Bild: Thomas Kuhlenbeck

Blitzschneller Speicher

Flash-Grundlagen, Teil 1: Von SLC bis QLC

Elektronen lassen sich ohne weitere Stromversorgung dauerhaft einsperren. Sind genug in einer Zelle gefangen, hat man ein Bit gespeichert. Doch damit hat die Entwicklung des in jeder SSD oder Speicherkarte anzutreffenden Flash-Speichers erst begonnen.

Von Tim Niggemeier

Das ist schnell wie der Blitz, dachte sich der Legende nach ein Entwickler bei Toshiba im Jahr 1984 bei den Arbeiten an einem neuen Speichertyp. Daraus entstand der Name für die Technik, auf der wir heute alle unsere Daten speichern: Flash hat sich zur dominierenden Festpeichertechnik für Anwendungen entwickelt, bei denen es auf Geschwindigkeit ankommt.

Billionen von Speicherzellen bringen die Hersteller dafür auf einen Silizium-Wafer auf. Im weiteren Herstellungsprozess schneiden sie die Wafer in kleine Teile; daraus entstehen Speicherbausteine für SSDs, Speicherkarten, Smartphones oder Geräte der Unterhaltungselektronik.

Wer heute von Flash-Speicher spricht, meint fast immer NAND-Flash. Für Spezialfälle existieren zwar andere Speichertechniken, etwa NOR-Flash oder 3D XPoint, aber NAND-Flash erlaubt die kostengünstigste Massenproduktion.

Funktionsweise von NAND-Flash

Eine NAND-Flash-Zelle ähnelt im Aufbau einem Feldeffekttransistor. Dieser leitet beim Anlegen einer Spannung am Gate, durch die sich ein elektrisches Feld zwischen dem Gate und dem Substrat (p) bildet. Das Feld erzeugt eine negative Raumladungszone unter dem Oxid, wodurch sich eine leitfähige Zone zwischen den beiden n-Gebieten „Source“ und „Drain“ bildet.

Um aus einem Feldeffekttransistor eine Speicherzelle zu machen, bringt man in die Oxidschicht zwischen dem Gate und dem Substrat eine elektrisch leitende Schicht ein, die vom isolierenden Oxid umgeben und nicht angeschlossen ist. Diese Lage nennt man Floating Gate.

Solange das Floating Gate keine Ladung trägt, verhält sich die Speicherzelle beim Lesezugriff wie ein Feldeffekttransistor: Durch das Anlegen einer positiven Spannung am Control Gate bildet sich zwischen Source und Drain ein Kanal und der Transistor leitet. Dies entspricht einer logischen Eins. Trägt das Floating Gate jedoch eine negative Ladung in Form freier Elektronen, wird das elektrische Feld des Control Gates kompensiert. Es wirkt dann kein elektrisches Feld mehr auf das Substrat, und es bildet sich kein Kanal. Der Transistor sperrt trotz einer positiven Spannung zwischen Control Gate und Substrat. Dies repräsentiert eine logische Null. In diesem Beispiel kann das Floating Gate also zwei Zustände einnehmen: geladen und ungeladen. Damit kann die Zelle ein Bit speichern.

Um die Information in einer NAND-Flash-Zelle zu speichern, also freie Elektronen in das Floating Gate einzubringen, müssen diese die Isolationsschicht über dem Substrat überwinden. Hier kommt der Fowler-Nordheim-Tunneleffekt zur Anwendung: Legt man am Control Gate eine Spannung an, die viel höher ist als die Schaltspannung des Transistors, entsteht ein so starkes elektrisches Feld, dass Elek-

tronen die Potenzialbarriere des Isolationsmaterials durchtunneln können. Das Tunneln ist ein quantenmechanischer Effekt, der es den Elektronen ermöglicht, eine Potenzialbarriere zu überwinden, die höher ist als die Energie der Elektronen. Einmal in das isolierte Floating Gate eingebracht, bleiben die Elektronen und somit die gespeicherte Information erhalten, ohne dass eine Versorgungsspannung anliegen muss oder die Zellen regelmäßig aufgefrischt werden müssen, wie es bei SRAM oder DRAM der Fall ist.

Um die Programmierung der Zelle wieder zu löschen, die Elektronen also zu entfernen, nutzt man das gleiche Verfahren mit umgekehrter Polarität. Der Polaritätswechsel beeinflusst jedoch alle Speicherzellen in der Umgebung, wodurch das Löschen nicht so feingranular erfolgen kann wie das Programmieren. Dies ist der Grund, weshalb sich nur ganze Flash-Blöcke löschen lassen, Programmieren aber in kleinen Einheiten möglich ist.

SLC, MLC, TLC, QLC

Bei den ersten NAND-Flashes konnte man nur zwischen den Zuständen programmiert (Floating Gate negativ geladen) und gelöscht (Floating Gate neutral) unterscheiden. Für diesen 1-Bit-pro-Zelle-Speicher hat sich die Bezeichnung Single Level Cell (SLC) etabliert.

Der Begriff Level bezieht sich auf die Anzahl der Bits und nicht, wie man vermuten könnte, auf die Anzahl der Zustände. Eine kostengünstigere Fertigung erreichten die Hersteller durch immer weitere Reduzierung des Technologieknotens (der kleinsten Strukturgröße) und damit der benötigten Siliziumfläche.

Mit kleineren Strukturgrößen nahm die Zugriffsgeschwindigkeit zu. Gleichzeitig nahm die Zuverlässigkeit der Speicherzellen ab, da sich die Anzahl der Elektronen im Floating Gate reduzierte und der Abstand zu Nachbarzellen geringer wurde, wodurch sich die Ladungen und elektrischen Felder der Zellen gegenseitig stärker beeinflussten.

Unterhalb einer Strukturgröße von 20 Nanometer begann man, die Speicherdichte durch die Erhöhung der Anzahl der Bits pro Zelle zu erhöhen; die Multi Level Cell (MLC) war geboren. Hier kann jede Flash-Zelle vier Zustände annehmen, die durch die Anzahl der Elektronen im Floating Gate bestimmt ist. Entsprechend kann MLC zwei Bits pro Zelle speichern, wodurch MLC die doppelte Speicherdichte von SLC pro Fläche erreicht. Vermutlich hatte niemand die weitere Erhöhung der möglichen Zustände vorausgesehen, denn sonst wäre Dual Level Cell der passendere Begriff gewesen.

Die nächsten Schritte waren die Triple Level Cell (TLC) mit einer Unterscheidung von acht Zuständen und drei Bit pro Zelle und die Quad Level Cell (QLC) mit 16 Zuständen für vier Bit. Penta Level Cells (PLC) mit 5 Bits Speicherkapazität pro Zelle sind in der Entwicklung, bei diesen muss der Detektor 32 Zustände (2^5) unterscheiden.

Für die Speicherung von mehr als einem Bit pro Zelle sind mehr als zwei Potenziale des Floating Gates notwendig; der Controller muss dafür eine genau definierte Menge an freien Elektronen in das Floating Gate einbringen. Dazu wird die Programmierspannung mehrfach mit kurzen Pulsen angelegt; nach jedem Programmierpuls wird das erreichte Potenzial ausgelesen. Ist die notwendige Anzahl an Elektronen im Floating Gate erreicht, stoppt der Programmiervorgang.

Ausgelesen werden die Potenziale im Floating Gate durch stufenweise Erhöhung der Spannung am Control Gate, und zwar so lange, bis der Transistor leitet.

Und 3D-Flash?

Mit 14 Nanometer Strukturgröße war das Ende der Miniaturisierung erreicht, eine weitere Reduktion unwirtschaftlich. Die NAND-Hersteller begannen stattdessen, die vorher in einer Ebene angeordneten Speicherzellen übereinander zu stapeln.

Die Strukturen wurden wieder größer, es passten auch wieder mehr Elektronen in das Floating Gate und der Abstand zu den Nachbarzellen vergrößerte sich, wodurch die gegenseitige Beeinflussung durch elektrische Felder abnahm. Zudem verkürzten sich die Signalwege durch den Stapel, was einen Geschwindigkeitsvorteil brachte.

Der Kanal zwischen Source und Drain ist bei 3D-Flash zylinderförmig und von ringförmigen Floating Gates und Control Gates sowie den Oxidlagen umschlossen, die Verkettung der Zellen geschieht vertikal. Aktuell ist der höchste 3D-NAND als Doppelstapel mit 176 Lagen (2×88 Lagen) verfügbar.

Der letzte große Technologieschritt, der bei den Flash-Herstellern gleichzeitig mit oder erst nach der Einführung von 3D erfolgte, war der Wechsel von Floating Gate auf Charge Trap: Das elektrisch leitende und dicke Floating Gate aus Polysilizium wurde durch eine dünne Lage aus nichtleitendem Siliziumnitrit ersetzt. Die eingebrachten Elektronen können sich in der Charge-Trap-Schicht nicht mehr frei bewegen. Entsprechend wirken sich Abnutzungsschäden im Tunneloxid weniger stark aus. Zudem beeinflussen sich durch die fehlende Beweglichkeit der Elektronen und die geringe Dicke der Charge-Trap-Schicht die Nachbarzellen weniger.

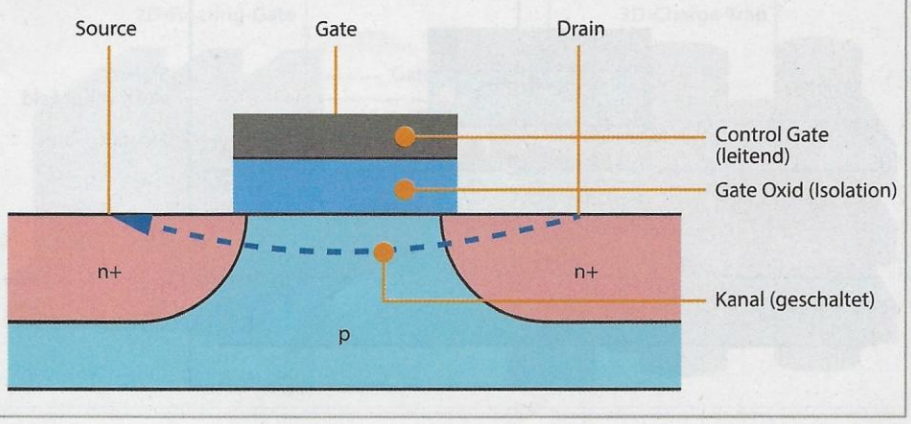
Geringere Oxid-Dicken und damit einhergehende kleinere Spannungen reduzieren die Beeinflussung weiter. Damit konnten die Hersteller zum einen die Zellgröße wieder schrumpfen und zum anderen die Programmiergeschwindigkeit erhöhen.

Verwaltungsgrößen

Die kleinste zum Lesen und Programmieren adressierbare Einheit bei NAND-Flash

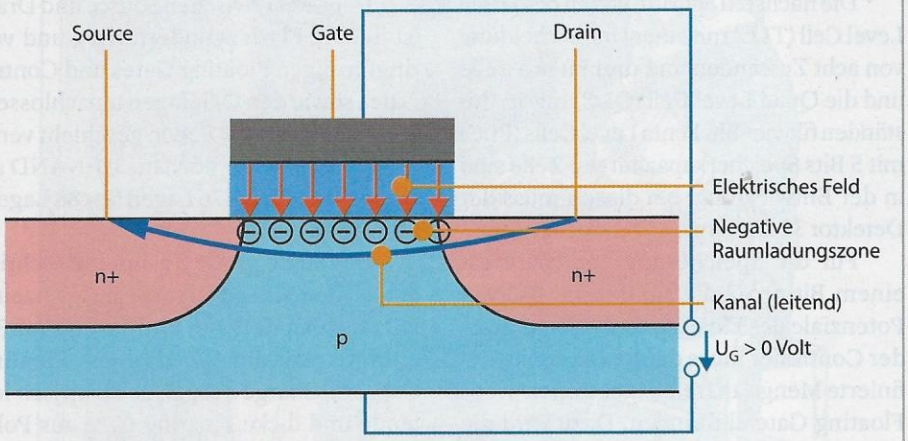
Feldeffekttransistor

NAND-Flash beruht auf dem Prinzip eines Feldeffekttransistors. Eine elektrische Spannung zwischen dem Gate und dem Substrat steuert die Leitfähigkeit eines Kanals zwischen Source und Drain.



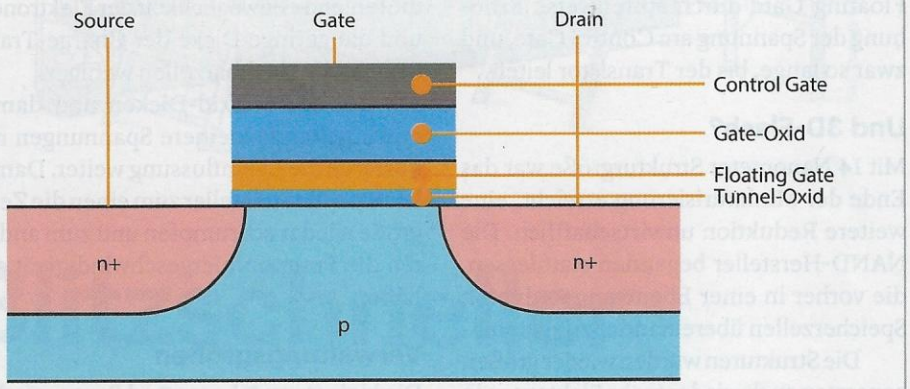
Feldeffekttransistor – leitend

Eine positive Spannung am Gate erzeugt ein elektrisches Feld; unter der Oberfläche des Siliziumsubstrates entsteht dadurch ein Elektronenüberschuss. Der Transistor leitet.



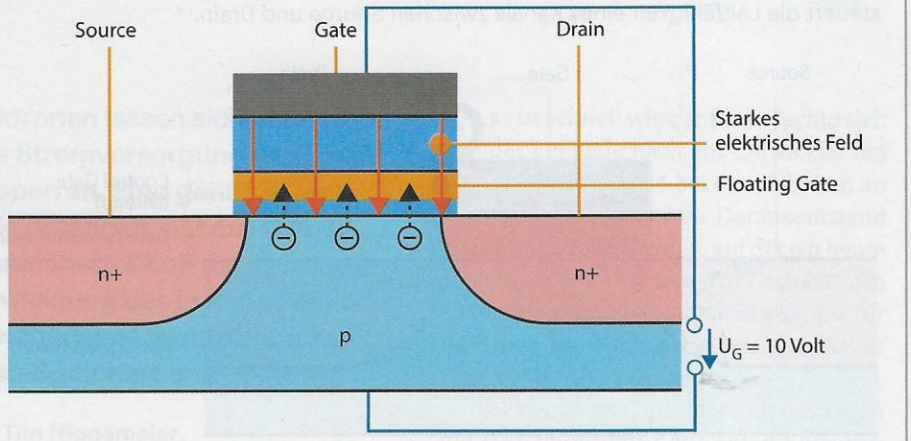
Flash-Zelle

Durch das Hinzufügen eines Floating Gates und des Tunnel-Oxids verwandelt sich der Transistor in eine Speicherzelle. Er verhält sich wie zuvor und leitet beim Anlegen einer Spannung am Gate.



Programmierung

Bei der Programmierung der Speicherzelle tunneln Elektronen durch Anlegen einer hohen Spannung durch die Isolationsschicht hindurch zum Floating Gate und verbleiben dort.



ist die Page. Üblich ist eine Größe von 16 KByte. Überwiegend werden heute alle Bits in einer Speicherzelle gleichzeitig programmiert; bei Charge-Trap-TLC sind es etwa 48 KByte zugleich.

Mehrere Pages bilden einen Block, der die kleinste löschbare Einheit darstellt. Ein Block hat eine Größe zwischen 4 und 72 MByte. Tausende von Blöcken sind zu Planes zusammengefasst; pro Plane kann nur ein Block zeitgleich gelesen, beschrieben oder gelöscht werden.

Zwei oder vier Planes fassen die Hersteller auf einem Flash-Die zusammen. Dies erhöht den Durchsatz, da der Controller auf mehrere Blöcke gleichzeitig zugreifen kann. Der endgültige Flash-Baustein besteht aus einem Stapel von 2 bis 16 Flash-Dies.

Auslese bei der Produktion

NAND-Flash wird wie alle Halbleiter auf Silizium-Wafern produziert. Bei der Wafer-Herstellung können die Hersteller das Produkt in zwei verschiedene Richtungen optimieren: entweder auf Geschwindigkeit für den Enterprise- und Consumer-Markt oder auf lange Lebensdauer und großen Temperaturbereich, etwa für den Automobilsektor.

Nach der Produktion folgt noch auf dem Wafer ein erster Test des Speichers. Dazu kontaktiert man die späteren Dies mit Nadeln und führt generelle Funktionstests sowie einen ersten Schreib-Lese-Test durch.

Das Ergebnis ist eine grobe Einteilung in verschiedene Qualitäten, sogenannte Bins. Zeigen die Tests, dass der Flash-Speicher vermutlich die Anzahl an erwarteten Zyklen nicht erreichen wird, oder eine so große Anzahl an schlechten Blöcken besitzt, dass er die spezifizierte Kapazität nicht erreicht, führt dies zur Abwertung.

Die Wafer mit den meisten Dies der höchsten Qualitätsstufen verbleiben in der Regel beim NAND-Hersteller für eigene Produkte oder gehen an Produzenten von Industrie- oder Enterprise-Speicherlösungen. Die darunterliegende Qualitätsstufe steckt meist in Consumer-SSDs. Danach folgen Wechselmedien wie (Micro)SD und andere Speicherkarten. Der Rest, oder anders gesagt, der Abfall, landet in USB-Sticks.

Da die NAND-Flash-Hersteller die besten Wafer selbst verarbeiten, verwerfen sie die wenigen schlechten Flash-Dies der Wafer oder verkaufen sie nach dem Sägen des Wafers und dem Absammeln der besten Dies weiter.

Fehlerkorrektur

NAND-Flash ist auf kostengünstige Produktion ausgelegt. Dabei sind nicht sämtliche Speicherzellen funktionstüchtig wie dies beispielsweise bei DRAM der Fall ist; einen hundertprozentig fehlerfreien NAND-Chip gibt es nicht. Das ist jedoch nicht weiter schlimm: Eine Fehlerkorrektur sorgt dafür, dass man die echten Nutzdaten aus den abgespeicherten und bitweise verfälschten Daten im Flash regenerieren kann. Dazu ist jede Page im NAND etwas grösser als für die Nutzdaten nötig. In diese zusätzlichen Bits speichert der Controller die errechneten Redundanzinformationen über die Nutzdaten und Statusbits. Beim Lesen kann er so die defekten Bits aus dem kompletten Datenpaket erkennen und reparieren.

Die Fehlerkorrektur kompensiert Speicherzellen, die bereits seit der Produktion nicht funktionieren, während der Lebensdauer ausfallen oder aufgrund einer fehlerhaften Isolationsschicht ihre gespeicherte Information schnell wieder verlieren. Während bei SLC und Strukturgrößen von über 40 Nanometer noch die Korrekturmöglichkeit von einem Bit pro 512 Byte ausreichte, wurde unter 30 Nanometer bald das Zwölffache nötig.

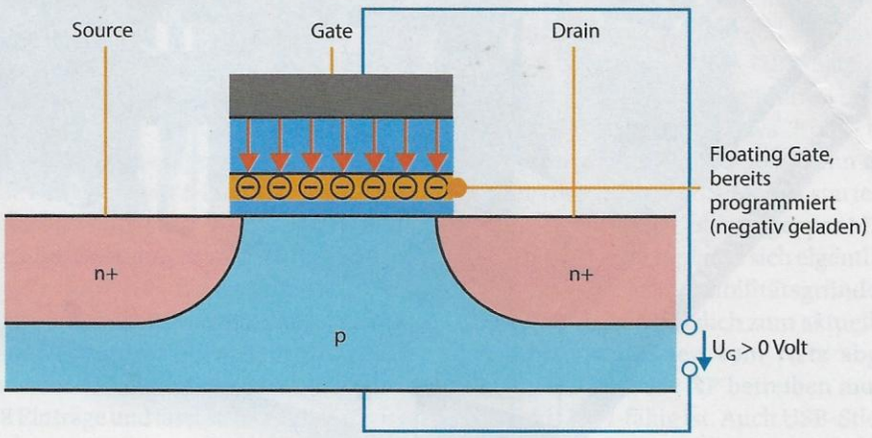
Um bei MLC die nötige Zuverlässigkeit sicherzustellen, muss die Fehlerkorrektur bis zu 40 Bits pro 1 KByte wiederherstellen können. Bei TLC und QLC sind schon mindestens 120 Bits Korrekturfähigkeit pro 1 KByte notwendig. Damit musste auch die Menge an zusätzlichem Platz in der Page steigen. Man kann deshalb nicht rückwirkend eine mächtige Fehlerkorrektur mit einem älteren Flash kombinieren.

Auch der Code der Fehlerkorrektur wechselte über die Zeit. Eine 40-Bit-Fehlerkorrektur vom Typ BCH (Bose Chaudhuri Hocquenghem) ist einfach in Hardware realisierbar. Eine moderne 120-Bit-LDPC-Korrektur (Low Density Parity Code) besteht aus einer Hardware- und einer Softwarekomponente. Bei einer höheren Anzahl von fehlerhaften Bits wechselt sie dann in die deutlich langsamere Software-Unterstützung. Dadurch ist ein LDPC-Code aber erheblich mächtiger als BCH und kann noch Daten korrigieren, die mit BCH verloren wären.

Flash altert vor allem durch die Benutzung. Doch auch stromlos gelagerte SSDs halten ihre Daten nicht über Jahre hinweg. In einem der kommenden Hefte lesen Sie, warum die Daten verloren gehen und was Sie dagegen tun können. *(ll@ct.de)* c't

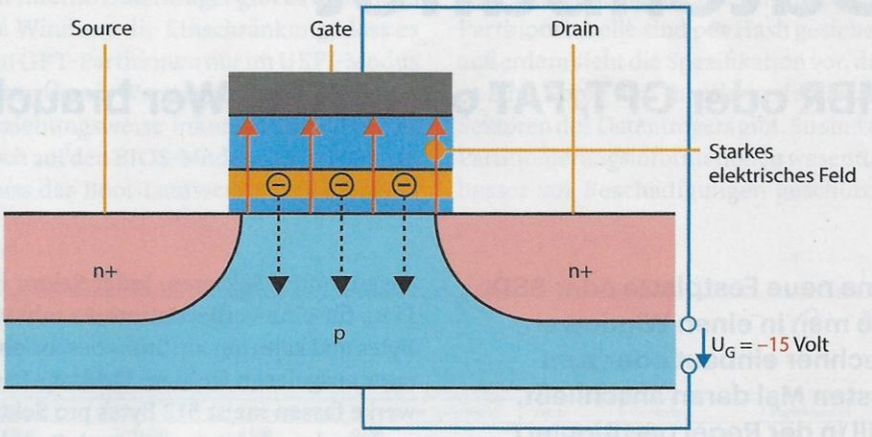
Programmierte Zelle

Die negative Ladung der Elektronen im isolierten Floating Gate einer Zelle kompensiert das elektrische Feld des Gate. Dadurch bildet sich kein Kanal.



Löschen

Der Richtungswechsel der hohen Programmierspannung zwischen Control Gate und Substrat entfernt die Elektronen wieder aus dem Floating Gate, die Zelle wird dadurch gelöscht.



2D-Flash versus 3D-Flash

Flache Flash-Zellen nehmen deutlich mehr Siliziumfläche ein als die moderneren hoch bauenden 3D-Zellen.