Technische Universität Dresden

Modellierung und Charakterisierung des elektrischen Verhaltens von haftstellen-basierten Flash-Speicherzellen

Dipl.-Ing. Thomas Melde

von der Fakultät Elektrotechnik und Informationstechnik der Technischen Universität Dresden

zur Erlangung des akademischen Grades eines

Doktoringenieurs

(Dr.-Ing.)

genehmigte Dissertation

Vorsitzender:	Prof. Dr. rer. nat. J. W. Bartha		
Gutachter:	Prof. Dr.-Ing. T. Mikolajick	Tag der Einreichung:	03.12.2009
	Prof. Dr.-Ing. habil. Dipl.-Math. B. Meinerzhagen	Tag der Verteidigung:	01.09.2010

Kurzfassung

Die rasante Miniaturisierung der Floating-Gate-Struktur in NAND-Flash-Speichern kommt zunehmend an ihre Grenzen. Daher besteht ein großes Interesse, einerseits diese Technologie weiterzuentwickeln, aber auch alternative Speichertechnologien zu untersuchen, die diese ersetzen könnten. Am weitesten Fortgeschritten ist die Entwicklung von haftstellen-basierten Speicherzellen. Bei dieser Speichertechnologie erfolgt die Ladungsspeicherung in einem haftstellen-reichen Dielektrikum, welches zwischen zwei isolierenden Schichten eingebettet ist. Diese Arbeit befasst sich mit der Untersuchung von haftstellen-basierten Strukturen und deren Anwendung in skalierten NAND-Speicherzellenfeldern.
Ein Ziel dieser Arbeit ist es, die Funktionalität solch einer Zelle tief greifend zu verstehen. Entscheidend hierbei ist, wie sich die Ladung in der Speicherschicht anordnet. Hierzu wurden Simulationen durchgeführt die das elektrische Verhalten anhand einfacher physikalischer Modelle nachbilden. Durch den Vergleich der Simulationsergebnisse mit Messungen war es möglich, zum Beispiel die Ladungsträgerinjektion besser zu verstehen. Parallel dazu wurde mittels gezielter Auswahl von Messungen an verschiedenen Schichtstapeln auf die Ladungsverteilung geschlossen. Basierend auf diesem Ansatz konnten Effekte erklärt werden, wie zum Beispiel der speicherschichtdickenabhängige Ladungsverlust.
Weiterhin befasst sich die Arbeit mit dem Programmier- und Löschverhalten von MOS-Transistoren mit einer haftstellen-basierten Speicherschicht. Durch die Ladungsspeicherung in einem Dielektrikum ist die Bildung einer inhomogenen Ladungsverteilung begünstigt, was zu einem unerwarteten Verhalten des Speicherelements führen kann. Ursache kann etwa eine über die Weite der Gateelektrode hinausreichende Speicherschicht sein, verursacht durch eine nicht optimale Ätzung des Al_2O_3-Topoxids. Zudem wurde untersucht, wie sich die Materialwahl des Topoxids und des Gateelektroden-Materials auf das elektrische Verhalten der Speicherzellen auswirkt. Hierbei konnte gezeigt werden, dass die besten elektrischen Eigenschaften erzielt werden, wenn reines Al_2O_3 in H_2-Atmosphäre bei 1100°C kristallisiert wird. Das Gateelektrodenmaterial mit den günstigsten Eigenschaften im Hinblick auf Ladungshaltung und Löschbarkeit ist amorphes Tantalnitrid mit einem hohen Kohlenstoff-Anteil.

Die Integration in NAND-Speicherfeldern bedingt eine Strukturierung mit einer Grabenoxid-Isolation. Die Untersuchung von verschiedenen Strukturformen hat verdeutlicht, dass die Zellstruktur möglichst nah an der ideal flachen Form liegen muss, um ein optimales elektrisches Verhalten zu erzielen. Es hat sich zudem gezeigt, dass bei TANOS-Zellen eine Schädigung der Metall-Gateelektrode während der Al_2O_3-Ätzung auftritt. Der hierbei hervorgerufenen Verschlechterung der elektrischen Eigenschaften, im Besonderen der Ladungshaltung, kann durch die Integration einer Kapselungsschicht entgegengewirkt werden. Weiterhin führt die starke Miniaturisierung bei der Anwendung in NAND-Speichern zu unerwünschten Nebeneffekten. So wird zum Beispiel gezeigt, dass es unter Löschbedingungen an den Speicherzellen zu einem Programmieren der Auswahltransistoren kommen kann.
Die durchgeführten Untersuchungen bilden einen weiteren Schritt in Richtung des Verständnisses von haftstellen-basierten Speicherzellen. Zudem konnte nachgewiesen werden, dass diese Technologie auch auf stark skalierten Strukturen anwendbar ist. Es werden verschiedene Optimierungen vorgestellt, die in ihrer Gesamtheit zu einer Speicherzelle führen, die nah an die gestellten Anforderungen heranreicht.

Abstract

The rapid scaling of the floating-gate structure used in non volatile stand alone NAND flash memories is reaching a limit in the near future. Therefore, a large interest is rising up to either push the limit to smaller dimensions or to bring up an alternative technology. The most promising candidate for replacing floating-gate in the next generations is the charge trapping memory device. This memory technology utilizes the effect that a dielectric with an high trap density placed in between two insulating layers can store charges over a long time period. The thesis focuses on the electrical characterisation, modelling and integration of charge trapping memory devices into highly scaled memory arrays.

One objective of this work is to understand the fundamental functionality of the charge trap stack. The most important aspect is the analysis of the vertical charge distribution in the storage layer after program and erase operation and its impact on the retention. For this, charge injection simulations have been carried out to describe the electrical behaviour by the use of simplified physical models. Comparing the simulation results and measurement data, also the injection mechanisms could be better understood. In addition, the charge distribtion was deduced from a measurement analysis of different charge trapping stacks and dimensions. Based on this approach, the charge loss dependent on the storage layer thickness could be explained.

Furthermore, this thesis adresses the electrical characteristic of MOS memory transistors containing a charge trapping storage layer. Due to the possibility of lateral nonunifomity of the charge distribution unexpected device behaviour has been observed. Such a non-ideal behaviour was found for an extended charge storage layer in length direction of the memory cell. This structural issue is the result of a not optimised alumina etch process. Additionally, the influence of the top oxide composition and gate material on the electrical characteristic was investigated. It is demonstrated that the best performance is achieved by annealing pure Al_2O_3 in H_2 ambient at 1100°C. Regarding the gate material, amorphous TaN with an high carbon content is the recommended choice.

The integration of the memory cell into NAND-arrays involves the structuring with shallow trench isolation. This structure introduces inhomogenous field distributions to the memory. Investigations on memory cells with different STI step heights in experiments and simulations are demonstrating that the nearly planar device lead to the best performance. An etch damage of the metal gate side walls is observed for a conventional gate patterning process. As a result, the electrical performance, especially the retention, is adversely affected. However, this integration issue can be strongly improved by the introduction of an encapsulation liner. The shrinkage to small dimension for a NAND product also results in unwanted side effects. It is observed that the charge trap stack containing select devices of a NAND string are programmed at erase with conventional settings.

The presented investigations depict a major improvement in the understanding of the charge trap stack functionality. Moreover, it is demonstrated that this technology approach can be scaled down to 48 nm. Different optimisations regarding the stack and materials are presented. The result is a memory device nearly reaching the requirements needed for a product.

Inhaltsverzeichnis

1 Einleitung

Die Einführung von Flash-Speichern hat innerhalb weniger Jahre die Nutzung digitaler Medien revolutioniert. Heutzutage ist der moderne Mensch ohne mobiles MP3-Musik-Abspielgerät oder digitale Fotografie nicht mehr vorstellbar. Ein Fakt, der zu der rasanten Erfolgsgeschichte beigetragen hat, ist die schnelle Weiterentwicklung von Flash-Speichern hin zu immer höherer Kapazität bei gleichzeitig sinkenden Preisen. Die aktuell marktübliche Floating-Gate Speicherzelle (dt. schwebende/ potentialfreie Gateelektrode) hat es den Herstellern auf einfache Weise möglich gemacht, die Zellen immer weiter zu skalieren, ohne große Änderungen in der Technologie durchführen zu müssen. Dadurch wurde ein schnelles Wachstum des Marktes ermöglicht. Die große Attraktivität der relativ einfachen Herstellung hat aber gleichzeitig auch den Druck auf die Hersteller erhöht, folgende Generationen schneller einzuführen, um profitabel fertigen zu können. Die Entwicklung ist zum Zeitpunkt dieser Arbeit bei Strukurgrößen von circa 35 nm Transistorlänge und -weite angekommen, wie in Abb. 1.1 gezeigt wird [1, 2].

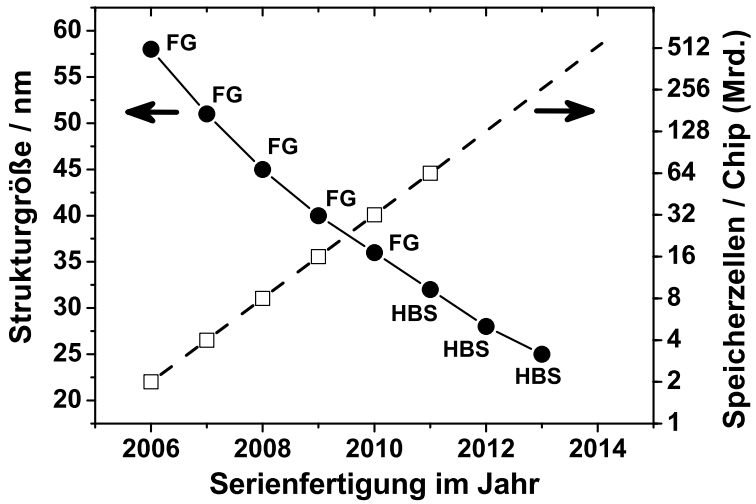

Abbildung 1.1: Durch die ITRS bestimmten Kennwerte für die Strukturgrößen und die Anzahl Speichertransistoren pro Speicherchip; FG = Floating-Gate; HBS = haftstellen-basierte Speicher; entsprechend [1]

Die bisherige und weitere Entwicklung wird durch die 'International Technology Roadmap for Semiconductor' (ITRS) wiedergegeben. Diese spiegelt die von verschiedenen Herstellern beabsichtigten Pläne wider. Es wird in Abb. 1.1 gezeigt, dass gegenwärtig Produkte mit 16 Milliarden Speicherzellen gefertigt werden. Die Anzahl an Speicherzellen hat sich in den letzten Jahren jedes Jahr nahezu verdoppelt. Dieser Trend kann aber auf Grund von physikalischen Grenzen bei der verwendeten Floating-Gate Technologie nicht weiter fortgesetzt werden [3, 4]. In Abb. 1.2a ist die Struktur einer Floating-Gate Speicherzelle schematisch dargestellt. Um eine funktionale Zelle mit dem aktuell verwendeten Aufbau zu bekommen, ist es zwingend erforderlich, dass

die Fläche zwischen der Steuerelektrode und dem Floating-Gate so groß wie möglich ist. Dies wird durch eine Erweiterung der Steuerelektrode zwischen den Speicherzellen gewährleistet, die das Floating-Gate umschließt, wie Abb. 1.2a veranschaulicht. Versucht man nun die Struktur immer kleiner zu machen, ist man bei der Verringerung der Isolationsschichtdicken begrenzt, da ansonsten die gespeicherte Ladung zu schnell verloren geht. Daher lassen sich prinzipiell nur die Weite des Floating-Gate und der Steuerelektroden-Erweiterung d verringern. Unter Programmierbedingungen kann es bei zu dünner Erweiterung zur Verarmung kommen. Wird die Dicke d zu sehr verringert (< 15 nm) kommt es zur kompletten Verarmung und die Erweiterung verliert ihre Funktion [5]. Eine mögliche Option dies zu umgehen, besteht in einer planaren Floating-Gate Zelle ohne Elektrodenerweiterung, wie in Abb. 1.2b illustriert.

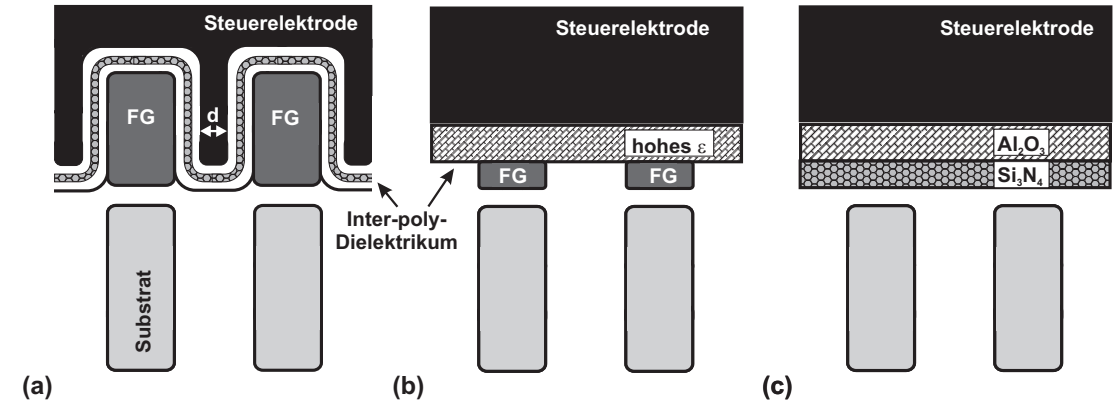

Abbildung 1.2: Drei mögliche Konzepte für hochintegrierte Flash-Speicher; (a) klassisches Floating-Gate-Konzept, (b) planare Floating-Gate-Zelle und (c) TANOS-Konzept mit haftstellen-basierter Speicherschicht

Für eine funktionale Speicherzelle ist es aber notwendig, die Kopplung zwischen der Steuerelektrode und dem Floating-Gate aufrecht zu erhalten. Die Funktion der Elektrodenerweiterung übernimmt in dem Fall einer planaren Zelle ein Material mit hoher Dielektrizitätskonstante ϵ_r, wie zum Beispiel Zirkoniumoxid oder Hafniumoxid. Dadurch ist sichergestellt, dass die Kopplung zwischen den Elektroden groß genug ist. Allerdings haben diese Materialien den Nachteil, dass sie im Vergleich zu Siliziumoxid (SiO_2) eine erhebliche größere Zahl an Fehlstellen haben. Damit ist die Bildung eines Leckpfades sehr stark begünstigt und die Wahrscheinlichkeit, dass die gespeicherte Ladung und somit die Information verloren geht, sehr groß. Haftstellenbasierte Speicherzellen sind deutlich unempfindlicher gegenüber solchen Leckpfaden. Da die Ladung nur im Bereich des Leckpfades gestört ist, geht nicht die gesamte Information verloren, wie im Falle von Floating-Gate. Dort entlädt ein Leckpfad die komplette Zelle. Der Aufbau einer haftstellen-basierten Speicherzelle ist sehr ähnlich der einer planaren Floating-Gate Speicherzelle, wie es durch Abb. 1.2b,c verdeutlicht wird. Allerdings wird als Speicherschicht ein Dielektrikum mit einer hohen Dichte an Fehlstellen verwendet, wie zum Beispiel Siliziumnitrid (Si_3N_4). Um haftstellenbasierte Speicherzellen ausreichend löschen zu können, ist es wie bei einer planaren Floating-Gate Speicherzelle notwendig, ein Dielektrikum mit hohem ϵ_r zu verwenden. Als am Besten geeignet hat sich Aluminiumoxid (Al_2O_3) herausgestellt. Aufgrund der genannten Grenzen für die Skalierung der Floating-Gate Struktur wird derzeit angenommen, dass haftstellen-basierte Speicher unterhalb einer Zellgröße von circa 30

nm eingeführt werden. Ein weiterer großer Vorteil der einfachen Strukturierung ist die Umsetzung in 3-dimensionalen Strukturen, die eine beachtenswerte Erhöhung der Speicherdichte nach sich ziehen [6, 7]. Um dieses neue Zellkonzept einzuführen, ist es aber notwendig, die Funktionsweise einer haftstellen-basierten Speicherzelle zu verstehen.
In dieser Arbeit werden zunächst im **Kap. 2** die Voraussetzungen für das Verständnis der untersuchten Strukturen geschaffen. Darin werden der Kondensator und Transistor eingeführt, sowie ein erster Überblick über die Funktionsweise von Speicherzellen und Zellenfeldern gegeben. Für die Untersuchung der Speicherzellen sind entsprechende Charakterisierungsmethoden notwendig, die auch in diesem Kapitel erläutert werden.
Daran schließt sich das **Kap. 3** an, in dem zunächst ein Einblick in die physikalischen Grundlagen von Haftstellen in dielektrischen Schichten gegeben wird. Aufbauend darauf erfolgt die Analyse von Messdaten mittels Injektionssimulationen, die das Programmieren und Löschen beschreiben. Hierbei werden verschiedene Modifikationen vorgestellt, mit denen es möglich ist, eine bessere Übereinstimmung von Simulation und Messung zu erzielen. Es erfolgt zudem eine Betrachtung der vertikalen Ladungsverteilung in der Speicherschicht mit Hilfe der Simulationen. Aber auch durch eine passende Auswahl an Messungen ist es möglich, die Ladungsverteilung qualitativ abzuleiten. Dies wird anhand von Messungen an SONOS- und TANOS-Strukturen in diesem Kapitel gezeigt.
Bei einer haftstellen-basierten Speicherzelle hat die Materialwahl der einzelnen Schichten einen großen Einfluss auf die elektrischen Eigenschaften und wird daher in **Kap. 4** betrachtet. Es wird untersucht, inwieweit die Wahl des Topoxids, der Speicherschicht und der Gateelektrode das Programmier- und Löschverhalten beeinflussen. Außerdem werden verschiedene Aspekte der Strukturierung von kleinen Speicherzellen und ihr Einfluss auf das elektrische Verhalten analysiert.
Abschließend wird in **Kap. 5** gezeigt, dass es möglich ist, haftstellen-basierte Speicherzellen in Zellenfeldern mit höchster Dichte zu implementieren. Es werden verschiedene Störmechanismen mit strukturellem und algorithmischem Hintergrund genauer betrachtet. Zudem werden Vorschläge unterbreitet, wie es möglich ist, die negativen Auswirkungen der Störmechanismen zu unterdrücken.

2 Grundlagen aktiver Halbleiterelemente

2.1 Die MOS-Struktur

2.1.1 Grundlagen des MOS-Kondensators

Bei der MOS-Struktur handelt es sich um einen Schichtstapel, bestehend aus einer Metallelektrode, einem isolierenden Oxid und einer Silizium-Substrat-Elektrode. Die Metallelektrode wird in den meisten Fällen durch ein hochdotiertes Silizium ersetzt, welches nahezu metallisches Verhalten besitzt. Abbildung 2.1a zeigt eine solche Struktur. Das Isolatormaterial ist im Allgemeinen aufgrund seiner einfachen Fertigung Siliziumoxid (SiO_2), wird aber immer mehr durch hoch-ϵ Materialien (mit hoher Dielektrizitätskonstante ϵ, engl. high-k) verdrängt.

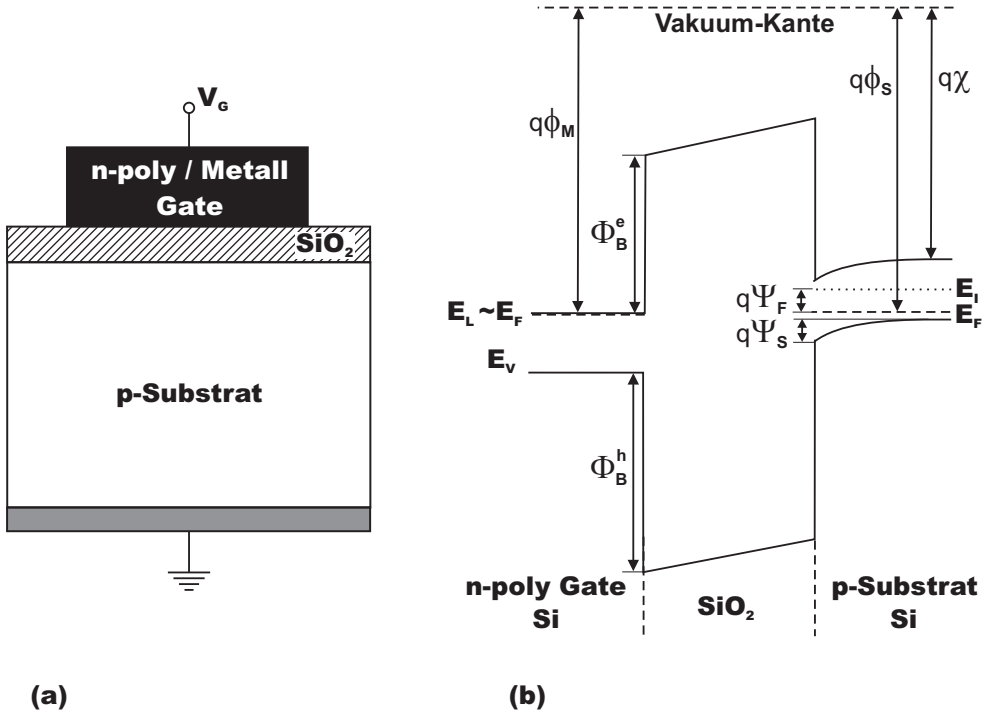

Abbildung 2.1: (a) Aufbau einer MOS-Kapazität und (b) Bändermodell für Leitungs- und Valenzband mit den charakteristischen Größen bei einem p-Substrat und n-poly Gate und $V_G = 0$ V

Ein großer Vorteil dieser Struktur ist deren Einfachheit. Es reicht aus, wenn man einen Silizium-Wafer mit dem zu analysierenden Isolatormaterial ganzflächig beschichtet und im Anschluss mit einem geeigneten Verfahren kleine Kondensator-Elektroden aufbringt [8]. Mit wenigen Prozessschritten hat man eine hinreichende

Probe geschaffen. Daher eignet sie sich zur einfachen Charakterisierung zum Beispiel von Isolator-Materialien, aber auch von haftstellen-basierten Speicherschichten. Der Fokus der heutigen Forschung liegt bei der Suche nach geeigneten Isolations-Materialien für Logik-MOSFET-Transistoren [9] und für die Kapazitäten in DRAM-Speicherbausteinen [10]. Aufgrund der frei wählbaren Strukturgröße ist es möglich, eine Vielzahl von Messungen, wie zum Beispiel Leckstrommessungen durchzuführen, die an skalierten Strukturen nicht vorgenommen werden können.
Für das Verständnis des spannungsabhängigen Verhaltens der MOS-Struktur ist eine Betrachtung des Bändermodells hilfreich. Das in Abb. 2.1b gezeigte Schema gibt den Bandverlauf für die in (a) gezeigte Struktur wieder. E_L und E_V bezeichnen das Valenz- beziehungsweise das Leitungsband im Halbleiter. Da die Gateelektrode generell sehr hoch dotiert ist, bekommt sie annähernd metallisches Verhalten, da das Ferminiveau (E_F) der Energie des Leitungsbandes entspricht oder leicht darüber liegt (entarteter Halbleiter). Dieses Verhalten wird genutzt, um die später durchgeführten Simulationen zu vereinfachen. Dabei wird die Verarmung in der poly-Silizium Gateelektrode nicht berücksichtigt. In der gezeigten Struktur befinden sich die Ferminiveaus von Substrat und Gateelektrode auf dem gleichen Potential. Dies bedeutet, dass keine äußere Spannung angelegt ist. Aber wegen der unterschiedlichen Dotierung, die in einer unterschiedlichen Position des Fermi-Niveaus resultiert, kommt es zu einem Spannungsabfall über dem Oxid, repräsentiert durch die Bandverkippung. Dieser Spannungsabfall lässt sich über die Austrittsarbeiten der beiden Elektroden und die Bandverbiegung ($q\psi_S$) berechnen. Die Austrittsarbeit $q\phi_M$ gibt die Energiedifferenz zwischen Ferminiveau und Vakuumenergie für die Gateelektrode wieder. Entsprechend ist $q\phi_S$ die Austrittsarbeit für die Substrat-Elektrode, wie in Abbildung 2.1b veranschaulicht. Die Differenz $q\phi_{MS}$ lässt sich über [11]

$$q\phi_{MS} = q\left(\phi_M - \phi_S\right) = q\phi_M - \left(q\chi + \frac{E_L - E_V}{2} + q\psi_F\right) \tag{2.1}$$

berechnen und entspricht dem Spannungsabfall über dem Oxid. Die Elektronenaffinität ($q\chi$) beschreibt die Energiedifferenz zwischen Leitungsband des Halbleiters und der Vakuumenergie. Der Einfluss der Dotierung wird im zweiten Teil von Gl. 2.1 durch das Fermipotential $q\psi_F$ im Volumen des Halbleiters abgebildet, wie in Abb. 2.1b gezeigt.
Legt man nun eine äußere Spannung an die Kondensator-Struktur an, kann man die Bandverbiegung $q\psi_S$ des Substrats im Fall der Verarmung verändern. Daraus resultiert, dass sich die durch Bandverbiegung bestimmte Substratkapazität $C_s^{'}$ ändert. Die gemessene Gesamtkapazität zwischen Gate- und Substratkontakt $C_g^{'}$ teilt sich dann auf in die konstante Oxidkapazität $C_{ox}^{'}$ und die in Reihe geschaltete Substratkapazität $C_s^{'}$. Daher ergibt sich $C_g^{'}$ zu:

$$\frac{1}{C_g^{'}} = \frac{1}{C_{ox}^{'}} + \frac{1}{C_s^{'}} \tag{2.2}$$

Die Kapazitäten sind auf die Einheitsfläche bezogen und werden entsprechend $C_x^{'} = \frac{C_x}{A}$ umgerechnet. Die Oxidkapazität lässt sich aus der Schichtdicke d_{ox} und der zugehörigen Dielektrizitätskonstante ϵ_{ox} berechnen:

$$\frac{1}{C_{ox}^{'}} = \frac{\epsilon_{ox}}{d_{ox}} \tag{2.3}$$

Die Bestimmung der Substratkapazität ist erheblich komplexer und hängt von einer Vielzahl unterschiedlicher Parameter ab. Am Einfachsten lässt sich das Verhalten der Substratkapazität anhand der im folgenden Abschnitt vorgestellten Kapazitäts-Spannungs-Kennlinie erläutern.

2.1.2 Kapazitäts-Spannungs-Kennlinie

Bei der Kapazitäts-Spannungsmessung (kurz C(V)-Messung) wird die Kapazität in Abhängigkeit von der Spannung gemessen, welche über die gesamte Kondensatorstruktur abfällt. Die C(V)-Messung unterscheidet man in Hochfrequenz- (HF) und Niederfrequenz- (NF) Messung. Eine einfache und sichere Messmethode für HF-C(V)-Messungen ist die Bestimmung der Kapazität mittels einer Kleinsignal-Wechselspannung. Die Auswertung erfolgt anhand einer Messung des komplexen Widerstandes. Die Kapazität lässt sich dann anhand eines einzustellenden Modells ermitteln, welches das Ersatzschaltbild der untersuchten Struktur berücksichtigt. NF-C(V)-Kurven lassen sich sowohl mit der bereits genannten Methode messen, als auch mit der sogenannten 'quasi-statischen Methode'. Dabei wird der Ladestrom während eines kleinen Spannungssprungs ermittelt und dann inkrementell die Kapazität errechnet. Wird von außen an die Gateelektrode eine Spannung V_G angelegt, teilt sich diese entsprechend

$$V_G = \psi_S + V_{ox} + V_{FB} \tag{2.4}$$

auf. V_{FB} ist die sogenannte Flachbandspannung und gibt die Spannung wieder, die angelegt werden muss, damit das Oxid feldfrei ist und keine Bandverbiegung $\psi_S = 0$ V auftritt. Bestimmt wird V_{FB} durch die Austrittsarbeitsdifferenz ϕ_{MS} und Ladung, welche in der dielektrischen Schicht gespeichert ist.
Wird im Fall eines p-dotierten Substrats eine negative Spannung an die Gateelektrode angelegt, kommt es zur Anreicherung von Löchern an der Substratoberfläche. In diesem Fall spricht man von Akkumulation und ψ_S ist kleiner 0. Wählt man nun eine sehr hohe negative Spannung V_G nähert sich die gemessene Kapazität der Oxidkapazität C'_{ox} an. Verdeutlicht wird dies anhand der in Abb. 2.2 gezeigten C(V)-Kurven.

Erhöht man nun V_G, wird die Löcherdichte an der Oberfläche immer weiter verringert. Ist die wirksame Löcherdichte kleiner als die Löcherdichte p_0 bei thermischem Gleichgewicht, spricht man von Verarmung. Es bildet sich nun eine Raumladungszone mit der Weite w_v aus. Die Kapazität C'_{RLZ} dieser Raumladungszone entspricht C'_S und lässt sich durch Gl. 2.5 [12] berechnen. Die relative Dielektrizitätskonstante des Halbleiters, im betrachteten Fall Silizium, ist ϵ_{Si}.

$$C'_S \cong C'_{RLZ} = \sqrt{\frac{\epsilon_{Si}\, q\, p_0}{2\psi_S}} = \frac{\epsilon_{Si}}{w_v} \tag{2.5}$$

Mit zunehmender angelegter Gleichspannung wird die Weite der Raumladungszone w_v immer größer und die Kapazität C'_{RLZ} demzufolge immer kleiner. Dadurch wird auch die messbare Gesamtkapazität C'_G kleiner, wie es in Abb. 2.2 veranschaulicht ist. Der weitere Verlauf der Messkurve wird durch die Minoritätsladungsträger, im Fall von p-dotiertem Substrat den Elektronen, bestimmt. Wird ein Mess-Kleinsignal ausreichend kleiner Frequenz angelegt (< 20 Hz), können die generierten Minoritätsladungsträger dem Signal folgen. Ist die Spannung über der Halbleiterkapazität groß

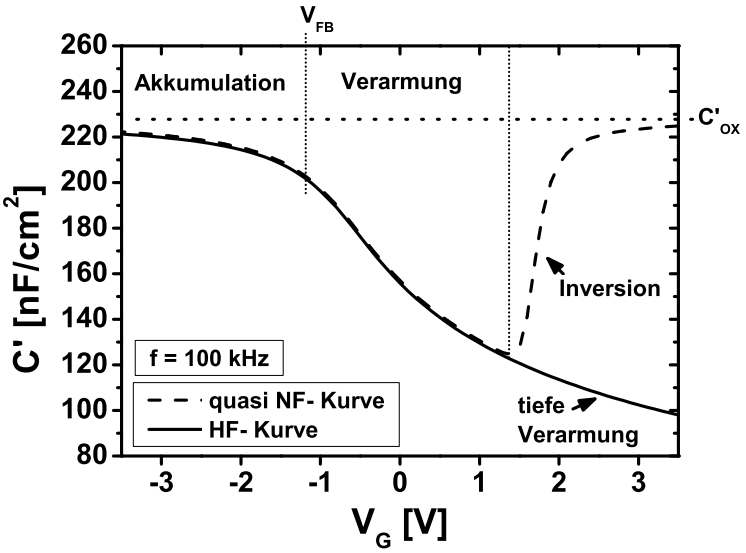

Abbildung 2.2: Gemessene HF- und quasi NF-C(V)-Kurve einer TANOS-Kapazität mit p-Substrat; ONA: 4.5/6/12 nm; A = $2.24e^{-4}$ cm^2; die quasi NF-C(V)-Kurve wurde mit n+-Kontakt als Inversionsladungsträger-Quelle gemessen

genug, werden mehr Minoritäten generiert, als Majoritätsladungsträger vorhanden sind, und man spricht von Inversion [8]. Diese tritt ein, wenn $\psi_S > \psi_F$ ist, sprich die Zahl freier Majoritäten ist kleiner als die Anzahl freier Minoritätsladungsträger. Anschaulich dargelegt sind beide Größen in Abb. 2.1b. Der Aufbau der Inversionsladung erfolgt an der Oberfläche und die Substratkapazität verschwindet wieder. Denn die Inversionsladung an der Oberfläche folgt dem an der Gateelektrode eingespeisten Kleinsignal und blendet daher die durch Verarmung gebildete Substratkapazität wieder aus. Dadurch steigt die gemessene Kapazität $C_g^{'}$ der NF-C(V)-Messung mit steigender Spannung an und sättigt nahe $C_{ox}^{'}$, wie in Abb. 2.2 durch die gestrichelte Linie gezeigt wird. Es ist nun wieder nur die Oxidkapazität messbar. Die Berechnung der NF-Halbleiterkapazität erfolgt mit (ohne Herleitung):

$$C_{S,NF}^{'}(\psi_s) \equiv \sqrt{\frac{q\,\epsilon_s \beta}{2}} \cdot \frac{p_o\,(1 - e^{-\beta\psi_s}) + n_o\,(e^{\beta\psi_s} - 1)}{F(\beta\psi_s)} \tag{2.6}$$

Die Funktion $F(\beta\psi)$ wird durch

$$F(\beta\psi) = \left[p_0\left(e^{-\beta\psi} + \beta\psi - 1\right) + n_0\left(e^{\beta\psi} - \beta\psi - 1\right)\right]^{\frac{1}{2}} \tag{2.7}$$

beschrieben und gibt ein Teil der gelösten Poisson-Gleichung im Silizium mit den Gleichgewichtsladungsträgerdichten n_0 und p_0 wieder.
Wird nun die Kleinsignal-Messfrequenz erhöht, kommt es zu dem Effekt, dass die Ladungsträger in der Inversionsschicht nicht mehr dem Kleinsignal folgen können. Es können innerhalb einer Periode des angelegten Wechselstroms nur eine begrenzte Zahl an Minoritätsladungsträgern generiert beziehungsweise durch Rekombination abgebaut werden. Bei einer sehr hohen Frequenz (HF) kann die Inversionsladung nicht mehr dem Kleinsignal folgen und die erforderliche Ladung wird durch Vergrößerung des Verarmungsgebietes im Substrat aufgebaut. In diesem Fall spricht man von Verarmung. Hierbei kommt es zu einem Gleichgewichtszustand, der durch die Messfrequenz

und Generations-Rekombinationsrate des Halbleiters bestimmt ist. Daraus ergibt sich die Menge der dem Wechselsignal folgenden Minoritäten im Inversionsbereich. Der resultierende Kurvenverlauf befindet sich zwischen der gezeigten Kurve für Inversion und tiefer Verarmung in Abb. 2.2. Wird zu der hohen Kleinsignalfrequenz auch die Gatespannung V_G schnell erhöht, kann sich im Halbleiter kein Gleichgewichtszustand mehr einstellen. Dadurch kommt es zu einer weiteren Verarmung über den eigentlichen Gleichgewichtszustand (Minimum der NF-C(V)-Kurve) hinaus. In diesem Fall kommt es zur tiefen Verarmung, und es wird eine immer kleiner werdende Kapazität gemessen. Dies kann auch bei der gemessenen Probe (durchgezogene Linie) aufgrund des hochwertigen Substrats beobachtet werden.
Bei der Messung der NF-C(V)-Kurve in Abb. 2.2 wurde eine Möglichkeit genutzt, die es erlaubt, auch mit einer HF-Messung den Bereich der Inversion richtig zu messen. Hierzu wird ein Kontakt am Rand der Kapazität hinzugefügt, welcher als Minoritätsladungsträger-Quelle fungiert. Im Fall von p-Substrat ist dies ein n-Kontakt, der während der Messung mit dem Substrat kurzgeschlossen ist. Kommt man mit der Messung in den Bereich der Inversion, stellt der Kontakt eine Elektronenquelle dar. Die Elektronen stehen innerhalb kurzer Zeit in nahezu unbegrenzter Anzahl zur Verfügung und ermöglichen die Ausbildung der Inversionsschicht ohne Beschränkungen durch die Eigenschaften des Halbleiters. Dieses Verhalten entspricht einer Messung bei kleiner Messfrequenz, bei der es keinen Einfluss auf die Messung durch die Rekombinations- und Generationsrate gibt.

2.2 Der MOS-Feldeffekt-Transistor

Bei dem MOS-Feldeffekt-Transistor (MOSFET) handelt es sich prinzipiell um die gleiche Struktur, wie ein MOS-Kondensator, wobei der steuerbare Widerstand der Inversionsschicht genutzt wird. Als Anschluss dieses Widerstandes werden im Fall von p-Substrat zwei n-Gebiete integriert (Source / Drain), zwischen denen durch die Inversionsschicht ein Strom fließen kann. Diese hochdotierten Gebiete werden noch für einen zuverlässigen Anschluss mit Metallkontakten versehen. In Abb. 2.3 ist eine solche Struktur dargestellt und auch der leitende Inversionskanal eingezeichnet.
Das Bauelement ist bestimmt durch dessen physikalische Größen Länge (L) und Weite (W), sowie den angelegten Spannungen V_G (Gatespannung), V_S (Sourcespannung), V_D (Drainspannung) und der Spannung am Substratanschluss (V_B). Bei den folgenden Betrachtungen wird V_S mit 0 V festgelegt und für alle anderen Spannungen als Referenzpotential verwendet. Die wichtigste Größe zur Beschreibung eines MOSFET ist dessen Schwellspannung V_T (T = engl. threshold, Schwellwert), die durch

$$V_T = 2\psi_F + V_{FB} + \frac{Q'_{ges}}{C'_{ox}} = 2\psi_F + V_{FB} + \gamma\sqrt{2\psi_F - V_B} \tag{2.8}$$

beschrieben wird [12]. Die Schwellspannung bezeichnet die Spannung, ab der starke Inversion vorliegt, beschrieben durch die Größe $2\psi_F$. Ab diesem Wert ist ein vollständig leitfähiger Kanal ausgebildet. Zudem berücksichtigt $\frac{Q'_{ges}}{C'_{ox}}$ den Einfluss von Ladung im Oxid und an der Substrat-Oxid-Grenzfläche auf die Erzeugung des leitfähigen Kanals. Eine einfache Abschätzung für das Stromkriterium, bei dem der Strom fließt, anhand dessen sich die Schwellspannung ablesen lässt, zeigt Gl. 2.9. Darin

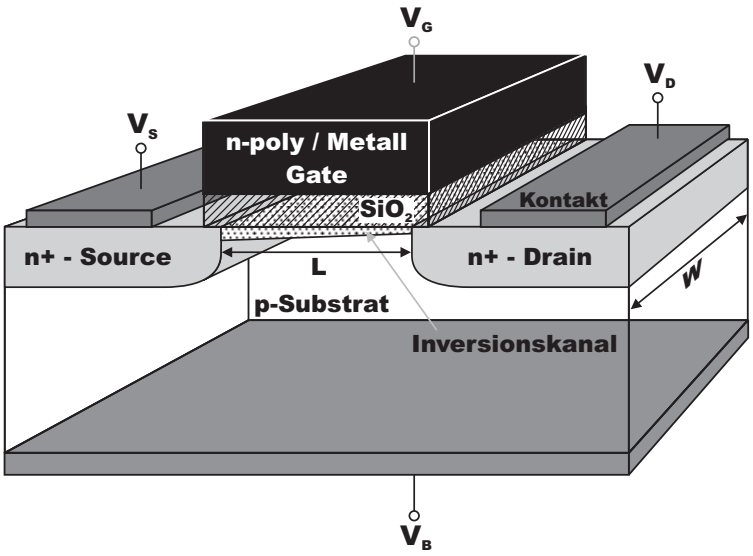

Abbildung 2.3: Aufbau eines MOS-Feldeffekttransistors mit dessen charakteristischen Größen Länge (L) und Weite (W); der leitende Inversionskanal ist eingezeichnet, der die beiden n^+-dotierten Kontaktgebiete von Source und Drain verbindet

werden auch die Transistor-Abmessungen berücksichtigt.

$$I_D\left(V_T\right) = 100\ nA \cdot \frac{W}{L} \tag{2.9}$$

Weiterhin wird der Begriff Einsatzspannung verwendet. Dieser Wert beschreibt die Spannung V_G, ab der sich ein signifikanter Stromfluss einstellt und durch $\psi_S > \psi_F$ definiert ist [13].
Bei der normalen Anwendung von MOS-Transistoren erfolgt ein Kurzschluss des Source- und Substrat-Anschlusses. Ist dies nicht der Fall, muss der Einfluss der Spannung V_B, wie in Gl. 2.8 gezeigt, berücksichtigt werden. Der Parameter γ wird als Substratsteuerfaktor bezeichnet und ist eine Größe, die durch die Oxiddicke d_{ox}und die Dotierstoffkonzentration N_A des Substrates bestimmt ist.

2.2.1 Transistor-Kennlinien

Damit ein Stromfluss zwischen den beiden n-Gebieten entstehen kann, muss eine Spannung $V_D > 0$ V angelegt werden. In diesem Fall liegt ein Nichtgleichgewicht vor und es fließt ein Strom. In Abb. 2.4 wird die Abhängigkeit des Drain-Stromes von der angelegten Gatespannung gezeigt.
Die Kennlinie in Abb. 2.4a zeigt, dass bei niedrigen Gatespannungen immer noch ein sehr kleiner Leckstrom I_L fließt. Bei der Betrachtung von Einzeltransistoren muss dieser Strom nicht beachtet werden. Aber bei der Verwendung in großen integrierten Schaltungen und bei großen Speicherzellenfeldern kann dieser nicht mehr vernachlässigt werden (siehe Kap. 4.4) [14, 15]. Daran schließt sich der Unterschwellspannungsbereich an. Dies ist der Bereich, in dem bereits Inversion vorliegt, aber die Bandverbiegung ψ_S kleiner als das Kriterium für die starke Inversion von $2\psi_F$ ist. Dieser Bereich gibt auch Aufschluss über die Ladungsverhältnisse in einer Speicherzelle

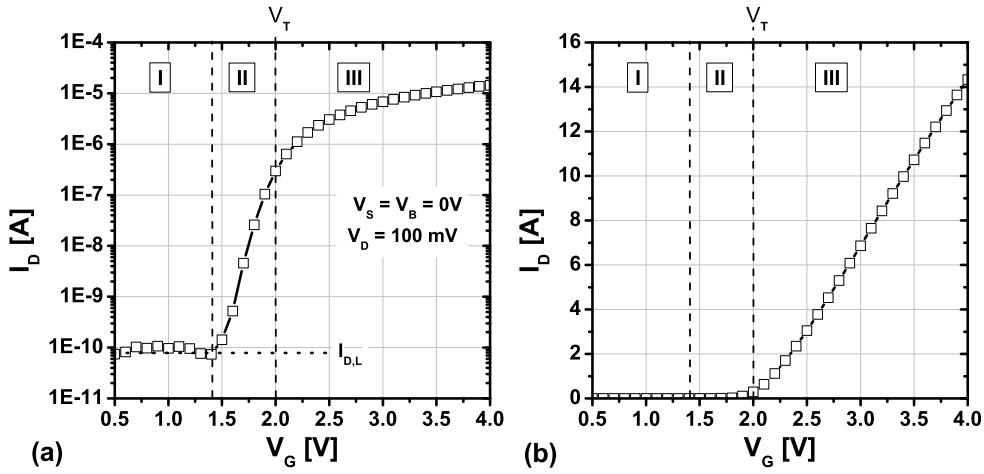

Abbildung 2.4: Gemessene Transistor-Übertragungskennlinie einer 5x5 μm SONOS-Speicherzelle, (a) mit logarithmischer und (b) in linearer Auftragung von $I_{D,L}$; eingezeichnet sind die Schwellspannung V_T und der Transistorleckstrom $I_{D,L}$; Bereich I Sperrbereich, II Bereich der Unterschwellspannung und III der eingeschaltene Zustand

und wird daher in Kap. 4.1 noch einmal genauer betrachtet. In der linearen Darstellung von Abb. 2.4b konnte in diesem Bereich noch keine Veränderung beobachtet werden. Aber oberhalb der Schwellspannung V_T befindet sich der Transistor im eingeschalteten Zustand und es kann unter der Bedingung einer kleinen Drainspannung V_D ein nahezu linearer Verlauf der Kennlinie beobachtet werden. Dieses Verhalten ist darauf zurückzuführen, dass sich der Transistor im sogenannten Ohm'schen Bereich arbeitet. In diesem Fall ist die Ladungsträgerdichte über die Länge des Transistors nahezu konstant und er verhält sich wie ein Ohm'scher Widerstand. Der Stromfluss von Source zu Drain errechnet sich gemäß:

$$I_{D,lin} = \beta \left(V_G - V_T\right) \cdot V_D. \tag{2.10}$$

Deutlich ist die lineare Abhängigkeit des Transistorstromes von V_G und V_D zu erkennen. Dieses lineare Verhalten wird auch in speziellen Verstärkern genutzt. Bei der Größe β handelt es sich um den Übertragungsleitwert (engl. transconductance), einer bauteil-spezifischen Größe, welche gemäß

$$\beta = \frac{\epsilon_{ox}\mu}{d_{ox}} \cdot \frac{W}{L} = \mu C_{ox} \frac{W}{L} \tag{2.11}$$

beschrieben wird. μ ist die Ladungsträgerbeweglichkeit, in dem gezeigten Fall für Elektronen. Gl. 2.11 zeigt zudem eine wichtige Beziehung bei der Dimensionierung und dem Vergleich verschieden großer Transistoren. Es wird deutlich, dass der Strom durch den Transistor durch die Beziehung aus Länge und Weite $\frac{W}{L}$ bestimmt wird. Die Abhängigkeit von der Drainspannung unterteilt man in die drei Bereiche

$$\begin{array}{lll} V_D \ll V_{D,sat} & \text{linearer Bereich ,} \\ V_D \lesssim V_{D,sat} & \text{nichtlinearer Bereich ,} \\ V_D > V_{D,sat} & \text{Sättigungsbereich ,} \end{array} \tag{2.12}$$

wobei

$$V_{D,sat} \cong V_G - V_T \tag{2.13}$$

die Spannung der einsetzenden Stromsättigung repräsentiert. In Abb. 2.5a sind die drei Bereiche anhand des Ausgangskennlinienfeldes schematisch dargestellt. An den Ohm'schen Bereich schließt sich der nichtlineare Bereich an (II). Dieser ist dadurch gekennzeichnet, dass sich der bestehende Inversionskanal aufgrund der Drainspannung immer mehr zu einem Dreieck verformt.

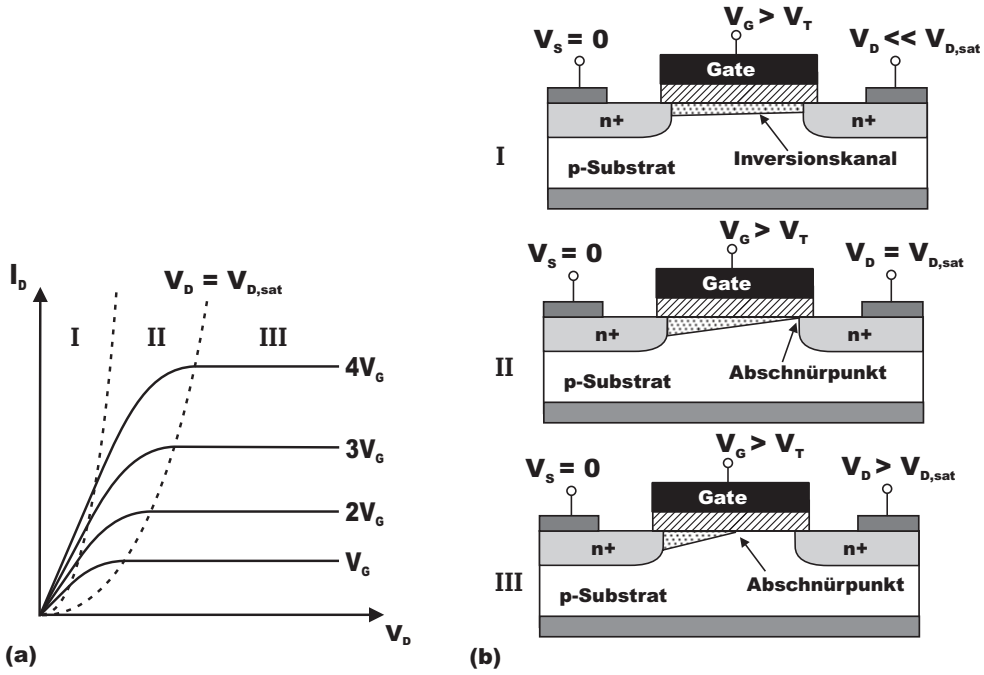

Abbildung 2.5: (a) Ausgangskennlinienfeld eines MOS-Transistors mit dem linearen Bereich (I), dem nichtlinearen Bereich (II) und dem Sättigungsbereich (III); in (b) ist jeweils die Form der Inversionsschicht für den entsprechenden Bereich illustriert

Die angelegte Drainspannung führt dazu, dass das Oberflächenpotential ortsabhängig ist und sich dadurch die Dicke der Inversionsschicht in Richtung Drain verringert. Das Oberflächenpotential ergibt sich zu [13]:

$$\phi_{Ob} = V_G - V_T - \phi(x)\,. \tag{2.14}$$

Am drain-seitigen Ende ist $\phi(x) = V_D$. Die Abhängigkeit des Stromes von der Drainspannung bekommt nun noch eine nichtlineare quadratische Abhängigkeit, wie Gl. 2.15 verdeutlicht.

$$I_{D,nl} = \beta\left((V_G - V_T)\cdot V_D - \frac{V_D^2}{2}\right) \tag{2.15}$$

Wird nun die Drainspannung so gewählt, dass diese genau der Differenz $V_G - V_T = V_{D,sat}$ entspricht, wird das Kanalpotential ϕ_{Ob} gleich 0. Somit liegt an der Drain keine Inversion mehr vor und der Kanal ist abgeschnürt, wie in Abb. 2.5b gezeigt. Wird nun die Spannung V_D weiter erhöht, verschiebt sich der Punkt der Abschnürung (engl. pinch-off) immer weiter in Richtung Source, wie in Abb. 2.5b III dargestellt. In dem Stück zwischen dem Abschnürpunkt und der Drain bildet sich ein schmaler oberflächennaher Kanal mit hoher Ladungsträgerdichte aus. Dieser Kanal lässt genau den Strom passieren, welcher sich bis zum Abschnürpunkt ergeben hat. Da eine

weitere Erhöhung von V_D nun keine Änderung des Transistorstromes mehr bewirkt, befindet sich dieser im Bereich der Sättigung von Abb. 2.5a III. Diese idealisierte Betrachtung führt zu der Gleichung

$$I_{D,sat} = \frac{\beta}{2} \left(V_G - V_T\right)^2 \tag{2.16}$$

für den Stromfluss in der Sättigung. Reale Transistoren zeigen aber nicht das Verhalten einer idealen Stromquelle, da sich der effektive Kanal aufgrund der Verschiebung des Abschnürpunktes verkürzt. Dadurch nimmt der Strom durch den Transistor bei Erhöhung der Drainspannung weiter leicht zu.

2.2.2 Kurzkanal-Effekt

Betrachtet man allerdings skalierte Zellen, so hat die Länge des Kanals einen Einfluss auf das Zellverhalten. Abbildung 2.6 zeigt die Veränderung der physikalischen Kanallänge L durch die Source/Drain-Verarmungsgebiete zu L'.

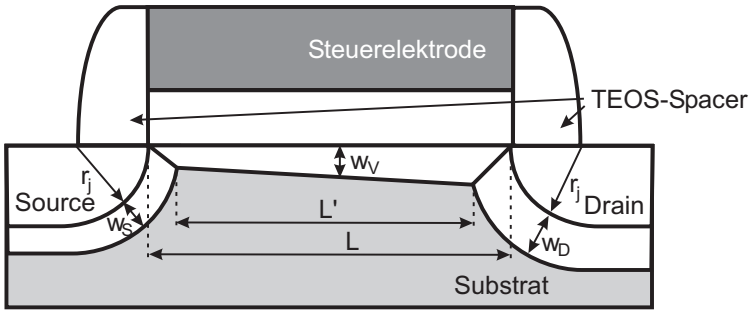

Abbildung 2.6: aus [12], Abmessungen der relevanten Größen, die zur Betrachtung des Kurzkanaleffektes nötig sind, $V_D > 0$

Die Verkürzung der effektiven Kanallänge hat einen Einfluss auf die Verarmungsladung $Q_V^{'}$ im Kanal, da diese zum Teil durch die Verarmungszonen der Source/Drain-Gebiete erzeugt wird. Für einen Langkanaltransistor ergibt sich:

$$Q_V^{'} = qAN_AW_V, \tag{2.17}$$

wohingegen sich die Fläche A bei einem Kurzkanaltransistor verkleinert. Durch eine dreieckförmige Näherung der Überlappgebiete zu Source/Drain ergibt sich [12]:

$$Q_V^{'} = qN_Aw_V \left(\frac{L + L^{'}}{2}\right) \cdot W. \tag{2.18}$$

Eine wichtige Vereinfachung ist die Betrachtung bei kleiner Drain-Spannung, wodurch $w_S = w_D = w_V$ genähert werden kann. Die erst einmal unbekannte Größe $L^{'}$ lässt sich über eine einfache Dreiecksnäherung berechnen. Hierbei wird die Summe aus Source/Drain-Diffusionsweite r_j und Dicke des Verarmungsgebietes w_V mit w_V ins Verhältnis gesetzt um aus der physikalischen Kanallänge L die scheinbare Länge $L^{'}$ zu berechnen, wodurch sich:

$$L^{'} = L - 2\left(\sqrt{r_j^2 + 2w_Vr_j} - r_j\right) \tag{2.19}$$

ergibt. Setzt man dies nun in die Gleichung 2.8 zur Berechnung der Schwellspannung ein, lässt sich die Änderung derer in Abhängigkeit der Kanallänge mit:

$$\Delta V_T = -\frac{qN_A w_V r_j}{C_{ox} L}\left(\sqrt{1+\frac{2w_V}{r_j}}-1\right) \tag{2.20}$$

beschreiben. Die komplette Herleitung kann in Sze [12] und Yau [16] nachgelesen werden. Das Ergebnis ist eine Verringerung der Schwellspannung, die umso größer wird, desto kleiner die physikalische Länge L wird. Zudem wird gezeigt, dass sich die Schwellspannung V_T ändert, wenn bei einer konstanten physikalischen Kanallänge L die Diffusionsweite r_j vergrößert wird.

2.2.3 Gate-induzierter-Drain-Leckstrom (GIDL)

Die durch die Skalierung bedingte Zunahme der Felder führt wiederum zu Effekten im Transistor, die nicht beabsichtigt sind. Zum Beispiel tritt bei Transistoren mit dünnen Oxiden und relativ hoch dotierten Source-/Drain-Gebieten ein Leckstrom von der Drain in das Substrat auf. Dieser Effekt beruht darauf, dass zu dem intrinsischen Feld des p-n-Übergangs auch noch das durch die Drainspannung erzeugte Feld und das durch die Gateelektrode induzierte Feld hinzukommen. Das Ergebnis ist ein gatespannungs-abhängiger Leckstrom wie in Abb. 2.7a gezeigt wird.

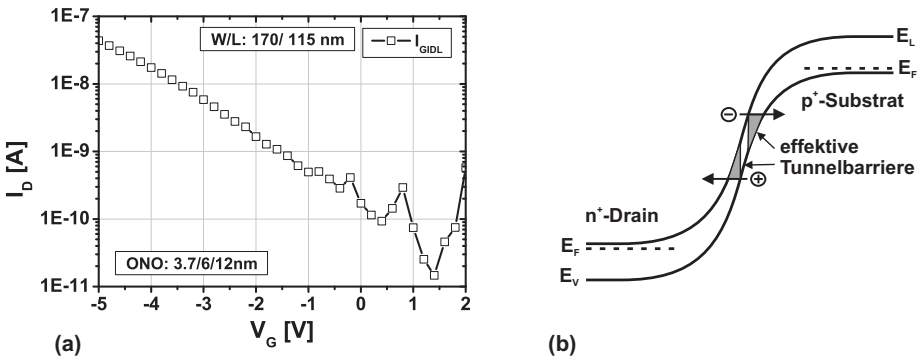

Abbildung 2.7: (a) GIDL-Strom bei negativer Gatespannung, gemessen bei $V_D = 1.5V$; (b) Bandverlauf in einem p-n-Übergang bei hohem elektrischen Feld und der Band zu Band Tunnelprozess mit den zugehörigen Tunnelbarrieren

Durch die Überlagerung der Felder, verursacht durch Drain- und Gatespannung, kommt es in der Sperrrichtung der p-n-Diode zu einem signifikanten Stromfluss. Bei hohen elektrischen Feldern im Silizium von $> 10^6\ V/cm$ kann der Stromfluss einmal aufgrund von Tunneln oder aber auch durch Lawinen-Multiplikation verursacht sein. Beim GIDL handelt es sich um einen Band zu Band Tunnelprozess, der in Abb. 2.7b illustriert wird. Durch die starke Verbiegung der Bänder entsteht eine dreiecksförmige Barriere, wodurch ein Tunneln, entsprechend dem Fowler-Nordheim-Tunneln (Kap. 2.3.2.1), wirksam wird. Dieser Mechanismus, der normalerweise unerwünscht ist, bildet die Grundlage des NROM-Speicherkonzepts [17] und wird auch in Zener-Dioden ausgenutzt [18]. Weiterhin ist er Ursache für den Störmechanismus der äußeren Speichertransistoren, vorgestellt in Kap. 5.2.2 . Wertet man die Spannung bei einem festen Stromkriterium wie z.B. I = $1e^{-10}$ A aus, spricht man von der Band zu Band Tunnelspannung V_{BTB}.

2.3 Nichtflüchtige Festkörperspeicher

Das Gebiet der Halbleiterspeicher gliedert sich in die wesentlichen drei Gruppen DRAM (engl. dynamic random access memory), SRAM (engl. static random access memory) und die nichtflüchtigen Speicher auf. In dem Gebiet der nichtflüchtigen Speicher gibt es wiederum eine Vielzahl an verschiedenen Typen, wie Abb. 2.8 verdeutlicht.

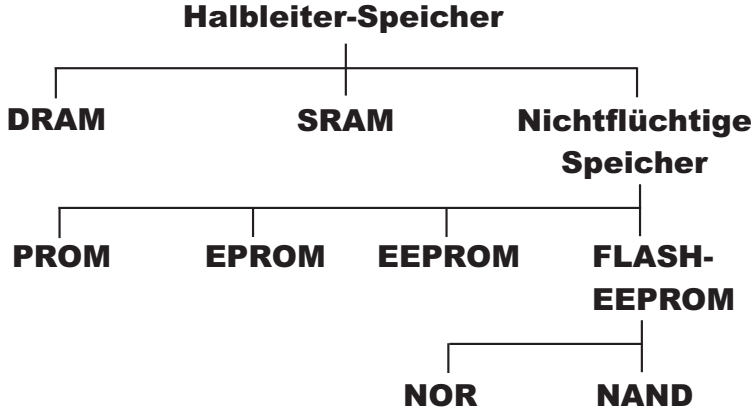

Abbildung 2.8: Übersicht über die Einteilung von Halbleiter-Speichern mit einer Einordnung der untersuchten NAND-Flash-Speicher

Der einfachste Typ ist das PROM, welches schon ab Fertigung programmiert ist. Als Steigerung ist das EPROM vom Nutzer programmierbar, aber kann entweder gar nicht mehr (OTP, one time programmable), oder nur mit UV-Licht gelöscht werden. Das EEPROM wiederum kann auch elektrisch gelöscht werden. Um die Schreib-/Lese-Geschwindigkeit zu erhöhen, wurde als weitere Verfeinerung das Flash-EEPROM eingeführt. Beim einfachen EEPROM ist meist ein Auswahltransistor der eigentlichen Speicherzelle vorgeschaltet, um eine versehentliche Veränderung des Speicherinhalts zu vermeiden. Dies erhöht aber den Platzbedarf, der bei Flash-Speichern ein entscheidendes Maß darstellt und somit weggelassen wird. Andererseits werden beim Flash-EEPROM nicht nur einzelne Byte elektrisch gelöscht, sondern ganze Blöcke aus Speicherzellen. Es ergibt sich durch die große Parallelisierung ein erheblicher Zeitgewinn, der wiederum die Speichergeschwindigkeit erhöht. Eine weitere Unterteilung ist anhand der Architektur des Zellenfeldes möglich. Als die wichtigsten Architekturen haben sich die Anordnung in NOR and NAND herauskristallisiert, die in Kap. 2.4 noch einmal genauer erläutert werden. Diese Arbeit befasst sich mit der Untersuchung der NAND-Architektur.

2.3.1 Speicherprinzip

Nichtflüchtige Halbleiter nutzen aus, dass Ladung in einem Dielektrikum zu einer Verschiebung der Schwellspannung führt, wie in Gl. 2.8 bereits gezeigt wurde.

$$V_T = 2\psi_F + V_{FB} + \frac{Q'_{ges}}{C'_{ox}}$$

Injiziert man gezielt Ladung in eine Speicherschicht, welche sich im Oxid des Transistors oder aber auch des Kondensators befindet, ist es möglich verschiedene Zustände zu erzeugen, welche im nachhinein beim dem Lesevorgang getrennt werden können.

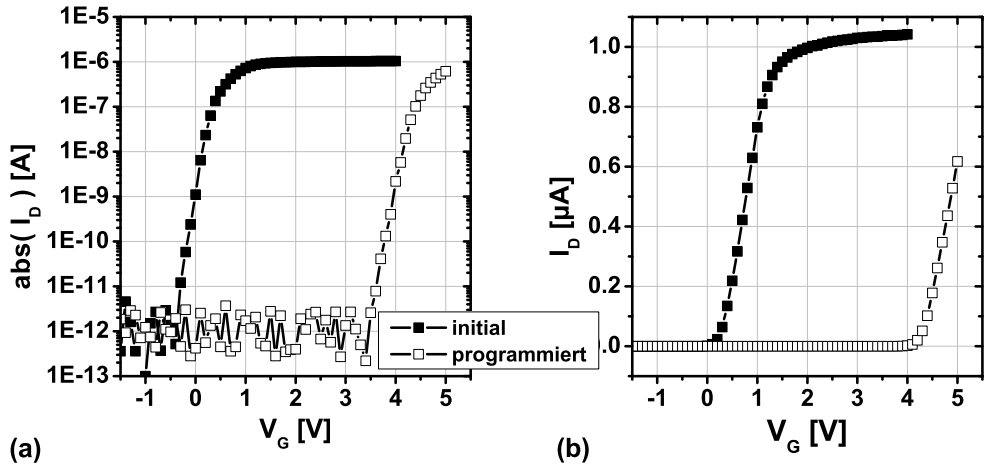

Abbildung 2.9: Kennlinien einer NAND-Speicherzelle, einmal logarithmisch (a) und in linearer Darstellung (b); $V_D = 0.7$ V; Speicherzellgröße: 48x48 nm

In Abb. 2.9 ist gezeigt, wie sich die Kennlinie eines Transistors aus einem NAND-String nach dem Programmieren verschoben hat. Es kommt dabei zu einer parallel-Verschiebung, weil die injizierte Ladung im Normalfall keine Auswirkung auf die weiteren Transistorparameter hat. Bei der Anwendung in Speichern ist die Speicherdichte ein zentraler Faktor. Normalerweise hat man nur einen gelöschten und einen programmierten Zustand und man spricht von einer single-level Zelle (SLC). Programmiert man mehrere Zustände in eine Speicherzelle, wie in Abb. 2.10 gezeigt, lassen sich mehrere Bit pro Zelle speichern. Man spricht in diesem Fall von einer multi-level Speicherzelle (MLC).

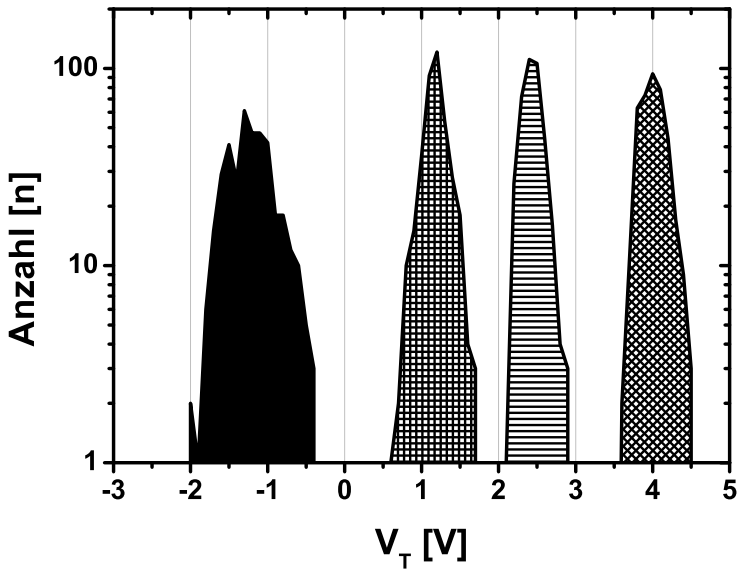

Abbildung 2.10: Verteilung der Schwellspannung für ein NAND-Speicherfeld, welches zuvor vom gelöschten Zustand (schwarz) auf drei verschiedene Speicherzustände programmiert wurde (schraffiert)

Dies erhöht zusätzlich die Speicherdichte, daher versucht man heutzutage immer mehr Zustände zu programmieren. Es wurden bereits Produkte angekündigt, die bis zu 15 unterschiedliche Programmierniveaus (4 bit/Zelle) aufweisen [19]. Allerdings leidet

die Zuverlässigkeit der gespeicherten Informationen, da die Abstände zwischen den programmierten Verteilungen kleiner werden und es daher schneller zu Lesefehlern kommen kann.
Im folgenden Abschnitt werden zunächst die Mechanismen erläutert, die zum Betrieb einer Speicherzelle notwendig sind. Daran schließt sich die Vorstellung der verschiedenen Konzepte von nichtflüchtigen Speichern an.

2.3.2 Schreib- und Löschmechanismen

Zum Betrieb der Speicherzelle sind Mechanismen notwendig, die dazu führen, dass Ladung durch das Tunneloxid in die Speicherschicht gelangt. Man unterteilt die Mechanismen anhand der Charakteristika in 'heiße Ladungsträger' und 'quantenmechanisches Tunneln'.

2.3.2.1 Quantenmechanisches Tunneln

Beim Tunneln handelt es sich um einen quantenmechanischen Prozess, bei dem ein Ladungsträger eine Potentialbarriere durchdringt. Bei Halbleiterspeichern bildet das SiO_2 die Potentialbarriere mit einer Höhe von 3.1 eV für Elektronen. Ist die Anzahl an Ladungsträgern groß genug und wird durch ein angelegtes elektrisches Feld eine Vorzugsrichtung vorgegeben, gibt es eine gewisse Wahrscheinlichkeit die Barriere zu durchdringen. Abb. 2.11 zeigt verschiedene Tunnelvorgänge, die abhängig von der Form der Potentialbarriere sind.

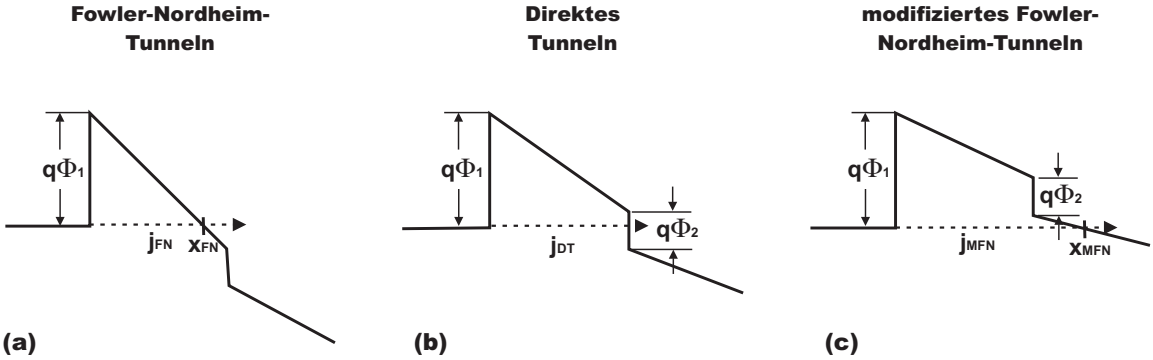

Abbildung 2.11: Darstellung der Tunnel-Barrierenform anhand des Leitungsbandes für (a) Fowler-Nordheim-Tunneln, (b) direktes Tunneln und (c) modifiziertes FN-Tunneln

Handelt es sich um eine dreieckförmige Barriere, wie in Abb. 2.11a gezeigt, so spricht man von Fowler-Nordheim Tunneln (FN-Tunneln). Der Tunnelstrom bei FN-Tunneln hängt aufgrund des im Oxid bei x_{FN} liegenden Injektionspunktes nicht von der Oxiddicke, sondern nur von der Feldstärke ab. Die Berechnung der Tunnelstromdichte lässt sich nach [20, 21] anhand von

$$J_{FN}\left(E_{ox}\right) = A_T E_{ox}^2 exp\left(-\frac{B_T}{E_{ox}}\right) \tag{2.21}$$

durchführen. Die Konstanten A_T und B_T sind beschrieben durch:

$$A_T = \frac{q^3\, m_e}{16\,\pi\,\hbar\, m'\,\phi_1},\ B_T = \frac{4}{3}\frac{\left(2\,m'\right)^{1/2}}{q\,\hbar}\phi_1^{3/2} \tag{2.22}$$

und sind im wesentlichen durch die effektive Masse $m^{'}$ und die Barrierenhöhe ϕ_1 bestimmt. Verringert man die elektrische Feldstärke und ist das Oxid ausreichend dünn ($d_{ox} < 4nm$), befindet sich der Injektionspunkt nicht mehr im Leitungsband. In diesem Fall spricht man von direktem Tunneln und der Ladungsträger tunnelt durch eine trapezförmige Barriere, wie in Abb. 2.11b illustriert. Die Feldbedingung ist gegeben durch:

$$\frac{q\left(\phi_1 - \phi_2\right)}{d_{ox}} < E_{ox} < \frac{q\phi_1}{d_{ox}}. \tag{2.23}$$

Die Neigung des Bandes hat nun keinen so großen Einfluss auf die Barriere. Die Neigung repräsentiert die Feldstärke, was bedeutet, dass die Abhängigkeit von der Feldstärke stark reduziert ist. Berechnen lässt sich die Tunnelstromdichte in vereinfachter Form für direktes Tunneln mittels [22, 23]

$$J_{DT}\left(E_{ox}\right) = \frac{A_T E_{ox}^2}{\left[1 - C_T^3\left(E_{ox}\right)\right]^2} exp\left[\frac{B_T}{E_{ox}}\left(1 - C_T^3\left(E_{ox}\right)\right)\right] \tag{2.24}$$

berechnen. Die Konstanten A_T und B_T sind entsprechend denen der FN-Berechnung (Gl. 2.22) und C_T ergibt sich zu

$$C_T = \sqrt{1 - \frac{qE_{ox}d_{ox}}{\phi_1}}. \tag{2.25}$$

Bei Verwendung eines Stapels mit verschiedenen Dielektrika, wie im Fall von haftstellen-basierten Speichern, gibt es einen weiteren Tunnelmechanismus. Nun tunneln die Ladungsträger komplett durch das erste Dielektrikum und auch einen Teil durch das zweite Dielektrikum. Da sich jetzt wieder eine dreieckförmige Barriere ausbildet, kommt es zu einer mit dem FN-Tunneln vergleichbaren Feldabhängigkeit und man spricht daher von modifiziertem Fowler-Nordheim-Tunneln (MFN-Tunneln). Ein sehr einfacher analytischer Ausdruck, der alle drei Tunnelmechanismen in einer Formel vereinigt ist von Beguwala [24] gegeben worden:

$$J(FN, DT, MFN) = \frac{q^3}{8\pi h}\frac{1}{\left[g\left(E_{ox}, E_{ni}\right)\right]^2} exp\left[-\frac{f\left(E_{ox}, E_{ni}\right)}{3\hbar q}\right]. \tag{2.26}$$

Der Temperatur-Kompensationsterm kann vernachlässigt werden und ist nicht mit aufgeführt. Die in den Unterfunktionen $f\left(E_{ox}, E_{ni}\right)$ und $g\left(E_{ox}, E_{ni}\right)$ enthaltenen Wurzelterme werden 0 gesetzt, wennn sie kleiner als 0 werden. Berechnet werden die Funktionen mit

$$\begin{aligned} f\left(E_{ox}, E_{ni}\right) = \; & 4\frac{\sqrt{2m_{ox}}}{E_{ox}}\left[\left(q\phi_1\right)^{3/2} - \left(q\phi_1 - qE_{ox}d_{ox}\right)^{3/2}\right] \\ & + 4\frac{\sqrt{2m_{ni}}}{E_{ni}}\left(q\phi_1 - q\phi_2 - qE_{ox}d_{ox}\right)^{3/2} \end{aligned} \tag{2.27}$$

und

$$\begin{aligned} g\left(E_{ox}, E_{ni}\right) = \; & \frac{1}{E_{ox}}\left[\left(q\phi_1\right)^{3/2} - \left(q\phi_1 - qE_{ox}d_{ox}\right)^{1/2}\right] \\ & + \frac{1}{E_{ni}}\left(q\phi_1 - q\phi_2 - qE_{ox}d_{ox}\right)^{1/2}. \end{aligned} \tag{2.28}$$

Die bestimmenden Parameter sind die Barrierenhöhe ϕ_1, der Bandabstand der Dielektrika ϕ_2 und das elektrische Feld E_{ox}. Diese Formel erlaubt es auf einfache Weise

den Tunnelstrom von haftstellenbasierten Speicherzellen zu berechnen und wird daher auch in den, im Kap. 3.2, gezeigten Simulationen angewandt. Weiterhin wird in Abb. 3.3b die Kennlinie gezeigt, die sich ergibt, wenn der Tunnelstrom in Abhängigkeit des elektrischen Feldes für die verschiedenen Tunnel-Modi aufgetragen wird.

2.3.2.2 Heiße Ladungsträger

Bei dem Tunnelvorgang durchdringen die Ladungsträger die Barriere aufgrund einer gewissen Tunnel-Wahrscheinlichkeit. Eine andere Möglichkeit eine Potentialbarriere zu überwinden, ist eine größere kinetische Energie aufzubauen als die Barrierenhöhe ϕ_1. Der Aufbau von kinetischer Energie von Elektronen ist möglich, wenn der Transistor im Sättigungsbereich betrieben wird (siehe Kap. 2.2.1). Ab dem Sättigungspunkt im Kanal bewegen sich die Elektronen mit Sättigungsgeschwindigkeit. Die durch das laterale Feld hinzugefügte Energie wird nun in kinetische Energie der Ladungsträger umgesetzt. Auf dem Weg zur Drain kommt es immer wieder zu Kollisionen mit anderen Elektronen, so dass einige Elektronen sehr hohe kinetische Energien erhalten. Daher auch der Ausdruck 'heiße Ladungsträger' (engl. hot-carrier). Die Energie einzelner Ladungsträger kann so groß werden, dass sie die Barriere des Tunneloxids von 3.1 eV überwinden können. Speziell für die Elektronen, die im Kanal erzeugt werden, gibt es die Bezeichnung 'channel hot electrons' (CHE), gleichbedeutend mit 'heiße Elektronen aus dem Kanal'. In Abb. 2.12a sind die Energie und das elektrische Feld im Kanal eines Transistors für zwei verschiedene Längen simuliert.

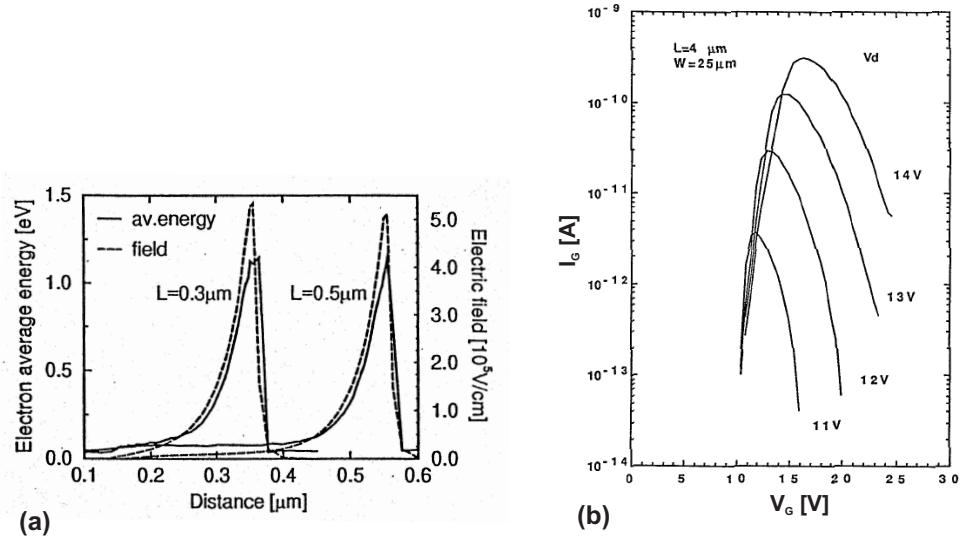

Abbildung 2.12: (a) Elektrisches Feld und die Energie der Elektronen im Kanal von Source (links) zur Drain (rechts) für $V_G = V_D/2 = 1.5V$ und $d_{ox} = 10nm$, aus [25]; (b) Strom zum Gate in Abhängigkeit der Drain- und Gatespannung, aus [26]

Deutlich ist zu erkennen, dass sich eine starke Konzentration der hochenergetischen Elektronen auf das Gebiet nah der Drain ergibt. Wird nun zusätzlich noch eine ausreichend große Spannung an das Gate angelegt, entsteht ein Stromfluss durch das Oxid. Abbildung 2.12b zeigt den Strom in Abhängigkeit von der Gate- und Drainspannung. Es wird beobachtet, dass der Strom zuerst ansteigt und nach einem Maximum, welches bei $V_G - V_T \approx V_D/2$ liegt, wieder stark abnimmt. Im Normalfall wird dieser Be-

triebsfall vermieden, weil er in einer Schädigung des Oxids resultiert und unbeabsichtigte V_T-Verschiebungen herbeiführt [27]. Im Vergleich zum Tunneln werden deutlich kleinere Gatespannungen benötigt, was bei kleinen Speichern ein deutlicher Vorteil ist, weil die Ladungspumpen kleiner ausfallen können und damit Chipfläche gespart wird. Ein Nachteil ist allerdings die geringe Ladungsträger-Injektionseffizienz von circa 1:1000. Dies bedeutet, dass circa 1000 Elektronen durch den Kanal fließen, wenn ein Elektron in die Speicherschicht gelangt. Durch den hohen nötigen Kanalstrom wird die Parallelität beim Programmieren eingeschränkt, da nur ein begrenzter Gesamtstrom für den Speicherchip in den meist mobilen Anwendungen zur Verfügung steht. Daher ist für diesen Injektionsmechanismus die Speicher-Schreibgeschwindigkeit limitiert. Beim Tunneln hingegen ist die Effizienz nahezu 1 und es kann mit großer Parallelität programmiert werden.

2.3.3 SONOS-Struktur

Bei der SONOS-Struktur wird das Oxid zwischen Si-Substrat und Si-Gate durch einen Schichtstapel aus Siliziumoxid-Siliziumnitrid-Siliziumoxid ersetzt. Die beiden Oxidschichten dienen der Isolation von der Speicherschicht, dem SiN. Es hat sich bereits früh gezeigt, dass sich SiN als Ladungsspeicherschicht eignet [28], obwohl es sich auch um eine Isolatorschicht handelt. Ursache ist die große Dichte an energetisch günstig liegenden Haftstellen, die in Kap. 3.1 genauer betrachtet werden. Ursprünglich bestand die Struktur nur aus zwei Schichten, einem Tunneloxid und der darüberliegenden SiN-Speicherschicht. Dieser Schichtstapel wurde dementsprechend MNOS-Struktur (Metall-Si$\underline{N}$-Si$\underline{O_2}$-Si) genannt [28,29] . Bei der weiteren Entwicklung wurde eine weitere Oxidschicht, das Topoxid, auf der Speicherschicht hinzugefügt, um die Ladungshaltung zu verbessern [30] . In Abb. 2.13a ist der Aufbau einer SONOS-Struktur für eine Transistor-Speicherzelle gezeigt.

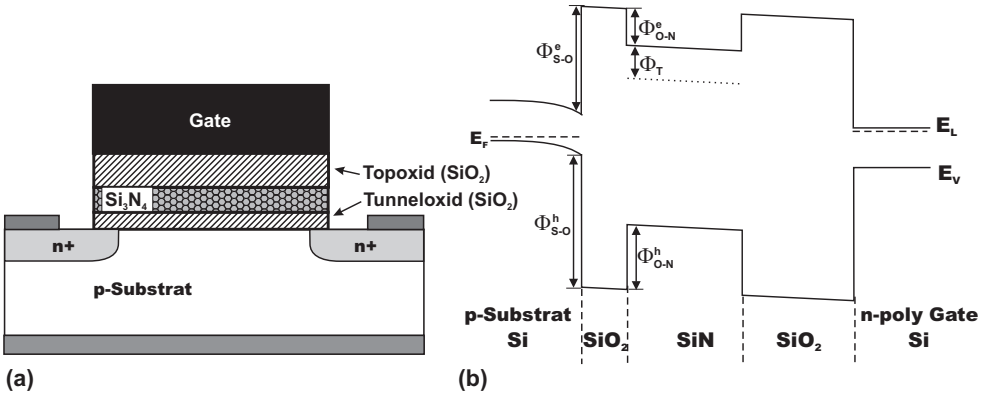

Abbildung 2.13: (a) Schichtstapel und (b) das entsprechende Bänderdiagramm einer SONOS-Struktur, ein mögliches Energieniveau ist eingezeichnet, auf dem sich gespeicherte Elektronen befinden können

Die Bandstruktur zeigt Abb. 2.13b, wobei auch das Energieniveau der Elektronenhaftstellen eingezeichnet ist. Dieses befindet sich circa $\phi_T = 0.8 - 1.6eV$ unterhalb der Leitungsbandkante [31, 32]. In einem bekannten Anwendungsfall [33] wird die SONOS-Struktur mit Schichtdicken von ≈1.8 nm für das untere Oxid, 9 nm für das Nitrid und 4 nm für das Topoxid ausgeführt. Diese Schichtwahl resultiert in kleinen

benötigten Spannungen und durch das dünne untere Oxid, auch Tunneloxid genannt, ist es möglich ausreichend zu löschen. Damit einher gehen aber ein begrenztes maximal erreichbare V_T-Verschiebung, sowie eine begrenzte Fähigkeit, Ladung dauerhaft zu speichern. Auf Grund dessen wird der SONOS-Schichtstapel zur Zeit nur in Verbindung mit einem Differenzverstärker betrieben [33]. Denn vergrößert man die Dicke des Tunneloxids um die Ladungshaltung zu verbessern, kommt es dazu, dass sich das Löschen verschlechtert. Dieser Effekt ist darauf zurückzuführen, dass die Barrierenhöhe ϕ^h_{S-O} von Löchern mit 3.8 eV deutlich größer ist als die Barriere ϕ^e_{S-O} für Elektronen mit 3.1 eV [34]. Bei einer SONOS-Struktur bestehen sowohl Gateelektrode als auch Substrat aus Silizium. Wird bei einem p-Substrat programmiert, befindet sich dieses in Inversion. Die Inversionsschicht ist die Quelle für die Elektronen, die in die Speicherschicht tunneln. Wird aber gelöscht, tritt die Injektion von Elektronen auf der Gateseite auf [35]. Liegt auf beiden Injektionsseiten FN-Tunneln vor, kommt es dazu, dass auch bei einer negativen Spannung programmiert wird, wie Abb. 2.14 anhand der durchgezogen Kurve verdeutlicht.

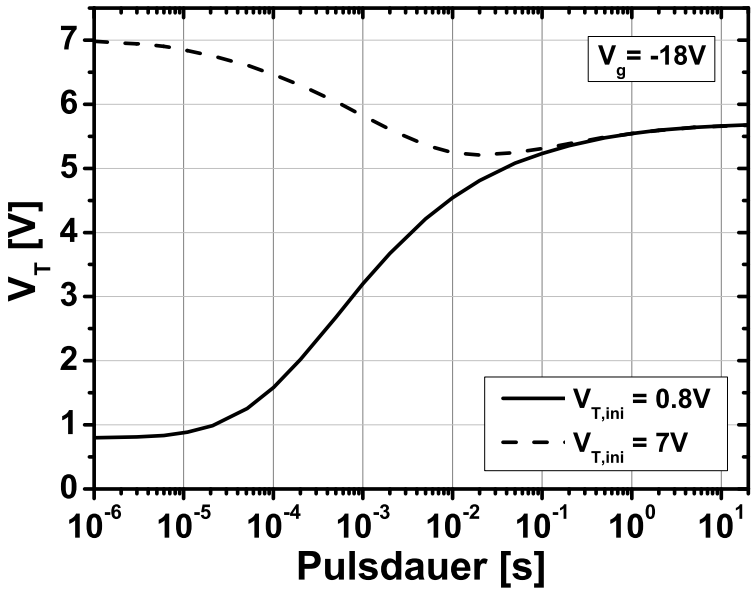

Abbildung 2.14: Löschcharakteristik eines SONOS-Transistors bei einer negativen Gatespannung für zwei verschiedene Ausgangszustände, einmal einer frischen Zelle (durchgezogene Linie) und einer auf $V_T = 7$ V programmierten Zelle (gestrichelte Linie), L/W = 5x5 μm, ONO: 2.8/9/9 nm

Nach einer entsprechenden Zeit verschieben sich die Feldverhältnisse im SONOS-Schichtstapel so, dass der Löcher- und Elektronenstrom auf einen vergleichbaren Wert jeweils zu- und abnehmen. Dann findet keine effektive Ladungsspeicherung mehr statt und die Programmierung sättigt auf einem stabilen Niveau. Bei SONOS mit einem dünnen Tunneloxid wird aus diesem Grund ausgenutzt, dass der Tunnelstrom bei direktem Tunneln größer ist als bei FN-Tunneln und vergleichbaren Feldern. Damit lässt sich das Problem der Elektroneninjektion von der Gateelektrode unterdrücken, gleichzeitig leidet aber die Ladungshaltung. Ein Ausweg aus dem Kompromiss zwischen Löschbarkeit und Ladungshaltung stellt die TANOS-Struktur dar.

2.3.4 TANOS-Struktur

Die Bezeichnung TANOS ergibt sich aus den verwendeten Materialien. Angefangen bei der Gateelektrode bis zum Substrat sind das Tantalnitrid, Aluminiumoxid, Siliziumnitrid, Siliziumoxid und das Siliziumsubstrat. Bei der Verwendung dieser Material-Kombination wird der Stapel auf der Gateelektrodenseite so modifiziert, dass im Löschvorgang die Elektroneninjektion stark reduziert ist. Ein Vergleich für die unterschiedlichen Gateelektroden-Materialien erfolgt in Kap. 4.3. Die Änderung des Topoxid-Materials hin zu Aluminiumoxid bewirkt eine effektive Reduktion des elektrischen Feldes, da Aluminiumoxid eine größere Dielektrizitätskonstante ϵ_r aufweist. Diese ist im Vergleich zu SiO_2 (ϵ_r= 3.9) mit 9.5 ungefähr 2,5 mal so groß. Demzufolge ist aufgrund der konstanten elektrischen Flussdichte D

$$E = \frac{D}{\epsilon_0 \epsilon_r} \tag{2.29}$$

das elektrische Feld im Al_2O_3 der ungeladenen TANOS-Speicherzelle nur circa 40% dessen vom SiO_2-Tunneloxid. In Abb. 2.15a wird dies anhand dem Vergleich einer Struktur mit SiO_2 und Al_2O_3 gleicher equivalenter Oxiddicke (EOT), veranschaulicht. Die Normierung resultiert in einem vergleichbaren Feld im Tunneloxid und somit gleichen Löschbedingungen. Damit verbunden ist aber auch eine Vergrößerung der Al_2O_3-Oxiddicke, damit der Spannungsabfall über dem Topoxid konstant bleibt.

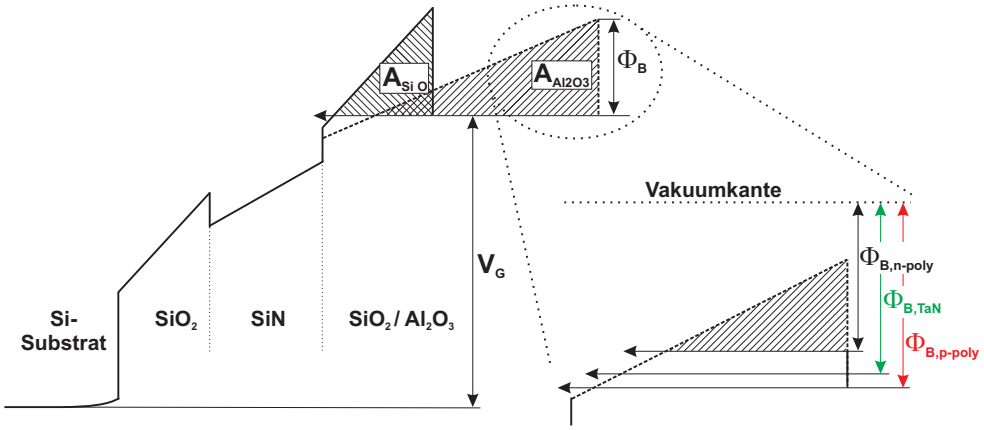

Abbildung 2.15: Leitungsbandausschnitt aus dem Bänderdiagramm für einen Schichtstapel mit SiO_2 - (durchgezogen) und Al_2O_3 -Topoxid (gestrichelt) unter Löschbedingungen; der vergrößerte Bildausschnitt verdeutlicht den Einfluss der Gateelektrodenaustrittsarbeit auf die Tunnel-Barrierenfläche

Vergleicht man die Tunnelbarriere für die Elektronen, ergibt sich ein Flächenverhältnis von $A_{SiO2} : A_{Al2O3} \approx 1{:}2$ zugunsten des Aluminiumoxids. Dies ist trotz der geringeren Barrierenhöhe der Fall, die zu n^+-dotiertem Silizum bei Aluminiumoxid ϕ_B= 2.8 eV beträgt, im Vergleich zu 3.1 eV bei SiO_2. Daraus resultiert eine überproportionale Abnahme des Elektroneninjektionsstromes, da entsprechend Gl. 2.21 der Tunnelstrom nichtlinear mit dem elektrischen Feld abnimmt.
Eine weitere Einflussgröße auf die Injektion von Elektronen ist das Material der Gateelektrode selbst. Denn unterschiedliche Dotierung und die Nutzung anderer Materialien ändert die Austrittsarbeit. In Abb. 2.15 wird deutlich, dass zum Beispiel hoch dotiertes n-Silizium eine kleinere Austrittsarbeit als p-dotiertes Silizium aufweist.

Das Resultat ist ein deutlich größerer Injektionsstrom vom Gate für das n^+-dotierte Silizium und demnach ein schlechteres Löschverhalten. In Tab. 2.1 sind die Austrittsarbeiten einiger Gateelektroden aufgeführt.

Tabelle 2.1: Austrittsarbeiten verschiedener Gateelektroden auf SiO_2 [36, 37]

Material	Austrittsarbeit
n^+-Si	4.05 eV
Al	4.1 eV
TaN	4.7 eV
TiN	4.8 eV
p^+-Si	5.1 eV
Au	5.1 eV
Pt	5.7 eV

Es zeigt sich, dass p^+-dotiertes Silizium eine im Vergleich sehr hohe Austrittsarbeit aufweist. Allerdings ist dies nur ein Wert, der bei kleinen elektrischen Feldern gültig ist, wie in Kap. 4.3.1 aufgezeigt wird. Sind die Felder zu groß, kommt es zur Verarmung und die wirksame Austrittsarbeit wird kleiner. Als ideales Elektrodenmaterial haben sich TaN und TiN herausgestellt, da diese sehr temperaturbeständig sind. Eine Untersuchung des elektrischen Verhaltens dieser Materialien erfolgt in Kap. 4.3.2. Des Weiteren erfolgt die Beschreibung der betrachten Schichtstapel mit Hilfe von Abkürzungen, die in Tab. 2.2 erläutert werden.

Tabelle 2.2: Übersicht über die verwendeten Abkürzungen der Schichtstapel haftstellen-basierter Speicherzellen

Abkürzung	Schichtstapel
ONO ; SONOS	Si-Substrat / SiO_2 / Si_3N_4 / SiO_2 / poly-Si-Elektrode
ONA ; SANOS	Si-Substrat / SiO_2 / Si_3N_4 / Al_2O_3 / poly-Si-Elektrode
ONA ; TANOS	Si-Substrat / SiO_2 / Si_3N_4 / Al_2O_3 / Metall-Elektrode

2.3.5 Floating-Gate-Struktur

Im Gegensatz zu der haftstellen-basierten Speicherschicht der SONOS-Struktur befindet sich bei der Floating-Gate Struktur eine leitende Schicht zwischen den isolierenden Oxidschichten. Man spricht hierbei auf Substratseite vom Tunneloxid. Zwischen der Speicherelektrode und der Steuerelektrode befindet sich das Interpoly-Dielektrikum (IPD). Die Steuerelektrode wird auch als Control-Gate (CG) bezeichnet. Der schematische Aufbau ist noch einmal in Abb. 2.16a illustriert.
Für das Tunneloxid wird reines SiO_2 als Isolatorschicht verwendet, an deren Defektdichte sehr große Anforderungen gestellt werden. Tritt nur ein schwacher Leckpfad auf, wird das gesamte Floating-Gate entladen und die gespeicherte Information geht verloren. Für das Interpoly-Dielektrikum wird eine Schicht wie bei der SONOS Struktur bestehend aus zwei SiO_2-Schichten mit einer eingebetteten SiN-Schicht verwendet, wie in Abb. 2.16a gezeigt. Dieser Schichtstapel wird angewandt, da es sich hierbei

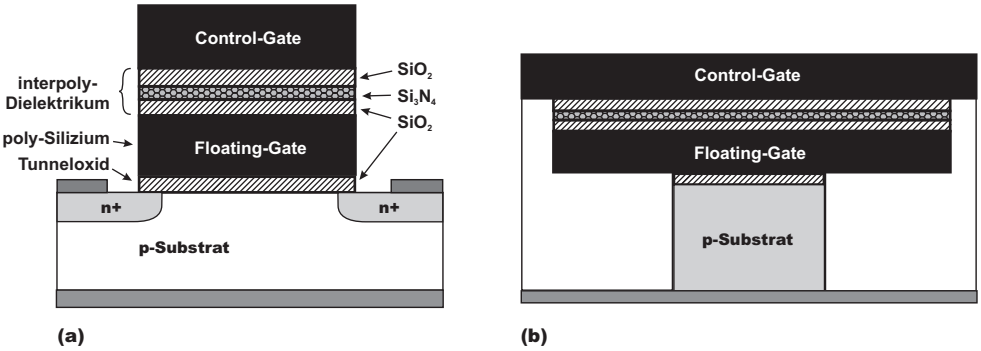

Abbildung 2.16: (a) Schematischer Aufbau einer Floating-Gate (FG) Speicherzelle in Wortleitungsrichtung und in (b) in Bitleitungsrichtung, Kopplung zwischen Control-Gate (CG) und Floating-Gate wird verbessert, wenn sich die Fläche zwischen den Elektroden im Vergleich zum Kanal durch Verlängerung des Floating-Gates vergrößert

um eine gute Elektronenbarriere handelt, die verhindert, dass beim Programmieren Elektronen zum Control-Gate abfließen. Ist dies trotzdem der Fall, werden die Elektronen in der SiN-Schicht gespeichert und das Nachfließen weiterer Elektronen wird vermindert. Die Funktionsweise der Floating-Gate Struktur basiert auf der deutlich größeren kapazitiven Kopplung zwischen Control- und dem Floating-Gate im Vergleich zur Kopplung zum Substrat, $C_{IPD} > C_{Tun}$. Bereits in der Einleitung wurde die Möglichkeit vorgestellt, dass man die Kopplung zwischen Control- und Floating-Gate durch eine Oberflächenvergrößerung verbessern kann. Hierzu wird das Control- um das Floating-Gate herumgeführt. Dies ist aber nur bei hoch-integrierten Strukturen notwendig. Eine einfachere Möglichkeit stellt die Verbreiterung des Floating-Gates, wie in Abb. 2.16b gezeigt, dar. Der über dem Fülloxid liegende Bereich wird hierbei als Flügel (engl. wing) bezeichnet. Hierdurch ergibt sich eine bessere Kopplung zum Control-Gate, ausgedrückt durch den Control-Gate-Koppelfaktor α_c:

$$\alpha_c = \frac{C_{IPD}}{C_{Ges}} = \frac{C_{IPD}}{C_{IPD} + C_{Tun}}. \tag{2.30}$$

Dieser Faktor dient auch der Berechnung des Floating-Gate Potentials, welches sich durch Gl. 2.31 berechnen lässt:

$$V_{FG} = \alpha_{CG} V_{CG} + \frac{Q_{FG}}{C_{Ges}}. \tag{2.31}$$

Ist α_c größer als 0.5, resultiert eine stärkere Kopplung zum Gate und der Spannungsabfall der angelegten Gatespannung erfolgt hauptsächlich über dem Tunneloxid. Erfolgt nun eine Programmierung durch eine ausreichend große Spannung am Control-Gate, tunneln die Ladungsträger auf das Floating-Gate und werden als Floating-Gate Ladung Q_{FG} gespeichert. Sie können dieses nicht weiter in Richtung Control-Gate verlassen, weil das Feld im IPD deutlich kleiner ist und kein ausreichender Tunnelstrom entsteht. Die gespeicherte Ladung verschiebt nun nach Gl. 2.31 bei konstanter Spannung V_{CG} die Floating-Gate Spannung V_{FG}. Daraus resultiert eine dauerhafte Verschiebung der Schwellspannung V_T, dem gewünschten Speichereffekt.

2.4 Speicherarchitekturen

Die Speicherarchitektur beschreibt die Anordnung der Speicherzellen in großen Zellenfeldern. Hierfür gibt es eine Vielzahl von Möglichkeiten, die unterschiedlichen Anforderungen genügen. Durchgesetzt hat sich die Anordnung in NOR für Speicher mit schnellem wahlfreien Zugriff auf beliebige Byte, wie es zum Beispiel für Programmspeicher notwendig ist. Für eine Datenspeicherung hingegen spielt der Preis und demzufolge die Dichte an Speicherzellen die größte Rolle und es hat sich daher die NAND Architektur durchgesetzt. Es gibt aber auch eine Vielzahl anderer Anordnungen, wie zum Beispiel AND [38] oder als Teil von nichtflüchtigen SRAMs [33]. Im folgenden Abschnitt werden die NOR und NAND-Architektur vorgestellt und im Anschluss dargelegt, welche Effekte beim Programmieren von NAND berücksichtigt werden müssen.

2.4.1 NOR

Bei der NOR-Architektur sind die Speicherzellen parallel geschaltet, wie in Abb. 2.17a illustriert. Es gibt neben dem gezeigten Konzept noch weitere Verschaltungsmöglichkeiten, wie zum Beispiel das virtual-ground-NOR Konzept [39]. Das Auslesen der Speicherzellen erfolgt, indem eine Wortleitung (WL) auf eine Auslesespannung gelegt wird, die sich zwischen programmiertem und gelöschtem Zustand befindet. Alle anderen Wortleitungen erhalten eine Spannung, bei der die Zellen sicher ausgeschaltet sind. Ist nun die ausgewählte Speicherzelle programmiert und das V_T liegt über der Auslesespannung, fließt kein Strom von der Bitleitung (BL) zur Sourceleitung (SL). Im gelöschten Fall fließt ein Strom und die Information kann ausgelesen werden.

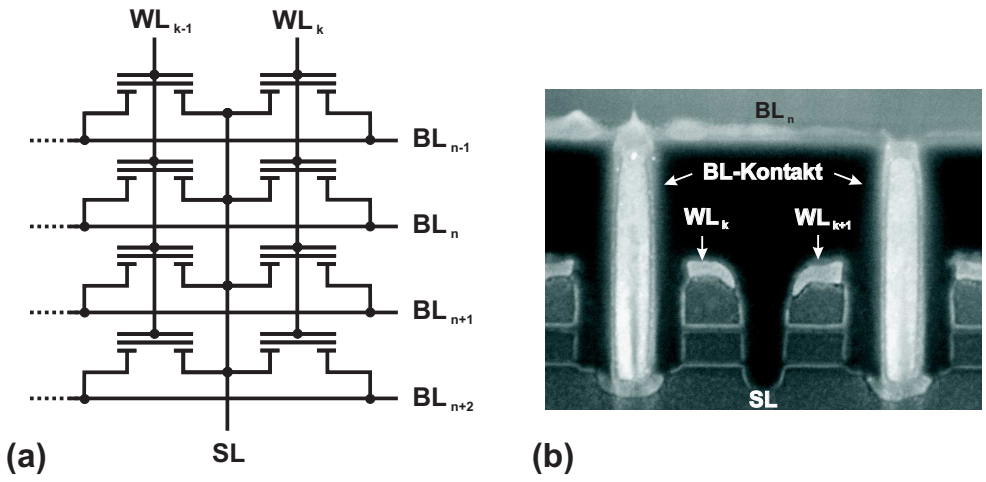

Abbildung 2.17: (a) Verschaltung der Speicherzellen bei der NOR-Architektur; angesteuert werden die Zellen über die Wortleitung (WL) und der Stromfluss erfolgt über die Bitleitung (BL) zur Sourceleitung (SL); (b) TEM-Aufnahme eines NOR-Flashspeicher Produktes der 65nm-Generation von INTEL [40]

Aufgrund der parallelgeschalteten Speicherzellen resultiert eine ODER-Verknüpfung. NOR ergibt sich aus der Tatsache, dass eine programmierte Zelle keinen Stromfluss aufweist und demnach das Leseergebnis negiert wird. Ein Thema dem hierbei Beachtung geschenkt werden muss, ist die Berücksichtigung des Stromes der abgeschalteten

parallelen Speicherzellen. Ist nur eine dieser Zellen leitend, weil sie zum Beispiel unter die Abschaltspannung gelöscht wurde, wird das Ausleseergebnis verfälscht. Hierfür gibt es zwei mögliche Ursachen, einmal das Löschen unter die Abschaltspannung (engl. over-erase) [41] oder der Leckstrom $I_{D,L}$ der vielen parallelen Zellen ist zu groß. Die Messschaltung wertet in diesem Fall einen Strom aus, obwohl der ausgewählte Transistor nicht leitet.
In Abb. 2.17b ist die Realisierung auf einem Produkt gezeigt. Es wird deutlich, dass für jede Speicherzelle ein Kontakt für die Verknüpfung mit der Bitleitung notwendig ist. Dies hat einerseits den Vorteil, dass die Zelle direkt angesprochen werden kann und daraus eine hohe Lesegeschwindigkeit für wahlfreien Zugriff resultiert. Andererseits wird aber immer ein gewisser Raum benötigt, um diesen Kontakt zu platzieren. Dies erhöht den Platzbedarf pro Speicherzelle und die Flächeneffizienz sinkt.

2.4.2 NAND

Eine deutlich größere Flächeneffizienz weist die NAND-Architektur auf. In Abb. 2.18 ist ausschnittsweise der Schnitt durch eine NAND-Kette (engl. NAND-string) gezeigt. Es befinden sich k+1 Speicherzellen in Reihe geschaltet mit zwei Auswahltransistoren zwischen dem Sourceleitungs- und dem Bitleitungskontakt. Begonnen hat man mit 4 Speicherzellen [42], wobei heutige Produkte bis zu 64 Speicherzellen enthalten können [43]. Damit verteilt sich der Anteil von nicht speicherrelevanter Fläche auf eine große Anzahl von Speicherzellen und die Flächeneffizienz wird besser. Erkauft wird dieser Gewinn durch eine deutlich aufwendigere Beschaltung mit Auswahltransistoren und einer geringeren effektiven Lesegeschwindigkeit, weil die Speicherzellen in einer NAND-Kette seriell ausgelesen werden müssen. Durch eine große Parallelisierung lässt sich der Geschwindigkeitsnachteil egalisieren, allerdings wächst damit auch die Größe des zu beschreibenden oder zu lesenden Bereiches, der sogenannten 'Page size' [3, 44]. Dadurch werden heute Datenraten von 100 MB/s für Lesen und Schreiben erreicht [44].
Zum Auslesen einer Speicherzelle müssen alle anderen Speicherzellen der NAND-Kette so geschaltet werden, dass sie leitend sind. Es wird eine Spannung angelegt, die den Strom passieren lässt, man spricht daher auch von der 'Pass-Spannung'. Diese liegt im Normalfall ungefähr 2.5 V über dem maximal möglichen V_T. Bei einer Multi-Level-Zelle mit max. 4 V V_T ergeben sich 6.5 V Pass-Spannung. Die Auswahltransistoren haben ein festes V_T von $\approx$ 0.7 V und werden deswegen mit der Versorgungsspannung von 3.3 V eingeschalten. Nun ist ein Stromfluss durch die Kette möglich und es wird an die ausgewählte Zelle entsprechend NOR eine Lesespannung angelegt. Aufgrund der kleinen Zellen, die sich in heutigen Produkten befinden, ist der Strom so klein, dass er nicht mehr direkt gemessen werden kann ($\approx$ 1 μA). Aus diesem Grund wird eine zuvor geladene Kapazität durch den Stromfluss entladen. Anschließend erfolgt die Bestimmung der Spannung an der Kapazität, wodurch sich der Zustand der Speicherzelle ablesen lässt [45]. Gelöscht werden die Speicherzellen, indem eine positive Spannung an das Substrat der Speicherzellen angelegt wird. Hierbei wird ein größerer Bereich parallel gelöscht, der größer sein kann, als der zuvor beschriebene Programmierbereich (Erase-Block). Das gezielte Programmieren in einem NAND-Zellenfeld ist weitaus komplexer, weil bei dem Anlegen der Programmierspannung ein versehentliches Programmieren der Nachbarzellen vermieden werden muss. Die entsprechende Konfiguration zum fehlerfreien Programmieren wird

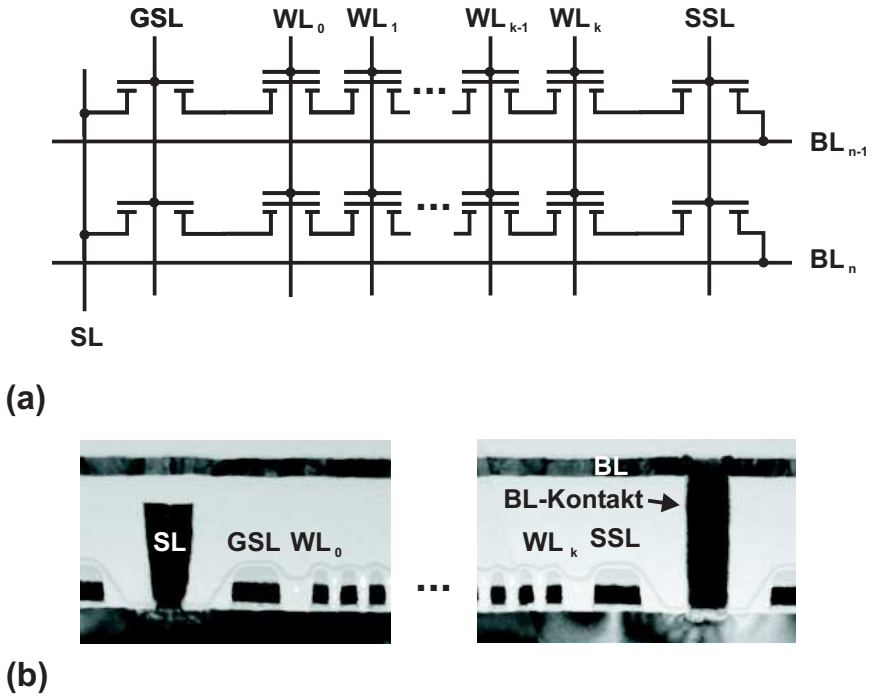

Abbildung 2.18: (a) Verschaltung der Speicherzellen bei der NAND-Architektur, angesteuert werden die Zellen über die Wortleitung (WL) und der Stromfluss erfolgt über die Bitleitung (BL) zur Sourceleitung (SL); (b) TEM-Aufnahme von einer untersuchten Struktur mit den Abmessungen von 48x48 nm

im folgenden Abschnitt erläutert. Da sich diese Arbeit auf die NAND-Architektur konzentriert, werden im Folgenden die wesentlichen Störmechanismen, die in einem NAND-Speicherfeld beim Programmieren auftreten, erklärt.

2.4.3 Störmechanismen beim Programmieren eines NAND-Zellenfeldes

In Speicherzellenfeldern sind eine Vielzahl von Speicherelementen so verbunden, dass sich im Kreuzungspunkt zweier Leitungen die gewählte Zelle befindet. Legt man nun eine Spannung an eine Leitung an, sieht nicht nur die zu programmierende Speicherzelle diese Spannung, sondern auch alle Nachbarn entlang dieser Leitung. Dies ist unabhängig von der Architektur, da sowohl bei NOR als auch bei NAND zum Beispiel die Wortleitungen immer mehrere Speicherzellen überdecken (siehe Abb. 2.17a und 2.18a). Dadurch kann es zu unbeabsichtigtem Programmieren von Nachbarzellen (engl. disturb) kommen [46]. In Abb. 2.19a sind die Zellen markiert, an denen ein ungewolltes Programmieren beim gezielten Programmieren der mittleren Speicherzelle auftritt.
Das Programmieren von einer NAND-Speicherzelle erfolgt, indem an der entsprechenden WL die Programmierspannung V_{Prog} angelegt wird. Damit die gewünschte Zelle programmiert wird, muss ein Kanalpotential von 0V vorliegen, wie in Abb 2.19b I veranschaulicht. Dies wird erreicht, indem die benachbarten Zellen bis zum bitleitungsseitigen Auswahltransistor (SSL) mit V_{Pass} eingeschaltet sind. Dieser Auswahltransistor (SSL) ist aufgrund der angelegten Spannung von 3 V eingeschaltet und leitet

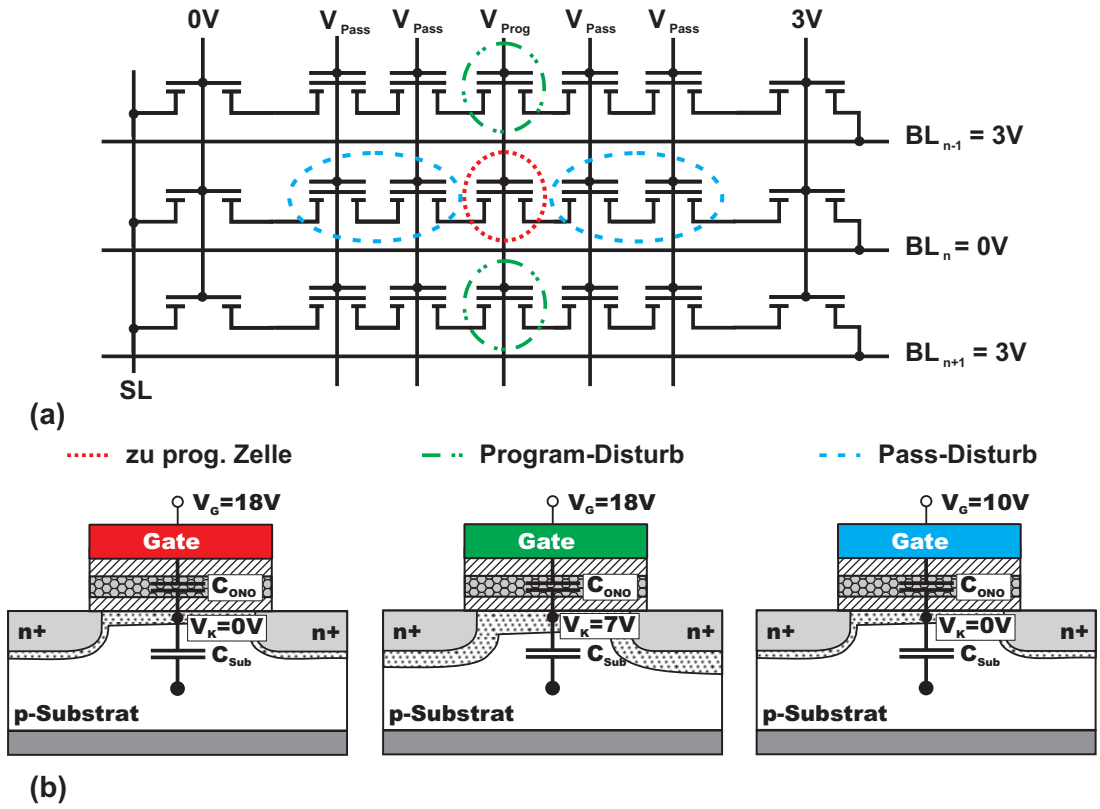

Abbildung 2.19: (a) Zu programmierende Zelle in einem NAND-Speicher (rot) und durch Pass-Disturb (blau) bzw. Program-Disturb (grün) gestörte Speicherzellen; (b) Potentiale an den Speicherzellen unter Programmierbedingungen, I zu programmierende Zelle, II Program-Disturb, III Pass-Disturb

die am Bitleitungskontakt angelegten 0V weiter. Demnach sieht die zu programmierende Speicherzelle eine Spannung V_{Prog} und alle anderen Speicherzellen in der NAND-Reihe sehen eine Spannungsdifferenz von V_{Pass} (blau markiert). Im gezeigten Beispiel in Abb. 2.19b sind das entsprechend 18 V und 10 V. Durch die angelegte Spannung V_{Pass}, an den nicht zu programmierenden Zellen, kann es aber zu einer Verschiebung des V_T's kommen. Man nennt diesen Störmodus entsprechend der angelegten Spannung Pass-Disturb.
Aber die Programmierspannung wird auch von den grün markierten Zellen der Nachbar-NAND-Reihen gesehen. Damit diese nicht programmiert werden, wird ein Effekt ausgenutzt, der 'channel-boosting' genannt wird und ein durch kapazitive Kopplung angehobenes Kanalpotential bezeichnet. Es wird an den NAND-Reihen, die nicht programmiert werden sollen, eine Bitleitungsspannung von 3 V angelegt, wie in Abb. 2.19a gezeigt. Dann ist der source-seitige Auswahltransistor (GSL) so beschalten, dass dieser nicht mehr leitet, wie es bei 0 V Bitleitungsspannung der Fall wäre. Da sich auch der SSL-Auswahltransistor, wegen angelegten 0 V im gesperrten Zustand befindet, ist das Kanalgebiet dieser NAND-Reihen mit keinem festen Potential mehr verbunden. Nun kommt es entsprechend dem kapazitiven Teiler aus Substratkapazität C_{Sub} und Gesamtspeicherzellkapazität C_{ONO}:

$$\frac{V_K}{V_{Pass}} \approx \frac{C_{ONO}}{C_{ONO} + C_{Sub}} \tag{2.32}$$

zur Ausbildung eines Kanalpotentials V_K, das größer als 0 V ist [47]. Messungen

haben gezeigt, dass sich ein Kanalpotential bis zu 7 V aufbauen kann. Zellen, die sich in solch einer NAND-Reihe befinden und auf der Programmier-WL liegen, sehen demnach eine Spannung entsprechend der Differenz aus Kanalpotential und Programmierspannung $V_{Prog} - V_K$. Für das Beispiel in Abb. 2.19b ergibt sich demnach eine Spannung von 11 V, klein genug um noch nicht zu programmieren. Entsprechend der angelegten Gate-Spannung wird dieser Modus Program-Disturb genannt. Die Spannungsdifferenz zur Programmierspannung wird umso kleiner, je höher die angelegte Pass-Spannung gewählt wird. Das Ergebnis ist eine geringere V_T-Verschiebung, wie in Abb. 2.20 anhand einer Messung gezeigt wird. Das maximal erreichbare Kanalpotential ist aber durch Leckströme, vorrangig durch Band-zu-Band Tunneln am Auswahltransistor, begrenzt. Demzufolge kann die erreichbare Spannung durch weitere Größen, wie die Kontaktdotierung beeinflusst werden (siehe Kap. 4.4). Weiterhin begrenzen die Leckströme auch die mögliche Gesamtpulslänge, da sie das Kanalpotential mit der Zeit absenken. Es ist demzufolge eine genaue Abstimmung zwischen den Leckströmen und den Pulslängen notwendig. Eine Erhöhung der Spannung V_{Pass} wiederum führt dazu, dass die nicht zu programmierenden Zellen in der NAND-Reihe, wie die zu programmierende Zelle, eine immer höhere Gatespannung sehen. Das Ergebnis ist ein mit der Pass-Spannung ansteigendes V_T, wie Abb. 2.20 verdeutlicht. Bei den Messungen wurde jeweils berücksichtigt, dass die Zeiten der angelegten Spannungen ungefähr den Bedingungen entsprechen, wie sie in Produkten auftreten. Wird dies alles berücksichtigt, bekommt man im gezeigten Fall ein mögliches Fenster für die Wahl der Pass-Spannung von 5 V - 13 V. Zugrunde gelegt wurde die Bedingung, dass das V_T der beeinflussten Zelle nicht über -2 V gelangt.

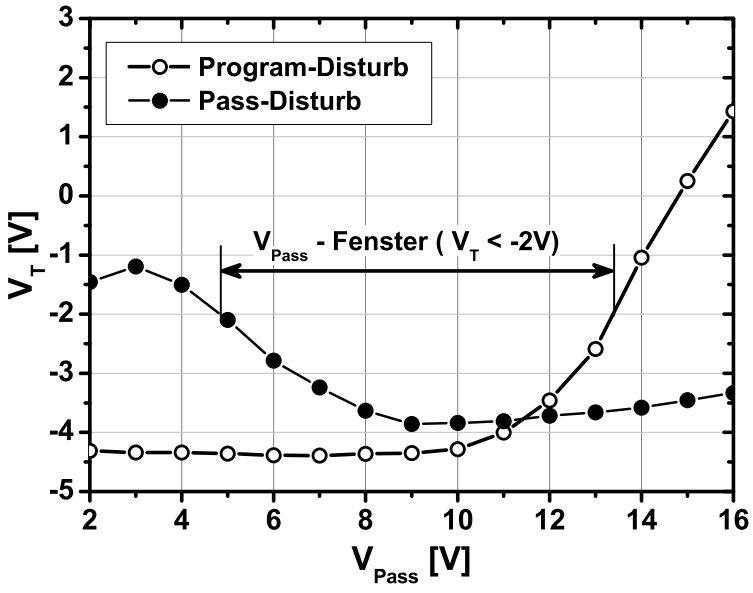

Abbildung 2.20: V_T der beeinflussten Zelle nach Program-Disturb (geschlossene Symbole) und nach Pass-Disturb (offene Symbole), es ergibt sich ein Fenster für die Wahl der Pass-Spannung, die Messung wurde an FG-NAND-Strukturen durchgeführt

2.5 Charakterisierungsmethoden von Halbleiter-Speicherelementen

2.5.1 Inkrementelle Gatespannungs-Programmierung

Bei der inkrementellen Gatespannungs-Programmierung (engl. incremental step pulse programming, ISPP) wird bei einer konstanten Pulszeit die Pulsspannung schrittweise erhöht, wie in Abb. 2.21a gezeigt [48]. Zudem wird gleichzeitig veranschaulicht, wie der zeitliche Verlauf einer Schwellspannungsbestimmung aussehen kann. Diese erfolgt immer zwischen den Programmierpulsen. Bei den untersuchten Proben ist ein sinnvoller Wert für die Schrittweite 0,5 V. Der verwendete Spannungsbereich hängt stark vom Schichtstapel ab und liegt bei den Standard-TANOS-Zellen mit einem EOT $\approx$ 13 nm zwischen 10 - 20 V bei einer Pulszeit von 100 μs.
Die in Abb. 2.21b dargestellten Programmiercharakteristiken zeigen einen typischen Verlauf einer Messkurve. Nachdem der Programmiervorgang eingesetzt hat, ist der Kurvenverlauf nahezu linear. Die Steigung der Kurve wird als Programmiersteigung bezeichnet (engl. program-slope). Wird eine negative Spannung an das Gate angelegt, spricht man von der Löschsteigung. Die Steigung bei haftstellen-basierten Halbleiter-Speichern liegt im Bereich von 0,5 - 1 V/V und hängt vorrangig von den Schichtdicke ab. Weitergehende Untersuchungen werden in Kap. 3.2.3 diskutiert. Im Gegensatz dazu ist die Programmier-Steigung bei Floating-Gate Speicherzellen immer nahezu 1 [5, 49]. Eine Änderung der Pulszeit bewirkt eine Verschiebung auf der Gatespannungsachse. Wie in Abb. 2.21b verdeutlicht, resultiert eine Erhöhung der Pulszeit in einem Einsetzen des Programmierens bei niedrigeren Spannungen.

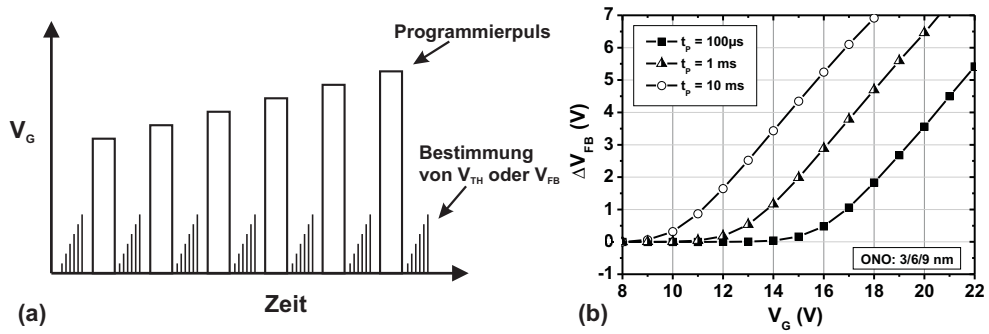

Abbildung 2.21: a) Vergrößerung der Programmierpulsspannung bei konstanter Pulsdauer, zwischen den Programmierpulsen wird das V_T oder V_{FB} der untersuchten Struktur bestimmt; b) exemplarische Messkurven einer SONOS-Speicherzelle für drei verschiedene Pulszeiten mit konstanter Steigung

2.5.2 Transiente Programmierung

Man spricht von einer transienten Programmierung, wenn bei einer konstanten Spannungsamplitude schrittweise die Pulszeit erhöht wird. Der zeitliche Ablauf eines Programmiervorgangs ist in Abb. 2.22a veranschaulicht. Die Programmierpulslänge nimmt im gezeigten Fall logarithmisch zu. Dadurch erreicht man eine günstigere Verteilung der Messpunkte bei logarithmischer Auftragung, wie Abb. 2.22b verdeutlicht. Zwischen den Pulsen erfolgt, wie bei der inkrementellen Gatespannungs-

Programmierung, die Bestimmung der Schwellspannung, um den Verlauf des Programmierens aufzuzeichnen.

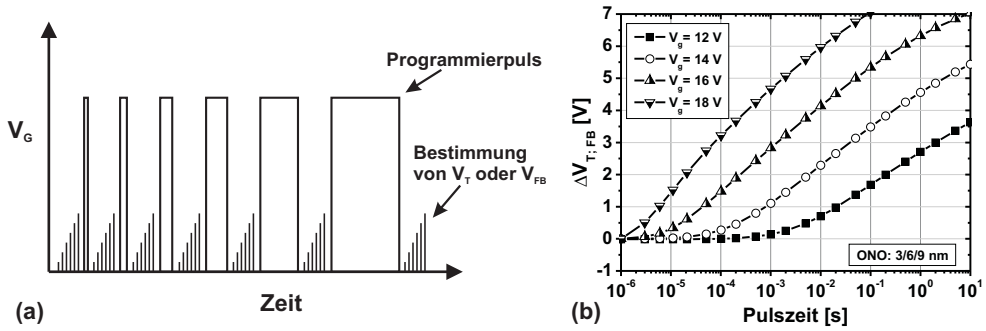

Abbildung 2.22: a) Programmierung mit zunehmender Pusldauer bei konstanter Programmierspannung, zwischen den Programmierpulsen erfolgt die Bestimmung von V_T oder V_{FB} ; b) Verhalten einer SONOS-Speicherzelle bei 4 verschiedene Programmierspannungen

Im Gegensatz zur inkrementellen Gatespannungsprogrammierung ist die Steigung der Programmierkurve nicht konstant, wie in Abb. 2.22b gezeigt. Sie variiert sowohl in Abhängigkeit von der eingeprägten Spannung als auch von der Pulsdauer. Eine höhere Spannung resultiert in einem schnelleren Einsetzen des Programmiervorgangs und verschiebt die Kurve zu kürzeren Pulszeiten. Dieses Verfahren wird im allgemeinen dafür verwendet, verschiedene Proben miteinander zu vergleichen. Denn es ist einfach möglich, bei gleicher Pulszeit und Programmier- bzw. Löschspannung, die erreichten V_T's gegenüberzustellen. Prinzipiell ist es mit verringerter Genauigkeit möglich, die Charakteristiken von transientem Programmieren in inkrementelles Programmieren umzurechnen und umgekehrt.

2.5.3 Messung des Ladungsverlustes

Die in Kap. 2.5.1 und 2.5.2 vorgestellten Verfahren untersuchen die Injektion von Ladungsträgern in die Speicherschicht. Für eine nichtflüchtige Speicherzelle ist aber die Haltezeit der gespeicherten Information ein weiteres wichtiges Kriterium. Daher muss auch eine Betrachtung des Verlusts der gespeicherten Ladung über die Lagerzeit erfolgen. Der Ladungsverlust kann über die Verschiebung der Flachbandspannung bei einer Kapazität bzw. der Schwellspannung eines Transistors ermittelt werden. Erste Betrachtungen des Ladungsverlustes innerhalb weniger Sekunden wurden bereits 1973 durchgeführt [50]. Hierbei erfolgte die Auswertung mittels Darstellung auf einem Oszilloskop anhand einer rückgekoppelten Kapazitätsmessung. Der dafür nötige Kalibrieraufwand ist heutzutage nicht mehr nötig, da es möglich ist, innerhalb kurzer Zeit eine komplette Bauelement-Kennlinie aufzuzeichnen. Aus der Kennlinie kann dann in zeitlichen Abständen die Spannungsverschiebung extrahiert werden [51, 52]. Ein Diagramm, welches solch einen Kurvenverlauf zeigt, ist in Abb. 2.23 dargestellt. Es wird für drei unterschiedliche Programmierzustände jeweils bis zu einer Zeit von 8 h die Schwellspannung gemessen.
Dieses Verfahren ist günstig bei langen Zeiten zwischen den Messpunkten, da für die Messung eine gewisse Zeit, in der Größenordnung von einigen 100 ms, benötigt wird. Zudem kann durch die Messung der kompletten Kennlinie festgestellt werden, ob die

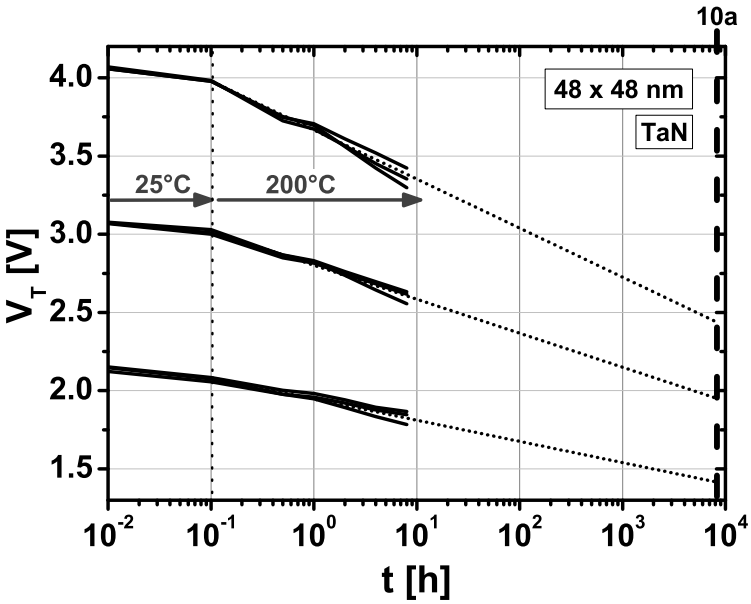

Abbildung 2.23: Ladungsverlust für drei Programmierzustände zwischen 2 und 4 V V_T bei einer Temperung von 200°C und eine Interpolation auf 10 Jahre, Darstellung des Mittelwertes aus jeweils 40 gemessenen Strukturen auf drei vergleichbaren Wafern; Größe 48x48 nm; ONA: 5/6/12 nm

Probe degradiert. Dies zeigt sich durch eine Veränderung der Kennlinie und muss bei der Auswertung berücksichtigt werden. Für die Charakterisierung bei kleinen Zeiten gibt es die Möglichkeit, den Strom eines Transistors bei einer festen Spannung zu messen. Bei bekannter Kennlinie kann die Änderung des Stroms daraufhin in eine Schwellspannungsänderung umgerechnet werden [53]. Bei einer großen Steigung der Unterschwellspannuungscharakteristik ist dieses Verfahren sehr empfindlich und ermöglicht eine Betrachtung ab einigen μs. Dies ist zum Beispiel für die Betrachtung von Random Telegraph Noise (RTN) erforderlich [54]. Für Produkte fordert man eine Informationsspeicherung von mindestens 10 Jahren. Daher wird normalerweise bei der Temperatur der Spezifikation gemessen und auf 10 Jahre extrapoliert, so wie es in Abb. 2.23 dargestellt ist. Ein Nachteil der Extrapolation bei der Spezifikation sind die unter Umständen zu kleinen V_T-Verschiebungen im Betrachtungszeitraum. Eine weitere Methode, um eine hinreichend genaue Vorhersage der Informations-Speicherzeit durchführen zu können, ist die Verwendung eines Beschleunigungsmodells [55], welches die Extrapolation durch eine Betrachtung bei erhöhter Temperatur durchführt. Dadurch kann die Zeit der Zuverlässigkeitsuntersuchung deutlich reduziert werden.

2.5.4 Zyklenfestigkeit

Ein Kriterium, welches bei Speicherzellen eine elementare Größe darstellt, ist die Fähigkeit, möglichst oft beschreib- und löschbar zu sein. Man spricht hierbei von Zyklenfestigkeit, engl. endurance. In Abb. 2.24 ist der Verlauf des gelöschten und programmierten Zustandes bis zu einer Zyklenzahl von 10.000 dargestellt.
Das Programmieren und Löschen erfolgt mit konstanten Pulsen. Hierdurch erhält man eine erste Aussage über eventuell auftretende Effekte, die zu einer Verschlechterung der Zelleigenschaften führen. Für die Zunahme der Schwellspannung kann zum

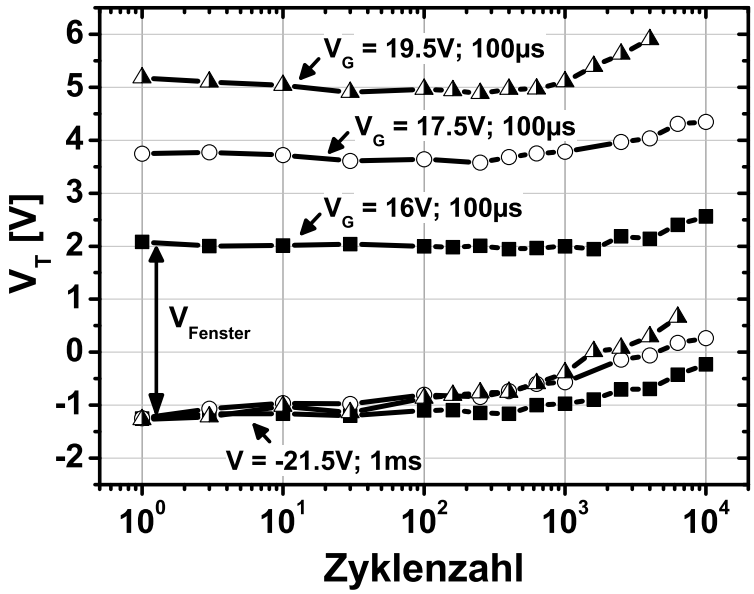

Abbildung 2.24: Darstellung des programmierten und gelöschten V_T's in Abhängigkeit der Programmier-/ Löschzyklen für drei verschiedene Programmierspannungen; Darstellung des Mittelwertes von 10 Speicherzellen; 48x48nm; TaN Gate; ONA: 5/6/12 nm

Beispiel die Speicherung von Elektronen im Tunneloxid veranwortlich sein [56, 57]. Dadurch kommt es zu einer parallelen Verschiebung von beiden Zuständen hin zu höheren V_T's. Dabei ist deutlich zu sehen, dass auch die Größe des Programmier-/ Löschfensters $V_{Fenster}$ einen Einfluss hat. Wird das Fenster vergrößert, kommt es zu einer stärkeren Oxidschädigung und die zyklen-abhängige V_T-Verschiebung wird größer.

3 Defektbasierte Ladungsspeicherung in dielektrischen Schichten

Bereits im Jahre 1969 wurde durch Wallmark [58] und Frohman [28] gezeigt, dass eine Schicht aus Siliziumnitrid (Si_3N_4) Ladung dauerhaft speichern kann, wenn sich darunter eine dünne Siliziumoxidschicht befindet. Es handelt sich dann um eine MNOS-Struktur (Metall-Siliziumnitrid-Siliziumoxid-Siliziumsubstrat). Allerdings sind bis heute nicht alle Effekte, die in Schichtstapeln mit einer Siliziumnitrid-Speicherschicht auftreten, verstanden und erklärt. Im Folgenden soll auf verschiedene Untersuchungen zu dem physikalischen Hintergrund eingegangen werden. Darauf erfolgt eine Betrachtung der Ladungsverteilung in der Speicherschicht, einerseits mit Hilfe von Injektionssimulationen und andererseits anhand von Messungen.

3.1 Physikalische Grundlagen von Haftstellen

Wie bereits erwähnt, konnte durch Messungen an MNOS-Schichtstapeln gezeigt werden, dass Ladung in Siliziumnitrid dauerhaft gespeichert werden kann. Dieser Speichermechanismus wurde als ein Prozess interpretiert, bei dem die durch Tunneln injizierte Ladung in diskreten Haftstellen der Speicherschicht eingefangen wird [58]. Da die durch die energetische Lage dieser Haftstellen aufgebaute Potentialbarriere ausreichend groß ist, kommt es zu einer dauerhaften Ladungsspeicherung. Mit Hilfe verschiedener Analysemethoden, wie zum Beispiel Elektronenenergieverlustspektroskopie (EELS) [59] und Elektron-Spin-Resonanz-Messungen (ESR) [60] konnten die Energien der Haftstellen genauer untersucht und zugeordnet werden. Grundsätzlich handelt es sich bei den Haftstellen um Defekte in der Struktur des amorphen Siliziumnitrids.
Die Betrachtungen haben gezeigt, dass für die Speicherung von Löchern und Elektronen zwei verschiedene Bindungstypen verantwortlich sind. Robertson [61] weist mit Hilfe von Röntgenphotoelektronenspektroskopie (XPS) nach, dass sich Punktdefekte sowohl an Silizum- als auch an Stickstoff-Atomen befinden. Diese besitzen die Fähigkeit Elektronen einzufangen und abzugeben. Weiterhin hat Kamigaki [60] mit Hilfe von ESR-Messungen gezeigt, dass eine mögliche Elektronen-Haftstelle aus einem Verbund eines Si-Atoms in Verbindung mit drei N-Atomen ($N_3 \equiv Si^0$) gebildet wird. Dies wird in Abb. 3.1a verdeutlicht. Hierbei hat das vierte Si- Valenzelektron eine Tendenz zum Einfangen eines freien Elektrons, wodurch sich die Struktur $N_3 \equiv Si^-$ ergibt. Im Gegensatz dazu wird gezeigt, dass die Haftstellen für Löcher bevorzugt durch einen Verbund aus vier Si-Atomen gebildet werden $Si_3 \equiv Si^0$. Das Atom mit dem freien Elektron gibt unter Umständen sein Elektron ab und geht in die positiv geladene Struktur $Si_3 \equiv Si^+$ über. In Abb. 3.1b wird der Übergang vom neutralen in den positiven Zustand illustriert. Dies stellt aber nur einen kleinen Ausschnitt der möglichen Bindungszustände und Haftstellen dar.

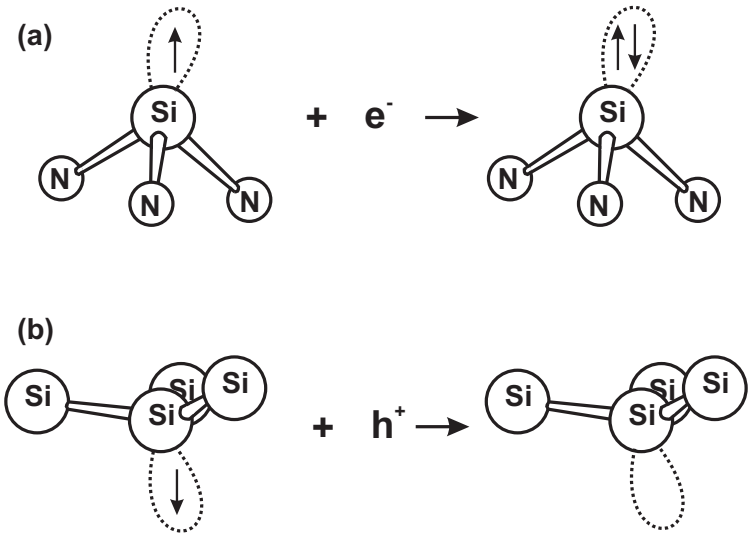

Abbildung 3.1: a) Struktur $N_3 \equiv Si^0$, für eine Elektronenhaftstelle und b) äquivalente Struktur für Löcher $Si_3 \equiv Si^0$

Die ersten theoretischen Betrachtungen zu möglichen Energiezuständen von Haftstellen wurden durch Robertson [62] durchgeführt. Hierbei wird gezeigt, dass energetisch flache und tiefe Haftstellen existieren. Diese Simulationsdaten wurden durch Messungen von Belyi [32] bestätigt. Die Untersuchung der Haftstellen-Energien ist nach wie vor von großem Interesse, da sie für die Vorhersage der Ladungshaltung elementar sind. Daher werden auch für spezielle Anwendungen, wie die NROM-Speicherzelle, dementsprechende Betrachtungen durchgeführt [63]. Die in Abb. 3.1 eingezeichnete Eigenschaft, dass sich beim Einfang eines Elektrons verschiedene Spinrichtungen ausbilden müssen, wurde durch Gritsenko [59] untersucht und bestätigt. Zudem wurde gezeigt, dass eine Haftstelle nicht nur die zwei Zustände, besetzt und nicht besetzt besitzt, sondern Löcher und Elektronen binden kann. In diesem Fall spricht man von einer amphoteren Haftstelle. Die bisherigen Betrachtungen wurden an stöchiometrischem Siliziumnitrid durchgeführt. Ändert man zum Beispiel durch Zugabe von Sauerstoff die Zusammensetzung der Speicherschicht oder sättigt Bindungen mit Wasserstoff ab, ergeben sich abweichende Energieverteilungen [60,64–66]. So wird gezeigt, dass der Einbau von Sauerstoff die Haftstellendichte reduziert [67], was sich wiederum direkt auf die elektrischen Eigenschaften der Speicherzelle auswirkt [66]. Das Ergebnis der reduzierten Haftstellendichte ist ein langsameres Programmieren und Löschen, sowie ein kleineres Programmier-/ Lösch-Fenster. Dies kann durch eigene Untersuchungen bestätigt werden, welche in Abb. 3.2a gezeigt sind. Auch der zu erwartende Effekt energetisch tieferer Haftstellen, bei denen sich die Ladungshaltung verbessert, konnte durch eigene Messungen nachgewiesen werden. Das elektrische Ergebnis wird in Abb. 3.2b gezeigt, wobei der Unterschied nicht so groß ist, wie erwartet. Das Ergebnis ist aber möglicherweise durch das dünne Tunneloxid der Proben beeinflusst. Unter Berücksichtigung eines nahezu vergleichbaren Ladungsverlustes ist das reine SiN bei Programmieren bzw. Löschen deutlich besser. Aus diesem Grund wird reines SiN in der Anwendung als Speicherschicht bevorzugt.
Eine Möglichkeit, die Anzahl energetisch flacher Löcher-Haftstellen zu erhöhen, ist eine Anreicherung der Speicherschicht mit Silizium [68]. Diese hat zwar die Eigenschaft, besonders gut zu Löschen, kann Elektronen jedoch schlechter dauerhaft spei-

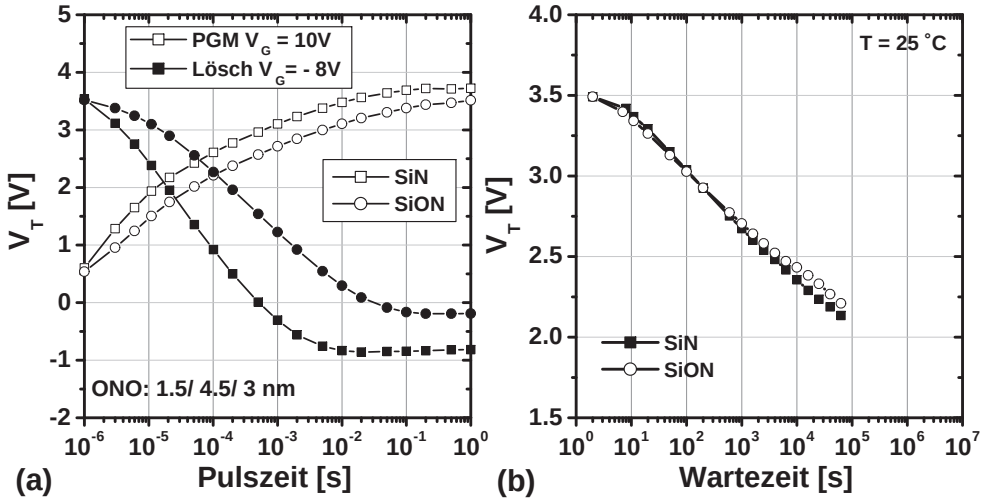

Abbildung 3.2: Vergleich zweier SONOS-Speichertransistoren mit SiN und oxidreicher Si-ON Speicherschicht; (a) Programmier- und Löschverhalten; (b) Messung der Ladungshaltung; 5x5μm Transistoren

chern [66, 69–71]. Aufgrund der im Vergleich zu stöchiometrischem Siliziumnitrid schlechteren Ladungshaltung im programmierten Zustand stellt die siliziumreiche Speicherschicht keine Alternative dar.

3.2 Betrachtung der vertikalen Ladungsverteilung mit Hilfe von Simulationen

Um die Funktionsweise von haftstellen-basierten Speichern genauer zu verstehen, ist es notwendig, die Injektion in und den Ladungsverlust aus der Speicherschicht mit Hilfe von Simulationen zu beschreiben. Dies ist die einzige Möglichkeit das Verhalten der Ladung in der Speicherschicht nachzuvollziehen. Denn die einzige Größe, die von außen an den zu untersuchenden Strukturen gemessen werden kann, ist die Schwell- bzw. Flachbandspannung. Daraus lässt sich aber noch keine Aussage über die Verteilung und Dichte der Ladungsträger ableiten. Die durch injizierte Ladungsträger hervorgerufene Spannungsverschiebung berechnet sich nach [34] durch:

$$\Delta V_{FB} = -\Delta Q_N \left[\frac{x_{bo}}{\epsilon_{bo}} + \frac{x_{ni} - \bar{x}_{ni}}{\epsilon_{ni}} \right]. \tag{3.1}$$

Die Größen x_{bo} und x_{ni} beschreiben jeweils die Dicken von Tunneloxid und der Siliziumnitrid-Speicherschicht. Die weiteren Teilgrößen, Änderung der Gesamtladung ΔQ_N und Ladungsschwerpunkt $\bar{x}_{ni}$ sind durch Integrale bestimmt:

$$\Delta Q_N = \int_0^{x_{ni}} \Delta\rho_N(x)\, dx \tag{3.2}$$

$$\bar{x}_{ni} = \frac{\int_0^{x_{ni}} x\rho_N(x)\, dx}{Q_N}. \tag{3.3}$$

Dadurch, dass das Integral von Ladungsänderung und -schwerpunkt die Spannungsänderung bestimmt, gibt es keinen direkten Zusammenhang zwischen den

Größen. Eine Möglichkeit, trotzdem eine Betrachtung der Ladungsverteilung und Dichte vorzunehmen, ist die Anwendung von Simulationen. Durch Libsch [34] wurde ein Modell mit amphoteren Haftstellen vorgestellt, mit dem man gut das Programmier- und Löschverhalten einer SONOS-Struktur beschreiben kann. Hierbei wird die Ladungsinjektion über Tunnelprozesse modelliert. Die Beschreibung des Ladungstransports innerhalb der Nitrid-Speicherschicht erfolgt anhand der Ladungsträgerdrift bei Sättigungsgeschwindigkeit mit Hilfe der Strom-Kontinuitätsgleichung, getrennt jeweils für Leitungs- und Valenzband. Die Ladungsspeicherung wird durch einen Shockley-Reed-Hall-Prozess modelliert, wobei die Ladungsträgerdichten in Leitungs- und Valenzband berücksichtigt werden. Die verwendeten Simulationsparameter sind in Tab. 3.1 aufgeführt und entsprechen den in [72] verwendeten Werten. Die elektrischen Daten anhand derer ein Vergleich zwischen Simulation und Messung erfolgt, wurden stets an großen 5x5 μm Speicher-Transistoren gemessen. Dadurch kann ein Einfluss der Effekte, die durch die Struktur selbst bedingt sind, nahezu vernachlässigt werden.

Tabelle 3.1: Übersicht über die in den Simulationen verwendeten Parameter

Schicht	Parameter	Wert
Tunneloxid, Topoxid, SiO_2	Energie Leitungsband[a] ϕ^e_{S-O}	3,1 eV
	Energie Valenzband[a] ϕ^h_{S-O}	3,8 eV
	ϵ_r	3,9
	effektive Elektronenmasse m'	0,23 m_e
	effektive Löchermasse m'	0,6 m_e
Speicherschicht, Si_3N_4	Energie Leitungsband[a]	3,1 eV
	ϵ_r	7,5
	effektive Elektronenmasse m'	0,42 m_e
	effektive Löchermasse m'	0,42 m_e
	Einfangquerschnitt Elektronen, unbesetzt σ^0_n	1e^{-15} cm^2
	Einfangquerschnitt Elektronen, einfach besetzt σ^+_n	1e^{-16} cm^2
	Einfangquerschnitt Löcher, unbesetzt σ^0_p	1e^{-13} cm^2
	Einfangquerschnitt Löcher, einfach besetzt σ^-_p	1e^{-14} cm^2
	Haftstellendichte N_T	2,7e^{19} cm^3
Topoxid, Al_2O_3	Energie Leitungsband[a]	2,6 eV
	Energie Valenzband[a]	4,3 eV
	ϵ_r	10
	effektive Elektronenmasse m'	0,23 m_e
	effektive Löchermasse m'	0,6 m_e
Gateelektrode	n^+-poly Silizium Dotierung	1e^{20} cm^3
	Austrittsarbeit TaN	4,5 eV

[a] relativ zum entsprechenden Band in Silizium

3.2.1 Berücksichtigung der Tunneloxid-Siliziumnitrid-Grenzfläche

Bei dem Vergleich von Messung und Simulation hat sich stets gezeigt, dass bei den berechneten Programmierkurven Abschnitte unterschiedlicher Programmiersteigung existieren, wie Abb. 3.3a veranschaulicht. Die Simulation wurde in dem untersuchten Fall an den Beginn der gemessenen Programmierkurve angepasst. Die gemessene Kurve besitzt eine konstante Steigung von circa 1.5 V/dec. Die Simulation hingegen zeigt ein dreistufiges Verhalten [72]. Zu Beginn des Vergleichs in Abb. 3.3a ist sowohl bei der Messung als auch bei der Simulation eine vergleichbare Steigung zu beobachten. Kurz nach dem Einsetzen des Programmiervorgangs ändert sich die Steigung aber auf einen Wert von 2.4 V/dec, um dann ab einer Pulsdauer von 5 ms wieder auf den ursprünglichen Wert von 1.5 V/dec zurückzugehen.

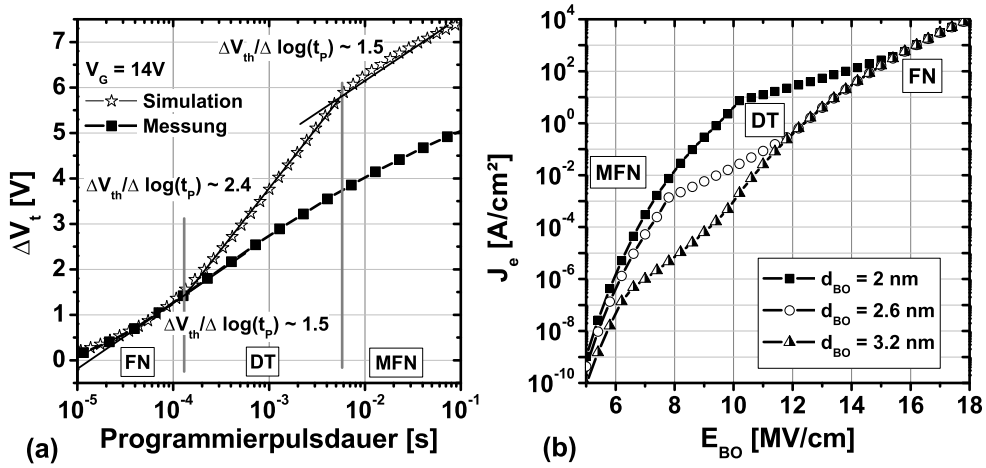

Abbildung 3.3: (a) Vergleich von Simulation und Messung für eine TANOS-Struktur mit ONA = 3.1/9/12 nm; (b) Abhängigkeit der Stromdichte vom Feld über dem Tunneloxid für die auftretenden Tunnelmechanismen

Diese Beobachtung kann damit erklärt werden, dass die Injektion der Elektronen durch verschiedene Tunnelmechanismen bestimmt ist. Abbildung 3.3b zeigt die Tunnelstromdichte in Abhängigkeit der Feldstärke, wie sie in der Simulation angenommen wird. Deutlich ist die größere Feldabhängigkeit der Fowler-Nordheim-Mechanismen (MFN, FN) im Vergleich zum direkten Tunneln (DT) zu erkennen. Dieses feldabhängige Verhalten wirkt dann direkt auf den Kurvenverlauf der Programmierkurve und resultiert in der beobachteten dreistufigen Charakteristik. Wichtig hierbei ist, dass das Verhalten nur bei Schichtstapeln beobachtet wird, die ein Tunneloxid mit einer Dicke kleiner 4 nm besitzen. Ist die Tunneloxiddicke größer, kommt es in den relevanten Spannungsbereichen nicht zum Übergang in den Bereich des direkten Tunnelns. In den Simulationen werden analytische Gleichungen zur Beschreibung des Tunnelvorgangs verwendet [24]. Diese haben den Vorteil einfacher Implementierung und hoher Rechengeschwindigkeit. Nachteilig ist aber der nicht stetige Verlauf der Kurven, da es sich nur um Näherungen handelt. Durch die Erkenntnisse aus den Simulationen war es aber möglich zu erkennen, dass die erwartete Steigung des direkten Tunnelns von 2.4 V/dec bei den Messungen nicht beobachtet werden kann. Eine mögliche Erklärung für das Verhalten ist der nicht ideale Verlauf des Leitungsbands an der Grenzfläche zwischen Tunneloxid und der Siliziumnitrid-Speicherschicht. Für direktes Tunneln wird angenommen, dass es sich

um einen abrupten Übergang handelt. In der Realität kommt es aber zu einer leichten Durchmischung der Schichten, woraus geschlussfolgert werden kann, dass das gering feldabhängige direkte Tunneln nur einen kleinen Beitrag leistet. In Abb. 3.4a zeigt die schwarze Kurve den theoretisch richtigen Verlauf des Leitungsbandes für den untersuchten Schichtstapel. Die eingezeichnete graue Linie deutet den erwarteten Bandverlauf an, bei dem es keinen abrupten, sondern einen kontinuierlichen Übergang des Leitungsbandes zwischen den verschiedenen Schichten gibt.

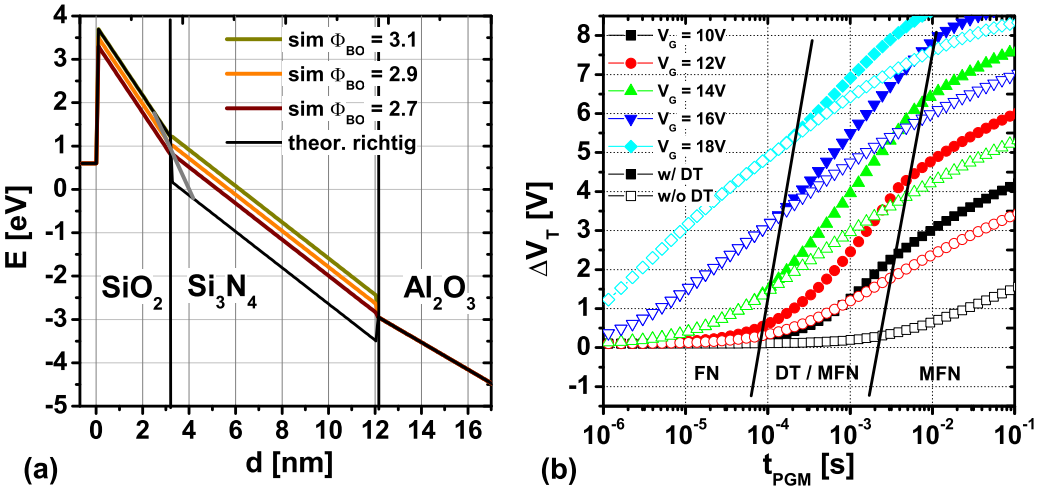

Abbildung 3.4: (a) Bandverlauf für eine theoretisch korrekte Simulation (schwarz) und bei Unterdrückung des direkten Tunnelns (farbig); die graue Linie gibt den Verlauf für eine Durchmischung der Schichten an; (b) Simulationsergebnisse mit und ohne unterdrücktem direkten Tunneln für eine TANOS-Struktur mit ONA = 3.1/9/12 nm für 5 versch. Programmierspannungen

Dieser Verlauf lässt sich aber nicht mehr mit den verwendeten analytischen Gleichungen beschreiben. Daher wurde eine physikalisch nicht korrekte Modifikation am Bandverlauf vorgenommen, um dieses Verhalten nachzuvollziehen. Um das direkte Tunneln zu unterdrücken, wurde das Leitungsband des Si_3N_4 an das Leitungsband von SiO_2 angepasst. Dies ist durch die farbigen Graphen in Abb. 3.4a wiedergegeben. Eine Gegenüberstellung der Simulationsergebnisse für die beiden Fälle '*direktes Tunneln wird unterdrückt*' und '*direktes Tunneln wird berücksichtigt*' erfolgt in Abb. 3.4b. Deutlich ist der dreistufige Kurvenverlauf bei Berücksichtigung des direkten Tunnelns auch bei verschiedenen Programmierspannungen zu sehen. Im Gegensatz dazu ist ein kontinuierlicher Kurvenverlauf bei den Simulationsergebnissen mit stufenloser Grenzfläche zu beobachten, so wie er auch bei Messungen beobachtet wird. Zum besseren Verständnis ist der jeweils dominierende Tunnelprozess mit eingetragen. Zu Beginn des Programmierens liegt bei allen gezeigten Spannungen FN-Tunneln vor. Durch die injizierte Ladung nimmt das Feld im Tunneloxid ab. Dadurch kommt es zum Übergang zu direktem Tunneln oder MFN-Tunneln, je nachdem ob die Grenzschicht berücksichtigt wurde oder nicht. Bei langen Programmierzeiten (> 10 ms) ist das Feld im Tunneloxid so klein, dass nur noch MFN-Tunneln auftritt. Ein direkter Vergleich zwischen Simulation mit Berücksichtigung der Grenzfläche und Messung erfolgt in Abb. 3.5. Der Vergleich zeigt deutlich, zu welch einer Verbesserung die Berücksichtigung des unterdrückten direkten Tunnelns führt. Die simulierten Kurven stimmen nahezu perfekt mit den gemessenen Kurven überein.

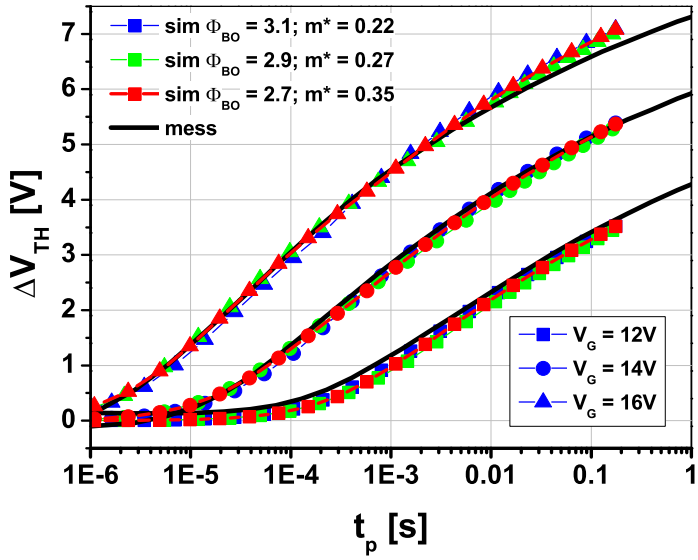

Abbildung 3.5: Vergleich von Simulation und Messung für eine TANOS-Struktur mit ONA = 3.1/9/12 nm, Bandverlauf entsprechend Abb. 3.4

Weiterhin wurde untersucht, welchen Einfluss die energetische Lage des Leitungsbandes hat. Dies beruht auf der Tatsache, dass durch den kontinuierlichen Übergang die wirksame Tunnelbarriere um einen kleinen Beitrag reduziert wird. Der Bandverlauf für verschiedene Barrierehöhen ist durch die farbigen Kurven in Abb. 3.4a gezeigt. Damit die Programmierkurven, welche in Abb. 3.5 dargestellt sind, den Messungen entsprechen, musste die effektive Elektronenmasse im Tunneloxid angepasst werden. Es wird veranschaulicht, dass die Verringerung der Barrierenhöhe eine Erhöhung der effektiven Elektronenmasse erfordert. Wird eine Barrierenhöhe Φ_e^B von 2.9 eV gewählt, stimmt die effektive Masse nahezu mit dem theoretisch bestimmten Wert von $m_e^* = 0.3$ überein. Daher entspricht die Reduktion der Tunnelbarriere durch den kontinuierlichen Übergang für die untersuchte Probe ungefähr einer Verringerung der Oxid-Barrierenhöhe um 0.2 eV vom theoretischen Wert von 3.1 eV.

3.2.2 Berücksichtigung des Injektionspunktes bei modifiziertem FN-Tunneln

Eine weitere Verbesserung bei der Übereinstimmung zwischen Simulation und Messung konnte erzielt werden, wenn der Injektionspunkt bei modifiziertem FN Tunneln berücksichtigt wird. Die erste Anwendung des verwendeten Modells erfolgte an Proben mit dünnem Tunneloxid (≈ 2 nm) und niedrigen Programmierspannungen [34]. In diesem Fall kommt es selbst nach langem Programmieren nur zum Übergang von FN-Tunneln zu direktem Tunneln, nicht aber zu modifiziertem FN-Tunneln. Die weiteren untersuchten Proben hatten Tunneloxid-Dicken von 3 und mehr nm. Dabei stellt sich nach einer von der Programmierspannung abhängigen Zeit modifiziertes FN-Tunneln ein, wie im vorangegangenen Kapitel gezeigt. Ein wichtiger Punkt ist bei der Betrachtung, dass sich die Position an dem die Ladungsträger in das Leitungsband des Nitrids tunneln, nicht mehr an der Grenzfläche zum Tunneloxid befindet. Dieser Schnittpunkt aus Tunnelstrecke und Nitrid-Leitungsband befindet sich innerhalb der Schicht, wie in Abb. 3.6a veranschaulicht. Bereits Imanaga und Aozasa [73] zeigten,

dass diese Tiefenbetrachtung in den Simulationen berücksichtigt wurde. Allerdings untersuchten sie nicht, inwieweit sich dadurch eine Verbesserung der Simulationsergebnisse ergibt.

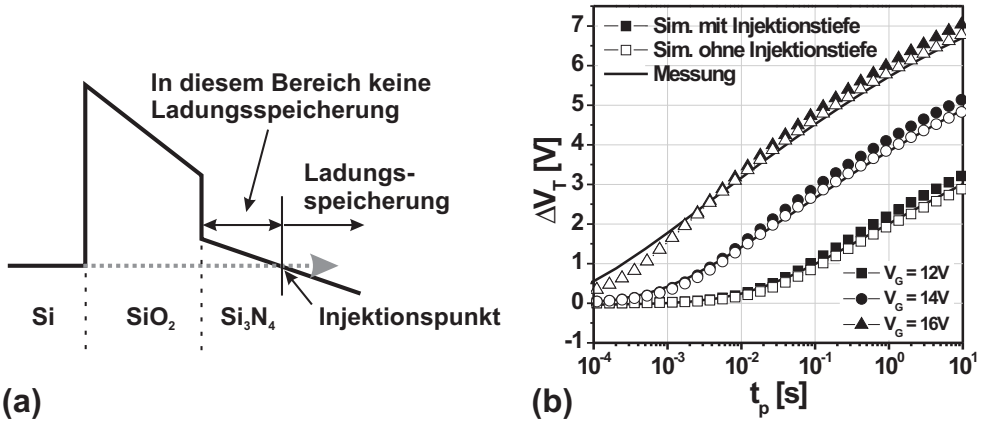

Abbildung 3.6: (a) Erklärung anhand des Bändermodells, welcher Bereich der Speicherschicht bei MFN nicht bzw. beladen wird ; (b) Simulation mit und ohne Injektionstiefe im Vergleich zur Messung; ONO = 3.1/9/9 nm

Eine Herausforderung für die Simulation stellt im Fall des modifiziertem FN-Tunnelns der Speicherschicht-Bereich zwischen Tunneloxid und Injektionspunkt dar. Dieses Gebiet wird durchtunnelt und es erfolgt keine Ladungsspeicherung, wie in Abb. 3.6a gezeigt. Gelangen zum Beispiel beim Programmieren die Ladungsträger im Valenzband in diesen Bereich, kann es zu einer unkontrollierten Akkumulation und Simulationsinstabilität kommen. In Abb. 3.6b ist ein Vergleich durchgeführt, welcher zeigt, wie sich die Berücksichtigung des Injektionspunktes auswirkt. Wird in der Simulation angenommen, dass die injizierten Ladungsträger immer an der $SiO_2 - Si_3N_4$-Grenzfläche in das Leitungsband gelangen, zeigt sich eine größere Abweichung von der Messkurve. Daher kann geschlussfolgert werden, dass die Simulation eine weitere Verbesserung erfährt, wenn der theoretisch erwartete Injektionspunkt in den Simulationen berücksichtigt wird. Es ist auch zu erkennen, dass erst mit dem Wechsel auf den modifizierten FN-Tunnelprozess eine korrekte Angleichung an die Messkurven erfolgt. Die Berücksichtigung des Injektionspunktes hat aber auch einen Einfluss auf die Ladungsverteilung, da sich bei unterschiedlichen Spannungen und Ladungszuständen auch unterschiedliche Positionen für den Ladungseintritts in das Nitrid-Leitungsband ergeben. In Abb. 3.7a ist gezeigt, wie sich die Elektronendichte über die Zeit entwickelt. Bei der gewählten Spannung von 14 V ist bereits zu Beginn des Beladungsvorganges der Injektionspunkt innerhalb der Nitridschicht. Demnach wird der Bereich der Speicherschicht bis zu einer Tiefe von 1 nm nicht mit Elektronen beladen. Mit zunehmender Zeit wird weiter Ladung gespeichert und die Elektronendichte nimmt zu. Damit einher geht eine durch die Ladung induzierte Feldreduktion auf der Tunneloxidseite. Diese Veränderung des elektrischen Feldes resultiert in einer Verschiebung der Bandstruktur und folglich der Lage des Elektronen-Injektionspunktes. Dieser wandert durch das abnehmende Feld immer tiefer in die Speicherschicht. Dadurch wird nur noch der Bereich dahinter weiter beladen und die Ladungsverteilung davor bleibt unverändert. Betrachtet man die Beladung zu unterschiedlichen Zeitpunkten während des Programmierens, wie sie in Abb. 3.7a erfolgt, wird dieses Verhalten noch einmal deutlich.

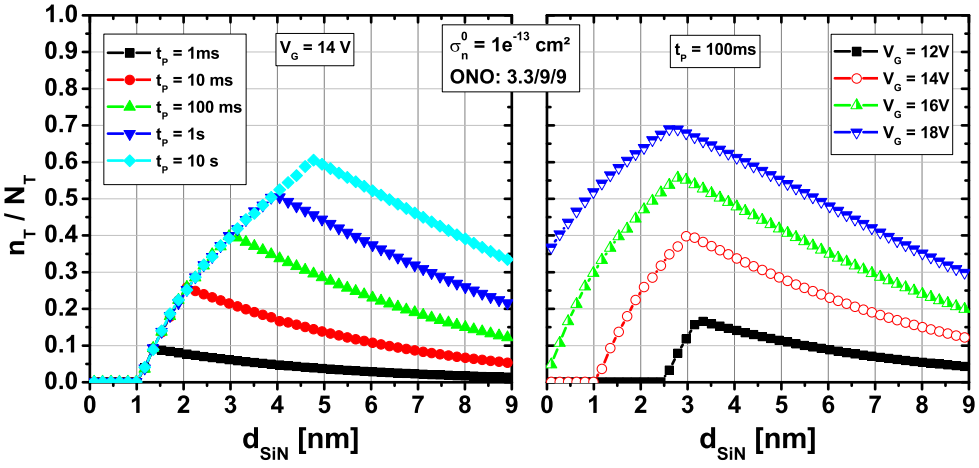

Abbildung 3.7: (a) Ladungsverteilung bei einer festen Spannung und zunehmender Pulszeit (b) Verhalten bei konstanter Pulszeit und zunehmender Spannung

Demnach entsteht eine einhüllende Kurve für die Elektronendichte, von der mit zunehmender Pulsdauer ein immer größer werdender Teil beladen wird. Die Form dieser Einhüllenden hängt wiederum vom Feld und somit der angelegten Spannung ab. Betrachtet man die Ladungsverteilung für die Beispiel SONOS-Struktur bei einer festen Pulszeit ergibt sich ein anderes Bild, wie in Abb. 3.7b gezeigt. Mit zunehmender Programmierspannung verschiebt sich der initiale Injektionspunkt und die Hüllkurve wandert weiter in Richtung Tunneloxid. Bei einer Spannung von 16 V erfolgt eine Änderung des Ablaufs der Tunnelmechanismen und demzufolge auch der Hüllkurve. Ab dieser Spannung wird noch der Übergang von FN-Tunneln zu modifiziertem FN-Tunneln beobachtet. Ist die Spannung kleiner, tritt kein direktes Tunneln mehr auf und der gesamte Programmiervorgang erfolgt durch MFN-Tunneln. Allerdings erfolgt der Übergang DT zu MFN schon nach der Injektion einer kleinen Ladungsmenge und die Hüllkurve ist denen bei kleineren Spannungen ähnlich. Die Hüllkurve ähnelt unter diesen Bedingungen einem Dreieck. Ist die Programmierspannung größer als 16 V, wird eine erhebliche Ladungsmenge nah am Tunneloxid gespeichert und die Dreiecks-Form der Hüllkurve wird auf dieser Seite 'abgeschnitten'. Die unterschiedliche Ladungsmenge, welche bei verschiedenen Spannungen gespeichert wurde, bedingt, dass der Injektionspunkt zu einem festen Zeitpunkt in eine vergleichbaren Tiefe der Speicherschicht verschoben wurde.
Mit Hilfe dieser Analyse ist es auch möglich, die unterschiedlichen Programmiermodi ISPP (Kap. 2.5.1) und transiente Programmierung (Kap. 2.5.2) zu vergleichen. Deutlich ist der Unterschied bei der Entwicklung der Ladungsverteilung zwischen den Programmiermodi zu erkennen. Es ist daher zu erwarten, dass die unterschiedlichen Ladungsverteilungen für transientes Programmieren und ISPP zu unterschiedlichem elektrischem Verhalten in Abhängigkeit von variierender Pulsdauer bzw. Pulsspannung führen.

3.2.3 Einfluss des Ladungsträger-Einfangquerschnittes auf die Programmiersteigung

Bereits in der Arbeit von Furnemont [74] ist aufgefallen, dass die simulierten Programmierkurven unterschiedlicher Programmierspannung weiter auseinander liegen, als die gemessenen Kurven. Verdeutlicht wird dieser Effekt durch die Pfeile in Abb. 3.8, wobei ein Vergleich von Simulation und Messung für einen SONOS-Schichtstapel gezeigt wird. Es wurden für die Simulation in Abb. 3.8 jeweils die Parameter für das Speichernitrid aus [75] genommen. Eine Untersuchung von Furnemont [74] befasst sich mit diesem Effekt.

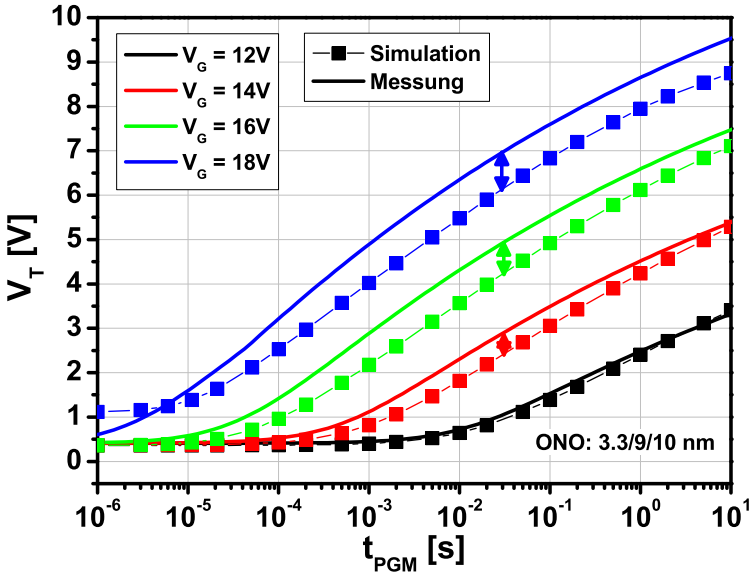

Abbildung 3.8: Vergleich von Simulation und Messung einer SONOS-Struktur für vier versch. Programmierspannungen; Abstand zwischen Simulation und Messung ist durch Pfeile angedeutet; Simulation an $V_G = 12$ V angepasst

Hierbei wurde gezeigt, dass eine bessere Übereinstimmung von Simulation und Messung erzielt werden kann, wenn ein Teil der injizierten Ladung nicht gespeichert wird und diese die Speicherschicht durch das Topoxid wieder verlässt (engl. fly-through). Bei dem verwendeten Simulationsmodell kann die Wahrscheinlichkeit, dass ein Ladungsträger gespeichert wird, über den Einfangquerschnitt gesteuert werden. Die Programmiersteigung s ist der Parameter, welcher angibt, wie groß der Abstand zwischen den Kurven unterschiedlicher Programmierspannung ist. Aus diesem Grund wurde der Einfangquerschnitt in den Simulationen variiert. Gleichzeitig wurde analysiert, wie sich die Programmiersteigung verhält. Die Simulationsergebnisse in Abb. 3.9 zeigen, dass die Programmiersteigung s mit größer werdendem Einfangquerschnitt σ_n^0 zunimmt.
Es wird bei einer Größe von $\sigma_n^0 = 1e^{-13}\,cm^2$ die maximal mögliche Steigung von $s = 1\,V/V$ erreicht. Dies ist auch der Wert, der in den Publikationen angegeben wird, welche sich nur mit dünnen Schichtstapeln befassen [34, 73]. Allerdings wird bei der Anwendung dieses Wertes schon bei einer Programmierspannung von 18 V eine Abweichung von mehr als 1 V zur Messung erreicht, wie Abb. 3.9 verdeutlicht. Für den Vergleich wurde ein TANOS-Stapel mit einem 3.1 nm dicken Tunneloxid gewählt, um eher den Anforderungen für Flash-Speicher gerecht zu werden. Die Simulation

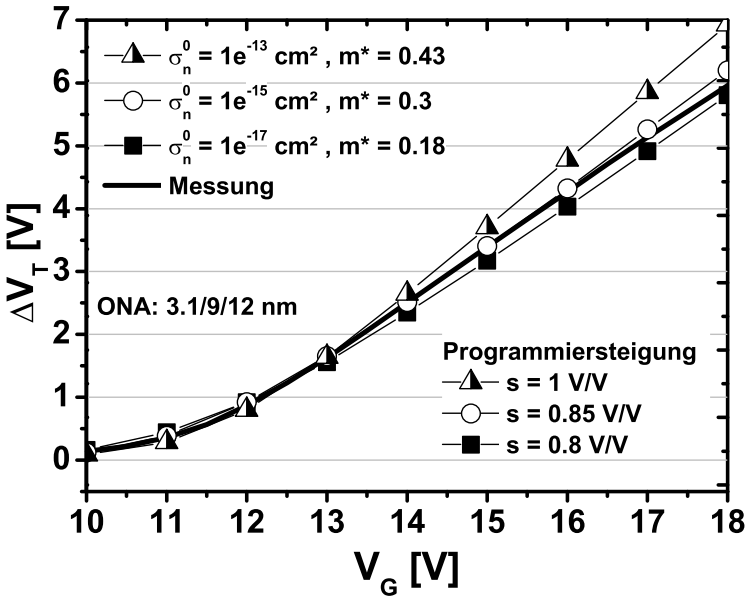

Abbildung 3.9: Untersuchung der Programmiersteigung in Abhängigkeit des Ladungsträger-Einfangquerschnitts und Vergleich mit der Messung; $t_P = 100\ \mu$s

erfolgte unter Berücksichtigung der Anpassung der Simulation auf das Einsetzen des Programmierens. Hierzu ist es notwendig, die effektive Masse so zu variieren, dass der Tunnelstrom auf der Injektionsseite so groß ist, dass die Programmierkurven vergleichbar sind. Die jeweils verwendeten effektiven Massen sind in Abb. 3.9 angegeben. Verringert man nun den Einfangquerschnitt, nähert sich die Simulationskurve der gemessenen Kurve an, wobei ein immer kleiner werdender Anteil der injizierten Ladung gespeichert wird. Eine genauere Betrachtung des Verhältnisses von ein- und ausströmenden Ladungsträgern erfolgt im folgenden Kapitel. Die nicht gespeicherte Ladung verlässt die Speicherschicht über das Topoxid, wobei eine Berücksichtigung der Tunnelbarriere in dem Modell nicht erfolgt. In dem gegebenen Beispiel ist die größte Übereinstimmung für einen Wert von $\sigma_n^0 = 1e^{-15}\,cm^2$ abzulesen. Daher wurde dieser Wert auch für alle weiteren durchgeführten Simulationen verwendet. Dies gilt auch für die Simulationen in den vorangegangenen Kapiteln. Wie gut die Simulation mit der Messung übereinstimmt, wurde bereits in Abb. 3.5 gezeigt, wobei in den Simulationen der genannte Einfangquerschnitt von $\sigma_n^0 = 1e^{-15}\,cm^2$ verwendet wurde.

3.2.4 Einfluss des Ladungsträger-Einfangquerschnittes auf die Ladungsverteilung

Die Variation des Einfangquerschnittes hat auch einen Einfluss auf die Ladungsverteilung. Bei dem verwendeten Modell wird der Elektronenstrom $J_{n,in}$, welcher in ein Raumelement der Simulation fließt, in zwei Größen aufgespalten.

$$J_{n,in} = J_{n,aus} + J_{n,element} \tag{3.4}$$

Der Strom $J_{n,aus}$ beschreibt die Ladungsträger, die nicht gespeichert werden und in das nächste Raumelement fließen. Die zweite Komponente ist der Strom $J_{n,element}$, welcher die Zu- und Abnahme der Elektronen in den Haftstellen und im Leitungsband beschreibt. Das Gleiche gilt respektive für die Löcher im Valenzband, es wird

der Einfachheit wegen aber nur das Leitungsband betrachtet. In Gl. 3.5 ist die Beschreibung für die Berechnung der Kontinuitätsgleichung gezeigt. n_c beschreibt die Ladungsträgerdichte im Leitungsband, N_T gibt die Haftstellendichte wieder und V_D ist die Driftgeschwindigkeit der Elektronen.

$$J_{n,element} = \frac{\delta J_n(x,t)}{\delta x} = q\left[\frac{\delta}{\delta t} n_c(x,t) + N_T(x)\,\alpha_T(x,t)\,v_D n_c(x,t)\right] \tag{3.5}$$

$$\alpha_T(x,t) = \sigma_n^0 - \sigma_n^0 f^- + \left(\sigma_n^+ - \sigma_n^0\right) f^+ \tag{3.6}$$

Betrachtet man nun den Einfluss des Einfangquerschitts σ_n^0 auf die Größe α_T, so zeigt sich eine lineare Abhängigkeit beider Parameter. Nimmt σ_n^0 zu, ergibt sich auch eine Vergrößerung von α_T. Diese wiederum bewirkt in Gl. 3.5 eine Vergrößerung des zweiten Summanden in der Klammer. Schlussendlich ergibt sich eine Erhöhung der Ladungsmenge in dem Raumelement, welche zum größten Teil in den Haftstellen gespeichert wird. Dieses Ergebnis zeigt sich auch in den simulierten Ladungsverteilungen von Abb. 3.10. Bei einem großen Einfangquerschnitt wird relativ viel Ladung gespeichert und der Strom im Leitungsband nimmt von Raumelement zu Raumelement stark ab.

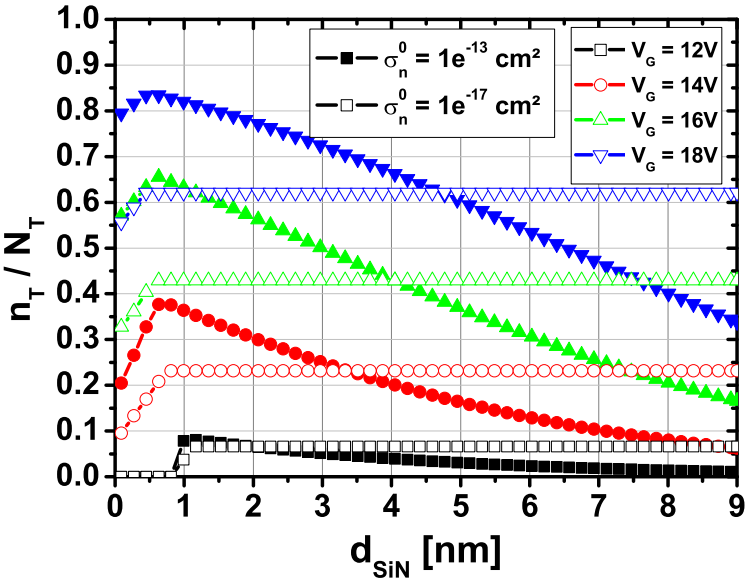

Abbildung 3.10: Vergleich der Ladungsverteilung nach einem Programmierpuls t_P = 100 ms bei der Simulation mit untersch. Einfangquerschnitten σ_n^0 ; ONA = 3.1/9/14 nm

Dies resultiert in einer starken Abnahme der besetzten Haftstellen, je weiter man in Richtung Topoxid gelangt. Denn die Ladungsmenge, welche aus dem Leitungsband in die Haftstellen abgezweigt wird, ist durch den Strom, und demzufolge durch die Ladungsträgerdichte n_c im Leitungsband, bestimmt. Wird im umgekehrten Fall nur eine kleine Ladungsmenge abgezweigt, verändert sich die Stromdichte im Leitungsband über die Dicke des Nitrides nur wenig. Dadurch bedingt sich bei kleinem Einfangquerschnitt nur ein geringer Unterschied der gespeicherten Ladung über die Dicke des Nitrids. Daraus resultiert ein flacher, nahezu konstanter Verlauf der Ladungsdichte, wie in Abb. 3.10 durch die Kurven mit den offenen Symbolen gezeigt wird.

Hierbei entsteht aber ein weiterer Punkt, dem Beachtung geschenkt werden muss. Wird ein kleiner Querschnitt gewählt, kommt es nur zu einer geringen Abnahme des injizierten Stromes über das Nitrid. Ein großer Teil der Ladung gelangt an die Grenzfläche zum Topoxid. In dem verwendeten Modell kann diese Ladung direkt abfließen, entsprechend dem „flythrough"-Modell nach [74]. Demzufolge fließt ein großer Teil der Ladung durch das Topoxid ab und trägt nicht zum Verhalten der Speicherzelle bei. In Abb. 3.11 wird veranschaulicht, wie sich der Einfangquerschnitt auf die Speichereffizienz η_s in Gl. 3.7 auswirkt.

$$\eta_s = 1 - \frac{J_{BO}}{J_{TO}} \tag{3.7}$$

Die Speichereffizienz beschreibt, mit welcher Wahrscheinlichkeit ein Elektron, dass durch das Tunneloxid in die Speicherschicht tunnelt, gespeichert wird.

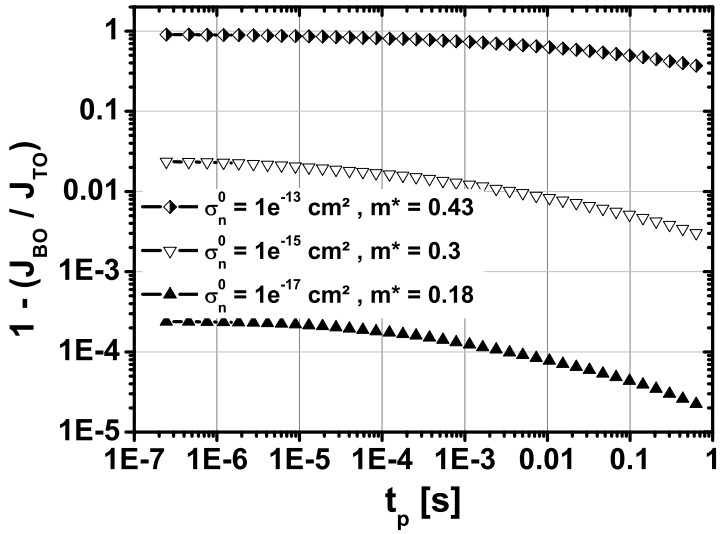

Abbildung 3.11: Speichereffizienz η_s während des Programmierens für drei versch. Einfangquerschnitte σ_n^0

Bei einem Einfangquerschnitt von $\sigma_n^0 = 1e^{-13}\, cm^{-2}$ wird zu Beginn des Programmiervorgangs nahezu jedes injizierte Elektron gespeichert, was durch einen Wert von nahezu 1 für η_s angezeigt ist. Mit zunehmender Programmierpulsdauer wird immer mehr Ladung in der Speicherschicht gespeichert, was dazu führt, dass der Einfang weiterer Ladung abnimmt. Der Einfluss durch die Ladungsspeicherung führt zu einer Reduktion der Wahrscheinlichkeit am Ende des Programmiervorgangs auf circa 1/10 der ursprünglichen Einfangwahrscheinlichkeit. Die Veränderung des Einfangquerschnitts um den Faktor 100 zeigt, dass eine Vergrößerung in einer Zunahme der Speichereffizienz η_s von circa 120 resultiert. Betrachtet man den vorgeschlagenen Wert für σ_n^0 von $1e^{-15}\, cm^{-2}$, so wird deutlich, dass zu Beginn des Programmiervorgangs nur jedes 40. Elektron gespeichert wird. Das bedeutet, dass 98% der injizierten Elektronen durch den kompletten Schichtstapel hindurchgehen und keinen Beitrag zur Schwellspannungsverschiebung leisten.

3.3 Ableitung der vertikalen Ladungsverteilung aus Messungen

Für die Betrachtung der vertikalen Ladungsverteilung ist es nicht nur möglich Simulationen zu verwenden, sondern man kann auch bei geeigneter Wahl von Proben und Messung eine qualitative Abschätzung des Profils vornehmen. Im ersten Abschnitt wird anhand von SONOS/TANOS-Strukturen mit unterschiedlicher Speicherschichtdicke eine Herleitung der Ladungsverteilung vorgenommen. Im zweiten Abschnitt wird mit Hilfe eines speziellen Auswerteverfahrens für die Langzeit-V_T-Verschiebung, eine Ableitung, gültig für alle haftstellenbasierten Speicherzellen, durchgeführt.

3.3.1 Variation der Speicherschichtdicke bei SONOS

Bei der Untersuchung des Ladungsverlustes von SONOS-Schichtstapeln ist aufgefallen, dass sich ein unterschiedliches Verhalten ergibt, wenn man mit unterschiedlicher Polarität auf die gleiche Schwellspannung programmiert. Die Besonderheit bei SONOS-Strukturen, dass sie nur bedingt bzw. gar nicht löschbar sind, wird hierbei ausgenutzt. In Abb. 3.12 ist die Flachbandspannungsverschiebung gezeigt, welche sich für das Programmieren mit positiver und negativer Gatespannung ergibt.

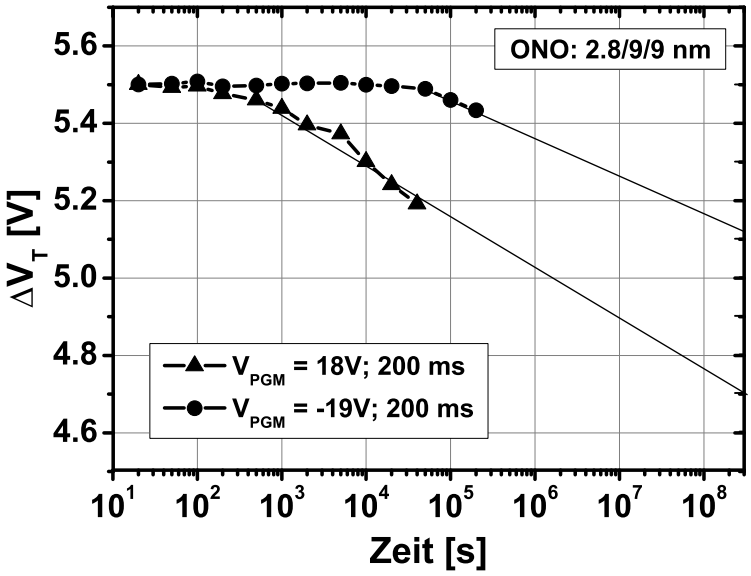

Abbildung 3.12: Vergleich der Ladungshaltung eines SONOS-Schichtstapels nach dem Programmieren mit einer positiven (Dreiecke) oder negativen (Kreise) Gatespannung; ONO = 2.8/9/9 nm

Interessanterweise zeigt sich, dass ein Programmieren mit negativer Gatespannung in einer deutlichen Verschiebung des Beginns der Flachbandspannungsverschiebung resultiert. Für die untersuchte Probe ergibt sich eine Verschiebung um mehr als zwei Zeitdekaden. Dieses Verhalten kann damit begründet werden, dass durch die unterschiedlichen Polaritäten, die Ladungsträger einmal von der Substrat- und das andere Mal von der Gateseite injiziert werden. Dies ist möglich, da ähnliche Tunnelbarrieren bei den symmetrischen Schichten der SONOS-Struktur vorliegen. Der variierende Ladungsverlust kann nun damit begründet werden, dass sich unterschiedliche Ladungsverteilungen ausbilden. Werden mit positiver Gatespannung Elektronen vom Substrat

in die Speicherschicht injiziert, ergibt sich ein dem Tunneloxid nahes Ladungsprofil, wie in Abb. 3.13a schematisch gezeigt. Eine solche Ladungsverteilung würde sich unter der Annahme ausbilden, dass die Ladung mit großer Wahrscheinlichkeit von Haftstellen eingefangen wird. Dies entspricht einem großen Einfangquerschnitt bei den in dem vorangegangenen Kapitel durchgeführten Simulationen. Erfolgt nun die Injektion von der Gateelektrode, bildet sich eine an die Topoxidseite gespiegelte Verteilung aus, welche in Abb. 3.13b dargestellt ist. Für den Ladungsverlust der untersuchten SONOS-Struktur ist maßgeblich die Komponente durch das Tunneloxid verantwortlich. Aufgrund der großen Dicke des Topoxids kann davon ausgegangen werden, dass der Ladungsverlust in diese Richtung vernachlässigbar ist.

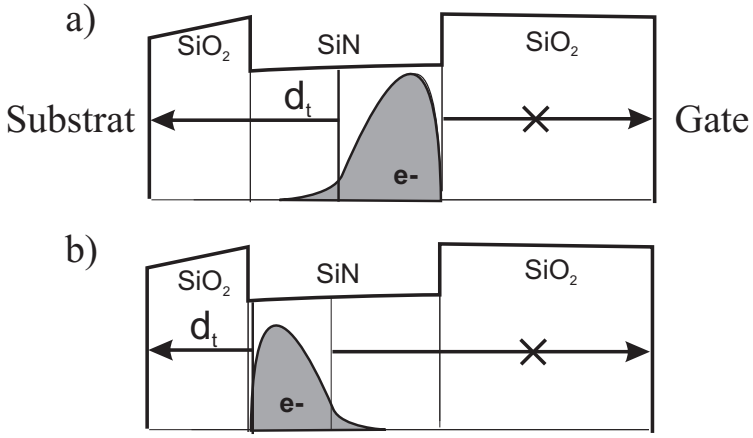

Abbildung 3.13: schematische Darstellung der Ladungsverteilung in der Speicherschicht nach dem Programmieren mit positiver Spannung (a) und negativer Spannung (b); es variiert die Tunneldistanz d_t

Betrachtet man nun die Tunnelstrecke d_t der injizierten Ladungsträger, so ist offensichtlich, dass die Tunnelstrecke für das Programmieren mit negativer Gatespannung für die gegebenen Ladungsverteilungen deutlich größer ist. In Gl. 3.8 ist die Berechnung der Tunnelwahrscheinlichkeit für das Tunneln aus einer Haftstelle durch das Tunneloxid in das Leitungsband des Substrats gezeigt [76].

$$\tau_1 = \tau_0 exp\left[2\Phi_{ox} d_{ox} + 2\Phi_N \left(d_t - d_{ox}\right)\right] \tag{3.8}$$

τ_1 ist die Gesamttunnelzeitkonstante und τ_0 beschreibt eine Zeitkonstante in der Größenordnung $10^{-12} - 10^{-14}\ s$. Beide Größen beschreiben die Häufigkeit, mit der ein Elektron versucht der Haftstelle zu entfliehen. Φ_{ox} und Φ_N beschreiben die Höhe der jeweiligen Barriere von der energetischen Lage der Haftstelle aus gesehen. Eine größere Tunnelstrecke d_t resultiert in einer größeren Tunnelzeit τ_1. Dies bedeutet eine geringere Tunnelwahrscheinlichkeit und daher einen Ladungsverlust, welcher erheblich kleiner ausfällt. Der zeitliche Unterschied ist dabei über die Differenz der beiden Tunneldistanzen d_t bestimmt. Die Annahme, dass die Ladung lokal auf der Injektionsseite gespeichert wird, kann durch eine Langzeitmessung einer Probe mit dünnerer Speicherschicht unterstützt werden. Hierzu wurde ein Vergleich der Programmierpolarität an einer SONOS-Probe mit 2.5 nm Nitridschichtdicke durchgeführt.
Die in Abb. 3.14 dargestellte Messung verdeutlicht, dass der Abstand zwischen den Messkurven unterschiedlicher Programmierpolarität kleiner geworden ist. Das Einsetzen des eigentlichen Verlustmechanismus ist auf eine Zeitdekade reduziert. Somit sind

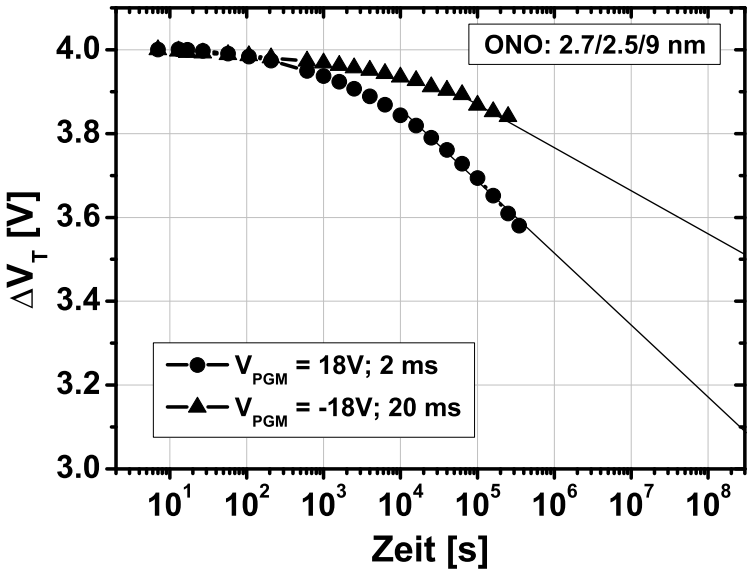

Abbildung 3.14: Messung der Ladungshaltung für eine SONOS-Struktur mit ONA = 2.7/2.5/9 nm nach dem Programmieren mit einer positiven (Kreise) und negativen (Dreiecke) Gatespannung

die beiden Ladungsverteilungen durch die geringere Dicke näher aneinander gerückt und bestätigen die vorangegangene, empirische Betrachtung. Der unterschiedliche weitere Verlauf der Kurven ist auf ein unterschiedliches Profil der vertikalen und energetischen Verteilung zurückzuführen. Eine Ableitung der vertikalen Verteilung ist jedoch nicht so einfach, da sich der Ladungsverlust aus einer Überlagerung der energetischen Lage und der Tunneldistanz der entsprechenden Ladung ergibt [77].
Die empirische Betrachtung des Ladungsverlustes zeigt bezüglich der Ladungsverteilung ein anderes Verhalten als die Simulationen in den vorangegangenen Kapiteln. Die Simulationen haben gezeigt, dass sich die beste Übereinstimmung zu den Messungen ergibt, wenn eine möglichst homogene Ladungsverteilung angenommen wird. Dieses Ergebnis steht im Widerspruch zu den Betrachtungen aus den Ladungsverlust-Messungen, die auf eine lokale Ladungsspeicherung schließen. Ursache für die Diskrepanz können die unterschiedlichen Ansätze der Extraktion sein, da bei den Simulationen die Programmiercharakteristiken und bei der empirischen Betrachtung der Ladungsverlust analysiert wurden. Da jeweils andere Mechanismen die Charakteristiken bestimmen, kann es zu dieser Differenz kommen. Weiterhin ist es Aufgabe von weiterführenden Arbeiten diesen Sachverhalt aufzuklären.

3.3.2 Variation der Speicherschichtdicke bei TANOS

Es wurde auch untersucht, inwieweit sich die Speicherschichtdicke auf das Verhalten bei TANOS-Strukturen auswirkt. Die Besonderheit im Vergleich zu SONOS ist der Einfluss durch das Al_2O_3-Topoxid. Dieses ist im Gegensatz zum SiO_2 bei SONOS aufgrund seiner Fehlstellendichte als nicht ideal isolierend anzusehen [78]. Daher ist es interessant, wie sich diese zusätzliche Leckstrom-Komponente bemerkbar macht. Für zwei unterschiedliche Tunneloxiddicken ist das Ergebnis für eine Temperung von 2 h 200°C in Abb. 3.15 gezeigt. Interessanterweise weisen die gemessenen Proben eine starke Abhängigkeit von der Speicherschichtdicke auf. So nimmt die V_T-Verschiebung

bei 5 nm Tunneloxid und einer Dickenänderung von 3 auf 9 nm Si_3N_4 von 120 auf 620 mV zu. Der Einfluss der Tunneloxiddicke ist bei 3 nm Speicherschichtdicke vernachlässigbar. Bei 9 nm ist die Verschiebung für das dünne Tunneloxid nahezu doppelt so hoch. Die starke Abhängigkeit von der Speicherschichtdicke kann auch wieder mit der in Kap. 3.3.1 gezeigten Ladungsverteilung, welche sich auf der Injektionsseite ausbildet, erklärt werden.

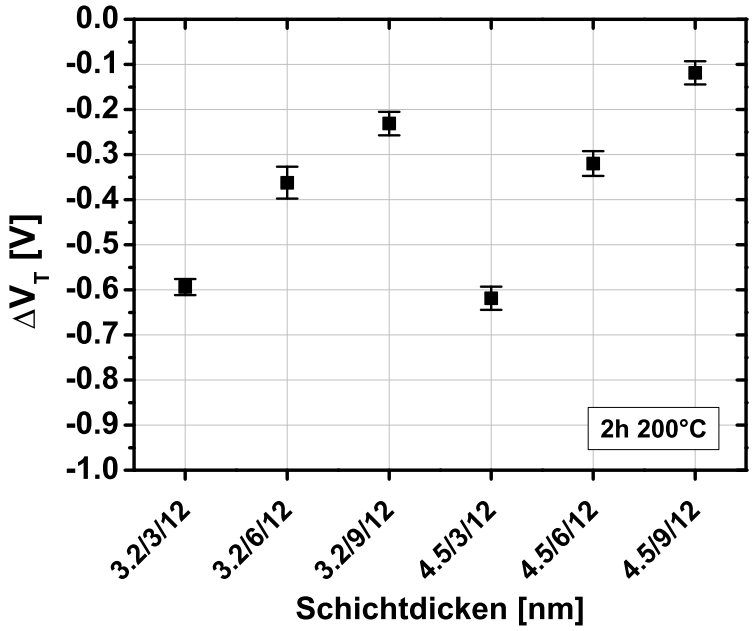

Abbildung 3.15: Messung des Ladungsverlustes nach einer Temperzeit von 2h bei 200°C für drei versch. Nitriddicken und zwei Tunneloxiddicken; $V_{T,PGM} = 5$ V

Demzufolge wird auch bei der Betrachtung der Schichtdickenabhängigkeit bei TANOS eine lokale Ladungsspeicherung auf der Injektionsseite angenommen. Wird eine dicke Speicherschicht verwendet, bildet sich eine Ladungsverteilung aus, wie sie in Abb. 3.16a illustriert ist. In diesem Fall ist die V_T-Verschiebung durch den Ladungsverlust über das Tunneloxid dominiert. Daher ist in Abb. 3.15 auch ein Unterschied bei den 9 nm Si_3N_4-Gruppen zu beobachten. Aufgrund der geringeren Dicke des Tunneloxids ist der Ladungsverlust bei der 3.2/9/12 nm Gruppe größer als bei der vergleichbaren Gruppe mit 5 nm Tunneloxid. Wird die Speicherschichtdicke auf 6 nm verringert, ist ein Profil wie in Abb. 3.16b gezeigt, zu erwarten. Die Tunneldistanz d_t wird dabei in einem Maße reduziert, dass der Einfluss des Al_2O_3-Topoxids nicht mehr vernachlässigt werden kann. Daher kommt es zu einer Angleichung der V_T-Verschiebung durch die Zunahme des gezeigten Verlustmechanismus. Die immer noch vorhandene kleine Differenz entsteht dadurch, da es sich für diese Abmessung um eine Überlagerung von Tunneln durch das Tunneloxid und Verlust durch das Topoxid handelt. Eine weitere Reduktion der Dicke auf 3 nm führt dazu, dass die Ladungsverteilung bis fast an die Topoxid-Grenzfläche reicht, wie Abb. 3.16c verdeutlicht. Dadurch dominiert der Verlust durch das Al_2O_3 und der Einfluss durch das Tunneloxid verschwindet. Unterstützt wird diese Betrachtung durch die Analyse der V_T-Verschiebung bei Raumtemperatur und verschiedenen Programmierzuständen.
Die in Abb. 3.17a gezeigte Messung der V_T-Verschiebung über der Zeit verdeutlicht die Abhängigkeit des Ladungsverlustes von der Speicherschichtdicke. Wird nun durch

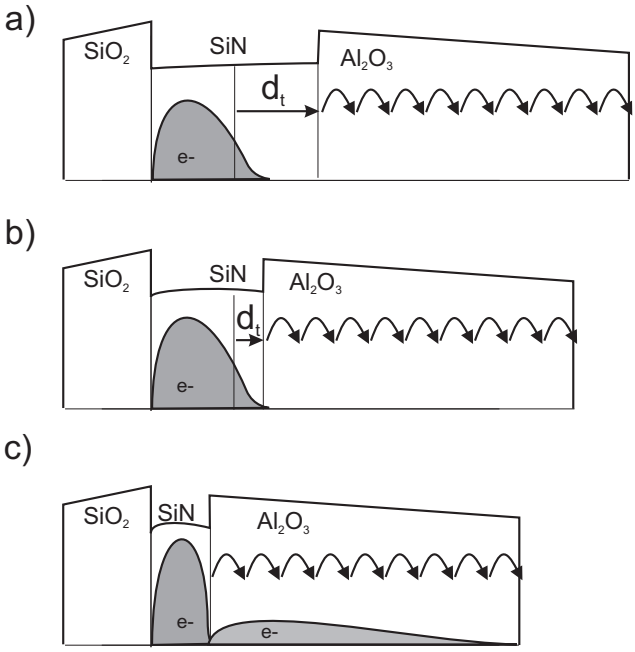

Abbildung 3.16: Qualitative Ladungsverteilung für (a) 9 nm, (b) 6 nm und (c) 3 nm Nitriddicke, es ist der Hauptleckpfad bei einer TANOS-Stuktur durch das Al_2O_3 illustriert

das Programmieren auf einen höheren Programmierzustand mehr Ladung injiziert, ergibt sich ein Bild wie in Abb. 3.17b dargestellt. Es kommt zu einer nahezu vergleichbaren V_T-Verschiebung für die beiden dünneren Speicherschichtdicken. Bezieht man das Ergebnis aus Abb. 3.17a mit ein, kommt man zu dem Schluss, dass bei einer Programmierung auf 4.5 V ΔV_T die Ladungsträgerdichte für 3 und 6 nm an der Topoxid-Grenzfläche vergleichbar ist. Somit bewirkt die Programmierung der 6 nm Gruppe auf 4.5 V ΔV_T eine deutliche Änderung des Ladungsprofils in Richtung der Topoxid-Grenzfläche.

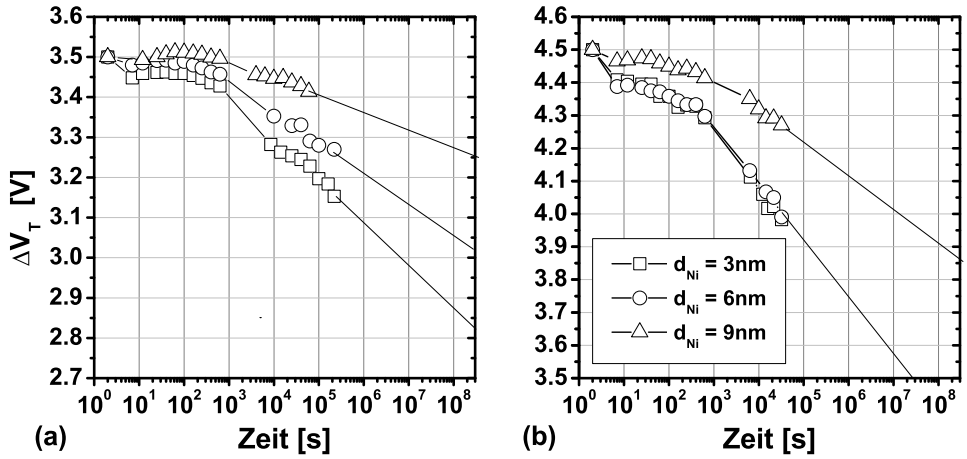

Abbildung 3.17: V_T-Verschiebung für TANOS-Zellen mit 3 Nitriddicken nach einer Programmierung auf ein (a) $\Delta V_T = 4.5$ V und (b) $\Delta V_T = 3.5$ V; $d_{BO} = 5$ nm; $d_{TO} = 12$ nm; T = 25°C

Dieses Ergebnis unterstützt wiederum die Annahme, dass es sich bei der Ladungsspeicherung um einen lokal konzentrierten Prozess handelt. Wobei sich mit zunehmender Programmierung die Ladung in Richtung Topoxid-Grenzfläche ausdehnt. Steigt die Konzentration an Ladungsträgern an der Topoxid-Grenzfläche an, entsteht ein großer Strom, welcher durch das Topoxid fließt. Daher kommt es zu einer nicht mehr vernachlässigbaren Zahl an Ladungsträgern, die im Topoxid gespeichert werden, wie in Abb. 3.16c angedeutet. Da Al_2O_3 aber nur relativ flache Haftstellen besitzt [78], kommt es zu einem schnellen Verlust der in dieser Schicht gespeicherten Ladung. Die schnelle Verschiebung zwischen dem ersten und zweiten Messpunkt in Abb. 3.17 repräsentiert die schnelle Entladung. Es zeigt sich auch, dass die dünne 3 nm Speicherschicht eine größere Stufe aufweist, als die beiden dickeren Speicherschichtgruppen, wenn man wie in Abb. 3.17b auf 3.5 V ΔV_T programmiert. Erhöht man die injizierte Ladungsmenge, wird auch der Sprung größer, wie ein Vergleich zwischen Abb. 3.17a und b verdeutlicht. Damit bestätigt sich die Annahme, dass sich die Ladungsverteilung mit zunehmender Ladungsinjektion in Richtung des Topoxids ausbreitet. Dort angekommen, findet eine teilweise Ladungsspeicherung im Al_2O_3 statt. Die Ladung befindet sich in flachen Haftstellen und entlädt sich innerhalb kurzer Zeit.

3.3.3 Betrachtung des Ladungsneutralpunktes

Die Betrachtung des Ladungsverlustes erfolgt normalerweise, indem man eine frische Speicherzelle auf einen definierten Zustand programmiert. Im Anschluss misst man nach einer möglichst zeitnahen Messung die V_T-Verschiebung bei Raumtemperatur und anschließend noch einmal nach einer Temperung von 2 h 200°C. Wird allerdings die Prozedur um einen initialen Löschvorgang erweitert, tritt eine zusätzliche Komponente auf. Diese resultiert in einer stärkeren V_T-Verschiebung, wie in Abb. 3.18 gezeigt wird.

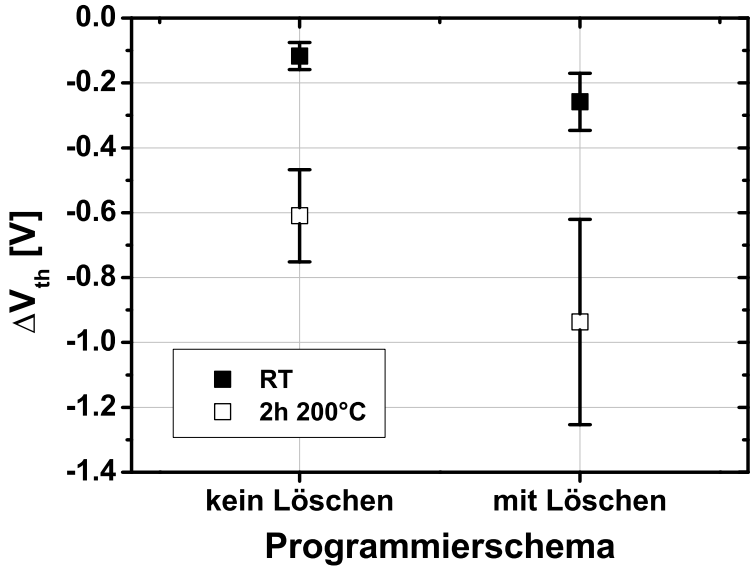

Abbildung 3.18: V_T-Verschiebung nach 2 h bei Raumtemperatur und 200°C, einmal bei einer frischen programmierten Zelle oder einer Zelle, die zuvor einmal gelöscht wurde; gemessen an 48x48 nm Speicherzellen mit TiN Gateelektrode; ONA = 5/6/12 nm

Normalerweise werden bei einer optimierten TANOS-Zelle V_T-Verschiebungen nach 2 h 200°C in der Größenordnung 600 mV gemessen, wenn sie auf ein V_T von 4 V programmiert wurde. Wird aber vor dem Programmieren die Zelle zuerst gelöscht, so ergibt sich für den gezeigten Fall in Abb. 3.18 eine Erhöhung der V_T-Verschiebung um circa 50%. Mögliche Ursachen hierfür könnten eine veränderte Ladungsverteilung oder aber auch eine Degradation des Tunneloxids während des Löschvorgangs sein [79]. Um dies zu untersuchen, wurde eine Methodik entwickelt, mit deren Hilfe die V_T-Verschiebung genauer analysiert werden kann. Hierzu wurden die Speicherzellen auf verschiedene Zustände programmiert und im Anschluss in mehreren Temper-Schritten die V_T-Verschiebung gemessen. Trägt man nun die V_T-Verschiebung nach Temperung über dem programmierten Zustand auf, ergibt sich ein Bild, wie in Abb. 3.19 gezeigt. Für die weitere Beschreibung wird diese Darstellung Hoffmann-Diagramm[1] genannt.

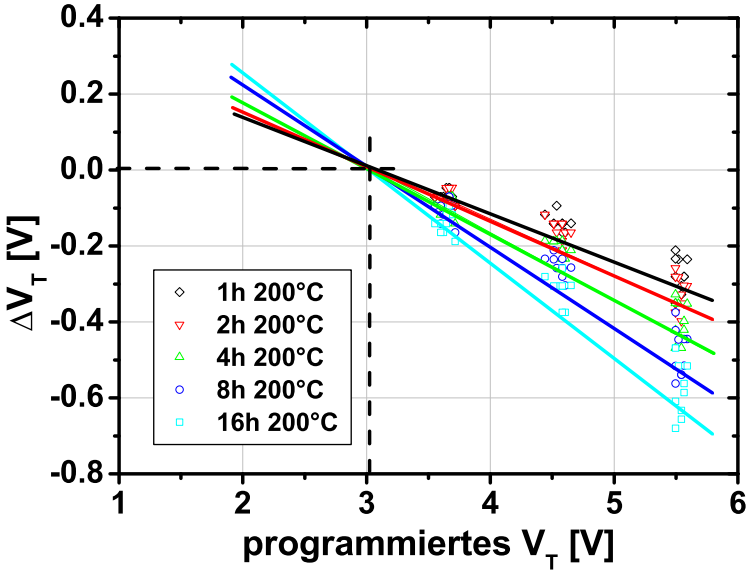

Abbildung 3.19: Hoffmann-Diagramm für eine 5x5 μm Speicherzelle mit TaN Gateelektrode; ONA = 5/6/12 nm

Wird eine Speicherzelle auf ein höheres V_T programmiert, erhöht sich gleichzeitig auch der Ladungsverlust und somit die Verschiebung. Dieser Effekt bleibt auch bei der Betrachtung für verschiedene Temperzeiten gleich, wie in Abb. 3.19 dargestellt. Interessant ist aber, dass sich ein definierter Schnittpunkt ergibt, wenn man die Ausgleichsgeraden für verschiedene Temperzeiten und eine größere Anzahl von Zellen aufträgt. Dieser Punkt repräsentiert den Zustand, an dem es zu keiner Verschiebung der Schwellspannung kommt. Aus diesem Grund wird der Punkt als Neutralpunkt bezeichnet. Er unterscheidet sich deutlich vom frischen Zustand nach der Fertigung, denn das initiale V_T für die gezeigte Zelle liegt bei circa 1.6 V. Diese Betrachtung des Ladungsverlustes ist auch auf skalierte Zellen anwendbar, wie das Hoffmann-Diagramm in Abb. 3.20a verdeutlicht.
Aufgrund des unterschiedlichen initialen Zustandes verschiebt sich bei der skalierten Speicherzelle aber das V_T des Neutralzustandes auf 1.1 V. Wenn wir nun zu der ursprünglichen Betrachtung für die erhöhte V_T-Verschiebung nach zuvor erfolgtem Programmieren zurückkehren und das Ergebnis in dem genannten Diagramm auftragen,

[1] nach dem Messingenieur, welcher die zugehörigen Messungen durchgeführt hat und bei der Auswertung behilflich war

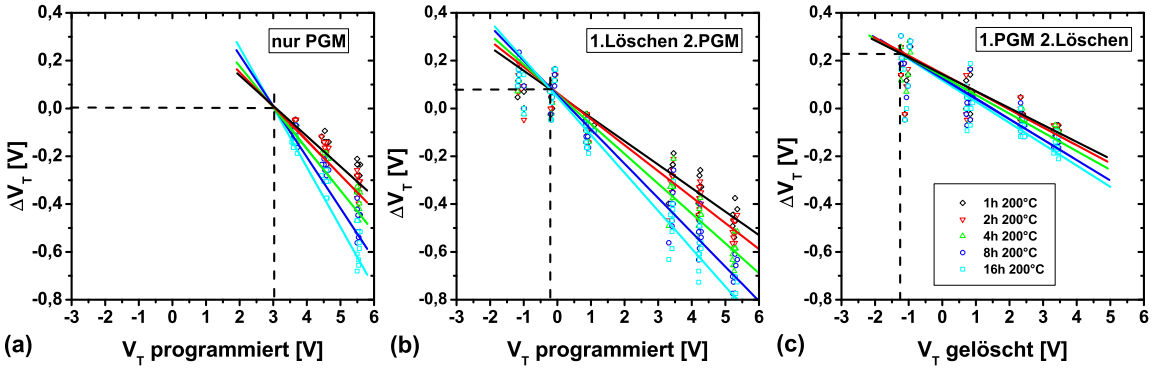

Abbildung 3.20: Hoffmann-Diagramme einer 5x5 μm Speicherzelle mit TaN Gateelektrode (a) für nur programmierte Zellen, in (b) sind die Zellen zuvor gelöscht und in (c) sind die Zellen nach dem Programmieren leicht gelöscht; ONA = 5/6/12 nm

zeigt sich das in Abb. 3.20b gezeigte Ergebnis. Es handelt sich hier um die gleichen Zellen wie in Abb. 3.20a, nur dass zuvor einmal gelöscht wurde. Dieser Löschvorgang resultiert in einer erheblichen Verschiebung des Neutralpunktes auf ein V_T von -0.2 V. Weiterhin verringert sich auch die Steigung der Ausgleichsgeraden. Die Überlagerung aus beiden Vorgängen zeigt aber schlussendlich einen höheren Ladungsverlust bei den betrachteten Programmierzuständen größer als 3 V. Dies wird bei dem Vergleich mit den Ergebnissen für nur programmierte Zellen in Abb. 3.20a deutlich. Die starke Verschiebung des Neutralzustandes erklärt demnach den erhöhten Ladungsverlust, wenn nur einmal zuvor gelöscht wurde. Eine Degradation des Tunneloxids kann aufgrund der sehr kurzen Stresszeit nahezu ausgeschlossen werden.
Eine mögliche Erklärung für die Verschiebung des Ladungs-Neutralpunktes ist die Injektion von Löchern während des Löschvorgangs, welche in der Speicherschicht verbleiben, auch in dem Fall, dass anschließend programmiert wird. Wird die Speicherzelle im frischen und ungeladenen Zustand programmiert, ergibt sich eine Ladungsverteilung wie in Abb. 3.21a illustriert.

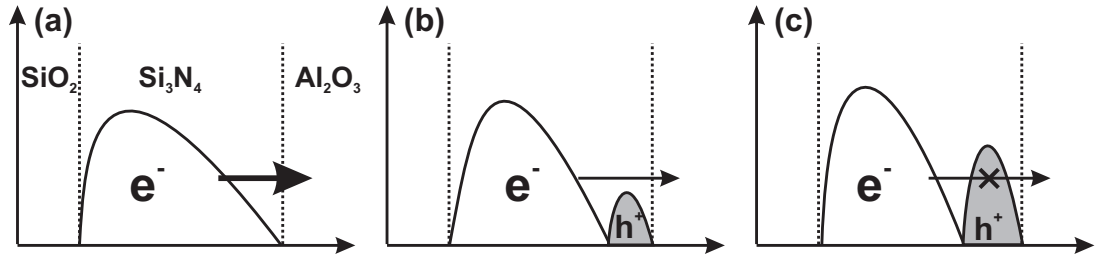

Abbildung 3.21: Erwartete Ladungsverteilung, die sich einstellt, wenn (a) nur programmiert wurde, (b) gelöscht und anschließend programmiert wurde oder (c) programmiert und nachfolgend gelöscht wurde

Aufgrund der günstigen Injektionsbedingungen werden nur Elektronen injiziert und die Anwesenheit von Löchern kann vernachlässigt werden. Werden zuvor Löcher durch einen Löschvorgang injiziert und anschließend programmiert, so ergibt sich eine Ladungsverteilung entsprechend Abb. 3.21b. Wie bereits in Kap. 3.3.2 gezeigt wurde, ist bei den untersuchten TANOS-Schichtstapeln mit dickem Tunneloxid der

Hauptleckpfad das Topoxid. Daher muss der Elektronendichte nah an der Grenzfläche zum Topoxid das Hauptaugenmerk geschenkt werden. Es wird für das Programmieren nach Löschen gezeigt, dass sich eine Löcherverteilung an der Topoxid-Grenzfläche ausbildet. Da die Tunnelbarriere für Löcher im Vergleich zu Elektronen größer ist, wird angenommen, dass diese sich erheblich schlechter aus der Speicherschicht entfernen lassen. Dies wird auch durch die schlechte Löschbarkeit von SONOS-Speicherschichten deutlich [80]. Berücksichtigt man dies für Löcher, kann man annehmen, dass sich beim Programmieren nach Löschen eine Elektronenverteilung neben einer Löcherverteilung, wie in Abb. 3.21b gezeigt, einstellt. Durch die zusätzliche Anwesenheit der Löcher ergibt sich eine Verschiebung des Neutralpunktes, entsprechend den Beobachtungen bei den Messungen. Die durch die Verschiebung der Elektronen-Ladungsverteilung vergrößerte Tunneldistanz resultiert in einem effektiv geringeren Ladungsverlust, welcher durch die Steigung der Ausgleichsgeraden in Abb. 3.20 wiedergegeben wird. Allerdings ist der beobachtete Ladungsverlust aufgrund der Verschiebung des Neutralpunktes größer. Die weitere Verbesserung des Ladungsverlustes für das Löschen nach dem Programmieren ist auf die gleiche Weise erklärbar. Durch die Injektion von Löchern in die zuvor durch das Programmieren erzeugte Elektronenverteilung ergibt sich eine größere Dichte von Löchern nah am Topoxid. In diesem Fall wird die Elektronen-Tunneldistanz im Vergleich zum Programmieren nach Löschen noch größer, wie in Abb. 3.21c verdeutlicht ist. Die größere Löcherdichte sorgt für eine weitere Verschiebung des Neutralpunktes in negativer V_T-Richtung. Zudem erhöht die größere Breite der Löcherverteilung die Tunnelstrecke und somit auch die Steigung der Ausgleichsgeraden. Eine genaue Auswertung der Ausgleichsgeraden-Steigung erfolgt in Abb. 3.22. Wobei die Daten, welche in Abb. 3.22a dargestellt sind, eine Auswertung der in Abb. 3.20 gezeigten Messwerte repräsentieren. Die Steigung selbst sagt aus, wie groß der Ladungsverlust ausgehend vom Neutralpunkt ist. Somit muss, wie bereits erwähnt, auch der Neutralpunkt für eine Absolutbetrachtung der V_T-Verschiebung berücksichtigt werden. Deutlich ist die Abhängigkeit der Steigung von dem Programmierschema erkennbar. Wie zuvor bereits erörtert, ist die geringste Steigung bei der Gruppe mit Löschen nach Programmieren zu beobachten. Wird die Zelle aus dem frischen Zustand programmiert, ist die Steigung am größten.

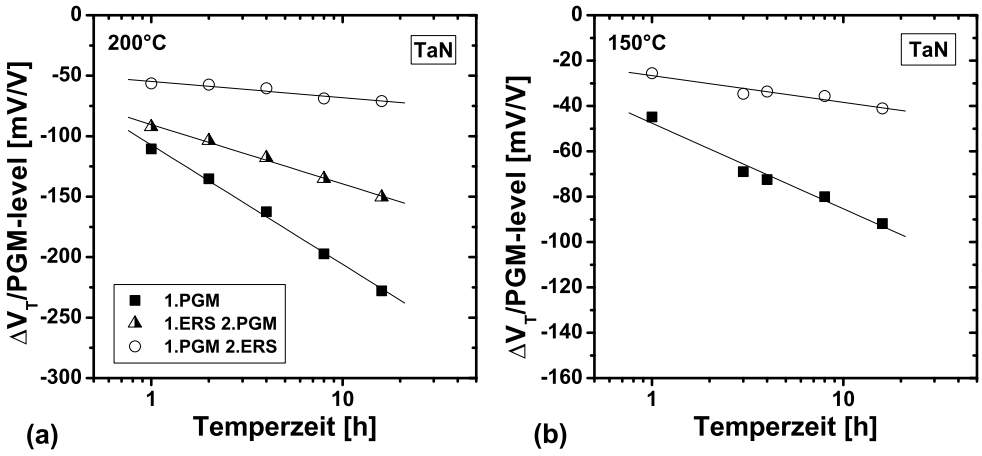

Abbildung 3.22: Steigung der Ausgleichsgeraden aus den Hoffmann-Diagrammen in Abb. 3.20 für die drei untersuchten Programmiermodi auf 48x48 nm Speicherzellen (a) bei 200°C und (b) bei 150°C

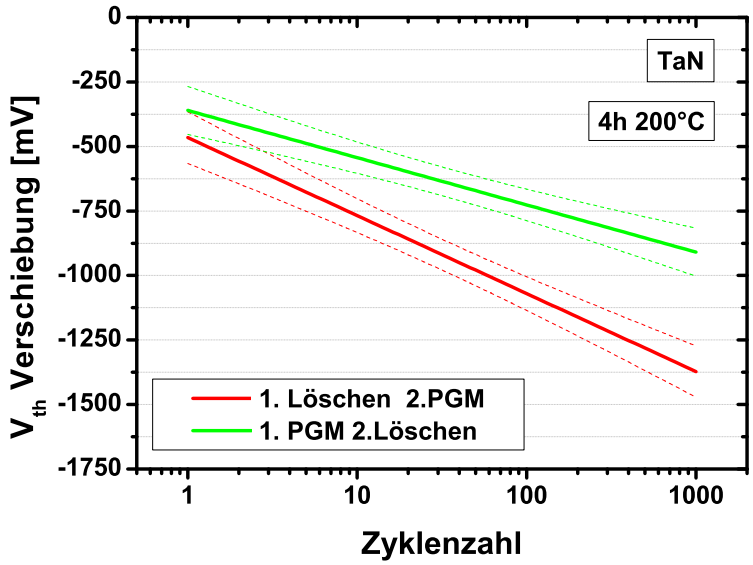

Abbildung 3.23: Mittel- und σ-Wert der V_T-Verschiebung über der Zyklenzahl für die Modi Löschen/Programmieren und Programmieren/Löschen; 48x48 nm Speicherzellen mit TaN Gateelektrode; ONA = 5/6/12 nm

Dieses Verhalten ist auch bei unterschiedlichen Temperaturen gleich, wie eine Auswertung einer Messung bei 150°C in Abb. 3.22b zeigt. Die Steigung ist wegen der niedrigeren Temperatur kleiner, aber das qualitative Verhalten in Abhängigkeit des Programmiermodus bleibt bestehen.
Weiterhin ist interessant, ob dieser Effekt auch nach einer größeren Zahl von Programmier- und Löschvorgängen bestehen bleibt. Aus diesem Grund wurden die Zellen mit einer unterschiedlichen Zahl an Zyklen vorbehandelt und anschließend der Ladungsverlust gemessen. Das Ergebnis in Abb. 3.23 untermauert das beobachtete Verhalten und die entsprechende Erklärung. Mit zunehmender Zyklenzahl wird der Abstand der V_T-Verschiebung zwischen den Programmiermodi immer größer. Dies kann damit erklärt werden, dass sich während der Löschvorgänge immer mehr Löcher ansammeln, welche während des Programmierens nicht kompensiert werden können. Die Zunahme des Ladungsverlustes beruht vorrangig auf einem immer größer werdenden initialen Ladungsverlust, schon kurz nach dem Programmieren. Dieses Verhalten entspricht dem in Kap. 3.3.2 beschriebenen Verlust kurz nach dem Programmieren. Die angenommene Akkumulation von Ladung im Al_2O_3-Topoxid verstärkt sich mit zunehmender Zyklenzahl entsprechend einer Programmierung auf einen höheren Programmierzustand. Wird nach dem Programmieren noch einmal gelöscht, wird diese Ladungsmenge deutlich reduziert und der beobachtete Ladungsverlust fällt geringer aus. Als weiterer Faktor kommt hinzu, dass während der Zyklen aufgrund der sehr hohen elektrischen Felder auch das Tunneloxid degradiert. Dadurch werden zusätzlich Fehlstellen gebildet, die zu einer Verschlechterung der Ladungshaltung führen [81].

4 Elektrisches Verhalten einer haftstellen-basierten Speicherzelle

In diesem Kapitel wird die Auswirkung auf das elektrische Verhalten untersucht, welches sich durch die Realisierung einer Speicherzelle basierend auf einem Transistor ergibt. Zunächst erfolgt eine Betrachtung, wie sich eine inhomogene Ladungsverteilung auf die Charakteristik auswirkt. Anschließend werden die Einflüsse betrachtet, die sich durch eine geänderte Materialwahl von Top-Dielektrikum und der Gateelektrode ergeben. Die Auswirkung einer veränderten Kanal- und Kontaktdotierung wird zum Ende des Kapitels erläutert.

4.1 Auswirkung von inhomogen verteilter Ladung in der Speicherschicht

Um eine hohe Zuverlässigkeit der Speicherzellen sicher zu stellen, ist es von elementarer Bedeutung, dass die Ladungsverteilung in der Speicherschicht homogen ist. Im Gegensatz zur Floating-Gate Zelle ist es in einer haftstellen-basierten Speicherzelle möglich, durch Feldüberhöhungen, wie sie an Strukturkanten entstehen, unterschiedlich viel Ladung zu injizieren. Daher befasst sich der folgende Abschnitt mit der Auswirkung inhomogener Ladungsverteilung in Weitenrichtung und Längenrichtung auf die Transferkennlinie einer haftstellen-basierten Speicherzelle.

4.1.1 Betrachtung in Weitenrichtung

Gibt es Bereiche mit unterschiedlicher Ladungsinjektion, kommt es zu einer Ausbildung von Bereichen mit unterschiedlicher Schwellspannung. Betrachtet man die Auswirkung einer inhomogenen Ladungsverteilung in Weitenrichtung, so kann dies durch eine Parallelschaltung von Transistoren mit unterschiedlichem V_T, gleicher Länge und entsprechender Weite dargestellt werden. Für den einfachsten Fall, dass zwei Bereiche unterschiedlicher injizierter Ladung und entsprechender Schwellspannung existieren, ergibt sich die in Abb. 4.1a gezeigte V_T-Verteilung. Die Ersatzschaltung ist in Abb. 4.1b dargestellt. Durch die Parallelschaltung der Bereiche unterschiedlicher Schwellspannung addiert sich der Strom, den die jeweiligen Bereiche durch die angelegte Gatespannung leiten. Der Teilstrom ist proportional zum Verhältnis Gesamtweite W zur Weite W_X des Bereiches X. Die Kennlinien sind durch die unterschiedlichen Schwellspannungen auf der Spannungsachse verschoben. Durch das Weitenverhältnis wiederum ergeben sich unterschiedliche Sättigungsströme. Ist nun das Weitenverhältnis entsprechend groß (> 10) und die Schwellspannung des schmaleren Bereiches kleiner, so ergibt sich ein zweistufiges Verhalten der Kennlinie, wie in Abb. 4.2 verdeutlicht. Hierbei wurde eine gemessene Kennlinie genommen, um 1 V

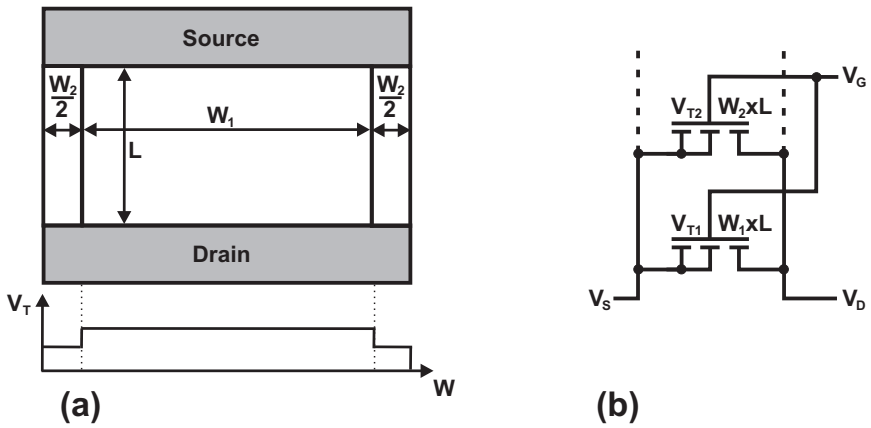

Abbildung 4.1: (a) Ausrichtung der Bereiche unterschiedlicher Schwellspannung; (b) Ersatzschaltung, die einen Transistor durch mehrere parallelgeschaltete Transistoren variierender Weiten und unterschiedlicher Schwellspannung ersetzt

in negative Gatespannungsrichtung verschoben und der Strom auf ein hundertstel skaliert, um das Verhalten zu illustrieren.

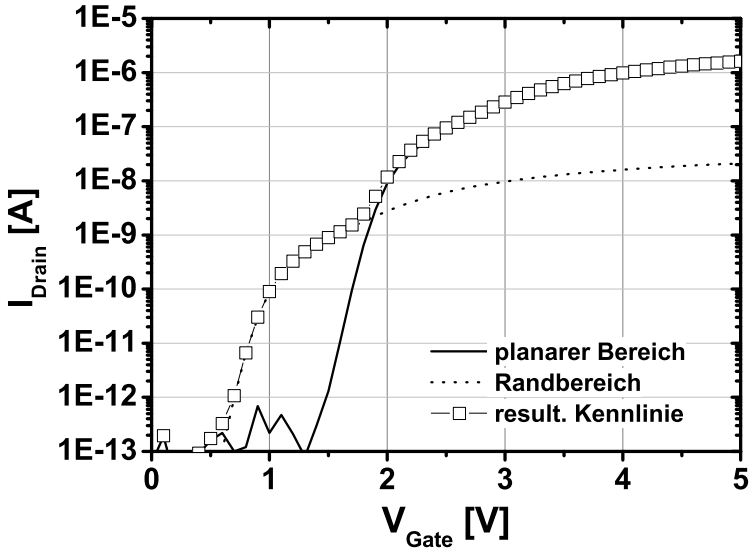

Abbildung 4.2: Transferkennlinien eines Transistors mit zwei Bereichen unterschiedlicher Schwellspannung und die zu messende resultierende Kennlinie (Kasten), der schmalere Bereich (punktiert) besitzt eine niedrigere Schwellspannung

Zu Beginn der Untersuchungen wurden speziell hergestellte Speicherzellen auf Basis des Buried-Bitline-Prozesses betrachtet [82]. Aufgrund ihrer Herstellung hat diese Speicherzelle den Nachteil, dass die Steuerelektrode an allen vier Seiten eine Kante zur Speicherschicht aufweist. Dies führt zu Feldüberhöhungen im Randbereich der Zelle und resultiert in einem abweichenden Programmierverhalten in diesem Bereich. Man spricht daher auch von einem 'corner-device'. Abbildung 4.3a zeigt die Transferkennlinie einer 5x5 μm großen Zelle während des Programmierens. Es ist deutlich ersichtlich, dass ein überwiegender Teil schnell programmiert und ein kleiner Teil mit deutlich weniger Strom langsamer programmiert. Dadurch wird die Unterschwellspannungscharakteristik beeinflusst und es können aus Parametern, die darauf, aufbauen

keine Informationen gewonnen werden. Abbildung 4.3b zeigt die Transferkennlinie einer modifizierten Zelle, bei der die Komponente mit niedriger Programmiergeschwindigkeit ausgeschaltet werden konnte. Dadurch ergibt sich ein Verhalten, welches nur durch eine Verschiebung der Kennlinie auf der Spannungsachse gekennzeichnet ist. Eine genauere Untersuchung der Zellen mit Hilfe von SEM Aufnahmen führt zur Erklärung des Effektes.

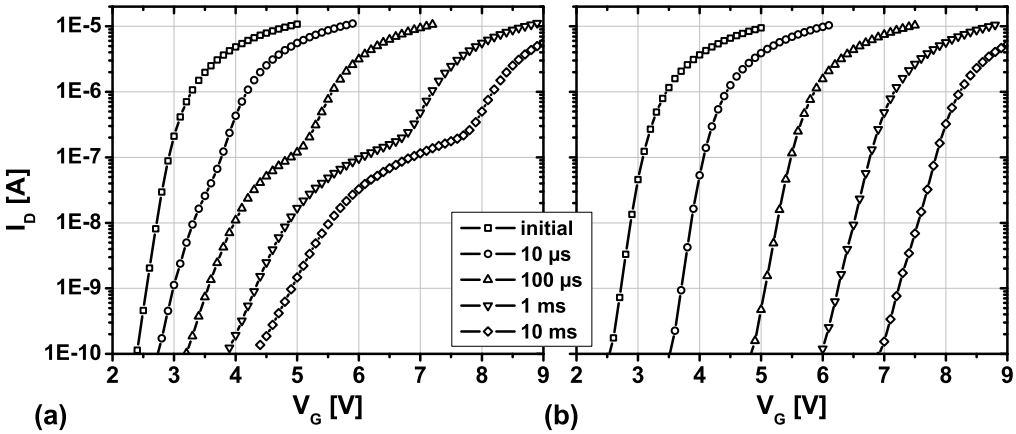

Abbildung 4.3: Transferkennlinien während des Programmierens nach den angegebenen Pulszeiten mit (a) Standardprozess zuzüglich einer Speichernitrid-Ätzung und (b) Standardprozess mit einer zusätzlich abgeschiedenen Oxidschicht

Abbildung 4.4a zeigt den Randbereich der Zelle mit einem starken Randeffekt und Abbildung 4.4b eine Zelle mit unterdrücktem Randeffekt. Die Zellen unterscheiden sich durch einen variierenden Abstand einer Nitrid-Kapselungsschicht zum Substrat. Bei der Probe mit einem messbaren Randtransistor befindet sich die Kapselungsschicht nah an der Steuerelektrode und im Bereich hoher elektrischer Felder. Es kommt daher während des Programmierens zu einer Injektion von Elektronen vom Substrat in die Einkapselungsschicht.

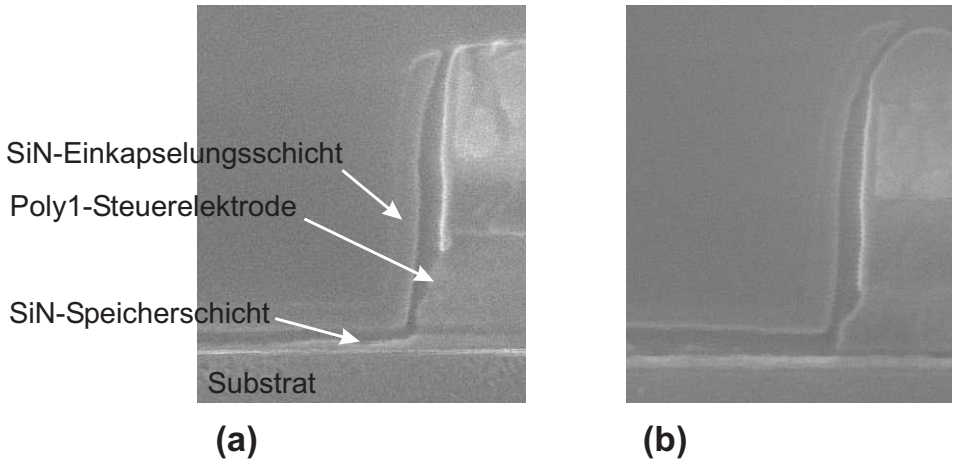

Abbildung 4.4: SEM Aufnahme des Randbereiches von Zellen einmal mit dem (a) Standardprozess und zusätzlicher Ätzung der Speicherschicht und (b) Standardprozess mit zusätzlicher Oxidschicht; für einen besseren Kontrast wurden die Proben durch eine Oxidätzung dekoriert

Dies resultiert in einem leitfähigen Kanal, der sich neben dem durch die Zellgeometrie definierten Kanal befindet und einen parasitären Transistor bildet. Möglich ist dies, da sich auch neben der eigentlichen Zelle Substrat befindet, welches zum Leiten gebracht werden kann. Die Wirkung der injizierten Ladung auf den leitfähigen Kanal ist durch den relativ großen Abstand der Schicht reduziert. Dadurch ist die V_T-Verschiebung im Vergleich zum eigentlichen Speichertransistor kleiner. In Abb. 4.5 ist die Änderung der Schwellspannung des Randbereiches im Vergleich mit dem planaren Transistor dargestellt. Die Auswertung der Schwellspannungs-Verschiebung des Randtransistors erfolgt bei einem Drainstrom von $1e^{-9}$ A und für den planaren Bereich bei $1e^{-6}$ A.

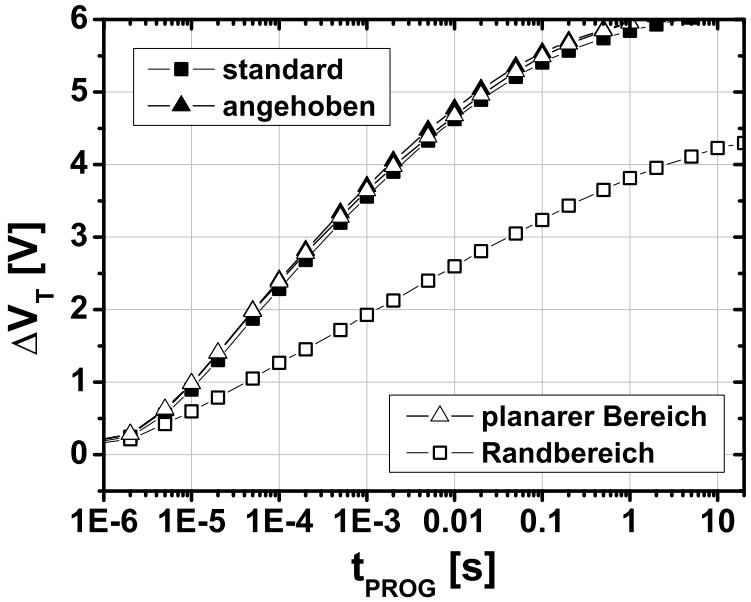

Abbildung 4.5: V_T-Verschiebung des Randbereichs und des planaren Transistors, einmal für den Standardprozess (Dreiecke) und andererseits für die angehobene Nitridschicht (Vierecke)

Die eigentlichen Transistoren verhalten sich identisch, nur der parasitäre Transistor zeigt durch den großen Abstand der Speicherschicht eine geringere Änderung der Schwellspannung in Abhängigkeit von der Pulslänge. Ist die Schwellspannung des parasitären Transistors größer, wird dieser durch den eigentlichen Transistor überlagert und hat keinen Einfluss auf die Transferkennlinie. Dieser Effekt kann gezielt genutzt werden, um etwaige Randtransistoren zum Beispiel durch entsprechende Kanalimplantationen zu unterdrücken [83].

4.1.2 Betrachtung in Längenrichtung

Eine inhomogene Ladungsverteilung in Längenrichtung wird im NROM Konzept [17] gezielt genutzt, um zwei lokal voneinander getrennte Bits zu speichern. Zum besseren Verständnis dieses Zellkonzept gibt es verschiedene theoretische Betrachtungen, die eine Auswirkung von lokalisierter Ladung auf den Unterschwellspannungsbereich untersuchen [84–86]. Im Prinzip entspricht die lokalisierte Ladungsspeicherung in Längenrichtung einer Reihenschaltung von zwei Transistoren mit unterschiedlicher Schwellspannung, entsprechend der injizierten Ladung. Allerdings stimmt diese Vereinfachung nur näherungsweise. In diesem Fall würde der Bereich mit der größten

Schwellspannung die Transferkennlinie bestimmen und eine Veränderung des Swings wäre nicht zu beobachten. Shappir [87] hat ein Modell zur Beschreibung des Swing in Abhängigkeit der Weite des gespeicherten Ladungspaketes vorgeschlagen. Er umgeht das Problem der Reihenschaltung durch Einführung eines Übergangsbereiches im Kanal, der nur teilweise invertiert ist. In Abb. 4.6 ist dies für den Lesefall schematisch dargestellt. Im Bereich der Speicherzelle in der keine Ladung gespeichert ist, befindet sich der Kanal in dem niederohmigen Zustand der Inversion bei einer entsprechend gewählten Lesespannung. Durch die Ladung befindet sich ein Stück des Kanals der Länge L_{Ladung} in Verarmung und das Stück mit der Länge $\alpha \cdot x_v$ weist eine verringerte Dicke der Inversionsschicht auf.

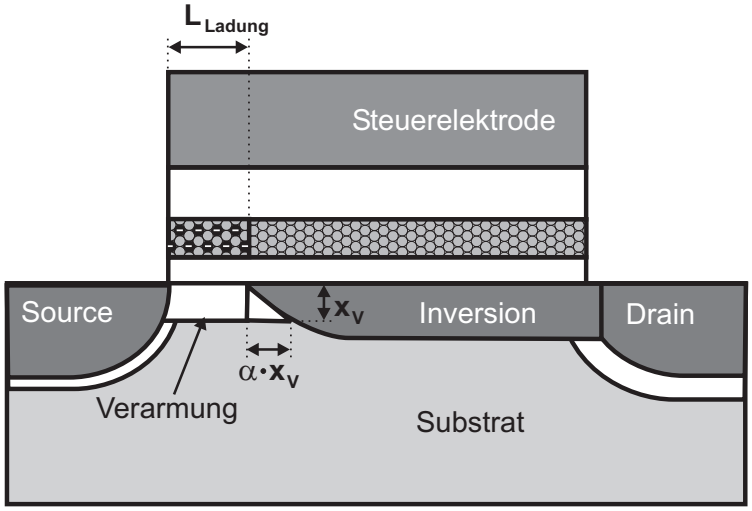

Abbildung 4.6: Schematische Darstellung für die Ausbildung des Kanals während des Lesens einer Speicherzelle mit lokaler Ladungsspeicherung; Darstellung der notwendigen Parameter für die Berechnung des Swings nach [87]

Beide Bereiche dominieren durch ihren relativ hohen Widerstand die Transferkennlinie, wobei sich folgende Beziehung zwischen Gatespannung und Oberflächenpotential ergibt [86]:

$$\phi_{Ob} \equiv V_G - V_{FB} = \phi_s + \frac{qN_B}{C_i L_{Ladung}} \left(x_v L_{Ladung} + \frac{\alpha x_v^2}{2} \right). \tag{4.1}$$

Daraus lässt sich eine Gleichung ableiten, die einen linearen Koeffizienten vor dem Oberflächenpotential erhält [87]:

$$\phi_{Ob} = \phi_s + \frac{\alpha \epsilon_{Si}}{C_i L_{Ladung}} \phi_s + \frac{qN_B x_v}{C_i} = (1 + \beta_L)\, \phi_s + \frac{qN_B x_v}{C_i} \tag{4.2}$$

mit

$$\beta_L = \frac{\alpha \epsilon_{Si}}{C_i L_{Ladung}}. \tag{4.3}$$

β_L gibt das Längenverhältnis von Verarmungszone zu der Zone mit einsetzender Inversion wieder. Es wird gezeigt, dass durch eine Verkleinerung der Ladungspaketlänge der Faktor β_L größer wird. Durch Einsetzen der Gl. 4.2 in die Gleichung zur Bestimmung des Swings erhält man

$$S = ln10 \frac{d\phi_{Ob}}{d\,(lnI_D)} \cong 2.3 \frac{kT}{q} \left[\frac{(1 + \beta_L)\, C_i + C_V\,(\phi_s)}{C_i} \right]. \tag{4.4}$$

Es folgt somit, dass eine Vergrößerung von β_L, respektive ein kleineres Ladungspaket, zu einer Verschlechterung des Swings führt. Diese Näherung ist allerdings nur gültig, wenn das Ladungspaket eine ausreichend große Differenz der Schwellspannung, im Vergleich zum restlichen Transistor, induziert. Das beruht auf der Annahme, dass sich der Rest des Kanals in Inversion befindet, was bei der Betrachtung sichergestellt sein muss. Eine genauere Analyse mit Hilfe von 2D-Transistor-Simulationen wird von O. Klar [86] durchgeführt. Es wird hierbei untersucht, welchen Einfluss die Lage und Größe eines Ladungspaketes auf die Transferkennlinie einer Speicherzelle haben. Es zeigt sich bei den untersuchten Zellen, dass ein Ladungspaket von $N_B = 1 \cdot 10^{19}\ cm^{-3}$ der Länge 35 nm zu einem maximalen Swing von 230 mV/dec führt. Ein Vergrößerung des Ladungspaketes verringert den Einfluss des Faktors β_L aus dem zuvor beschriebenen Modell und die Auswirkung auf den Swing nimmt ab. Bei einer Verkürzung wiederum nimmt der Einfluss der Ladung auf den Kanal ab und der Swing wird wieder kleiner. Erhöht man die injizierte Ladung, verschiebt sich das Maximum hin zu kürzeren Längen. Bei $4 \cdot 10^{19}\ cm^{-3}$ liegt der maximale Swing von 550 mV/dec bei einer Weite des Ladungspaketes von 20 nm.
Die Untersuchungen beziehen sich jeweils nur auf den Fall der Elektroneninjektion. Bei den untersuchten Proben traten allerdings Schwierigkeiten im Fall des Löschens auf. Abbildung 4.7a zeigt die Transferkennlinien einer solchen Speicherzelle mit einer Dimension von L/W = 180 nm/5 μm, wodurch sich der relativ große Sättigungsstrom ergibt.

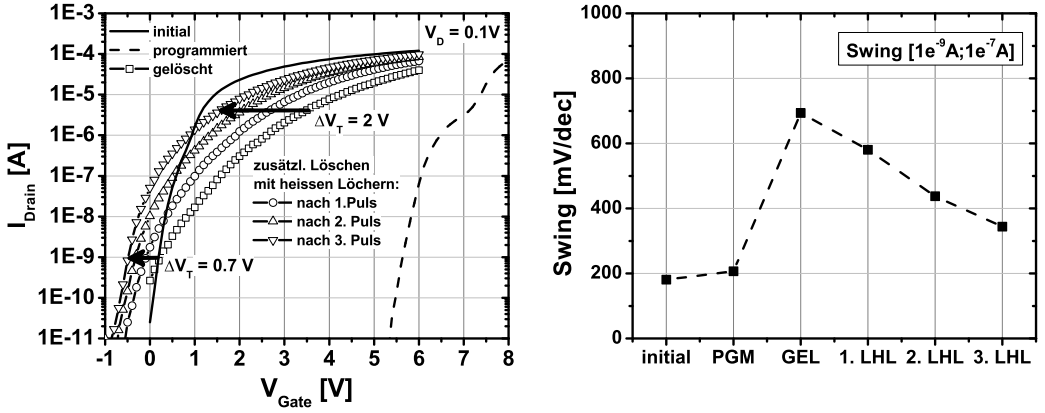

Abbildung 4.7: (a) Transferkennlinien einer TANOS-Zelle (3.5/9/12 nm) initial, nach dem Programmieren und Löschen mit FN-Tunneln und anschließendem Löschen mit heißen Löchern (LHL); (b) sub-V_T Swing der einzelnen Zustände (PGM = Programmiert, GEL = Gelöscht)

Bei der Kennlinie des programmierten Zustandes zeigt sich wieder ein parasitärer Transistor in Längenrichtung, da an vergleichbaren Strukturen gemessen wurde, die bereits im vorangegangen Kap.4.1.1 analysiert wurden. Nach dem Löschen mit FN-Tunneln durch einen Puls von -22 V und 20 ms stellt sich die durch Boxen dargestellte Kennlinie ein. Deutlich ist die Abnahme des sub-V_T Swings zu sehen, welcher in 4.7b zur besseren Verdeutlichung zusätzlich ausgewertet wurde. Es erfolgt ein Sprung von 200 mV/dec auf 700 mV/dec. Es wird aus Abb.4.7b deutlich, dass nach einem Löschpuls mit heißen Löchern (LHL) der sub-V_T Swing kleiner wird. Dies ist auch dadurch erkennbar, dass sich die Kennlinie in Abb.4.7a während der Löschpulse (Kreise → Dreiecke) der initialen Charakteristik annähert. Beim Löschen mit FN-

Tunneln erfolgt eine Löcherinjektion vom Kanal und Elektronen werden von der Steuerelektrode emittiert. Aufgrund der Struktur kommt es an den Kanten der Elektrode zu Feldüberhöhungen und der Bereich nahe der Source und der Drain erfährt eine erhöhte Elektroneninjektion. Dieser Bereich kann daher nur bedingt gelöscht werden und hat eine höhere Schwellspannung entsprechend der Funktionsweise des NROM-Konzepts [17]. Eine Möglichkeit, diese Differenz aufzuheben, ist die Injektion von Löchern nahe der Source und Drain mittels heißer Löcher, wie im Experiment von Abb.4.7 gezeigt. In diesem Fall wird die Schwellspannung dem planaren Bereich angepasst und der sub-V_T Swing wird verringert. Das erwartete Verhalten wird durch die Messung bestätigt und zeigt, dass der Bereich über den Anschlussgebieten nicht so gut gelöscht werden kann. Dieses Verhalten verhindert eine genaue Untersuchung des Löschverhaltens, welches ein störungsfreie Speicherzelle zeigen würde. Weiterhin wurde die Abhängigkeit der Löschcharakteristik bei einer Variation der Topoxiddicke untersucht. Die Ergebnisse sind im folgenden Kap. 4.2.1 dargelegt.

4.2 Auswirkungen von Al_2O_3-Topoxid auf das Zellverhalten

Bereits in Kap. 2.3.4 wurde gezeigt, dass die Wahl das Topoxids einen großen Einfluss auf das Verhalten der Speicherzelle besitzt. Aus diesem Grund wurden verschiedene Modifikationen vorgenommen, mit dem Ziel, die Eigenschaften des Isolators zu verbessern. Das betrachtete Aluminiumoxid ist in der Form von Korund am bekanntesten. In diesem Fall liegt das Al_2O_3 im trigonalen Kristallsystem α-Al_2O_3 vor. Es besitzt eine sehr große Härte und findet demzufolge als Schleifmittel weite Verbreitung. Eine Herausforderung stellt in diesem Zusammenhang die Strukturierung von Speicherzellen dar, welche zuerst betrachtet wird. Im Anschluss werden der Einfluss von Prozessparametern sowie die Zugabe von SiO_2 untersucht.

4.2.1 Ätzflanke des Al_2O_3-Topoxids

Die anisotrope Ätzung von Aluminiumoxid stellt eine große Herausforderung aufgrund seiner hohen chemischen und thermischen Beständigkeit von Al_2O_3 dar [88]. Eine Möglichkeit Aluminiumoxid zu ätzen, ist eine Vorschädigung mittels Ionenimplantation und anschließendem nasschemischem Abtrag [89]. Alternativ kann man große Strukturen ohne Maske auch mittels hochenergetischem Laserlicht generieren [90].
Weiterhin besteht die Möglichkeit, mit Hilfe von einem BCl_3/Argon/N_2 Plasma zu ätzen [91–93]. Allerdings zeigt sich bei der Ätzung mit Raumtemperatur eine relativ starke Ätzflanke mit einem Winkel von circa 55 Grad [94]. Im Rahmen dieser Arbeit wurden zu Beginn Versuche mit einer Ätzkammer durchgeführt, welche diese Bedingungen aufweist. Aus den genannten Gründen kam es dabei zu einer relativ starken Ätzflanke mit einem Winkel von circa 45 Grad, wie Abb.4.8 verdeutlicht. Durch die Ätzflanke kommt es zu einer größeren effektiven Weite des Siliziumnitrides. Dies resultiert in einer etwas längeren Speicherzelle und einem Bereich, der nur in einem begrenzten Maße durch die Steuerelektrode beeinflusst wird. Beim Programmieren solch eines Transistors gibt es keine Probleme, da eine Löcherinjektion von der Steuerelektrodenkante vernachlässigbar ist, wie Abb. 4.9a zeigt. Wird allerdings

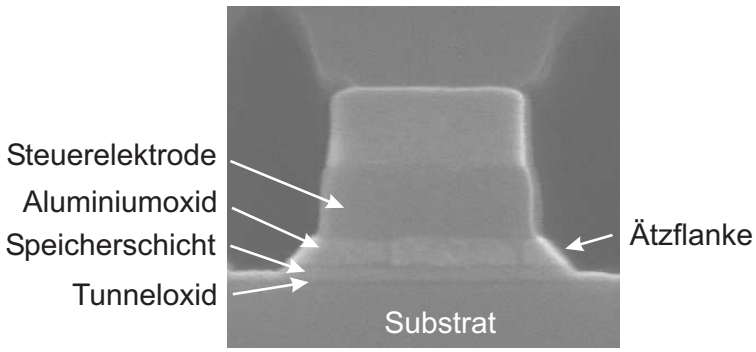

Abbildung 4.8: SEM-Aufnahme einer SANOS-Zelle mit deutlich zu erkennender Ätzflanke, verursacht durch einen nicht optimierten Aluminiumoxid Ätzprozess

versucht, eine solche Zelle vom ladungsfreien Zustand aus zu Löschen, ist ein Anstieg der Schwellspannung zu beobachten. Dieses Verhalten wird in Abb. 4.9a durch die geschlossenen Symbole verdeutlicht.

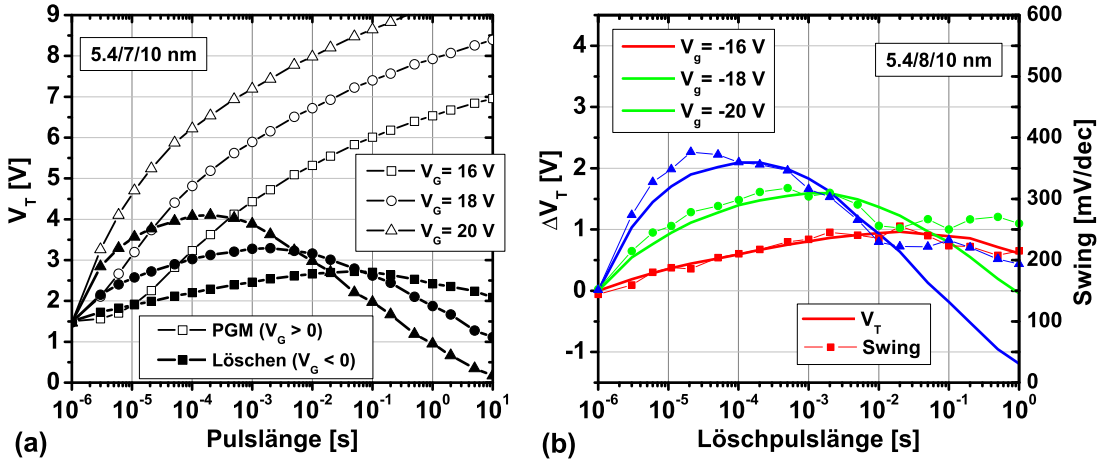

Abbildung 4.9: (a) Programmier- (offene Symbole)und Löschcharakteristik (geschlossene Symbole) einer SANOS-Zelle; (b) Vergleich des sub-V_T Swings mit der Löschcharakteristik während des Löschvorgangs

Bei genauerer Betrachtung des Löschvorgangs mit einer zusätzlichen Auswertung des sub-V_T Swings zeigt sich, dass es sich bei dem Spannungsanstieg um ein Auswerteartefakt handelt. Ein Vergleich zwischen sub-V_T Swing und Schwellspannung während des Löschens ist in Abb. 4.9b veranschaulicht. In diesem Diagramm wird gezeigt, dass der Anstieg der Schwellspannung nur durch die Abnahme der Unterschwellspannungssteigung zustande kommt. Der Einfluss der Unterschwellspannungssteigung auf die Auswertung des V_T's wurde bereits in Abb. 4.7 dargestellt. Die sich ergebende Löschcharakteritik kann durch eine Elektroneninjektion von der Steuerelektrode während des Löschens erklärt werden [95]. Veranschaulicht wird dieses Verhalten in Abb. 4.10.
Bei kurzen Pulszeiten erfolgt zunächst eine überwiegende Speicherung von Elektronen im Randbereich der Zelle. Die hohen Felder an der Elektrodenkante sind Ursache für diesen Vorgang und durch die starke Ladungskonzentration kommt es zu einer Zunahme des sub-V_T Swings, wie in Kap. 4.1.2 beschrieben. Wird der Löschvorgang

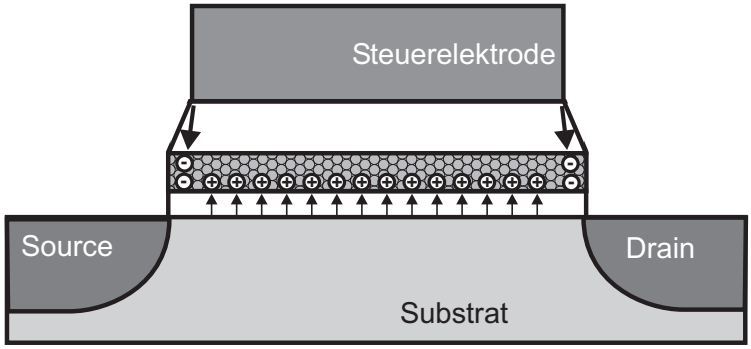

Abbildung 4.10: Schematische Darstellung der Elektroneninjektion im Randbereich der Speicherschicht beim Löschen mit ausgeprägter Al_2O_3-Ätzflanke

fortgesetzt, nimmt auch die Löcherinjektion weiter zu. Hierbei erfolgt eine bevorzugte Löcherinjektion im Bereich der injizierten Elektronen, da diese das Feld über das Tunneloxid erhöhen. Daraufhin wird die Elektroneninjektion durch die Löcherinjektion kompensiert und die Zunahme des sub-V_T Swings und der Schwellspannung enden. Mit zunehmender Pulsdauer setzt der eigentliche Löschvorgang ein und die lokal nahe Source/Drain gespeicherten Elektronen werden durch Löcher kompensiert. Dadurch wird der sub-V_T Swing wieder kleiner und die Schwellspannung folgt dem Verlauf des sub-V_T Swings unabhängig von der angelegten Löschspannung, wie in Abb. 4.9 gezeigt wird. Der sub-V_T Swing kehrt wieder ungefähr zum Ausgangswert zurück und sättigt dann, ein Zeichen dafür, dass die Gateinjektion nahezu ausgeglichen wurde. Ab diesem Zeitpunkt, der bei -20 V ungefähr bei 10 ms liegt, ist das Löschverhalten zu beobachten, welches bei einer störungsfreien Speicherzelle zu erwarten ist. Die Weite der neben der Steuerelektrode überstehenden Speicherschicht kann durch die Dicke des Aluminiumoxids variiert werden. Dies beruht auf der Tatsache, dass die Weite des maskierenden Topoxids durch den Flankenwinkel in Kombination mit der Dicke bestimmt ist. Abbildung 4.11a veranschaulicht die Abhängigkeit der Schwellspannungsänderung von der Dicke der Aluminiumoxid Schichtdicke. Es werden drei Dicken zwischen 6 und 14 nm bei einer Löschspannung von -20 V verglichen. Bei einem Flankenwinkel von 45 Grad ergeben sich somit Weiten in der Größenordnung zwischen 6 und 14 nm, die nur begrenzt durch die Steuerelektrode gesteuert werden können.

Zwei Effekte bestimmen das Verhalten in Abhängigkeit der Topoxiddicke. Einerseits maskiert die Ätzflanke die Source-/Drain-Implantation, welche später durch thermische Prozesse eine gewisse Strecke unter den Schichtstapel diffundiert. Wird Ladung von der Steuerelektrodenkante in die Speicherschicht unterhalb der Ätzflanke injiziert, befindet diese sich teilweise über dem Kontakt. Bei einer Toppoxiddicke von 6 nm wird der größte Teil der Elektronen oberhalb des Kontaktes gespeichert. Durch die geringe Wirkung auf den Kanal vergrößert sich der sub-V_T Swing nicht so stark. Eine zunehmende Topoxiddicke vergrößert den Anteil der Ladung über dem Kanal und die Wirkung wird immer größer, was sich in einer deutlichen Zunahme des sub-V_T Swings und der entsprechenden Schwellspannungsverschiebung äußert. Ein weiterer Effekt ergibt sich durch den wirksamen Hebelarm, mit dem die injizierte Ladung die Oberflächenspannung des Kanals bestimmt. Vergrößert man die Dicke des Topoxids bei gleichbleibender Dicke von Tunneloxid und Speicherschicht, vergrößert sich auch die Wirkung gespeicherter Ladung. Abbildung 4.11b verdeutlicht das er-

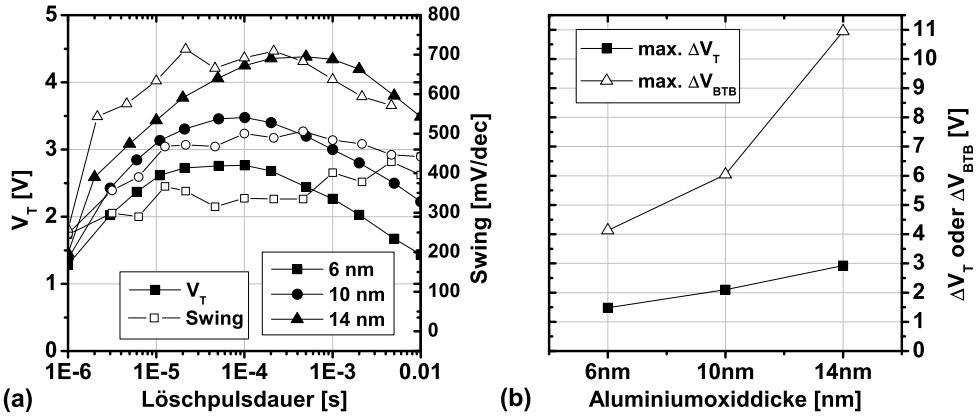

Abbildung 4.11: Abhängigkeit der V_T-Verschiebung beim Löschen für drei verschiedene Al_2O_3-Dicken, (a) vergleicht die V_T-Verschiebung mit dem sub-V_T Swing und (b) vergleicht die maximale Verschiebung bei dem Löschen am Kondensator (offene Symbole) und Transistor (geschlossene Symbole)

wartete Verhalten mit Hilfe der maximalen Verschiebung von gemessener Einsatz- und Band-zu-Band-Tunnel-Spannung (V_{BTB}). V_{BTB} ist sensitiv gegenüber Ladung, die sich direkt über dem pn-Übergang des entsprechenden n-Kontaktes befindet [96]. Bestimmt wird V_{BTB} durch die Messung des GIDL Stromes (Kap. 2.2.3) und der Auswertung bei einem definierten Leckstrom. Beide Messgrößen verhalten sich wie erwartet proportional zur Dicke des Topoxids. Es wird auch deutlich, dass es sich um eine große Ladungsmenge, aufgrund der großen V_{BTB}-Verschiebung, handeln muss, die zu der Verschiebung des V_T's führt. Für 14 nm Al_2O_3-Topoxid ist eine Verschiebung von V_{BTB} um 11 V zu beobachten.

Da es sich bei dem Effekt um eine Degradation der Transferkennlinie eines Transistors handelt, dürfte der Effekt bei einem Kondensator nicht zu beobachten sein. Abbildung 4.12 zeigt den Vergleich von Kondensator und Transistor für eine Löschspannung von -20 V für zwei verschiedene Ausgangsbeladungen.

Es ist deutlich zu erkennen, dass der Kondensator zu Beginn des Löschens nahezu keine Änderung der Flachbandspannung zeigt, wohingegen der Transistor, wie bereits berichtet wurde, eine starke Änderung der Schwellspannung aufweist. Weiterhin ist auffällig, dass die Löschkurven der unterschiedlichen Strukturen nicht übereinander liegen, wie man es erwarten würde, wenn das Verhalten nur durch den Schichtstapel bestimmt ist. Erklärt werden kann dies durch eine nicht reversible Verschiebung der Schwellspannung um circa 1 V für die untersuchten Strukturen, hervorgerufen durch die Degradation der Transferkennlinie. Dies gilt sowohl für eine programmierte Zelle, als auch für eine unbehandelte Zelle. Die Degradation des geschriebenen Zustandes wird sichtbar durch den Sprung der Schwellspannung zu Beginn des Löschvorganges, gezeigt in Abb. 4.12. Der Effekt kann unterdrückt werden, indem man den Bereich begrenzter Steuerbarkeit so klein wie möglich macht. Dies wird erreicht, indem man den gesamten Zellrand so senkrecht wie möglich strukturiert. Eine deutliche Verbesserung des Ätzflankenwinkels kann durch Verwendung einer Plasmaätzung in Verbindung mit erhöhter Wafertemperatur (250°C) erzielt werden [97–99], wie Abb. 4.13 anhand einer SEM-Aufnahme verdeutlicht.

Hierbei wird gezielt die anisotrope chemische Ätzkomponononente verstärkt. Dies geschieht durch eine schnellere chemische Reaktion und durch die erhöhte Tempera-

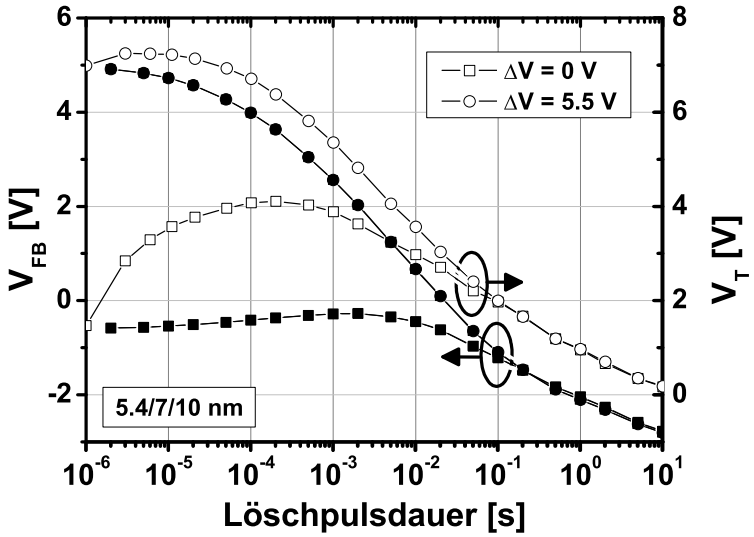

Abbildung 4.12: Vergleich der Löschkurven am Kondensator (geschlossene Symbole) und Transistor (offene Symbole) für zwei unterschiedliche Ausgangszustände an einer SANOS-Struktur; V_G = -20 V

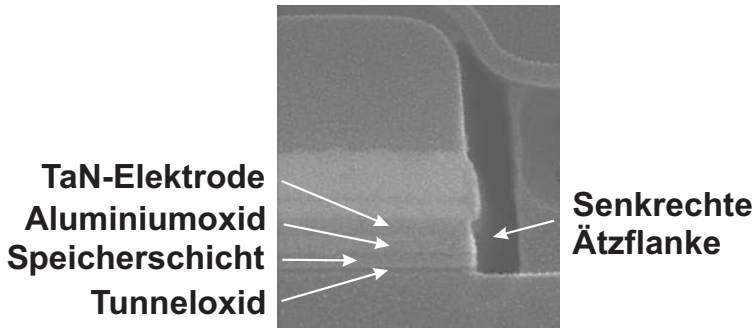

Abbildung 4.13: SEM-Aufnahme eines NAND-Auswahltransistors, bei dem das Aluminiumoxid mit einer erhöhten Temperatur von 350°C geätzt wurde und man dadurch ein nahezu senkrechtes Ätzprofil erhält

tur entstehen gasförmige Ätzprodukte die leichter abtransportiert werden können. Verwendete Ätzgase sind BCl_3/Cl_2 [100, 101] und HBr [102]. Ätzprodukte sind zum Beispiel Aluminiumchlorid ($AlCl_3$), welches eine Sublimationstemperatur von 196°C besitzt [100] und Aluminiumbromid mit einer Siedetemperatur von ≈260°C [103]. Es ist ersichtlich, dass der Ätzprozess verbessert wird, wenn die Oberflächentemperatur größer als 250°C ist und gasförmige Reaktionsprodukte entstehen. Denn diese können nur als Gase effektiv abtransportiert werden. Ist dies der Fall, kommt es zu keiner Wechselwirkung zwischen den Reaktionsprodukten und dem Ätzprozess und es resultiert eine steilere Ätzflanke.

4.2.2 Al_2O_3-Topoxid Abscheidebedingungen

Im Rahmen dieser Arbeit wurde das Aluminiumoxid mit einem Prozess durch Atomlagenabscheidung (ALD) erzeugt. Als Reaktionsgas, auch Präkursor (engl. precursor) genannt, wurde Trimethylaluminium (TMA) als Aluminiumquelle und O_3 bzw. H_2O als Sauerstoffquelle verwendet. Die Schichten wurden auf einer Anlage

von Tokyo Electron Limited (TEL) des Types FORMULA bei einer Temperatur von 300°C abgeschieden. Die Abscheiderate hängt zu Beginn stark von dem Oberflächenmaterial und der Konditionierung ab [104, 105]. Nachdem sich eine geschlossene Monolage gebildet hat, beträgt die Abscheiderate auf Siliziumnitrid pro Zyklus 9.11 Å. Ein wichtiger Aspekt bei der ALD-Abscheidung ist die Zyklenzeit. Bei der verwendeten Standardabscheidung einer nach Temperung etwa 12 nm dicken Aluminiumoxid-Schicht werden 157 Zyklen benötigt. Ein Zyklus besteht aus den 4 Teilschritten:

- 1. Stickstoffspülung,
- TMA Reaktion,
- 2. Stickstoffspülung und
- Oxidation mit O_3 bzw. H_2O.

Bei einer Zyklenzeit von $\approx$ 4 Minuten, für eine Oxidationsdauer von 3 Minuten, ergibt sich eine Gesamtprozessdauer von 628 Minuten, dies entspricht circa $10\frac{1}{2}$ Stunden. Für eine großtechnische Fertigung ist eine solche Prozessdauer, trotz der Tatsache, dass mehr als 100 Wafer parallel prozessiert werden, nicht akzeptabel. Daher wurde untersucht, inwieweit sich eine Verkürzung der Zyklenzeit auf das elektrische Verhalten der Speicherzellen auswirkt. Der Einfluss auf Programmieren und Löschen ist in Abb. 4.14a gezeigt.

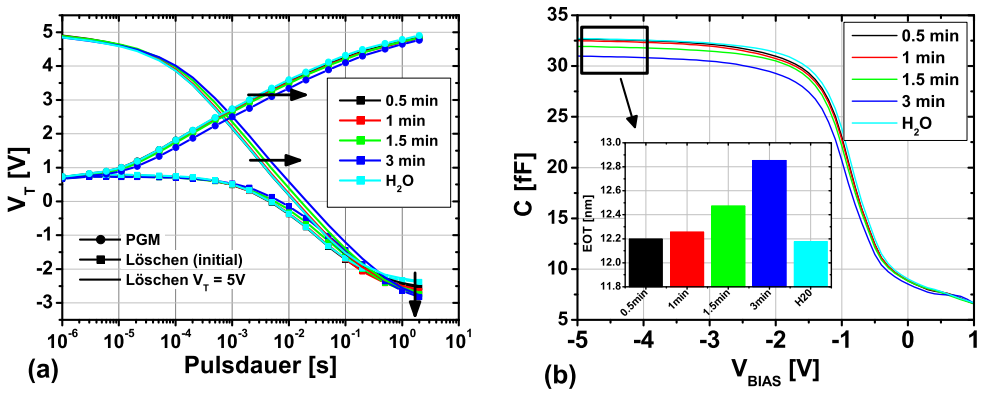

Abbildung 4.14: (a) Programmier- und Löschverhalten von SANOS-Zellen mit 2.5/9/10 nm Schichtdicken und variierten Oxidaktionszeiten während der ALD-Aluminiumoxid-Abscheidung; (b) gemessenen CV-Kurven und extrahierten Schichtdicken der in (a) gemessenen Proben

Der gemessene Unterschied zwischen den verschiedenen Proben ist nur klein. Ab einer Pulsdauer von 1,5 Minuten verschiebt sich die Programmier- und Löschcharakteristik hin zu längeren Pulszeiten. Wie Abb. 4.14b allerdings verdeutlicht, ist dies nicht auf Effekte durch die Abscheidung zurückzuführen. Es wird gezeigt, dass alle Proben nahezu vergleichbare äquvalente Oxiddicken (EOT) aufweisen, außer die Gruppen 1,5 min und 3 min Oxidationsdauer. Deren EOT ist größer und somit sind die wirksamen Tunneloxidfelder kleiner. Die Berechnung des EOT's erfolgt nach folgender Gleichung:

$$EOT\,(nm) = \frac{\epsilon_0 \epsilon_{SiO_2} A}{C_{meas}} \tag{4.5}$$

mit A = 11534 μm^2, ϵ_{SiO_2}=3.9 . Das Resultat ist ein langsameres Programmieren und Löschen, aber auch eine leicht niedrigere Löschsättigung. Das wiederum ist auf ein günstigeres Verhältnis von Kanal-Löcherinjektion und Steuerelektroden-Elektroneninjektion, durch das niedrigere Feld im Al_2O_3 zurückzuführen. Zudem enthalten die Messungen in Abb. 4.14 auch einen Vergleich von O_3 und H_2O Präkursor. Es ist ersichtlich, dass die Dicke der H_2O Gruppe gut mit der 30 s O_3 Gruppe übereinstimmt. Ein Vergleich der elektrischen Resultate dieser Gruppen zeigt einen vernachlässigbaren Unterschied. Als Ursache für die Änderung der elektrisch wirksamen Oxiddicke wurde die bei der Abscheidung erzeute SiO_2-Schicht auf dem S_3N_4 erkannt [106, 107].

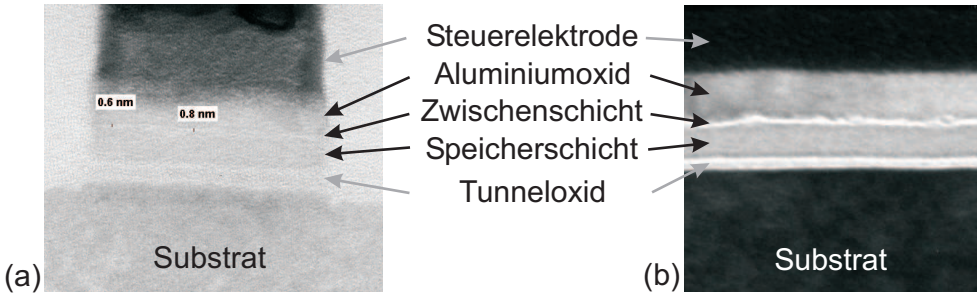

Abbildung 4.15: TEM-Aufnahmen von (a) einem SANOS-Stapel mit Al_2O_3 abgeschieden mit H_2O-Präkursor und (b) einem TANOS-Stapel mit Al_2O_3 abgeschieden mit O_3-Präkursor 3 min

Abbildung 4.15 zeigt die Schichtdicken für eine Abscheidung mit H_2O und O_3 Oxidant mit einer Oxidationszeit von 3 min. Deutlich ist eine dickere Zwischenschicht bei der Prozessierung mit O_3 zu erkennen. Die Abhängigkeit der Schichtdicke von der Pulsdauer lässt sich mit der längeren wirksamen Oxidationszeit des SiN erklären. Zu Beginn der Abscheidung wird für eine kurze Pulszeit relativ schnell eine geschlossene Al_2O_3-Schicht abgeschieden und eine SiN-Oxidation wird abgeschnürt. Es lässt sich aus den elektrischen Messungen und den TEM-Aufnahmen schließen, dass H_2O mit 3 min und O_3 mit 30 s Oxidationszeit ein ähnliches Verhalten bezüglich der Grenzfläche zeigen sollten. Generell wird von Kim [108] festgestellt, dass das Aluminiumoxid mit Ozon einen geringeren Leckstrom hat, was auch mit der geringen Konzentration an Verunreinigungen liegen kann, wie von Jakschik gezeigt [109]. Dies hat einen großen Einfluss auf die Ladungshaltung. Allerdings war bei den untersuchten Proben das Tunneloxid so dünn, dass eine Klassifizierung der Abscheidebedingung hinsichtlich Leckstromverhalten nicht möglich war. Prinzipiell unterscheidet sich die Abscheidung mit O_3 Präkursor in einer dünneren SiO_2-Grenzfläche zu Si oder SiN [110], die sich zu Beginn bildet und das elektrische Verhalten verbessert [106,111].

Eine weitere Größe, die die Eigenschaften des Aluminiumoxids bestimmt, ist die Temperatur eines Temperungsschrittes, der im Anschluß an die Abscheidung erfolgt. Jakschik [112] zeigt, dass ein Temperungsschritt mit einer Temperatur größer $\approx$ 920°C zu einer Kristallisation des Aluminiumoxids führt. Elektronenmikroskop-Aufnahmen bestätigen dieses Verhalten, wie Abb. 4.16 illustriert. Deutlich ist die nach 850°C Temperung noch amorphe Aluminiumoxid-Schicht in Abb. 4.16a von der kristallinen Schicht in b zu unterscheiden, welche bei einer Temperatur von 950°C getempert wurde. Es ist daher zu erwarten, dass sich auch die elektrischen Eigenschaften bei

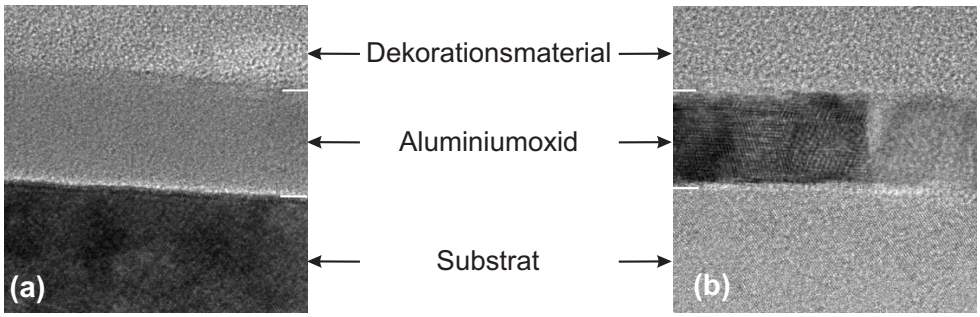

Abbildung 4.16: TEM-Aufnahmen von Aluminiumoxid bei (a) 850°C und (b) 950°C getempert

diesem Phasenübergang ändern, zudem sich bei der Kristallisation eine Schrumpfung um $\approx 10\%$ einstellt. Afanase'ev [113] zeigt, dass sich mit zunehmender Temperatur der Leckstrom durch das Dielektrika verringert. Setzt die Kristallisation ein, kommt es zu einer sprunghaften Abnahme des Leckstroms [114]. Cacciato [115] begründet dies durch eine Erhöhung des Leitungsband-Energieniveaus, worin eine Reduktion der Elektroneninjektion vom Gate resultiert. Abbildung 4.17 verdeutlicht, dass auch das Löschverhalten der gemessenen Proben auf die Temperbedingungen reagiert.

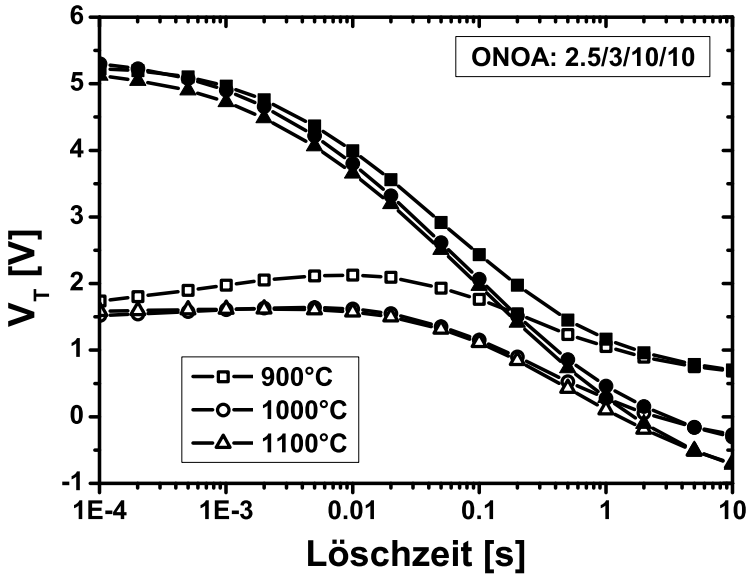

Abbildung 4.17: Löschverhalten von 5x5 μm großen SANOS-Zellen mit einer SiO_2 Schicht zwischen Speicherschicht und Aluminiumoxid für frische Zellen (offene Symbole) und $V_T = 5$ V (geschlossene Symbole); Schichtdicken: 2.5/3/10/10 nm (ONOA), das Aluminiumoxid wurde bei der genannten Temperatur 20 s formiert

Die Probe mit der niedrigsten Temperatur offenbart ein klar schlechteres Löschverhalten als die zwei anderen Vergleichsgruppen. Sowohl die Löschsättigung liegt höher, als auch die Löschgeschwindigkeit ist ein wenig langsamer. Das Ergebnis korreliert mit den Beobachtungen von Afanase'ev [113], wobei das langsamere Löschen und die Löschsättigung durch eine stärkere Elektroneninjektion von der Steuerelektrode erklärt werden kann. Eine weitere Erhöhung der Temperatur bewirkt noch einmal eine kleine Verbesserung des Löschverhaltens. Eine Betrachtung des Leckstromverhaltens

während kleiner Felder im Fall der Ladungshaltung konnte aufgrund der prozessierten Schichtdicken nicht durchgeführt werden.
Ein wichtiger Aspekt für die Prozessierung ist der Übergang von amorpher zu kristalliner Phase. Aluminiumoxid in amorpher Phase ist leicht durch Nassprozesse, auch Wasser, zu entfernen. Dies erfordert eine große Vorsicht bei der Prozessierung [116]. Im kristallinen Zustand hingegen handelt es sich um ein extrem thermisch und chemisch beständiges Material.
Ein weitere Einflussgröße auf die Eigenschaften des Aluminiumoxids, ist die Atmosphäre in der die Kristallisation stattfindet. Hierzu wurden zwei Versuche mit verschiedenen Gasen durchgeführt. Die Messung des programmierten Zustands im ersten Experiment, dessen Messdaten in Abb. 4.18a dargestellt sind, zeigen für die verschiedenen Gase keine Abhängigkeit.

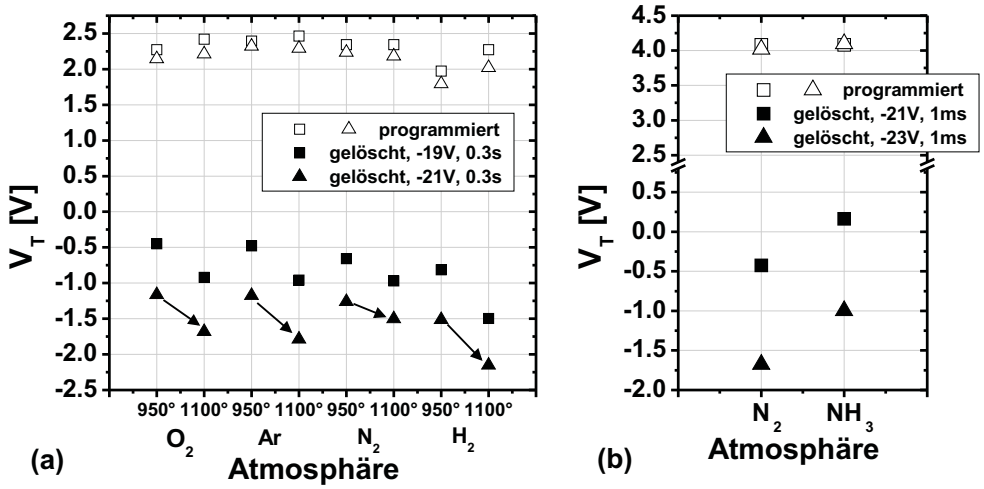

Abbildung 4.18: Abhängigkeit des Programmier-V_T's (geschl. Symbole) und des gelöschten Niveaus von der Temper-Atmosphäre bei zwei Experimenten; 48x48 nm Zellen; Gatematerial: TaN; ONA = 5/6/12 nm

Nur die H_2-angereicherte Atmosphäre liegt $\approx$ 200 mV unterhalb der anderen Gruppen. Es wird auch deutlich, dass die Atmosphäre vor allem einen Einfluss auf die Löschsättigung besitzt. So wurde bereits zuvor gezeigt, dass eine Erhöhung der Temperatur für die Kristallisation zu einer Verbesserung des Löschsättigungsniveaus führt. Es zeigt sich auch, dass das niedrigste Sättigungsniveau unter Verwendung einer Temper-Atmosphäre erreicht wird, die mit circa 10 % Wasserstoff angereichert ist. Dieses Verhalten ist unabhängig der Formierungstemperatur zu beobachten. Der Gewinn durch die H_2-Temperung bei 1100°C ist ein $\approx$ 400 mV niedrigeres Löschniveau, bei annähernd vergleichbarer Programmiergeschwindigkeit. Eine Erklärung für diese Beobachtung ist die Eigenschaft von Wasserstoff offene Bindungen abzusättigen. Diese offenen Bindungen befinden sich in der Schicht und an den Korngrenzen, da es sich um ein μ-kristallines Material handelt. Dadurch werden Leckpfade im Aluminiumoxid reduziert, und demzufolge auch die Injektion vom Gate über die Defekte unterdrückt. Die verringerte Gateinjektion wiederum bewirkt ein verbessertes Löschen, wie in Kap. 2.3.4 erläutert. Die anderen drei Gase aus dem ersten Experiment, Argon, Stickstoff und Sauerstoff haben ein vergleichbares Löschverhalten. Daraus kann geschlußfolgert werden, dass das sehr stabile Al_2O_3 durch diese Gase nicht beeinflusst wird. Die Untersuchungen von Jeon [117] zeigen ein entgegengesetztes Verhalten. Es

wird erläutert, dass mit einer Temperung in Sauerstoff-Atmosphäre das niedrigste Löschniveau erreicht wird. Allerdings sind die Temperaturen für das Tempern mit maximal 900°C noch in einem Bereich, bei dem es nur bedingt zu einer Kistallisation kommt.
In einem zweiten Experiment wurde untersucht, ob die Bildung einer oberflächennahen Nitridschicht, durch die Applikation einer NH_3-Atmosphäre zu einer Verbesserung des Löschverhaltens führt. Die durch Chen [118] gezeigte Verbesserung durch die Änderung der Schichteigenschaften kann nicht beobachtet werden. Im Vergleich zu einer Stickstoff-Atmosphäre wird das erreichte Löschniveau der untersuchten Speicherzellen bei gleicher Pulszeit deutlich zu höheren V_T's verschoben. Die Messergebnisse sind in Abb. 4.18b dargestellt. Die Ursache für die Verschlechterung des Löschverhaltens ist die gebildete Nitridschicht. Diese hat ein kleineres ϵ_r als die Aluminiumoxidschicht. Dadurch wird die Ladungsinjektion von der Gateelektrode größer und die Löschgeschwindigkeit reduziert (siehe Kap. 2.3.4). Weiterhin besteht ein Unterschied bei der verwendeten Gateelektrode. Die bei Chen verwendete Poly-Silizium Elektrode hat bei der Verwendung auf einer Al_2O_3-Schicht eine von der Schichtzusammensetzung abhängige Austrittsarbeit (siehe Kap. 4.2.3), anders als die im Experiment verwendete TaN-Elektrode. Daher ist eine Verbesserung mit einer Poly-Elektrode zu erwarten, im Gegensatz zur Metallelektrode in dem Experiment. Weiterhin wurden deutlich niedrigere Temperaturen für die Formierung verwendet, wodurch sich auch ein Einfluss ergibt. Abschließend kann festgehalten werden, dass eine Wasserstoff angereicherte Atmosphäre mit einer Temperung bei 1100°C zu den besten Löschergebnissen führt.

4.2.3 Al_2O_3-Topoxid unter Zugabe von SiO_2

Die Anreicherung von Aluminiumoxid mit Silizium ist in einem weiten Bereich möglich und existiert in einer Vielzahl von möglichen Kristallstrukturen [119]. Aluminiumoxid besitzt die Fähigkeit bis zu einem Anteil von circa 15 Atomprozent (At%) Silizium auf Zwischengitterplätzen aufzunehmen. Zudem zeigt Lanza [114], dass das eingebaute Si auch Störstellen an Korngrenzen des Al_2O_3 absättigt. Ein Anteil von 15 At% Silizium entspricht in etwa einer Abscheidepulsrate von Al:Si = 2:1. Dabei entspricht das Kristallisationsverhalten nahezu dem von Al_2O_3 und auch die Schrumpfungsrate nach dem Formierschritt mit 1100°C 20 s bleibt bei $\approx 10\%$. Erhöht man den Anteil von Silizium weiter, beginnt der Einbau von Silizium in die Kristallstruktur und es bilden sich Silikate. Um den Einfluss des Siliziumgehaltes auf die Eigenschaften von Speicherzellen zu untersuchen wurde ein Versuch durchgeführt, bei dem der Gehalt von Silizium zwischen 5 At% (Al:Si = 5:1) und 45 At% (Al:Si = 1:4) variiert wurde. Durch die Zugabe von Si ändert sich auch die relative dielektrische Konstante ϵ_r, welche für die Gruppen errechnet wurde und in Tab. 4.1 gezeigt ist.
Das ϵ_r wurde mit der Formel

$$\epsilon_{r,TO} = \frac{d_{TO} \cdot \epsilon_{SiO}}{d_{EOT} - d_{BO} - d_{SiN}\frac{\epsilon_{SiO}}{\epsilon_{SiN}}} \tag{4.6}$$

errechnet, wobei folgende Parameter vorgegeben werden, d_{BO} = 2.5 nm, d_{SiN} = 8 nm und d_{TO} = 8 nm. Es bestätigen sich die Beobachtungen von Gusev [120], der eine relative Dielektrizitätskonstante ϵ_r größer 9 für Aluminiumoxid extrahiert hat [116]. Erhöht man nun den Anteil an Silizium bis zu einem Prozentsatz von 10

Tabelle 4.1: Abhängigkeit des Si-Gehaltes und der relativen Dielektrizitätskonstante ϵ_r von dem Abscheideverhältnis (AV), ermittelt aus dem EOT

AV (Al:Si)	At%-Si	ϵ_r, TO	EOT(nm)
1:0	0%	10.2	9.7
5:1	5%	9.9	9.8
3:1	10%	9.3	10
1:1	25%	7.2	11
1:4	45%	4.4	13.7

At%, ändert sich zunächst das ϵ_r nur geringfügig. Dies ist darauf zurückzuführen, dass sich in diesem Bereich das eingebaute Silizium auf Zwischengitterplätzen oder an den Korngrenzen befindet und zunächst nur einen kleinen Einfluss auf die elektrischen Eigenschaften hat. Setzt allerdings die Silikatbildung ein, nimmt das ϵ_r rapide ab und nähert sich schnell dem Wert 3.9 von SiO_2 an. Bei einer gleichen physikalischen Dicke resultiert die Abnahme der Dielektrizitätskonstante des Topoxids in einem kleineren elektrischen Feld im Tunneloxid. Das Ergebnis ist eine deutliche Abnahme der Programmiergeschwindigkeit, wie Abb. 4.19a verdeutlicht.

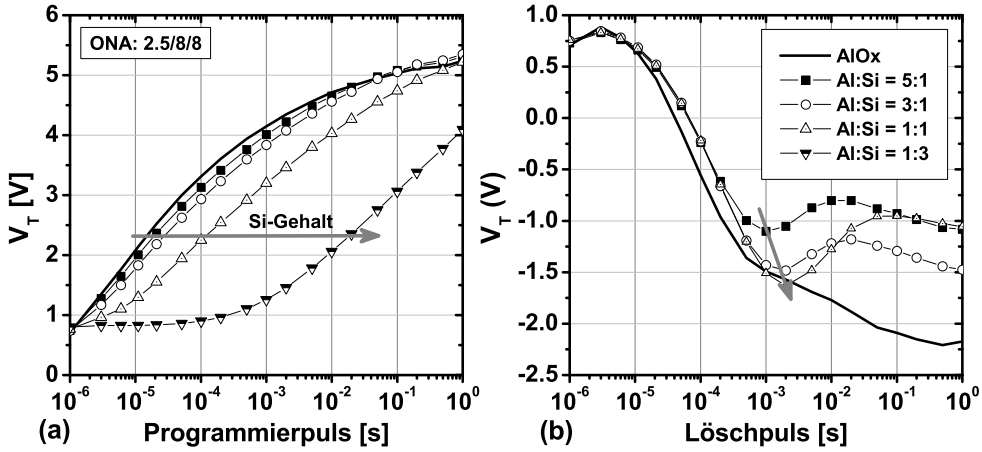

Abbildung 4.19: (a) Programmier- und (b) Löschcharakteristik von 5x5 μm großen SANOS-Zellen mit Aluminiumoxid, welches einen variierenden Si-Gehalt aufweist, angegeben ist das Abscheide-Pulsverhältnis Al:Si

Die Änderung in der Programmiercharakteristik lässt sich einfach auf den Unterschied in der equivalenten elektrischen Stapeldicke (EOT) zurückführen. Beim Löschen ist eine Auswertung schwieriger. Grundsätzlich verschlechtert die Zugabe von SiO_2 zu Al_2O_3 die Löschcharakteristik. Interessanterweise sättigt die Gruppe mit dem niedrigsten Siliziumanteil auf dem höchsten Niveau, wie Abb. 4.19b zeigt. Eine zunehmende Beimischung verbessert wieder das Löschverhalten, wobei die Gruppe mit einem Abscheidverhältnis von 1:3 nicht gezeigt wird, da das Löschen dort sehr langsam erfolgt und ein Vergleich nicht sinnvoll ist.
Das Löschverhalten ist durch zwei Effekte geprägt. Einerseits muss es eine Änderung der Austrittsarbeit an der p^+-poly Steuerelektrode geben, da sich die Löschsättigungsniveaus verschieben. Andererseits muss auch das effektive Feld im gesamten Schichtstapel, welches mit zunehmendem Siliziumgehalt abnimmt, einen Einfluss ha-

ben. Bei dem Übergang von Metalloxiden zu poly-Silizium tritt ein Effekt auf der darin resultiert, dass die Austrittsarbeitsdifferenz ϕ^e_{S-O} nicht den Werten entspricht, die sich aus getrennten Betrachtungen für die Materialien ergeben. Zu Beginn wurde dieser Effekt auf Ladung an der Grenzfläche zurückgeführt [121]. Später wurde der Effekt 'Fermi-Level-pinning' genannt [122–124]. In den Betrachtungen zu diesem Effekt wurde festgestellt, dass sich die Austrittsarbeit der Silizium-Gateelektrode in Richtung der Bandmitte verschiebt, wenn sich der Anteil von SiO_2 an der Elektrodengrenzfläche verringert [122]. Angewendet auf die Messung bedeutet das, dass sich ein höherer Siliziumanteil positiv auf die Austrittsarbeit auswirkt und somit eine niedrigere Löschsättigung zu beobachten ist. Dieses Verhalten wird durch die Messung in Abb. 4.19b wiedergegeben. Reines Al_2O_3 zeigt ein davon abweichendes Verhalten, was darauf zurückgeführt wird, dass die Grenzfläche nicht durch zusätzliche Si-Si Bindungen gestört ist. Schaut man sich das Verhalten während eines Zyklentests an, stellt sich auch ein interessantes Ergebnis ein. Abbildung 4.20 zeigt für die vier bereits beim Löschen untersuchten Gruppen den gelöschten und programmierten Zustand nach einem konstanten Puls in Abhängigkeit von der Zyklenzahl.

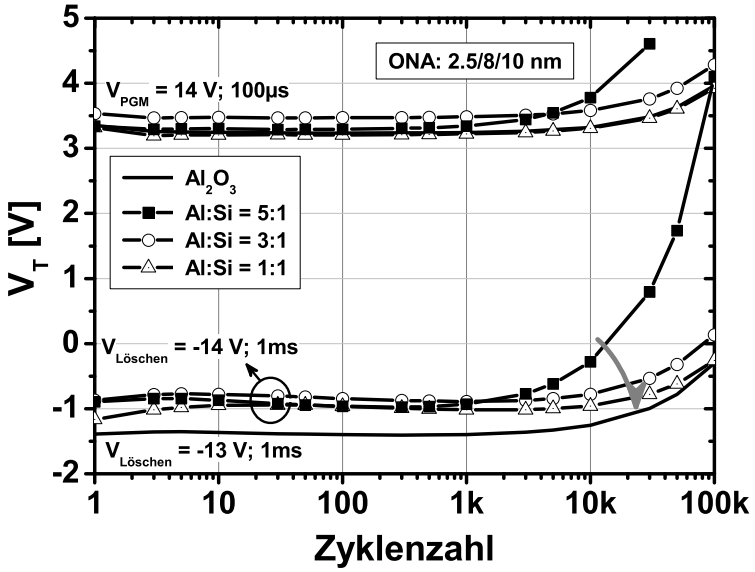

Abbildung 4.20: Zyklen-Abhängigkeit des gelöschten und programmierten V_T's vom Gehalt an Si im Al_2O_3 für 4 verschiedene SANOS-Proben

Für den programmierten Zustand ist der Unterschied nicht groß, wenn man von der Gruppe mit dem kleinsten Siliziumgehalt (5:1) absieht. Der Unterschied im gelöschten Zustand ist größer. Dabei fällt auf, dass das reine Al_2O_3 trotz niedrigerer Löschspannung am tiefsten löscht. Betrachtet man die anderen Gruppen, so wird deutlich, dass ein zunehmender Siliziumgehalt die Degradation deutlich verringert. Offensichtlich reduziert zunächst eine geringe Beimischung von Silizium die elektrische Stabilität des Aluminiumoxids. Wird der Silizium-Anteil erhöht, kommt es zu einer Anhäufung von Silizium an den Korngrenzen, wodurch dort freie Bindungen abgesättigt werden und damit die Grenzfläche stabilisiert wird. Es resultiert eine bessere elektrische Widerstandsfähigkeit und demzufolge Zyklenfestigkeit. Weiterhin vergrößert ein höherer Si-Anteil die elektrisch wirksame Dicke durch die Verringerung von ϵ_r und somit auch die elektrischen Felder im Tunneloxid während des Löschens, welche für die Degradation verantwortlich sind [125] (siehe Kap. 2.3.4).

Abschließend kann festgehalten werden, dass eine Beimischung von Silizium in das Al_2O_3 zu keiner Verbesserung der elektrischen Eigenschaften führt.

4.2.4 Integration einer SiO_2-Zwischenschicht

Eine Alternative zur Beimischung von SiO_2 stellt die Strukturierung des Topoxids mittels zweier getrennter Schichten aus Al_2O_3 und SiO_2 dar [126–129]. Man versucht hierdurch die Vorteile beider Schichten zu vereinen. Abbildung 4.21 verdeutlicht den Aufbau.

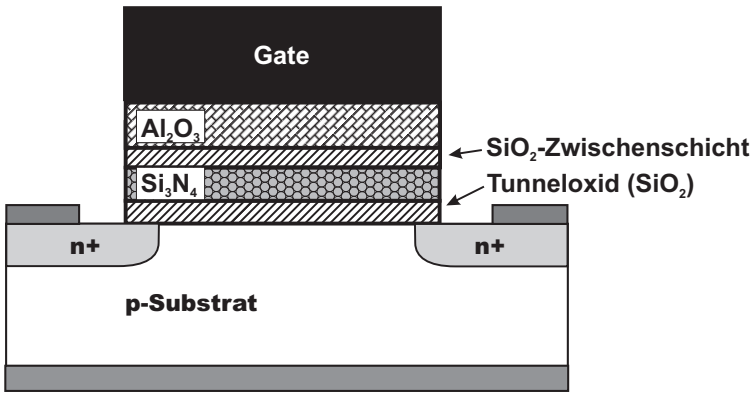

Abbildung 4.21: Aufbau einer TANOS-Speicherzelle mit einer SiO_2-Zwischenschicht für die Verbesserung der Ladungshaltung

Man nimmt das Al_2O_3 auf der Elektrodenseite, damit die Elektroneninjektion von dieser unterdrückt wird, wie in Kap. 2.3.4 beschrieben. Andererseits wurde in Kap. 3.3.2 gezeigt, dass der Hauptleckpfad von TANOS-Strukturen durch das Al_2O_3-Topoxid führt. Daher befindet sich bei der untersuchten Struktur die SiO_2-Zwischenschicht direkt an der Speicherschicht, damit ein Ladungsverlust über das Topoxid aufgrund der höheren Elektronenbarriere reduziert wird (engl. sealing-layer). Einerseits besitzt das SiO_2 eine geringere Fehlstellendichte und unterdrückt damit den Ladungsverlust durch das relativ schnelle Tunneln über diese Fehlstellen (engl. trap-assisted tunneling - TAT). Andererseits wird das direkte Tunneln durch die höhere Elektronenbarriere reduziert. Diese theoretische Betrachtung wird durch den in Abb. 4.22 durchgeführten Vergleich einer einfachen TANOS-Struktur mit einer TANOS-Zelle, die eine Zwischenschicht von 3.5 nm enthält, bestätigt. Die Ladungshaltung auf großen Speicherzellen kommt mit einer V_T-Verschiebung von 40 mV nach 2 h bei 200°C in die Größenordnung, wie sie auch bei der Floating-Gate Struktur beobachtet werden. Die normale TANOS-Struktur zeigt für die großen Zellen einen um 200 mV größeren Verlust auf und bestätigt damit die Annahme, dass die Ladung zu einem großen Teil über das Topoxid abfließt. Aber die Integration dieser Schicht bringt nicht nur Vorteile, wie Abb. 4.22b bei der Betrachtung der Löschtransienten offenbart. Der Einbau einer SiO_2-Zwischenschicht verschlechtert die Löschcharakteristik deutlich. Es wurden zum Vergleich Gruppen mit vergleichbarem EOT ausgewählt, was sich darin äußert, dass der Beginn des Löschvorgangs nahezu identisch ist. Allerdings beginnt die Löschkurve für die Gruppe mit Zwischenschicht schon nach kurzer Zeit zu sättigen. Die große Speicherzelle läßt sich innerhalb einer akzeptablen Löschzeit nicht unter ein V_T von 1 V löschen. Dieses Phänomen der begrenzten Löschbarkeit wurde bei vergleichbaren

Experimenten nicht in dieser Deutlichkeit beobachtet [126, 128] und erfordert demnach weitergehende Untersuchungen. Eine mögliche Erklärung für den Unterschied zu den anderen Experimenten ist die Durchführung auf Basis von Transistoren, während die vergleichbaren Daten auf Kondensatoren erstellt wurden.

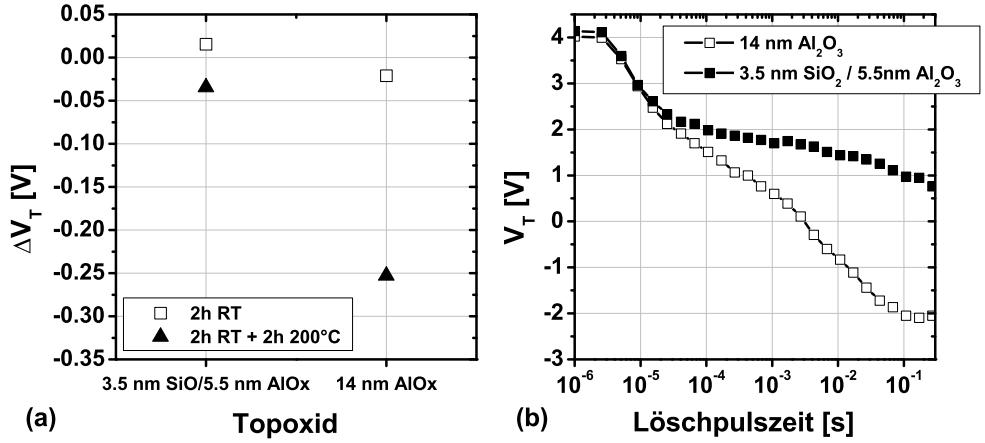

Abbildung 4.22: Vergleich der Ladungshaltung (a) und Lösch-Transienten (b) von 5x5 μm großen TANOS-Speicherzellen, einmal mit und ohne SiO_2-Zwischenschicht; bei der Ladungshaltung zuvor auf 5 V V_T programmiert

4.3 Auswirkung des Steuerelektrodenmaterials auf das Zellverhalten

Die Steuerelektrode ist ein elementarer Bestandteil einer MOS-Struktur, welche einen großen Einfluss auf das Zellverhalten ausübt. Lange Zeit war es ausreichend, ein n^+-dotiertes poly-Silizium zu verwenden. Durch die schnell fortschreitende Miniaturisierung ist aber eine Verwendung von metallischen Elektroden unabdingbar. Diese haben den Vorteil, dass man bei entsprechender Materialwahl durch deren Austrittsarbeit die Schwellspannung des Transistors einstellen kann. Bei Verwendung von poly-Silizium und Einstellung mit Hilfe der Implantationsdosis würde auch die Leitfähigkeit und demzufolge die RC-Konstante der Gateelektrode verschlechtert. Bei der Anwendung in Mikroprozessoren ist dies ein Grund, welcher die maximale Taktfrequenz bestimmt. Die Einführung von Metalloxiden als Isolatoren mit hohen dielektrischen Konstanten impliziert eine Verwendung von metallischen Elektroden aufgrund der besseren Materialverträglichkeit. Im folgenden Abschnitt wird zunächst die Anwendung von poly-Silizium diskutiert. Daran schließt sich eine Betrachtung von metallischen Gateelektroden, im speziellen TiN und TaN, an.

4.3.1 Poly-Silizium-Gateelektrode

Bei poly-Silizium handelt es sich um ein abgeschiedenes Silizium, welches in einer μ-kristallinen Form vorliegt. Das Material ist aufgrund seiner einfachen Prozessierbarkeit nach wie vor das bevorzugte Elektrodenmaterial. Bei der entsprechenden Implantation von Dotierstoffen kann die Schwellspannung für PMOS und

NMOS-Transistoren in Logikschaltungen auf einfache Art und Weise angepasst werden [130,131]. Dies beruht auf dem Effekt, dass die Schwellspannung eines Transistors auch durch die Austrittsarbeitsdifferenz Φ_{MS} zwischen Siliziumsubstrat und der Gateelektrode bestimmt ist. Die Gl. 4.7 zur Bestimmung der Schwellspannung zeigt für ein ladungsfreies Oxid diesen Zusammenhang:

$$V_T = \Phi_{MS} + 2\Psi_B + \frac{\sqrt{2\epsilon_s q N_A (2\Psi_B)}}{C_{ox}}. \tag{4.7}$$

Im Fall von Speicherzellen mit poly-Elektrode ist es günstiger, eine p^+-Dotierung zu wählen. Da die Austrittsarbeit größer ist, lassen sich die Zellen durch die erhöhte Elektronenbarriere an der Gateelektrode tiefer löschen, wie bereits in Kap. 2.3.4 erläutert. Dies wird in Abb. 4.23, durch einen Vergleich von n^+- und p^+-poly Elektroden veranschaulicht.

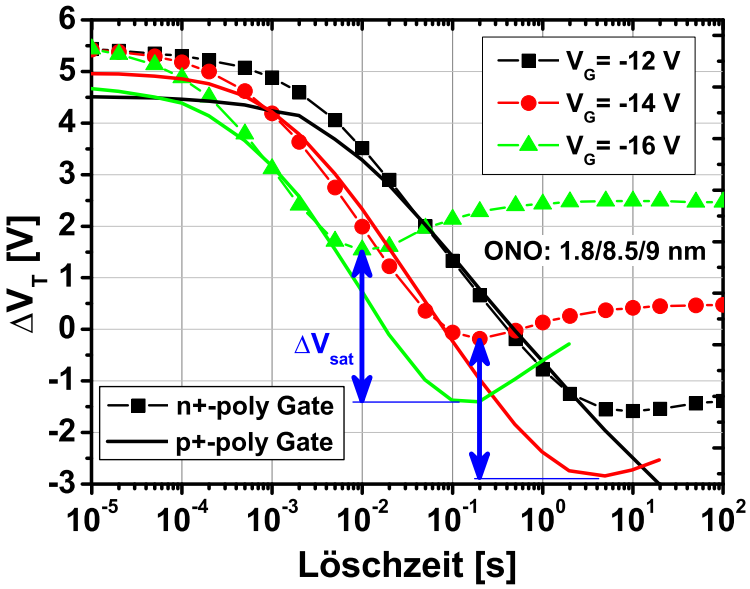

Abbildung 4.23: Vergleich des Löschverhaltens von einer p^+-poly (mit Punkten) und n^+-poly Gateelektrode (dick) für drei verschiedene Löschspannungen; ΔV_{sat} bezeichnet die Spannungsdifferenz des jeweiligen Löschsättigungsniveaus

Abbildung 4.23 zeigt das typische Löschverhalten einer SONOS-Struktur. Die verhältnismäßig gute Löschbarkeit der Zellen basiert auf der geringen Dicke des Tunneloxids von 1,8 nm. Zu Beginn des Löschvorganges unterscheiden sich die unterschiedlich dotierten Gateelektroden nicht. Es wird deutlich, dass erwartungsgemäß die Elektrode mit höherer Austrittsarbeit auf einem niedrigeren Niveau sättigt. Die größere Tunnelbarriere für Elektronen von der Gateelektrode ist die Erklärung hierfür. Denn die Elektroneninjektion vom Gate hebt die Wirkung der durch das Tunneloxid injizierten Löcher auf, und es kommt zu einer Löschsättigung. Der Abstand der Löschsättigung ΔV_{sat} zwischen den verschieden dotierten Elektroden bleibt hingegen konstant. Historisch gewachsen, erfolgt die Prozessierung der p^+-poly Elektrode mittels Abscheidung eines undotierten Siliziums mit anschließender Dotierung. Für eine ausreichend hohe Dotierung, zur Minimierung der Verarmung, sind mehrere Prozessschritte erforderlich. Eine Alternative, die zu einer beträchtlichen Reduktion der Schrittzahl führt, ist eine Abscheidung von hoch dotiertem Silizium (in-situ). Ein weiterer Vorteil ist die

große Homogenität, die eine Diffusion durch folgende thermische Prozesse verkleinert. In Abb. 4.24 wird ein Vergleich einer Probe mit dotiertem und in-situ abgeschiedener poly-Silizium Elektrode durchgeführt.

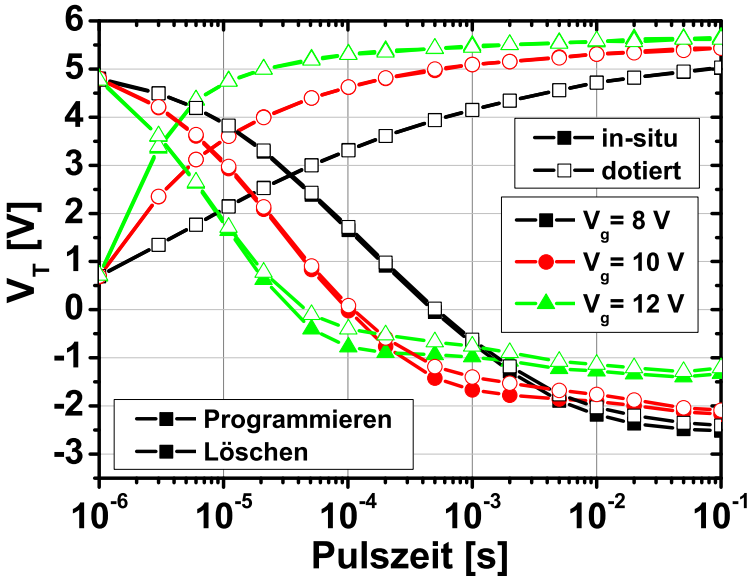

Abbildung 4.24: Vergleich des Löschverhaltens von einer Gateelektrode mit dotiertem p^+-poly (offene Symbole) und dotiert abgeschiedenem (in-situ) p^+-poly Silizium (geschlossene Symbole) für drei verschiedene Spannungen; ONA: 2.5/8/10 nm

Das Programmierverhalten der beiden Elektrodenvarianten unterscheidet sich nicht, da das Programmieren durch die Gateelektrode nahezu nicht beeinflusst wird. Beim Löschen zeigt sich ein kleiner Unterschied zu Beginn der Löschsättigung. Hier verläuft die Löschkurve der in-situ p^+-poly Gateelektrode circa 300 mV unter der dotierten Elektrode. Somit hat die bereits bei der Abscheidung dotierte Elektrode einen geringen Vorteil. Bei langen Zeiten wird der Abstand wieder geringer, aber dieser Zeitbereich ist für den Betrieb von Speicherzellen nicht mehr interessant.
Ein entscheidender Nachteil von einer poly-Gateelektrode ist deren nicht ideal metallisches Verhalten. Werden entsprechend große Felder durch eine angelegte Spannung in die Struktur eingebracht, kommt es bei entsprechender Polarität zu einer Verarmung der Gateelektrodenoberfläche [132]. Diese wiederum führt im Falle einer p^+-Dotierung zu einer Verringerung der effektiven Austrittsarbeit, wenn eine Löschspannung angelegt wird. Im Gegensatz dazu würde man theoretisch bei einer n^+-Dotierung eine Erhöhung der Austrittsarbeit erwarten. Diesem Verhalten kann auch nicht durch eine Erhöhung der Implantationsdosis entgegengewirkt werden, da folgende thermische Prozesse durch Diffusion zu einer Verringerung der wirksamen Dosis an der Oberfläche führen. Die folgende Tabelle 4.2 zeigt die Bandverbiegung und Tiefe der Verarmungszone für zwei Feldstärken.
Die theoretische Betrachtung basiert auf den in Sze [12] hergeleiteten Formeln, wobei sich die Weite der Zone mit reduzierter wirksamer Dotierung durch

$$w_{ver} = \sqrt{\frac{2\phi_s\epsilon_s}{qN_A}} \tag{4.8}$$

Tabelle 4.2: Bandverbiegung und Dicke des Verarmungs- bzw. Inversionsgebietes w_{ver} in hochdotiertem Silizium für eine Oxidfeldstärke von 10 und 15 MV/cm; inv ≅ Inversion; s _ inv ≅ schwache Inversion; ver ≅ Verarmung

$N_A(cm^{-3})$	$E_{SiO_2} = 10MV/cm$	$E_{SiO_2} = 15MV/cm$
$5e^{18}$	1.16 eV (inv) $w_{ver} = 17.4nm$	1.18 eV (inv) $w_{ver} = 17.6nm$
$1e^{19}$	1.17 eV (inv) $w_{ver} = 12.4nm$	1.19 eV (inv) $w_{ver} = 12.5nm$
$5e^{19}$	0.73 eV (s_inv) $w_{ver} = 5.6nm$	1.2 eV (inv) $w_{ver} = 5.6nm$
$1e^{20}$	0.37 eV (ver) $w_{ver} = 2.2nm$	0.82 eV (s_inv) $w_{ver} = 3.3nm$
$2e^{20}$	0.2eV (ver) $w_{ver} = 1.15nm$	0.42 eV (ver) $w_{ver} = 1.66nm$
$4e^{20}$	0.11 eV (ver) $w_{ver} = 0.6nm$	0.23 eV (ver) $w_{ver} = 0.87nm$

ergibt.
Es wird verdeutlicht, dass selbst bei den größtmöglichen Dotierungen von 1 - 2 e^{20} cm^{-3} [133] noch eine Bandverbiegung von ≈ 300 mV auftritt. Für eine p^+-dotierte Gateelektrode bedeutet dies im Fall des Löschens, dass die wirksame Austrittsarbeit von theoretischen 5.1 eV um diesen Betrag auf etwa 4.8 eV verringert wird. Tabelle 4.2 zeigt zudem, dass sich bei Erhöhung der Feldstärke, und somit der Gatespannung, die Bandverbiegung vergrößert. Gemäß der Abhängigkeit der Löschsättigung von der Elektroneninjektion an der Gateelektrode beobachtet man eine Verschiebung des Sättigungsniveaus in Abhängigkeit der Gatespannung. Ein Vergleich der Löschcharakteristik einer Struktur mit p^+-poly Elektrode und einer mit metallischer TaN Elektrode, bei der keine Bandverbiegung auftritt, verdeutlicht diesen Effekt. Es ist gezeigt, dass mit zunehmender negativer Gatespannung die Löschsättigung zeitiger, im Bezug auf die Löschpulszeit und auf einem höheren Niveau einsetzt [35, 134, 135]. Dadurch ergibt sich eine Verschiebung des Löschsättigungsniveaus um 4.5 V bei einer Erhöhung der Gatespannung von -12 V auf -20 V. Im Gegensatz dazu ändert sich bei der TaN-Metallelektrode die Löschsättigung nur um circa 0.7 V.
Dadurch nimmt auch die Löschsättigungs-Spannungsdifferenz ΔV_{sat} mit erhöhter Gatespannung zu. Die Änderung bei der Metallelektrode lässt sich darauf zurückführen, dass sich bei einem höheren Feld, durch eine stärkere Bandverkippung, die wirksame Elektronen-Tunnelbarriere verkleinert.

4.3.2 Metall-Gateelektrode

Es wurde bereits im vorangegangenen Abschnitt ausgeführt, dass Metall-Gateelektroden Vorteile gegenüber einer poly-Silizium-Gateelektrode haben. Aus diesem Grund gab es in den vergangenen Jahren große Anstrengungen geeignete Materialien zu finden [136–138]. Ein wichtiger Aspekt bei der Integration ist die thermische Stabilität, die eine Vielzahl von Materialien ausschließt [139]. Eine Möglichkeit die Problematik der thermischen Stabilität zu umgehen, ist die Abscheidung

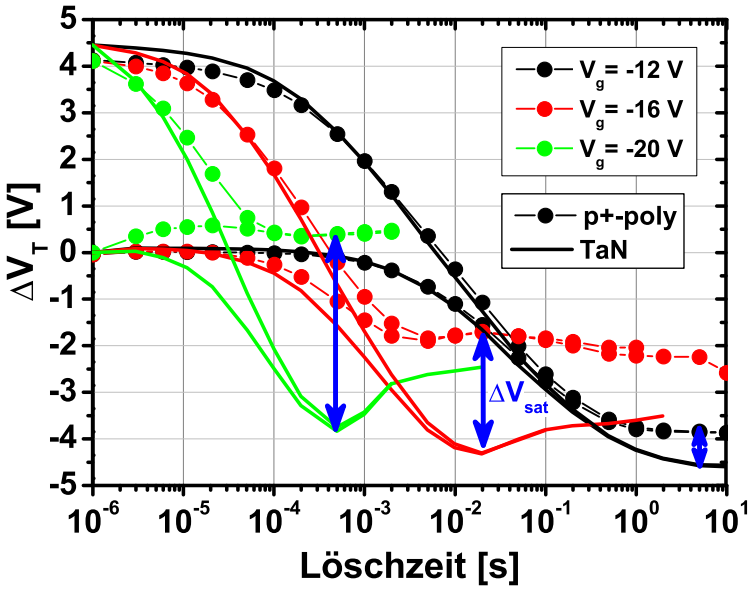

Abbildung 4.25: Vergleich des Löschverhaltens einer Speicherzelle mit den Schichtdicken ONA: 2.8/9/12 nm für drei verschiedene Löschspannungen; ΔV_{sat} bezeichnet die Spannungsdifferenz des jeweiligen Löschsättigungsniveaus zwischen der p^+-poly (mit Punkten) und TaN Gateelektrode (dick)

der Gateelektrode nachdem sämtliche Hochtemperaturprozesse, wie zum Beispiel die Source/Drain-Aktivierung, durchgeführt wurden (replacement-Gate) [140, 141]. Eine Umsetzung des replacement-Gate-Prozesses für Speicherzellen ist nicht interessant, da dies die Prozesskomplexität erheblich steigern würde. Bei der Anwendung als Gateelektrode in Speicherzellen haben sich daher die beiden thermisch stabilen Materialien Tantalnitrid (TaN) und Titannitrid (TiN) herauskristallisiert. In den folgenden Abschnitten wird die Anwendung dieser Elektroden mit unterschiedlichen Abscheidevarianten untersucht.

4.3.2.1 Tantalnitrid aus chemischer Gasphasenabscheidung

Bei Tantalnitrid (TaN) handelt es sich um eine keramische Verbindung, welche in mehreren Modifikationen vorkommt (Ta_2N;TaN;Ta_5N_6;Ta_3N_5; ...) [142–144]. Hauptsächliche Anwendung findet TaN, ebenso wie TiN, als Diffusionsbarriere für die Kupfermetallisierung von Hochgeschwindigkeits-Logikschaltkreisen [145–147]. TaN ist für Speicherzellen neben der thermischen Stabilität interessant, weil es eine relativ große Austrittsarbeit in der Größenordnung 4.8 eV besitzt [133, 136, 148, 149]. Besonders zu beachten ist die Tatsache, dass es sich bei abgeschiedenem TaN um eine poröse Schicht handelt, die stark zur Oxidation neigt [144]. Daher ist es außerordentlich wichtig, die Schicht im Anschluss an die Abscheidung mit einer Schutzschicht zu versiegeln. In unserem Fall wurde eine auf Silan basierende Kapselungsschicht abgeschieden.
Das verwendete TaN wurde auf einer Anlage von TEL mit der Bezeichnung TRIAS abgeschieden. Es handelt sich um eine Einzel-Wafer-Beschichtungsanlage für Metalle und Halbmetalle auf Basis der chemischen Gasphasenabscheidung (engl. chemical vapor deposition - CVD). Hierbei wird zwischen zwei Betriebsmodi unterschieden. Es kann einmal ein kontinuierlicher Abscheideprozess mit einer hohen Abscheiderate

verwendet werden, der im Folgenden mit CVD abgekürzt wird. Als weitere Möglichkeit kann ein gepulster Prozess verwendet werden, bei dem die Abscheiderate kleiner ist (p-CVD). Die Prozesstemperatur kann von 550°C bis 640°C variiert werden.

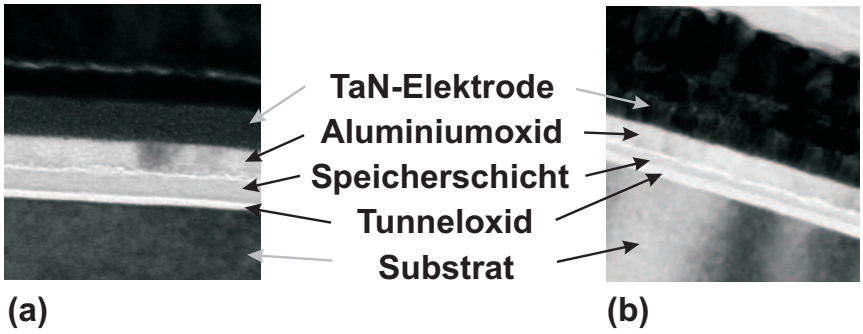

Abbildung 4.26: (a) TEM-Aufnahme einer TANOS-Struktur, wobei das TaN mit einer gepulsten CVD abgeschieden wurde und amorph ist; (b) gleiche Struktur mit einem kristallinen TaN, welches durch eine kontinuierliche CVD abgeschieden wurde

Ein Vergleich der beiden Abscheidemethoden erfolgt in Abb. 4.26 anhand von TEM-Aufnahmen. Es ist ersichtlich, dass bei der gepulst abgeschiedenen TaN-Elektrode die Schicht zum Ende des Fertigungsprozesses noch amorph ist, wohingegen die kontinuierlich abgeschiedene Schicht kristallin ist. Weitergehende Untersuchungen haben gezeigt, dass die kontinuierliche Abscheidung zu einer Kristallisation während der Abscheidung führt. Im Fall des gepulst abgeschiedenen TaN haben auch auf die Abscheidung und die Formierungstemperung folgende Hochtemperaturprozesse zu keiner Kristallisation geführt. Eine Erklärung für das Verhalten ist die unterschiedliche Materialzusammensetzung der Schichten, wie sie in Tab. 4.3 gezeigt wird.

Tabelle 4.3: Elementbestimmung mittels ERDA (elastische Rückstreuungsanalyse mit hochenergetischen Schwerionen) und spezifischer elektrischer Widerstand des abgeschiedenen TaN für die betrachteten Abscheideverfahren

Probe	Schicht dicke (nm)	Ta (at%)	O (at%)	N (at%)	C (at%)	H (at%)	ρ ($\mu\Omega$cm)
p-CVD (550°C)	21	24	25	20	22	8.5	59.95
CVD (550°C)	24	35	19	40	3	3	5.78

Als Ursache für die Unterdrückung der Kristallisation bei der gepulsten Abscheidung kann der hohe Anteil von Kohlenstoff in der Schicht angegeben werden [149]. Der eingebaute Kohlenstoff stammt von den verwendeten Reaktionsgasen, da diese einen hohen Kohlenstoffanteil besitzen. Das Vorhandensein von Kohlenstoff in der Schicht unterscheidet auch die chemische maßgeblich von der physikalischen Gasphasenabscheidung [142, 150]. Ein weiterer Effekt des hohen Kohlenstoffanteils ist ein deutlich erhöhter spezifischer elektrischer Widerstand ρ, wie Tab. 4.3 verdeutlicht. Diese Werte wurden bei den untersuchten Proben aus dem gemessenen Schichtwiderstand R_S berechnet. Dadurch vergrößert sich bei der gepulsten Abscheidung die RC-Konstante des Gateanschlusses, was unter Umständen zu Problemen bei kurzen Pulsen bzw. langen Leitungen führen kann. Die Dichte der Schichten ändert sich aber nicht und liegt

immer im Bereich von 8 g/cm. Aufgrund der unterschiedlichen Materialphasen ist zu erwarten, dass sich auch das elektrische Verhalten der Proben unterscheidet, wie auch Cacciato zeigt [151]. Jedoch konnte in den durchgeführten Untersuchungen auf großen Zellen mit einer Dimension von 5x5 μm nur eine vernachlässigbare Abhängigkeit des elektrischen Verhaltens beobachtet werden. Betrachtet man allerdings stark skalierte Speicherzellen mit einer Größe von 48x48 nm ergibt sich ein Bild für das Programmier- und Löschverhalten gemäß Abb. 4.27.

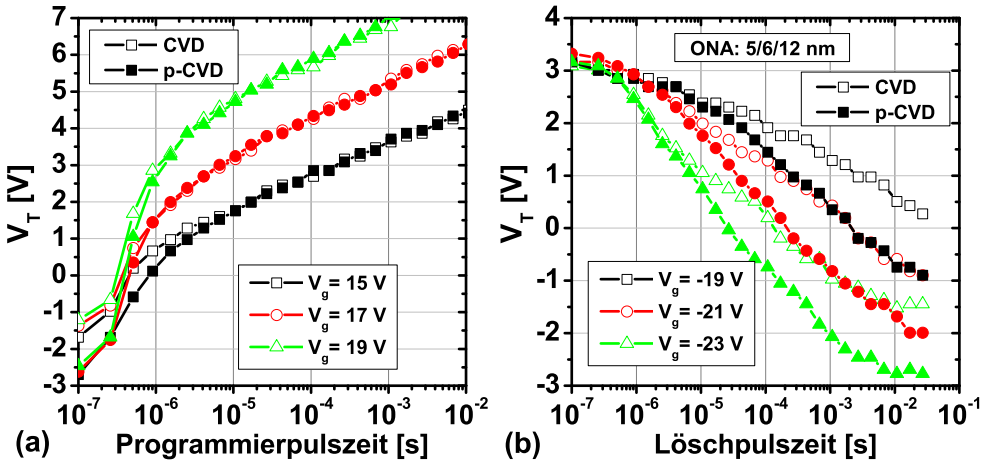

Abbildung 4.27: (a) Programmiercharakteristiken einer representativen Speicherzelle für kontinuierliche (offene Symbole)und gepulste (geschlossene Symbole) CVD-TaN Abscheidung bei drei Programmierspannungen; (b) zeigt die entsprechenden Löschcharakteristiken; $d_{TaN} = 20\ nm$

Der Vergleich der Programmiercharakteristiken in Abb. 4.27a verdeutlicht, dass die verschiedenen Schichtstapel gut miteinander vergleichbar sind, da die Unterschiede minimal sind. Nur zu Beginn unterscheiden sich die Kurven, da die Messungen für die gepulste Abscheidung von einem tieferen Löschniveau beginnen. Im Gegensatz dazu unterscheiden sich die Löschcharakteristiken, gezeigt in Abb. 4.27b, deutlich. Zu Beginn des Löschvorgangs ist das Verhalten gleich, da das Löschen bei allen drei Spannungen bei der gleichen Pulsdauer einsetzt. Im weiteren Verlauf löscht das gepulst abgeschiedene TaN deutlich besser. So ergibt sich ein Überschneiden der Löschcharakteristik mit -21 V und gepulster Abscheidung und der Kurve für -23 V und Standard-Abscheidung. Zudem setzt die Löschsättigung für die Standard-Abscheidung bei circa -1.5 V V_T ein und die Gruppe mit gepulstem TaN löscht bis zu einem V_T von -3 V. Dies verdeutlicht das bessere Löschverhalten des TaN mit gepulster Abscheidung. Als Ursache für den Unterschied kommen zwei Effekte in Frage. Das Tantalnitrid zeigt ein unterschiedliches Verhalten hinsichtlich des Ätzangriffes während der Strukturierung des Aluminiumoxids [152]. Eine genauere Analyse und Diskussion zum Ätzangriff findet in Kap. 5.1.3 statt. Durch den Ätzangriff kann es einmal zu einer Änderung des Materials selbst kommen. Dies wird allerdings durch Messungen an großen Zellen wiederlegt, da diese keine Abhängigkeit von dem Abscheideverfahren zeigen. Oder die Gateelektrode wird so modifiziert, dass sich entsprechend Kap. 4.1 eine inhomogene Ladungsverteilung ergibt, die das Löschen beeinflusst. Eine zusätzliche Auswertung mit der Betrachtung des sub-V_T Swings zeigt Abb. 4.28.
Es wird deutlich, dass der sub-V_T Swing zu Beginn des Löschens unabhängig von der Abscheidung bei circa 250 mV/dec liegt. Aber die Zunahme des sub-V_T Swing

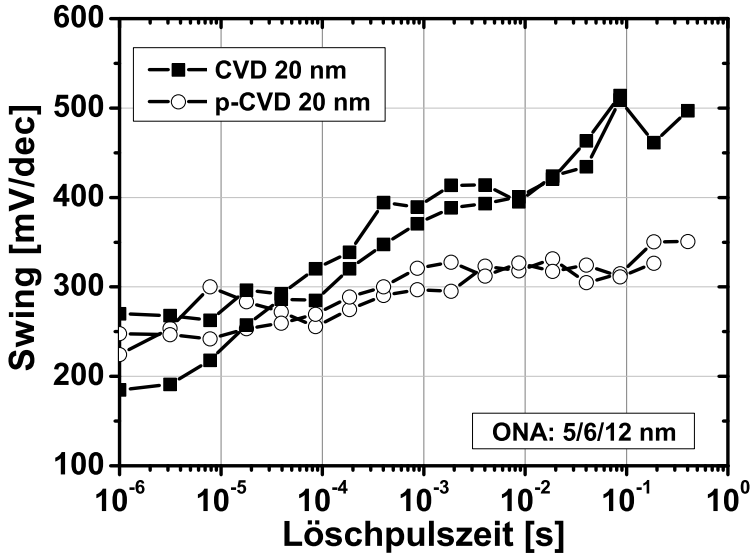

Abbildung 4.28: Auswertung des sub-V_T Swings während des Löschvorgangs auf 48x48nm Zellen für kontinuierliche(geschlossene Symbole) und gepulste (offene Symbole) CVD-TaN Abscheidung für jeweils zwei Proben

ist für die kontinuierliche Abscheidung deutlich stärker. Dieses Ergebnis bestätigt die Annahme, dass es sich um eine Veränderung der Gateelektrode handelt, die zu einer Änderung der Elektroden-Randbereiche führt. Dies resultiert in einem Verhalten, vergleichbar mit einer Ätzflanke im Aluminiumoxid. Paul [152] weist in seinen Untersuchungen nach, dass die Ätzung des Aluminiumoxids zu einer Schädigung des Tantalnitrids führt. Hierbei wird durch die Ätzgase vorrangig der Tantal-Anteil der Elektrode reduziert. Das Ergebnis ist ein Gemisch aus Stickstoff, Sauerstoff und Kohlenstoff mit dielektrischem Verhalten. Dadurch wird die effektive Länge der Elektrode verringert und es stellt sich eine ähnliche Struktur wie eine Speicherzelle mit Ätzflanke im Aluminiumoxid ein. Interessanterweise ist das kristalline Material empfindlicher als die amorphe Schicht. Eine Ursache hierfür könnte sein, dass der Ätzangriff mit einem Mal ganze Körner des Kristalls verändert. Hingegen kann bei der amorphen Schicht von einem kontinuierlichen Umbau der Schicht ausgegangen werden. Betrachtet man nun die Dicke der Schicht, ergibt sich für amorphes CVD-TaN das in Abb. 4.29 gezeigte Bild.

Das elektrische Verhalten wird erheblich besser, wenn die Dicke der Schicht vergrößert wird. Dies gilt sowohl für das Löschen als auch die Datenhaltung, wie Abb. 4.29b zeigt. Eine mögliche Erklärung für das beobachtete Verhalten ist ein abnehmendes Ausbleichen der Gateelektrode. Denn während des Ätzvorganges ist die Dichte der Radikale, welche zum Ausbleichen führen, konstant. Vergrößert man nun die Schichtdicke, wird die Oberfläche der Schicht größer und die erreichbare Tiefe kleiner. Da bei kristallinem TaN ganze Körner durch das Ausbleichen betroffen sind, sollte eine Verbesserung mit der Schichtdicke nicht zu beobachten sein, wie Abb. 4.30 bestätigt. Um die Verbesserung, die durch die Schichtdicke erreicht wird herauszustellen, ist es günstig die entscheidenden Parameter erreichtes gelöschtes V_T und den Ladungsverlust gegenüber zustellen. Trägt man die Parameter, wie in Abb. 4.30 dargestellt, auf, ist es möglich zu erkennen, ob man den Gewinn bei einem Parameter mit Einbußen bei dem anderen Parameter erkauft.

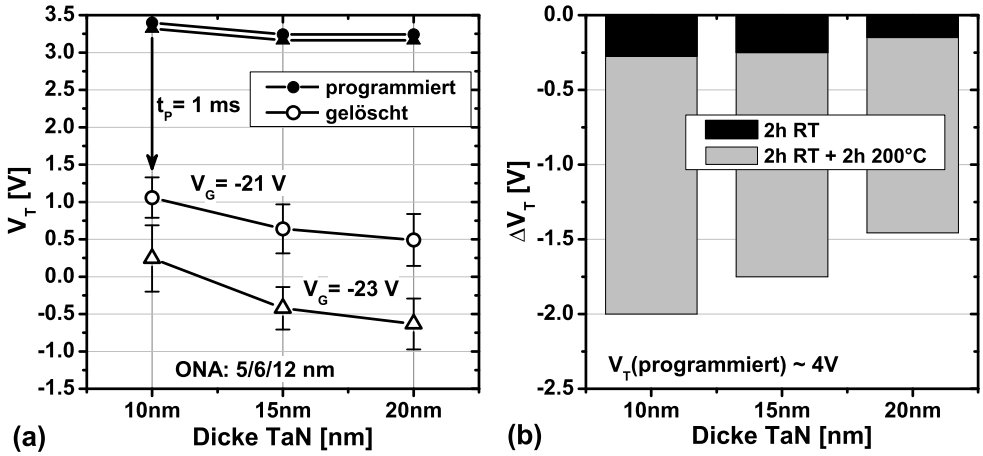

Abbildung 4.29: (a) Programmierter und gelöschter Zustand für zwei Löschspannungen in Abhängigkeit der TaN-Schichtdicke; (b) entsprechend der V_T-Verlust nach 2 h Raumtemperatur (RT) und nach 2 h RT zuzüglich 2 h 200°C

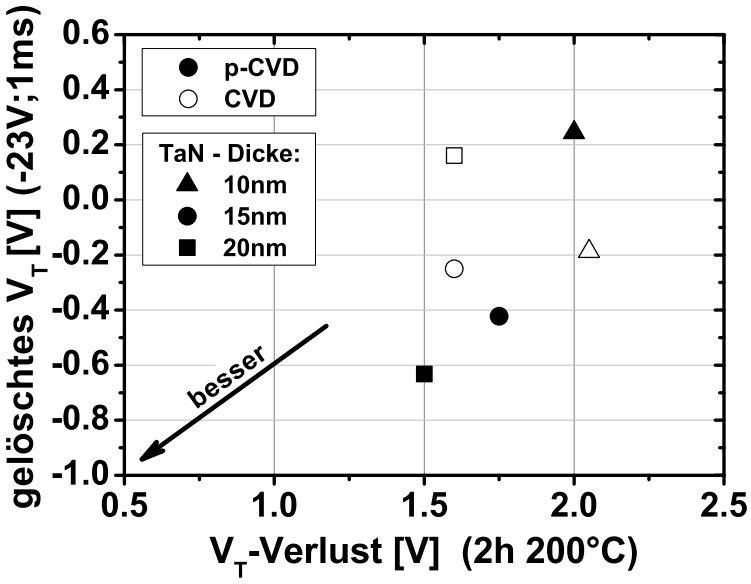

Abbildung 4.30: Gegenüberstellung des gelöschten Niveaus nach einem Löschpuls von −23 V, 1 ms und der durch Ladungsverlust verursachten V_T-Verschiebung nach einer Temperung von 2 h mit 200°C

Wie bereits erläutert, ist eine klare Abhängigkeit für das kristalline CVD-TaN nicht zu erkennen. Denn die Schichtdicke von 15 nm zeigt sowohl hinsichtlich Ladungsverlust als auch erreichtes Löschniveau das beste Verhalten. Hingegen ist beim gepulst abgeschiedenen amorphen TaN ein klarer Trend hin zur Abhängigkeit von der Dicke zu erkennen. Bei letzterem tritt mit zunehmender Schichtdicke eine Verbesserung der Ladungshaltung von 2 V auf 1.5 V V_T-Verlust ein. Weiterhin verbessert sich in diesem Fall das erreichte Löschniveau von positiven 0.2 V auf -0.6 V für einen Löschpuls von -23 V und 1 ms Länge. Daher ist eine Dicke von 20 nm für die Abscheidung zu wählen. Einen Einfluss auf die Austrittsarbeit der Gateelektrode in Abhängigkeit der Schichtdicke, wie durch Alshareev und Choi [37, 153] beschrieben, kann aufgrund der gewählten Dicken nicht beobachtet werden. Eine Änderung der Austrittsarbeit durch

die Fernwirkung der darüberliegenden Schicht wird erst ab Dicken von kleiner 5 nm beobachtet.
Bei der Gasphasenabscheidung von TaN werden verschiedene Präkursoren, ähnlich der ALD-Abscheidung, mit jeweils einer Tantalquelle und einer Stickstoffquelle benötigt. Als Tantalquelle kommt zum Beispiel $TaCl_5$ und TAIMATA (tert-amylimido-trisdim-ethylamidotantalum) [154] zur Anwendung. Als Stickstoffquelle dienen Ammoniak oder Stickstoff. Das Mischungsverhältnis von Tantal zu Stickstoff kann bei der Gasphasenabscheidung durch den Gasfluss während der Abscheidung variiert werden. Zudem gibt es auch Reaktionsgase, die sowohl als Tantal- als auch als Stickstoffquelle dienen. Zu nennen ist hierbei TBTEMT (tert-butylimido-tris-ethylmethylamido-tantalum) [150, 155], das auch für eigene Untersuchungen verwendet wurde. Ein Vergleich des elektrischen Verhaltens der verschiedenen Präkursoren erfolgt in Abb. 4.31.

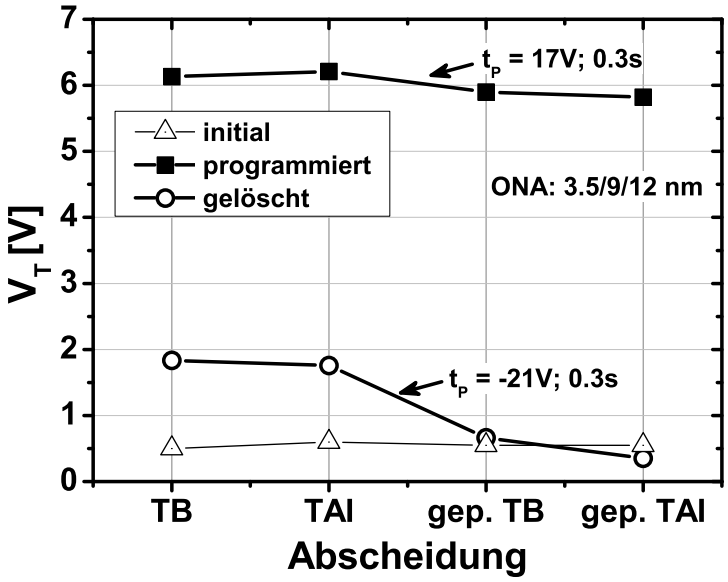

Abbildung 4.31: Programmier- und Löschcharakteristik für die zwei Präkursoren TBTEMT (TB) und TAIMATA (TAI), jeweils für Standard und gepulste (gep.) CVD-TaN Abscheidung

Der programmierte Zustand unterscheidet sich für alle vier untersuchten Gruppen nur minimal, da die Gateelektrode in diesem Modus nur einen geringen Einfluss auf das Zellverhalten besitzt. Am Augenscheinlichsten ist der bereits gezeigte Unterschied zwischen Standard und gepulster CVD-Abscheidung für das Löschen. Es bestätigt sich auch bei diesen Proben das günstigere Verhalten für die gepulste Abscheidung. Ein Einfluss des Präkursors für die kontinuierliche Abscheidung kann nicht festgestellt werden. Hingegen resultiert der TAIMATA Präkursor bei der gepulsten Abscheidung in einem circa 300 mV niedrigeren Löschniveau. Das generell schlechte Löschverhalten ist auf eine noch nicht optimierte Speicherzelle zurückzuführen.

4.3.2.2 Tantalnitrid aus physikalischer Gasphasenabscheidung

Zusätzlich zur chemischen Abscheidung mittels Präkursoren ist es möglich, Schichten mittels der sogenannten physikalischen Gasphasenabscheidung (engl. physical vapor deposition - PVD) zu prozessieren. Hierbei werden die abzuscheidenden Atome durch

Stickstoffatome, welche in einem Plasma beschleunigt wurden, in die Kammer freigesetzt, wo sie auf Oberflächen absorbiert werden. In den untersuchten Proben wurde das TaN aus einem massiven Tantalblock und Stickstoffspülung während der Abscheidung erzeugt.
Die erste Untersuchung fand an Proben statt, die bei der Firma Applied Materials Inc. (AMAT) abgeschieden wurden. Das TaN wurde auf einer AMAT ENDURA2 bei einem Druck von $7.2e^{-8}$ mbar abgeschieden. Durch Variation des Stickstoffflusses konnte das Verhältnis von Tantal zu Stickstoff variiert werden, wie Tab. 4.4 veranschaulicht.

Tabelle 4.4: Übersicht über die Abscheidebedingungen und resultierende Schichtzusammensetzung für das bei Applied Materials Inc. abgeschiedene PVD-TaN

Gruppe	Verhältnis (Ta : N)	N_2-Fluss (sccm)	Abscheiderate ($\AA$/s)
2:1	1.6 : 1	30	1.52
1:1	1 : 1	55	1.17
1:2	0.2 : 1	80	0.4

Es wird gezeigt, dass das Tantal-zu-Stickstoff-Verhältnis in einem weiten Bereich variiert werden kann. Die Schichten sind zum Ende des Prozesses unabhängig von der Schichtzusammensetzung kristallin. Dies ist auch auf den vernachlässigbaren Kohlenstoffgehalt in der TaN Schicht zurückzuführen. Als weiterer Parameter bei den Untersuchungen wurde auch die Schichtdicke mit 10 und 17 nm variiert. Die Betrachtung des elektrischen Verhaltens auf skalierten Zellen wird in Abb. 4.32 dargestellt.

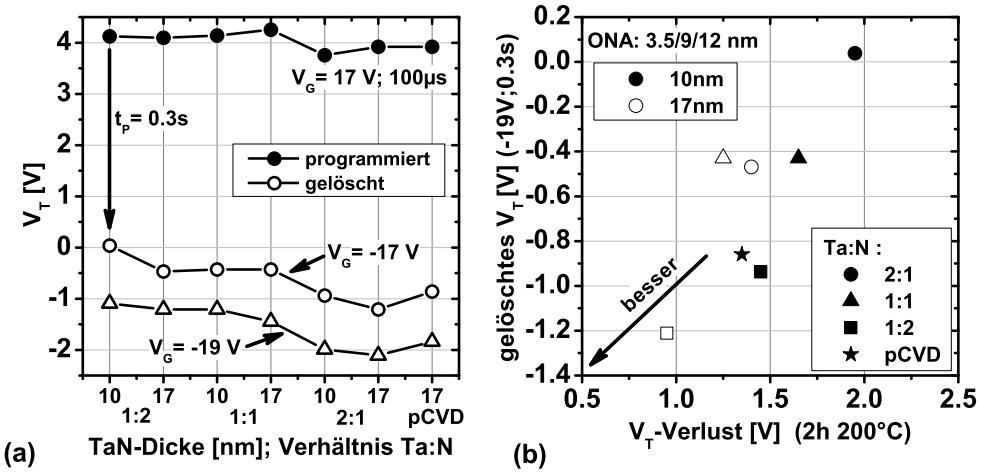

Abbildung 4.32: (a) Programmier- und Lösch-V_T für die angegebenen Spannungen und die in Tab. 4.4 genannten Gruppen mit zwei Schichtdicken (b); Gegenüberstellung von der Ladungshaltung und des erreichten Löschniveaus; Zellgröße: 48x48 nm

Die Abscheidung des PVD TaN weist eine starke Abhängigkeit von der Zusammensetzung und der Schichtdicke auf. Abbildung 4.32a verdeutlicht dies anhand des gelöschten Zustands nach einer Löschpulszeit von 300 ms. Die Gruppe 1:2 besitzt

einen hohen Tantalgehalt und zeigt die geringste Löschgeschwindigkeit. Erhöht man den Anteil von Stickstoff in der TaN-Schicht, verschiebt sich das erreichbare Löschniveau zu immer niedrigeren Spannungen. Bei einem Verhältnis von 2:1 verbessert sich das erreichte Lösch-V_T um circa 1 V. Erhöht man zudem noch die Schichtdicke, ist ein weiterer Gewinn von 200 mV zu beobachten. Eine Erklärung für die Verbesserung des elektrischen Verhaltens ist auch wieder mit Hilfe des Ausbleicheffektes möglich, da es sich um kleine Speicherzellen handelt. Eine Schicht mit geringem Tantalgehalt wird schneller auf einen minimalen Tantalgehalt reduziert als eine Schicht mit sehr hohem Tantalanteil. Aus diesem Grund ist nach erfolgter Aluminiumoxidätzung für die Gruppe 2:1 noch ausreichend Tantal in der Schicht, um keine dielektrische Schicht zu bilden, was wiederum die Homogenität der Ladungsverteilung in der Speicherzelle verbessert. Der Dickeneffekt bestätigt das bereits bei der CVD-Abscheidung beobachtete Verhalten, dass eine dickere Schicht in einem besseren Verhalten resultiert. Die Gegenüberstellung von Ladungshaltung und erreichtem Löschniveau in Abb. 4.32b bestätigt die Verbesserung durch eine Erhöhung des Tantalanteils in der Elektrodenschicht. Neben der Verbesserung des bereits diskutierten Löschverhaltens wird auch der Ladungsverlust bei Temperung mit zunehmendem Tantalgehalt reduziert. Auch hier kann die Auswirkung auf das elektrische Verhalten mit der inhomogenen Ladungsspeicherung durch das Ausbleichen der Gateelektrode erklärt werden. Durch die Rekombination der gespeicherten Elektronen mit den vom Gate in den Randbereich der Speicherzelle injizierten Löchern kommt es zu einer schnelleren Verschiebung der Schwellspannung [69] bei der Betrachtung der Ladungshaltung. Es ist auch eine Gruppe mit p-CVD zum Vergleich gezeigt. Die PVD Gruppe mit dem höchsten Tantalgehalt zeigt für die 10 nm Schichtdicke ein vergleichbares Verhalten wie die Probe mit p-CVD-Abscheidung. Die Gruppe mit dicker Schicht ist sowohl bei der Ladungshaltung als auch beim erreichten Lösch-V_T besser.
Ein weiterer Versuch zur PVD-Abscheidung erfolgte mit einer Anlage der Firma Singulus. Hierbei wurde die Elektrodenschicht auf einer TRIMARIS-Kammer nach dem sogenannten LDD-Verfahren (Linear Dynamic Deposition) abgeschieden [144]. Hierbei fährt der Wafer unter einer länglichen Elektrode mit dem abzuscheidenden TaN hindurch. Es erfolgt dabei nur eine Abscheidung im Bereich der Elektrode. Diese wiederum gliedert sich in mehrere Teilkomponenten. Ein Teil ist zum Beispiel dafür da, die Ionen zu erzeugen, die zur Abscheidung des gewünschten Materials führen. Der Kammerdruck während der Abscheidung ist im Vergleich zur Abscheidung auf der AMAT Anlage deutlich höher und liegt bei circa 1.0-3.6e^{-3} mbar. Die untersuchten Gruppen sind in Tab. 4.5 aufgezeigt.

Tabelle 4.5: Schichtdicke und Abscheidebedingungen der Gateelektrode, hergestellt auf der TRIMARIS Kammer

Gruppe	Druck (1e^{-3} mbar)	Schicht-dicke (nm)	N_2-Fluss (sccm)	Ar-Fluss (sccm)
1.0	1.0	10	47.5	82.5
1.8-	1.8	10	45	175
1.8	1.8	10	60	160
1.8+	1.8	10	75	145
1.8/17	1.8	17	75	145

Es wurden die Parameter Druck, Gasfluss-Verhältnis von Stickstoff zu Argon und die abgeschiedene Schichtdicke variiert. Der Einfluss des Druckes während der Abscheidung ist klein [144]. Hingegen ändert eine Verschiebung des Gasflusses die Schichtzusammensetzung bezüglich des Verhältnisses der kubischen zur hexagonalen TaN-Phase. Eine Erhöhung des Stickstoffflusses von 45 sccm zu 75 sccm verschiebt das Verhältnis TaN_{hex} :TaN_{kub} von 1:1 nach 1:2. Die Betrachtung der Ergebnisse für das Programmieren und Löschen erfolgt in Abb. 4.33a.

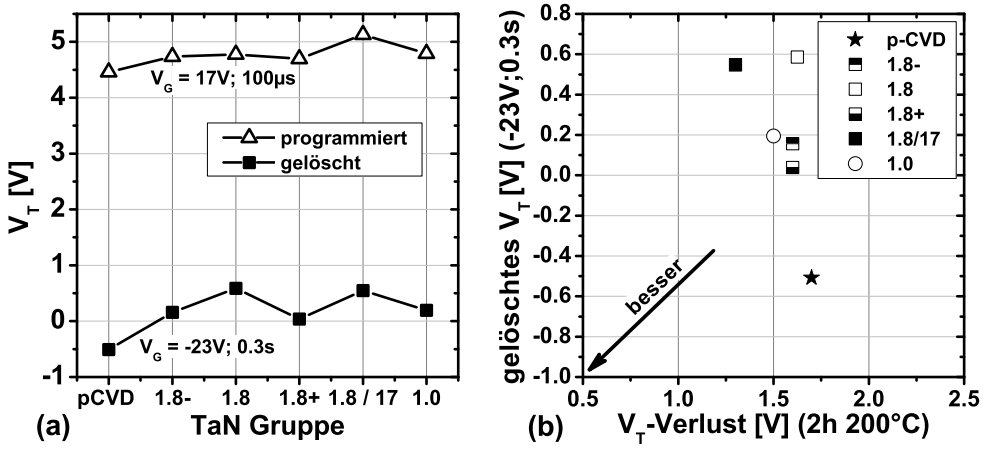

Abbildung 4.33: Programmier- und Löschcharakteristiken für die zwei Präkursoren TBTEMT (TB) und TAIMATA (TAI), jeweils für Standard und gepulste (gep.) CVD-TaN Abscheidung

Generell zeigen die Zellen im Vergleich zur p-CVD Referenz ein schlechteres Verhalten. Eine Änderung zu einem niedrigeren oder höheren Stickstoff-Gasfluss verbessert jeweils das Löschniveau um 500 mV bei nahezu konstantem Programmierniveau. Eine Verbesserung durch eine Erhöhung der Schichtdicke wie bei den bereits vorgestellten Abscheidungen kann bei dieser Abscheidung nicht beobachtet werden. Eine Verringerung des Druckes während der Abscheidung durch eine Reduktion des Gasflusses resultiert in einer ähnlichen Verbesserung wie die Änderung des Gasfluss-Verhältnisses. Da die Ladungshaltung im Vergleich zu p-CVD TaN nicht besser ist, abgesehen von der dicken TaN-Schicht, kann man schlussfolgern, dass diese Abscheidung für die Anwendung in Speicherzellen noch optimiert werden muss.

4.3.2.3 Titannitrid aus physikalischer Gasphasenabscheidung

Neben Tantalnitrid ist Titannitrid (TiN) ein weiteres Elektrodenmaterial, welches eine hohe thermische Stabilität bei vergleichbaren elektrischen Eigenschaften besitzt. Dass das elektrische Verhalten von TiN gut mit dem des TaN übereinstimmt, verdeutlicht Abb.4.34a. Bei dem untersuchten TiN handelt es sich um eine kristalline Schicht.
Die Programmiertransienten zeigen für drei verschiedene Spannungen nur minimale Abweichungen zwischen TiN und TaN. Betrachtet man aber die Löschtransienten in Abb. 4.34b, beobachtet man ein günstigeres Löschverhalten für das TiN. Bei den Löschspannungen von -19 V und -21 V beträgt die Differenz bis zu einer Zeit von 10 ms circa 400 mV. Bei längeren Zeiten wird das höhere Löschsättigungsniveau von TaN deutlich. Bei -23 V wird die Differenz zwischen den Löschkurven mit zunehmender Zeit immer größer und die Kurve des TaN sättigt bei -4 V V_T. Die Kurve

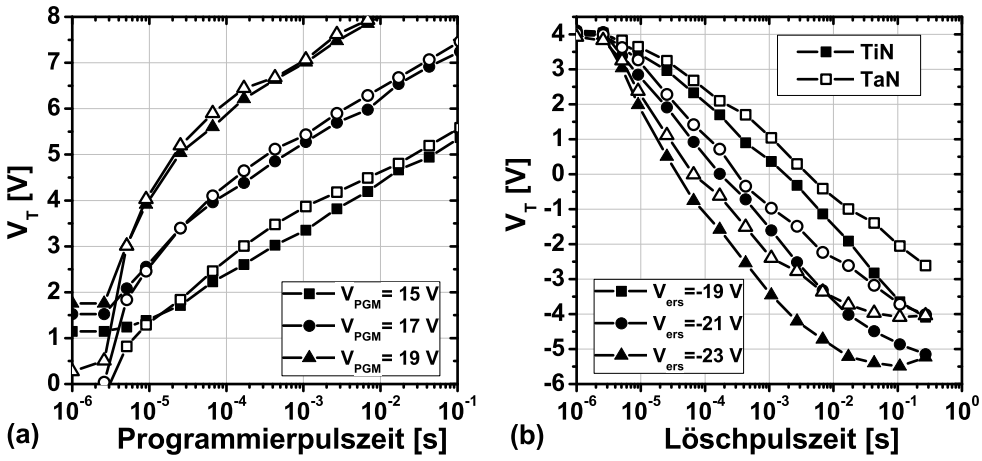

Abbildung 4.34: Vergleich von 48x48 nm Zellen gleichen Schichtstapels ONA: 5/6/12 nm mit einmal TaN- (offene Symbole) oder TiN-Gateelektrode (geschlossene Symbole); (a) Programmieren und (b) das Löschen für jeweils drei versch. Spannungen

des TiN hingegen läuft bis zu einem V_T von -5.5 V und steigt dann wieder leicht an. Den Unterschied in der Löschsättigung lässt sich einmal durch eine unterschiedliche Austrittsarbeit der Elektrodenmaterialien erklären. Wie durch Choi [37] gezeigt wird, hat TaN eine etwas niedrigere Austrittsarbeit im Vergleich zu TiN. Dies resultiert im Fall des TiN in einem kleineren Elektronenstrom von der Gateelektrode während das Löschens und somit in einem tieferen Löschniveau [36]. Eine weitere mögliche Erklärung für das verbesserte Löschverhalten sind im Aluminiumoxid befindliche negative Ladungen. Jeon [117]zeigt, dass bei höheren Temper-Temperaturen als 650°C negative feste Ladungen im Aluminiumoxid eingebaut werden. Somit kann man darauf schließen, dass auch in dem untersuchten Fall mit Temperbedingungen von 1100°C, 20 s identische Bedingungen herrschen. Allerdings erklärt dies noch nicht den Unterschied zwischen den Materialien. Eine noch nicht betrachtete Größe ist die mechanische Verspannung (Stress), die auf das Al_2O_3 wirkt und dessen Eigenschaften beeinflusst. So kann mechanischer Stress Bindungen aufbrechen und feste Ladungen generieren. In Abb. 4.35a wird ein Vergleich des mechanischen Stresses für TiN und TaN Gateelektroden nach verschiedenen Prozessschritten durchgeführt. Die Stressmessung erfolgt durch die Messung der Durchbiegung des Wafers nach den Prozessschritten, wodurch sich aus der Wafer- und Schichtdicke, sowie der Richtung der Durchbiegung, der Stress berechnen lässt [156, 157]. Man unterscheidet zwischen tensilem (Zug) und kompressivem (Druck) Stress. Ein negatives Vorzeichen steht für einen kompressiven Stress.

Es wird einmal durch die gefüllten Balken der Stress direkt nach der Abscheidung des Gateelektroden-Materials gezeigt. Es ist ein circa 8 mal so großer Schichtstress, der auf das Al_2O_3 wirkt, für TiN im Vergleich zu TaN zu beobachten. Der hohe Wert für den Filmstress bei TiN wir durch [158] bestätigt. Dieser Wert wird durch die folgenden Schichten für die Bildung der Wortleitung reduziert, wobei am Ende der Zellenfeldstrukturierung (schraffierte Balken) TiN einen mehr als doppelt so hohen Stress ausübt. Diese Differenz kann eine weitere mögliche Ursache für das unterschiedliche Löschverhalten sein. Ein weiteres Indiz, welches den Einfluss auf die Zelleigenschaften unterstützt, ist der in Abb. 4.35b gezeigte temperaturbeschleunigte Ladungsverlust.

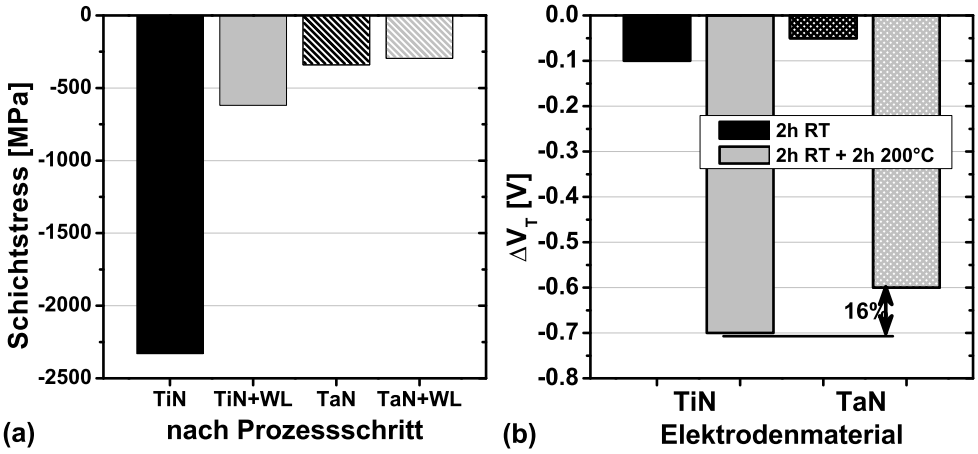

Abbildung 4.35: (a) Vergleich der mechanischen Spannung von TiN und TaN Gateelektroden zu verschiedenen Zeitpunkten der Prozessierung; (b) Auswirkung der mechanischen Spannung auf die Ladungshaltung von 48x48 nm Speicherzellen

Arghavani [159] berichtet in seiner Veröffentlichung, dass ein größerer kompressiver Stress den Leckstrom durch ein Oxid erhöht. Bezieht man dieses Ergebnis auf den Vergleich TaN zu TiN, kann man schlussfolgern, dass der höhere Ladungsverlust bei TiN auf den höheren Stress, welcher auf die Oxide wirkt, zurückführen ist.
Weiterhin war von Interesse, ob eine Zugabe von Silizium zu TiN zu einer Verbesserung des elektrischen Verhaltens führt. Es wurde anhand von Kondensator-Messungen nachgewiesen, dass eine Zugabe von Silizium eine höhere Austrittsarbeit ergibt und in einer amorphen Schicht resultiert [136, 160]. Es ist außerdem zu erwarten, dass eine amorphe Schicht einen geringeren Schichtstress erzeugt. Daher wurden Messungen an Proben mit gepulst abgeschiedenem TiN durchgeführt, entsprechend dem gepulst abgeschiedenem TaN. Während der Abscheidung wird auf eine Siliziumzugabe umgeschaltet und eine entsprechende Pulszahl Silizium abgeschieden. Dabei variiert der Anteil linear von 4% Si (6p) und 16% Si (24p). Die elektrischen Ergebnisse für Programmieren und Löschen sind in Abb. 4.36a dargestellt. Betrachtet man das Programmieren, ist kein großer Unterschied bei einer Zugabe von Silizium festzustellen. Eine Änderung des Programmierverhaltens ergibt sich erst bei sehr hohen Austrittsarbeiten [36], dies ist hier aber nicht der Fall. Die Abweichung für die TiN-Gruppe bei langen Programmier-Pulszeiten in Abb. 4.36a ist auf strukturelle Effekte zurückzuführen. Viel deutlicher aber ist die Abhängigkeit des Löschverhaltens von dem Siliziumgehalt. Zu Beginn des Löschvorgangs ist der Unterschied zwischen den Gruppen noch nicht groß, aber mit zunehmender Löschzeit wird der Unterschied immer größer. Daraufhin wird auch das Löschsättigungsniveau mit zunehmendem Siliziumgehalt immer mehr in positive Richtung verschoben. Dies deutet eindeutig auf eine Verringerung der Austrittsarbeit hin. Eine mögliche Erklärung ist die Bildung von Titansilizid, welches eine geringere Austrittsarbeit von ≈4.4 eV [161] im Vergleich zu ≈4.8 eV von TiN aufweist. Ein solcher Effekt wurde auch für TaN mit zunehmendem Siliziumgehalt beobachtet [160]. Zudem ist die Ladungshaltung durch Zugabe von Silizium deutlich verschlechtert, wie Abb. 4.36b zeigt. Demzufolge ist eine Beimischung von Silizium zu TaN oder TiN nicht zielführend und verschlechtert das elektrische Verhalten.

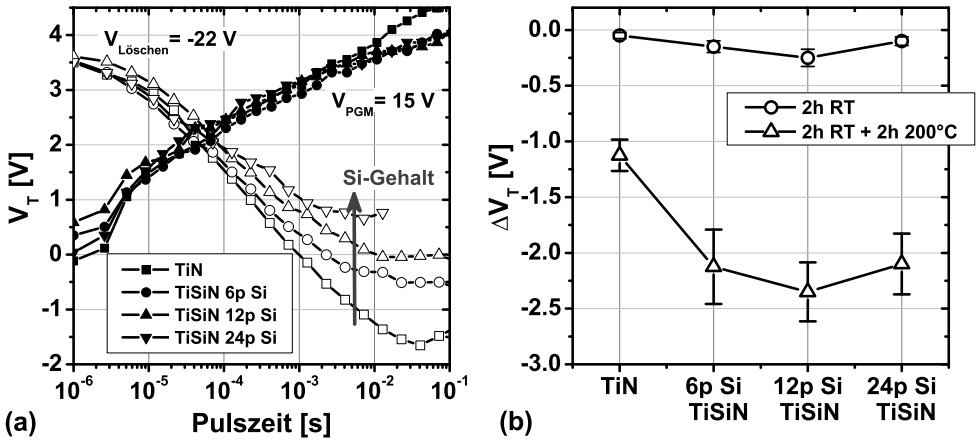

Abbildung 4.36: Betrachtung des Einflusses einer Si-Beimischung bei der Abscheidung einer TiN-Gateelektrode; (a) Einfluss auf das Programmier- und Löschverhalten, (b) Auswirkung auf die Ladungshaltung einmal bei RT und andererseits nach einer Temperung von 2 h bei 200°C

4.3.2.4 Bestimmung der Austrittsarbeit

Um die Literaturwerte der Austrittsarbeit mit den eigenen Proben zu vergleichen, wurden Untersuchungen von TaN, abgeschieden mit Gasphasenabscheidung, durchgeführt. Hierfür wurde ein Vergleich mit n^+-poly Gateelektroden durchgeführt. Man verwendet für die Bestimmung der Austrittsarbeit Proben, bei denen die effektive Oxiddicke variiert wurde [137, 160]. Es wird auf einer durch Ätzung abgestuften SiO_2-Schicht eine Al_2O_3-Schicht aufgebracht (engl. terraced oxide), wie in Abb. 4.37b dargestellt.

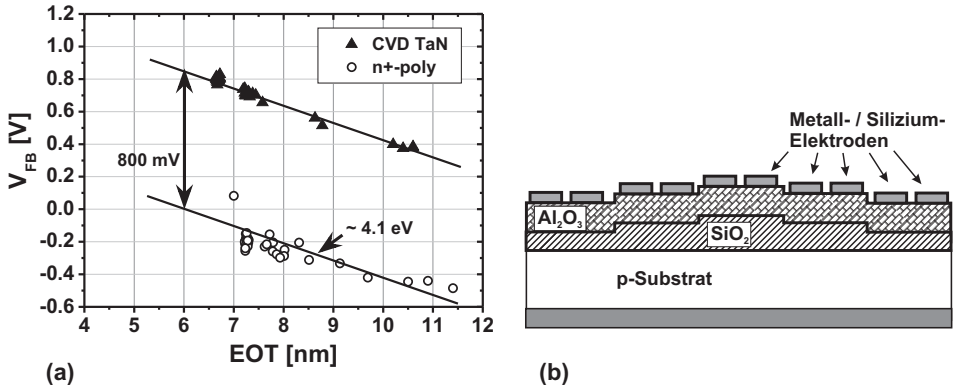

Abbildung 4.37: (a) Austrittsarbeit-Extraktion an Oxid-Wafern mit abgestufter Schichtdicke für eine n-poly Gateelektrode (Dreiecke) und eine TaN-Gateelektrode (Kreise); (b) schematische Darstellung eines Wafers mit abgestufter Oxiddicke und einfachen MOS-Kondensatoren

Unter der Annahme, dass sich die im Aluminiumoxid und an der Grenzfläche der beiden Oxide eingebaute Ladung nicht ändert, hat nur noch die im SiO_2 eingebaute Ladung einen Einfluss. Gemäß Gl. 4.7, erweitert um den Einfluss durch die im Oxid

gespeicherte Ladung ergibt sich [8, 162]:

$$V_{FB} - \Phi_{ms} = -\frac{Q_f \cdot d_{SiO_2}}{\epsilon_{SiO_2}}. \tag{4.9}$$

Q_f bezeichnet die feste Oxidladung, deren Beitrag durch die Dicke der SiO_2 Schicht, beschrieben durch den Parameter d_{SiO_2}, variiert. Es sollte also eine lineare Abhängigkeit der Flachbandspannung V_{FB} von der Oxiddicke zu beobachten sein, da die Dicke der Al_2O_3-Schicht konstant ist. In Abb. 4.37a wird das Ergebnis für eine n^+-poly und TaN Elektrode gezeigt. Gut ist die nahezu lineare Abhängigkeit der Flachbandspannung von der äquivalenten Oxiddicke zu erkennen, was auch die Annahmen bestätigt. Eine exakte Bestimmung der Austrittsarbeit ist nur möglich, wenn auch die Beiträge der Ladungen im Al_2O_3 und an der Dielektrika-Grenzfläche bekannt sind. Dies erfordert aber einen erheblichen Probenaufwand, der nur einen geringen Wissensgewinn bringt. Denn bei einer hinlänglich bekannten Größe für die Austrittsarbeit, wie zum Beispiel 4.1 eV für n^+-Silizium [12], lässt sich gemäß [8]

$$(V_{FB})_1 - (V_{FB})_2 = (\Phi_{ms})_1 - (\Phi_{ms})_2 \tag{4.10}$$

die Austrittsarbeit der zweiten Elektrode abschätzen. Die beobachtete Differenz von TaN zu n^+-poly beträgt 800 mV. Es wird somit für die aufgebrachte p-CVD TaN Elektrode eine Austrittarbeit von 4.9 eV auf Al_2O_3 ermittelt. Dieser Wert stimmt gut mit den in der Literatur beobachteten Werten von circa 4.8 eV überein [133, 136, 148, 149].

4.4 Einfluss von Kanal- und Source/Drain-Dotierung

4.4.1 Kontaktdotierung

Die Source/Drain-Gebiete sind bei einem MOS-Transistor die Bereiche, die Elektronen für den leitfähigen Inversionskanal zur Verfügung stellen. Die Schwellspannung eines Transistors wird entsprechend Gl. 2.8 berechnet.

$$V_T = 2\psi_F + V_{FB} + \frac{Q'_{ges}}{C'_{ox}}$$

Die Gleichung beschreibt die Spannung, bei der genau die Inversion einsetzt und ein leitfähiger Kanal gebildet wird, representiert durch die zweifache Fermi-Potentialdifferenz ψ_F. Es zeigt sich, dass die Kontaktdotierungen in der Berechnung der Schwellspannung nicht enthalten sind. Im Prinzip bestimmen sie somit nicht das Verhalten des Transistors, wenn er ausreichend groß ist und keine Kurzkanal-Effekte, wie in Kap. 2.2.2 beschrieben, aufweist. Aber bei den von uns untersuchten Zellen ist aufgrund der geringen Größe eine Abhängigkeit zu beobachten. In dem Versuch wurden drei verschiedene Gruppen mit unterschiedlichen Arsen-Implantationsbedingungen hergestellt. Die Diffusionsweite wurde einmal durch die Dosis zwischen $2e^{13}$ und $5e^{13}$ variiert und andererseits durch eine dünne Oxidschicht, mit TEOS abgeschieden, der Abstand der Implantation zur Steuerelektrode vergrößert. Die Anordnung dieser Schicht ist in Abb. 2.6 illustriert und TEOS steht für die chemische Basis Tetraethylorthosilikat mit der sehr konformale Schichten hergestellt werden können. Das elektrische Ergebnis wird in Abb. 4.38 gezeigt.

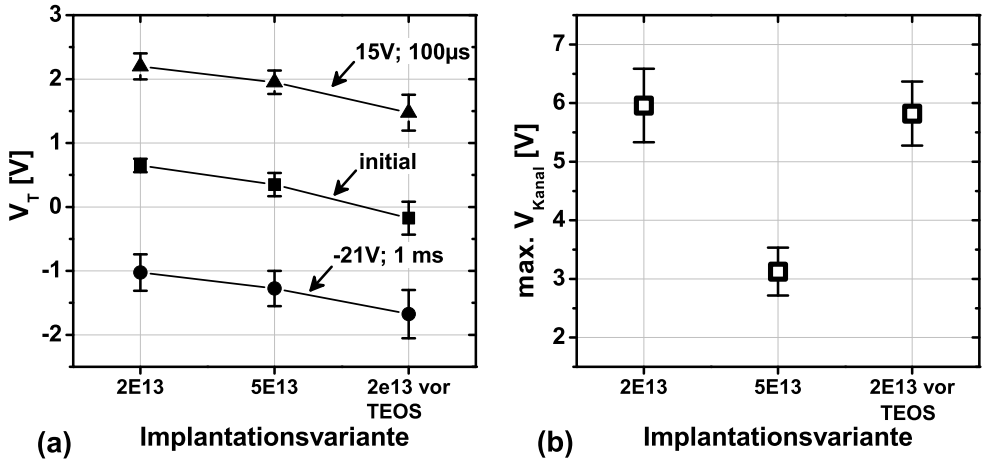

Abbildung 4.38: (a) Initialer, nach Programmieren und dem gelöschten Zustand gemessenes V_T von drei Gruppen unterschiedlicher Kontaktprozessierung, (b) errechnetes maximal erreichbares Kanalpotential während des Programmier-Inhibits für die gleichen Gruppen wie (a); 48x48 nm Speicherzellen

Es zeigt sich, wie erwartet, dass die unterschiedliche Behandlung der Kontaktimplantation einen Einfluss auf die initiale Schwellspannung hat. Für die Gruppe $2e^{13}$ mit der kleinsten Diffusionsstrecke r_j ist auch das V_T am höchsten, wie durch Gl. 2.20 bereits aufgezeigt wurde. Eine Implantation vor Abscheidung der zusätzlichen TEOS-Schicht verkürzt den Abstand des Implantationsgebietes zur Steuerelektrodenkante und das Arsen kann weiter in den Kanal diffundieren. Diese Variante resultiert in der niedrigsten Schwellspannung der drei Gruppen und hat somit die größte Unterdiffusion. Ein weiterer Aspekt wird in Abb. 4.38a deutlich. Die Kontaktimplantation hat keinen weiteren Einfluss auf das Verhalten der Speicherzellen. Es kommt zu einer Parallelverschiebung der programmierten und gelöschten Zustände relativ zur initialen Schwellspannung unter gleichen Bedingungen. Dies eröffnet eine Möglichkeit die Transistoren so einzustellen, dass der Betrieb von haftstellen-basierten Speicherzellen optimiert wird. Es ist möglich, den Einsatzbereich der Zelle hin zu niedrigeren Löschspannungen mit einer höheren Kontaktimplantation zu verschieben. Ein niedrigeres initiales V_T resultiert in einer kleineren Löschdifferenz zu einem gelöschten Absolutwert, wodurch sich niedrigere Löschspannungen ergeben. Ein Resultat niedrigerer Löschspannungen ist eine verbesserte Zuverlässigkeit, wie in Kap. 4.4.2 gezeigt wird.

Ein weiterer Aspekt ist die Verkürzung der Kanallänge L' durch Unterdiffusion. Dies führt zu einer Ausblendung des Speicherschichtbereiches über den Kontakten. Veranschaulicht wird dieser Effekt in Abb. 4.39. Die zunehmende Unterdiffusion schirmt die äußeren Bereiche ab, da diese keinen Einfluss mehr auf den Inversionskanal haben. Dort injizierte Ladung verliert somit den Einfluss auf das Zellverhalten.

Das Ausblenden der Randbereiche kann nun gezielt genutzt werden, um den in Kapitel 4.2.1 beschriebenen störenden Kanteneffekt zu unterdrücken. Anhand von Messungen an einer speziellen Teststruktur mit Auswahltransistoren lässt sich dieses Verhalten gut beobachten. Abbildung 4.40a verdeutlicht den Einfluss durch eine Betrachtung der Löschtransienten.

Es ist eine starke Abhängigkeit der Löschsättigung von der Impantation zu sehen. Es

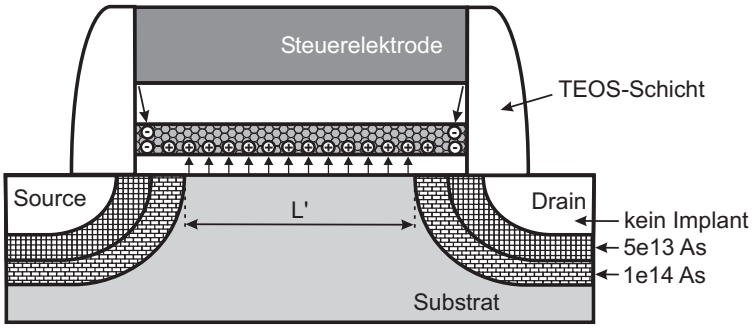

Abbildung 4.39: Auswirkung einer Erhöhung der Source/Drain-Implantationsdosis auf den Randbereich der Speicherzelle, dadurch ergibt sich eine Verkürzung der effektiven Kanallänge L' und ein Ausblenden des Bereichs, der durch die Steuerelektroden-Injektion gestört ist

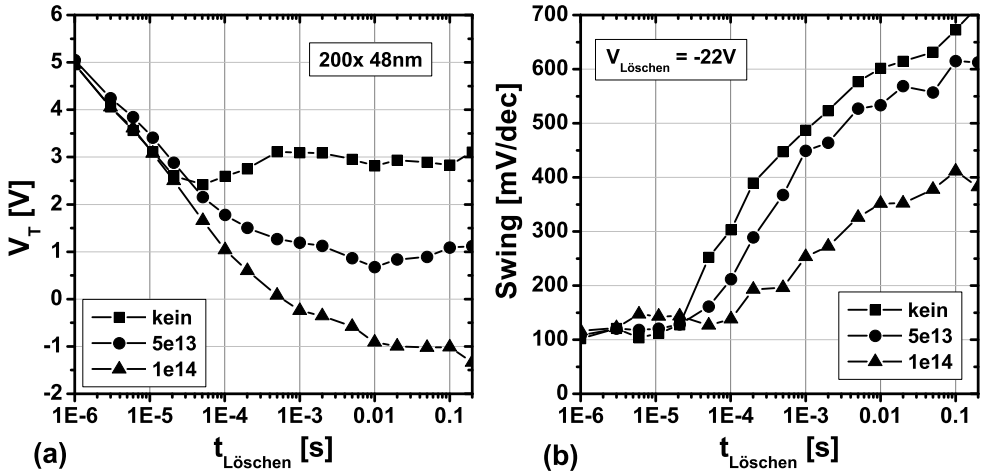

Abbildung 4.40: (a) Löschtransienten für drei Kontaktimplantationsdosen bei einer Löschspannung von -22V; (b) Entwicklung des sub-V_T Swings der drei Gruppen während des Löschens

wurde bereits dargelegt, dass dies nicht auf einen Effekt zurückgeführt werden kann, der durch den Schichtstapel induziert ist. Daher muss es sich um einen parasitären Effekt handeln, der einen Einfluss auf die Schwellspannung hat. In Abb. 4.40b ist die Entwicklung des sub-V_T-Swings während des Löschens dargestellt. Deutlich ist eine schnelle Zunahme des sub-V_T Swings bei einer niedrigen Kontaktdotierung zu beobachten. Der Anstieg des sub-V_T-Swings kann wieder mit einer inhomogenen Injektion am Rand der Steuerelektrode erklärt werden. Wird der Einfluss dieses Bereichs verringert, verbessert sich auf der einen Seite der sub-V_T-Swing und andererseits ist es möglich die Speicherzelle tiefer zu löschen. Es wird daher nachgewiesen, dass es sich bei der Löschsättigung um ein Auswerteartefakt handelt, welches durch sub-V_T-Swing-Degradation hervorgerufen wird.
Die Integration in NAND-Ketten birgt allerdings eine zusätzliche Komponente, die eine Erhöhung der Kontaktimplantation begrenzt. Und zwar wurde durch Lee [163] festgestellt, dass das Verfahren zur Unterdrückung des unerwünschten Programmierens (Program-Inhibit) (Kap. 2.4.3), welches zum gezielten Programmieren einzelner Zellen benötigt wird, eine Abhängigkeit zeigt. Es wird verdeutlicht, dass das erreichbare Kanalpotential durch Leckstromkomponenten eingeschränkt wird. In dem Ver-

such wurde nur eine Modifikation der Kontaktimplantation durchgeführt und das Ergebnis kann damit korreliert werden. Abbildung 4.38b zeigt das errechnete Kanalpotential aus einer Messung, die den Inhibit untersucht. Es wird deutlich, dass eine Erhöhung der Implantationsdosis das erreichbare Kanalpotential reduziert. Komponenten, die zur Reduktion beitragen, sind in Abb. 4.41 schematisch für eine NAND-Reihe unter Inhibit-Bedingungen dargestellt.

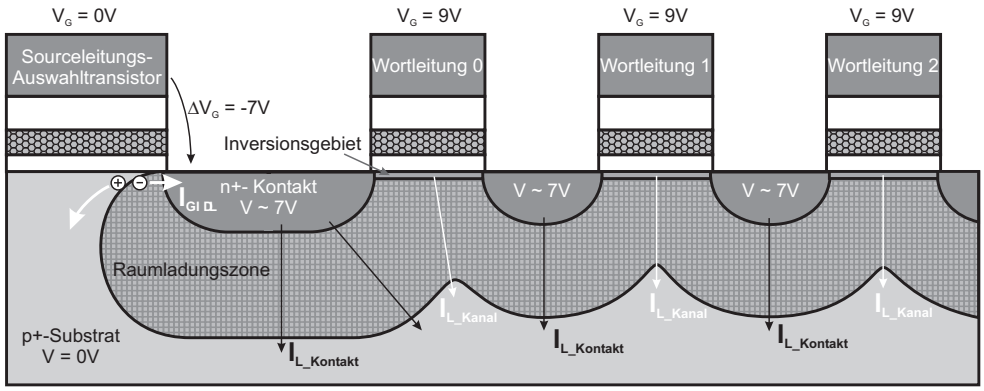

Abbildung 4.41: Potentiale und Stromkomponenten für einen Ausschnitt eines NAND-Strings im Inhibit-Fall dargestellt

Es gibt die drei Leckstromkomponenten I_{GIDL}, I_{L_Kanal} und $I_{L\_Kontakt}$, wobei sich der Gesamtleckstrom zu:

$$I_L = 2 \cdot I_{GIDL} + n \cdot I_{L\_Kanal} + (n+1) \cdot I_{L\_Kontakt} \tag{4.11}$$

ergibt. n bezeichnet die Anzahl der Speicherzellen in einem NAND-String, I_{L_Kanal} bezeichnet die Leckstromkomponente, welche durch die Raumladungszone des Kanals abfließt und $I_{L\_Kontakt}$ den entsprechenden Kontaktleckstrom. I_{GIDL} ist die Leckstromkomponente welche durch die relativ zur Drainspannung negative Gatespannung einen Tunnelstrom an der Substratoberfläche am Auswahltransistor induziert. Dies ist die wichtigste Komponente der Leckströme, welche vorrangig am Drain-Gebiet des sourceleitungs-seitigen Auswahltransistor entsteht. Im Fall des Programmierens ist dessen Steuerelektrode auf 0 V geschalten und hat damit die größere Potentialdifferenz im Vergleich zum Bitleitungs-Auswahltransistor, welcher auf 3 V geschalten ist.

4.4.2 Kanaldotierung

Die Kanaldotierung bestimmt die Schwellspannung der Speicherzellen, wie sich aus Gl. 2.8 ergibt:

$$V_T = V_{FB} + 2\phi_F + \frac{\sqrt{2\epsilon_s q N_A (2\phi_F)}}{C_{OX}}$$

Erhöht man die Kanaldotierung N_A, so verschiebt sich auch das V_T in positive Richtung. Dieses Verhalten ist genau gegensätzlich zum Kurzkanaleffekt der Kontaktimplantation bei kleinen Transistoren. Eigene Untersuchungen bestätigen dieses Verhalten, wie Abb.4.42 veranschaulicht.
Es wird deutlich, dass das Verhalten, dem der Kontaktimplantation entspricht. So ist der programmierte und gelöschte Zustand annähernd eine Parallelverschiebung des

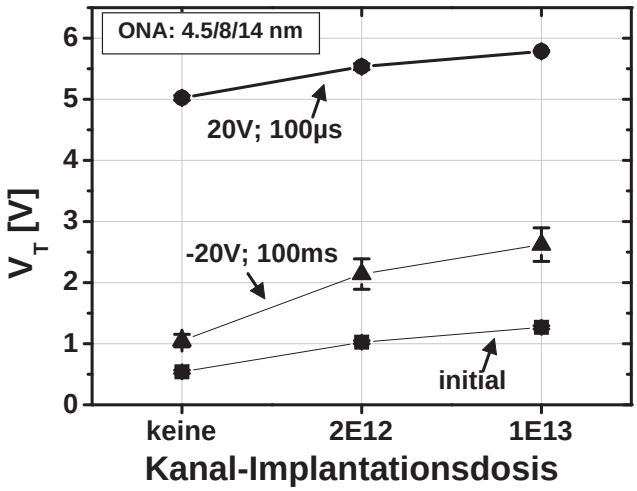

Abbildung 4.42: Abhängigkeit der initialen, programmierten und gelöschten Schwellspannung in Abhängigkeit von drei Kanaldotierungen, gemessen an SANOS -Speicherzellen

initialen Zustandes. Wie bereits bei der Kontaktimplantation angesprochen, wird die Zuverlässigkeit von haftstellenbasierten Zellen gesteigert, wenn die initiale Schwellspannung verringert werden kann. Eine Möglichkeit den Kanal dementsprechend zu modifizieren, ist die Implantation einer Spezies, die an der Oberfläche einen Bereich bildet, der die Majoritäten verringert oder im Extremfall sogar invertiert. Man spricht dann von einem Transistor mit vergrabenem Kanal (engl. buried-channel). Dadurch wird das initiale V_T in negative Richtung verschoben. Dies wird deutlich in Abb. 4.43a, wobei eine inverse Kanalimplantation mit Arsen durchgeführt wurde. Durch das niedrigere initiale V_T wird eine niedrigere Löschspannung benötigt, um das Ziel-Löschniveau zu erreichen, wie in bb. 4.43b gezeigt. Die niedrigere Löschspannung wiederum verringert erfolgreich die Degradation während des Zykelns.

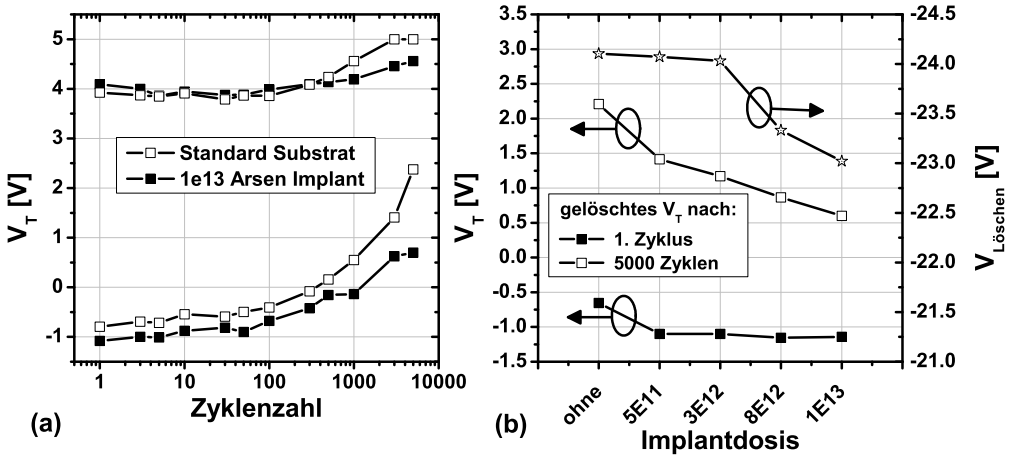

Abbildung 4.43: (a) Entwicklung der Schwellspannung während des Zykelns bei konstanter Programmier- und Löschspannung für einen Standard- und einen invers dotierten Kanal, (b) erreichtes Lösch-V_T nach einem Zyklus (geschlossene Box) und 5000 Zyklen (offene Box) bei konstantem Löschpuls (1 ms) mit zuvor bestimmter Löschspannung (Sterne) bei zunehmender inverser Kanalimplantation für 48 nm Zellen

Interessanterweise ist der Einfluss auf die Programmierspannung relativ klein, diese unterscheidet sich maximal um 0.4 V. Es wird deutlich, dass die Gruppe mit dem Standard-Prozess beim ersten Löschvorgang aufgrund einer Spannungsbegrenzung von -24 V nicht bis auf das Zielniveau -1 V gelöscht werden kann. Alle Gruppen mit einer inversen Implantation erreichen das Löschniveau, wobei mit zunehmender Dosis die Löschspannung und Degradation kleiner werden. Ein Vergleich der beiden Randgruppen ohne und $1e^{13}$ zeigt Abb. 4.43a. Hierbei wird noch einmal der Unterschied deutlich. Interessant ist in diesem Zusammenhang die Auswirkung auf die Ladungshaltung. Durch die Kanalimplantation wird das V_T einer frischen Zelle verringert. Will man nun auf das gleiche Niveau programmieren, ist mehr Ladung notwendig, da ein größeres ΔV_T nötig ist. Es ist somit zu erwarten, dass die Gruppe mit der höchsten inversen Dotierung die größte V_T-Verschiebung aufweist. Die untersuchten Proben hatten aufgrund von einem noch nicht optimal eingestellten Prozess einen sehr großen und dementsprechend nicht sinnvoll vergleichbaren Ladungsverlust.

5 Integration in eine stark skalierte NAND Architektur

5.1 Auswirkung struktureller Effekte auf die Speicherzelle

Eine effektive Nutzung als Speicher ist nur möglich, wenn möglichst viel Speicherkapazität auf möglichst kleinem Raum untergebracht wird. Damit verbunden ist, dass das Speicherelement so klein wie möglich sein muss. Dementsprechend verschieben sich auch die Größen, die das Verhalten der Speicherzelle bestimmen. Vergleicht man zum Beispiel den Verlauf des elektrischen Feldes einer haftstellen-basierten Speicherzelle unterschiedlicher Dimension, ergeben sich selbst bei einfachen planaren Strukturen, wie in Abb. 5.1 gezeigt, große Unterschiede.

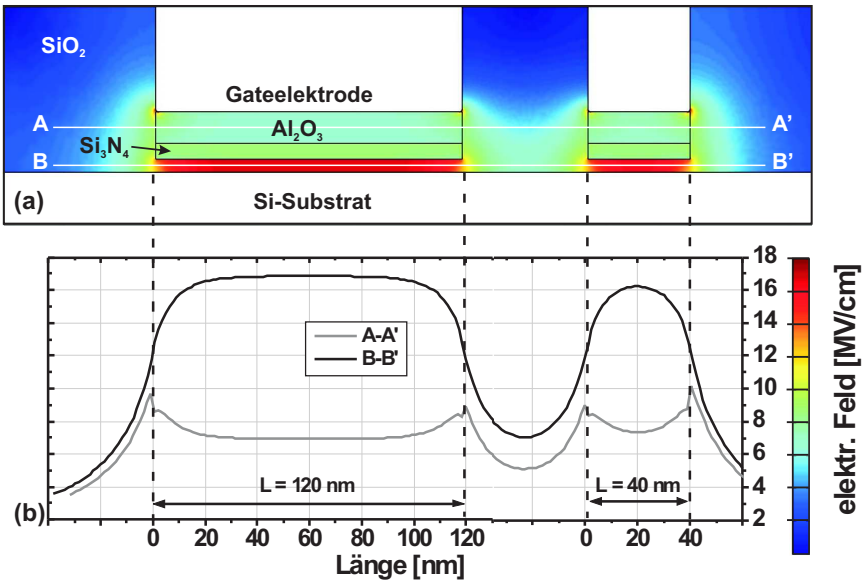

Abbildung 5.1: (a) Vergleich der elektrischen Felder in einer relativ großen und kleinen Struktur; (b) Feldverlauf im Tunneloxid (schwarz) und Topoxid (grau)

Es wird veranschaulicht, dass sich bei einer ausreichend langen Speicherzelle ein Bereich ausbilden kann, welcher ein konstantes Feld aufweist, wie in Abb. 5.1b verdeutlicht. Sowohl bei der langen als auch bei der kurzen Zelle gibt es einen weiteren Bereich am Transistorrand, bei dem das Feld im Tunneloxid kleiner ist, als der ideale Fall in der Mitte der großen Zelle. Bei der kleinen 40 nm langen Zelle führt dieser Bereich von Feldinhomogenität dazu, dass in der Transistormitte das erwartete Feld von 17 MV/cm nicht erreicht wird [83, 164]. Der Bereich welcher durch Feldinhomogenitäten beeinflusst wird, ist grob durch das Verhältnis von EOT zu Transistorlänge

bestimmt. Für den betrachteten Fall von 13 nm EOT kann eine Länge von circa 40 nm für den Bereich reduzierten Feldes extrahiert werden. Daraus ergibt sich ein Verhältnis von EOT:Länge = 1:3. Einen weiteren Effekt verdeutlicht der graue Graph, welcher das Feld im Topoxid zeigt. Bei der kleinen Zelle wird das Feld durch die Feldinhomogenitäten erhöht. Damit verringert sich das durch die Materialwahl beabsichtigte Feldverhältnis von Bottomoxid zu Topoxid. Demzufolge ist eine Verschlechterung des elektrischen Verhaltens bei zunehmender Verkleinerung zu erwarten, was auch durch Lue [83] bestätigt wurde. Bei den Feldspitzen an dem Übergang von Al_2O_3 zu SiO_2 handelt es sich um artifizielle Effekte der Simulation, die so nicht in der Realität auftreten.

5.1.1 STI-Stufenhöhe

Die Integration von Speicherzellen in der NAND-Architektur macht es notwendig, benachbarte Transistor-Reihen elektrisch voneinander zu trennen. Hierfür wird eine Grabenoxid-Isolation (engl. shallow trench isolation - STI) eingefügt [165]. Erst die Einführung einer solchen Isolation ermöglicht einen zuverlässigen Betrieb der NAND-Architektur, da benachbarte Transistoren unterschiedliche Kanalpotentiale besitzen können. Die unterschiedlichen, teils hohen Kanalpotentiale werden benötigt, um gezielt Zellen von benachbarten Transistorketten beschreiben zu können [166]. Die Tiefe der Gräben ist bestimmt durch die Raumladungszone, die sich im hochdotierten Bereich der Kontakte ausbildet und liegt im Bereich von 250 nm. Eine Verdeutlichung der Struktur erfolgt in Abb. 5.2a anhand eines schematischen Schnittes entlang einer Wortleitung.

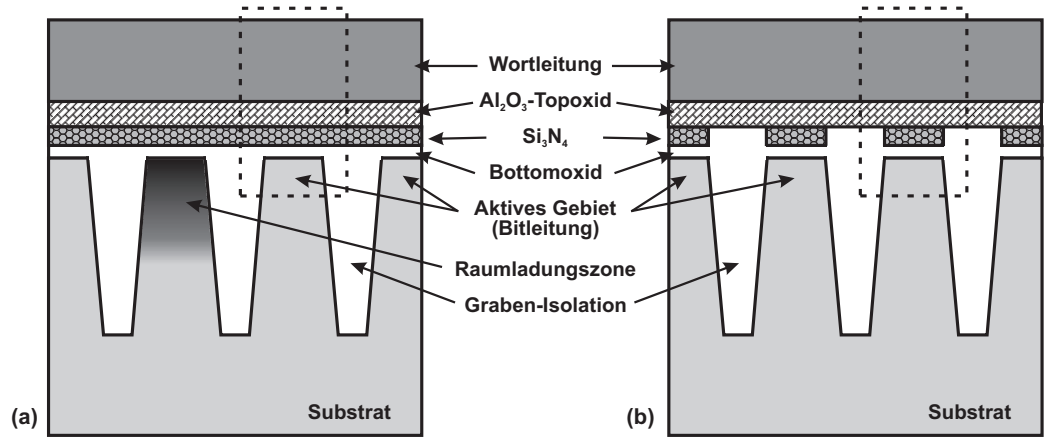

Abbildung 5.2: Schematischer Aufbau eines NAND-Speichers entlang einer Wortleitung; (a) mit durchgängiger Speicherschicht; (b) eine bei der Grabenoxid-Isolation strukturierte Speicherschicht (self-aligned) ; die gestrichelten Boxen zeigen die Ausdehnung einer Speicherzelle und in (a) zusätzlich die Ausdehnung einer Raumladungszone

Die isoliert liegenden aktiven Gebiete, in denen sich das Inversionsgebiet befindet, sind deutlich zu erkennen. Um dies noch anschaulicher zu machen, wurde in einem aktiven Gebiet die Ausdehnung einer möglichen Raumladungszone, die sich unterhalb des leitfähigen Kanals ausbildet, dargestellt. in Abb. 5.2b wird eine weitere Möglichkeit der Strukturierung von haftstellenbasierten NAND-Speichern veranschaulicht.

Hierbei wird während der Ätzung der Isolationsgräben auch die Speicherschicht strukturiert, wodurch sich eine Verbesserung der Zuverlässigkeit ergibt [69]. Allerdings ist der Aufwand zur Herstellung einer solchen Speicherzelle im Vergleich zu einer Speicherzelle mit durchgängiger SiN-Schicht deutlich größer. Daher soll eine Speicherzelle mit einer durchgängigen Speicherschicht genauer analysiert werden. Die in Abb. 5.2a gezeigte Strukturierung repräsentiert einen idealen Fall. Bei der Herstellung kommt es immer zu leichten Schwankungen in der Höhe des Füll-Oxides, welches in die Isolationsgräben abgeschieden wird. Die mögliche Geometrie, welche sich durch Variation der Füll-Oxidhöhe ergibt, ist in Abb. 5.3 gezeigt.

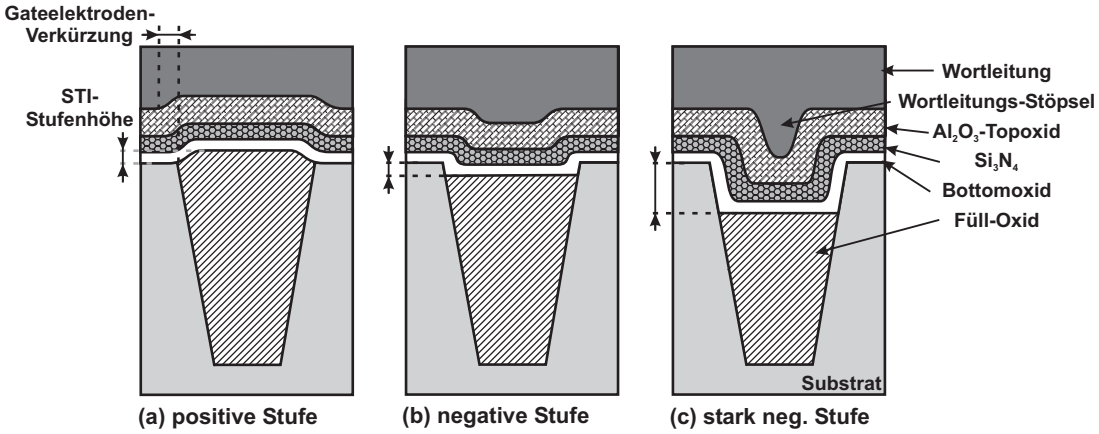

Abbildung 5.3: Schematischer Aufbau eines NAND-Speichers entlang einer Wortleitung; (a) mit einer positiven STI-Stufe; (b) mit einer leicht negativen und (c) mit einer stark negativen STI-Stufe

Es ist deutlich zu erkennen, dass die Füllhöhe des Grabenoxids die resultierende Geometrie bestimmt. Lässt man das Fülloxid höher als das aktive Gebiet stehen, spricht man von einer positiven Stufe, wie in Abb. 5.3a gezeigt. Befindet sich die Oberkante des Oxids unterhalb des aktiven Gebietes spricht man von einer negativen Stufe. Durch das Absenken der dielektrischen Schichten in den Graben kommt es zur Bildung eines Stöpsels an der Wortleitung, wie Abb. 5.3c verdeutlicht. Bei einer ausreichend weiten Öffnung des STI-Grabens ist es möglich, diesen Stöpsel bis weit in den Graben hinein ragen zu lassen. In dem Fall steuert die Gateelektrode nicht nur das Gebiet auf der Oberseite, sondern auch auf der Seitenwand des aktiven Gebietes. Ist das Verhältnis Seitenwand zu Oberseite deutlich größer als 1, spricht man von einem FinFET [167, 168]. Dies zeigt auch, dass ein weiterer Aspekt bei der Betrachtung unterschiedlicher Stufenhöhen berücksichtigt werden muss. Und zwar verschiebt die Stufenhöhe die Kopplung zwischen Gateelektrode und Kanal des Transistors. Die wirksame Elektrodenfläche ist für eine positive Stufe reduziert. In Abb.5.3a wird gezeigt, dass die Stufe den Abstand zum Kanal vergrößert und das Gebiet gleicher physikalischer Dicke kleiner als die Breite des aktiven Gebietes ist. Daher kommt es zu einer effektiven Verkürzung der Gateelektrode. Bei einer negativen Stufe wird wie bereits erwähnt, der wirksame Kanal auf die Flanken des aktiven Gebietes verlängert. Ein Indikator, welcher bei konstanter Geometrie des aktiven Gebietes dieses Verhalten widerspiegelt, ist die Transkonduktanz g_m:

$$g_m = \frac{\Delta I_D}{\Delta V_G}. \tag{5.1}$$

Diese gibt an, wie groß die Stromänderung im Kanal I_D bei einer Änderung der Steuerelektrodenspannung V_G ist. Der Parameter g_m wird um so größer, um so weiter der Kanal des Transistors ist. Somit ist zu erwarten, dass auch bei einer Variation der Stufenhöhe dieser Parameter eine Abhängigkeit zeigt. Die Entwicklung von g_m während des Programmiervorganges eines Auswahltransistors der Dimension L/W = 200x48 nm ist in Abb. 5.4 dargestellt.

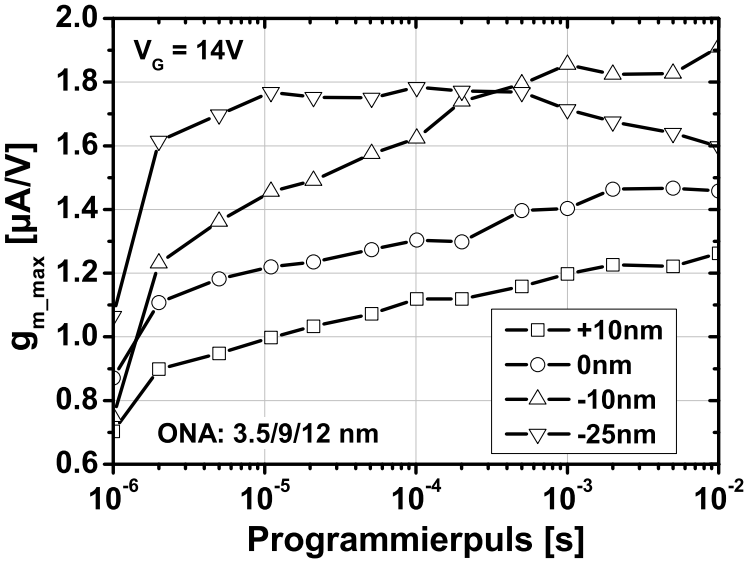

Abbildung 5.4: Vergleich des Verlaufs der maximalen Transkonduktanz g_m während des Programmierens von 200x48 nm TANOS-Zellen für verschiedene STI-Stufenhöhen

Anfänglich korreliert das Verhalten in Abhängigkeit der Stufenhöhe nicht klar miteinander. Beginnt man aber die Zellen zu programmieren, gibt es sofort eine klare Aufspaltung. Die positive Stufe mit dem kleinsten effektiven Kanal hat die niedrigste Transkonduktanz. Die Kopplung Gate zu Kanal wird immer besser und bei der Stufe -25nm, welche den größten Einfluss auf die Seitenflanke besitzt, ist g_m am größten. Mit zunehmender Zeit wird g_m immer größer, da sich durch Feldinhomogenitäten die Ladung so verteilt, dass die Kopplung immer besser wird. Allerdings wird auch deutlich, dass bei starker Feldinhomogenität, wie sie bei der -25 nm Gruppe vorliegen, es auch wieder zu einer Reduktion der Transkonduktanz kommen kann. Zum besseren Verständnis, wie inhomogen das Feld bei welcher Wahl der Stufe ist, wurde mit Hilfe von Feldsimulationen der Strukturen ein Vergleich durchgeführt. Für die drei exemplarischen Fälle +10 nm, 0 nm und -25 nm erfolgt die Darstellung der Felder in Top-und Tunneloxid in Abb. 5.5a. Die Felder wurden einerseits 2 nm oberhalb des Siliziums für das Tunneloxid und andererseits 5 nm unterhalb der Gateelektrode für das Topoxid extrahiert.
Bei der Betrachtung der Felder im Tunneloxid ist klar zu erkennen, dass mit dem Übergang von positiver zu negativer Stufenhöhe die Inhomogenität zunimmt [164]. In Tabelle 5.1 sind das minimale und maximale Feld im Bereich des aktiven Gebietes für Tunnel- und Topoxid aufgeschlüsselt. Es wird für das Tunneloxid gezeigt, dass sich das Verhältnis E_{Min}/E_{Max} von 94% für eine positive Stufe auf 78% für eine negative Stufe verschlechtert. Somit ist die Feldüberhöhung im Tunneloxid für die negative Stufe am größten. Weiterhin ist zu beobachten, dass die minimalen Felder im Tunneloxid für

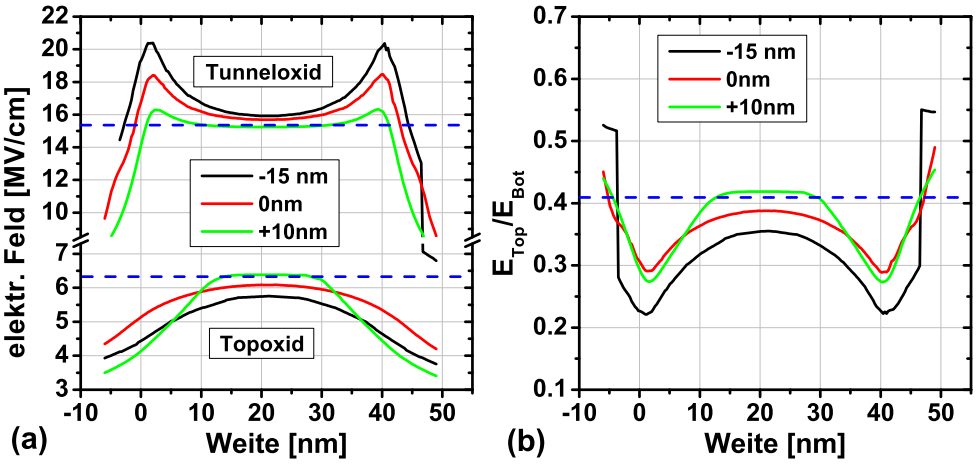

Abbildung 5.5: (a) Analyse des Feldverlaufs beim horizontalen Schnitt durch das Tunneloxid (oben) und Topoxid (unten) für versch. STI-Stufenhöhen in WL-Richtung, die blauen gestrichelten Linien zeigen das Feld einer idealen planaren Struktur; (b) zeigt das Feldverhältnis von Tunnel- zu Topoxid; Weite der simulierten Zelle ist 40 nm; simuliert wurde 2-dimensional

0 nm und -25 nm über dem, theoretisch für den 1-dimensionalen Fall, erwarteten Wert liegen, welcher durch die blaue Linie markiert ist. Der Feldverlauf im Topoxid wiederum hängt stark von der Form der Gateelektrode ab. Liegt eine positive Stufe vor, so ist die Elektrode konkav geformt und das Feld ist über dem aktiven Gebiet konzentriert, wohingegen im Bereich außerhalb der eigentlichen Speicherzelle ,auf Grund der größeren Entfernung, das Feld stark abnimmt. Ist die Stufe negativ, findet keine Konzentration über dem aktiven Gebiet statt, aber durch die konvexe Form kommt es auch zu einer leichten Reduktion der Felder im Bereich neben dem aktiven Gebiet. Die Simulationsergebnisse in Tab. 5.1 zeigen, dass die Felder im Topoxid für eine positive Stufe am größten und für eine negative Stufe am kleinsten sind. Dieses Ergebnis wird auch durch die in Abb. 5.6 a-c gezeigte Schattierung verdeutlicht.

Tabelle 5.1: Betrachtung der Felder einer 40 nm TANOS Speicherzelle mit einer 2-dimensionalen Feldsimulation für verschiedene Stufenhöhen und Zellkonzepte; minimales (E_{Min})und maximales Feld E_{Max} im Bereich der Speicherzelle, das Verhältnis minimal zu maximal η, sowie die Gesamtfeldhomogenität

Variante	Tunneloxid			Topoxid			Gesamt
	E_{Max} MV/cm	E_{Min} MV/cm	η_{Tun} %	E_{Max} MV/cm	E_{Min} MV/cm	η_{Top} %	$\eta_{Tun} \cdot \eta_{Top}(\%)$
+10 nm	16.27	15.23	93.6	6.38	4.13	64.7	60.6
0 nm	18.43	15.66	85	6.1	5.11	83.8	71.2
-25 nm	20.45	15.91	77.8	5.76	4.43	76.9	59.8
SA	17.4	15.33	88.1	6.08	4.98	81.9	72.1
SSA	15.7	15.3	97.4	6.28	6.24	99.4	96.8

Betrachtet man die Kombination aus den Feldern in Tunnel- und Topoxid stellt sich heraus, dass die nahezu planare Struktur die homogenste Feldverteilung aufweist. So

ist das Produkt aus den Felddifferenzen in den Oxiden 71% für die Stufe 0 und circa 60% für positive beziehungsweise negative Stufe. Die Bereiche der Feldüberhöhung im Tunneloxid und Feldverringerung im Topoxid liegen im gleichen Bereich und es kommt zu einer Verbesserung des Feldverhältnisses von E_{Bot}/E_{Top}, wie in Abb. 5.5b gezeigt wird. Dadurch ist eine Verbesserung der Programmiergeschwindigkeit, als auch eine Verbesserung der Löschsättigung zu erwarten.

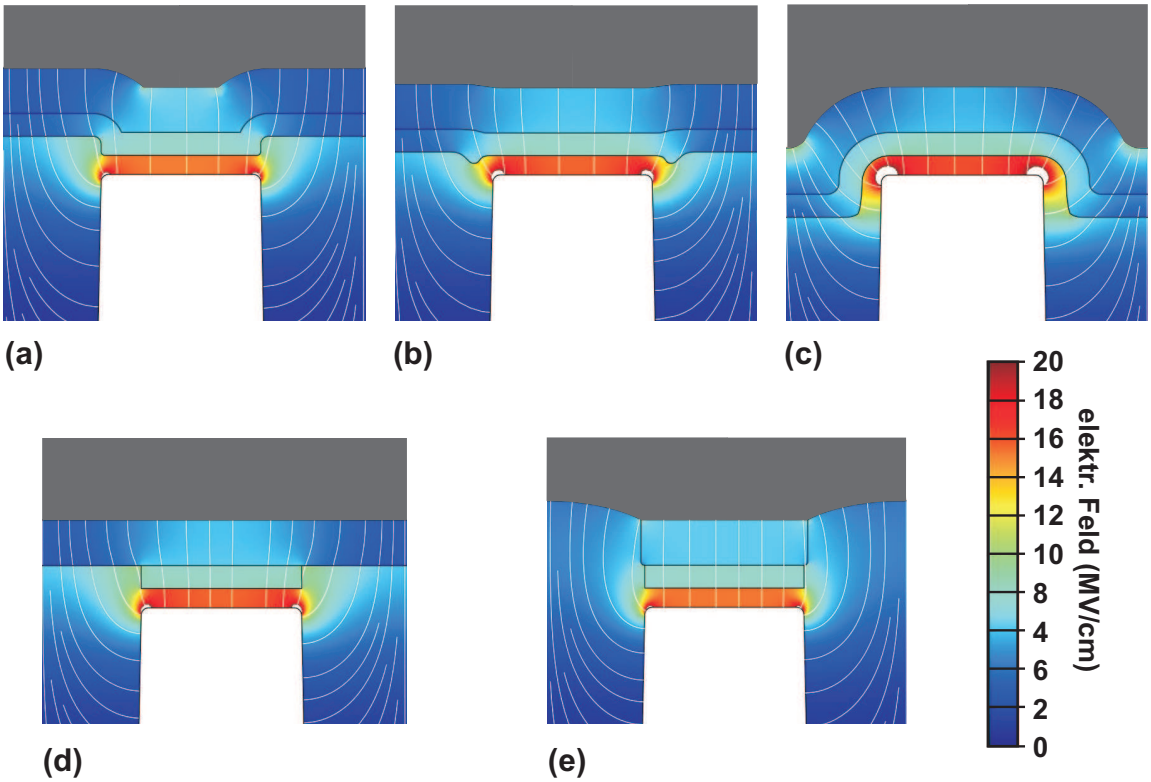

Abbildung 5.6: Darstellung der Feldstärke und Feldlinien für einen Schnitt in WL-Richtung von TANOS-Speicherzellen mit unterschiedlicher Stufenhöhe (a) +10 nm, (b) 0 nm, (c) -10 nm und den Integrationskonzepten mit strukturierter Speicherschicht (d) bzw. mit strukturiertem Topoxid und Speicherschicht (e), $V_G = 22$ V

Zusätzlich wurde eine Betrachtung für die Fälle einer Speicherzelle mit strukturierter Speicherschicht (self-aligned: SA) und zusätzlich strukturiertem Topoxid (super-self-aligned: SSA) durchgeführt. Eine Veranschaulichung der SA-Zelle erfolgt in Abb. 5.2d und bei der SSA-Zelle ist das Topoxid wie die Speicherschicht strukturiert, was in Abb. 5.6e gezeigt ist. In der Standardzelle sind beide Schichten durchgängig unter der Wortleitung vorhanden. Interessant ist, dass eine Strukturierung der Speicherschicht zu keinem Gewinn für die Homogenität der Felder, im Vergleich zur Stufe 0nm, führt. Wird aber auch das Topoxid räumlich begrenzt, wie es bei der SSA Zelle der Fall ist, ergibt sich eine starke Konzentration der Felder auf den Bereich der eigentlichen Speicherzelle. Dies resultiert in einer nahezu homogenen Feldverteilung, wobei die Gesamtfelddifferenz gerade einmal 3% beträgt. Somit könnte man schließen, dass die Zelle ein sehr gutes Programmier- und Löschverhalten haben müsste. Schaut man sich aber die Felder im Tunnel- und Topoxid an, stellt man fest, dass sich das Feld, im Vergleich zu den anderen Gruppen im Tunneloxid verringert und im Topoxid erhöht hat. Dies führt zu einem ähnlichen Verhalten, wie die positive

Stufe, und der Gewinn, der durch die Streufelder für eine neutrale und negative Stufe auftritt, ist nicht vorhanden. Demzufolge ist nicht zu erwarten, dass eine solche Struktur elektrisch besser als die Standardstruktur mit durchgehender Speicherschicht ist. Eine elektrische Charakterisierung der SA- und SSA-Zelle war nicht möglich, es soll aber das theoretische Potential der Zellen betrachtet und aufgezeigt werden.
Wie bereits in Kap. 4.1.2 beschrieben, ist das Löschen gegenüber inhomogener Feldverteilung sehr sensibel. Daher erfolgt in Abb. 5.7 eine Betrachtung des Löschens für die vier untersuchten STI-Stufenhöhen.

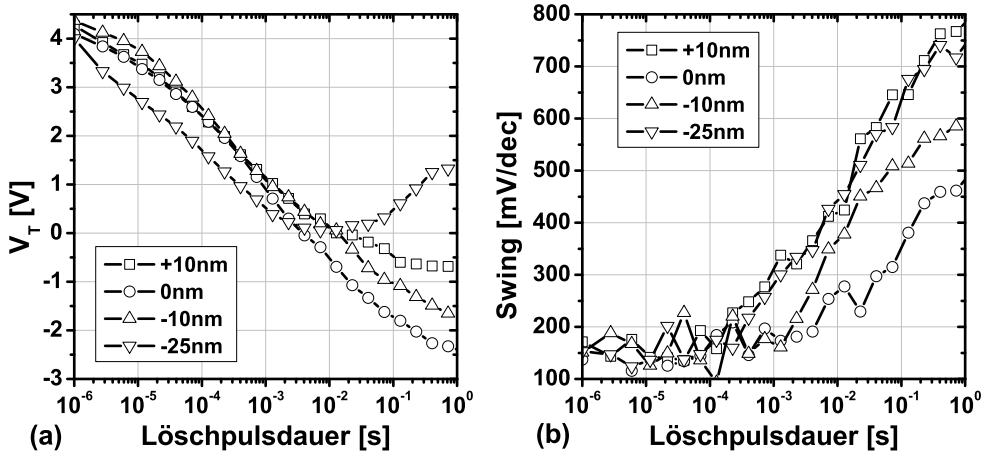

Abbildung 5.7: (a) Analyse der Löschcharakteristik für die vier untersuchten STI-Stufenhöhen und (b) Verlauf des Swings während des Löschvorgangs, 48x48 nm Speicherzellen

Zu Beginn des Löschvorganges ist die V_T-Verschiebung bei allen Gruppen, abgesehen von der stark negativen Stufe, gut vergleichbar. Die stark negative Stufe zeigt eine größere V_T-Verschiebung zu Beginn des Löschvorgangs. Dies ist darauf zurückzuführen, dass durch die höchste Feldüberhöhung (Tab. 5.1)am Tunneloxid die Löcherinjektion stark erhöht ist. Dadurch kommt es zu dem beobachteten schnelleren Löschen. Betrachtet man den weiteren Verlauf der Löschkurven, wird deutlich, dass die Speicherzelle mit der Stufenhöhe 0 nm das niedrigste V_T erreicht. Dieses Ergebnis bestätigt, dass das Löschen stark von der Homogenität der Felder im aktiven Zellbereich bestimmt ist. Denn alle Gruppen, welche auf einem höheren V_T-Niveau sättigen, zeigen eine inhomogenere Feldverteilung als die 0 nm Gruppe. Eigentlich ist anhand des Feldverhältnisses von Topoxid zu Tunneloxid (Abb. 5.5b)zu erwarten, dass die stark negative Stufe aufgrund des günstigsten Verhältnisses auf dem niedrigsten Niveau sättigt. Aber der sich bei der Integration ausbildende Gate-Stöpsel resultiert in einer konkaven Unterseite, die zu einer Feldüberhöhung in diesem Bereich führt. In Abb. 5.8a ist ein Ausschnitt des relevanten Bereiches gezeigt. Betrachtet man nun den Feldverlauf entlang einer Feldlinie in Abb. 5.8b, wird deutlich, dass das Feld an Substrat und Gateelektrode gleich groß ist. Der ausgewählte Weg entlang einer Feldlinie ist ein Teil des aktiven Gebiets der Speicherzelle. Sind im unbeladenen Zustand die Felder gleich groß, bedeutet dies, dass es nicht möglich ist, diesen Bereich weit genug zu löschen. Denn unabhängig der Polarität erfolgt eine überwiegende Elektroneninjektion aufgrund der kleineren Tunnelbarriere. Dieser Bereich wird demnach ab einem gewissen Zeitpunkt nicht mehr gelöscht, sondern programmiert. Daraus resultiert eine Zunahme des V_T's und Swings, wie in Abb. 5.7 gezeigt.

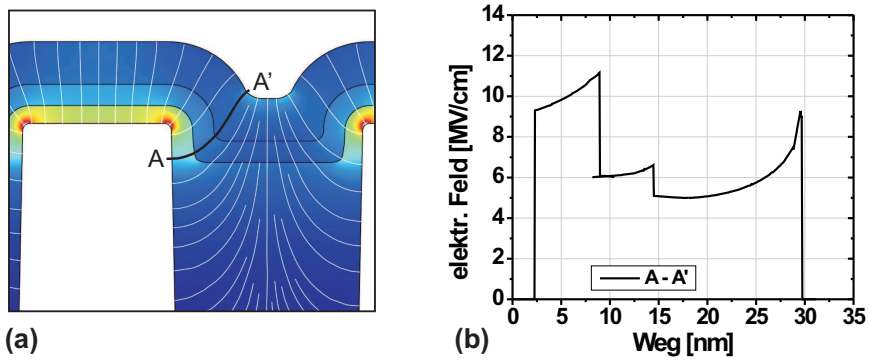

Abbildung 5.8: (a) Detailierter Ausschnitt der Feldverteilung und Feldlinien für eine stark negative STI-Stufe (b) Feldverhältnisse zwischen Gate-Stöpsel und aktivem Gebiet anhand des Feldverlaufs entlang der Feldlinien A-A'

5.1.2 Größenvergleich des elektrischen Verhaltens

Die Betrachtung der Felder lässt nun nicht klar erkennen, ob sich eine kleine Speicherzelle von einer großen Speicherzelle im elektrischen Verhalten unterscheidet. Eine große Struktur, wie ein Kondensator, repräsentiert hierbei das Verhalten einer nahezu idealen eindimensionalen Struktur. Hingegen hat die Betrachtung einer skalierten Speicherzelle in Längenrichtung gezeigt, dass sich eine im Vergleich zu großen Zellen negative Auswirkung auf die Feldverhältnisse ergibt. Es wird einmal das Feld im Tunneloxid verringert und im Topoxid erhöht. Wiederum beeinflusst die STI-Stufenhöhe in Weitenrichtung, je nach gewählter Stufe, die Felder positiv (Erhöhung des Feldes im Tunneloxid) oder negativ. Es wird angesichts der besten elektrischen Ergebnisse in Abb. 5.9 ein Vergleich einer 48 nm Zelle mit einer Stufenhöhe von 0 nm mit einer 5x5 μm großen Speicherzelle durchgeführt.

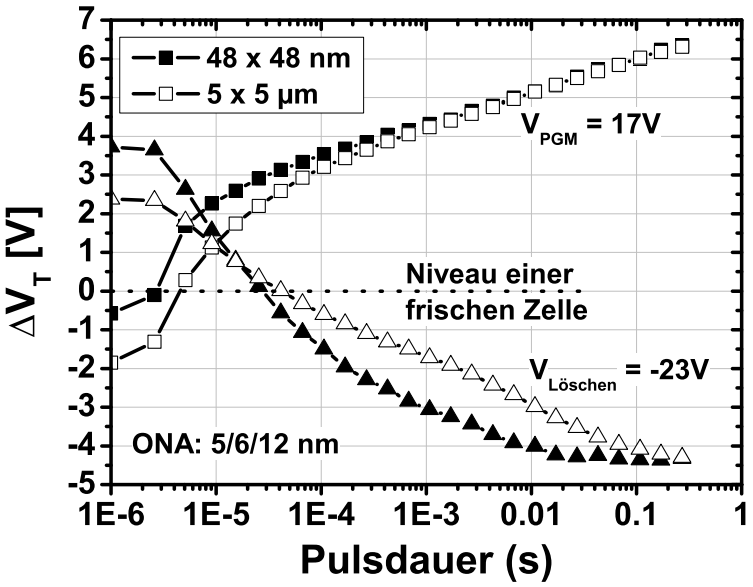

Abbildung 5.9: Vergleich einer 48x48 nm und einer 5x5 μm großen TANOS-Speicherzelle; die Datenpunkte repräsentieren den Mittelwert von 4 gemessenen Transistoren

Der Größenvergleich zeigt, dass das Programmieren nach einer gewissen Annäherungsphase identisch ist. Der Geschwindigkeitsunterschied zu Beginn des Programmierens kann mit den Feldüberhöhungen an der STI-Kante erklärt werden, siehe Abb. 5.5. Diese resultieren in einer stärkeren Elektroneninjektion und demzufolge schnellerer V_T-Verschiebung. Allerdings ist beim Programmieren der Effekt zu beobachten, dass schnell programmierende Komponenten durch langsamere überdeckt werden, wie in Kap. 4.1.1 beschrieben. Dies führt nach einer Zeit von ≈1 ms zu einem Ineinanderlaufen der Programmierkurven. Ab diesem Zeitpunkt bestimmt der planare Bereich der skalierten Speicherzelle den Programmiervorgang. Ein ähnliches Verhalten bestimmt auch das Löschen der kleinen Zelle. Zu Beginn dominieren die Feldüberhöhungen im STI-Kantenbereich die V_T-Verschiebung. Demzufolge löscht die skalierte Zelle deutlich schneller als die vergleichbare große Zelle. Zu einem ähnlichen Zeitpunkt wie beim Programmieren, beginnt der planare Bereich das Verhalten zu bestimmen. Nun ändert sich die Steigung auf einen ähnlichen Wert, wie die der großen Zelle. Der weitere Verlauf der Löschcharakteristiken ist parallel, bis beide Kurven auf einem vergleichbaren Niveau sättigen. Demzufolge kann geschlussfolgert werden, dass die skalierte Zelle durch die Feldüberhöhungen im Zeitbereich bis 1 ms einen Vorteil gegenüber der großen Zelle besitzt. Bei längeren Zeiten ist das Programmierverhalten gleich, aber der Vorteil beim Löschen bleibt bestehen. Eine Verschlechterung, wie sie durch die Längenbetrachtung aufgezeigt wurde, kann nicht beobachtet werden.

5.1.3 Integration einer Elektroden-Kapselungsschicht

Es wurde bereits in Kap. 4.3.2.1 darauf eingegangen, dass es bei der Strukturierung des Aluminiumoxids zu einer Schädigung der TaN-Elektrode kommt. In Abb. 5.10a wird eine Übersicht über die Struktur zum Zeitpunkt der Aluminiumoxidätzung gegeben. Außerdem ist der durch Paul [152] präsentierte Ätzangriff der TaN-Elektrode eingezeichnet.

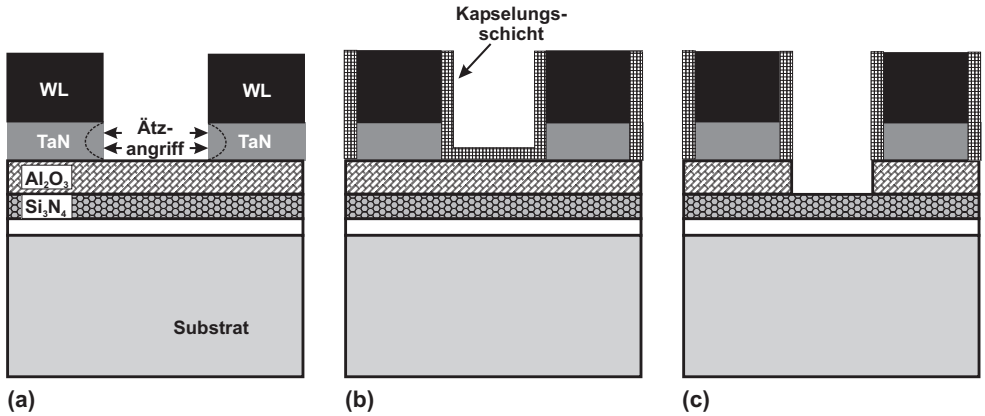

Abbildung 5.10: (a) Ausschnitt aus einer NAND-Reihe vor der Aluminiumoxidätzung, wobei auch die Schädigung der Gateelektrode verdeutlicht wird, (b) Integration einer Kapselungsschicht zum Schutz der Gateelektrode und (c) Zustand nach der Aluminiumoxidätzung für eine Integration mit Kapselungsschicht

Eine Möglichkeit diese zu unterbinden, ist das Abdecken des empfindlichen Elektrodenmaterials mit einer dünnen dem Ätzangriff resistenten Schicht. Abbildung

5.10b zeigt die Integration mit einer solchen Kapselungsschicht vor dem Beginn der Aluminiumoxidätzung. Durch die Integration einer Kapselungsschicht entsteht eine größere Weite des Aluminiumoxids im Vergleich zur Gateelektrode, wie Abb. 5.10c verdeutlicht. Dadurch ist ein Angriff auf die empfindlichen Schichten nicht mehr möglich und deren Eigenschaften werden nicht beeinflusst. Ein weiterer Effekt, ist die Abschattung der Source/Drain-Implantation, wodurch sich eine etwas längere Speicherzelle ergibt. Die Länge ist die ursprüngliche Länge zuzüglich 2x der Kapselungsschichtdicke. Für ein besseres Verständnis, wie eine solche Kapselungsschicht funktioniert, wurde ein Versuch durchgeführt, bei dem die Dicke dieser Schicht variiert. Das elektrische Verhalten, hinsichtlich Programmieren und Löschen wird in Abb. 5.11a dargestellt.

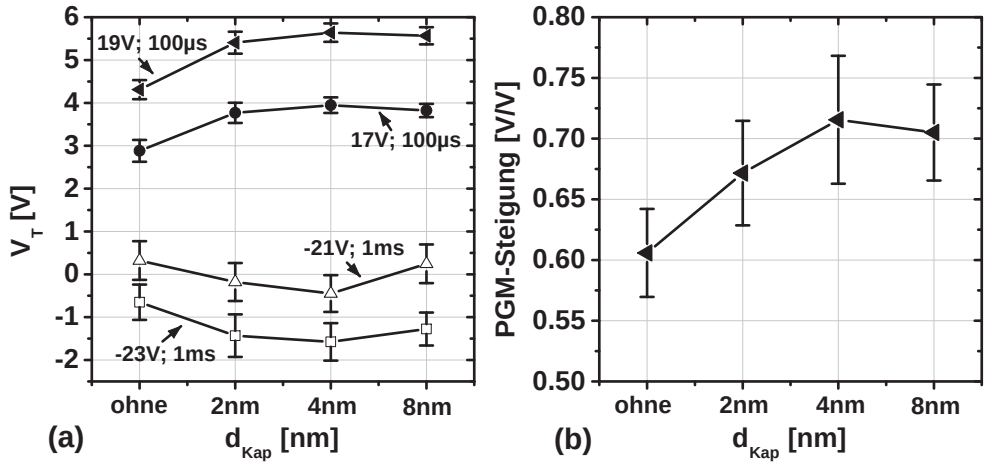

Abbildung 5.11: (a) Erreichtes Programmier- bzw. Löschniveau in Abhängigkeit der Kapselungsschichtdicke t_P= 100 μs für das Programmieren und 1 ms für Löschen bei jeweils zwei Spannungen, (b) Auswertung der ISPP-Steigung

Der Einbau einer Kapselungsschicht verbessert sowohl die erreichten Programmier- als auch Löschniveaus, wie Abb. 5.11 zeigt. So beträgt der Unterschied für das Programmieren mit und ohne Schutzschicht bei 19 V und 100 μs Pulsdauer mehr als 1 V. Dieser Unterschied kommt zustande, weil auch die in Abb. 5.11 b gezeigte Programmiersteigung eine starke Abhängigkeit von der Kapselungsschichtdicke aufweist. Es wird hierbei bei zunehmender Schichtdicke die Programmiersteigung besser, bis sie bei einer Dicke von 4nm sättigt. In Kap. 3.2.3 wurde gezeigt, dass die Programmiersteigung von dem Verhältnis des Injektionsstroms durch das Tunneloxid und dem Leckstrom durch das Topoxid bestimmt wird [72]. Erhöht man den Leckstrom durch das Topoxid, verschlechtert sich im Gegenzug die Zahl der gespeicherten Ladung und pro Schritt Spannungserhöhung ist die V_T-Verschiebung kleiner. Aus der Beobachtung der Programmiersteigung, kann nun geschlossen werden, dass ohne beziehungsweise bei schmaler Kapselungsschicht ein Leckpfad durch das Topoxid existiert. Da die Schichtdicke einen Einfluss hat, kann angenommen werden, dass sich dieser Leckpfad am Rand der Zelle befindet, wie in Abb. 5.12 gezeigt [169].
Es stellt sich nach der Ätzung des Speichernitrids eine Situation, wie in Abb. 5.12a, ein, wenn gleichzeitig die Kapselungsschicht wieder mit entfernt wurde. Für diesen Fall ist ein durch Ätzschädigungen generierter Leckpfad nicht mit der Gateelektrode verbunden. Dadurch wird der Leckstrom durch das Topoxid verringert und die Programmiersteigung wird größer. Bei dem in Abb. 5.12b gezeigten Fall für eine Ätzung

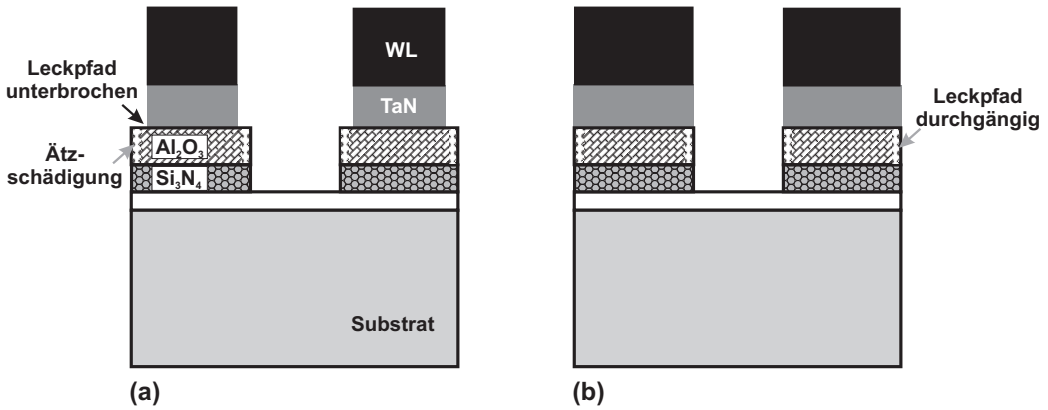

Abbildung 5.12: Schematische Darstellung des Prinzips der Kapselungsschicht; (a) Leckpfad im Al_2O_3 durch die laterale Einschnürung der Gateelektrode unterbrochen; (b) ohne Kapselungsschicht mit durchgängigem Leckpfad

ohne Kapselungsschicht befindet sich der Leckpfad zwischen Speicherschicht und Gateelektrode. Somit kann die injizierte Ladung direkt abfließen. Ein solcher Leckpfad würde auch die Ladungshaltung beeinflussen. Die Ergebnisse der Messung nach einer Temperung in Abb. 5.13a bestätigen diese Annahme.

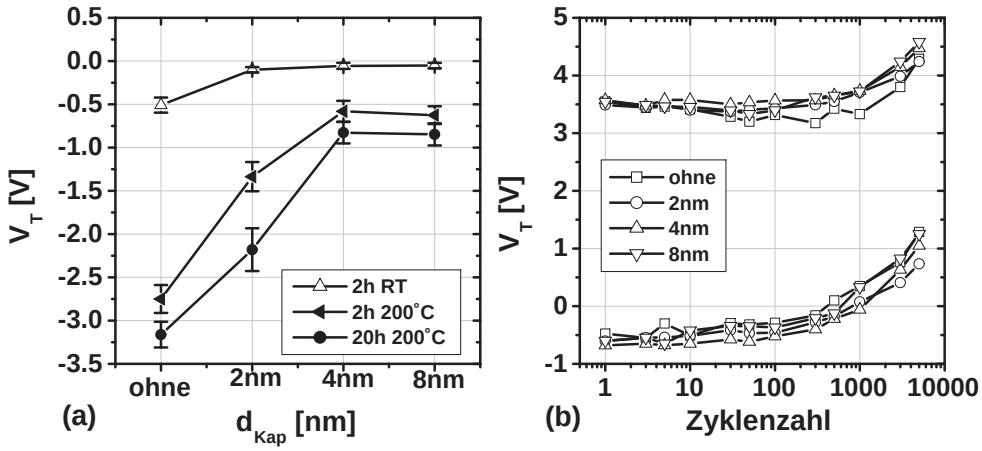

Abbildung 5.13: Auswirkung der Dicke der Kapselungsschicht auf die Ladungshaltung in (a) und auf die Zyklenfestigkeit in (b)

Die Gruppe ohne Liner zeigt einen sehr starken Ladungsverlust, der in einer V_T-Verschiebung von über 2.5 V nach 2 h 200 °C resultiert. Schnürt man den Leckpfad zunehmend ab, sättigt der Ladungsverlust bei einer Schichtdicke von 4 nm. Eine weitere Erhöhung verbessert die Ladungshaltung nicht, wie Abb. 5.13 verdeutlicht, was die präsentierte Annahme untermauert.
Allerdings wurde bei der Betrachtung des Löschens festgestellt, dass eine weitere Erhöhung der Kapselungsschichtdicke über 4 nm die Löschcharakteristik wieder verschlechtert. Dazu wurde eine genauere Analyse der Kennlinien durchgeführt, die in Abb. 5.14 dargestellt ist. Die Betrachtung des Swings in Abb. 5.14a verdeutlicht, dass die Gruppe ohne Liner einen wesentlich schlechteren Swing hat.
Dies deutet auf eine inhomogenere Ladungsverteilung im Vergleich zu den Gruppen mit Kapselungsschicht hin. Ist eine solche Schicht integriert, ist der Swing vergleichbar und degradiert auch während des Löschvorgangs nur minimal. Dies begründet

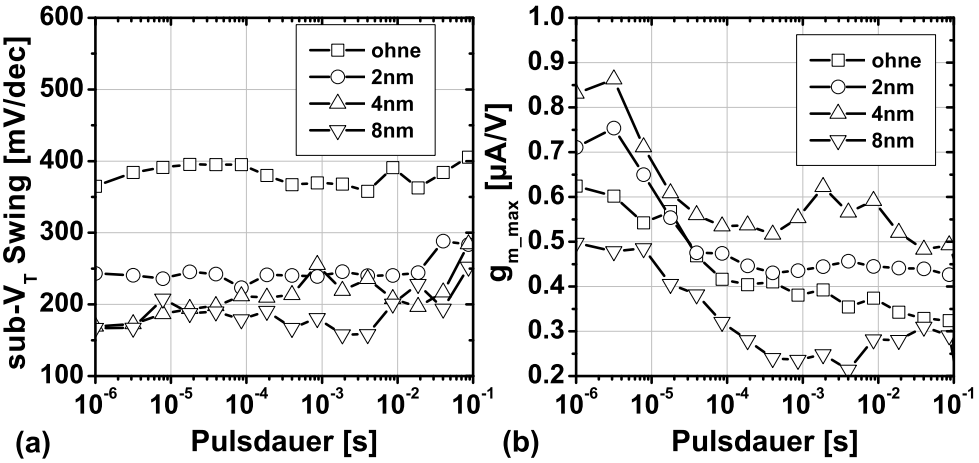

Abbildung 5.14: Analyse des Verlaufs von Swing (a) und g_{m_max} (b) während des Löschens von 48x48 nm TANOS Zellen mit V_G = -23 V; dargestellt sind jeweils die Kurven die untersuchten vier Kapselungsschichtdicken

aber noch nicht das schlechtere Verhalten der dicken Kapselungsschicht. Dies erklärt sich durch die Betrachtung der Transkonduktanz in Abb. 5.14b. Es zeigt sich, dass die 4 nm Gruppe das größte und die Gruppe mit 8 nm das kleinste g_m aufweist. Eine mögliche Ursache hierfür kann die Kopplung zwischen Kanal und Gateelektrode sein. Ein Effekt, der durch die Kapselungsschicht auftritt, ist die Abschirmung der Source/Drain-Implantation durch die über die Gateelektrode ragende Al_2O_3-Schicht (siehe Abb. 5.12a). Dadurch kommt es zu einer Verlängerung der wirksamen Kanallänge bei gleichbleibender Länge der Gateelektrode. Der Inversionskanal in der Nähe des Implantationsgebietes ist folglich mit zunehmender Kapselungsschichtdicke schlechter ausgebildet und begrenzt den Stromfluss. Dies resultiert wiederum in einer Degradation der Kennlinie und bei der Betrachtung mit einem festen Stromkriterium entsteht eine Verschlechterung des Programmierverhaltens. Weiterhin ist auch ein geringer Einfluss durch die Änderung der Speicherzelllänge möglich. Denn die Strukturbreite der Gateelektrode ist für die untersuchten Gruppen mit einer Kapselungsschicht gleich. Dementsprechend nimmt mit zunehmender Kapselungsschichtdicke auch die Breite des Al_2O_3 und der Speicherschicht zu, wie es in Abb. 5.10 gezeigt wurde. Die Speicherzelle ist dann effektiv länger und es kann zu einem zusätzlichen Einfluss durch diese Längenänderung auf die elektrischen Eigenschaften kommen.

5.2 Störmechanismen beim Betrieb von stark skalierten NAND-Speichern

Die Betrachtung der Speicherzellen im Einzelnen gibt Aufschluss darüber, ob das Speicherkonzept selbst tauglich ist, in einem zukünftigen Produkt eingesetzt zu werden. Hinzu kommt aber, dass eine Integration in ein großes Zellenfeld neben geometrischen Effekten auch Störungen auftreten, die strukturell und algorithmisch bedingt sind. Im folgenden Abschnitt werden zwei Effekte vorgestellt, die bei der Integration in NAND-Strukturen berücksichtigt werden müssen. Während der zuerst vorgestellte Mechanismus sich nur auf haftstellen-basierte Speicherzellen bezieht, tritt der zweite Effekt bei allen stark skalierten NAND-Speichern auf.

5.2.1 Unerwünschte Programmierung der Auswahltransistoren bei TANOS NAND-Speichern

Die Speicherzellen müssen für die Anwendbarkeit in einem Produkt auch eine ausreichend große Zahl an Programmier- und Löschzyklen überdauern können. Dazu werden die Zellen mit Bedingungen getestet, die einem Betrieb in einem Produkt entsprechen. Die angelegten Spannungen während des Lösch-/ Programmier- und Lesevorgangs sind in Tab. 5.2 gezeigt.

Tabelle 5.2: Spannungsbedingungen an einer NAND-Struktur während der Betriebs-Modi Lesen, Programmieren (PGM) und Löschen

	WL16	WLn	WL1;WL32	SSL	GSL	Source	BL	Substrat
Lesen	-3 - 6V	6.5V	6.5V	3V	3V	0V	0.7V	0V
PGM	17V	8V	8V	3V	0V	0V	0V	0V
Löschen	-22V	-22V	-22V	0V	0V	0V	0V	0V

Programmiert wird bei der Untersuchung der Zyklenfestigkeit zur Vereinfachung nur die WL16. Dies wird dadurch erreicht, dass nur diese Wortleitung die Programmierspannung von 17 V erhält, während alle anderen Wortleitungen auf eine Spannung von 8 V geschaltet werden. Im Gegensatz dazu werden alle Zellen einer NAND-Reihe parallel gelöscht, da normalerweise das Löschen über die Wanne erfolgt. Bei unseren Teststrukturen ist dies nicht möglich und das Löschen wird durch das Anlegen einer negativen Spannung an den Wortleitungen durchgeführt. In dem Fall wird das Wannenpotential auf 0 V gelegt. Während eines Lesevorgangs sind alle Transistoren einer NAND-Reihe eingeschaltet. Die Wortleitungen erfahren eine Spannung von 6.5 V und die die Auswahltransistoren (SSL, GSL) werden auf 3 V geschaltet. Dann ist es möglich von der untersuchten Zelle WL16 die Transferkennlinie aufzunehmen und die Schwellspannung bei einer Bitleitungsspannung von 0,7 V zu bestimmen. In logarithmisch skalierten Abständen der Zyklenzahl wird dann die Schwellspannung gemessen und deren Verlauf betrachtet, wie ihn Abb. 5.13b zeigt.
Bei der Untersuchung der Zyklenfestigkeit von NAND-Strukturen hat sich ergeben, dass es nach einer bestimmten Zahl an Programmier-/ Löschvorgängen zu einem Ausfall der NAND-Reihe kommen kann [170]. Eine erste Betrachtung hat dann aufgedeckt, dass solche NAND-Reihen keinen Strom mehr leiten. Dies geschieht, obwohl alle Transistoren eine Spannung erhalten, bei der sie im Normalfall leiten. Um das Verhalten besser zu verstehen, wurde in definierten Abständen der Programmier-/ Löschvorgang unterbrochen und die Kennlinie von den Speicherzellen an den Wortleitungen 1, 16, 32, sowie den beiden Auswahltransistoren gemessen.
Nach dem ersten Löschzyklus kann ein normales Verhalten mit einer Sättigung des Stroms durch die NAND-Reihe bei $\approx 1\mu$A beobachtet werden, wie Abb. 5.15a verdeutlicht. Dieser Sättigungsstrom ist für alle gemessenen Kennlinien gleich.
Weiterhin ist eine Abhängigkeit der Kennliniensteilheit von der Lage des gemessenen Transistors in der NAND-Reihe zu sehen. Dieser Effekt ist damit zu begründen, dass sich durch den Widerstand der in Reihe geschaltenen Transistoren unterschiedliche Spannungsbedingungen ergeben. So hat der Transistor nah an der Sourceleitung am Source nahezu das Potential von 0 V. Betrachtet man aber einen Transistor auf der Bitleitungsseite, so hat er eine Spannung größer 0 V am Sourcekontakt, resultierend

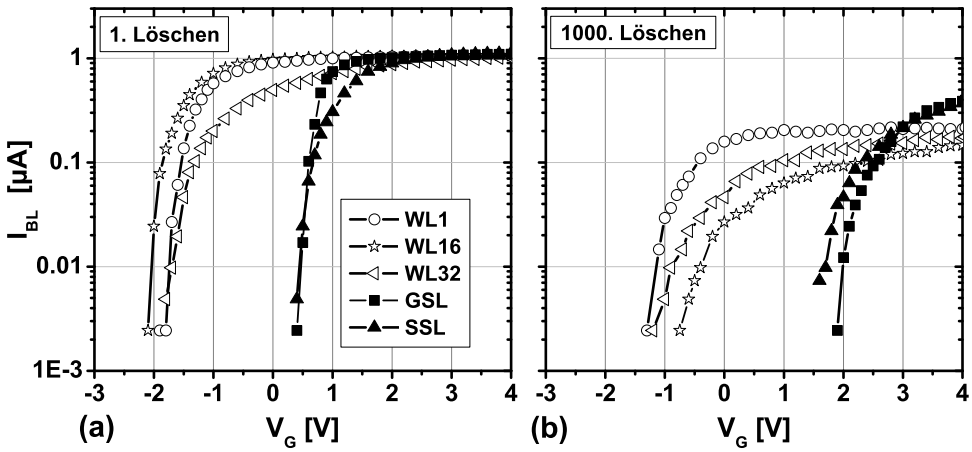

Abbildung 5.15: Transferkennlinien von den Wortleitungen 1, 16, 32 und der Auswahltransistoren auf Sourceseite (GSL) und der Seite des Bitleitungskontaktes (SSL) nach dem (a) 1. Löschvorgang und (b) 1000. Löschvorgang; Zellgröße 48x48 nm, Länge der Auswahltransistoren 200 nm

aus dem Widerstand der Transistorreihe in Verbindung mit dem Stromfluss durch die Transistorreihe. Dadurch kommt es zu einer virtuellen Substratspannung V_S, die dazu führt, dass sich die Kennliniensteilheit ändert [171]. Unter den gegebenen Bedingungen mit einer negativen Spannung V_S wird die Steilheit schlechter.
Betrachtet man die Kennlinien nach dem tausendsten Löschvorgang, stellt sich ein anderes Bild dar, so wie in Abb. 5.15b verdeutlicht. Die Kennlinien der Speicherzellen sind durch die Degradation während der Zyklen ein Stück zu höheren V_T's verschoben. Aber auch die Kennlinien der Auswahltransistoren sind auf einen höheren Wert verschoben. Bei einem Stromkriterium von 0,1 μA kann eine Verschiebung um 1,4 V extrahiert werden. Allerdings wurde in Tab. 5.2 gezeigt, dass an diese keine Spannungen angelegt werden, die zu einer Verschiebung der Schwellspannung führen können. Es werden maximal 3 V während des Lesens beziehungsweise Programmierens angelegt. Da diese Spannung auch während der Messung der Kennlinie anliegt, begrenzt der Schnittpunkt aus Kennlinie und der Auswahltransistor-Spannung von 3 V den Strom durch die NAND-Reihe. Dadurch ergibt sich für die Speicherzellen eine Begrenzung des Sättigungsstroms, der bei einer weiteren Verschiebung der Auswahltransistor-Kennlinie in einem nicht mehr messbaren Stromfluss resultiert. Somit kann geschlussfolgert werden, dass die beobachteten Lesefehler auf eine Verschiebung der Auswahltransistor-Schwellspannung zurückzuführen sind. In Abb. 5.16 ist die Abhängigkeit der Schwellspannungs-Verschiebung von der Löschpulsdauer dargestellt. Wie bereits gezeigt, verschiebt sich die Schwellspannung mit zunehmender Zyklenzahl. Dieser Effekt ist unabhängig von der Löschpulsdauer. Allerdings kann beobachtet werden, dass mit zunehmender Pulsdauer auch die Verschiebung verstärkt wird. Ein Effekt, der auf heißen Ladungsträgern basiert, wie der in Kap. 5.2.2 beschriebene Störeffekt, kann demnach ausgeschlossen werden.

Denn ein solcher, auf Ausgleichsströmen basierender, Effekt würde schon nach wenigen μs seine Programmierwirkung auf Grund eines ausgeglichenen Kanalpotentials einstellen. Es liegen hierbei Spannungsbedingungen vor, die nicht mit denen in Kap. 5.2.2 vergleichbar sind. Es wird demzufolge von einem Tunnelmechanismus ausgegangen, der zum Programmieren der Auswahltransistoren führt. Allerdings kann es

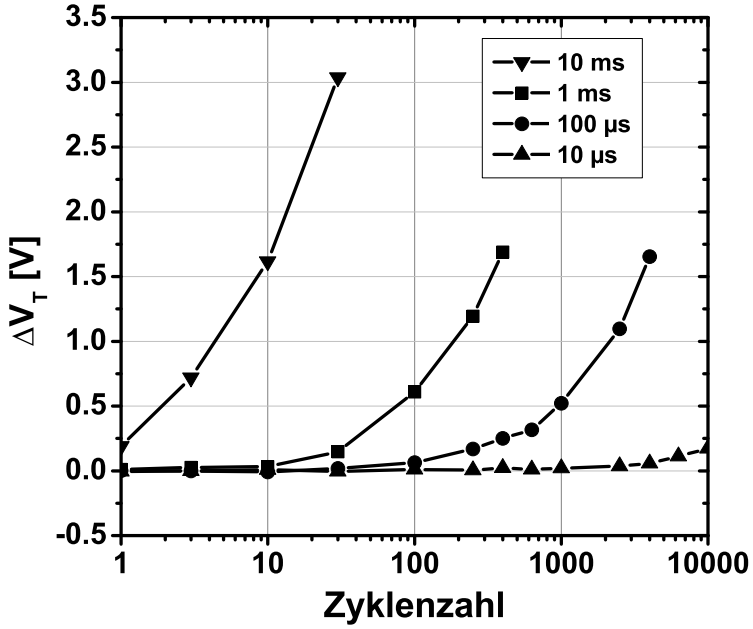

Abbildung 5.16: Verschiebung der Schwellspannung vom GSL-Auswahltransistor über die Zyklenzahl in Abhängigkeit von der Löschpulszeit; Abstand zur 1. Wortleitung = 80 nm; $V_{WL\_Löschen}$ = -22 V

sich nicht um einen dem Programmieren ähnlichen Tunnelvorgang handeln, da sich gezeigt hat, dass die Verschiebung während der Löschphase eines Zyklus auftritt. Die einzige mögliche Quelle für Elektronen während des Löschens ist die Gateelektrode der äußersten Wortleitung.

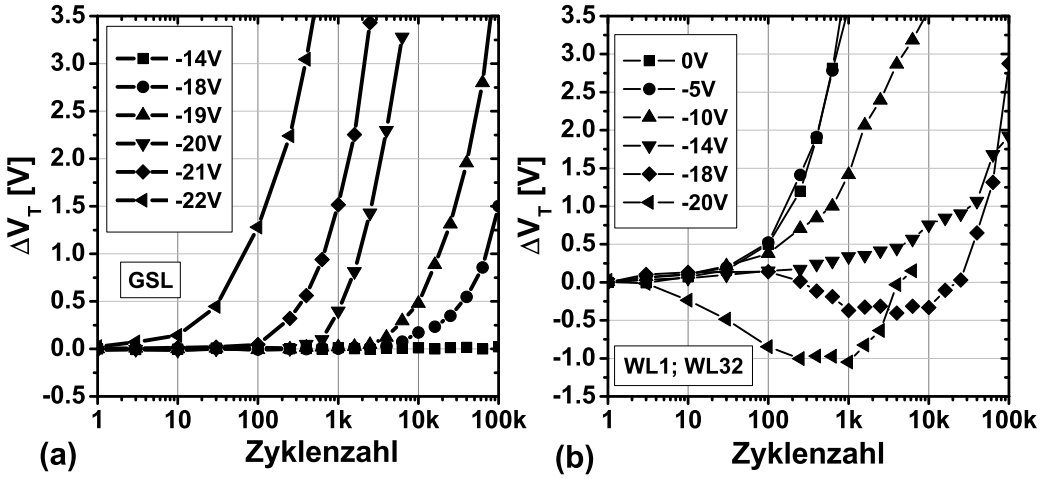

Abbildung 5.17: Betrachtung der V_T-Verschiebung für einen Auswahltransistor (a) und die äußeren Wortleitungen während der Programmier- /Löschvorgänge in Abhängigkeit der Löschspannung auf den Rand-Wortleitungen; Abstand GSL - WL = 80 nm

Um diese Quelle genauer zu untersuchen, wurde die Spannung auf den äußeren Wortleitungen variiert und alle anderen Wortleitungen mit der normalen Spannung von -22 V beschalten. Die Auswirkung auf die Schwellspannungsverschiebung der Auswahltransistoren ist in Abb. 5.17a dargestellt. Es zeigt sich, dass eine Verringerung

der Löschspannung auf den äußeren Wortleitungen eine Verbesserung herbeiführt. Dies geht soweit, dass bei einer Löschspannung von -14 V bis zu einer Zyklenzahl von 100.000 keine Verschiebung der Schwellspannung mehr beobachtet werden kann. Dieses Ergebnis bestätigt die Annahme, dass die äußeren Wortleitungen als Quelle der Ladungsträger, welche in der Speicherschicht der Auswahltransistoren gespeichert werden, fungieren. Geht man noch einen Schritt weiter, so kann man den Effekt bestätigen, indem man die Rand-Wortleitungen wie einen Auswahltransistor betreibt. Hierbei sollte ein vergleichbares Verhalten wie bei den Auswahltransistoren beobachtet werden. Aus diesem Grund wurde die Löschspannung auf den inneren Wortleitungen bei -22 V konstant gehalten und die Spannung auf den äußeren Wortleitungen variiert. Die Messergebnisse in Abb.5.17 zeigen zwei sich überlagernde Effekte. Der Verlauf ist oberhalb einer Spannung von -14 V auf den äußeren Wortleitungen durch das eigentliche Zellverhalten bestimmt. Es zeigt sich erst eine Abnahme der Schwellspannung, wie es bei normalem Löschen beobachtet wird. Im Anschluss nimmt die Schwellspannung wieder zu, welche auf Zelldegradation zurückgeführt werden kann. Verringert man die Löschspannung unter -14 V, kommt es zu einer Änderung der Charakteristik und der beschriebene Injektionseffekt von der benachbarten Wortleitung ist zu beobachten. Dieser führt wieder zu einem Anstieg der Schwellspannung. Diese Schwellspannungsänderung, durch eine Injektion von der Nachbar-Wortleitung hervorgerufen, sättigt bei einer Wortleitungsspannung von $\approx$ -5 V. Das Verhalten kann damit erklärt werden, dass zwischen den Zellen ein gewisses laterales Feld existieren muss, um die Ladungsträger in die Speicherschicht der äußeren Wortleitung zu transportieren. Oberhalb einer Löschspannung von -14 V ist das laterale Feld für einen effektiven Transport zu klein. Erhöht man das laterale Feld durch eine Verringerung der Spannung auf den äußeren Wortleitungen während des Löschens, wird die Injektion effizienter, wodurch eine stärkere Verschiebung der Schwellspannung resultiert. Damit kann das bei den Auswahltransistoren beobachtete Verhalten bestätigt werden.

Für eine Untersuchung der Ladungs-Transportwege wurde eine Simulation der elektrischen Felder im Bereich zwischen äußerer Wortleitung und Auswahltransistor durchgeführt. Elektrische Felder und die entsprechenden Feldvektorlinien sind in Abb. 5.18 gezeigt. Es werden wichtige Faktoren aufgezeigt, die schlussendlich gemeinsam dazu führen, dass dieser Störmechanismus zum Tragen kommt. Es wird durch die Farbgebung in Abb. 5.18 veranschaulicht, dass das Feld auf der Seitenflanke der Gateelektrode im Bereich von 8 MV/cm ist. Es ist somit groß genug für eine effektive Elektronen-Emmission von der Oberfläche. Elektronen welche nun in das Leitungsband des Fülloxides injiziert wurden, werden aufgrund des lateralen Feldes beschleunigt. Die Bewegung entlang der Feldlinien führt zu einem Ladungstransport entweder in die Speicherschicht des Auswahltransistors oder das Substrat. Ein Ladungstransport über größere Entfernungen ist bei entsprechenden Feldbedingungen möglich, wie in [118] nachgewiesen wird. Allerdings ist auch zu erwarten, dass eine Erhöhung des Abstandes zu einer Verringerung der transportierten Ladung führt. Denn das Fülloxid ist nicht ideal defektfrei und es werden auf dem Transportweg Ladungen eingefangen. Weiterhin verschieben sich die Feldlinien auf der Wortleitungs-Seitenwand so, dass das Feld auf der Oberfläche kleiner wird und weniger Ladungen emittiert werden. Die erwartete Verbesserung stellt sich ein, wie Abb. 5.19a verdeutlicht. Es wird gezeigt, dass es bei einem Abstand von 50 nm bereits nach 100 Programmier-/ Lösch-Zyklen zu einer Schwellspannungsverschiebung von mehr als 1 V kommt.

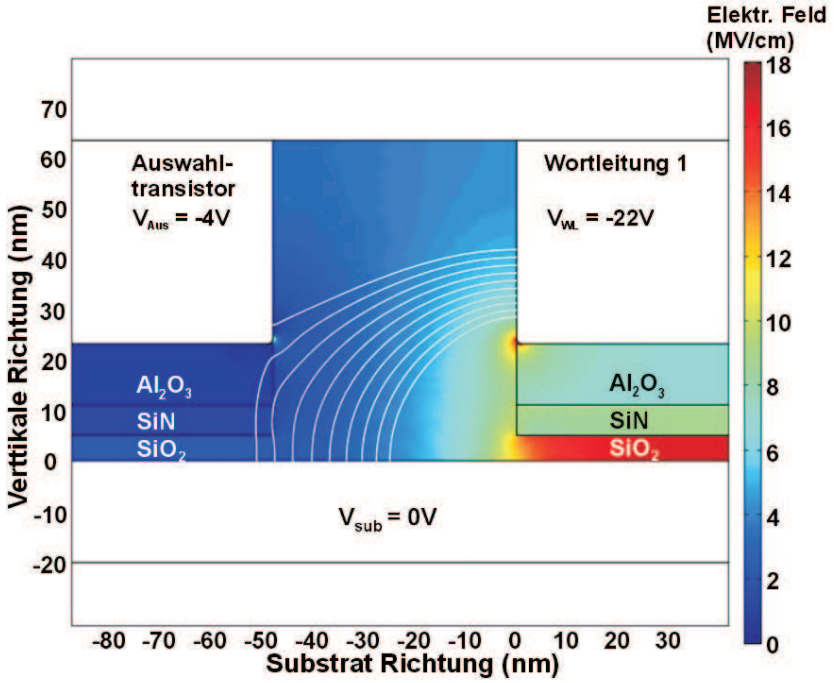

Abbildung 5.18: Analyse des Feldes zwischen Auswahltransistor (links) und äußerster Speicherzelle (rechts) anhand von 2D-Feldsimulationen, gezeigt ist der Transportweg der Ladungsträger, der dem Verlauf der Feldlinien entspricht

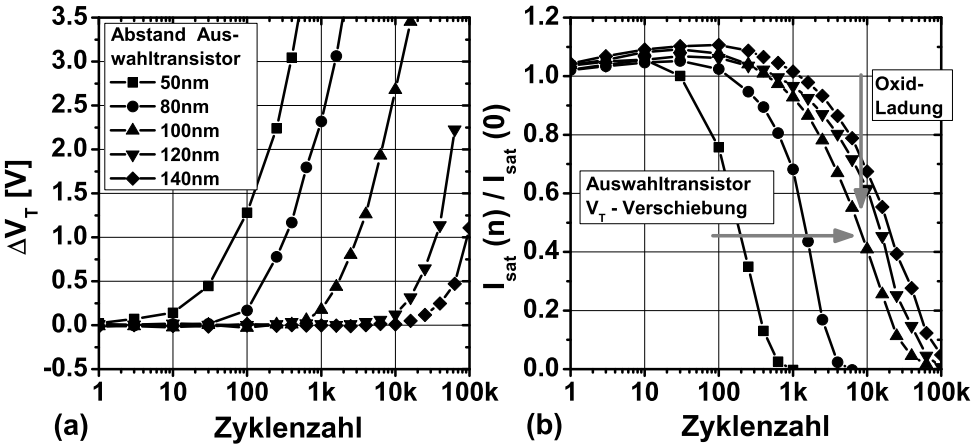

Abbildung 5.19: (a) Abhängigkeit der V_T-Verschiebung des Auswahltransistors während des Zykelns bei einer Änderung des Abstandes zwischen Auswahltransistor und äußerer WL, (b) normierter Verlauf des NAND-Reihen Sättigungsstromes über der Zyklenzahl, zudem eingezeichnet sind der Einfluss der Auswahltransistor V_T-Verschiebung und der im Oxid gespeicherten Ladung

Wird der Abstand vergrößert, verschiebt sich die Änderung der Schwellspannung hin zu höheren Zyklenzahlen. Bei einem Abstand von 140 nm wird eine Verschiebung um 1 V erst nach 100.000 Zyklen erreicht und stellt eine erhebliche Verbesserung dar. Der Einfluss auf den Sättigungsstrom der NAND-Reihe zeigt ein anderes Bild, welches in Abb. 5.19b zu sehen ist. Wie erwartet verschiebt sich die Kurve, die die Ab-

nahme des Sättigungsstromes zeigt, mit zunehmendem Abstand auch hin zu höheren Zyklenzahlen. Aber ab einem Abstand von ≈100 nm verlangsamt sich diese Verschiebung. Dieser unerwartete Effekt kann auf den starken Stress zurückgeführt werden, dem zum Beispiel das Fülloxid ausgesetzt ist. Durch Abb.5.18 wird deutlich, dass ein Großteil der Feldlinien im Substrat endet. Da auch das Feld an der Austrittsstelle auf der Gateelektrode größer ist, fließt ein nicht vernachlässigbarer Strom durch das Oxid. Dieser Strom schädigt das Oxid und es kommt zur Speicherung von Ladungsträgern. Nach [56] kommt es zu Beginn des Stresses zu einer vorübergehenden Speicherung von Löchern an der Oxid-Grenzfläche. In der Betrachtung des NAND-Reihen-Sätttigungsstromes äußert sich dies durch einen Anstieg des Stromes bei kleinen Zyklenzahlen. Die anschließend einsetzende Speicherung von Elektronen im Bereich des Oxides zwischen Auswahltransistor und äußerer Wortleitung führt zu einer Verringerung des Stromes. Dieser Effekt ist vergleichbar mit dem Anstieg der Flachbandspannung bei einer MOS-Struktur.
Ein möglicher Weg, um den Effekt zu unterdrücken, ist demnach nicht die Vergrößerung des Abstandes. Auch, weil aus Produktsicht eine Vergrößerung zu viel Platz benötigen würde und die Flächeneffizienz des Speicherzellenfeldes reduziert. Es hat sich aber gezeigt, dass eine Reduktion der Spannung auf den äußeren Wortleitungen den Störmechanismus reduziert, wie es in Abb. 5.17 demonstriert wird. Diesen Effekt auf den Störmechanismus kann man nutzen, um ohne großen Aufwand diesen zu reduzieren. Die Wirkung auf Auswahltransistor und äußere Wortleitung führt zu einem Spannungsbereich, in dem die Verschiebung der Schwellspannung jeweils am kleinsten ist. Dies wird durch Diagramm 5.20 illustriert.

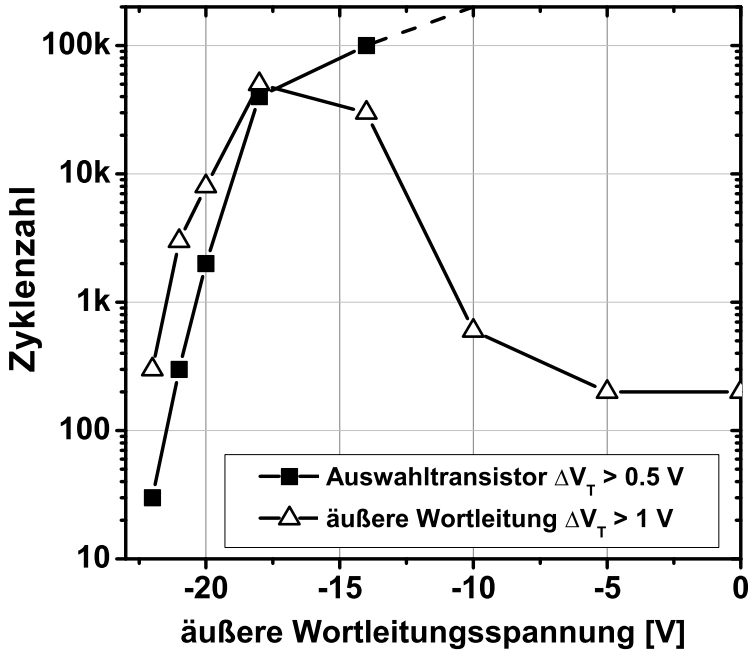

Abbildung 5.20: Bestimmung der maximal erreichbaren Zyklenzahl in Abhängigkeit der Löschspannung auf den äußeren WL, berücksichtigt wird die V_T-Verschiebung von Auswahltransistor (geschlossene Symbole) und äußerer WL (offene Symbole)

Es wird gezeigt, dass sich ein Fenster im Bereich von -14 V bis -18 V bei einer WL-Löschspannung von -22 V ergibt. Durch eine Wahl der Spannung auf den äuße-

ren Wortleitungen in diesem Bereich bleiben die Schwellspannungsverschiebungen so klein, dass bis 30.000 Zyklen kein nennenswerter Einfluss auf den Zellstrom zu erwarten ist. Der Betrieb der äußeren Wortleitungen mit abweichenden Spannung zur Unterdrückung von Störeffekten entspricht dem Dummy-Wortleitungs-Konzept [43,172,173]. Die Anwendung dieses Konzepts zur Unterdrückung des GIDL-Störmechanismus, wie in Kap. 5.2.2 vorgestellt, ermöglicht es auch im Löschfall die dafür vorgesehenen Strukturen zu verwenden.
Ein anderer Ansatz zur Unterdrückung dieses Störphänomens ist die Entfernung der Speicherschicht im Auswahltransistor. Dann ist die Schwellspannung des Auswahltransistors nicht mehr verschiebbar und es gibt keinen Einfluss auf den Sättigungsstrom mehr. Allerdings wird es weiterhin eine Begrenzung der kumulativen Löschzeit durch die Ladungsspeicherung im Fülloxid geben.

5.2.2 Unerwünschte Programmierung der äußeren Speichertransistoren bei erhöhtem Kanalpotential

Ein weiterer Effekt der sich durch die zunehmende Verringerung der Strukturgrößen ergibt, ist das durch Joo [174] im Jahr 2005 vorgestellte unerwünschte Programmieren der äußeren Speichertransistoren eines NAND Strings. Es wird hierbei beobachtet, dass sich das V_T der am Select-Transistor liegenden Speicherzelle verschiebt, wenn man dort eine Spannung anlegt, welche normalerweise nicht zu einer Verschiebung führt. Besonders ist dabei, dass das Programmieren in einem String erfolgt, der nicht programmiert werden soll und demzufolge ein hohes Kanalpotential aufweist (siehe Kap. 2.4.3). Dieser Effekt wird in Abb. 5.21a anhand einer Spannung zwischen 10 und 11 V verdeutlicht. Normalerweise programmieren FG-Speicherzellen erst ab einer Spannung von circa 17 V. Aber im gezeigten Fall beginnt die dem String-Select-Transistor (SSL) am nächsten liegende Wortleitung (W/L = 0) bereits ab 10 V eine signifikante Verschiebung zu zeigen.

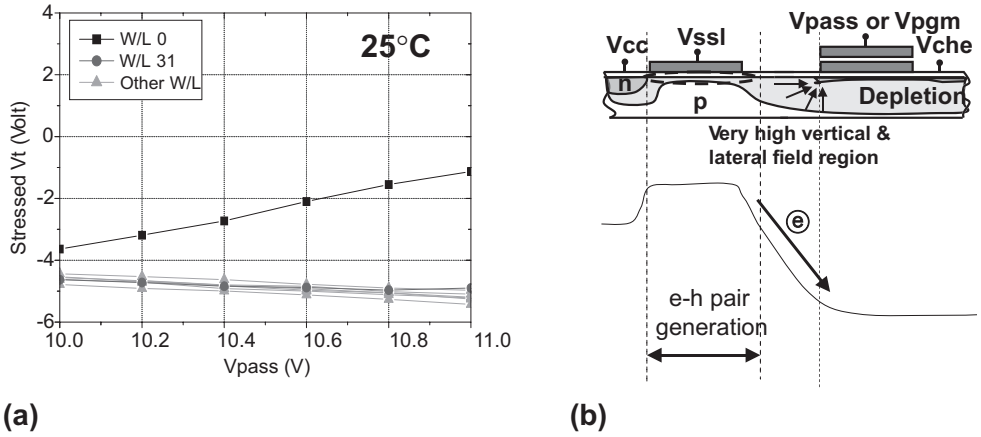

Abbildung 5.21: aus [175]; (a) Gemessene Verschiebung der Schwellspannung von 1. WL im Vergleich zu den anderen Wortleitung für verschiedene Pass-Spannungen; (b) Mechanismus der heißen Elektronen, die zum ungewollten Programmieren führen

Erklärt wird der Effekt damit, dass sich aufgrund der Strukturverkleinerung ein Effekt, wie bei den in Kap. 2.3.2.2 beschriebenen heißen Ladungsträgern, einstellt. Mit

der Verkürzung des Abstandes zwischen SSL und der ersten Speicherzelle erhöht sich das laterale Feld so weit, dass die Ladungsträger ausreichend Energie erhalten, um das Tunneloxid der ersten Speicherzelle zu überwinden. Dieses Verhalten wird in Abb. 5.21b veranschaulicht. Die Quelle der Elektronen ist der GIDL-Strom (siehe Kap. 2.2.3) des Transistors. Dieser kommt dadurch zustande, dass der Kanal der Speicherzelle aufgrund des Inhibits auf einem Potential von ≈ 7 V (V_{che}) liegt. Gleichzeitig ist das Gate des Auswahltransistors auf 3.3 V V_{SSL} geschaltet. Dadurch ergibt sich eine effektiv negative Gatespannung und es kommt zur Elektronen-Lochpaar Generation im p-n-Übergang das Auswahltransistors. Die Löcher fließen in das Substrat ab, während die Elektronen anschließend in Richtung des hohen Potentials im Kanal der Speicherzellen abfließen. Sie werden dann durch das Feld, welches durch die Potentialdifferenz zwischen Substrat und angehobenem Kanalpotential existiert, beschleunigt und es besteht die Möglichkeit in die erste Speicherzelle zu gelangen. Die Möglichkeiten den Effekt zu unterdrücken, sind ähnlich dem im vorangegangen Kapitel, da auch dieser Effekt auf einem zu großen lateralen Feld beruht. Daher verringert eine Vergrößerung des Abstandes zwischen SSL und Speicherzelle den Effekt. Eine geschickte Lösung den Störmechanismus zu berücksichtigen, ist die Reduktion der Zahl an Programmierzuständen auf den äußeren Speichertransistoren und damit eine größere Störtoleranz [2]. Dies ist eine Verbesserung des von Park [172] vorgeschlagenen dummy-WL Konzepts. Hier werden anstatt von Speicherzellen zwei zusätzliche Transistoren mit eingebaut, die später aber nicht als Speicherzelle genutzt werden. Diese werden so betrieben, dass sie immun gegen den Störmechanismus sind und werden nicht zur Informationsspeicherung genutzt. Zum Beispiel besteht dann ein NAND-String nicht aus 32 Wortleitungen, sondern aus 34 zuzüglich der Auswahltransistoren. Zunächst ist die Flächeneffizienz bei Strukturen größer 40 nm schlechter als eine Struktur mit einem größeren Abstand zwischen äußerer WL und Auswahltransistor. Dieser Abstand sollte für eine effiziente Unterdrückung des Störmechanismus größer 100 nm sein. Nimmt man das Konzept mit zusätzlicher WL, wird 3 mal die Weite der kleinsten Strukturgröße benötigt. Daher ergibt sich ein Flächengewinn für Strukturen mit weniger als ≈ 35 nm.

6 Zusammenfassung und Ausblick

6.1 Zusammenfassung

In dieser Arbeit wurden haftstellen-basierte Speicherzellen hinsichtlich ihrer Funktionalität mit Hilfe von Simulationen und ausgesuchten elektrischen Messungen analysiert. Es ergab sich basierend auf Simulationsdaten, dass die Grenzfläche zwischen Tunneloxid und Siliziumnitrid-Speicherschicht nicht als ideal betrachtet werden kann. Besser ist für die Berechnung der Tunnelströme einen Bereich an der Grenzfläche zwischen SiO_2 und SiN anzunehmen, in dem ein gleitender Übergang der unterschiedlichen Leitungsbandenergien erfolgt. Aufbauend darauf muss auch beachtet werden, dass in der Simulation die Injektionstiefe, ab der die Elektronen gespeichert werden können, berücksichtigt wird. Durch diese beiden Maßnahmen kann eine deutliche Verbesserung der Übereinstimmung zwischen Simulation und Messung erzielt werden. Die Abhängigkeit von der Gatespannung kann verbessert werden, wenn man den Einfangquerschnitt für die Ladungsträger anpasst. Es hat sich gezeigt, dass für eine gute Übereinstimmung ein verhältnismäßig kleiner Wert von circa $1e^{-15}\ cm^2$ gewählt werden muss. Dadurch stellt sich eine sehr homogene vertikale Ladungsverteilung und es fließt ein relativ großer Strom über das Topoxid ab.
Im Gegensatz zu den Simulationsergebnissen deuten die Messergebnisse auf eine vertikal lokalisierte Ladungsspeicherung hin. Sowohl bei SONOS- als auch bei TANOS-Schichtstapeln wurde beobachtet, dass die Ladungshaltung stark durch die Siliziumnitriddicke bestimmt ist. Durch das Programmieren mit negativer Gatespannung bei SONOS wird eine deutlich bessere Ladungshaltung erzielt, was nur über eine lokale Ladungsspeicherung auf der Injektionsseite der Ladungsträger erklärbar ist. Bei TANOS kommt mit der Ladungsspeicherung im Aluminiumoxid eine weitere zu beachtende Komponente hinzu. Aber auch hier lässt sich aufgrund des nitridschichtdicken-abhängigen Ladungsverlustes ableiten, dass die Ladung in der Speicherschicht inhomogen verteilt ist. Dieses Ergebnis wird durch die Betrachtung des Ladungsneutralpunktes untermauert. Abhängig vom Programmiermodus ergeben sich bei identischer Probe und gleichem V_T nach dem Tempern unterschiedliche V_T-Verschiebungen. Die Ausbildung von unterschiedlich stark lokalisierten Ladungsverteilungen für Löcher und Elektronen erklärt die Beobachtungen. Weiterhin wird durch die Messungen deutlich, dass die gespeicherten Löcher trotz intensiver Elektroneninjektion nicht mehr aus der Speicherschicht entfernt beziehungsweise neutralisiert werden können.
Andererseits wurde der Einfluss der geometrischen Verhältnisse, sowie der Einfluss von der Materialwahl im Schichtstapel und der Gateelektrode untersucht. So hat sich gezeigt, dass die Wahl der Gateelektrode stark vom Material und bei Metallelektrode von dem Abscheideverfahren abhängig ist. Als bestes Gateelektrodenmaterial für die TANOS-Zelle hat sich TaN aus einer gepulsten CVD-Abscheidung herausgestellt. Weiterhin wurde deutlich, dass reines Aluminiumoxid mit einer Temperung von 1100°C zu den besten Ergebnissen führt. Zudem ist zu beachten, dass im Aluminium-

oxid möglichst keine Überlappung zur Gateelektrode auftritt. Dies kann zum Beispiel durch die Aluminiumoxid-Ätzung eingestellt werden und bei einer senkrechten Flanke wird das elektrische Verhalten der Speicherzelle deutlich verbessert. Ferner kann es bei der Prozessierung des Schichtstapels zu einer Schädigung des Gateelektrodenmaterials kommen. Die Einführung einer Kapselungsschicht, welche eine Schädigung verhindert, resultiert in einer erheblichen Verbesserung aller elektrischer Parameter stark skalierter TANOS-Speicherzellen. Besonders stark ist die Verbesserung der Ladungshaltung, was auf eine Unterdrückung eines Leckpfades an der Aluminiumoxid-Seitenflanke zurückgeführt werden konnte. Weiterhin konnte bei dem Vergleich von TiN und TaN eine Korrelation zwischen Ladungshaltung und mechanischer Verspannung gefunden werden. Hierbei zeigt sich durch die bei der Abscheidung von TiN induzierte größere mechanische Spannung ein größerer Ladungsverlust.
Bei der Integration in Speicherzellenfeldern ergeben sich neue Gesichtspunkte die berücksichtigt werden müssen. So kommt aufgrund der inhomogenen Felder, welche durch die STI-Strukturierung induziert werden, zu einer effektiven Verbesserung der elektrischen Eigenschaften. Diese Verbesserung tritt aber nur auf, wenn es sich um eine quasi planare Strukturierung handelt. Wird die STI-Stufe auf eine positive oder negative Stufe eingestellt, kommt es zu einer Verschlechterung, vor allem der Lösch-Geschwindigkeit und -Sättigung. Aber auch Streufelder in Längenrichtung führen aufgrund der geringen Strukturgröße zu einer Verschlechterung der Zelleigenschaften. Insgesamt ist aber bei der Überlagerung beider feldinduzierter Effekte eine leichte Verbesserung gegenüber großen Zellen hinsichtlich der Löschgeschwindigkeit zu beobachten. Weiterhin ist von großem Interesse, ob der Speicherschichtstapel auch in den Auswahltransistoren verwendet werden kann. Es hat sich gezeigt, dass bei den untersuchten Proben ein unerwünschtes Programmieren dieser erfolgt. Es konnte nachgewiesen werden, dass es sich um einen Injektionsmechanismus handelt, der von den äußeren Wortleitungen bei Löschbedingungen ausgeht und zu der beobachteten Programmierung führt. Die Programmierung führt zu einer Reduktion des Lesestroms und im Extremfall zum Ausfall der NAND-Reihe.

6.2 Ausblick

In dem Bearbeitungszeitraum dieser Arbeit konnten erhebliche Fortschritte bei der Entwicklung von haftstellenbasierten Speicherzellen erzielt werden. So konnte zum Beispiel die Ladungshaltung durch eine entsprechende Auswahl der Materialien und Integrationskonzepte für die TANOS-Integration von 2 V nach 2 h bei 200°C Temperung auf 500 mV verbessert werden. Gleichzeitig konnten weitere Eigenschaften, wie Programmieren und Löschen mit verbessert werden. Allerdings konnten bis zum gegenwärtigen Zeitpunkt nicht alle elektrischen Eigenschaften der Floating-Gate Technologie erreicht werden. Dennoch bietet TANOS die Möglichkeit mit geringeren Spannungen zu Programmieren und ist daher für zukünftige Speicherzellgrößen besser geeignet. Es wird vielfach die Lithografie als begrenzende Größe für die weitere Miniaturisierung angegeben. Allerdings ist mit Hilfe des sogenannten 'Double Patterning'-Ansatzes [176] diese Grenze auf unter 25 nm verschoben. In diesem Konzept wird mit Hilfe von integrativen Maßnahmen eine Verdoppelung der Strukturen bei halbierter Strukturweite erzielt. Demnach ist das Zellkonzept selbst die begrenzende Größe. Durch eine weitere Optimierung der TANOS-Struktur ist es möglich, im Ge-

gensatz zum Floating-Gate Konzept, diese den verschiedenen Anforderungen anzupassen. So gibt es Speicher, von denen eine hohe Zuverlässigkeit bei gleichzeitig hoher Zyklenzahl gefordert ist. Für die Gewährleistung dieser Bedingungen werden die Zellen mit nur einem Programmierzustand betrieben (SLC). Hierfür kann die einfache TANOS-Struktur verwendet werden, da die Datenhaltung für diesen Fall ausreichend ist. Wird allerdings der MLC-Betrieb genutzt, sind erhöhte Anforderung an die Ladungshaltung gestellt. Aussichtsreiche Konzepte, die auch eine Anwendung von haftstellenbasierten Speichern für diese Anforderungen erlauben, sind die Modifikation des Tunneloxids [128] und die Integration einer Schicht aus SiO_2 zwischen Nitrid-Speicherschicht und Al_2O_3-Topoxid [126]. Die Modifikation der Tunnelbarriere verbessert zwar erheblich die Löschgeschwindigkeit, allerdings sind die Zuverlässigkeit und Reproduzierbarkeit noch nicht auf einem produkttauglichen Niveau. Auch die Einführung einer Zwischenschicht verbessert nicht nur die Ladungshaltung, sondern verschlechtert im Gegenzug das erreichbare Löschniveau [177]. Hierfür sind weitergehende Untersuchungen notwendig, die diese Beobachtungen hinreichend erklären. Auch die in dieser Arbeit aufgetretene Diskrepanz zwischen Injektionssimulation und qualitativer Betrachtung des Ladungsverlustes erfordert weitergehende Untersuchungen. Kann die Simulation und Messung besser in Einklang gebracht werden, wird das Verständnis für haftstellenbasierte Speicherzellen weiter verbessert. Dann ist es auch möglich Optimierungen nicht nur empirisch, sondern auch gezielt durchzuführen. Dann steht der Integration in 3-dimensionale Strukturen nichts mehr im Weg und es können kostengünstige Speicher höchster Dichte gefertigt werden [6, 7].

Danksagung

Ein besonderer Dank gilt Prof. Dr.-Ing. Thomas Mikolajick der mir die Möglichkeit gab, meine Doktorarbeit zu schreiben und auch in schwierigeren Momenten stets eine Lösung für die problemlose Fertigstellung meiner Arbeit gefunden hat.

Florian Beug danke ich sehr für die von Anfang an sehr angenehme Zusammenarbeit und die immer hilfreichen Diskussionen und Hilfe bei der Bearbeitung der vielen Themen im Zusammenhang mit meiner Doktorarbeit.

Lars Bach und Jan Paul haben einen großen Anteil an der Entwicklung hin zur funktionalen 48nm TANOS-Zelle und haben mit Ihrem Engagement meine Doktorarbeit in der endgültigen Form erst möglich gemacht.

Raik Hoffmann und Rico Reichelt möchte ich für die Hilfe bei der Bearbeitung von Messaufgaben und dem Aufbau der Labview-Messumgebung, sowie den teils notwendigen komplexen Messungen mit Hilfe des Rifle-Messsystems danken.

Christoph Ludwig danke ich für seinen kritischen Blick auf die Ergebnisse und den daraus folgenden sehr hilfreichen Anregungen für die Bearbeitung meiner Doktorarbeit.

Karl-Heinz Küsters und Alexander Ruf, die stellvertretend für Qimonda Dresden, meine Arbeit gefördert haben, danke ich für Ihr Vertrauen und die Unterstützung, wodurch die Bearbeitung der Doktorarbeit stets problemlos fortgeführt werden konnte. Außerdem wurde mir durch die Zeit bei Qimonda Dresden die Möglichkeit gegeben, eine Vielzahl von Prozessen kennenzulernen, die für die Bearbeitung von Projekten in der Industrie unabdingbar sind.

Vielen Dank an die Kollegen aus der Zellgruppe, Stephan Riedel, Torsten Müller, Timm Höhr, Nigel Chan, Matthias Strassburg und Mark Isler für die sehr gute Zusammenarbeit und für das geduldige Beantworten der vielen Fragen.

Ulrike Bewersdorff, Armin Tilke und Roman Knoefler ist es zu verdanken, dass die skalierte TANOS-Zelle einen Stand erreicht hat, der weltweit mit führend ist. Sie haben auch die Umsetzung der vielen, zum Teil sehr aufwendig umzusetzenden Ideen erst möglich gemacht.

Ricardo, Ulf Kotarsky, Michael Laube, Jens Hassmann und allen Anderen, die bei der Qimonda Flash Integration mitgewirkt haben, danke ich für das herausragende Arbeitsklima und der in fast jeder Situation guten Stimmung im Büro.

Vielen Dank an die Mitarbeiter vom Fraunhofer CNT, besonders Malte Czernohorsky und Volkhard Beyer für die Unterstützung und angenehme Zusammenarbeit zum Ende meiner Promotionszeit.

Den Kollegen von der TU Braunschweig Prof. Meinerzhagen und Frau Kuligk möchte ich für die sehr hilfreiche Zusammenarbeit und angenehmen Diskussionen zu den Injektions-Simulationen danken.

Frau Oestreich, Markus Haverkamp, Jonas Schönlebe und Katja, sowie dem Frei-

berger Kollegium vom ESM habe ich eine sehr schöne Zeit an der Universität in Freiberg zu verdanken.

Und ein abschließender, besonderer Dank gilt meiner Freundin Susi und meinen Eltern Angelika und Harald, die mich die ganze Zeit unterstützt und immer zu mir gehalten haben.

Lebenslauf

Ausbildung

1987 - 1992	21.Grundschule „Anton Saefkow“, Dresden
1992 - 1999	Gymnasium Dresden-Gruna Abitur, allgemeine Hochschulreife
2000 - 2006	Technische Universität Dresden Studium an der Fakultät Elektrotechnik Fachrichtung Nachrichtentechnik mit Vertiefung Hoch/Höchstfrequenztechnik

beruflicher Werdegang

1997 - 1998	Institut für Makromolekulare Chemie der TU Dresden Jahresarbeit über Butylacrylat-Styrol-Dispersionen
September 2001	F&S Prozessautomation, Dohna Industriepraktikum
Apr. 2002 - Sep. 2004	Fraunhofer Institut für Integrierte Schaltungen, Dresden Hilfswissenschaftlicher Assistent
Okt. 2004 - März 2005	Diehl Avionik Systeme, Überlingen Industriepraktikum im EMV-Labor
Apr. 2005 - Juli 2005	Infineon Flash, Product Engineering, Dresden Industriepraktikum
Sep. 2005 - März 2006	Infineon Flash, Predevelopment, Dresden Diplomarbeit zu alternativem Flash-Speicherkonzept
Juni 2006 - März 2009	Qimonda Flash GmbH & Co. OHG, Predevelopment, Dresden Doktorand in der Speicherzellentwicklung
Apr. 2009 - Okt. 2009	Fraunhofer-Center Nanoelektronische Technologien (CNT), Dresden wiss. Mitarbeiter

Apr. 2009 - Dez. 2009	TU Bergakademie Freiberg Institut für Elektronik- und Sensormaterialien wiss. Mitarbeiter
Okt. 2009 -	Nanoelectronic Materials Laboratory (NaMLab), TU Dresden wiss. Mitarbeiter

Symbol- und Abkürzungsverzeichnis

Symbole

α	Winkel des Bereichs verringerter Inversion bei der Betrachtung lokaler Ladungsspeicherung
α_T	Speicherwahrscheinlichkeit für Elektronen
α_c	Control-Gate-Koppelfaktor
β	Transistor-Übertragungsleitwert
β_L	Längenverhältnis von Verarmungszone zu Inversionszone, bei inhomogen beladener Speicherzelle
ϵ_r	relative Dielektrizitätskonstante, 3.9 für SiO_2, 9.5 für Al_2O_3
ϵ_r	relative Dielektrizitätskonstante
η_s	Speichereffizienz
γ	Substratsteuerfaktor ($\gamma = \sqrt{2\epsilon_{ox} q N_A}/C'_{ox}$)
μ	Ladungsträgerbeweglichkeit
ϕ^e_{S-O}	Barrierenhöhe von Si-SiO_2, ≈3.1 eV
ϕ^h_{S-O}	Barrierenhöhe von Si-SiO_2, ≈3.8 eV
ϕ_{Ob}	Oberflächenpotential im Kanal des Transistors
σ^+_n	Einfangquerschnitt einer zweifach neg. geladenen Haftstelle
σ^0_n	Einfangquerschnitt einer einfach neg. geladenen Haftstelle
C'_{IPD}	normierte Kapazität zwischen Control- und Floating-Gate
C'_{Tun}	normierte Kapazität zwischen Substrat und Floating-Gate
C_{ONO}	Kapazität des MOS-Schichtstapels
C_{Sub}	Kapazität zwischen Inversionskanal und Substrat
D	elektrische Flussdichte
d_t	Ladungsträger-Tunneldistanz
E_L	Energieniveau des Leitungsbandes (eV)

E_V	Energieniveau des Valenzbandes (eV)
E_{ox}	elektrisches Feld im Oxid
f^+	Besetzungsdichte für eine positiv geladene Haftstelle
f^-	Besetzungsdichte für eine negativ geladene Haftstelle
g_m	Transkonduktanz eines MOS-Transistors
$I_{D,L}$	Transistor-Leckstrom
I_{sat}	Sättigungsstrom einer NAND-Reihe
J	Stromdichte
J_{BO}	Tunnelstromdichte durch das Tunneloxid
J_{TO}	Tunnelstromdichte durch das Topoxid
L	Länge des Transistors
$L^{'}$	wirksame Transistor-Kanallänge
L_{Ladung}	Länge des Ladungspaketes
$m^{'}$	effektive Masse; materialabhängig
m_e	Elektronenmasse; $9.1095 \cdot 10^{-31}$ kg
N_A	Dotierstoffkonzentration bei p-dotiertem Halbleiter
n_c	Ladungsträgerdichte im Leitungsband
N_B	Ladungsträgerdichte $(1/cm^3)$
N_T	Haftstellendichte
r_j	Diffusionsweite der Transistor-Kontaktgebiete
S	sub-V_T-Steigung der Transistor-Übertragungskennlinie, sub-V_T Swing
s	Programmiersteigung
V_B	Spannung des Substratanschlusses (Bulk)
V_D	Drainspannung
v_D	Sättigungs-Driftgeschwindigkeit von Elektronen
V_G	Gatespannung
V_S	Sourcespannung
V_T	Transistor-Schwellspannung

V_{BTB}	ausgewertete Spannung des GIDL-Stromes bei einem festen Stromkriterium
V_{CG}	Control-Gate Spannung
V_{FB}	Flachbandspannung
$V_{Fenster}$	Programmier-/ Löschfenster beim Zyklentest
V_{FG}	Floating-Gate Spannung
V_K	Spannung im Kanal der NAND-Reihe
V_{Pass}	Pass-Spannung
V_{Prog}	Programmier-Spannung
W	Weite des Transistors
x_V	Dicke der Verarmungszone bei der Betrachtung lokaler Ladungsspeicherung
x_{FN}	Injektionspunkt bei FN-Tunneln im Leitungsband des Oxids
ALD	Atomlagenabscheidung, engl. atomic layer deposition
At%	Atomprozent
BL	Bitleitung, Drain-Anschluss der Speicherzelle
CG	Control-Gate
CVD	chemische Gasphasenabscheidung, engl. chemical vapor deposition
DRAM	dynamischer Speicher mit freiem Lesezugriff (engl. dynamic random access memory)
DT	direktes Tunneln
EEPROM	elektrisch programmier- und löschbarer Nur-Lese-Speicher (engl. electrically erasble and programmable read-only memory)
EOT	equivalente Oxiddicke, bezieht die Dicke des Stapels auf die Dicke einer einfachen SiO_2-Schicht
EOT	äquivalente Oxiddicke, entspricht der elektrischen Dicke von SiO_2
EPROM	elektrisch programmierbarer Nur-Lese-Speicher (engl. electrically programmable read-only memory)
FG	Floating-Gate

FN	Fowler-Nordheim-Tunneln
GSL	Auswahltransistor auf der Sourceleitungsseite der NAND-Reihe
HBS	haftstellen-basierte Speicherzelle
HF	Hochfrequenz
IPD	interpoly-Dielektrikum
ISPP	inkrementelle Gatespannungs-Programmierung, engl. incremental step pulse programming
LHL	Löschen mit heißen Löchern
MFN	modifiziertes Fowler-Nordheim-Tunneln
MLC	multi-level Speicherzelle, drei bzw. mehrere programmierte Zustände
MNOS	Metall-Siliziumnitrid-Siliziumoxid-Siliziumsubstrat
MOS	Metall-Oxid-Silizium
MOSFET	Metall-Oxid-Silizium-Feldeffekt-Transistor
NAND	Speicherarchitektur, bei der die Speicherzellen in Reihe geschalten sind
NF	Niederfrequenz
NMOS	MOS-Transistor mit einem p-Substrat, die Leitung im Inversionskanal erfolgt durch Elektronen
NOR	Speicherarchitektur, bei der die Speicherzellen parallel geschalten sind
ONA	Schichtstapel bestehend aus SiO_2-Tunneloxid, Si_3N_4 und Al_2O_3-Topoxid
ONO	Schichtstapel bestehend aus SiO_2-Tunneloxid, Si_3N_4 und SiO_2-Topoxid
p-CVD	gepulste CVD
PGM	Programmieren
PMOS	MOS-Transistor mit einem n-Substrat, die Leitung im Inversionskanal erfolgt durch Löchern
PROM	programmierbarer Nur-Lese-Speicher (engl. programmable read-only memory)

PVD	physikalische Gasphasenabscheidung, engl. physical vapor deposition
RT	Raumtemperatur
SA	self-aligned; steht bei TANOS für eine Speicherschicht, die sowohl in Längen- als auch Weitenrichtung räumlich begrenzt ist
SEM	Rasterelektronenmikroskop; engl. Scanning Electron Microscope
SL	Sourceleitung
SLC	single-level Speicherzelle, ein programmierter Zustand
SRAM	statischen Speicher mit freiem Lesezugriff (engl. static random access memory)
SRAM	statischer, flüchtiger Speicher mit wahlfreiem Zugriff, engl. static random access memory
SSA	super-self-aligned; steht dafür, dass bei TANOS die Speicherschicht und das Topoxid sowohl in Längen- als auch Weitenrichtung räumlich begrenzt sind
SSL	Auswahltransistor auf der Bitleitungsseite der NAND-Reihe
STI	engl. shallow trench isolation, Grabenoxid-Isolation
TAIMATA	tert-amylimido-trisdim-ethylamidotantalum, Präkursor für TaN
TaN	Tantalnitrid
TBTEMT	tert-butylimido-tris-ethylmethylamido-tantalum, Präkursor für TaN
TEL	Tokyo Electron Limited, Hersteller von Anlagen für die Halbleiterfertigung
WL	Wortleitung, bezeichnet normalerweise ein Leiterbahn, die mit den Gateelektroden verbunden ist

Literaturverzeichnis

[1] A. Tilke, F. Beug, T. Melde, and R. Knoefler, "Speichern ohne fluchtgefahr," *Physik-Journal*, , no. 4, pp. 33–38, April 2009.

[2] R. Zeng, N. Chalagalla, D. Chu, D. Elmhurst, M. Goldman andC. Haid, A. Huq, T. Ichikawa, J. Jorgensen, O. Jungroth andN. Kajla, and R. Kajley, "A 172mm2 32gb mlc nand flash memory in 34nm cmos," *in Proc. of the Int. Sol.-State Circuits Conf.*, pp. 236–237, 2009.

[3] K. Kim and J. Choi, "Future outlook of nand flash technology for 40nm node and beyond," *in Proceedings of the NVSM Workshop*, pp. 9–11, 2006.

[4] Kirk Prall, "Scaling non-volatile memory below 30 nm," *in Proceedings of the NVSM Workshop*, pp. 5–10, August 2007.

[5] N. Chan, M. F. Beug, R. Knoefler, T. Mueller, T. Melde, M.Ackermann, S. Riedel, M. Specht, C. Ludwig, and A. T. Tilke, "Metal control gate for sub-30nm floating gate nand memory," *in Proc. of the Non-Vol. Mem. Tech. Symp.*, pp. 82–85, 2008.

[6] Y. Fukuzumi, Y. Matsuoka, M. Kito, M. Kido, M. Sato, H. Tanakaand Y. Nagata, Y. Iwata, H. Aochi, and A. Nitayama, "Optimal integration and characteristics of vertical array devices for ultra-highdensity, bit-cost scalable flash memory," *in IEDM Tech. Dig.*, pp. 449–452, 2007.

[7] J. Kim, A. J. Hong, M. Ogawa, S. Ma, E.B. Song, Y.-S. Linand J. Han, U-In Chung, and K. L. Wang, "Novel 3-d structure for ultra high density flash memory with vrat (vertical-recess-array-transistor)and pipe (planarized integration on the same plane)," *in Proceedings of the Symp. on VLSI Tech.*, pp. 122–123, June 2008.

[8] E. H. Nicollian andJ. R. Brews, *MOS (Metal Oxide Semiconductor) Physics and Technology*, John Wiley & Sons, wiley classic library edition, 2003.

[9] A. I. Kingon, J. P . Maria, and S. K. Streiffer, "Alternative dielectrics to silicon dioxide for memory and logic devices," *Nature*, vol. 406, no. 6799, pp. 1032–1038, 2000.

[10] E. Gerritsen, N. Emonet, C. Caillat, N. Jourdan, M. Piazza andD. Fraboulet, B. Boeck, A. Berthelot, S. Smith, and P. Mazoyer, "Evolution of materials technology for stacked-capacitors in 65 nm embedded-dram," *Solid-State Electron.*, vol. 49, pp. 1767–1775, 2005.

[11] M. F. Beug, *Charakterisierung von EEPROM Tunneloxiden mittels transienter Strom- undKapazitätsmessungen*, Ph.D. thesis, Universität Hannover, 2004.

[12] S. M. Sze and Kwok K. Ng, *Physics of semiconductor devices*, John Wiley & Sons, 3rd edition, 2007.

[13] Joachim Goerth, *Baulemente und Grundschaltungen*, B.G.Teubner, Stuttgart, Leipzig, 1999.

[14] T. Yamamoto, T. Kubo, T. Sukegawa, E. Takii, Y. Shimamune andN. Tamura, T. Sakoda, M. Nakamura, H. Ohta, T. Miyashita andH. Kurata, S. Satoh, M. Kase, and T. Sugii, "Junction profile engineering with a novel multiple laser spike annealingscheme for 45-nm node high performance and low leakage cmos technology," *in IEDM Tech. Dig.*, pp. 143–146, 2007.

[15] Y. Polansky, A. Lavan, R. Sahar, O. Dadashev, Y. Betser andG. Cohen, E. Maayan, and B. Eitan et. al., "A 4b/cell nrom 1gb data-storage memory," *in Proc. of the Int. Sol.-State Circuits Conf.*, pp. 132–133 ; 644, 2006.

[16] L. D. Yau, "A simple theory to predict the threshold voltage of short-channel igfet's," *Solid-State Electron.*, vol. 17, pp. 1059–1063, 1974.

[17] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer, and D. Finzi, "Nrom: A novel localized trapping, 2-bit nonvolatile memory cell," *IEEE Electron Device Lett.*, vol. 21, no. 11, pp. 543–546, 2000.

[18] C. Zener, "A theory of the electrical breakdown of solid dielectrics," *Proc. R. Soc. Lond. A*, vol. 145, pp. 523–529, July 1934.

[19] C. Trinh, N. Shibata, T. Nakano, M. Ogawa, J. Sato, Y. Takeyamaand K. Isobe, B. Le, F. Moogat, N. Mokhlesi, K. Kozakai andP. Hong, T. Kamei, K. Iwasa, and J. Nakai et. al., "A 5.6mb/s 64gb 4b/cell nand flash memory in 43nm cmos," *in Proc. of the Int. Sol.-State Circuits Conf.*, pp. 246–248, 2009.

[20] M. Lenzlinger and E. H. Snow, "Fowler-nordheim-tunneling into thermally grown sio2," *J. Appl. Phys.*, vol. 40, no. 1, pp. 278–283, 1968.

[21] Z. A. Weinberg, "On tunneling in metal-oxide-silicon structures," *J. Appl. Phys.*, vol. 7, no. 53, pp. 5052–5056, 1982.

[22] K. F. Schuegraf, C. C. King, and C. Hu, "Ultra-thin silicon dioxide leakage current and scaling limit," *in Proceedings of the Symp. on VLSI Tech.*, pp. 18–19, 1992.

[23] M. Depas, B. Vermeire, P. W. Mertens, and R. L. von Meirhaeghe andM. M. Heyns, "Determination of tunneling parameters in ultrathin oxide layer poly-si/sio_2/sistructures," *Solid-State Electron.*, vol. 38, no. 8, pp. 1465–1471, 1995.

[24] M. M. E. Beguwala and T. L. Gunckel, "An improved model for the charging characteristics of a dual-dielectric(mnos) nonvolatile memory device," *IEEE Trans. Electron Devices*, vol. ED-25, no. 8, 1978.

[25] P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, *Flash memories*, Springer, 2nd edition, 1999.

[26] W. D. Brown and J. E. Brewer, *Nonvolatile Semiconductor Memory Technology, A Comprehensive Guide to Understandingand Using NVSM Devices*, IEEE Press, 345 East 47th Street, New York, 1998.

[27] C. E. Blat, E. H. Nicollian, and E. H. Poindexter, "Mechanism of negative-bias-temperature instability," *J. Appl. Phys.*, vol. 69, no. 3, pp. 1715–1720, 1991.

[28] D. Frohman-Bentchkowsky and M. Lenzlinger, "Charge transport and storage in metal-nitride-oxide-silicon (mnos) structures," *J. Appl. Phys.*, vol. 40, no. 8, pp. 3307–3319, 1969.

[29] S.-I. Minami and Y. Kamigaki, "New scaling guidelines for mons nonvolatile memory devices," *IEEE Electron Device Lett.*, vol. 38, no. 11, pp. 2519–2526, November 1991.

[30] P. C. Y. Chen, "Threshold-alterable si-gate mos devices," *IEEE Trans. Electron Devices*, vol. 24, no. 5, pp. 584–586, 1977.

[31] H. Aozasa, I. Fujiwara, A. Nakamura, and Y. Komatsu, "Analysis of carrier traps in si_3n_4 in oxide/nitride/oxide for metal/oxide/nitride/oxide/siliconnonvolatile memory," *Jap. J. Appl. Phys.*, vol. 38, no. 3A, pp. 1441–1447, 1999.

[32] V. I. Belyi and A. A. Rastorguyev, "A new view on the nature of electron levels in amorphous silicon nitride," *journal of chem. for sustain. devel.*, vol. 8, no. 1-2, pp. 13–20, 2000.

[33] M. Fliesler, D. Still, and J.-M. Hwang, "A 15ns 4mb nvsram in 0.13μ sonos technology," *in Proceedings of the NVSM Workshop*, pp. 83–85, 2008.

[34] F. R. Libsch and M. H. White, "Charge transport and storage of low programming voltage sonos/monos memorydevices," *Solid-State Electron.*, vol. 33, no. 1, pp. 105–126, 1990.

[35] H.Bachhofer, H.Reisinger, E. Bertagnolli, and H. von Philipsborn, "Transient conduction in multidielectric silicon-oxide-nitride-oxide-semiconductorstructures," *J. Appl. Phys.*, vol. 89, no. 5, march 2001.

[36] S. Jeon, J. H. Han, J. Lee, S. Choi, H. Hwang, and C. Kim, "Impact of metal work function on memory properties of charge-trap flashmemory devices using fowler-nordheim p/e mode," *IEEE Electron Device Lett.*, vol. 27, no. 6, pp. 486–488, Juni 2006.

[37] K. Choi, H. N. Alshareef, H. C. Wen, H. Harris, H. Luan andY. Senzaki, P. Lysaght, P. Majhi, and B. H. Lee, "Effective work function modification of atomic-layer-deposited-tan filmby capping layer," *Appl. Phys. Lett.*, vol. 89, no. 032113, 2006.

[38] Y. Sasago, H. Kurata, T. Arigane, K. Otsuga, T. Kobayashi andY. Ikeda, T. Fukumura, S. Narumi, A. Sato, T. Terauchi andM. Shimizu, O. Tsuchiya, and K. Furusawa, "90-nm-node multi-level ag-and type flash memory with cell size

of true $2f^2$/bit and programming throughput of 10 mb/s," *in IEDM Tech. Dig.*, pp. 823–826, 2003.

[39] E. Maayan, R. Dvir, J. Shor, Y. Polansky, Y. Sofer, I. Bloomand D. Avni, B. Eitan, Z. Cohen, M. Meyassed, Y. Alpern andH. Palm, E.S. v Kamienski, P. Haibach, D. Caspary, and S. Riedeland R. Knoefler, "A 512 mb nrom flash data storage memory with 8 mb/s data rate," *in Proc. of the Int. Sol.-State Circuits Conf.*, vol. 1, pp. 100–101, 2002.

[40] G.Servalli, D.Brazzelli, E.Camerlenghi, G.Capetti, S.Costantiniand C.Cupeta, D.DeSimone, A.Ghetti, T.Ghilardi, P.Gulli andM.Mariani, A.Pavan, and R.Somaschini, "A 65nm nor flash technology with $0.042\mu m^2$ cell size for high performancemultilevel application," *in IEDM Tech. Dig.*, pp. 849–852, 2005.

[41] S. Yamada, T. Yamane, K. Amemiya, and K. Naruke, "A self-convergence erase for nor flash eeprom using avalanche hot carrierinjection," *IEEE Trans. Electron Devices*, vol. 43, no. 11, pp. 1937–1941, 1996.

[42] F. Masuoka, M. Momodomi, Y. Iwata, and R. Shirota, "New ultra high density eprom and flash eeprom with nand structure cell," *in IEDM Tech. Dig.*, pp. 552–555, 1987.

[43] K. Kanda, M. Koyanagi, T. Yamamura, K. Hosono, M. Yoshiharaand T. Miwa, Y. Kato, A. Mak, S. L. Chan, F. Tsai, and R. Cerneaand et. al., "A 120 mm^2 16gb 4-mlc nand flash memory with 43nm cmos technology," *in Proc. of the Int. Sol.-State Circuits Conf.*, pp. 430–431, 625, 2008.

[44] D. Nobunaga, E. Abedifard, F. Roohparvar, J. Lee, E. Yu andA. Vahidimowlavi, and M. Abraham et. al., "A 50nm 8gb nand flash memory with 100mb/s program throughput and 200mb/sddr interface," *in Proc. of the Int. Sol.-State Circuits Conf.*, pp. 426–427, 625, 2008.

[45] J.-K. Kim, K. Sakui, S.-S. Lee, Y. Itoh, S.-C. Kwon, K. Kanazawaand K.-J. Lee, H. Nakamura, K.-Y. Kim, and T. Himeno et. al., "A 120-mm ^{2} 64-mb nand flash memory achieving 180 ns/byte effective programspeed," *IEEE J. Solid-State Circuits*, vol. 32, no. 5, pp. 670–680, 1997.

[46] J. E. Brewer and M. Gill, *Nonvolatile Memory Technologies with Emphasis on Flash*, J. Wiley and Sons, Hoboken, New Jersey, 2008.

[47] B.-H. Suh K.-D. Suh, Y.-H. Lim, J.-K. Kim, Y.-J. Choi, and Y.-N.Koh et. al., "A 3.3v 32mb nand flash memory with incremental step pulse programming scheme," *in Proc. of the Int. Sol.-State Circuits Conf.*, pp. 128–129,350, 1995.

[48] H.-T. Lue, T.-H. Hsu, S.-Y. Wang, E.-K. Lai, K.-Y. Hsieh andR. Liu, and C.-Y. Lu, "study of incremental step pulse programming (ispp) and sti edge effect ofbe-sonos nand flash," *in Proceedings of the IRPS*, pp. 693–694, 2008.

[49] R. Duane, M. F. Beug, and A. Mathewson, "Novel capacitance coupling coefficient measurement methodology for floatinggate nonvolatile memory devices," *IEEE Electron Device Lett.*, vol. 26, no. 7, 2005.

[50] B. H. Yun, "Direct display of electron back tunneling in mnos memory capacitors," *Appl. Phys. Lett.*, vol. 23, no. 3, pp. 152–153, 1973.

[51] P. J. Mc Worther, S. L. Miller, and T. A. Dellin, "Modelling the memory retention characteristics of silicon-nitride-oxide-siliconnonvolatile transistors in a varying thermal environment," *J. Appl. Phys.*, vol. 68, no. 4, pp. 1902–1909, 1990.

[52] T.-H. Hsu, H.-T. Lue, E.-K. Lai, J.-Y. Hsieh, and S.-Y. Wang andL.-W. Yang et. al., "A high-speed be-sonos nand flash utilizing the field-enhancement effectof finfet," *in IEDM Tech. Dig.*, pp. 913–916, 2007.

[53] K. Sonoda, K. Ishikawa, T. Eimori, and O. Tsuchiya, "Discrete dopant effects on statistical variation of random telegraph signalmagnitude," *IEEE Trans. Electron Devices*, vol. 54, no. 8, pp. 1918–1925, 2007.

[54] P. Fantini, A. Ghetti, A. Marinoni, G. Ghidini, and A. Viscontiand A. Marmiroli, "Giant random telegraph signals in nanoscale floating-gate devices," *IEEE Electron Device Lett.*, vol. 28, no. 12, pp. 1114–1116, 2007.

[55] M. Janai, B. Eitan, A. Shappir, E. Lusky, I. Bloom, and G. Cohen, "Data retention reliability model of nrom nonvolatile memory products," *IEEE Trans. Device Mater. Rel.*, vol. 4, no. 3, pp. 404–415, 2004.

[56] Y. Nissan-Cohen, J. Shappir, and D.Frohmann-Bentchkowsky, "Characterization of simultaneous bulk and interface high-field trappingeffects in sio2," *in IEDM Tech. Dig.*, pp. 182–185, 1983.

[57] J. D. Lee, J. H. Choi, D. Park, and K. Kim, "Data retention characteristics of sub-100 nm nand flash memory cells," *IEEE Electron Device Lett.*, vol. 24, no. 12, 2003.

[58] J. T. Wallmark and J.H. Scott, "Switching and storage characteristics of mis memory transistors," *RCA review*, vol. 30, pp. 335–365, 1969.

[59] V. A. Gritsenko, Hei Wong, J. B. Xu, R. M. Kwok, I. P. Petrenkoand B.. A. Zaitsev, Y. N. Morokov, and Y. N. Novikov, "Excess silicon at the silicon nitride/thermal oxide interface in oxide-nitride-oxidestructures," *J. Appl. Phys.*, vol. 86, no. 6, pp. 3234–3240, 1999.

[60] Y. Kamigaki, S. i. Minami, and H. Kato, "A new portrayal of electron and hole traps in amorphous silicon nitride," *J. Appl. Phys.*, vol. 68, no. 6, pp. 2211–2215, 1990.

[61] J. Robertson, "The electronic properties of silicon nitride," *philos. magazine B*, vol. 44, no. 2, pp. 215–237, 1981.

[62] J. Robertson and M. J. Powell, "Gap states in silicon nitride," *Appl. Phys. Lett.*, vol. 44, no. 4, pp. 415–417, 1983.

[63] E. Lusky, Y. Shacham-Diamand, A. Shappir, I. Bloom, and B. Eitan, "Traps spectroscopy of the si3ni4 layer using localized charge-trapping nonvolatilememory device," *Appl. Phys. Lett.*, vol. 85, no. 4, pp. 669–671, 2004.

[64] H. J. Stein and H. A. R. Wegener, "Chemically bound hidrogen in cvd si3n4: Dependence on nh3/sih4 ratio andon annealing," *journal on sol.-state sci. and tech.*, vol. 124, no. 6, pp. 908–912, 1977.

[65] V. J. Kapoor and S. B. Bibyk, "Energy distribution of electron trapping defects in thick-oxide mnos structures," *in proc. of Conf. on The Physics of MOS insul.*, pp. 117– 121, 1980.

[66] G. van den Bosch, A. Furnemont, M. B. Zahid, R. Degraeve, L.Breuil, A. Cacciato, A. Rothschild, C. Olsen, and U. Ganguly andJ. van Houdt, "Nitride engineering for improved erase performance and retention of tanosnand flash memory," *in Proceedings of the NVSM Workshop*, pp. 128–129, 2008.

[67] S. J. Wrazien, Y. Zhao, J. D. Krayer, and M. H. White, "Characterization of sonos oxynitride nonvolatile semiconductor memory devices," *Solid-State Electron.*, vol. 47, pp. 885–891, 2003.

[68] T. H. Kim, I. H. Park, J. D. Lee, H. C. Shin, and B.-G. Park, "Electron trap densitiy distribution of si-rich silicon nitride extractedusing the modified negative charge decay model of silicon-oxide-nitride-oxide-siliconstructure at elevated temperatures," *Appl. Phys. Lett.*, vol. 89, no. 063508, 2006.

[69] C. Kang, J. Choi, J. Sim, C. Lee, Y. Shin, J. Park, J.Sel, S. Jeon, Y. Park, and K. Kim, "Effects of lateral charge spreading on the reliability of tanos (tan/alo/sin/oxide/si)nand flash memory," *in Proceedings of the IRPS*, pp. 167–170, 2007.

[70] J. S. Sim, J. Park, C. Kang, W. Jung, Y. Shin, J. Kim andJ. Sel, C. Lee, S. Jeon, Y. Jeong, Y. Park, and J. Choi andW. S. Lee, "Self aligned trap-shallow trench isolation scheme for the reliability oftanos (tan/alo/sin/oxide/si) nand flash memory," *in Proceedings of the NVSM Workshop*, pp. 110–111, 2007.

[71] C. Sandhya, U. Ganguly, N. Chattar, C. Olsen, S.M. Seutter andL. Date, R. Hung J.M. Vasi, and S. Mahapatra, "Effect of sin on performance and reliability of charge trap flash (ctf)under fowler-nordheim tunneling program/erase operation," *IEEE Electron Device Lett.*, vol. 30, no. 2, pp. 171–173, 2009.

[72] T. Melde, M. F. Beug, N. Chan, L. Bach, S. Riedel, and C. Ludwigand T. Mikolajick, "Accurate program simulation of tanos charge trapping devices," *in Proc. of the Non-Vol. Mem. Tech. Symp.*, pp. 90–94, 2008.

[73] S. Imanaga and H. Aozasa, "Modeling of nonvolatile memory operation of polysilicon-oxide-nitride-oxide-semiconductorand analysis of program characteristics dependent on the trap distributionin the silicon-nitride layer," *Jap. J. Appl. Phys.*, vol. 43, no. 8A, pp. 5186–5198, 2004.

[74] A. Furnémont, M. Rosmeulen, A. Cacciato, L. Breuil, K. De Meyerand H. Maes, and J. Van Houdt, "A consistent model for the sanos programming operation," *in Proc. of the Non-Vol. Mem. Tech. Symp.*, pp. 96–97, 2007.

[75] T. Melde, M. F. Beug, L. Bach, S.Riedel, C. Ludwig, and T. Mikolajick, "Nitride thickness scaling limitations in tanos charge trapping devices," *in Proc. of the NVSMW / ICMTD 2008*, pp. 130–132, 2008.

[76] C. Svensson and I. Lundström, "Trap-assisted charge injection in mnos structures," *J. Appl. Phys.*, vol. 44, no. 10, pp. 4657–4663, 1973.

[77] R. Degraeve, M. Cho, B. Govoreanu, B. Kaczer, M.B. Zahid, J. Van Houdt, M. Jurczak, and G. Groeseneken, "Trap spectroscopy by charge injection and sensing (tscis): a quantitative electrical technique for studying defects in dielectric stacks," *in IEDM Tech. Dig.*, pp. 1–4, December 2008.

[78] M. Specht, M. Städele, S. Jakschik, and U. Schröder, "Transport mechanisms in atomic-layer-deposited al2o3 dielectrics," *Appl. Phys. Lett.*, vol. 84, no. 16, pp. 3076–3078, april 2004.

[79] D. K. Schroder and J. A. Babcock, "negative bias temperature instability: Road to cross in deep submicron siliconsemiconductor manufacturing," *J. Appl. Phys.*, vol. 94, no. 1, pp. 1–18, 2003.

[80] S. Choi, M. Cho, H. Hwang, and J. W. Kim, "Improved metal-oxide-nitride-oxide-silicon-type flash device with high-k dielectrics for blocking layer," *J. Appl. Phys.*, vol. 94, no. 8, pp. 5408–5410, 2003.

[81] J. D. Lee, J. H. Choi, D. Park, and K. Kim, "Degradation of tunnel oxide by fn current stress and its effects on dataretention characteristics of 90-nm nand flash memory cells," *in Proceedings of the IRPS*, pp. 497–501, 2003.

[82] N. Nagel, T. Muller, M. Isler, V. Pissors, J.-U. Sachse andD. Manger, D. Caspary, S. Parascandola, and D. Olligs et.al., "A new twin flash cell for 2 and 4 bit operation at 63nm feature size," *in Proc. of Int. Symp. on VLSI-TSA*, pp. 1–2, 2007.

[83] H. T. Lue, T. H. Hsu, S. C. Lai, Y. H. Hsiao, W. C. Peng andC. W. Liao, Y. F. Huang, S. P. Hong, M. T. Wu, F. H. Hsu andN. Z. Lien, S. Y. Wang, L.W. Yang, T. Yang, K.C. Chen, K.Y.Hsieh, Rich Liu, and Chih-Yuan. Lu, "Scaling evaluation of be-sonos nand flash beyond 20 nm," *in Proceedings of the Symp. on VLSI Tech.*, pp. 116–117, 2008.

[84] E. Lusky, Y. Shacham-Diamand, I. Bloom, and B. Eitan, "Characterization of channel hot electron injection by the subthreshold slopeof nrom device," *IEEE Electron Device Lett.*, vol. 22, no. 11, pp. 556–558, November 2001.

[85] L. Larcher, G. Verzellesi, P. Pavan, E. Lusky, and i. Bloom andB. Eitan, "Impact of programming charge distribution on threshold voltage and subthresholdslope of nrom memory cells," *IEEE Electron Device Lett.*, vol. 49, no. 11, pp. 1939–1946, november 2002.

[86] Oliver Klar, *Charakterisierung und Modellierung von Ladungseinfangmechanismen in dielektrischenSpeicherschichten*, Ph.D. thesis, Universität Erlangen-Nürnberg, 2008.

[87] A. Shappir, Y. Shacham-Diamand, E. Lusky, I. Bloom, and B. Eitan, "Subthreshold slope degradation model for localized-charge-trapping basednon-volatile memory devicesan analytical model is presented for the subthresholdslope degradation of localized-charge-trapping based nonvolatile," *Solid-State Electron.*, vol. 47, pp. 937–941, Oktober 2003.

[88] A. Crunteanu, P. Hoffmann, M. Pollnau, and Ch. Buchal, "Comparative study on methods to structrure sapphire," *Applied Surface Science*, vol. 208/209, pp. 322–326, 2003.

[89] X. Dongzhu, Z. Dezhang, P. Haochang, X. Hongjie, and R. Zongxin, "Enhanced etching of sapphire damaged by ion implantation," *Journal of Physics D: Applied Physics*, vol. 31, no. 14, pp. 1647–1651, 1998.

[90] S.I. Dolgaev, A.A. Lyalin, A.V. Simakin, and G.A. Shafeev, "Fast etching of sapphire by a visible range quasi-cw laser radiation," *applied surface science*, vol. 96-98, pp. 491–495, april 1996.

[91] D.W. Kim, C.H. Jeong, K.N. Kim, H.Y. Lee, H.S. Kim, Y.J.Sung, and G.Y. Yeom, "High rate sapphire (al2o3) etching in inductively coupled plasmas usingaxial external magnetic field," *Thin Solid Films*, vol. 435, pp. 242–246, 2003.

[92] S.-M. Koo, D.-P. Kim, K.-T. Kim, and C.-I. Kim, "The etching properties of al2o3 thin films in n2/cl2/bcl3 and ar/cl2/bcl3gas chemistry," *Materials Science and Engineering B*, vol. 118, no. 1-3, pp. 201–204, april 2005.

[93] Y.J. Sung, H.S. Kim, Y.H. Lee, J.W Lee, S.H. Chae, and Y.J. Parkand G.Y. Yeom, "High rate etching of sapphire wafer using cl2/bcl3/ar inductively coupledplasmas," *Mater. Sci. Eng. B*, vol. 82, pp. 50–51, 2001.

[94] Sun Jin Yun, Alexander Efremov, Mansu Kim, Dae-Won Kim, JungWook Lim, Yong-Hae Kim, Choong-Heui Chung, Dong Jin Park, and Kwang-HoKwon, "Etching characteristics of al2o3 thin films in inductively coupled bcl3/arplasma," *Vacuum*, vol. 82, no. 11, pp. 1198–1202, June 2008.

[95] M. F. Beug, T. Melde, M. Isler, L. Bach, M. Ackermann, S. Riedeland K. Knobloch, and C. Ludwig, "Anomalous erase behaviour in charge trapping memory cells," *in Proceedings of the NVSM Workshop*, pp. 121–123, may 2008.

[96] E. Lusky, Y. Shacham-Diamand, G. Mitenberg, A. Shappir, and I. Bloomand B. Eitan, "Investigation of channel hot electron injection by localized charge-trappingnonvolatile memory devices," *IEEE Trans. Electron Devices*, vol. 51, no. 3, pp. 444–451, 2004.

[97] J. H. Hwang and X. Chen, "Plasma heating of a substrate with subsequent high temperature etching," patent, März 2004.

[98] R. Wise, W. Yan, Y. Zhang, N. Gani, N. Sun, M. Shen, and T.Lill, "High-k etch performance for next-generation logic gate stacks," *solid state technology*, Dezember 2008.

[99] M. Hélot, T. Chevolleau, L. Vallier, O. Joubert, E. Blanquetand A. Prisch, P. Mangiagalli, and T. Lill, "Plasma etching of hfo2 at elevated temperatures in chlorine-based chemistry," *J. Vac. Sci. Technol.*, vol. 24, no. 1, pp. 30–40, Januar 2006.

[100] K. Pelhos, V. M. Donnelly, A. Kornblit, M. L. Green, R. B. Van Dover, L. Manchanda, Y. Hu, M. Morris, and E. Bower, "Etching of high- kdielectric zr(1-x)/al(x)/o(y) films in chlorine-containingplasmas," *J. Vac. Sci. Technol.*, vol. 19, pp. 1361–1366, Juli 2001.

[101] T. Banjo, M. Tsuchihashi, M. Hanazaki, M. Tuda, and K. Ono, "Effects of o2 addition on bcl3/cl2 plasma chemistry for al etching," *Jap. J. Appl. Phys.*, vol. 36, no. 7B, pp. 4824–4828, Juli 1997.

[102] T. Maeda, H. Ito, R. Mitsuhashi, A. Horiuchi, T. Kawahara andA. Muto, T. Sasaki, K. Torii, and H. Kitajima, "Selective dry etching of hfo2 in cf4 and cl2/hbr-based chemistries," *Jap. J. Appl. Phys.*, vol. 43, no. 4B, pp. 1864–1868, Apirl 2004.

[103] Merck Schuchardt OHG, *Sicherheitsdatenblatt*, http://www.merck-chemicals.com/documents/sds/emd/deu/de/8010/801068.pdf, 2006.

[104] R. T. Brewer, M.-T. Ho, K. Z. Zhang, L. V. Goncharova, D. G.Starodub, T. Gustafsson, and Y. J. Chabal, "Ammonia pretreatment for high- k dielectric growth on silicon," *Appl. Phys. Lett.*, vol. 85, no. 17, pp. 3830–3832, october 2004.

[105] M. M. Frank, Y. J. Chabal, and G. D. Wilk, "Nucleation and interface formation mechanisms in atomic layer depositionof gate oxides," *Appl. Phys. Lett.*, vol. 82, no. 26, pp. 4758–4759, june 2003.

[106] Y. Chang, F. Ducroquet, E. Gautier, O. Renault, J. Legrand andJ.F. Damlencourt, and F. Martin, "Surface preparation and post thermal treatment effects on interface propertiesof thin al2o3 films deposited by ald," *Insulating Films on Semiconductors*, , no. 13, 2003.

[107] J. K. Schaeffer, S. B. Samavedam, D. C. Gilmer, V. Dhandapani andP. J. Tobin, J. Moga, B.-Y. Nguyen, B. E. White, S. Dakshina-Murthyand R. S. Rai, Z.-X. Jiang, R. Martin, M. V. Raymond, M. Zavalaand L. B. La, J. A. Smith, R. Garcia, D. Roan, and M. Kottke andR. B. Gregory, "Physical and electrical properties of metal gate electrodes on hfo2 gatedielectrics," *J. Vac. Sci. Technol. B*, vol. 21, pp. 11–17, 2003.

[108] J. B. Kim, D. R. Kwon, K. Chakrabarti, Chongmu Lee, and K. Y. Ohand J. H. Lee, "Improvement in al2o3 dielectric behavior by using ozone as an oxidant forthe atomic layer deposition technique," *J. Appl. Phys.*, vol. 92, no. 11, pp. 6739–6742, Dezember 2002.

[109] S. Jakschik, U. Schroeder, T. Hecht, D. Krueger, G. Dollingerand A. Bergmaier, C. Luhmann, and J. W. Bartha, "Physical characterization of thin ald-al2o3 films," *Appl. Surf. Sci.*, vol. 211, pp. 352–359, Februar 2003.

[110] S. Jakschik, U. Schroeder, T. Hecht, G. Dollinger, and A. Bergmaierand J.W. Bartha, "Physical properties of ald-al2o3 in a dram-capacitor equivalent structurecomparing interfaces and oxygen precursors," *Mat. Sci. and Eng. B*, , no. 107, pp. 251–254, 2004.

[111] J. Lützen, A. Birner, M. Goldbach, M. Gutsche, T. Hecht andS. Jakschik, A. Orth, A. Sänger, U. Schröder, H. Seidl andB. Sell, and D. Schumann, "Integration of capacitor for sub-100-nm dram trench technology," *in Proceedings of the Symp. on VLSI Tech.*, pp. 178–179, 2002.

[112] S. Jakschik, U. Schroeder, T. Hecht, M. Gutsche, and H. Seidl andJ. W.Bartha, "Crystallization behavior of thin ald-al2o3 films," *Thin Solid Films*, vol. 425, no. 1-2, pp. 216–220, Februar 2003.

[113] V. V. Afanas'ev, A. Stesmans, B. J. Mrstik, and C. Zhao, "Impact of annealing-induced compaction on electronic properties of atomic-layer-depositedal2o3," *Appl. Phys. Lett.*, vol. 81, no. 9, pp. 1678–1680, August 2002.

[114] M.Lanza, M.Porti, M.Nafria, X.Aymerich, G.Benstetter, E.Lodermeier, H.Ranzinger, G.Jaschke, S.Teichert, L.Wilde, and P.Michalowski, "Crystallization and silicon diffusion nanoscale effects on the electrical properties of al2o3 based devices," *conf. of Insul. Films on Semicon.*, , no. 16, Juli 2009.

[115] A. Cacciato, A. Furnémont, L. Breuil, J. De Vos, L. Haspeslagh, and J. Van Houdt, "Effect of al2o3 morphology on the erase saturation performance in sanos-typememory cells," *in Proc. of the Int. Conf. on Memory Tech. and Design*, pp. 217–220, 2007.

[116] J. H. Lee, K. Koh, N. I. Lee, M. H. Cho, Y. K. Kim, J. S.Jeon, K. H. Cho, H. S. Shin, M. H. Kim, K. Fujihara, H.K. Kang, and J. T. Moon, "Effect of polysilicon gate on the flatband voltage shift and mobility degradationfor ald-al203 gate dielectric," *in IEDM Tech. Dig.*, 2000.

[117] S. Jeon and C. Kim, "The effect of fixed oxide charge in al2o3 blocking dielectric on memoryproperties of charge trap flash memory devices," *Electrochem. Solid-State Lett.*, vol. 9, no. 8, pp. 265–267, 2006.

[118] B. Chen, "Highly reliable superflash embedded memory scaling for low power soc," *VLSI-TSA*, pp. 108–109, april 2007.

[119] K.K. Strelov and I.D. Kashcheev, "Phase diagram of the system al2o3 - sio2," *Refractories and Industrial Ceramics*, vol. 36, no. 8, pp. 244–246, august 1995.

[120] E. P. Gusev, M. Copel, E. Cartier, I. J. R. Baumvol, and C. Krugand M. A. Gribelyuk, "High-resolution depth profiling in ultrathin al2o3 films on si," *Appl. Phys. Lett.*, vol. 76, no. 2, pp. 176–178, Januar 2000.

[121] K. Torii, Y. Shimamoto, S. Saito, O. Tonomura, M. Hiratani andY. Manabe, M. Caymax, and J. W. Maes, "The mechanism of mobility degradation in misfets with al2o3 gate dielectric," *in Proceedings of the Symp. on VLSI Tech.*, 2002.

[122] C. C. Hobbs, L. R. C. Fonseca, A. Knizhnik, V. Dhandapani, S.B. Samavedam, W. J. Taylor, J. M. Grant, L. G. Dip, D. H. Triyosoand R. I. Hegde, D. C. Gilmer, R. Garcia, and D. Roan et. al., "Fermi-level pinning at the polysilicon/metal-oxide interface - part ii," *IEEE Trans. Electron Devices*, vol. 51, no. 6, pp. 978–984, Juni 2004.

[123] G. Pourtois, A. Lauwers, J. Kittl, L. Pantisano, B. Sorée andS. De Gendt, W. Magnus, M. Heyns, and K. Maex, "First-principle calculations on gate/dielectric interfaces: on the originof work function shifts," *Insulating Films on Semiconductors*, , no. 14, pp. 272–279, 2005.

[124] H. Y. Yu, Chi Ren, Yee-Chia Yeo, J. F. Kang, X. P. Wang andH. H. H. Ma, Ming-Fu Li, D. S. H. Chan, and D.-L. Kwong, "Fermi pinning-induced thermal instability of metal-gate work functions," *IEEE Electron Device Lett.*, vol. 25, no. 5, Mai 2004.

[125] G. Van den Bosch, L. Breuil, A. Cacciato, A. Rothschild, M. Jurczak, and J. Van Houdt, "Investigation of window instability in program/erase cycling of tanos nand flash memory," *in Proc. of the Int. Mem. Workshop*, pp. 81–82, May 2009.

[126] L. Breuil, A. Furnemont, A. Rothschild, G. van den Bosch, A.Cacciato, and J. van Houdt, "Improvement of tanos nand flash performance by the optimization of a sealinglayer," *in Proceedings of the NVSM Workshop*, pp. 126–127, 2008.

[127] M. Chang, Y. Ju, J. Lee, S. Jung, H. Choi, M. Jo, and S. Jeonand H. Hwang, "Impact of oxygen incorporation at the si_3n_4/al_2o_3 interfaceon retention characteristics for nonvolatile memory applications," *Appl. Phys. Lett.*, vol. 93, pp. 022101, 2008.

[128] S. H. Lai, H. T. Lue, C. W. Liao, Y. F. Huang, M. J. Yang andY. H. Lue, T. B. Wu, and J. Y. Hsieh et. al., "An oxide-buffered be-manos charge-trapping device and the role of al_2o_3," *in Proceedings of the NVSM Workshop*, pp. 101–102, 2008.

[129] M. Bocquet, G. Molas, L. Perniola, X. Garros, J. Buckley andM. Gély, J.P. Colonna, and H. Grampeix et. al., "Impact of a hto/al2o3 bi-layer blocking oxide in nitride-trap non-volatilememories," *Solid-State Electron.*, vol. 53, pp. 786–791, 2009.

[130] J. Hayden, F.Baker, S. Ernst, B. Jones, J. Klein, M. Lienand T. McNelly, T. Mele, H. Mendez, B. Y. Nguyen, L. Parrilloand W. Paulson, J. Pfiester, F. Pintchovski, Y.-C. See, R.Sivan, B. Somero, and E. Travis, "A high-performance sub-half micron cmos technology for fast srams," *in IEDM Tech. Dig.*, pp. 417–420, 1989.

[131] C. Y. Wong, J. Y.-C. Sun, Y. Taur, C.S. Oh, and R. Angelucci andB. Davari, "Doping of n+ and p+ polysilicon in a dual-gate cmos process," *in IEDM Tech. Dig.*, pp. 238–241, 1988.

[132] C.-H. Choi, P. R. Chidambaram, R. Khamankar, C. F. Machala andZ. Yu, and R. W. Dutton, "Dopant profile and gate geometric effects on polysilicon gate depletionin scaled mos," *IEEE Trans. Electron Devices*, vol. 49, no. 7, pp. 1227–1231, 2002.

[133] Y.-T. Hou, M.-F. Li, T. Low, and D.-L. Kwong, "Metal gate work function engineering on gate leakage of mosfets," *IEEE Trans. Electron Devices*, vol. 51, no. 11, pp. 1783–1789, 2004.

[134] C.-H. Lee, K.-C. Park, and K. Kim, "Charge-trapping memory cell of sio2 /sin/high-k dielectric al2o3 with tanmetal gate for suppressing backward-tunneling effect," *Appl. Phys. Lett.*, vol. 87, no. 073510, 2005.

[135] C. H. Lee, K. I. Choi, M. K. Cho, Y. H. Song, and K. C. Park andK. Kim, "A novel sonos structure of sio2/sin/al2o3 with tan metal gate for multi-gigabit flash memeries," *in IEDM Tech. Dig.*, 2003.

[136] J.K. Schaeffer, C. Capasso, L.R.C. Fonseca, S. Samavedam, D.C.Gilmer, Y. Liang, S. Kalpat, B. Adetutu, H.-H. Tseng, Y.Shiho, A. Demkov, R. Hegde, W.J. Taylor, R. Gregory, J.Jiang, E. Luckowski, M.V. Raymond, K. Moore, D. Triyoso andD. Roan, B.E. White Jr., and P.J. Tobin, "Challenges for the integration of metal gate electrodes," *in IEDM Tech. Dig.*, 2004.

[137] P. Majhi, H.C. Wen, H. Alshare'ef, K. Choi, R. Harris, P.Lysaght, H. Luan, Y. Senzaki, S. C. Song, B.H. Lee, and C. Ramiller, "Evaluation and integration of metal gate electrodes for future generationdual metal cmos," *Int. Conf. on Integ. Circuits and Tech.*, pp. 69–72, 2005.

[138] C. Ren, D. S. H. Chan, X. P. Wang, B. B. Faizhal, M.-F. Li andY. C. Yeo, A. D. Trigg, A. Agarwal, N. Balasubramanian, J.S. Pan, P. C. Lim, A. C. H. Huan, and D.-L. Kwong, "Physical and electrical properties of lanthanide-incorporated tantalum nitridefor n-channel metal-oxide-semiconductor field-effect transistors," *Appl. Phys. Lett.*, vol. 87, no. 073506, 2005.

[139] Y. Momiyama, H. Minakata, and T. Sugii, "Ultra-thin ta2o5/sio2 gate insulator with tin gate technology for 0.1ummosfets," *in Proceedings of the Symp. on VLSI Tech.*, pp. 135–136, 1997.

[140] Atsushi Yagishita, Tomohiro Saito, Kazuaki Nakajima, Seiji Inumiyaand Yasushi Akasaka, Yoshio Ozawa, Katsuhiko Hieda, YoshitakaTsunashima, Kyoichi Suguro, Tsunetoshi Arikado, and Katsuya Okumura, "High performance damascene metal gate mosfet's for 0.1 um regime," *IEEE Trans. Electron Devices*, vol. 47, no. 5, pp. 1028–1034, 2000.

[141] J. Pan, C. Woo, M.-V. Ngo, J. Xie, D. Matsumoto, D. Murthyand J.-S. Goo, Q. X.iang, and M.-R. Lin, "The effect of annealing temperatures on self-aligned replacement (damascene)tacn/tan-stacked gate pmosfets," *IEEE Trans. Electron Devices*, vol. 51, no. 4, pp. 581–586, April 2004.

[142] M. H. Tsai, S. C. Sun, C. E. Tsai, S. H. Chuang, and H. T. Chiu, "Comparison of the diffusion barrier properties of chemical-vapor-depositedtan and sputtered tan between cu and si," *J. Appl. Phys.*, vol. 79, no. 9, pp. 6932–6938, 1996.

[143] R. Sreenivasana, T. Sugawara, K. C. Saraswat, and P. C. McIntyre, "High temperature phase transformation of tantalum nitride films depositedby plasma enhanced atomic layer deposition for gate electrode applications," *Appl. Phys. Lett.*, vol. 90, no. 102101, 2007.

[144] M. Kozlowska, R. Oechsner, M. Pfeffer, A.J. Bauer, E. Meissnerand L. Pfitzner, H. Ryssel, W. Maass, J. Langer, B. Ocker andS. Schmidbauer, and J.-P. Gonchond, "Properties of tan thin films produced using pvd linear dynamic depositiontechnique," *Int. Conf. on Solid Films and Surf.*, , no. 14, 2008.

[145] S. Riedel, S.E. Schulz, J. Baumann, M. Rennau, and T. Gessner, "Influence of different treatment techniques on the barrier properties ofmocvd tin against copper diffusion," *Microelectron. Engineering*, vol. 55, pp. 213–218, 2001.

[146] A. Kim, A. J. Kellock, and S. M. Rossnagel, "Growth of cubic-tan thin films by plasma-enhanced atomic layer deposition," *J. Appl. Phys.*, vol. 92, no. 12, pp. 7080–7085, 2002.

[147] J. Y. Kim, S. Seo, D. Y. Kim, H. Jeon, and Y. Kim, "Remote plasma enhanced atomic layer deposition of tin thin films using metalorganicprecursor," *J. Vac. Sci. Technol. A*, vol. 22, no. 1, pp. 8–12, 2004.

[148] D. Gu, S. K. Dey, and P. Majhi, "Effective work function of pt, pd, and re on atomic layer deposited hfo2," *Appl. Phys. Lett.*, vol. 89, no. 082907, 2006.

[149] J. K. Schaeffer, C. Capasso, R. Gregory, D. Gilmer, L. R. C.Fonseca, M. Raymond, C. Happ, M. Kottke, S. B. Samavedam andP. J. Tobin, and B. E. White, "Tantalum carbonitride electrodes and the impact of interface chemistry ondevice characteristics," *Appl. Phys. Lett.*, vol. 101, no. 014503, 2007.

[150] M. Lemberger, S. Thiemann, A. Baunemann, H. Parala, R.A. Fischerand J. Hinz, A.J. Bauer, and H. Ryssel, "Mocvd of tantalum nitride thin films from tbtemt single source precursoras metal electrodes in cmos applications," *Surf. and Coat. Tech.*, vol. 201, pp. 9154–9158, 2007.

[151] A. Cacciato, L. Breuil, G. Van den bosch, O. Richard, A. Rothschildand A. Furnémont, H. Bender, J. A. Kittl, and J. Van Houdt, "Effect of top dielectric morphology and gate material on the performance of nitride-based flash memory cells," *Material Research Society*, Symposium F, spring 2008.

[152] J. Paul, V. Beyer, P. Michalowski, M. F. Beug, L. Bach, M.Ackermann, S. Wege, and A. Tilke, "Tan metal gate damage during high-k (al2o3) high-temperature etch," *microelectronics engineering*, 2008.

[153] H. N. Alshareef, H. C. Wen, H. R. Harris, K. Choi, H. F. Luanand P. Lysaght, P. Majhi, and B. H. Lee, "Modulation of the work function of silicon gate electrode using thin taninterlayers," *Appl. Phys. Lett.*, vol. 87, no. 052109, 2005.

[154] J. W. Hong, K. I. Choi, Y. K. Lee, S. G. Park, S. W. Lee andJ. M. Lee, S. B. Kang, G. H. Choi, S. T. Kim, and U-I. Chung andJ. T. Moon, "Characteristics

of paald-tan thin films derived from taimata precursor forcopper metallization," *Int. Interconnect Technology Conf.*, Juni 2004.

[155] S. G. Park, Y. K. Lee, S. B. Kang, H. S. Jung, S. J. Doh andJ. H. Lee, J. H. Choi, G. H. Kim, G. H Choi, and U I. Chung andJ. T. Moon, "Performance improvement of mosfet with hfo2-al2o3 laminate gate dielectricand cvd-tan metal gate deposited by taimata," *in IEDM Tech. Dig.*, 2003.

[156] G. Gerald Stoney, "The tension of metallic films deposited by electrolysis," *Nature*, pp. 172–175, Januar 1909.

[157] J. A. Ruud, A. Witvrouw, and F. Spaepen, "Bulk and interface stresses in silver-nickel multilayered thin films," *J. Appl. Phys.*, vol. 74, no. 4, pp. 2517–2523, 1993.

[158] B.K. Tay, X. Shi, H.S. Yang, H.S. Tan, Daniel Chua, and S.Y.Teo, "The effect of deposition conditions on the properties of tin thin filmsprepared by filtered cathodic vacuum-arc technique," *Surf. and Coat. Tech.*, , no. 111, pp. 229–233, 1999.

[159] R. Arghavani, N. Derhacobian, V. Banthia, M. Balseanu, N. Ingleand H. M'Saad, S. Venkataraman, E. Yieh, Z. Yuan, L.-Q. Xiaand Z. Krivokapic, U. Aghoram, K. MacWilliams, and S. E. Thompson, "Strain engineering to improve data retention time in nonvolatile memory," *IEEE Trans. Electron Devices*, vol. 54, no. 2, pp. 362–365, Februar 2007.

[160] H. C. Wen, H. N. Alshareef, H. Luan, K. Choi, P. Lysaght andH. R. Harris, C. Huffman, G. A. Brown, G. Bersuker, P. Zeitzoffand H. Huff, P. Majhi, and B. H. Lee, "Systematic investigation of amorphous transition-metal-silicon-nitride electrodesfor metal gate cmos applications," *in Proceedings of the Symp. on VLSI Tech.*, , no. 4A-3, pp. 46–47, 2005.

[161] P. Xuan and J. Bokor, "Investigation of nisi and tisi as cmos gate materials," *IEEE Electron Device Lett.*, vol. 24, no. 10, pp. 634–637, October 2003.

[162] A. S. Grove, *Physics and Technology of Semiconductor Devices*, John Wiley & Sons, 1967.

[163] C.-H. Lee, J. Choi, Y. Park, C. Kang, B.-I. Choi, H. Kimand H. Oh, and W.-S. Lee, "Highly scalable nand flash memory with robust immunity to program disturbanceusing symmetric inversion-type source and drain structure," *in Proceedings of the Symp. on VLSI Tech.*, pp. 118–119, 2008.

[164] H.-T. Lue, T.-H. Hsu, S.-Y. Wang, Y.-H. Hsiao, E.-K. Lai andL. W. Yang, T. Yang, K.-C. Chen, K.-Y. Hsieh, R. Liu, and C.-Y.Lu, "Study of local trapping and sti edge effects on charge-trapping nand flash," *in IEDM Tech. Dig.*, pp. 161–163, 2007.

[165] S. Aritome, S. Satoh, T. Maruyama, H. Watanabe, S. Shuto andG. J. Hemink, R. Shirota, S. Watanabe, , and F. Masuoka, "A 0.67um^2 self-aligned shallow trench isolation cell (sa-sti cell) for3v-only 256mbit nand eeproms," *in IEDM Tech. Dig.*, pp. 61–64, 1994.

[166] R. Kirisawa, S. Aritome, R. Nakayama, T. Endoh, and R. Shirota andF. Masuoka, "A nand structured cell with a new programming technology for high reliable5v-only flash eeprom," *in Proceedings of the Symp. on VLSI Tech.*, pp. 129–130, 1990.

[167] Peiqi Xuan, Min She, Bruce Harteneck, Alex Liddle, and Jeffrey Bokorand Tsu-Jae King, "Finfet sonos flash memory for embedded applications," *in IEDM Tech. Dig.*, 2003.

[168] M. Specht, U. Dorda, L.Dreeskornfeld, J.Kretz, F. Hofmann andM.Städele, R.J.Luyken, W.Rösner, H.Reisinger, E. Landgraf andT.Schulz, J.Hartwich, R. Kömmling, and L.Risch, "20 nm tri-gate sonos memory cells with multi-level operation," *in IEDM Tech. Dig.*, p. 2004, 2004.

[169] M. F. Beug, T. Melde, J. Paul, U. Bewersdorff-Sarlette, M. Czernohorskyand V. Beyer, R. Hoffmann, K. Seidel, D. A. Löhr, L. Bach andR. Knoefler, and A. Tilke, "Improvement of 48 nm tanos nand cell performance by introduction of a removableencapsulation liner," *in Proc. of 1st International Memory Workshop*, pp. 88–89, 2009.

[170] T. Melde, M. F. Beug, L. Bach, A. T. Tilke, R. Knoefler andU. Bewersdorff-Sarlette, V. Beyer, M. Czernohorsky, and J. Paul andT. Mikolajick, "Select device disturb phenomenon in tanos nand flash memories," *IEEE Electron Device Lett.*, vol. 30, no. 5, pp. 568–570, 2009.

[171] J.-C. Guo, M.-C. Chang, C.-Y. Lu, C.-H. Hsu, and S.-S. Chung, "Transconductance enhancement due to back bias for submicron n-mosfet," *IEEE Trans. Electron Devices*, vol. 42, no. 2, pp. 288–294, Februar 1995.

[172] K.-T. Park, S. C. Lee, J.-S. Sel, J. Choi, and K. Kim, "Scalable wordline shielding scheme using dummy cell beyond 40 nm nand flashmemory for eliminating abnormal disturb of edge memory cell," *Jap. J. Appl. Phys.*, vol. 46, pp. 2188–2192, 2007.

[173] T. Futatsuyama, N. Fujita, N. Tokiwa, Y. Shindo, T. Edahiroand T. Kamei, H. Nasu, M. Iwai, K. Kato, Y. Fukuda, and N. Kanagawaand et. al., "A 113mm^2 32gb 3b/cell nand flash memory," *in Proc. of the Int. Sol.-State Circuits Conf.*, pp. 242–243, 2009.

[174] S.-J. Joo, H.-J. Yang, H.-S. Kim, K.-H. Noh, H.-G. Lee, W.-S.Woo, J.-Y. Lee, and M.-K. Lee et. al., "Abnormal disturb mechanism of sub 100nm nand flash," *ext. abstracts of 2005 Int. conf. on sol. state dev. and mat.*, pp. 628–629, 2005.

[175] S. J. Joo, H. J. Yang, K. H. Noh, H. G. Lee, W. S. Woo, J.Y. Lee, M. K. Lee, W. Y. Choi, K. P. Hwang, H. S. Kim, S.Y. Sim, S. K. Kim, H. H. Chang, and G. H. Bae, "Abnormal disturbance mechanism of sub-100nm nand flash memory," *Jap. J. Appl. Phys.*, vol. 45, no. 8A, pp. 6210–6215, 2006.

[176] M. F. Beug, S. Parascandola, T. Hoehr, T. Müller, R. Reicheltand L. Müller-Meskamp, P. Geiser, T. Geppert, and L. Bach et. al., "Pitch fragmentation induced odd/even effects in a 36 nm floating gate nandtechnology," *in Proc. of the Non-Vol. Mem. Tech. Symp.*, pp. 77–81, 2008.

[177] M. F. Beug, T. Melde, M. Czernohorsky, V. Beyer, J. Paul, R. Hoffmann, U. Bewersdorff-Sarlette, R. Knoefler, and A. T. Tilke, "Analysis of tanos memory cells with sealing oxide containing blocking dielectric," *IEEE Trans. Electron Devices*, vol. 57, no. 7, pp. 1590–1596, 2009.