Aus diesem Grund basieren die folgenden Aussagen auf der Annahme, dass die Speicherzelle zuvor gezykelt wurde (einige 10 Zyklen reichen bereits aus, um einen deutlichen Unterschied zu einer nicht gezykelten Zelle zu beobachten). Bei ungezykelten Zellen sind Ergebnisse ähnlich zu jenen in Abschnitt 3.4.3 zu messen. Es ist zu folgern, dass für ungezykelte Zellen auch bei STI-begrenzten Zellen die thermische Emission dominiert.

Wie es durch eine laterale Bewegung der Löcher zu einer Absenkung der einprogrammierten Einsatzspannungsdifferenz kommen kann, soll anhand der Zeichnungen in Abbildung 3.18 erläutert werden.

Beim Programmieren werden Elektronen in die Nitridschicht des ONO injiziert. Die eingeschossene Elektronenverteilung ist in A) blau dargestellt. Nehmen wir die Unbeweglichkeit der Elektronen an und lassen die weiter oben besprochenen Verlustmechanismen außer Acht, so ändert sich an dieser Elektronenverteilung nach dem ersten Programmieren nichts.

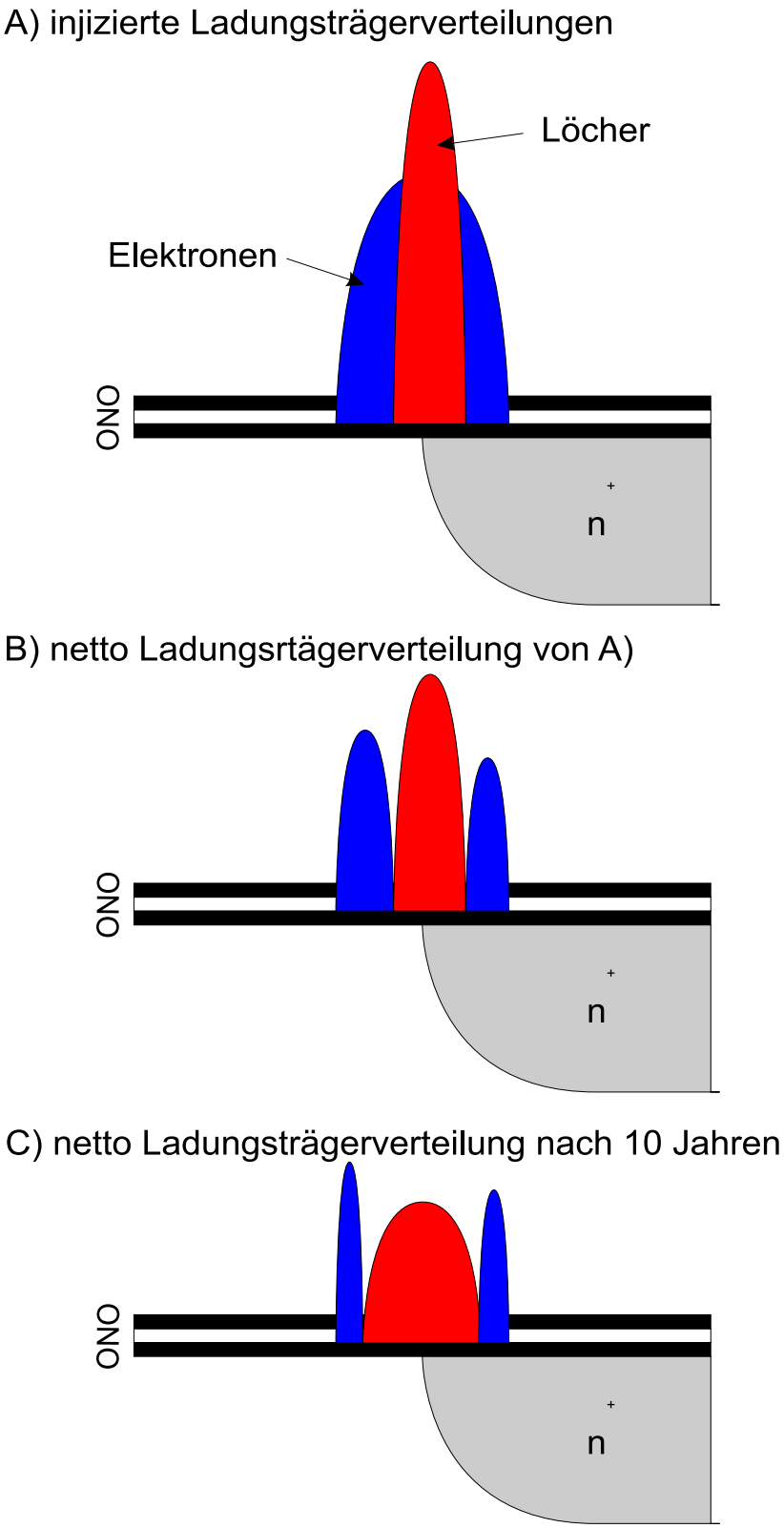


Abbildung 3.18: Schematische Darstellung der Ladungsträgerverteilungen im ONO

Diese Annahme ist durch Messungen insoweit gedeckt, dass der Verlust einer einmal programmierten Zelle nach einer Lagerung von einer Stunde bei $200°C$ nur ca. $50mV$ beträgt. Das ist eine Größenordnung unter Messungen an gezykelten Zellen.
Beim Löschen werden nun Löcher in die Nitridschicht injiziert. Dies geschieht so lange, bis die Elektronenladung elektrisch kompensiert ist, bis also wieder die ursprüngliche Einsatzpannung der Zelle hergestellt ist. Das bedeutet nicht, dass so viele Löcher, wie zuvor Elektronen injiziert werden, mit diesen rekombinieren und der physikalische Ausgangszustand wiederhergestellt ist. Der Einschuss der Löcher ist ein selbstjustierender Vorgang. Durch das lokale Feld der Elektronen im Nitrid werden die Löcher besonders in diese Region gezogen. Wir nehmen also an, dass die Verteilung der Löcher schmaler ist, als jene der Elektronen. Dies ist in Teil A) von Abbildung 3.18 durch die rote Verteilung dargestellt. In Teil A) sind die injizierten Verteilungen von Elektronen und Löchern übereinander abgebildet.
In Teil B) ist die Nettoverteilung abgebildet. Es bleiben Verteilungen von Elektronen und Löchern übrig. In dem Bereich, in dem sich die Verteilungen in A) überschnitten haben, ist es für die Zelle nicht von Bedeutung, ob die Ladungsträger rekombinieren, oder ob sie sich nur gegenseitig völlig nach außen neutralisieren, zumal dies nicht einfach messtechnisch unterscheidbar ist.
Nehmen wir also an, dass sich eine Situation, wie in B) ergibt. Es ist klar, dass für die genaue Lage der Verteilungen zueinander die Lagen der Injektionsmaxima von Elektronen und Löchern von entscheidender Bedeutung sind. Bei einer solchen Art der Ladungsverteilungen im ONO bestehen Dipolkräfte zwischen Elektronen und Löchern. Diese Kräfte verursachen Ladungsträgerbewegungen im ONO.
Legt man eine derartige Ladungsträgerverteilung in der Nitridschicht und deren zeitliche Veränderung zu Grunde, so lassen sich viele experimentelle Ergebnisse anschaulich und leicht erklären. So lässt sich ein Absinken der Einsatzspannung einer programmierten Zelle durch die Diffusion von Löchern, die über dem Source- bzw. Drain-Gebiet gespeichert sind, hin zum Kanal erklären. Die Diffusion wird durch das Dipolmoment zwischen den Löchern und der Elektronenverteilung links von selbigen begünstigt. Zugleich können Löcher auch

nach rechts wandern, dies wird jedoch nur in extrem geringen Maße über die Einsatzspannung erfasst. Die Ausdiffusion der Löcherverteilung kann sogar zu einer Absenkung der Einsatzspannung führen, selbst wenn ihr Schwerpunkt nach rechts wandert, da mit der Einsatzspannung nur die Ladungen erfasst werden, die über dem Kanalbereich gespeichert sind.

Betrachtet man die Dipolkräfte als den treibenden Mechanismus für die Ladungsumverteilung und für die Änderung der Einsatzspannung, so lassen sich auch Phänomene auf einem großen Bereich der Zeitskala erklären. In einem sehr kurzen Zeitabschnitt, direkt nach dem Einschuss neuer Ladungsträger, ergibt sich ein schnelles Abfallen der Einsatzspannung, danach wird der Prozess immer langsamer. Das lässt sich im Experiment beobachten. Das Absinken der Einsatzspannung wird, wie zu Beginn dieses Kapitels erörtert, als Ladungsverlust bezeichnet. Dieser ist für die Informationsspeicherung im Produkt von besonderer Bedeutung ist.

Durch den Ladungsverlust ergibt sich nach z.B. zehn Jahren Lagerung der Zelle oder nach einer äquivalenten Heizzeit ein Zustand, wie er in Teil C) der Abbildung 3.18 dargestellt ist. Wie erläutert, werden nur die Löcher als bewegliche Ladungsträger betrachtet, ihre Verteilung verbreitert sich und zieht in diesem Fall eine Absenkung der Einsatzspannung nach sich. Wird dieses Experiment nach einer größeren Anzahl von Zyklen ausgeführt, so ist auch der gemessene Verlust größer, da sich größere Reservoirs von Ladungsträgern gebildet haben.

Nimmt man dieses Modell zur Erläuterung der Informationshaltung (Konservierung eines programmierten V_{th} Zustands), so muss man erkennen, dass es keinen offensichtlichen, besten Weg der Zelloptimierung gibt.

Selbst bei ausschließlicher Betrachtung des Langzeitverlustes in einer gezykelten Zelle gestaltet sich eine Optimierung schwer. Um die Zelldimensionen physikalisch kleinst möglich zu machen, sind schmale Ladungsträgerverteilungen erstrebenswert. Dadurch werden allerdings auch die wechselwirkenden Kräfte größer und somit der Ladungsverlust. Zudem muss man sich bemühen, die Verteilungen von Elektronen und Löchern möglichst gut zur Deckung zu bringen, damit sich nicht unnötig große Ladungsmengen im ONO ansammeln.

Nimmt man die unvermeidbaren Schwankungen bei einem großen Ensemble von Zellen mit in Betracht, so muss es Ziel sein, eine robuste Zelle zu bekommen. Es hilft also nicht, eine gute Deckung mit sehr schmalen Verteilungen an einer Einzelzelle zu erreichen, wenn dafür das Zellensemble sehr sensibel auf Schwankungen reagiert.

Begründung für die Anordnung der Ladungspakete

Wir haben gesehen, dass sich durch eine laterale Bewegung von Löchern der Einsatzspannungsverlust einer programmierten NROM-Zelle erklären lässt. In Abbildung 3.18 werden zur Veranschaulichung Ladungspakete und deren qualitative Entwicklung dargestellt. Es stellt sich die Frage, warum muss die Lage dieser Pakete zueinander so und nicht anders sein.

Zuerst wird der Bereich über dem Kanal betrachtet. Dieser ist für die Einsatzspannung der NROM-Zelle maßgeblich. Die Programmierbarkeit von NROM bedeutet, dass eine Elektronenverteilung über dem Kanalbereich injiziert wird. Einen wichtigen Hinweis, dass die Löschverteilung nicht deckungsgleich mit der Programmierverteilung sein kann, liefert ein Blick auf das Verhalten einer Zelle beim Zykeln, siehe Abbildung 4.18. Die detaillierte Beschreibung des Experiments ist in Abschnitt 4.7 erläutert. Es wird nur eine Seite der Zelle gezykelt, dennoch steigt mit zunehmender Zyklenanzahl die Einsatzspannung des benachbarten Bits deutlich an. Dies bedeutet, dass sich Elektronen kanalseitig ansammeln. In der Raumvorstellung von Abbildung 3.18 liegt die Ansammlung auf der linken Seite. Diese Elektronen verursachen durch Nebensprechen eine Erhöhung der Einsatzspannung der anderen Seite der NROM-Zelle. Würden sich zum Kanal hin Löcher ansammeln, müsste die Einsatzspannung des benachbarten Bits sinken. Wären die Verteilungen deckungsgleich, so würde man keine Einsatzspannungsänderung des benachbarten Bits beobachten.

Zudem muss über dem Kanalbereich zumindest noch ein Teil einer Löcherverteilung liegen. Wäre dies nicht der Fall, könnte die Zelle nicht gelöscht werden. Betrachten wir die untere Darstellung der Abbildung 4.18, so sehen wir, dass die zum Löschen notwendige Drain-Spannung mit steigender Zykelzahl ebenfalls ansteigt. Dies bedeutet, dass immer

mehr Löcher injiziert werden. Es sammeln sich folglich Löcher über dem n^+-Gebiet an, wo sie keine Auswirkung auf die Einsatzspannung haben.
Die Existenz eines weiteren Elektronenpakets rechts von der Löcherverteilung über dem n^+-Gebiet lässt sich nicht über Messungen der Einsatzspannung zeigen. Monte Carlo Simulationen von Ingrosso et al., [27], liefern jedoch deutliche Hinweise auf eine Elektronenladung rechtsseitig der Löcher.
Die qualitative Anordnung der Ladungspakete zueinander dient der Erklärung des Einsatzspannungsverlusts bei NROM und ist zudem durch experimentelle Ergebnisse untermauert.

3.4.6 Zwei-Transistor-Modell für eine programmierte NROM-Zelle

Ausgangspunkt für die Entwicklung eines Modells zur Beschreibung einer programmierten NROM-Zelle ist ein Experiment, das weiteren Aufschluss über die Ladungsverteilung im ONO gibt. Das hier vorgestellte Zwei-Transistor-Modell bietet eine Möglichkeit, den Programmiervorgang von NROM-Zellen durch eine Verbreiterung der injizierten Elektronenverteilung zu beschreiben.
Gewöhnlich werden NROM-Zellen so programmiert und gelöscht, dass eine vorgegebene Einsatzspannung erzielt wird. Zum Erreichen dieses Zieles wird die Drain-Spannung stufenweise erhöht. Auf diese Weise wird zwar stets die gewünschte Einsatzsspannung erzielt, aber Effekte, die Aufschlüsse über die Entwicklung der Ladungsträger im ONO zulassen, werden teilweise überdeckt. Aus diesem Grund wird hier eine Zelle mit konstanten Programmier- und Löschbedingungen gezykelt. Das Ergebnis dieses Experiments ist in Abbildung 3.19 dargestellt.
Es ist zu beobachten, dass die Einsatzspannung nach ca. 10 Zykeln beginnt anzusteigen. Dies gilt sowohl für den programmierten, als auch für den gelöschten Zustand. Es ist wichtig anzumerken, dass die ursprüngliche Einsatzspannung der Zelle bei $2,3$ Volt liegt. Die Bedingungen wurden bewusst so eingestellt, dass die ursprüngliche Einsatzspannung beim ersten Löschen um ca. $100mV$ unterschritten wurde. Hierdurch wird sichergestellt,

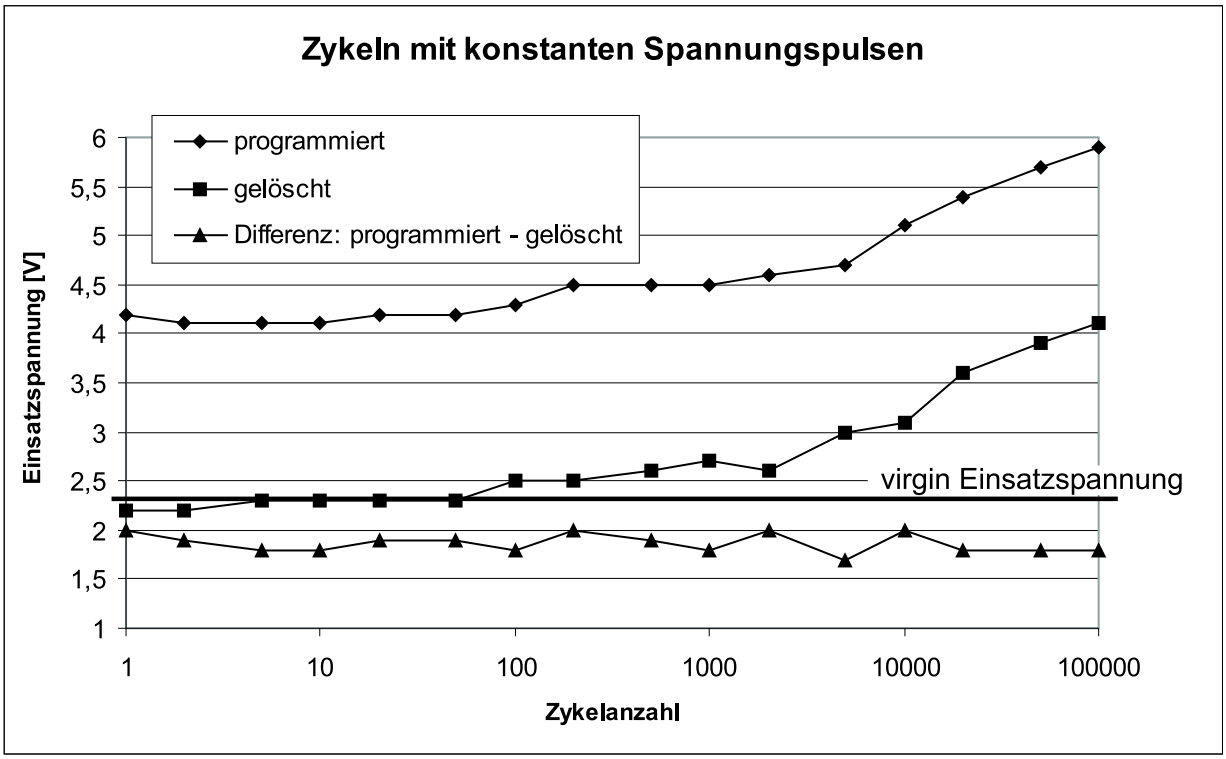


Abbildung 3.19: Einsatzspannungsentwicklung beim Zykeln einer NROM-Zelle mit konstanten Programmier- und Löschbedingungen

dass der spätere Anstieg nicht durch zu schwach eingestellte Löschpulse verursacht wird. Neben dem Ansteigen der Einsatzspannung mit zunehmender Zykelzahl ist bemerkenswert, dass das Fenster zwischen programmiertem und gelöschtem Zustand faktisch über 100.000 Zykeln konstant bleibt.

Das Ansteigen des programmierten Zustandes soll modelliert werden, um Einblick in die Entwicklung der Elektronenverteilung im ONO zu gewinnen. Hierzu wird die NROM-Zelle durch zwei Transistoren beschrieben, dies ist in Abbildung 3.20 zu sehen.

Die effektive Kanallänge einer programmierten NROM-Zelle lässt sich in zwei Bereiche unterteilen: in einen Bereich, über dem keine Ladungsträger injiziert wurden, dieser wird durch den Transistor T1 beschrieben und in einen zweiten Bereich, über dem Elektronen gespeichert sind, selbiger wird durch den Transistor T2 beschrieben. Daraus ergibt sich das, in der unteren Hälfte der Abbildung 3.20, dargestellte Modell. Die Beschaltung ist

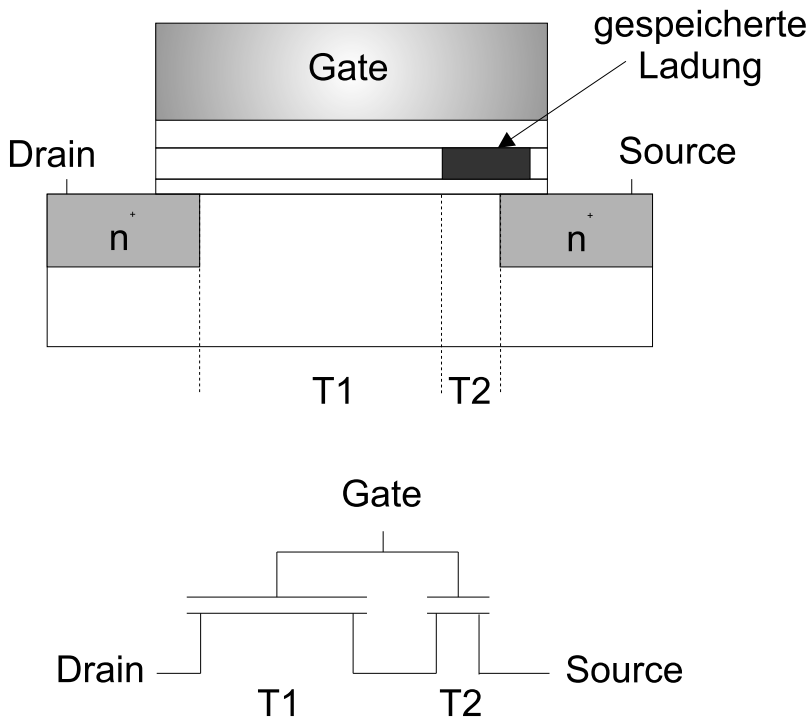


Abbildung 3.20: Schematische Darstellung einer programmierten NROM-Zelle und ihrer Modellierung durch zwei Transistoren (T1 und T2). Source und Drain sind zum Auslesen der programmierten Seite angegeben.

zum Auslesen des programmierten Bits angegeben, die vorhergehende Programmierung geschieht unter Vertauschung von Source und Drain. Eine Beschreibung einer NROM-Zelle durch zwei Transitoren ist bei Chang et al., [7], zur Evaluierung der Lesegeschwindigkeit zu finden.

Hier soll gezeigt werden, dass sich die Erhöhung der Einsatzspannung durch eine Verbreiterung der injizierten Elektronenverteilung beschreiben lässt. Über einem größeren Anteil des Kanals sind somit Ladungen gespeichert. Dies entspricht im Zwei-Transistor-Modell einer Verlängerung der Kanallänge des Transistors T2 (bei entsprechender Verkürzung von T1). Die Gesamtlänge beträgt $200nm$, dies entspricht der effektiven Kanallänge der betrachteten Zelle.

Alternativ könnte man versuchen, die Erhöhung der Einsatzspannung durch ein Ladungspaket konstanter Breite mit steigender Ladungsträgerdichte zu beschreiben. Dies ent-

spräche einer kontinuierlichen Einsatzspannungserhöhung von T2. Jedoch ist dies aus einer Reihe von Gründen, zumindest für sehr hohe Einsatzspannungen, physikalisch nicht naheliegend:

- Die Trapdichte im Siliziumnitrid ist begrenzt.
- Bei sehr hoher lokaler Ladungsdichte im Nitrid verursacht selbige ein starkes Feld, was einer weiteren Injektion an dieser Stelle entgegenwirkt.
- Bei zu hoher Ladungsdichte wird das Feld über der unteren Oxidschicht des ONO so stark, das Elektronen direkt wieder aus der Nitridschicht in den Kanal tunneln können.
- Die nicht programmierte Seite sollte nicht so stark beeinflusst werden, wie dies bei der Messung in Abbildung 3.19 zu beobachten ist.

Aus diesen Gründen wird für das Zwei-Transistor-Modell vereinfachend ein Ladungspaket konstanter Dichte angenommen, dass sich räumlich ausdehnt.
Als Voraussetzung wird der Fall einer „virgin“ Zelle betrachtet. Deren Transferkennlinie kann mit Hilfe der in Kapitel 2 eingeführten Gleichungen ((2.23) ff) beschrieben werden. Es wird eine Anpassung ohne und eine mit Kanallängenmodulation durchgeführt. Von Interesse ist hier nur der Sättigungsbereich, da mit einer hohen Source-Drain-Spannung zum Detektieren nur einer Seite der Zelle gearbeitet wird.
Ohne Kanallängenmodulation, jedoch unter Berücksichtigung von Kurzkanaleffekten und Drain induced barrier lowering, ergibt sich aus den Gleichungen (2.26),(2.23),(2.29),(2.51):

$$\begin{aligned} I_{DS} &= \frac{W\mu C'_{ox}}{2L\left(1+d_1\frac{\gamma}{2\sqrt{\phi_0+U_{SB}}}\right)}\cdot \\ &\quad \left(U_{GS}-U_{FB}-\phi_0-\gamma\sqrt{\phi_0+U_{SB}}+2\beta_1\frac{\epsilon_{Si}\cdot d_{ox}}{\epsilon_{ox}\cdot L}\left(\phi_0+U_{SB}+\beta_2 U_{DS}\right)\right)^2 \end{aligned} \tag{3.4}$$

Unter zusätzlicher Berücksichtigung der Kanallängenmodulation wird die rechte Seite der

Gleichung (3.4) mit dem Faktor (vgl. Gl.(2.24), Gl.(2.30), Gl.(2.38))

$$\frac{1}{1-\frac{\sqrt{\frac{2\epsilon_0\epsilon_{Si}}{qN_A}}\left[\sqrt{\phi_D+\left(U_{DS}-U'_{DS}\right)}-\sqrt{\phi_D}\right]}{L}} \tag{3.5}$$

multipliziert.

Das Ergebnis dieser beiden Berechnungen wird in Abbildung 3.21 mit der gemessen Kurve verglichen.

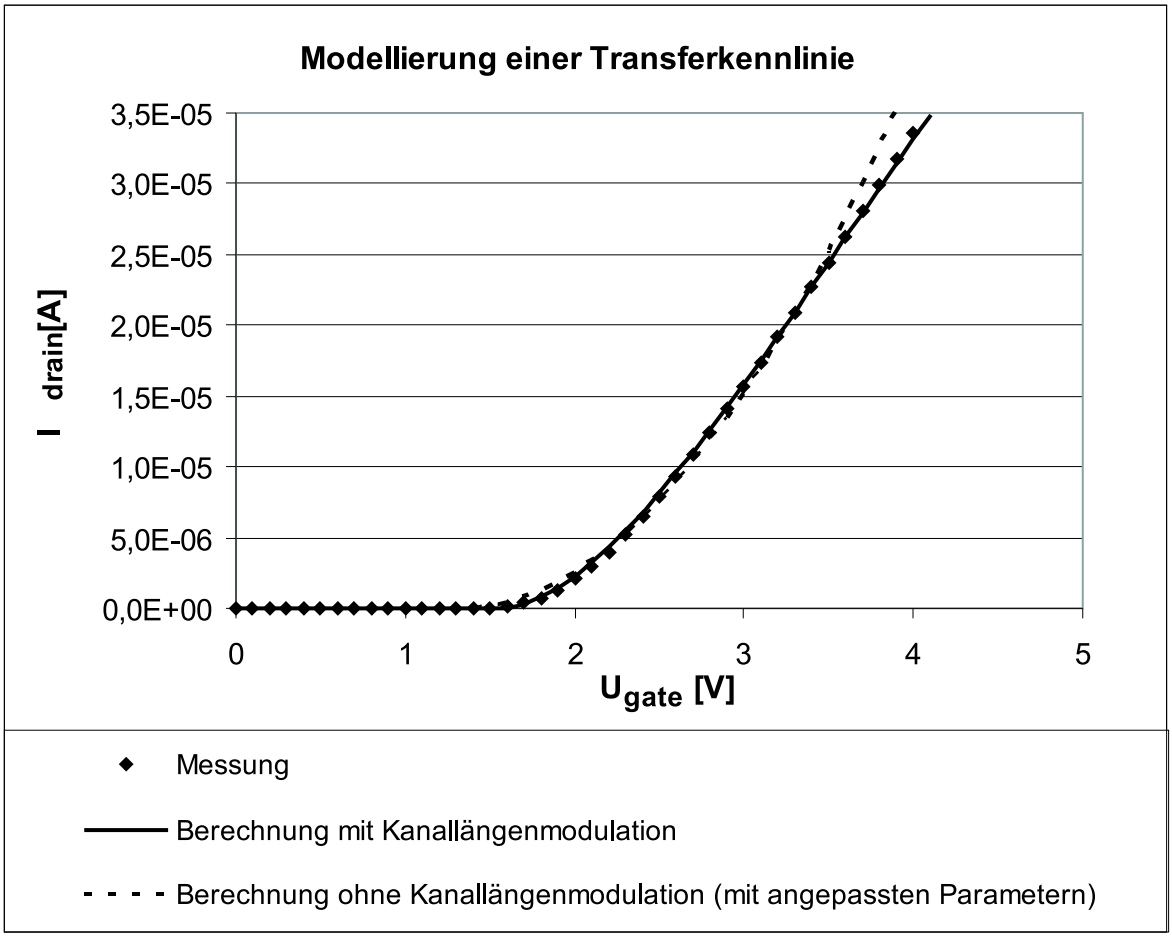


Abbildung 3.21: Gemessene und modellierte Transferkennlinien einer „virgin“ NROM-Zelle

Es lässt sich eine gute Übereinstimmung der beiden Berechnungen mit der Messung feststellen. Man erkennt, dass die Berechnung ohne Kanallängenmodulation nur für hohe Gate-Spannungen zu hohe Ströme vorhersagt. Des weiteren wird hier zur Modellierung eine leicht

erniedrigte Kanaldotierung verwendet. Hierdurch wird die Erniedrigung der Einsatzspannung durch die Kanallängenmodulation kompensiert.
Abbildung 3.21 zeigt, dass die Gleichungen (3.4) und (3.5) geeignet sind, die hier untersuchten NROM-Zellen zu beschreiben. Für das Zwei-Transistor-Modell aus Abbildung 3.20 wird zur einfacheren Berechenbarkeit die Gleichung (3.4) ohne Kanallängenmodulation verwendet.
Die Zelle wird gemäß der Abbildung so betrieben, dass sich die Einsatzspannung der rechten Seite erhöht. Dies entspricht einer Vergrößerung des Bereichs von T2. Beim Auslesen dieses Spannungshubs wird die Source (T2) auf $0V$ gelegt und die Drain (T1) auf $1,6V$. Das Source-Potenzial von T1 muss somit noch berechnet werden, um den Strom, der durch T1 und T2 fließt, ermitteln zu können. Hierzu werden für T1 und T2 Gleichungen der Form (3.4) angesetzt, diese werden gleichgesetzt. Nach einigen Umformungen erhält man eine Gleichung, die U_{DS} von T1 als Funktion von U_G angibt. Hieraus lassen sich alle notwendigen Spannungen von T1 und T2 berechnen. Somit lässt sich I_{DS} des Zwei-Transistor-Modells als Funktion von U_G berechnen. Die Länge von T2, die die Breite der injizierten Ladung über dem Kanal repräsentiert, dient als Fittingparameter, um die verschiedenen Programmierzustände zu beschreiben.
Zur Modellierung des Zykelexperiments aus Abbildung 3.19 wird für T1 eine Langkanaleinsatzspannung von $2V$ und für T2 eine solche von $8V$ eingestellt. Die Einsatzspannung von T1 resultiert aus der Anpassung an den unprogrammierten Zustand. Die Einsatzspannung von T2 ist eine Annahme, sie beruht auf Verteilungsbreitenangaben aus der Literatur für einen Einsatzspannungshub von ca. $2V$,[59]. Dies hat jedoch keine Auswirkung auf die qualitative Beschreibung mit Hilfe des Zwei-Transistor-Modells. Es soll die Verbreiterung der Elektronenverteilung plausibel gemacht werden und nicht deren exakte Breite bestimmt werden. Die Gesamtlänge von T1 und T2 beträgt $200nm$. Vor dem ersten Programmieren hat T2 die Länge 0. Das erste Programmieren mit einer Einsatzspannungserhöhung von $2V$ lässt sich durch eine Länge von $20nm$ von T2 beschreiben. Die gesamten Messungen und Modellierungen sind in Abbildung 3.22 dargestellt.
Es sind die gemessenen Transferkurven nach x-Zyklen und die zugehörigen Berechnungen

mit Angabe der Länge des programmierten Transistoranteils T2 abgebildet. Nach 100.000 Zyklen hat dieser sich auf $39nm$ ausgedehnt und folglich fast verdoppelt.
Diese Abbildung zeigt, dass sich die Zunahme der Einsatzspannung gut durch das Zwei-Transistor-Modell beschreiben lässt. Dies legt nahe, dass sich die Elektronenladung bei dem in Abbildung 3.19 dargestellten Zykelexperiment immer weiter in den Kanal hin ausdehnt.

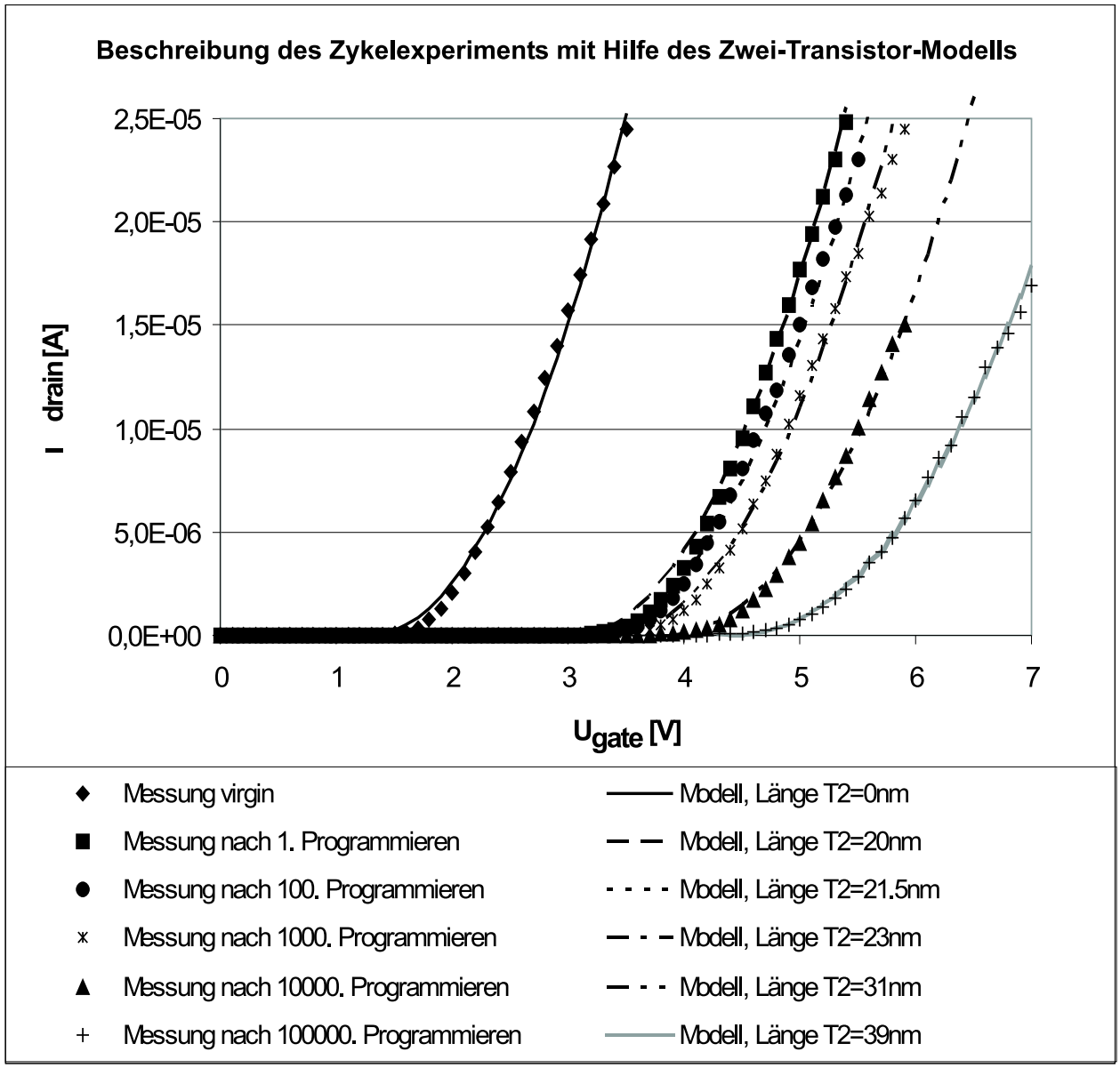


Abbildung 3.22: Modellierung der Entwicklung des programmierten Zustands aus dem in Abb. 3.19 dargestellten Versuch durch das Zwei-Transistor-Modell

Kapitel 4

Experimentelle Evaluierung des STI - Konzepts

In diesem Kapitel werden experimentelle Ergebnisse an STI-begrenzten NROM-Speicherzellen vorgestellt. Zum einen wird generell die Funktionalität des STI-Konzeptes gezeigt, zum anderen werden typische Eigenschaften so gearteter Speicherzellen untersucht. Dabei werden klare Abhängigkeiten aufgezeigt.
Alle in diesem Kapitel präsentierten Experimente sind ausschließlich an STI-begrenzten NROM-Zellen durchgeführt worden.

4.1 Einsatzspannungen und Transferkennlinien

Hier werden zunächst grundlegende Größen für NROM-Zellen betrachtet. Diese Kenngrößen sind auch für Transistoren Standard und aus Kapitel 2 bekannt.
Zuerst betrachten wir eine Transferkennlinie einer STI-Zelle, Abbildung 4.1. Es handelt sich um eine typische Zelle ohne Pocket-Implantation.
Die Kurve ist nicht außergewöhnlich. Im Vergleich zu einem Transistor ähnlicher Generation fällt jedoch auf, dass der Swing schlechter ist. Die hier gemessene Zelle hat einen Swing von $\sim 125mV/Dekade$. Der hohe Swing beruht in erster Linie auf dem vergleichsweise

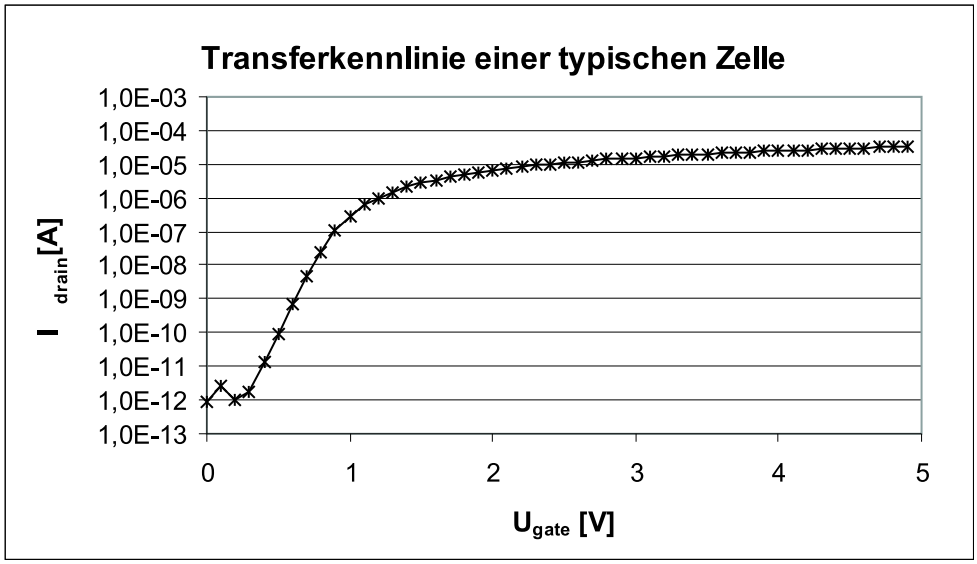


Abbildung 4.1: Transferkennlinie einer typischen STI-begrenzten NROM-Zelle

dicken ONO-Stapel, zudem wird er zusätzlich durch Kurzkanaleffekte erhöht.
Um dies genauer zu untersuchen, betrachten wir Einsatzspannungen verschieden großer Zellen. Da NROM auf der Manipulation der Einsatzspannung beruht, ist diese in jedem Fall eine wesentliche Größe. Im Abbildung 4.2 ist daher die Einsatzspannung verschiedener Zellen über deren gezeichnete Länge aufgetragen.
Die Weiten der Zellen sind selbstverständlich identisch. Jeder Messpunkt stellt eine Mittelung über drei Zellen dar. Die in dieser Arbeit behandelten Zellen haben effektive Längen um die $150nm$. Man sieht deutlich, dass die Zellen ohne Pocket-Implantation sich in diesem Bereich deutlich im roll-off befinden. Ihre Einsatzspannungen sind deutlich niedriger im Vergleich mit Langkanalzellen. Dies ist eine Erklärung für den schlechten Swing der Zelle aus Abbildung 4.1. Abhilfe kann die in Abschnitt 3.2 erläuterte Pocket-Implantation schaffen. Die Absenkung der Einsatzspannung dieser Zellen ist wesentlich geringer.
Betrachten wir nun den Weiten roll-off in Abbildung 4.3.
Diese Messungen wurden an Zellen ohne Pocket-Implantation durchgeführt. Auch hier sieht man deutlich, dass Zellen mit typischen Dimensionen um $120nm$ im roll-off liegen. Ihre Einsatzspannungen sind klar abgesenkt. Dieses Schmalkanalverhalten ist typisch für

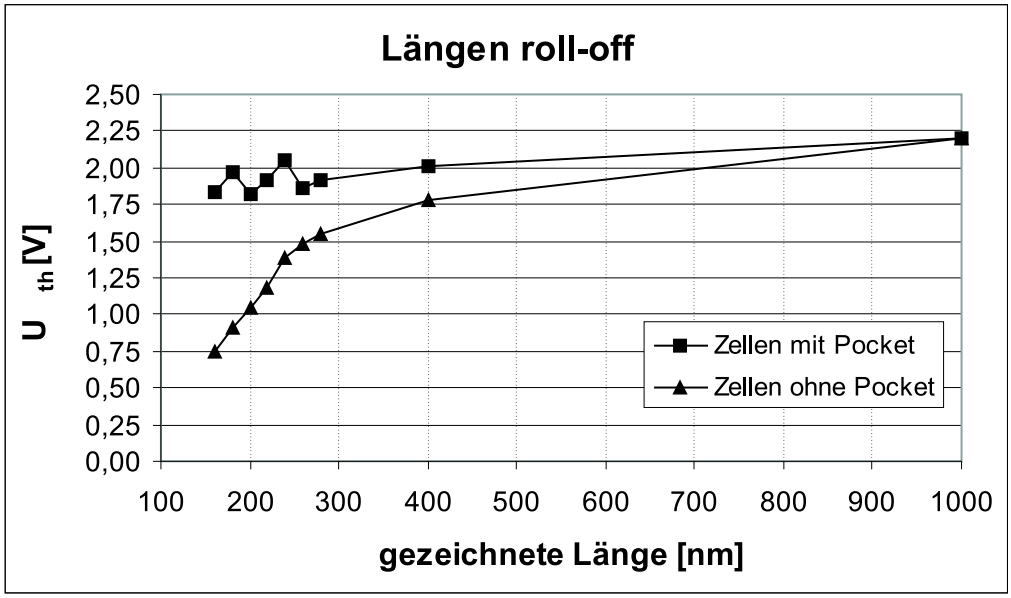


Abbildung 4.2: Längen roll-off

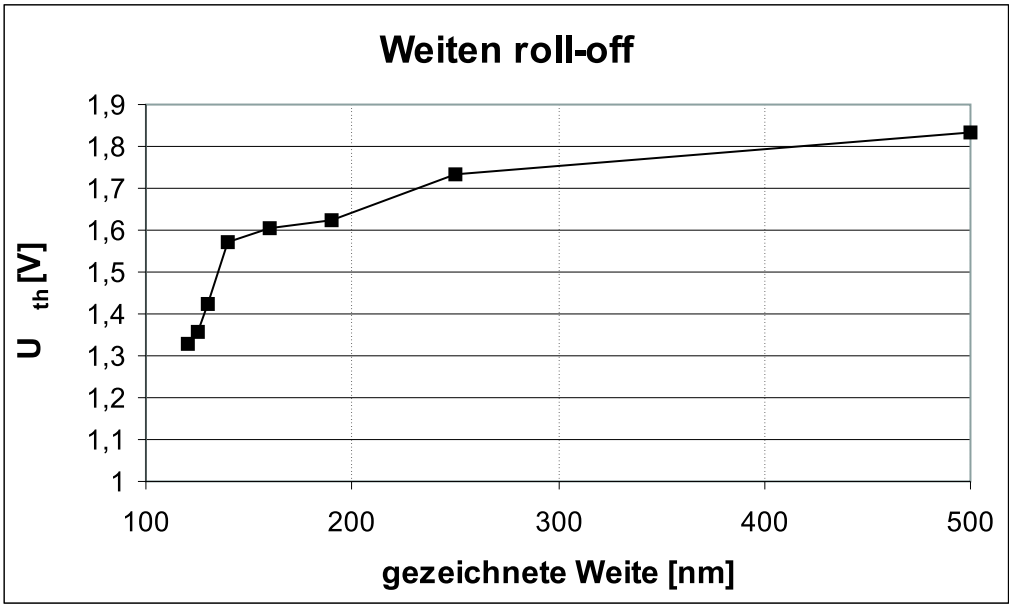


Abbildung 4.3: Weiten roll-off

STI-begrenzte Bauelemente. Es entspricht den Erwartungen aus Abschnitt 2.3.4.
Insgesamt ist festzustellen, dass bei den in dieser Arbeit behandelten Zellen deutlich Miniaturisierungseffekte zur Geltung kommen.

4.2 Kanaldotierung

Für die Beurteilung von Messergebnissen an NROM-Zellen ist es wichtig, Informationen über die Kanaldotierung zu haben. Es gibt eine Vielzahl von Möglichkeiten, die Dotierung eines Transistors und somit auch einer NROM-Zelle zu bestimmen. Die genauesten Methoden sind jedoch nicht zerstörungsfrei und zudem sehr aufwendig. Daher wird hier ein elektrisches Verfahren verwendet, dass leicht in der Anwendung ist und daher alltagstauglich.
Bei diesem Verfahren wird die Kanaldotierung über den Substratsteuerfaktor (Gleichung (2.17)) in einem iterativen Prozess bestimmt. Es wird die Einsatzspannung einer Zelle (Gleichung (2.23)) bei unterschiedlicher Substrat-Spannung, U_{SB}, gemessen.
Als Gesamtausdruck für die Einsatzspannung ergibt sich aus den beiden Gleichungen:

$$U_{th} = U_{FB} + \phi_0 + \frac{1}{C'_{ox}}\sqrt{(2qN_A\epsilon_0\epsilon_{Si})\,(\phi_0 + U_{SB})} \tag{4.1}$$

Es ist zu beachten, dass nach Gleichungen (2.20) und (2.21) gilt:

$$\phi_0 \sim \ln\frac{N_A}{n_i} \tag{4.2}$$

Gleichung (4.1) wird nach $\sqrt{N_A}$ umgeformt, es ergibt sich:

$$\sqrt{N_A} = \frac{U_{th} - \overbrace{U_{FB} - \phi_0}^{TermA}}{\frac{1}{C'_{ox}}\sqrt{2q\epsilon_0\epsilon_{Si}\left(2\underbrace{\phi_0}_{TermB} + U_{SB}\right)}} \tag{4.3}$$

An dieser Gleichung erkennt man sofort, warum dieses Verfahren iterativ funktioniert. Auf der rechten Seite steht zweimal ϕ_0 (in Term A und Term B), also ein Ausdruck, der selber von N_A abhängig ist. Diese Abhängigkeit ist jedoch sehr schwach, da N_A normalerweise

mindestens fünf Größenordnungen größer ist als die Intrinsicdichte n_i. Der natürliche Logarithmus aus dem Quotienten (Gl. (4.2)) schwankt also nur wenig und somit auch ϕ_0. Es wird zuerst ein Wert für N_A angenommen, mit dessen Hilfe ϕ_0 berechnet wird. Somit sind alle Größen auf der rechten Seite bekannt. U_{th} wird in Abhängigkeit von U_{SB} gemessen. In die flächenbezogene Oxidkapazität, C'_{ox}, geht die Dicke des ONO-Stapels ein. Sie wird zuvor in eine equivalente Oxiddicke umgerechnet.

Nun wird Zähler über Nenner aufgetragen. Die Steigung dieser Kurve, bei richtiger Messung ergibt sich eine Gerade, liefert die Wurzel der Kanaldotierung. Nach Quadrierung erhält man N_A selbst. Da nur die Steigung der Kurve von Interesse ist und nicht die absolute Lage in y-Richtung, kann der Term A aus Gleichung (4.3) von vorne herein vernachlässigt werden, er stellt nur einen Offset dar.

Mit dem neuen Wert für die Kanaldotierung geht man in die nächste Iterationsschleife. Hier wird ϕ_0 mit Hilfe des neuen N_A errechnet. Dieses Verfahren konvergiert nach spätestens drei Iterationen, wenn der zuerst angenommene Wert für N_A nicht um mehrere Grössenordnungen neben dem wirklichen Wert liegt. Dies geschieht in der Realität jedoch nicht, da man auf Grund der implantierten Dosis stets eine Vorstellung von der Kanaldotierung hat.

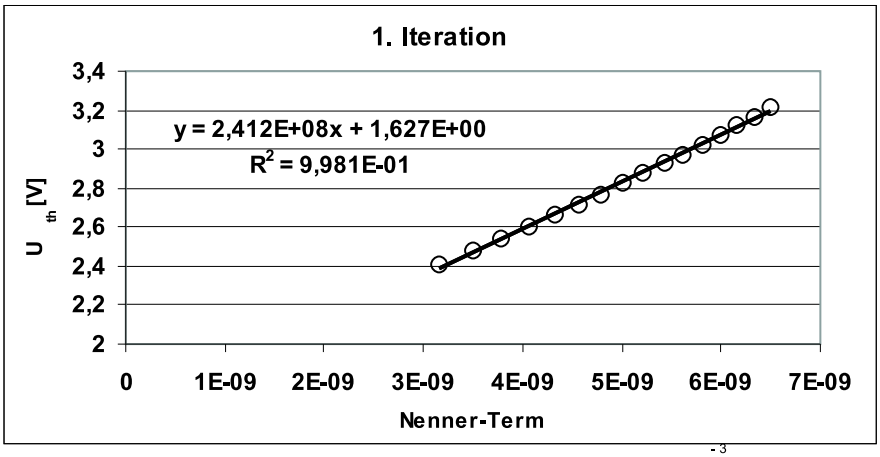
Ein Beispiel aus der Praxis ist in Abbildung 4.4 zu sehen.

Hier wurde mit Absicht ein Startwert ausgewählt, der weit vom Zielwert entfernt liegt, über eine Größenordnung. Man sieht jedoch, dass das Verfahren trotzdem schnell zum Ziel führt.

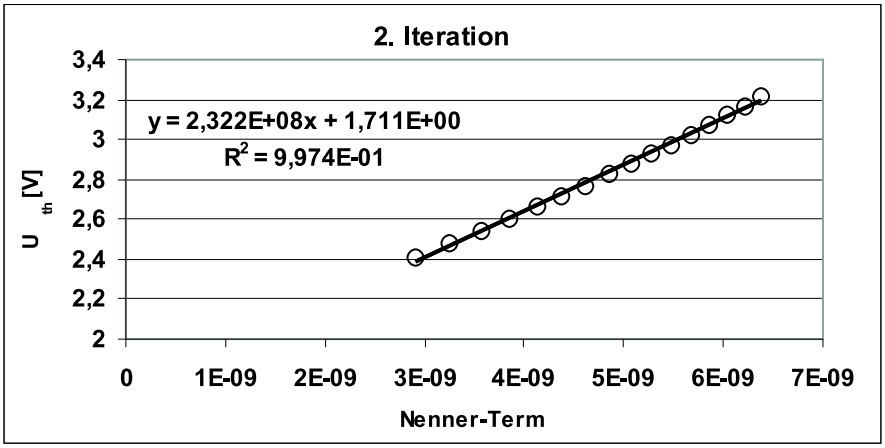
	N_A
angenommener Anfangswert	$1 \cdot 10^{18} cm^{-3}$
1. Iteration	$5,82 \cdot 10^{16} cm^{-3}$
2. Iteration	$5,39 \cdot 10^{16} cm^{-3}$
3. Iteration	$5,38 \cdot 10^{16} cm^{-3}$

Tabelle 4.1: Iterationsverfahren zur Bestimmung der Kanaldotierung

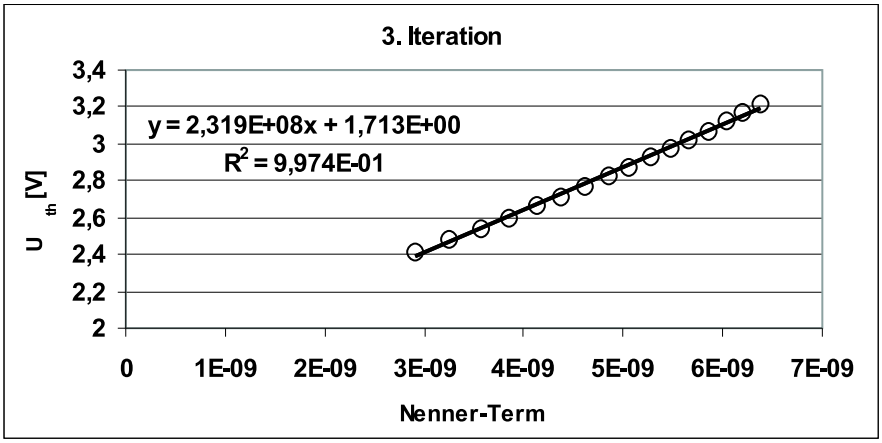
Startwert für N_A : 1E18 cm^{-3}



Ergebnis der 1. Iteration: N_A = 5,82E16 cm^{-3}



Ergebnis der 2. Iteration: N_A = 5,39E16 cm^{-3}



Ergebnis der 3. Iteration: N_A = 5,38E16 cm^{-3}

Abbildung 4.4: Iterationsverfahren zur Bestimmung der Kanaldotierung

Bei der Messung der Kanaldotierung mit diesem Verfahren muss auf die Wahl einer geeigneten Transistorgeometrie geachtet werden. Diese Forderung beruht auf den Annahmen des Verfahrens, ausgedrückt in Gleichung (4.1). In dieser Formel werden keine Kurzkanal- bzw. Schmalkanal-Effekte berücksichtigt. Daher führt die Messung an sehr kleinen Zellen zu Ergebnissen, deren absolute Werte nicht die physikalische Realität widerspiegeln. Daher sollten für die Bestimmung der Wannendotierung große Strukturen verwendet werden (einige μm Kanallänge $\times$ einige μm Kanalweite).
Wenn jedoch nur Zellen untereinander verglichen werden sollen und man sich bewusst ist, dass die absoluten Werte nicht korrekt sind, so kann man dieses Verfahren auch bei kleinen Strukturen einsetzen.

4.2.1 Weiten-Effekt

Dieses iterative Messverfahren wird nun zur Bestimmung der Kanaldotierung auf Zellen mit unterschiedlicher Kanalweite angewendet, um den Verlust von Bor ins STI zu verifizieren. Für diesen Versuch werden Zellen mit gleicher effektiver Kanallänge verwendet. Da diese Länge jedoch deutlich kleiner $1\mu m$ ist, sind die Absolutwerte nicht exakt. Da der Fehler auf Grund des Kurzkanal-Effekts aber bei allen Zellen gleichermaßen auftritt, ist eine Vergleichbarkeit gewährleistet.
Die Ergebnisse sind in Abbildung 4.5 dargestellt.
Es ist deutlich zu erkennen, dass für schmalere Zellen eine geringere Konzentration des Wannendotierstoffs gemessen wird. Diese Beobachtung deckt sich mit den Erwartungen für Zellen mit shallow trench isolation. Zudem steht dieses Ergebnis in gutem Einklang mit den Messungen der Einsatzspannungen an verschieden weiten NROM-Zellen.

4.3 Programmierkurven

Aus dem Programmierverhalten einer NROM-Speicherzelle lässt sich viel über ihre Eigenschaften und ihre Güte aussagen.

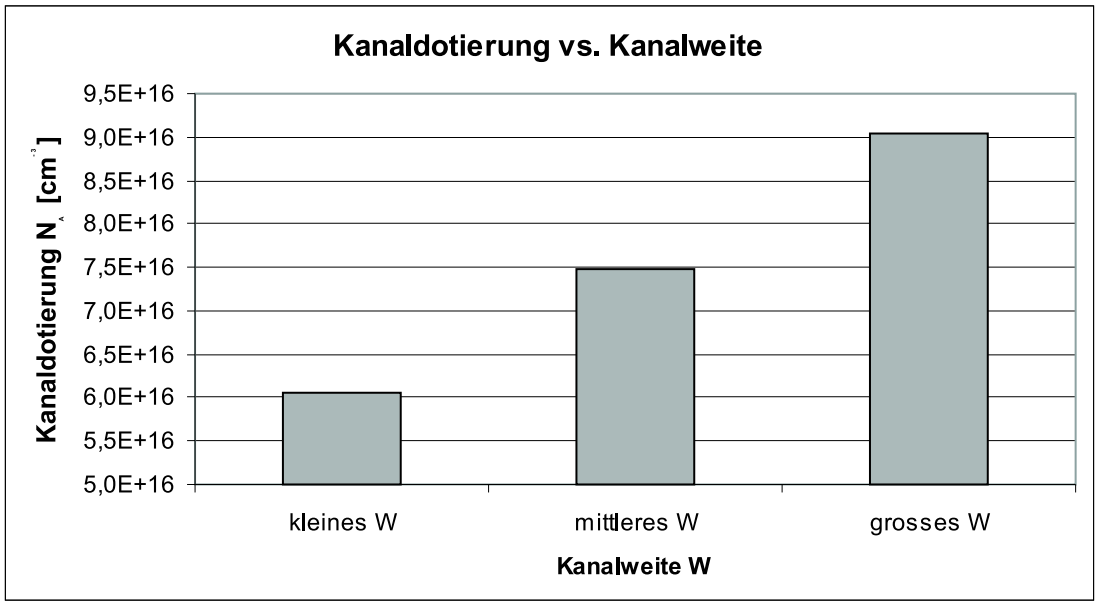


Abbildung 4.5: Kanaldotierung vs. Kanalweite

Die Injektion von Elektronen geschieht durch Anlegen von Spannungs-Pulsen an Gate und Drain. Grundsätzlich bieten sich zwei Vorgehensweisen für die Programmierung an:

- bei Wahl einer konstanten Drain-Puls-Spannung werden die Gate-Puls-Spannungen schrittweise erhöht,
- bei Wahl einer konstanten Gate-Puls-Spannung werden die Drain-Puls-Spannungen schrittweise erhöht.

Hier wird die zweite Möglichkeit näher betrachtet. Es wird an das Gate immer die gleiche Spannung angelegt und eine Reihe von ansteigenden Spannungen an die Drain. Dies wird so lange durchgeführt, bis eine gewünschte Einsatzspannungsverschiebung erreicht ist.

4.3.1 Längen-Effekt

Die verwendeten Gate-Spannungen liegen üblicherweise im Bereich von $8V$ bis $11V$. In Abbildung 4.6 sind derartige Programmierkurven zu sehen. Es wird die Erhöhung der Einsatzspannung über die Drain-Spannung aufgetragen.

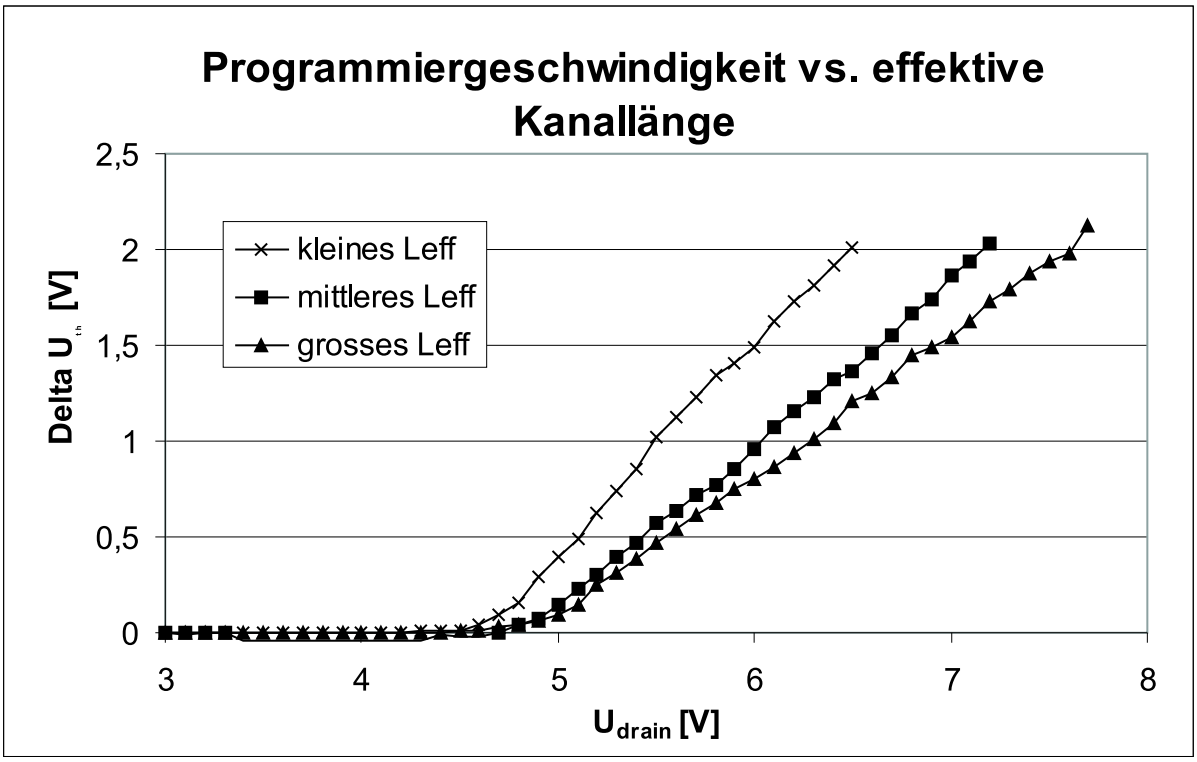


Abbildung 4.6: Programmiergeschwindigkeit in Abhängigkeit von der effektiven Kanallänge

Wichtig ist anzumerken, dass die Dauer der Spannngspulse an der Drain konstant ist. Es sind drei Messkurven von Zellen aufgetragen, die sich nur durch die effektive Kanallänge unterscheiden. Die Messergebnisse zeigen deutlich, dass sich Zellen mit kürzerer effektiver Kanallänge leichter programmieren lassen. Die Begründung liegt in der höheren lateralen Feldstärke. Für gleiche Drain-Source-Spannung, U_{DS}, ergibt sich bei kürzerer Kanallänge eine größere laterale Feldstärkenkomponente. Dies hat eine stärkere Beschleunigung der Elektronen zur Folge. Somit setzt die Ladungsträgerinjektion bereits bei niedrigerer Drain-Source-Spannung ein.

Diese Trends sind positiv für zukünftige Generationen von NROM-Technologien, da sie die Möglichkeit eröffnen, mit dem Verkleinern der physikalischen Dimensionen auch die benötigten Spannungen zu erniedrigen.

4.3.2 Weiten-Effekt

Will man Speicherzellen miniaturisieren, so geschieht dies nicht nur in Längsrichtung, sondern auch in Weitenrichtung. Aus diesem Grund wird hier auch das Verhalten von extrem schmalen NROM-Zellen betrachtet. Die Messergebnisse an Zellen mit gleicher Länge, jedoch mit unterschiedlichen Weiten sind in Abbildung 4.7 zu sehen.

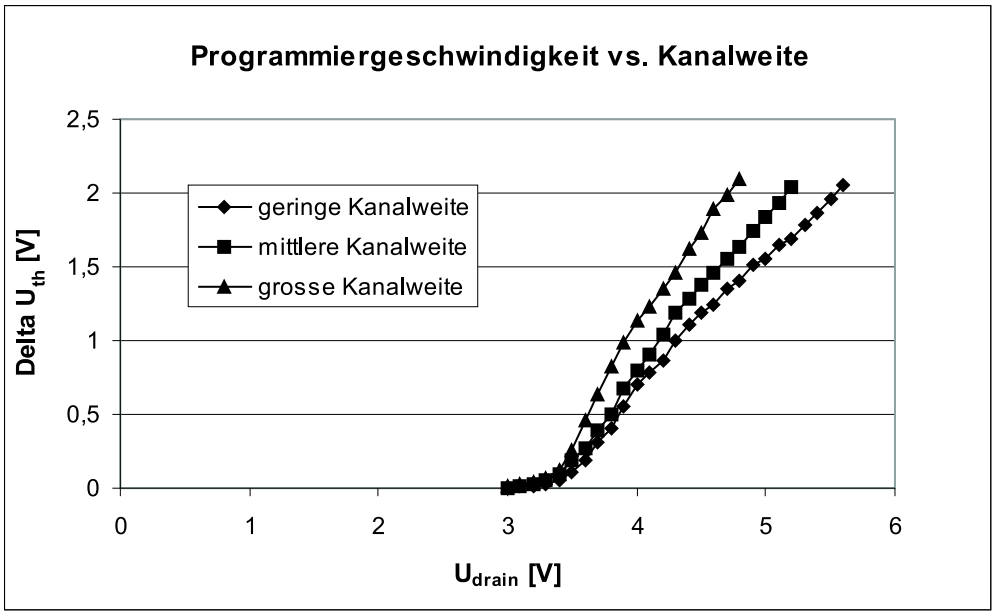


Abbildung 4.7: Programmiergeschwindigkeit in Abhängigkeit von der Kanalweite

Es ist deutlich zu erkennen, dass sehr schmale Zellen schwerer zu programmieren sind als dies für weite Zellen der Fall ist. Weitere Messungen haben gezeigt, dass die notwendige Drain-Spannung beim Programmieren für noch breitere Zellen sättigt. Diese Beobachtung, dass sich Zellen mit unterschiedlicher Kanalweite unterschiedlich schnell programmieren lassen, war nicht zu erwarten. Bei gleichartigen Zellen mit verschiedener Weite sollte sich die Programmiergeschwindigkeit nicht unterscheiden, solange die Zellen elektrisch gut angeschlossen sind und genügend Strom zur Verfügung steht.
Die verlangsamte Programmierung bei besonders schmalen Zellen lässt sich folglich nur mit den Auswirkungen von Schmalkanaleffekten erklären (Abschnitt 2.3.4). Sinkt die p-

Dotierung zum Rand der Zelle hin ab, so hat dies Einfluss auf das Verhalten der Inversionsschicht beim Lesen der Einsatzspannung der Zelle. Genauer wird dies im Rahmen des Nebensprechenes in Abschnitt 4.5.2 betrachtet. Aus der niedrigeren Dotierung am Rand folgt, dass mehr Ladungsträger injiziert werden müssen, um die gleiche U_{th} Verschiebung zu erzielen. Diese Erklärung wird durch die Messung gestützt. Höhere Drain-Spannungen bedeuten bei gleicher Kanallänge, mehr eingeschossene Elektronen.

4.4 Löschkurven

Der Löschvorgang ähnelt dem Programmieren. Es werden ebenfalls Pulse an Gate und Drain angelegt. Da die beim Programmieren injizierten Elektronen kompensiert werden sollen, werden beim Löschen so lange Löcher injiziert, bis eine gewünschte, geringere Einsatzspannung erreicht ist.
Das Gate wird folglich auf negatives Potential gelegt, üblicherweise im Bereich zwischen $-6V$ und $-12V$. Dann werden positive Spannungspulse auf die Drain gegeben. Somit kommt es durch die starke Bandverbiegung im Bereich des pn-Übergangs zu einer Generation von Löchern, von denen ein geringer Teil in die Nitridschicht des ONO-Stapels injiziert wird.
Eine typische Löschkurve ist in Abbildung 4.8 dargestellt.
Hier wurde mit einer Gate-Spannung von $-6.5V$ gearbeitet. Die Zelle wurde zuvor so programmiert, dass die Einsatzspannung um $2V$ angehoben ist. Daher ist auf der y-Achse die Änderung der Einsatzspannung gegenüber dem ursprünglichen Wert aufgetragen. Dies ist sinnvoll, da Zellen, die unterschiedlich stark programmiert werden, sich auch unterschiedlich beim Löschen verhalten. Dies ist leicht einsichtig, da das lokale Feld der injizierten Elektronen einen starken Einfluss auf den Löcher-Einschuss hat. So lässt sich eine stark programmierte Zelle zu Beginn einfacher löschen als eine schwach programmierte Zelle. Daher wird stets die ursprüngliche Einsatzspannung als Referenz verwendet. Auf der x-Achse ist die Spannung der angelegten Drain-Pulse aufgetragen. Die Kurve zeigt das Absinken der Einsatzspannung vom programmierten Zustand aus durch Anlegen von Löschpulsen,

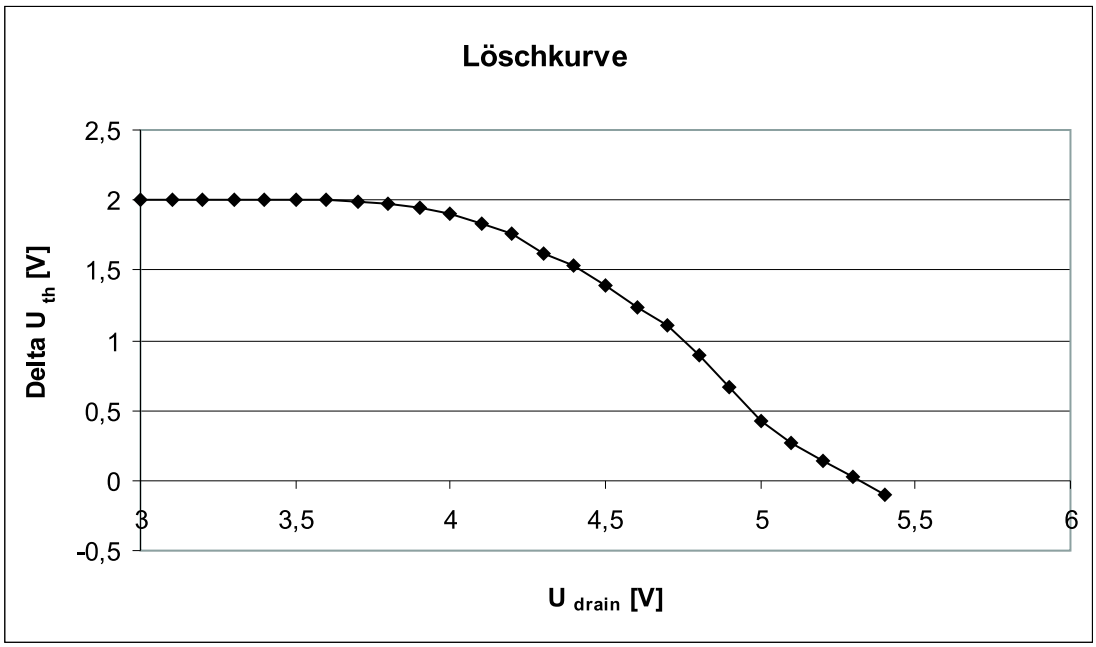


Abbildung 4.8: relatives Absinken der Einsatzspannung beim Löschen

bis wieder die ursprüngliche Einsatzspannung erreicht ist.

4.4.1 Längen-Effekt

Es wurden Zellen untersucht, die sich nur durch die effektive Kanallänge unterschieden. Die Puls-Spannungen, die an die Drain angelegt werden müssen, um die Zellen zurück zu ihrer ursprünglichen Einsatzspannung zu löschen, sind in Grafik 4.9 abgebildet.
Das Ergebnis ist eindeutig. Lange Zellen lassen sich schlechter löschen als kurze Zellen, dies ist ähnlich dem Programmieren. Dieses Untersuchungsergebnis beim Löschen war nicht so klar zu erwarten, wie das beim Programmieren. Das Löschen findet lokal statt, es sollte also in erster Näherung nicht von der Kanallänge abhängen. Dazu betrachten wir die Löschkurve einer sehr langen Zelle in Abbildung 4.10.
Der Unterschied zu der Kurve in Abbildung 4.8 ist sehr deutlich, die Löschkurve der langen Zelle flacht zum Ende hin stark ab. Dies bedeutet, dass immer mehr Löcher injizert werden, die nur noch einen geringen Einfluss auf die Einsatzspannung besitzen. Das Alignement

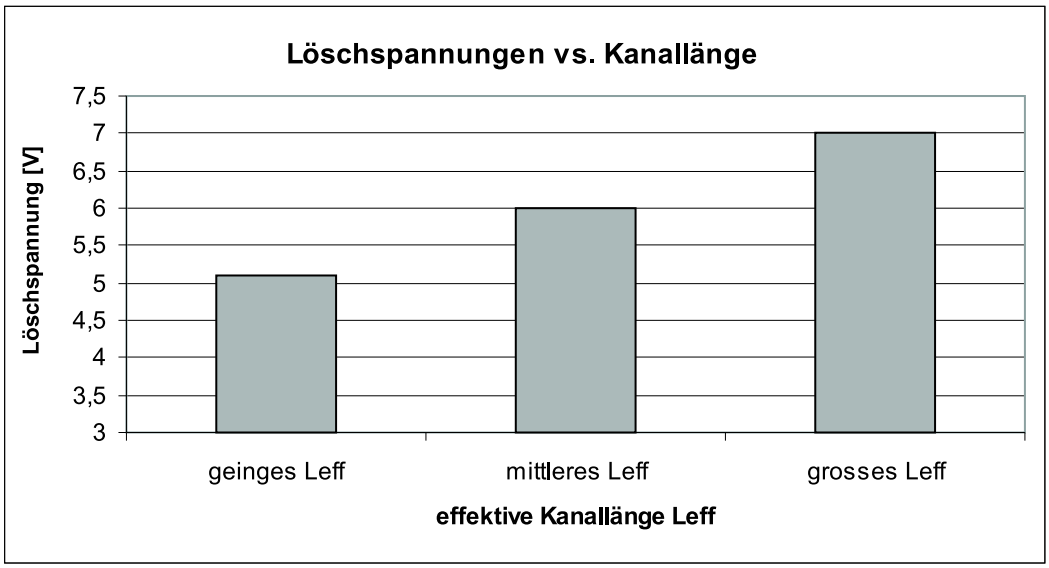


Abbildung 4.9: Löschspannungen in Abhängigkeit von der effektiven Kanallänge

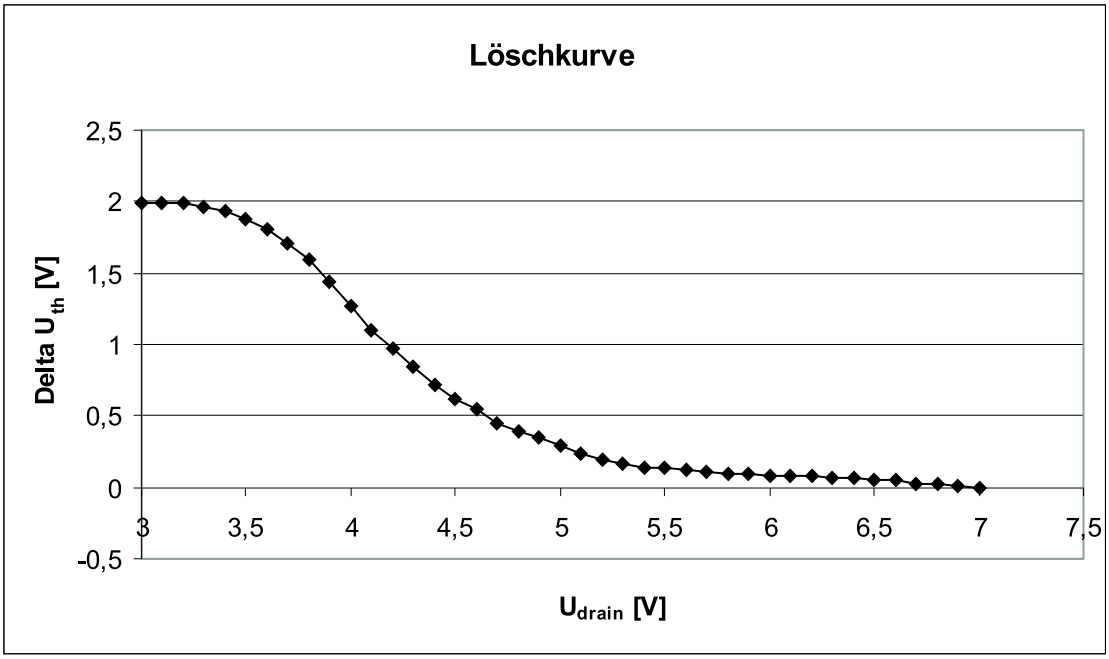


Abbildung 4.10: Löschverlauf einer sehr langen Zelle

von Löchern und Elektronen ist folglich schlecht.
Als Resultat ergibt sich, dass nicht primär die Injektion an sich schwerer wird für längere Zellen, sondern dass die relative Lage von Löchern zu Elektronen schlechter wird. Dies spiegelt sich in den benötigten Drain-Puls-Spannungen wider. Eine Löschkurve, die zum Ende hin stark abflacht, kann Ausdruck einer zu breiten Elektroneninjekton sein, es muss nicht Ausdruck eines Löschproblems sein.

4.5 Nebensprechen

Unter dem Nebensprechen in einer NROM-Speicherzelle versteht man die Auswirkungen des Programmierzustandes der einen Seite auf die andere Seite der Zelle. Da in bisherigen Konzepten stets ein Bit durch den Programmierzustand einer Seite charakterisiert wird, ist die Verwendung der Begriffe 'Seite' und 'Bit' austauschbar.
Werden in eine Seite der NROM-Zelle Elektronen injiziert, so steigt die Einsatzspannung, die durch ein Lesen in umgekehrter Betriebsrichtung (Vertauschen von Source und Drain) bestimmt wird. Beträgt diese gewünschte Erhöhung der Einsatzspannung beispielsweise $2V$, so steigt die Einsatzspannung des anderen Bits ungewollt ebenfalls an. Dieses unerwünschte Nebensprechen liegt normalerweise in der Größenordnung von wenigen $100mV$. Da man die beiden Bits unabhängig voneinander behandeln möchte, ist das Nebensprechen gleichbedeutend mit einer Reduktion des Fensters, das durch die Differenz von programmiertem zu nicht programmiertem Zustand beschrieben wird.

Es gibt einige Parameter, durch die das Nebensprechen in der Zelle direkt beeinflusst wird, wie z.B.:

- Position bzw. Breite der eingeschossenen Ladungsverteilungen im ONO
- Kanaldotierung
- effektive Kanallänge
- Kanalweite

4.5.1 Position bzw. Breite der eingeschossenen Ladungsverteilungen im ONO

Intuitiv ist es leicht einzusehen, dass ein größerer physikalischer Abstand der Ladungspakete im ONO zu einem geringeren Nebensprechen führt.
Folglich lautet die Fragestellung, ob es möglich ist, eine solche Verbesserung im Nebensprechen durch Abstandsveränderung der Ladungspakete zu erzielen, ohne die Speicherzelle physikalisch zu verändern.
Ansatzpunkt bilden die in Abschnitt 2.3.1 besprochenen Überlegungen zur Kanallängenmodulation. Wenn man die Strecke l_{km}, also den Abstand zwischen pinch-off Punkt und dem metallurgischen pn-Übergang verkürzt, so werden die Elektronen erst näher an der Drain heiß und können in das Nitrid injiziert werden. Somit bauen sich die Elektronenverteilungen weiter außen in der Zelle auf und ihr physikalischer Abstand nimmt zu. Daher ist eine Verringerung der Auswirkung des Nebensprechens zu erwarten.
Gleichung (2.30) zeigt, dass eine Verkleinerung der Differenz $(U_{DS} - U'_{DS})$ zur erwünschten Reduktion von l_{km} führt. Beim Programmieren sind die Gate-Source-Spannung und die Drain-Source-Spannung die entscheidenden Einflussgrößen (es wird mit $U_{SB} = 0V$ gearbeitet). Den gewünschten Hebel bietet U_{GS} und dies in zweierlei Hinsicht. Zum einen wird bei erhöhter Gate-Source-Spannung eine geringere Drain-Source-Spannung zum Programmieren benötigt und zum anderen steigt U'_{DS}. Nach Gleichung (2.24) gilt $U'_{DS} \sim (U_{GS} - U_{th})$.

Die Ergebnisse dieses Experiments sind in Abbildung 4.11 veranschaulicht. Jede Kurve stellt einen Programmiervorgang dar. Auf den Achsen sind die Änderungen der Einsatzspannungen der beiden Bits in einer Zelle dargestellt. Bit 1 wird programmiert, die resultierende Erhöhung der Einsatzspannung ist auf der x-Achse aufgetragen. Das unerwünschte Hochlaufen des benachbarten Bits ist auf der y-Achse aufgetragen.
Betrachtet man das Hochlaufen von Bit 2, wenn die Einsatzspannung von Bit 1 um 2 Volt angehoben ist, so sieht man deutlich, dass das Nebenspechen für höhere Spannungen U_{GS} beim Programmieren schwächer ist. Dies entspricht den Erwartungen.

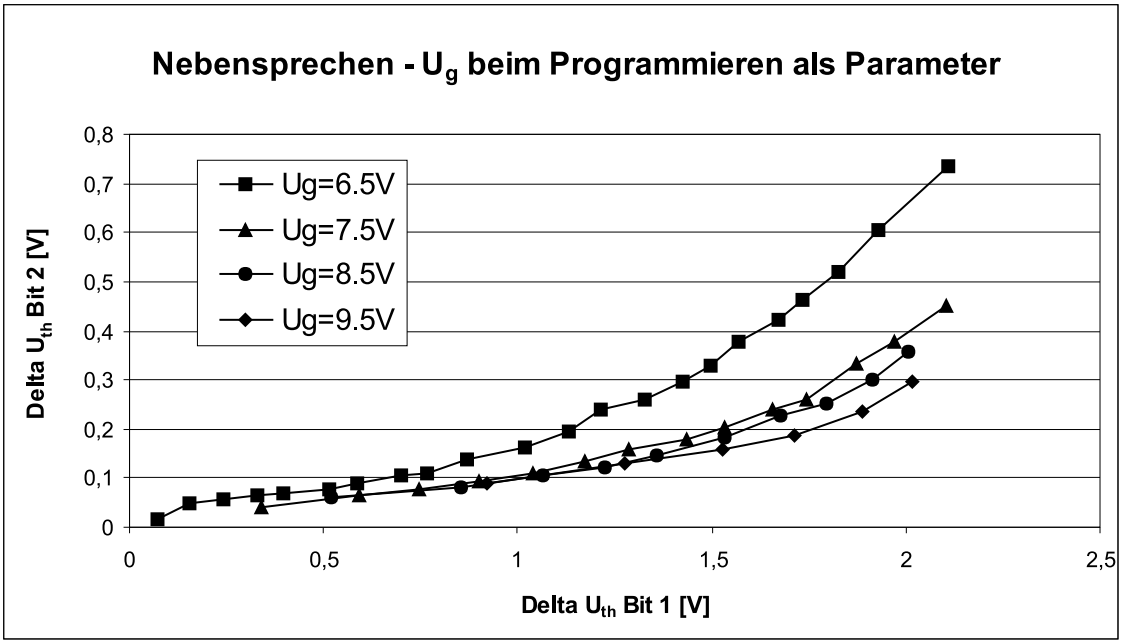


Abbildung 4.11: Nebensprechen - U_G beim Programmieren als Parameter

4.5.2 Kanaldotierung

Die Auswirkung der beiden Seiten einer NROM-Zelle aufeinander hängt direkt mit der Kanaldotierung zusammen.

Für die Untersuchung dieses Effekts wurden drei Zellen herangezogen, die sich jeweils nur durch die Wannendotierung unterscheiden. Die p-Dotierungen liegen in einem Bereich zwischen $1,7 \cdot 10^{17} cm^{-3}$ (niedriges N_A) und $5,5 \cdot 10^{17} cm^{-3}$ (hohes N_A).

Die Dimensionen der gemessenen Zellen betragen:

- effektive Kanallänge: $\sim 150nm$
- Weite: $\sim 200nm$

Die Ergebnisse dieses Experiments sind in Abbildung 4.12 dargestellt.

Die Messwerte sind genauso aufgetragen, wie es bei Grafik 4.11 der Fall ist. Für ein Programmierfenster von $\Delta U_{th} = 2V$ ergeben sich folgende Werte:

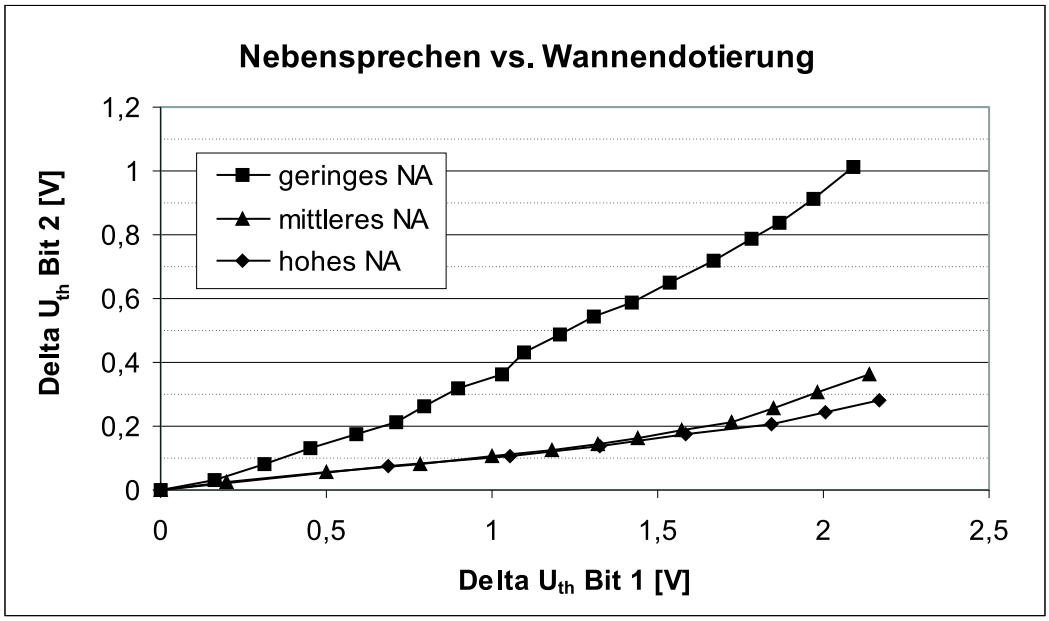


Abbildung 4.12: Nebensprechen vs. Wannendotierung

Es ist eine deutliche Abhängigkeit des Nebensprechens von der Wannendotierung, N_A, zu erkennen. Mit steigender Wannendotierung nimmt das Nebensprechen in der Zelle ab. Die extrem schlechte Trennung der beiden Bits bei der als niedrig bezeichneten Wanne zeigt deutlich, dass diese Dotierung für eine Zelle mit einer effektiven Kanallänge von nur $\sim 150nm$ nicht mehr geeignet ist. Im Bereich besser geeigneter Wannendotierungen fallen die Unterschiede geringer aus.

Wie lässt sich diese Beobachtung erklären? Die naheliegendste Erklärung liegt in der höheren Kanaldotierung, durch die ein Transistorverhalten erzielt wird, das einem Langkanal-

Wannendotierung	Nebensprechen
geringes N_A	$935mV$
mittleres N_A	$312mV$
hohes N_A	$240mV$

Tabelle 4.2: Nebensprechen vs. Wannendotierung

bauelement ähnlicher ist. Diese Begründung ist jedoch noch sehr allgemein. Um ein besseres Verständnis der NROM-Zelle zu erreichen, wollen wir die Ursachen hier genauer betrachten.

Eine geringere Wannendotierung ist gleichbedeutend mit einem geringeren Substratsteuerfaktor (siehe Gleichung (2.17)). Dies hat eine unmittelbare Auswirkung auf die Ladungsmenge, die beim Programmieren injiziert werden muss, um eine bestimmte Einsatzspannungsverschiebung zu erzielen. Bei niedrigerem N_A sind mehr Elektronen notwendig, als bei höherem N_A.

Zu dieser Fragestellung wurden von der Simulationsgruppe in München bei Infineon um Patrick Haibach Simulationen durchgeführt. Maßgeblich haben Reiner Hagenbeck und Frank Lau daran mitgearbeitet. Die Ergebnisse sind in Abbildung 4.13 dargestellt.

Es ist eine typische NROM-Zelle im STI-Konzept simuliert worden. Zum Vergleich ist die Simulation für zwei unterschiedlich dotierte Wannen durchgeführt worden. Die effektive Kanallänge ist für beide Zellen identisch. Dann ist in der Nitridschicht ein negatives Ladungsträgerpaket platziert worden. Dies bildet die beim Programmieren injizierten Elektronen nach. Die Ladungsmenge wird genauso angepasst, dass sich für Lesebedingungen eine Einsatzspannungserhöhung von $2V$ ergibt. Dies ist für beide Wannen durchgeführt worden, da natürlich ihre ursprünglichen Einsatzspannungen unterschiedlich sind.

Bei dieser Anpassung der Ladungspakete zeigt sich als wichtige Erkenntnis, dass für eine höher dotierte Wanne weniger negative Ladung in der Nitridschicht benötigt wird, um eine Einsatzspannungsverschiebung von $2V$ zu erzielen, als dies für eine niedriger dotierte Wanne der Fall ist. Die Erklärung ergibt sich nach genauer Betrachtung der Simulation.

In Abbildung 4.13 ist der Lese-Fall für die beiden um $\Delta U_{th} = 2V$ programmierten Zellen dargestellt. Der Ausschnitt ist auf der Source-Seite für diesen Lese-Fall, was für das vorherige Programmieren die Drain-Seite war. Aufgetragen ist nun der Elektronenstrom für den Fall $U_G = U_{th}$, also gerade für jenen Punkt, an dem der Transistor aufschaltet. Es ist deutlich zu sehen, dass der einsetzende Stromfluss für die niedrigere Wannendotierung (linke Seite) deutlich weiter in der Tiefe liegt. Bei höherem N_A fließen die Elektronen sehr viel näher am Gate entlang und somit sehr viel näher an der injizierten Elektronenladung

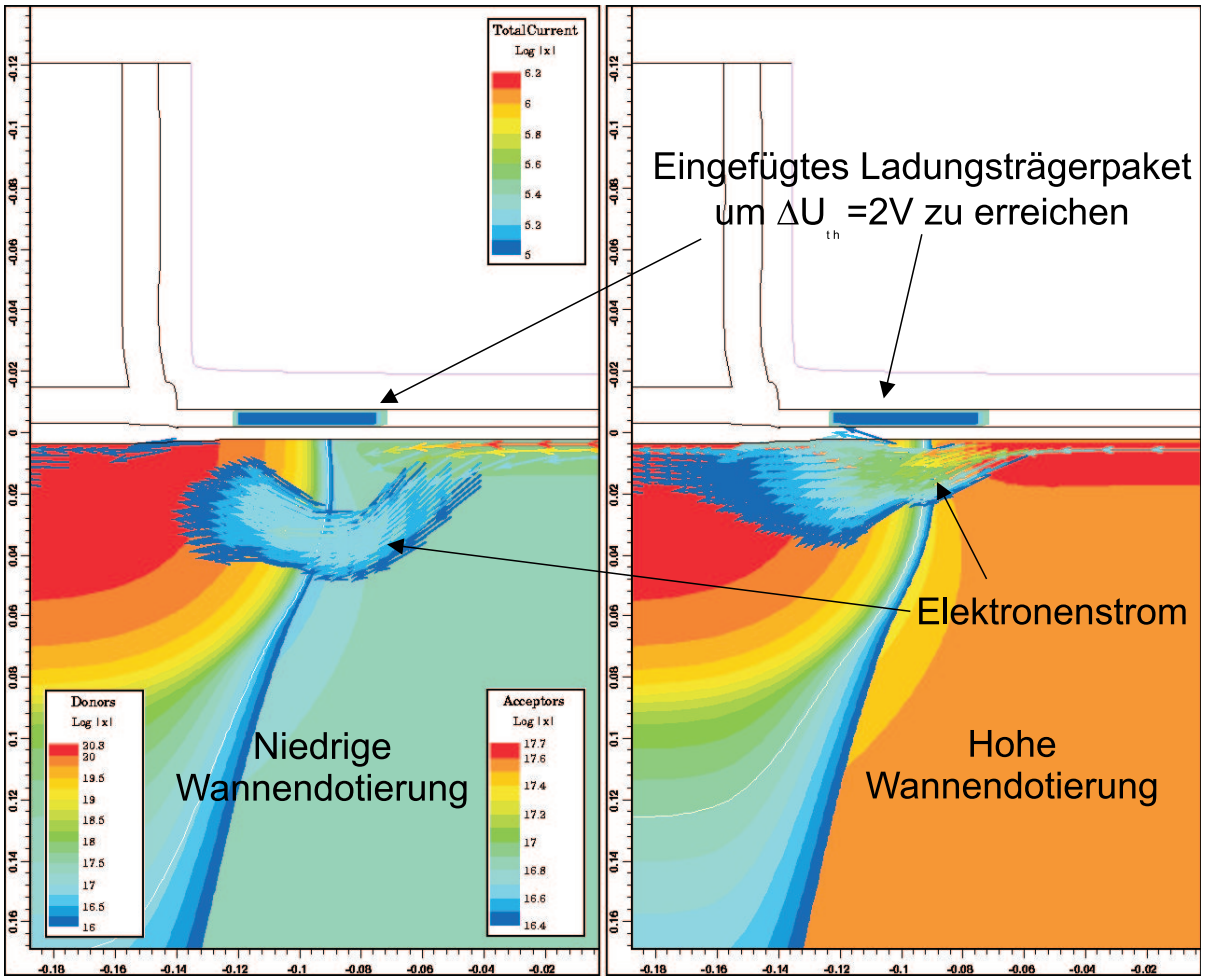


Abbildung 4.13: Simulation der Lesestromverteilung für verschiedene Wannendotierungen, [22].

in der Nitridschicht.

Die durchgeführte Simulation bestätigt die Erklärung. Für eine geringeres N_A müssen die Ladungsträger im ONO über eine geometrisch sehr viel größere Strecke (in eine größere Tiefe) die Öffnung eines Stromkanals unterdrücken. Somit wird eine größere Gesamtladung benötigt. Dies führt ebenfalls zu einer stärkeren Beeinflussung des benachbarten Bits, also zu einer Verschlechterung des Nebensprechens.

4.5.3 Effektive Kanallänge

Da der Markt immer höhere Speicherdichten nachfragt und somit immer kleinere Zellabmessungen notwendig werden, ist in erster Linie die untere Grenze, die durch die effektive Kanallänge gesetzt wird, von Interesse. Bei zu kurzen effektiven Kanallängen verschlechtert sich das Nebensprechen in der Speicherzelle. Die naheliegendste Limitierung ist durch die Breite der injizierten Ladungsträgerverteilungen gegeben. Sie müssen stets physikalisch getrennt bleiben. Die hieraus resultierende Grenze gibt Shapppir et al., [59], mit $70nm$ an. Es ist anzumerken, dass dies noch nicht experimentell untermauert wurde.
Die Limitierung aus der Trennung der Ladungspakete ist aus deren Breite leicht zu ermitteln. Schwieriger wird es in einem Bereich, in dem zwar die physikalische Trennung bereits sicher ist, aber die elektrische Trennung problematisch wird. Verringert man die effektive Kanallänge bei gleichbleibender Kanaldotierung N_A, so wird das Nebensprechen schlechter (siehe vorherigen Abschnitt). Für einen gewissen Kanallängenbereich kann dies durch höhere Dotierungen kompensiert werden.
In Abbildung 4.14 ist die experimentelle Verifikation an zwei NROM-Zellen, die bis auf die effektive Kanallänge vergleichbar sind, zu sehen. Bei der kürzeren Zelle ist das Nebensprechen um den Faktor zwei schlechter.

4.5.4 Kanalweite

Bei Verwendung von Modellen, wie sie in Abschnitt 3.4 vorgestellt wurden, werden Weiteneffekte nicht berücksichtigt. All diese Modelle beruhen auf einer zweidimensionalen Vorstellung. Auf dieser Grundlage kann man folglich nicht erwarten, dass das Nebensprechen von der Kanalweite abhängig ist.
Zur Überprüfung wurden Messungen an Zellen mit unterschiedlicher Weite durchgeführt. Die Ergebnisse sind in der Abbildung 4.15 abgebildet.
Es ist ein eindeutiger Trend zu erkennen. Für breite Zellen sind die Auswirkungen der beiden Bits aufeinander deutlich geringer, als dies bei schmalen Zellen der Fall ist. Es ist jedoch auch zu sehen, dass dieser Effekt für größere Weiten gegen einen fixen Wert läuft

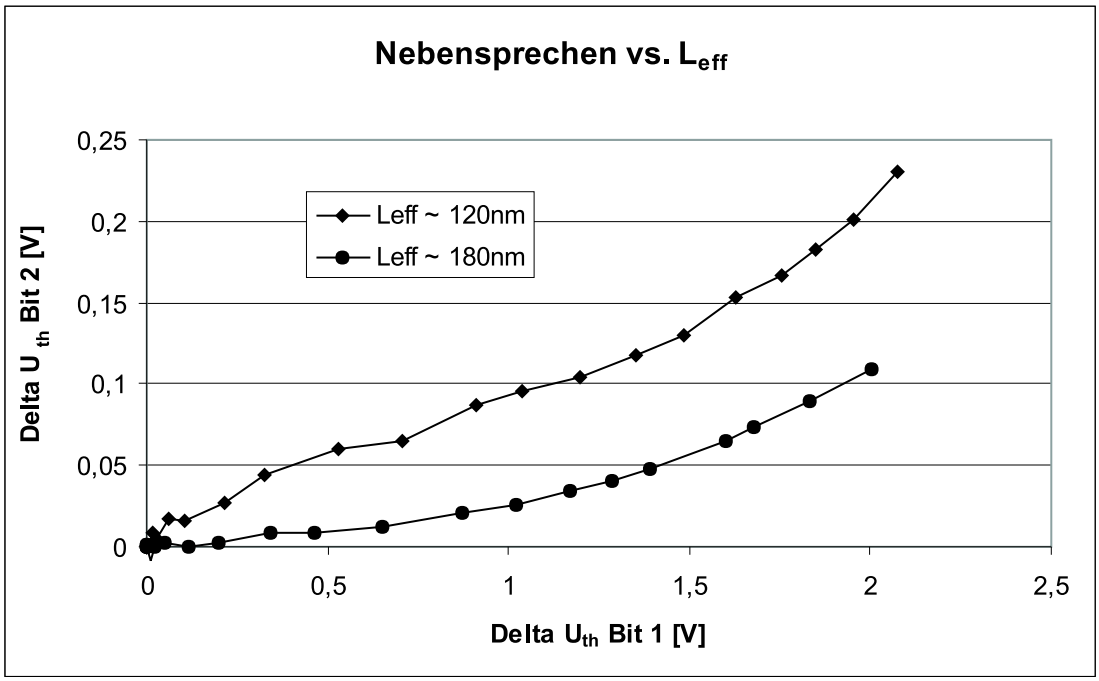


Abbildung 4.14: Nebensprechen vs. effektive Kanallänge

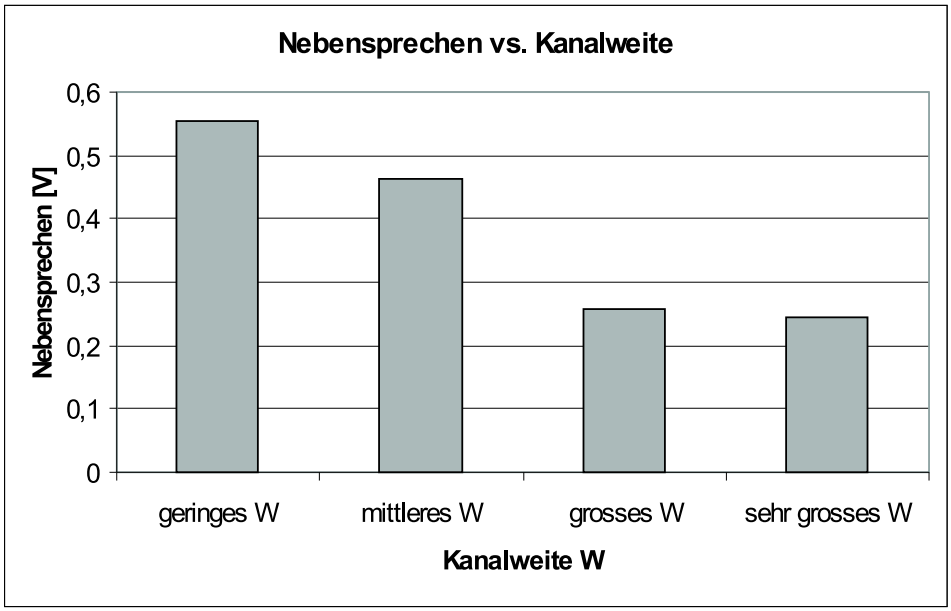


Abbildung 4.15: Nebensprechen vs. Kanalweite

und nicht kontinuierlich weiter sinkt.
Der deutliche Weiten-Effekt zeigt, dass für sehr schmale NROM-Zellen nicht weiterhin angenommen werden kann, dass nur die Längsrichtung betrachtet werden muss. Der Schmalkanaleffekt für STI-begrenzte Zellen wurde bereits in Abschnitt 2.3.4 erläutert. Er besagt, dass die Einsatzspannung für schmale Zellen sinkt. Dies wird u.a. durch eine Erniedrigung der Kanaldotierung am Rand erklärt. Nehmen wir diese Erklärung und die Erkenntnisse aus Abschnitt 4.5.2, so folgt, dass am Rand mehr Ladungsträger injiziert werden müssen. Das starke Nebensprechen in schmalen Zellen lässt sich folglich auf die hohe Anzahl der injizierten Elektronen zurückführen. Dies unterlegen zum einen die Erkenntnisse über die Kanaldotierungsabhängigkeit und zum anderen die Programmierkurven aus Abschnitt 4.3.2. Die Ergebnisse zeigen deutlich, dass es für ein aggressives Verkleinern der Kanalweite dringend notwendig ist, die Auswirkungen des Weiteneffekts kontrollieren zu können.

4.6 Punch-Messungen

Bei den hier durchgeführten Punch-Messungen werden Source und Gate auf $0V$ gelegt und die Drain-Spannung erhöht. Gemessen wird der Strom an der Source. Starkes Punchen ist unerwünscht, da dies sich negativ z.B. auf das Löschen auswirkt. Wenn Zellen gelöscht werden sollen und benachbarte Zellen punchen, so baut sich nicht die gewünschte Spannung an der Drain zum Löschen auf, somit wird das Löschen verlangsamt.
Eine Punch-Messung an Zellen mit unterschiedlicher effektiver Kanallänge ist in Abbildung 4.16 zu sehen.

Es ist das erwartete MOS-Transistorverhalten zu beobachten. Bei kurzen Zellen ist der Source-Strom größer. Eine analoge Messung wurde mit negativer Spannung am Gate durchgeführt. Bei diesen abgeänderten Bedingungen ist der Strom gesunken. Dies ist ein klares Indiz für ein Punchen an der Oberfläche.

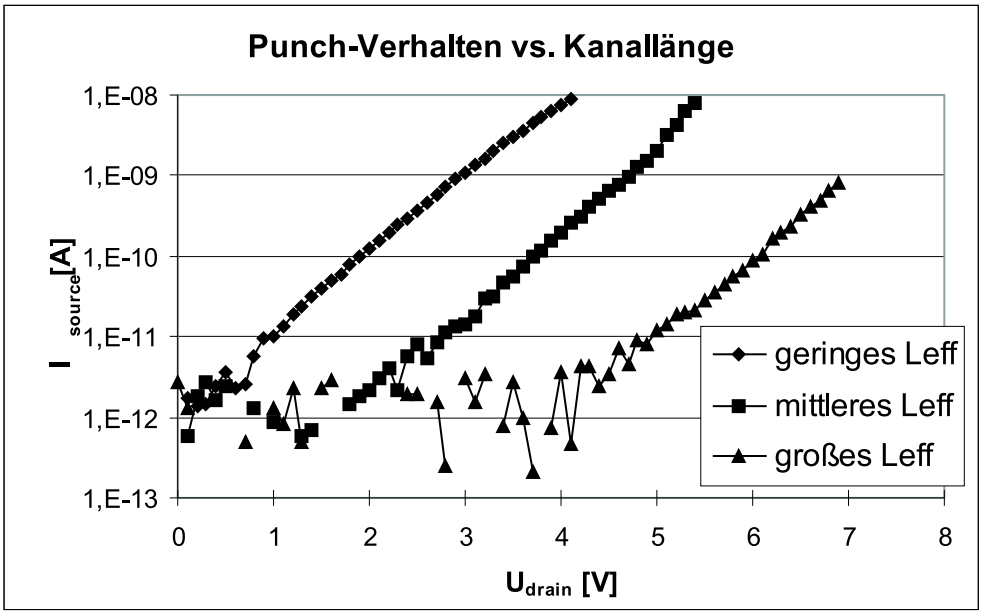


Abbildung 4.16: Punch-Verhalten in Abhängigkeit von der effektiven Kanallänge

4.7 Zyklen - Messungen

Das Zykeln von Speicherzellen ist in vielerlei Hinsicht von grosser Bedeutung. Zum einen ist es für den wirtschaftlichen Nutzen von Speichern wichtig, dass diese sich möglichst oft programmieren und löschen lassen. Zum anderen ist die messtechnische Beobachtung einer Zelle beim bzw. nach dem Zykeln von großem Interesse, um das physikalische Verständnis für die NROM-Zelle zu verbessern.
Bei den bisher vorgestellten Messungen wurden, solange nicht anders gekennzeichnet, stets Zellen verwendet, die nicht zuvor gezykelt wurden. Im Verlauf vieler Programmier- und Löschvorgänge verändern sich jedoch einige Eigenschaften der NROM-Zellen. Dies wird durch die nun vorgestellten Messungen deutlich.
Die Messungen sind wie folgt durchgeführt:

- Anhand von Transferkennlinien werden Stromkriterien für den programmierten und den gelöschten Zustand festgelegt.

- Mit Hilfe dieser festgelegten Kriterien wird ein Bit der Zelle gezykelt, dabei werden nach jedem Programmieren bzw. Löschen die Leseströme beider Bits gemessen. Zudem werden die Spannungen, die für die Programmier- bzw. Löschpulse benötigt werden, gespeichert.

Für das Bestimmen der Stromkriterien reicht es aus, an einer Zelle in ihrem ursprünglichen Zustand eine Transferkennlinie zu messen. Aus der Transferkennlinie lassen sich alle notwendigen Größen gewinnen. Es wird eine fixe Gate-Spannung gewählt, bei der der Lesestrom gemessen wird. Nun kann man durch Verschieben der I-U Kurve in linearer Auftragung die Stromkriterien bei der gewählten Gate-Spannung so festlegen, dass sich das gewünschte Programmierfenster ergibt. Bei der Messung, deren Ergebnisse in Abbildung 4.17 dargestellt sind, wurde für den gelöschten Zustand ein Stromkriterium von $60\mu A$ und für den programmierten Zustand ein Kriterium von $10\mu A$ festgesetzt.
Nun wird mit einem geeigneten Algorithmus sichergestellt, dass die Zelle nach den festgelegten Stromkriterien gezykelt wird. Dies bedeutet, dass beim Programmieren so lange die Spannung der Pulse an der Drain erhöht wird, bis beim anschließenden Lesen ein Strom von $10\mu A$ unterschritten wird. Beim Löschen wird entsprechend vorgegangen, bis ein Strom von $60\mu A$ überschritten wird.
Während der Messung werden stets die Leseströme an beiden Bits gemessen, auch wenn nur ein Bit gezykelt wird. Zudem werden die Spannungen der Programmier- und Löschpulse mitprotokolliert, dies ermöglicht zusätzliche Aussagen über das Verhalten der gemessen Zelle.
Die Ergebnisse einer solchen Messungen sind in den Abbildungen 4.17 und 4.18 dargestellt. Bevor wir zum Vergleich der Ergebnisse kommen, wird zunächst nur die erste Messung betrachtet.
In der oberen Darstellung von Abbildung 4.17 sind die gemessenen Ströme über die Anzahl der Programmier- und Löschvorgänge für beide Bits aufgetragen. Die schwarzen Symbole stellen das gezykelte Bit und die grauen Symbole das nicht gezykelte Nachbarbit dar. Es fällt sofort auf, dass beim Löschen nicht bis auf den Ausgangswert des Stromes und da-

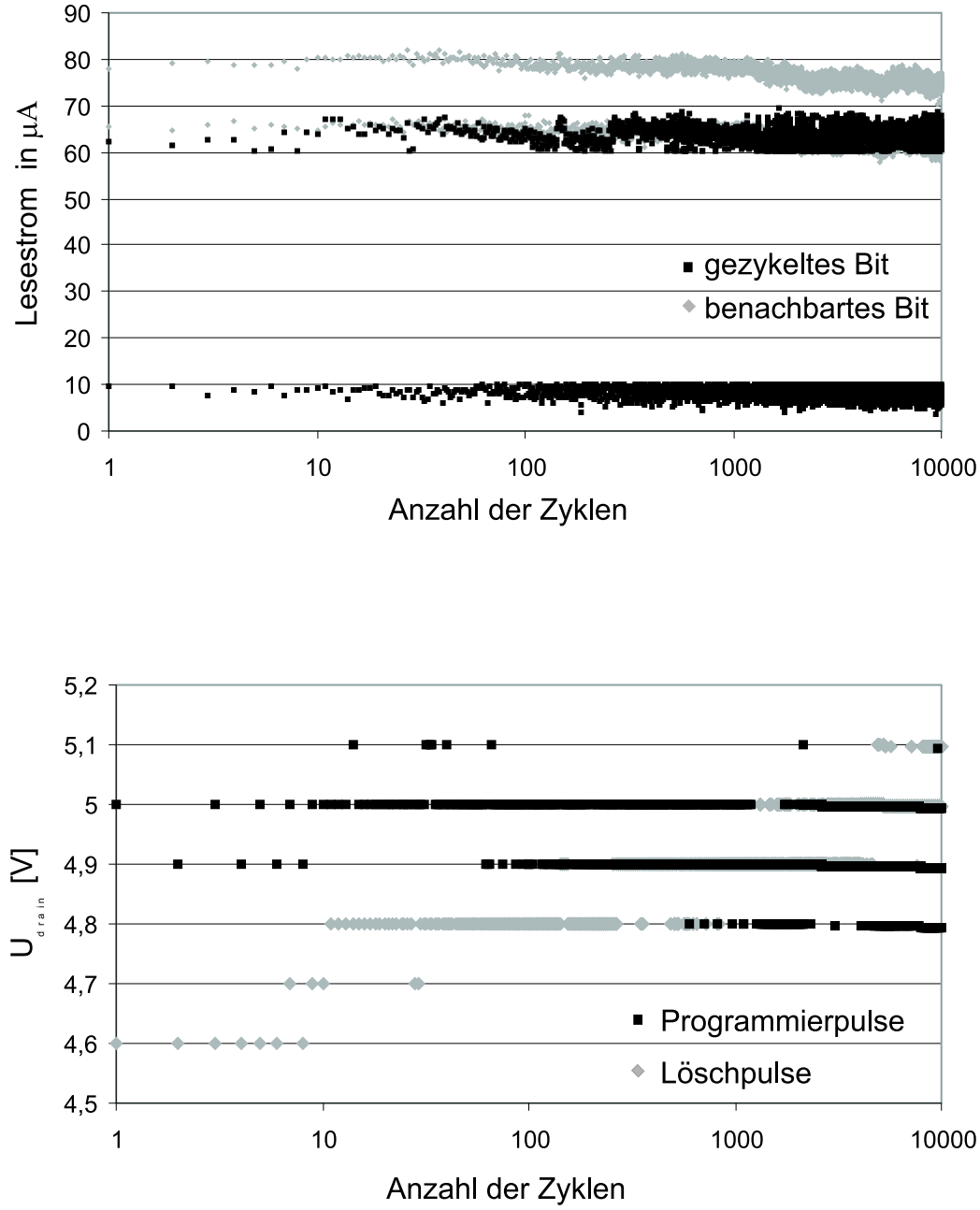


Abbildung 4.17: Zyklen-Messung an einer Zelle mit mittlerer Wannendotierung; oben: Leseströme im Verlauf der Messung; unten: Spannungen der Programmier- bzw. Löschpulse

mit nicht bis zur ursprünglichen Einsatzspannung zurückgelöscht wird. Das Nachbarbit, das sich im ursprünglichen Zustand befindet, liefert einen Strom der größer ist, als das Löschkriterium von $60\mu A$. Diese Festlegung des Fensters beim Zykeln ist willkürlich.
Weiter kann festgestellt werden, dass es auch für das Nachbarbit (das nicht gezykelte Bit) zwei voneinander unterscheidbare Zustände bei der Strommessung gibt. Die Ursache hierfür ist das Nebensprechen, das in Abschnitt 4.5 behandelt wurde. Ist das gezykelte Bit gerade programmiert, so steigt hierdurch auch die Einsatzspannung des benachbarten Bits, und somit sinkt für dieses der Lesestrom. Dies verursacht die beiden Linien von Punkten für das Nachbarbit. Zudem fällt auf, dass der Lesestrom für dieses Bit in der Tendenz mit steigender Anzahl von Zyklen leicht absinkt.
Das Absinken des Lesestroms kann auf zwei verschiedene Weisen erklärt werden. Eine mögliche Erklärung ist, dass sich über die Dauer der Zyklen Elektronen über dem Kanal weit weg von der primär injizierten Verteilung angesammelt haben. Dies ist das Bild der Sekundärelektronen-Injektion. Solche Elektronen beeinflussen das benachbarte Bit stark, da ihre geometrische Entfernung nicht so groß ist, wie die der übrigen Elektronen. Zudem müssen sie durch eine größere Zahl von Löchern kompensiert werden, damit das gezykelte Bit wieder das festgelegte Stromkriterium erfüllt.
Die zweite mögliche Erklärung ist eine allgemeine Verbreiterung der Ladungsverteilungen während des Zykelns. Nehmen wir Abschnitt 3.4.5 als Basis, so kommt es wegen der unterschiedlichen Verteilungsbreiten von Elektronen und Löchern zu einer Verbreiterung der Verteilungen, da injizierte Elektronen und Löcher nicht stets rekombinieren, sondern stets so lange Ladungen injiziert werden, bis eine gewünschte Einsatzspannung erreicht ist. Wird die Elektronenverteilung zum Kanal hin bereiter, so wird sich auch ein Reservoir von Löchern näher zum pn-Übergang hin aufbauen. Durch eine solche Anhäufung von Ladungsträgern wird folglich auch das benachbarte Bit stärker beeinträchtigt.
Bei der Betrachtung des gezykelten Bits sieht man, dass alle Punkte unter $10\mu A$ bzw. oberhalb von $60\mu A$ liegen, somit hat der Algorithmus funktioniert. Die Abgrenzung dieses Fensters ist sehr scharf nach innen, nicht jedoch nach außen. Durch die stufenweise Erhöhung der Drain-Pulse kann es natürlich zu einem leichten Überprogrammieren bzw.

Überlöschen kommen.
In der unteren Darstellung von Abbildung 4.17 sind die Spannungen der Programmier- und Löschpulse über die Anzahl der Zyklen aufgetragen. Es ist eine klare Tendenz zu erkennen. Die Spannungen der Programmierpulse nehmen ab, wohingegen die Spannungen der Löschpulse ansteigen. Dies ist ein deutlicher Hinweis auf die Veränderung der Elektronen- und Löcherverteilungen in der Nitridschicht. Wir nehmen wiederum das Modell aus Abschnitt 3.4.5 als Ausgangspunkt der Betrachtung. Obwohl die Stromkriterien, und somit die Bedingungen für die Einsatzspannung, für den programmierten und den gelöschten Zustand konstant sind, ändert sich das Programmier- und Löschverhalten. Das Löschen wird schwerer, es müssen höhere Drain-Spannungen angelegt werden, um die gleiche Wirkung zu erzeugen. Zugleich wird das Programmieren leichter, es werden geringere Spannungen benötigt.
Dies deutet auf eine Anhäufung von Löchern hin. Der Schwerpunkt dieser Verteilung liegt über dem Drain-Gebiet. Würde er im Kanalbereich liegen, so wäre die elektrische Wirksamkeit auf die Einsatzspannung der Löcher sehr hoch, und es könnte sich kein größeres Reservoir bilden. Diese Ansammlung von Löchern hat eine Änderung des lokalen elektrischen Feldes zur Folge. Elektronen erfahren eine stärkere Anziehung, somit genügen zum Programmieren geringere von außen angelegte Spannungen. Zugleich wird die Abstoßung von Löchern verstärkt, was eine weitere Löcherinjektion erschwert. Folglich steigen die Löschspannungen mit steigender Anzahl von Zyklen.

4.7.1 Vergleich der Degradation nach Zyklen

Eine Messung analog zu der in Abbildung 4.17 wurde an einer zweiten NROM-Zelle durchgeführt. Das Resultat ist in Abbildung 4.18 dargestellt.
Die beiden Zellen unterscheiden sich nur durch unterschiedliche Wannendotierungen. Die bereits zuvor besprochene Zelle (Abb. 4.17) hat eine mittlere Wannendotierung, wohingegen die Zelle aus Abbildung 4.18 eine sehr viel höhere Wannendotierung hat.
Beim Vergleich der beiden Ergebnisse fallen zwei wesentliche Unterschiede auf. Zum einen

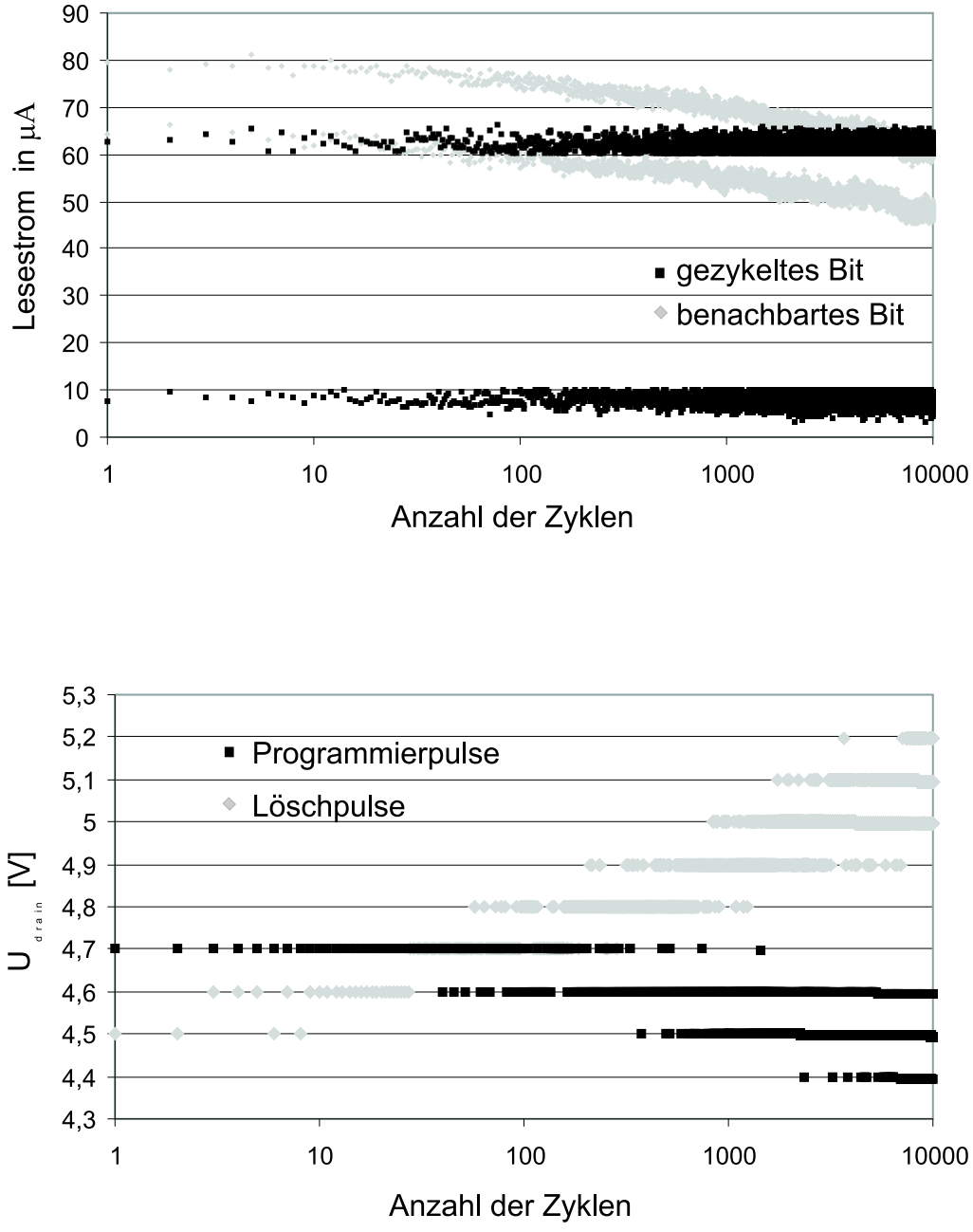


Abbildung 4.18: Zyklen-Messung an einer Zelle mit sehr hoher Wannendotierung; oben: Leseströme im Verlauf der Messung; unten: Spannungen der Programmier- bzw. Löschpulse

ist die Degradation des ungezykelten Bits im zweiten Fall (Abb. 4.18) wesentlich stärker. Der Lesestrom des benachbarten Bits sinkt deutlich stärker mit steigender Zyklenzahl ab. Zum anderen ist beim Vergleich der Degradation der Puls-Spannungen ein deutlich größerer Hub für die Zelle mit hohem N_A zu erkennen.
Beide Beobachtungen sind eng miteinander verknüpft. Die stärkere Degradation der Puls-Spannungen für Programmieren und Löschen weist auf größere Ladungsträgermengen hin, die sich in der Nitridschicht anhäufen. Diese größere Anzahl von Ladungsträgern führt zu einer Verbreiterung der Verteilungen. Beides zusammengenommen hat ein stärkeres Nebensprechen zur Folge. Somit ist eine stärkere Degradation des benachbarten Bits zu beobachten.
Erklären lässt sich dieses Verhalten mit der Injektion von Sekundärelektronen (siehe Kapitel 2.3.5). Das Auftreten von Sekundärelektronen nimmt mit steigender Wannendotierung zu, somit werden Elektronen jenseits der normalen Verteilung weiter über dem Kanal injiziert. Diese müssen an anderer Stelle durch eine größere Anzahl von Löchern elektrisch kompensiert werden. So kommt das beobachtete Verhalten zu Stande.
Die beiden hier gezeigten Zellen verhalten sich während der ersten Programmier- und Löschvorgänge sehr ähnlich. An Hand von Messungen an nicht gezykelten Zellen ist eine solche Unterscheidung sehr schwer. Betrachtet man jedoch zusätzlich die hier gezeigten Messungen, so ist klar, dass die Zelle mit sehr hoher Wannendotierung in ihrem Verhalten schlechter ist. Diese Art von Messungen bietet somit eine notwendige, zusätzliche Beurteilungsgrundlage für Qualität einer Zelle.

4.8 LVZ - Ladungsverlust nach Zykeln

Die Informationshaltung ist ein ganz wesentliches Merkmal einer Speicherzelle. Aus diesem Grund ist es wichtig zu zeigen, dass NROM-Zellen, die auf dem STI-Konzept basieren, ausreichend gute Eigenschaften auch bei diesem Parameter aufweisen.
Um die Eigenschaft hier zu untersuchen, wird eine Zelle 10.000 Mal programmiert und gelöscht. Danach wird ein Bit der Zelle programmiert, während das andere Bit gelöscht

bleibt. Nun wird die Zelle bei erhöhter Temperatur gelagert, um einen Zeitraum von z.B. 10 Jahren zu simulieren. Nach dieser Lagerung bei hoher Temperatur wird die Einsatzspannung des zuvor programmierten Bits gemessen. Die relevante Größe ist die Erniedrigung der Einsatzspannung während der Lagerung, sie wird hier mit LVZ (Ladungsverlust nach Zykeln) abgekürzt.
Das Ergebnis einer solchen Messung an einer STI-begrenzten Zelle ist in Abbildung 4.19 zu sehen.

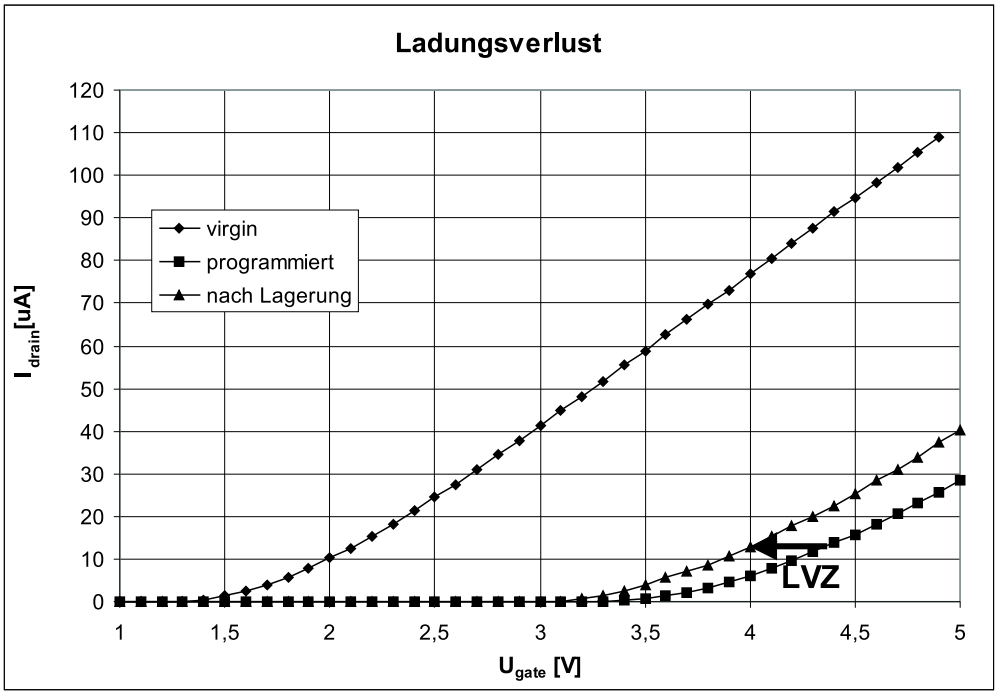


Abbildung 4.19: Ermittlung des Ladungsverlusts

Hier ist zusätzlich die Kennlinie im ursprünglichen Zustand der Zelle abgebildet (Rauten-Symbole). Diese dient dem Vergleich, wie stark die Zelle programmiert wurde, und wie stark der Verlust an Einsatzspannung während der Lagerung zu gewichten ist. Die Kurve mit den quadratischen Symbolen repräsentiert den Zustand nach dem Programmieren und die Kurve mit den dreieckigen Symbolen den Zustand nach Programmieren und Lagerung bei erhöhter Temperatur. Die Differenz der beiden letzten Kurven gibt das Absinken der

Einsatzspannung an und somit den LVZ. Die hier abgebildete Zelle weist einen Verlust von $350mV$ auf. Dies ist bezogen auf eine Programmierung von $\Delta U_{th} = 2V$ ein sehr gutes Ergebnis. Es zeigt die Funktionalität einer STI-begrenzten NROM-Zelle.

LVZ für unterschiedliche Wannedotierungen

Der gleiche Versuch wurde mit Zellen, die sich nur durch die Wannendotierung unterscheiden durchgeführt. Das Resultat ist in Abbildung 4.20 dargestellt.

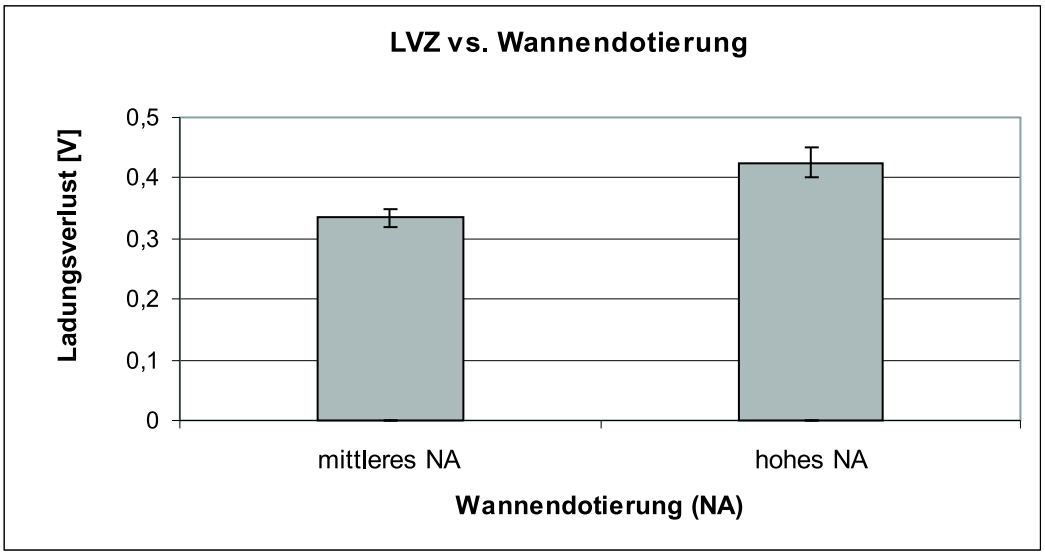


Abbildung 4.20: Ladungsverlust in Abhängigkeit von der Wannendotierung, N_A, gemessen an je 2 Zellen.

Für die Zellen mit mittlerer Dotierung wurden Werte von $320mV$ und $350mV$ gemessen. Bei hoher Bor-Dotierung wurden $400mV$ und $450mV$ erzielt. Wird die Wannenkonzentration zu hoch gewählt, so verschlechtert sich das LVZ-Verhalten. Dies ist nach den Ergebnissen der vorherigen Abschnitte nicht überraschend. Die Wannendotierung kann jedoch auf Grund anderer Effekte, wie z.B. des Nebensprechens, nicht beliebig abgesenkt werden.

4.9 Beweglichkeit von Ladungsträgern im ONO

Zweck der in diesem Abschnitt vorgestellten Experimente ist die Überprüfung des Modells, welches Ladungsverlust durch laterale Bewegung von Löchern erklärt (3.4.5). Für dieses Modell ist es von entscheidender Bedeutung, dass sich Löcher sehr viel einfacher in der Nitridschicht des ONO-Stapels bewegen können als Elektronen.

Zwei Experimente werden durchgeführt. Im ersten Versuch wird die Einsatzspannung von „virgin“ NROM-Zellen durch Injektion von Elektronen um $\sim 1V$ angehoben. Für den zweiten Versuch werden ebenfalls Zellen in ihrem ursprünglichen Zustand verwendet, sie werden Löschbedingungen ausgesetzt, bis die Einsatzspannung $\sim 450mV$ unter dem Ausgangszustand liegt. Es wird hier keine Differenz von $1V$ erreicht, da die Grenzen der Betriebsspannungen erreicht sind. Nach dieser Behandlung werden sowohl die Speicherzellen, in die nur Elektronen injiziert wurden, wie auch die Speicherzellen, in die nur Löcher injiziert wurden, bei 200°C eine Stunde lang gelagert. Im Anschluss werden die Einsatzspannungen der Zellen wiederum gemessen.

In Abbildung 4.21 sind die Ergebnisse für die programmierten Zellen zu sehen.

Auf der x-Achse ist die ursprüngliche Einsatzspannung aufgetragen. Die y-Achse zeigt die Einsatzspannung zu verschiedenen Phasen des Experiments. Es wurden vier Zellen gemessen. Somit liegen die „virgin“ U_{th}'s auf einer Geraden. Desweiteren sind die Einsatzspannungen der beiden Bits jeder Zelle nach dem Programmieren von Bit1 und nach der Lagerung bei erhöhter Temperatur aufgetragen. Das benachbarte Bit (Bit2) wird stets gemessen, um eine zusätzliche Aussage über die Veränderung der Ladungsträgerverteilungen zu erhalten.

Wie bereits erwähnt, wurde ein analoges Experiment für die ausschließliche Injektion von Löchern durchgeführt. Das Resultat ist in Abbildung 4.22 dargestellt.

Es ist zu betonen, dass diese Speicherzellen von ihrem ursprünglichen Zustand aus überlöscht werden. Es hat nie eine Injektion von Elektronen stattgefunden.

Kommen wir nun zum Vergleich der Ergebnisse. Bei dieser Untersuchung sind in erster Linie die Änderungen der Einsatzspannungen von Interesse und weniger deren Absolutwerte.

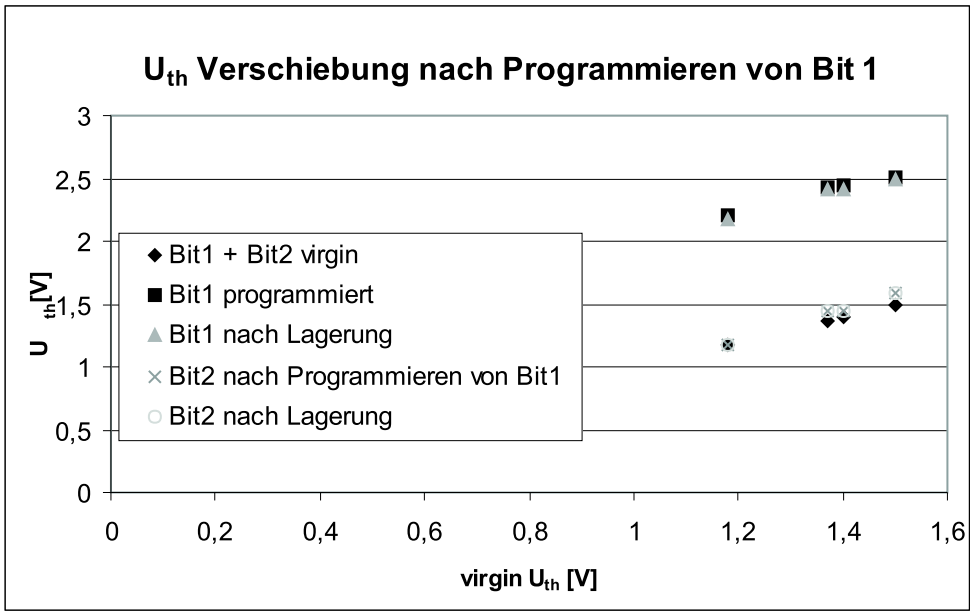


Abbildung 4.21: Beweglichkeit von gespeicherten Elektronen

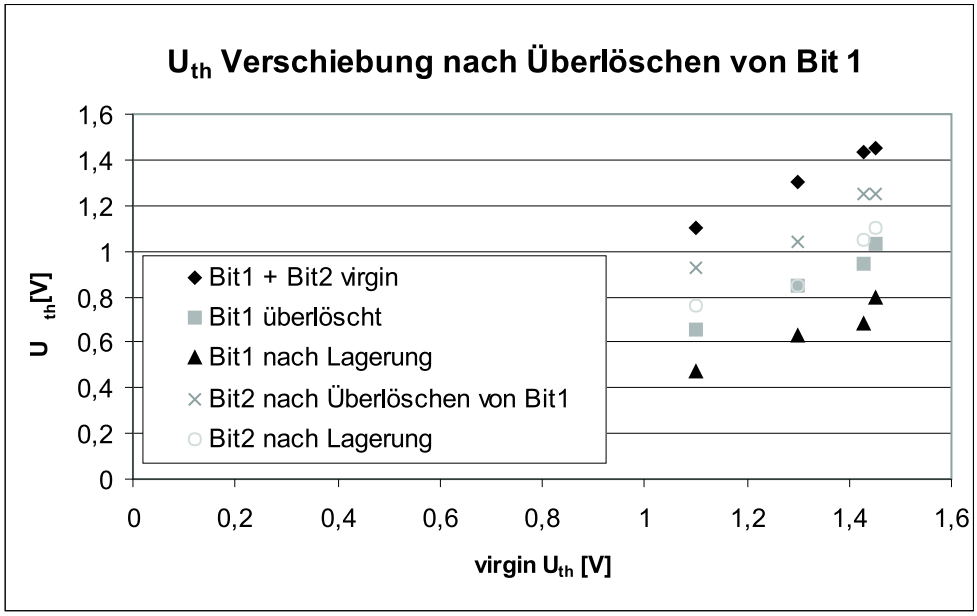


Abbildung 4.22: Beweglichkeit von gespeicherten Löchern

Da die Unterschiede zwischen den je vier gemessenen Zellen sehr gering sind, werden im Weiteren nur berechnete Mittelwerte betrachtet. Die Zusammenfassung der wesentlichen Messergebnisse ist in Tabelle 4.3 wiedergegeben.

Elektroneninjektion (vgl. Abb. 4.21)		
	ΔU_{th} nach Programmierung von Bit1	ΔU_{th} durch Lagerung (200°C, 1h)
Bit1	$1,04V$	$-0,03V$
Bit2	$0,07V$	$< 0,01V$
Löcherinjektion (vgl. Abb. 4.22)		
	ΔU_{th} nach Überlöschen von Bit1	ΔU_{th} durch Lagerung (200°C, 1h)
Bit1	$-0,45V$	$-0,23V$
Bit2	$-0,20V$	$-0,18V$

Tabelle 4.3: Ergebnisse zur Ladungsträgerbeweglichkeit

Zwei prinzipielle Unterschiede fallen sofort ins Auge. Zum einen wird beim Versuch zur Elektroneninjektion das benachbarte Bit fast gar nicht beeinflusst, wohingegen die Auswirkungen bei der Löcherinjektion für das Nachbarbit gravierend sind. Zum anderen ist die Differenz der Einsatzspannungen beider Bits bei den überlöschten Zellen nach der Lagerung bei erhöhter Temperatur sehr viel größer, als bei den programmierten Zellen.
Beim Programmieren ändert sich fast nur die Einsatzspannung der gewünschten Bits. Nach der Lagerung ist für das programmierte Bit ein Verlust von $30mV$ zu verzeichnen, was sehr wenig ist. Das benachbarte Bit ändert seine Einsatzspannung nur unwesentlich. Daraus resultiert, dass weder eine erhebliche Anzahl von Elektronen aus dem ONO entweicht, noch dass sich die Verteilung der Elektronen merklich verändert.
Die überlöschten Zellen verhalten sich anders. Bei der Injektion der Löcher ändert sich die Einsatzspannung des benachbarten Bits erheblich. Dies deutet auf eine breite Verteilung der Löcher hin. Betrachtet man das Verhalten nach der Lagerung, so stellt man fest, dass nicht nur die Einsatzspannung des überlöschten Bits weiter absinkt, sondern dass auch die Einsatzspannung des benachbarten Bits weiter absinkt und dies fast im gleichen Ma-

ße. Die Änderungen nach der Lagerung lassen sich durch eine Wanderung von Löchern, die sich über dem Drain-Gebiet angesammelt haben, in Richtung Kanal erklären. Es ist naheliegend, dass das Maximum der injizierten Löcherverteilung über dem n^+ - Gebiet liegt. Zerfließt diese Verteilung, so gelangen mehr Löcher über den Kanal und entfalten eine stärkere Auswirkung auf die Einsatzspannung. Daher sinkt die Einsatzspannnung des überlöschten Bits. Eine Wanderung von Löchern weiter in den Kanal hat zugleich auch ein weiteres Absinken der Einsatzspannung des benachbarten Bits zu Folge. Diese Auswirkungen sind klar im Experiment zu erkennen.
Die Gegenüberstellung der beiden Versuche liefert ein klares Ergebnis, die Elektronen bewegen sich sehr viel weniger als die Löcher.

4.10 Temperaturabhängigkeit von NROM-Zellen

In diesem Abschnitt wird die Temperaturabhängigkeit von NROM-Zellen betrachtet. Messkurven für drei verschiedene Programmierzustände und je drei Temperaturen sind in Abbildung 4.23 dargestellt.

Für diesen Versuch wurde ein Zelle verwendet, die in zwei Stufen programmiert wurde. So können Schwankungen zwischen verschiedenen Speicherzellen ausgeschlossen werden. Wie erwartet, steigen die Ströme im Bereich der schwachen Inversion für steigende Temperaturen an. Hier unterscheiden sich die verschiedenen Programmierzustände nicht prinzipiell.
Da die Informationsspeicherung bei NROM über die Einstellung der Einsatzspannung erfolgt, wird deren Temperaturabhängigkeit genauer betrachtet. Hierzu ist in Abbildung 4.24 die Wurzel des Drain-Stromes über die Gate-Spannung aufgetragen.

Man sieht, dass die Temperaturabhängigkeit im unprogrammierten Zustand deutlich geringer ist als in programmierten Zuständen. Die extrahierten Werte in Tabelle 4.4 unterstreichen dies.

Um einen Vergleichswert für den unprogrammierten Fall zu erhalten, werden die Gleichungen eines normalen MOS-Transistors betrachtet. Ohne Berücksichtigung von Kurzkanalef-

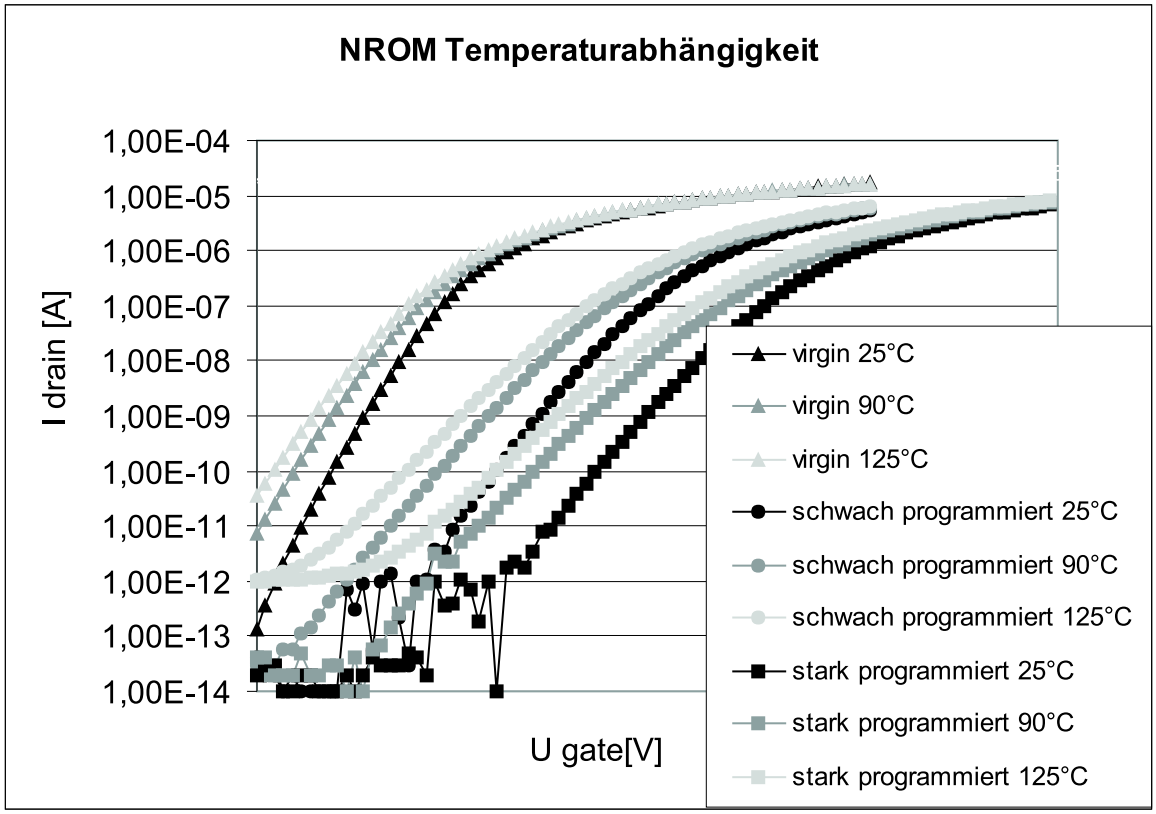


Abbildung 4.23: Temperaturverhalten einer NROM-Speicherzelle für verschiedene Programmierzustände im logarithmischen Maßstab

fekten, erhält man für die Temperaturabhängigkeit der Einsatzspannung, [24]:

$$\frac{dU_{th}}{dT} = \left(\frac{1}{T}\left(\phi_F(T) - \frac{W_g(T)}{2q}\right) - \frac{3k}{2q}\right)\left(2 + \frac{1}{C'_{ox}}\sqrt{\frac{qN_A\epsilon_0\epsilon_{Si}}{\phi_F(T)}}\right) \tag{4.4}$$

Die Änderung des Bandabstandes lässt sich näherungsweise angeben mit, [33]

$$\Delta W_g = -2.4 \cdot 10^{-4}\frac{eV}{K} \cdot \Delta T \tag{4.5}$$

Die Fermispannung ist nicht nur über die Temperaturspannug ϕ_t abhängig von T, sondern auch über die Intrinsicdichte. Diese ist gegeben durch:

$$n_i(T) = \sqrt{N_C N_V} \cdot \exp\left(-\frac{W_g}{2kT}\right) \tag{4.6}$$

Die äquivalenten Zustandsdichten N_C und N_V haben eine Temperaturabhängigkeit der Form $T^{\frac{3}{2}}$. Für sie werden bei $T = 300K$ die Werte $N_C = 2,8 \cdot 10^{-19}cm^{-3}$ und $N_V =$

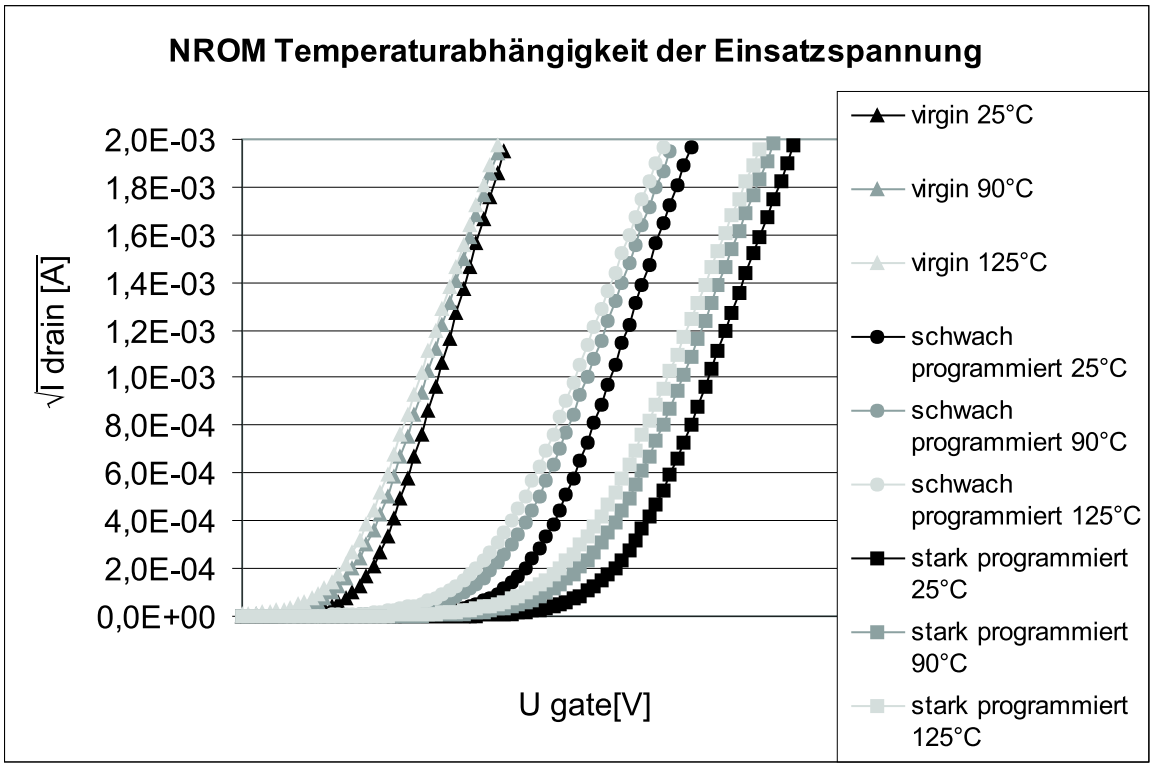


Abbildung 4.24: Temperatursensitivität der Einsatzspannung einer NROM-Speicherzelle für verschiedene Programmierzustände im logarithmischen Maßstab

$1,04 \cdot 10^{-19} cm^{-3}$ angenommen, [24]. Zur Bestimmung der Wannendotierung wird ein Fit der „virgin“ Kurve, siehe Abschnitt 3.4.6, vorgenommen.

Somit lässt sich die Temperaturabhängigkeit zu $-1.8 mV/°C$ berechnen. Dies ist für den Mittelwert des betrachteten Temperaturintervalls bei $T = 75°C$ gerechnet. Berücksichtigt man die Einfachheit der verwendeten Formeln und die effektive Kanallänge von $150nm$ des verwendeten Devices, so ist die Übereinstimmung gut.

Wie lässt sich die erhöhte Temperatursensitivität der programmierten Zustände erklären? Diese Frage kann hier noch nicht abschließend beantwortet werden. Im Folgenden werden mögliche Erklärungen besprochen.

Gleichung (4.4) liefert hier keine ausreichende Erklärung. Ein sehr stark vereinfachender Ansatz wäre, die erhöhte Einsatzspannung der programmierten Zustände durch eine

	U_{th} Verschiebung in $mV/°C$
„virgin“	-2.1
schwach programmiert	-3.1
stark programmiert	-3.0
berechneter Wert für „virgin“	-1.8

Tabelle 4.4: Temperaturabhängigkeit der Einsatzspannung

höhere Wannendotierung zu beschreiben. Gleichung (4.4) zeigt zwar, dass die Temperaturabhängigkeit mit steigender Wannendotierung zunimmt, jedoch ist diese Zunahme sehr viel geringer als der beobachtete Effekt. Bei näherem Hinsehen ist verständlich, dass dieser Ansatz physikalisch nicht zu rechtfertigen ist. Die lokal injizierten Elektronen im ONO haben eine völlig andere Wirkung auf den Ladungsträgertransport in der Zelle als eine erhöhte Wannendotierung. Dies verdeutlicht die Simulation in Abbildung 4.25.

Zur Nachbildung der programmierten Zelle ist ein Ladungsträgerpaket in der Simulation eingefügt. Die gewählten Bedingungen sind nahe der Einsatzspannung. Es ist deutlich zu erkennen, dass der Elektronenstrom in die Tiefe ausweicht. Es fließt kein oberflächennaher Strom dort, wo Ladungsträger im ONO platziert sind.

Eine anschauliche Erklärung basiert auf der Betrachtung der schematisch dargestellten Raumladungszonen bei Lesebedingungen in Abbildung 4.26.

Durch die injizierte Ladung beginnt der Strom nicht wie bei einer unprogrammierten Zelle an der Oberfläche zu fließen, sondern in der Tiefe. Wenn der Abstand d_m gegen Null geht, beginnt ein Strom zu fließen. Dies basiert auf der gleichen Überlegung wie das Zwei-Transistor-Modell in Abschnitt 3.4.6. Hier wird die programmierte NROM-Zelle durch zwei Transistoren modelliert. Der Kanalbereich, über den Elektronen im ONO gespeichert sind, wird durch einen sehr kurzen Transistor mit sehr hoher Einsatzspannung repräsentiert. Dieser Transistor ist durch extreme Kurzkanaleffekte gekennzeichnet. Die Temperaturabhängigkeit eines solchen Elements ist deutlich größer als die eines Langkanaltransistors. Als Indikator dient die Degradation des Swing, diese ist ausführlich in der Literatur be-

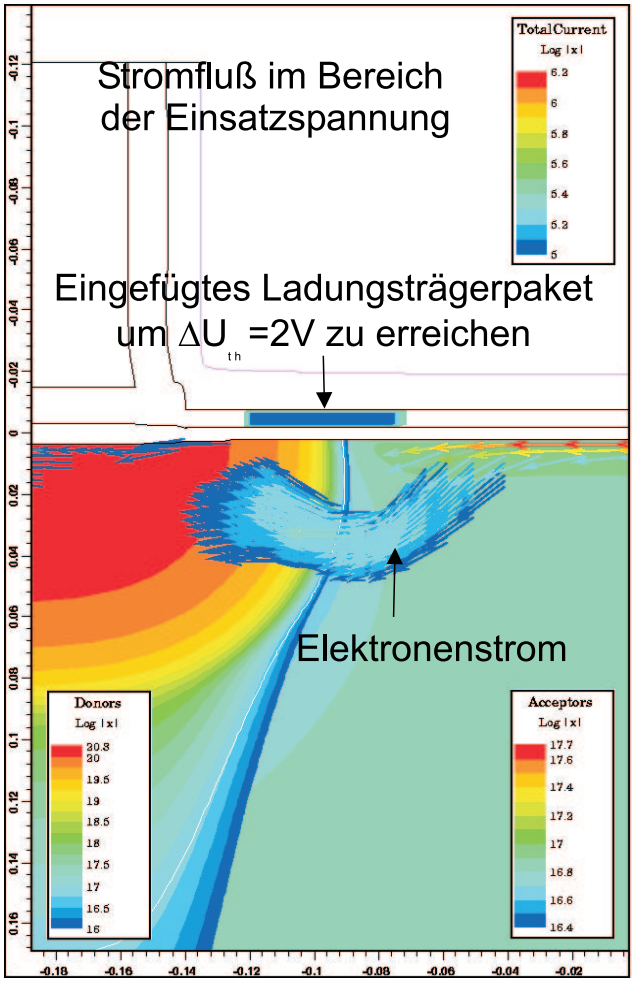


Abbildung 4.25: Simulation der Lesestromverteilung einer programmierten Zelle im Bereich der Einsatzspannung, [22].

handelt worden, [59]. Die Verringerung des Swing für höhere Programmierzustände sieht man in Abbildung 4.23. Aus diesem Grund steigt die Temperaturabhängigkeit der Einsatzspannung von programmierten NROM-Speicherzellen gegenüber unprogrammierten Zellen deutlich an.

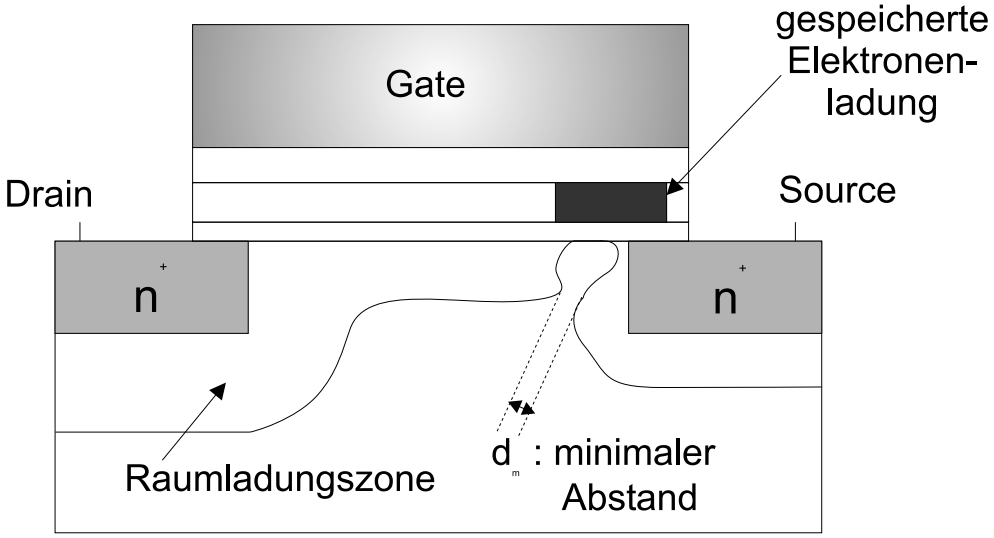


Abbildung 4.26: Erklärungsmöglichkeit für die erhöhte Temperaturabhängigkeit von programmierten NROM-Zellen.

Kapitel 5

Multilevel NROM

In diesem Kapitel wird eine neue Betriebsweise für NROM vorgestellt, die besonders geeignet für eine Multilevel-Anwendung ist. Diese neue Betriebsweise hat zudem weitere wesentliche Vorteile. Sie reduziert die negativen Auswirkungen des Nebensprechens in einer NROM Zelle drastisch und zieht aus dem Nebensprechen als solches sogar Nutzen. So ist eine effektive Vergrößerung des verwendbaren Einsatzspannungsfensters möglich. Bisher wird das Nebensprechen als wesentliches Hindernis für die Miniaturisierung von NROM gesehen, diese Hürde wird überwunden.
Zur Erläuterung der neuen Betriebsweise, und um ihre Vorteile zu belegen, wird zunächst ein Vergleich mit dem herkömmlichen NROM-Betrieb für zwei Bits pro Zelle durchgeführt. Abschließend wird am Beispiel von drei Bits pro Zelle die Multilevel-Tauglichkeit des neuen Betriebsschemas gezeigt.

5.1 Herkömmliche Betriebsweise für NROM

Um die Neuerung einfacher erklären zu können, wird zuvor die herkömmliche Betriebsweise bei NROM (siehe z.B. [5]) veranschaulicht. Hier werden zwei Bits pro Zelle gespeichert. Zur Informationsspeicherung werden die Einsatzspannungen der beiden Seiten einer Zelle verwendet, schematisch ist dies in Abbildung 5.1 dargestellt.

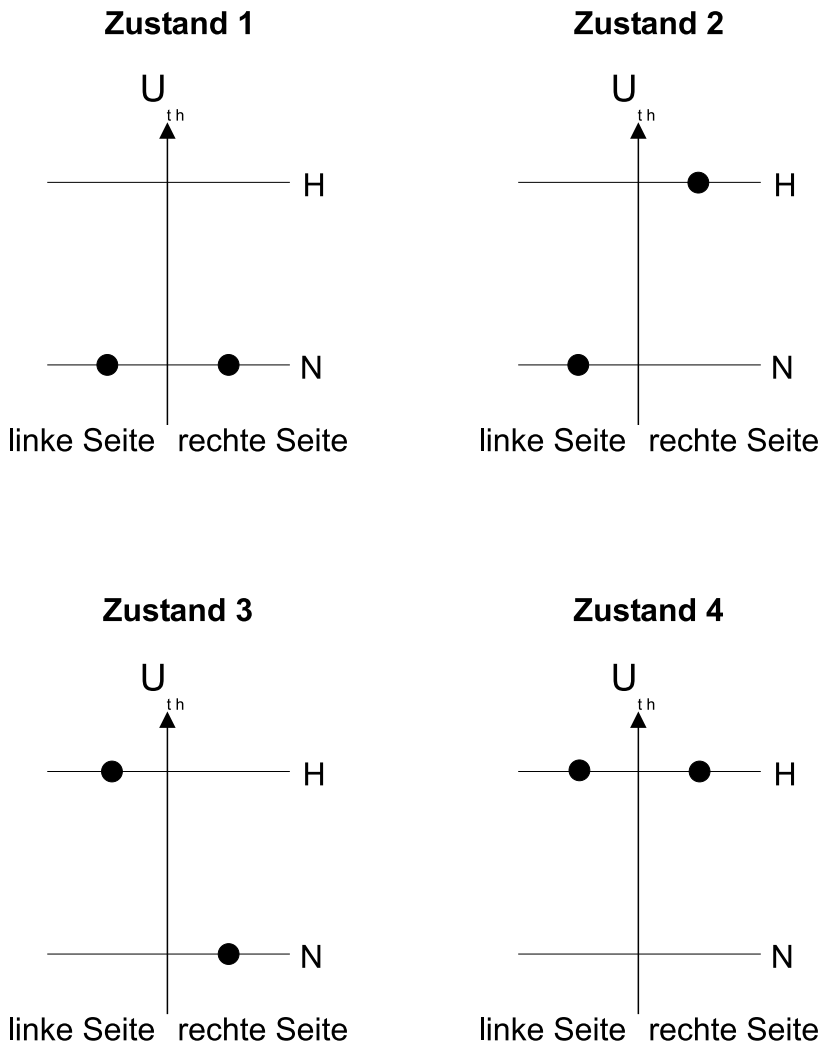


Abbildung 5.1: Gebräuchliche Betriebsweise für zwei Bits pro Zelle bei NROM

Es sind die vier Zustände, die die zwei Bits definieren, dargestellt. Für jeden Zustand sind die Einsatzspannungen der beiden Seiten einer NROM-Zelle aufgetragen. Es sind je zwei Niveaus definiert, niedrige Einsatzspannug (N) und hohe Einsatzspannung (H). Die Differenz zwischen H und N liegt üblicherweise im Bereich zwischen ein und zwei Volt. Man sieht, dass jede der beiden Seiten für sich entweder eine hohe oder eine niedrige Einsatzspannung hat. Die vier Kombinationen ergeben die Zustände. Hier kann also die Bedeutung von einem Bit mit einer physikalischen Seite gleichgesetzt werden. Das Bit auf jeder Seite kann unabhängig von der benachbarten Seite betrachtet werden. Dies ist die

gebräuchliche Betriebsweise für zwei Bits pro Zelle bei NROM.

Diese Definition hat einen inhärenten Nachteil, es gibt Zustände bei denen eine große Differenz zwischen den Einsatzspannungen der beiden Seiten existiert (Zustand 2 und Zustand 3). Dies ist problematisch, da es mindestens zwei Mechanismen gibt, die das Fenster zischen H und N reduzieren. Zum einen sinkt der H-Zustand bei langer Lagerzeit ab. Dies ist unvermeidlich. Zum anderen steigt beim Programmieren der einen Seite auf das H-Niveau die Einsatzspannung der anderen Seite, die auf dem N-Niveau bleiben soll, mit an. Dies wird als Nebensprechen bezeichnet und wird ausführlich in Abschnitt 4.5 behandelt. Somit verliert man nochmal einige hundert Millivolt dadurch, dass die Einsatzspannung der Seite, die auf dem N-Niveau verbleiben soll, ansteigt. Dies wird in Abbildung 5.2 noch einmal deutlich.

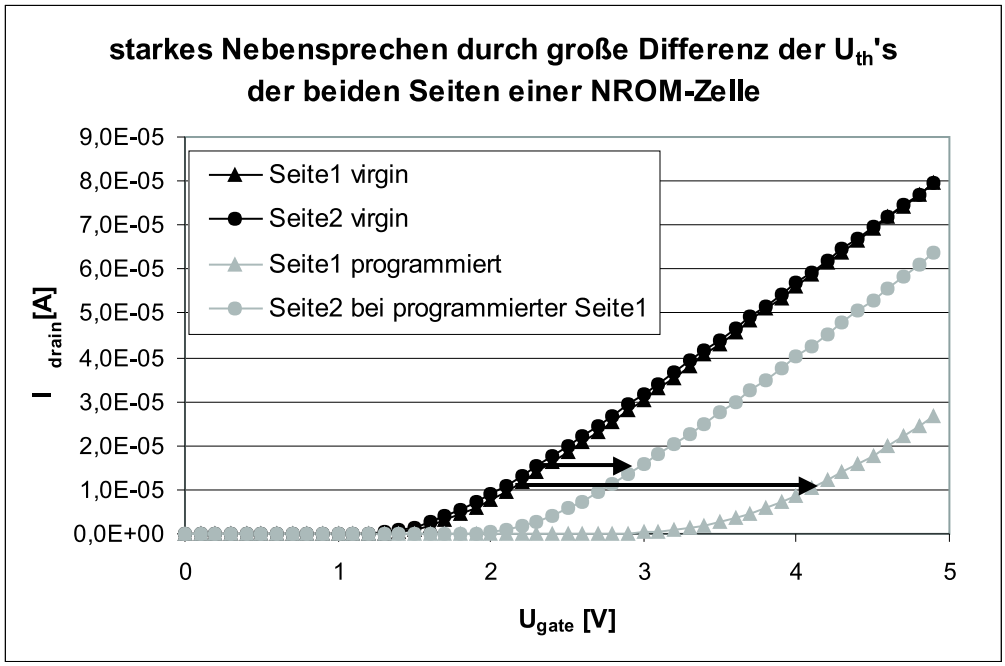


Abbildung 5.2: Auswirkung des Nebensprechens bei der herkömmlichen Betriebsweise für NROM

Es sind die Transferkennlinien für beide Seiten der Zelle abgebildet. Die mit „virgin“ bezeichneten Kurven sind an einer Zelle im ursprünglichen Zustand gemessen. Danach wurde

die Seite 1 programmiert. Die Erhöhung der Einsatzspannung von Seite 1 ist also gewollt, dabei lässt es sich aber nicht verhindern, dass die Einsatzspannung von Seite 2 ebenfalls ansteigt. Dies ist eine Störung. Überträgt man diesen Effekt in die Darstellungsweise von Abbildung 5.1, so gelangt man zu der Veranschaulichung in Abbildung 5.3.

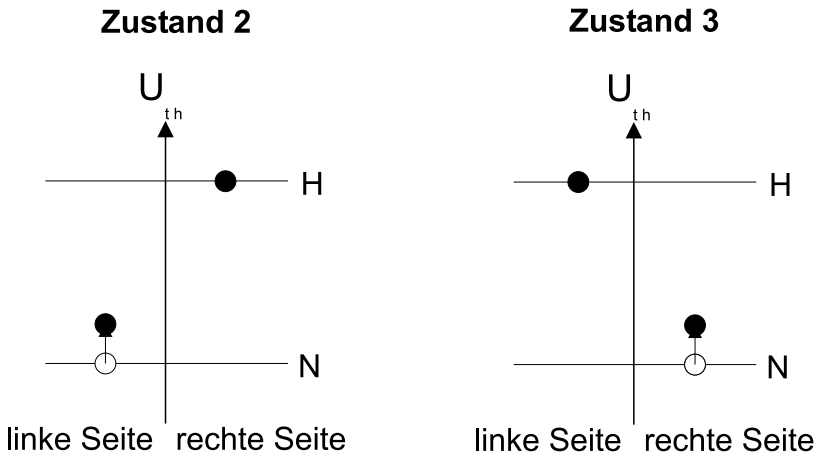


Abbildung 5.3: Veranschaulichung des Nebensprechens auf die Zustände 2 und 3 aus Abbildung 5.1

Durch das Nebensprechen wird der reale Unterschied zwischen H- und N-Niveau reduziert. Je größer die Differenz der Einsatzspannungen der beiden Seiten in einer Zelle ist, desto größer ist auch die Auswirkung des Nebensprechens. Daher sind die Zustände 2 und 3 ungünstig.

Das Nebensprechen gilt als wesentliches Hindernis bei NROM, um die effektive Kanallänge in Zukunft deutlich zu reduzieren. Je kürzer die Kanallänge, desto weiter nähern sich die Ladungen der beiden Seiten einander an, folglich wird das Nebensprechen immer stärker. Dies ist ein Problem für zukünftige Zell-Generationen.

5.2 Multilevel-Betrieb für NROM

Die neue Betriebsweise reduziert die Auswirkung des Nebensprechens drastisch und ist multilevel-tauglich. Zudem bietet sie noch weitere neue Möglichkeiten zur Verbesserung

der Informationshaltung. Es ist Grundidee dieser Betriebsweise, die Einsatzspannungsunterschiede der beiden Seiten einer NROM-Zelle so gering wie möglich zu halten.
Aus diesem Ansatz folgt die, in Abbildung 5.4 dargestellte, neue Betriebsweise, die nachfolgend am Beispiel für zwei Bits pro Zelle bei NROM veranschaulicht wird.

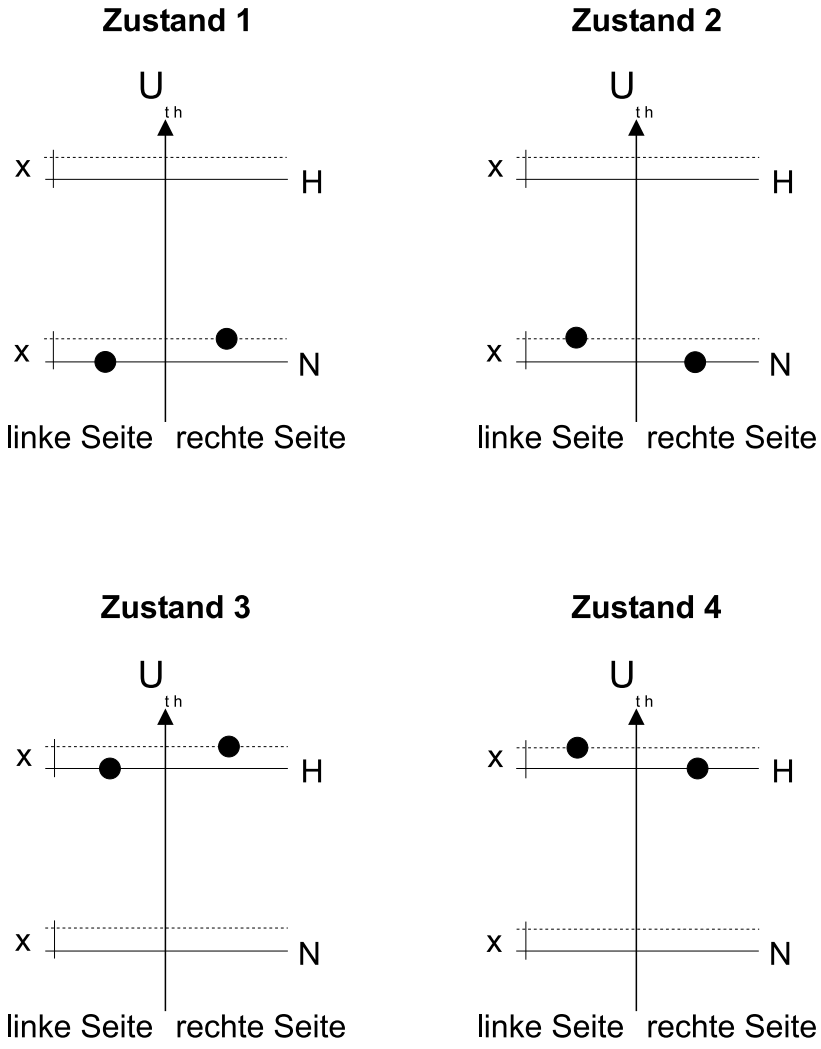


Abbildung 5.4: Neuer Multilevel-Betrieb für zwei Bits pro Zelle bei NROM

Das hohe und niedrige Niveau werden jeweils beibehalten. Zudem werden zwei neue Niveaus eingeführt, die je um einen Abstand x (z.B. $300mV$) oberhalb von N bzw. H liegen. Bei den neuen vier Zuständen sind die physikalischen Seiten nicht mehr identisch mit den Bits. Zum Auslesen müssen beide Seiten der Zelle betrachtet werden. Die erste Information

steckt in der Differenz der Einsatzspannungen beider Seiten; entweder ist die Einsatzspannung der rechten oder der linken Seite höher. Die zweite Information ergibt sich aus der absoluten Lage der beiden Seiten. Entweder sind die Einsatzspannungen beider Seiten auf einem hohen oder auf einem niedrigen Niveau. So lassen sich, wie in Abbildung 5.4, vier Zustände definieren, wobei nie eine große Differenz in den Einsatzspannungen der beiden Seiten auftritt. Somit wird die negative Auswirkung des Nebensprechens fast völlig unterdrückt.
Eine wichtige Annahme, die hinter der neuen Definition steht, liegt darin, dass bei langjähriger Lagerung der Verlust auf beiden Seiten der Zelle gleichförmig vonstatten geht. Es darf trotz der geringen Differenz der beiden Seiten nicht vorkommen, dass sich die relative Lage der beiden Seiten zueinander umkehrt. Die Validität dieser Annahme wird durch Experimente untermauert.
Der Multilevel-Betrieb geht noch einen Schritt weiter. Er minimiert nicht nur die negative Auswirkung des Nebensprechens, er zieht sogar direkten Nutzen aus dem Nebensprechen in der Zelle.
Resultat:

- geringerer Ladungsverlust (LVZ)
- Vergrößerung des nutzbaren U_{th}-Fensters

Diese messbare Verbesserung hängt direkt mit der Definition der Zustände zusammen. Betrachtet man die Zustände 2 und 3 der herkömmlichen Betriebsweise, so wird nur eine Seite der Zelle programmiert. In diese Seite muss eine Elektronenladung injiziert werden, die einen Einsatzspannungshub von z.B. $2V$ verursacht. Der Ladungsverlust dieser Seite ist von der Menge der injizierten Ladungsträger abhängig, je mehr Elektronen, desto größer der LVZ. Für einen vergleichbaren Multilevel-Betrieb liegt das H-Niveau $2V$ über dem N-Niveau. Den kritischen Fall für den Ladungsverlust stellen Zustände mit hohen Einsatzspannungen dar, also die Zustände 3 und 4 aus Abbildung 5.4. Beide Seiten haben hier hohe Einsatzspannungen. Hier genügt auf jeder Seite eine injizierte Ladungsmenge, die einem U_{th}-Hub von $\sim 1.5V$ entspricht. Die Ladungen jeder Seite bewirken über das

Nebensprechen eine Erhöhung der Einsatzspannung auf der jeweils anderen Seite. So wird trotz geringerer Ladungsmengen der gewünschte U_{th}-Hub von $\sim 2V$ gemessen. Auf Grund der geringeren Ladungsmenge im Vergleich mit dem herkömmlichen Betrieb fällt der Ladungsverlust geringer aus. Hier eröffnen sich neue Optimierungswege.
Der geringere LVZ bei gleichem gemessenen U_{th}-Hub und die Ausnutzung des Nebensprechens können zu einer Vergrößerung des nutzbaren U_{th}-Fensters verwendet werden.
Die neue Betriebsart bietet zwei weitere Möglichkeiten, eine vergrößerte Spanne zur Unterscheidung von H- und N-Niveau zu erzielen.
Die erste Möglichkeit besteht in einem zusätzlichen Leseschritt. Beim Auslesen der Einsatzspannung einer Seite wird üblicherweise eine Drain-Source-Spannung im Bereich zwischen $1V$ und $2V$ gemessen. Dies ist notwendig, um eine Trennung der beiden Seiten beim Lesen zu erzielen. Dadurch, dass bei der neuen Betriebsweise die Einsatzspannungen beider Seiten immer entweder gleichzeitig hoch oder gleichzeitig niedrig sind, bietet sich eine neue Möglichkeit H und N zu unterscheiden. Es wird ein zusätzliches Lesen bei niedriger Drain-Source-Spannung (z.B. $0,1V$) eingeführt. Ein solches Lesen ist nicht nur auf die injizierte Ladung einer Seite sensitiv, sondern auf die Ladungen beider Seiten. Da bei der neuen Betriebsweise, im Gegensatz zur herkömmlichen, stets beide Seiten das generelle Niveau gemeinsam haben, ergibt sich bei einem solchen Leseschritt ein größerer Unterschied zwischen H- und N-Niveau.
Die zweite Möglichkeit ist schaltungstechnischer Natur. Für die neue Betriebsweise muss stets die Differenz in der Zelle bewertet werden, sowie die absolute Lage der Einsatzspannungen. Zur Bewertung der Differenz ist ein zweistufiges Auslesen notwendig, da eine Zelle nicht in beide Richtungen gleichzeitig gelesen werden kann. Also muss die Information jeder Seite durch das Beladen einer Kapazität zwischengespeichert werden. Dies bietet die Möglichkeit, dass zur Bewertung der absoluten Lage der Einsatzspannung der Summenstrom von beiden Seiten detektiert wird. Hierdurch wird der Stromunterschied, also auch das effektive U_{th}-Fenster, zwischen H- und N-Niveau annähernd verdoppelt. Damit kann die Informationshaltung in der Zelle weiter verbessert werden.

5.3 Experimenteller Vergleich der Betriebsweisen für zwei Bits pro Zelle

Dieser Abschnitt dient der Überprüfung und Bewertung der neuen Betriebsweise für zwei Bits pro Zelle.
Um einen Vergleichspunkt für die neue Methode zu bekommen, sind NROM-Zellen nach herkömmlicher Betriebsweise auf dem gleichen Wafer vermessen worden. Für diese Zellen wurde ein Nebensprechen von $300mV$ gemessen. Nach $10K$ Zyklen wurde ein Ladungsverlust von $600mV$ bestimmt, wobei das Fenster $1,6V$ (Differenz zwischen programmiertem und gelöschtem Zustand) betrug. Subtrahiert man Nebensprechen und LVZ, so bleiben nur $\sim 700mV$ übrig.
Kommen wir nun zur Überprüfung der neuen Methode. Auch hier werden die Zellen zuerst $10K$ gezykelt. Danach werden sie gemäß der Definition der Zustände aus Abbildung 5.4 programmiert. Die Größe x wird zu $300mV$ gewählt. Der Abstand zwischen hohen bzw. niedrigen Zuständen des N- und des H-Niveaus ist $1,6V$. Es werden nur die Zustände eins und drei betrachtet, da die beiden anderen Zustände spiegelsymmetrisch zu diesen sind. Das Verhalten des hohen Zustandes (H-Niveau) ist in Abbildung 5.5 zu sehen.

Das erste wesentliche Resultat dieser Messung ist, dass die beiden Seiten ihre relative Lage zueinander, auch nach der Lagerung bei erhöhter Temperatur (200°C, 1h), beibehalten. Die Information, die aus der relativen Lage der beiden Seiten gewonnen werden muss, bleibt erhalten. Das zweite positive Ergebnis ist, dass der LVZ der beiden Seiten im Bereich $400 \ldots 450mV$ liegt, und somit geringer ist, als der Wert für die herkömmlich behandelten Vergleichszellen mit $600mV$. Dies bestätigt die Betrachtungen aus Abschnitt 5.2.
Die Ergebnisse der Messung zum N-Niveau sind in Abbildung 5.6 dargestellt.

Für das N-Niveau ist genauso, wie zuvor für das H-Niveau, festzustellen, dass die relative Lage der Einsatzspannungen der beiden Seiten zueinander erhalten bleibt. Der Verlust der Einsatzspannungen im N-Niveau ist im Bereich zwischen $250mV$ und $400mV$.
Betrachten wir wiederum die Auswirkung auf die verbleibende Fenstergröße. Da die Zustände

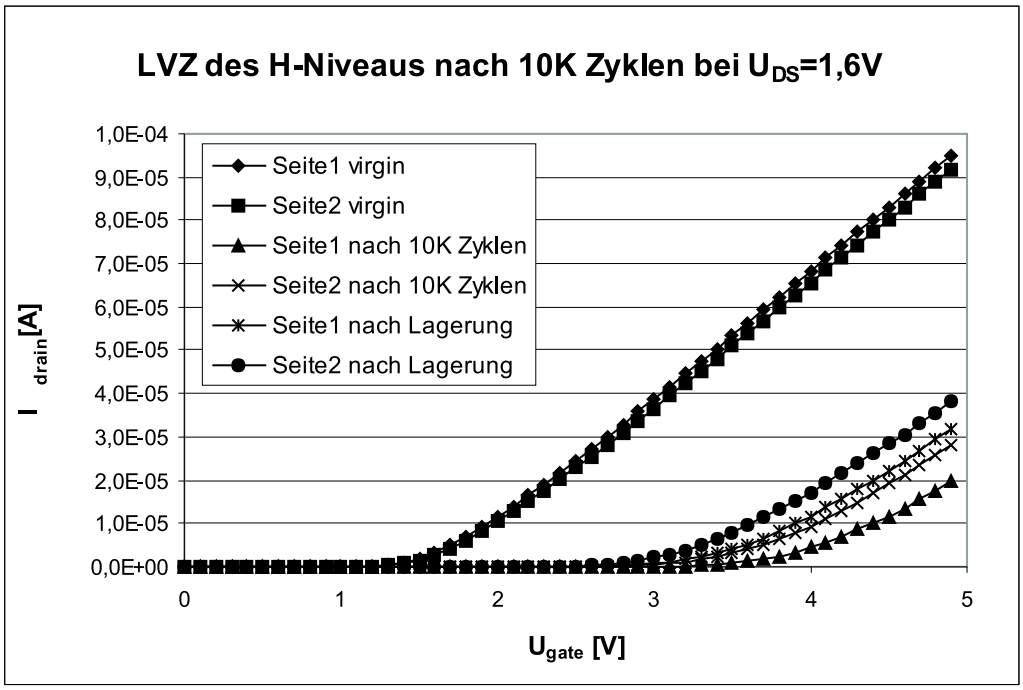


Abbildung 5.5: LVZ des H-Niveaus bei Multilevel-Betrieb für zwei Bits pro Zelle

alle definiert eingestellt werden und keine erheblichen Differenzen zwischen den Einsatzspannungen der beiden Seiten einer Zelle liegen, entfällt das Nebensprechen als Verlustmechanismus. Zudem beträgt der maximale LVZ des H-Niveaus nach Lagerung nur $450mV$. Für die verbleibende Fenstergröße wird das N-Niveau nach Zykeln, jedoch vor Lagerung, mit dem H-Niveau nach Lagerung verglichen. Nimmt man nun die minimale Differenz zwischen je höheren bzw. niedrigeren Zuständen der beiden Niveaus, so verbleibt eine Fensteröffnung von $\sim 1V$. Aus den im Text erwähnten Zahlen käme man auf einen Wert, der sogar um $150mV$ höher läge. Die Diskrepanz wird dadurch verursacht, dass es beim Einstellen der Einsatzspannungen des N-Niveaus zu Werten kommen kann, die leicht oberhalb der Zielwerte liegen. Trotzdem verbleibt ein Fenster, das um $300mV$ oder $42,86\%$ größer ist als bei der herkömmlichen Methode. Dies zeigt die Funktionalität und den Nutzen des Multilevel-Betriebs am Beispiel von zwei Bits Pro Zelle.

Als weitere Option im Rahmen der neuen Betriebsweise wurde, zusätzlich zu den obigen

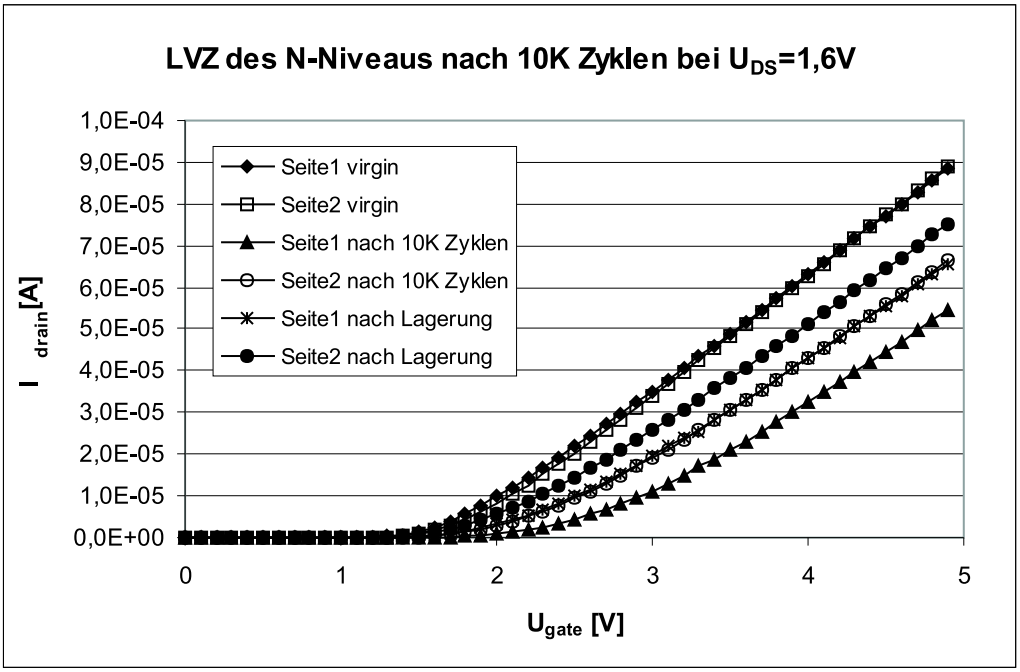


Abbildung 5.6: LVZ des N-Niveaus bei Multilevel-Betrieb für zwei Bits pro Zelle

Messungen, das Lesen bei niedriger Drain-Source-Spannung durchgeführt. Die Kurven sind in Abbildung 5.7 zu sehen.

Man sieht, dass die Unterschiede beim Auslesen der beiden Seiten für ein Niveau sehr gering sind. Dies ist Folge der geringen Spannung U_{DS}. Der Ladungsverlust des H-Niveaus ist mit $\sim 700mV$ deutlich größer als er bei hoher Spannung U_{DS} ist. Dafür misst man jedoch auch ein mit $\sim 2V$ deutlich größeres Fenster. Zudem wirkt sich leichtes Überprogrammieren des N-Niveaus auf Grund der mittelnden Wirkung der niedrigen Drain-Source-Spannung nicht so stark aus. Somit verbleibt ein Fenster von $\sim 1,3V$. Dies ist wiederum eine deutliche Verbesserung. Im Vergleich zum herkömmlichen Fall ist das Fenster um $600mV$ oder $85,7\%$ größer.

Ein weiterer Vorteil aus der Messung bei niedriger Drain-Source-Spannung ist, dass es nicht mehr notwendig ist, jeweils die Seiten mit höherer oder niedrigerer Einsatzspannung für die beiden Niveaus zu vergleichen. Es ist egal, in welche Richtung gelesen wird. Bei der Angabe

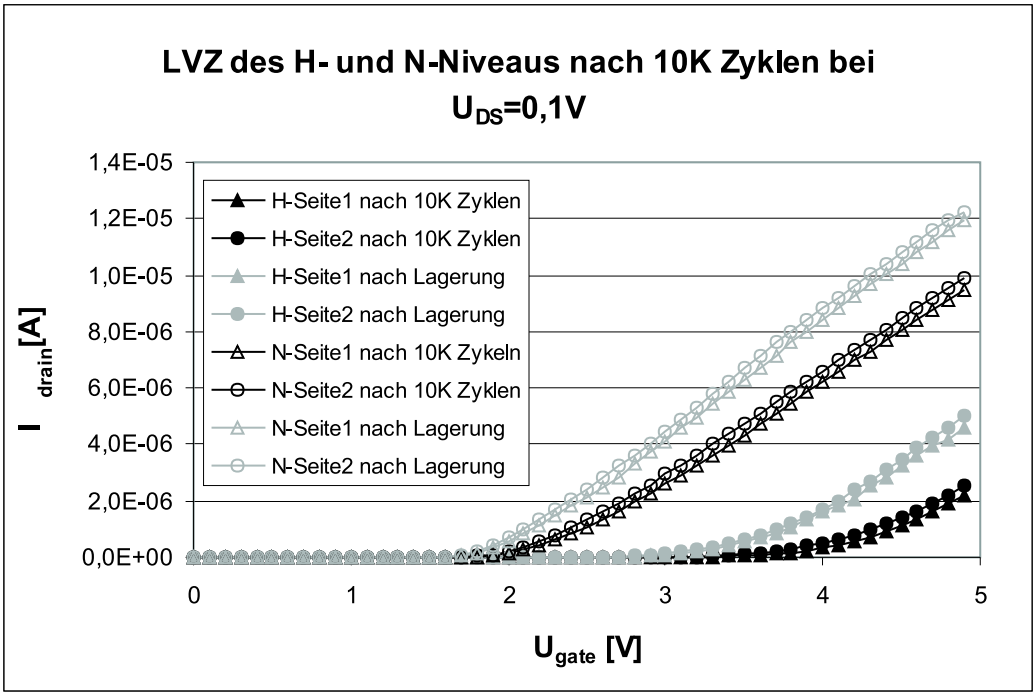


Abbildung 5.7: Margin Gain durch zusätzlichen Leseschritt bei niedriger Drain-Source-Spannung

der verbleibenden Fenstergröße von $\sim 1,3V$ ist dies berücksichtigt. Sie wurde zwischen den Kurven 'H-Seite2 nach Lagerung' und 'N-Seite1 nach 10K Zyklen' aus Abbildung 5.7 bestimmt.

Die Experimente bestätigen den prognostizierten Gewinn durch die neue Betriebsweise bei zwei Bits pro Zelle.

5.4 Multilevel am Beispiel von drei Bits pro Zelle

Nachdem gezeigt wurde, dass die Multilevel-Betriebsweise funktioniert und im Vergleich mit der herkömmlichen Betriebsweise deutliche Vorteile bietet, wird nun die Funktionstauglichkeit für höhere Speicherdichten untersucht.

Die Multilevel-Tauglichkeit wird hier am Beispiel von drei Bits pro Zelle untersucht. Für drei Bits pro Zelle wird der in Abbildung 5.4 dargestellte Multilevel-Betrieb lediglich an-

gepasst. Dies ist in Abbildung 5.8 zu sehen.

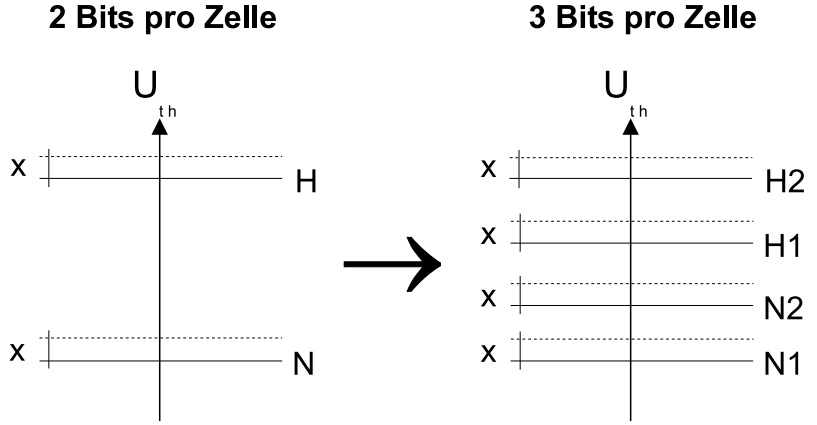


Abbildung 5.8: Von 2 Bits zu 3 Bits pro Zelle mit Multilevel-Betrieb

Anstelle von zwei generellen Niveaus (N und H) werden nun vier generelle Niveaus verwendet (N1, N2, H1, H2). Wie für zwei Bits pro Zelle kann für jedes dieser Niveaus, entweder die rechte Seite oder die linke Seite der Zelle, eine geringfügig höhere Einsatzspannung als die andere Seite besitzen. Vier generelle Niveaus bedeuten folglich, dass die Zelle nun acht mögliche voneinander verschiedene Zustände besitzt, sich also drei Bits speichern lassen.
Es ist naheliegend, dass sich eine NROM-Zelle auf die zuvor beschriebene Weise programmieren lässt. Es schließt sich allerdings die Frage an, ob die langzeitige Informationshaltung unter diesen Bedingungen noch gewährleistet werden kann. Aus diesem Grund wurden verwendete Zellen zuvor 10.000 Mal programmiert und gelöscht. Danach wurden vier Zellen mit dem Zustand je eines generellen Niveaus beschrieben. Der zweite Zustand eines Niveaus braucht aus Symmetriegründen nicht gesondert betrachtet werden. Die Kennlinien der programmierten Zellen sind in Abbildung 5.9 dargestellt.
Seite 1 wurde je zu einer etwas höheren Einsatzspannung programmiert. Entscheidend für die Informationshaltung ist nun, dass sowohl die relative Lage der beiden Seiten zueinander erhalten bleibt, als auch, dass es nicht zu einer Überschneidung der Zustände verschiedener genereller Niveaus kommt.
Da der Einsatzspannungsverlust mit der Programmierhöhe korreliert, wird der Abstand

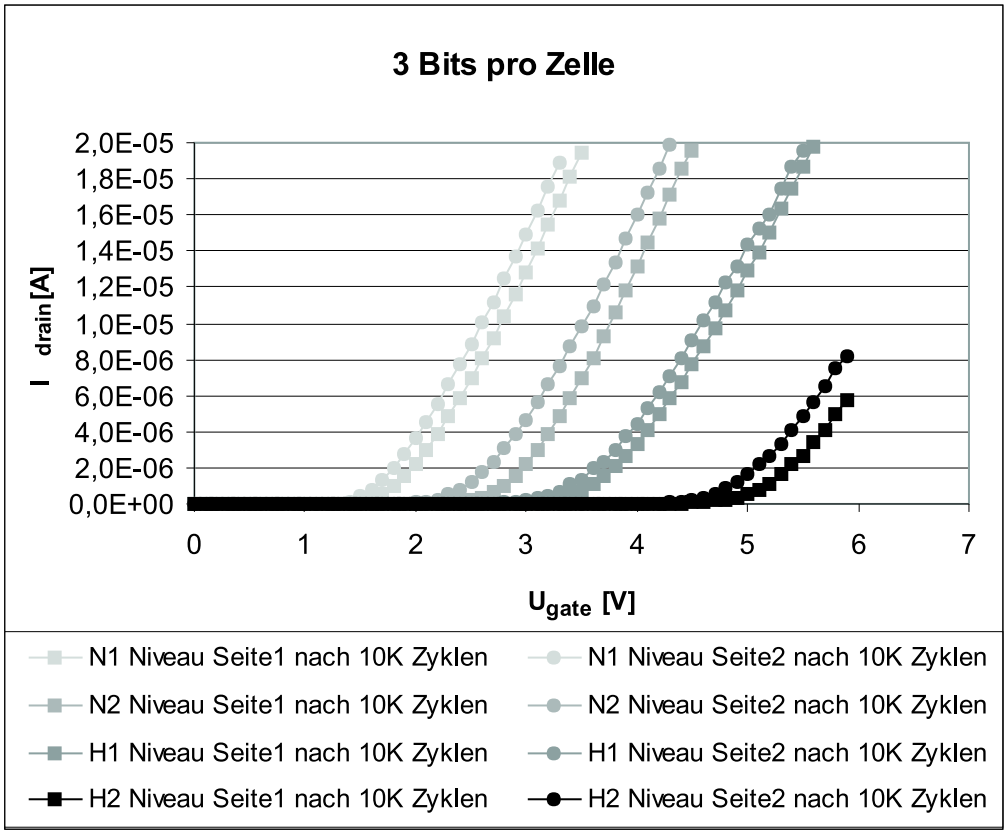


Abbildung 5.9: Transferkurven für 3 Bits pro Zelle

zwischen den beiden obersten Niveaus (H1 und H2) größer gewählt, als die Abstände zwischen den tiefer liegenden Niveaus.
Anschließend wurden die Zellen bei erhöhter Temperatur gelagert. Die Messkurven nach Lagerung sind, zusätzlich zu denen direkt nach Programmierung, in Abbildung 5.10 zu sehen.

Die leeren Symbole sind die Messkurven, die nach Lagerung aufgenommen wurden. Es ist zu erkennen, dass die Veränderung beim höchsten Niveau, wie erwartet, am größten ist. Die Einsatzspannungen der beiden unteren Niveaus ändern sich nur sehr geringfügig. Bei allen Niveaus bleibt die relative Lage der beiden Seiten erhalten. Zudem bleiben alle vier Niveaus klar getrennt.
Die Informationshaltung ist also auch bei drei Bits pro Zelle gewährleistet. Der Multilevel-

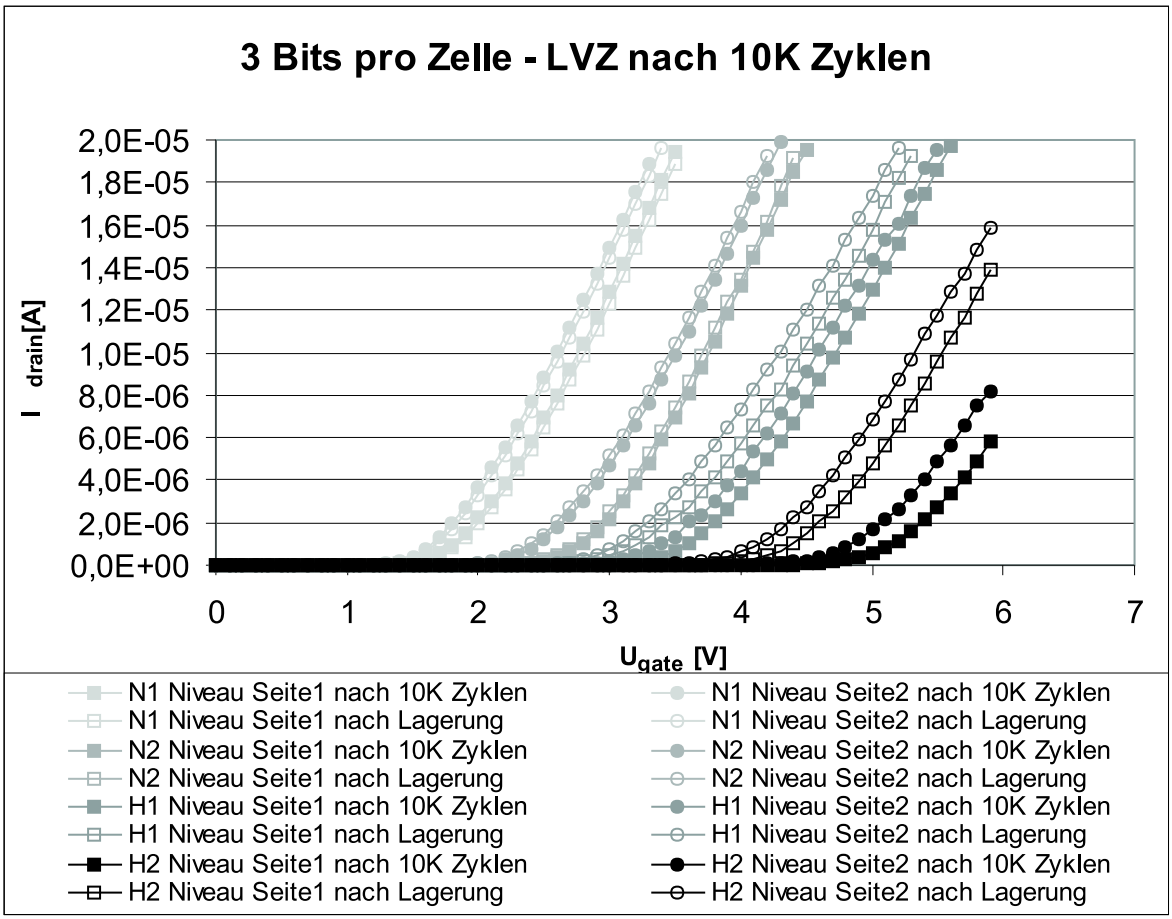


Abbildung 5.10: LVZ bei 3 Bits pro Zelle

Betrieb ist folglich geeignet, die Speicherdichte zu erhöhen.

Kapitel 6

Zusammenfassung und Ausblick

Die Arbeit ist in vier grosse inhaltliche Blöcke unterteilt. Zuerst werden die zum Verständnis notwendigen physikalischen Grundlagen in Kapitel 2 besprochen. Ausgehend von der MOS-Struktur über den MOSFET werden dann besonders sub-μm Effekte behandelt, die für die NROM-Zelle von grosser Bedeutung sind. Zudem wird ein erster Einblick in die prinzipielle Funktionsweise von NROM-Speicherzellen gegeben.
Im zweiten Block, Kapitel 3, wird detailliert auf zwei Zell-Konzepte für NROM eingegangen. Es wird ein bereits aus der Literatur bekanntes , konventionelles Modell dargestellt und bewertet. Desweiteren wird ein neuartiges Konzept eingeführt und besprochen. Es handelt sich um das erste Konzept mit shallow trench isolated NROM-Zellen. Die Änderungen und neuen Möglichkeiten, die sich durch dieses Konzept ergeben, werden beleuchtet. Es zeigt ein deutlich höheres Miniaturisierungspotential im Vergleich mit bisherigen Konzepten.
Neben der Behandlung dieser beiden Konzepte werden wesentliche Eigenschaften für die NROM-Zelle untersucht, Trap-Eigenschaften des Nitrids im ONO-Stapel und Modelle, die den Ladungsverlust erklären. Bei letzterem wird für die weitere Arbeit ein Modell gewählt, das auf der lateralen Bewegung von Ladungsträgern beruht, da dieses am besten mit den Messergebnissen in Einklang steht. Zudem wird gezeigt, dass sich eine programmierte NROM-Zelle durch ein Zwei-Transistor-Modell beschreiben lässt.
Den dritten Block bildet die experimentelle Evaluierung von STI-begrenzten NROM-Zellen

in Kapitel 4. Es wird eine Vielzahl von Messergebnissen vorgestellt. Dies geschieht, um die Bedeutung von Einflussparametern, wie z.B. Kanallänge und -weite, herauszuarbeiten. Desweiteren wird zum ersten Mal gezeigt, dass programmierte NROM-Zellen eine erhöhte Temperatursensitivität der Einsatzspannung gegenüber nicht programmierten Zellen aufweisen. Das Fazit aller Messungen ist, dass die STI-begrenzte NROM-Speicherzelle funktionstüchtig ist. Sie eröffnet somit eine neue Zukunftsperspektive für die NROM-Technologie. Im vierten Block wird ein neuer Multilevel-Betrieb für NROM vorgestellt. Durch das neuartige Betriebsschema kann die Einsatzspannungsdifferenz zwischen den beiden Seiten in einer Zelle gering gehalten werden. Das Nebensprechen in der Zelle wird zur Verbesserung der Informationshaltung ausgenutzt, es ist nicht länger ein unerwünschter Störmechanismus. Die Vorteile werden experimentell für zwei Bits pro Zelle nachgewiesen. Das Nebensprechen wird als Hindernis für die weitere Miniaturisierung von NROM überwunden. Darüber hinaus macht der neue Betrieb NROM multilevel tauglich. Die Funktionstauglichkeit wird am Beispiel von drei Bits pro Zelle nachgewiesen.

Literaturverzeichnis

[1] H. Aozasa, I. Fujiwara, A. Nakamura, and Y. Komatsu. Analysis of Carrier Traps in Si_3N_4 in Oxide/Nitride/Oxide for Metal/Oxide/Nitride/Oxide/Silicon Nonvolatile Memory. *Japan Journal of Applied Physics*, 38(3A):1441–1447, 1999.

[2] P.C. Arnett and Z.A. Weinberg. A Review of Recent Experiments Pertaining to Hole Transport in Si_3N_4. *IEEE Transactions on Electron Devices*, ED-25(8):1014–1018, 1978.

[3] J.J. Barnes, K. Shimohigashi, and R.W. Dutton. Short-channel MOSFETs in punch-through current mode. *IEEE Transactions on Electron Devices*, ED-26:446–453, 1979.

[4] S. Biesemans and K. de Meyer. Analytical calculation of subthreshold slope increase in short-channel MOSFETs by taking drift component into account. *Japan Journal of Applied Physics*, 34, 1995.

[5] Ilan Bloom, Paolo Pavan, and Boaz Eitan. $NROM^{TM}$ - a new technology for non-volatile memory products. *Solid-State Electronics*, (46):1757–1763, 2002.

[6] J.R. Brews, W. Fichtner, E.H. Nicolian, and S.M. Sze. Generalized guide for MOSFET miniaturization. *IEEE Electron Device Letters*, EDL-1:2–3, 1980.

[7] Y.-W. Chang, T.-C. Lu, S. Pan, and C.-Y. Lu. Modeling for the 2nd-Bit Effect of Nitride-Based Trapping Storage Flash EEPROM Cell Under Two-Bit Operation. *IEEE Electron Device Letters*, 25:95–97, 2004.

[8] K. Chen, C. Hu, P. Fang, M.R. Lin, and D.L. Wollesen. Predicting CMOS Speed with Gate Ocide and Voltage Scaling and Interconnect Loading Effects. *IEEE Transaction on Electron Devices*, 44(11):1951–1957, 1997.

[9] K. Chen, H.C. Wann, J. Duster, P.K. Ko, and C. Hu. MOSFET carrier mobility model based on gate oxide thickness, threshold and gate voltage. *J. Solid-State Electronics*, 39(10):1515–1518, 1996.

[10] W.-J. Cho, R. Kosugi, J. Senzaki, K. Fukuda, K. Arai, and S. Suzuki. Study on electron trapping and interface states of various gate dielectric materials in 4H-SiC metal-oxide-semiconductor capacitors. *Applied Physics Letters*, 77(13):2054–2056, 2000.

[11] D. Dürand and M. Kroker. Wirksamer Cocktail - Mit Biochips und superschnellen Prozessoren erschließt sich Halbleiterhersteller Infineon die Boommärkte Medizin und das vernetzte Auto. *WirtschaftsWoche*, (36):73–74, 2003.

[12] F. Driussi, R. Iob, D. Esseni, L. Selmi, R. van Schaijk, and F. Widdershoven. Spectroscopic analysis of trap assisted tunneling in thin oxides by means of substrate hot electron injection experiments. *IEDM Tech. Dig.*, 2002.

[13] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi. Can NROM, a 2 Bit, Trapping Storage NVM Cell, Give a Real Challenge to Floating Gate Cells? *International Conference on Solid State Devices and Materials*, pages 522–523, 1999.

[14] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi. NROM: A Novel Localized Trapping, 2-Bit Nonvolatile Memory Cell. *IEEE Electron Device Letters*, 21:543, 2000.

[15] C.P. Chang et al. A highly manufacturable corner rounding solution for $0.18\mu m$ shallow trench isolation. *in IEDM Tech. Dig.*, page 380, 1997.

[16] D. Frohman-Bentchkowsky and A.S. Grove. *IEEE Transactions on Electron Devices*, ED-16:108, 1969.

[17] S.D. Ganichev, E. Ziemann, W. Prettl, I.N. Yassievich, A.A. Istratov, and E.R. Weber. Destinction between the Poole-Frenkel and tunneling models of electric-field-stimulated carrier emission from deep levels in semiconductors. *Physical Review B*, 61(15):10361–10365, 2000.

[18] D. Gitlin, J. Karp, and B. Moyzhes. Dangling bonds with 'negative Hubbard U': Physical model for degradation of SiO_2 gate dielectric under voltage stress. *Journal of Applied Physics*, 92(12):7257–7260, 2002.

[19] A. Godoy, J.A. López-Villanueva, J.A. Jiménez-Tejada, A. Palma, and F. Gámiz. A simple subthreshold swing model for short channel MOSFETs. *J. Solid-State Electronics*, 45:391–397, 2001.

[20] N. Goldman and J. Frey. Electron energy distribution for calculation of gate leakage current in MOSFETs. *Solid-State Electronics*, 31:1089–1092, 1988.

[21] V.A. Gritsenko, H. Wong, J.B. Xu, R.M. Kwok, I.P. Petrenko, B.A. Zaitsev, Y.N. Morokov, and Y.N. Novikov. Excess silicon at the silicon nitride/thermal oxide interface in oxide-nitride-oxide structures. *Journal of Applied Physics*, 86(6):3234–3240, 1999.

[22] Patrick Haibach, Rainer Hagenbeck, and Frank Lau. unveröffentlicht, CL PTD SIM in München (Infineon), 2003.

[23] James A. Hayes. Insulated gate field-effect transistor read-only memory cell. U.S. Patent 4,173,766, 1979.

[24] Kurt Hoffmann. *Systemintegration: vom Transistor zur großintegrierten Schaltung*. Oldenbourg Wissenschaftsverlag, München; Wien, 2003.

[25] F.S. Hsu, R.S. Muller, C. Hu, and P-K. Ko. A simple punchthrough model for short-channel MOSFETs. *IEEE Transactions on Electron Devices*, ED-30:1354–1359, 1983.

[26] C. Hu. Hot-carrier effects. In *Advanced MOS Device Physics*, volume 18, pages 119–160, VLSI Electronics, New York, 1989. Academic Press, N. G. Einspruch and G. Gildenblat.

[27] Gianluca Ingrosso, Luca Selmi, and Enrico Sangiorgi. Monte Carlo Simulation of Program and Erase Charge Distributions in $NROM^{TM}$ Devices. *ESSDERC*, 2002.

[28] J.-W. Jung, J.-M. Kim, J.-H. Son, and Y. Lee. Dependence of Subthreshold Hump and Reverse Narrow Channel Effect on the Gate Length by Suppression of Transient Enhanced Diffusion at Trench Isolation Edge. *Japan Journal of Applied Physics*, 39(4B):2136–2140, 2000.

[29] Wilhelm Jutzi. *Digitalschaltungen - Eine Einführung*. Springer-Verlag, Berlin Heidelberg, 1995.

[30] B. Kaczer, F. Crupi, R. Degraeve, Ph. Roussel, C. Ciofi, and G. Groesenecken. Observation of hot-carrier-induced nFET gate-oxide breakdown in dynamically stressed CMOS circuits. *IEDM Tech. Digest*, 2002.

[31] Y. Kamigaki, S. Minami, and H. Kato. A new portrayal of electron and hole traps in amorphous silicon nitride. *Journal of Applied Physics*, 68(5):2211–2215, 1990.

[32] V.J. Kapoor and S.N.B. Bibyk. *The Physics of MOS Insulators*. Pergamon, Oxford, U.K., 1980.

[33] F. X. Kärtner. *Halbleiterbauelemente*. Lecturenotes of the Universität Karlsruhe (TH), 2000.

[34] F.M. Klaasen. Review of physical models for MOS transistors. *Process and Device Modelling for Integrated Circuit Design*, 1977.

[35] D.T. Krick and P.M. Lenahan. Nature of the dominant deep trap in amorphous silicon nitride. *Physical Review B*, 38(12):8226–8229, 1988.

[36] K.J. Kuhn, D. Mei, I. Post, and J. Neirynck. Scaling challenges for 0.13μm generation shallow trench isolation. In *IEEE International Symposium on Semiconductor Manufacturing*, San Jose, California, 2001.

[37] K. Kurosawa, T. Shibata, and H. Iizuka. A new bird's beak free field isolation technology for VLSI devices. *in IEDM Tech. Dig.*, page 384, 1981.

[38] M.S. Liang, J.Y. Choi, P.K. Ko, and C. Hu. Inversion-layer capacitance and mobility of very thin gate-oxide MOSFET's. *IEEE Transactions on Electron Devices*, ED-33:409, 1986.

[39] F.-T. Liou and S.-O. Chen. Evidence of Hole Flow in Silicon Nitride for Positive Gate Voltages. *IEEE Transactions on Electron Devices*, ED-31(12):1736–1741, 1984.

[40] C.W. Liu and T.X. Hsieh. Analytic modeling of the subthreshold behavior in MOSFET. *J. Solid-State Electronics*, 44:1707–1710, 2000.

[41] E. Lusky, I. Bloom, G. Cohen, B. Eitan, Y. Shacham-Diamand, and A. Shappir. Retention Loss Characteristics of Localized Charge-Trapping Devices. In *IEEE International Electron Device Meeting*, 2003.

[42] E. Lusky, Y. Shacham-Diamand, I. Bloom, and B. Eitan. Characterization of Channel Hot Electron Injection by th Subthreshold Slope of $NROM^{TM}$ Devices. *IEEE Electron Device Letters*, 22:556, 2001.

[43] E. Lusky, Y. Shacham-Diamand, I. Bloom, and B. Eitan. Electron Discharge Model of Locally-Trapped Charge in Oxide-Nitride-Oxide (ONO) Gate for $NROM^{TM}$ Nonvolatile Semiconductor Memory Devices. *Ext. Abst. 2001 Conf. Solid State Devices and Materials*, page 534, 2001.

[44] E. Lusky, Y. Shacham-Diamand, I. Bloom, and B. Eitan. Electrons Retention Model for Localized Charge in Oxide-Nitride-Oxide (ONO) Dielectric. *IEEE Electron Device Letters*, 23(9):556–558, 2002.

[45] E. Maayan, R. Dvir, J. Shor, Y. Sofer, I. Bloom, D. Avni, B. Eitan, Z. Cohen, M. Meyassed, Y. Alpern, H. Plam, E. Stein v. Kamienski, P. Haibach, D. Caspary, S. Riedel, and R. Knofler. A 512Mb NROM Data Storage Memory with 8MB/s data rate. In *International Solid-State Circuits Conference*, San Francisco, 2002.

[46] S. Manzi. Electronic processes in silicon nitride. *Journal of Applied Physics*, 62(8):3278–3284, 1987.

[47] T. Maruyama and R. Shirota. The low electric field conduction mechanism of silicon oxide-silicon nitride-silicon oxide interpoly-Si dielectrics. *Journal of Applied Physics*, 78(6):3912–3914, 1995.

[48] G. Merckel. CAD models for MOSFETS. *Process and Device Modelling for Integrated Circuit Design*, 1977.

[49] G. Merckel. Short channels - scaled down MOSFETs. *Process and Device Modelling for Integrated Circuit Design*, 1977.

[50] G. Merckel, J. Borel, and N.Z. Cupcea. An accurate large-signal MOS tansistor model for use in computer-aided design. *IEEE Transactions on Electron Devices*, ED-19:681–690, 1972.

[51] MIT Artificial Intelligence Laboratory, http://www.ai.mit.edu/people/tk/tks/tcon.html. *Thermal Properties of Materials*, 2004.

[52] Dieter A. Mlynski. *Elektrodynamik*. Manuskript an der Universität Karlsruhe, 7. edition, 1999.

[53] K.A. Nasyrov, V.A. Gritsenko, M.K. Kim, H.S. Chae, S.D. Chae, W.I. Ryu, J.H. Sok, J.-W. Lee, and B.M. Kim. Charge Transport Mechanism in Metal-Nitride-Oxide-Silicon Structures. *IEEE Electron Device Letters*, 23:336–338, 2002.

[54] S. Ogura, P.J. Tsang, W.W. Walker, D.L. Critchlow, and J.F. Shepard. Design and characteristics of the lightly doped drain-source (LDD) insulated-gate field-effect transistor. *IEEE Transactions on Electron Devices*, ED-27:1359–1366, 1980.

[55] D. Qian and D.J. Dumin. A comprehensive physical model of oxide wearout and breakdown explaining the fluence, time, field, and thickness dependence of trap generation. *Electrochemical Society Proceedings*, 6:11–25, 1999.

[56] V.G. Reddi and C.T. Sah. Source to Drain Resistance Beyond Pinchoff in Metal-Oxide-Semiconductor Transistors (MOST). *IEEE Transactions on Electron Devices*, ED-12:108, 1969.

[57] Y. Roizin, M. Gutman, E. Aloni, V. Kairys, and P. Zisman. Retention Characteristics of $microFLASH^{TM}$ Memory (Activation Energy of Traps in the ONO Stack). *Non-Volatile Sem. Memory Workshop*, page 128, 2001.

[58] A.G. Sabnis and J.T. Clemens. Characterization of electron velocity in the inverted $\langle 100 \rangle$ Si surface. *in IEDM Tech. Dig.*, pages 18–21, 1979.

[59] A. Shappir, Y. Shacham-Diamand, E. Lusky, I. Bloom, and B. Eitan. Subthreshold slope degradation model for localized-charge-trapping based non-volatile memory devices. *Solid-State Electronics*, 47:937–941, 2003.

[60] N. Shigyo and R. Dang. Analysis of anomalous subthreshold current in a fully recessed oxide MOSFET using a three-dimensional device simulator. *IEEE Transactions on Electron Devices*, ED-32:441, 1985.

[61] N. Shigyo and T. Hiraoka. A review of narrow-channel effect for STI MOSFET's: A difference between surface- and buried-channel cases. *J. Solid-State Electronics*, 43:2061–2066, 1999.

[62] R.A. Stuart and W. Eccleston. Punchthrough currents in chort-channel M.O.S.T. devices. *Electronics Letters*, 9:586–588, 1973.

[63] E. Suzuki and Y. Hayashi. Carrier conduction and trapping in metal-nitride-oxide-semiconductor structures. *Journal of Applied Physics*, 53(10):8880, 1982.

[64] E. Suzuki, K. Miura, Y. Hayasi, R.-P. Tsay, and D.K. Schroder. Hole and Electron Current Transport in Metal-Oxide-Nitride-Oxide-Silicon Memory Structures. *IEEE Transactions on Electron Devices*, 36(6):1145–1149, 1989.

[65] M. Tao, D. Park, S.N. Mohammad, D. Li, A.E. Botchkerav, and H. Morkoç. Electrical conduction in silicon nitrides deposited by plasma enhanced chemical vapour deposition. *Philosophical Magazine B*, 73(4):723–736, 1996.

[66] R.R. Troutman. VLSI limitations from drain-induced barrier lowering. *IEEE Journal of Solid-State Circuits*, EC-14:383–391, 1979.

[67] W.J. Tsai, S.H. Gu, N.K. Zous, C.J. Liu, C.C. Liu, C.H. Chen, T. Wang, S. Pan, and C.Y. Lu. Data Retention Behavior of a SONOS Type Two-Bit Storage Flash Memory Cell. *IEDM Tech. Digest*, page 719, 2001.

[68] W.J. Tsai, S.G. Hu, N.K. Zous, C.C. Yeh, C.C. Liu, C.H. Chen, T. Wang, S. Pan, and C.Y. Lu. Cause of Data Retention Loss in a Nitride-Based Localized Trapping Storage Flash Memory Cell. *IEEE 40th Annual International Reliability Physics Symposium*, pages 34–38, 2002.

[69] Yannis Tsividis. *Operation and Modeling of the MOS Transistor*. McGraw-Hill, Singapore, second edition, 1999.

[70] Allan T. Mitchell und Bert R. Riemenschneider. non-volatile semiconductor memory. U.S. Patent 5,168,334, 1992.

[71] R.C. Varschney. Simple theory for threshold voltage modulation in short-channel MOS transistor. *Electronic Letters*, 9:600–602, 1973.

[72] E.M. Vogel, D.-W. Heh, J.B. Bernstein, and J.S. Suehle. Impact of the Trapping of Anode Hot Holes on Silicon Dioxide Breakdown. *IEEE Electron Device Letters*, 23(11):667–669, 2002.

[73] C.T. Wang. An improved hot-electron-emission model for simulation the gate-current characteristic of MOSFETs. *Solid-State Electronics*, 31:229–231, 1988.

[74] J. Willer, C. Ludwig, J. Deppe, C. Kleint, S. Riedel, J.-U. Sachse, M. Krause, R. Mikalo, E. Stein v. Kaminski, S. Parascandola, T. Mikolajick, J.-M. Fischer, M. Isler, K.-H. Küsters, I. Bloom, A. Shapir, E. Lusky, and B. Eitan. 110nm nrom technology for code and data flash products. *IEEE Symposium on VLSI Technology*, pages 76–77, 2004.

[75] Stanley Wolf. *Silicon Processing for VLSI Era; Volume 3 - The Submicron MOSFET*. Lattice Press, Sunset Beach, California, 1995.

[76] L. D. Yau. A simple theory to predict the threshold voltage of chort-channel IGFETs. *Solid-State Electronics*, 17:1059–1063, 1974.

[77] L.D. Yau. Arguments For Electron Conduction in Silicon Nitride. *IEEE Electron Device Letters*, EDL-5(8):318–321, 1984.