Andreas Weber

Charakterisierung von Leckstrompfaden in DRAM Speicherzellen und deren Reduktion

Cuvillier Verlag Göttingen

Charakterisierung von Leckstrompfaden in DRAM Speicherzellen und deren Reduktion

Vom Promotionsausschuss der
Technischen Universität Hamburg-Harburg
zur Erlangung des akademischen Grades
Doktor-Ingenieur
genehmigte Dissertation

von

Andreas Weber
aus Heidenheim a.d. Brenz

2007

Qimonda Dresden GmbH & Co. OHG
Technische Universität Hamburg-Harburg, Institut für Nanoelektronik

Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.
1. Aufl. - Göttingen : Cuvillier, 2007
Zugl.: (TU) Hamburg-Harburg), Univ., Diss., 2007

978-3-86727-199-8

Erster Gutachter:	Prof. Dr. Wolfgang Krautschneider
Zweiter Gutachter:	Prof. Dr. Wolfgang Albrecht
Tag der mündlichen Prüfung:	12.01.2007

© CUVILLIER VERLAG, Göttingen 2007
Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

Alle Rechte vorbehalten. Ohne ausdrückliche Genehmigung des Verlages ist es nicht gestattet, das Buch oder Teile daraus auf fotomechanischem Weg (Fotokopie, Mikrokopie) zu vervielfältigen.
1. Auflage, 2007
Gedruckt auf säurefreiem Papier

978-3-86727-199-8

Inhaltsverzeichnis

Liste der Symbole und Abkürzungen **vii**

1 Einleitung **1**

2 DRAM Grundlagen **7**

2.1 Zellaufbau 7

2.2 Speichermatrix 10

2.3 Elektrisches Funktionsprinzip 16

2.3.1 Blockschaltbild 17

2.3.2 Befehlssatz 17

2.3.3 Leseoperation 19

2.3.4 Schreiboperation 20

2.3.5 Verstärkungsvorgang 21

2.3.6 Refresh 23

2.4 Mode-Register und Testmodes 24

2.5 Redundanz 24

3 DRAM Data Retention **25**

3.1 Definition und typische Verteilung der Retentionzeiten 25

3.2 Retention-Formel 27

3.3 Ursachen für kurze Haltezeiten 31

4 Monte Carlo Analyse der Retentionverteilung **33**

4.1 Das Simulationsmodell 34

4.2 Parameterverteilungen 35

4.3 Simulation der Leckstromverteilung 37
4.4 Rücksimulation der Retentionverteilung 38
4.4.1 Einfluss der Kondensatorkapazität 39
4.4.2 Einfluss der Bitleitungslänge 39
4.4.3 Einfluss des Differenzverstärker-Offsets 42

5 Leckstrompfade im DRAM **47**
5.1 Überblick 47
5.2 pn-Leckströme 48
5.2.1 GIDL 49
5.2.2 Junction-Leakage 50
5.3 Transistor-Unterschwellenleckströme 51
5.3.1 SubVt-Leckstrom 51
5.3.2 Deep-SubVt Leckstrom 54
5.3.3 Vertikaler Parasitärer Leckstrom 54
5.3.4 SubSTI Leckstrom 54
5.4 Leckströme durch Dielektrika 55
5.4.1 Node-Leckstrom 58
5.4.2 Gateoxid Leckstrom 59
5.4.3 Passing-WL Leckstrom 60
5.5 Übersicht der Spannungsabhängigkeiten 60

6 Charakterisierungsmethoden **63**
6.1 Messungen im Wafer-Kerf 63
6.1.1 Einzel-Strukturen 65
6.1.2 Parallel-Strukturen 66
6.2 Wafer-Tests 68
6.2.1 Prefuse-Test 68
6.2.2 Shmoo-Tests 68
6.3 Baustein-Tests 69
6.4 Einzelzell-Analysen 70

6.4.1 Charakterisierungsmethode 71
6.4.2 Messapparatur 76
6.5 Zusammenfassung 80

7 Ergebnisse zur elektrischen Charakterisierung 81
7.1 Retentionkurven 81
7.1.1 „0"- und „1"-Retention 82
7.1.2 Temperatur-Abhängigkeit 85
7.1.3 V_{BB}-Abhängigkeit 88
7.1.4 V_{NWLL}-Abhängigkeit 88
7.1.5 V_{BL}-Abhängigkeit 90
7.2 Einzelzell-Analysen 93
7.2.1 Temperaturabhängigkeit 93
7.2.2 Spannungsabhängigkeit der Retentionzeiten 94
7.2.3 Spannungsabhängigkeit der Aktivierungsenergien 97
7.2.4 Aktivierungsenergien entlang der Retentionkurve 99

8 Theoretische Betrachtung & Abschätzung 103
8.1 Einfaches Modell zur Abschätzung 103
8.2 Leckstrom des idealen pn-Übergangs 104
8.3 Thermische Generation (SRH) 105
8.3.1 Rekombinationsgleichung 105
8.3.2 Generation im pn-Übergang 110
8.4 Tunnelunterstützte Generation (TFE) 113
8.5 Trapunterstütztes Tunneln (TAT) und Band-zu-Band-Tunneln (BTB) 117
8.6 Zusammenfassung 119

9 Experimentelle Verifikation 121
9.1 Mögliche Maßnahmen zur Tailverbesserung 121
9.1.1 E-Feldreduktion im kondensatorseitigen pn-Übergang 122
9.1.2 Passivierung von Zuständen in der Bandlücke 124
9.2 Experimentelle Verifikation durch Passivierung mit Fluor 125

9.2.1 Experimente 125
9.2.2 Ergebnisse der Implantationsexperimente 126
9.2.3 Diskussion der experimentellen Ergebnisse 127
9.2.4 Aktivierungsenergie-Analyse 129

10 Zusammenfassung **133**

Literaturverzeichnis **137**

Liste der Symbole und Abkürzungen

Symbol	Beschreibung	Einheit
A	Fläche eines pn-Übergangs	cm^2
A_S	Grenzfläche eines pn-Übergangs zum Oxid	cm^2
BP	Buried Plate	
BS	Buried Strap	
BTB	Band-To-Band Tunneling	
DT	Deep Trench	
$c_{n,p}$	Elektron- und Locheinfangskoeffizient	$cm^3 s^{-1}$
CAS	Column Address Strobe	
C_{BL}	Bitleitungskapazität	F
C_{BLBL}	Bitleitungskopplungskapazität	F
C_S	Speicherkapazität	F
C_{SA}	Kapazität des Leseverstärkers	F
CL	CAS latency = Zeit zwischen Spaltenadressierung und Verfügbarkeit der Daten an den externen Kontakten	Taktzyklen
DRAM	Dynamic Random Access Memory	
DDR	Double Data Rate	
DIBL	Drain Induced Barrier Lowering	
DQ	Data Query = Datenpin	
$D_{N,P}$	Diffusionskoeffizienten von Elektronen und Löchern	$m^2 s^{-1}$
D_{it}	Interfacetrapdichte	$cm^{-2} eV^{-1}$
$e_{n,p}$	Elektron- und Lochemissionskoeffizient	s^{-1}
E_a	Aktivierungsenergie	eV
E_c	Leitungsbandkante	eV
E_F	Fermi-Energie	eV
E_G	Silizium-Bandlücke	eV
E_i	Intrinsische Energie	eV
E_v	Valenzbandkante	eV

F	DRAM 1/2 pitch = kleinste lithographische Strukturbreite einer DRAM Technologie (z.B. $F = 90\,nm$ -> Wortleitungsbreite und Wortleitungsabstand $\sim 90\,nm$)	nm
F	elektrische Feldstärke	V/cm
GC	Gate Conductor (Wortleitungsebene)	
GIDL	Gate Induced Drain Leakage	
I_{DS}	Drain-Source-Strom eines MOSFET	A
I_G	Gate-Strom eines MOSFET	A
I_{ideal}	idealer Leckstrom eines pn-Übergangs	A
I_{SRH}	Shockley Read Hall Generationsstrom	A
I_{TFE}	Generationsstrom durch Thermionic Field Emission	A
I_{BTB}	Band-zu-Band Tunnelstrom	A
ITRS	International Technology Roadmap for Semiconductors	
JEDEC	Joint Electron Device Engineering Council	
L	Kanallänge eines MOSFET	cm
$L_{N,P}$	Diffusionslänge von Elektronen und Löchern	cm
LOCOS	Local Oxidation Of Silicon	
m_{Si}	Elektronenmasse im Silizium	
MINT	Merged Isolation and Node Trench	
MOSFET	Metal Oxide Semiconductor Field Effect Transistor	
n	Elektronendichte	cm^{-3}
n_i	intrinsische Ladungsträgerdichte	cm^{-3}
n_t	Konzentration mit Elektronen besetzter Trapzentren	cm^{-3}
N_A	Akzeptorendichte	cm^{-3}
N_c	Zustandsdichte im Leitungsband	cm^{-3}
N_D	Donatorendichte	cm^{-3}
N_v	Zustandsdichte im Valenzband	cm^{-3}
p	Löcherdichte	cm^{-3}
p_t	Konzentration nicht besetzter Trapzentren	cm^{-3}
PWL	Passing Wordline	
R	Rekombinationsrate von Löchern und Elektronen	s^{-1}
RAS	Row Address Strobe	
ROR	Row Address Only Refresh	
R_s	Serienwiderstand	Ω
R_{SD}	SD-Widerstand eines MOSFET	Ω
S	Unterschwellsteigung eines MOSFET	V/dec
SIA	Semiconductor Industry Association	
SRAM	Static Random Access Memory	
STI	Shallow Trench Isolation	

t_{ox}	Oxiddicke	cm
t_{RCD}	RAS-to-CAS Delay = Zeit zwischen Zeilen-und Spaltenadressierung	ns
t_{RP}	Row Precharge Time = Vorladezeit der Bitleitungen	ns
t_{RAS}	Activate-to-Precharge Time = Zeit zwischen der Aktivierung einer WL und dem erneuten Vorladen der Bitleitungen der selben Bank	ns
t_{RC}	RAS Cycle Time = Zeit zwischen dem Anlegen von zwei Zeilenadressen in derselben Bank (Summe aus t_{RP} und t_{RAS})	ns
t_{CAS}	Column Address Strobe Time	ns
t_{Ret}	Retentionzeit	ms
T	Temperatur	K
TAT	Trap Assisted Tunneling	
TEAS	TEOS mit Arsen-Zusatz	
TEOS	auf Tetra-Ethyl-Ortho-Silicate (TEOS) basiertes SiO_2-Abscheidungsverfahren	
TFE	Thermionic Field Emission	
V	Sperrspannung am pn-Übergang	V
V_{BB}	Potenzial der p-Wanne (back bias)	V
V_{bi}	eingebaute Spannung eines pn-Übergangs	V
V_{BL}	Bitleitungspotenzial	V
V_{BLL}	Bitleitungspotenzial im *low*-Zustand	V
V_{BLH}	Bitleitungspotenzial im *high*-Zustand	V
V_{BLEQ}	Vorladepotenzial der Bitleitungen (bitline equalize)	V
V_{DS}	DS-Spannung eines MOSFET	V
V_{FB}	Flachbandspannung	V
V_G	Gate-Spannung eines MOSFET	V
V_{GS}	GS-Spannung eines MOSFET	V
V_{NWLL}	Wortleitungsspannung bei ausgeschaltetem Transistor (negative wordline low)	V
V_{PL}	Potenzial der äußeren Kondensatorplatte (plate)	V
V_{PP}	Wortleitungsspannung bei eingeschaltetem Transistor	V
V_S	Potenzial des Speicherkondensators	V
V_{SB}	SB-Spannung eines MOSFET	V
V_t	Einsatzspannung eines MOSFET	V
v_{th}	thermische Geschwindigkeit	cm/s
V_j	Volumen der Verarmungszone eines pn-Übergangs	cm^3
W	Breite der Verarmungszone eines pn-Übergangs	cm

WL	Write Latency = Zeit zwischen der Spaltenadressierung und dem Anlegen von Daten zum Schreiben	Taktzyklen
μ_p	Löcher-Mobilität	cm^2/Vs
μ_n	Elektronen-Mobilität	cm^2/Vs
σ_n	Elektronen-Einfangsquerschnitt	cm^2
σ_p	Löcher-Einfangsquerschnitt	cm^2
$\tau_{n,p}$	Minoritätslebensdauer von Elektronen bzw. Löchern	s
ϕ_F	Fermipotenzial	V
ϕ_M	Metall-Austrittsarbeit	V
ϕ_S	Halbleiter-Austrittsarbeit	V
ϕ_s	Oberflächen-Potenzial	V

Kapitel 1

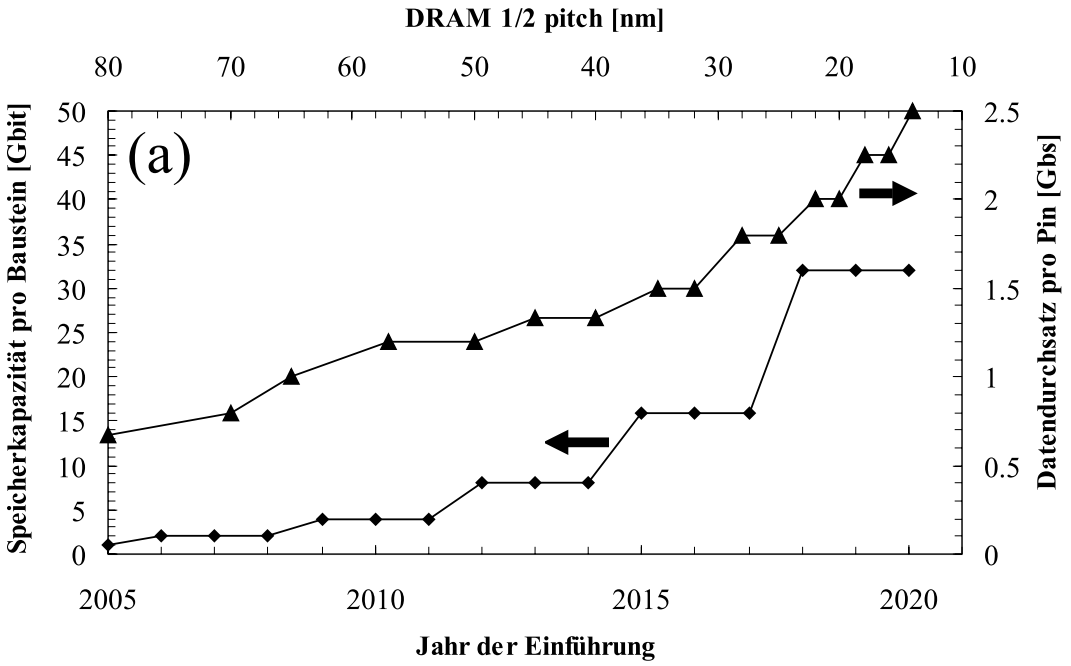
Einleitung

Nach der Erfindung des Transistors durch B. Shockley, J. Bardeen und W. Brattain im Jahre 1947 legten D. Kahng und M.M. Atalla mit dem ersten industriell hergestellten MOSFET im Jahr 1960 den Grundstein für einen Industriezweig, der sich in den letzten Jahrzehnten rasant wie kein anderer entwickelte. Schon 1959 reichte J. Kilby die Anmeldung für sein Patent für integrierte Schaltkreise („Solid Circuit made of Germanium") ein. R. Dennard erfand 1966 den *Dynamic Random Access Memory* (DRAM) und erhielt 1968 als Forscher am IBM Watson Research Lab das Patent dafür [Den68]. Bereits im Jahre 1971 führte Intel den ersten 1-Transistor DRAM mit einer Kapazität von $2\,kBit$ ein [Den84]. IBM stellte den ersten im heutigen Sinne als Personal Computer zu bezeichnenden Rechner (Modell 5150) am 12. August 1981 der Öffentlichkeit vor. Dieser enthielt bereits einen $16\,kByte$ großen Arbeitsspeicher der auf 8 einzelne ICs mit je $16\,kBit$ Kapazität aufgeteilt war. Seit dieser Zeit folgt die Mikroelektronik *Moore's law* [Moo65], nachdem sich die Anzahl der auf einem IC integrierten Bauelemente alle zwei Jahre verdoppelt. Dies hat dazu geführt, dass bereits heute Speicher-ICs mit $1\,GBit$ auf dem Markt sind. Technologisch wird dieses rasante Wachstum der Bauelementeanzahl pro IC dabei hauptsächlich durch Strukturverkleinerung (*shrinking*) getragen. Da die Produktionskosten pro Fläche in erster Näherung konstant blieben, konnte der Bitpreis somit exponentiell fallen und die Mikroelektronik hielt Einzug in unser tägliches Leben. Von dort ist sie heute nicht mehr wegzudenken (z.B. PC, PDA, MP3-Player, Digitalkamera, Handy ....). Der DRAM konnte sich aufgrund der sehr einfachen Speicherzelle, die hohe Integrations- und Speicherdichten und damit eine kostengünstige Produktion erlaubt, bis heute in vielen Bereichen gegenüber anderen Speichertechnologien (z.B. SRAM) behaupten. Mit dem heutigen Stand ist jedoch noch kein Ende erreicht und die Entwicklung geht gemäß *Moore's Law* weiter. Seit 1998 schreibt ein aus der *Semiconductor Industry Association (SIA)* hervorgegangenes Konsortium bestehend aus Experten aus Industrie und Forschung die Anforderungen und Erwartungen an die Mikroelektronik für die nächsten 15 Jahre in der *International Technology Roadmap for Semiconductors* [ITR05] nieder. Ziel der Roadmap ist es, die zur Si-

cherung des zukünftigen Wachstums der Mikroelektronik notwendigen Anregungen zur Innovation und Investition zu geben. Abbildung 1.1a zeigt die in der aktuellen Version vorhergesagte Entwicklung für den *Commodity*-DRAM, d.h. dem in Massenproduktion befindlichen Standardspeicher, bis zum Jahr 2020. Demzufolge wird die Speicherkapazität von Standardspeicherchips bis zum Jahr 2010 bereits auf $4\,GBit$ anwachsen und mit einem *DRAM 1/2 pitch (F)* von $60\,nm$ gefertigt werden. Der Wert F misst im DRAM die halbe Wiederholungslänge der kleinsten Strukturen, also die Hälfte von Wortleitungsbreite plus Wortleitungsabstand. Langfristig sehen die Experten ein Anwachsen der Kapazität bis auf $32\,GBit$ im Jahr 2020, welche in $14\,nm$-Technologie gefertigt werden soll. Ebenso wie die Kapazität soll auch der Datendurchsatz stark anwachsen, um die immer schnelleren Prozessoren ausreichend mit Daten versorgen zu können. Dabei ist vor allem ein höherer Parallelisierungsgrad im DRAM-Design gefordert, da die Lese- und Schreibzeiten von Generation zu Generation nicht in dem dazu notwendigen Maße reduziert werden können. Bereits in der Vergangenheit hat dies immer neue Standards erfordert und wir stehen heute kurz vor der Einführung des DDR3-Standards in die Massenproduktion. Abbildung 1.1b zeigt die Design-Entwicklung der letzten Jahre. Neben dem *Commodity*-Bereich gibt es noch weitere Spezialprodukte, wie z.B. Grafikspeicher oder Speicher für mobile Anwendungen, deren Anforderungen noch höher liegen (siehe [ITR05] für Details).

Der Nachteil des DRAMs lässt sich aus dem Namen ableiten: die in den Zellen gespeicherte Information bleibt nur für kurze Zeit erhalten und muss durch aufwendige Mechanismen ständig aufgefrischt werden (*Refresh*). Die maximale Zeit zwischen zwei *Refreshes* ist laut ITRS bis ins Jahr 2020 mit $64\,ms$ spezifiziert. In dieser Zeit darf keine einzige der vielen Speicherzellen auf einem Chip ihre gespeicherte Information verlieren. Bei den heute verfügbaren Speicherchips mit 10^9 Speicherzellen dürfen deshalb selbst $6\,\sigma$-Streuungen der Haltezeiten diesen Wert nicht unterschreiten und im Jahr 2020 müssen dann sogar $6.5\,\sigma$-Werte berücksichtigt werden. In der Praxis liegen die durchschnittlichen Haltezeiten selbst bei Temperaturen über $85\,°C$ noch Größenordnungen über den $64\,ms$ der Spezifikation. Das Problem dabei ist, dass selbige Haltezeiten einer sehr breiten und bisher nicht vollständig verstandenen Verteilung unterliegen, deren äußere Enden ($6\,\sigma$-Werte) die Spezifikation unterlaufen. Deshalb enthalten heutige Chips bereits Reparaturmöglichkeiten, durch welche einige schlechte Zellen durch Redundanz ersetzt werden können. Dies ist wirtschaftlich natürlich nur in begrenztem Umfang möglich und die Anzahl der durch Redundanz ersetzten Zellen muss klein gehalten werden. Um die laut ITRS weiter anwachsenden Bitzahlen realisieren zu können, ist es deshalb unumgänglich die Ursache für die breite Verteilung zu verstehen. Ohne Verständnis und Verbesserung der Retentionverteilung wird die Speicherindustrie die vorhergesagte ITRS-Roadmap nicht einhalten können.

Daraus folgt direkt die Motivation dieser Arbeit. Ziel ist es, die Ursache für die breite Verteilung der Haltezeiten eines Chips zu verstehen und daraus mögliche Verbesserungen abzuleiten, zu verifizieren und zu implementieren.



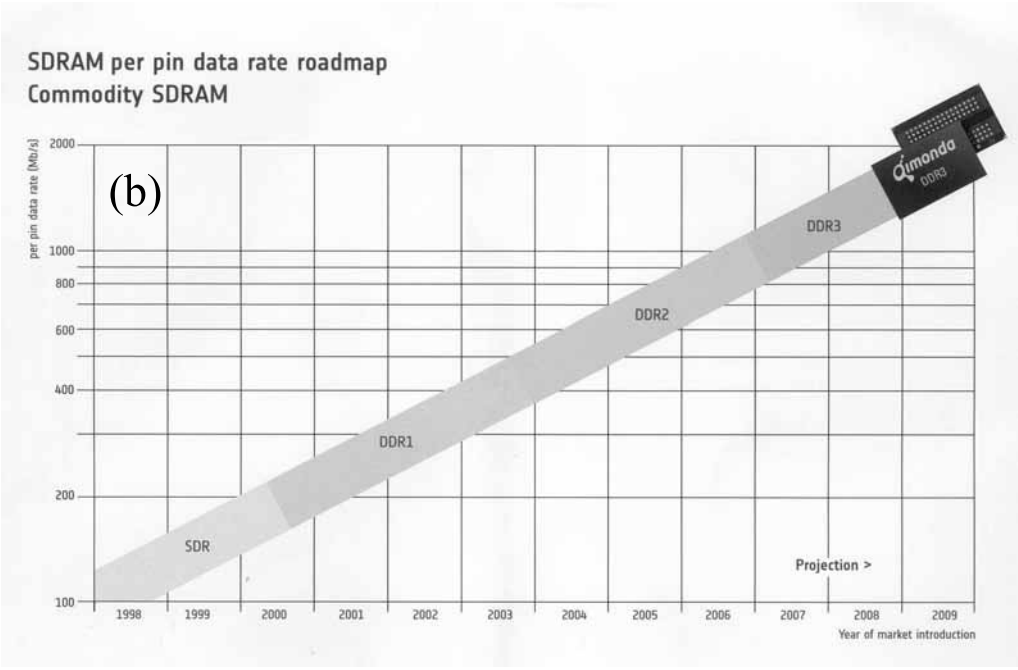


Abbildung 1.1: (a)Vorhergesagte Entwicklung der DRAM-Kapazität pro Chip und des Datendurchsatzes [ITR05]. (b) DRAM-Standards der letzten Jahre.

Gliederung der Arbeit

Die vorliegende Arbeit ist so aufgebaut, dass der Leser von den Grundlagen ausgehend hin zu immer detaillierteren Teilaspekten geführt wird. Dabei setzen die einzelnen Kapitel aufeinander auf. Am Ende hat der Leser alle notwendigen Kenntnisse, um die Ursache für die breite Retentionverteilung zu verstehen und die vorgeschlagenen Verbesserungen nachvollziehen zu können. Aufgrund der großen Breite des Themas werden nicht alle Details angesprochen, um den Blick auf das Wesentliche zu lenken und den Überblick nicht zu verlieren.

Die Arbeit ist folgendermaßen aufgebaut:

Nach der Einleitung in das Thema in Kapitel 1 schildert Kapitel 2 den Aufbau der 1- Transistor Speicherzelle und erklärt die grundlegenden Schreib-/Lese-Funktionen im DRAM. Die Arbeitsweise der Leseverstärker (*Sense Amps*), die als Differenzverstärker ausgeführt sind, wird erläutert und dabei die *refresh*-Operation erklärt.

Kapitel 3 führt in die spezielle Problematik dieser Arbeit ein. Die Retentionzeit t_{Ret} wird als maximale Haltezeit ohne Informationsverlust zwischen zwei *Refreshes* definiert. Mit Hilfe des *Charge-Sharing* Prinzips wird eine sehr einfache Formel (die so genannte Retention-Formel) formal hergeleitet, die im Weiteren als einfaches Modell dient. Neben den Zell-Leckströmen gehen noch weitere Parameter der Speicherzelle wie z.B. Bitleitungs- und Speicherkapazität, Differenzverstärker-Offset usw. in die Formel ein. In dem einfachen Modell findet auch die kapazitive Kopplung zwischen benachbarten Bitleitungen Beachtung.

Die Haltezeiten t_{Ret} der Zellen eines Speicherchips sind jedoch nicht alle identisch, sondern unterliegen einer breiten Verteilung. Unter Verwendung der Monte Carlo Technik und einem eigens dafür entwickelten MATLAB-Programm wird in Kapitel 4 der Einfluss der einzelnen Parameter auf die gesamte Retentionverteilung beispielhaft untersucht. Das Programm erlaubt die Simulation der Retentionverteilung unter veränderten Parametern, wie z.B. doppelte bzw. halbierte Bitleitungslänge oder vergrößerter Speicherkapazität. Als Hauptursache für die breite t_{Ret}-Verteilung stellt sich die Verteilung der Leckströme heraus.

Deshalb fasst Kapitel 5 die möglichen Leckstrompfade der untersuchten Speichertechnologie und deren Spannungsabhängigkeiten zusammen.

Kapitel 6 beschäftigt sich mit Charakterisierungstechniken unter Beachtung der besonderen Randbedingungen, die sich aus der sehr geringen Auftrittswahrscheinlichkeit von Tailzellen für die Charakterisierung ergibt. Es stellt sich heraus, dass Teststrukturen für Tailuntersuchungen nicht eingesetzt werden können und die in der Arbeit zur Lösung der Charakterisierungsschwierigkeiten entwickelte Einzelzellcharakterisierung wird vorgestellt.

In Kapitel 7 werden die Ergebnisse der elektrischen Charakterisierung an DRAM Speicherbausteinen vorgestellt. Die Temperaturabhängigkeit der Retentionzeit (ausgedrückt durch die Aktivierungsenergie) erlaubt dabei Rückschlüsse auf Leckstrompfade und Mechanismen. Generationsleckströme im während der Haltezeit in Sperrrichtung geschalteten kondensatorseitigen pn-Übergang (später auch einfach *Junction* oder *Node* genannt) erweisen sich als Hauptursache für den Informationsverlust.

Durch Vergleich der Messdaten mit theoretischen Überlegungen und Abschätzungen kann in Kapitel 8 der grundlegende Mechanismus benannt werden. Daraus ergeben sich Ansätze zur Verifikation des Modells in Kapitel 9, die schließlich durch Fertigungsversuche erprobt wurden und zu einer abschließenden Verbesserung führten.

Anmerkung zu den Einheiten:
Retentionzeiten und Fehlerzahlen gehören zu den wichtigsten Parametern einer DRAM Technologie. Sie erlauben Rückschlüsse auf Produktivität und Ausbeute. Aus diesem Grund werden in Veröffentlichungen die Retentionzeiten meist einheitenlos und Fehlerzahlen in normierter Form angegeben. Ich bitte um Verständnis, dass auch in dieser Arbeit keine Angaben gemacht werden können, die Rückschlüsse auf Ausbeute und Produktivität der Speichertechnologie von Qimonda erlauben. Durch die ganze Arbeit hindurch wurde dieser Vorgabe seitens Qimonda Rechnung getragen. Durch diese Einschränkung gehen jedoch keine Zusammenhänge verloren, die für das physikalische Verständnis der hier dargestellten Problematik erforderlich sind.

Kapitel 2

DRAM Grundlagen

Das Grundprinzip der dynamischen Speicherzelle hat sich seit ihrer Erfindung 1966 bis heute nicht verändert. Noch immer werden Speicherzellen durch Transistor- und Kondensatorelemente in hochintegrierten Schaltungen auf Siliziumbasis realisiert. Jede Speicherzelle repräsentiert ein einzelnes Bit in Form einer logischen Null oder Eins. Die Kenntnis des Aufbaus und der Funktion eines Speicherchips ist für das weitere Verständnis der Arbeit unbedingt erforderlich. Darauf soll in diesem Kapitel im nötigen Detail eingegangen werden. Darüber hinausgehende Detailinformationen geben z.B. [Wid96, Ito01, Kee01].

2.1 Zellaufbau

Die DRAM-Zelle ist eine vom Prinzip einfache Speicherzelle. Sie besteht nur aus zwei Bauelementen: einem Transistor und einem Kondensator. Die Information wird dabei auf dem Kondensator in Form von zwei Ladungszuständen gespeichert. Die Ansteuerung der Zelle geschieht über den Transistor, der als Schalter fungiert. Er kann die Ladung im Kondensator isolieren oder zum Ein- und Auslesen eines Datums einen elektrisch leitenden Pfad öffnen. Abbildung 2.1 zeigt das Ersatzschaltbild einer DRAM Speicherzelle. Das *Gate* des Transistors ist mit der Wortleitung (WL) verbunden. Liegt der Pegel dieser Signalleitung auf „low", dann befindet sich der Transistor im hochohmigen Zustand. Die Ladung des Kondensators ist isoliert und bleibt gespeichert. Zum Schreiben oder Lesen der Speicherzelle wird der Signalpegel der WL auf „high" angehoben. Der Kanal des Transistor ist dann leitfähig und verbindet den Kondensator mit der Bitleitung (BL). Beim Schreiben gleicht sich die Ladung des Kondensators entsprechend dem Pegel der Bitleitung an, auf der die zu schreibende Information liegt. Beim Lesen verteilt sich die im Kondensator gespeicherte Ladung auf die nach dem Öffnen der Wortleitung parallel geschalteten Kapazitäten der Bitleitung und des Speicherkondensators. Das Potenzial der

Bitleitung steigt bzw. fällt dabei je nach Ladungszustand des Kondensators und signalisiert dadurch, ob eine „1" oder eine „0" gespeichert war. Aufgrund des großen Kapazitätsunterschiedes zwischen Bitleitung und Zellkondensator (ungefähr 5:1 bei modernen DRAMs) entwickelt sich auf der Bitleitung nur eine sehr kleine Potenzialänderung, die anschließend mittels eines Differenzverstärkers auf den vollen Informationspegel verstärkt werden muss. Der Lesevorgang wird in Abschnitt 2.3.5 im Detail erklärt werden.

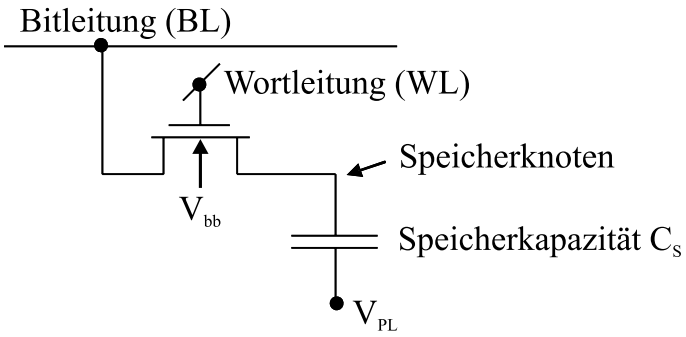


Abbildung 2.1: Ersatzschaltbild einer DRAM Speicherzelle. Die Zelle besteht nur aus zwei Bauelementen: einem Kondensator und einem Transistor. Die Information wird als Ladungszustand auf dem Kondensator gespeichert. Durch Aktivieren der Wortleitung, die mit dem *Gate* des Transistors verbunden ist, wird eine leitende Verbindung zwischen Speicherknoten und Bitleitung hergestellt, über die Ladung gelesen bzw. geschrieben werden kann.

Realisierung in Silizium

Die Realisierung der Bauelemente einer Speicherzelle im Silizium hat sich mit der Entwicklung hin zu kleineren Strukturgrößen stark verändert. Bis einschließlich der $1\,MBit$ Generation konnte der Kondensator in planarer Form an der Oberfläche des Siliziums realisiert werden. Mit weiter abnehmender Zellfläche stand auch weniger Kondensatorfläche A_S zur Verfügung und die drohende Abnahme der Speicherkapazität C_S gemäß

$$C_S = \epsilon\epsilon_0 \frac{A_s}{t_{diel}} \tag{2.1}$$

konnte aufgrund von Leckströmen nicht weiter durch geringere Dielektrikadicken t_{diel} ausgeglichen werden. Da über die verschiedenen Speichergenerationen hinweg die Kondensatorkapazität konstant mindestens $25\,fF$ betragen muss, wurden ab der $4\,MBit$ Generation andere „nicht-planare" Kondensatoren, die trotz weiterer Zellflächenreduktion eine gleichbleibende Kapazität erlauben, integriert. Dafür gibt es prinzipiell zwei Möglichkeiten, welche die weltweit führenden Speicherhersteller in zwei Lager unterteilt. Der Kondensator kann entweder in einem Graben (-> *trench*-DRAM) oder über den Auswahltransistoren (-> *stacked*-DRAM) strukturiert werden (siehe Abbildung 2.2). Um die Kapazität weiterhin konstant halten zu können werden darüber hinaus in kommenden DRAM-Technologien so genannte *high-k*-Materialien mit höheren Dielektrizitätskonstan-

ten wie z.B. Aluminiumoxid oder Hafniumoxid Anwendung finden. Außerdem werden verschiedene Techniken zur Oberflächenvergrößerung, wie z.B. HSG (hemispherical silicon grains) oder die nasschemische Aufweitung der Gräben in der Tiefe (DT-bottle) Verwendung finden.

In der Vergangenheit gab es viele Diskussionen über die Vor- und Nachteile von *stacked* wie auch *trench*-DRAM. Fakt ist, dass bis heute keines der beiden Konzepte als eindeutiger Sieger hervor ging und sich gegenwärtig beide DRAM-Konzepte in der Massenproduktion befinden. Da diese Arbeit in Zusammenarbeit mit Qimonda Dresden GmbH & Co. OHG durchgeführt wurde, erfolgten alle Untersuchungen ausschließlich an der von Qimonda produzierten *trench*-DRAM Technologie.

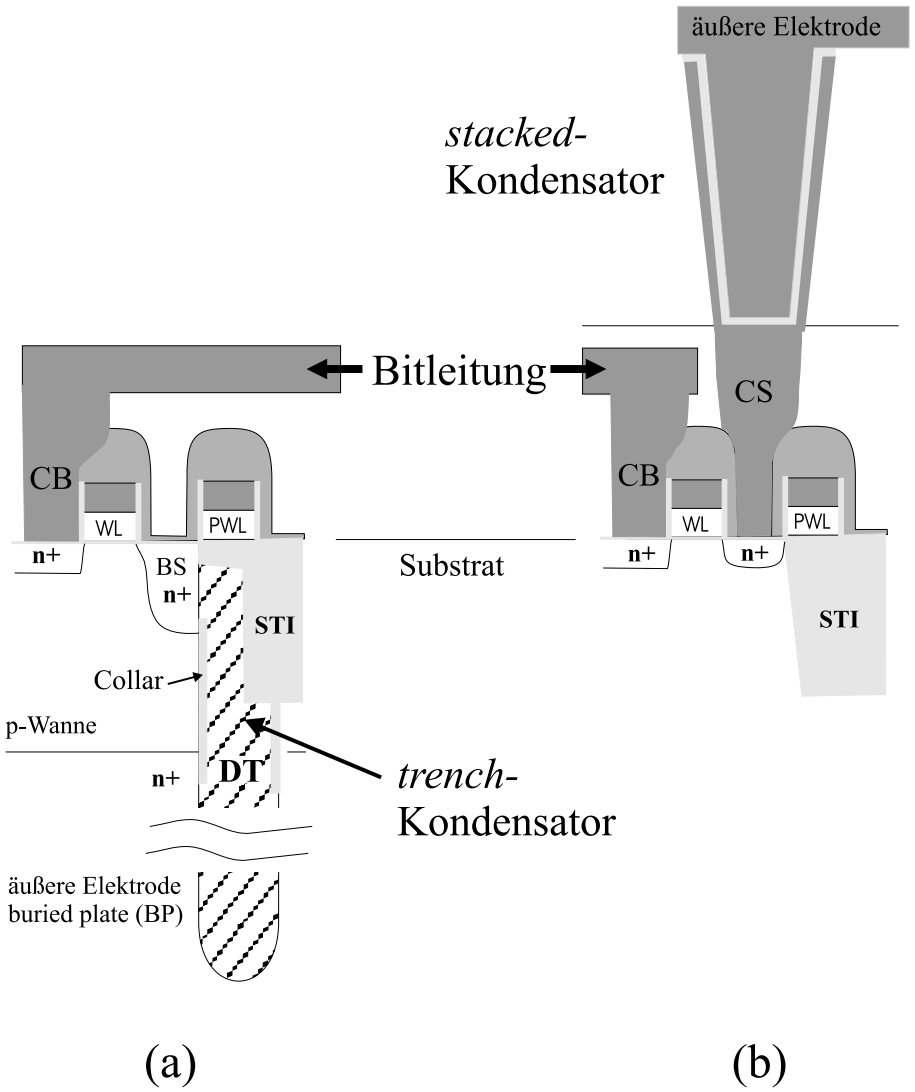


Abbildung 2.2: Vergleich der beiden Hauptintegrationsvarianten im Querschnitt. (a) Beim *trench*-DRAM wird der Speicherkondensator in einem tiefen Graben (DT) ausgebildet, während er bei (b) *stacked*-DRAM oberhalb der Transistoren entsteht.

Die Qimonda $110\,nm$-Technologie basiert auf der so genannten BEST Zellstruktur [Nes93, Bro95], die der in Abbildung 2.2a gleicht. Eine Speicherzelle setzt sich aus einem Auswahltransistor und einem *trench*-Speicherkondensator (kurz DT für *deep trench*) zusammen. Die DT-Innenelektrode besteht aus Arsen-dotiertem Polysilizium, die äußere Elektrode ist durch die vergrabene Platte (kurz BP für *buried plate*) gegeben, die durch Ausdiffusion aus einer Opfer-TEAS-Schicht bzw. durch Gasphasendotierung vor der Abscheidung des Kondensatordielektrikums entsteht und durch zusätzliche tiefe P-Implantationen verbunden wird. Die BP umschließt das gesamte Zellenfeld und isoliert dadurch die p-Wanne vom Substrat. Das Dielektrikum des Kondensators wird durch eine Nitridabscheidung gefolgt von einer teilweisen Aufoxidation gebildet (NO Dielektrikum). Im oberen Teil ist der DT von einer dickeren Oxidschicht umgeben. Dieser Oxidkragen wird *Collar* genannt und bildet das Gateoxid eines parasitären vertikalen Transistors (siehe Abschnitt 5.3.3). Die elektrische Verbindung vom DT-Inneren zum Kanal des Auswahltransistors ist durch einen vergrabenen Anschluss (kurz BS für *buried strap*) realisiert. Dieser wird durch Arsen-Ausdiffusion aus dem DT und flache P-Ionenimplantation nach der Gatestrukturierung gebildet. Beim Auswahltransistor der BEST-Zelle handelt es sich um einen planaren asymmetrischen Transistor mit Poly/WSi Gate-Stapel. Abbildung 2.3 zeigt Rasterelektronenmikroskop-Aufnahmen einer Zelle, wie sie in der vorliegenden Arbeit untersucht wurde.

2.2 Speichermatrix

Die große Anzahl von Speicherzellen eines DRAM-ICs sind in einer Matrix aus Zeilen und Spalten angeordnet. Dabei bilden die Wortleitungen die Zeilen und die Bitleitungen die Spalten. Jede Zelle in der Matrix ist somit durch Kombination einer Wortleitungs- und einer Bitleitungsadresse eindeutig bestimmt. Um wertvolle Siliziumfläche zu sparen und die Produktionskosten pro Chip kleinstmöglich zu halten, werden die Speicherzellen immer so dicht wie nur möglich angeordnet. Eine Anordnung, welche den durchschnittlichen Platzbedarf pro Zelle auch nur ein wenig reduziert, kann aufgrund der hohen Zellzahl zu einer signifikanten Verkleinerung der Gesamtgröße des Chips und damit zu einer erheblichen Kostenreduzierung führen. Beim *trench*-DRAM werden die Zellen gegenwärtig im Wesentlichen auf zwei Arten angeordnet: *Merged Isolation and Node Trench* (MINT) [Ken92] und *Checkerboard* (CKB). Abbildung 2.4 zeigt REM-Aufnahmen von Speicherfeldern beider Layouts in Draufsicht. Sichtbar sind die durch das STI getrennten aktiven Gebiete (AA), d.h. die Bereiche in denen sich die Auswahltransistoren befinden. Die Metallisierungslagen, Gate-Bahnen (GC) sowie das STI-Oxid wurden zu diesem Zweck entfernt. Der ursprüngliche Verlauf der Wortleitungen ist angedeutet. Die Bitleitungen verlaufen senkrecht dazu über den aktiven Gebieten. In den ovalen dunklen Bereichen befand sich vor der Präparation das Collar-Oxid, wodurch die *trench*-Kondensatoren markiert wer-

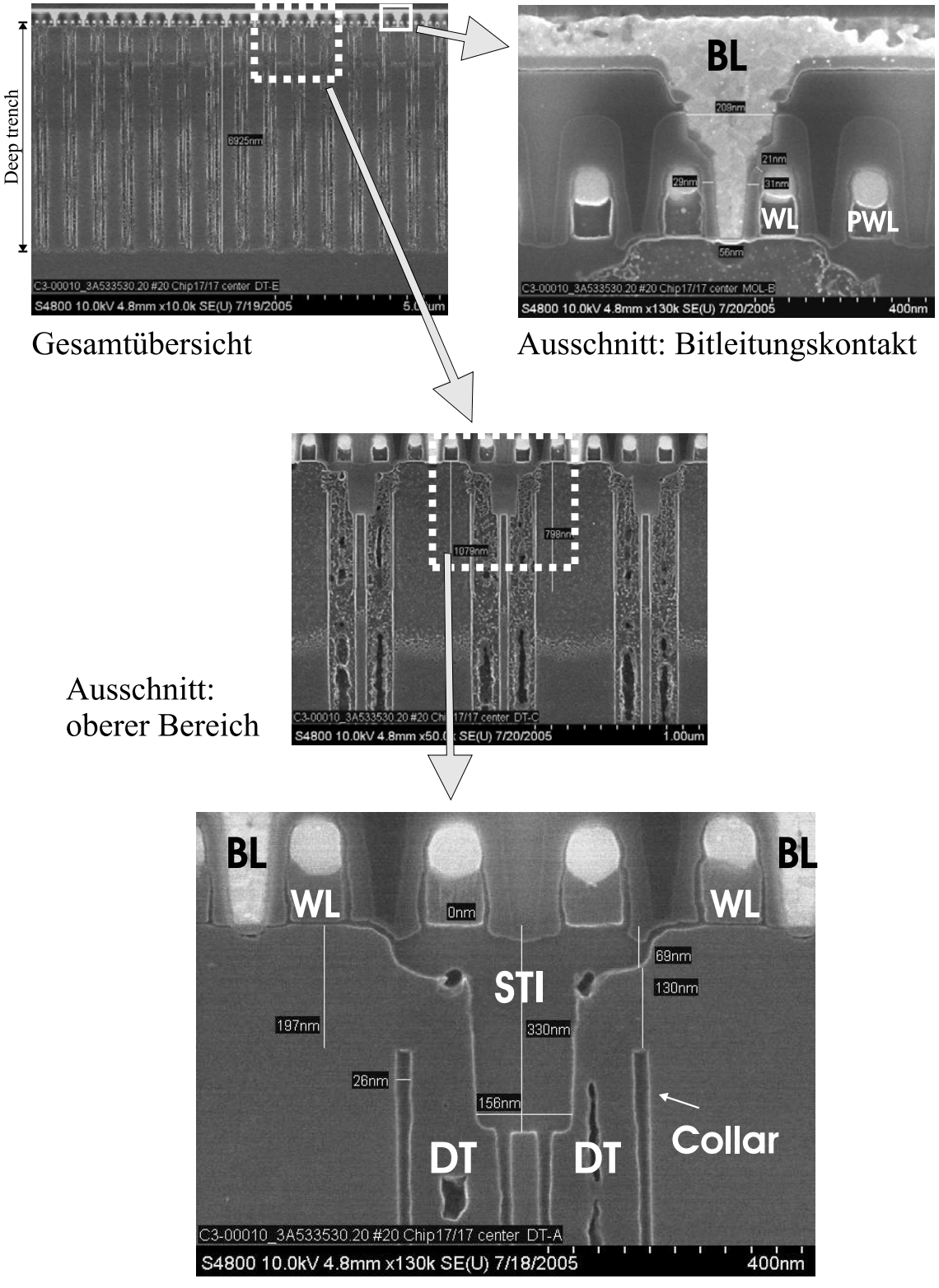


Abbildung 2.3: Rasterelektronenmikroskopische Aufnahmen einer Qimonda DRAM Zelle in 110 nm-Technologie. Beim MINT-Design liegen jeweils die DTs zweier benachbarter Zellen dicht zusammen.

den. Beim *MINT*-Layout sind auf jedem aktivem Gebiet (AA) zwei Zelltransistoren integriert, die sich einen BL-Kontakt (BC) in der Mitte des aktiven Gebietes teilen. Jeweils zwei *trench*-Kondensatoren von Zellen benachbarter AAs bilden ein so genanntes MINT-Pärchen. Beim CKB-Layout befindet sich auf jedem AA nur ein Transistor, wodurch die doppelte Anzahl von Bitleitungskontakten notwendig ist. Ein Vorteil des CKB besteht darin, dass die Kondensatoren über die Fläche gleich verteilt sind, d.h. in jede Richtung den gleichen Abstand zueinander haben. Dadurch können Kondensatoren mit einer größeren Oberfläche und somit höherer Kapazität gebaut werden.

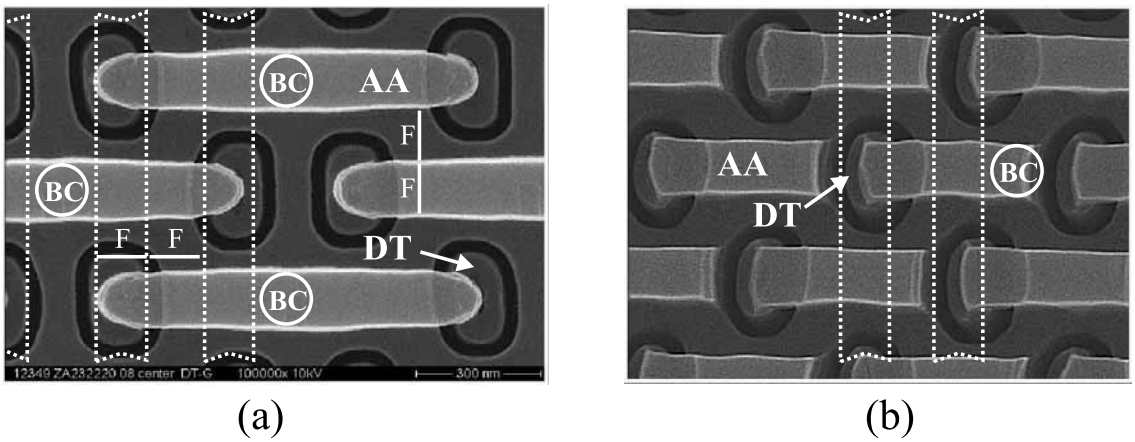


Abbildung 2.4: REM Aufnahmen der Speichermatrix in Draufsicht. Die Metallisierungslagen, Gate-Bahnen und das STI-Oxid wurden in der Präparation entfernt. (a) *Merged Isolation and Node Trench* (MINT) wurde bis einschließlich der $110\,nm$-Generation eingesetzt, (b) *Checkerboard* (CKB) ist seit der $90\,nm$ *trench*-Technologie das bevorzugte Layout.

Beide Matrix-Layouts von Abbildung 2.4 sind für das so genannte *Folded-Bitline* Konzept ausgelegt. Daneben existiert noch das so genannte *Open-Bitline* Konzept. Abbildung 2.5 zeigt das Schema beider Konzepte im Vergleich. Beim *Folded-Bitline* Konzept laufen die zwei Bitleitungen eines Differenzverstärkers (SA) parallel nebeneinander über das Speicherfeld. Übersprecheffekte und elektrische Störungen betreffen dadurch beide Datenleitungen gleichermaßen und fallen aufgrund der differenziellen Verstärkung nicht ins Gewicht. Da jedoch immer nur Zellen einer der Bitleitungen eines Bitleitungspaares gelesen werden können (die andere Bitleitung muss als Referenz auf Vorladeniveau bleiben), darf eine Wortleitung immer nur Zellen jeder zweiten Bitleitung öffnen. Deshalb muss über die Grundfläche jeder Zelle neben der Wortleitung (WL) dieser Zelle auch eine zweite passive Wortleitung (PWL) geführt werden (siehe Abbildung 2.5a). Dies führt zu dem Nachteil einer minimalen Zellfläche von $8\,F^2$ (x: $2 \cdot WL = 4\,F$, y: $1 \cdot BL = 2\,F$), wobei F der so genannte *DRAM 1/2-Pitch* ist (=die Hälfte der kleinsten lithografisch herstellbaren periodischen Struktur, d.h. Abstand der WLs und BLs). Beim *Open-Bitline*-Layout (Abbildung 2.5b) laufen die zwei Bitleitungen eines Verstärkers in verschiedene Richtungen und damit Zellenfelder. Durch das Öffnen einer Wortleitung werden dadurch automatisch nur Zellen an einer der zwei Bitleitungen eines Verstär-

kers gelesen. Es muss keine zweite Wortleitung über die Grundfläche einer Zelle geführt werden und die minimale Fläche pro Speicherzelle kann auf ungefähr $6\,F^2$ reduziert werden. Damit verbunden sind jedoch Schwierigkeiten beim Layout der Differenzverstärker, welche nun dichter zusammenrücken (ein Verstärker alle zwei anstatt vier Bitleitungen) sowie ein schlechteres Signal/Rausch-Verhältnis. Unter Experten war es deshalb lange Zeit umstritten, ob dieses Layout einen stabilen Speicherbetrieb zulässt. Dass dies möglich ist hat der amerikanische Hersteller Micron gezeigt, wofür das Unternehmen 2004 den *Semiconductor Insights Award* erhielt. Die $6\,F^2$ Micron DRAM-Zelle ist seither in der Volumenfertigung. Auch Samsung hat kürzlich Speicherbausteine im $6\,F^2$-Layout angekündigt [Oh05]. Die in dieser Arbeit untersuchten $110\,nm$-Speicherchips von Qimonda wurden im MINT-Design und *Folded-Bitline* Konzept gefertigt.

Einfluss der Nachbarzellen

Um Kosten einzusparen, geht die Entwicklung immer weiter zu kleineren Strukturbreiten F. Gleichzeitig muss durch die kleinen Abstände bedingtes elektrisches Übersprechen verhindert werden. Die künftigen Speicherprodukte werden bezüglich gegenseitiger Beeinflussung immer grenzwertiger werden. Deshalb müssen auch beim Test nicht nur die einzelnen Zellen, sondern auch deren Nachbarzellen mit in Betracht gezogen werden. Das bedeutet, dass der Ladungszustand der Nachbarzellen Einfluss auf die untersuchte Zelle haben kann. Abbildung 2.6 zeigt einen Ausschnitt aus einem Zellenfeld in MINT-Architektur. Demzufolge hat eine MINT-Zelle drei nächste Nachbarn: eine Zelle an der gleichen und jeweils eine an den zwei benachbarten Bitleitungen. Beim CKB-Layout ist jeder Speicherkondensator von vier gleichweit entfernten Nachbarn umgeben, die über die benachbarten Bitleitungen gelesen bzw. geschrieben werden (nicht gezeigt).

Adress- und Datenscrambling

Aus schaltungstechnischen Gründen sind die Zellen in der Praxis nicht einfach in der Reihenfolge ihrer elektrischen Adresse in der Speichermatrix angeordnet, sondern es findet intern ein *Address-Scrambling* statt, d.h. die Leitungen sind in einer anderen als der physikalischen Reihenfolge „verdrahtet". Abbildung 2.6b zeigt den Unterschied zwischen physikalischer und elektrischer Adresse anschaulich. Die Wortleitungen sind darin bezüglich ihrer elektrischen Adresse benannt. Würde man diese in ihrer Reihenfolge aktivieren (WL0, WL1, WL2, WL3), so werden nacheinander die gespeicherten Informationen aus den Kondensatoren DT2, DT1, DT3 und DT4 ausgelesen. Physikalisch gesehen ist die Reihenfolge jedoch DT1, DT2, DT3, DT4. Um in der physikalischen Reihenfolge zu schreiben oder zu lesen, müssen die Wortleitungen deshalb in der Reihenfolge WL1, WL0, WL2 und WL3 angesprochen werden. Neben dem *Scrambling* für Adressen existiert

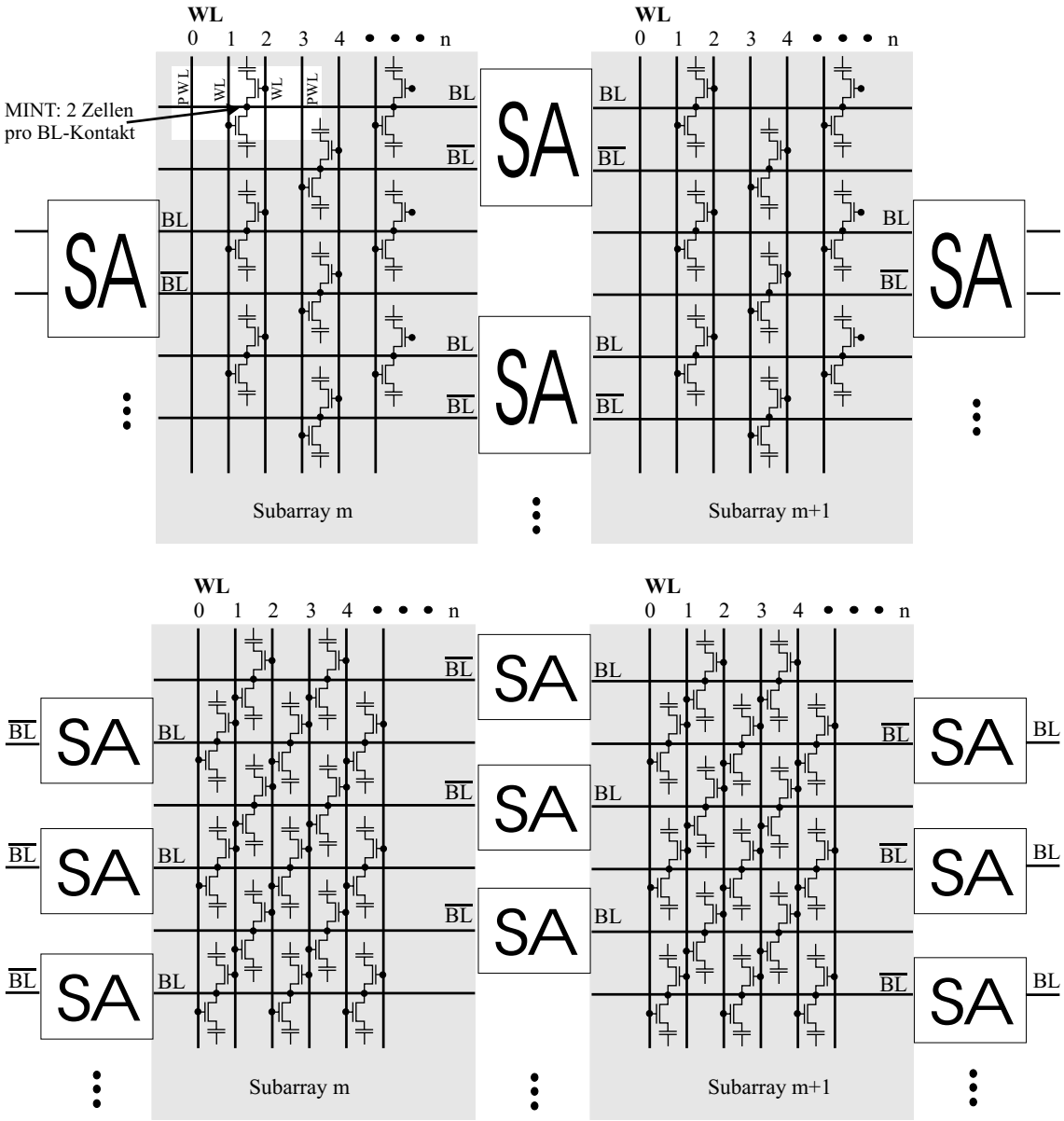


Abbildung 2.5: (a) *Folded-Bitline*-Design: Die beiden Bitleitungen eines Differenzverstärkers (SA) laufen parallel über das Speicherfeld. Dadurch wirken elektrische Störungen auf beide Bitleitungen und fallen aufgrund der differenziellen Verstärkung nicht ins Gewicht. Der Nachteil besteht in dem größeren Flächenbedarf pro Zelle von $8\,F^2$. (b) *Open-Bitline*-Design: Die beiden Bitleitungen eines SAs laufen in unterschiedliche Richtungen. Dieses Design erfordert die kleinere Zellfläche von nur $6\,F^2$, bietet jedoch Nachteile hinsichtlich Signal/Rausch-Verhältnis und SA-Design.

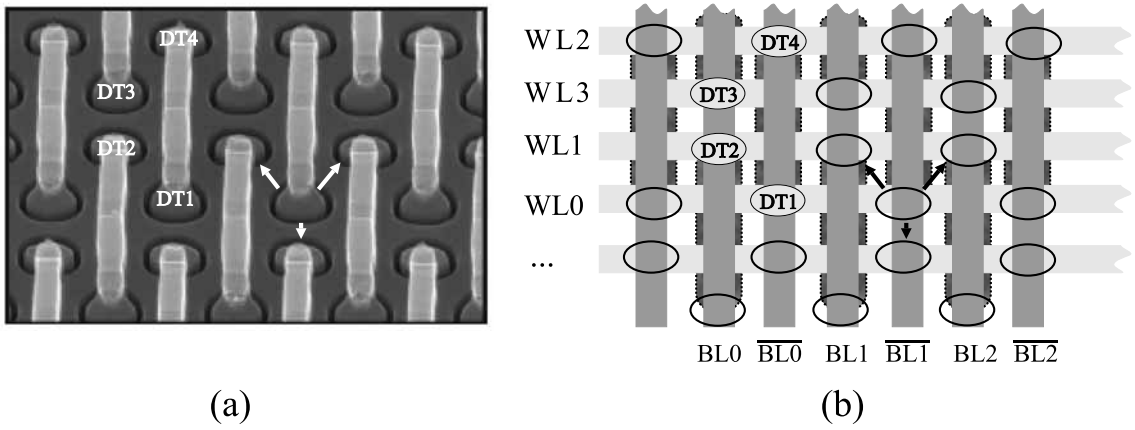
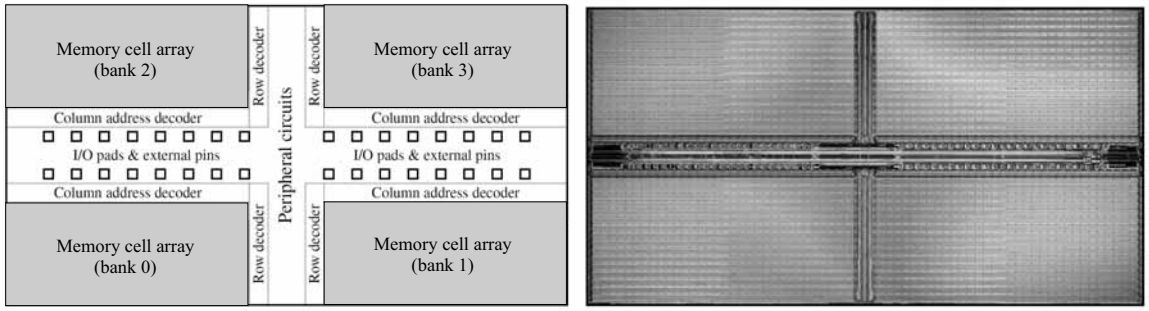


Abbildung 2.6: Übersicht der Nachbarschaftsbeziehungen und des Scramblings. (a) REM Aufnahme und (b) Layout-Skizze.

auch ein *Scrambling* für Daten. Dabei werden logische Daten abhängig von der Zelladresse physikalisch inverse abgebildet, d.h. für bestimmte Zelladressen wird eine logische „1" als physikalische „0" bzw. umgekehrt gespeichert. Für diese Arbeit sind die schaltungstechnischen Hintergründe des Adress- und Datenscramblings weniger von Interesse. Wichtig für die elektrische Charakterisierung von Leckströmen sind jedoch die physikalischen Adressen und Daten, d.h. um Nachbarschaftsbeziehungen und Ladungszustand des Kondensators korrekt zu berücksichtigen, müssen Adress- und Datenscrambling bei der Charakterisierung berücksichtigt werden. Ein *Descrambler* sorgt für die richtige Umrechnung von physikalischen Datentopologien und Adressen auf die elektrischen Äquivalente. Die *Scramble*- und *Descramble*-Funktionen gehören nicht zur Spezifikation eines Speicherbausteines und werden von den Speicherherstellern geheim gehalten.

Chip-Layout

Abbildung 2.7 zeigt beispielhaft das Layout eines modernen DRAM-Bausteins. Das Speicherfeld wird meist in vier oder acht Speicherbänke unterteilt, die parallel und unabhängig voneinander betrieben werden können. In der kreuzförmigen Fläche zwischen den Speicherbänken sind die Peripherie-Schaltkreise, wie z.B. Adressdekoder, Ein-Ausgabe-Schaltkreise und Spannungsgeneratoren untergebracht. Darüber hinaus befinden sich auch die Anschlusspads in diesem Bereich. Besonders vorteilhaft an dieser symmetrischen Anordnung sind die kurzen und ähnlich langen Signalleitungen.



(a) (b)

Abbildung 2.7: Typisches Layout eines DRAM-Speicherbausteins. (a) Aufbauskizze: Die Speichermatrix ist in vier Speicherbänke unterteilt. In dem kreuzförmigen Bereich zwischen den Bänken sind Peripherie-Schaltungen und Kontaktpads untergebracht. (b) optisches Bild zum Vergleich.

2.3 Elektrisches Funktionsprinzip

Die Speichermatrix des DRAMs setzt sich intern je nach IC-Typ aus mehreren parallel betriebenen Speicherfeldern zusammen. Diese Aufteilung muss nicht unbedingt der Aufteilung in Speicherbänke entsprechen. Die Aufteilung ist der Organisation zu entnehmen, z.B. kann ein $512\,MBit$ Chip als $128x4$, $64x8$ oder $32x16$ organisiert sein. Die erste Zahl in der Organisationsangabe gibt die Größe eines einzelnen Speicherfeldes in $MBit$ an, während die zweite Zahl für die Anzahl der parallelen Speicherfelder steht. Letztere steht gleichzeitig für die Datenwortbreite des ICs, d.h. die Anzahl der angeschlossenen Datenleitungen. Aus einer größeren Datenwortbreite folgt bei gleicher Taktfrequenz ein höherer Datendurchsatz. Gleichzeitig können aufgrund der verfügbaren Datenwortbreite des Prozessors und damit des Datenbuses auf dem Motherboard (heute meist $64\,bit$) nur weniger Chips zu einem Speichermodul aufgebaut werden, wodurch sich die maximal mögliche Speichergröße pro Modul reduziert. Deshalb werden bei Grafikspeichern, die vor allem einen sehr hohen Datendurchsatz benötigen, $x16$ oder sogar $x32$ Organisationen benutzt, während für Servermodule, die für möglichst große Speicherdichten ausgelegt sind, $x4$ organisierte Speicherchips eingesetzt werden. Die Aufteilung in Zeilen und Spalten wird durch den Parameter *Refresh Cycle* in den Datenblättern angegeben. Die *Refresh-Cycle* Angabe bezeichnet die Anzahl der Zeilen bzw. Wortleitungen, die regelmäßig aufgefrischt werden müssen (Ref. $1k$, $2k$, $4k$, $8k$). Je kleiner der *Refresh*-Parameter, desto größer ist die Verfügbarkeit des Chips, da die für den *Refresh* einer Zeile benötigte Zeit durch die Technologie bestimmt ist.

2.3.1 Blockschaltbild

Abbildung 2.8 zeigt ein vereinfachtes Blockschaltbild eines Speicherchips. Im Mittelpunkt steht die Speichermatrix, die im Allgemeinen in mehrere Bänke unterteilt ist. Jede Speicherbank besitzt eigene Zeilen- und Spaltendekoder, einen Vorladeschaltkreis und Leseverstärker. Der Chip ist mit der Außenwelt durch drei „Leitungsbündel" verbunden:

- Adressleitungen ($A0 - A12$, $BS0$, $BS1$), welche zum Adresspuffer führen
- Datenleitungen ($DQ0 - DQ7$), welche zum Datenpuffer führen
- Steuerleitungen ($\overline{CS}$, $\overline{RAS}$, $\overline{CAS}$, $\overline{WE}$), welche Befehle an die interne Steuerung übermitteln.

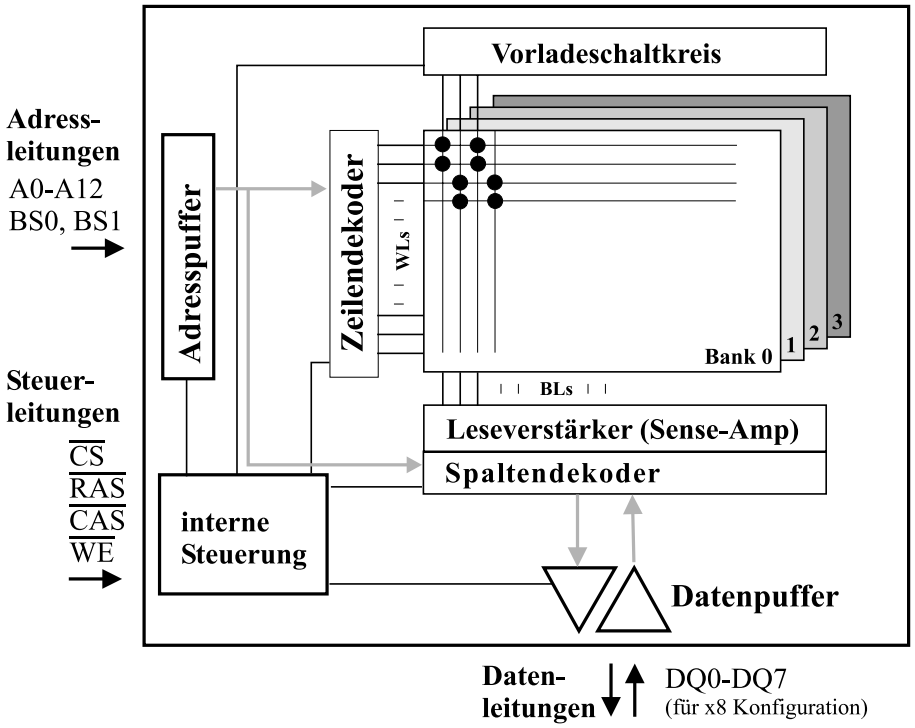


Abbildung 2.8: Vereinfachtes Blockschaltbild eines DRAM-Speicherchips.

2.3.2 Befehlssatz

Durch Steuersignale werden dem DRAM-Chip eine Reihe von Befehlen übermittelt. Die Kodierungen der wichtigsten Steuerbefehle sind in Tabelle 2.1 zusammengefasst und werden im Folgenden kurz erläutert.

Activate öffnet die durch die anliegende Zeilenadresse bestimmte Wortleitung und startet den Verstärkungsvorgang

Read wählt die Spalte der anliegenden Spaltenadresse aus der geöffneten Zeile aus und leitet das Datum von den *Sense-Amps* an die Peripherie weiter

Write holt Daten vom den DQ-Pins ab und führt sie den ausgewählten Zellen zu

Precharge alle Wortleitungen werden geschlossen; alle Bitleitungen werden auf $V_{BLH}/2$ vorgeladen

NOP Der Befehl führt keine Funktion aus. Der Status des Speichers bleibt unverändert, d.h. die Zeitspanne des vorhergehenden Befehls wird um einen Zyklus verlängert.

Befehl	Abkürzung	$\overline{CS}$	$\overline{RAS}$	$\overline{CAS}$	$\overline{WE}$
Deselect	DES	H	X	X	X
Activate	ACT	L	L	H	H
Read	Read	L	H	L	H
Write	Write	L	H	L	L
Precharge	PRE	L	L	H	L
No Operation	NOP	L	H	H	H

Tabelle 2.1: Kodierung der wichtigsten DRAM Befehle durch die Steuersignale $\overline{CS}$, $\overline{RAS}$, $\overline{CAS}$ und $\overline{WE}$. L=low, H=high, X=nicht relevant.

Darüber hinaus existieren noch weitere Befehle, die vor allem zur Erhöhung des Datendurchsatzes dienen bzw. den *Refresh* steuern (siehe z.B. [Qim05]). Die Befehle sind nicht unabhängig voneinander, sondern müssen in einer bestimmten Reihenfolge aufeinander folgen. Die Befehlsfolgen und zeitlichen Abhängigkeiten für die Schreib- und Leseoperationen werden in den folgenden Abschnitten beschrieben werden. Sowohl das Lesen als auch das Schreiben sind hierbei vereinfacht dargestellt. In der Realität gibt es darüber hinaus noch eine Reihe von Feinheiten wie z.B. *burst*, *bank interleave*, *prefetch* usw., die an dieser Stelle jedoch nicht von großer Bedeutung sind (siehe dazu z.B. [Ito01]). Eine Zusammenfassung typischer Werte für die wichtigsten *Timing*-Parameter gibt Tabelle 2.2. Die darin angegebenen Wertebereiche entsprechen der Entwicklung vom *Single Data Rate (SDR)* im Jahr 2000 bis zu den aktuellen *Double Data Rate II (DDR2)* Bausteinen im Jahr 2006. Interessant daran ist, dass in dieser Zeit zwar die Chip-Taktfrequenzen um den Faktor vier, die maximalen Datenraten sogar um den Faktor acht angestiegen sind, die *Row Cycle Time* t_{RC}, die ein Maß für die Zugriffszeit im ungünstigsten Fall zweier unabhängiger Speicherzugriffe innerhalb der selben Bank darstellt, jedoch lediglich von $70\,ns$ auf $55\,ns$ reduziert werden konnte. Das bedeutet, dass das eigentliche Lesen einer Zelle nur unwesentlich schneller geworden ist und die enorme Erhöhung der Datenrate nur von höherer Parallelität und anderen Verbesserungen lebt.

Parameter	Beschreibung	Wert [ns]
t_{CK}	Clock Cycle time	$2.5 - 10$
t_{RAS}	Row address strobe time	$30 - 50$
t_{RC}	Row Cycle time	$55 - 70$
t_{RCD}	Row-column delay time	$10 - 30$
t_{RP}	Row precharge time	$10 - 30$
t_{WR}	Write recovery time	$10 - 30$

Tabelle 2.2: Typische *Timing*-Parameter für Speicherchips. Der Wertebereich beinhaltet Single Data Rate (SDR) DRAM vom Jahr 2000 bis zu aktuellen Double Data Rate II (DDR2-800) DRAMs im Jahr 2006.

2.3.3 Leseoperation

Der zeitliche Ablauf einer für DDR Speicherbausteine typischen Leseoperation ist in Abbildung 2.9 dargestellt. In der obersten Zeile sind die Befehle für jeden Taktzyklus mit Länge t_{CK} angegeben. Im Ausgangszustand sind alle Wortleitungen geschlossen und die Bitleitungen auf das Potenzial $V_{BLH}/2$ vorgeladen. Der Lesevorgang beginnt mit dem Bereitstellen der Zeilenadresse an den Adresseingängen. Durch den Befehl *Activate* wird das Steuersignal $\overline{RAS}$ low und signalisiert das Anliegen einer gültigen Zeilenadresse. Diese wird in den Adresspuffer des DRAMs eingelesen und an den internen Zeilendekoder weitergeleitet. Dieser dekodiert die Zeilenadresse und aktiviert die entsprechende Zeile(=Wortleitung). Das Auslesen und Verstärken aller Zellen dieser Zeile wird dadurch gestartet (Details zur Verstärkung werden in Abschnitt 2.3.5 beschrieben). Nach der in der Spezifikation definierten Zeit t_{RCD} ist die Zeile verstärkt und der *Read*-Befehl kann gegeben werden. Abhängig von der Taktfrequenz und der spezifizierten t_{RCD} müssen gegebenenfalls *NOP*-Befehle eingefügt werden, die keine eigene Funktionalität besitzen. Beim *Read*-Befehl geht das Steuersignal $\overline{CAS}$ auf low und signalisiert die zwischenzeitlich an den Adresseingängen anliegende Spaltenadresse. Der Adresspuffer liest die Adresse ein und leitet sie diesmal an den Spaltendekoder weiter. Dieser wählt die gewünschte Spalte aus der inzwischen ausgelesenen und verstärkten Zeile aus und leitet das Datum an den Ausgangsdatenpuffer weiter, welcher das Datum schließlich am DQ-Pin zur Verfügung stellt. Das serielle Einlesen der Zeilen- und Spaltenadressen wird mit *Address-Multiplexing* bezeichnet. Dadurch sind weniger Adress-Pins notwendig und Kosten können eingespart werden. Auf die Lesegeschwindigkeit hat das *Address-Multiplexing* keine Auswirkung, da der DRAM-Aufbau ohnehin das Öffnen einer Zeile zeitlich vor der Auswahl der Spalte erfordert. Die Zeit zwischen dem *Read*-Befehl und Ausgabe des Datums wird in vielfachen der Taktzykluszeit t_{CK} gemessen und mit CAS-Latency (CL) bezeichnet. Mit dem *Precharge*-Befehl, der frühestens nach der Zeit t_{RAS} auf den *Activate*-Befehl folgen darf, wird der Lesezyklus abgeschlossen. Die Wortleitung wird wieder geschlossen und die Bitleitungen auf $V_{BLH}/2$ vorgeladen. Da die Wortlei-

tung während des gesamten Verstärkungsvorgangs geöffnet war, wurden auch die Informationen/Ladungen in den Zellen wieder auf das volle Signal angehoben. Die notwendige Vorladezeit, vor deren Ablauf kein neuer Lesezyklus starten darf, wird mit *Row Precharge time* t_{RP} bezeichnet. Die Mindestzeit bevor wieder eine Wortleitung der gleichen Bank aktiviert werden kann wird mit *Row Cycle time* t_{RC} bezeichnet und ergibt sich aus der Summe von t_{RAS} und t_{RP}. Die *Row Cycle time* t_{RC} entspricht der Lesezeit im ungünstigsten Fall.

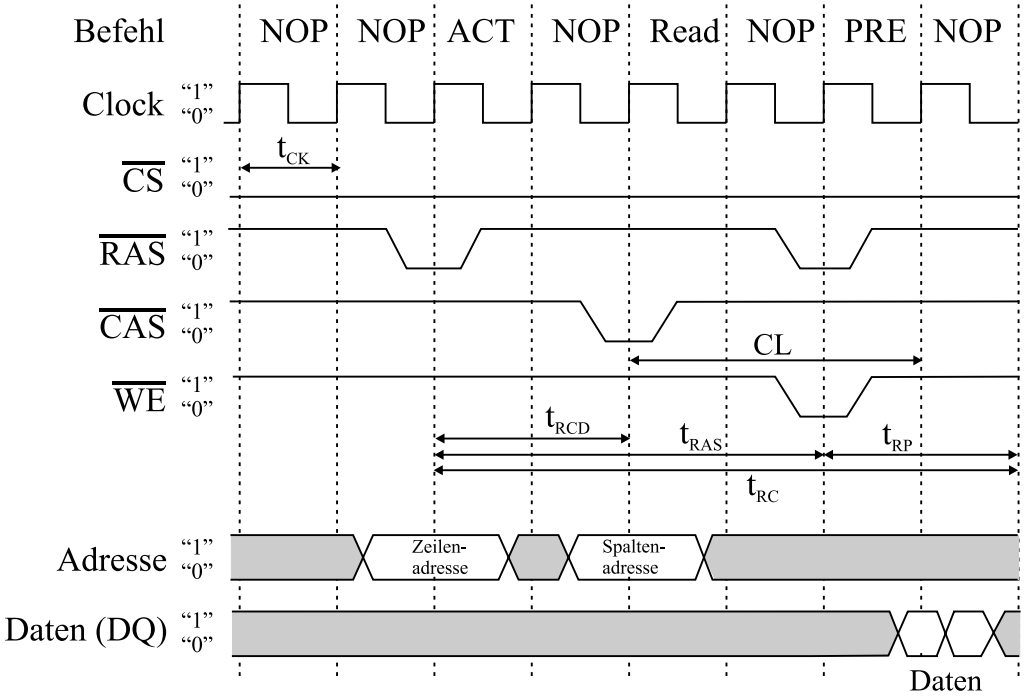


Abbildung 2.9: Zeitlicher Ablauf einer einfachen Leseoperation bei einem DDR Speicherchip.

2.3.4 Schreiboperation

Der Schreibvorgang unterscheidet sich nur wenig vom Lesevorgang. Lediglich der *Read*-Befehl wird durch ein *Write* ersetzt und es müssen etwas andere Timings beachtet werden. Vor dem Schreibzugriff müssen alle Bitleitungen auf $V_{BLH}/2$ vorgeladen sein. Der *Activate*-Befehl signalisiert das Anliegen der Zeilenadresse, öffnet die zugehörige Wortleitung und startet die Verstärkung der Speicherzeile. Jeder Schreibzugriff beginnt also mit dem Lesen der adressierten Zeile. Nach der Zeit t_{RCD} folgt der *Write*-Befehl für die nun anliegende Spaltenadresse. Der *Write*-Befehl unterscheidet sich durch den Übergang von $\overline{WE}$ auf low vom *Read*. Dem Chip wird dadurch signalisiert, dass nach der Zeit t_{DQSS} die zu schreibenden Daten an den Eingängen abgeholt werden können. Der Dateneingangspuffer liest das anliegende Datum und führt es über den Spaltendekoder der entsprechenden Spalte zu. Das zuvor durch *Activate* gelesene Signal der Zelle wird

einfach überschrieben. Da das Schreiben im Fall des Wechsels der enthaltenen Information ein Umladen der BLs benötigt, ist der Schreibzyklus der zeitlich längste Zyklus im DRAM. Erst nach der Zeitspanne t_{WR} darf der *Precharge*-Befehl auf das Einlesen der Daten von den Datenpins folgen. Dadurch wird die Zeile wieder geschlossen und das Vorladen der Bitleitungen gestartet, welches die Zeit t_{RP} benötigt. Erst nachdem das Vorladen beendet ist, kann erneut auf die Speichermatrix zugegriffen werden.

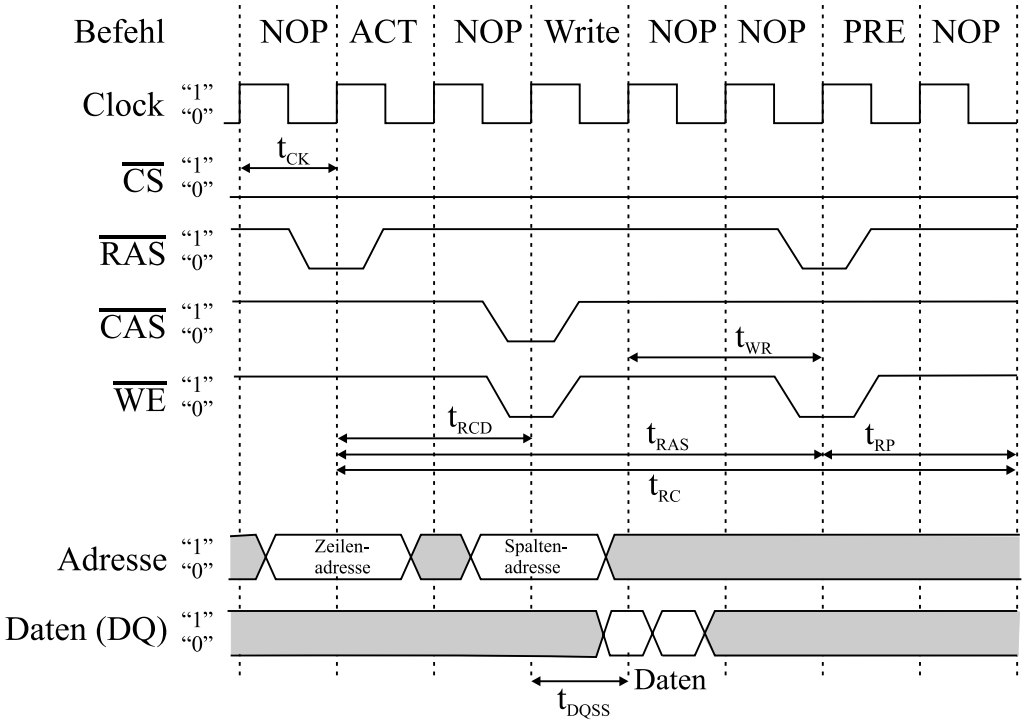


Abbildung 2.10: Zeitlicher Ablauf einer einfachen Schreiboperation bei einem DDR-Speicher.

2.3.5 Verstärkungsvorgang

Wie schon bei der Beschreibung des Zellaufbaus in Abschnitt 2.1 erwähnt, ist das Lesen und Schreiben aufgrund des großen Kapazitätsunterschiedes zwischen Bitleitung und Zellkondensator mit etwas mehr Aufwand als nur dem Öffnen des Gates verbunden. In diesem Abschnitt wird der Verstärkungsvorgang beim Lesen beschrieben. Bevor Zellen ausgelesen werden können, müssen die Bitleitungen auf das Potenzial $V_{BLH}/2$ ($= 0.75\,V$) vorgeladen sein. Durch den *Activate*-Befehl wird die zur Zeilenadresse gehörende Wortleitung durch Anlegen von V_{PP} geöffnet und damit die Verstärkung gestartet. Die Auswahltransistoren der Speicherzeile schalten durch. Dadurch werden Bitleitungskapazität und Zellkapazität C_S parallel geschaltet und die gespeicherten Ladungen verteilen sich neu. Im Falle einer zuvor gespeicherten „1" ($= 1.5\,V$ auf C_S) fließen Elektronen von der Bitleitung in den Zellkondensator und im Falle einer „0" ($= 0\,V$ auf C_S) vom Zellkondensator auf die Bitleitung. Das Potenzial der Bitleitung wird somit etwas höher falls eine

„1" und etwas niedriger falls eine „0" gespeichert war. Aufgrund des großen Kapazitätsunterschiedes ist diese Potenzialänderung der Bitleitung jedoch sehr gering ($< 150\,mV$) und muss noch auf den vollen Signalpegel verstärkt werden ($V_{BLL} = 0\,V$ für „0" bzw. $V_{BLH} = 1.5\,V$ für „1"). Die dazu nötige hochempfindliche Verstärkung wird mit Hilfe der so genannten *Sense-Amps* durch eine differenzielle Verstärkungstechnik realisiert. Dabei wird das Potenzial der Bitleitung BL, an der gelesen wird, mit einer Referenzbitleitung $\overline{BL}$, an der nicht gelesen wird und deshalb ihr Vorladepotenzial $V_{BLH}/2$ behält, verglichen. Die beiden Bitleitungen BL und $\overline{BL}$ sind dazu an die Eingänge des Differenzverstärkers (*Sense-Amp*) angeschlossen. Dieser besteht aus zwei kreuzgekoppelten CMOS-Invertern, deren Eingänge (Gates der Transistoren) jeweils vom Ausgang des anderen abgeleitet sind und dadurch ein bistabiles System bilden (siehe Abbildung 2.11a). Abbildung 2.11b zeigt den zeitlichen Spannungsverlauf beim Lesen einer „1" aus einer Zelle an BL qualitativ. Die Steuersignale SN und SP liegen zum Zeitpunkt, an dem die Wortleitung geöffnet wird, noch auf $V_{BLH}/2$, wodurch die Spannungsversorgung des Inverterpaares ausgeschaltet ist und alle vier Transistoren sperren. Die eigentliche Verstärkung wird erst durch den Übergang des Steuersignals SN von $V_{BLH}/2$ nach GND gestartet, der wenige Nanosekunden (*Signal Development Time*) nach dem Öffnen der Wortleitung stattfindet. Da das Potenzial von BL durch das Lesen einer „1" aus einer Zelle an BL zu Beginn der Verstärkung um das Zellsignal über dem Potenzial von $\overline{BL}$ liegt (ca. $150\,mV$) und die Gatespannung von T2 bestimmt, erreicht T2 beim Übergang des SN-Signals die Schwellenspannung zeitlich vor T1 und beginnt $\overline{BL}$ zu entladen. Da T1 die Gatespannung wiederum von $\overline{BL}$ ableitet, kann dieser nicht einschalten und BL behält ihr Potenzial bei. Durch diese Rückkopplung öffnet T2 während T1 geschlossen bleibt. $\overline{BL}$ wird dadurch nach und nach bis auf V_{BLL} entladen. Durch den anschließenden Übergang der Signalleitung SP von $V_{BLH}/2$ nach V_{BLH} werden auch die pFETs des *Sense-Amps* aktiviert. Das Potenzial von $\overline{BL}$ ist zu dem Zeitpunkt bereits deutlich niedriger als das von BL. Da $\overline{BL}$ die Gatespannung von T3 liefert, schaltet dieser durch und beginnt BL in Richtung V_{BLH} aufzuladen. Durch die Rückkopplung bleibt T4 weiterhin geschlossen. Am Ende des Prozesses wird ein stabiler Zustand mit V_{BLH} auf BL und V_{BLL} auf $\overline{BL}$ erreicht. Da während des gesamten Verstärkungsvorgangs die Wortleitung aktiviert blieb, wurde auch der Kondensator der gelesenen Zelle wieder vollständig auf V_{BLH} aufgeladen. Die durch das Auslesen zunächst verloren gegangene Kondensatorladung wurde wieder vollständig erneuert. Dieser Vorgang wird mit *Refresh* bezeichnet.

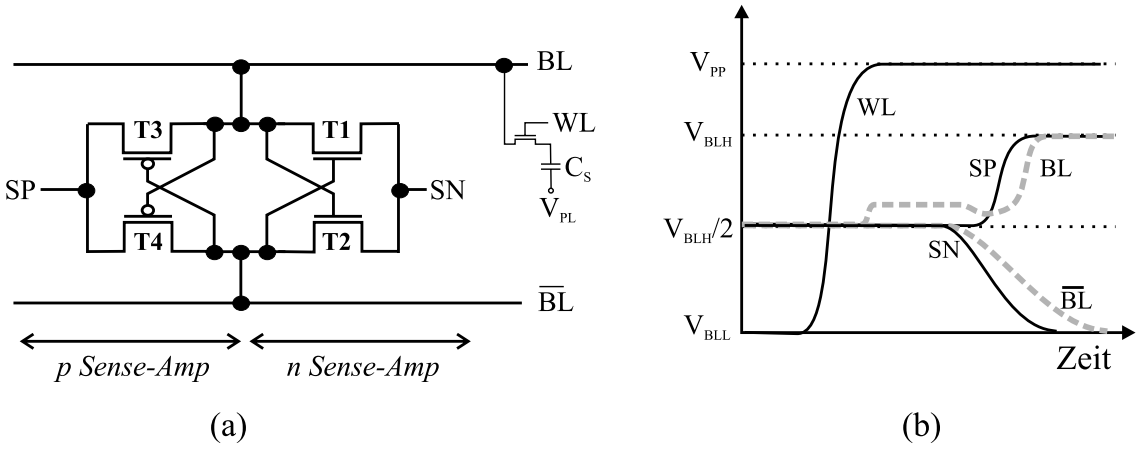


Abbildung 2.11: (a) Schaltskizze eines Leseverstärkers (*Sense-Amps*), (b) Spannungsverlauf beim Lesen einer „1" aus einer Zelle an Bitleitung BL.

2.3.6 Refresh

Da in jeder mikroelektronischen Schaltung Leckströme existieren, geht die Ladung auf dem Kondensator und mit ihr die gespeicherte Information mit der Zeit verloren. Mögliche Leckstrompfade und -mechanismen im DRAM werden in dieser Arbeit untersucht und noch ausgiebig diskutiert werden. Die gespeicherte Information muss also in bestimmten Zeitintervallen dynamisch aufgefrischt werden, um nicht verloren zu gehen. Daher kommt das „Dynamic" in der Abkürzung DRAM (***Dynamic*** *Random Access Memory).* Für standard DRAM Anwendungen verlangt die JEDEC Spezifikation einen so genannten *Refresh* aller Zellen innerhalb von $64\,ms$. Das bedeutet, dass jede Zelle die gespeicherte Informationen unter allen spezifizierten Betriebsbedingungen mindestens für diese Zeit halten muss. Wie schon in den vorherigen Abschnitten angedeutet, muss zum Refresh einer Speicherzeile diese lediglich aktiviert (Activate-Befehl) werden. Nach dem anschließenden erneuten Vorladen (Precharge-Befehl) steht der Speicher wieder für andere Zugriffe zur Verfügung. Während der $64\,ms$ Refresh-Periode müssen also alle WLs der Speichermatrix mindestens einmal geöffnet werden. Bei der traditionellen DRAM Konfiguration mit 4096 WLs erfordert dies durchschnittlich alle $15.6\,\mu s$ einen Refreshzyklus. Bei einer typischen Refreshdauer von ca. $50\,ns$ steht der Speicher somit für $4096 * 50\,ns = 205\,\mu s$ der $64\,ms$ Refresh-Periode (bzw. 0.32 % der Zeit) nicht anderweitig zur Verfügung. Bei heutigen Chips, die aufgrund der hohen Kapazität meist $32\,k$ WLs besitzen, werden deshalb bis zu acht Wortleitungen gleichzeitig geöffnet (zwei pro Bank und vier Bänke pro Chip), um die gleiche Verfügbarkeit zu gewährleisten.

2.4 Mode-Register und Testmodes

Das *Mode-Register* ist ein kleiner Speicher in dem aktuelle Betriebseinstellungen wie z.B. Burstlänge, Bursttyp und CAS-Latency des DRAM Chips abgelegt werden. Es muss vor Inbetriebnahme eines Speicherchips geschrieben werden und behält seine Information bei, bis es erneut geschrieben bzw. die Stromversorgung abgeschaltet wird. Viele der möglichen Einstellungen im *Mode-Register* sind industrieweit spezifiziert. Darüber hinaus können über das *Mode-Register* Testmodes eingeschaltet werden, die nicht genormt sind und von den Herstellern geheim gehalten werden. Angesteuert werden sie durch eine Art Zahlenschloss über eine Reihe von so genannter *Testmode-Entry Codes*. Die Testmodes dienen vor allem erweiterten Testmöglichkeiten, wie z.B. der Programmierung von Onchip-Spannungsgeneratoren. Außerdem gibt es Adress- und Datenkompressions-Modes, die höhere Parallelität beim Test und dadurch Kosteneinsparungen erlauben. Eine weitere Art von Testmode ist der so genannte *Burn-In*, in dem Stressbedingungen direkt auf dem Chip erzeugt werden können, um diesen künstlich zu altern und potenzielle Frühausfälle beim Kunden zu vermeiden. Vor allem von der Programmierung der internen Betriebsspannungen über Testmodes wurde in dieser Arbeit ausgiebig Gebrauch gemacht.

2.5 Redundanz

Auf einem einzigen DRAM Chip werden derzeit bis zu 10^9 Speicherzellen realisiert. Es ist heutzutage nahezu unmöglich Speicherchips zu fertigen, die von vornherein absolut fehlerfrei sind. Schon winzigste Partikel können Bereiche mit mehreren Zellen unbrauchbar machen. Die Chipausbeute (*Yield*) bestimmt jedoch den Erfolg und das Bestehen im schwer umkämpften Speichermarkt. Um eine gewisse Anzahl Fehler tolerieren zu können, werden Speicherchips heutzutage mit einer gewissen Anzahl redundanter Speicherelemente gefertigt. Noch auf dem Wafer werden alle Chips kontaktiert und getestet. Fehlerhafte Zellen werden identifiziert und die betroffenen Wort- und Bitleitungen bei dem anschließenden *Fusing* abgeschaltet und durch redundante Leitungen ersetzt. Dazu werden Sicherungen (*Fuses*) in so genannten Fusebänken gezielt mit einem Laser zerschossen. Durch diesen Austauschmechanismus kann die Ausbeute erheblich gesteigert werden.

Kapitel 3

DRAM Data Retention

3.1 Definition und typische Verteilung der Retentionzeiten

Im Kapitel 2 wurde die Funktion und der Aufbau eines DRAM Speicherbausteins beschrieben. Es wurde bereits darauf hingewiesen, dass die gespeicherte Information aufgrund von Leckströmen verloren geht, wenn nicht eine periodische Auffrischung erfolgt. Die maximale Zeit, nach der die gespeicherte Information ohne Auffrischung noch korrekt ausgelesen werden kann, definiert die *DRAM Data Retention* (t_{Ret}). Nach der JEDEC Spezifikation muss für Standard-DRAM eine Retentionzeit von $64\,ms$ bei $70\,°C$ garantiert werden. Bei Speichern für mobile Anwendungen liegt die Anforderung sogar bei bis zu $160\,ms$. Da keine einzige Zelle des gesamten Speicherchips die Information in dieser Zeit verlieren darf, wird die Retentionzeit eines Speicherbausteins durch die schlechteste Zelle definiert. Wie alle Parameter bei Produktionsprozessen mit großer Stückzahl, unterliegt auch die Retentionzeit der vielen Speicherzellen eines Chips einer statistischen Streuung. Da die Retentionzeit der schlechtesten Zelle eines Speicherchips ausschlaggebend ist, spielt die Breite der Verteilung eine wesentliche Rolle. Abbildung 3.1 zeigt die so genannte *Retentionverteilung* oder *Retentionkurve* eines $256\,MBit$ DDR Speicherbausteins. In der *Retentionkurve* wird die kumulative Wahrscheinlichkeit in Quantilen der Standardnormalverteilung gegenüber der Retentionzeit aufgetragen. Diese Art der Auftragung wird in der Statistik auch mit *Wahrscheinlichkeitsplot* bezeichnet. Durch die Umrechnung der kumulativen Wahrscheinlichkeit in Quantile der Normalverteilung und lineare Auftragung derselben, wird die Skala der y-Achse derart verzerrt, dass eine normalverteilte Größe als Gerade im Diagramm erscheint. Ist wie in Abbildung 3.1 die x-Achse logarithmisch skaliert, d.h. die aufgetragene Größe entsprechend transformiert, entspricht eine Gerade im Diagramm einer Lognormal-Verteilung. Die Retentionkurve des Speicherbausteins in Abbildung 3.1 zeigt zwei gerade Teilabschnitte. Demzufolge kann sie nicht durch eine einfache Lognormal-Verteilung, sondern muss als Mischverteilung aus zwei Lognormal-Verteilungen beschrieben werden. Die Existenz zweier Unterverteilungen ist

aus der Literatur bekannt (z.B. [Ham98]). Der gerade Teilabschnitt mit niedriger Retentionzeit wird *Retention-Tail* genannt, der mit höherer Retentionzeit *Retention-Main*. Die allermeisten Zellen (> 99.999%) liegen in der Mainverteilung und die durchschnittliche Retentionzeit liegt im zweistelligen Sekundenbereich. Die meisten Zellen übertreffen die Spezifikation somit um ein Vielfaches. Der Übergang zwischen Tail und Main findet bei einer kumulativen Wahrscheinlichkeit von ungefähr $-4\,\sigma$ statt. Das bedeutet, dass weniger als 10^{-5} aller Zellen in die Tailverteilung fallen. Lediglich das äußerste untere Ende der Tailverteilung ist ausbeuterelevant ($< -5\,\sigma$). Es sind also nur sehr wenige Zellen, die außergewöhnlich schlechte Retention haben. Dennoch begrenzen diese wenigen „Ausreißer" die Funktionalität des ganzen Chips. Die Zellen der Tailverteilung unterscheiden sich elektrisch somit wesentlich von den Zellen der Mainverteilung. Strukturell können selbst mit Raster- und Tunnelelektronenmikroskopie (REM und TEM) jedoch keine Unterschiede zwischen Tail- und Mainzellen festgestellt werden, sodass die Ursache für die Tailverteilung bisher nicht eindeutig geklärt ist. Es ist von höchstem wirtschaftlichen Interesse den physikalischen Grund für die Tailverteilung herauszufinden. Diese Arbeit soll dazu einen Beitrag liefern.

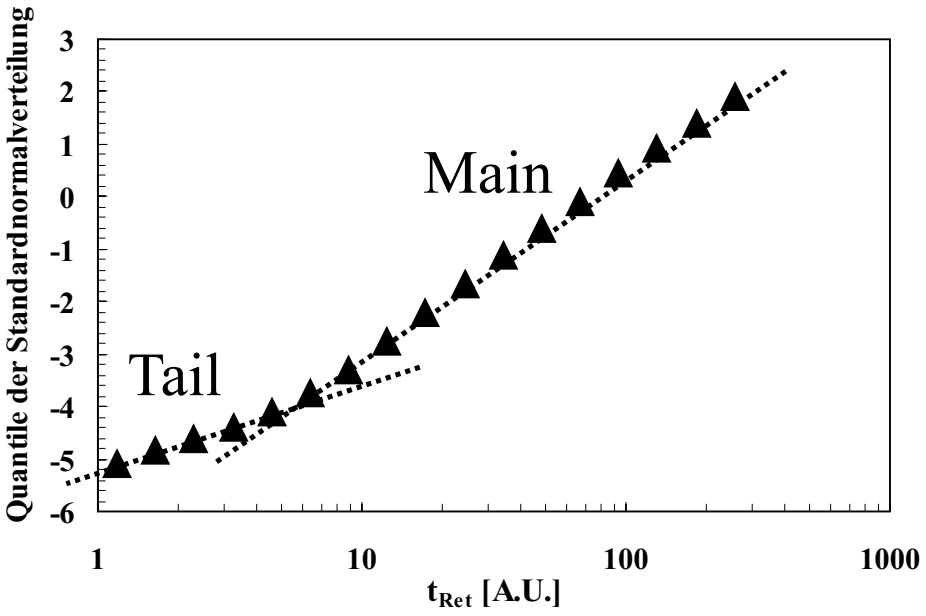


Abbildung 3.1: Retentionkurve eines $256\,MBit$ DDR Speicherchips bei $85\,°C$.

3.2 Retention-Formel

In diesem Abschnitt soll geklärt werden, wodurch die Retentionzeit einer Speicherzelle bestimmt wird. Dazu wird die so genannte *Retention-Formel* [Hum03], die einen Zusammenhang zwischen der Retentionzeit t_{Ret} und den Zell- sowie Sensingparametern darstellt, anschaulich hergeleitet und motiviert. Bei der Herleitung muss der Auslese- und Verstärkungsvorgang nochmals genauer betrachtet werden. Abbildung 3.2 zeigt dazu den Zustand vor und nach dem Öffnen der Wortleitung für einen idealisierten Zelltransistor. Nach dem Vorladen von Bitleitung und Leseverstärkerschaltkreis und vor dem Aktivieren der Wortleitung ist die Gesamtladung in dem durch den Zelltransistor isolierten System gegeben durch

$$Q_{vorher} = C^*_{BL}\frac{V_{BLH}}{2} + C_S V_{BLH} P_w - Q_{Leak}(t) \tag{3.1}$$

Der erste Term beschreibt die Ladung der auf $V_{BLH}/2$ vorgeladenen Bitleitung (C^*_{BL} beinhaltet dabei die Kapazität C_{SA} des Leseverstärkers). Im zweiten Term steht die ursprünglich in die Zelle geschriebene Ladung. Dabei wird durch den Faktor P_w mit $0 \leq P_w \leq 1$ berücksichtigt, dass die Zellkapazität aufgrund der exponentiellen Ladekurve eines Kondensators in endlicher Schreibzeit nicht auf die volle Spannung V_{BLH} aufgeladen werden kann. Schließlich entspricht $Q_{Leak}(t)$ der Ladung, die seit dem Schreiben der Zelle durch Leckströme verloren gegangen ist. Diese ist gegeben durch

$$Q_{Leak}(t) = \int_0^t I_{Leak}(t')dt' \tag{3.2}$$

wobei t die Zeit seit dem Schließen der Wortleitung darstellt. Auf die Leckstrompfade selbst soll in Kapitel 5 genauer eingegangen werden. Das Potenzial V_S der inneren Kondensatorelektrode fällt dadurch mit der Zeit ab.

Wird die Wortleitung zum Auslesen der Zelle aktiviert, schaltet der Zelltransistor durch und verbindet die Bitleitung mit dem Zellkondensator. Alle Kapazitäten sind nun parallel geschaltet und die gespeicherte Ladung verteilt sich neu (siehe Abbildung 3.2b). Über der Bitleitung fällt nun die Spannung V_{BL} ab und für die Gesamtladung gilt

$$Q_{nachher} = (C^*_{BL} + C_S) \cdot V_{BL} \tag{3.3}$$

Das Öffnen der Wortleitung dauert nur wenige Nanosekunden und in erster Näherung geht in dieser Zeit keine Ladung durch Leckströme verloren. Die Gesamtladungen vor und nach dem Öffnen sind deshalb gleich

$$\begin{array}{rcl} Q_{vorher} & = & Q_{nachher} \\ C^*_{BL}\frac{V_{BLH}}{2} + C_S V_{BLH} P_w - Q_{Leak}(t) & = & (C^*_{BL} + C_S) \cdot V_{BL} \end{array} \tag{3.4}$$

Auflösen nach V_{BL} liefert das Potenzial der gelesenen Bitleitung, wenn zur Zeit t nach dem letzten Schreiben gelesen wird.

$$V_{BL} = \frac{1}{C^*_{BL} + C_S}\left[C^*_{BL}\frac{V_{BLH}}{2} + C_S V_{BLH} P_w - Q_{Leak}(t)\right] \tag{3.5}$$

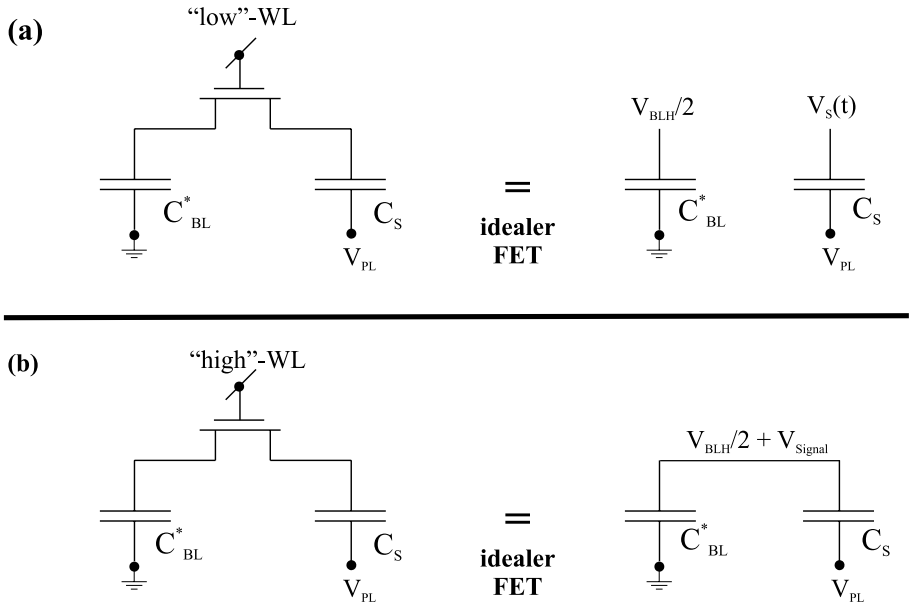


Abbildung 3.2: Auslesen einer Speicherzelle. (a) Vor dem Öffnen der Wortleitung ist die Bitleitungskapazität isoliert von der Zellkapazität. Aufgrund von Leckströmen verliert der Zellkondensator mit der Zeit Ladung und die Kondensatorspannung V_S fällt ab. (b) Nach dem Öffnen der Wortleitung verteilen sich die Ladungen aus den nun parallel geschalteten Kapazitäten neu und die Spannung der Bitleitung ändert sich um V_{Signal}. Je nach Ladungszustand des Zellkondensators zum Zeitpunkt der WL-Aktivierung ist die resultierende Spannungsänderung positiv oder negativ.

Wie in Kapitel 2 beschrieben, wird die Spannung an der Bitleitung V_{BL} relativ zur Spannung der komplementären Bitleitung $V_{\overline{BL}} = V_{BLH}/2$ bewertet. Zwischen dem Bitleitungspaar liegt zum Startzeitpunkt der Verstärkung im idealen Fall die Differenzspannung ΔV_{ideal} an:

$$\Delta V_{ideal} = V_{BL} - V_{\overline{BL}} = \frac{1}{C^*_{BL} + C_S}\left[C^*_{BL}\frac{V_{BLH}}{2} + C_S V_{BLH} P_w - Q_{Leak}(t)\right] - \frac{V_{BLH}}{2} \tag{3.6}$$

In der Realität müssen jedoch noch weitere Effekte berücksichtigt werden, die sich auf die Spannungsdifferenz auswirken. Erstens kann wie beim Schreiben der Zelle auch beim Lesen der Ladungsaustausch in endlicher Zeit nicht vollständig abgeschlossen werden und zweitens reagiert die komplementäre Bitleitung durch kapazitive Kopplung auf die Spannungsänderung der gelesenen Bitleitung. Die endliche Lesezeit wird durch einen Faktor P_r berücksichtigt, der je nach Entladezeit einen Wert aus dem Wertebe-

reich $0 \leq P_r \leq 1$ annehmen kann. Der kapazitiven Kopplung zwischen den Bitleitungen wird durch den Faktor P_n mit selbigem Wertebereich Rechnung getragen. Eine Abschätzung für P_n wird im nächsten Abschnitt gegeben. Beide Effekte bewirken eine Reduktion der maximalen Spannungsdifferenz ΔV zwischen den Eingängen des Sense-Amps zum Startzeitpunkt der Verstärkung und es gilt:

$$\begin{aligned} \Delta V &= P_r \cdot P_n \cdot \Delta V_{ideal} \\ &= P_r P_n \left(\frac{1}{C^*_{BL}+C_S} \left[C^*_{BL} \frac{V_{BLH}}{2} + C_S V_{BLH} P_w - Q_{Leak}(t) \right] - \frac{V_{BLH}}{2} \right) \end{aligned} \tag{3.7}$$

Um eine gespeicherte „1" noch richtig auslesen zu können, ist eine vom jeweiligen Differenzverstärker minimale Differenzspannung V_{SA} notwendig. Für $\Delta V = V_{SA}$ ist dadurch die maximale Verlustladung $Q_{Leak,max} = Q_{Leak}(t_{Ret})$ und damit auch die Retentionzeit $t = t_{Ret}$ definiert. Ist der zeitliche Leckstromverlauf bekannt, kann die Retentionzeit t_{Ret} aus

$$Q_{Leak,\,max} = Q_{Leak}(t_{Ret}) = \int_0^{t_{Ret}} I_{Leak}(t')dt' \tag{3.8}$$

berechnet werden. In der Praxis ist dies aufgrund unterschiedlichster Leckstrompfade und Leckstrommechanismen mit verschiedensten Spannungsabhängigkeiten (mehr dazu in Kapitel 5) jedoch nicht möglich. Deshalb benutzt man stattdessen den zeitlich gemittelten Leckstrom $\overline{I_{Leak}}$. Es gilt dann

$$Q_{Leak,\,max} = t_{Ret}\overline{I_{Leak}} \tag{3.9}$$

Einsetzen der Gleichung 3.9 und $\Delta V = V_{SA}$ in Gleichung 3.7 führt schließlich zur so genannten Retention-Formel

$$t_{Ret} = \frac{C_S}{\overline{I_{Leak}}} \left[(V_{BLH} \cdot P_w - \frac{V_{BLH}}{2}) - \frac{V_{SA}}{P_r \cdot P_n} (\frac{C_S + C^*_{BL}}{C_S}) \right] \tag{3.10}$$

BL-BL Kopplung

In Gleichung 3.7 wurde der Faktor P_n eingeführt, welcher die kapazitive Kopplung zwischen benachbarten Bitleitungen näherungsweise berücksichtigen soll. Beim Auslesen einer Zelle ändert sich aufgrund dieser das Potenzial nicht nur auf der gelesenen, sondern auch auf der komplementären Bitleitung. Wird in der Nachbarschaft der komplementären Bitleitung eine physikalische „1" gelesen, wandert das Potenzial dieser ein wenig in Richtung „1"; wird dagegen eine „0" gelesen, so wandert das Potenzial etwas in Richtung „0" mit. Die Differenzspannung ΔV zwischen Bitleitung und komplementärer Bitleitung wird verringert. P_n hängt also von der gelesenen Datentopologie ab. Da die Retentionkurve in Abbildung 3.1 durch Lesen bzw. Schreiben einer „1" Daten-Topologie entsteht,

soll P_n speziell für diesen Fall betrachtet werden. Beim Lesen einer Zeile ergibt sich dann die in Abbildung 3.3 dargestellte Situation.

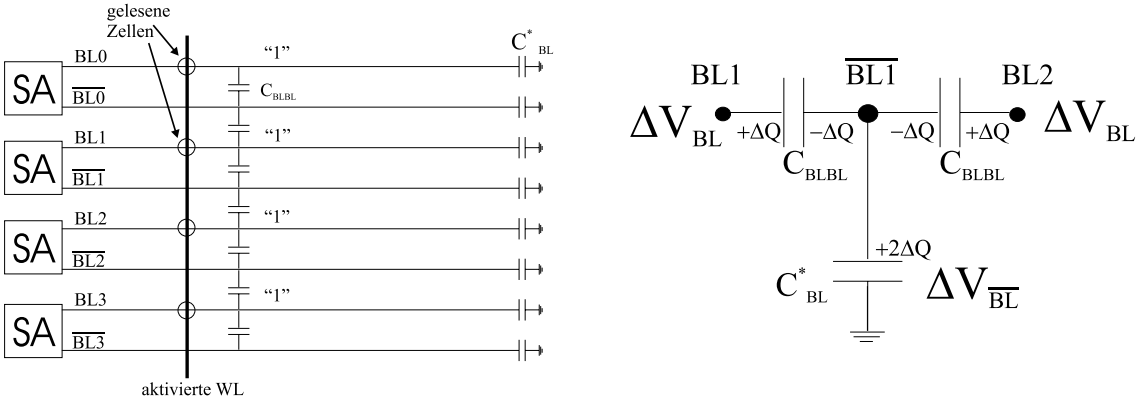


Abbildung 3.3: Lesen einer Zeile aus einem Speicherfeld, dessen Zellen zuvor alle mit einer physikalischen „1" geschrieben wurden.

Vor dem Öffnen der Wortleitung liegen alle Bitleitungen auf dem Vorladepotenzial $V_{BLH}/2$ und werden floatend geschaltet. Das Aktivieren einer Wortleitung öffnet aufgrund des *Folded-Bitline*-Konzepts nur Zellen der Bitleitungen BLx, während keine Zellen an den komplementären Bitleitungen $\overline{BLx}$ geöffnet werden. An den geöffneten Bitleitungen BLx findet der Ladungsausgleich zwischen Zellkondensator und Bitleitung statt und die Spannung V_{BL} dieser steigen ausgehend von $V_{BLH/2}$ um ΔV_{BL}, während die komplementären Bitleitungen $\overline{BL}$ nur kapazitiv darauf reagieren können. Die mit der Spannungsänderung verbundene Ladungsänderung $+\Delta Q$ auf den BLs erzeugt auf der anderen Elektrode der Kopplungskapazität C_{BLBL} die Ladung $-\Delta Q$. Da die $\overline{BL}$s floatend sind, muss diese Ladungsänderung durch $+\Delta Q$ auf der Bitleitungskapazität C^*_{BL} ausgeglichen werden. Die Folge ist eine entsprechende Spannungsänderung der Nachbar-Bitleitungen im Verhältnis der Kapazitäten (kapazitiver Spannungsteiler).

$$\Delta V_{\overline{BL}} = \frac{2\Delta Q}{C^*_{BL}} = \frac{2C_{BLBL}}{C^*_{BL}} \cdot \Delta V_{BL} \tag{3.11}$$

Die letztendliche Spannungsdifferenz ΔV zwischen den zwei Bitleitungen des Differenzverstärkers direkt vor dem Start der Verstärkung ist dementsprechend kleiner und gegeben durch

$$\Delta V = \Delta V_{BL} - \Delta V_{\overline{BL}} = \Delta V_{BL} \cdot (1 - \frac{2C_{BLBL}}{C^*_{BL}}) = \Delta V_{ideal} \cdot \underbrace{(1 - \frac{2C_{BLBL}}{C^*_{BL}})}_{P_n} \tag{3.12}$$

Die ohne Kopplung erzielte Spannungsdifferenz ΔV_{ideal} wird durch die Kopplung um den Faktor P_n reduziert.

Zur Verringerung des Signalverlustes durch kapazitive Kopplung werden im modernen

DRAM so genannte *Bitleitungs-Twists* angewendet (siehe Abbildung 3.4), d.h. die Bitleitungen werden in speziellen Mustern gekreuzt, damit Störungen auf beide Bitleitungen eines Verstärkers gleichermaßen wirken können. In den DRAM Bausteinen dieser Arbeit wurde ein Bitleitungs-Twist gemäß Abbildung 3.4b verwendet. Durch das Überkreuzen der Leitungen kann das Übersprechen um 50% reduziert werden [Min99]. P_n wird dadurch zu:

$$P_n = (1 - \frac{C_{BLBL}}{C^*_{BL}}) \tag{3.13}$$

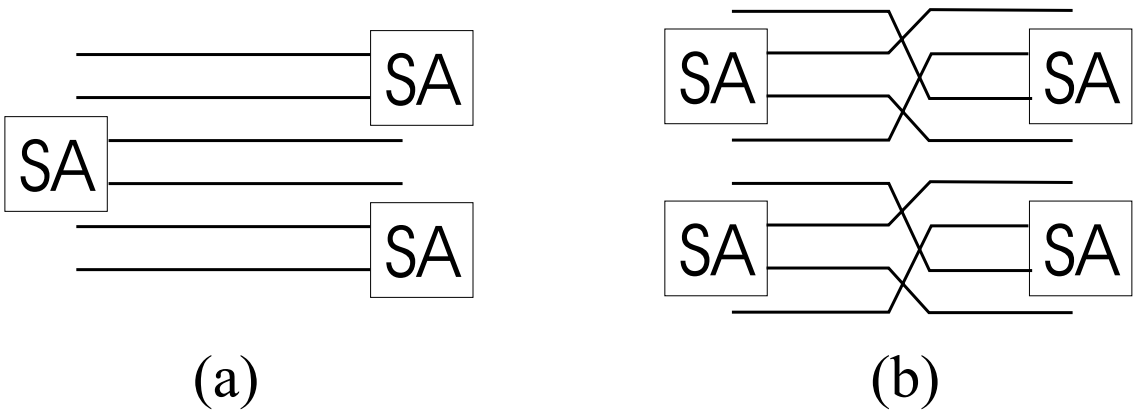


Abbildung 3.4: Bitleitungs-Layout zur Reduktion der BL-BL-Kopplung. (a) einfaches Layout (b) Bitleitungs-Twist der in dieser Arbeit untersuchten Speicherbausteine. Für weitere Twist-Möglichkeiten siehe z.B. [Min99, Kee01].

3.3 Ursachen für kurze Haltezeiten

Betrachtet man die Retention-Formel, so ergeben sich verschiedene Möglichkeiten, die zu sehr kurzen Haltezeiten führen. Die Retention-Formel stellt ein Produkt aus zwei Termen dar:

$$t_{Ret} = \underbrace{\frac{C_S}{\overline{I_{Leak}}}}_{Term\,1} \underbrace{\left[(P_w \cdot V_{BLH} - \frac{V_{BLH}}{2}) - \frac{V_{SA}}{P_r \cdot P_n} \overbrace{(\frac{C_S + C^*_{BL}}{C_S})}^{Transfer-Ratio} \right]}_{Term\,2} \tag{3.14}$$

Wird einer der Terme sehr klein, verkürzt sich die Retentionzeit massiv. Man unterscheidet die folgenden drei Fälle:

Leakage

Leckströme verringern die Ladung im Zellkondensator. Ein großer Leckstrom $\overline{I_{Leak}}$ geht in den Nenner von Term 1 ein und verringert dadurch die Retentionzeit. Eine größe-

re Kondensatorkapazität C_S hält mehr Ladungen bereit, die durch Leckströme verloren gehen dürfen, und die Entladezeit/Retentionzeit steigt.

Signal Margin

Das Produkt aus Offset des Differenzverstärkers V_{SA} und *Transfer-Ratio* bestimmt den zweiten Teil des zweiten Terms. Durch besonders großes V_{SA} bzw. *Transfer-Ratio* kann der zweite Term klein werden und fast unabhängig vom Gesamtleckstrom der Zelle zu sehr kurzen Retentionzeiten führen. Liegt V_{SA} über dem Zellsignal ohne Ladungsverlust ($t_{Ret} = 0$), d.h. benötigt der Leseverstärker mehr Spannungshub als die Zelle selbst bei sofortigem Wiederauslesen auf der Bitleitung erzeugt, kann die Information zu keinem Zeitpunkt korrekt gelesen werden. Die Retention-Formel liefert in diesem Fall physikalisch unsinnige negative Retentionzeiten, die als *Signal Margin*-Fehler gedeutet werden müssen.

Bitline Coupling

Eine weitere Möglichkeit für kurze Haltezeiten entsteht dadurch, dass der Term 2 aufgrund kleiner Schreib- und Lesefaktoren $P_{w,r}$ klein wird. Die Ursache liegt in zu hohen Serienwiderständen von Bitleitungskontakt, Zelltransistor und BS-Anschluss. Sind die Schreib/Lese-Zyklen zu kurz gewählt, werden die Zellkondensatoren nicht vollständig aufgeladen bzw. ausgelesen. Dieser Effekt ist vor allem für niedere Temperaturen von Bedeutung. Die in dieser Arbeit untersuchte Problematik besteht in besonders kurzen Retentionzeiten bei Temperaturen von ungefähr $85°C$ und ist deshalb grundsätzlich von *Bitline Coupling* zu unterscheiden. Um *Bitline Coupling* in den Untersuchungen zu minimieren, wurden bei allen Messungen entspannte Timings angewandt, sodass die Zellkondensatoren immer voll aufgeladen werden können und *Bitline Coupling* keine Rolle spielt.

Kapitel 4

Ergebnisse zur Monte Carlo Analyse der Retentionverteilung

Die in Abschnitt 3.2 hergeleitete Retention-Formel gibt den Zusammenhang zwischen den Designgrößen eines Speicherchips und der Retentionzeit einer einzelnen Speicherzelle an. Durch einfaches Einsetzen der Design-Werte folgt daraus zunächst eine für alle Zellen eines Speicherchips identische Retentionzeit. Das steht jedoch im Widerspruch zur gemessenen Retentionkurve, die viele Größenordnungen überspannt (siehe Abschnitt 3.1). Für jede Speicherzelle müssen deshalb unterschiedliche Parameterwerte angenommen werden, d.h. auch alle Parameter unterliegen selbst gewissen Verteilungen. Bedingt durch die große Anzahl von Speicherzellen (z.B. $512\,Mbit$) müssen Werte berücksichtigt werden, die mit einer Wahrscheinlichkeit von nur $1/512\,M \sim 2 \cdot 10^{-9}$ auftreten, d.h. am Rande eines $6\,\sigma$ Bereiches liegen. Zum Verständnis der Retentionverteilung muss zunächst untersucht werden, wie sich die einzelnen Parameter und deren Verteilungen auf diese auswirken. Einfache „worst-case" Betrachtungen zur Berechnung der kürzesten Retentionzeit, also das Einsetzen der $6\,\sigma$-Werte für alle Parameter, sind sicherlich unrealistisch, da die kombinierte Auftrittswahrscheinlichkeit einer unabhängigen Parameterkombination sich zu $\sim (10^{-9})^{Parameteranzahl}$ berechnet und deshalb auf einem Speicherchip nicht beobachtbar ist. Mit Hilfe der Monte Carlo Methode kann dieses Problem umgangen werden und realistischere Aussagen sind möglich. Dazu werden für jeden Parameter der Retention-Formel Zufallszahlen gemäß dessen Verteilung generiert und in die Retention-Formel eingesetzt. Durch 10^9-maliges Wiederholen kann auf diese Weise die Verteilung innerhalb eines Speicherchips einfach „ausgewürfelt" werden.

Im ersten Abschnitt dieses Kapitels wird das zur Simulation verwendete Modell vorgestellt. In Abschnitt 4.2 werden dann möglichst realistische Verteilungen für die Parameter der Retention-Formel bestimmt. Teilweise wurden diese aus Messungen gewonnen, teilweise müssen sie geschätzt werden. Unbekannt ist die Verteilung der Gesamtleckströme, da Leckströme an Einzelzellen nicht direkt messbar sind. Die Retentionzeit kann jedoch

sehr genau bestimmt werden. Deshalb wird in Abschnitt 4.3 in einem ersten Simulationsschritt die Leckstromverteilung aus der Retentionverteilung unter Annahme realistischer Verteilungen für alle anderen Größen bestimmt. Die auf diese Weise erhaltene Leckstromverteilung stellt das wichtigste Ergebnis dieses Kapitels dar. Sie wird in Abschnitt 4.4 verwendet, um in weiteren Simulationen die Auswirkung von Parameteränderungen (z.B. C_S, C_{BL}) auf die Retentionverteilung zu untersuchen. Aus den rücksimulierten Retentionverteilungen kann mit Hilfe der Kenntnis über verfügbare Redundanz der Einfluss dieser Parameter auf die Chipausbeute abgeschätzt werden.

4.1 Das Simulationsmodell

Die Simulation basiert auf der im vorherigen Kapitel abgeleiteten Retention-Formel (Gleichung 4.1). Die Gesamtkapazität der Bitleitung C^*_{BL} wurde darin in die beinhalteten Einzelkapazitäten C_{BL}, C_{BLBL} und C_{SA} aufgeschlüsselt. Um Aufwand und Rechenzeit zu sparen, werden für die Spannung V_{BLH}, die Kapazität des Differenzverstärkers C_{SA} sowie die Schreib-Lesefaktoren P_w und P_r konstante Werte angenommen. Dies ist zulässig, da die eingesetzten gemessenen Retentionkurven mit entspannten Schreib- und Lesezeiten aufgenommen wurden und dadurch Streuungen in diesen Parametern weniger ins Gewicht fallen. Weitere Annahmen sind die statistische Unabhängigkeit aller Parameter sowie die isolierte Betrachtung aller Speicherzellen. Letzteres bedeutet, dass die Zugehörigkeit von mehreren Speicherzellen zu ein und derselben Bitleitung vernachlässigt und in der Simulation für jede Zelle ein komplett unabhängiger Parametersatz angenommen wird. Diese Vorgehensweise verringert den Rechenaufwand erheblich und stellt für grundlegende qualitative Untersuchungen keine Einschränkung dar.

$$\overline{I_{Leak}} = \frac{C_S}{t_{Ret}} \left[(V_{BLH} \cdot P_w - \frac{V_{BLH}}{2}) - \frac{V_{SA}}{P_r \cdot P_n} (\frac{C_S + \overbrace{C_{BL} + 2C_{BLBL} + C_{SA}}^{C^*_{BL}}}{C_S}) \right] \tag{4.1}$$

Für den Bitleitungskopplungsterm wird ein einfacher Twist angenommen:

$$P_n = 1 - \frac{C_{BLBL}}{C_{BL} + 2C_{BLBL} + C_{SA}} \tag{4.2}$$

4.2 Parameterverteilungen

In Tabelle 4.1 sind die in die Leckstromsimulation eingehenden Parameterverteilungen aufgelistet. Hier werden repräsentative Werte eines $512\,Mbit$-Speicherchips herangezogen. Es muss jedoch darauf hingewiesen werden, dass die Parameter von der Architektur und dem Design des Speicherchips abhängen und deshalb für andere Speicherprodukte davon abweichen können. Außerdem besitzen die meisten Parameter eine Abhängigkeit von der Position des Chips auf dem Wafer, sodass auch dadurch Abweichungen entstehen. Die Verteilung der Kapazitäten C_S, C_{BL} und C_{BLBL} einzelner Speicherzellen eines Speicherchips können nicht direkt gemessen werden. Es wird angenommen, dass die Kapazitäten einer Normalverteilung mit Wahrscheinlichkeitsdichte

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2) \tag{4.3}$$

folgen. Die Mittelwerte μ können durch Messungen an hochgradig parallelen Teststrukturen bestimmt werden (siehe Kapitel 6). Die Standardabweichungen σ werden für die Kapazität des Speicherkondensators auf 10% und die Bitleitungskapazitäten auf 5% geschätzt. Die Verteilung der Offsets V_{SA} des Differenzverstärkers ist hauptsächlich durch die V_t-Streuung der zwei n-FETs im Differenzverstärker bestimmt. Sie kann durch so genannte „Signal-Margin" Messungen bestimmt werden. Es ergibt sich eine Normalverteilung mit einem typischen Mittelwert von $\mu = -10\,mV$ und einer Standardabweichung von $\sigma = 10\,mV$. Dabei ist der Mittelwert durch die so genannte „Capacitive Imbalance" zu leicht negativen Spannungen hin verschoben. Die Ursache dafür liegt in der beim Auslesen im Vergleich zur komplementären Bitleitung um die Speicherkapazität der gelesenen Zelle vergrößerten Kapazität der gelesenen Bitleitung. Der Differenzverstärker wird dadurch in Richtung „1" vorgespannt, sodass diese bevorzugt wird und auch bei leicht negativem Spannungshub auf der gelesenen Bitleitung noch eine „1" detektiert werden kann.

Größe	Verteilungsfunktion	Parameter	Bemerkung
C_S	Normalverteilung	$\mu = 30\,fF$, $\sigma = 3\,fF$	μ gemessen, σ geschätzt
C_{BL}	Normalverteilung	$\mu = 56\,fF$, $\sigma = 2.8\,fF$	μ gemessen, σ geschätzt
C_{BLBL}	Normalverteilung	$\mu = 22\,fF$, $\sigma = 1.1\,fF$	μ gemessen, σ geschätzt
C_{SA}	konstant	$20\,fF$	geschätzt
V_{SA}	Normalverteilung	$\mu = -10\,mV$, $\sigma = 10\,mV$	μ, σ gemessen
V_{BLH}	konstant	$1.5\,V$	Designwert
P_r	konstant	0.95	Designwert
P_w	konstant	0.95	Designwert
t_{Ret}	Mischverteilung	nicht angegeben (A.U.)	Fit an gemessene Verteilung

Tabelle 4.1: In die Monte Carlo Simulation eingesetzten Parameterverteilungen.

Die gemessene Verteilung der Retentionzeiten setzt sich aus zwei geraden Teilabschnitten zusammen, wenn die Quantile der Standardnormalverteilung gegenüber dem Logarithmus der Retentionzeit aufgetragen werden (siehe Kreuze in Abbildung 4.1). Um für die Simulation Zufallszahlen gemäß der Retentionverteilung generieren zu können, muss diese zunächst parametrisiert werden. Dazu wird die Wahrscheinlichkeitsdichte p_{tRet} als Mischverteilung aus zwei Unterverteilungen angesetzt:

$$p_{tRet} = \alpha * f_1 [x \mid \mu_1, \sigma_1] + (1 - \alpha) * f_2 [x \mid \mu_2, \sigma_2] \tag{4.4}$$

Hierbei sind

$$f_{1,2} [x \mid \mu, \sigma] = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(lnx-\mu)^2}{2\sigma^2}} \tag{4.5}$$

Wahrscheinlichkeitsdichten zweier Lognormalverteilungen mit Medianwerten $\mu_{1,2}$ und Standardabweichungen $\sigma_{1,2}$. Die Konstante α stellt einen Gewichtungsfaktor zwischen den zwei Verteilungen dar und wurde zu $1.2 \cdot 10^{-5}$ bestimmt. Der Gewichtungsfaktor α markiert im Wesentlichen den Übergangspunkt zwischen der Tail- und Mainverteilung (Knick in der Retentionverteilung). Die Medianwerte sowie die Standardabweichungen wurden durch einen iterativen Fitprozess an die gemessene Verteilung nach der Methode der kleinsten quadratischen Abweichungen mit Hilfe von MATLAB ermittelt. Abbildung 4.1 zeigt die gemessene Retentionverteilung (Kreuze) zusammen mit der angefitteten Mischverteilung (durchgezogene Linie) in der kumulierten Darstellung.

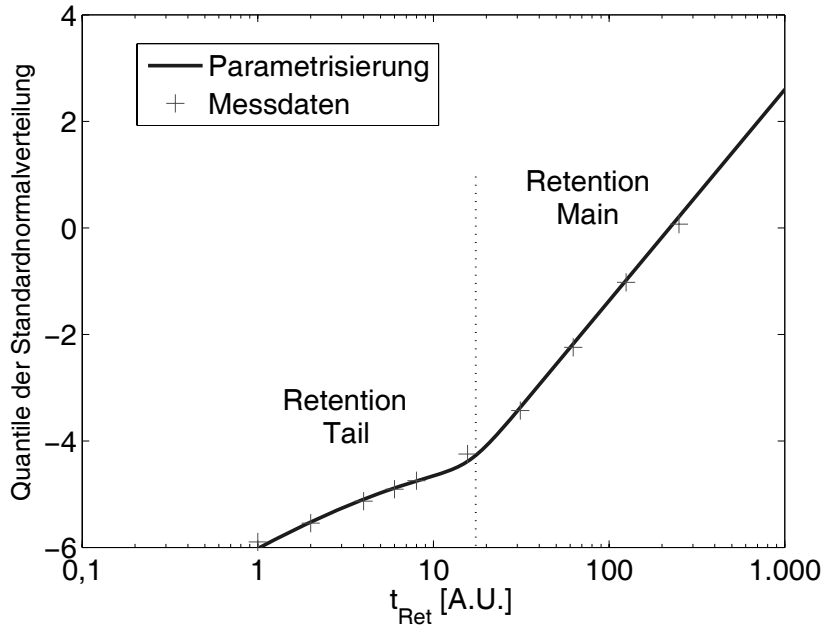


Abbildung 4.1: Kreuze: Gemessene Retentionverteilung eines $512\,Mbit$-Speicherchips. **Linie:** Parametrisierung der Verteilung durch eine Mischverteilung bestehend aus zwei Lognormal-Verteilungen.

4.3 Simulation der Leckstromverteilung

Mit Hilfe der im vorherigen Abschnitt bestimmten Verteilungen kann durch Monte Carlo Simulation, d.h. durch wiederholtes Erzeugen von Zufallszahlen für die bekannten Parameter entsprechend deren Verteilung und Einsetzen in Formel 4.1, die Leckstromverteilung gewonnen werden. Die Zufallszahlen wurden mit Hilfe der in MATLAB zur Verfügung stehenden Zufallszahl-Generatoren erzeugt. Die „Güte" der Zufallszahlen wurde für alle Parameter durch Überprüfung ihrer Verteilung und deren Kenngrößen sichergestellt. Die Handhabung der großen bei der Simulation auftretenden Datenmengen wurde dadurch gelöst, dass die Zielvariable in einzelne Wertebereiche aufgeteilt und in der Simulation nur die Häufigkeiten pro Bereich gespeichert wurden. Aus den Häufigkeiten kann durch Normierung auf die Gesamtzellenzahl die Wahrscheinlichkeitsdichte und daraus durch Integration die in Abbildung 4.2 gezeigte kumulative Leckstromverteilung (Kreuze) erhalten werden. Der Leckstrom zeigt demnach genau wie die Retentionverteilung zwei Unterverteilungen. Wiederum kann die resultierende Verteilung durch eine Mischverteilung aus zwei Lognormalverteilungen mit dem Gewichtungsfaktor $\alpha = 1.2 \cdot 10^{-5}$ angenähert werden (durchgezogene Linie). Die Breite der Leckstromverteilung entspricht der Breite der Retentionverteilung, d.h. vom Median bei $0\,\sigma$ bis zum äußersten Rand der Verteilung bei $6\,\sigma$ ungefähr 2.5 Größenordnungen. Die große Breite der Retentionverteilung spiegelt sich demzufolge in der Breite der Leckstromverteilung wider. Es kann weiterhin ausgeschlossen werden, dass die Retentiontailverteilung alleine durch eine ungünstige Kombination der Eingangsparameter entsteht. Vielmehr hat sie ihren Ursprung in der Verteilung der Leckströme selbst. Deshalb widmet sich diese Arbeit beginnend mit dem Kapitel 5 vor allem den Leckstrompfaden in DRAM-Zellen und den hierzu grundlegenden Mechanismen. Zuvor wird im folgenden Abschnitt jedoch darauf eingegangen, in wieweit die Retentionkurve durch die anderen Eingangsparameter beeinflusst werden kann.

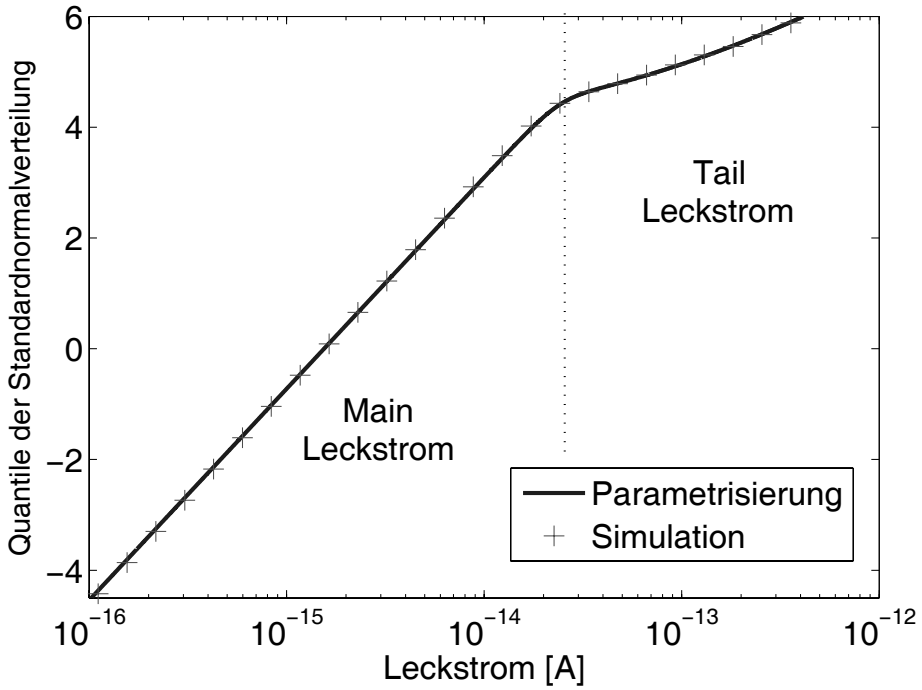


Abbildung 4.2: Gesamtleckstromverteilung als Ergebnis der Monte Carlo Simulation. Main-Leckströme liegen im fA-Bereich, Tail-Leckströme $1 - 2.5$ Größenordnungen darüber.

4.4 Rücksimulation der Retentionverteilung

Der Einfluss der Speicher- und Bitleitungskapazitäten sowie des Differenzverstärker-Offsets auf die Retentionkurve und insbesondere die Fehlerzahl an der Reparaturgrenze kann durch Rücksimulation der Retentionkurve aus der nun bekannten Leckstromverteilung analysiert werden. Da die Leckstromverteilung hier als empirisch ermittelte Größe eingeht, ist eine Rücksimulation nur mit solchen veränderten Parametern zulässig, die den Leckstrom selbst nicht verändern (z.B. C_S, C_{BL}, V_{SA}). Simulationsgleichung ist jetzt die Retention-Formel in ihrer ursprünglichen Form:

$$t_{Ret} = \frac{C_S}{\overline{I_{Leak}}} \left[(V_{BLH} \cdot P_w - \frac{V_{BLH}}{2}) - \frac{V_{SA}}{P_r \cdot P_n} (\frac{C_S + C_{BL} + 2C_{BLBL} + C_{SA}}{C_S}) \right] \tag{4.6}$$

4.4.1 Einfluss der Kondensatorkapazität

Abbildung 4.3 zeigt simulierte Retentionkurven für verschiedene mittlere Kondensatorkapazitäten C_s. Das Verhältnis μ/σ wurde bei allen Simulationen konstant gehalten. Durch Rücksimulation mit der ursprünglichen mittleren Kapazität von $30\,fF$ wird auch die ursprüngliche Retentionverteilung näherungsweise reproduziert, wodurch die Vorgehensweise legitimiert ist. Für kleinere mittlere Speicherkapazität verschiebt sich die Retentionkurve hin zu kleineren Retentionzeiten und für größere Kapazitäten zu größeren Retentionzeiten. Die Form der Verteilung bleibt dabei erhalten. Durch Extraktion der Ausfallzahl an der Reparaturgrenze und Normierung auf den Nominalfall von $30\,fF$ entsteht Abbildung 4.3b. Die Reparaturgrenze liegt an einem hier nicht genauer spezifizierten Punkt im Retentiontail. Aus der Abbildung kann für Kapazitäten kleiner $30\,fF$ eine Reduktion der Fehlerzahl um ungefähr 10% pro fF Speicherkapazität abgelesen werden. Für Kapazitäten größer $30\,fF$ wird die Kurve zunehmend flacher, das entspricht einer Fehlerreduktion von weniger als 10% pro fF. Dieses Ergebnis deckt sich mit experimentellen Erfahrungen.

4.4.2 Einfluss der Bitleitungslänge

Auch die Bitleitungslänge und damit die Anzahl der Speicherzellen pro Bitleitung hat Einfluss auf die Retentionverteilung. Um Chipfläche durch Reduzierung der notwendigen Anzahl von Differenzverstärkern zu sparen, möchte man möglichst viele Zellen an eine Bitleitung anschließen. Das Problem dabei ist, dass mit der Bitleitungslänge auch deren Kapazität steigt. Dadurch verschlechtert sich das Verhältnis von Speicherkapazität zu Gesamtkapazität und das Zellsignal beim Lesen verkleinert sich. *Signal Margin* Fehler (siehe Abschnitt 3.3) sind die Folge. Abbildung 4.4 zeigt simulierte Retentionkurven für verschiedene Bitleitungslängen. Zur Simulation wurden die Mittelwerte von C_{BL} und C_{BLBL} bei konstantem Verhältnis μ/σ variiert. Die Kapazität der Differenzverstärker C_{SA} blieb dabei konstant, da mit der Bitleitungslänge zwar die Anzahl benötigter Verstärker abnimmt, nicht jedoch deren Einzelkapazitäten. Eine Verdoppelung der Bitleitungskapazitäten zeigt sich zuerst in einer geringfügig flacheren Steigung der Mainverteilung. Im Tailbereich der Retentionkurve ist keine Veränderung sichtbar. Bei erneuter Verdoppelung wird die Mainverteilung nochmals flacher, außerdem treten nun sehr viele Fehler im Tail mit sehr kurzer Retentionzeit auf. Dies zeigt sich in einem nahezu waagrechten Auslaufen der Tailverteilung. Es handelt sich dabei um *Signal Margin* Fehler. Im Gegensatz dazu führt eine Halbierung der Bitleitungslänge zu keiner nennenswerten Veränderung der Retentionkurve. Die simulierten Kurven fallen mit der Kurve der nominellen Bitleitungslänge zusammen.

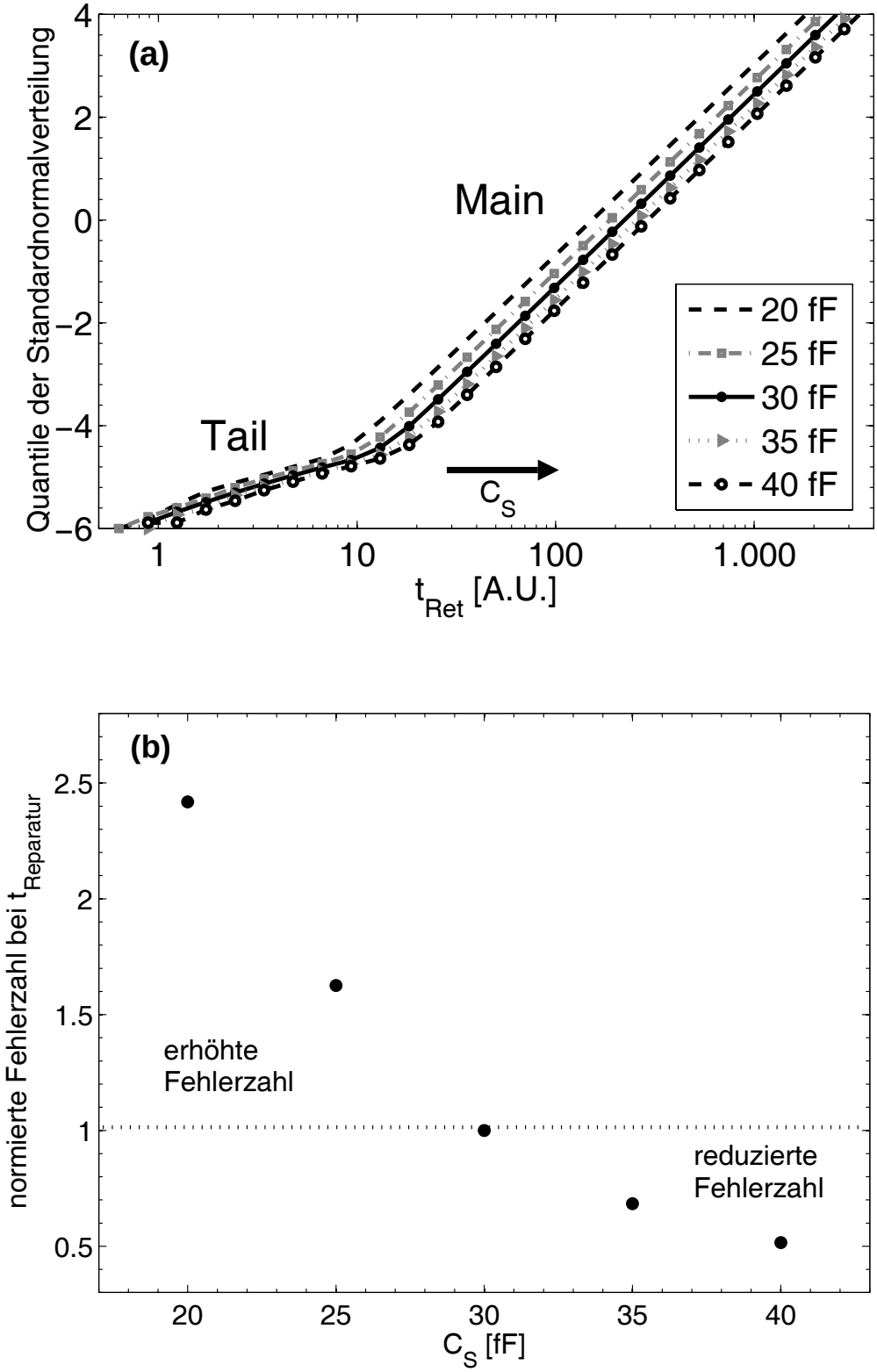


Abbildung 4.3: Einfluss der Speicherkapazität auf die Retentionverteilung. (a) Eine größere Speicherkapazität bewirkt eine Verschiebung der gesamten Verteilung hin zu größeren Retentionzeiten. Die Form der Retentionkurve bleibt dabei erhalten. (b) Die Fehlerzahl an der Reparaturgrenze nimmt mit größerer Kapazität exponentiell ab.

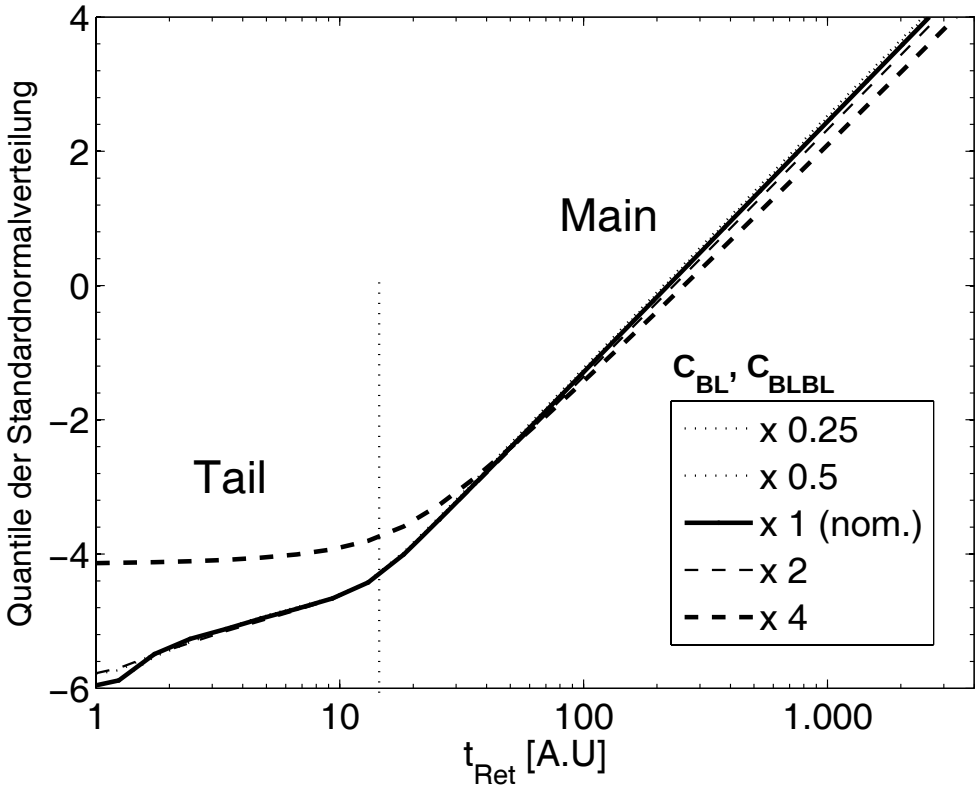


Abbildung 4.4: Einfluss der Bitleitungslänge/-kapazitäten auf die Retentionverteilung. Die Mittelwerte der Kapazitäten C_{BL} und C_{BLBL} wurden in der Rücksimulation um die in der Legende genannten Faktoren variiert. Das Verhältnis μ/σ wurde konstant gehalten.

4.4.3 Einfluss des Differenzverstärker-Offsets

Auch die V_{SA}-Verteilung hat Einfluss auf die Retentionverteilung. Genau wie die Bitleitungskapazitäten geht V_{SA} in den Term in eckigen Klammern der Retention-Formel ein (Gleichung 4.6). Unterschiedliche Mechanismen können zu einer Veränderung der V_{SA}-Verteilung führen. Zum Beispiel resultiert eine größere V_t-Streuung der Transistoren in den Differenzverstärkern in einer größeren Standardabweichung der V_{SA}-Verteilung, während eine größere Kapazität der Speicherkondensatoren durch die *Capacitive Imbalance* zu einer Verschiebung der gesamten Verteilung zu negativen V_{SA}-Spannungen führt. Im Folgenden werden der Mittelwert μ und die Standardabweichung σ unabhängig voneinander variiert, um ein grundlegendes Verständnis der Abhängigkeiten zu erhalten.

Veränderung des Mittelwerts

Eine Verschiebung des Mittelwertes der V_{SA}-Verteilung bei konstanter Standardabweichung wirkt sich für alle Zellen gleichermaßen aus (siehe Term in eckigen Klammern in Gleichung 4.6). Abbildung 4.5a zeigt simulierte Retentionkurven für verschiedene mittlere V_{SA}-Werte. Durch eine Verschiebung des Mittelwertes in positive Spannungsrichtung wird der Differenzterm in der Retentiongleichung kleiner und die Retentionzeiten aller Zellen kürzer. Umgekehrt verbessern sich die Retentionzeiten für einen negativeren Offset. Eine Veränderung des Mittelwertes der V_{SA}-Verteilung hat somit ähnliche Auswirkungen auf die Retentionverteilung wie die Variation der Speicherkapazität, nämlich eine Parallelverschiebung. Dies kann auch dadurch begründet werden, dass sowohl eine höhere Speicherkapazität als auch ein negativerer Offset netto zu einer größeren verfügbaren Ladung führen. Im ersten Fall nimmt der Kondensator mehr Ladungen auf, im zweiten Fall wird das Bewertungslevel abgesenkt, wodurch zum erfolgreichen Auslesen ein höherer Ladungsverlust zulässig ist. Die Auswertung an der Reparaturgrenze ist in Abbildung 4.5b zu sehen. Vom Nominalfall bei $-10\,mV$ ausgehend reduziert sich die Fehlerzahl ungefähr um 20% pro $10\,mV$ kleinerer Offsetspannung, während sie für größere Offsets um ungefähr 40% pro $10\,mV$ anwächst.

Es muss an dieser Stelle bemerkt werden, dass sich die Retentionzeit einer gespeicherten „0" bezüglich einer V_{SA} Verschiebung gerade invers zur „1" verhält und deshalb V_{SA} nicht beliebig abgesenkt werden kann. Wie im nächsten Kapitel genauer erläutert werden wird, ist eine „0" jedoch weniger von Leckströmen betroffen, sodass eine leicht negative Verschiebung der V_{SA}-Verteilung die „1" verbessert und nicht sofort zu „0"-Fehlern führt.

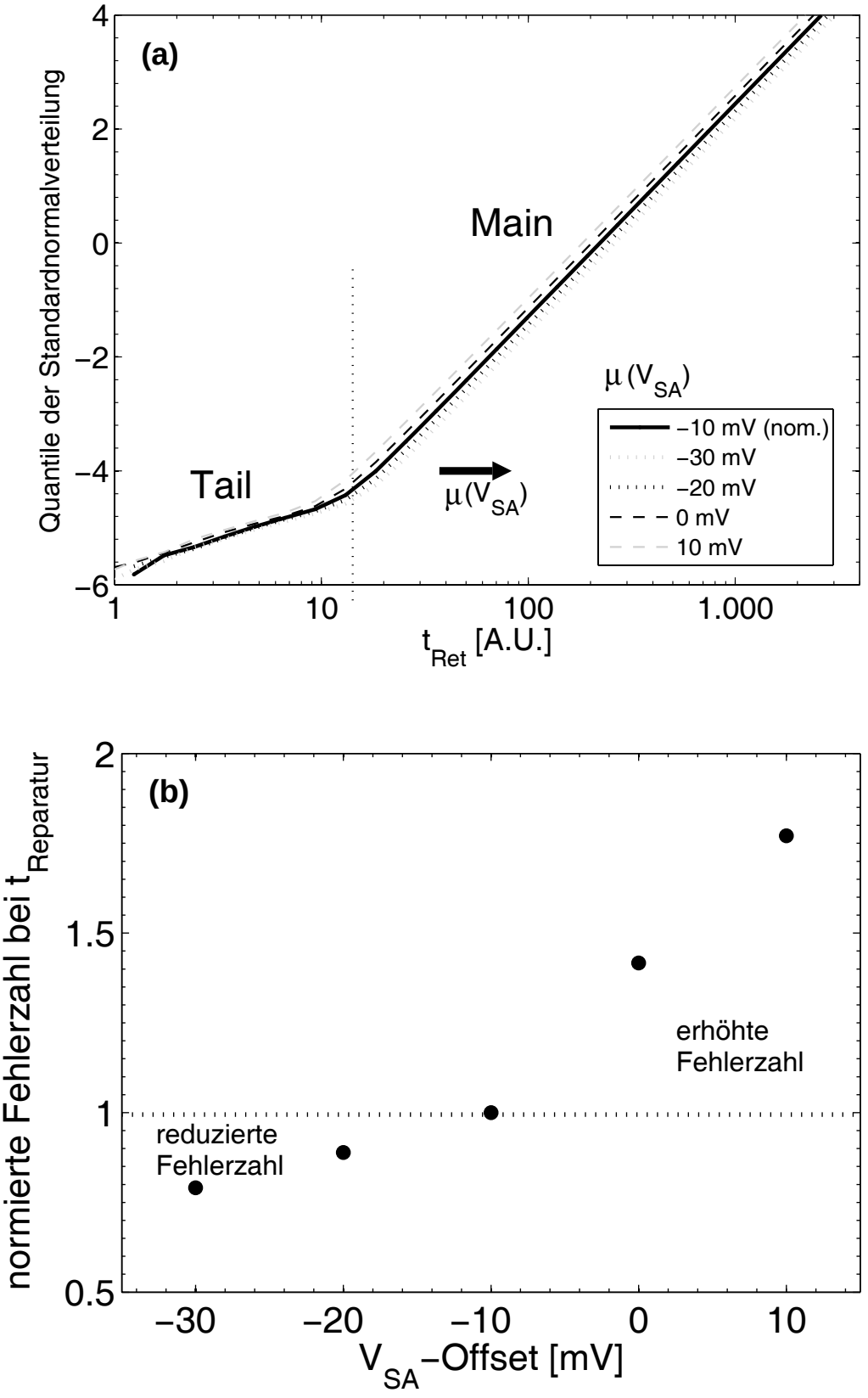


Abbildung 4.5: Einfluss des Mittelwertes der V_{SA}-Verteilung (a) auf die Retentionkurve, (b) auf die normierte Fehlerzahl an der Reparaturgrenze. Für positiveres $\mu(V_{SA})$ nimmt die Fehlerzahl zu.

Veränderung der Standardabweichung

Eine breitere V_{SA}-Verteilung wirkt sich in der Simulation ähnlich dem Fall höherer Bitleitungskapazitäten aus. Der Differenzterm in der Retention-Formel wird für einen gewissen Anteil der Zellen null und es entstehen *Signal Margin* Fehler (siehe Abbildung 4.6). Ein waagrechtes Auslaufen der Retentionverteilung im Tail ist die Folge.

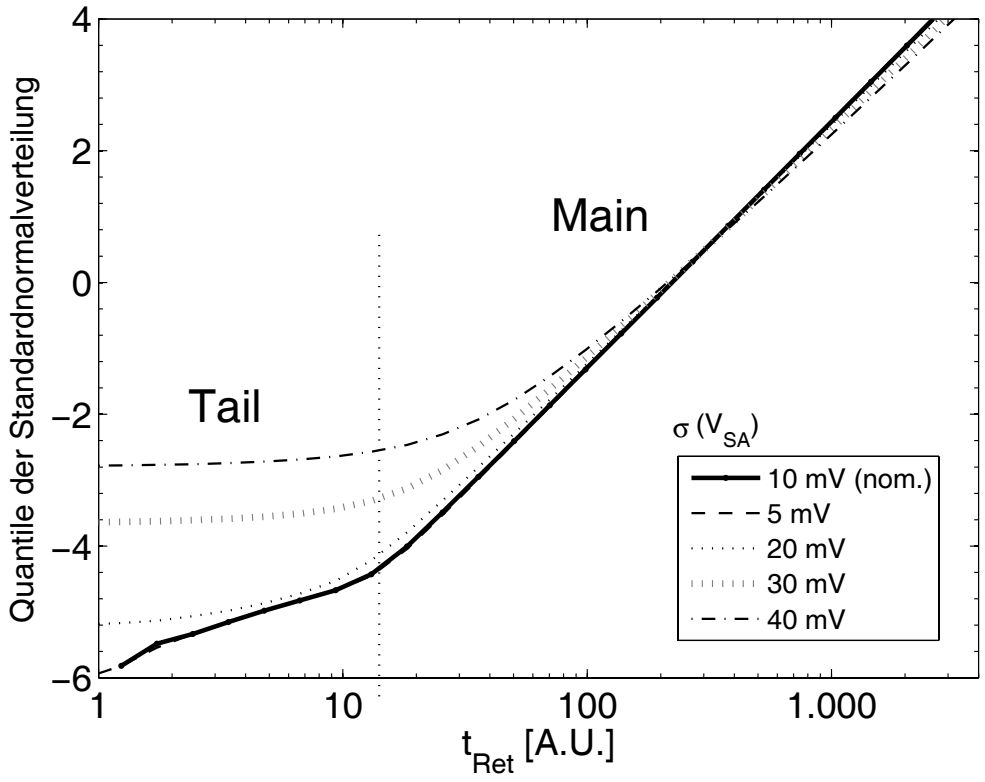


Abbildung 4.6: Einfluss der Standardabweichung der V_{SA}-Verteilung auf die Retentionkurve. Für größere Standardabweichungen treten zunehmend *Signal-Margin* Fehler auf.

Zusammenfassung:

Auf Grundlage der in Kapitel 3.2 eingeführten Retention-Formel wurde durch Monte Carlo Simulationen der Einfluss einiger eingehender Größen auf die Retentionverteilung untersucht. Dazu wurde zunächst aus einer gemessenen Retentionverteilung unter Annahme realistischer Parameterverteilungen die nicht direkt messbare Verteilung der Leckströme simuliert. Diese zeigt genau wie die Retentionverteilung zwei gerade Teilabschnitte im kumulativen Wahrscheinlichkeits-Plot und hat annähernd die gleiche Verteilungsbreite. Die breite Verteilung der Retentionzeiten ist deshalb hauptsächlich durch die Verteilung der Leckströme bestimmt und nicht etwa durch 6σ-Streuungen der anderen Parameter. Insbesondere ist die Aufspaltung in Main- und Tailverteilungen nicht durch ungünstige Parameterkombinationen zu erklären.

In weiteren Simulationen wurde der Einfluss der Speicherkapazität, der Bitleitungslänge und des Differenzverstärker-Offsets auf die Retentionverteilung qualitativ analysiert. Alle Parameter zeigen Einfluss auf die Verteilung und die für die Chipausbeute wichtige Fehlerzahl an der Reparaturgrenze. Der Haupthebel zur Verbesserung der Retentionverteilung liegt jedoch im Verständnis und der Reduktion der Leckströme selbst. Deshalb werden im folgenden Kapitel die möglichen Leckstrompfade in einer DRAM-Zelle betrachtet. Mit Hilfe der Einzelzellcharakterisierung in Kapitel 7 können Details der grundlegenden Mechanismen untersucht werden, die nach einer theoretischen Betrachtung in Kapitel 8 schließlich eine experimentelle Verifikation in Kapitel 9 erlauben.

Kapitel 5

Leckstrompfade im DRAM

Die Monte Carlo Simulationen in Kapitel 4 haben gezeigt, dass die große Breite der Retentionverteilung nicht durch Streuung der in die Retention-Formel eingehenden Parameter erklärbar ist. Die Verteilung kann nur durch eine breite Verteilung der Gesamt-Leckströme wiedergegeben werden. Der erste Teil dieses Kapitels gibt einen Überblick über die in einer *trench*-DRAM Zelle auftretenden Leckstrompfade. In den darauffolgenden Abschnitten wird auf diese detaillierter eingegangen. Das Kapitel endet mit einer Zusammenfassung der Spannungsabhängigkeiten der Leckstrompfade als Grundlage für die Charakterisierung in Kapitel 6.

5.1 Überblick

Es gibt eine ganze Reihe von möglichen Leckstrompfaden in einer DRAM Zelle. Abbildung 5.1 stellt eine Übersicht der Leckstrompfade dar. Auch in der deutschen Fachwelt sind dabei englische Bezeichnungen üblich und sollen deshalb auch in dieser Arbeit Verwendung finden. Die Leckstrompfade können gemäß ihrer physikalisch grundlegenden Mechanismen in drei Kategorien unterteilt werden:

1. **pn-Leckströme**

   (1) Gate Induced Drain Leakage (GIDL): vom kondensatorseitigen Gate/Drain-Überlappgebiet in die p-Wanne

   (2) Junction Leakage (JL): von BS-Anschlussgebiet in die p-Wanne

2. **Unterschwellen-Leckströme**

   (3) SubVt: vom BS-Anschlussgebiet zum Bitleitungskontakt

   (4) Vertical-Parasitic: vom BS-Anschlussgebiet zur Kondensatoraußenelektrode (BP)

   (5) SubSTI: vom BS zum BS oder Bitleitungskontakt eines benachbarten aktiven Gebietes

3. **Leckströme durch Dielektrika**

   (6) Node-Leakage: Leckstrom durch das Dielektrikum des Kondensators

   (7) Gateoxide-Leakage: Leckstrom durch das Gateoxid des Auswahltransistors

   (8) Passing-Wordline-Leakage: Leckstrom vom Kondensator durch das STI-Oxid in die darüberliegende passive Wortleitung

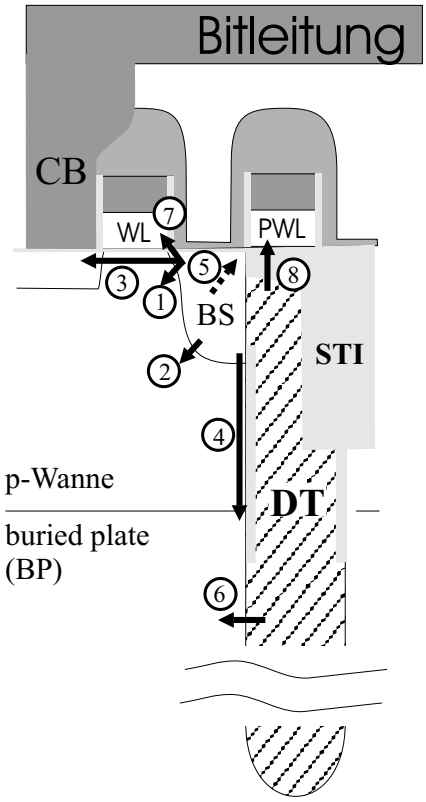


Abbildung 5.1: Übersicht der Leckstrompfade einer DRAM Zelle: (1) *GIDL*, (2) *Junction Leakage*, (3) *SubVt*, (4) *Vertical-Parasitic*, (5) *SubSTI*, (6) *Node*, (7) *Gateoxid*, (8) *Passing-Wordline*-Leakage.

5.2 pn-Leckströme

In einer DRAM-Speicherzelle gibt es drei pn-Übergänge (siehe Abbildung 5.2a). Alle sind im Speicherbetrieb zu jeder Zeit in Sperrrichtung geschaltet und es fließt ein unvermeidlicher Diodensperrstrom. Zum Informationsverlust tragen jedoch nur Leckströme bei, die zur Ladungsänderung auf der inneren Elektrode des Kondensators führen. Die Leckströme der pn-Übergänge (2) und (3) erhöhen somit den Gesamtstromverbrauch des Speicherchips jedoch ändern sie nicht die Ladung des Kondensators und sind deshalb für die Retention der Speicherzellen bedeutungslos. Im Folgenden wird daher nur der kondensatorseitige pn-Übergang vom BS-Anschluss zur p-Wanne weiter betrachtet. Wie schon bemerkt wurde, ist der pn-Übergang sowohl für eine gespeicherte „1" also auch für eine

gespeicherte „0" in Sperrrichtung gepolt. Der Sperrstrom reduziert deshalb das Potenzial der inneren Kondensatorelektrode in beiden Fällen. Im Falle der gespeicherten „1" führt dies zu einer Zerstörung der Information, während die „0" durch den Leckstrom gestärkt wird.

Der kondensatorseitige pn-Übergang kann seinerseits in zwei Bereiche unterteilt werden. In den Bereich des Überlapps zwischen dem Gate und dem n+ Gebiet (Drain) und in den Bereich außerhalb der elektrischen Reichweite des Gates (siehe Abbildung 5.2b). Im ersten Fall wird der pn-Übergang durch das Gate des Transistors beeinflusst. Dies entspricht der Konfiguration einer „gated diode" und der in diesem Bereich entstehende Leckstrom wird „gate induced drain leakage" (GIDL) genannt. Im zweiten Fall handelt es sich um einen einfachen pn-Übergang, der nur von der Sperrspannung selbst abhängt. Dieser wird mit „Junction Leakage" bezeichnet.

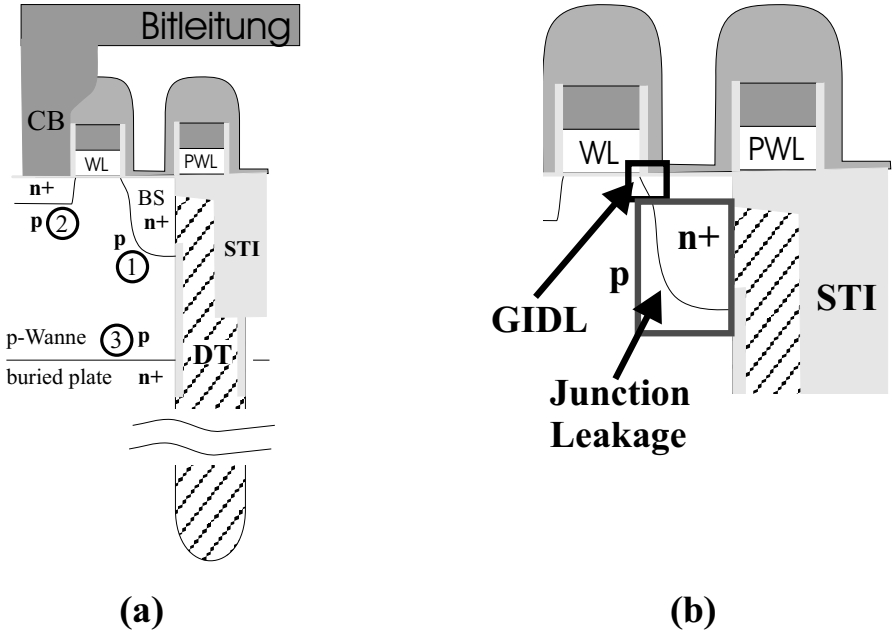


Abbildung 5.2: (a) In Sperrrichtung gepolte pn-Übergänge einer DRAM Speicherzelle: Von der p-Wanne zum Buried Strap (1), zum Bitleitungskontakt (2) und zur Kondensator-Außenelektrode (3). Nur der Leckstrom von (1) führt zu Informationsverlust. (b) Unterteilung des Leckstroms von pn-Übergang (1) in den durch das Gate beeinflussten *GIDL* und einfachen *Junction Leakage*.

5.2.1 GIDL

Abbildung 5.3 zeigt den GIDL-Leckstrom einer Speicherzelle schematisch. Im ausgeschalteten Zustand des Zelltransistors bewirkt die Wortleitungs-Spannung V_{NWLL} eine starke Bandverbiegung im Gate-Drain Überlappbereich, die im einfachen Falle einer MOS Kapazität zur Inversion der Oberfläche führen würde. Im Falle der „gated diode" Konfiguration werden an der Oberfläche ankommende Minoritätsladungsträger (hier

Löcher) jedoch gleich in die p-Wanne abgezogen, da das Potenzial für diese dort niedriger ist. Dadurch fließt ein Strom in die p-Wanne und eine Inversionsschicht kann nicht ausgebildet werden. Man spricht in diesem Fall von tiefer Verarmung. Durch Störstellen verursachte Zustände innerhalb der Bandlücke im Bereich der tiefen Verarmung können Ladungsträgerpaare generieren, wodurch der Leckstrom drastisch erhöht wird. Dieser Aspekt wird in Kapitel 8 genauere Betrachtung finden. Moduliert werden kann die Ladungsträger-Generationsrate und damit auch der GIDL-Leckstrom durch die elektrische Feldstärke im Gate/Drain Überlappbereich, welche von der Gatespannung im ausgeschalteten Zustand V_{NWLL} und der Spannung des Speicherkondensators V_S abhängt.

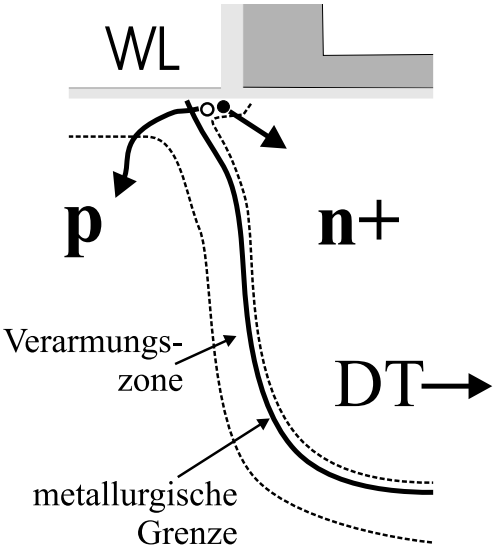


Abbildung 5.3: Schematische Darstellung des GIDL-Leckstroms. Ladungsträger-Paare entstehen in der Überlapp-Region durch Tunneleffekte und werden durch das elektrische Feld getrennt. Die Generationsrate wird durch das elektrische Feld bestimmt, welches durch die Gatespannung V_{NWLL} moduliert wird.

5.2.2 Junction-Leakage

Ähnlich zum GIDL führt die Generation von Ladungsträgern in der Verarmungszone des kondensatorseitigen pn-Übergangs zu einem Fluss von Elektronen in das Anschlussgebiet der inneren Kondensatorelektrode (Abbildung 5.4). Außerhalb des Einflussgebiets des Gates wird der Generationsstrom *Junction Leakage* genannt und hängt im Gegensatz zum GIDL nur von der Potenzialdifferenz der n- und p-Seite ab. Zur elektrischen Charakterisierung dienen das Potenzial der p-Wanne V_{BB} und das Potenzial des Speicherknotens V_S. Dabei gilt generell, dass eine höhere Potenzialdifferenz einen größeren Leckstrom verursacht. Mögliche Generations-Mechanismen und deren Spannungsabhängigkeiten werden in Kapitel 8 behandelt.

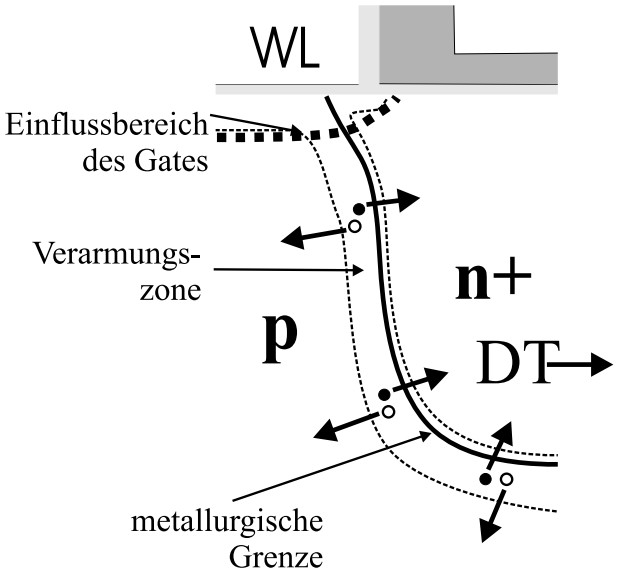


Abbildung 5.4: Schematische Darstellung von Junction-Leakage. Dabei handelt es sich um Generationsströme, die in der Verarmungszone des kondensatorseitigen pn-Übergangs außerhalb des Einflussgebiets des Gates entstehen.

5.3 Transistor-Unterschwellenleckströme

Trotz der vermeintlichen 1-Transistor-Zelle gibt es in einer DRAM Zelle eigentlich insgesamt drei MOS-Transistoren (siehe Abbildung 5.5). Es handelt sich dabei um den Auswahltransistor (1), einen vertikalen parasitären Transistor (2) und eine transistorartige Struktur vom BS-Gebiet der betrachteten Zelle zum BS-Gebiet bzw. dem Bitleitungskontakt des benachbarten aktiven Gebietes (3). Nur der Auswahltransistor wird als Transistor mit Schaltfunktion betrieben. Bei den beiden anderen handelt es sich um parasitäre Transistoren, die durch das Integrationsschema der Zelle bedingt sind und für die ordnungsgemäße Funktion der Speicherzelle zu jeder Zeit sicher abgeschaltet sein müssen.

5.3.1 SubVt-Leckstrom

Im Speicherzustand einer DRAM-Zelle ist der Auswahltransistor ausgeschaltet und befindet sich im Unterschwellenbereich. Typische Sättigungsströme $I_{DS,sat}$ liegen im Bereich von $20 - 30\,\mu A$ und zur Informationserhaltung notwendige Off-Ströme I_{OFF} im Bereich von wenigen fA. Der Off-Strom muss somit ungefähr neun Größenordnungen unter dem On-Strom liegen. Geht man von einer typischen Unterschwellensteigung von $100\,mV/dec$ aus und berücksichtigt etwas Vorhalt aufgrund von statistischen Streuungen der Einsatzspannungen, so folgt daraus, dass die Gatespannung im ausgeschalteten Zustand V_{NWLL} ungefähr $1\,V$ unter der Einsatzspannung liegen muss. Abbildung 5.6

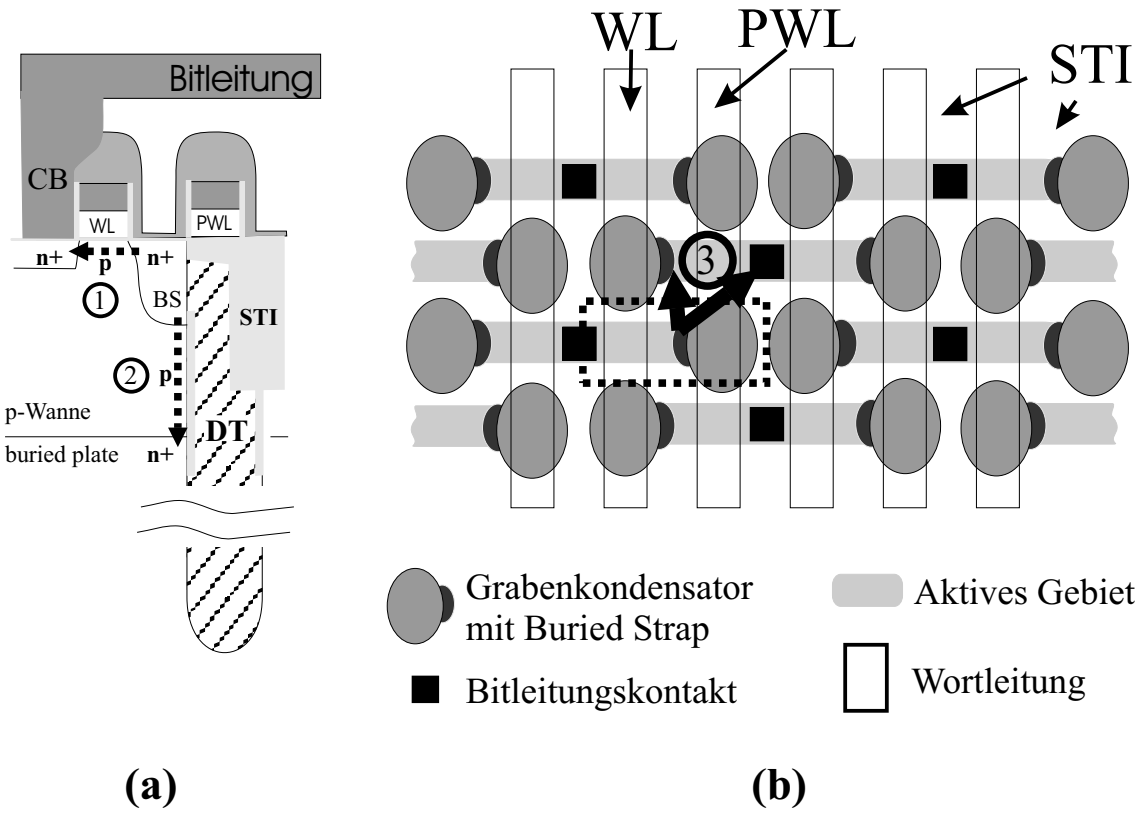


Abbildung 5.5: Transistorstrukturen in einer DRAM Zelle.(a) Schnitt senkrecht zu den Wortleitungen, (b) Sicht von oben auf das Zellenfeld. Die Speicherzelle aus (a) ist in (b) gestrichelt dargestellt. Neben dem Auswahltransistor (1) existiert ein vertikaler parasitärer MOSFET (2) und eine transistorartige Struktur zum benachbarten aktiven Gebiet mit dem STI als Gateoxid (3). Die Unterschwellenströme der drei Transistoren tragen zum Entladestrom des Speicherkondensators bei.

zeigt den Verlauf des *SubVt*-Pfades in einer Speicherzelle schematisch. Da es sich beim Auswahltransistor um einen Kurzkanaltransistor handelt, wirkt sich das Potenzial der Bitleitung über den DIBL-Effekt auf die Einsatzspannung und somit auch auf den Unterschwellenstrom einer betrachteten Zelle aus. Die Bitleitung dieser Zelle kann je nach Betriebszustand verschiedene Potenziale annehmen. Erfolgt kein Zugriff auf Zellen des Bitleitungspaares, so liegen beide Leitungen auf dem Vorladezustand $V_{BLH}/2$, wird dagegen von einer Zelle der gleichen Bitleitung eine „1" gelesen oder in diese geschrieben, so liegt das Potenzial V_{BLH} an und der Unterschwellenstrom der betrachteten Zelle wird während dieser Zeit durch Verringerung des DIBL-Effekts reduziert. Entsprechend liegt beim Lesen bzw. Schreiben einer „0" an der gleichen Bitleitung deren Potenzial auf V_{BLL} und der Unterschwellenleckstrom wird erhöht. Aufgrund des differenziellen Verstärkungsprinzips wirkt sich das Lesen/Schreiben von Zellen der komplementären Bitleitung invers dazu aus. Der DIBL-Effekt kann somit zur Charakterisierung der SubVt-Anfälligkeit eingesetzt werden. Eine andere Möglichkeit besteht in der Variation der Gatespannung im ausgeschalteten Zustand (V_{NWLL}). Dabei erhöht bzw. erniedrigt sich der SubVt-Leckstrom mit der Gatespannung gemäß der Unterschwellensteigung S. Außerdem kann der Substratsteuereffekt zur Charakterisierung herangezogen werden. Eine erniedrigte Spannung der p-Wanne (V_{BB}) erhöht die Schwellenspannung des Transistors und reduziert dadurch bei gleicher Gatespannung den Leckstrom gemäß der Unterschwellensteigung. Entsprechend wird der SubVt-Leckstrom durch erhöhtes V_{BB} vergrößert.

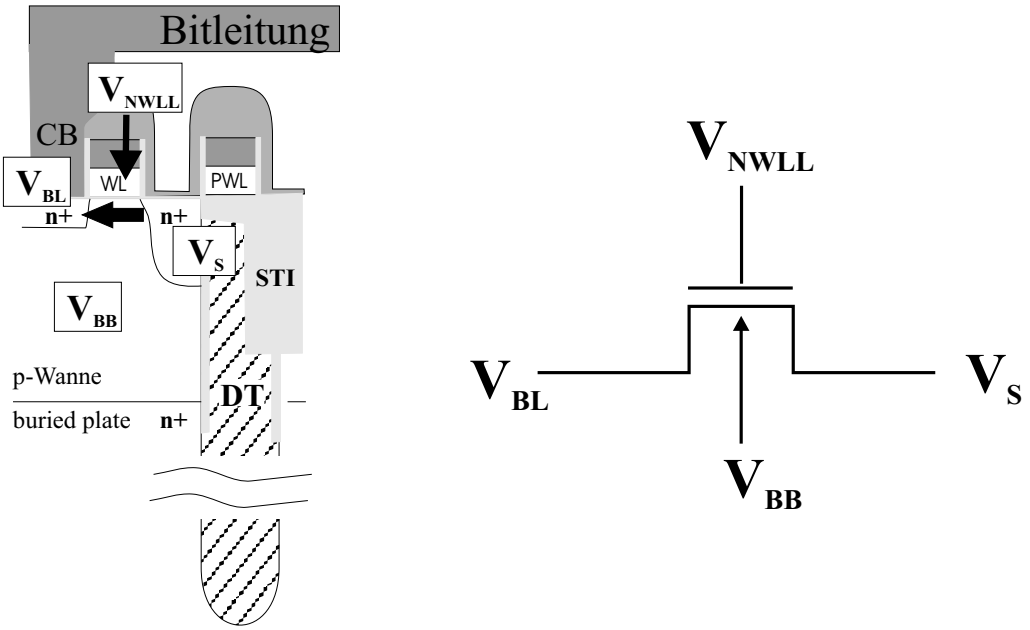


Abbildung 5.6: Unterschwellenleckstrom des Auswahltransistors.

5.3.2 Deep-SubVt Leckstrom

Mit *Deep-SubVt* wird der Leckstrompfad entlang der Seitenflächen des aktiven Gebietes tiefer im STI und außerhalb des Einflussbereiches des Gates bezeichnet. Auch dieser Pfad ist ein Unterschwellenleckstrom einer parasitären Transistorstruktur. Dabei bilden der Bitleitungskontakt und der Buried Strap die S/D-Gebiete, die STI-Isolation das Gateoxid und der Kondensator der Nachbarzelle das Gate. Eine gespeicherte „1" in der Nachbarzelle schaltet den parasitären Transistor an und erhöht den Deep-SubVt-Leckstrom, während eine „0" diesen reduziert. Dadurch kann durch gezieltes Schreiben einer „0" in die Nachbarzellen überprüft werden, ob der *Deep-SubVt* Pfad die Retentionzeit einer Zelle maßgeblich bestimmt. Verbessert sich die Retentionzeit einer Zelle beim Übergang von „1"-en in den Nachbarzellen zu „0"-en, so ist Deep-SubVt dominant.

5.3.3 Vertikaler Parasitärer Leckstrom

Der Oxidkragen (Collar) im Kondensator bildet das Gateoxid eines weiteren parasitären Transistors mit dem *Buried Strap* Ausdiffusionsgebiet als Drain und der *Buried Plate* als Source (siehe Abbildung 5.7). Die innere Kondensatorelektrode bildet das Gate und ist mit der Drain leitend verbunden. Die Einsatzspannung V_t des Transistors wird maßgeblich durch die Dicke des Collar-Oxids bestimmt. Diese muss so gewählt werden, dass der parasitäre Transistor in allen Betriebszuständen sicher abgeschaltet ist. Wie bei allen parasitären Transistoren kann der Substratsteuereffekt zur Charakterisierung herangezogen werden. Desweiteren wird der Leckstrom durch Variation der S/D-Spannung (DIBL-Effekt), d.h. Variation der Spannungen V_{PL} oder V_{BLH} moduliert. Bei letzterem Fall ist zu beachten, dass gleichzeitig die Gate/Source-Spannung verändert wird.

5.3.4 SubSTI Leckstrom

Leckströme zwischen benachbarten aktiven Gebieten (AA) werden mit *SubSTI* bezeichnet, da sie unter der Shallow Trench Isolation (STI) zum benachbarten aktiven Gebiet fließen. Auch der SubSTI-Leckstrom ist als Unterschwellenstrom eines parasitären Transistors zu verstehen. Dabei ist die Drain das BS-Gebiet der untersuchten Zelle, die Source ist entweder durch das BS-Gebiet einer Nachbarzelle (Abb. 5.8a C->D) oder durch den Bitleitungskontakt des Nachbar-AAs (Abb. 5.8a C->E) gegeben. Das STI-Oxid selbst bildet das Gateoxid des parasitären Transistors. Die Passing-Wordline (PWL) oder Ladungen im STI-Oxid können zu einem Inversionsstrom entlang der Si/STI-Grenzfläche führen (siehe Abb. 5.8b und c).

Charakterisiert werden kann SubSTI-STI durch unterschiedliche Datentopologien. Wird in alle Zellen eine „1" geschrieben, ist der Leckstrompfad C->D aufgrund des gleichen

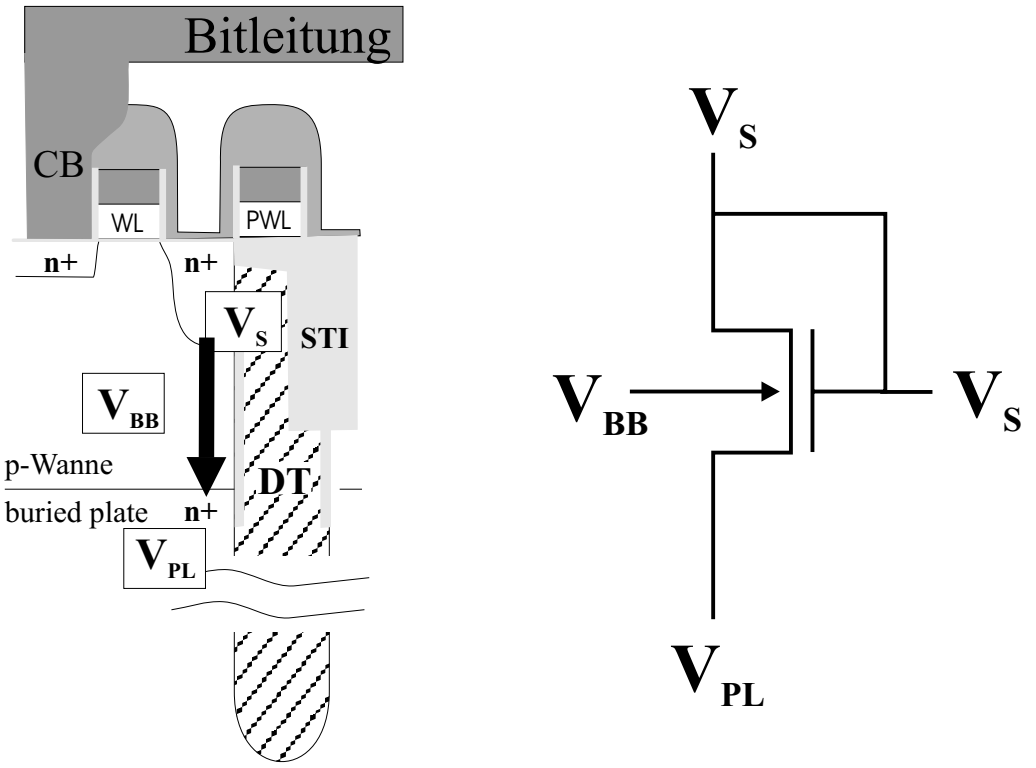


Abbildung 5.7: Unterschwellenleckstrom des vertikalen parasitären Transistors.

Potenzials an Source und Drain ausgeschaltet. Wird dagegen in benachbarte Zellen „0" geschrieben ist die S/D-Potenzialdifferenz maximal. Der Leckstrom von C->E wird durch Zugriffe auf Zellen an der Nachbarbitleitung und deren komplementären Bitleitung moduliert. Der benachbarte Bitleitungskontakt kann während des Speicherbetriebs dadurch die Spannungszustände V_{BLH}, V_{BLEQ} und V_{BLL} annehmen. Für Testzwecke werden alle Bitleitungen während der Retentionpause auf $V_{BLL} = GND$ geschaltet.

5.4 Leckströme durch Dielektrika

Bei einer idealen MOS-Kapazität fließt kein Strom durch das Dielektrikum. In der Realität besitzt jedes Dielektrikum jedoch eine endliche Leitfähigkeit. Dabei gibt es prinzipiell zwei Möglichkeiten durch die Ladungsträger zu einem Strom durch das Dielektrikum beitragen können. Zum einen können Ladungsträger mit genügend hoher Energie (heiße Ladungsträger) ins Leitungsband (Elektronen) bzw. Valenzband (Löcher) des Dielektrikums gelangen und dadurch zur Leitung beitragen und zum anderen können bei sehr dünnen Dielektrika Tunneleffekte zu einem Strom durchs Dielektrikum führen. Im Ladungserhaltungszustand des DRAM spielen Effekte durch heiße Ladungsträger keine Rolle, da Ladungsträger nur relativ geringe thermische Energien besitzen. Deshalb soll hier nur auf Tunneleffekte eingegangen werden. Im Folgenden werden verschiedene Mechanismen kurz vorgestellt.

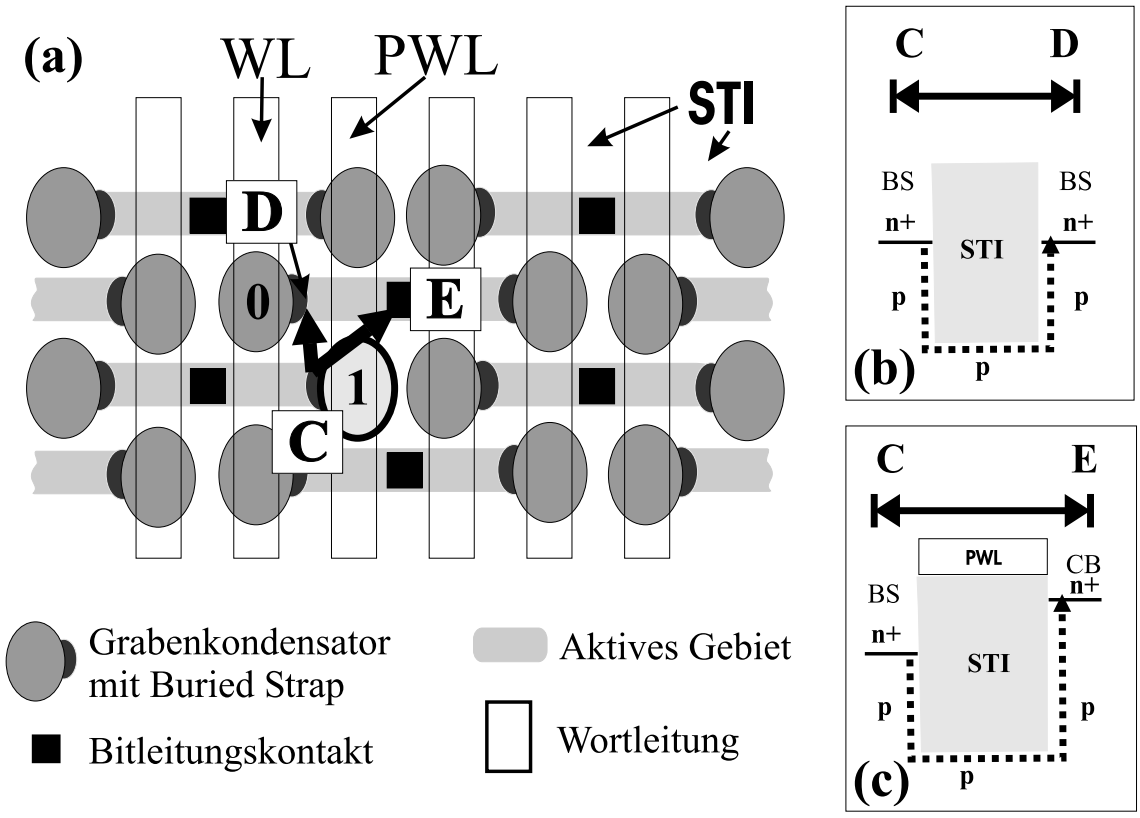


Abbildung 5.8: (a) Schematische Darstellung des Zellenfeldes in Draufsicht. Die SubSTI-Leckstrompfade sind durch Pfeile markiert. (b) Schnitt senkrecht zur Oberfläche entlang des Pfades von C nach D. (c) Schnitt senkrecht zur Oberfläche entlang des Pfades von C nach E.

Fowler-Nordheim-Tunneln

Bei einer MOS-Kapazität in Inversion können Elektronen aus der Inversionsschicht des Substrats in das Leitungsband des Dielektrikums tunneln (siehe Abbildung 5.9a). Die Tunnelwahrscheinlichkeit ist dabei von der Barrierendicke abhängig, die wiederum vom elektrischen Feld im Dielektrikum abhängt. Ein höheres elektrisches Feld führt zu größerer Bandverbiegung und damit zu einer geringeren Barrierendicke. Die Tunnelwahrscheinlichkeit steigt dadurch mit dem elektrischen Feld an. Die komplette Fowler-Nordheim-Theorie ist sehr komplex. Unter Vernachlässigung der endlichen Temperatur und der Barrierenerniedrigung durch Spiegelladungen kann die FN-Stromdichte durch Lösen der Schrödingergleichung für eine dreiecksförmige Barriere mit Hilfe der WKB-Näherung berechnet werden (z.B. [Tau98]) :

$$J_{FN} = A_F \cdot F_{ox}^2 \cdot exp(-\frac{B}{F_{ox}}) \tag{5.1}$$

mit

$$A_F = \frac{q^3}{16 \cdot \pi^2 \hbar \Phi_{ox}} \tag{5.2}$$

$$B = \frac{4 \cdot \sqrt{2 \cdot m^*}}{3 \cdot \hbar \cdot q} \Phi_{ox}^{\frac{3}{2}} \tag{5.3}$$

ϕ_{ox} ist hierbei die Barrierenhöhe in eV. Für SiO_2 ergeben sich daraus die Werte $A_F \approx 1.25 \cdot 10^{-6}\, A/V^2$ und $B \approx 240\, MV/cm$. Selbst für ein Oxidfeld $F_{ox} = 8\, MV/cm$, welches nur unter *Burn-In* Bedingungen erreicht wird, ergibt sich daraus eine Fowler-Nordheim-Tunnelstromdichte von lediglich $J_{FN} \approx 7 \cdot 10^{-6}\, A/cm^2$, welche um Größenordnungen unter den Tail-Leckströmen liegt.

Direktes Tunneln

Für sehr dünne Oxide (~kleiner $3\, nm$) können Elektronen aus dem invertierten Silizium direkt durch das verbotene Band des Oxids ins Leitungsband der n^+-Elektrode tunneln. Daher sind Transistoren mit Gateoxiden von weniger als $3\, nm$ für DRAM Anwendungen nicht akzeptabel. Im Fall von direktem Tunneln ist die Barriere trapezförmig und es gibt keinen einfachen analytischen Zusammenhang zwischen Stromdichte und elektrischem Feld [Tau98, S.96]. Numerische Rechnungen zeigen jedoch, dass das direkte Tunneln im Vergleich zum Fowler-Nordheim-Tunneln weniger vom elektrischen Feld abhängt. Bei kleinen Spannungen und dünnen Oxiden dominiert deshalb das direkte Tunneln [Wol02, S.102].

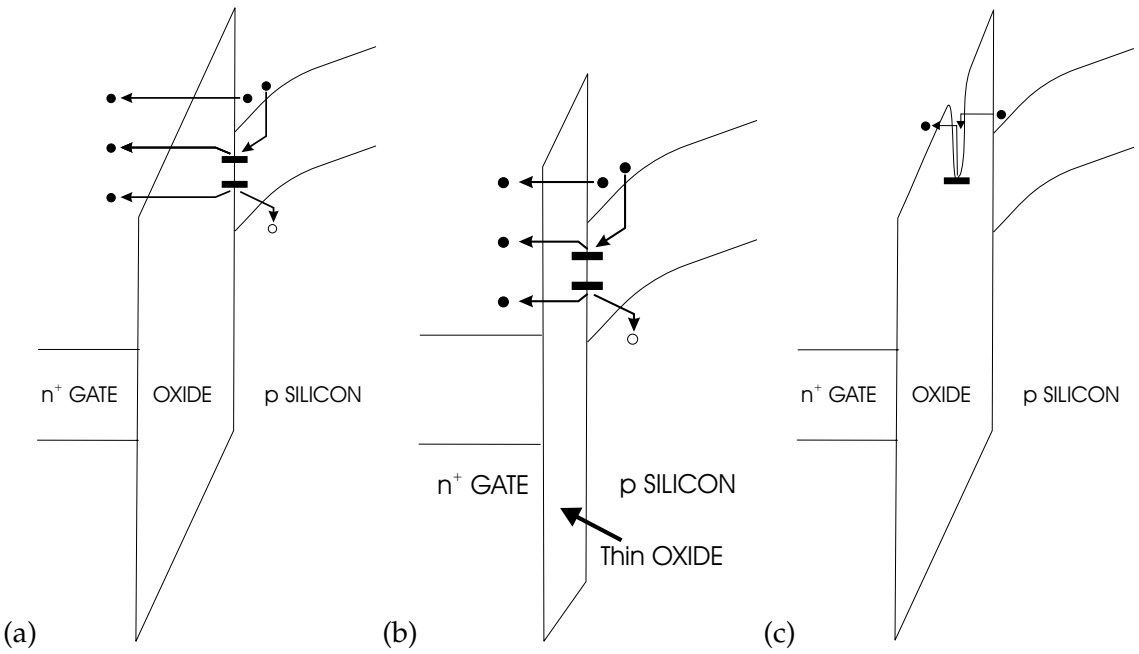


Abbildung 5.9: Leckstrommechanismen durch ein Dielektrikum. (a) Fowler-Nordheim-Tunneln, (b) direktes Tunneln, (c) Pool-Frenkel Emission. Ladungsträger können dabei sowohl aus der Inversionsschicht im p-Silizium, als auch aus besetzten Grenzflächen-Zuständen zum Tunnelstrom beitragen.

Poole-Frenkel-Tunneln

Wenn Ladungsträger zu Traps im Dielektrikum tunneln und von dort durch thermische Anregung ins Leitungsband des Dielektrikums gelangen, spricht man vom Pool-Frenkel-Mechanismus. Besonders für high-k Materialien (z.B. Al_2O_3) ist dieser Mechanismus von Bedeutung, da diese im Vergleich zu SiO_2 mehr Trapzentren aufweisen. Die Abhängigkeit der Stromdichte J vom elektrischen Feld F und der Temperatur T ist gegeben durch [Sze81, S. 403]:

$$J \propto F \cdot exp(-\frac{q(\Phi_{ox} - \sqrt{q \cdot F/\pi\epsilon_i})}{k_B \cdot T}) \tag{5.4}$$

5.4.1 Node-Leckstrom

Der Leckstrompfad durch das Dielektrikum des Zellkondensators wird als *Node-Leakage* bezeichnet. Im Falle einer gespeicherten „1" gelangen dadurch Elektronen von der *Buried Plate* zur inneren Kondensatorelektrode (Abbildung 5.10a); im Falle einer gespeicherten „0" anders herum (Abbildung 5.10b). Als grundlegende Mechanismen kommen die im vorhergehenden Abschnitt beschriebenen Mechanismen Fowler-Nordheim, Direktes und Pool-Frenkel-Tunneln in Frage. Defektprobleme, wie Unregelmäßigkeiten im Dielektrikum (z.B. Pinholes oder lokale Abdünnungen), können ebenfalls zu erhöhten

Leckströmen führen. Zur Charakterisierung wird die Feldabhängigkeit der Mechanismen ausgenutzt. Dazu wird das elektrische Feld übers Dielektrikum durch Variation der Spannung V_{PL} bzw. V_S moduliert.

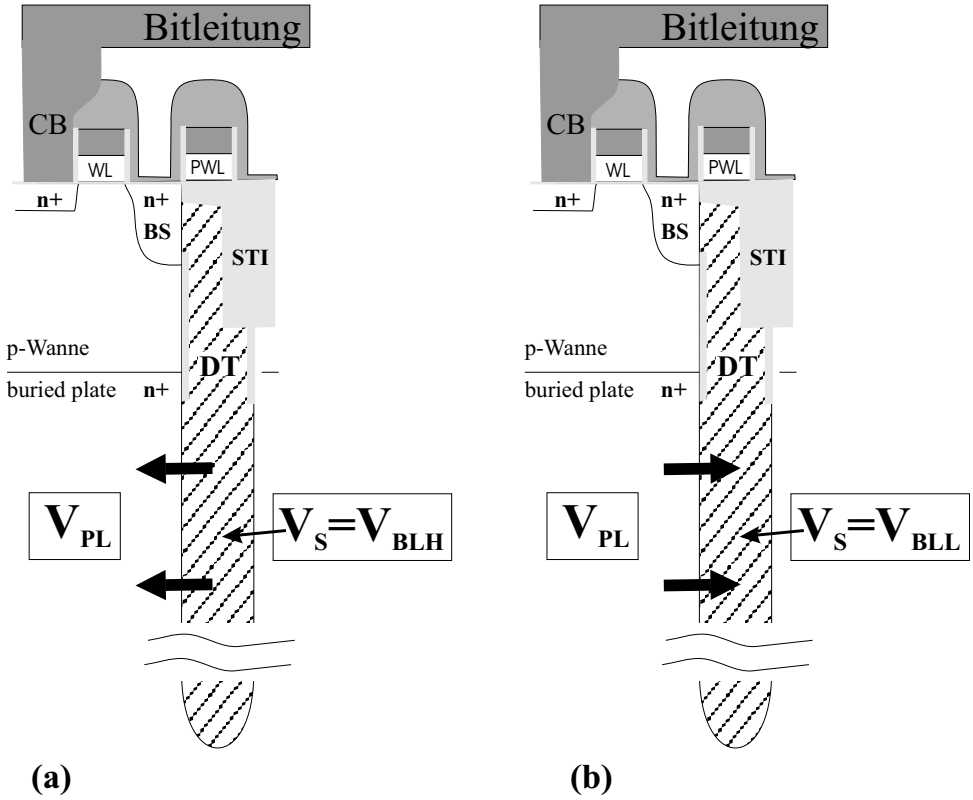


Abbildung 5.10: Der Node-Leckstrom ist aufgrund von $V_{PL} = V_{BLH}/2$ symmetrisch bezüglich den zwei Speicherzuständen. (a) gespeicherte „1", (b) gespeicherte „0". D.h. der Leckstrom durchs Node verschlechtert beide Speicherzustände gleichermaßen.

5.4.2 Gateoxid Leckstrom

Die Qualität des Gateoxids wird im so genannten *Burn-In* Test, der Bestandteil der Standard-Speichertests ist, überprüft. Ziel dabei ist, Degradationsmechanismen, die durch Defekte und Unregelmäßigkeiten im Gateoxid bedingt sind, zu beschleunigen und dadurch spätere Ausfälle beim Kunden zu verhindern. Dadurch kann die ausgelieferte Qualität erheblich gesteigert werden. Im *Burn-In* werden dazu alle Speicherbausteine unter erhöhten Temperaturen und Spannungsbedingungen betrieben. Man spricht von Temperatur- und elektrischem *Stress*. Der bezüglich Leckströmen durchs Gateoxid kritischste Betriebszustand im DRAM ist dabei nicht der Speicherzustand, d.h. der Auszustand des Transistors, sondern das Schreiben bzw. Lesen der Speicherzelle, da hierbei größere Oxidfelder auftreten. Es kann davon ausgegangen werden, dass alle Defekte, die unter Normalbedingungen einen Leckstrom im Tailbereich treiben können, unter *Burn-In* Bedingungen zum Durchbruch führen würden und entsprechende Bausteine

verworfen werden. Leckströme durchs Gateoxid kommen deshalb nicht als Ursache für die Retention-Tailverteilung in Frage und werden im Folgenden nicht weiter betrachtet.

5.4.3 Passing-WL Leckstrom

Die Oxiddicken zwischen der inneren Kondensator-Elektrode und der darüber verlaufenden Passing-WL sind zu dick ($> 30\,nm$), als dass Ströme durch Tunnelmechanismen zu nennenswerten Ladungsverlusten führen könnten. Durch Schwankungen der Oxiddicke zwischen der Passing-WL und der inneren Elektrode des darunter liegenden Speicherkondensators kann es zu erhöhten Leckströmen kommen. Strukturelle Untersuchungen mittels SEM und TEM von besonders schlechten Zellen aus der Tailverteilung zeigen im Allgemeinen keine erkennbaren Defekte, sodass dieser Leckstrompfad als Ursache für die breite Retentionverteilung im Folgenden nicht weiter in Betracht gezogen wird.

5.5 Übersicht der Spannungsabhängigkeiten

In den bisherigen Abschnitten dieses Kapitels wurden die existierenden Leckstrompfade und deren grundlegenden Mechanismen vorgestellt. Die den jeweiligen Leckstrom beeinflussenden internen Spannungen wurden angesprochen. Die Schwierigkeit der Charakterisierung besteht darin, dass die verschiedenen Leckstrompfade nicht separiert betrachtet werden können, da alle internen Spannungen gleichzeitig mehrere Leckstrompfade ansprechen. An dieser Stelle sollen die Abhängigkeiten der Leckstrompfade von den internen Spannungen eines Speicherbausteins nochmals zusammengefasst werden. Der Tabelle 5.11 kann die qualitative Leckstromänderung bei Abweichung von den nominellen Spannungswerten bzw. bei Änderung der Datentopologie entnommen werden. Die Leckströme einer Speicherzelle können jedoch nicht direkt gemessen werden. Anstatt dessen wird die Änderung der Retentionzeit t_{Ret} bestimmt, welche sich in erster Näherung inversproportional zum Gesamtleckstrom verhält (Details zum Messverfahren siehe Kapitel 6). Um den anfänglichen Ladungszustand der Speicherzellen nicht durch veränderte Schreib-/Lesebedingungen zu beeinflussen und die t_{Ret}-Messung dadurch zu verfälschen, dürfen die internen Spannungen bei der Charakterisierung nur während der Haltezeit variiert werden. Es muss sichergestellt werden, dass immer unter nominellen Bedingungen gelesen und geschrieben wird. Im Gegensatz zu den Spannungen V_{NWLL}, V_{BB} und V_{PL} ist die Spannung des Kondensators V_S zeitabhängig und kann nicht von außen aufgeprägt werden, d.h. kann nicht zur Charakterisierung genutzt werden. V_S nimmt mit der Zeit aufgrund der Entladung des Kondensators durch die zu untersuchenden Leckströme ab, welche ihrerseits wiederum durch die geringere Kondensatorspannung reduziert werden. Die Retentionzeit t_{Ret} ist gerade die Zeit, in der V_S von $V_S(t = 0) = V_{BLH}$ auf die minimale zum korrekten Lesen notwendige Spannung abgefallen ist.

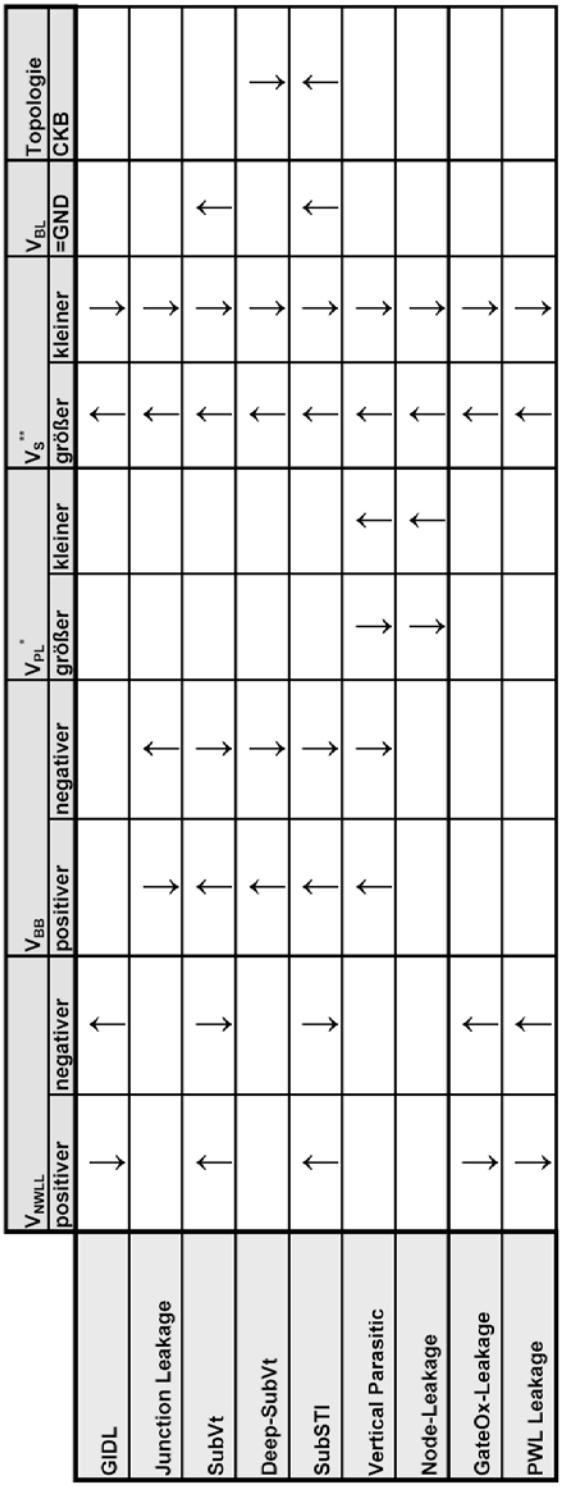


Abbildung 5.11: Spannungsabhängigkeiten der Leckstrompfade. Die Tabelle gibt die Leckstromänderung relativ zum Leckstrom bei Nominalbedingungen eines passiven „1"-Retentiontests an.

Kapitel 6

Charakterisierungsmethoden

Die Charakterisierung von Speicherzellen des Retentiontails stellt ganz besondere Anforderungen an die Messmethodik. Der Anteil der Tailzellen gemessen an den Speicherzellen eines modernen DRAM-Chips beträgt nur $\sim 10^{-5}$ bzw. $-4\,\sigma$ in Quantilen der Standardnormalverteilung (siehe Kapitel 3.1). Es muss also ein sehr kleiner Anteil der Speicherzellen gezielt untersucht werden können, während der größte Teil die Anforderungen um Größenordnungen übertrifft und von geringerem Interesse ist. Der Betrag der Gesamtleckströme liegt für den Großteil aller Zellen im Bereich weniger fA, der Leckstrom der Tailzellen jedoch $2-3$ Größenordnungen darüber. Trotz des relativ großen Leckstromunterschieds entstehen aufgrund der sehr geringen Auftrittswahrscheinlichkeit bei der Charakterisierung Schwierigkeiten. In diesem Kapitel sollen die bestehenden Standard-Charakterisierungsmöglichkeiten im DRAM-Prozess kurz erläutert werden. Abbildung 6.1 zeigt die verschiedenen Charakterisierungsebenen in der Prozessreihenfolge im Überblick. Die einzelnen Ebenen werden in den folgenden Abschnitten angesprochen und ihre Limitierungen hinsichtlich der Charakterisierung von Tailzellen werden diskutiert. In Abschnitt 6.4 wird als ein Ergebnis dieser Arbeit die Charakterisierung von einzelnen Zellen auf vollständig prozessierten Speicherbausteinen beschrieben. Dadurch ist es möglich die Probleme der Standardmethoden bei der Tailcharakterisierung zu umgehen und neue Informationen über die Leckstrommechanismen zu erhalten.

6.1 Messungen im Wafer-Kerf

Beim so genannten „Kerf" handelt es sich um den Waferbereich, der beim späteren Zersägen des Wafers zu Chips verloren geht (siehe Abbildung 6.2). Diese Fläche wird für Teststrukturen genutzt, die zur Extraktion verschiedener technologisch wichtiger Parameter dienen. Die Kerf-Strukturen können bereits nach der Strukturierung der ersten Metall-Lage vermessen werden und liefern daher schon sehr früh im Prozess wichtige

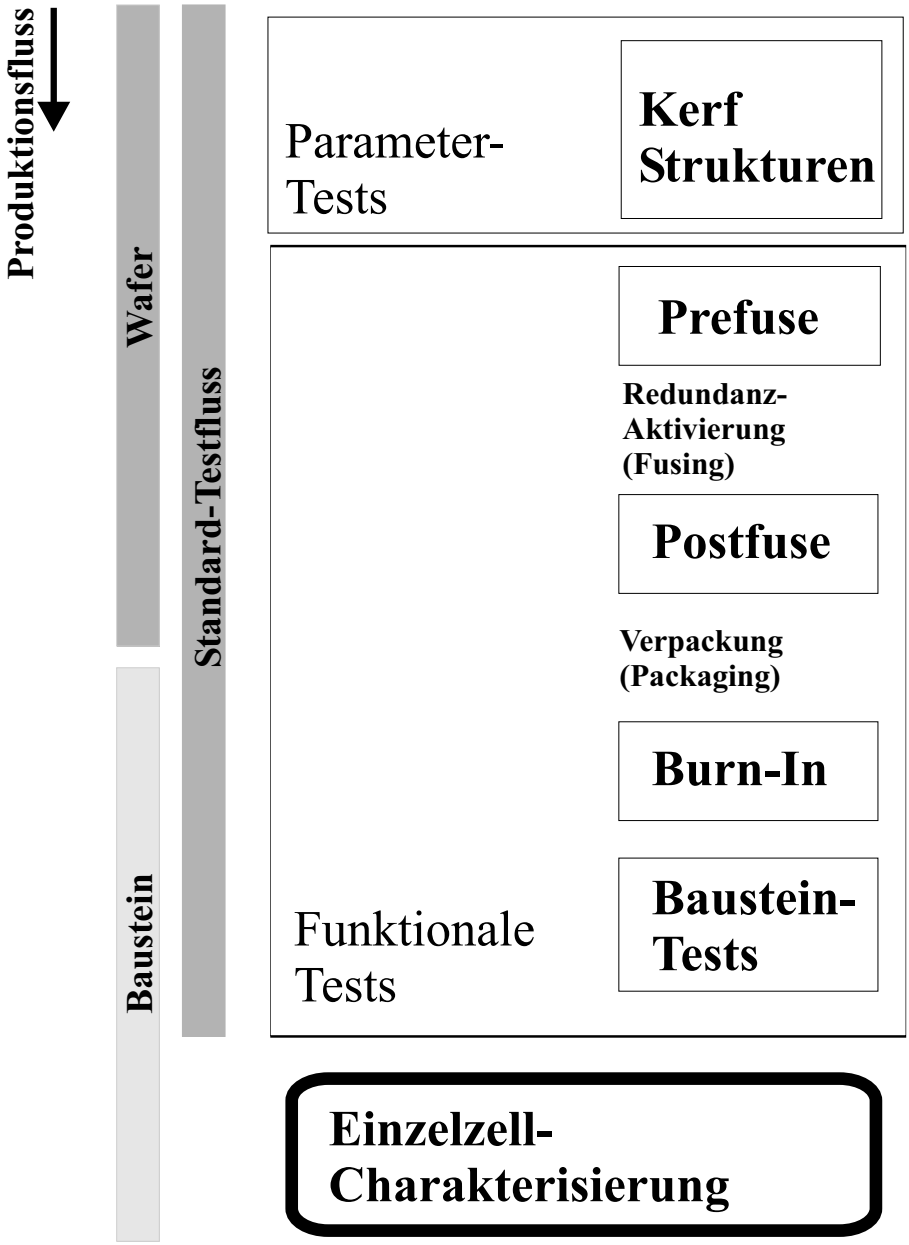


Abbildung 6.1: Testfluss von DRAM Speicherbausteinen im Verlauf des Produktionsprozesses. In dieser Arbeit werden mit einer speziell entwickelten Methode einzelne Zellen fertig aufgebauter Speicherbausteine elektrisch charakterisiert.

Daten, während die eigentlichen Chips noch nicht funktional sind. Auch stehen Kerf-Daten bei der Einführung neuer Technologien schon lange zur Verfügung, bevor überhaupt erste funktionelle Tests durchgeführt werden können. Dadurch können Parameter wie Schicht- und Kontaktwiderstände, Einsatzspannungen usw. kontrolliert und eingestellt werden. Über den Wafer werden mehrere (bis zu ca. 100) gleichartige Teststrukturen verteilt, um die Variation über den Wafer kontrollieren zu können. Teststrukturen können prinzipiell in die zwei Kategorien *Einzel- und Parallelstrukturen* unterteilt werden, die im Folgenden besprochen werden.

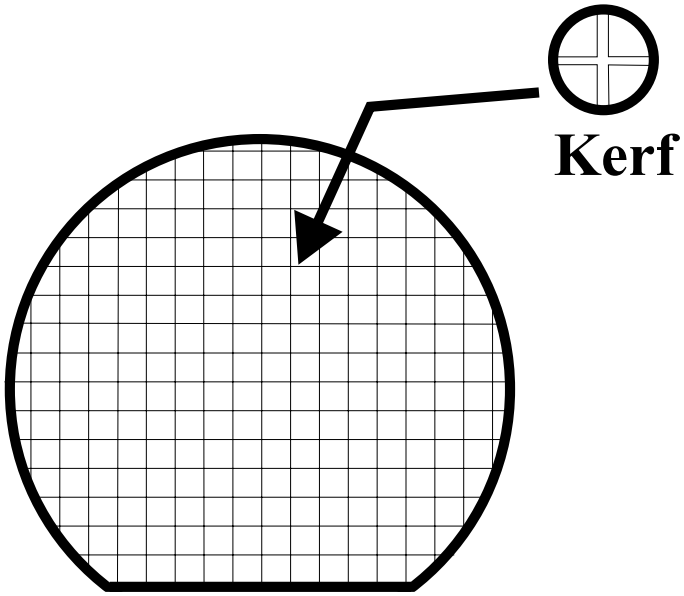


Abbildung 6.2: In den Sägeschlitzen (*Kerf*) eines Wafers werden Teststrukturen zur Parameterkontrolle integriert.

6.1.1 Einzel-Strukturen

Bei *Einzel-Strukturen* handelt es sich um Teststrukturen bestehend aus einzelnen Transistoren oder Kontakten, die mittels Zuleitungen und Kontaktpads in jeder Metallisierungsebene durch Nadeln kontaktiert und elektrisch vermessen werden können. Dadurch können Parameter wie z.B. V_t, S, I_{On} und Überlapp-Kapazitäten der verschiedenen Transistortypen eines Chips sowie Schicht- und Kontaktwiderstände bestimmt werden. Das I_{Off} einzelner Transistoren kann dabei im Allgemeinen nicht direkt gemessen werden, da dieses meist im sub-fA Bereich liegt. Deshalb wird der Off-Strom aus der Einsatzspannung V_t und der Unterschwellensteigung S berechnet und beinhaltet demnach nur *SubVt*-Ströme, dagegen bleiben Ströme in die p-Wanne (*GIDL*- und *Junction Leakage*) unberücksichtigt. Eine wichtige Rolle in Bezug auf Retention nimmt

die Teststruktur des Auswahltransistors ein. Leider ist in der *trench*-Zelle genau mit dieser Struktur ein Problem verbunden. Die *trench*-Zelle besitzt im Gegensatz zum *stacked*-Konzept keinen Node-seitigen Substratkontakt, der über die Metallisierung direkt zugänglich wäre. Bei der Auswahltransistor-Teststruktur muss deshalb die Node-Seite über einen länglichen Trench zum Nachbar-AA verbunden werden. Dadurch existiert inherent ein struktureller Unterschied zwischen Teststruktur im Kerf und dem Transistor im Speicherfeld, der zu Interpretationsbedarf beim Übertragen von Teststrukturmessungen auf reale DRAM-Zellen führt. Außerdem sind Einzel-Strukturen zur Charakterisierung der für Retentiontailzellen verantwortlichen Leckstrompfade und Mechanismen aus weiteren Gründen nicht geeignet. Die Erklärung dafür liegt in der sehr geringen Auftrittswahrscheinlichkeit von Tailzellen ($< 1 \cdot 10^{-5}$) und der damit vergleichsweisen geringen Anzahl von Teststrukturen pro Wafer. Unter der hypothetischen Annahme, dass Tailzellen sich von Mainzellen durch zu große Off-Ströme des Zelltransistors unterscheiden, folgt, dass 10^5 Einzeltransistor-Strukturen vermessen werden müssen, um einen Zelltransistor mit stark erhöhtem und zugleich direkt messbarem Off-Strom zu finden. Dieser könnte dann weiter charakterisiert werden. Bei üblicherweise ungefähr 100 Teststrukturen pro $300\,mm$ Wafer entsprechen die 10^5 Einzeltransistoren 1000 Wafern bzw. 40 Losen. In Anbetracht des dafür nötigen Zeitaufwandes und der endlichen Wahrscheinlichkeit von Kontaktproblemen bei der Messung sowie anderen Defekten der Struktur, erscheint dies ungeeignet. Deshalb können durch Einzeltransistor-Messungen immer nur Aussagen über durchschnittliche („Main"-Transistoren) gewonnen werden. Aus der Schwellenspannung V_t und der Unterschwellensteigung S berechnete Off-Ströme beinhalten lediglich den *SubVt*-Strom und vernachlässigen *GIDL* sowie *Junction Leakage.*

6.1.2 Parallel-Strukturen

Eine zweite Art von Kerf-Teststrukturen dient zur Defektkontrolle und zur Bestimmung gemittelter Werte für kleine Ströme und Kapazitäten, wie z.B. die Kapazität des Speicherkondensators oder der Bitleitungen. Dazu werden viele gleichartige Strukturen entweder parallel (Zellen) oder in Reihe (Kontakte) geschaltet. Geeignet sind solche Strukturen vor allem zur Bestimmung der Defektdichten für Kurzschlüsse und fehlenden Kontakten. Weshalb auch diese Strukturen zur Untersuchung von Tailzellen nicht geeignet sind, soll im Folgenden erläutert werden. Für eine Überschlagsrechnung wird dazu eine Tailwahrscheinlichkeit von $1 \cdot 10^{-5}$ und ein gegenüber Mainzellen um den Faktor 300 erhöhter Leckstrom für Tailzellen angenommen. Die Werte entsprechen der maximalen Wahrscheinlichkeit und Leckstromerhöhung, um in der Abschätzung den bestmöglichen Fall zu betrachten. Abbildung 6.3 zeigt den Einfluss der Arraygröße auf die Anzahl zur Tailzellencharakterisierung benötigter Strukturen und die maximale Gesamtstromreduktion bei hypothetischen Bedingungen, welche Tailzellen zu Mainzellen umwandeln, oh-

ne bestehende Main-Zellen in irgendeiner Form zu beeinflussen. Aufgrund der geringen Auftrittswahrscheinlichkeit von Tailzellen werden für kleine Arraygrößen deshalb viele Strukturen benötigt, um überhaupt Tailzellen enthaltende Strukturen zu finden. Im anderen Extremfall einer sehr großen Arraygröße sind zwar in nahezu jeder Struktur auch Tailzellen enthalten, jedoch ist deren Beitrag zum Gesamtstrom zu gering als dass eine potenzielle Umwandlung zu Mainzellen sicher im Gesamtleckstrom beobachtet werden könnte. Zur Charakterisierung von Tailzellen muss deshalb ein Kompromiss zwischen den zwei Extremfällen gefunden werden. Unter realistischen Randbedingungen von ungefähr 100 verfügbaren Strukturen pro Wafer und der Anforderung, dass sich der Gesamtleckstrom um mindestens 10% ändern soll, falls es gelingt die enthaltenen Tailzellen zu Mainzellen umzuwandeln, kann der Abbildung 6.3 eine optimale Array-Größe von 1000-3000 Zellen entnommen werden.

Die kleinsten verfügbaren Teststrukturen im Kerf der hier untersuchten Produkte haben eine Arraygröße von 10^6 Zellen und können daher nicht zur Charakterisierung herangezogen werden. Außerdem kann wie bei Einzel-Strukturen auch bei den Parallel-Strukturen der DT nicht direkt kontaktiert werden. Der für Retention interessante Fall einer gespeicherten physikalischen „1" mit geschlossenem Gate ist somit nicht zugänglich. Daraus folgt direkt, dass z.B. *GIDL* nicht an solchen Strukturen untersucht werden kann.

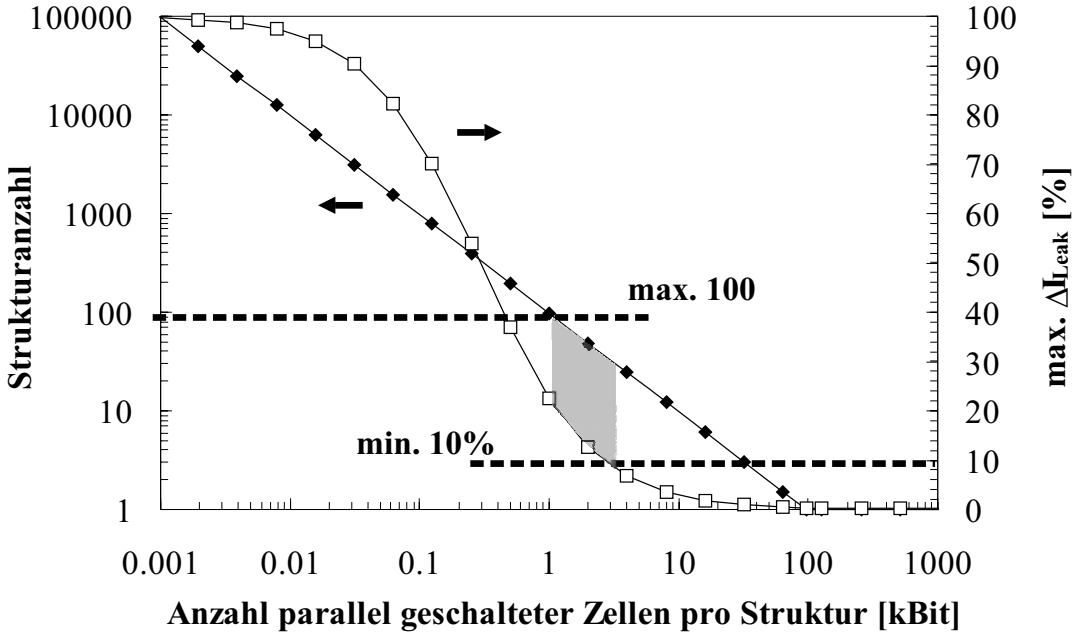


Abbildung 6.3: Linke Achse: Notwendige Strukturanzahl, um eine Struktur mit mindestens einer Tailzellen zu finden. Rechte Achse: maximaler Effekt auf den Gesamtleckstrom, falls enthaltene Tailzellen zu Mainzellen gemacht werden ohne den Leckstrom enthaltener Mainzellen zu beeinflussen. Zur Charakterisierung von Tailzellen ist eine Strukturgröße von 1000-3000 Zellen am Besten geeignet.

6.2 Wafer-Tests

6.2.1 Prefuse-Test

Der erste Test der Speicherbausteine selbst ist der so genannte *Prefuse*-Test. Dabei werden die Speicherchips direkt auf dem Wafer mittels Nadelkarten kontaktiert und betrieben. Nichtfunktionale Chips können in einer Reihe von Tests identifiziert und aus Kostengründen vom weiteren Aufbau zu Bausteinen ausgeschlossen werden. Zur Ausbeutesteigerung ist es darüber hinaus möglich eine gewisse Anzahl von defekten Speicherzellen in der nachfolgenden Redundanzaktivierung durch redundante Zellen zu ersetzen. Dies geschieht durch gezieltes Zerstören von Sicherungen (Fuses) mit Hilfe eines Laserstrahls. Die Retentionzeiten der Wafer-Tests liegen je nach Speicherhersteller und Speicherprodukt bei bis zu $400\,ms$. Die getesteten Retentionzeiten liegen über der Produktspezifikation, um einer Degradation durch den Komponentenaufbau sowie Temperaturungenauigkeiten beim Test vorzuhalten. Neben den nominellen werden auch leicht variierte Randbedingungen, wie z.B. erhöhte oder erniedrigte Versorgungsspannungen oder Temperaturen, getestet. Der genaue Testablauf ist sehr komplex, unterscheidet sich von Produkt zu Produkt und unterliegt ständiger Anpassung. Details können an dieser Stelle deshalb nicht gegeben werden. Die *Prefuse*-Tests sind für die Volumenproduktion optimiert, d.h. es müssen in möglichst kurzer Testzeit alle Zellen, die bei spezifizierten Bedingungen zu einem späteren Zeitpunkt zum Ausfall des Chips führen können, gefunden und falls möglich anschließend durch Redundanz ersetzt werden. Aufgrund der relativ hohen Testkosten wird für die Volumenproduktion so wenig wie möglich getestet und lediglich die Fuseinformation wird gespeichert.

6.2.2 Shmoo-Tests

Sind bei der Entwicklung neuer Produkte detaillierte Analysen notwendig, wird bei Wafer-Tests oft ein Parameter über einen größeren Bereich auch außerhalb den nominellen Betriebsbedingungen variiert und gegen die Anzahl der Ausfallzellen aufgetragen. Man nennt eine Reihe von Tests mit einem variierenden Parameter auch „Shmoo". Die Retentionkurve ist nichts anderes als ein Shmoo mit der Retentionzeit als Parameter (siehe z.B. Abbildung 3.1). Ein weiterer wichtiger Shmoo-Parameter ist die Gate-Spannung des Auswahltransistors im ausgeschalteten Zustand (V_{NWLL}). Diese muss für das Produkt so gewählt werden, dass alle Transistoren sicher abschalten, jedoch keine zu hohen elektrischen Felder im Gate/Drain Überlapp entstehen, die zu erhöhtem *GIDL* führen. Dementsprechend zeigt der V_{NWLL}-Shmoo zwei Teiläste, die als *GIDL*- und *SubVt*-Ast bezeichnet werden (siehe Abbildung 6.4). Für negativeres V_{NWLL} nimmt der *GIDL*-Leckstrom zu, während *SubVt* abnimmt. Zwischen den zwei Bereichen existiert ein Minimum der Fehlerzahl aufgrund der zwei gegenläufigen Leckstrommechanismen (siehe Kapitel 5).

Die Problematik bei Shmoo-Tests besteht darin, dass nur die Änderung der Fehlerzahl bei variierten Parametern dargestellt wird. Die Zelladressen sind dagegen unbekannt und deshalb kann anhand Shmoos z.B. nicht festgestellt werden, ob in den Bereichen des *GIDL*- und *SubVt*-Ast die selben oder unterschiedliche Zellen zur Fehlerzahl beitragen. Deshalb sind auch Shmoo-Tests zur Charakterisierung von Tailzellen nur bedingt geeignet.

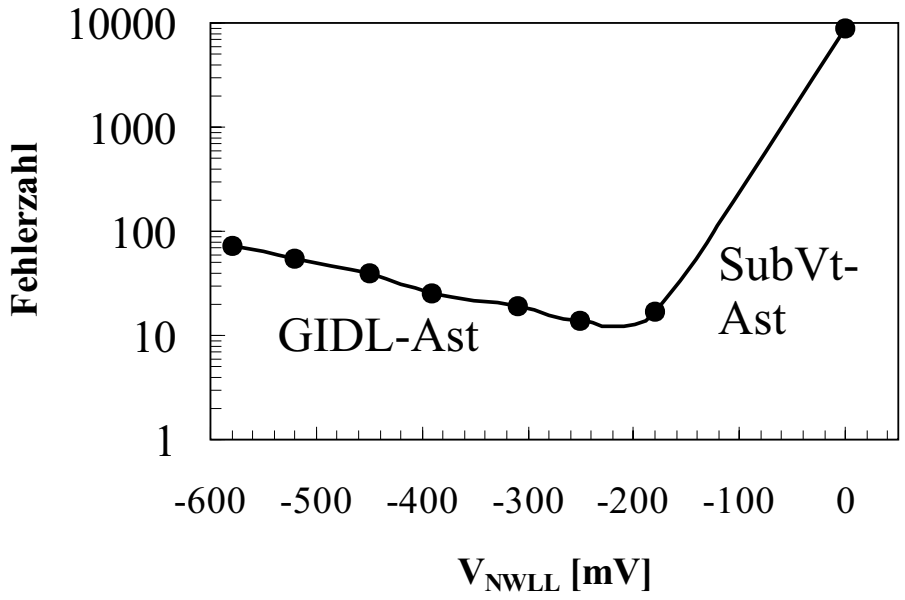


Abbildung 6.4: V_{NWLL}-Shmoo mit einer Retentionzeit in der Tailverteilung.

6.3 Baustein-Tests

Nach dem Aufbau zu Bausteinen durchläuft ein Speicherchip noch eine Reihe weiterer elektrischer Tests. Im so genannten *Burn-In* Test werden die Bausteine bei erhöhten Temperaturen und Spannungen betrieben. Aus der Literatur ist bekannt, dass die Degradation von Dielektrika dadurch beschleunigt werden kann (siehe z.B. [Wol95]). Weiterhin kann die typische Ausfallrate integrierter Schaltkreise in drei Bereiche unterteilt werden: *early failure*, *use*, *wearout*. Dabei ist die Ausfallrate besonders zu Beginn (early failure) und am Ende (wearout) der Lebenszeit besonders hoch. In der Zeit dazwischen (use) liegt sie konstant auf dem niedrigsten Niveau (siehe Abbildung 6.5). Ziel des *Burn-In* Tests ist es, die Speicherchips künstlich zu altern, um alle *early failure*-Ausfälle schon im Labor zu verwerfen und diese vom Kunden fern zu halten. Dabei muss ein Speicherbaustein bereits beim Ausfall einer einzigen Zelle verworfen werden, da diese im Allgemeinen nicht mehr ausgetauscht werden kann. Deshalb gehen manche Speicherpro-

duzenten dazu über, zusätzlich zur Redundanzaktivierung per Laser, einzelne elektrisch programmierbare Sicherungen (*e-fuses*) zu integrieren, um auch auf Bausteinbasis noch einzelne Zellen austauschen zu können. Zusätzlich zu den *early failures* aufgrund von Defekten im Oxid, werden in den Baustein-Tests noch weitere Fehlerklassen aussortiert. Das Spektrum reicht von rein mechanischen Defekten durch den Verpackungsprozesses über degradierte Kontakte, Veränderungen in den Metallisierungsebenen bis hin zu späten Retentionausfällen. Die getesteten Retentionzeiten der Baustein-Tests sind kleiner als bei Prefuse und damit dichter an der Produktspezifikation.

Anders als bei den *Prefuse*-Tests ist das Ergebnis der Baustein-Tests keine Fehlerzahl pro Chip, sondern es gibt nur noch die beiden Zustände: *Chip Fail* oder *Chip Pass*. Um Verpackungskosten zu sparen, wird versucht mögliche Bausteinausfälle bereits bei der Prefuse-Messung zu erkennen und auszusortieren (*Screening*). Idealerweise sind die Fehlerraten bei den Baustein-Tests dadurch relativ gering und für statistisch relevante Bewertungen von Fertigungsversuchen werden deshalb große Mengen von Speicherbausteinen benötigt.

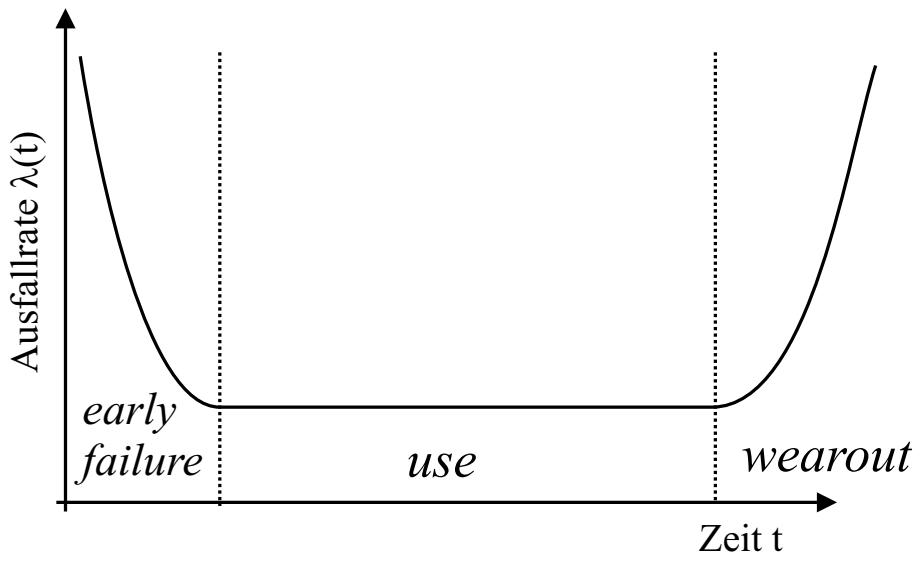


Abbildung 6.5: Typische Zeitabhängigkeit der Ausfallrate von integrierten Schaltungen.

6.4 Einzelzell-Analysen

Da alle in den vorherigen Abschnitten beschriebenen Verfahren zur Charakterisierung von Tailzellen weniger geeignet sind, musste als Teil dieser Arbeit eine Möglichkeit gefunden werden, um die genannten Schwierigkeiten zu umgehen. Dazu wurden Einzelzellen vollständig aufgebauter Speicherbausteine auf Tail-Mechanismen hin untersucht. Besonders die Bestimmung der Temperaturabhängigkeit gibt Aufschluss über die vorherrschenden Leckstrommechanismen (mehr dazu in den folgenden Kapiteln). Die ge-

ringe Auftrittswahrscheinlichkeit von Tailzellen stellt bei dieser Methode keine Limitierung dar, da diese aus den $512\,MBit$ Speicherzellen eines Bausteins gezielt ausgewählt werden können. Dadurch ist deren Adresse bekannt und durch die sehr genaue Messung der Retentionzeiten können auch Leckströme im fA-Bereich akkurat bestimmt werden. Ein Nachteil der Einzelzell-Charakterisierung ist die komplexe Messtechnik und die damit verbundenen langen Messzeiten. Die Methode kann prinzipiell auch anhand Speicherchips auf Wafern durchgeführt werden, die Charakterisierung von Speicherbausteinen bringt hinsichtlich der Messtechnik jedoch erhebliche Vorteile mit sich. So entstehen bei den nötigen Langzeitmessungen erheblich weniger Kontaktprobleme und darüber hinaus kann für einzelne Bausteine deren Temperatur exakter kontrolliert werden. Einzelzell-Analysen zählen nicht zu den Standardmethoden einer Produktionsumgebung und mussten inklusiv aller Mess- und Auswerteprogramme speziell für diese Arbeit entwickelt werden.

Die untersuchten Speicherbausteine durchliefen vor der Einzelzellcharakterisierung den kompletten Produktionsprozess sowie alle Produktionstests. Dadurch können nicht Retention bedingte Fehlerklassen bereits ausgeschlossen werden. Der Teil des Retentiontails unterhalb der Reparaturgrenze wurde im Produktionsprozess durch Redundanz ersetzt und steht somit nicht mehr zur Verfügung. Die Charakterisierung erfolgt am verbleibenden Retentiontail. Es kann davon ausgegangen werden, dass sich die physikalischen Ursachen der verbleibenden Tailzellen nicht grundlegend von den durch Redundanz ersetzten unterscheidet, da sich auch die Retentionverteilungen von Chips auf Wafern und fertigen Speicherbausteinen sonst nicht unterscheiden.

Im Folgenden soll zunächst die Methode der Einzelzell-Analysen beschrieben werden. Im Prinzip kann diese auf allen Standard-Speichertestern implementiert werden. Das speziell in dieser Arbeit verwendete System unter Verwendung eines MOSAID 3480 Speichertesters wird im anschließenden Abschnitt 6.4.2 genauer beschrieben.

6.4.1 Charakterisierungsmethode

Im Folgenden soll die Methode genauer erläutert werden. In Abbildung 6.6 sind die drei Teilschritte der in dieser Arbeit etablierten Charakterisierungstechnik aufgelistet. In einem ersten Schritt werden die Adressen der für die Charakterisierung interessanten Zellen eines Speicherbausteins ermittelt. Danach erfolgt die Messung der Retentionzeiten unter verschiedenen Spannungs- und Temperaturbedingungen, gefolgt von der Berechnung der Aktivierungsenergien aus der Temperaturabhängigkeit. Die einzelnen Charakterisierungsschritte werden im Folgenden einzeln betrachtet werden.

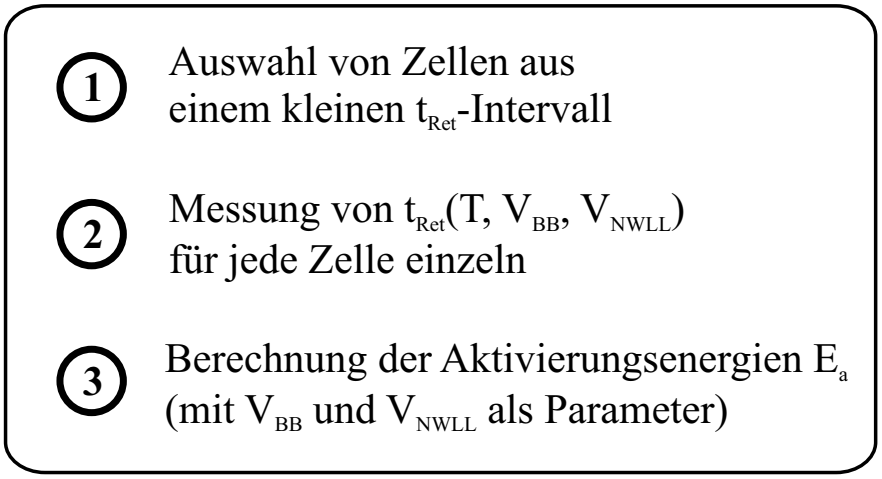


Abbildung 6.6: Prinzip der Analyse von Einzelzellen.

1. Auslesen von Zelladressen

Im ersten Schritt werden die für die genauere Charakterisierung interessanten Zellen identifiziert. Dazu werden zwei Retentiontests mit geringfügig unterschiedlichen Retentionzeiten um den gewünschten Punkt der Retentionkurve herum durchgeführt. Die physikalischen Fehleradressen, bestehend aus X-Adresse, Y-Adresse und Daten-Pin (DQ), werden aus dem Fehlerspeicher für beide Tests getrennt extrahiert. Eine Excel-Arbeitsmappe erstellt durch Vergleich der extrahierten Adressen eine Liste der im Intervall ausgefallenen Zellen und schreibt diese in eine einfache Textdatei, welche im nächsten Schritt als Eingabedatei dient.

2. Messung der Einzelzell-Retentionzeiten

Im zweiten Schritt werden nacheinander die Retentionzeiten der in der Eingabedatei aufgelisteten Zellen ermittelt. Darüber hinaus kann eine Liste von Temperaturen und Testbedingungen angegeben werden. Zur Automatisierung der langen Messreihen dient ein in C++ geschriebenes Programm, welches als *Dynamic Link Library* (DLL) in die Testersoftware eingebunden wird. Eine manuelle Messung aller Einzelwerte ist zu aufwendig. Je nach Zellanzahl, gemessener Temperaturen und Testbedingungen kann die Messzeit pro Speicherchip mehrere Tage erfordern.

Abbildung 6.7 zeigt den schematischen Ablauf des Messprogramms. In der äußersten Programmschleife wird die Temperatur gesteuert, da dies der langsamste Prozess ist und über eine lange Zeit während der Einzelzellmessungen möglichst konstant gehalten werden muss. Nach Temperaturwechsel muss eine zusätzliche Wartezeit (*soak time*)

eingehalten werden, bis sich das gesamte Testsystem im thermischen Gleichgewicht befindet. Der Grund dafür ist vor allem die thermische Kapazität des Testers. Die minimal notwendige Wartezeit wurde durch wiederholte Messung der Retentionzeit einer Einzelzelle zu 10 min bestimmt und in allen weiteren Messungen eingehalten. Erst nach Ablauf dieser Wartezeit wird mit der eigentlichen Messung begonnen. In der zweiten Schleife werden die Einzelzellen der Liste sukzessive abgearbeitet. Zur Reduzierung der Testzeit wird das getestete Speicherfeld auf 1024 WLs * 512 BLs um die untersuchte Speicherzelle herum eingeschränkt. Dadurch sind die Nachbarschaftsbeziehungen immer noch gewährleistet, während die Testzeit jedoch durch Verkürzung der Gesamtschreib- und Lesezeiten dramatisch reduziert wird. Die Reduktion der Arraygröße wirkt sich nicht messbar auf die Retentionzeiten selbst aus. Die dritte Programmschleife legt verschiedene Randbedingungen, wie Spannungen oder Daten-Topologien fest. Der eigentliche Kern des Messprogramms ist die Messung der Retentionzeit selbst. Da man beim Testen von Speicherbausteinen nur Pass/Fail Informationen erhält, basiert die Messung der Retentionzeit auf einem Intervallhalbierungsverfahren. Dabei muss der Retentiontest sukzessiv mit verschiedenen Retentionzeiten wiederholt werden, bis der Pass/Fail Übergang mit der geforderten Genauigkeit gefunden wurde. Abbildung 6.8 zeigt den Testalgorithmus zur Bestimmung der Retentionzeit schematisch.

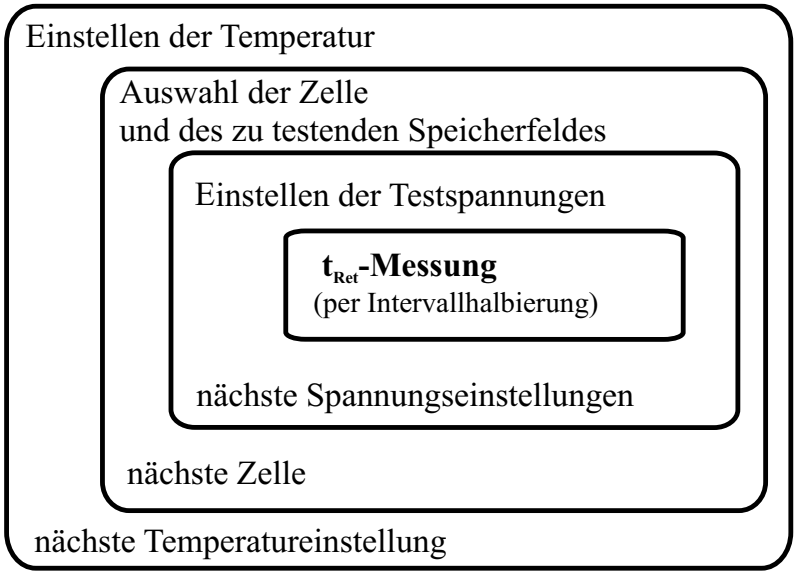


Abbildung 6.7: Schematischer Programmablauf der Einzelzellcharakterisierung

Zunächst wird eine physikalische „0" ins ganze Speicherfeld geschrieben, um für alle Zellen den gleichen Ausgangszustand herzustellen. Danach folgt das Schreiben einer physikalischen „1". Nach dem Schreiben wird ein schneller ROR-Refresh bestehend aus den Speicherbefehlen *Activate* und *Precharge* durchgeführt, um alle Zellen innerhalb eines kleinen Zeitfensters erneut voll aufzuladen und vor der Retentionpause den gleichen Ladungszustand zu garantieren. Kurz vor der eigentlichen Retentionpause wer-

den optional die internen Spannungen getrimmt. Bei Speicherbausteinen geschieht dies durch Einstellen der On-Chip Spannungsgeneratoren über Testmodes, da die internen Betriebsspannungen nicht mehr von außen angelegt werden können. Dadurch ist die Spannung nicht frei wählbar, sondern es steht nur eine gewisse Anzahl von im Design implementierten Trim-Stufen zur Verfügung. Anschließend kommt die eigentliche Retentionpause, während der die Bitleitungen für passive Tests auf $V_{BLH}/2$ gehalten werden, die Wortleitungen auf V_{NWLL} liegen und kein Zugriff auf die Speichermatrix erfolgt. Für *SubVt*-Tests können die Bitleitungen per Testmode auf GND geschaltet werden, um zusätzlichen Source/Drain-Stress zu erzeugen. Vor dem Auslesen der Einzelzelle werden die Standardspannungsbedingungen wieder hergestellt und erneut ein schneller ROR-Refresh durchgeführt. Dabei findet die eigentliche Bewertung für alle Zellen innerhalb eines kurzen Zeitintervalls statt. Bei Retentiontests mit verschiedenen internen Spannungen ist darauf zu achten, dass diese nur während der Retentionpause vom Nominalwert abweichen, ansonsten wird der Schreib- bzw. Lesevorgang beeinflusst und das Ergebnis verfälscht. Bei den V_{NWLL}-Tests ist diese Bedingung automatisch erfüllt, da V_{NWLL} nur bei geschlossenem Transistor, also in der Retentionpause, anliegt. Die bestimmten Retentionzeiten werden zusammen mit den Zelladressen und den Testbedingungen in ein Ergebnis-File geschrieben.

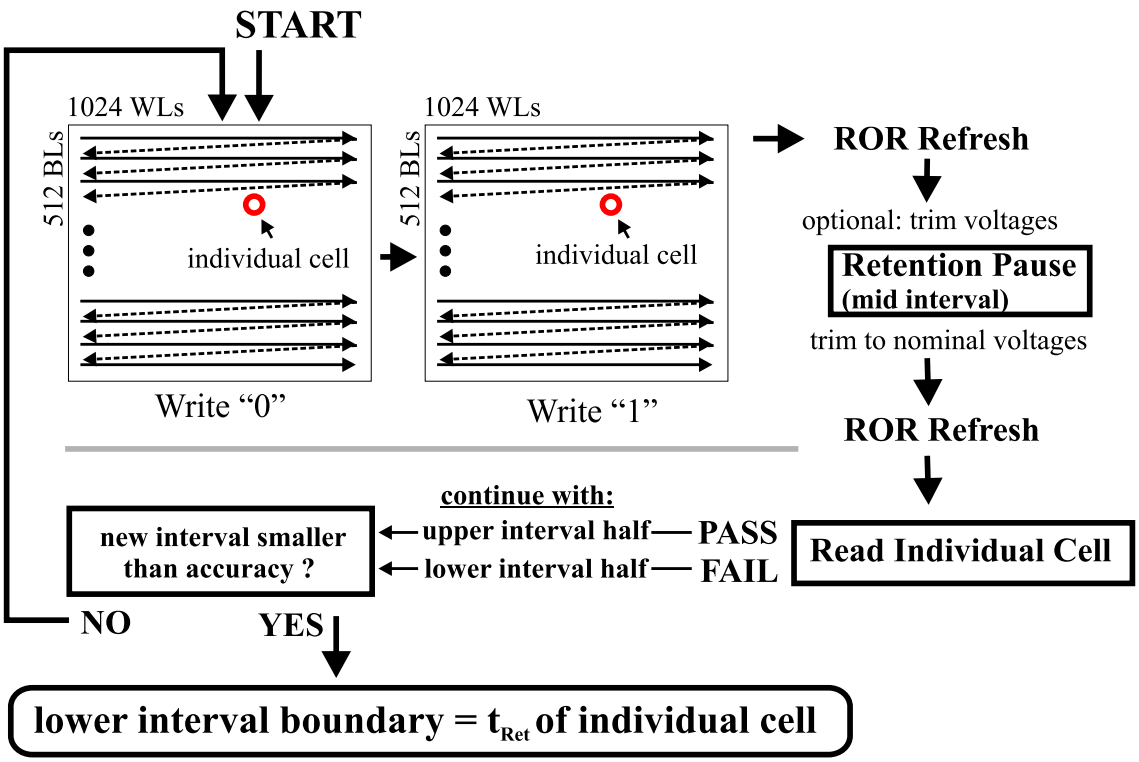


Abbildung 6.8: Intervallhalbierung zur Messung der Retentionzeit einer Einzelzelle direkt auf einem Speicherbaustein (Abbildung aus [Web06a]).

3. Berechnung der Aktivierungsenergien

Das aus den Messungen im zweiten Schritt erhaltene Ergebnis-File wird von einer Excel-Arbeitsmappe eingelesen. Für jede untersuchte Zelle und Spannungsbedingung wird unter Ausnutzung der gemessenen Temperaturabhängigkeit der Retentionzeit die Aktivierungsenergie E_a durch einen exponentiellen Fit gemäß Gleichung 6.1 an die Daten ermittelt. E_a stellt dabei ein Maß für die Temperaturabhängigkeit dar. Ein großer Wert für E_a steht für eine hohe, während ein kleiner Wert für eine geringe Temperaturabhängigkeit steht. Abbildung 6.9 zeigt die Retentionzeit gegenüber der Temperatur für zwei Beispielzellen im so genannten Arrhenius-Plot, bei dem eine Abhängigkeit gemäß Gleichung 6.1 als Gerade wiedergegeben wird. Die Datenpunkte liegen für beide Zellen jeweils auf einer Geraden, wodurch eine exponentielle Temperaturabhängigkeit entsprechend dem Arrhenius-Gesetz (Gleichung 6.1) hervorgeht. Da der Leckstrom I_{Leak} in erster Ordnung inversproportional zur Retentionzeit ist, besitzt dieser die gleiche Aktivierungsenergie (Gleichung 6.2).

$$t_{Ret} = const \cdot exp(\frac{E_a}{k_B T}) \tag{6.1}$$

$$I_{Leak} \propto \frac{1}{t_{Ret}} \propto const' \cdot exp(-\frac{E_a}{k_B T}) \tag{6.2}$$

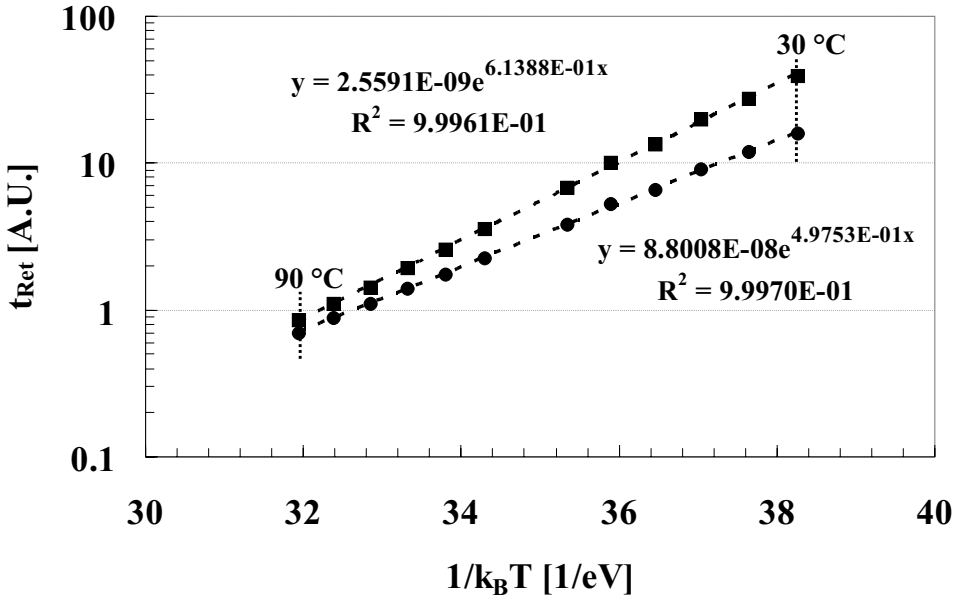


Abbildung 6.9: Temperaturabhängigkeit der Retentionzeit für zwei Einzelzellen aus dem Tailbereich der Retentionkurve. Über den gesamten Temperaturbereich von $30\,°C$ bis $90\,°C$ liegen alle Datenpunkte einer Zelle im Arrhenius-Plot auf einer Geraden.

Als Ergebnis der Einzelzellcharakterisierung sind nun für alle Zellen die genaue Position in der Retentionverteilung, die Spannungssensitivitäten der Retentionzeiten sowie die Aktivierungsenergie und deren Spannungsabhängigkeiten bekannt. Durch Vergleich mit theoretischen Überlegungen (siehe dazu Kapitel 8) können dadurch Aussagen über den Leckstrompfad und dessen Hauptmechanismen gemacht werden .

6.4.2 Messapparatur

In diesem Abschnitt wird das verwendete Testsystem für die Einzelzell-Analysen kurz vorgestellt. Es ist keine spezielle Hardware erforderlich, sodass die Methode prinzipiell auch für andere Speichertester implementiert werden kann.

Speichertester

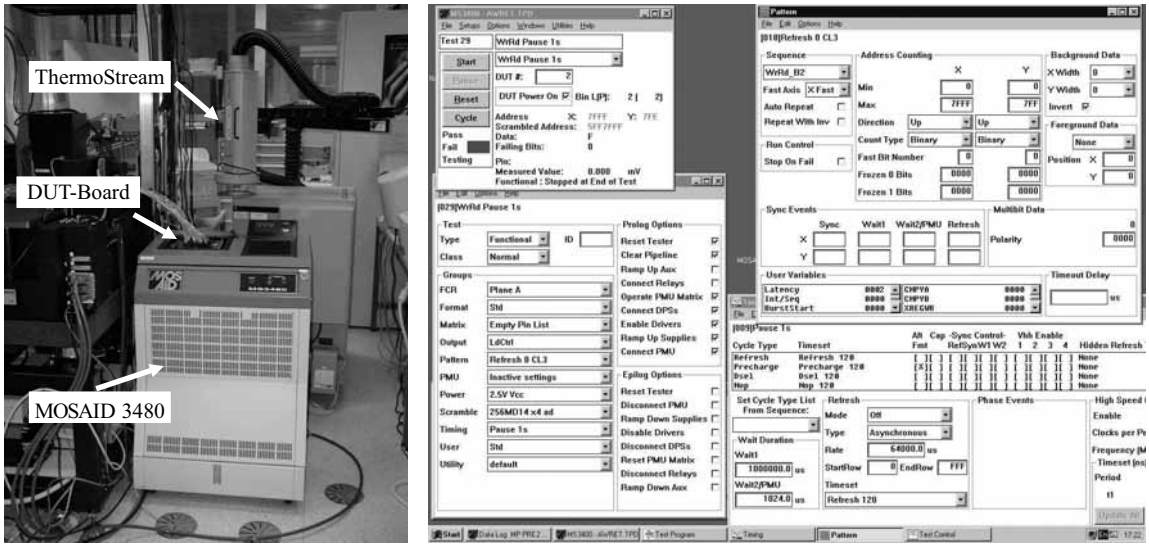
Für die Messungen an fertig aufgebauten DRAM Speicherbausteinen wurde ein MOSAID 3480 Speichertester herangezogen (Abbildung 6.10a). Der Tester wird von einer Windows basierten Software kontrolliert (Abbildung 6.10b), die über in C++ programmierte *Dynamic Link Libraries* (DLLs), so genannte *Userexits*, erweiterte Testfunktionalität zulässt. Die detaillierte Beschreibung des notwendigen Setups zur Speicheransteuerung am MOSAID 3480 geht über den Umfang dieser Arbeit hinaus und wird daher hier nicht weiter angesprochen. Das experimentelle Setup zur Messung von Einzelzellen an Komponenten ist anspruchsvoll und detailreich. Für Details zur Speicheransteuerung wird auf die DRAM Datenblätter bzw. die umfangreichen Bedienungsanleitungen des MOSAID Speichertesters verwiesen.

Bausteine und DUT-Board

Abbildung 6.11 zeigt die Pinbelegung eines DDR-DRAMs für die zwei aktuell gängigsten Gehäuse TSOP66 (*Thin Small Outline Package*) und FBGA (*Fine Ball Grid Array*). In der vorliegenden Arbeit wurden Bausteine beider Bauarten untersucht. Dazu musste speziell für die FBGA Bausteine ein DUT-Board, das den Baustein aufnimmt und die Kontakte zum Tester herstellt, gefertigt werden. Über den Baustein wird der „Rüssel" des ThermoStream (Temptronics) gestülpt, durch welchen Luft (ca. $4\,L/s$) mit definierter Temperatur zugeführt wird. Dabei verhindern spezielle Isoliermatten, dass der Luftstrom unkontrolliert seitlich entweichen kann.

Temperatursteuerung

Der ThermoStream (Temptronics) wird über ein GPIB-Interface von der in die Testersoftware integrierten DLL angesteuert. Bei Temperaturwechsel meldet dieser das Erreichen



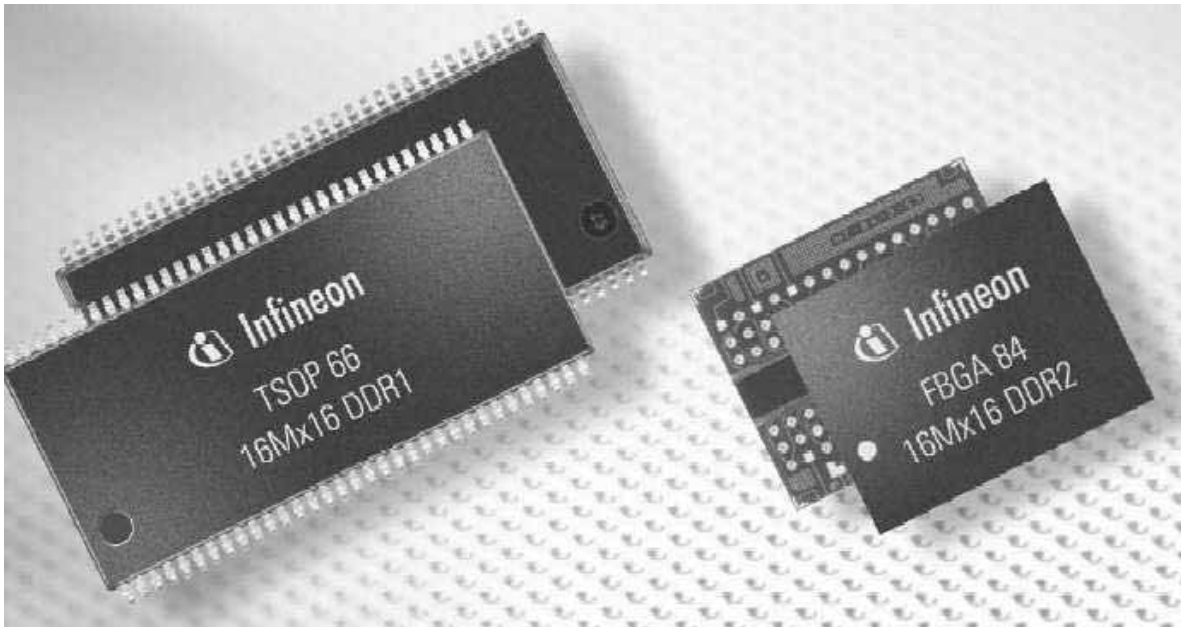
(a) (b)

Abbildung 6.10: (a) MOSAID 3480 Speichertester. Das DUT-Board nimmt den Speicherbaustein auf. Über dieses wird der „Rüssel" des ThermoStream gestülpt. (b) Beispielhaft vier der vielen Fenster der Tester-Software.

der Temperatur an die Software zurück, welche vor Start der Messung die eingestellte Wartezeit (*soak time*) verstreichen lässt, damit sich das gesamte System bestehend aus Tester und ThermoStream thermisch stabilisieren kann. Die minimal notwendige Wartezeit wurde durch wiederholte Messung der Retentionzeit einer Einzelzelle zu 10 min bestimmt und in allen Messungen eingehalten.

Messgenauigkeit & Temperaturstabilität

Bei der Messung der Retentionzeit von Einzelzellen wurde das Abbruchkriterium des Intervallhalbierungsverfahren auf $1\,ms$ gesetzt, sodass der dadurch entstehende systematische Fehler gegenüber den gemessenen Retentionzeiten vernachlässigt werden kann. Durch die Wahl eines reduzierten Speicherfeldes von lediglich 1024 WLs um die untersuchte Speicherzelle herum, wurde das Zeitintervall, in dem alle Zellen vor der Retentionpause aufgefrischt bzw. nach der Pause bewertet werden, auf die Zeit des dazu verwendeten ROR-Refreshes von $1024 \cdot 240\,ns \approx 0.25\,ms$ eingeschränkt. Da der ROR-Refresh alle WLs jeweils in der gleichen Reihenfolge aktiviert, ist der tatsächliche Fehler noch kleiner und kann ebenfalls gegenüber den gemessenen Retentionzeiten vernachlässigt werden. Die verbleibende Hauptfehlerquelle liegt in der Temperaturstabilität zwischen verschiedenen Messungen. Um den Fehler bei der Temperatureinstellung zu minimieren, wurde dem System jeweils genügend Zeit gegeben (*soak time*), um ein thermisches Gleichgewicht zu erreichen. Die dafür benötigte Zeit betrug weniger als 10 Minuten. Zur Abschätzung der durch die Temperaturregelung verursachten Schwankungen nach dem



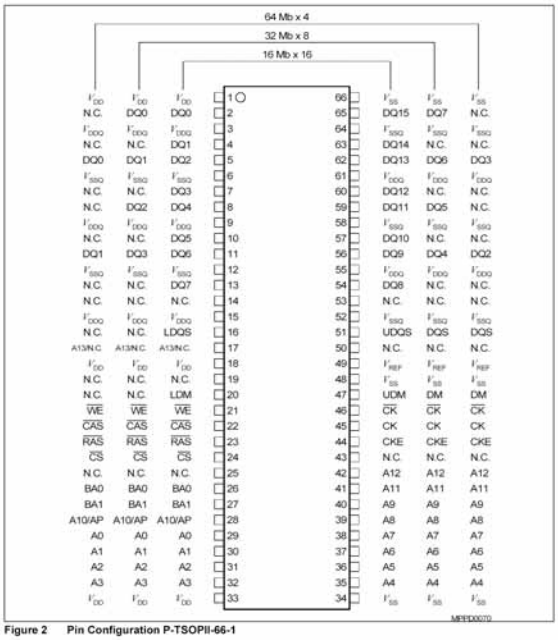


Figure 2 Pin Configuration P-TSOPII-66-1

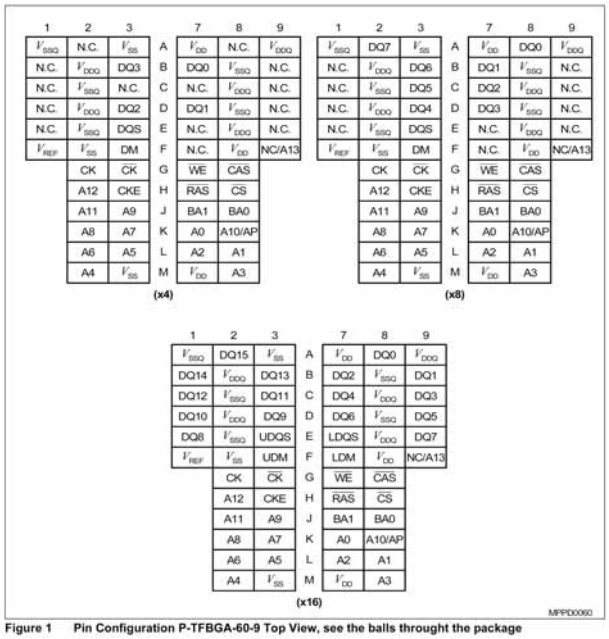


Figure 1 Pin Configuration P-TFBGA-60-9 Top View, see the balls throught the package

(a) (b)

Abbildung 6.11: Pin-Belegung von DDR DRAMs. (a) TSOP66 (*Thin Small Outline Package*) Gehäuse, (b) FBGA (*Fine Ball Grid Array*) Gehäuse.

Erreichen des Gleichgewichts und deren Auswirkungen auf die Retentionmesswerte und die daraus berechneten Aktivierungsenergien, wurde die Retentionzeit einer Einzelzelle aus dem Tailbereich bei verschiedenen Temperaturen jeweils 1000 Mal in einer langen Testreihe bestimmt (siehe Abbildung 6.12). Die dafür nötige Messzeit betrug insgesamt mehr als 24 Stunden, sodass auch langsame Temperaturschwankungen erkannt werden können. Da die Retentionzeit exponentiell von der Temperatur abhängt, stellt diese eine sehr empfindliche Messgröße zur Beurteilung der Temperaturstabilität dar. In Abbildung 6.12 wurden neben der jeweils am ThermoStream eingestellten Temperatur der Mittelwert und die Standardabweichung der Retentionmesswerte angegeben. Demzufolge kann die Retentionzeit mit einer Standardabweichung von nur 0.8% bestimmt werden. Zur Abschätzung der daraus resultierenden Ungenauigkeit in der Aktivierungsenergie wurde diese aus den Werten der Messreihe 1000 Mal berechnet. Für die betrachtete Zelle ergibt sich eine Aktivierungsenergie von $0.3676 \pm 0.0036\,eV$. Dies entspricht einem $1\,\sigma$-Fehler der Aktivierungsenergie von ungefähr 1%. Bei der nun bestimmten Aktivierungsenergie entspricht der beobachtete Retentionmessfehler einer maximalen Temperaturschwankung von $\pm 0.24\,°C$.

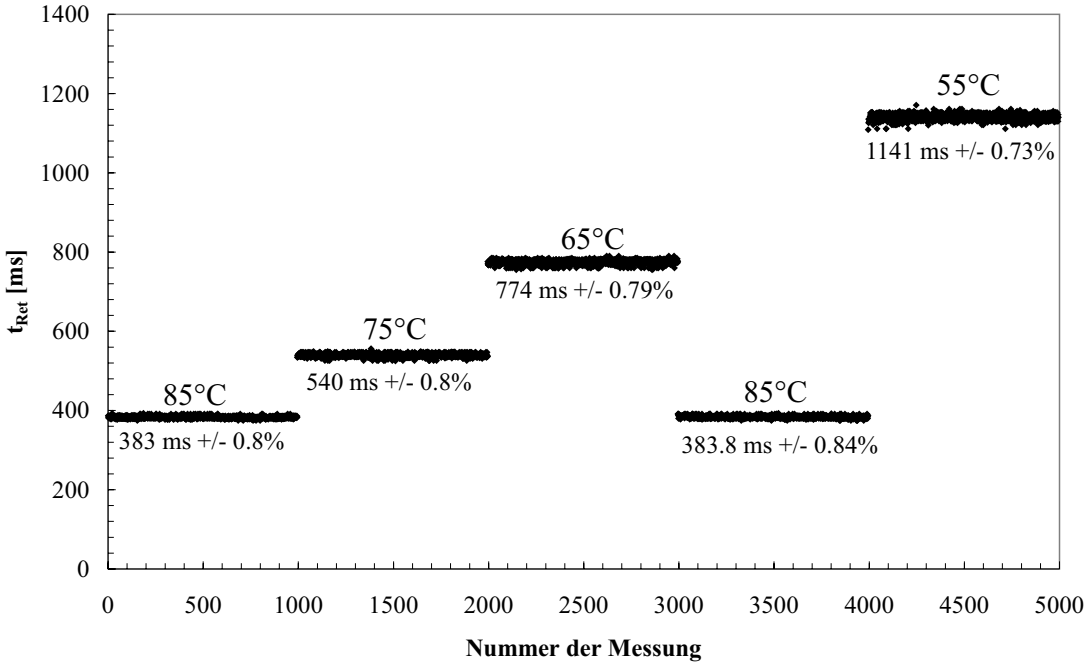


Abbildung 6.12: Zur Bewertung der Messgenauigkeit wurde die Retentionzeit einer Einzelzelle aus dem Tailbereich für verschiedene Temperaturen jeweils 1000 Mal bestimmt. Diese Messreihe benötigte mehr als 24 Stunden. Neben der jeweiligen Temperatur sind der Mittelwert und die Standardabweichung für jede Temperaturstufe angegeben.

6.5 Zusammenfassung

Die Retentiontail-Charakterisierung anhand Standard-Produktionstests ist mit Schwierigkeiten verbunden, die durch die sehr geringe Auftrittswahrscheinlichkeit von Tailzellen bedingt sind. An Einzel-Strukturen können nur Zellen charakterisiert werden, die dem Main der Verteilung entsprechen, da die Wahrscheinlichkeit für das vorliegen einer Tailzelle viel zu gering ist. Bei Parallel-Strukturen gibt es abhängig von der Anzahl parallel geschalteter Zellen einen kleinen Bereich zwischen ungefähr 1000 und 3000 Zellen, in dem Tailzellen eingeschränkt charakterisiert werden können. Bei größeren parallelen Arrays wird der Tailleckstrom durch die vielen enthaltenen Mainzellen derart „verdünnt", dass kein messbarer Effekt im Gesamtleckstrom zu beobachten ist. Für künftige Technologien mit noch höheren Speicherdichten und noch geringerem Tailanteil, wird der Bereich in dem eine Charakterisierung prinzipiell möglich ist zunehmend kleiner werden. Die Standard Wafer-Tests sind zur Charakterisierung ebenfalls nicht geeignet, da zwar die Fehlerzahl jedoch nicht die Fehleradressen bekannt sind, wodurch wichtige Informationen verloren gehen. Die in dieser Arbeit entwickelte und hier beschriebene Methode der Einzelzellcharakterisierung ist in der Lage, alle genannten Schwierigkeiten zu umgehen. Besonders hervorzuheben ist, dass die Methode unabhängig vom Zahlenverhältnis zwischen Main-und Tailzellen ist und deshalb auch für alle zukünftigen Technologien eingesetzt werden kann.

Kapitel 7

Ergebnisse zur elektrischen Charakterisierung

Ein Ergebnis aus Kapitel 4 war, dass die Retentionverteilung hauptsächlich durch die Leckströme bestimmt wird. Nachdem in Kapitel 5 die Leckstrompfade im DRAM vorgestellt und im letzten Kapitel die Messmethodik erläutert wurde, folgt in diesem Kapitel die elektrische Charakterisierung mit der Zielsetzung die dominierenden Leckstrompfade der Tail- und Mainverteilung zu bestimmen. Dazu wird im ersten Abschnitt das Verhalten der Retentionverteilung unter Temperatur- und Spannungsvariationen untersucht. Die Charakterisierung durch Einzelzellmessungen im zweiten Teil erlaubt detaillierter in die Zelle zu sehen und zusätzliche Informationen über zugrundeliegende Mechanismen zu gewinnen.

7.1 Retentionkurven

Die einfachste Art der Retentioncharakterisierung besteht in der Messung der Spannungs- und Temperaturabhängigkeiten. Durch Vergleich mit den in Kapitel 5 zusammengefassten Spannungsabhängigkeiten der Leckströme können Rückschlüsse auf die dominanten Leckstrompfade gemacht werden. Alle Untersuchungen in diesem Abschnitt wurden an fertig aufgebauten Speicherkomponenten durchgeführt, um den Vorteil geringerer Kontaktprobleme bei Komponenten gegenüber Wafermessungen für die zeitlich sehr langen Messreihen zu nutzen. Die an Bausteinen gemessenen Retentionkurven unterscheiden sich von Wafermessungen dadurch, dass alle Zellen mit Retentionzeiten kleiner der Reparaturzeit durch Redundanz ersetzt wurden und deshalb die ersten Fehler erst oberhalb der Reparaturgrenze auftreten.

7.1.1 „0"- und „1"-Retention

Wenn bisher von Retention gesprochen wurde, dann war damit stets die Haltezeit einer gespeicherten „1" gemeint. Auch in der Literatur wird meist nur vom Speicherzustand „1" gesprochen. Jedoch kann auch eine gespeicherte „0" durch Leckströme verloren gehen. Für grundlegende Untersuchungen müssen daher beide Fälle betrachtet werden. Abbildung 7.1 zeigt die Retentionkurven bei $85\,°C$ für beide Speicherzustände zum Vergleich in einem Diagramm. Die beiden Retentionkurven unterscheiden sich deutlich voneinander. Erstens liegt die Fehlerzahl der „1" zu allen Retentionzeiten deutlich über der Fehlerzahl der „0". Zweitens treten „0"-Fehler erst bei im Vergleich zur „1" deutlich höheren Retentionzeiten auf. Da diese bereits deutlich über der Reparaturgrenze liegen, ist die „0" für die Chipausbeute nicht limitierend und rechtfertigt die Vorgehensweise in Standard-Produktionsmessungen nur die „1" zu betrachten. Der Unterschied in den beiden Retentionkurven kann dadurch erklärt werden, dass nicht alle Leckstrompfade gleichermaßen zum Datenverlust beider Zustände führen. In Kapitel 5 wurden die Pfade in die zwei Klassen symmetrisch und asymmetrisch eingeteilt. Die asymmetrischen Leckströme, die aus den pn-Leckströmen *Junction Leakage* und *GIDL* bestehen, bewirken eine Erniedrigung des in der Speicherzelle gespeicherten Potenzials. Für eine gespeicherte „1" bedeutet dies eine Verschlechterung der gespeicherten Information, während dies für eine „0" eine Verbesserung darstellt. Die symmetrischen Leckstrompfade *Vertical Parasitic*, *Node Leakage*, *SubVt*, *DeepSubVt*, und *SubSTI* bewirken für beide Zustände eine Änderung des gespeicherten Potenzials in Richtung V_{PL} und damit immer eine Verschlechterung der gespeicherten Information. Der gravierende Unterschied in den Retentionkurven der beiden Speicherzustände deutet bereits darauf hin, dass die bezüglich des Speicherzustands symmetrischen Leckstrompfade *Vertical Parasitic*, *Node Leakage*, *SubVt*, *DeepSubVt*, und *SubSTI* gegenüber den *pn-Leckströmen* einen betragsmäßig kleinen Anteil des Gesamtleckstroms ausmachen.

Eine weitere Beobachtung, die in der Literatur nicht zu finden ist, besteht in der Sättigung der Fehlerzahlen für sehr hohe Retentionzeiten (im Bereich von Minuten bis Stunden). Abbildung 7.1 zeigt den üblicherweise gemessenen Bereich grau schattiert. Der eingeschränkte Messbereich ist dadurch motiviert, dass für produktive Zwecke alle benötigten Informationen beinhaltet sind. Für diese ist nur die „1" von Bedeutung und Retentionzeiten im Minuten- oder sogar Stundenbereich sind weit von der Spezifikation bei $64\,ms$ entfernt. Meist werden keine Kurven sondern nur Fehlerzahlen an der Reparaturgrenze gemessen. Trotzdem sind die vollständigen Kurven für grundsätzliche Untersuchungen sehr interessant. So geht die „0"-Kurve bei einer kumulativen Wahrscheinlichkeit von in diesem Fall -3.3 σ in Sättigung. Demzufolge kann ein Großteil der Speicherzellen (99.952%) die Information „0" überhaupt nie verlieren. Interessanterweise fällt gerade der komplementäre Anteil bezüglich der „1" nie aus (Sättigung bei $3.3\,\sigma$). Eine mögli-

che Erklärung dafür ist, dass sich nach einer gewissen Zeit in den Speicherzellen eine von $0\,V$ verschiedene Spannung einstellt. Abhängig vom Offset des Differenzverstärkers wird diese Information als „0" oder „1" gewertet. Experimentell simuliert werden kann diese „Endspannung" durch so genannte *Signal Margin*-Messungen. Dabei werden alle Zellen mit einer variierten Spannung beschrieben und sofort anschließend ohne Retentionpause wieder ausgelesen. Durch wiederholte Messung über den Spannungsbereich ergibt sich Abbildung 7.2. Daraus kann ein Potenzial von ungefähr 0.52 bis $0.54\,V$ bei einer Ausfallrate von $3.3\,\sigma$ für die „1" und gleichzeitig $-3.3\,\sigma$ für die „0" entnommen werden. Stellt sich also in allen Zellen dieser Potenzialbereich ein, würde die Verteilung der Differenzverstärker-Offsets genau die beobachteten Sättigungen erklären.

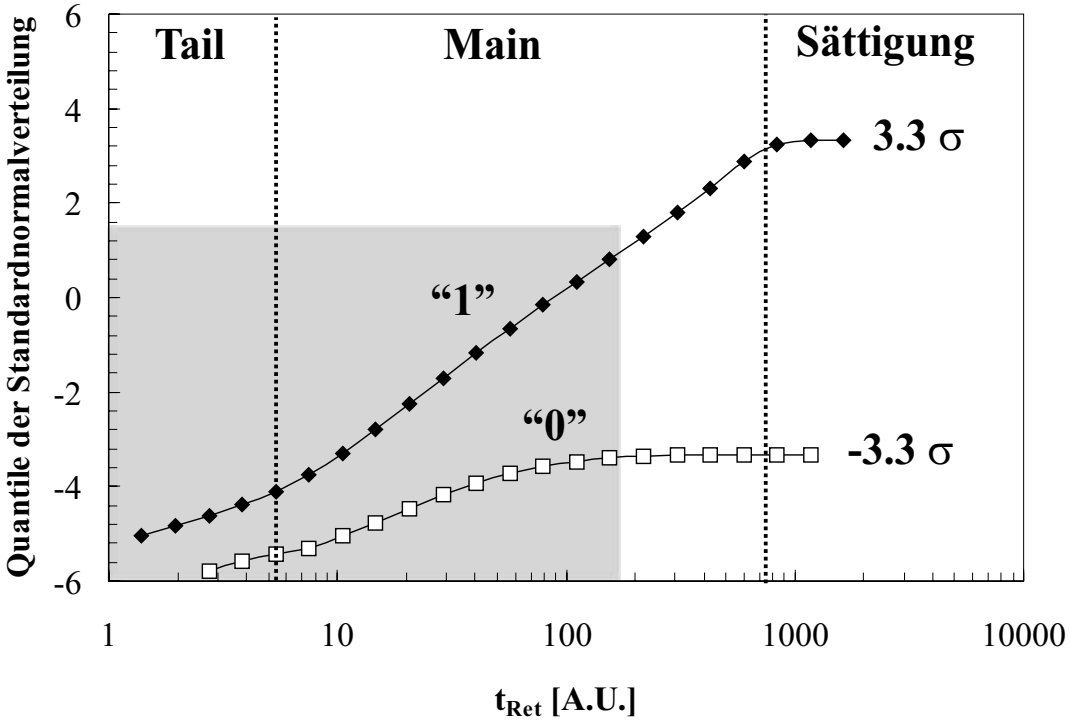


Abbildung 7.1: Vollständige „1" und „0"-Retentionkurven eines Speicherchips bei $85\,°C$. Die Haltezeiten wurden dabei bis in den Stundenbereich variiert. In der Literatur wird üblicherweise nur die „1"-Retention im schattierten Bereich betrachtet.

Zur Motivation einer von $0\,V$ verschiedenen Endspannung zeigt Abbildung 7.3 den zeitlichen Verlauf der symmetrischen und asymmetrischen Komponenten des Gesamtleckstroms. Der aus *Junction Leakage* und *GIDL* bestehende asymmetrische Anteil liefert für alle Zeiten und Spannungen nur eine Stromkomponente von der inneren Kondensatorelektrode in die p-Wanne, während der symmetrische Anteil das Vorzeichen bei $V_S = V_{PL}$ wechselt. Die genaue Form des Stromverlaufes ist für die Betrachtung hier nicht relevant. Speziell im oberen Spannungsbereich gibt es für den asymmetrischen Anteil Abweichungen, die in späteren Kapiteln noch genau diskutiert werden. Zur Beschreibung

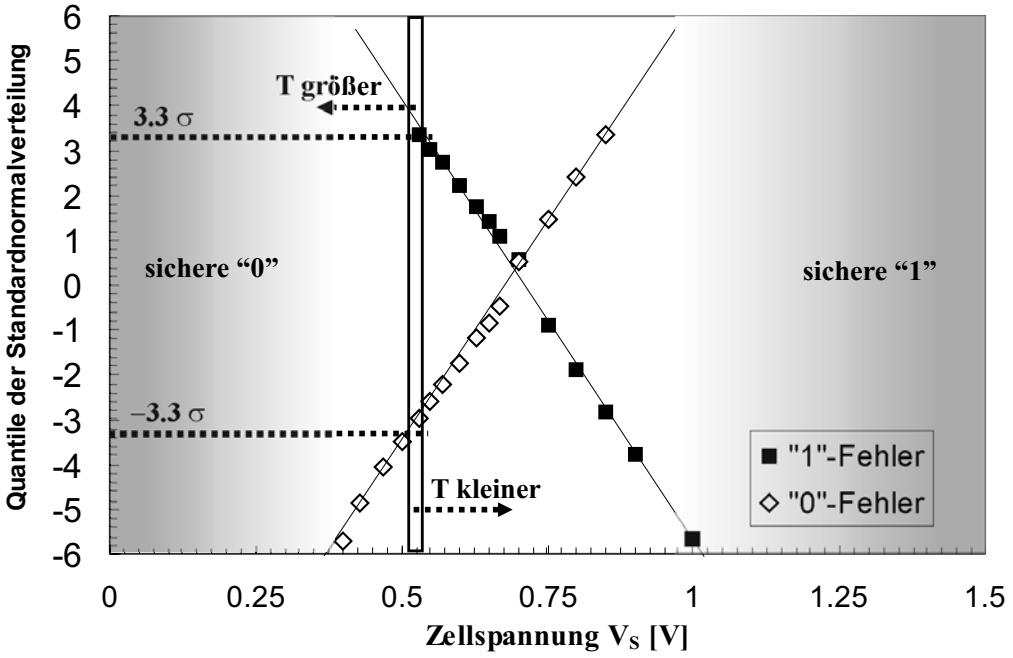


Abbildung 7.2: *Signal Margin* Messungen. Ungefähr $0.52 - 0.54\,V$ Zellspannung führen zu einem kumulativen Ausfall von $-3.3\,\sigma$ für die „0" und $3.3\,\sigma$ für die „1". Die Streuung um die per Hand eingezeichneten Geraden werden auf die Messmethode zurückgeführt, sind hier aber für ein qualitatives Verständnis ohne Bedeutung.

der zwei Ladungszustände muss die Richtung der Zeit-Achse vertauscht werden. Die „0" beginnt zum Schreibzeitpunkt mit $V_S = 0\,V$ und der Stromverlauf entwickelt sich von links nach rechts. Für eine „1" startet die Zellspannung mit $1.5\,V$ und der Stromverlauf wird von rechts nach links wiedergegeben. Für beide Fälle gibt es im unteren Spannungsbereich einen Spannungswert, bei dem sich zu- und abfließende Ströme gerade ausgleichen. Bei diesem Spannungswert angekommen bleibt die Spannung für alle Zeiten unverändert und der Offset des jeweiligen Differenzverstärkers entscheidet über die Fehlerwertung. Es muss an dieser Stelle angemerkt werden, dass die Abbildung den Verlauf der Ströme in einer speziellen Zelle skizziert. Tatsächlich streuen alle Leckstromkomponenten von Zelle zu Zelle (für alle Komponenten existiert eine Verteilung), sodass eigentlich breit gestreute Kurvenbündel eingezeichnet werden müssten. Für lange Retentionzeiten scheint sich dennoch ein relativ schmales Band von Endspannungen einzustellen, was auf für geringe Zellspannungen kleinere Leckstromstreuungen hindeutet.

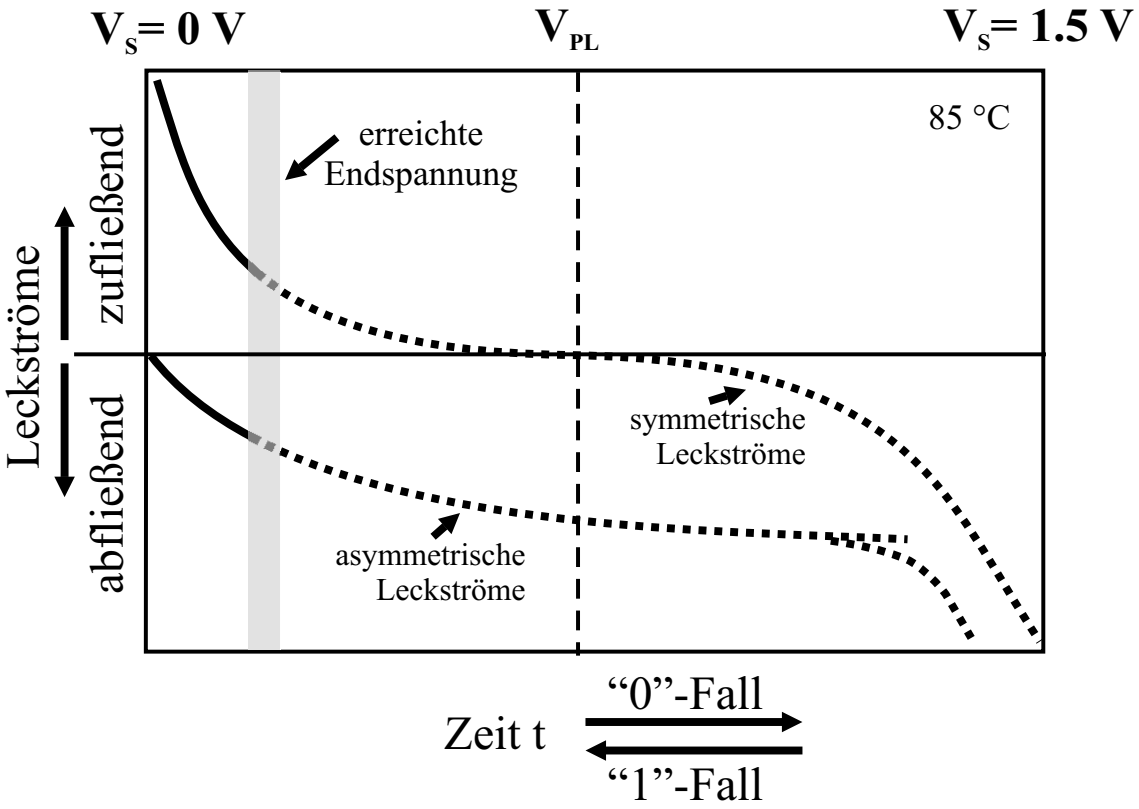


Abbildung 7.3: Zeitlicher Stromverlauf und erreichte Endspannung in einer Speicherzelle. Der Differenzverstärker entscheidet abhängig von der erreichten Spannung und des Offsets, ob die Endspannung als Fehler gewertet wird.

7.1.2 Temperatur-Abhängigkeit

Im vorigen Abschnitt wurde der Unterschied zwischen „0" und „1"-Retention auf die unterschiedliche Gewichtung von symmetrischen und asymmetrischen Leckstrompfaden

zurückgeführt. Da die asymmetrischen Anteile stark temperaturabhängig sind, während lediglich *Node-Leakage* als Teil des symmetrischen Anteils temperaturunabhängig ist, müssen sich daraus auch unterschiedliche Temperaturverhalten der Retentionkurven für beide Zustände ergeben.

Temperaturabhängigkeit der „1"-Retention

Die Temperaturabhängigkeit der „1"-Retentionkurve ist in Abbildung 7.4 dargestellt. Die Retentionkurve verschiebt sich mit zunehmender Temperatur in erster Näherung als Ganzes hin zu kürzeren Retentionzeiten. Der in allen Zellen dominante Leckstrom nimmt also stark mit der Temperatur zu. Wäre unzureichendes Schreiben bzw. Lesen aufgrund von Serienwiderständen die Ursache für die breite Verteilung der Haltezeiten, würde aufgrund der Temperaturabhängigkeit des Widerstandes in Halbleitern ein umgekehrtes Verhalten erwartet werden. Dies ist eine Bestätigung dafür, dass die Verteilung der Schreib- und Lesefaktoren $P_{w,r}$ in den Monte Carlo Simulationen von Kapitel 4 unberücksichtigt bleiben durften. Als Faustregel für die Temperaturabhängigkeit gilt, dass $\Delta T = 10\,°C$ in der Retentionzeit ungefähr einem Faktor 1.7 im Tail und einem Faktor 2 im Main entsprechen. Der Retention-Main unterliegt also einer größeren Temperaturabhängigkeit als der Tail. Dies entspricht einer größeren Aktivierungsenergie für den Main und einer kleineren Aktivierungsenergie für den Tail der Verteilung.

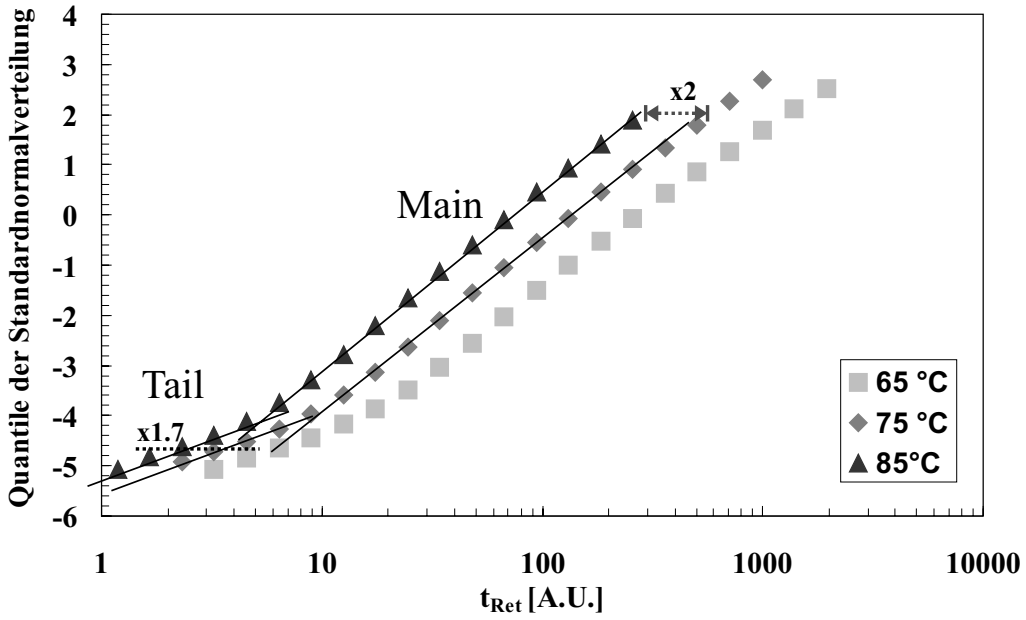


Abbildung 7.4: Temperaturabhängigkeit der „1"-Retentionkurve. Faustregel: $\Delta T = 10\,°C$ entsprechen einer Verschiebung um ungefähr einen Faktor 1.7 im Tail und 2 im Main der Verteilung.

Temperaturabhängigkeit der „0"-Retention

Abbildung 7.5 zeigt die Temperaturabhängigkeit für den „0"-Speicherzustand. Wie erwartet sind in der Temperaturabhängigkeit klare Unterschiede zur „1"-Kurve erkennbar. Ausgehend von der $85\,°C$-Kurve schmiegt sich die Verteilung für kleinere Temperaturen einer temperaturunabhängigen Geraden an. Die bei $85\,°C$ auf $-3.3\,\sigma$ limitierte Sättigungsfehlerzahl wächst dabei an und die Verteilung schmiegt sich auch für hohe Retentionzeiten der temperaturunabhängigen Gerade an. Bedingt durch beide Effekte schneiden sich alle Kurven nahezu in einem Punkt. Da die Unterschwellenleckströme ebenfalls temperaturabhängig sind, bleibt als Ursache für die temperaturunabhängige Leckstromkomponente nur *Node-Leakage* übrig. Aus der Ausfallzeit der schlechtesten Zelle, die den maximalen *Node-Leakage* bezeichnet und weit über der Spezifikation liegt, folgt, dass *Node-Leakage* trotz der sehr breiten Verteilung für alle Zellen klein gegenüber den *pn-Leckströmen* ist.

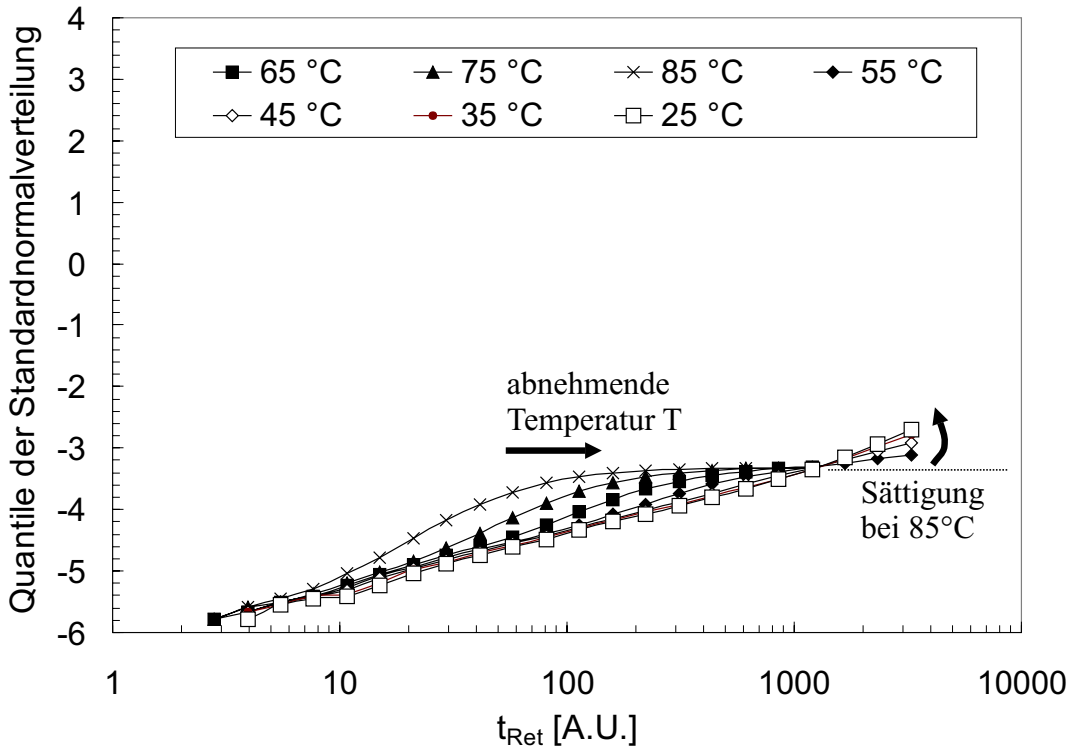


Abbildung 7.5: Temperaturabhängigkeit der „0" - Retentionkurve

Temperaturabhängigkeit der Sättigungsbereiche

Die Lage des Sättigungsbereiches hängt vom Gegenspiel der symmetrischen und asymmetrischen Leckstrompfade ab. Wie schon diskutiert besitzen diese unterschiedliche Temperaturabhängigkeiten. Geht man gedanklich zu extrem hohen Temperaturen, so ist zu

erwarten, dass ab einer gewissen Temperatur die *pn-Leckströme* alle symmetrischen Anteile überwiegen und deshalb alle Zellen eine Endspannung nahe $0\,V$ erreichen. Gemäß der Verteilung der Differenzverstärker-Offsets (Abbildung 7.2) würde dies zum Ausfall aller Zellen führen und die Mainverteilung der „1"-Retention würde sich bis zu $6\,\sigma$ gerade fortsetzen ohne in Sättigung zu gehen. Im Gegensatz dazu würden keine Zellen jemals bezüglich der „0" ausfallen. Im Fall sehr kleiner Temperaturen sind alle temperaturabhängigen Leckströme gegenüber *Node-Leakage* zu vernachlässigen. Dies würde, wenn auch für sehr hohe Retentionzeiten, zu einer Endspannung aller Zellen nahe V_{PL} führen und es würde bedingt durch die Verteilung der Differenzverstärker-Offsets zu einer maximalen Ausfallrate der „1" von ungefähr $-1\,\sigma$ und der „0" von $+1\,\sigma$ kommen. Leider konnten Messungen für beide Extremfälle nicht durchgeführt werden, da einerseits die Temperatur durch die Möglichkeiten des Testers auf maximal $110\,°C$ begrenzt ist und andererseits für kleine Temperaturen die nötigen Messzeiten im Bereich von mehreren Tagen liegen (in einer Produktionsumgebung sind solch lange Testzeiten im Hinblick auf die Spezifikation bei $64\,ms$ einfach nicht zu rechtfertigen).

7.1.3 V_{BB}-Abhängigkeit

Abbildung 7.6 zeigt die V_{BB}-Abhängigkeit der Retentionkurve. Dabei zeigt sich, dass die kumulative Fehlerzahl sowohl im Main als auch im Tail mit negativerem V_{BB}, d.h. größerer Sperrspannung des pn-Übergangs und somit höherem elektrischen Feld in der Verarmungszone, zunimmt. Der größte Effekt ist dabei in der Mainverteilung zu beobachten. Im Tail sind nur geringfügige Auswirkungen zu sehen und der Sättigungsbereich bleibt unverändert. Die Spannungsabhängigkeit des Leckstroms entspricht der von *Junction Leakage* und ist entgegengesetzt zur Abhängigkeit der Unterschwellenleckströme *SubVt*, *Deep-SubVt*, *SubSTI* und *Vertical Parasitic*. *Junction Leakage* dominiert somit den Main der Retentionverteilung.

7.1.4 V_{NWLL}-Abhängigkeit

Die Abhängigkeit der Retentionkurve von der Wortleitungsspannung V_{NWLL} ist verglichen mit der V_{BB}-Abhängigkeit deutlich komplexer (siehe Abbildung 7.7a). Im Main der Verteilung kann für verschiedene Wortleitungsspannungen keine signifikante Veränderung beobachtet werden. Die Kurven sind nahezu deckungsgleich. Im Tail nimmt die Fehlerzahl mit positiverem V_{NWLL} zuerst ab und dann schlagartig wieder zu. Grund dafür sind zwei gegenläufige Effekte. Ein positiveres V_{NWLL} reduziert zum einen das elektrische Feld im Überlappbereich zwischen Gate und Drain. Dadurch wird der GIDL-Leckstrom kleiner und Zellen wandern vom Tail in das Maingebiet. Zum anderen wird der *SubVt-Leckstrom* gemäß der Unterschwellensteigung exponentiell mit positiverem

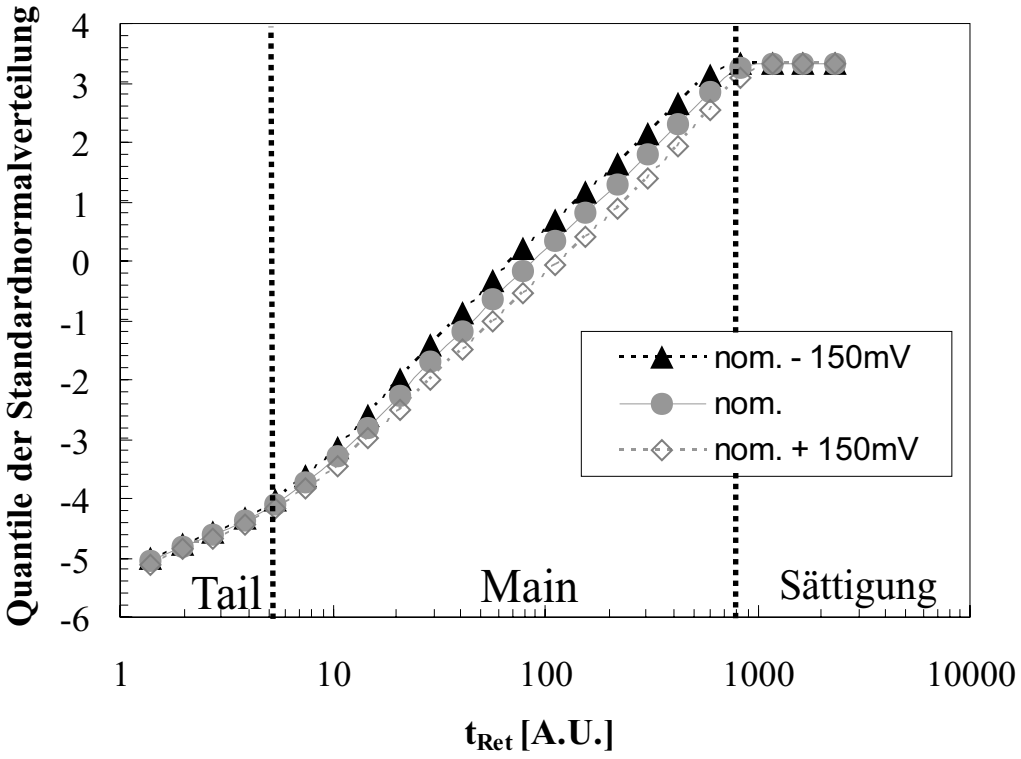


Abbildung 7.6: Retentionkurven bei variierten V_{BB}-Spannungen. Der Haupteinfluss zeigt sich im Bereich der Mainverteilung, während der Einfluss auf Tail- und Sättigungsbereich geringer ist.

V_{NWLL} erhöht und Zellen wandern aus dem Main der Verteilung in den Tail. Anhand Retentionkurven-Messungen kann nicht festgestellt werden, ob es sich bei den hinzukommenden Fehlerzellen um die ursprünglichen Tailzellen oder andere Zellen handelt, da die Adressen unbekannt sind. Dass es tatsächlich andere Zellen sind, zeigen Einzelzellmessungen im zweiten Teil dieses Kapitels. Typische Unterschwellensteigungen heutiger Zelltransistoren liegen im Bereich von ungefähr $100\,mV/dec$. Die Gatespannung V_{NWLL} wurde in der Kurvenschar der Abbildung 7.7a insgesamt um $600\,mV$ variiert. Daraus folgt eine Erhöhung des Unterschwellenleckstroms um sechs Größenordnungen. Jedoch erst nach knapp vier Dekaden Stromerhöhung wird die Verschlechterung im Tail der Verteilung sichtbar. Demzufolge sind unter Nominalbedingungen *SubVt*-Leckströme um Größenordnungen kleiner als die Leckströme der Tailzellen. Außerdem muss die Verteilung der *SubVt*-Leckströme sehr breit sein, da der Main der Verteilung bei positiverem V_{NWLL} nahezu unverändert bleibt. Der äußerste Rand der SubVt-Verteilung wandert „unbemerkt" durch den Main und wird dann schlagartig im Tail sichtbar. Abbildung 7.7b zeigt die auf Nominalbedingungen normierte Fehlerzahl für $\Delta V_{NWLL} = +340\,mV$, um den in den Retentionkurven nur schwer zu erkennenden Effekt sichtbar zu machen. Es zeigt sich, dass die Fehlerzahl im Tail und Anfang des Maingebietes abnimmt, obwohl der $SubVt$-Leckstrom dadurch um grob 3.5 Größenordnungen erhöht wird. Demzufolge überwiegen GIDL-Leckströme in diesem Bereich den Unterschwellenleckstrom. Es muss jedoch angemerkt werden, dass obwohl ein positiveres V_{NWLL} in diesem Fall eine deutliche Verbesserung verspricht, in der Praxis dies nicht ausgenutzt werden kann. Grund dafür sind hinsichtlich einiger Bedingungen (*SubVt*-Stress, elektrisches Übersprechen ...) höhere Anforderungen im Speicherbetrieb gegenüber den Retentionkurvenmessungen. Da $SubVt$-Fehler im Tail „schlagartig" in großer Anzahl auftreten und die Reparaturmöglichkeiten schnell übersteigen, muss V_{NWLL} derart gewählt werden, dass jegliche V_t-Variationen, die z.B. durch die Position des Chips auf dem Wafer, Wafer-zu-Wafer oder Los-zu-Los Streuungen entstehen, zu keinem Ausbeuteverlust führen. Die nominellen Betriebsspannungen sind bereits dahingehend optimiert.

7.1.5 V_{BL}-Abhängigkeit

Die Spannung des Bitleitungskontaktes V_{BL} während der Retentionpause beeinflusst die Leckstrompfade *SubVt, Deep-SubVt* und *SubSTI*. Zu Testzwecken kann diese mittels Testmodes auf $0\,V$ gesetzt werden. Dadurch liegt im Falle einer gespeicherten „1" eine erhöhte Source/Drain- Spannung an und Unterschwellenleckströme des Auswahltransistors sowie *SubSTI*-Ströme werden erhöht. Wie bei zu geringer Gatespannung V_{NWLL} werden zusätzliche Fehler im Tail der Verteilung sichtbar, der Retentionmain bleibt davon jedoch unverändert (siehe Abbildung 7.8). Da der Unterschied in der kumulativen Kurve nur schwer erkennbar ist, wurde in Abbildung 7.8 zusätzlich die Fehlerzahl bei $V_{BL} = 0\,V$, normiert auf die Standardmessung bei $V_{BL} = V_{BLH}/2$, aufgenommen.

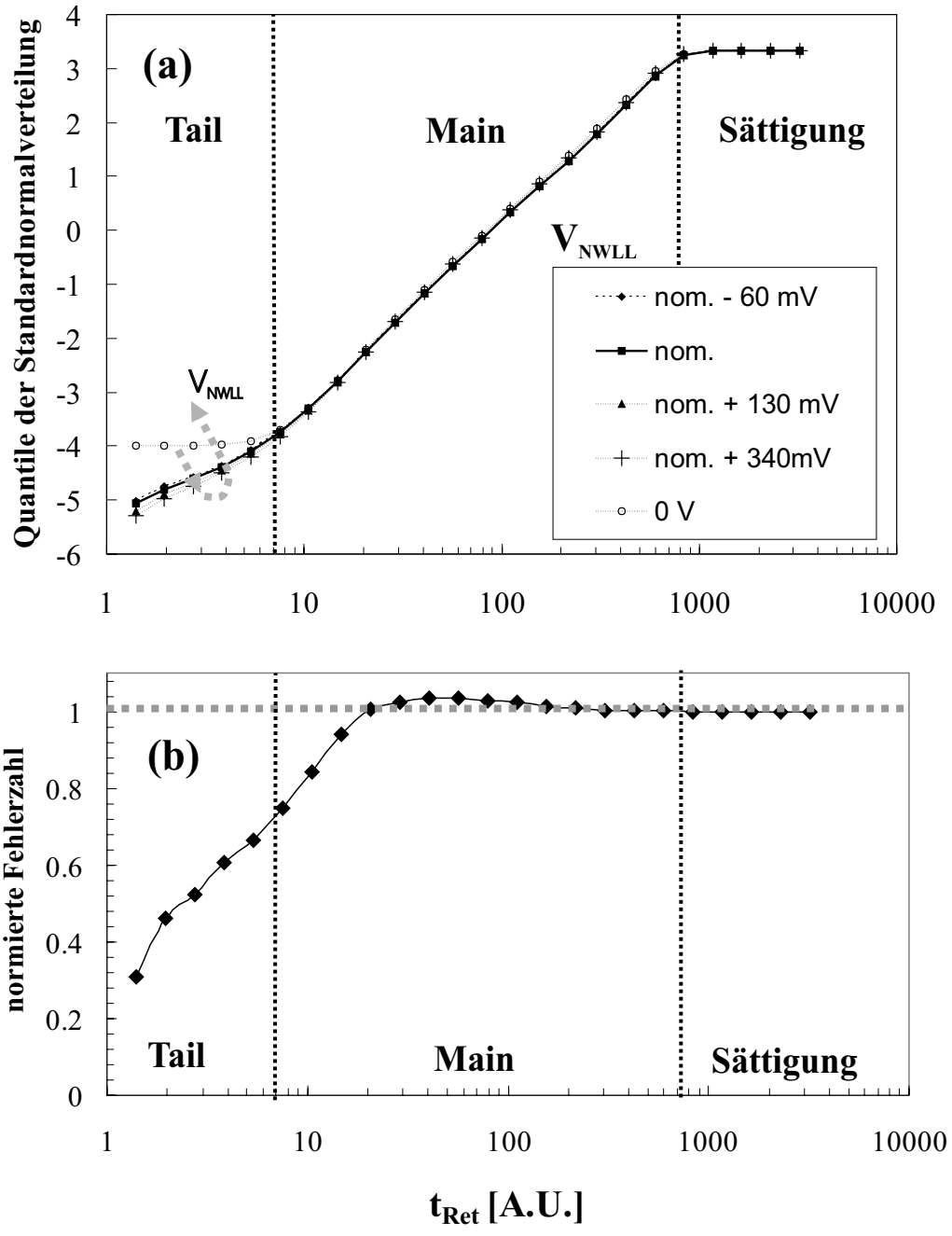


Abbildung 7.7: (a) Retentionkurven einer Komponente mit verschiedenen Gatespannungen V_{NWLL} während der Pause. (b) normierte Fehlerzahl für $\Delta V_{NWLL} = +340\,mV$.

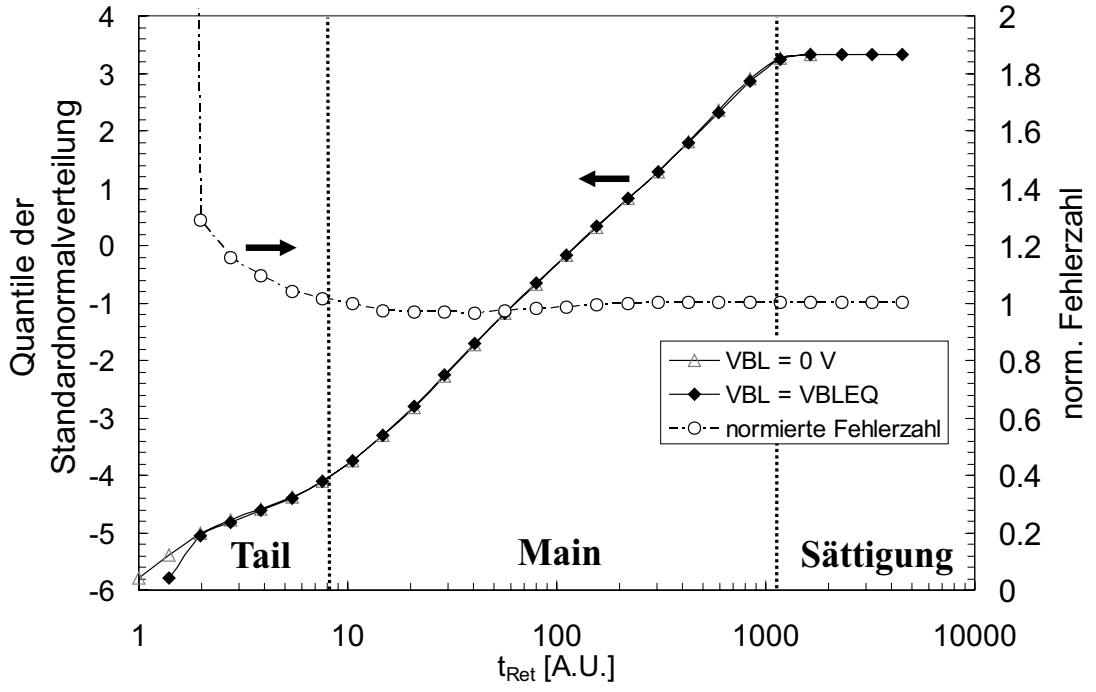


Abbildung 7.8: Vergleich der Retentionkurven mit Bitleitungsspannungen $V_{BL} = 0\,V$ und $V_{BL} = V_{BLH}/2$ während der Pause. Die Abbildung zeigt beide Retentionkurven sowie das Verhältnis der Fehlerzahlen.

Zusammenfassung der Retentionkurven-Messungen

Anhand von Retentionkurven für beide Speicherzustände wurde gezeigt, dass nur der „1"-Zustand limitierend für die Retentionzeit eines Speicherbausteines ist. Mehr als 99.9% aller Zellen fallen bezüglich des „1"-Zustandes aus, während nur der komplementäre Anteil jemals bezüglich der „0" ausfällt. Dadurch konnte gezeigt werden, dass die asymmetrischen Leckstrompfade (*Junction Leakage, GIDL*) die symmetrischen (*SubVt, DeepSubVt, SubSTI, Vertical Parasitic, Node Leakage*) überwiegen. Auch die Temperaturabhängigkeit der maximalen Fehlerzahl im Sättigungsbereich passt zu diesem Modell. Für die „1"-Retention zeigt die Zunahme der Fehlerzahl mit negativerem V_{BB}, dass über die gesamte Verteilung *Junction Leakage* die Unterschwellenleckströme *SubVt, DeepSubVt, SubSTI, Vertical Parasitic* dominiert. Dabei ist der Einfluss auf die Mainverteilung wesentlich größer als auf den Tail. Es wurde klargestellt, dass Unterschwellenleckströme zwar bei Retentionmessungen keine wesentliche Rolle spielen, im aktiven Fall, d.h. zum Beispiel für $V_{BLH} = GND$ während der Retentionpause, aber zusätzliche Fehler im Tail auftreten können. Die V_{NWLL}-Abhängigkeit bestätigt einerseits, dass *SubVt*-Leckströme unter Nominalbedingungen um Größenordnungen unter den für Tailzellen notwendigen Strömen liegen müssen, andererseits im Tail eine starke GIDL-Abhängigkeit vorhanden ist. Insgesamt zeichnet sich ab, dass für die Retentionverteilung bei $85\,°C$ hauptsächlich die Leckstrompfade *Junction Leakage* und *GIDL* ausschlaggebend sind.

7.2 Einzelzell-Analysen

Die direkte Charakterisierung von Einzelzellen auf einem Speicherbaustein bietet gegenüber Messungen von Retentionkurven einige Vorteile. Da im Gegensatz zu den Messungen im letzten Abschnitt hierbei die Zelladressen bekannt sind, können Effekte durch Vertauschen der Ausfallreihenfolge nicht auftreten. Dadurch ist auch bei veränderten Randbedingungen sichergestellt, dass die ursprüngliche Speicherzelle Gegenstand der Charakterisierung bleibt. Die Einzelzellmethode ist eine sehr aufwendige und deshalb unübliche Charakterisierungstechnik, die speziell in dieser Arbeit weiterentwickelt wurde. Das eigens dafür geschriebene Messprogramm wurde in Kapitel 6 vorgestellt. Einzelzell-Analysen werden als Erweiterung zum vorherigen Abschnitt eingesetzt und erlauben die Gewinnung von zusätzlichen Informationen über grundlegende Leckstrommechanismen.

7.2.1 Temperaturabhängigkeit

Durch die Bestimmung der Retentionzeiten von Einzelzellen nach der in Kapitel 6 beschriebenen Methode, kann die Temperaturabhängigkeit der Retentionzeit und damit auch des dominanten Leckstrommechanismus genau bestimmt werden. Im Gegensatz zur in der Literatur üblichen Bestimmung der Aktivierungsenergien aus Retentionkurven (siehe z.B. [Ham98]), erhält man dadurch pro Retentionzeit eine breite Verteilung der Aktivierungsenergien anstatt eines einzelnen Wertes. Das Problem der konventionellen Methode besteht darin, dass durch das Ablesen bei einer bestimmten Wahrscheinlichkeit für jede Temperatur die Retentionzeit einer anderen Zelle entnommen wird. Abbildung 7.9 zeigt die kumulative Verteilung der Aktivierungsenergien einer Stichprobe von Zellen mit Retentionzeiten innerhalb eines kleinen Intervalles nahe der Reparaturgrenze. Jeder Messpunkt im Diagramm repräsentiert die Aktivierungsenergie einer Zelle. Im kumulativen Plot ergeben sich zwei um ungefähr $0.05\,eV$ getrennte Teilabschnitte. Die Aktivierungsenergien unterliegen demnach keiner einfachen Normalverteilung, sondern müssen als Mischverteilung bestehend aus zwei Normalverteilungen mit unterschiedlicher Gewichtung angesetzt werden. Die Unterverteilung bei niedrigeren Aktivierungsenergien beinhaltet ca. 78% der getesteten Zellen. Die verbleibenden 22% gehören zu einer Unterverteilung mit höherem Median. Diese Aufspaltung der Aktivierungsenergien in zwei Unterverteilungen kann nur durch Einzelzellmessungen beobachtet werden und lässt zwei verschiedene Mechanismen vermuten. Die Bestimmung der Aktivierungsenergie aus Retentionkurven hätte anstatt der Verteilung nur einen einzelnen Wert bei ungefähr $0.47\,eV$ ergeben.

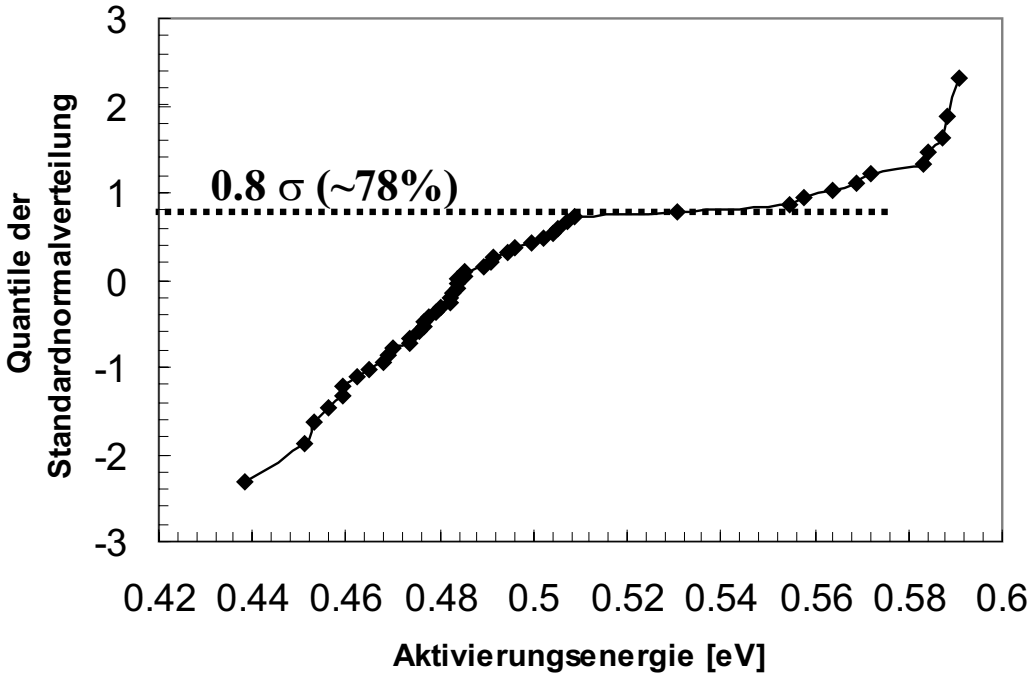


Abbildung 7.9: Aktivierungsenergien einer Stichprobe von Zellen mit Retentionzeiten nahe der Reparaturgrenze. Es ergibt sich eine Mischverteilung bestehend aus zwei Normalverteilungen.

7.2.2 Spannungsabhängigkeit der Retentionzeiten

Um der Hypothese verschiedener Mechanismen nachzugehen, wurden die Retentionzeiten der selben Stichprobe von Speicherzellen spannungsabhängig untersucht. Dabei konnte für die Unterverteilung mit niederen Aktivierungsenergien eine klare Korrelation zur Abhängigkeit der Retentionzeit von der Gatespannung V_{NWLL} festgestellt werden. Abbildung 7.10a zeigt dazu die Verteilung der Aktivierungsenergien und 7.10b die Verteilung der prozentualen Retentionzeitänderungen für den Fall einer um $70\,mV$ erhöhten Gatespannung. Zur Visualisierung der Korrelation wurden in beiden Teilabbildungen Zellen, die zur Verteilung mit höheren Aktivierungsenergien gehören, mit offenen Symbolen und Zellen, die der Unterverteilung mit niederen Aktivierungsenergien angehören, mit gefüllten Symbolen dargestellt. Aus der Betrachtung der offenen Symbole in beiden Diagrammen ist ersichtlich, dass die Retentionzeiten der Zellen mit hohen Aktivierungsenergien innerhalb der Messgenauigkeit nicht durch die Änderung der Gatespannung verändert werden. Die Zellen aus der niederen Unterverteilung reagieren dagegen sehr empfindlich auf die veränderte Gatespannung. Dabei wird die Retentionzeit für positivere Gatespannungen, d.h. kleinerem elektrischen Feld im Gate/Drain Überlapp, größer und damit der Leckstrom kleiner, was der Charakteristik von *GIDL* entspricht.

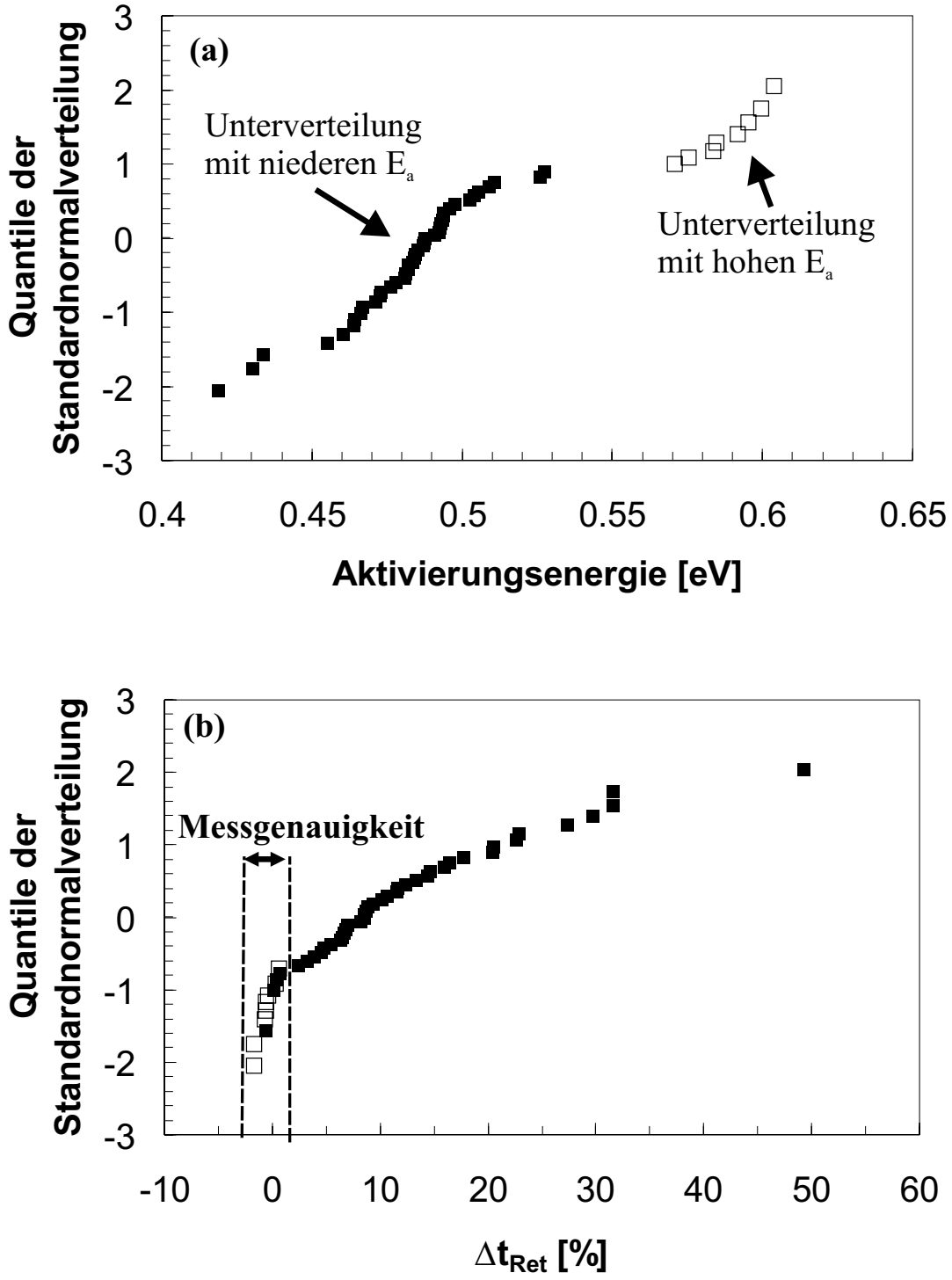


Abbildung 7.10: Korrelation zwischen den Unterverteilungen der Aktivierungsenergien und der Spannungsabhängigkeit der Retentionzeiten. Zur besseren Darstellung wurden die Zellen der Unterverteilung mit hohen Aktivierungsenergien in beiden Diagrammen mit offenen und die mit niedrigen Aktivierungsenergien mit gefüllten Symbolen gezeichnet.

Abbildung 7.11 zeigt zur Veranschaulichung des Effektes einer veränderten Gatespannung die durch Davinci 3d-Simulation gewonnene elektrische Feldverteilung für eine Schnittebene längs der Mitte des aktiven Gebietes. Die Positionen des Gates (GC), der Isolation zu den Nachbarzellen (STI), des Grabenkondensators (DT) und des Oxidkragens am oberen DT-Ende (Collar) sind darin skizziert. Die linke Hälfte der Abbildung 7.11 zeigt die prozentuale Änderung des elektrischen Feldes bei einer Änderung der Gatespannung um$+340\,mV$. Die größte Feldänderung entsteht nahe der Siliziumoberfläche. Dort nimmt die Feldstärke um bis zu 25% ab. Da nicht nur die prozentuale Änderung, sondern auch der Betrag der Feldstärke selbst von Bedeutung ist, zeigt der rechte Teil der Abbildung 7.11 die Feldverteilung bei nominellen Spannungsbedingungen zum Vergleich. Die maximalen elektrischen Felder treten hierbei zum einen Nahe der Oberfläche direkt unterhalb des Gate-Spacers sowie an der Grenzfläche des Buried Straps zum Collar auf. Der Bereich unterhalb von $100\,nm$ sieht den Einfluss der Änderung von V_{NWLL} nicht. Deshalb müssen die Defekte, die zu Leckströmen führen, welche durch V_{NWLL} beeinflussbar sind, in einer maximalen Tiefe von $100\,nm$ liegen. Der Großteil der untersuchten Zellen nahe der Reparaturgrenze leidet demnach an Generationsleckströmen, die durch Defekte nahe der Oberfläche im G/D-Überlappbereich verursacht werden.

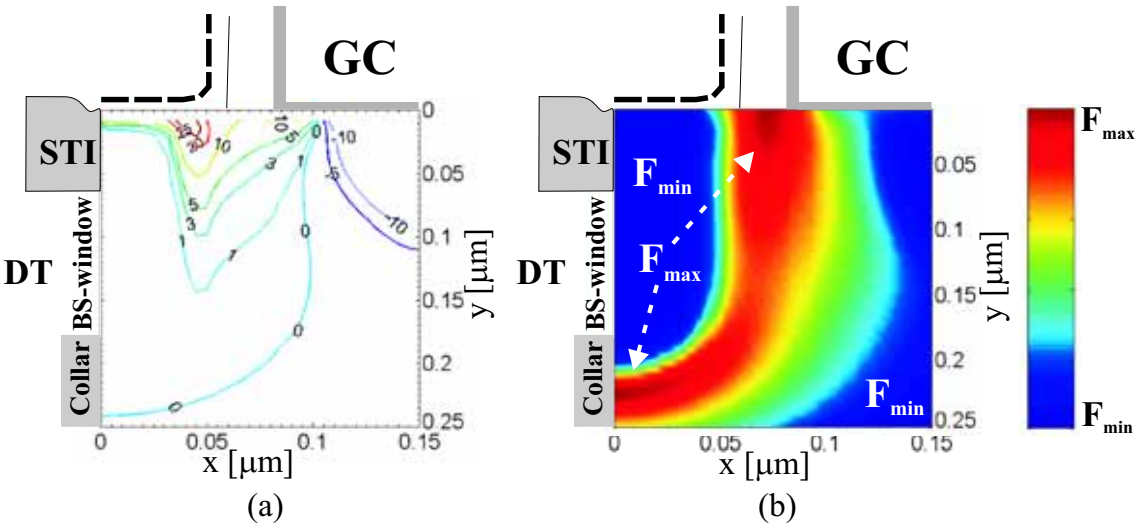


Abbildung 7.11: (a) Prozentuale Feldänderung bei Änderung der Gatespannung V_{NWLL} um $+340\,mV$. Die Gatespannung beeinflusst das elektrische Feld bis in eine Tiefe von ungefähr $100\,nm$. (b) Feldverteilung bei Nominalbedingungen.

Das elektrische Feld im Bereich des kondensatorseitigen pn-Übergangs hängt zusätzlich von der Spannung der p-Wanne V_{BB} ab. Abbildung 7.12 zeigt die durch eine Änderung von V_{BB} hervorgerufene Feldänderung. Im Gegensatz zur V_{NWLL} bedingten Feldänderung wirkt sich V_{BB} gleichmäßiger auf das Feld entlang des Buried Straps aus. Die maximale Feldreduktion beträgt dabei 5 % und erstreckt sich ab einer Tiefe von ca. $30\,nm$ bis zum Collar. Direkt an der Siliziumoberfläche ist die Feldänderung wesentlich kleiner als bei Variation der Gatespannung, sodass die beiden Spannungen im wesentlichen

unterschiedliche Siliziumbereiche ansprechen und deshalb der Ort der zugrundeliegenden Defekte durch deren Abhängigkeit eingeschränkt werden kann. Eine Korrelation zur Retentionzeit ist im Fall einer V_{BB}-Änderung jedoch nicht erkennbar.

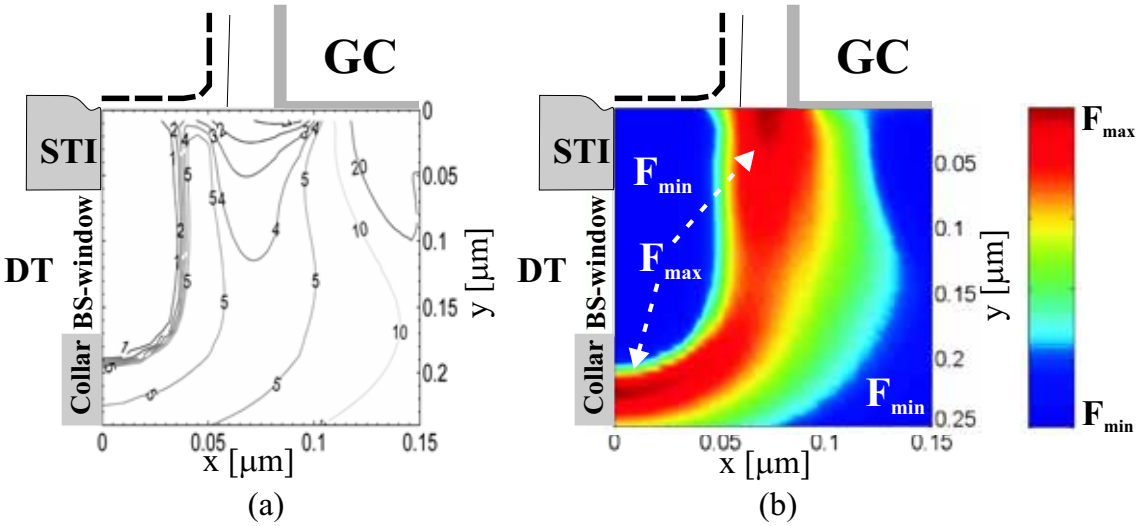


Abbildung 7.12: (a) Prozentuale Feldänderung durch Anhebung der p-Wannenspannung V_{BB} um $+150\,mV$. (b) Feldverteilung bei Nominalbedingungen.

7.2.3 Spannungsabhängigkeit der Aktivierungsenergien

Aus den vorherigen Abschnitten geht hervor, dass Tail-Leckströme durch Generation im kondensatorseitigen pn-Übergang entstehen. Die Korrelation zur Spannungsänderung zeigte weiterhin, dass der Hauptteil der für solch große Generationsströme verantwortlichen Defekte im Einflussbereich des Gates liegen müssen. Durch Untersuchung der Feldabhängigkeit der Aktivierungsenergien kann nun noch detaillierter zwischen möglichen Generationsmechanismen unterschieden werden (theoretische Hintergründe folgen in Kapitel 8). Dazu zeigt Abbildung 7.13 die Abhängigkeit der Aktivierungsenergien bezüglich der Spannungen V_{BB} und V_{NWLL}. Abbildung 7.13a zeigt die Änderung der Aktivierungsenergieverteilung für $\Delta V_{NWLL} = +340\,mV$. Der Vergleich der hohen Unterverteilung (offene Symbole) für beide Gatespannungen ergibt keine Änderung der Aktivierungsenergien, während sich die Aktivierungsenergien der Zellen aus der niederen Unterverteilung (gefüllte Symbole) für eine positivere Gatespannung, also kleinerem elektrischem Feld, hin zu höheren Aktivierungsenergien verschieben. Für die Abhängigkeit von V_{BB} (Abbildung 7.13b) ergibt sich gerade das inverse Bild. Die Unterverteilung mit höheren Aktivierungsenergien verschiebt sich für kleineres Feld hin zu höheren Werten und die der niederen Unterverteilung bleiben unverändert.

Insgesamt kann also festgestellt werden, dass alle Zellen eine feldabhängige Aktivierungsenergie aufweisen. Dabei wird die Aktivierungsenergie mit abnehmendem elektrischen Feld größer.

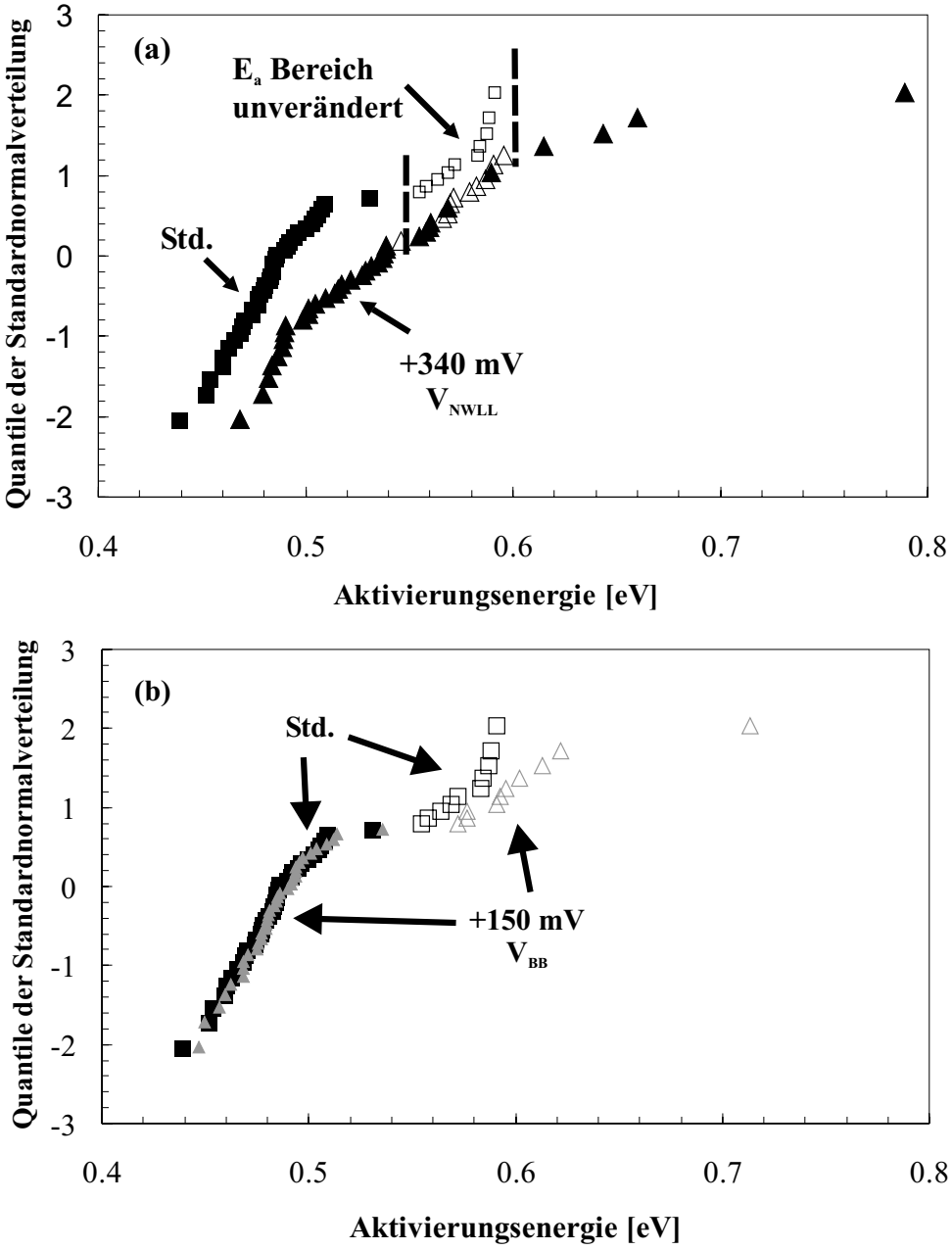


Abbildung 7.13: Abhängigkeit der Aktivierungsenergien vom elektrischen Feld bzw. den internen Spannungen V_{NWLL} und V_{BB}.

7.2.4 Aktivierungsenergien entlang der Retentionkurve

Bisher wurde nur die Aktivierungsenergieverteilung einer Stichprobe von Zellen nahe der Reparaturgrenze betrachtet. Durch weitere Stichproben von Zellen bei unterschiedlichen Retentionzeiten entlang der Retentionkurve und Bestimmung der Aktivierungsenergieverteilungen kann der Zusammenhang zwischen dem Betrag des Leckstroms und der Aktivierungsenergie hergestellt werden. Außerdem kann unter Zuhilfenahme theoretischer Betrachtungen der Aktivierungsenergien eine Aussage über den vorherrschenden Mechanismus der Tailverteilung getroffen werden. Die hierfür untersuchten Bereiche der Retentionverteilung sind in Abbildung 7.14 mit A, B und C markiert. Der Bereich A stellt die in den vorhergehenden Abschnitten untersuchte Stichprobe dar. Der Bereich B liegt am Übergang von der Tailverteilung zur Mainverteilung und der Bereich C am unteren Ende der Mainverteilung.

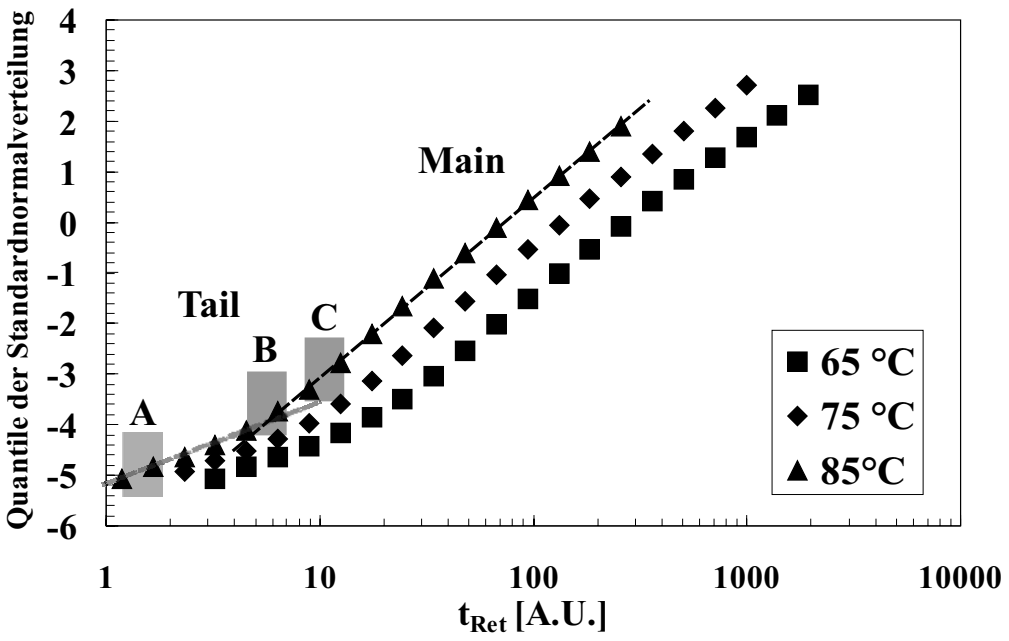


Abbildung 7.14: Für Messungen der Aktivierungsenergieverteilung wurden Stichproben von Zellen aus den Bereich A, B und C der Retentionkurve bei $85\,°C$ ausgewählt. Abbildung 7.15 zeigt die Aktivierungsenergieverteilungen der Stichproben.

Die kumulativen Verteilungen der Aktivierungsenergien für die drei Stichproben sind in Abbildung 7.15 dargestellt. Die Verteilung der Aktivierungsenergien verschiebt sich insgesamt mit längeren Retentionzeit der Stichprobe hin zu höheren Aktivierungsenergien. Zusätzlich geht die Verteilung von einer Mischverteilung im Tailbereich in eine Normalverteilung in der Mainverteilung über. Dies zeigt sich am kumulativen Anteil der Unterverteilung mit niederen Aktivierungsenergien. Am Punkt A hatte die niedere Unterverteilung einen Anteil von 78% der Stichprobe. Im Punkt B weniger als 10 % und

im Punkt C weniger als 0,01%. Da der niedrigere Anteil der Verteilung *GIDL* ausmacht, kann daraus geschlossen werden, dass *GIDL* im Tail eine große Rolle spielt und in der Mainverteilung vernachlässigbar ist.

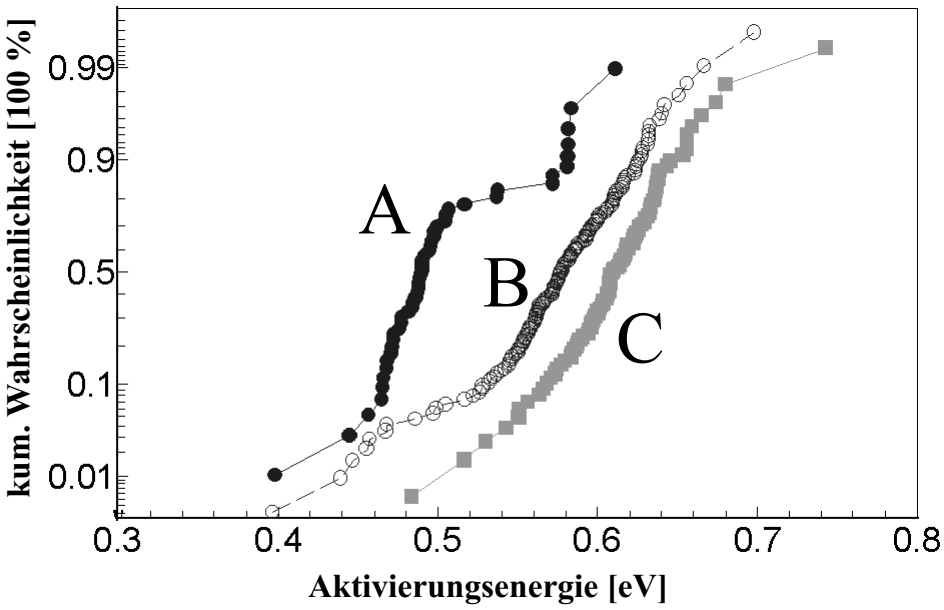


Abbildung 7.15: Verteilung der Aktivierungsenergien für Stichproben mit verschiedenen Retentionzeiten (siehe Abbildung 7.14). Die Anteil der Unterverteilung mit niedrigen Aktivierungsenergien verschwindet beim Übergang von der Tail zur Mainverteilung. Insgesamt verschiebt sich die Verteilung der Aktivierungsenergien für höhere Retentionzeiten hin zu höheren Werten.

Zusammenfassung der Ergebnisse aus Einzelzell-Messungen

Durch Einzelzellmessungen können Vertauschungseffekte durch die Kenntnis der Zelladresse ausgeschlossen werden. Dadurch kann die Aktivierungsenergie einer Zelle und damit des dominanten Leckstroms genauer bestimmt werden. Es ergibt sich für eine Stichprobe von Zellen nahe der Reparaturgrenze ein breite Verteilung der Aktivierungsenergien ($0.4\,eV - 0.6\,eV$), die sich aus zwei Unterverteilungen zusammensetzt und somit auf verschiedene Mechanismen und Pfade schließen lässt. Es muss betont werden, dass die breite Verteilung nur durch Einzelzellmessungen erhalten werden kann. Die konventionelle Bestimmung aus Retentionkurven liefert nur einen Wert, der am unteren Ende der tatsächlichen Verteilung liegt. Der größere Anteil der Zellen (78%) gehört zu einer Normalverteilung mit niedrigem Median. Alle Aktivierungsenergien von Zellen nahe der Reparaturgrenze sind spannungsabhängig, wobei die niedere Unterverteilung auf V_{NWLL} und die höhere Unterverteilung auf V_{BB} reagiert. Aus Stichproben entlang der Retentionkurve folgen für höhere Retentionzeiten größere Aktivierungsenergien. Außerdem nimmt der niedere V_{NWLL}-sensitive Anteil entlang des Tails ab und verschwindet beim Übergang von Tail zu Main ganz. Zusammen mit der Spannungsabhängigkeit der niederen Aktivierungsenergieverteilung folgt daraus, dass GIDL den Tail der Retentionverteilung dominiert und im Main nur eine geringe Rolle spielt.

Kapitel 8

Ergebnisse der theoretischen Betrachtung & Abschätzung

Aus der elektrischen Charakterisierung in Kapitel 7 gingen die Leckstrompfade *Junction Leakage* und *GIDL* als für die Retentionverteilung bestimmend hervor. Dieses Kapitel beschäftigt sich mit den theoretischen Hintergründen der Leckstrommechanismen im pn-Übergang. Im Speziellen spielt die Temperaturabhängigkeit der verschiedenen Mechanismen eine zentrale Rolle zur Identifikation der wichtigsten Prozesse. Der Leckstrom eines realen pn-Übergangs kann durch eine Reihe zu benennender Effekte um Größenordnungen über dem des idealen Übergangs liegen. Es werden die pn-Leckstrommechanismen vorgestellt und deren Spannungs- und Temperaturabhängigkeiten besprochen. Für alle Mechanismen wird der Betrag des maximalen Leckstroms mit Hilfe eines vereinfachten Modells abgeschätzt. Um den in Kapitel 7 bestimmten Aktivierungsenergien Mechanismen zuordnen zu können, müssen für alle bekannten Mechanismen diese rechnerisch bestimmt werden.

8.1 Einfaches Modell zur Abschätzung

Für eine grobe Abschätzung der Leckströme der verschiedenen Mechanismen in den folgenden Abschnitten wird der kondensatorseitige pn-Übergang durch einen einfachen abrupten pn-Übergang mit homogenen Dotierstoffverteilungen von $N_A = 6 \cdot 10^{17} cm^{-3}$ und $N_D = 5 \cdot 10^{18}\, cm^{-3}$ ersetzt. Die Länge des Übergangs ergibt sich aus der Länge des *Buried Straps* (y-Richtung: $\sim 210\, nm$, x-Richtung: $\sim 60\, nm$) zu $270\, nm$ (vgl. Abbildung 8.1). Die Breite entspricht der Breite des aktiven Gebietes ($110\, nm$), die senkrecht zur Zeichnungsebene verläuft. Die Breite der Verarmungszone wird im Mittel mit $W = 60\, nm$ angenommen. Aus diesen Bemaßungen ergibt sich eine Diodenfläche $A = (270 \cdot 110)\, nm^2 \approx 3 \cdot 10^{-10}\, cm^2$ und ein Volumen der Verarmungszone $V_j = A \cdot W = 3 \cdot 10^{-10}\, cm^2 \cdot 60\, nm \approx$

$1.8 \cdot 10^{-15}\, cm^3$. Die Grenzfläche des Verarmungsgebietes zum Oxid (A_S) errechnet sich aus der doppelten Fläche der Verarmungszone in Abbildung 8.1 (Schnittflächen zum STI-Oxid vor und hinter der Schnittebene) plus zwei $60\, nm$ breite und $110\, nm$ lange Streifen senkrecht zur Zeichnungsebene (entlang des Collars und des Gateoxids direkt unterhalb des Gate-Spacers). Die Interfacefläche ist demnach $A_S = (2 \cdot 270 \cdot 60 + 2 \cdot 60 \cdot 110)\, nm^2 = 4.56 \cdot 10^{-10}\, cm^2$.

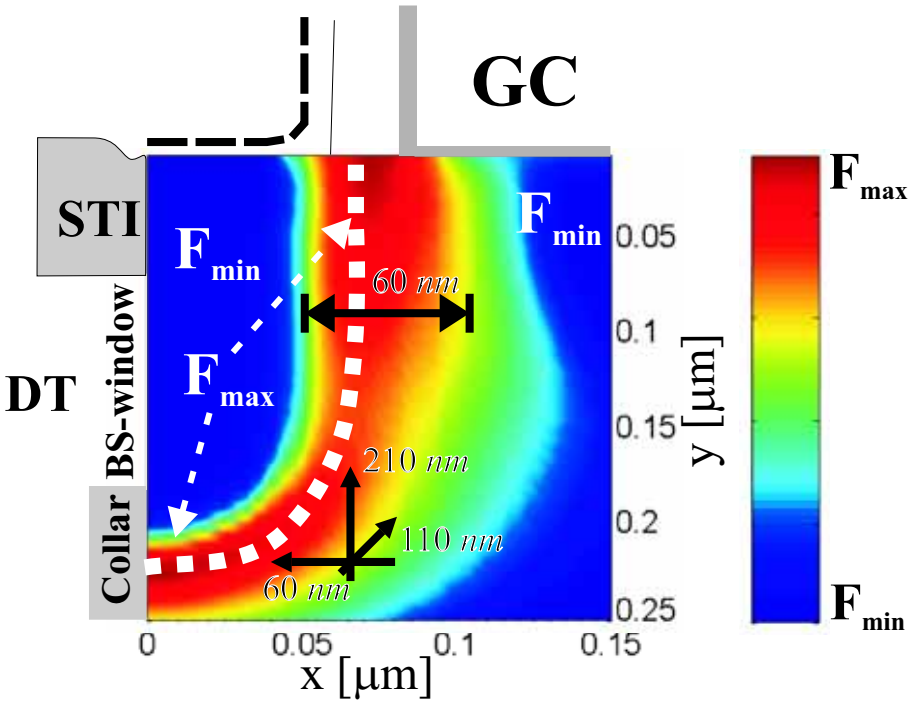


Abbildung 8.1: Vereinfachtes Modell des kondensatorseitigen pn-Übergangs: abrupter Übergang mit einer Fläche $A = 270 \cdot 110\, nm^2 \approx 3 \cdot 10^{-10}\, cm^2$ und einem Volumen der Verarmungszone von $V_j = 270 \cdot 110 \cdot 60\, nm^3 \approx 1.8 \cdot 10^{-15}\, cm^3$.

8.2 Leckstrom des idealen pn-Übergangs

Selbst beim idealen pn-Übergang fließt ein Leckstrom in Sperrrichtung. Dieser ergibt sich für Sperrspannungen größer als ein paar $k_B T/q$ durch Lösen der Stromgleichungen für Diffusion- und Driftstrom unter Vernachlässigung von Generation und Rekombination in der Verarmungszone (siehe z.B. [Pie96, S. 249]):

$$I_{ideal} = -qA\left(\frac{D_N}{L_N}\frac{n_i^2}{N_A} + \frac{D_P}{L_P}\frac{n_i^2}{N_D}\right) \tag{8.1}$$

$D_{N,P}$ und $L_{N,P}$ sind hierbei die Diffusionskonstanten für Elektronen und Löcher sowie deren Diffusionslängen, $N_{A,D}$ die Dotierstoffkonzentrationen von Akzeptoren und Donatoren, q die Elementarladung und n_i die intrinsische Ladungsträgerdichte. Der Leckstrom eines idealen pn-Übergangs in Sperrrichtung ist demzufolge spannungsunabhän-

gig und proportional zur Diodenfläche A. Für die Dotierstoffkonzentrationen unseres vereinfachten Modells ergeben sich die Beweglichkeiten $\mu_n = 180\,cm^2/Vs$ und $\mu_p = 210\,cm^2/Vs$ (siehe z.B. [Pie03, S.185,186]). Daraus ergeben sich über die Einstein-Relation $D_{N,P} = \mu_{n,p} k_B T/q$ die Diffusionskonstanten bei $85\,°C$ und unter Annahme der kleinsten typischen Minoritätslebensdauer im einkristallinem Silizium $\tau_{n,\,p} = 1 \cdot 10^{-6}$ [Pie03, S.163] deren Diffusionslängen. Die intrinsische Ladungsträgerdichte bei $85\,°C$ ist näherungsweise $n_i = 1 \cdot 10^{13}\,cm^{-3}$. Einsetzen in Gleichung 8.1 ergibt den maximalen Leckstrom des kondensatorseitigen pn-Übergangs in Abwesenheit von Generation aus der Verarmungszone in Höhe von $I_{ideal} \approx 0.03\,fA$. Dieser ist eine Größenordnung kleiner als die kleinsten aus Retentionmessungen bestimmten Gesamtleckströme (vgl. Abschnitt 4.3). Zusätzliche Leckstrompfade bzw. Mechanismen überwiegen den idealen Leckstrom in praktisch allen Speicherzellen.

Temperaturabhängigkeit

Da die Temperaturabhängigkeit in Gleichung 8.1 von $n_i^2 \propto exp(-E_G/k_B T)$ dominiert wird, wobei E_G die Bandlückenbreite ist, ergibt der Vergleich mit der Arrhenius-Gleichung $I_{Leak} \propto exp(-E_a/k_B T)$ die Aktivierungsenergie für den Diffusionsstrom im idealen pn-Übergang $E_{a,\,ideal} = E_G = 1.12\,eV$. Diese liegt weit über allen gemessenen Aktivierungsenergien, wodurch der geringe Anteil des Diffusionsstromes am Gesamtleckstrom nochmals bestätigt wird.

8.3 Thermische Generation (SRH)

Jede Generation von Ladungsträgern in der Verarmungszone führt zu einem zusätzlichen Sperrstrom. Zustände in der Verarmungszone mit Energien innerhalb der Bandlücke sind nach Shockley, Read und Hall Grund für stark erhöhte Ladungsträgergeneration [Sho52, Hal52]. In diesem Abschnitt wird zuerst die Rekombinationsgleichung aufgrund ihrer grundlegenden Bedeutung und zur Illustration der bei Generationsströmen vorherrschenden Prozesse formal hergeleitet (siehe z.B. auch [Sze81, Pie96]). Danach findet diese Anwendung auf den in Sperrrichtung gepolten pn-Übergang. Im Weiteren wird die Temperaturabhängigkeit und deren Abhängigkeit vom elektrischen Feld untersucht.

8.3.1 Rekombinationsgleichung

In Halbleitern können Ladungsträger über die in Abbildung 8.2 dargestellten thermischen Prozesse ihre Lage bezüglich der Bänder ändern. Zu jedem Zeitpunkt finden alle Prozesse mit einer durch die Fermi-Verteilung bestimmten Wahrscheinlichkeit statt.

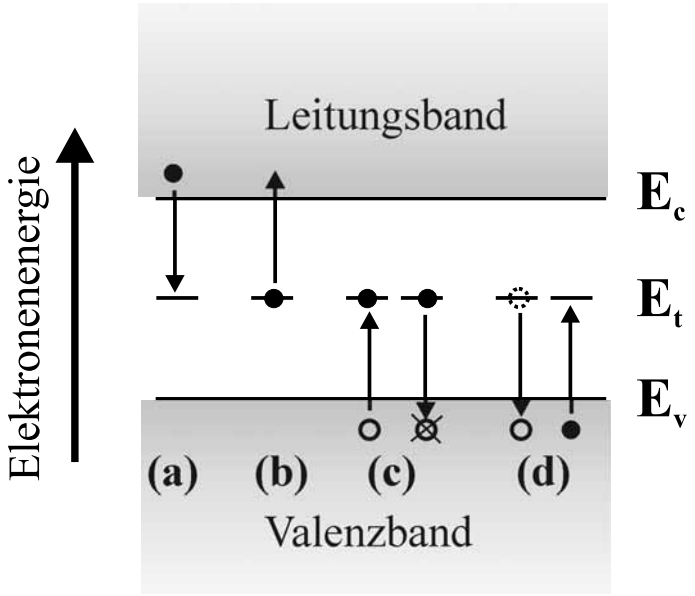


Abbildung 8.2: Thermische Übergänge in Halbleitern (nach [Hal51, Hal52, Sho52])

(a) Elektroneneinfang, (b) Elektronenemission, (c) Locheinfang, (d) Lochemission

Mit Hilfe der in Abbildung 8.2 aufgezeigten Prozesse kann die zeitliche Änderung der Elektronenkonzentration im Leitungsband sowie der Löcherkonzentration im Valenzband aufgestellt werden.

$$\frac{\partial n}{\partial t} = \left.\frac{\partial n}{\partial t}\right|_{(a)} + \left.\frac{\partial n}{\partial t}\right|_{(b)} \tag{8.2}$$

$$\frac{\partial p}{\partial t} = \left.\frac{\partial p}{\partial t}\right|_{(c)} + \left.\frac{\partial p}{\partial t}\right|_{(d)} \tag{8.3}$$

Dabei hängt die Wahrscheinlichkeit für Elektroneneinfang (Prozess a) von der Elektronendichte im Leitungsband und der Konzentration freier Trapzentren ab. Ist eines von beidem zu klein, geht die Elektroneneinfangrate gegen null. Verdoppelt sich eines von beidem, so verdoppelt sich auch die Einfangrate. Daraus folgt die direkte Proportionalität zur Elektronendichte n im Leitungsband und zur Konzentration nicht gefüllter Trapzustände p_t. Die Proportionalitätskonstante wird als Elektroneneinfang-Koeffizient c_n bezeichnet.

$$\left.\frac{\partial n}{\partial t}\right|_{(a)} = -c_n p_t n \tag{8.4}$$

Dagegen ist die Emission von Elektronen (Prozess b) proportional zu der Konzentration besetzter Trapzentren n_t und den freien Zuständen im Leitungsband. Letzteres stellt

aufgrund der Vielzahl von Zuständen im Leitungsband jedoch keine Einschränkung dar und die Emissionsrate ist deshalb direkt proportional zu n_t. Die Proportionalitätskonstante wird als Emissionskoeffizient e_n bezeichnet.

$$\left.\frac{\partial n}{\partial t}\right|_{(b)} = e_n n_t \tag{8.5}$$

Aus analogen Überlegungen folgen auch für die Prozesse c) und d) die entsprechenden Übergangsraten:

$$\left.\frac{\partial p}{\partial t}\right|_{(c)} = -c_p n_t p \tag{8.6}$$

$$\left.\frac{\partial p}{\partial t}\right|_{(d)} = e_p p_t \tag{8.7}$$

Die Definition der Rekombinationsraten für Elektronen r_n und Löcher r_p als die zeitliche Abnahme von Ladungsträgern in den entsprechenden Bändern ergibt unter Verwendung der bisher aufgestellten Beziehungen:

$$r_n \equiv -\frac{\partial n}{\partial t} = -\left(\left.\frac{\partial n}{\partial t}\right|_{(a)} + \left.\frac{\partial n}{\partial t}\right|_{(b)}\right) = c_n p_t n - e_n n_t \tag{8.8}$$

$$r_p \equiv -\frac{\partial p}{\partial t} = -\left(\left.\frac{\partial p}{\partial t}\right|_{(c)} + \left.\frac{\partial p}{\partial t}\right|_{(d)}\right) = c_p n_t p - e_p p_t \tag{8.9}$$

Für weitere Betrachtungen müssen die in Abbildung 8.3 gezeigten zwei Gleichgewichtsfälle unterschieden werden. Im *thermischen Gleichgewicht* (Abbildung 8.3a) sind die jeweils entgegengesetzten Teilprozesse, d.h. Elektronenemission und Elektroneneinfang sowie Lochemission und Locheinfang, gleich wahrscheinlich und die Elektronen- und Lochrekombinationsraten r_n und r_p sind null, da sich die Teilprozesse jeweils paarweise aufheben. Dadurch fließt kein gerichteter Strom und überall stellt sich eine zeitlich konstante Ladungsträgerdichte ein. Durch Einsetzen in Gleichungen 8.8 und 8.9 können Relationen zwischen Emissions- und Einfangskoeffizienten abgeleitet werden:

$$e_{n0} = \frac{c_{n0} p_{t0} n_0}{n_{t0}} = c_{n0} n_1 \tag{8.10}$$

$$e_{p0} = \frac{c_{p0} n_{t0} p_0}{p_{t0}} = c_{p0} p_1 \tag{8.11}$$

Der Index 0 bezeichnet hier den thermischen Gleichgewichtsfall. Die Konstanten

$$n_1 = \frac{p_{t0} n_0}{n_{t0}} \tag{8.12}$$

$$p_1 = \frac{n_{t0}p_0}{p_{t0}} \tag{8.13}$$

können mit Hilfe der Besetzungswahrscheinlichkeit von Zuständen in der Bandlücke n_{t0}/N_t, welche durch die Fermi-Verteilung beschrieben wird,

$$\frac{n_{t0}}{N_t} = \frac{1}{1+\exp(\frac{E_t-E_F}{k_BT})} \tag{8.14}$$

und den Ladungsträgerdichten n_0 und p_0, die sich aus der Lage des Fermi-Niveaus bezüglich der intrinsischen Energie E_i ergeben,

$$n_0 = n_i \exp(\frac{E_F - E_i}{k_BT}) \tag{8.15}$$

$$p_0 = n_i \exp(\frac{E_i - E_F}{k_BT}) \tag{8.16}$$

einfach berechnet werden zu

$$n_1 = \frac{p_{t0}n_0}{n_{t0}} = (\frac{N_t - n_{t0}}{n_{t0}})n_0 = (\frac{N_t}{n_{t0}} - 1)n_0 = n_i \exp(E_t - E_i)/k_BT \tag{8.17}$$

$$p_1 = \frac{n_{t0}p_0}{p_{t0}} = (\frac{n_{t0}}{N_t - n_{t0}})p_0 = (1/(\frac{N_t}{n_{t0}} - 1))p_0 = n_i \exp(E_i - E_t)/k_BT \tag{8.18}$$

Dadurch können die Emissionskoeffizienten aus den Rekombinationsgleichungen eliminiert werden:

$$r_n \equiv -\left.\frac{\partial n}{\partial t}\right|_{R-G} = c_n(p_t n - n_t n_1) \tag{8.19}$$

$$r_p \equiv -\left.\frac{\partial p}{\partial t}\right|_{R-G} = c_p(n_t p - p_t p_1) \tag{8.20}$$

Im Fall des *dynamischen Gleichgewichts* (Abbildung 8.3b) heben sich im Gegensatz zum *thermischen Gleichgewicht* die komplementären Prozesse nicht auf ($r_n,\ r_p \neq 0$). Trotzdem müssen sich zeitlich konstante Ladungsträgerdichten einstellen. Dies gilt insbesondere auch für die Anzahl der belegten Traps:

$$\frac{dn_t}{dt} = -\frac{\partial n}{\partial t} + \frac{\partial p}{\partial t} = r_n - r_p = 0 \tag{8.21}$$

Daraus folgt, dass die Rekombinationsraten für Elektronen und Löcher betraglich gleich sein müssen ($r_n = r_p$). Deshalb begrenzt immer die kleinere Rate der beiden Teilprozesse (d.h. der unwahrscheinlichere Teilprozess) die Wahrscheinlichkeit für den Gesamtprozess der Rekombination bzw. Generation. Unter der Annahme, dass die Rekombinati-

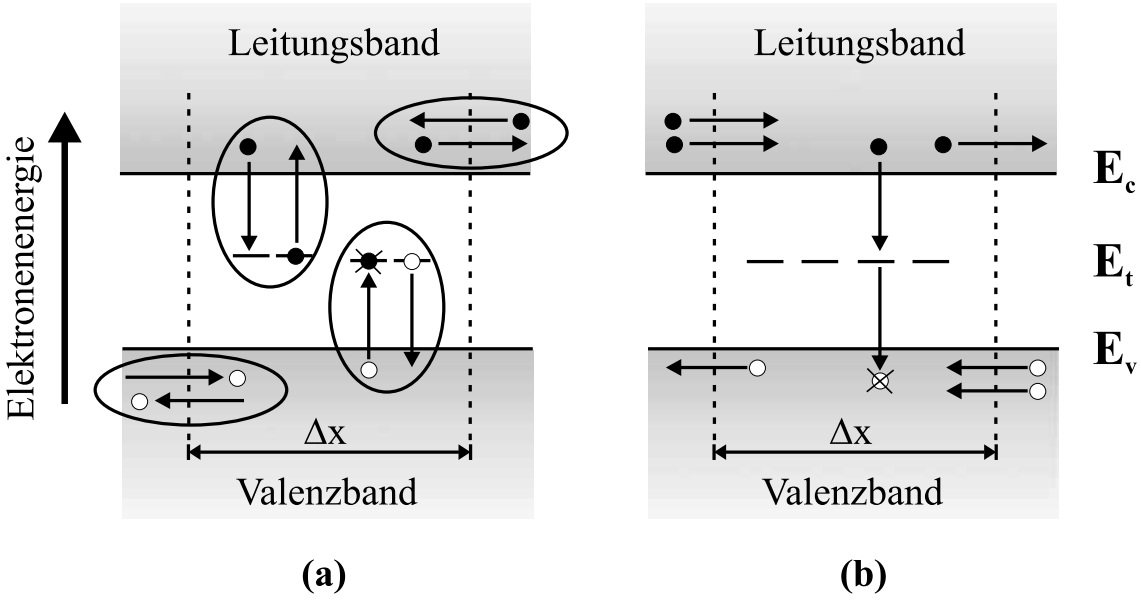


Abbildung 8.3: Vergleich zwischen (a) *thermischem Gleichgewicht* und (b) *dynamischem Gleichgewicht* (nach [Pie03, S.148]). Im *thermischen Gleichgewicht* gleichen sich die eingekreisten Teilprozesse paarweise aus, während im *dynamischen Gleichgewicht* die Konstanz der Ladungsträgerdichten n, p und n_t durch Ausgleichseffekte gänzlich verschiedener Prozesse einsteht.

onsraten des *thermischen Gleichgewichts* auch für den Fall des *dynamischen Gleichgewichts* ihre Gültigkeit behalten, kann n_t durch Gleichsetzen der beiden Gleichungen 8.19 und 8.20 bestimmt werden zu:

$$n_t = \frac{c_n N_t n + c_p N_t p_1}{c_n(n + n_1) + c_p(p + p_1)} \tag{8.22}$$

Einsetzen in Gleichung 8.19 ergibt dann die Rekombinationsrate im *dynamischen Gleichgewicht*:

$$R = r_n = r_p = \frac{np - n_i^2}{\frac{1}{c_p N_t}(n + n_1) + \frac{1}{c_n N_t}(p + p_1)} \tag{8.23}$$

Diese stellt eine sehr wichtige Gleichung in der Mikroelektronik dar. Ein positiver Wert steht für Rekombination, wohingegen ein negativer Wert für die Generation von Ladungsträgern steht. Durch Einsetzen der Minoritätslebensdauer $\tau_{n,p} = 1/(\sigma_{n,p} v_{th} N_t) = 1/(c_{n,p} N_t)$ in die Rekombinationsgleichung, erhält man noch zwei weitere oft in der Literatur verwendete Formen der Rekombinationsgleichung:

$$R = \frac{pn - n_i^2}{\tau_p(n + n_1) + \tau_n(p + p_1)} \tag{8.24}$$

$$R = \frac{\sigma_p \sigma_n v_{th} N_t (pn - n_i^2)}{\sigma_n(n + n_i exp(\frac{E_t - E_i}{k_B T})) + \sigma_p(p + n_i exp(\frac{E_i - E_t}{k_B T}))} \tag{8.25}$$

8.3.2 Generation im pn-Übergang

Die im letzten Abschnitt hergeleitete Rekombinationsrate gilt zunächst überall. Der für diese Arbeit interessante Fall ist die Generation von Ladungsträgern in der Verarmungszone des kondensatorseitigen pn-Übergangs einer DRAM-Zelle. Abbildung 8.4 zeigt die Situation vereinfacht im Banddiagramm. Bei der Generation eines Ladungsträgerpaares wird ein Elektron ins Leitungsband und ein Loch ins Valenzband emittiert. Das elektrische Feld der Verarmungszone trennt das neu entstandene Ladungsträgerpaar sofort räumlich und verhindert dadurch den Rekombinationsprozess. Das Elektron bewegt sich durch das Driftfeld zur n-Seite, das Loch zur p-Seite des Übergangs. Insgesamt entsteht pro Generationsprozess ein zusätzliches Elektron auf der n-Seite und ein zusätzliches Loch auf der p-Seite, das gleichbedeutend mit dem Transport eines Elektrons von der p zur n-Seite oder mit einem Strom in Sperrrichtung ist. Hervorzuheben ist hierbei, dass für einen Sperrstrom beide Generationsteilprozesse, d.h. sowohl Elektron- als auch Lochemission notwendig sind.

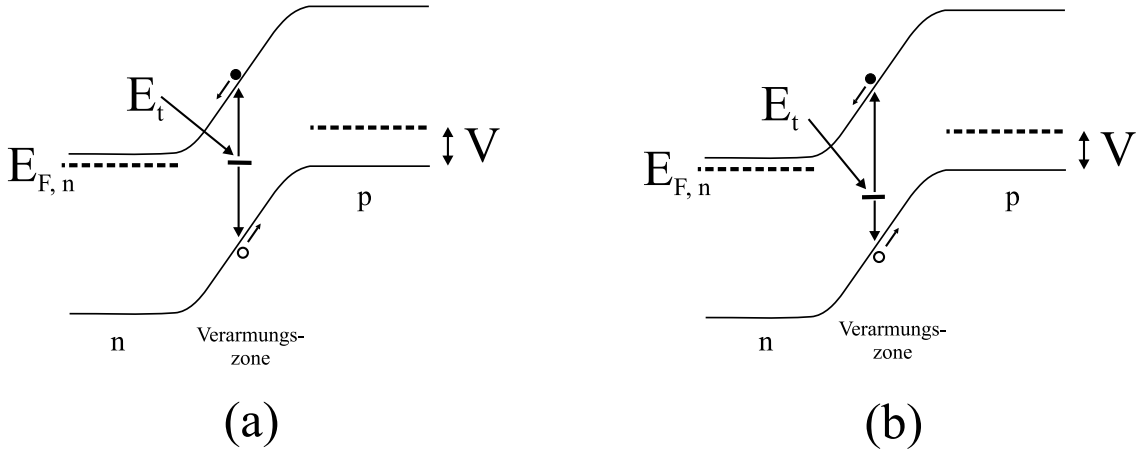


Abbildung 8.4: Banddiagramm für SRH-Generationsstrom mit Trap (a) in Bandlückenmitte, (b) außerhalb der Bandlückenmitte.

Der Generationsstrom für den vereinfachten Fall eines stufenartigen pn-Übergangs kann für Sperrspannungen größer als wenige k_BT/q einfach berechnet werden. In diesem Fall können die Ladungsträgerdichten in der Raumladungszone vernachlässigt werden und Gleichung 8.24 vereinfacht sich zu

$$R = -\frac{n_i^2}{\tau_p n_1 + \tau_n p_1} \tag{8.26}$$

Das negative Vorzeichen zeigt, dass in der Verarmungszone die Generation von Ladungsträgern die Rekombination überwiegt. Der Generationsstrom kann daraus durch Integration über die Raumladungszone berechnet werden:

$$\begin{aligned} I_{SRH} &= qA\int_{-x_p}^{x_n} Rdx = qAWR = -qAWn_i \frac{1}{\frac{\tau_p n_1}{n_i}+\frac{\tau_n p_1}{n_i}} \\ &= \frac{-qAWn_i}{\tau_p \exp(E_t-E_i)/k_BT+\tau_n \exp(E_i-E_t)/k_BT} \end{aligned} \tag{8.27}$$

mit der Breite der Verarmungszone

$$W = \sqrt{\frac{2\epsilon_{Si}}{q}\frac{(N_a+N_d)}{N_aN_d}(V_{bi}-V)} \tag{8.28}$$

und der eingebauten Spannung

$$V_{bi} = \frac{k_BT}{q} ln(\frac{N_aN_d}{n_i^2}) \tag{8.29}$$

Der Generationsstrom ist demnach proportional zur Breite der Verarmungszone, welche wiederum proportional zur Wurzel der Sperrspannung V ist. Die Verarmungszone wird außerdem mit zunehmenden Dotierstoffkonzentrationen kleiner. Der maximale Generationsstrom ergibt sich für ein Trap in der Mitte der Bandlücke. Durch Einsetzen der Werte für das vereinfachte Modell (siehe Abschnitt 8.1) sowie $\tau = 1 \cdot 10^{-6}\, s$ und $n_i = 1 \cdot 10^{13}\, cm^{-3}$ ergibt sich ein maximaler SRH-Leckstrom eines Traps bei $85\,°C$ von $I_{SRH,\,max} = -q{\cdot}A{\cdot}W{\cdot}n_i/2{\cdot}\tau_{n,\,p} \approx 1.44\, fA$. Der Leckstrom fällt exponentiell mit dem energetischen Abstand des verursachenden Traps von der Bandlückenmitte ab. Befinden sich mehrere Traps im pn-Übergang einer Zelle, müssen die Generationsströme addiert werden. Es wird vermutet, dass nur relativ wenige Traps den Gesamtleckstrom in den pn-Übergängen heutiger Speicherzellen bestimmen. Diese Aussage rührt von typischen Interfacezustandsdichten von einigen $10^{10}\, cm^{-2}$ (abhängig von Oxidations- und Annealbedingungen) und einer Grenzfläche der Verarmungszone zum Oxid von $A_S \approx 5{\cdot}10^{-10}\, cm^2$ her. Zustandsdichten im kristallinem Silizium selbst liegen üblicherweise darunter.

Temperaturabhängigkeit

Betrachtet man zunächst den Generationsstrom für Störstellen in der Bandlückenmitte, also $E_t = E_i$, so verschwinden die temperaturabhängigen Terme im Nenner der Gleichung 8.27 und die Temperaturabhängigkeit des Generationsstromes reduziert sich auf die von n_i. Diese kann direkt aus der Gleichung für n_i zu $E_G/2$ abgelesen werden. Liegen die Störstellen außerhalb der Bandlückenmitte, muss auch der Nenner in Betracht gezogen werden. Je nachdem in welcher Bandlückenhälfte sich die Störstelle befindet, kann jeweils einer der beiden Exponentialterme im Nenner vernachlässigt werden. Der verbleibende Term bestimmt die Temperaturabhängigkeit mit. In diesem Fall können die Exponenten der Exponentialfunktionen von Zähler und Nenner zusammengefasst werden und es ergibt sich eine Aktivierungsenergie von $E_{a,\,SRH} = E_G/2 + |E_t - E_i|$. Die geringe Temperaturabhängigkeit von V_{bi} und damit der Verarmungszonen-

breite W kann demgegenüber vernachlässigt werden. Numerische Rechnungen mittels MATLAB bestätigen die Abschätzung (siehe Abbildung 8.5). Die Abbildung zeigt die berechnete Aktivierungsenergie gegenüber der Position eines Traps in der Bandlücke. Die niedrigste auftretende Aktivierungsenergie entsteht durch ein Trap in Bandlückenmitte und entspricht betraglich der halben Bandlückenbreite. Für Traps außerhalb der Bandlückenmitte nimmt die Aktivierungsenergie genau um den Abstand des Traps zur Bandlückenmitte zu. Im Banddiagramm von Abbildung 8.4 kann die Aktivierungsenergie des Leckstroms somit direkt der Länge des längeren Teilprozesses (Elektron- oder Lochemission) entnommen werden. Die physikalische Begründung dafür liegt darin, dass für einen resultierenden Leckstrom beide Teilprozesse notwendig sind. Der Teilprozess mit der größeren benötigten thermischen Anregung und damit der kleineren Wahrscheinlichkeit limitiert den Gesamtprozess und bestimmt dadurch auch dessen Temperaturabhängigkeit. Die anschauliche Betrachtungsweise im Banddiagramm wird im Folgenden noch mehrfach verwendet werden.

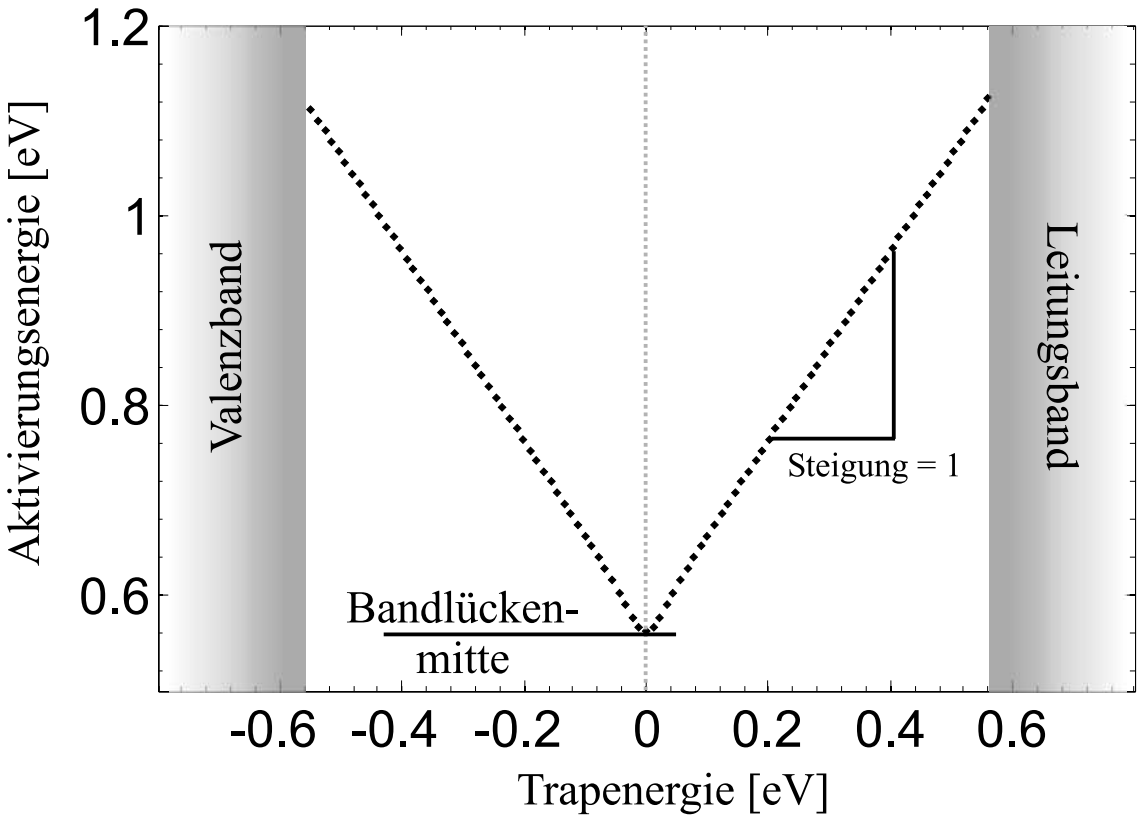


Abbildung 8.5: Abhängigkeit der Aktivierungsenergie des SRH-Generationsstroms von der Trapenergie. Die kleinste auftretende Aktivierungsenergie entsteht für ein Trap in der Bandlückenmitte und entspricht betraglich der halben Bandlückenbreite.

Die Sperrspannung V geht in den SRH-Generationsstrom nur durch die Verarmungszonenbreite W ein. Eine höhere Sperrspannung vergrößert somit das Volumen der Verarmungszone und dadurch auch den Gesamtleckstrom. Die Temperaturabhängigkeit und

deshalb auch die Aktivierungsenergie ist jedoch unabhängig von der Sperrspannung. Der Zusammenhang kann entweder den Formeln entnommen werden oder ist direkt im Banddiagramm sichtbar. Dazu zeigt Abbildung 8.6 das Banddiagramm für zwei unterschiedliche Sperrspannungen $V_2 > V_1$. Da SRH nur in thermischer (=vertikale) Generation besteht und der vertikale Bandabstand von der Spannung unabhängig ist, folgt daraus auch die Spannungsunabhängigkeit der Aktivierungsenergie im SRH-Fall.

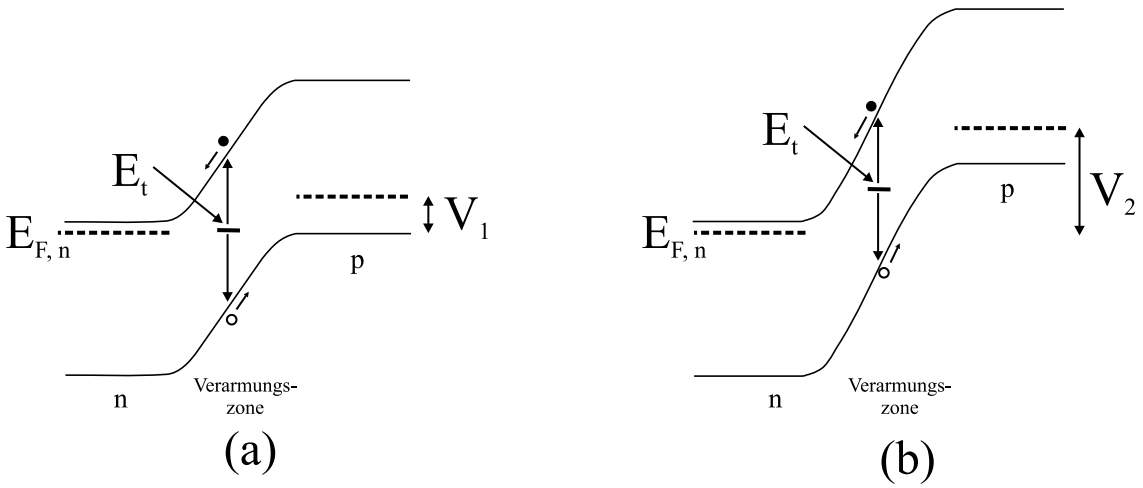


Abbildung 8.6: SRH-Generation im Banddiagramm für verschiedene Sperrspannungen und damit verschiedenen elektrischen Feldern in der Verarmungszone. Da SRH nur in thermischer (=vertikale) Generation besteht und der vertikale Bandabstand von der Spannung unabhängig ist, ist auch die Aktivierungsenergie im SRH-Fall spannungsunabhängig.

8.4 Tunnelunterstützte Generation (TFE)

Mit der Strukturverkleinerung müssen auch die Verarmungszonen der pn-Übergänge kleiner dimensioniert werden, um Kurzkanaleffekten vorzubeugen. Dazu sind zunehmend höhere Dotierungen notwendig, die unweigerlich zu höheren elektrischen Feldstärken F führen. Typische Feldstärken im DRAM liegen mittlerweile im Bereich von $(2-8)\cdot 10^5\, V/cm$. Experimentell werden sehr hohe Leckströme beobachtet, die durch thermische Generation alleine nur schwer erklärt werden können. Neue Effekte bedingt durch die hohen Feldstärken scheinen eine Rolle zu spielen. *Thermionic Field Emission* (TFE), ein um Tunneleffekte erweitertes Generations-Modell, wurde in [Hur92] vorgestellt. Abbildung 8.7 zeigt die Situation im Banddiagramm. Anders als bei der SRH-Generation besteht TFE aus einer nur teilweisen thermischen Generation bis zu einem Punkt P gefolgt von Tunneln durch die verbleibende Barriere ins Leitungsband bzw. Valenzband. Aufgrund der großen Anzahl von freien Zuständen in den Bändern, ist dabei das Tunneln alleine durch die Barrierendicke beschränkt.

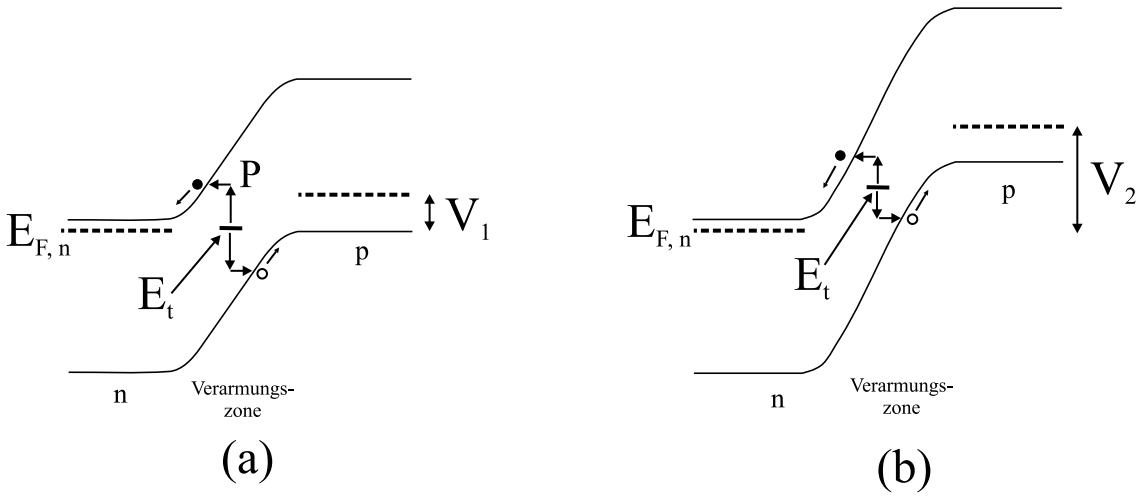


Abbildung 8.7: TFE-Generation im Banddiagramm für unterschiedliche Felder.

Rechnerisch geht die Tunnelkomponente durch eine feldabhängige Minoritätslebensdauer in die Rekombinations-Generations-Formel ein [Hur92]:

$$\tau_{TFE,n,p} = \frac{\tau_{n,p}}{1 + \Gamma_{n,p}} \tag{8.30}$$

Der Feldfaktor Γ kann für elektrische Felder kleiner $9 \cdot 10^5\, V/cm$ (also für DRAM typische Felder) analytisch bestimmt werden zu

$$\Gamma_{n,p} = 2\sqrt{3\pi}\frac{|F|}{F_\Gamma} \exp\left[(\frac{F}{F_\Gamma})^2\right] \tag{8.31}$$

Hierbei ist F die elektrische Feldstärke und F_Γ gegeben durch

$$F_\Gamma = \frac{\sqrt{24 m^* (k_B T)^3}}{q\hbar} \tag{8.32}$$

Das Einsetzen der modifizierten Lebensdauer in die Stromgleichung 8.27 zeigt, dass der SRH-Generationsstrom dadurch um den Faktor $1 + \Gamma$ erhöht wird.

$$I_{TFE} = (1 + \Gamma(F)) \cdot I_{SRH} \tag{8.33}$$

Abbildung 8.8 zeigt den Verlauf des Feldfaktors Γ bei $85\,°C$ in Abhängigkeit von der elektrischen Feldstärke F. Für typische Feldwerte im DRAM ergeben sich Werte für Γ zwischen 1 und 200. Dabei ist hervorzuheben, dass Γ ab einer kritischen Feldstärke $F_\Gamma = 4.8 \cdot 10^5\, V/cm$ stark zunimmt. Bei dieser kritischen Feldstärke hat Γ schon einen Wert von $\Gamma_{n,p}(F_\Gamma) = 2\sqrt{3\pi}e \approx 16.7$ erreicht. Nach Gleichung 8.33 überträgt sich dies direkt auf den Leckstrom. Das bedeutet konkret, dass Leckströme aus dem Mainbereich (fA-Bereich) durch den tunnelunterstützten Mechanismus TFE bis in den Tailbereich (zwei bis dreistelliger fA-Bereich) erhöht werden können.

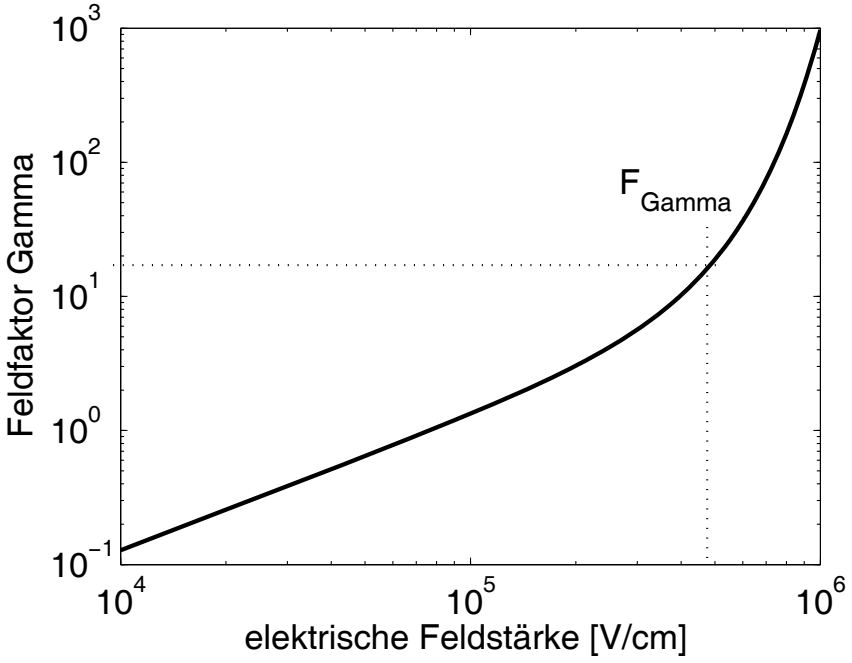


Abbildung 8.8: Feldfaktor Γ bei $85\,°C$ in Abhängigkeit vom elektrischen Feld.

Temperaturabhängigkeit

Abbildung 8.7 zeigt die Situation im Banddiagramm für zwei verschiedene Sperrspannungen $V_2 > V_1$. Im Gegensatz zur SRH-Generation beeinflusst das elektrische Feld den Emissionspfad von Elektronen und Löchern. Da die Wahrscheinlichkeit für den Tunnelanteil bei der Emission nur von der Barrierendicke abhängt, führt deren Modulation zu spürbaren Veränderungen. Vergrößert sich die Sperrspannung über den pn-Übergang, folgt daraus eine größere Feldstärke in der Verarmungszone, die sich durch eine größere Bandverbiegung ausdrückt. Dadurch wird die für die Ladungsträger zu durchtunnelnde Barriere dünner (vgl. Abbildung 8.7a und b). Deshalb setzt das Tunneln bereits nach geringerer thermischer Anregung ein und eine erniedrigte Aktivierungsenergie ist die Folge. Die Relation zwischen Trapposition, Aktivierungsenergie und elektrischem Feld kann in einem Diagramm dargestellt werden (siehe Abbildung 8.9). Für ein elektrisches Feld von $1 \cdot 10^5\, V/cm$ liegen die Aktivierungsenergien sehr nahe am Grenzfall des feldunabhängigen SRH-Generationsstroms. Mit zunehmenden elektrischen Feldern schiebt sich die gesamte Kurve hin zu kleineren Aktivierungsenergien. Dabei treten für Zustände in der Bandlückenmitte und elektrischen Feldern von $8 \cdot 10^5\, V/cm$ Aktivierungsenergien bis runter zu $\sim 0.15\, eV$ auf. Die verwendete analytische Näherung für den Feldfaktor Γ aus [Hur92] gilt für Felder kleiner $\sim 9 \cdot 10^5\, V/cm$. Für größere Felder werden andere Mechanismen zunehmend wichtig (siehe nächster Abschnitt).

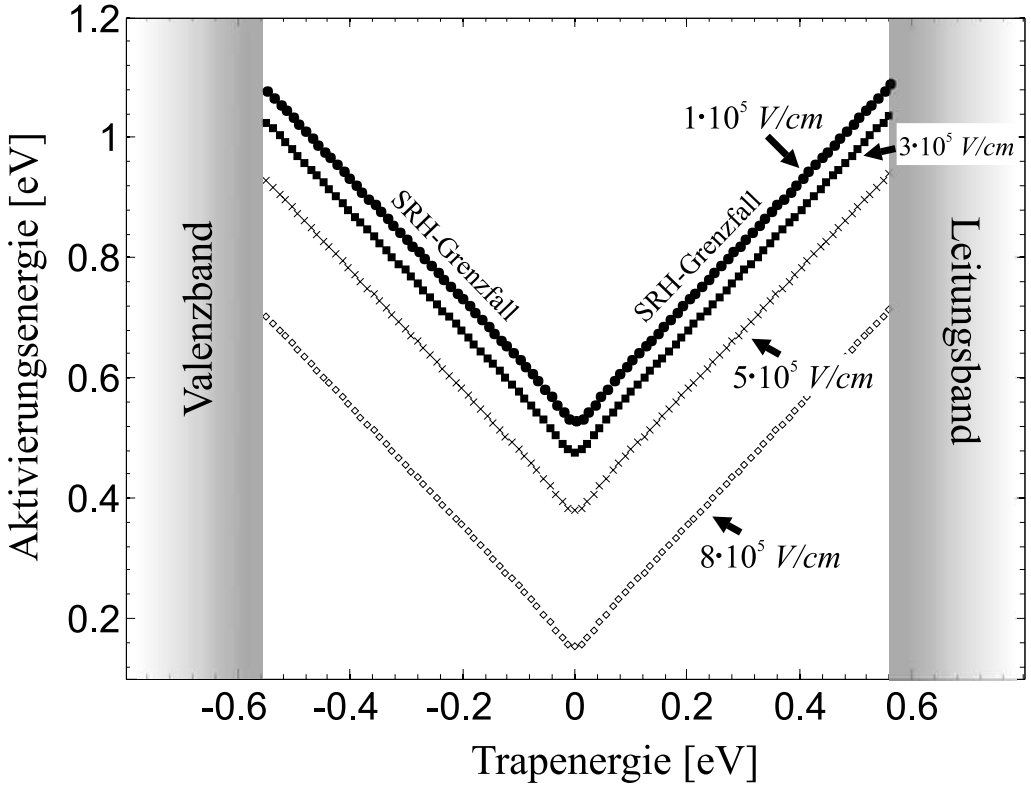


Abbildung 8.9: Aktivierungsenergie in Abhängigkeit von der Trap-Position bei verschiedenen elektrischen Feldstärken.

Analytisch berechnet werden kann die Aktivierungsenergie durch Differentiation der logarithmierten Stromgleichung 8.33 nach $1/k_BT$. Da $\Gamma(F)$ im interessanten Feldbereich größer 1 ist, kann Gleichung 8.33 zu $I_{TFE} = (1 + \Gamma(F)) \cdot I_{SRH} \approx \Gamma(F) \cdot I_{SRH}$ vereinfacht werden und die Ableitung ergibt:

$$\begin{aligned} E_{a,TFE} &= \frac{\partial \ln I_{TFE}}{\partial(\frac{1}{k_BT})} = E_{a,SRH} - \frac{\partial}{\partial(\frac{1}{k_BT})} \ln(\Gamma(F)) \\ &= E_{a,SRH} - \tfrac{3}{2} k_B T - 3 k_B T (\tfrac{F}{F_\Gamma})^2 \end{aligned} \tag{8.34}$$

Gleichung 8.34 bietet eine Möglichkeit das maximale elektrische Feld in den Speicherzellen nach oben abzuschätzen. Unter der Annahme, dass Tailzellen durch TFE-Leckströme entstehen (beobachtete Aktivierungsenergien legen dies nahe), folgt für ein Trap mit Trapenergie in der Bandlückenmitte, das sich am Ort des höchsten elektrischen Feldes befindet, die kleinste Aktivierungsenergie. Umgekehrt kann aus der kleinsten beobachteten Aktivierungsenergie unter diesen Annahmen mit Hilfe von Gleichung 8.34 das maximale elektrische Feld berechnet werden.

8.5 Trapunterstütztes Tunneln (TAT) und Band-zu-Band-Tunneln (BTB)

Für elektrische Felder größer $\sim 8 \cdot 10^5\, V/cm$ sind die Bänder so stark verbogen, dass der horizontale Abstand im Banddiagramm sehr klein wird. Dadurch wird zuerst Trapunterstütztes-Tunneln (TAT) und bei noch höheren Feldern direktes Tunneln (BTB) von Elektronen aus dem Valenzband der p-Seite ins Leitungsband der n-Seite möglich. Die Situation im Banddiagramm für beide Tunnelmechanismen ist in Abbildung 8.10 schematisch dargestellt. TAT entsteht aus TFE (siehe letzten Abschnitt) durch zunehmende Felderhöhung bis die Barriere für das generierte Ladungsträgerpaar sehr klein wird und ein Strom durch direktes zweistufiges Tunneln ohne thermische Anregung (nur horizontale Komponenten im Banddiagramm) fließen kann. Aus der anschaulichen Betrachtungsweise der Aktivierungsenergie im Banddiagramm als längste vertikale Komponente folgt direkt ein sehr niedriger Wert der Aktivierungsenergie ($< 0.15\, eV$). Kommen sich die Bänder noch näher können die Elektronen direkt vom Valenzband der p-Seite ins Leitungsband der n-Seite tunneln und es fließt ein großer BTB-Tunnelstrom. Dieser ist näherungsweise gegeben durch [Tau98, S.94]:

$$I_{BTB} = A \cdot J_{b-b} = A \cdot \frac{\sqrt{2m^*}q^3 F \cdot V}{4\pi^3 \hbar^2 E_G^{1/2}} exp(-\frac{4\sqrt{2m^*}E_G^{3/2}}{3qF\hbar}) \tag{8.35}$$

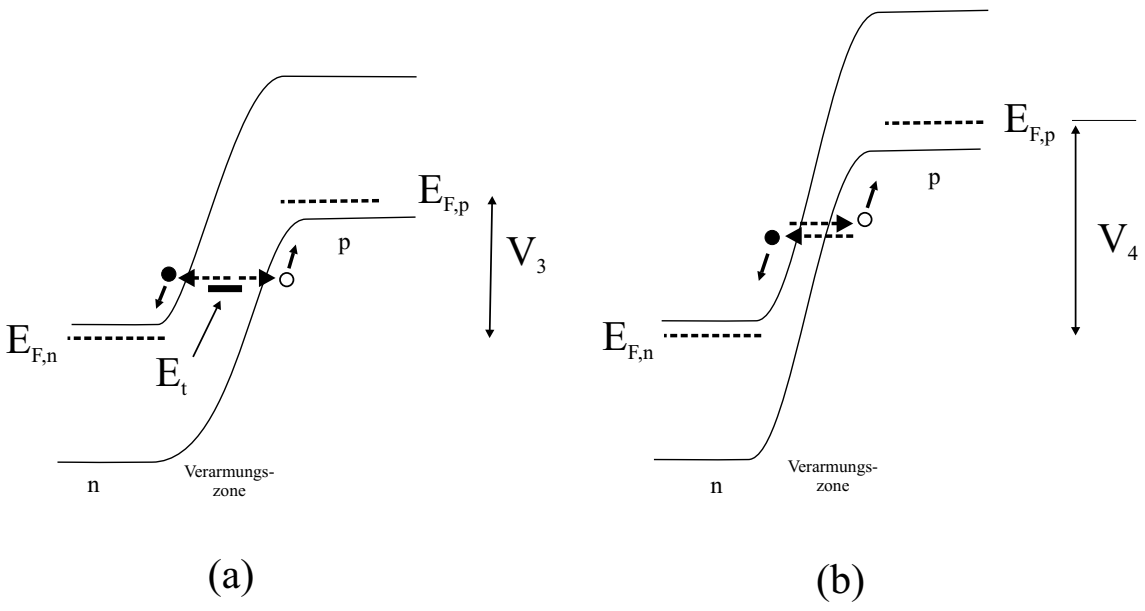


Abbildung 8.10: (a) Trapunterstütztes Tunneln (TAT) (b) Band-zu-Band Tunneln (BTB).

Leckstromabschätzung & Temperaturabhängigkeit

Abbildung 8.11 zeigt neben dem BTB-Tunnelstrom die Aktivierungsenergie in Abhängigkeit von der elektrischen Feldstärke im für DRAM relevanten Bereich. Da der Tunnelstrom offensichtlich nicht exakt dem Arrhenius-Gesetz folgt (vgl. Gleichung 8.35), wurden die Aktivierungsenergien im Bereich von $65\,°C - 85\,°C$ bestimmt. Nach dieser Abschätzung ist BTB erst ab einer Feldstärke von ca. $1.2 \cdot 10^6\, V/cm$ in der Lage Leckströme im Bereich der Tailverteilung (zweistelliger fA-Bereich) zu generieren. Für kleinere Felder spielt BTB keine Rolle. Zu bemerken ist hier, dass in Gleichung 8.35 nicht nur die nominalen Felder sondern auch durch inhomogene Dotierstoffverteilungen oder Metallpräzipitaten lokal erhöhte Felder eingehen.

Was die Temperaturabhängigkeit betrifft, so ist die Aktivierungsenergie eines Tunnelstroms praktisch gleich null. Da jedoch die Bandlücke E_G und die eingebaute Spannung V_{bi} in Gleichung 8.35 eine gewisse Temperaturabhängigkeit besitzen, ist auch die Aktivierungsenergie für TAT und BTB von null verschieden. Im Falle von BTB ist keine einfache Arrhenius-Abhängigkeit des Leckstroms gegeben (siehe Gleichung 8.35).

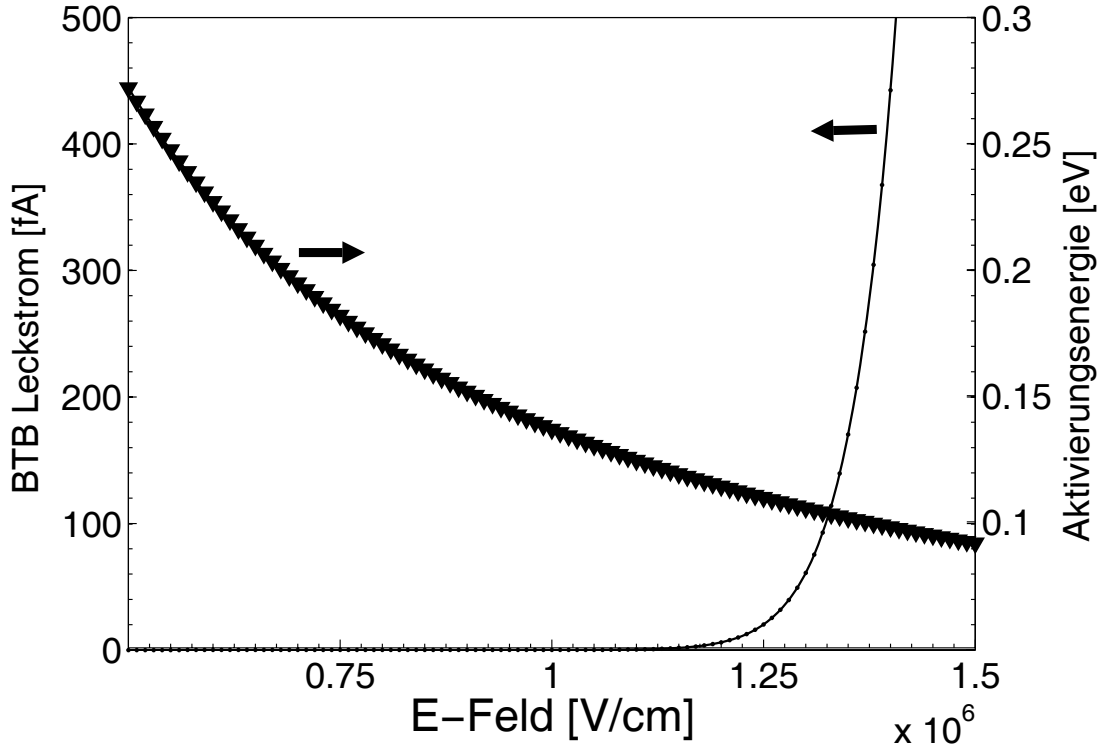


Abbildung 8.11: BTB-Tunnelstrom nach Gleichung 8.35. In der Simulation wurde das elektrische Feld bei konstanter äußerer Spannung V variiert.

8.6 Zusammenfassung

In diesem Kapitel wurden die verschiedenen pn-Leckstrommechanismen besprochen und bezüglich deren Spannungs- und Temperaturabhängigkeit analysiert. Aktivierungsenergien geben dabei einen wichtigen Hinweis auf den zugrundeliegenden Mechanismus. Durch das einfache Modell eines abrupten pn-Übergangs mit zum kondensatorseitigen pn-Übergang entsprechenden Abmessungen konnten die durch verschiedene Mechanismen entstehenden Leckströme grob abgeschätzt werden. Genaue Rechnungen sind nicht möglich, da sowohl die Anzahl der Zustände in der Bandlücke sowie deren energetische Verteilung und Wirkungsquerschnitte nicht bekannt sind. Für den unvermeidbaren Diffusionsstrom eines idealen pn-Übergangs ergab die Abschätzung Leckströme, die weit unter den kleinsten über Retentionmessungen bestimmten Leckströmen liegen. Auch die Aktivierungsenergie für den Diffusionsstrom liegt mit $1.12\,eV$ deutlich über den gemessenen Werten (siehe Kapitel 7). Der Diffusionsstrom ist demnach in keinem Bereich der Retentionkurve von Bedeutung. SRH-Generation durch Traps im Bereich der Verarmungszone führt dagegen zu Leckströmen, die den Diffusionsstrom weit übertreffen. Ein einzelnes Trap in der Bandlückenmitte generiert einen Leckstrom im Bereich der Mainverteilung (fA-Bereich). Die Traps können dabei sowohl im Volumen als auch an der Grenzfläche zum SiO_2 sitzen, da jeder pn-Übergang zwangsweise irgendwo an die Oberfläche kommen muss. Aus typischen Interfacezustandsdichten folgt eine zweistellige Anzahl Traps im Bereich der Verarmungszone pro Speicherzelle. Die Zustandsdichte ist in der Bandlückenmitte am geringsten und der Leckstrom fällt exponentiell mit dem energetischen Abstand der Traps von der Bandlückenmitte ab. Die Faltung aus Trapanzahl und Trapenergieverteilung stellt eine mögliche Erklärung für die Retention-Mainverteilung dar. Gemessene Aktivierungsenergien von $\approx 0.68\,eV$ im Main untermauern diese Erklärung. Tailzellen können durch SRH-Leckströme nicht erklärt werden, da zum einen im Gegensatz zum SRH-Mechanismus die gemessenen Aktivierungsenergien feldabhängig sind und zum anderen derart hohe Leckströme nur als Summe relativ vieler Traps in der Bandlückenmitte erklärbar wären, was sehr unwahrscheinlich ist. Beim TFE-Mechanismus sind die Aktivierungsenergien jedoch wie gefordert spannungsabhängig und liegen auch betragsmäßig im gemessenen Bereich ($0.4 - 0.6\,eV$). Außerdem macht der Feldfaktor Γ mit Werten bis in den dreistelligen Bereich genau den Unterschied im Leckstrom zwischen Main und Tail aus. Die Mechanismen TAT und BTB können aufgrund der zu niedrigen Aktivierungsenergien nicht als Erklärung für den Retentiontail herangezogen werden. Desweiteren sind für nennenswerten Band-zu-Band Tunnelstrom elektrische Felder weit über den aus Prozesssimulationen bekannten Nominalfelder notwendig.

Schlussfolgerung: Durch Vergleich von theoretischen Betrachtungen und den Ergebnissen der elektrischen Charakterisierung aus Kapitel 7 werden Interface-Traps als Ursache für den Retentionmain angesehen. Tailzellen sind durch einzelne Traps mit Energien nahe der Bandlückenmitte und am Ort der größten elektrischen Felder durch den TFE-Mechanismus erklärbar. Dabei können auch Fluktuationen des elektrischen Feldes, z.B. durch statistische Dotierstoffverteilungen oder Felderhöhung im Umfeld von Metall-Präzipitaten eine Rolle spielen. In Kapitel 9 folgt die experimentelle Verifikation dieser These durch gezielte Trapreduzierung im Gate/Drain Überlapp-Bereich.

Kapitel 9

Experimentelle Verifikation der dominanten Zellleckströme und deren selektive Verbesserung

Aus der Charakterisierung in den vorhergehenden Kapiteln ging hervor, dass die Retentionverteilung hauptsächlich durch Generationsströme im pn-Übergang des Speicherknotens bestimmt ist und die meisten zu Tail-Leckströmen führenden Traps im Einflussbereich des Gates liegen müssen. Durch die Feldabhängigkeit der Aktivierungsenergien wurde außerdem gezeigt, dass für Tailzellen die Generationsströme durch Tunneleffekte verstärkt werden. In diesem Kapitel wird das erarbeitete Modell für den Retentiontail experimentell verifiziert und gleichzeitig eine Möglichkeit zur Verbesserung der Tailverteilung vorgestellt.

9.1 Mögliche Maßnahmen zur Tailverbesserung

Der im Retentiontail vorherrschende Mechanismus TFE setzt sich aus zwei Komponenten zusammen. Der thermisch angeregten Ladungsträgeremission ausgehend von Traps in der Bandlücke folgt ein Tunnelprozess durch eine feldabhängige Barriere in die Bänder (siehe Abschnitt 8.4). Für kleine elektrische Felder ist die Barriere zu dick und kann nicht durchtunnelt werden. Ein Generationsstrom kann dann nur durch wesentlich höhere thermische Anregung bis ins Leitungsband entstehen und ist wegen der damit verbundenen geringeren Wahrscheinlichkeiten kleiner als bei teilweiser thermischer Anregung gefolgt von Tunneln. Gerade das Auftreten eines Traps am Ort großer elektrischer Felder führt deshalb zu besonders großen Leckströmen, die wiederum zu Tailzellen führen. Andersherum führt die Unterbindung einer der Komponenten zu einer deutlichen Reduktion des Leckstroms. Daraus ergeben sich prinzipiell zwei Ansatzpunkte zur Leckstrom-

reduktion. Erstens kann durch Reduktion der elektrischen Feldstärke die Bandverbiegung verringert, die Tunnelbarriere damit verbreitert und schließlich die Tunnelwahrscheinlichkeit reduziert werden. Der Leckstrom wird dadurch im besten Fall (keine Tunnelkomponente mehr) vom TFE-Mechanismus auf standard SRH-Generation zurückgeführt und damit um den Feldfaktor Γ (siehe Kapitel 8) reduziert. Der zweite erheblich schwierigere Ansatz liegt in der Passivierung bzw. Elimination der für Tail-Leckströme verantwortlichen Zustände in der Bandlücke selbst, wodurch der Generationsleckstrom ganz unterbunden wird und die elektrische Feldstärke unverändert bleiben kann (solange BTB-Tunneln keine Rolle spielt). Beide Ansätze führen zur Reduktion der Fehleranzahl an der für die Chipausbeute relevanten Reparaturgrenze. Da jedoch vielerlei zusätzliche Randbedingungen im DRAM beachtet werden müssen, sollen die bestehenden Verbesserungsmöglichkeiten im Folgenden genauer diskutiert werden.

9.1.1 E-Feldreduktion im kondensatorseitigen pn-Übergang

Aufgrund der exponentiellen Abhängigkeit der Tunnelkomponente vom elektrischen Feld hat die Minimierung der elektrischen Felder in der Umgebung des kondensatorseitigen pn-Übergangs der Speicherzelle oberste Priorität. Dies kann auf verschiedene Arten erreicht werden.

Geringere Dotierstoffkonzentrationen

Da die elektrischen Feldstärken mit den Dotierstoffkonzentrationen wachsen, werden letztere bei der Technologieentwicklung so niedrig wie es alle weiteren Randbedingungen zulassen, gewählt. Bei sich in der Volumenproduktion befindlichen DRAM Technologien kann durch einfache Reduktion von Dotierstoffkonzentrationen deshalb meist keine weitere Retentionverbesserung ohne Ausbeuteverluste aufgrund anderer Randbedingungen erzielt werden. Beispielsweise wachsen mit Reduktion der Kanaldotierung die Verarmungszonen der Source/Drain-Gebiete, wodurch sich diese einander annähern und die effektive Kanallänge des Auswahltransistors verkürzen. Kurzkanaleffekte führen zu stark erhöhten Unterschwellenleckströmen, die ihrerseits zum schnellen Ladungsverlust führen. Ein weiteres Problem reduzierter Dotierstoffkonzentrationen liegt in der damit verbundenen Erhöhung der Serienwiderstände, wodurch Speicherzellen bei den sehr kurzen im DRAM üblichen Schreibzeiten nicht bis zur vollen Spannung aufgeladen werden können. Es steht weniger Ladung für Leckströme zu Verfügung und die gespeicherte Information geht dadurch schneller verloren. Eine Möglichkeit die Dotierstoffkonzentrationen zu reduzieren ohne in Probleme mit Kurzkanaleffekten zu laufen sind dreidimensionale Transistorstrukturen. Dabei wird der Transistorkanal unter Ausnutzung der dritten Dimension ausgebildet und dadurch von der minimalen Strukturgröße entkoppelt. In den letzten Jahren sind verschiedene Ansätze, wie z.B. RCAT

[Kim03, Kim05], SRCAT [Oh05], STAR [Jan05], SGT[Goe02], planare Transistoren mit erhöhten Source/Drain Gebieten und EUD[Mue05], vorgestellt worden. Käuflich erhältlich sind derzeit DRAMs mit RCAT-Auswahltransistoren (Samsung), sowie erhöhten S/D Gebieten (Micron). Ein Nachteil der Auswahltransistoren mit 3-dimensional verlängerten Kanallängen liegt in den damit verbundenen geringeren On-Strömen, welche jedoch z.B. für schnelle Grafikspeicher Voraussetzung sind. Eine Zusammenfassung der aktuellen Herausforderungen für die Entwicklung bis zu einer Strukturgröße von $40\,nm$ gibt [Mue05].

Die allermeisten Zellen heutiger DRAMs übertreffen die notwendigen Haltezeiten um Größenordnungen. Deshalb muss an dieser Stelle nochmals darauf hingewiesen werden, dass aufgrund relativ weniger Zellen mit besonders hohen Leckströmen (weniger als $1 \cdot 10^{-6}$ aller Zellen) gegenwärtig neue 3d-Transistoren entwickelt werden müssen, die in anderer Hinsicht planaren Auswahltransistoren deutlich unterlegen sind. Um das elektrische Feld in wenigen Zellen zu reduzieren, treibt man also einen gewaltigen Entwicklungsaufwand, der obendrein für den Großteil der Zellen Nachteile mit sich bringt.

Dickeres Gateoxid im Bereich des Gate/Drain Überlapps

Durch eine kurze Oxidation nach der Strukturierung des Gatestapels wird das Gateoxide von der Seite her in Richtung Kanalmitte aufgedickt. Bekannt ist diese schnabelartige Verdickung (*bird's beak*) des Gateoxids von der bis zu $0.8\,\mu m$ Speicher-Technologien benutzten LOCOS-Isolation. In heutigen Technologien mit Strukturen $< 100\,nm$ ist die LOCOS-Isolation nicht mehr üblich, da der dabei entstehende *bird's beak* den gesamten Transistor unterlaufen würde. Anstatt dessen dient ein geätzter Graben (STI) zur Isolation zwischen benachbarten Strukturen. Der *bird's beak* führt zu einer verringerten Gatekapazität sowie einer kleineren Bandverbiegung im G/D-Überlapp. Die Reduktion von GIDL-Leckströmen durch die geringere Bandverbiegung wird durch eine verkürzte effektive Kanallänge und damit verbundenen Kurzkanaleffekten sowie durch eine geringere Kanalkontrolle mit geringeren On-Strömen kompensiert. Durch weitere Optimierung des *bird's beak* sind demzufolge keine revolutionären Verbesserungen zu erwarten.

Verringerte Betriebsspannungen

Die Verringerung der Betriebsspannungen zur Reduktion der Leckströme liegt nach Betrachtung der Spannungsabhängigkeiten in Kapitel 5 zugegebenermaßen auf der Hand. Darüber hinaus ist dies fester Bestandteil der Roadmap für die nächsten Jahre. Der Grund für die zukünftige Verringerung der Spannungen liegt vor allem in der Notwendigkeit den Stromverbrauch bei steigender Speichergröße konstant zu halten bzw. für Low-Power-Anwendungsgebiete (z.B. DRAM für Handy und PDA ...) zu senken. Auch hier

gibt es weitere Randbedingungen, die diese auf den ersten Blick einfache Lösung erheblich schwieriger machen. Durch verringerte Betriebsspannungen wird einerseits weniger Ladung im Kondensator gespeichert und andererseits, was noch viel wichtiger ist, nimmt das beim Lesen auf der Bitleitung ausgebildete Signal ab. Dadurch müssen die Differenzverstärker verbessert werden, während gleichzeitig durch die immer kleineren minimalen Strukturgrößen weniger Platz zur Verfügung steht und außerdem elektrische Übersprecheffekte durch immer geringere Leitungsabstände weiter zunehmen und das Signal-Rauschverhältnis dramatisch verschlechtern. Es entstehen somit enorme Anforderungen an das Design der Differenzverstärker. Letztendlich nehmen zwar Leckströme mit den kleineren Betriebsspannungen ab, gleichzeitig entstehen jedoch schwerwiegende Probleme beim Lesen und Bewerten der gespeicherten Information.

9.1.2 Passivierung von Zuständen in der Bandlücke

Durch die Passivierung von Traps kann im Gegensatz zur generellen Feldreduktion die Retention von Tailzellen ohne negative Auswirkungen auf die Gesamtheit der Zellen gezielt verbessert werden. Da für die untersuchte Qimonda $110\,nm$ Technologie die höchsten elektrischen Felder im Bereich des G/D-Überlapps liegen und der Hauptmechanismus (TFE) durch das elektrische Feld verstärkt wird, führen Traps in diesem Bereich zu besonders hohen Leckströmen. Auch bei anderen aktuellen Technologien sind die maximalen Felder im selben Bereich zu finden, sodass die Ergebnisse durchaus übertragbar sind. Über den mikroskopischen Ursprung der Traps ist wenig bekannt, da sie sich aufgrund der sehr geringen Auftrittswahrscheinlichkeit den Standard-Charakterisierungsmethoden an relativ großflächigen Teststrukturen entziehen (siehe Kapitel 6). Erst durch die in dieser Arbeit speziell entwickelten Einzelzellcharakterisierung direkt auf Speicherbausteinen können zusätzliche Informationen gewonnen werden. Generell wird zwischen Zuständen am Si/SiO_2-Interface und Defekten im Siliziumvolumen unterschieden. Während Volumendefekte konkrete Energieniveaus aufweisen, sind Grenzflächenzustände kontinuierlich über die Bandlücke verteilt [Pie03]. Ein mikroskopisches Bild für Interfacezustände sind meist „dangling bonds", d.h. ungesättigte Si-Bindungen am Interface. Für Volumendefekte kommen einerseits Verunreinigungen (z.B. Metallatome) und andererseits intrinsische Punktdefekte, wie z.B. *Interstitial-* oder *Vacancykomplexe,* die bei der Prozessierung entstehen, in Frage. Es ist anzunehmen, dass sowohl Interfacezustände als auch Defekte im Volumen Leckströme erzeugen können, die zu Tailzellen führen.

9.2 Experimentelle Verifikation durch Passivierung mit Fluor

Zur Reduktion von Interface-Zuständen werden in vielen CMOS Prozessen so genannte *Post Metallization Anneals* (PMA) bei $400 - 450\,°C$ in Wasserstoff oder Formiergas eingesetzt (siehe z.B. [Wol00, S. 293]). Es wird angenommen, dass in Anwesenheit von Aluminium atomarer Wasserstoff entsteht, der zu den Grenzflächen diffundiert und dort Interfacezustände absättigt. Auch im DRAM sind bereits derartige PMA-Anneals im Prozessfluss enthalten. An anderer Stelle wurde berichtet, dass Fluor aufgrund der im Vergleich zu Wasserstoff größeren Bindungsenergie mit Silizium, Interfacetraps stabiler als Wasserstoff absättigen kann (siehe [Mog97, Pic04]). Fluor kann aufgrund der hohen Reaktivität jedoch nicht wie Wasserstoff durch Anneals in einer geeigneten Atmosphäre eingebracht werden. Darüber hinaus besitzt Fluor ein sehr kompliziertes und noch nicht vollständig erforschtes Verhalten im Silizium. Sowohl für CMOS Schaltungen vorteilhafte wie auch negative Einflüsse werden berichtet. Einen Überblick über das Wissen zum Verhalten von Fluor gibt [Pic04].

In diesem Abschnitt soll ein Weg gefunden werden, die stabile Passivierung von Traps durch Fluor ohne negative Folgen im DRAM Prozess einzusetzen.

9.2.1 Experimente

Eine Technologie wird im allgemeinen durch Prozessvariationen (Splitlose) optimiert, wobei im besten Fall nur ein Prozessparameter variiert und die Auswirkung auf eine Vielzahl von Kenngrößen (Yields, notwendige Reparatur ...) analysiert wird. Grundgedanke der Versuche in diesem Kapitel ist, die Ursache für TFE-Leckströme im G/D-Überlapp, nämlich die Traps an dieser Stelle, gezielt und sehr stabil mit Fluor zu passivieren und den Tail-Leckstrom dadurch abzuschalten. Dazu wurden eine Reihe von Fertigungsversuchen auf $512M$ und $1G$ DRAMs in $110\,nm$ Technologie durchgeführt und anschließend detailliert analysiert. Die Schwierigkeit besteht in der Art und Weise der Fluoreinbringung und in der Bestimmung des besten Zeitpunkts im DRAM-Gesamtprozess. Eine frühzeitige Fluoridierung ist nicht erfolgsversprechend, da einerseits die zu passivierenden Traps erst spät im Prozess entstehen können und das Fluor aufgrund hoher Diffusivität nicht „vorgehalten" werden kann und andererseits auch SiF Bindungen in Hochtemperaturschritten aufgebrochen werden können. Deshalb muss Fluor möglichst spät im Gesamtprozess und erst nach allen Hochtemperaturschritten eingebracht werden. Aufgrund der hohen Diffusivität von Fluor in Silizium verteilt sich eine eingebrachte Fluormenge sehr schnell und muss deshalb möglichst nahe an die zu passivierenden Traps im G/D-Überlappbereich gebracht werden. In den Experimenten wurde Fluor deshalb durch Implantation, die eine effiziente, kostengünstige und leicht zu kontrollierende Möglichkeit darstellt, eingebracht. Die letztmögliche Stelle im DRAM-Prozess, an

der Fluor durch Implantation nahe an das G/D-Überlappgebiet gebracht werden kann, ist nach der Strukturierung des Gatestapels. Abbildung 9.1 zeigt die Zellstrukturen zum Zeitpunkt der Fluor-Implantation. Im Detail bieten sich dabei zwei Stellen an, an denen die pn-Übergangsgebiete für Implantationen zugänglich sind. Erstens nach der Strukturierung des Gatestapels und der Seitenwandoxidation und zweitens nach Abscheidung und Strukturierung des Nitridspacers (siehe Abbildung 9.1a und b). Um herauszufinden, wie nahe das Fluor an die zu passivierenden Traps gebracht werden muss, wurde in einem dritten Versuch die BS-Seite der Zelle während der Implantation durch eine zusätzliche Lackmaske abgedeckt. Das Fluor kann dadurch nur BL-seitig ins Substrat eindringen (Abbildung 9.1c).

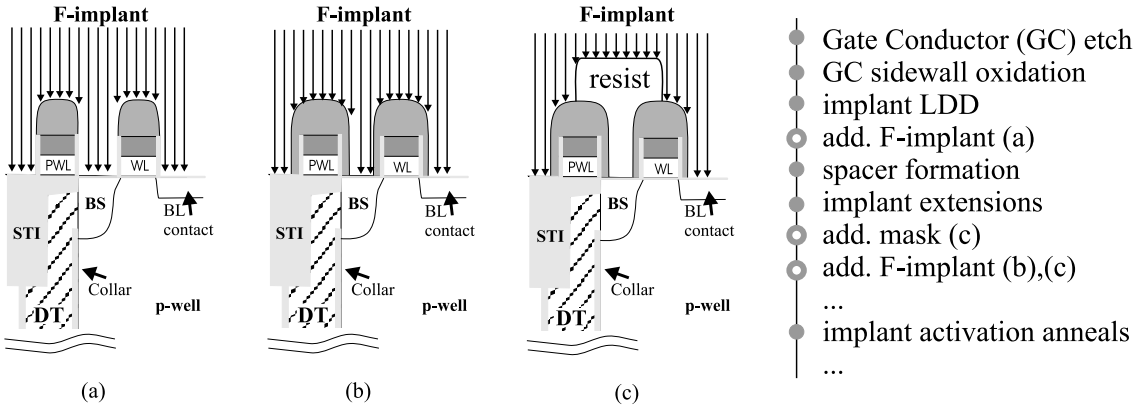


Abbildung 9.1: Fluor Implantationsexperimente: (a) Vor Gatespacer, (b) nach Gatespacer und (c) nur bitleitungsseitig nach Gatespacer. Bei Versuch (c) wird die Kondensatorseite der Zelle durch eine Lackmaske abgedeckt (Grafik aus [Web06b]).

9.2.2 Ergebnisse der Implantationsexperimente

Zur Bewertung der Fertigungsversuche wurde die auf eine standardprozessierte Referenzgruppe normierte Fehlerzahl (FC) an der Reparaturgrenze herangezogen. Ein Fehlerverhältnis kleiner eins entspricht einer Verbesserung, ein Verhältnis größer eins einer Verschlechterung der Retention-Tailverteilung. Abbildung 9.2 fasst die verwendeten Implantationsenergien und Fluor-Dosen zusammen. Die Implantationsenergie betrug in allen Versuchen $10\,keV$ und implantiert wurde senkrecht zur Waferoberfläche. Jeder Datenpunkt basiert auf Grundlage mehrerer $300\,mm$ Wafer mit jeweils ca. 500 Speicherchips. Die besten Splitgruppen wurden auf einer größeren Anzahl von Losen mit jeweils 25 Wafern verifiziert, wodurch die Ergebnisse als statistisch sicher angenommen werden können. Ein wichtiges Ergebnis der Versuche ist, dass der Zeitpunkt der Fluor-Implantation von entscheidender Bedeutung ist. Für den Fall der Implantation vor der Spacerstrukturierung (Versuch a) konnte keine Retentionverbesserung festgestellt werden, während bei den Experimenten nach der Spacerstrukturierung (Versuch b) eine Reduzierung des Fehlerzahl abhängig von der Implantationsdosis gezeigt werden konnte.

OVERVIEW FLUORINE IMPLANT EXPERIMENTS

Scheme	Energy [keV]	Dose [1/cm^2]	normalized retention FC
Experiment 1 (pre spacer)	10	3e13	0.97
	10	1e14	1.13
Experiment 2 (post spacer)	10	3e13	1.07
	10	7e13	0.82
	10	1e14	0.69
	10	1.3e14	0.6
	10	1.7e14	0.73
	10	2e14	0.79
	10	3e14	1.19
Experiment 3 (BL side only)	10	1e14	0.95

Abbildung 9.2: Fluor-Implantationsexperimente. Die Tabelle zeigt die zur Standardprozessierung (ohne F-Implant) normierte Fehleranzahl an der Reparaturgrenze. Retentionvorteile ergeben sich bei Implantation nach der Nitridspacer-Formierung (Tabelle aus [Web06b]).

Die Abhängigkeit der Retentionfehler an der Reparaturgrenze von der Implantationsdosis ist in Abbildung 9.3 grafisch dargestellt. Für Implantationsdosen größer $\sim 4{\cdot}10^{13}\,cm^{-2}$ nimmt die Fehlerzahl mit der Dosis ab. Die minimale Fehlerzahl ist bei $1.3 \cdot 10^{14}\,cm^{-2}$ erreicht. Bei noch höherer Dosis nimmt die Fehlerzahl wieder zu, bis schließlich bei ungefähr $2.5 \cdot 10^{14}\,cm^{-2}$ kein Vorteil durch die Implantation mehr vorhanden ist. Insgesamt kann der durchlaufene Dosisbereich in Abbildung 9.3 in Abschnitte mit erhöhter bzw. erniedrigter Fehlerzahl unterteilt werden. Das Prozessfenster für eine Verbesserung umspannt knapp eine Größenordnung in der Fluor-Dosis. Demgegenüber liegt die Genauigkeit in der Implantationsdosis heutzutage im niederen Prozentbereich, sodass eine Implantationsdosis mit FC Verbesserung leicht einzustellen ist. Zu bemerken ist weiterhin, dass nur der Retentiontail durch die Fluorimplantation deutlich verbessert werden konnte, während sich die Mainverteilung nur unwesentlich ändert (hier nicht gezeigt).

9.2.3 Diskussion der experimentellen Ergebnisse

Eine mögliche Erklärung für die Dosisabhängigkeit gibt der durch die Implantation selbst verursachte Schaden in der kondensatorseitigen Verarmungszone. Abbildung 9.4 zeigt zur Verdeutlichung die mittels Davinci simulierte Feldverteilung für eine Schnittebene senkrecht zur Wortleitung und mittig durch das aktive Gebiet des Auswahltransistors. Die Wortleitung (GC), der Grabenkondensator (DT), die Grabenisolation (STI) sowie die Fluor-Implantationsgebiete sind für die beiden Implantationsversuche vor und

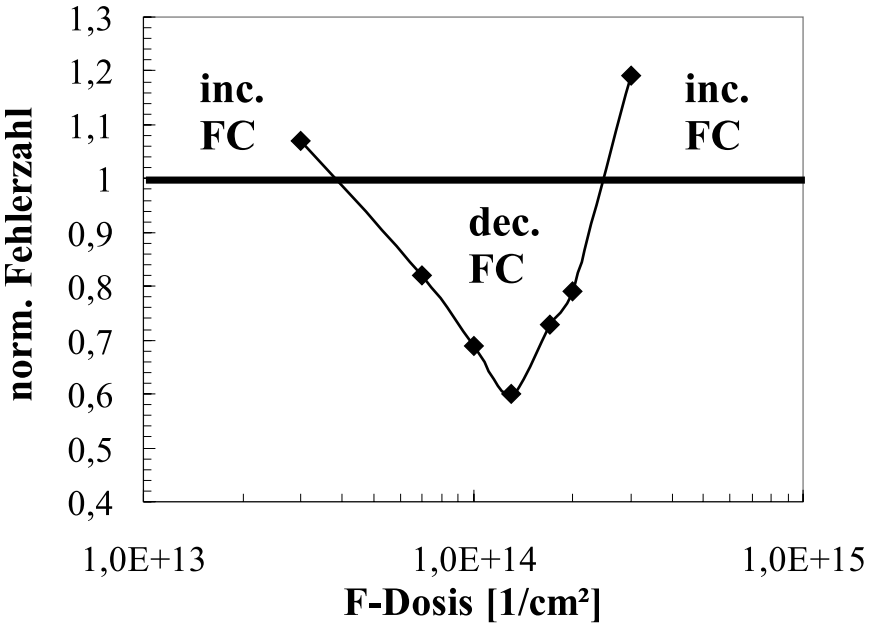


Abbildung 9.3: Abhängigkeit der zu standardprozessierten Chips normierten Retentionfehlerzahl von der F-Dosis. Fluor wurde für alle Dosen mit einer Energie von $10\,keV$ nach der Spacer-Strukturierung implantiert.

nach Spacerformierung skizziert. Für den Fall der Implantation vor Nitridspacerstrukturierung (siehe Abbildung 9.4a) wird direkt in Gebiete mit dem höchsten elektrischen Feld implantiert, während bei der Implantation nach Spacer (9.4b) hohe Feldregionen durch den Spacer geschützt sind. Durch die Implantation selbst entstehen - wenn auch mit sehr geringer Wahrscheinlichkeit - neue Defekte, die zu zusätzlichen Tailzellen führen können und dadurch die Fehlerzahl wieder in die Höhe treiben. Mit zunehmender F-Dosis erhöht sich durch laterale Streuung zunehmend auch die Wahrscheinlichkeit dafür, dass auch für die Implantationsvariante nach Spacerformierung Schäden im Bereich hoher elektrischer Felder entstehen. Ab einer gewissen Dosis überwiegen schließlich zusätzlich entstandene Defekte die passivierende Wirkung und es kommt sogar zu einer insgesamt höheren Fehlerzahl.

Die Implantationsvariante (c), bei der nur bitleitungsseitig implantiert wurde, liefert ein weiteres sehr wichtiges Ergebnis. Im Gegensatz zu den vorhergehenden Experimenten ist hierbei kein Retentionvorteil beobachtbar. Im Umkehrschluss kann die beobachtete Verbesserung der Retentionfehler bei beidseitiger Implantation, auf die kondensatorseitige Implantation von Fluor zurückgeführt werden. Dies deutet darauf hin, dass das Fluor sehr nahe an die zu passivierenden Stellen gebracht werden muss, um wirksam zu sein. Die Implantation in den nur ca. $200\,nm$ entfernten Bitleitungskontakt führt bereits zu keiner Verbesserung mehr. Die Erklärung dafür liefert die hohe Diffusivität von Fluor in Silizium zusammen mit einer minimalen benötigten Fluorkonzentration am Ort der zu

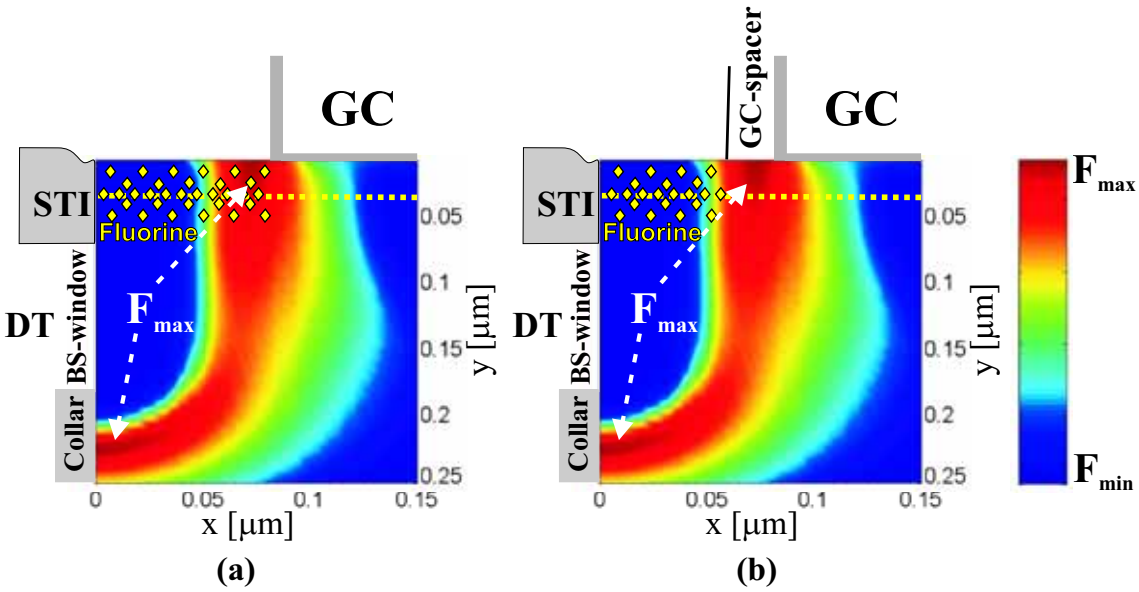


Abbildung 9.4: Simulierte Feldverteilung für eine Schnittebene senkrecht zur Wortleitung und mittig durch das aktive Gebiet des Auswahltransistors. Das mit F implantierte Gebiet ist für die beiden Fälle (a) vor Spacer-Strukturierung und (b) nach Spacer-Strukturierung skizziert.

passivierenden Traps. Die implantierte Fluormenge verteilt sich bei einer Temperung so schnell im Silizium, dass die Fluorkonzentration an der Kondensatorseite zu gering ist, um dort Defekte mit genügend hoher Wahrscheinlichkeit zu passivieren.

9.2.4 Aktivierungsenergie-Analyse

Die in Kapitel 6 beschriebene Methode der Aktivierungsenergie-Analyse auf Einzelzellenbasis gibt als einzige Methode Aufschluss darüber, welche Leckstrompfade und Mechanismen durch die Fluorimplantation reduziert werden können. Abbildung 9.5 zeigt die Verteilung der Aktivierungsenergien von Zellen nahe der Reparaturgrenze für standardprozessierte und F-implantierte Chips. Jeder Datenpunkt stellt die Aktivierungsenergie des Leckstroms einer Speicherzelle dar. Zur Verbesserung der Statistik wurden Messdaten mehrerer Chips überlagert. Die beiden resultierenden Verteilungen unterscheiden sich deutlich voneinander. Dabei verläuft die Kurve für F-implantierte Speicherzellen über den gesamten Energiebereich hinweg unterhalb der Verteilung für standardprozessierte Zellen. Sowohl für sehr kleine wie auch für große Energien nähern sich die Verteilungen einander an. Im Bereich zwischen $\sim 0.46\,eV$ und $\sim 0.52\,eV$ spalten die beiden gemessenen Verteilungen auf. In der kumulativen Darstellung von Abbildung 9.5 ist für beide Kurven die Wahrscheinlichkeit bezogen auf die jeweilige Gesamtfehlerzahl dargestellt. Die erzielten 30% Fehlerreduktion der F-implantierten Gruppe (vergleiche Tabelle 9.2) werden in der Darstellung nicht ersichtlich und der Effekt der Fluorimplantation wird dadurch unterschätzt.

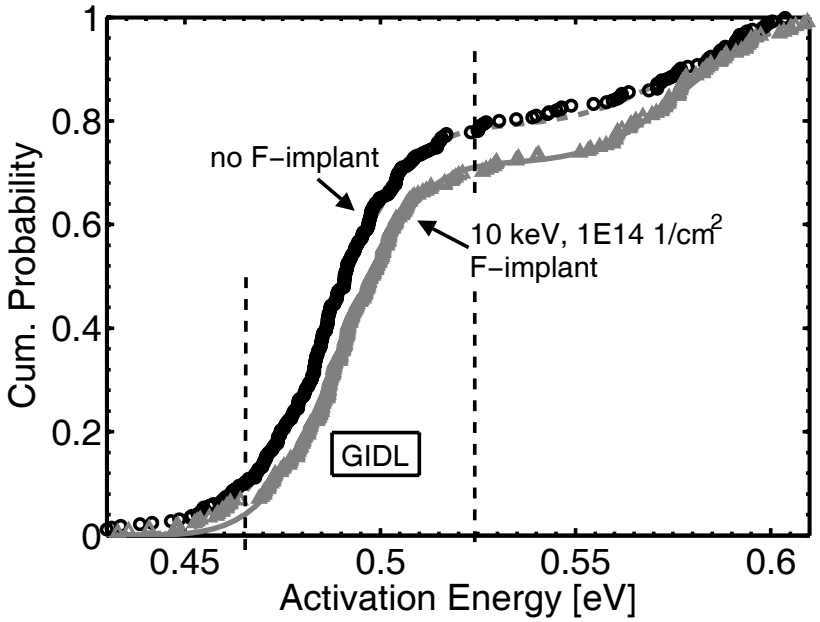


Abbildung 9.5: Kumulative Verteilung der Aktivierungsenergien der schlechtesten Zellen für standard und F-implantierte Speicherbausteine. Fluor wurde nach der Spacer-Strukturierung mit $10\,keV$ und einer Dosis von $1 \cdot 10^{14}\,cm^{-2}$ implantiert (Grafik aus [Web06b]).

Um die Wirkung des Fluors besser zu visualisieren, wurde eine einfache Monte Carlo Simulation der dazugehörigen Histogramme durchgeführt. Dazu wurde in einem ersten Schritt die kumulative Aktivierungsenergieverteilung F_{Ea} durch einen Mischverteilungsansatz bestehend aus zwei Normalverteilungen angenähert:

$$F_{Ea} = \alpha * F_1\left[x \mid \mu_1, \sigma_1\right] + (1-\alpha) * F_2\left[x \mid \mu_2, \sigma_2\right] \tag{9.1}$$

Hierbei sind $F_{1,2}$ die kumulativen Verteilungsfunktionen von Normalverteilungen mit den Mittelwerten $\mu_{1,2}$ und den Standardabweichungen $\sigma_{1,2}$

$$F\left[x \mid \mu, \sigma\right] = \int_0^x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x'-\mu)^2}{2\sigma^2}}\, dx' \tag{9.2}$$

und α der Gewichtungsfaktor zwischen den zwei Verteilungskomponenten. Die durchgezogenen Linien in Abbildung 9.5 entsprechen den zwei auf diese Weise modellierten Verteilungen. Die Übereinstimmung zwischen statistischen Modell und den Messdaten ist sehr gut. Im zweiten Schritt kann aus diesem statistischen Modell unter Berücksichtigung der Unterschiede in der Fehlerzahl ein Histogramm simuliert werden (siehe Abbildung 9.6). Dazu wurden für die Standardkomponenten 100000 Zufallszahlen generiert und für die F-implantierten Bausteine entsprechend der Verbesserung der Retentionfehlerzahl von 30% nur 70000 Zufallszahlen. Durch die Monte Carlo Simulation des Histogramms mit großer Samplezahl, die weit über der tatsächlichen Fehlerzahl liegt, wird der Einfluss des Fluors auf die Verteilung der Aktivierungsenergien deutlich.

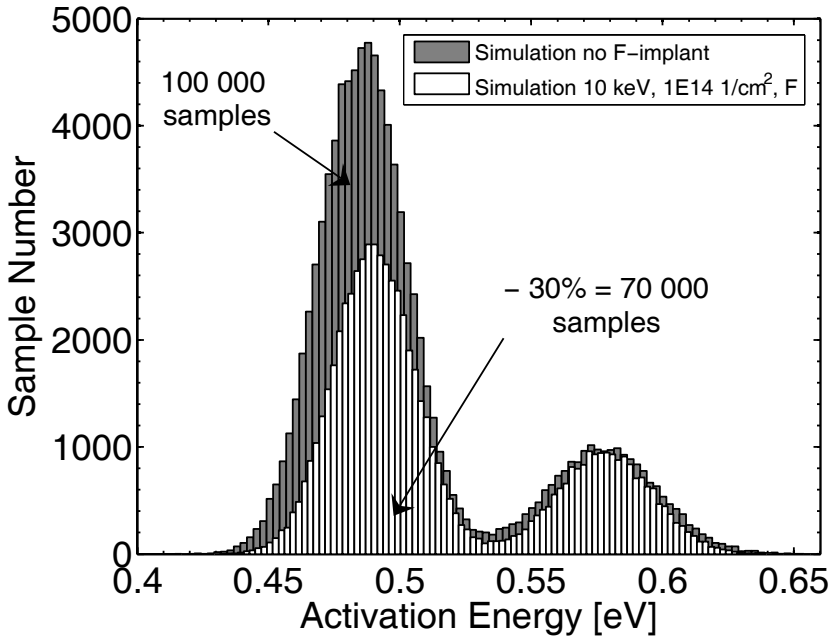


Abbildung 9.6: Durch Monte Carlo Simulation unter Berücksichtigung der erzielten Retentionverbesserung von 30% erhaltene Verteilung der Aktivierungsenergien für Zellen nahe der Reparaturgrenze. Es wird deutlich, dass durch die Fluorimplantation erheblich weniger Zellen mit Leckströmen mit niedriger Aktivierungsenergie auftreten (Grafik aus [Web06b]).

Der Vergleich der Histogramme in Abbildung 9.6 zeigt, dass die Reduktion der Retentionfehler durch Fluorimplantation auf das alleinige Verschwinden von Zellen mit Leckströmen, deren Aktivierungsenergien zu der Unterverteilung mit niedrigem Mittelwert gehören, aus der Standard-Aktivierungsenergieverteilung zurückzuführen ist. Genau dieser Anteil wurde in Kapitel 7.2 GIDL-Leckströmen zugesprochen. Dadurch konnte gezeigt werden, dass implantiertes Fluor wie beabsichtigt Traps im Bereich des Implantationsgebietes passiviert und sich die Zahl der Retentionfehler reduzieren lässt. Darüber hinaus bestätigen die Experimente die Richtigkeit des vorgeschlagenen Modells.

Zusammenfassung

Durch Fertigungsversuche auf Basis von $512M$ und $1G$ Speicherbausteinen in $110\,nm$ Technologie konnte gezeigt werden, dass durch Implantation von Fluor nach der Gatespacer-Strukturierung eine Reduzierung der Tailretentionfehler um bis zu 40% erreicht werden kann. Anhand aufwendiger temperaturabhängiger Messungen von Einzelzellen wurden die Aktivierungsenergieverteilungen für die beste Implantationsvariante bestimmt und mit Hilfe einer einfachen Monte Carlo Rechnung die zugehörigen Histogramme simuliert. Durch Vergleich der Aktivierungsenergiehistogramme von standardprozessierten und Fluor-implantierten Speicherbausteinen kann die Verbesserung auf Elimination von Speicherzellen mit niedrigen Aktivierungsenergien zurückgeführt wer-

den. Diese Aktivierungsenergien konnten in den vorherigen Kapiteln mit trapunterstütztem GIDL korreliert werden. Es konnte gezeigt werden, dass der Implantationszeitpunkt eine entscheidende Rolle für eine mögliche Verbesserung spielt. Darüber hinaus muss Fluor sehr nahe an die zu passivierende Stelle gebracht werden. Letztendlich entscheidet das Wechselspiel zwischen Trappassivierung durch Fluor und zusätzlichem Implantationsschaden über die resultierende Retentionfehlerzahl.

Kapitel 10

Zusammenfassung

Da die Retentionzeit einer Speicherzelle durch eine Vielzahl unterschiedlicher Faktoren und Prozesse beeinflusst wird, stellte sich das Thema „Retention" als ganz besonders interessant und umfassend aber auch herausfordernd heraus. Zu dessen erfolgreicher Bearbeitung ist vielerlei unterschiedlichstes Wissen notwendig. Angefangen mit dem physikalischem Verständnis der im Halbleiter vorgehenden Prozesse über DRAM-Funktionalität, -Integration und -Technologie bis hin zu detaillierten Aspekten der Speichertester-Programmierung und Monte Carlo Techniken. Das Thema „Retention" ist sehr vielseitig und konnte in dieser Arbeit sicherlich nicht in vollem Umfang behandelt werden. Dennoch ließen sich entscheidende Teilaspekte genauer beleuchten und die daraus gewonnenen Erkenntnisse werden sicherlich in die Entwicklung neuer Technologien einfließen. Darüber hinaus sind die Ergebnisse dieser Arbeit aufgrund der ständig weiter steigenden Anforderungen für zukünftige DRAM-Generationen von grundlegender Bedeutung. Diese können folgendermaßen zusammengefasst werden:

Modellierung & Simulation der Retentionverteilung

Die so genannte Retention-Formel wurde formal hergeleitet und beschreibt den Zusammenhang zwischen der Retentionzeit einer Speicherzelle und weiteren DRAM technologiespezifischen Größen, wie z.B. Bitleitungs- und Speicherkapazitäten, Leseverstärker-Offset sowie dem nicht direkt messbaren Leckstrom einer Zelle. Die in die Retention-Formel eingehenden Größen sind nicht für alle Speicherzellen eines Chips identisch, sondern unterliegen vielerlei Prozessschwankungen, sodass für jeden Parameter in der Realität nicht konkrete Werte, sondern ganze Verteilungen vorliegen. Ein im Zuge dieser Arbeit in MATLAB erstelltes Programm erlaubt unter Verwendung ganzer Verteilungen für die Parameter, die Simulation der Retentionkurve mit Hilfe von Monte Carlo Techniken. Die breite Retentionverteilung, die aus den Unterverteilungen „Tail" und „Main" besteht, kann nicht durch reine statistische Kombination der Technologiepara-

meter erklärt werden. D.h. insbesondere, dass der Retentiontail nicht durch ungünstiges Zusammentreffen der Designparameter entsteht, sondern von einer ebenso breiten und unterteilten Verteilung der Zell-Leckströme herrührt. Diese nicht direkt messbare Verteilung der Leckströme kann mit Hilfe des Programms aus den anderen Verteilungen simuliert werden. Durch Rücksimulation können anschließend Auswirkungen verschiedener Design-Parameteränderungen auf die gesamte Retentionverteilung analysiert werden.

Analyse der Leckstrompfade

Mögliche Leckstrompfade der $110\,nm$ Qimonda DRAM-Zelle wurden analysiert und Möglichkeiten der elektrischen Charakterisierung aufgewiesen. Die Leckstrompfade können in *symmetrische* und *asymmetrische* Leckstrompfade unterteilt werden. Die symmetrischen Leckströme *SubVt*, *Deep-SubVt*, *Node* und *SubSTI Leakage* führen zur Degradation beider Speicherzustände „0" und „1", während die asymmetrischen Leckströme *Junction Leakage* und *GIDL* die „1" verschlechtern und die „0" verbessern. Durch Messungen von spannungs- und temperaturabhängigen Retentionkurven der beiden Speicherzustände bis zu sehr hohen Retentionzeiten, können die für die schlechtesten Zellen dominierenden Leckstrompfade bereits auf *GIDL* und *Junction Leakage* eingeschränkt werden. Für detailliertere Untersuchungen mussten neue Charakterisierungsmethoden entwickelt werden.

Einzelzellmessungen & Aktivierungsenergie-Analyse

Die sehr geringe Auftrittswahrscheinlichkeit von Tailzellen zusammen mit einem relativ zur Durchschnittszelle um einen Faktor $10 - 300$ erhöhten Leckstrom macht deren detaillierte Charakterisierung sehr schwierig. Wie in Kapitel 6 ausgeführt, sind deshalb konventionelle Parametermessungen an Teststrukturen nicht in der Lage, Aussagen über Tailzellen zu machen. Mit Standard-Messmethoden können immer nur „Durchschnittszellen" charakterisiert werden, welche die Anforderungen bei weitem übertreffen und keinen limitierenden Einfluss auf die Chipfunktionalität haben. Ein wesentliches Ergebnis dieser Arbeit ist die Entwicklung einer Methode, die die damit verbundenen Schwierigkeiten umgeht. Die Lösung besteht in der Untersuchung einzelner gezielt ausgewählter Speicherzellen, direkt auf fertigen Speicherbausteinen. Dazu mussten neue Testprogramme für den MOSAID 3480 Speichertester entwickelt werden. Ein DUT-Interfaceboard für die Kontaktierung neuer FBGA-Komponenten am Tester musste ebenfalls erst entwickelt werden. Mit dem Testaufbau ist es nun möglich den Einfluss verschiedener Betriebsspannungen und der Chiptemperatur auf die Retentionzeit jeder beliebigen Zelle des Speicherbausteins zu messen. Ein wichtiger Vorteil der Methode ist, dass sie unabhängig von der Speicherkapazität des Chips und somit auch für alle zu-

künftigen Technologien anwendbar ist. Die konventionelle Methode der Charakterisierung an parallelen Teststrukturen hat bereits heute ihre Grenzen erreicht und wird mit zunehmender Speicherkapazität immer weiter an Aussagekraft verlieren.

Die Temperaturabhängigkeit der Retentionzeit und damit des dominierenden Leckstroms einer Einzelzelle kann mit der entwickelten Methode akkurat bestimmt werden. Dadurch konnte gezeigt werden, dass im Gegensatz zu bisherigen Annahmen zu Zellen mit annähernd gleicher Retentionzeit nicht eine einzelne, sondern eine breite Verteilung von Aktivierungsenergien gehört. Die oft in der Literatur verwendete Methode der Bestimmung von Aktivierungsenergien aus Retentionkurven führt demnach zu falschen Ergebnissen. Für Zellen nahe der Reparaturgrenze konnten zwei Unterverteilungen festgestellt werden. Die Unterverteilung mit kleineren Aktivierungsenergien kann gemäß der Spannungsabhängigkeiten dem Leckstrompfad *GIDL* zugeordnet werden, während die mit höheren Werten ihren Ursprung außerhalb der Reichweite des Gates (also BS-Junction) haben muss. Darüber hinaus sind die Aktivierungsenergien beider Unterverteilungen von der angelegten Spannung (entweder Gate oder Substrat) und damit vom elektrischen Feld abhängig.

Modellbildung

Die Ergebnisse der Einzelzellcharakterisierung erlauben zusammen mit theoretischen Betrachtungen und Abschätzungen folgende Beschreibung der Retentionverteilung. Aktivierungsenergien um $0.68\,eV$ sowie eine fehlende Spannungsabhängigkeit der Aktivierungsenergie legen Generationsströme an Grenzflächenzuständen als Ursache für die Mainverteilung nahe. Grenzflächenzustände können minimiert werden (darauf beruht der Erfolg der Si-Technologie). Traps führen aber, sofern sie in den Bereich eines gesperrten pn-Übergangs gelangen, unweigerlich zu Leckströmen. Die Tailverteilung entsteht durch Traps, die sich am Ort besonders hoher elektrischer Felder befinden und deren Generationsstrom durch einen Tunnelmechanismus verstärkt wird (TFE). Dies folgt aus dem kleinen Wert gekoppelt mit der E-Feld-Abhängigkeit der Aktivierungsenergien. Da sich die höchsten elektrischen Felder in der untersuchten Qimonda $110\,nm$ Zelle gemäß E-Feld-Simulationen im G/D Überlapp befinden, sind auch hauptsächlich Defekte in diesem Bereich für besonders hohe Leckströme und damit den Retentiontail verantwortlich. Aber auch bei anderen (stacked) Technologien sind die höchsten Felder in diesem Bereich zu finden, sodass das Ergebnis allgemeingültig ist. Direktes Band-zu-Band Tunneln kann dagegen ausgeschlossen werden, da die beobachteten Aktivierungsenergien dafür zu hoch liegen.

Verifikation durch Fertigungsversuche

Da Tail-Leckströme gemäß dem vorgeschlagenen Modell durch feldunterstützte Generation von Ladungsträgern ausgehend von Zuständen in der Bandlücke entstehen, gibt es prinzipiell zwei Möglichkeiten die besonders hohen Leckströme zu reduzieren. Einerseits kann der erhöhte Leckstrom durch geringere elektrische Felder auf den konventionellen SRH-Generationsstrom reduziert werden, andererseits führt eine „Passivierung" der Zustände zum Abschalten des Leckstroms. Der erste Ansatz der Feldreduzierung am kondensatorseitigen pn-Übergang wird von allen Speicherherstellern in Form neuer 3d-Transistoren verfolgt (z.B. RCAT, STAR, EUD ...). Leider geht das meist zu Lasten anderer Randbedingungen wie On-Strom oder der Prozesskomplexität. Deshalb wurde in dieser Arbeit der zweite Weg der Eliminierung der Ursache selbst, nämlich der Zustände in der Bandlücke, ohne negativen Einfluss auf andere Randbedingungen, gewählt. Durch geschickte Einbringung von Fluor mittels Implantation an geeigneter Stelle im DRAM-Gesamtprozess konnte eine Verbesserung des Retentiontails abhängig von der Implantationsdosis von bis zu 40% erzielt werden. Das Fluor muss dazu möglichst spät und sehr nahe an die zu passivierenden Stellen im G/D-Überlapp gebracht werden. Die Verbesserung kann durch stabile $F - Si$ Bindungen erklärt werden, welche offene Bindungen (dangling bonds) an der $Si - SiO_2$ Grenzfläche absättigen und dadurch die Zustandsdichte reduzieren. Durch Messungen der Aktivierungsenergieverteilungen der schlechtesten Zellen von F-implantierten und Standard-Speicherbausteinen, konnte dies bestätigt werden. Das vorgeschlagene Modell konnte demnach durch „gezielte Verbesserung" verifiziert werden.

Ausblick

Die Retention ist in heutigen und bleibt auch in zukünftigen DRAM Generationen ein die Entwicklung bestimmendes Thema. Ohne weitere Verbesserung kann die ITRS Roadmap nicht erfüllt werden. Mit dieser Arbeit konnte ein Beitrag zum besseren Verständnis und sogar zur Verringerung der Problematik durch einen kostengünstigen Prozess im Frontend geleistet werden. Über die betrachteten Aspekte der Arbeit hinaus existieren noch weitere wichtige daran anknüpfende Themen, wie z.B. Retention-Degradation im Backend bzw. *Variable Retention Time* (VRT), die mit zunehmenden Speicherdichten weiter an Bedeutung gewinnen. Diese gilt es in der Zukunft zu einem vollständigen Bild zusammenzufassen und Lösungen dafür zu finden.

Literaturverzeichnis

[Bro95] Bronner G., Aochi H., Gall M., Gambino J., Gernhardt S., Hammerl E., Ho H., Iba J., Ishiuchi H., Jaso M., Kleinhenz R., Mii T., Narita M., Nesbit L., Neumueller W., Nitayama A., Ohiwa T., Parke S., Ryan J., Sato T., Takato H., Yoshikawa S. *A Fully Planarized 0.25 μm CMOS Technology for 256 Mbit DRAM and Beyond*. In *Symposium on VLSI Technology*, pp. 15–16. 1995.

[Den68] Dennard R. *Field-Effect Transistor Memory*. U.S. Patent, 1968.

[Den84] Dennard R. *Evolution of the MOSFET Dynamic RAM - A Personal View*. IEEE Trans. Electron Devices, 31(11), pp. 1549–1555, 1984.

[Goe02] Goebel B., Luetzen J., Manger D., Moll P., Muemmler K., Popp M., Scheler U., Schloesser T., Seidl H., Sesterhenn M., Slesazeck S., Tegen S. *Fully depleted Surrounding Gate Transistor (SGT) for 70nm DRAM and beyond*. In *IEDM Tech. Dig.*, pp. 275–278. 2002.

[Hal51] Hall R. *Germanium Rectifier Characteristics*. Phys. Rev., 83(228), 1951.

[Hal52] Hall R. *Electron-Hole Recombination in Germanium*. Phys. Rev., 87(387), 1952.

[Ham98] Hamamoto T., Sugiura S., Sawada S. *On the retention time distribution of dynamic random access memory (DRAM)*. IEEE Transactions on Electron Devices, 45(6), pp. 1300–1309, 1998.

[Hum03] Hummler K. *DRAM Technology and its Interaction with Design and Test*. Qimonda Internal Seminar, 2003.

[Hur92] Hurkx G., Klaassen D., Knuvers M. *A new recombination model for device simulation including tunneling*. IEEE Transactions on Electron Devices, 39(2), pp. 331–339, 1992.

[Ito01] Itoh K. *VLSI Memory Chip Design*. Springer-Verlag, 2001.

[ITR05] ITRS. *INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS (ITRS)*. www.itrs.net, 2005.

[Jan05] Jang M., Seo M., Kim Y., Cha S., Kim Y., Kim S., Rhee J., Cheong J., Jung T., Pyi S., Kim H., Jeong J., Park S., Hong S., Park S. *Enhancement of data retention time in DRAM using step gated asymmetric (STAR) cell transistors*. In *European Solid-State Device Research Conference*, pp. 189–192. 2005.

[Kee01] Keeth B., Baker R. *DRAM Circuit Design: A Tutorial*. IEEE Press, 2001.

[Ken92] Kenney D., Parries P., Pan P., Tonti W., Cote W., Dash S., Lorenz P., Arden W., Mohler R., Roehl S., Bryant A., Haensch W., Hoffmann B., Levy M., Yu A., Zeller C. *A buried-plate trench cell for a 64-Mb DRAM*. In *Symposium on VLSI Technology*, pp. 14–15. 1992.

[Kim03] Kim J., Lee C., Kim S., Chung I., Choi Y., Park B., Lee J., Kim D., Hwang Y., Hwang D., Hwang H., Park J., Kim D., Kang N., Cho M., Jeong M., Kim H., Han J., Kim S., Nam B. *The breakthrough in data retention time of DRAM using Recess-Channel-Array Trasistor (RCAT) for 88 nm feature size and beyond*. In *Symposium on VLSI Technology*, pp. 11–12. 2003.

[Kim05] Kim J., Woo D., Oh H., Kim H., Kim S., Park B., Kwon J., Shim M., Ha G., Song J., Kang N., Park J., Hwang H., Song S., Hwang Y., Kim D., Kim D., Huh M., Han D., Lee C., Park S., Kim Y., Lee Y., Jung M., Kim Y., Lee B., Cho M., Choi W., Kim H., Jin G., Park Y., Kim K. *The excellent scalability of the RCAT (recess-channel-array-transistor) technology for sub-70nm DRAM feature size and beyond*. In *Symposium on VLSI Technology*, pp. 33–34. 2005.

[Min99] Min D.S., Langer D. *Multiple twisted dataline techniques for multigigabit DRAMs*. IEEE Journal of Solid-State Circuits, 34(6), pp. 856–865, 1999.

[Mog97] Mogul H., Rost T., Lin D.G. *Advantages of LDD-only implanted fluorine with submicron CMOS technologies*. IEEE Transactions on Electron Devices, 44(3), pp. 388–394, 1997.

[Moo65] Moore G. *Cramming more components onto integrated circuits*. Electronics, 38(8), 1965.

[Mue05] Mueller W., Aichmayer G., Bergner W., Erben E., Hecht T., Kapteyn C., Kersch A., Kudelka S., Lau F., Luetzen J., Orth A., Nuetzel J., Schloesser T., Scholz A., Schroeder U., Sieck A., Spitzer A., Strasser M., Wang P.F., Wege S., Weis R. *Challenges for the DRAM Cell Scaling to 40nm*. In *IEDM Technical Digest*, pp. 336–339. 2005.

[Nes93] Nesbit L., Alsmeier J., Chen B., DeBrosse J., Faheyk P., Gall M., Gambino J., Gernhard S., Ishiuchi H., Kleinhenz R., Mandelman J., Mii T., Morikado M., Nitayama A., Parke S., Wong H., Bronner G. *A 0.6 μm^2 256 Mb trench DRAM*

cell with self-aligned BuriEd STrap (BEST). In *IEDM Technical Digest*, pp. 627–630. 1993.

[Oh05] Oh H., Kim J., Kim J., Park S., Kim D., Kim S., Woo D., Lee Y., Ha G., Park J., Kang N., Kim H., Hwang J., Kim B., Kim D., Cho Y., Choi J., Lee B., Kim S., Cho M., Kim Y. *High-density low-power-operating DRAM device adopting $6F^2$ cell scheme with novel S-RCAT structure on 80nm feature size and beyond*. In *European Solid-State Device Research Conference*, pp. 177–180. 2005.

[Pic04] Pichler P. *Intrinsic Point Defects, Impurities, and Their Diffusion in Silicon*. Springer-Verlag, 2004.

[Pie96] Pierret R. *Semiconductor Device Fundamentals*. Addison Wesley, 1996.

[Pie03] Pierret R. *Modular Series on Solid State Devices - Volume VI - Advanced Semiconductor Fundamentals*. Prentice Hall, 2003.

[Qim05] Qimonda. *Produktbeschreibung 512MBit DDR SDRAM (HYB25D512400C)*. 2005.

[Sho52] Shockley W., Read W. *Statistics of the Recombination of Holes and Electrons*. Phys. Rev., 87(5), pp. 835–842, 1952.

[Sze81] Sze S. *Physics Of Semiconductor Devices*. John Wiley & Sons, Inc., 1981.

[Tau98] Taur Y., Ning T. *Fundamentals of Modern VLSI Devices*. Cambridge University Press, 1998.

[Web06a] Weber A., Birner A., Krautschneider W. *Method of activation energy analysis and application to individual cells of 256Mb DRAM in 110nm technology*. Solid State Electronics, 50(4), pp. 613–619, 2006.

[Web06b] Weber A., Birner A., Krautschneider W. *Retention Tail Improvement for Gbit DRAMs through Trap Passivation confirmed by Activation Energy Analysis*. In *European Solid-State Device Research Conference*, pp. 250–253. 2006.

[Wid96] Widmann D., Mader H., Friedrich H. *Technologie hochintegrierter Schaltungen*. Springer-Verlag, 2 ed., 1996.

[Wol95] Wolf S. *Silicon Processing for the VLSI Era - Volume 3 - The Submicron MOSFET*. Lattice Press, 1995.

[Wol00] Wolf S., Tauber R. *Silicon Processing for the VLSI Era - Volume 1 - Process Technology*. Lattice Press, second ed., 2000.

[Wol02] Wolf S. *Silicon Processing for the VLSI Era - Volume 4 - Deep-Submicron Process Technology*. Lattice Press, 2002.

Danksagungen

An dieser Stelle möchte ich Prof. Dr. Wolfgang Krautschneider der Technischen Universität Hamburg-Harburg für die sehr gute Betreuung während dem gesamten Verlauf der Dissertation danken. Außerdem danke ich ganz besonders Prof. Dr. Wolfgang Albrecht für die Zweitbegutachtung dieser Arbeit.

Auf Seiten der Qimonda Dresden GmbH & Co. OHG danke ich Herrn Dr. A. Birner für die ausgezeichnete Betreuung vor Ort. Besonders die vielen interessanten Diskussionen während der Kaffeepausen trugen zum Erfolg der Arbeit bei.

Außerdem danke ich Herrn Dr. J. Lützen für die Möglichkeit, dieses interessante Thema im Rahmen des Qimonda-Doktorandenprogrammes und innerhalb seiner Abteilung *Technology Innovation (TIN)* bearbeiten zu können.

Ich danke D. Weinmann für viele aufschlussreiche Diskussionen, Dr. S. Slesazeck für die Beantwortung aller anfallenden Simulationsfragen sowie A. Danneberg für die Hilfe bei technischen Fragen und Problemen mit dem MOSAID Speichertester.

Ganz besonderer Dank gilt meinem Studienkollegen A. Gatto für das sehr aufmerksame Korrekturlesen der Arbeit.

Vor allem J. Lützen und S. Slesazeck sorgten während den letzten drei Jahren durch regelmäßiges Klettertraining in der XXL-Kletterhalle und den Klettergärten der Dresdner Umgebung für meinen körperlichen Ausgleich zur Arbeit. Ihnen und allen weiteren Mitkletterern besten Dank dafür.

Schließlich danke ich meiner Frau Chiharu für ihr Verständnis und die super Unterstützung sowie meinen Eltern Bernhard und Elfriede Weber, die mich durch meine gesamte Studienzeit hindurch unterstützt haben.

Auch allen Anderen, die mir während meiner Zeit als Doktorand hilfreich zur Seite standen und hier nicht namentlich erwähnt werden konnten, gilt mein herzlichster Dank.

DANKE

Andreas Weber

Büroanschrift:
Königsbrücker Str.180
01099 Dresden
Germany
+49 (0)351 886 7782
Andreas.Weber.drs@qimonda.com

Privatanschrift:
Eisenberger Str.5
01127 Dresden
Germany
+49 (0)351 811 2375
Andi_Weber@gmx.de

Persönliche Daten:

Geburtsdatum:	12. Februar 1976, Heidenheim a.d. Brenz, Germany
Familienstand:	verheiratet seit 19.09.2003 mit Chiharu Okada-Weber

Ausbildung:

Vordiplom:	Physik, Eberhard-Karls-Universität Tübingen, November 1998
	Informatik, Eberhard-Karls-Universität Tübingen, März 1999
Auslandsstudium:	Physik, San Francisco State University, USA, August 1999 - Juli 2000
Diplom:	Physik, Eberhard-Karls-Universität Tübingen, Februar 2003
	„*Fluxondynamik in annularen intrinsischen Josephson-Kontakten*"
Dissertation:	Elektrotechnik, Hamburg University of Technology
	„*Charakterisierung von Leckstrompfaden in DRAM Zellen und deren Reduktion*"

Berufserfahrung:

04/1999 - 08/1999	Praktikumsleitung in Technischer Informatik, Universität Tübingen
02/2000 - 08/2000	Student Lab Assistant, Thin Film Lab, San Francisco State University
03/1998 - 03/2003	Werkstudententätigkeiten bei IBM Deutschland Entwicklung GmbH
03/2003 - 09/2005	Promotionsstelle bei Infineon (später Qimonda), Dresden
seit 10/2005	Systemexperte, Integration, Qimonda, Dresden

Sonstiges:

07/1999 - 07/2000	Fulbright Stipendium
Hobbys:	Reisen, Sportklettern, Musik

Publikationen in Zusammenhang mit dieser Arbeit

Weber A., Birner A. und Krautschneider W., *Data Retention Analysis On Individual Cells Of 256Mb DRAM In 110nm Technology*, In *European Solid-State Device Research Conference*, pp. 185-188, 2005.

Weber A., Birner A. und Krautschneider W., *Method of activation energy analysis and application to individual cells of 256Mb DRAM in 110nm technology*, Solid State Electronics, Volume 50(4), pp. 613-619, 2006.

Weber A., Birner A. und Krautschneider W., *Retention Tail Improvement for GBit DRAMs through Trap Passivation confirmed by Activation Energy Analysis*, In *European Solid-State Device Research Conference*,pp. 250-253, 2006.

Patentanmeldungen mit Qimonda GmbH & Co. OHG.

Birner A., Lützen J., Schlösser T. und Weber A., *Verfahren zur Erzeugung eines Dielektrikums und Halbleiterstruktur*, (DE: 102004031453.5, USA: 11/167,946)

Birner A., Weber A. und Weis R., *Herstellungsverfahren für eine Halbleiterstruktur und entsprechende Halbleiterstruktur*, (DE: 102005037566.9, JP: 2006-175042, TW: 095117043, USA: 11/477,577)

Birner A., Ludwig F., Mothes K., Radecker J., Weber A., Wilson K., *Integrated Circuit formed on a semiconductor substrate*, (USA: 11/269,897)

Birner A., Stadtmüller M., Storbeck O., Weber A., Wieland P., *Method for fabricating an integrated circuit with a CMOS manufacturing process*, (USA: 11/270,820)

7 weitere Erfindungsmeldungen sind seitens Qimonda in Bearbeitung.