16.4 Prozeßintegration und Analytik

16.4.1 Technologierelevante Randbedingungen

Die Entwicklung einer neuen Speichergeneration erfolgt unter einer Reihe von Randbedingungen die unbedingt eingehalten werden müssen. Den größten Effekt hat die Festlegung des künftigen Gehäuses; in der Regel erfolgt dies 3-4 Jahre vor Markteintritt. Eine de jure amerikanische, de facto aber internationale Organisation namens JEDEC (Joint Electron Device Engineering Council) normt Gehäuse, Belegung der Pins (= Beinchen) sowie einige elektrische Daten.

Aus den Gehäuseabmessungen läßt sich die Chipgröße ableiten; aus der Chipgröße folgt die Größe der Speicherzelle. Die Speicherzelle enthält einen Transistor, einen Kondensator, die Bit- und Wortleitung sowie die Isolation zur Nachbarzelle. Als Aufgabe bleibt, diese Elemente so günstig als möglich anzuordnen, um zu möglichst entspannten "Designregeln" zu kommen. Die Designregeln legen die minimalen Strukturmaße und die zugehörigen Toleranzen fest. Mit der Definition der Designregeln ist gleichzeitig vorgegeben, was die Lithographie leisten muß; darüberhinaus lassen sich Anforderungen an die Maßhaltigkeit der Ätztechnik festlegen. Auch die Meßtechnik weiß damit, mit welcher Genauigkeit gemessen werden muß, denn nur Dimensionen, die auch gemessen werden können, lassen sich sinnvoll spezifizieren. Fig. 14 zeigt diese Wirkungskette am Beispiel des 16M DRAMs.

Aus der Festlegung der Versorgungs- und Arbeitsspannungen ergeben sich weitere Randbedingungen. Auch beim 16M DRAM wird man extern nicht von der 5V Norm abgehen; intern arbeiten einige Bereiche jedoch nur mit 3,3V. Zusammen mit der Minimalzahl von Ladungsträgern ($< 10^6$) die unbedingt noch benötigt werden, um bei der Umladung des Kondensators (d. h. beim Lesen und Schreiben) noch sicher erkennbare Signale zu erhalten, ergibt sich die minimale Kapazität des Speicherkondensators. Dieser Wert liegt ziemlich unabhängig von der jeweiligen Speichergeneration bei 35 fF - 50 fF (f = femto = 10^{-15}); Aufgabe am Rande: Wie mißt man diese Kapazität?

Die Kapazität C des Kondensators ist gegeben durch seine Fläche F_c sowie Dicke d und Dielektrizitätskonstante ε des Dielektrikums; es gilt:

$$C = \frac{\varepsilon \cdot \varepsilon_0 \cdot F_c}{d} \quad .$$

Nimmt man an, daß 30-40 % der Zellfläche des 16M DRAMs für den Kondensator zur Verfügung steht (d. h. F_c ca. 2µm²; vgl. Fig. 14) und SiO_2 als Dielektrikum dient ($\varepsilon = 3{,}8$); ergibt sich die Dicke des Dielektrikums für das 16M DRAM zu ca. 2 nm (also ca. 6 Atomlagen); die Feldstärke wäre ca. 16 MV/cm. Beim 4M DRAM wären die entsprechenden Zahlen d = 4 nm, $\mathcal{E}$ = 10 MV/cm.
Diese Dicken und Feldstärken sind mit keinem Material mehr beherrschbar; die Lösung liegt beim Vergrößern der Fläche durch Einbeziehen der dritten Dimension.

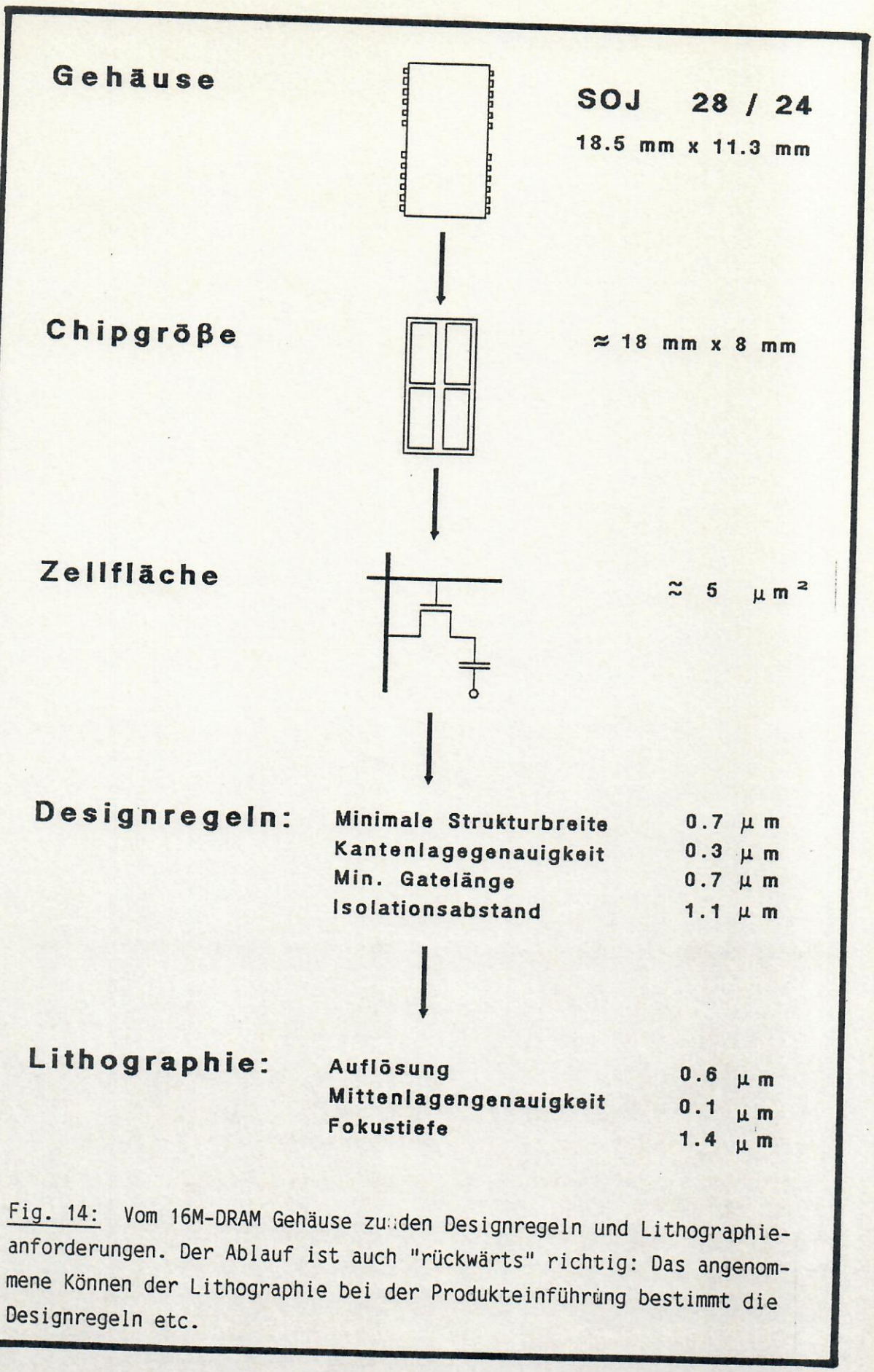


Fig. 14: Vom 16M-DRAM Gehäuse zu den Designregeln und Lithographieanforderungen. Der Ablauf ist auch "rückwärts" richtig: Das angenommene Können der Lithographie bei der Produkteinführung bestimmt die Designregeln etc.

So wie sich die absolut gesehen kleine Spannung von 5V bei Einbeziehung der Dimensionen eines Chips in extrem hohe Werte der elektrischen Feldstärke umsetzt (nicht nur im Kondensator-dielektrikum sondern auch in den Transistoren), transformieren sich die an sich kleinen Ströme im mA-Bereich zu extrem hohen Stromdichten in den Leiterbahnen von $10^5 A/cm^2$ - $10^6 A/cm^2$. (Zum Vergleich: Typische Haushaltgeräte liegen bei 10-100 A/cm^2).

Ähnliches gilt für die Verlustleistung. Sie liegt zwar im < 1 Watt Bereich, aber auf einer Fläche von ca. 1 cm^2. Eine Herdplatte liegt bei ca. 2 W/cm^2.

Da die Absolutwerte von Spannung und Strom von Generation zu Generation kaum sinken, würden sich die spezifischen Größen wie Feldstärke, oder Strom- und Leistungsdichte in jeder neuen Generation kräftig erhöhen. Die letzten Chips der Vor-Mega-Periode kamen jedoch den Grenzen der Physik schon ziemlich nahe, die Entwicklung der Megachips läuft daher ganz allgemein unter der Randbedingung Feldstärken, Strom- und Leistungsdichten trotz immer kleinerer Abmessungen halbwegs konstant zu halten.

16.4.2 Gesamtprozeß 1M-DRAM

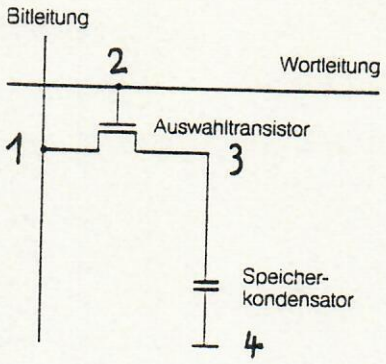
Das 1M-DRAM verkörpert die letzte Speichergeneration mit planarem Kondensator. Fig. 15b zeigt einen Querschnitt. Bei einer Chipfläche um 55mm^2, ist die Zellfläche ca. 40 μm^2; der Kondensator beansprucht davon 15 μm^2. Mit einer Dicke des Dielektrikums von ca. 10nm wird eine Kapazität von 50 fF gerade noch erreicht. Falls die volle Spannung von 5V am Kondensator anliegt, ist die Feldstärke im SiO_2-Dielektrikum ca. 5MV/cm; das ist gerade noch ein Faktor 2 unterhalb der "Tunnelgrenze". Denn selbst im idealen, 100% perfekten Dielektrikum tritt bei ca. 10V Stromfluß durch quantenmechanisches Tunneln der Ladungsträger auf. Ist das Dielektrikum nicht perfekt (weil z.B. an der Grenzfläche zum Si eine winzige Cu-Ausscheidung von ca. 3 nm sitzt, oder weil ein Partikel von 2 nm Durchmesser während der Oxidation auf der Scheibe lag), wird das Dielektrikum schon bei kleineren Spannungen durchlässig für Strom. Entlädt sich die gesamte Ladung des Kondensators von 10^6 Elektronen in ca. 10 nsec. über eine Schwachstelle von 1nm Durchmesser, ist die Stromdichte im Mittel

$$j = Q \cdot t = 1{,}6 \cdot 10^9 \; A/cm^2.$$

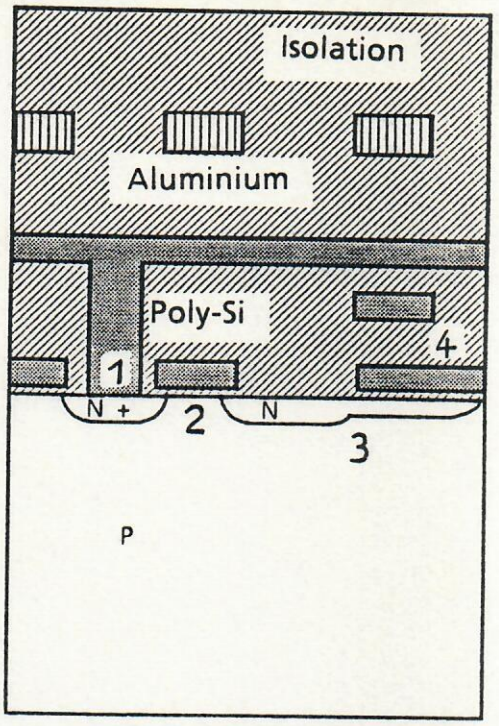
Auch falls die Entladung länger dauert, oder das "Loch" größer ist, ist diese Stromdichte und die damit verbundene Energiefreigabe jenseits von Gut und Böse; das Dielektrikum "schlägt durch" und wird unwiderruflich zerstört.

Beim 1M DRAM wurde die praktische Grenze der minimalen Dielektrikumsdicke erreicht; die Folgegenerationen mußten deshalb an dieser Stelle neuartige Strukturen einführen.

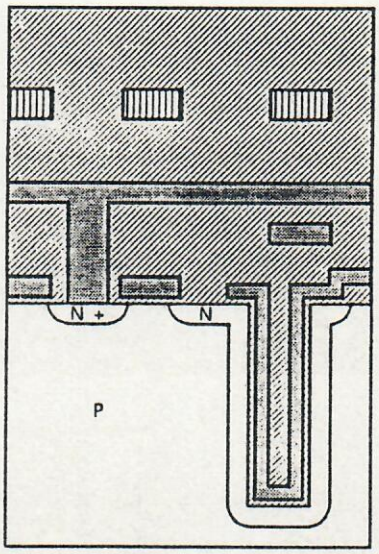
In Fig. 15a wird noch einmal der "Schaltplan" der Speicherzelle gezeigt. In der Querschnittszeichnung in Fig. 15b findet man die Teile dieses Schaltplans wieder, darüber hinaus einige Besonderheiten.



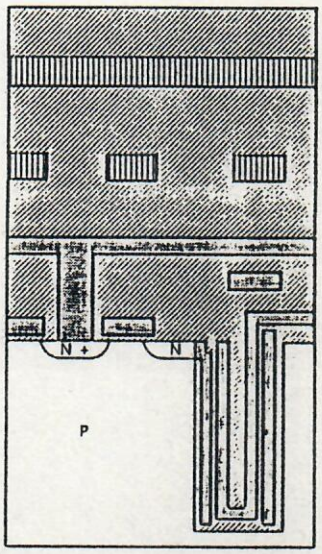
a) "Schaltbild" (vgl. Fig. 9)



b) 1M-DRAM



c) 4M-DRAM



d) 16M-DRAM

Fig. 15: Stark vereinfachte, schematische Querschnitte durch Kondensator und Auswahltransistor einer Speicherzelle. Die Größenverhältnisse sind lateral ungefähr richtig.

Was in Schaltplänen der konventionellen Elektrotechnik überhaupt nicht auftaucht, ist die elektrische Isolation zwischen den einzelnen Komponenten. Dies ist deshalb nicht der Fall, weil Einzelkomponenten (Drähte, Widerstände, Kondensatoren, Transistoren, ...) entweder sowieso isoliert sind oder weil die Komponente "Luft", ohne daß man sie extra einbauen muß, in der Regel für ausreichende Isolation sorgt.

In der integrierten Schaltungstechnik wird einem die Isolation leider nicht geschenkt. Der Trägerkörper Silicium isoliert Komponenten nicht ausreichend voneinander; eine elektrische Isolation ist deshalb ein platz- und prozeßschrittbeanspruchender Teil eines jeden Elements einer integrierten Schaltung. In Fig. 15b sind deshalb 5 der gezeigten Schichten der Isolation gewidmet; das sogenannte "LOCOS" Oxid beansprucht dabei einen beträchtlichen Teil der Zellflächen.

Erkennbar ist weiterhin der Kondensator mit seinen beiden Anschlüssen: Einmal ein n-dotiertes Gebiet im Si-Substrat (entsprechend dem Knoten 3 in Fig. 15a) und dem (nicht notwendigerweise auf Masse (= 0V)) liegenden Anschluß 4 in Fig. 15a. Punkt 4 ist allen 1 048 576 Kondensatoren gemeinsam; er besteht daher aus einer sich über das gesamte Zellenfeld erstreckenden, und an den richtigen Stellen "durchlöcherten" Poly-Silicium Platte. Der Anschluß des Kondensators an Source des Transistor erfolgt direkt; die beiden n-dotierten Gebiete überlappen sich. Das Gate des Transistors wird durch die Wortleitung aus poly-Si gebildet; sie berührt über der Gatefläche direkt das Gate-Dielektrikum; außerhalb der Gatefläche läuft sie über dickere Isolierschichten. In Fig. 15b ist z.B. über dem Kondensator die Wortleitung der beiden nach rechts versetzten, ober- und unterhalb der Papierebenen liegenden Transistoren zu erkennen.

Die senkrecht zur Wortleitung (in der Papierebene) verlaufende Bitleitung liegt über einer weiteren Isolierschicht. Der Knotenpunkt 1 wird gebildet, indem die isolierenden Schichten über dem Drain des Transistors durchgeätzt werden, so daß die Bitleitung hier bis auf das Silizium-Substrat herunterführt und den Transistor kontaktiert. Die Bitleitung besteht aus einem Sandwich aus poly-Si und $MoSi_2$-Silizid.

Die gesamte Länge der Wort oder Bitleitung beträgt ca. 25 m! Der Serienwiderstand R und die Streukapazität C dieser Leitungen sind deshalb außerordentlich wichtige Größen für die Schaltungstechnik des Speichers. Das Produkt R C bestimmt die Zeitkonstante und damit die Geschwindigkeit mit der ein- und ausgelesen werden kann. Die Streukapazität der Bitleitung ist besonders wichtig. Sie ist viel größer als die Kapazität des Speicherkondensators. Das hat zur Folge, daß beim Auslesen der Information (also dem Entladen der Speicherkondensators) die am Speicherkondensator anliegende Spannung im Verhältnis der Speicherkapazität zur Bitleitungskapazität geteilt wird; am Ausgang der Bitleitung, also am Leseverstärker, stehen deshalb nur einige 100mV zur Verfügung (vgl. 16.2.1).

Serienwiderstand und Kapazität der Bit- und Wortleitung muß also möglichst klein gehalten werden. Dabei sind viele Parameter zu optimieren, die teilweise gegenläufig sind. Eine kleine Kapazität kann z.B. durch eine möglichst kleine Querschnittsfläche der Leitungen und möglichst dicke isolierende Schichten erreicht werden - dies erhöht dann automatisch den Widerstand, beziehungsweise verschärft die Topologie.

Die Verwendung von Materialien mit kleinem spezifischem Widerstand (z.B. Al) für die Wortleitung ist nicht möglich, da nach Fertigstellung der Wortleitung Prozesse bei hoher Temperatur (> 900°C) durchgeführt werden müssen, Aluminium aber maximal mit 450°C belastet werden kann. Für die Bitleitung ist die Lage ähnlich, aber etwas entspannter. Ein möglicher Kompromiß besteht darin, die Wortleitung doppelt anzuführen: Einmal in poly-Silicium in den unteren Schichten und ganz zum Schluß nocheinmal mit Al.

Die Al-Leitung wird dabei nach ca. 100 Transistoren mit der poly-Si Leitung kurzgeschlossen (in Fig. 15b) nicht gezeigt).

Die Bitleitung wird als Poly-Silicium - Silizid (z.B. $MoSi_2$, $TaSi_2$, $TiSi_2$) Sandwich ausgeführt. Dies hat den Vorteil, daß die untere Poly-Si Schicht sich an den Kontaktstellen so verhält, wie man es von früher her (als es noch keine Silizide gab) kannte; die Silizid-Schicht den Schichtwiderstand aber um Größenordnungen senkt.

Als letzte Schicht kommt wieder eine Isolationsschicht; nämlich die Passivierung. Sie sorgt weniger für elektrische Isolation, sondern für die Isolation des gesamten Aufbaus von der "Umwelt", insbesondere Feuchtigkeit. Etwas Wasserdampf mit Spuren von Chlor (z.B. aus NaCl) reicht, um die Al-Leitungen in kürzester Zeit zu koordinieren. Elektrische Isolation zwischen den Al-Leitungen wäre eigentlich nicht erforderlich, jedoch ist die Feldstärke von 5V/1µm = 50.000 V/cm im Hinblick auf Kriechströme etc. nicht ganz zu vernachlässigen.

Eine weitere wichtige Eigenschaft des 1M-DRAMs wird aus Fig. 15b nicht mehr ersichtlich. Es ist dies die CMOS (= Complementary Metal Oxide Silicon) Technologie in der Peripherie als Innovation gegenüber zu den Vorgängergenerationen. Die CMOS-Technik erlaubt die Herstellung von sowohl n-MOS Transistoren (Kanal n-leitend; Substrat p-Typ) als auch von p-MOS Transistoren (Kanal p-leitend; Substrat n-Typ) auf einem Si-Substrat. Ist das Substrat, also der Wafer, p-leitend; müssen in einem ersten Schritt die Bereiche der späteren p-MOS Transistoren umdotiert werden auf n-leitend, - man erzeugt eine n-Wanne. Die Verfügbarkeit beider Transistoren erlaubt Schaltungen so auszuführen, daß die Verlustleistung drastisch sinkt gegenüber einer Schaltung mit gleicher Funktion aber nur einer Transistorsorte.

Auch in der Mikroelektronik gibt es nichts umsonst. Die geringere Verlustleistung wird erkauft durch eine erheblich höhere Prozeßkomplexität (Erzeugung der Wanne plus eine neue Transistorsorte bedingt ca. 5 zusätzliche Lithographieschritte

plus entsprechende Schichtabscheidungen und Ätzungen), höheren Platzbedarf für Isolation zwischen p- und n-Wannen, sowie einiger neuer Probleme (z.B. das "Zünden" von parasitären Thyristoren, bekannt als "latch-up" /17/). Ohne CMOS-Technik würde sich der Chip jedoch durch zu große Erwärmung praktisch selbst zerstören, so daß sowohl das 1M DRAM als auch alle nachfolgenden Generationen in dieser Technologie ausgeführt sind, vgl. Tabelle 2.

16.4.3 Gesamtprozeß 4M-DRAM

Die gegenüber dem 1M-DRAM erforderliche Verkleinerung der Zellfläche wird erreicht durch eine bessere Lithographie und durch platzsparende Anordnung der einzelnen Komponenten. Beide Maßnahmen sind nur durch neue und verbesserte Einzelprozesse machbar. Die dann herstellbaren kleineren Strukturen führen zu neuen Problemen, die wiederum neue Prozesse und Technologien erfordern.

Der Einfachheit halber wird die "neue" Lithographie und ihre Ausstrahlung auf andere Prozesse nicht weiter betrachtet. Die wichtigsten Stichworte zu platzsparenden Maßnahmen und den damit verbundenen Problemen sind:

- Grabenkondensator
- ONO Dreifachdielektrikum
- überlappender Bitleitungskontakt
- TiN Diffusionsbarriere
- Kontaktlochverrundung.

Neben diesen vier Prozeßkomplexen, die gegenüber dem 1M-DRAM neu sind, gibt es viele weitere Änderungen im Detail, auf die hier aber nicht eingegangen werden soll. Ein Querschnitt zeigt Fig. 15c, er ist direkt mit dem Querschnitt des 1M-DRAMs in Fig. 15b vergleichbar.

Grabenkondensator

Wie schon ausgeführt, reicht die für den Kondensator zur Verfügung stehende Fläche direkt nicht mehr aus, um die erforderliche Kapazität zu realisieren. Der Ausweg ist die Einbeziehung der dritten Dimension, entweder nach unten (Kondensator im "Trench") oder nach oben (Kondensator wird über den Transistor gezogen; "stacked" in Tabelle 1; hier nicht weiter behandelt, vgl. /6/).

Fig. 15c zeigt den Grabenkondensator des Siemens 4M-DRAMs. Die für den Kondensator zur Verfügung stehende Gesamtfläche vergrößert sich gegenüber der planar zur Verfügung stehenden Fläche um die Seitenwände und den Boden des Lochs; prinzipiell hat man nur den Kondensator des 1M-DRAMs in Fig. 15b in die Tiefe des Graben hineingefaltet. Die technologischen Probleme, die dieser Grabenkondensator mit sich bringt, sind jedoch beachtlich:

- Ätzung der Löcher (vgl. Fig. 13,21), Kontrolle der Lochtiefe
- Dotierung der Seitenwand (geht nicht mehr über Ionenimplantation, vgl. Tabelle 6).

- Ecken, Kanten, wechselnde Kristallorientierung unter dem kritischen Dielektrikum
- Herstellen der Polyelektrode im Graben
- Isolieren der Polyelektrode und komplette Auffüllung des Grabens
- Vermeidung von elektrischen Durchschlägen zwischen zwei Gräben.

Die Beherrschung nur dieses Prozeßkomplexes bedingt einen Aufwand von rund 100 Mannjahren und ca. 3-4 Jahre intensiver Arbeit.

ONO-Dielektrikum

Da die Wachstumsgeschwindigkeit und damit die Dicke eines thermischem Oxids sowohl von der Kristallorientierung als auch von der Topologie der Unterlage (konkave/konvexe Ecken) und eventuellen mechanischen Spannungen abhängt, ist eine gleichmäßige Dicke im gesamten Graben praktisch nicht zu erreichen. Aus diesem und einer Reihe von anderen Gründen wird ein Sandwich aus thermischem Oxid, abgeschiedenem Nitrid (CVD-Verfahren) und einem weiterem thermischem Oxid geformt - abgekürzt ONO. Damit sind die beiden extrem wichtigen Grundflächen zwischen Dielektrikum und Si-Substrat bzw. poly-Si erstmal unverändert, gleichzeitig sorgt das gleichmäßig dicke Nitrid aber für hohe Durchschlagfestigkeit (und erhöht etwas die Dielektrizitätskonstante).

Neben einer Fülle von wissenschaftlichen Fragen mit direkter Bedeutung für den Prozeß (z.B. Mechanismus der Ladungsträgertransports durch die ONO Schicht) wird der Dielektrikaprozeß jetzt komplexer (3 Prozeßschritte statt einem) und sehr viel schwerer kontrollierbar. Statt einer Schicht mit "bequem" meßbarer Dicke (ca. 10 nm) sind es jetzt drei Schichten mit schwer meßbaren Dicken (ca. 3 nm), die darüberhinaus noch dreidimensional geformt sind /18/.

Überlappende Bitleitungskontakte

Beim 1M-DRAM wurde der Bitleitungskontakt durch Ätzen einer entsprechenden Öffnung durch das isolierende Oxid hergestellt. Dabei muß der Abstand des Lochrandes zum Gate des Transistors so groß sein, daß auch unter ungünstigsten Umständen (alle Toleranzen summieren sich in eine Richtung) das Gate nicht angeätzt wird. Beim 4M-DRAM wurde durch einen komplizierten Prozeß die Gates der Transistoren in Oxid eingekapselt, das "Loch" für die Bitleitungskontakte kann dann so groß sein, daß es die Gates überlappt. Ein Sicherheitsabstand wird nicht mehr benötigt, das Bitleitungsraster kann um ca. 1 µm enger geführt werden. Erkauft wird dieser Platzgewinn durch ca. 40 zusätzliche Prozeßschritte. Fig. 16 zeigt den Prozeßablauf.

TiN Diffusionsbarriere

Obwohl in Fig. 15 keine direkten Kontakte zwischen den Al-Leiterbahnen und dem Si-Substrat zu sehen sind, gibt es an anderen Stellen der Schaltung davon sehr viele. Die durch sämtliche isolierende Schichten geätzten Kontaktlöcher haben

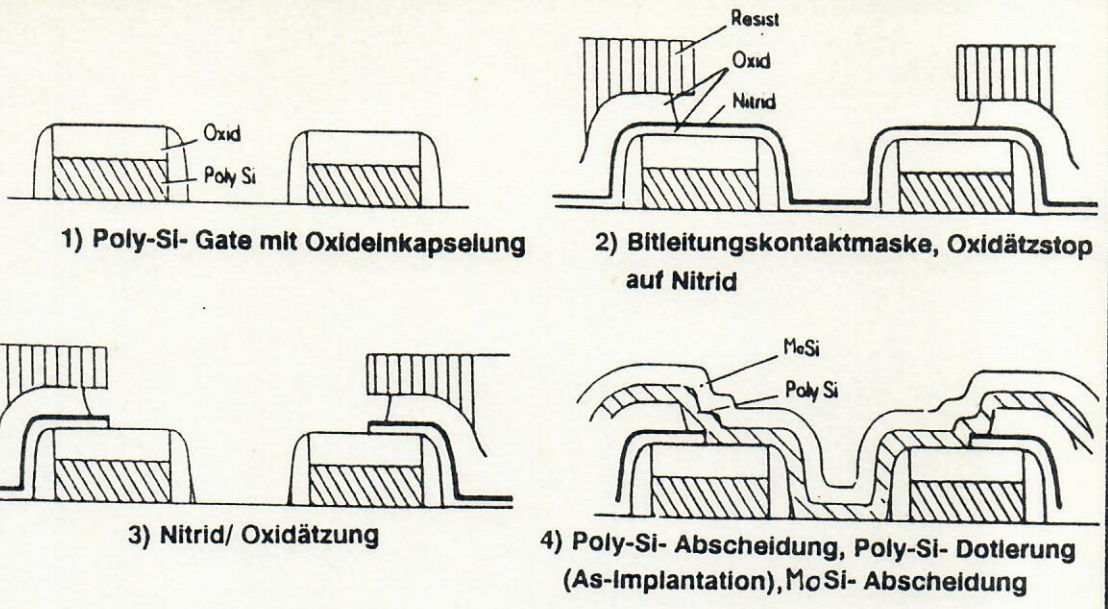


Fig. 16: Prozeßablauf für "FOBIC". Die Nitridschicht wird als "Ätzstop" benötigt /19/.

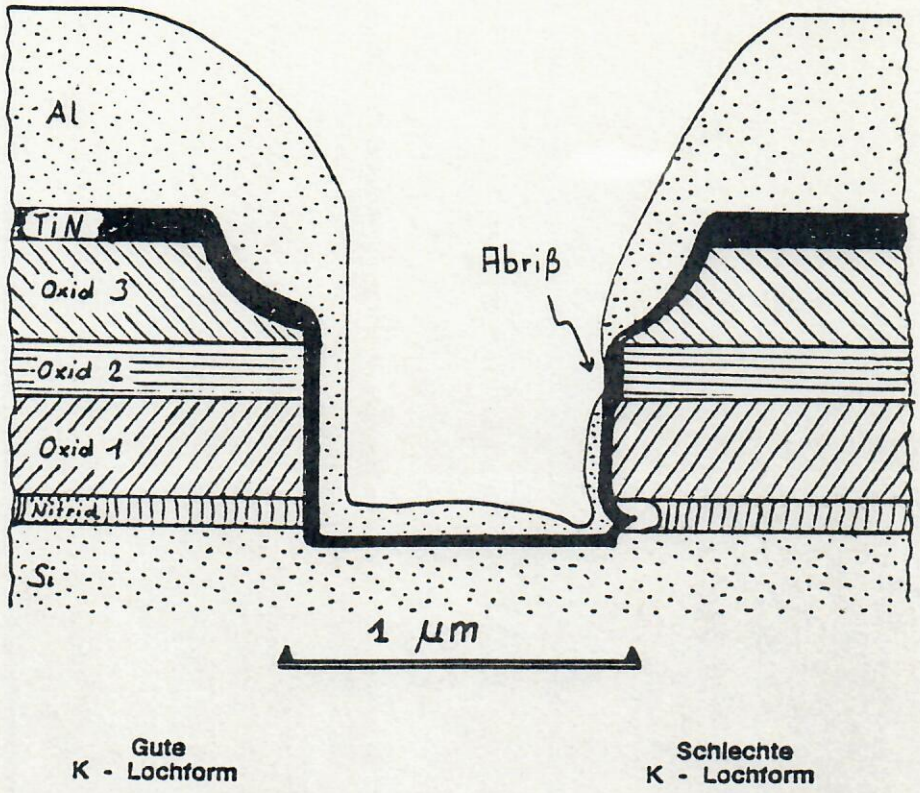


Fig. 17: Schematischer Schnitt durch ein 4M Kontaktloch. Die trotz Kontaktlochverrundung schlechte Kantenbedeckung des Aluminiums geht bei ungünstiger Geometrie auf Null.

Querschnitte von ca. 1 µm². An der Grenzfläche Si-Al laufen metallurgische Reaktionen ab (Eindiffusion von Si ins Al oder Ausscheidung von im Al schon enthaltenem Si), die den Kontaktwiderstand negativ beeinflussen. Bei den kleinen Kontaktlöchern des 4M-DRAMs sind die damit verbundenen hohen Kontaktwiderstände nicht mehr tolerierbar, es wird deshalb eine Zwischenschicht, z.B. aus TiN, eingezogen, die Reaktionen verhindert und den Kontaktwiderstand klein hält / /. Obwohl hier im Grunde nur ein zusätzlicher Prozeßschritt neu hinzukommt, entsteht Arbeit in Umfang von 15-20 Mannjahren. Beispielsweise ist zu klären: Optimale Herstellbedingungen der neuen Schicht; Anforderungen an die Reinheit der Ausgangsmaterialien; Haftungsfrage zwischen Schicht und Silicium, Oxid, Aluminium; Zuverlässigkeits-/Alterungsverfahren; Strukturierung durch Ätzen; Temperaturbeständigkeit; etc. Einen Schnitt durch das Kontaktloch mit TiN- und Al-Schicht zeigt Fig. 17.

Kontaktlochverrundung

Ohne die in Fig. 17 schon gezeigte Kontaktlochverrundung würde beim Aufbringen des Aluminiums durch Sputterverfahren nicht mehr genügend Al an die Seitenwände und auf den Boden des Kontaktlochs gelangen /21/. Nur durch das Abrunden der Kanten in einem technisch schwierigen Prozeß gelingt es, noch genügend Al in das Kontaktloch einzubringen. Wie schon die TiN-Diffusionsbarriere ist dieser Schritt eine Konsequenz, und nicht eine Maßnahme der Strukturverkleinerung.

Alles in allem erhöht sich gegenüber dem 1M-DRAM die Zahl der Prozeßschritte um ca. 150 auf ca. 400.

16.4.4 Gesamtprozeß 16M-DRAM

Neben den immer erforderlichen Fortschritten der Lithographie müssen wiederum viele davon betroffene Prozesse verfeinert werden. Wesentliche Innovationen um Platz zu sparen oder die mit kleineren Dimensionen verbundenen Probleme zu lösen, sind zum Beispiel:

- Al-Metallisierung in zwei Lagen
- Reduzierung der internen Spannung in einzelnen Bereichen auf 3,3 V
- Bessere Isolation zwischen den Trenches
- Auffüllung des Kontaktlochs.

Die beiden ersten Punkte führen direkt zu Platzeinsparungen und sollen nicht weiter behandelt werden, sie sind in dem Querschnitt von Fig. 15d aber nicht enthalten.

Trenchisolation

Beim 4M-DRAM ist der "elektrische" Abstand zwischen zwei Gräben gegeben durch den geometrischen Abstand minus der Tiefe der Raumladungszone; er beträgt ca. 1,5 µm. Die Feldstärke bei 5V Potentialdifferenz zwischen zwei Gräben liegt entsprechend bei ca. 30 000 V/cm. Da Silicium kein Isolator ist, wäre die Feldstärke beim 16M-DRAM so hoch, daß ein elektrischer

Durchbruch zwischen zwei Gräben eintreten könnte. Eine mögliche Abhilfe besteht, wie in Fig. 16d gezeigt, in einer Auskleidung des Grabens mit isolierendem Oxid. Als Kondensatorelektroden dienen dann zwei Polysilicium-Schichten. Die damit verbundenen Probleme sind erheblich; z.B. wächst das thermische Oxid für die untere Lage des ONO-Dielektrikums jetzt nicht mehr auf einkristallinem, sondern auf (rauhem) poly-Si. Außerdem muß jetzt ein extra Kontakt zwischen der poly-Si Elektrode und dem Transistoranschluß im einkristallinem Si erzeugt werden.

Kontaktlochauffüllung

Selbst die beim 4M eingeführte Kantenverrundung des Kontaktlochs reicht jetzt nicht mehr aus, um genügend Aluminium beim Sputtern in das Kontaktloch zu bringen. Das Kontaktloch muß jetzt mit einem leitenden Material ausgefüllt werden. Ein möglicher Weg ist die CVD-Abscheidung von Wolfram (funktioniert, im Gegensatz zu den meisten anderern Metallen auch bei relativ niedrigen Temperaturen); gefolgt von einer "Rückätzung" des Wolframs in den Gebieten außerhalb des Kontaktloches /22/. Die damit verbundenen Probleme und Arbeitsaufwände betragen etwa das 2-3 fache des Aufwands für die TiN-Diffusionsbarriere.

Diese wenigen Beispiele sollen genügen, um ein Gefühl zu vermitteln, was es bedeutet, eine neue Speichergeneration zu entwickeln. Detailliertere Darstellungen finden sich in /23,24/.

16.4.6 Meßtechnik, Analytik und Simulation

Fertige Chips, die aus der Linie herauskommen, gehen in die sogenannte Prüftechnik; darüber wird in Kapitel 16.5 gesondert berichtet. Selbstverständlich wird auch in der Prüftechnik gemessen und analysiert; mit Meßtechnik und Analytik im engeren Sinn bezeichnen wir hier aber die Aktivitäten, die während der Entstehung des Chips oder nach Abschluß der Prüftechnik anfallen.

Während der Chipherstellung wird in der F+E-Linie laufend gemessen und kontrolliert; beim 16M DRAM kommen auf > 400 Prozeßschritte mehr als 140 Messungen und Kontrollen. Spitzenreiter ist dabei die Messung einer Schichtdicke (ca. 40-50 mal), gefolgt von Kontrollen und aufwendigen Messungen der Partikeldichte (20-30 mal); Linienbreitenmessungen und anderen Lithographiekontrollen (ca. 20 mal) sowie Messungen von Schichtwiderständen, Ätzkontrollen, Durchbiegungen, Reflektivitäten, etc.. Manche Kontrollen sind simpel (z.B. Grobkontrolle auf Partikel im Schräglicht), manche sind kompliziert und aufwendig (z.B. exakte Messung der Linien-breite in einem speziellen Rasterelektronenmikroskop), manche sind nicht existent und müssen erst entwickelt werden. Wie messe ich z.B. die drei einzelnen Schichtdicken des "ONO"-Dielektrikums des 4M-DRAMs (also ein Sandwich aus Oxid, Nitrid, Oxid) mit den Sollwerten: 1. Oxid: (5 +/- 0,5) nm; Nitrid: (8 +/- 1,5) nm; 2. Oxid: (3 +/- 0,5) nm mit der notwendigen Genauigkeit? Und was macht man beim 16M DRAM, wo Schichtdicken und Toleranzen noch

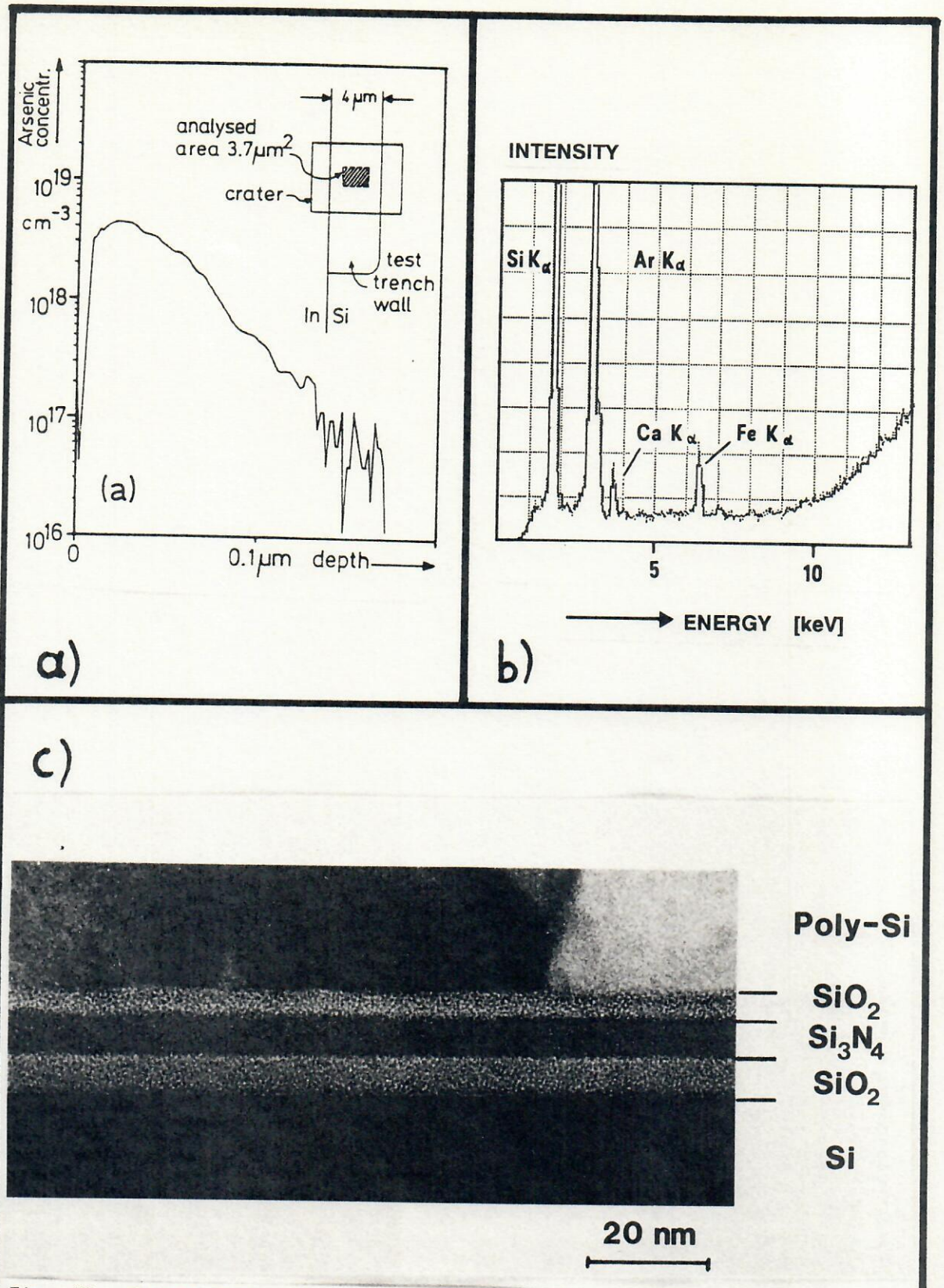


Fig. 18: Beispiele von Höchstleistungen moderner Analytik in der Mikroanalytik: a) Dotierungsprofil im Graben des 4M DRAMs /26/.
b) Detektion minimalster Kontamination durch Röntgenanalytik /28/.
c) Elektronenmikroskopischer Querschnitt durch das ONO-Dielektrikum /29/.

kleiner werden und das ganze Paket sich auf einer "rauhen" Unterlage befindet (nämlich auf poly-Si)?

Die Weiterentwicklung der vorhandenen Meßtechnik, aber auch die Entwicklung neuer Methoden gehört mit zu den Aufgaben der Prozeßentwickler. Letztendlich ist jeder Prozeß nur so gut wie seine Kontrollmöglichkeit.

Während die "in-line" Meßtechnik grundsätzlich die Scheiben weder zerstören noch die Prozesse beeinflussen darf, ist die physikalisch - chemische Analytik in der Regel zerstörend. Sie beantwortet Fragen wie z.B.:

- Konzentrationsprofil der Dotierstoffe.
- Konzentration von Zusatzstoffen in Schichten; z.B. B und P im Zwischenoxid, Cu im Aluminium.
- Konzentration und Art von Kontamination (z.B. Fe, Cu, Ni) im Silicium, in den Schichten, in den Medien oder an der Oberfläche.
- Genaue Geometrie und Topographie der Schichten.
- Gefüge und exakte Dimensionen von Strukturen, die einer direkten Messung nicht zugänglich sind (z.B. das dünne Dielektrika im Boden eines Grabens).
- Nachweis und Analyse von Kristallgitterdefekten.

Zum Einsatz kommt praktisch das gesamte Instrumentarium der physikalischen und chemischen Analytik. Neben den heutzutage schon eingesetzten Geräten und Methoden wie Rasterelektronenmikroskopie (REM), "Secondary Ion Mass Spectrometry" (SIMS), "Scanning Auger Microscopy" (SAM); "Fourier-Transform Infrared Spektrometry" ((FTIR); "Electron Spectroscopy for Chemical Analysis" (ESCA), Ionenchromatographie, etc., kommen auch besonders aufwendige oder teure Verfahren zunehmend zum Einsatz, z.B. die Transmission-Elektronenmikroskopie (TEM), "Rutherford Backscattering" (RBS) oder Neutronenaktivierungsanalyse (NAA).

Fig. 18 zeigt schlaglichtartig was einige dieser Methoden leisten.

Die Reihe der respektheischenden Bezeichnungen und Abkürzungen läßt sich fast beliebig verlängern. Wie diese Methoden im einzelnen funktionieren, was sie zu leisten vermögen und wie man sie einsetzt, füllt nicht nur Bände, sondern Bibliotheken; einen Einstieg in die Thematik bietet /25/. Interessant ist, welche Fragen, die bei der Chipentwicklung und -Herstellung auftreten, diese Methoden gar nicht, oder nicht schnell oder billig genug beantworten können:

- Wie ist das Dotierprofil an der Wand eines Grabens? (vgl. /26/).
- Identifiziere Art, Konzentration und Verteilung von Kontamination (Fe, Cu, Au, ...) in der Scheibe / /.
- Bestimme Ort und Zeit der Punkte, bei denen ein dünnes Dielektrikum bei angelegtem elektrischem Feld zuerst durchbricht.
- Bestimme, wie lange (in Jahren) eine Al-Leiterbahn der Strombelastung standhält.

- Bestimme Art und Zahl der im Ausgangsmaterial Silizium vorhandenen Kristalldefekte, die später als Keimbildner für Ausscheidungen dienen.
- Messe die Zahl von Partikeln mit Durchmesser < 0,1µm auf einer Scheibe.
- Bestimme die Konzentration der Dotieratome nach einer Implantation mit 1% Genauigkeit.

Auch diese Liste ist fast beliebig verlängerbar. Die offenen Probleme existieren, weil es Methoden mit der geforderten Empfindlichkeit/Ortsauflösung/Genauigkeit entweder überhaupt nicht gibt oder aber nicht in der erforderlichen Kombination, oder weil die Korrelation der gewünschten, einer direkten Messung nicht zugänglichen Zielgröße (z.B. Lebensdauer der Metallisierung) mit einer meßbaren Größe nicht hinreichend bekannt ist.

Die ungelösten analytischen und meßtechnischen Probleme der Halbleiterindustrie haben Forschung und Entwicklung bei Analytikern erheblich befruchtet. Es erscheinen laufend neue und oft außerordentlich ingeniöse Methoden, die eine Lücke bei der Chipentwicklung füllen (z.B. "Thermawave" /30/ und "ELYMAT" /27/); bestehende Methoden wurden verfeinert und der Fragestellung angepaßt (z.B. Rastermikroskope beginnen in den Linien die Lichtmikroskope zu verdrängen (allerdings bei Kosten für ein Gerät von bis zu $2 \cdot 10^6$DM)).

Analytik und Meßtechnik wird nicht nur während der Chipherstellung benötigt, sondern insbesondere auch danach. Die während der F+E Phase entstehenden Test- und Produktchips erfüllen (per Definition) nicht alle an sie gestellten Forderungen (sonst wäre die F+E Phase schon beendet). Meßtechnik und Analytik wird gebraucht, um festzustellen:

- ob die Ziele erreicht werden. Dies ist bei allen qualitätsrelevanten Fragen (Funktioniert alles noch nach 10 Jahren (oder 20 Jahren) Dauerbetrieb? (Sowohl am Nordpol als auch in der Sahara)) ein schwieriges Problem. Aussagen können offensichtlich nur durch beschleunigte Alterungsverfahren und Extrapolation gewonnen werden.
- Ob alle Strukturen nach Beendigung aller Prozesse auch wirklich noch so vorliegen wie geplant (ein Dotierprofil, z.B. ändert sich mit jedem Temperaturschritt).
- Was zu bei der Prüftechnik festgestellten Fehlfunktionen geführt hat. Dazu gehört die Lokalisierung des Fehlers und die Ursachenfindung.

Nimmt man an, daß in der heißen Phase kurz vor Aufnahme der Fertigung (der sog. Qualifikationsphase) pro Woche ungefähr 100-200 Scheiben mit ungefähr 100 Chips pro Scheibe aus der Linie herauskommen, heißt das, daß bei ca. 10% Ausbeute pro Woche 9.000-18.000 nicht funktionierende Chips anfallen, von denen man gerne wüßte, warum sie nicht funktionieren. Obwohl natürlich eine Ursache der Ausfallgrund für viele Systeme sein kann, zeigen die Zahlen doch, daß einige Arbeit für Analytiker anfällt.

Eine starke Analytik- und Meßtechnikgruppe, die neben modernsten Analysengeräten einige zig hochspezialisierte Experten (Physiker, Chemiker, Elektrotechniker) umfaßt, ist also unabdingbar bei der Entwicklung neuer Chips. Diese Gruppe muß nicht nur die anfallenden Aufgaben durchführen, sondern auch ständig Methoden verfeinern, sowie neue Methoden bewerten und eventuell einführen. Gleichzeitig ist darauf zu achten, daß man sich nicht zu Tode analysiert: Oberste Priorität hat das Beseitigen, nicht das Verstehen von Ausfallursachen.

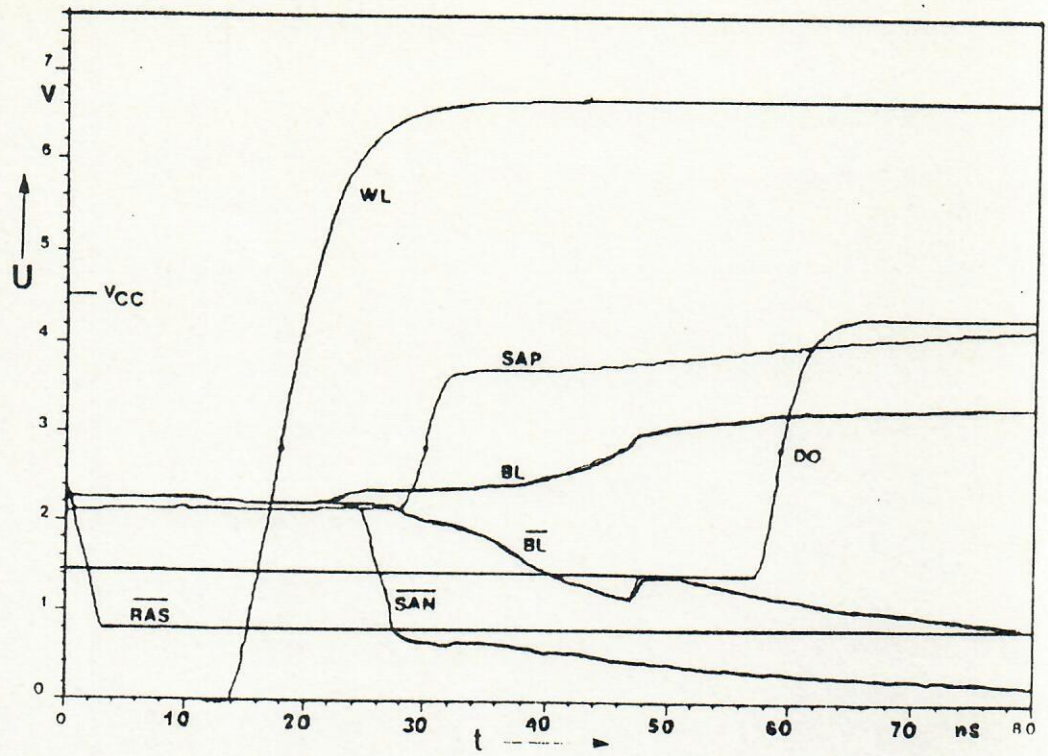
In den Kontext der Analyse- und Meßtechnik passen auch die Aktivitäten der "Simulanten", die mit aufwendigen mathematischen Modellen (häufig als Software kaufbar) z.B. Dotierstoffprofile nach zig Prozeßschritten ausrechnen und damit - falls man der Simulation traut - die Messung oder Analyse überflüssig machen.

Simuliert wird alles: Die Schaltungs- und Devicetechnik (prinzipielle Funktion, Schaltzeiten, Spannungsverläufe als Funktion der Zeit, etc.); die Einzelprozesse (z.B. Oxiddicke als Funktion der Kristallorientierung, des O_2- und H_2O-Partialdrucks und des Temperatur-Zeit Verlaufs), die Vorgänge im Silizium (z.B. Dotierprofile oder Sauerstoff-Ausscheidungsverhalten) und das Equipment (z.B. Strömungs-verhältnisse (und damit Schichtdickenhomogenität) in einem LPCVD-Ofen unter voller Berücksichtigung der Geometrie von Ofenrohr, Scheiben und Quarzboot (in dem die Scheiben stehen)). Fig. 19 zeigt ein Beispiel für ein wichtiges Simulationsergebnis. Wie die Analytik, ist auch die Simulation mehr als nur ein nützliches "Linienzubehör": Ohne weitgehende Simulation beim Design und in der Prozeßtechnik läßt sich heute kein moderner Chip mehr in endlicher Zeit entwickeln. Auch hier gilt, daß neben der Durchführung der direkten Aufgaben die Methoden ständig weiterentwickelt werden, müssen wobei insbesondere die physikalischen Modelle der zu simulierenden Phänomene eingebracht werden müssen. Dabei ist anzumerken, daß die Genauigkeit der Ergebnisse, und damit die Brauchbarkeit, heutzutage entscheidend vom Verständnis der Effekte höherer Ordnung abhängt, die man vor 10 Jahren noch kaum kannte (z.B. Verstärkung oder Schwächung der Diffusion eines Dotierstoffes durch die Anwesenheit eines anderen).

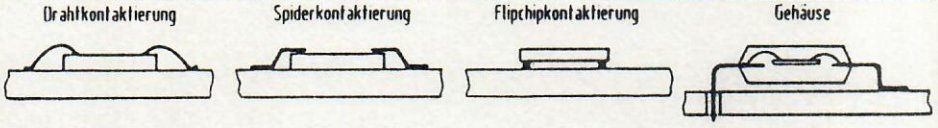
16.5 Prüftechnik und Montage

16.5.1 Prüftechnik

Nach Aufbringen der Passivierungsschicht und Freiätzen der Kontaktflächen können die Chips elektrisch vermessen werden. Es gibt eine Reihe von Meßmöglichkeiten: Der einfache Funktionstest prüft nur, ob der Chip macht, was er soll. Bei einem Speicher ist das verhältnismäßig einfach; bei einem komplizierten Logik-Chip jedoch sehr schwierig. Ein Analysetest mißt erheblich mehr, neben der direkten Funktion auch noch eine Vielzahl von Parametern, die den Designern und Prozeßtechnikern Aufschluß über das Verhalten einzelner Komponenten des Chips geben. Dazu befinden sich auf dem Wafer spezielle Teststrukturen, die sich z.B. im "Ritzrahmen", also dem Bereich zwischen zwei Chips, befinden. In der F+E Phase besteht der



Fig, 19: Beispiel für die Simulation der Spannung in einzelnen Leitungen eines 4M-DRAMs /31/. Die Rechenzeit auf einem Großrechner beträgt ca. 4 - 5hr.



Chipanschlüsse	Standardausführung	harte Höcker	weiche Höcker	entspr. Ausführung
Elektr. Prüfbarkeit vor dem Einbau	statisch prüfbar	vollständig prüfbar	statisch prüfbar	vollständig prüfbar
Substrat	Systemträger Schichtschaltungen Leiterplatten	Systemträger Leiterplatten Schichtschaltungen	Schichtschaltungen	Leiterplatten Schichtschaltungen
Verbindungsverf.	Einzelkontaktierung über Drähte	Komplettkontaktierg. auf Spiderband	Komplettlöten im Reflow-Verfahren	Komplettlöten in und auf Leiterplatten

Fig. 20: Kontaktierverfahren für Montagezwecke mit einigen Hauptmerkmalen /32/.

Chip sowieso hauptsächlich aus Teststrukturen; vom eigentlichen Produkt sind nur Teile enthalten. Ein 16M-DRAM Testchip enthält beispielsweise nur einen 4M-Speicher, den aber in 16M-Technologie. Der restliche Platz wird für Teststrukturen verwendet. Welche Parameter werden gemessen? Mehr als hundert, zum Beispiel: Einsatzspannungen der echten und der parasitären Transistoren; Leckströme aller Arten, Widerstände aller Arten (Bahn-, Schicht-, Kontakt- und Isolationswiderstände); geometrische Parameter (durch trickreiche Strukturen elektrisch meßbar). Das ganze dann möglicherweise noch als Funktion der angelegten Spannung (schließlich darf die Versorgungsspannung etwas schwanken) und mit Auswertung der Datenflut (Mittelwerte, Standardabweichung, graphische Darstellung, etc.). Nachdem der Chip montiert ist, kommen noch weitere Messungen dazu: Funktionsfähigkeit im gesamten Feld des spezifischen Einsatzbereichs (Spannungen, Temperatur, Luftfeuchtigkeit) - der prüftechnische Aufwand ist beachtlich. Zu bedenken ist, daß die zu messenden Parameter oft nicht ganz einfach zu messen sind: Ströme von 10^{-12}A bis 10^{-2}A; Zeiten im nsec-Bereich, Widerstände von 10 Ω bis 10 MΩ. Allein das Aufsetzen der Prüfnadeln (bei komplexen Produkt- oder Testchips mehrere hundert) auf die Kontaktflächen (muß auf ca. 10 µm genau erfolgen) ist nicht ganz trivial.

Funktions- und Analysetester sind dementsprechend Geräte, die mehrere 10^6DM kosten und zum Betrieb umfangreiche Software benötigen. Da der Prüfaufwand mit jeder neuen Generation massiv ansteigt, schiebt sich eine möglichst intelligente Prüftechnik bei der Chipentwicklung immer mehr in den Vordergrund. Bei komplexen Logikchips beträgt der Anteil an den Gesamtkosten bereits ca. 30%. Bei Speichern ist der Kostenanteil zwar noch etwas geringer, bei linearer Extrapolation aber wirtschaftlich schon beim 16M DRAM nicht mehr tragbar. Beim 16M DRAM müssen daher Maßnahmen ergriffen werden, die den Prüfaufwand deutlich reduzieren.

Der Prüftechnik-Ablauf einer DRAM-Scheibe sieht etwa folgendermaßen aus:

An einem ersten Tester wird mit den Teststrukturen im Ritzrahmen ein Parametertest durchgeführt. Danach erfolgt ein erster Funktionstest der Speicherchips. Dabei werden die Chips erkannt, die zwar nicht vollständig funktionieren, aber über die eingebaute Redundanz repariert werden können. Die Reparatur dieser Chips erfolgt anschließend am Laser-Cutter, der durch einen gezielten Schuß an der richtigen Stelle Leiterbahnen unterbricht und dadurch die redundanten Zellen aktiviert. Anschließend wird wiederum auf Funktion geprüft um sicher zu gehen, daß die Reparatur geklappt hat (es könnten ja die Redundanzzellen einen Defekt enthalten). Chips die jetzt nicht funktionieren, erhalten den ominösen roten Punkt.

Die Scheibe wird jetzt zersägt und die Chips in Gehäuse montiert (s. Kapitel 16.5.2). Anschließend wird erneut getestet - jetzt mit etwas einfacherer Technik, da man nicht mehr die komplizierten, anfälligen Kontaktnadeln braucht, sondern den IC in einen Sockel stecken kann. Anschließend wird der Chip gequält: Er kommt in den "Dampfkochtopf"; erhält einen "burn-

in" (d.h. wird bei hoher Temperatur einige Zeit betreiben), etc.. Dazwischen wird immer wieder neu auf Funktion geprüft. Die Chips die alle Torturen überleben werden schließlich verkauft.

Die Logistik der Prüftechnik ist außerordentlich komplex. Die Prüfzeiten pro Chip hängen von allen möglichen Einflüssen ab: Zahl der zu messenden Parameter und Qualität der Software (und Hardware). Mit zunehmender Erfahrung werden die Prüfprogramme in der Regel erheblich schneller; doch ist dies nur schwer in eine Kapazitätsplanung einzurechnen. Funktionstests können sehr schnell sein, solange die Ausbeute noch niedrig ist, da der Tester sofort aufhört, wenn er einen nicht reparaturfähigen Fehler findet. Bei hoher Ausbeute muß dann meistens bis zum Ende gemessen werden, die mittlere Testzeit pro Chip steigt an.

16.5.2 Montage

Früher oder später kommt jeder funktionierende Chip in ein Gehäuse. Das Gehäuse hat die Aufgabe, den Chip einerseits von Umwelteinflüssen (insbesondere Feuchtigkeit) hermetisch abzuriegeln, den elektrischen Kontakt zur Außenwelt herzustellen und die entstehende Wärme effektiv abzuführen. Darüberhinaus soll es möglichst klein, leicht steck- und lötbar sein und möglichst nichts kosten. Die letzte Forderung muß man unter dem Gesichtspunkt betrachten, daß Speicherchips früher oder später deutlich weniger als 20 DM kosten und damit in der gleichen Größenordnung liegen wie ein Keramiksubstrat für die besonders guten Keramikgehäuse.

Ein Speicherchip muß also in das billige Plastikgehäuse. Dabei sei nur ein Problem schlaglichtartig beleuchtet: Beim 4M- oder 16M DRAM bleibt für die Beinchen gerade noch ein halber Millimeter Platz innerhalb des Gehäuses. Das Hauptproblem ist aber, daß die thermischen Ausdehnungskoeffizienten von Plastik und Silizium um einen Faktor 4-5 verschieden sind. Dies führt fast unvermeidlich dazu, daß beim Abkühlen der Plastikmasse infolge der thermischen Spannungen entweder der Chip bricht oder das Gehäuse einen Riß bekommt. Nur durch ausgeklügelte Optimierung von Material und Verfahren läßt sich eine Montage in Plastikgehäusen realisieren. Unnötig zu betonen, daß die besten Materialien für die Montage (Plastikmasse, Keramikträger, etc.) aus Japan kommen. Fig. 20 zeigt einige Kontaktierverfahren für Montagezwecke.

Betrachtet man nicht Speicher, die nur 24 Beinchen haben, sondern komplexe Logik-Chips mit über 300 Anschlüssen, wird die Montagetechnik fast so anspruchsvoll wie das Herstellen des Chips. Außerordentlich wichtig ist dabei, daß die entstehende Wärme möglichst effizient abgeführt wird. Die Montage eines solchen Chip erfordert allein ca. 50 Prozeßschritte und hat damit ihre eigene Durchlaufzeit sowie Ausbeute- und Logistikprobleme.