**Vorlesung 9**

Folie 2

\*\*\*

Diese Folie zeigt die Klassifizierung von Speicherelementen

Folie 3

\*\*\*

Die Standardstruktur des Speicherblocks wird gezeigt. Eine Speichermatrix besteht aus Speicherzellen. Die Zeilen der Matrix stellen „Worte“ dar. Die Speicherzellen eines Worts sind an eine gemeinsame Readlinie (read line) angeschlossen. Wenn diese Linie geschaltet wird, senden die Speicherzellen ihre Signale (Werte) an die Bitlinien. Ein Adressendekoder setzt die Bitlinien. Wenn z.B. die Adresse = 8 ist, wird die Bitlinie 8 aktiviert. Ein Problem der einfachen Struktur ist es, dass die Höhe der Matrix oft viel größer als ihre Breite ist.

Folie 4

\*\*\*

Dieses Problem kann mit der Struktur in der Folie 4 behoben werden.

Folie 6

\*\*\*

Statische RAM oder SRAM Zellen basieren auf einem rückgekoppelten Buffer. ***Es handelt sich hier um eine positive Rückkopplung.*** Wenn wir auf gleichem Graph die Kennlinie des Buffers und die Kennlinie der Rückkopplung auftragen, erkennen wir, dass es drei Schnittpunkte gibt. Zwei davon entsprechen den stabilen Arbeitspunkten, logisch 1 und 0. SRAM Zelle bleibt in diesen Arbeitspunkten solange die Spanungsversorgung besteht.

Folie 7

\*\*\*

Folie zeigt den Schaltplan einer SRAM Zelle, oben vereinfacht und unten auf Transistorebene. Neben dem rückgekoppelten Buffer, der aus zwei Inverter besteht, sehen wir noch zwei Transistoren T1 und T2. Diese Transistoren verbinden die Knoten der SRAM Zelle mit den Bitlinien. Der Wert wird in der SRAM Zelle „differenziell“ gespeichert, negiert (QN) und nicht-negiert (Q). Es gibt ebenfalls zwei Bitlinien, eine für das negierte Signal (N) und eine für das normale (nicht-negierte) Signal (P).

Folie 8 und 9

\*\*\*

Funktionsweise der SRAM Zelle. Dicke Linie befindet im Zustand 1, dünne Linie im Zustand 0. Logisch 1 am PMOS Gate schaltet den Transistor aus. Logisch 0 am NMOS Gate schaltet den NMOS aus. Die ausgeschalteten Transistoren sind in der Abbildung (Folie 9) ausgeblendet. Die verbliebenen (eingeschalteten) Transistoren halten die Potentiale der Linien fest und verhindern, dass sich die Linien entladen.

Folie 10

SRAM Programmierung: Funktionsweise

Versuchen wir nun die SRAM Zelle zu programmieren. Die Annahme ist, dass ein NMOS Transistor stärker leitet als ein PMOS. Die RAM Zelle befindet sich im Zustand Q = 0. Versuchen wir die Zelle zu überschreiben. Dafür wird die Bitlinie P auf 1 gelegt und die Bitlinie N auf 0. Radlinie wird auf 1 geschaltet.

Folie 11

Sind zwei Biltinien zum Schreiben notwendig?

Ja!

Es wäre schwer nur mit einer Bitlinie 1 ins RAM Zelle zu schreiben. Das ist in Folie 11 illustriert. Transistor T2 leitet zu schwach wegen der kleinen Vgs und kann das Potential Q nicht über die Schwelle (etwa VDD/2) erhöhen. Tn1 hält das Potential niedrig.

Folie 12

Andere Bitlinie ermöglicht das Schreiben. NMOS T1 leitet stärker als PMOS Tp2 und zieht das Potential QN unter die Schwelle.

Folien 13 – 15

\*\*\*

Das führt zum “Umkippen” von RAM Zelle, der Zustand Q = 1/QN = 0 wird hergestellt.

Folie 16

Weitere Folien illustrieren den Lesevorgang.

Folie 17

Machen wir einige Annahmen: Bitlinien haben große Kapazitäten und sind anders aufgeladen als die RAM Zelle selbst.

Folie 18

Readlinie wird auf 1 geschaltet.

Folien 19 - 22

Auch im Fall vom Lesen könnten Probleme auftreten. Es kann z.B. passieren dass die RAM Zelle überschrieben wird. Das wird in den Folien 19 – 22 erklärt.

Folien 23 – 26

\*\*\*

Die Lösung des Problems ist es, die Bitlinien vorm Lesen auf logisch 1 aufzuladen – precharge. Folien 23 – 26 erklären den precharge-Vorgang.

Folie 27

Mithilfe eines Verstärkers (sense amp) kann das Lesezyklus beschleunigt werden.

Folien 28 – 30

Folien zeigen das Layout (die technische Zeichnung der Transistoren von oben betrachtet) einer RAM Zelle. Das Layout wird verkleinert indem man die Strukturen die kurzgeschlossen sind (z.B. der Substratkontakt oder die Knoten Q und QN) im Layout verbindet.

Folien 31 – 37

Folien zeigen die 3D-Darstellung des Layouts einer SRAM Zelle in einer 65nm Technologie. Wir sehen fünf Metallebenen und eine Polysilizium-Lage.

Folie 38 – 39

Folie 38 zeigt das Mikroskop-Foto der SRAM Zellen in einem Prozessor implementiert in einem 28nm Prozess.

Rechts sehen wir das schematische Layout, das dem Foto entspricht.

Folie 39 zeigt das Layout und den Schaltplan als Vergleich.

Folie 41

Dynamische RAM-Zellen (DRAM) basieren auf Speicherung von logischen Niveaus auf einem Kondensator. Wir zeigen zuerst eine DRAM Variante mit 3 Transistoren und einem Kondensator. Trd wird zum Lesen, Twr zum Schreiben und T1 als invertierender Verstärker verwendet. Die Zelle hat getrennte Read- und Writelinien.

Folien 42 - 44

Diese Folien illustrieren den Schreibvorgang. Writelinie (Wr) wird auf 1 geschaltet und der Kondensator aufgeladen – Logisch 1 wird gespeichert.

Folien 45 und 46

Eine dynamische RAM Zelle kann nicht lange ihren Zustand halten. Transistor Twr ist nicht perfekt „dicht“ wenn er ausgeschaltet ist, es einsteht ein Leckstrom und der Kondensator wird innerhalb typischerweise 1ms entladen.

Folien 47 - 50

Diese Folien illustrieren den Lesezyklus.

Linie Rd wird geschaltet (Folie 48). Transistor T1 dient als Verstärker. Seine Gate-Spannung ist hoch, weil der Kondensator aufgeladen ist (die Zelle ist im Eins-Zustand). Deswegen erzeugt der Transistor T1 Strom, der in die Bitlinie fließt. So kann der Zustand der DRAM Zelle gemessen werden. Abwesenheit des Stromes würde bedeuten, dass die Zelle im Null-Zustand ist.

Folie 51

Die Folie zeigt eine noch kleinere Realisierung der DRAM Zelle. Die Zelle enthält nur einen Transistor und einen Kondensator.

Folien 52 - 54

Die Folien zeigen den Schreibzyklus. Logisch 1 wird geschrieben.

Folie 55

Beim Lesen wird der Kondensator der DRAM Zelle stark entladen, da sich die gespeicherte Ladung auf die Kapazität der Bitlinie verteilt.

Folie 56 - 58

Diese Folien zeigen zeitlichen Verlauf von Signalen vC und vBL nach dem Einschalten von WrRd-Linie. Spannung am Kondensator vC sinkt vom VDD (logisch 1) auf ein Wert Vhigh, gegeben durch das Verhältnis von Kapazitäten C und CBL. Vhigh kann unter 100mV liegen. Potential vBL steigt von 0 auf Vhigh (Folie 57).

Ein Verstärker oder ein analoger Komparator wird benutzt um Vhigh als logisch 1 zu erkennen.

Folie 59 - 60

Komparator-Ausgang kann an die Bit-Linie zurück gekoppelt werden um den richtigen Spannungs-Niveau wieder herzustellen. Man nennt das refresh.

Folie 61

Die Folie zeigt wie eine DRAM Zelle typischerweise im Silizium implementiert wird. Es wird eine tiefe Elektrode (through silicon via) als Kondensator verwendet. Auf diese Weise nimmt die DRAM Zelle wenig Fläche. Ihr Kondensator hat große Kapazität und wird langsamer entladen und einfacher gelesen (da Vhigh größer ist).

Folien 62 – 64

\*\*\*

zeigen verschiedene DRAM Strukturen.

Folie 66 – 70

SRAM und DRAM Zellen behalten die gespeicherte Information nur solange Spannungsversorgung erhalten bleibt. Oft brauchen wir aber auch Speicherkomponente die Information permanent speichern. Bespiele sind Speicherkarten, Festlatten, usw.

Die einfachste Variante von einem Permanentspeicher ist PROM – programmable read only memory. Dieser Speicher kann nur einmal geschrieben, bzw. programmiert werden.

PROM basiert auf Metallverbindungen (fuses) die durch gezielte Anwendung von hoher Spannung verdampft werden. Solche Programmierung ist auf Folien 67 – 70 gezeigt. Widerstände R (pull down R) sind deutlich höher als die fuse-Widerstände.

Folie 71 – 74

Diese Folien illustrieren den Lesezyklus. Dioden verhindern dass die Bitlinien mit hohem Potential (b1 und b2) die Bitlinie mit niedrigem Potential beeinflussen. Widerstände R (pull down R) sind deutlich höher als die fuse-Widerstände.

Folie 76 - 78

Eine neuere Speichertechnolgie ist EEPROM (electrically erasable and prorgammable read only memory). Die EEPROM Zellen können elektrisch programmiert- und gelöscht werden. Eigentlich ist der Name „read only memory“ in dem Fall nicht ganz richtig, da EEPROM auch geschrieben werden kann. Das Schreiben ist aber nur mit höheren Spannungen möglich, nicht so schnell wie das Lesen und nur für eine begrenzte Zahl von Zyklen möglich. Wegen diesen Einschränkungen beim Schreiben kann man EEPROM read only memory nennen.

EEPROM basiert auf Transistoren mit variabler Schwellespannung als Speicherelementen. Folien 76 – 79 zeigen wie ein EEPROM gelesen wird. Transistoren mit „dickem“ Gate haben höhere Schwelle. Read-Linie wird auf 1 geschaltet – Folie 77. Transistoren mit niedrigerer Schwelle leiten und erzeugen logisch 0 auf den Bitlinien. Ein leitender Transistor hat niedrigeren Widerstand als der pullup-Widerstand R. Der Transistor mit hoher Schwelle leitet nicht. Der pullup-Widerstand R erzeugt 1 auf der Bitlinie.

Folie 79

Wie kann die Schwelle des Transistors verändert werden? Als Erinnerung: Transistorschwelle ist die Gate-Source Spannung, die benötigt wird um einen leitenden Kanal zwischen den Source und Drain zu erzeugen. Die EEPROM Transistoren haben ein zweites Gate (Floating-Gate) zwischen dem Haupt-Gate und dem Silizium-Substrat. Floating-Gate ist an keine Metallleitung angeschlossen. Trotzdem kann es aufgeladen werden. Wenn das Floating-Gate negativ aufgeladen ist, entsteht im Silizium unterhalb des Gates (im Kanalbereich) gleichwertige positive Ladung Q. Es entsteht auch zwischen dem Floating-Gate und dem Kanalbereich eine negative Spannung (plus-Punkt am Floating-Gate) VFG = Q/CFG. Die positive Ladung vergrößert die Schwelle im Falle eines NMOS Transistors da man eine zusätzliche Spannung am Gate VG = Q/Cox braucht um die positive Ladung zu entfernen. Die negative Spannung am Floating-Gate hat einen entgegengesetzten Effekt, aber da die Kapazität CFB größer als Cox ist, kann man diesen Effekt vernachlässigen.

Folie 80

Wie kann man Floating-Gate aufladen? Eine Möglichkeit basiert auf dem quantenmechanischen Tunneleffekt. Die Folie zeigt EEPROM-Struktur, einen Transistor mit Floating-Gate. Für einen Ausschnitt aus der Struktur zeigen wir das Energiediagramm. Das Energiediagramm links ist für Spannung Vbias = 0, und rechts für Vbias etwa 10V.

Im Fall ohne Biasspannung ist die Potentialbarriere für Elektronen zwischen dem Drain und dem Floating-Gate zu breit und Elektronen können durch die Barriere nicht „tunneln“. Sie können die Barriere auch nicht überwinden da sie zu wenig Energie haben.

Wenn man die Biasspannung einschaltet, verschieben sich die Energieniveaus wie im Diagramm rechts gezeigt wurde. Dadurch wird die Barriere für Elektronen aus dem Drain schmäler und sie können durch die Barriere in das Floating-Gate gelangen, obwohl sie kleinere kinetische Energie als die hohe der Barriere besitzen. Die Elektronen erzeugen dann eine negative Ladung im Floating-Gate indem sie mit Löchern rekombinieren.

Folie 81

Man kann den Tunneleffekt auch zunutze machen, um die negative Ladung aus dem Floating-Gate zu entfernen. Dafür würde man die Polarität der Biasspannung umdrehen, bzw. das Drain auf eine positive Spannung legen. In dem Fall können die Elektronen aus dem Valenzband des Floating-Gates in das Drain tunneln. Damit werden Löcher erzeugt, und die negative Ladung wird entfernt.

Folie 82

Man kann auch die so genannte „hot carrier injection“ benutzen um das Floating-gate negativ aufzuladen. Dabei ist es wichtig eine hohe Stromdichte im Kanalbereich zu erreichen. Transistor wird mit Vgs > Schwelle eingeschaltet. Vds-Spannung erzeugt den Strom. Durch die hohe Stromdichte im Kanalbereich nahe Drain, bekommen die Elektronen hohe kinetische Energie und überwinden die Barriere.

Folien 83 und 84

Folien 83 und 84 geben mehr Details über den Tunneleffekt.

Folie 85

Je nachdem wie Ladung im Floating-Gate erzegt oder gelöscht wird unterscheidet man zwischen einigen EEPROM Varianten.

Folien 87 - 90

Erste EEPROM Variante wird mittels Tunneleffekt gelöscht. Gates werden auf hohe Spannung (~10V) gelegt. Elektronen tunneln aus den Source-Bereichen in das Floating-Gate. Auf diese Weise werden alle Floating-Gates gleichzeitig negativ geladen. Die Schalter-Transistoren werden beim Löschen nicht benutzt.

Dünne Linien befinden sich auf niedrigen Potential.

Folien 91 - 97

Die Folien zeigen das Programmieren vom EEPROM. Es basiert ebenfalls auf dem Tunneleffekt. Beim Programmieren ist es wichtig, dass man einzelne Transistoren individuell programmieren kann, bzw. dass man die Ladungen an ihren Floating-Gates entfernen kann. Die Ladung wird nur dann entfernt, wenn die EEPROM Zelle hohe Spannung an Programmlinie und an Bitlinie hat. Das ist der Fall für die eingekreisten Zellen. Die Elektronen tunneln vom Floating-Gate in das Drain.

Folien 99 – 101

Zweite EEPROM Variante verwendet hot carrier injection fürs Programmieren. Da mit hot carrier injection negative Ladung am Floating-Gate entsteht, muss beim Löschvorgang die Ladung entfernt werden.

Folien 99 – 101 zeigen das Löschen. Es basiert auf dem Tunneleffekt, auf ähnliche Weise wie das Programmieren bei der Variante 1. Da die Zellen einen Schaltertransistor haben, könnte man das Löschen auch Zellenweise (selektiv) machen.

Folien 102 – 105

Diese Folien zeigen wie man eine Zelle (eingekreist) mit hot carrier injection programmieren kann, bzw. wie man ihr Floating-Gate aufladen kann.

Folien 107 – 114

\*\*\*

Dritte Variante funktioniert auf dem gleichen Prinzip wie die zweite, nur sie benutzt einfachere Zellen ohne den Schaltertransistor und ohne Programmlinie.

Folie 115

\*\*\*

Ein FLASH Speicher basiert auf dem gleichen Prinzip wie EEPROM. Der Hauptunterschied ist die Logik-Art.

EEPROMs verwenden Transistoren die NOR bilden – sie sind in parallel geschaltet. FLASH Speicher verwenden NAND Gates, die aus Transistoren mit Floating-Gate aufgebaut sind. Dadurch lassen sich die Zellen kleiner machen.

Folgender Text fasst die Eigenschaften von FLASH Speicher gut zusammen:

*Flash is a very popular term when it comes to storage media as it is used by portable devices like phones, tablets, and media players. Flash actually is an offspring of EEPROM, which stands for Electrically Erasable Programmable Read-Only Memory. The main difference between EEPROM and Flash is the type of logic gates that they use. While EEPROM uses the faster NOR (a combination of Not and OR), Flash uses the slower NAND (Not and AND) type. The NOR type is a lot faster than the NAND type but there is the matter of affordability as the former is significantly more expensive than the NAND type.*

*Another advantage of EEPROM over Flash is in how you can access and erase the stored data. EEPROM can access and erase the data byte-wise or a byte at a time. In comparison, Flash can only do so block-wise. In order to simplify the whole thing, individual bytes are grouped into a smaller number of blocks, which can have thousands of bytes in each block. This is a bit problematic when you only want to read or write to a single byte at a time; which is what’s typically needed in executing the code of a program. This is a reason why Flash cannot be used in electronic circuits that require byte-wise access to data. Data in Flash can also be executed, but it needs to be read as a whole and loaded into RAM beforehand.*

*EEPROM was designed to be read a lot more than it is written. This is in-line with programming for electronic circuits where you write to the chip a number of times while testing the program. Then, it is stored for good, only to be read every time the data is needed. This is not very suitable for storage media where data is routinely written and read.*

*In typical use, Flash is used mainly to refer to storage media and can range anywhere from a GB to hundreds of GB. In contrast, EEPROM is usually reserved for permanent code storage in electronic chips. Typical values range from kilobytes to a couple of megabytes*

*Summary:*

*1.Flash is just one type of EEPROM*

*2.Flash uses NAND type memory while EEPROM uses NOR type*

*3.Flash is block-wise erasable while EEPROM is byte-wise erasable*

*4.Flash is constantly rewritten while other EEPROMs are seldom rewritten*

*5.Flash is when large amounts are needed while EEPROM is used when only small amounts are needed*

 Ebenfalls folgender Text bringt viele interessante Infos:

*Ein Flash-Speicher besteht aus einer bestimmten, von der Speichergröße abhängigen Anzahl einzelner Speicherelemente. Die Bytes oder Worte (typisch durchaus bis 64 Bit) können einzeln adressiert werden. Dabei können sie in einigen Architekturen auch einzeln geschrieben werden, wogegen bei anderen nur größere Datenmengen auf einmal programmiert werden können. In der Regel ist die entgegengesetzte Operation, das Löschen, aber nur in größeren Einheiten, sogenannten Sektoren (meistens ein Viertel, Achtel, Sechzehntel usw. der Gesamtspeicherkapazität) möglich.*

*Flash-Speicher haben eine begrenzte Lebensdauer, die in einer maximalen Anzahl an Löschzyklen angegeben wird (10.000 bis 100.000 Zyklen für NOR-Flash und bis zu zwei Millionen für NAND-Flash). Dies entspricht gleichzeitig der maximalen Anzahl Schreibzyklen, da der Speicher jeweils blockweise gelöscht werden muss, bevor er wieder beschrieben werden kann. Diese Zyklenzahl wird Endurance (Beständigkeit) genannt. Verantwortlich für diese begrenzte Lebensdauer ist das Auftreten von Schäden in der Oxidschicht im Bereich des Floating-Gates, was das Abfließen der Ladung bewirkt.[3]*

*Eine andere wichtige Kenngröße ist die Zeit der fehlerfreien Datenhaltung, die Retention.*

*Ein weiterer Nachteil ist, dass der Schreibzugriff bei Flash-Speicher erheblich langsamer erfolgt als der Lesezugriff. Zusätzliche Verzögerungen können dadurch entstehen, dass immer nur ganze Blöcke gelöscht werden können.*

*Der Flash-Speicher speichert seine Informationen auf dem Floating-Gate. Bei einem Löschzyklus durchtunneln die Elektronen die Oxidschicht. Dafür sind hohe Spannungen erforderlich. Dadurch wird bei jedem Löschvorgang die Oxidschicht, die das Floating-Gate umgibt, ein klein wenig beschädigt (Degeneration). Irgendwann ist die Isolation durch die Oxidschicht nicht mehr gegeben, die Elektronen bleiben nicht mehr auf dem Floating-Gate gefangen, und die auf der Speicherzelle gespeicherte Information geht verloren. Der Defekt einer einzelnen Zelle macht einen Flash-Speicher jedoch noch lange nicht unbrauchbar.*

*Als nichtflüchtiges Speichermedium steht der Flash-Speicher in Konkurrenz vor allem zu Festplatten und optischen Speichern wie DVDs und Blu-ray-Discs.*

*Ein wesentlicher Vorteil liegt in der mechanischen Robustheit von Flash-Speicher. Demgegenüber sind Festplatten sehr stoßempfindlich (Head-Crash). Häufig ist die Lebensdauer der Steckkontakte (USB-Stecker) der limitierende Faktor.*

Folie 116 – 118

Die Folien zeigen die FLASH Struktur (acht 2-bit Worte) und wie die eingekreisten Zellen gelesen werden können.

Readlinie 2 wird auf ein Niveau zwischen den Schwelle-Spannungen von Transistoren mit geladenen und nicht geladenen Floating-Gates gelegt. Alle anderen Readlinien werden auf höheres Potential gelegt (z.B. 5V), so dass die entsprechenden Transistoren immer leiten (unabhängig davon ob ihre Gates aufgeladen sind). Auf diese Weise lässt sich messen, ob die Transistoren (angeschlossen an Readlinie 2) leiten oder nicht, bzw. welches Bit sie speichern.

Folien 119 – 122

\*\*\*

Diese Folien erklären den Löschvorgang. Tunneleffekt wird benutzt. Beim Löschen wird positive Spannung von etwa 10V an Transistorsubstrat gelegt. Elektronen tunneln vom Flaoting-Gate weg. Das Löschen kann auch selektiv gemacht werden, indem man eine leicht höhere Spannung an Readlinie legt (siehe readlinie 1). Die Transistoren in der Zeile 1 werden nicht gelöscht da die Spannung zwischen Substrat und Gate nicht ausreichend ist.

Folien 123 – 126

Diese Folien zeigen den Schreibvorgang. Auch hier wird der Tunneleffekt verwendet. Readlinie 1 wird auf etwa 10V gelegt und die anderen Readlinien auf 5V. Der eingekreiste Transistor T2 sieht eine hohe Spannung am Gate und 0V am Drain und sein Floating-Gate wird aufgeladen. Transistor T1 sieht zwar ein hohes Potential am Gate aber da sein Drain-Potential höher als 0V liegt, ist die Gate-Drain Spannung unzureichend fürs Tunneln.

Folien 128 - 149

Diese Folien sind nur als Info gedacht. Sie zeigen wie man eine RAM Struktur erweitern kann, so dass sie Information permanent speichert. Es werden Kondensatoren mit ferroelektrischem Dielektrikum verwendet. *Ferroelektrische Materialien können analog zu ferromagnetischen Materialien eine permanente elektrische Polarisation auch ohne externes elektrisches Feld besitzen. Durch ein externes Feld kann diese Polarisation in eine andere Richtung „umgeschaltet“ werden.*