ANGEWANDTE PHYSIK

Speichern ohne Fluchtgefahr

Für leistungsfähigere nichtflüchtige Flash-Speicher sind neue Konzepte nötig.

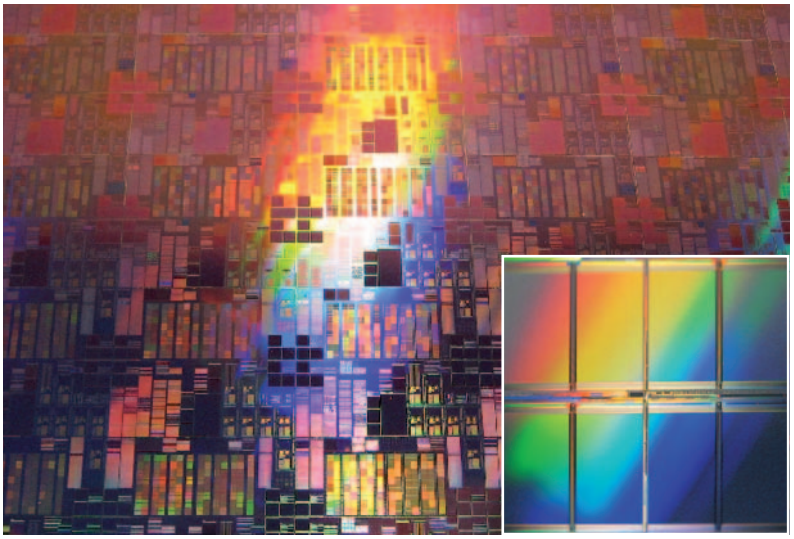
Armin Tilke, Florian Beug, Thomas Melde und Roman Knöfler

Seit ihrer Einführung vor 15 Jahren verdoppelt sich die Speicherdichte nichtflüchtiger Flash-Speicher kontinuierlich alle 12 Monate. Bislang gelang dies primär durch eine konventionelle Verkleinerung der Strukturen. Doch diese Skalierung ist in der letzten Zeit etwas ins Stocken geraten, weil sich der Lithographie große Herausforderungen stellen und vor allem, weil die Miniaturisierung klassischer Floating-Gate-Zellen an ihre Grenzen stößt.

Flash-Speicher begegnen uns fast überall im Alltag. In Handys, MP3-Playern, Digitalkameras oder USB-Sticks verbergen sich diese nichtflüchtigen Halbleiterspeicher, welche die Daten auch ohne Versorgungsspannung nicht verlieren. Der ehemalige Geschäftsführer von Samsung, Jae-sung Hwang, postulierte, dass sich die Speicherdichte von Flash-Speichern alle zwölf Monate verdoppelt. Analog zum bekannteren Mooreschen Gesetz für Mikroprozessoren [1] spricht man hier vom Hwangschen Gesetz. Bislang ließ sich diese Skalierung mit einer fortschreitenden Miniaturisierung der sog. Floating-Gate-Flash-Zellen erreichen.

Die Firma SanDisk brachte 1994 die erste kompakte Flash-Speicherkarte mit vier Megabyte Kapazität auf den Markt, 1998 stellte Sony den ersten Memory Stick vor. Wie erfolgreich diese Speichertechnik ist, zeigt die Tatsache, dass die typische Speichergröße von USB-Sticks mittlerweile im Bereich von zwei bis 16 Gigabyte liegt. Ein lange gehegter Traum vieler Computerbenutzer könnte bald in Erfüllung gehen, wenn auf Flash-Speichern basierende Halbleiterspeicher (Solid State Hard Disks, SSD) die konventionellen, auf magnetischen Effekten beruhenden Festplatten heutiger Computer nach und nach ersetzen. Für den Einsatz in Notebooks bringen konventionelle Festplatten Nachteile wie Stoßempfindlichkeit und vergleichsweise hohe Stromaufnahme mit sich. SSDs kommen gänzlich ohne mechanische Teile aus und verwenden heute bereits Siliziumchips, die bis zu acht Gigabyte auf einem Chip mit einer Fläche von nur 2,5 cm² speichern können [2].

Doch für Computeranwendungen muss garantiert sein, dass die Flash-Speicherzellen zuverlässig arbeiten und die gespeicherte Information über einen Zeitraum von mindestens zehn Jahren erhalten bleibt. Hier helfen zusätzliche Fehlerkorrektur-Codes, die in der Lage sind, die gespeicherten Daten bei Informationsverlust einzelner Zellen wieder herzustellen. Zudem sollten



Qimonda

Ausschnitt eines Wafers mit Teststrukturen zur Entwicklung von Charge-Trapping-Zellen bzw. eines Produktwafers mit jeweils 8 Milliarden Floating-Gate-NAND-Speicherzellen pro Chip (Inset).

Flash-Zellen bei 10 000 bis 100 000 Schreib-Lösch-Zyklen nicht ihre Speicherfähigkeit verlieren. Diese Anforderung kann je nach Art der Anwendung stark variieren. Für MP3-Player sind sicherlich geringere Zyklenzahlen nötig als für SSD-Speichermedien.[1)] Bei großen Flash-Speichern nutzt man den zur Verfügung stehenden Speicher gleichmäßig („wear levelling"), damit nicht einzelne Bereiche schnell hohe Zyklenzahlen erreichen, während andere ungenutzt bleiben.

Aus physikalischer und technologischer Sicht ist die derzeitige, konventionelle Floating-Gate-Technologie

1) Die maximal mögliche Zyklenzahl hängt auch davon ab, ob die Zellen zur Speicherung von einem Bit (Single Level Cell, SLC) oder mehreren Bits (Multi Level Cell, MLC) verwendet werden.

KOMPAKT

- Für die Speicherung nicht allzu großer Datenmengen haben sich mittlerweile Halbleiterspeicher durchgesetzt, die Metall-Oxid-Silizium-Feldeffektransistoren nutzen und ohne mechanische Bauteile auskommen (Floating-Gate-Zellen).
- Nutzt man in der Speicherzelle die NAND- statt der NOR-Architektur, so lässt sich die Kontaktfläche zwischen den Zellen einsparen und eine weitere Miniaturisierung erreichen.
- Noch höhere Speicherdichten versprechen sog. Charge-Trapping-Speicher, welche die Ladungen (und damit die Daten) in Störstellen einer dielektrischen Schicht speichern. Diese planaren Strukturen lassen sich prinzipiell weiter miniaturisieren.

Dr. Armin Tilke, Dr. Florian Beug, Thomas Melde und **Roman Knöfler**, Qimonda Dresden GmbH, Königsbrücker Straße 180, 01099 Dresden

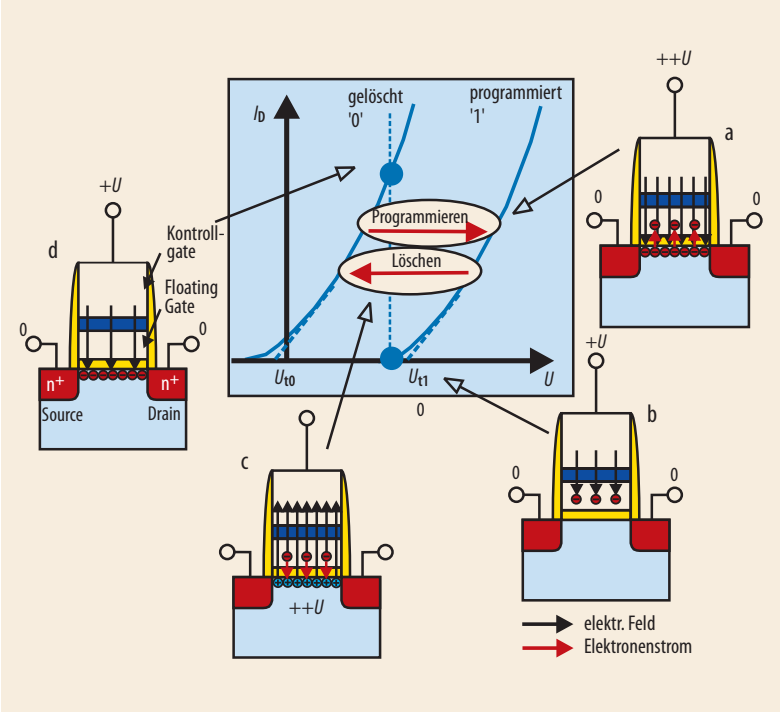


Abb. 1 Das Funktionsprinzip einer Floating-Gate-Flash-Speicherzelle (vgl. Text): Unterschieden wird zwischen Programmier- (a) und Löschvorgang (c). Im gelöschten Zustand (d) befinden sich keine Elektronen im Floating-Gate, im programmierten Zustand (b) ist das Floating-Gate geladen.

nicht beliebig weiter skalierbar. Neue Zellenkonzepte und Technologien sind notwendig, damit Hwangs Gesetz auch bis zu Strukturen unter 30 nm gültig bleibt.

Der Hauptvertreter nichtflüchtiger Speicher war über zwei Jahrzehnte die auf ferromagnetischen Prinzipien beruhende Festplatte. Neben dem Nachteil der Stoßempfindlichkeit benötigen rotierende Festplatten jedoch einen Elektromotor, der eine kleine Bauweise erschwert. Daher setzen sich seit rund zehn Jahren, vor allem bei nicht allzu großen Datenmengen, Halbleiterspeicher durch, die auf dem Speichern elektrischer Ladung in Metall-Oxid-Silizium-Feldeffekttransistoren beruhen und keine mechanischen Teile besitzen. Bei den Speicherzellen dient eine nicht kontaktierte polykristalline Siliziumschicht zwischen Gateoxid und Kontroll-Gate, das „Floating-Gate", als Speicherschicht für Elektronen. Im Prinzip arbeitet eine solche Zelle wie ein konventioneller Feldeffekttransistor [1].[2)]

Legt man eine positive Spannung an das Gate an, entsteht an der Grenzfläche zwischen Siliziumsubstrat und Gateoxid ein leitfähiger Inversionskanal aus Elektronen (**Abb. 1a**), charakterisiert durch die Einsatzspannung U_t. Bei sehr hoher positiver Spannung am Gate und wenn der Elektronenkanal auf Masse liegt, reicht die elektrische Feldstärke über das Gateoxid aus, damit Elektronen in das Floating-Gate tunneln können (Fowler-Nordheim-Tunneln). Die Zelle wird programmiert, indem das isolierte Floating-Gate Ladung speichert (**Abb. 1a**). Im programmierten Zustand schirmen die Elektronen im Floating-Gate einen Teil des Feldes, das durch eine positive Gatespannung entsteht, ab. Nun ist eine höhere Spannung am Kontroll-Gate anzulegen, um einen leitfähigen Inversionskanal zu erzeugen – die Einsatzspannung U_{t0} verschiebt sich zu einem höheren Wert U_{t1}. Der Transistor ist damit abgeschaltet, der Speicherzustand eine „1" (**Abb. 1b**). Durch eine am Substrat angelegte hohe positive Spannung lässt sich das Floating-Gate wieder entladen. Dadurch tunneln die Elektronen aus dem Floating-Gate in das Siliziumsubstrat (**Abb. 1c**). Das Floating-Gate ist gelöscht und die Einsatzspannung der Zelle wieder zu U_{t0} und dem Speicherzustand „0" verschoben (**Abb. 1d**). Beim Auslesen der Zelle wird eine Gatespannung zwischen U_{t0} und U_{t1} verwendet, sodass sich programmierte und unprogrammierte Zellen dadurch unterscheiden, dass sie leiten bzw. sperren.

Der Durchbruch von Floating-Gate-Flash-Speichern gelang Anfang der 1990er-Jahre, als eine neue Speicherarchitektur die Speicherdichte auf einem Chip enorm erhöhte. Während früher primär sog. NOR-Speicher zum Einsatz kamen, bei denen sich jede Speicherzelle direkt ansteuern lässt, dominieren inzwischen NAND-Speicher [4] den Markt (**Infokasten**). In diesen spart man die Fläche zur Kontaktierung der individuellen Zellen ein und minimiert somit die effektive Zellengröße.

2) Diese Technik wurde bereits 1967 bei Bell Labs vorgeschlagen [3].

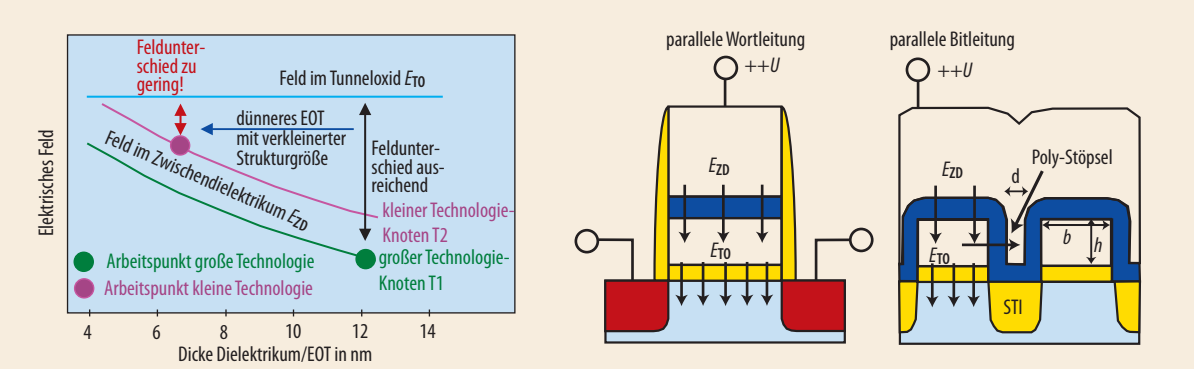


Abb. 2 Elektrisches Feld über das Tunneloxid E_{TO} bzw. das Zwischendielektrikum E_{ZD} (links) im programmierten Zustand: Der Unterschied zwischen E_{TO} und E_{ZD} wird mit jeder Technologiegeneration kleiner: Zum einen verringert sich damit die äquivalente Oxiddicke EOT. Durch Änderung der geometrischen Verhältnisse (Höhe h des Floating-Gates und Breite b, rechts) kann man die Felder im Zwischendielektrikum verändern und gelangt daher zum anderen für neue Technologiegenerationen auf andere Kurven für das elektrische Feld als Funktion des EOT.

Grenzen der Skalierbarkeit

Während des Programmiervorgangs liegen Spannungen um 15 Volt zwischen der selektierten und deren benachbarten Wortleitungen an. Bei immer kleineren Abmessungen müssen auch die Programmierspannungen sinken, denn sonst drohen laterale elektrische Durchbrüche zwischen benachbarten Wortleitungen.

Die elektrische Spannung, die am Kontroll-Gate anliegt, teilt sich auf in den Spannungsabfall über das Zwischendielektrikum (ZD), welches das Floating-Gate vom Kontroll-Gate isoliert, und den Spannungsabfall über das Tunneloxid. Das Tunneloxid muss dicker als ca. acht Nanometer sein, damit die Speicherzelle zuverlässig arbeitet. Bei dünneren Schichten wäre der Ladungsverlust durch störstellenunterstütztes Tunneln in einigen Speicherzellen so hoch, dass diese ihre Information in kurzer Zeit verlieren würden. Um mit verringerten Programmierspannungen immer noch eine für Fowler-Nordheim-Tunneln ausreichende Feldstärke über das Tunneloxid zu erreichen, gilt es daher, den Spannungsabfall über das Zwischendielektrikum zu verringern. Dies gelingt zunächst durch geometrische Tricks: Mit immer geringerer Breite b, aber gleichzeitig vergrößerter Höhe h des Floating-Gates und des Poly-Stöpsels gelingt es, die elektrostatische Kopplung zwischen Kontroll-Gate und Floating-Gate konstant und die elektrischen Felder im Zwischendielektrikum niedrig zu halten (**Abb. 2**). Diese geometrische Optimierung ist jedoch aufgrund von Strukturierungsproblemen inzwischen fast ausgereizt. Daher nutzt man Materialien mit hoher Dielektrizitätskonstante (high-k) wie Al_2O_3 als Zwischendielektrikum (ZD). Bei gleicher Dicke des ZDs lässt sich so eine geringere äquivalente Oxiddicke (EOT), d. h. auf die dielektrische Konstante von Siliziumdioxid normierte Dicke, erzielen. Damit verringert sich auch der Leckstrom. Allerdings verschlechtert sich das Verhältnis der elektrischen Felder, die über das ZD und das Tunneloxid im programmierten Zustand der Zelle abfallen. Gelangen beide Felder in dieselbe Größenordnung, ist es nicht mehr möglich zu programmieren, da der Leckstrom über das ZD dieselbe Größenordnung wie der programmierende Tunnelstrom hat. Damit fließt gleich viel Ladung in die Speicherzelle hinein wie hinaus (Programmiersättigung).

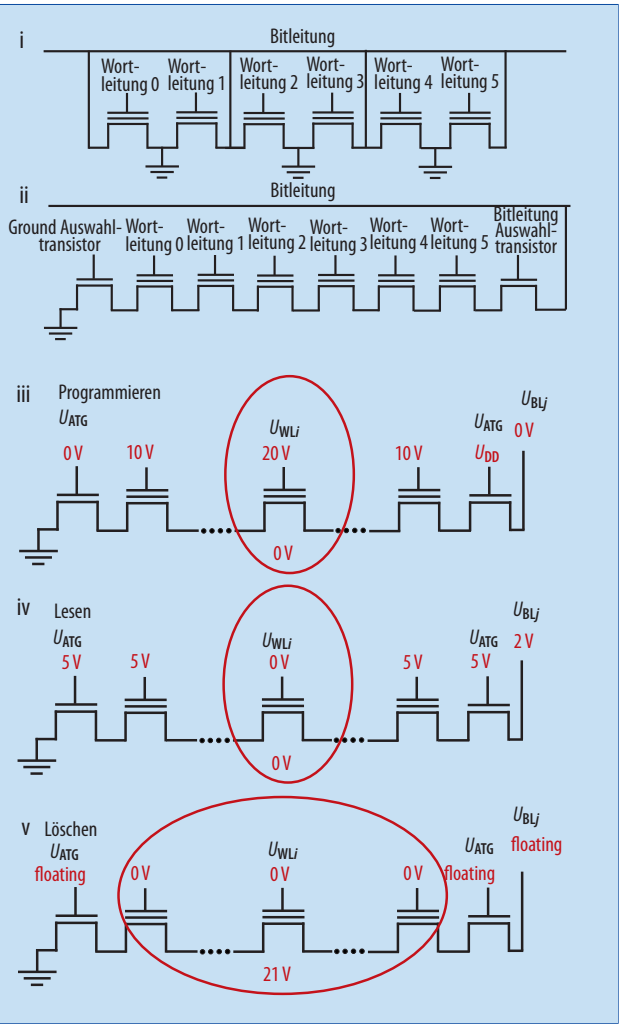
Beim Schnitt durch benachbarte Zellen parallel zur Bitleitung (**Abb. 2**, rechts) lässt sich ein weiteres Skalierungsproblem erkennen: Der Poly-Stöpsel zwischen benachbarten Floating-Gates wird so schmal, dass er seine Funktion zur elektrostatischen Kopp-

NOR- UND NAND-SPEICHER

In der Speicherarchitektur unterscheidet man **Wortleitungen**, d. h. Gateelektroden, die senkrecht zu den Transistoren verlaufen, und **Bitleitungen**, die durch die Transistoren selbst gebildet werden und wiederum senkrecht zu den Wortleitungen stehen. In einer **NOR-Anordung** der Flash-Speicherzellen (**Abb. i**) lässt sich jede Zelle einzeln ansteuern, wodurch sich eine hohe Lesegeschwindigkeit bei wahlfreiem Zugriff ergibt. Deshalb werden NOR Speicher verwendet, um Programme zu speichern und auszuführen. Da die Speicherzellengröße bei **NAND-Architekturen** [4] (**Abb. ii**) nur etwa 40 % derjenigen von NOR-Speichern beträgt, lassen sich damit in dichter gepackten Speicherzellenfeldern mehr Daten speichern. NAND-Flash-Speicher finden sich folglich in USB-Sticks, Speicherkarten und SSD-Festplatten (Solid State Drives) und sind schneller, wenn es darum geht, große und zusammenhängende Datenmengen zu schreiben und auszulesen.

Die Bezeichnungen NOR- und NAND-Flash leiten sich aus den logischen Verknüpfungen ab. Bei NOR sind die Speicherzellen parallel geschaltet, und es lässt sich jede Zelle einzeln ansteuern und auslesen. Bei NAND sind die Zellen in Serie geschaltet. Um eine einzelne Zelle im NAND-Strang zu lesen, müssen alle übrigen Zellen eingeschaltet sein. Durch alle Speicherzellen wird demnach „hindurch gelesen".

Für das Beschreiben einer NAND-Speicherzelle (**Abb. iii**) wird eine hohe positive Spannung U_{WLi} an die i-te selektierte Wortleitung gelegt. Um die 0 V Kanalpotential bis zu der zu programmierenden i-ten Zelle in den NAND-String zu transferieren, müssen alle Zellen im String eingeschaltet werden (hier mit 10 V an allen nicht zu programmierenden Wortleitungen). Während des Lesens (**Abb. iv**) werden alle Transistoren eines NAND-Strangs bis auf die auszulesende Zelle durch eine positive Spannung in einen leitfähigen Zustand geschaltet. Ist die selektierte Zelle leitfähig, ist Strom durch die Bitleitung messbar, sperrt sie, fließt kein Strom. Um alle Speicherzellen im NAND-String auszulesen, müssen alle Wortleitungen sukzessive ausgewählt und der Lesevorgang so oft durchgeführt werden, wie sich Speicherzellen im String befinden. Löschen ist möglich durch Anlegen einer hohen positiven Spannung an das Substrat (**Abb. v**). Durch diese Anordnung werden immer ganze Speichersektoren und damit auch alle Zellen im NAND-Strang gleichzeitig gelöscht.



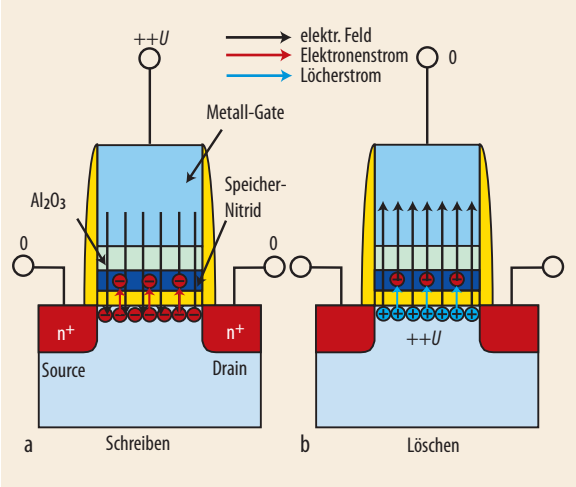


Abb. 3 Bei einer TANOS-Zelle besteht die Charge-Trap-Schicht aus SiN, das Kontroll-Gate aus TaN und das Zwischendielektrikum aus Al_2O_3. Beim Schreiben werden Elektronen im Charge-Trap-Nitrid gespeichert (a). Im Gegensatz zur Floating-Gate-Zelle findet das Löschen nun durch Löchertunneln in das Nitrid statt (b).

lung zwischen Floating- und Kontroll-Gate sowie zur Abschirmung benachbarter Zellen verliert. Dies ist darauf zurückzuführen, dass die Breite d des Stöpsels in die Größenordnung der Verarmungslänge des hochdotierten Polysiliziumgates von etwa 10 nm kommt. Es sind keine freien Ladungsträger mehr vorhanden und der Stöpsel wird isolierend. Die Kopplung wird schlechter, da sich die elektrisch wirksame Fläche zwischen Kontroll-Gate und Floating-Gate bei verarmtem Poly-Stöpsel verringert. Ein möglicher Ausweg ist die Einführung metallischer Materialien für das Kontroll-Gate [5].

Aus all diesen Gründen sind Floating-Gate-Flash-Speicher nicht beliebig skalierbar und erreichen zwischen 20 und 30 nm ihre Miniaturisierungsgrenze. Heute gefertigte Flash-Speicher sind bereits 40 bis 50 nm klein. Daher müssen innerhalb der nächsten fünf Jahre neue Flash-Speicherkonzepte, wie z. B. die TANOS-Charge-Trap-Zelle, Einzug halten.

Ladungen in der Falle

Anders als bei der Floating-Gate-Zelle lässt sich die Ladung auch in Störstellen einer dielektrischen Schicht speichern [6]. Die gespeicherte Ladung ist dann nicht frei beweglich wie im Polysilizium des Floating-Gates, sondern in Haftstellen der Isolatorschicht gefangen (engl. trapped). Aus diesem Grund spricht man von einem Charge-Trapping-Speicher [7].

Eine sehr effektive Speicherschicht ist Siliziumnitrid. Ersetzt man das Floating-Gate durch eine solche Siliziumnitridschicht, erhält man eine SONOS-Zelle (Polysilizium-Gate / Oxid-Zwischendielektrikum / Charge-Trap-Nitrid / Tunneloxid / Substrat) [8]. Dort geschieht das Programmieren durch Tunneln von Elektronen durch das Tunneloxid in die Nitrid-Speicherschicht. Das Löschen hingegen beruht aber nicht auf dem Zurücktunneln der Elektronen aus dem Nitrid. Die Elektronen befinden sich in tiefen Störstellen der Nitridschicht und können diese somit nur schwer wieder verlassen. Daher löscht man diese Zellen durch Tunneln von Löchern aus dem Substrat in die Nitridschicht, die mit den gespeicherten Elektronen rekombinieren. Diese kompensieren die gespeicherte Elektronenladung.

Obwohl die SONOS-Struktur sehr naheliegend ist und eine der ersten nichtflüchtigen Speichertechnologien darstellte, hat sie gravierende Nachteile. Aufgrund der im Vergleich zu Elektronen höheren Tunnelbarriere für Löcher muss das Tunneloxid so dünn sein, dass Löcher durch nur wenig feldabhängiges direktes Tunneln in die Charge-Trap-Schicht gelangen. Dadurch ist der Löcher-Tunnelstrom ausreichend groß, um die Speicherzelle effizient mit niedrigen Spannungen zu löschen. Der Nachteil eines dünnen Tunneloxids ist aber eine verminderte Fähigkeit, die Ladung dauerhaft zu halten.

Bei einer SONOS-Struktur mit dickem Tunneloxid ist die Elektronenbarriere zwischen Gate und Speicherschicht unter Löschbedingungen so klein, dass Elektronen während des Löschens vom Gate in die Speicherschicht nachtunneln können. Sind die Löcher-Injektionsströme vom Substrat und die Elektronen-Injektionsströme von der Gate-Elektrode gleich groß, endet der Löschvorgang (Löschsättigung). Dadurch erreichen die Speicherzellen, insbesondere bei dicken Tunneloxiden, nicht das erforderliche U_{t0}. Dieser Injektion von der Gate-Elektrode lässt sich mit einer größeren Elektronen-Tunnelbarriere durch das

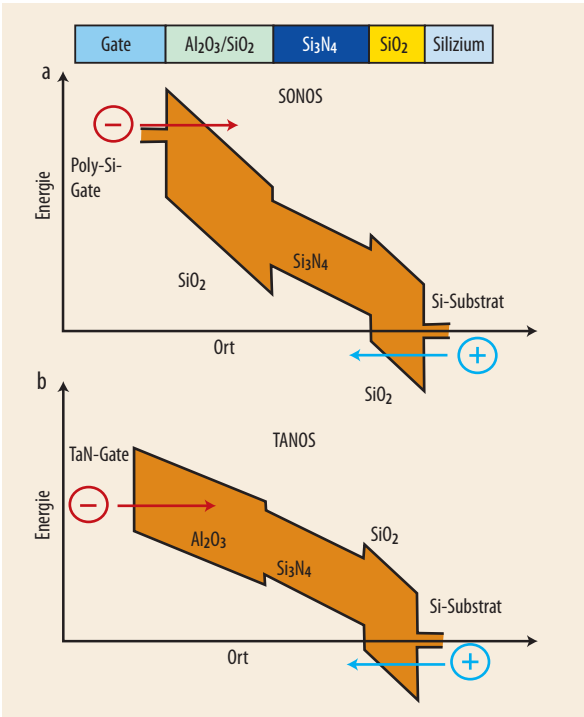


Abb. 4 Banddiagramme von SONOS-Zelle (a) und TANOS-Zelle (b). Die Strukturen unterscheiden sich durch das Zwischendielektrikum (SiO_2 bzw. Al_2O_3) sowie das Gatematerial (Poly-Si bzw. TaN). Da die Austrittsarbeit von TaN verglichen mit Polysilizium höher ist und zudem das elektrische Feld im Al_2O_3 geringer ist als im SiO_2, werden beim Löschen der TANOS-Zelle keine Elektronen vom Gate in das Nitrid injiziert (rote Pfeile). Dadurch wird die Löschsättigung vermieden.

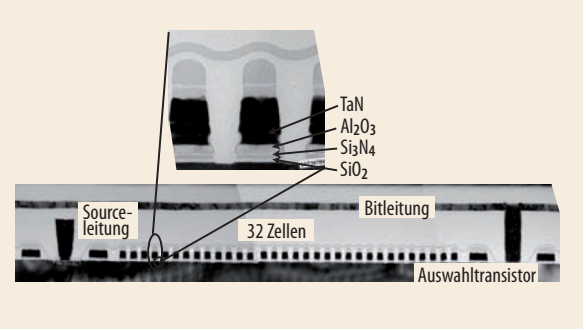


Abb. 5 Bei diesem NAND-Strang in TANOS-Technologie mit Strukturgrößen von 48 nm sind die 32 Speicherzellen sowie die Auswahltransistoren an beiden Enden des Strangs zu erkennen.

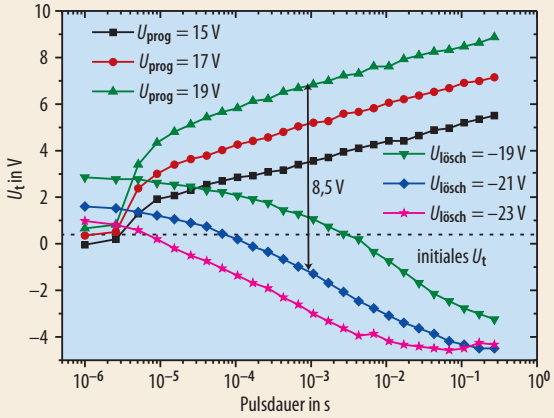


Abb. 6 Programmieren/Löschen für 48 nm TANOS-Zellen.

Zwischendielektrikum entgegenwirken. Hierfür gibt es zwei Optionen:

- Zum einen kann man das Feld im Zwischendielektrikum durch ein high-*k*-Material reduzieren. Wegen seiner geringen Defektdichte und moderatem *k*-Wert eignet sich z. B. Al_2O_3. Die Bandverkippung ist in dem Fall kleiner, und der Elektronentunnelstrom verringert sich.
- Ersetzt ein Elektrodenmaterial mit höherer Austrittsarbeit (z. B. TaN) das Silizium, reduziert sich der Strom durch die Gate-Elektroden abermals.

Die Kombination von Al_2O_3-Zwischendielektrikum und TaN-Gate bezeichnet man als TANOS-Zelle [9] (**Abb. 3**). Das Tunneloxid kann hier wesentlich dicker ausfallen als bei einer SONOS-Zelle. Erst dadurch gelangen die Speicherzeiten für die Ladungen und damit die Informationen in den Bereich von Jahren und ermöglichen den Einsatz der Zellen für den Mehrbit-Betrieb. Anhand der Bandkantendiagramme lassen sich die Vorteile von TANOS gegenüber SONOS erläutern (**Abb. 4**).

In Bezug auf die Skalierbarkeit haben Charge-Trapping-Speicher gegenüber konventionellen Floating-Gate-Zellen einen unschätzbaren Vorteil: Da sich die Elektronen auch noch in den tiefen Störstellen fangen lassen, wenn über dem Tunnel- und dem Zwischenoxid gleiche Felder anliegen, sättigt die Programmierung nicht (vgl. **Abb. 2**). Vor allem aber handelt es sich bei haftstellenbasierten Speicherzellen im Gegensatz zu Floating-Gate-Speichern um planare Strukturen. Daher ist die Skalierbarkeit strukturell deutlich weniger beschränkt.

Auch über die erwähnten physikalischen Grenzen der Skalierung konventioneller Flash-Zellen hinaus spielen lithographische Limitierungen eine immer stärkere Rolle. Mit konventioneller optischer Lithographie lassen sich Strukturgrößen bis etwa 50 nm realisieren. Mit sog. Immersions-Scannern, bei denen der im Vergleich zu Luft höhere Brechungsindex von Wasser zu einer höheren optischen Apertur und damit einer höheren Auflösung führt, sind Strukturen kleiner als 40 nm möglich.

Der technologisch nächste Schritt wäre die Einführung von extremer UV-Strahlung (EUV) mit einer Wellenlänge von 13,5 nm, die im Prinzip eine lithographische Auflösung um etwa 10 nm erlaubt. Allerdings ist diese Technologie noch nicht ausgereift. So fehlen leistungsstarke EUV-Quellen und geeignete Optiken.

Um reguläre Speicherfelder zu strukturieren, haben sich daher in der letzten Zeit sog. Doppelstrukturierungen etabliert [10]. Mithilfe von strukturellen Techniken (Spacer-Techniken) lässt sich dabei die Periode von lithografisch abgebildeten, regelmäßigen Linienfeldern verdoppeln und die Linienbreite halbieren. Diese Methode stellt jedoch enorm hohe Anforderungen an die Genauigkeit und Reproduzierbarkeit von Prozessschritten, wie Abscheidungen und Ätzungen. Zudem wird der gesamte Prozessfluss kompliziert und teuer.

Besseres Zellendesign nötig

Im Rahmen des europäischen Entwicklungsprojektes GOSSAMER[3] wird an verschiedenen Standorten der Halbleiterindustrie wie Dresden, Leuven (Belgien) und Agrate (Italien) an der Entwicklung hochperformanter TANOS-Zellen und NAND-Speicherfelder gearbeitet. **Abb. 5** zeigt ein Beispiel für eine solche, bei Qimonda in Dresden gefertigte Zelle, **Abb. 6** deren Verhalten beim Programmieren und Löschen [11]. Bei Pulszeiten von 10^{-3} s und typischen Programmier- und Löschspannungen von 19 V bzw. –21 V unterscheiden sich die Einsatzspannungen U_t des gelöschten und programmierten Zustands um 8,5 V. Die Flash-Zelle zeigt eine Zyklenfestigkeit bis etwa 10 000 Schreib/Löschzyklen, was normalen Anforderungen bereits gut entspricht.

Die momentan noch größten Probleme für den kommerziellen Einsatz heutiger TANOS-Zellen bereitet die im Vergleich zu Floating-Gate-Flash-Zellen relativ schlechte Ladungshaltung. Mit dem Verlust der im Nitrid gespeicherten Elektronen geht der Informationsgehalt der Zelle verloren. Um die Ladungshaltung zu untersuchen, werden Flash-Zellen bei hohen Temperaturen gehalten und danach gemessen. Durch die hohen Temperaturen geht die Ladung wesentlich schneller verloren, und die notwendige Messzeit verringert sich von Jahren auf einige Stunden. Bei einer Temperung über 2 Stunden bei 200°C verringert sich die Einsatzspannung des programmierten Niveaus

3) Dies steht für „Giga-scale oriented solid state flash memory for Europe", Projektnummer 214431. Mehr Infos unter www.fp-gossamer.eu.

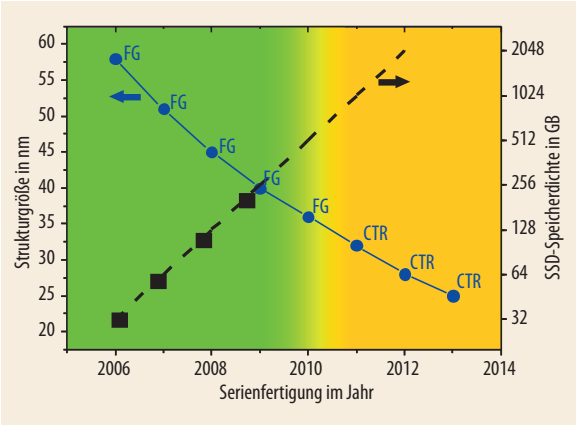


Abb. 7 Im Bereich von 30 nm wird die Skalierungsgrenze von Floating-Gate-Zellen erreicht, sodass ein Übergang zu Charge-Trapping-Zellen vorauszusehen ist. Den in der Fertigung befindlichen Strukturgrößen stehen die kommerziell etablierten Speicherdichten von SSDs gegenüber.

noch um etwa 15 % (600 mV). Hier ist es nötig, das Zellendesign weiter zu verbessern, insbesondere durch optimierte Speichermaterialien.

Neben einer reinen Strukturverkleinerung verdoppelt sich die Speicherdichte eines NAND-Speichers auch dadurch, dass jede Zelle zwei Bits speichert [12]. Durch Einbringen von unterschiedlich viel Ladung in die Zelle lassen sich vier verschiedene U_t-Niveaus programmieren, die zwei Bits (0,0) und (0,1) sowie (1,0) und (1,1) darstellen. Man kontrolliert die Ladungsmenge in einer Speicherzelle, indem man bei jeweils fester Programmierzeit sukzessive und in kleinen Schritten die Programmierspannung erhöht. Dadurch gelangt immer eine kleine Ladungsmenge in die Speicherzellen. Die Programmierspannung wird so lange erhöht, bis eine definierte Einsatzspannung überschritten ist. Moderne Floating-Gate-Speicher verwenden dieses Konzept bereits. An Produkten, die drei bis vier Bits pro Zelle speichern, wird bereits gearbeitet [2].

Ausblick

Seit der Einführung der ersten NAND basierten 32 GB-Festplatte im Jahr 2006 halten SSDs zunehmend Einzug in den Massenmarkt. Anfängliche Probleme wurden nach und nach behoben, und besonders die Schreib- und Lesegeschwindigkeiten haben sich deutlich verbessert. Aktuell sind für die Volumenproduktion bereits SSDs mit 256 GB und 512 GB Speicher und schnellen SATA II-Interfaces angekündigt. Die Speicherdichte für den Massenmarkt steigt direkt proportional zu den Innovationen und Kosteneinsparungen bei der Chipfertigung und folgt etwa Hwangs Gesetz, das jedoch allmählich durch technologische Hindernisse ins Stocken gerät. Zudem droht die notwendige Umstellung des Zellenkonzeptes. Glaubt man der Vorhersage der durch die großen, etablierten Halbleiterhersteller aufgestellten International Technology Roadmap for Semiconductors (ITRS), ist dieser Konzeptwechsel bereits in den nächsten Jahren zu erwarten (**Abb. 7**). Hier deutet sich an, dass für SSDs mit Speicherdichten im Terabyte-Bereich eben diese neuen Zellenkonzepte Einzug halten müssen.

Aus kommerzieller Sicht behindern zudem gegenwärtig die sehr stark gesunkenen Preise für NAND-Speicherchips Investitionen in neue Technologien. Aber auch zusätzliche Fragestellungen kommen auf die Flash-Industrie zu: Neue Materialien mit geringerem elektrischem Widerstand für die Wort- und Bitleitungsmetalle wie Kupfer sorgen in NAND-Chips für höhere Schaltgeschwindigkeiten. Ebenso werden geänderte Schaltungsarchitekturen sowie Ausleseverfahren die NAND-Speicher schneller und zuverlässiger machen.

Im Moment ist damit zwar ein Ende von Hwangs Gesetz nicht abzusehen, und der Siegeszug der NAND-Flash-Speicher wird sich fortsetzen. Die technologischen und physikalischen Herausforderungen für diese weitere Skalierung – hin zu Terabyte SSDs – werden allerdings deutlich zunehmen.

Literatur

[1] *A. Tilke, K. Schrüfer* und *Ch. Stapelmann,* Physik Journal, Juni 2007, S. 35
[2] *C. Trinh* et al., in ISSCC Proc. **246** (2009)
[3] *D. Kahng* und *S. M. Sze*, Bell. Syst. Techn. J. **46,** 1288 (1967)
[4] *R. Shirota* et al., VLSI Tech. Digest **33** (1988)
[5] *N. Chan* et al., in NVMTS Proc. **82** (2008)
[6] *H. A. R. Wegener* et al., IEEE IEDM Abstract **58** (1967)
[7] *T. Mikolajick* et al., VLSI TSA Techn. Digest **112** (2007)
[8] *Y. Hu* und *M. H. White*, Solid State Eletron. **36**, 1401 (1996)
[9] *C. H. Lee* et al., IEDM Tech. Digest (2003)
[10] *M. F. Beug* et al., in NVMTS Proc. **77** (2008
[11] *M. F. Beug* et al., International Memory Workshop, zur Publikation angenommen
[12] *M. Bauer* et al., ISSCC Techn. Digest **132** (1995)

DIE AUTOREN

Armin Tilke (FV Halbleiter) studierte Physik an der TU München und promovierte im Bereich der Silizium-Nanoelektronik am Center for Nano Science an der LMU München. Zwischen 1999 und 2003 arbeitete er bei Infineon Technologies in München, Dresden und East Fishkill (USA). Inzwischen entwickelt er bei Qimonda im Rahmen eines EU-Projektes neuartige TANOS NAND-Flash-Speicher. **Florian Beug** studierte an der Uni Hannover und Université Nice-Sophia Antipolis Physik und promovierte 2004 in Elektrotechnik. Ein Post-doc-Aufenthalt führte ihn ans NMRC/Tyndall Institute in Cork, Irland. Seit 2005 arbeitet er bei Infineon/Qimonda. **Thomas Melde** hat sein Studium der Elektrotechnik an der TU Dresden mit einer Diplomarbeit über die Simulation von nichtflüchtigen SONOS-Speichern abgeschlossen. Anschließend begann er eine Promotion bei Qimonda. **Roman Knöfler** arbeitet seit seinem Physikstudium in Chemnitz an der Skalierung von MEMS und Halbleitern, davon seit acht Jahren in den Entwicklungslaboren von Infineon und Qimonda an Standorten in Deutschland und den USA.