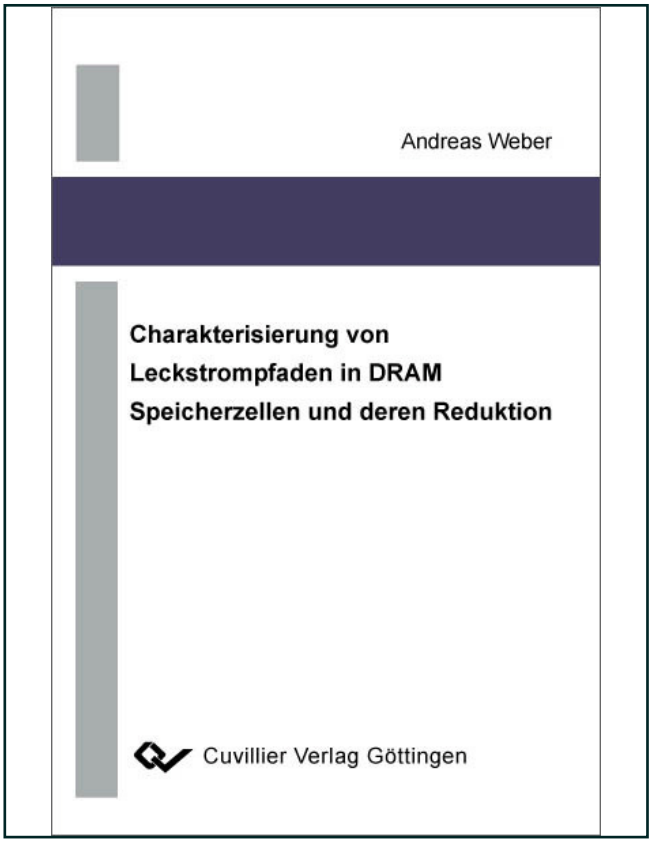
Andreas Weber (Autor)

Charakterisierung von Leckstrompfaden in DRAM Speicherzellen und deren Reduktion



https://cuvillier.de/de/shop/publications/1857

Kapitel 1

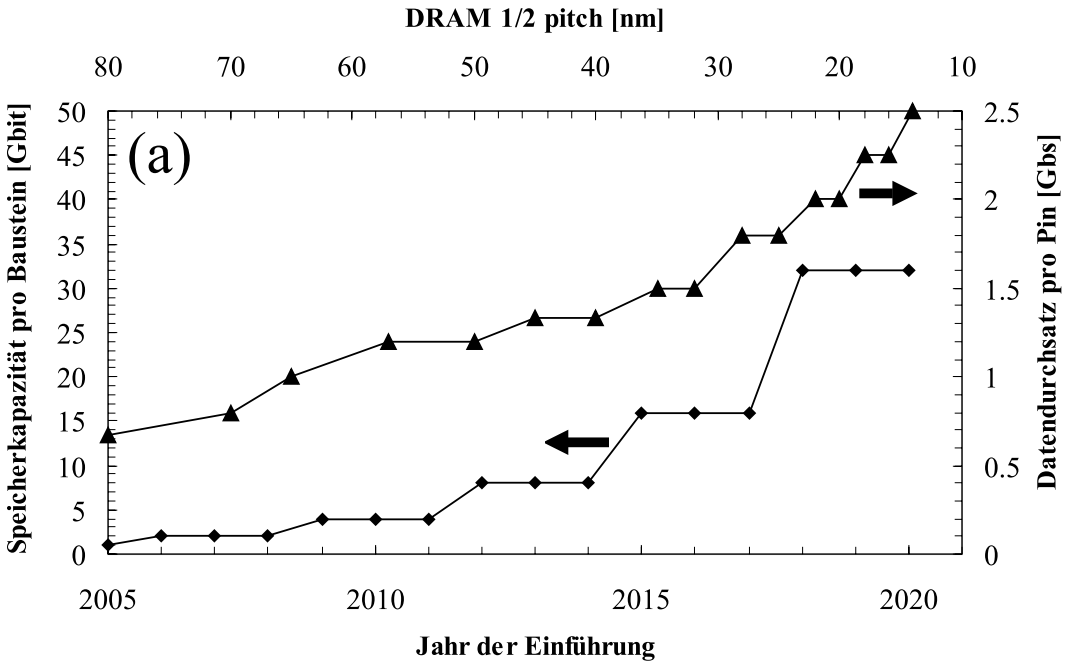
Einleitung

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

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

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

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



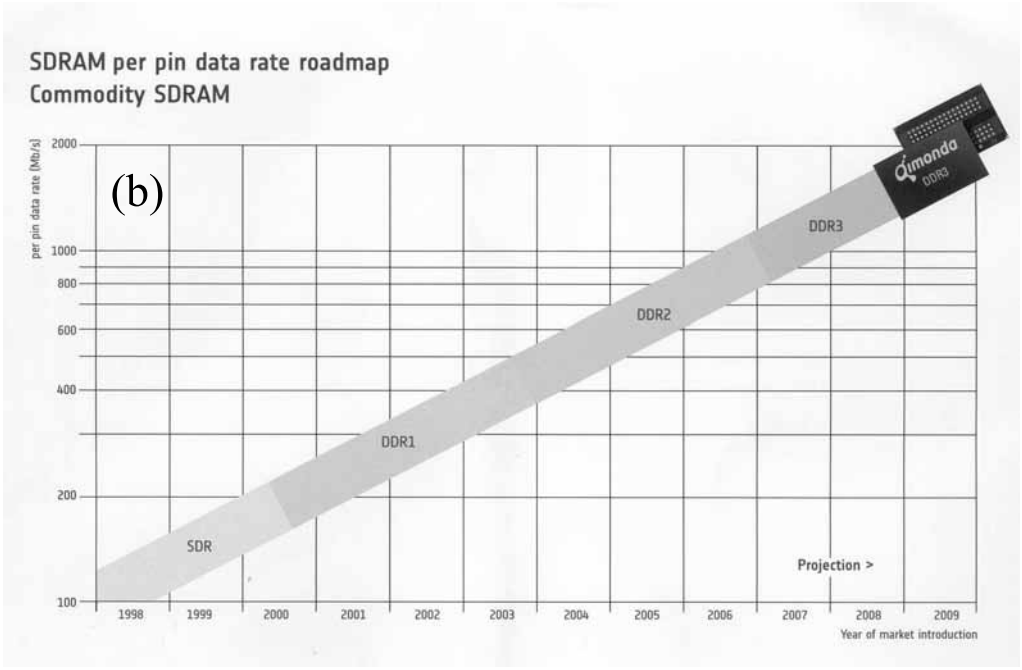


Abbildung 1.1: (a)Vorhergesagte Entwicklung der DRAM-Kapazität pro Chip und des Datendurchsatzes [ITR05]. (b) DRAM-Standards der letzten Jahre.

Gliederung der Arbeit

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

Die Arbeit ist folgendermaßen aufgebaut:

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

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

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

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

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

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

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

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

Kapitel 2

DRAM Grundlagen

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

2.1 Zellaufbau

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

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

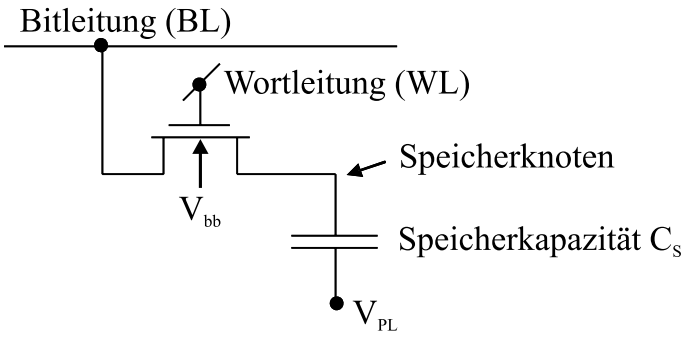


Abbildung 2.1: Ersatzschaltbild einer DRAM Speicherzelle. Die Zelle besteht nur aus zwei Bauelementen: einem Kondensator und einem Transistor. Die Information wird als Ladungszustand auf dem Kondensator gespeichert. Durch Aktivieren der Wortleitung, die mit dem *Gate* des Transistors verbunden ist, wird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

Realisierung in Silizium

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

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

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

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

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

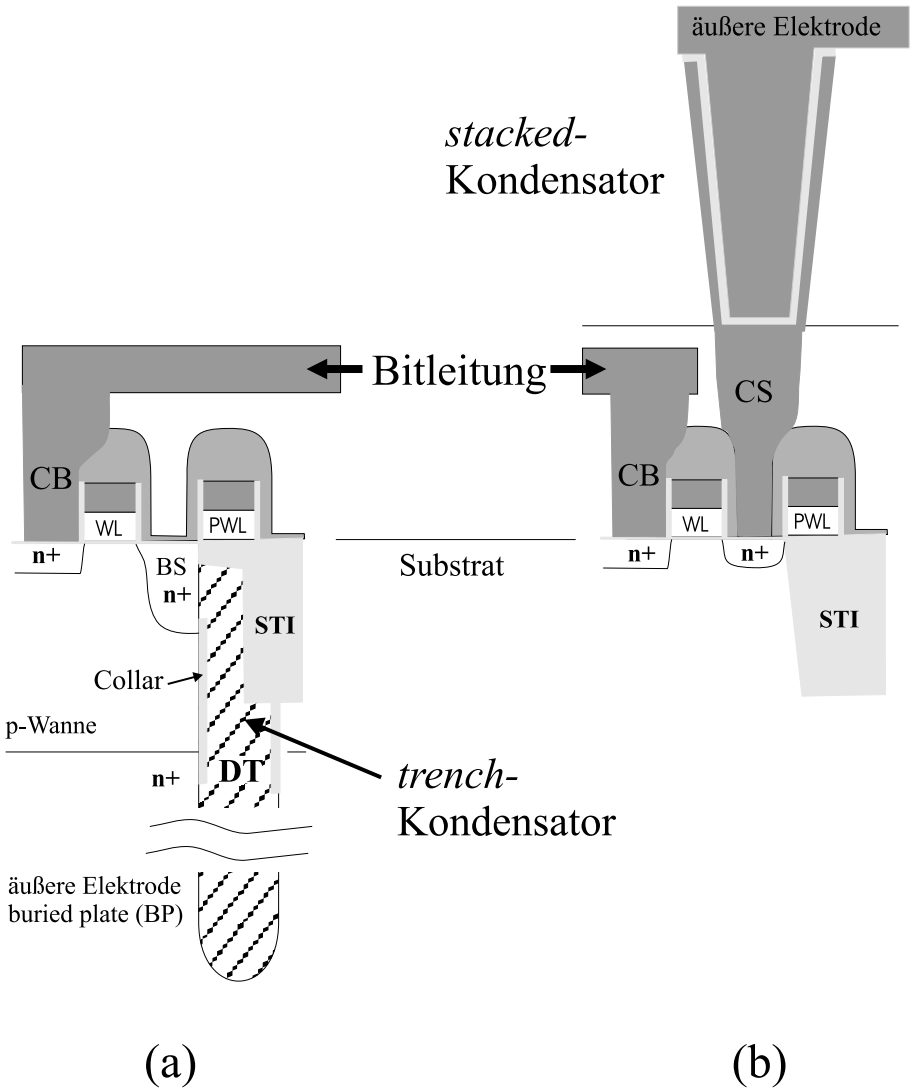


Abbildung 2.2: Vergleich der beiden Hauptintegrationsvarianten im Querschnitt. (a) Beim *trench*-DRAM wird der Speicherkondensator in einem tiefen Graben (DT) ausgebildet, während er bei (b) *stacked*-DRAM oberhalb der Transistoren entsteht.