Teilauszug aus meinem Speicherbuch

2 Halbleiterspeicher

2.1 Elektronikschaltungen

Als Grundschaltungen der digitalen Elektronik gelten die kombinatorische und die sequentielle Schaltung. Je nach Autor werden sie allerdings recht unterschiedlich bezeichnet. Die nebenstehende Tabelle gibt dazu eine Zusammenstellung mit den zugehörigen Theorien. Die wichtige *Speicherschaltung* erscheint bestenfalls als Teil der sequentiellen Schaltung und das, obwohl gerade sie die Mikroelektronik entscheidend vorangetrieben hat und es auch immer noch tut. Gemäß [VÖE] wird hier von der *kombinatorischen Schaltung* und der *Speicherschaltung* ausgegangen. Aus beiden lassen sich nämlich konsequent alle anderen digitalen Schaltungen ableiten.

Eine *kombinatorische* Schaltung faßt mehrere zeitabhängige Eingangssignale $f_{em}(t)$ (m=1 bis n) nach einer bestimmten logischen Funktion F_1 zu einem einzigen zeitabhängigen Ausgangssignal

$$f_a(t) = F_1(f_{e1}(t), f_{e2}(t), f_{e3}(t), ..., f_{em}(t))$$

zusammen, das im theoretischen Idealfall unverzüglich den Änderungen der Eingangssignale folgt. Es gibt also keinen Einfluß von äußeren Bedingungen und Signalen der Vergangenheit. Die kombinatorische Schaltung reagiert streng deterministisch auf die Eingangssignale. Für *Speicherschaltungen* (vgl. das einfachere Bild 1.7 S. 10) sind drei Signalgruppen zu unterscheiden:

- Eingangssignale $f_{em}(t)$ (m = 1 bis n),
- Ausgangssignale $f_{am}(t)$ (m = 1 bis n),
- Auslösesignale $f_{auf}(t_a)$ und $f_{wied}(t_W)$.

Darin bedeutet t_a den Auslösezeitpunkt für die Aufzeichnung und t_w den für die Wiedergabe. Somit gilt bezüglich der Eingangs- und Ausgangssignale:

$f_{am}(t) = f_{em}(t + \Delta T)$ mit $\Delta T = t_w - t_a$.

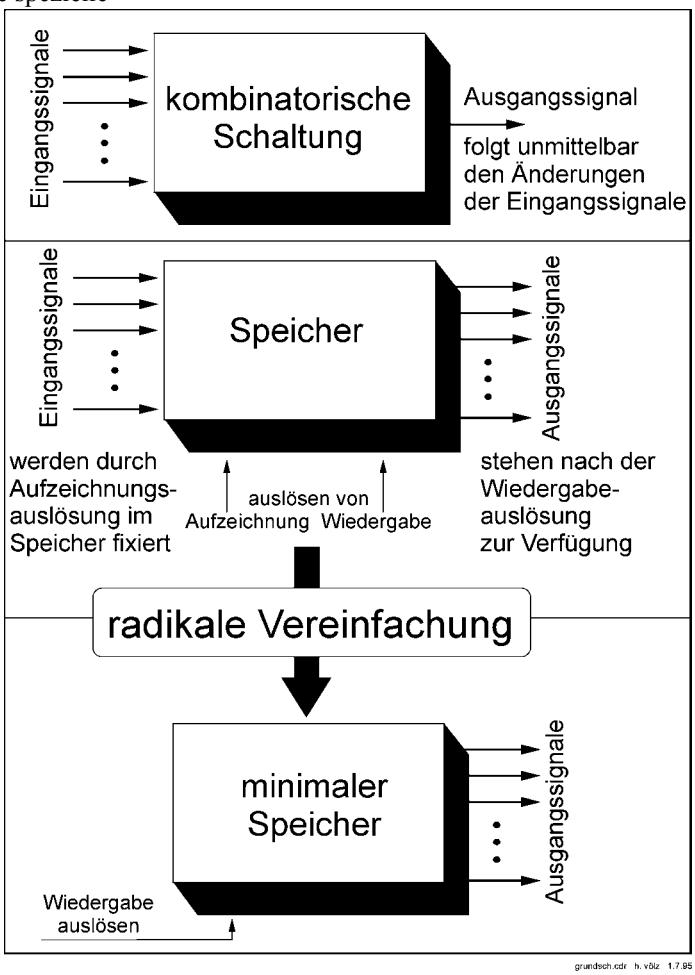
Synonyme der binären Technik z. T. mit etwas abweichender Bedeutung	
kombinatorische Schaltung	sequentielle Schaltung
Schaltnetz statische Logik binäre Schaltung Zuordner Codierer	Folge-Schaltung Schaltwerk dynamische Logik
Theorien dazu	
Schaltalgebra Boolesche-Algebra formale Logik Karnaugh-Diagramm	Automatentheorie Petrinetze

binschal.cdr h. völz 31.12.93

Rein funktionell kann daher ein Speicher als eine spezielle Verzögerungsschaltung betrachtet werden, die durch ihre Verzögerungszeit ΔT gekennzeichnet ist. Funktionell ist sie daher sogar einfacher als eine kombinatorische Schaltung, überbrückt sie doch „nur“ die Zeit zwischen der Entstehung der Eingangssignale und deren späterer Verwendung. Die Laufzeitketten auf CCD-Basis entsprechen inhaltlich genau dieser Betrachtung (S. 32). Dabei sind jedoch zwei wesentliche Speicherfunktionen unberücksichtigt:

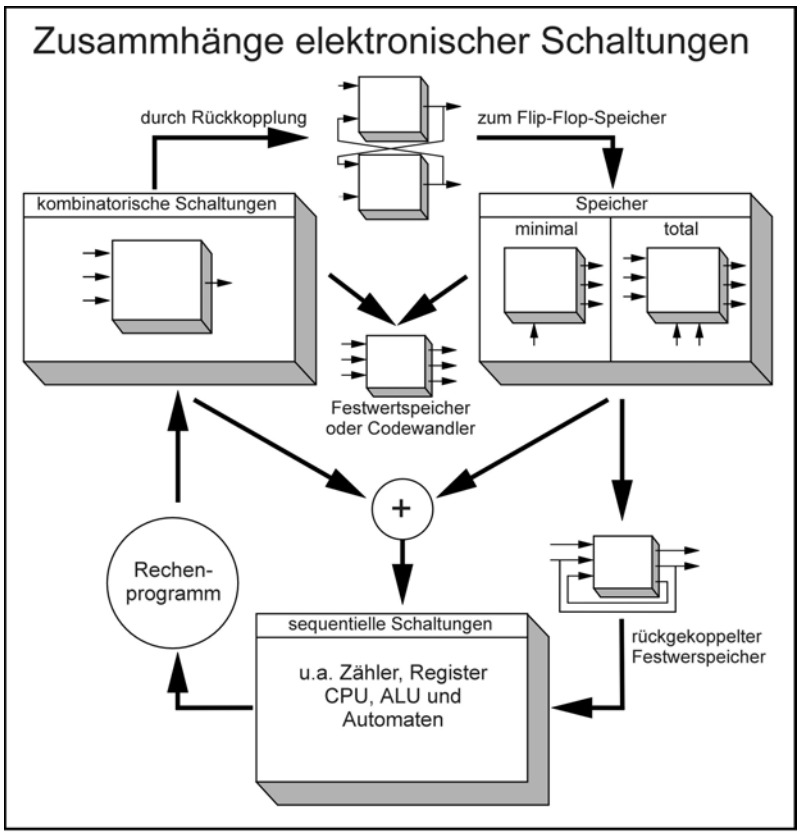
- Es gibt keine *Auslösesignale*. Somit sind die Startzeiten für Aufzeichnung und Wiedergabe – und ihr zeitlicher Abstand – nicht frei wählbar.
- Die *beliebig häufige* Wiederholung einer Aufzeichnung zu unterschiedlichen Wiedergabezeiten (Verzögerungen) ist nur mit zusätzlichem Aufwand über Rückkopplung zu erreichen.

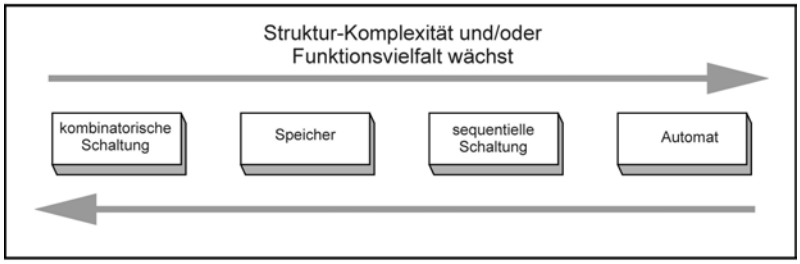
Trotz der funktionellen Einfachheit des allgemeinen Speichers ist noch eine Reduzierung zum *minimalen Speicher* (in der Elektronik das ROM read only memory) möglich (Bild 1 unten). Er besitzt keine Eingangssignale und keine Auslösung der Aufzeichnung. Seine einzige Funktion ist die Auslösung der Wiedergabe. Dennoch muß die Information irgendwie in den Speicher hineingekommen sein. Hierfür gibt es mehrere Möglichkeiten, die Auf S. 49 behandelt werden.



Nicht immer muß die Unsymmetrie zwischen Aufzeichnung und Wiedergabe so extrem wie beim minimalen Speicher sein. Übergänge sind u.a. PROM, EPROM, Flash-RAM usw. Auf sie wird Af S. 50ff eingegangen.

Zwischen den verschiedenen elektronischen Schaltungen existieren Querbeziehungen (Bild 2). Durch eine Rückkopplung aus zwei kombinatorischen Schaltungen entsteht ein *Flipflop*, der ein wichtiges Grundbauelement der Speicher ist. Der *Festwertspeicher* (minimaler Speicher) kann sowohl von den Speicherschaltungen abgeleitet werden als auch aus einer Zusammenschaltung von kombinatorischen Schaltungen entstehen. Seine Eingänge sind dann die Auslösesignale. Wird er als kombinatorische Schaltung benutzt, so entsprechen die Ausgänge mehreren gleichzeitig vorhandenen Funktionen. Dann kann er auch als Codewandler verwendet werden. Die Eingangssignale werden codiert zum Ausgang geführt. Eine *sequentielle Schaltung* kann gemäß Bild 2 einmal aus kombinatorischen Schaltungen und Speichern (z.B. Flipflops) aufgebaut werden. Andererseits läßt sie sich vom rückgekoppelten Speicher ableiten. Umgekehrt ermöglicht eine sequentielle Schaltung (Automat oder Rechner), mittels eines Programms eine kombinatorische Schaltung oder einen Speicher funktionell zu simulieren. Werden nach diesen Gesichtspunkten die Struktur- bzw. Funktions-K*omplexitäten* der digitalen elektronischen Schaltungen analysiert, so entsteht Bild 3.





2.2 Zur Systematik

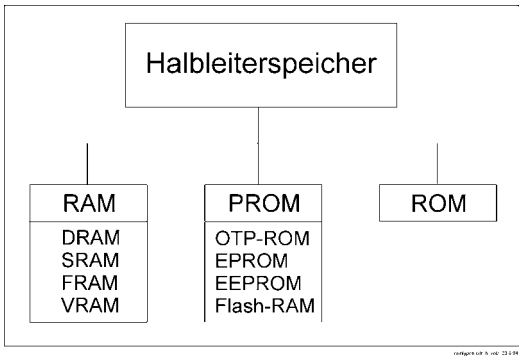
Elektrische bzw. elektronische Speicher existieren seit vielen Jahrzehnten. Am ältesten sind wohl das Telegraphenrelais mit zwei selbsthaltenden Stellungen und der Hebdrehwähler der Vermittlungstechnik. Noch in den 50er Jahren wurden die damals leistungsfähigsten Rechner mit ihnen gebaut. Erst später entstanden die Röhren- und Ferritkernspeicher. Die heutigen elektronischen Speicher arbeiten fast ausschließlich auf der Basis der Halbleitertechnik. Ihre Vielfalt ist sehr groß und läßt keine eindeutige *Klassifizierung* zu. Die folgende Grobeinteilung wird in den späteren Abschnitten vertieft.

- Speicher können unterschiedlich oft bzw. schnell mit *neuer Information* belegt werden. Dies führt zu den RAM, ROM, PROM usw.
- Elektronische Speicher beruhen auf unterschiedlichen Halbleiter*technologien*. Sie leiten sich meist aus den Schaltkreisfamilien der kombinatorischen Schaltungen ab, also einerseits bipolar z.B. TTL, I^2L, ECL und anderseits unipolar, vor allem nMOS, pMOS, CMOS. Experimentell existieren auch GaAs-Speicherzellen.
- Die *Speicherzelle* kann auf verschiedenen physikalischen Prinzipien beruhen, z.B.: Rückkopplung, Speicherkapazität oder periodisch umlaufende Information.
- Die *Adressierung* der einzelnen Speicherzellen kann z.B. direkt, reihen- und spaltenweise codiert oder sequentiell, gemäß RAS und CAS sowie Bit- und Wort-organisiert erfolgen.
- Von der Adressierung ist die *Organisationsform* zu unterscheiden. Daher gibt es u.a. Register, Buffer, Stacks, aber auch serielle, assoziative und holographische Speicher.
- Bezüglich einer optimalen *Zugriffszeit* sind neben den Technologien und Organisationsformen auch Einrichtungen, wie Cache, Buffer, serielle Speicher und Umlaufspeicher (wie z.B. CCD) entstanden.

- Zwischen *Leistungsverbrauch* je Bit und *Zugriffszeit* besteht ein enger Zusammenhang. Das Produkt ist genauso wie bei kombinatorischen Schaltungen ein typisches Kennzeichen der verwendeten Technologie. Die Zugriffszeit kann daher meist nur mit wachsender Leistung verkürzt werden.

2.2.1 Speichervarianten

Vom theoretischen Standpunkt sind zwei Speichervarianten ausreichend. Einmal wird für sich ändernde Daten der ständig neu beschreibbare Speicher RAM benötigt. Sein Name geht auf *random access memory* zurück, ist historisch bedingt und infolge der Übersetzung, etwa Speicher mit wahlfreien Zugriff, leicht mißverständlich. Die zweite Variante ROM (read only memory) ist dagegen für fixe Daten und zur sicheren Archivierung vorteilhaft. Technologische und ökonomische Gründe haben jedoch vielfältige Zwischenvarianten entstehen lassen. So führen die drei Hauptgruppen von Bild 4 zu der folgenden ergänzenden, alphabetisch geordneten, aber unvollständigen Aufstellung:



CDRAM	Cached DRAM, hat einen internen Cache und ähnelt dem EDRAM *).
DRAM	Dynamischer RAM.
EDRAM	Enhanced DRAM, enthält einen zusätzlichen SRAM mit der Wortbreite von 2048 Bit, der wie ein Cache arbeitet (15 ns erreicht) *).
EEPROM	auch E^2PROM, Electrical Erasable PROM, ist der Oberbegriff für mehrere Varianten, die Bit-, Wort-, Sektorweise oder vollständig löschbar und dann neu programmierbar sind. Er wird jetzt fast nur noch als Flash-RAM ausgeführt.
EPROM	Erasable PROM, im jetzigem Sprachgebrauch auf global mit UV-löschbare (in ca. 20 Minuten) und dann wiederbeschreibbare Bausteine eingeengt.
Flash-RAM	(Flash = Blitz) relativ neue EEPROM-Variante, die bevorzugt sektorweise oder vollständig löschbar und neu beschreibbar ist. Enthält vielfach schon die dafür notwendigen Algorithmen auf dem Speicherchip und ist wahrscheinlich die künftige Vorzugsvariante aller PROMs.
FRAM	Ferroelektrischer RAM, mit einer Hysterese, ähnlich wie bei magnetischem Material. Er existiert erst in Versuchsmustern, könnte aber bei Beherrschung der Technologie große Bedeutung gewinnen.
NV-RAM	Non volitile RAM ist eine Kopplung vom klassischen RAM und EEPROM. Der RAM-Teil wird mit seiner kurzen Zugriffszeit ständig vom Rechner verwendet. Zusätzlich wird in längeren Abständen sein jeweils aktueller Inhalt zur beständigen Speicherung in den EEPROM-Teil übertragen.
OTP-ROM	One Time Programmable ROM, der nur einmal vom Anwender programmierbar ist.
PROM	Programmable ROM, als Überbegriff für mehrere Produkte.
RAM	Random Access Memory.
RDRAM	DRAM nach der Firma Rambus, enthält internen 9-Bit-Bus für 500 MHz und vier 4K-SRAM als Cache *).
ROM	Read Only Memory, heute vorwiegend für Masken-ROM verwendet.
SDRAM	Synchronous DRAM, mit 100 MHz getaktet und eigenen Refresh *).
SRAM	Statischer RAM.
VRAM	Video-RAM mit der Möglichkeit des gleichzeitigen Schreibens und Lesens.

Die mit *) bezeichneten Bausteine sind sehr neu und ihre Zukunft ist noch unbestimmt. Sie wurden vor allem auf hohe Geschwindigkeit hin entwickelt und sind u.a. in [IRL] beschrieben.

Da praktisch alle Halbleiterspeicher bei Stromausfall ihre Information verlieren, sind zwei weitere Merkmale zur Klassifizierung gemäß der folgenden Tabelle nützlich:

- *Reversibiltät* im Sinne von neu mit Information zu belegen und
- *Beständigkeit* als Aussage dafür, daß die Information auch bei Abschaltung des Gerätes/Rechners erhalten bleibt.

	reversibel	Irreversibel
beständig	FRAM, SRAM mit Stützbatterie magnetische Verfahren	ROM, EPROM CD-ROM, WORM
unbeständig	DRAM (SRAM) CCD, Umlaufspeicher	kein Speichertyp

2.2.2 Halbleitertechnologien

Die Herstellung von Halbleitern begann mit dem Spitzentransistor und entwickelte sich über den gezogenen Transistor, den diffundierten Transistor und die Planartechnik zum Feldeffekttransistor. Hierbei entstanden gemäß den jeweiligen Gegebenheiten diverse logische Schaltkreisfamilien. Bei der *TTL*-Technik (transistor transistor logic) wird der leitende Transistor mit dem Sättigungsstrom betrieben. Dadurch verbraucht sie einerseits viel Strom und besitzt andererseits erhebliche Verzögerung beim Abbau der vielen Ladungsträger. Mit Weiterentwicklung der TTL-Technik konnten dann beachtliche Fortschritte erreicht werden. Dies macht auf

Gatter bezogen die folgenden Tabelle deutlich. Darin bedeuten H = high power, L = low power, S = schottky und A = advanced. Mit der TTL-Technik wurden auch die definierten logischen 5-Volt-Pegel eingeführt. Erst um 1991 wurde als zusätzliche Betriebsspannung 3,3 V international festgelegt. Dies war vor allem wegen der immer kleiner werdenden Abmessungen der Strukturen unumgänglich.

Generation	Typ	*T* in ns	*P* in mW	Produkt in pJ
I 1960	Standard TTL	10	10	100
	H-TTL	6	25	150
	L-TTL	33	1	33
II 1970	LS-TTL	10	2	20
	S-TTL	3	20	60
III 1980	AS-TTL	1,5	22	33
	ALS-TTL	4	1	4
	FAST-TTL	2	4	8
zum Vergleich	ECL	1	50	50
	I^2L	wählbar	wählbar	ca. 0,1

Die *ECL*-Technik (emitter coupled logic) ist durch eine Konstant-Stromquelle gekennzeichnet, an die je ein Emitter der Speicherzelle angeschlossen ist. Weil so der jeweils leitende Transistor deutlich unterhalb der Sättigung betrieben wird, ist ein schneller Ladungsträgerabbau möglich. Bei trotzdem hohem Stromverbrauch ist die ECL-Technik immer noch die schnellste Technologie. Weil auch die Hilfselektronik, wie Spalten- und Zeilenauswahl, in ECL-Technik realisiert werden muß, benötigt sie weiter viel Chipfläche. Daher wird sie nur mit kleinen Speicherkapazitäten für Spezialanwendungen für höchste Geschwindigkeitsansprüche eingesetzt.

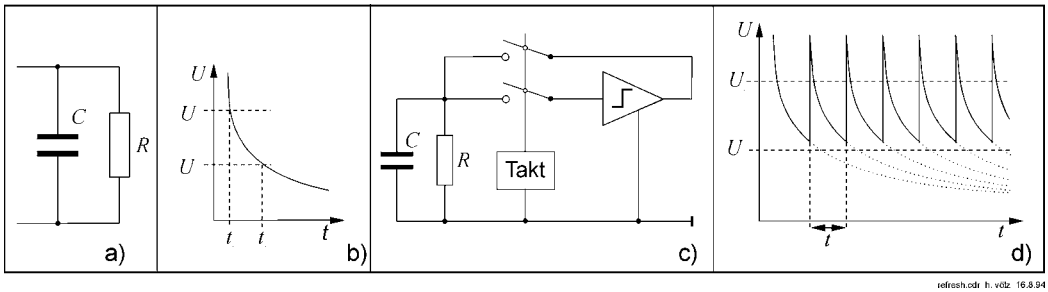
Bei der *I^2L*-Technik (integrated injection logic) erfolgt die Stromversorgung der Transistoren nicht einzeln durch eine Spannung, sondern global über die Injektion von Ladungsträgern durch einen pn-Übergang. Die so frei verfügbaren Ladungsträger werden in die jeweils „leitenden" Transistoren geführt. Folglich gibt es für I^2L-Bauelemente kein eigentliches Schaltbild, wodurch sie funktionell schwer verständlich sind (vgl. u.a. [VÖE]). Mit der Stärke des Injektionsstromes kann aber der Leistungsverbrauch und die Schaltzeit in weiten Grenzen verändert werden. Dabei bleibt in guter Näherung das Produkt aus Schaltzeit und Leistung konstant. Weiter ist eine hohe Packungsdichte erreichbar. Durch die großen Fortschritte der CMOS-Technik hat sie jedoch heute bei den Speichern keine Bedeutung mehr.

In der obigen Tabelle fehlen die *MOS*-Technologien (metall oxid silicium). Sie beruhen auf den FET (field effect transistor) mit einem sehr geringen Stromverbrauch. Beim schnellen Schalten benötigen sie so vor allem Umladeströme für die Kapazitäten, weshalb ihr Schaltzeit-Leistungs-Produkt etwa proportional mit der Taktfrequenz steigt (Bild 5). Neben den Schaltungen in p- bzw. n-Kanaltechnik hat die CMOS-Technik (complimentary) große Bedeutung erlangt. Bei ihr treten infolge der Kombination von n- und p-Kanaltechnik statisch nur noch Restströme (nA bis pA) auf. Sie ist folglich extrem leistungsarm. Ohne sie wären elektronische Armbanduhren, Taschenrechner usw. nicht möglich geworden.

Schließlich ist noch zu erwähnen, daß auch bereits SRAM auf *GaAs*-Basis realisiert wurden Sie erreichen eine sehr kurze Schaltzeit um 1 ns. Ihre Kapazitäten liegen bei einigen KBit.

2.3 Die Speicherzellen

Die funktionelle Grundlage aller Speicher ist eine Speicherzelle, die meist aus mehreren elektronischen Bauelementen, also Widerständen, Kondensatoren und Transistoren besteht. Im folgenden werden ihre unterschiedlichen physikalischen Prinzipien beschrieben.



2.3.1 Elektrische Kapazität

Die wohl einfachste elektronische Speicherzelle besteht aus nur einem Kondensator. Formal kann er geladen oder leer sein. Für die praktische digitale Anwendung sind jedoch zwei Spannungsschwellen u_0 und u_1 mit $U_0 < U_1$ entscheidend. Besitzt der Kondensator eine Spannung $U < U_0$, so ist die binäre „0" gespeichert, für $U > U_1$ dagegen die binäre „1". Der Bereich $U_0 \le U \le U_1$ ist binär nicht definiert. Er dient der sicheren Unterscheidung beider Werte und heißt verbotene Zone.

Die Speicherung im Kondensator ist durch Ladungsverluste zeitlich begrenzt. Dadurch kann jedem Kondensator mit der Kapazität *C* ein parallel geschalteter Widerstand *R* zugeordnet werden (Bild 6a). Mit der Zeitkonstanten

$$T = R \cdot C$$

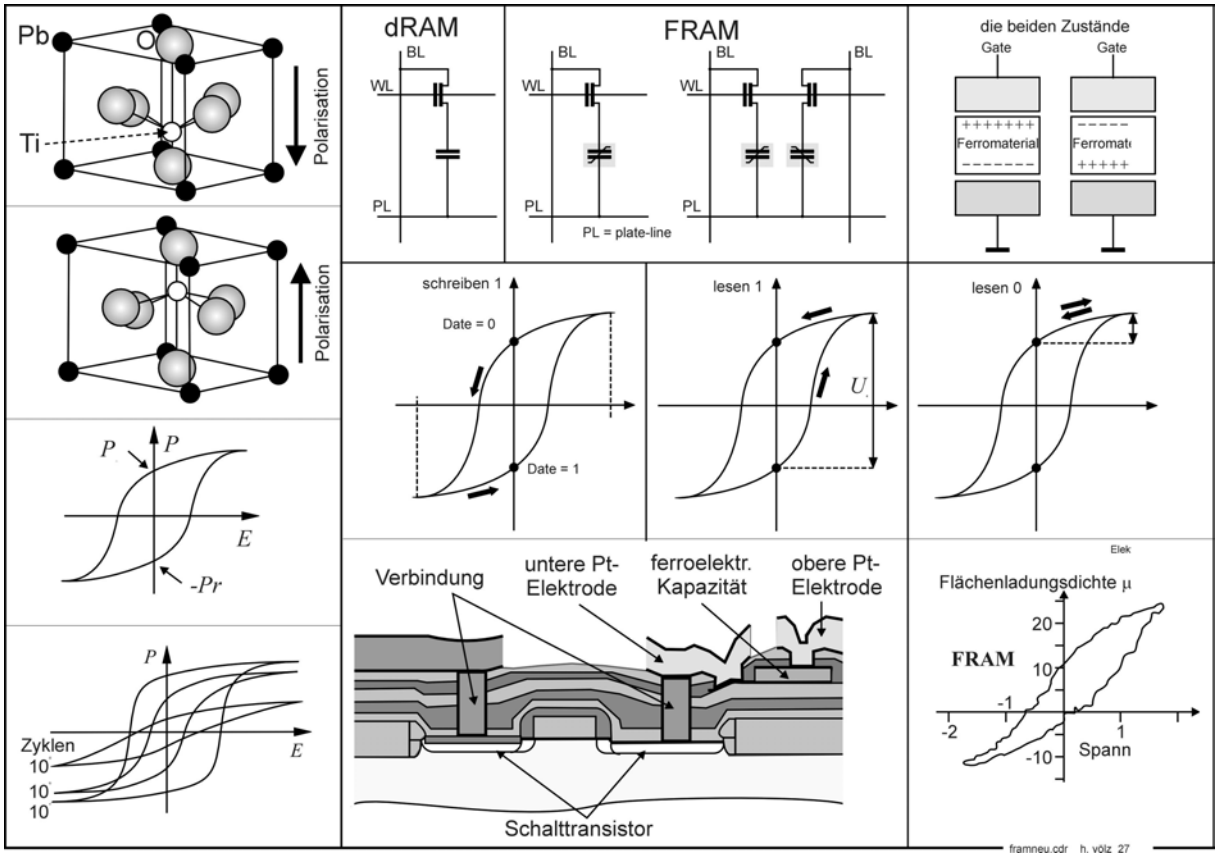
verliert der Kondensator Ladung. Bei einer Anfangsspannung $U_a > U_1$ gilt für den Zeitverlauf:

$$U = U_a \cdot e^{\frac{-t}{R \cdot C}}$$

Diesen Zusammenhang zeigt Bild 6b. Bei der Zeit t_1 wird also der binäre Wert 1 verlassen und bei t_2 wird der binäre Wert 0 angenommen. Damit kann dieser Kondensators den Wert 1 maximal für die Zeit t_1 speichern. Für die binäre 0 gibt es dagegen keine zeitliche Begrenzung. Mittels einer zusätzlichen Schaltung kann jedoch die Speicherzeit beliebig verlängert werden. Gemäß Bild 6c wird die Kondensatorspannung getaktet einem begrenzenden Verstärker zugeführt. Er erhöht sie ab einer Schwellspannung $U_0 < U_1$ auf U_a und führt sie wieder dem Kondensator zu. So ergibt sich für die logische 1 der zeitliche Verlauf von Bild 6d. Je kleiner U1 gewählt wird, desto länger kann die Taktzeit sein und desto aufwendiger wird der Verstärker. Falls im Kondensator die binäre 0 gespeichert ist, hält die Schaltung auch diesen Wert gegenüber Störungen fest. Bei den dynamischen RAM (S. 44ff) wird dieses Prinzip *refresh* genannt und meist von der CPU gesteuert.

2.3.2 Ferroelektika

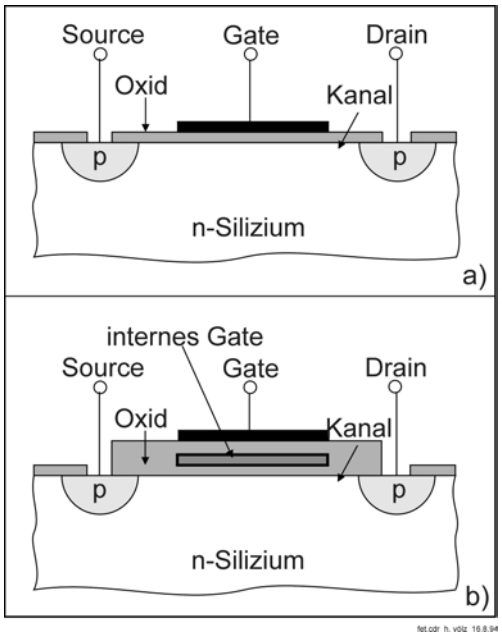
Eine andere Variante zur Erhaltung der Ladung eines Kondensators geht auf ferroelektrische Materialien zurück. Der Effekt wurde 1921 von Valasek am Seignettesalz gefunden. Wichtige Materialien sind heute $BaTiO_3$ (Bariumtitanat), KNO_3 und PZT-Keramiken. Die Kristallstruktur des PZT (Blei-Zirkon-Titan) $Pb(Ti\text{-}Zr)O_3$ zeigt Bild 7a. Das „verschiebbare“ Zentralatom Ti kann auch ganz oder teilweise durch Zr ersetzt werden. Seine beiden möglichen Lagen sind bezüglich der Mitte etwas verschoben und bewirken unterschiedliche Polarisationen. Über mehrere Kristallzellen tritt infolge von Energieschwellen zwischen der elektrischen Polarisation P und der elektrischen Feldstärke E eine Hysterese auf (Bild 7b), die weitgehend der magnetischen Hysterese ähnelt (vgl. S. 118ff). Nach dem Verschwinden der äußeren elektrischen Feldstärke verbleibt eine remanente elektrische Polarisation P_r bzw. $-P_r$. Sie ist teilweise über Jahrzehnte stabil und kann den binären Werten 1 und 0 zugeordnet werden. Die Curie-Temperatur, bei welcher der piezoelektrische Effekt verschwindet, liegt für die meisten Materialien deutlich über 100 °C, bei den PZT-Keramiken sogar um 380 °C. Während frühere Materialien Sättigungsspannungen von etwa 50 V benötigen, haben die PZT-Keramiken den Vorteil, mit ca. 2 V auszukommen und damit gut der Halbleitertechnik angepaßt zu sein. Ein Vorteil aller piezoelektrischen Materialien ist ihre hohe Dielektrizitätskonstante von mehreren Tausend. Dies führt zu großen Kapazitäten auf kleinem Raum. Die Schaltzeit der Keramiken liegt bei grob 1 ns. Leider ändert sich die Hysteresekurve – anders als beim Magnetismus – mit der Anzahl der Umpolungszyklen. Dies deutet Bild 7c an. Zur Zeit sind Materialien bekannt, die etwa 10^{10} bis 10^{12} Zyklen erlauben. Es werden Werte von besser als 10^{15} angestrebt. Ferner gibt es Probleme beim Einbringen des Ferroelektikums in den Halbleiter.



2.3.3 Feldeffekttransistor

Den prinzipiellen Aufbau eines Feldeffekttransistors (FET) zeigt Bild 8a. In einem n-leitenden Halbleitermaterial sind zwei p-Gebiete eingefügt, die als Source (Quelle) und Drain (Senke) bezeichnet werden. Zwischen ihnen

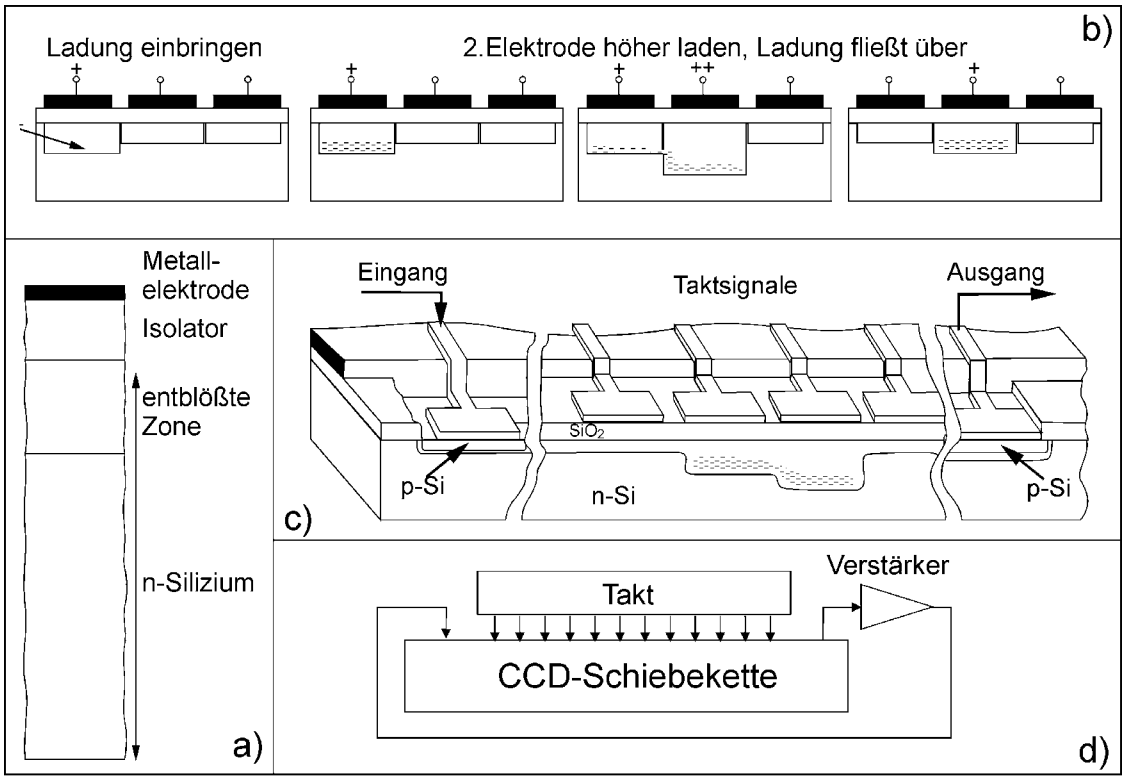
kann kein Strom fließen, da bei jeder Polung einer der pn-Übergänge gesperrt ist. Im Normalbetrieb wird nur der Drainanschluß im Sperrbetrieb gehalten. Zwischen beiden Anschlüssen und oberhalb des Silizium ist zunächst ein Isolator, z.B. SiO_2 oder Si_3N_4 aufgebracht. Darüber befindet sich ein Leiter, Metall oder Silizium, der Gate (Tor) heißt. Durch Änderung seines Potentials kann ein mehr oder weniger leitender Kanal zwischen Source und Drain erreicht werden. Mit verschiedenen Methoden ist es weiter möglich, einen leitenden oder nichtleitenden Zustand permanent zu erzwingen. Hierzu können z.B. Haftstellen im Isolator geschaffen werden, die leer oder mit Ladungsträgern belegt sein können. So erreicht der FET die beiden erforderlichen Zustände. Eine andere Möglichkeit besteht darin, ein zusätzliches, nicht nach außen geführtes, isoliertes Gate einzubringen (Bild 8b). Dann können hierauf Ladungsträger gebracht oder wieder entfernt werden. Die beiden Varianten der Speicherung sind erheblich komplizierter, als es auf den ersten Blick erscheint. Sie haben bei allen Varianten der EPROMs Bedeutung und werden dort genauer behandelt.



fet.cdr h. völz 16.8.94

2.3.4 Ladungsspeicher

Das Grundprinzip der Ladungsspeicherung im Halbleitermaterial kann mittels Bild 9a erklärt werden. Über n-leitendem Silizium befindet sich ein Isolator, meist SiO_2, und darüber eine Metall- oder Si-Elektrode. Die Struktur entspricht etwa dem Gate-Bereich eines Feldeffekt-Transistors (FET), jedoch ohne Source und Drain. Mit einer positiven Spannung an der Elektrode können die nahe gelegenen Majoritäts-Ladungsträger im Silizium vertrieben werden. So entsteht eine „entblößte“ Zone, ein sogenannter Potentialtopf. Seine Tiefe ist etwa der angelegten Spannung proportional. In den „Topf“ können Minoritäts-Ladungssträger eingebracht werden. Ihre Aufenthaltszeit ist u.a. durch thermische Effekte begrenzt. Daher entspricht der einzelne Topf weitgehend einem Kondensatorspeicher. Ähnlich dem Transport in einer Schiffsschleuse wird eingebrachte Ladung durch eine Kette mit vielen Töpfen bewegt. Bild 9b zeigt hierzu 4 getaktete Schritte. Zunächst wird der erste Topf abgesenkt. Dann wird dort die zu transportierende Ladung eingebracht. Anschließend wird der benachbarte Topf noch tiefer gemacht. Dadurch fließen die Minoritätsträger ähnlich einer Flüssigkeit dorthin. Wird nun der erste Topf wieder angehoben, so verbleibt die Ladung im mittleren Topf. Schließlich kann der Prozeß analog für den rechten Topf fortgesetzt werden. Insgesamt kann so die Ladung praktisch verlustfrei durch eine lange Kette von CCD-Zellen bewegt werden.



ccdkette.cdr h. völz 17.8.94

Entsprechend Bild 9c besitzt eine CCD-Kette außer den vielen linear angeordneten Kontakten an beiden Enden je einen pn-Übergang. Sie ähneln der Quelle (Source) bzw. dem Abfluß (Drain) beim FET. Mit dem Eingangs-pn-Übergang wird die Ladung eingebracht. Der Ausgangs-pn-Übergang wertet die ankommende Ladung aus und wandelt sie in eine Spannung zurück. Für eine Speicherung sind folglich vier Prozeßstufen erforderlich:

- *Einbringen* von Ladungen in die erste entblößte Zone (Aufzeichnung) z.B. mit einem pn-Übergang (p-Si),
- *Fortleitung* der Ladungen durch getaktete Potentiale an den aufeinander folgenden Elektroden,
- *Übernahme* der Ladungen (Wiedergabe) als Potential an einem Drain (Kollektor, p-Si),
- *Regenerieren* (Refresh) der geringen Ladungsverluste durch einen Verstärker und Rückführung der zugehörigen Ladung an den Eingang der Kette.

So ergibt sich gemäß Bild 9d eine ständig umlaufende Information, deren Menge durch die Anzahl der Töpfe gegeben ist. Es liegt also eine dynamische Speicherung vor. Derartige Speicher heißen daher Umlauf- oder Ladungskettenspeicher. Die technologische Bezeichnung für dieses Prinzip ist CCD (charge coupled device). Es hat unterschiedliche Anwendungen gefunden, wobei auch analoge Signale als genau definierte Ladungsmengen möglich sind. Beispiele sind die Speicherung eines kompletten Bildes beim Fernseher und in Kombination mit Lichtempfängern die CCD-Bildaufnahme-Matrizen bei Video-Kameras mit teilweise mehr als einer Million analoger Zellen.

Bei allen Laufzeitspeichern – es gab und gibt sie auf vielerlei Basis, z.B. Ultraschall in Quecksilberröhren bzw. Quarzblöcken, magnetische Bubble-Speicher und Schieberegister – besteht ein besonderer Zusammenhang zwischen den drei Größen:

Speicherkapazität – Zugriffszeit – Organisation.

Er wird durch das notwendige Einfügen der Refresh-Verstärker weiter modifiziert. So entstehen die vier Hauptvarianten nach Bild 10, wo auch die wichtigsten Formeln eingetragen sind. Der *Serpentinen*-Speicher ist dadurch gekennzeichnet, daß in die lange Kette mehrfach Refresh-Verstärker R eingeschaltet sind. Die *Einzelschleife* wird mit zwei Takten, je einem für den Refresh und für den Zugriff betrieben. Die Refresh-Rate wird, um Energie zu sparen, möglichst klein gehalten. Nur die Schleife, auf die zugegriffen wird, läuft für diese Zeit mit dem höheren Takt. Eine Weiterentwicklung stellt *LARAM* dar. Hier wird immer nur auf ein Register zugegriffen, während dessen alle anderen sogar vollständig ruhen. Daher ist nur ein Refresh-Verstärker notwendig. Völlig anders ist das SPS-Prinzip gestaltet. Mit einem Takt F_T werden nur die waagerechten Register um jeweils 1 Bit nach rechts verschoben. Sie erhalten also je ein neues Bit aus dem linken senkrechten Register und geben genau je eins an das rechte senkrechte Register weiter. Dann transportieren die senkrechten Register die Informationen mit einem *M*-fach höheren Takt um *M* Bit weiter. Dabei werden auch *M* Bit über den Refresh-Vestärker geführt. Im Informationsstrom sind daher die Bit der waagerechten Register ineinander verschachtelt.

2.3.5 Flipflop

Die binäre 1 und 0 kann durch Stromfluß und Nichtstromfluß in einem *Transistor* realisiert werden. Zur stabilen Erhaltung (Speicherung) der Zustände ist eine Rückkopplung mit einem zweiten Transistor erforderlich. Strukturell ist die zugehörige Schaltung nach Bild 11a recht einfach. Ihre Funktion ist jedoch schwer zu verstehen. Jeder Transistor besitzt eine S-förmige Kennlinie. Zunächst werde die Schaltung ohne die Rückkopplung, also ohne die Widerstände R_{r1} und R_{r2}, betrachtet. Dann gelten die beiden Kurven in Bild 11b, eine für die Spannungen an der Basis und am Kollektor von T_1 und die andere für Basis und Kollektor von T_2. Mit der Rückkopplung über die Widerstände R_{r1} und R_{r2} existieren drei Schnittpunkte, die als angenommen werden könnenn. Davon ist der mittlere instabil. Die kleinste Abweichung führt nämlich sofort zum Zustand „0“ oder „1“. Hierbei ist jeweils ein Transistor, T_1 oder T_2, leitend und der andere gesperrt. Der jeweilige Zustand kann aus dem Potential an den Ausgängen Q bzw. $\overline{Q}$ ständig abgeleitet werden.

Beim Einschalten eines Flipflop stellt sich immer einer der beiden stabilen Zustände spontan ein. Durch einen bewußt unsymmetrischen Aufbau ist es dabei möglich, einen bestimmten Anfangszustand zu erzwingen. Bei exakt symmetrischem Aufbau entscheidet jedoch der Zufall.

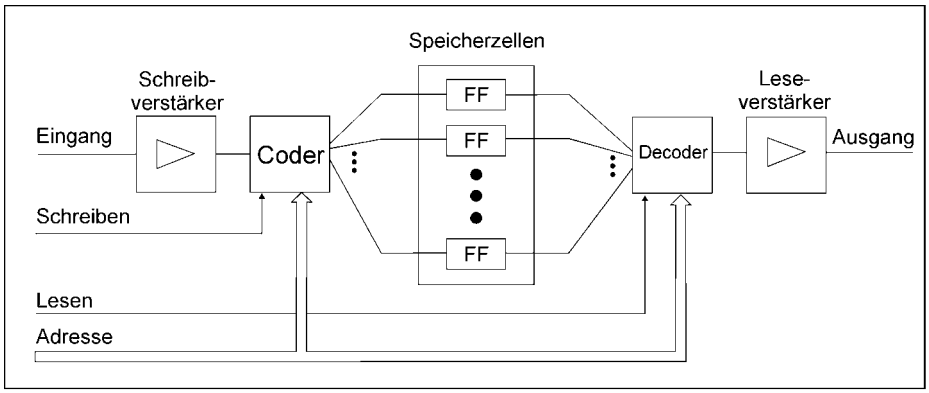
Auch bei der rückgekoppelten Schaltung kann durch eine Strom-Einspeisung an der Basis der jeweils gesperrte Transistor leitend gemacht werden. Dann tritt ein großer Spannungsabfall an seinem Kollektorwiderstand auf. Dadurch erhält der andere Transistor an seiner Basis so wenig Strom, daß er in den gesperrten Zustand übergeht und auch nach dem Abklingen der Stromeinspeisung dort verbleibt. Für eine vollständige Speicherschaltung sind also Zusatzschaltungen erforderlich, welche die zu speichernden Werte immer dem „richtigen“ Transistor zuleiten. Heute werden die Speicherzellen (Flipflop) meist von NOR- oder NAND-Gattern abgeleitet (Bild 11c und d). Dann stehen automatisch Eingänge zur Verfügung. Sie sind hier mit S von Setzen (set; Q=1) und R von Rücksetzen (reset; Q=0) bezeichnet. Allgemein existiert eine große Anzahl unterschiedlicher Flipflops. Auf sie wird auf S. 36f eingegangen. Systematisch sind alle Varianten z.B. in [VÖE] behandelt. Infolge der starken Rückkopplung erfolgt die Umschaltung zwischen den beiden Zuständen schlagartig. Bei den alten langsamen Schaltungen konnte dies noch deutlich hörbar gemacht werden. So ist lautnachbildend „Flipflop“ entstanden.

2.4 Randelektronik

Eine Speicherschaltung besteht aus vielen Speicherzellen und umfangreicher Randelektronik mit vielfältigen Aufgaben:

- Die *Adressierung* oder *Codierung* wählt einzelne oder eine bestimmte Auswahl von Speicherzellen an.
- Zum *Schreiben* (Aufzeichnen) muß der oder den angewählten Speicherzellen die gewünschte Information eingeprägt werden.
- Beim *Lesen* (Wiedergeben) wird von den Speicherzellen die dort gespeicherte Information geholt und an Ausgängen als Signal der richtigen Größe zur Verfügung gestellt.
- Zum Schreiben ist meist ein *Schreibverstärker* erforderlich, der aus den Eingangssignalen die notwendige Energie für die angewählten Speicherzellen erzeugt.
- Ein hochempfindlicher *Leseverstärker* in analoger Schaltungstechnik muß die oft sehr kleinen Signale der Speicherzellen selektieren und dann verstärken. Für ihn ist oft neben den Speicherzellen der größte Entwicklungsaufwand erforderlich.
- Häufig wird noch spezielle *Zusatz-Elektronik* integriert. So gibt es Schaltungen für Fehlererkennung oder -korrektur, Refresh, Spannungswandlung, Stromverbrauchsminderung in Pausen usw.

Den prinzipiellen Aufbau einer Speicherschaltung zeigt Bild 12. Der Coder (Codierer und Demultiplexer) und der Decoder (Decodierer, Multiplexer) realisieren die Adressierung der Speicherzellen und damit die Durchschaltung der Information gemäß dem Schreib- bzw. Lesesignal. Das Schreibsignal ermöglicht, daß Information nur zu bestimmten Zeitpunkten übernommen wird. Andernfalls könnten zufällige Änderungen am Eingang der Schaltung zu unerwünschten Veränderungen in den Speicherzellen führen. Das Lesesignal darf nur zu den Zeiten abgeleitet werden, wo der Speicherzustand stabil und eindeutig ist. Insbesondere müssen Umschaltimpulse der Codierung und Refresh-impulse für den Leseverstärker unterdrückt werden. Weiter ist es durch das Schreib- und Lesesignal möglich, den Ein- und Ausgang des Speicherschaltkreises an dieselben Anschlüsse (pin) zu legen. Dadurch wird die Pinzahl reduziert und der Betrieb vereinfacht.



Für die Anwendungen von Speichern werden unterschiedliche Wortbreiten benötigt. Sie entsprechen der Anzahl der gleichzeitig anzusprechenden Speicherzellen (Bit). Häufig ist die *Bitauswahl*, bei der jedesmal eine einzelne Speicherzelle ausgewählt wird. Als *Wortauswahl* kommen vor allem 4 (Nibble), 8 (Byte) oder 16 Speicherzellen in Betracht. Dann müssen auch entsprechend viele Eingangs-/Ausgangsleitungen verwendet werden.

Die Speicherschaltkreise befinden sich immer in einem Gehäuse mit einer begrenzten Anzahl von Anschlüssen. Insgesamt sind folgende Leitungen erforderlich:

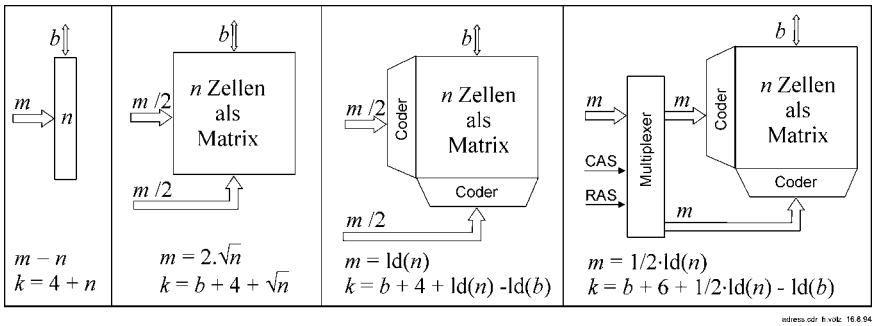
2	für die Betriebsspannung
2	für Schreiben/Lesen
b	für die Wortbreite
m	für die Adressierung

Sobald eine gewisse Speicherkapazität überschritten wird, machen den Hauptanteil die Adreßleitungen aus. In der Zukunft müssen sogar mehrere 10^9 Speicherzellen erreicht werden. Im Laufe der Entwicklung entstanden hierfür die 4 typischen Schaltungen gemäß Bild 13. In den zugehörigen Formeln treten die folgenden Parameter auf:

k	Anzahl der Kontakte (pins)
n	Anzahl der Speicherzellen, Speicherkapazität in Bit

Die *direkte* Adressierung ist nur für sehr kleine Speicher, also etwa unter 100 Bit brauchbar. Die einfache *Matrix*anordnung bietet nur etwa 10fach bessere Werte. Sie hat aber infolge der flächigen Chipstruktur dennoch für die Schaltkreisherstellung erhebliche Vorteile. Die *codierte Adreßwahl* macht davon Gebrauch, daß aus der Fülle der Speicheradressen immer nur jeweils genau eine Speicherzelle oder ein Speicherwort anzusprechen ist. Entsprechend der binären Technik können *x* Adreßleitungen genau 2^x Speicherzellen einzeln codieren. Umgekehrt gilt, daß für *n* Speicherzellen nur ld(*n*) Adreßleitungen erforderlich sind (ld ist der Zweierlogarithmus). Infolge dieses Zusammenhangs ist es nahezu belanglos, wie die Speicherzellen (eindimensional oder matrixförmig) organisiert sind. Deshalb ist im Bild 13 auch nur eine Variante

aufgenommen worden. Obwohl so bereits sehr große Speicherkapazitäten zu addressieren sind, hat sich für noch größere Kapazitäten die RAS-CAS-Variante (von row bzw. column address select) herausgebildet. Hier werden zeitlich nacheinander je die obere und untere Hälfte der Adressen dem Multiplexer zugeführt, und dort für die Spalten und Zeilen getrennt gespeichert. Erst danach erfolgt die Adressierung der Speicherzellen. Deshalb steht das 1/2 vor dem Logarithmus und die 6 bei der Addition (Bild 13). Hat der Schaltkreis die Wortbreite *b* und die Kapazität *n*, so braucht der Adreßdecoder nur *n*/*b* Adressen zu bilden. Deshalb sind in Bild 14 die Formeln für die Wortbreiten 1 und 8 Bit graphisch ausgewertet.



2.5 Einfache Speicher

In der Elektronik existieren vielerlei Schaltungen mit einem großen Anteil von Speicherschaltungen. Sie stellen einen gleitenden Übergang zu den eigentlichen kombinatorischen Schaltungen und Rechnerschaltkreisen (CPU, Arithmetik-Einheit usw.) dar. In diesem Abschnitt werden relativ einfache Strukturen mit hohem Speicheranteil behandelt.

2.5.5 Stack

Der Stack heißt auch Kellerspeicher. Er ist eine Erfindung der 50er Jahre und spielt in der Rechentechnik eine zentrale Rolle. Jeder Rechner verwendet mindestens einen Stack, zur Zwischenspeicherung von Daten des Betriebssystems. Für den Stack gilt das LIFO-Prinzip (Last In, First Out): Was zuletzt in diesen Speicher hineingegeben wurde, muß als erstes wieder herausgenommen werden. Schematisch kann er als ein reduziertes Vor-Rück-Schieberegister mit nur einem Ein- und Ausgang betrachtet werden (Bild 23 links). Wird ihm ein Eingangssignal zugeführt, so wandern alle bereits vorhandenen Werte einen FF nach rechts, und der neue Wert wird vom ersten FF übernommen. Wird aus dem Stack der zuletzt eingeschriebene Wert entnommen, so wandern die Zustände der FF um jeweils einen FF nach links. Es steht also anschließend der vorletzte Wert zur Verfügung. Besonders verständlich ist der Stack mit einem nur von oben zugänglichen Zettelkasten zu erklären. Auch ein Tellerständer, wie er in den Gaststätten üblich ist (Bild 23 rechts) ist ihm sehr ähnlich: In einem Gehäuse befindet sich eine Feder. Darauf werden die Teller gelegt. Je mehr Teller, desto stärker wird die Feder zusammengedrückt. So steht immer der zuletzt eingelegte Teller genau oben.
Die endliche Speicherkapazität begrenzt die von einem Stack aufnehmbare Datenmenge. Es muß also sein „Überlaufen“ verhindert werden.. Andererseits kann er auch leer (empty) sein und dann keine Daten mehr ausgeben. Stacks werden in der Rechentechnik nur ganz selten per Hardware realisiert. Oft wird er als RAM mit einen Vor-/Rückwärtszähler aufgebaut, häufig sogar nur per Software simuliert.

2.5.6 Cache

Der Begriff Cache kommt aus dem Englischen und kann etwa mit Versteck-Speicher übersetzt werden. Folglich besitzt der Cache vorwiegend Hilfefunktionen, die möglichst unsichtbar bleiben sollen. Da er funktionell meist mittels Software realisiert ist, werden hier nur wenige wesentliche Eigenschaften erklärt. Ausführlicher wird er u.a. in [BER, RHF], sowie bezüglich spezieller Details in [BUD] behandelt. Die Hauptanwendung des Cache ergibt sich aus der Tatsache, daß sich in den letzten zehn Jahren einerseits die CPU-Geschwindigkeit mehr als verzehnfacht hat, während bei den Speichern nur ein Geschwindigkeitsfaktor von etwa 1,7 erreicht wurde (Bild 24). Deshalb muß eine CPU oft auf Daten aus dem RAM warten. Bei einem Pentium mit 66 MHz Taktfrequenz und 60 ns DRAM sind bereits vier Wait-States (Taktzyklen der CPU) notwendig, bevor erneut auf den Speicher zugegriffen werden kann. Diesen Engpaß soll der Cache im Sinne der Speicherhierarchie von S. 13f mildern. Er muß dazu „vorausdenken“ (look ahead), was die CPU als nächstes benötigen wird und es sich schon zuvor aus dem RAM holen. Deshalb besteht ein Cache aus

- einem *Speicher* und
- dem Cache-Controller (als Programm oder Hardware), der den Datenaustausch steuert.

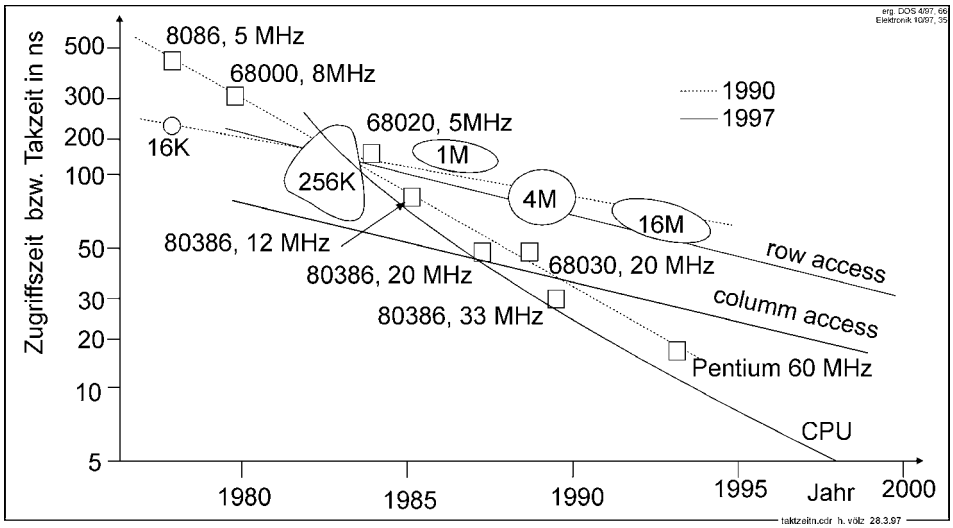
Der Speicher ist meist nicht sehr groß (16 bis 256K), dafür aber mit sehr schnellen (10 bis 30 ns) Speichern (SRAM) bestückt, die ähnlich Buffern betrieben werden. Funktionell befindet er sich z.B. zwischen der CPU und dem Rechnerbus und vermittelt so den Zugriff zum Arbeitsspeicher, zur Festplatte und zu anderen Einheiten (Bild 25a). Auch Festplatten, CD-ROM-Laufwerke und andere Einrichtungen besitzen oft intern einen Cache. Selbst DRAM-Bausteine mit Cache wurden schon marktreif realisiert.

Eine Besonderheit des Cache – vor allem bei den INTEL-Prozessoren – ist der TAG-Speicher. Er hält die höherwertigen Adreßteile der Segmentregister bereit. Zuweilen wird indiesem Zusammenhang auch fälschlicherweise von einem assoziativen Cache gesprochen.

Natürlich klappt das look ahead (Vorausschauen) auf die künftig erforderlichen Daten nicht immer. Die Trefferrate (Hitrate) ist daher ein typischer Wert für die Wirksamkeit eines Cache. Sie liegt meist über 80 %, zuweilen auch bei 95 % und mehr. Bild 25b zeigt prinzipiell, wie dieser Wert von der Größe des Caches abhängt. Einen vergleichenden, statistischen Überblick zur Wirksamkeit eines Cache gibt auch die folgende Tabelle. Daraus wird deutlich, daß der Cache vor allem eine ökonomische Variante gegenüber schnellen und daher teueren Arbeitsspeichern ist.

Betriebsart	relative Ausführungszeit
Lokaler schneller SRAM	1 (Bezugswert)
CacheTechnik 95 % Treffer	1,13
schnelle DRAM	1,53
Standard DRAM	1,92

Beim Cache besteht auch die Gefahr, daß z.B. bei Stromausfall oder Programmabsturz noch nicht alle Daten auf die Festplatte übertragen sind. Daher existiert in der Praxis die Möglichkeit unterschiedliche Modi für das Lesen und Schreiben einzustellen.



2.6 Schreibbare Speicher, RAM

Die wichtigsten Bausteine für den Arbeitsspeicher in Rechnern sind RAMs (random access memory). Sie ermöglichen gegenüber den Speicherbausteinen des nächsten Abschnittes ein ständiges Verändern ihres Inhalts. Die nun schon über Jahrzehnte anhaltende rasante Zunahme ihrer Kapazität ist prägend für die gesamte Mikroelektronik. Alle 3 bis 4 Jahre erfolgt eine Vervierfachung. Die RAMs existieren in den folgenden drei Hauptgruppen:

- *Bipolare RAMs*, also beispielsweise in ECL-, TTL-, I^2L-Technik, sind durch hohe Geschwindigkeit bei hohem Stromverbrauch und relativ kleine Kapazitäten gekennzeichnet.
- *Statische MOS-Speicher* sind meist durch die 6-Transistorenzelle gekennzeichnet. Sie sind relativ schnell, verbrauchen sehr wenig Energie, sind technologisch recht aufwendig und benötigen viel Fläche auf dem Chip.
- *Dynamische RAMs* beruhen auf der 1-Transistorzelle, führen zu den höchsten Speicherkapazitäten, erfordern dafür aber Zugeständnisse bei der Geschwindigkeit.

Entsprechend diesen Eigenschaften existieren in etwa die Einsatzgebiete von Bild 26. Die gezogenen Grenzen ändern sich natürlich im Laufe der Zeit.

Den prinzipiellen, internen Aufbau eines RAMs zeigt Bild 27a. Es ist eine Präzisierung gegenüber Bild 12 und bezieht sich auf die vorrangig verwendete Matrixorganisation. Sie wurde im Zusammenhang mit Bild 13 erklärt. Im Teilbild 27b ist die interne Anschaltung der Speicherzelle und der Schreib-Lese-Verstärker etwas detaillierter ausgeführt. Die Speicherzelle ist in diesem Fall eine statische CMOS-Variante. Auf sie und andere Typen wird im folgenden noch genauer eingegangen. Die Auswahl zwischen den verschiedenen Speicherzellen erfolgt durch vier MOS-FET. Sie realisieren gemeinsam die Ankopplung an die Bit-Leitungen B und $\overline{B}$. Wenn die Leitungen der Zeilen- und Spaltenauswahl den gültigen Pegel erhalten, werden die Q- und $\overline{Q}$-Pegel der Speicherzelle weitergeleitet. Die Bitleitungen aller Speicherzellen führen zum Schreib-Lese-Verstärker. Sein Leseteil beginnt mit einem Verstärker, der die Differenz beider Bit-Signale auswertet und verstärkt als Ausgangssignal bereitstellt. Vielfach folgt ihm noch ein Leistungsverstärker. Das Setzen (Schreiben) der angewählten Speicherzelle ist nur dann möglich, wenn die beiden MOS-FET im Schreib-Lese-Verstärker durch die Logik-Schaltung leitend sind. Dann wird das Data-In-Signal als normales und als negiertes Signal der Speicherzelle aufgezwungen.

2.6.1 SRAM

Ein statischer RAM (SRAM) erhält, so lange er mit Gleichstrom versorgt wird, seinen Speicherzustand aufrecht. Je nach der gewählten Technologie ist die Speicherzelle unterschiedlich aufgebaut. Beispiele hierzu zeigt Bild 28. Wegen der praktisch noch geringen Bedeutung wurde insbesondere die *GaAs*-Zelle nicht aufgenommen.

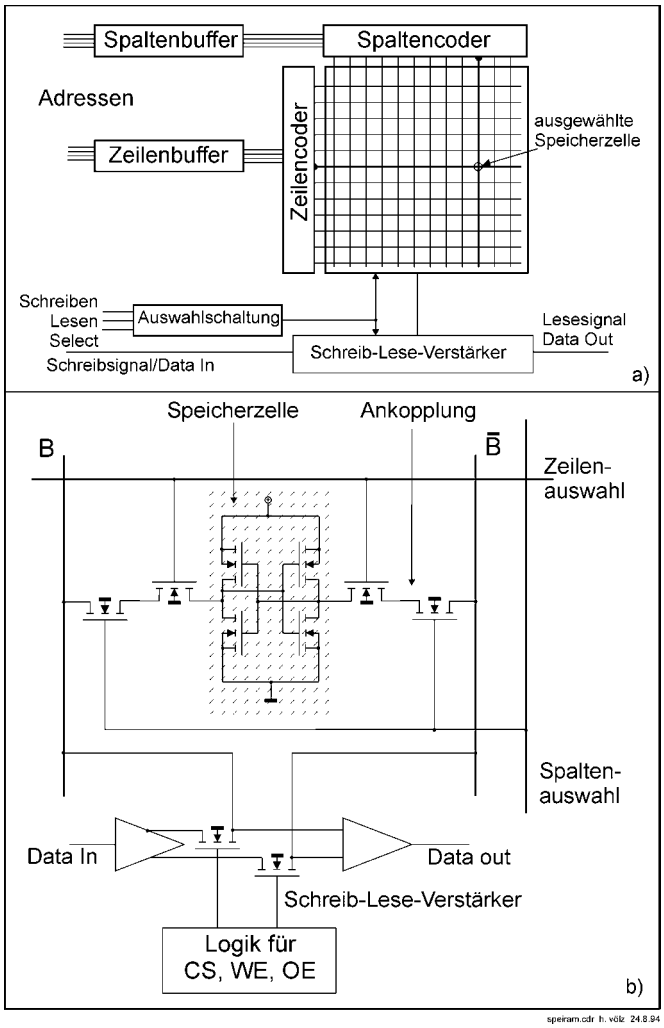
Für die *TTL*-Technik wurde die besonders einfache, klassische Variante ausgewählt. Die Mehrfach-Emitter werden hier zur zeilenweisen Wortauswahl (W) und auch zur spaltenweisen Bit-Auswahl (einschließlich Datenübergabe) verwendet. Auf die verschiedenen TTL-Technologien wurde bereits auf S. 29f eingegangen.

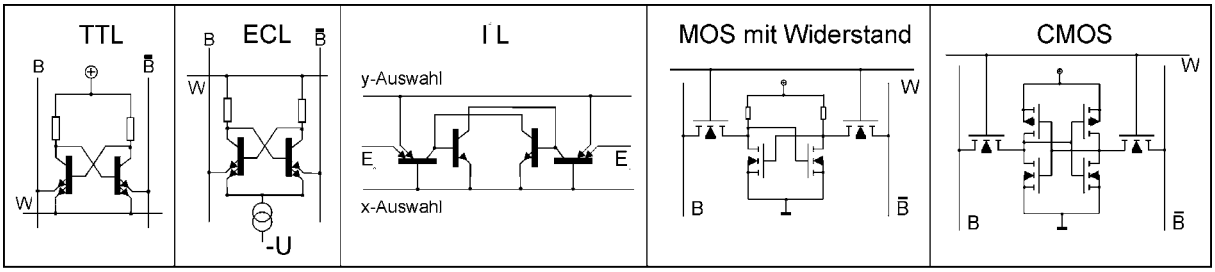
Bei der ECL-Zelle ist die Konstantstromquelle wesentlich. Sie begrenzt den Strom des Transistors unterhalb der Sättigung und ermöglicht so ein sehr schnelles Schalten.

Die I^2L-Zelle existiert eigentlich nur als Ganzheit im Halbleiter-Chip. Daher ist das gezeigte Schaltbild nur ganz formaler Natur und kann nicht die Wirkungsweise erklären.

Die einfache MOS-Technologie kann auf p- oder n-Basis betrieben werden. Hierbei befinden sich im einfachsten Fall *Arbeitswiderstände* zwischen Drain und Betriebsspannung. Sie werden jedoch häufig – ohne wesentliche Änderung der Eigenschaften – durch *FET* mit fester Gatespannung ersetzt. Hierdurch kann vielfach der Flächenbedarf der Speicherzelle reduziert werden. Außerdem läßt sich der Stromverbrauch je Zelle von µA auf nA senken.

Eine ganz wesentliche Verbesserung bei der Speicherzelle tritt dann ein, wenn die „oberen" FET komplementär zu den eigentlich aktiven, unteren FET (z.B. p- statt n-leitend) gewählt werden. So entsteht die *CMOS*-Speicherzelle. Durch diese Zusammenschaltung wird erreicht, daß immer nur zwei diagonale FET leitend sind, also links unten und rechts oben oder rechts unten und links oben. Insgesamt arbeitet dadurch diese Speicherzelle nur mit den Sperrströmen, also im pA-Bereich. Sie besitzt daher im Ruhezustand einen extrem kleinen Stromverbrauch. So können ganze Speicherchips im MBit-Bereich jahrelang mit nur einer kleinen Stütz-Batterie ihren Speicherzustand erhalten. Es existieren sogar Schaltkreise, wo diese Stütz-Batterie bereits fest in das Gehäuse des Schaltkreises integriert ist. Natürlich müssen diese Schaltkreise im Betrieb durch eine externe Stromquelle versorgt werden.





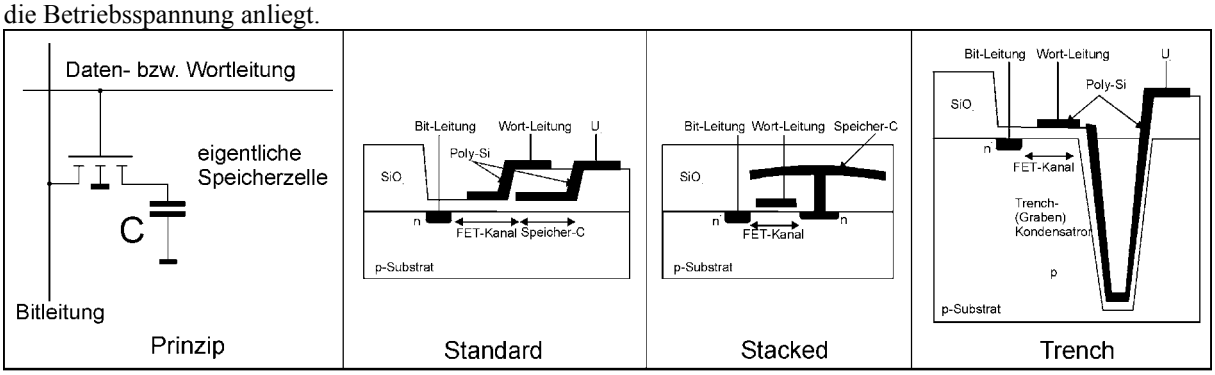
2.6.2 DRAM

Der *statische RAM* hat zwei Eigenschaften, die den Aufbau großer Speicher begrenzen:

- Er benötigt je Speicherzelle mehrere, meist 6 Transistoren und damit beachtliche *Waferfläche*.
- Abgesehen von den CMOS-Speichern benötigt er relativ *viel Energie*. So ist auch wegen der Wärmeabfuhr die Kleinheit der Speicherzelle begrenzt.

Bezüglich der Energie wurde immer wieder nach Auswegen gesucht. So ist es möglich, eine statische RAM-Zelle in den „Schlafzustand" zu versetzen. Hierzu wird die Betriebsspannung soweit herabgesetzt, daß gerade

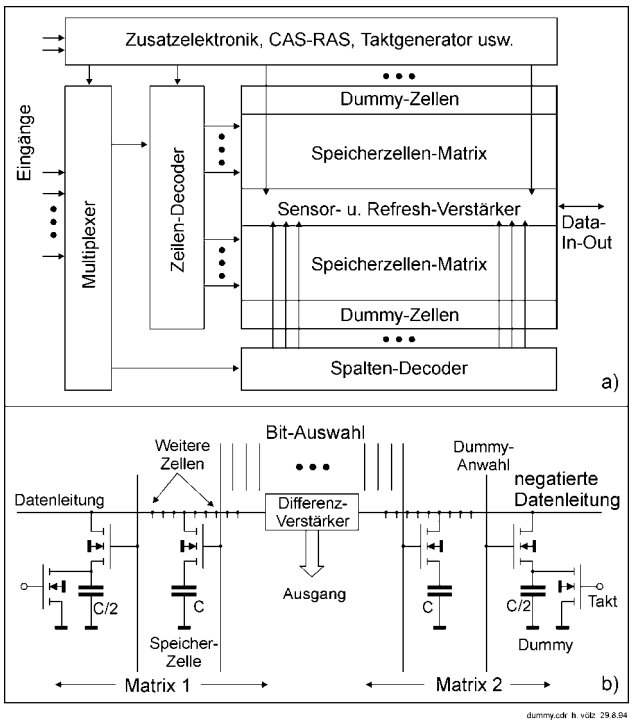
noch der aktuelle Speicherzustand stabil erhalten bleibt. Die volle Betriebsspannung wird nur dann angelegt, wenn Lese- oder Schreibzugriffe erfolgen. Für relativ kurze Zeiten im ms-Bereich kann sogar die Betriebsspannung ganz abgeschaltet werden. Der aktuelle Speicherzustand wird dann in den Schaltkapazitäten des Gate erhalten. Bei Rückkehr der Betriebsspannung nimmt dann der FF wieder den alten Zustand an. Es ist also eine Taktung der Speicherzellen erforderlich, und der Lese-Schreib-Zugriff kann nur dann erfolgen, wenn die Betriebsspannung anliegt.



dramzell.cdr h.völz 26.8.94

Alle genannten und viele weiteren Versuche haben schließlich zur *DRAM-Zelle* geführt. Ihr Prinzip zeigt Bild 29. Die Speicherung erfolgt hier nur noch im Kondensator *C*, und der Zugriff auf ihn erfolgt über den einen FET. Deshalb wird hier auch von der 1-Transistorzelle – als Gegensatz zur 6-Transitorzelle des SRAMs – gesprochen.

Jede elektrische Kapazität verliert, wie bereits auf S.30f behandelt, ständig Ladung. Deshalb ist für die 1-Transistorzelle ein ständiger Refresh – vgl. Bild 6 – erforderlich. Er besteht auch hier aus einem Lesen, Verstärken und erneutem Schreiben, ist also erheblich komplizierter als das einfache Ein- und Ausschalten der Betriebsspannung beim statischen RAM. Er muß außerdem zeitlich so ausgelegt sein, daß keine wesentliche Verzögerung für die Lese- und Schreibzyklen erfolgt.



dummy.cdr h. völz 29.8.94

Mit der Erhöhung der Speicherkapazität wurde die je Bit verfügbare Chipfläche und damit auch die elektrische Kapazität *C* immer kleiner. Heute sind nur noch 20 bis 100 fF (10^{-15} F) vorhanden. Bei einer Spannung *U* von maximal 5 V ergibt sich eine Ladung

$$Q = U \cdot C$$

von etwas über 10^{-13} C. Da die Ladung des Elektrons $1{,}6 \cdot 10^{-19}$ C beträgt, sind das nur noch eine Million Elektronen. Mit der noch stark wachsenden Speicherkapazität sinkt dieser Wert weiter und dürfte um die Jahrtausendwende, also beim GBit-Chip, nur noch bei einigen Tausend liegen. Dabei ist bereits berücksichtigt, daß etwa ab 1985 Lösungen entstanden, die bei kleiner Chipoberfläche möglichst große elektrische Kapazität erreichen lassen. Gegenüber der ursprünglichen Standard-Zelle haben sich die Stacked- (etwa Stapel, Haufen, verschachtelt) und die Trench-Technologie (Graben), durchgesetzt. Bei der *Trench*-Zelle wird durch anisotropes Ätzen ein tiefer Graben geschaffen und so für die Elektroden auch die Tiefe des Chips, die dritte Dimension, genutzt. In einigen Fällen wurde nicht nur die Kapazität, sondern zusätzlich auch der FET in die Tiefe des Grabens gelegt. So wird die Chipoberfläche optimal genutzt. Bei der *Stacked*-Methode wird die Kapazität dagegen dadurch vergrößert, daß die Elektroden flächig in die SiO_2-Isolation oberhalb des FET gelegt werden. Zuweilen kommt auch Al_2O_3 bzw. Ta_2O_5 statt SiO_2 mit den höheren Dielektrizitätskonstanten 8,5 bzw. 22 statt 3,9 zur Anwendung. Da jede Variante ihre eigene Technologie erfordert, haben sich zwei Lager herausgebildet, denen schwerpunktmäßig einerseits Fujitsu, Hitachi, Intel, NEC und Samsung beim Stacked-Typ und andererseits Motorola, Siemens, Texas Instruments und Toshiba bei der Trench-Variante angehören.

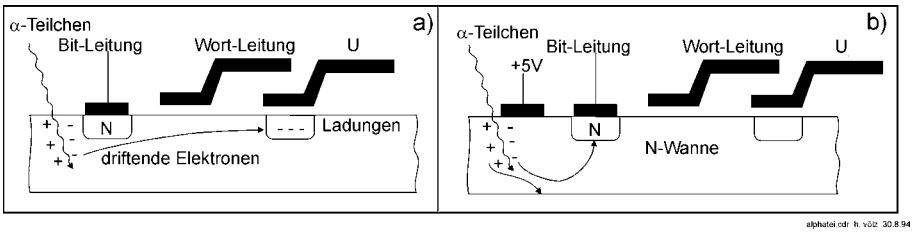
Zur Nutzung der DRAM-Zellen ist wesentlich mehr Aufwand zu treiben. Einiges hierzu demonstriert der Prinzip-Aufbau von Bild 30a. Die eigentliche Speichermatrix (grau unterlegt) besteht in diesem Fall (z.B. beim 16K-RAM) aus zwei symmetrisch angeordneten Teilmatrizen. Am Ende jeder Matrix sind entsprechend der

Spaltenanzahl zusätzliche Dummy-Zellen (von engl. Strohmann, Attrappe) angeordnet. Zwischen den beiden Hälften befindet sich der Sensor- und Refresh-Verstärker. Der Sinn dieser komplizierten Aufteilung kann mittels Bild 30b erklärt werden. Sie gilt für eine Spalte. Weiter ist in jeder Matrixhälfte nur eine Speicherzelle dargestellt. Die anderen sind durch Punkte und Geraden angedeutet. Diese Grundstruktur ist auf dem Chip also so oft vorhanden, wie es Spalten gibt. Der Sensor-Verstärker ist ein Differenzverstärker. Er selektiert immer die Potentialdifferenz zwischen der jeweils angewählten Speicherzelle auf der einen und der Dummy-Zelle auf der anderen Matrixseite. Hierdurch werden einmal vielfältige Unzulänglichkeiten unterdrückt, die sich gleichermaßen auf den gesamten Chip auswirken, und andererseits genügen selbst kleinste Potentialunterschiede zum Erkennen des aktuellen Zustands der Speicherzelle. Weiter ist es infolge der Parallelität der Spalten gemäß dem Aufbau von Bild 30b möglich, alle Speicherzellen einer Zeile mit einem Refresh gleichzeitig zu aktualisieren. Bei sehr großen Speicherschaltkreisen wird die Matrix in mehr als zwei Teilmatrizen zerlegt. Andernfalls wäre es nicht mehr möglich, die extrem kleinen Differenzsignale noch zuverlässig auszuwerten.

Durch die Zusatzeinrichtungen bei einem DRAM ergibt sich ungefähr die Flächenaufteilung der folgenden Tabelle.

Fläche	in %	Leistung	in %
Speichermatrix	50	Taktgeneratoren	60
Decoder	15	Leseverstärker	25
Taktgeneratoren	10	Ausgangstreiber	7
Leseverstärker	7	Sonstiges	8
Sonstiges	10		
Freifläche	8		

Bei der Entwicklung des 64K-DRAMs trat als ein entscheidendes Problem der strahlungsinduzierte „Soft-Error“ auf [HAL]. α-Strahlen mit einigen MeV lösen beim Durchdringen von Material, z.B. Silizium, innerhalb von wenigen ps etwa eine Million positive und negative Ladungsträger aus. Gemäß Bild 31a kann durch sie die Ladung des Kondensators verändert werden. Als Ursprung der Strahlung stellten sich geringste Mengen von α-Strahlern, z.B. Uran 238 und Thorium 235 in der Gehäuse-Keramik, im Lot usw. heraus. In sehr geringem Umfang ist auch die Höhenstrahlung wirksam. Durch höhere Reinheit der Materialien und neue Strukturen ist diese Problematik heute weitgehend überwunden. Beispiel dafür sind die N-Wanne im Bild 31b und eine Strahlenschutzschicht auf der Oberfläche der Chips. Die verbleibenden Fehler werden meist in Fit (failure in time = Fehler in 10^9 Betriebsstunden ~ 10^5 Jahre) angegeben. Für ein 4M-DRAM liegt der Wert heute unter 1000 Fit. Zur weiteren Senkung der Fehlerrate und zum Ausgleich von Strukturfehlern werden in Speicherschaltkreisen teilweise zusätzliche Speicherzellen und Algorithmen zur Fehlerkorrektur eingebaut. Dann sind Werte unter 1 Fit erreichbar (u.a. [RHF], [STI]).



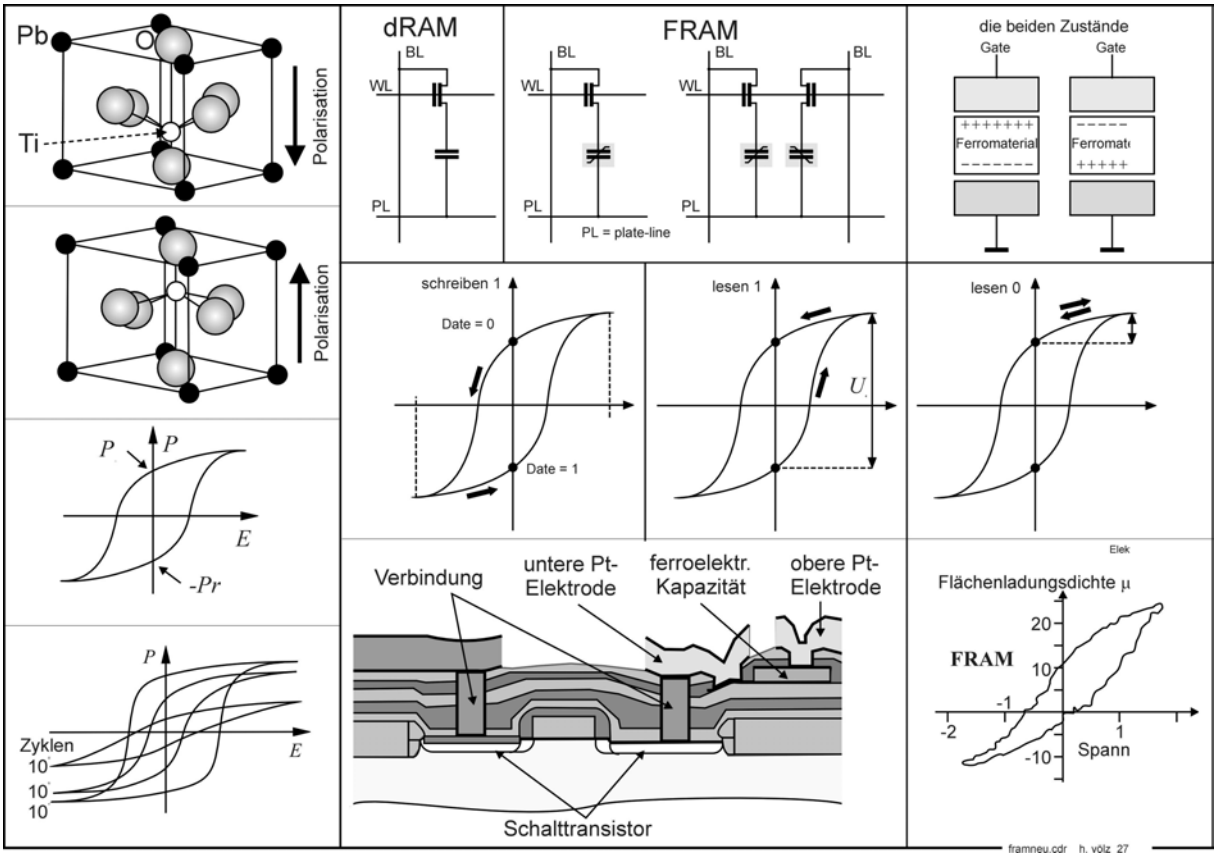
2.6.4 FRAM

Das Prinzip der ferroelektrischen Kapazität wurde bereits auf S. 31f beschrieben. Obwohl erste Speicher schon 1987 existierten, ist noch keine massenweise Produktion vorhanden. Lediglich von der Firma Ramtron werden z.Z. (Ende 94) 64 KBit-Speicher mit mehr als 10^{10} Schreibzyklen und Schreibzeiten von 400 ns angeboten. 256 KBit sind geplant [NN4]. Lizenzpartner von Ramtron sind Hitachi und Rohn. Da ein großer Markt entstehen könnte, betreiben auch andere Firmen Forschung. IBM und Texas Instruments verwenden BST (Barium-Strontiun-Titanid) und Symetix eine Y-1- bzw. Supralattice-Struktur als ferroelektrisches Material. Die einfache Prinzipschaltung einer FRAM-Zelle zeigt Bild 33a. Daraus leitet sich eine Struktur gemäß Bild 33d ab, die zugleich eine sich langfristig bewährende Variante sein könnte. Die PZT-Schichten müssen etwa 100 bis 300 nm betragen und werden epitaktisch durch CVD abgeschieden. Zwischendurch wurden auch Varianten erprobt, wie sie Bild 33b und c zeigen. Sie haben den Vorteil, daß ein direkter Zugriff zum ferroelektrischen Kondensator besteht und er daher auch getrennt zu ausgewählten Zeiten betrieben werden kann.

2.6.5 NVRAM

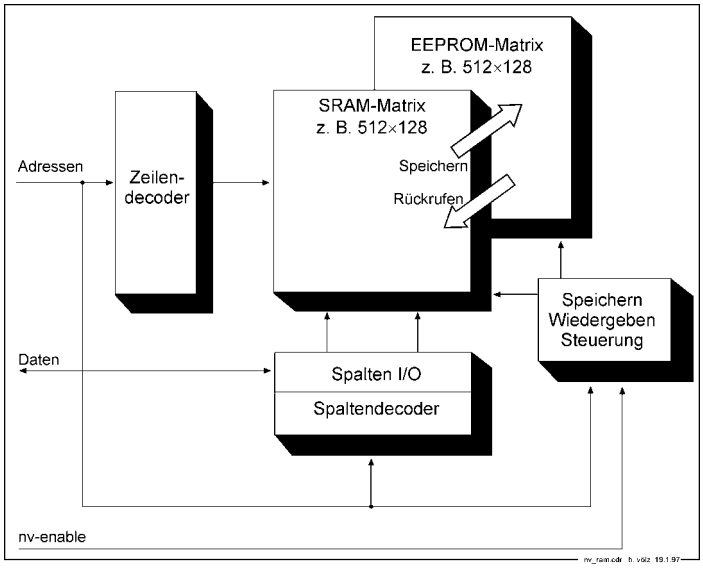
Ein non volatile RAM (NVRAM) ist ein „beständiger“ RAM. Er besteht aus zwei unterschiedlichen Speichern, einem SRAM (evtl. auch DRAM), der mit dem Rechner in der üblichen Weise zusammenarbeitet und einem kapazitätsgleichen EEPROM. Sinkt die Betriebsspannung unter einen vorgegebenen Wert oder treten andere Störungen auf, so wird der aktuelle Inhalt des RAMs in etwa 10 ms in den EEPROM übertragen. Von dort kann er jederzeit wieder zurückgelesen werden. Es gibt auch Betriebsarten, bei denen das Schreiben in den EEPROM in bestimmten Zeitabständen oder nach einer bestimmten Anzahl von Änderungen im RAM-Teil erfolgt. Infolge

der hohen Komplexität des NVRAMs existieren nur Speicher im Bereich von einigen KBit. Ihre Anwendung ist sehr selten.



2.7 Festwertspeicher

Unter dem Begriff Festwertspeicher wird eine größere Anzahl von Speichervarianten zusammengefaßt (Bild 34), die dadurch gekennzeichnet sind, daß sie zwar ständig gelesen aber nur selten oder gar nicht neu beschrieben bzw. gelöscht werden können. Ihre praktische Bedeutung ergibt sich allein schon daraus, daß 1990 weltweit rund $5 \cdot 10^8$ Festwertspeicher eingesetzt wurden. Neben den verschiedenen zur Anwendung kommenden Speicherzellen – Dioden, bipolare Transistoren und unipolare FET – sind hauptsächlich bezüglich der Programmierbarkeit die folgenden drei Typen zu unterscheiden:

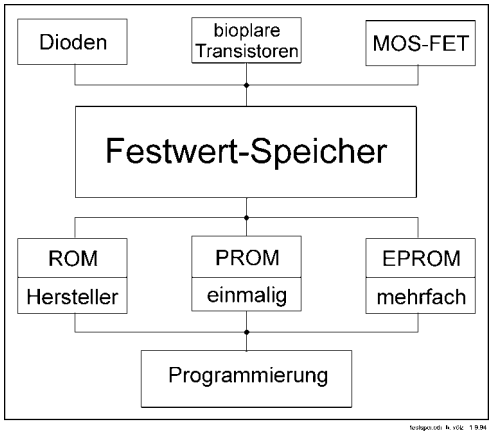


- Beim **ROM** wird der Speicherinhalt beim Herstellen der Bausteine fest eingeprägt.
- Die **PROM**s sind Bausteine, die der Anwender einmal, d.h. unveränderlich programmieren kann.
- Von den **EPROM**s existieren mit weiteren Untergruppen drei Hauptgruppen: UV-EPROM, EEPROM und Flash-RAM. Sie werden in je einem der folgenden Abschnitte getrennt behandelt.

Die Vielfalt der Festwertspeicher und ihre Abgrenzung gegenüber den RAMs folgt sowohl aus ökonomischen als auch praktischen Gesichtspunkten. Nur-Lese-Speicher (minimale Speicher gemäß S. 26f oder ROM) sind nämlich besonders einfach und wirtschaftlich herzustellen. Ihre Anwendung ist dort günstig, wo Information auf längere Sicht unverändert bereitstehen muß, also z.B. beim BIOS in Rechnern, den Fonts in Druckern und bei festen (Steuerungs-) Abläufen in Geräten. Da für ihre Entwicklung und Produktion ca. 4 Wochen erforderlich sind, kommen auch Anwendungen in Betracht, bei denen eine solche Verzögerung unwesentlich ist. Sie werden ausgetauscht, um „bugs“ zu beheben bzw. neuere oder veränderte Leistungen für ein Gerät zu realisieren (updating). Für kürzere Änderungszeiten eignen sich Varianten der EPROMs. Nur dort, wo ständig neue Information geschrieben werden muß, sind die vergleichsweise teuren RAMs erforderlich.

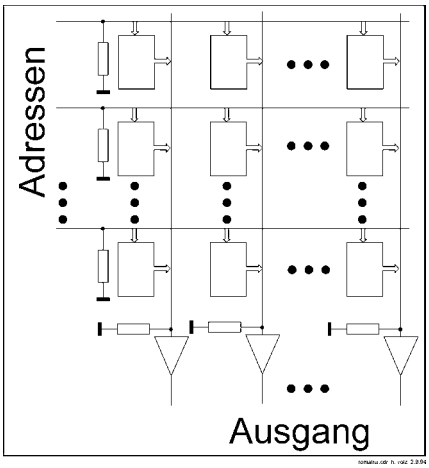
Eine recht deutliche Einteilung der Festwertspeicher ergibt sich einmal aus der Anzahl der möglichen Umprogrammierungen und zum anderen aus der Zeit, in der die neue Information eingebracht werden kann (Bild 35a). Eine andere Klassifizierung berücksichtigt auch die Speicherkapazität (Bild 35b).

Die ROMs sind genauso wie die RAMs matrixförmig aufgebaut (Bild 30a). An den Knotenpunkten befinden sich jedoch statt der Speicherzellen recht einfache Bauelemente, die eine mehr oder weniger feste Verbindung realisieren. Ferner sind Festwertspeicher vorwiegend in Byte- oder Wortbreite organisiert. So entsteht der Grundaufbau gemäß Bild 36. Die dort grau unterlegten Zellen realisieren den Speicherinhalt der Festwertspeicher. Ihr Aufbau und ihre Eigenschaften werden im folgenden behandelt.



2.7.1 ROM

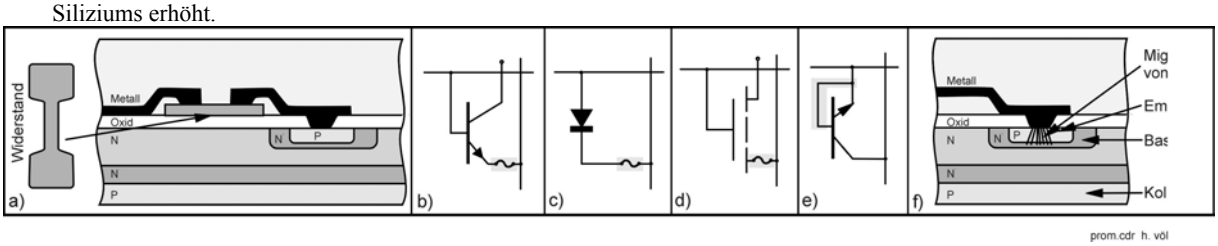
Die Speicherinhalte der ROMs werden direkt bei der Produktion erzeugt. Vorwiegend wird dazu der technologische Ablauf und die Struktur der Halbleiterschaltkreise so gestaltet, daß die Schaltkreise nahezu vollständig auf Lager vorgefertigt werden können. Der eigentliche Informationsinhalt wird dann mit der letzten Maske, die individuell zu schaffen ist, erzeugt. Deshalb weicht der Fertigungsprozeß der ROM-Schaltkreise etwas von der üblichen Maskenabfolge anderer Schaltkreise ab, die ja vor allem zur Erzielung der bestmöglichen Eigenschaften gestaltet wird. Für das Prinzip der letzten Maske sind vor allem die Möglichkeiten von Bild 37 bekannt. Die Stellen, wo die letzte Maske wirksam wird, sind darin hervorgehoben. Es existieren sowohl bipolare als auch unipolare Varianten. Durch die einfache Struktur der ROM-Speicherzelle ergeben sich ökonomische Vorteile, ein minimaler Flächenbedarf, große Speicherkapazitäten und beim bipolaren Aufbau, z.B. mit Dioden, auch kleine Zugriffszeiten. Außerdem werden ROMs nahezu als einzige Bausteine vollständig getestet. Ihr Haupt*nachteil* ist die etwa 4 Wochen dauernde Entwicklungs- und Produktionszeit. Infolge der hohen Akzeptanz von Videospielen und durch die Einführung neuer Produktionstechniken „Late Implantation“ sind in den letzten Jahren sowohl die Preise als auch die Entwicklungszeiten erheblich gesunken. So könnten künftig die ROMs bis zu einem fünfmaligen Austausch deutliche Vorteile gegenüber den anderen Festwertspeichern erlangen.



2.7.2 PROM

Heute werden unter PROMs jene Speicher verstanden, die der Anwender einmalig programmieren kann. Dazu gehören auch die OTP-ROMs (one time programmable). Grundsätzlich stehen drei Varianten zur Verfügung:

- **FS** von *fusable link*, also Schmelzsicherung. Hierbei wird in einem Stromkreis das als Speicherzelle dienende Element im Sinne einer Sicherung eingebaut, die bei der Programmierung zerstört wird.
- **AIM** von *avalanche induced migration* bedeutet soviel wie erzeugter *Material-Transport (Migration)* durch einen lawinenartigen Ladungsdurchbruch.
- **Antifuse** ist die Umkehrung des ersten Verfahrens FS. Zwischen zwei Metallisierungsschichten befinden sich kleine zylindrische Löcher, die mit amorphen Silizium ausgefüllt sind. Durch Stromstöße wird die Leitfähigkeit des amorphen Siliziums erhöht.



Die häufigste Struktur einer „Sicherung“ zeigt Bild 38a. Die größeren endständigen Flächen dienen der Kontaktierung, während der schmale Teil bei der Programmierung zerstört (verdampft) wird. Das Material ist meist NiCr – schmilzt bei 1450 °C – von etwa 0,1 × 0,2 mm^2 und 20 nm Dicke mit einem Widerstand von ca. 3 kΩ. Bei der Programmierung müssen die Stromstärke und die Impulslänge genau eingehalten werden. Sonst kann sich die Sicherung wieder zurückbilden oder das verdampfte Material anderweitig störend niederschlagen. Die typische Leistung beträgt 1,5 W/mm^2. Sie wird in mehreren Einzelimpulsen von etwa 20 mA zugeführt. Daher beträgt die Programmierzeit wenige Sekunden. Weil die Kontaktierung des NiCr mit dem Silizium etwas schwierig ist, sind einige Hersteller zu Titan-Wolfram, Platinsilizid oder polykristallinem Silizium übergegangen.

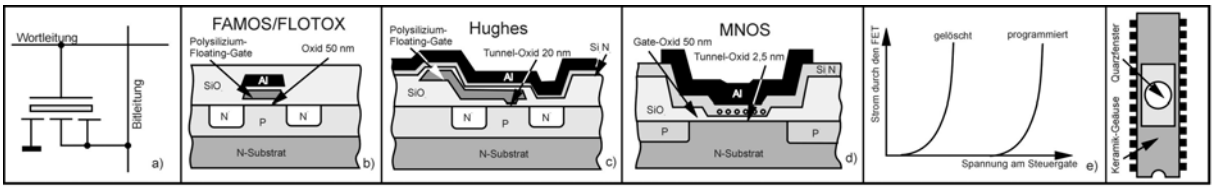
Die *Sicherung* kann in recht unterschiedliche Schaltungen eingefügt werden. Beispiele zeigen die Bilder 38b bis d als grau hinterlegte Stellen. Beim bipolaren Transistor und beim FET wird der unverbunden gezeichnete Anschluß nur bei der Programmierung zur besseren Steuerung des Stromflusses benutzt. Der rechte Teil von Bild 38a gilt für die Schaltung mit dem Transistor (b).

Die AIM-Zelle (Bild 38e und f) geht von einem Transistor aus, dessen Emitter mit Aluminium kontaktiert ist. Die Basis ist nicht nach außen geführt, sondern ist lediglich als Wanne ausgebildet. Die Strecke Kollektor-Basis-Emitter ist daher zunächst für beide Stromrichtungen gesperrt. Zur Programmierung wird eine Spannung so zwischen Kollektor und Emitter angelegt, daß die Teilstrecke Kollektor-Basis leitend wird. Die Spannung wird bis zum Durchbruch der Emitter-Basis-Strecke erhöht. Bei wiederholten kurzzeitigen Impulsen (ca. 200 mA; 7 µs) wandert das Aluminium des Kontaktes in die Strecke Emitter-Basis und überbrückt sie leitend. Damit ist die Zelle programmiert, was im Ersatzschaltbild 38e durch die grau hinterlegte Brücke angedeutet ist.

2.7.3 EPROM

Die EPROMs sind nach der einfachen Schaltung von Bild 39a aufgebaut und existieren zumindest in den drei Varianten des Bildes 39b bis d. Ihre besonderen Eigenschaften sind:

- Zwischen dem Steuergate und dem Kanal eines FET werden Ladungen festgehalten, die den jeweiligen Speicherzustand kennzeichnen. Dies kann durch ein zweites, frei schwebendes Gate (floating gate) oder durch mikroskopische Haftstellen an Grenzflächen zwischen Gate und Kanal erfolgen.
- Das Programmieren erfolgt durch Tunnelung von Ladungsträgern zum schwebenden Gate oder zu den Haftstellen.
- Das Löschen erfolgt global für alle Speicherzellen mit UV-Licht und dauert etwa 20 Minuten. Daher ist ein teures keramisches Gehäuse mit Quarzfenster erforderlich (Bild 39f).



eprom.cdr !

Den Grundaufbau einer FAMOS-Zelle (floating gate avalanche injektion MOS) zeigt Bild 39b. INTEL nennt seine entsprechenden Speicherzellen FLOTOX (floating gate tunnel oxid). Eine etwas abgewandelte Zelle der Firma Hughes zeigt Bild 39c. In beiden Zellen existiert ein „floating gate". Es ist vollständig isoliert, hat also auch keine Anschlüsse nach außen. Die beiden Speicherzustände des Gate sind dadurch gekennzeichnet, daß es ladungsfrei oder geladen sein kann. Hierdurch ergeben sich die beiden Kennlinien von Bild 39e, wobei die Abszisse der Spannung am Steuergate entspricht. Bei der Anwahl der Zelle – mittlere Spannung am Steuergate – existieren so die beiden Zustände leitend oder gesperrt. Die Kapazität des floating gate liegt ähnlich wie beim DRAM bei 100 fF, die Ladung beträgt etwa ein pC, was wieder nur wenigen Millionen Elektronen entspricht. Anders als beim DRAM muß die Ladung hier jedoch über viele Jahre erhalten bleiben. Es wird ein Ladungsverlust von nur 10 % innerhalb von 10 Jahren – alle 10 Minuten ein Elektron – erreicht. Das ist ein Leckstrom von rund 10^{-22} A. Darin ist auch die Wirkung von α-Strahlung enthalten. Durch den Aufbau gelangen die freigesetzten Ladungspaare wesentlich schwerer zum floating gate.

Zur *Programmierung* von EPROMs wird eine Spannung um 12 V zwischen Source und Drain gelegt. Hierdurch nehmen die Ladungsträger im Kanal soviel Energie auf, daß sie zum floating gate tunneln können. Ab 1993 wird diese Schreibspannung durch spezielle Wandler auf dem Chip aus den üblichen 5V erzeugt. Beim Programmieren wird das Durchbruchgebiet stark erwärmt. Deshalb werden Impulse von 10 ms, seit etwa 1989 von 0,1 ms, angewendet und nach jedem Impuls die Speicherzelle des EPROMs gewechselt. Infolge der hohen Belastung kann ein EPROM nur 10^4- bis 10^6-mal umprogrammiert werden.

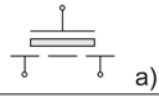
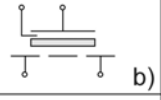
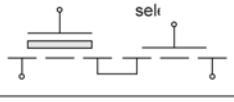
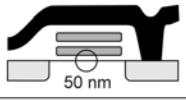
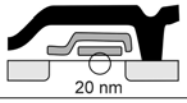
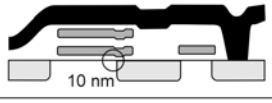
Das *Löschen* des EPROMs erfolgt mit UV-Licht. Hierdurch nehmen die Ladungsträger auf dem floating gate soviel Energie auf, daß sie zum leitend gewordenen Silizium des Kanals abwandern können. Das floating gate wird also entladen. Um unbeabsichtigtes Entladen zu vermeiden, muß das Quarz-Fenster nach der Programmierung lichtdicht verschlossen werden. Selbst mittleres Sonnenlicht kann bereits den Verlust einzelner Bit bewirken.

Eine etwas abweichende Struktur besitzt das MNOS-EPROM (metal nitrid oxid semiconductor) von Bild 39d. Die Kristall-Gitter von SiO_2 und Si_3O_4 sind so unterschiedlich, daß sich an ihrer Grenze viele Haftstellen ausbilden. Hier können sich ähnlich wie beim floating gate Elektronen anlagern.

Die Vorteile von EPROMs liegen bei der Entwicklung. Mit der zu verwendenden Software kann bis zur letzten Minute gewartet werden. Sie kann auch später noch leicht, schnell und teilweise sogar vor Ort geändert werden. Ihre *Nachteile* sind das schwere und teure Keramik-Gehäuse, das umständliche Löschen (20 min) und das Herausnehmen aus dem Gerät zum Umprogrammieren. In den letzten Jahren hat sich die Produktion auf die Flash-RAMs (s.u.) verlagert. Dadurch steigt der Preis der EPROMs bereits wieder. Heute ist daher ihr Einsatz nur noch dann sinnvoll, wenn mehr als 3mal umprogrammiert wird. In der Praxis werden EPROMs ohnehin meist nur wie Masken-ROMs ausgetauscht. Es hat den Anschein, daß der EPROM in einigen Jahren als Exote betrachtet werden wird. Die geschichtliche Entwicklung der EPROMs zeigt Bild 40.

2.7.4 EEPROM und Flash-RAM

Wird der Abstand vom floating gate zum Kanal wesentlich verringert, so ist es möglich, mittels Tunnelung von heißen Elektronen eine *elektrische* Löschung zu erreichen. Das langsame und aufwendige Löschen mit UV kann also entfallen. So entstand der elektrisch löschbare EEPROM = E^2PROM. Dennoch sind weiterhin hohe, impulsförmige Spannungen notwendig. Da sich durch den kleinen Gate-Abstand die Leckströme vergrößern, muß ein zweiter FET mit dem select gate in Reihe geschaltet werden (Bild 41c, auch FLOTOX-Zelle genannt). Er ermöglicht es, das floating-gate-system von den anderen Zellen ab- bzw. anzukoppeln. Dadurch kann sogar jede Speicherzelle einzeln geschrieben und gelöscht werden. Natürlich wird so eine deutlich größere Fläche für die Speicherzelle benötigt.

Name	UV-EPROM	Flash-RAM	EEPROM
Schaltbild			
Struktur			
Löschen	UV	Tunnelung	Tunnelun
Schreiben	heiße Elektronen	heiße Elektronen	Tunnelun

Eine Vereinfachung des EEPROMs stellt der 1988 entwickelte Flash-RAM – auch Flash-EPROM oder Flash-EEPROM genannt – dar (Bild 41). Die Bezeichnung Flash (engl. Blitz) soll die Möglichkeit des schnellen Löschens gegenüber dem UV-EPROM ausdrücken. Durch eine andere Formgebung dient das Steuergate zugleich als select gate.

Die *Vorteile* des Flash-RAMs bestehen vor allem darin, daß er im Gerät programmierbar ist, typischerweise 10^6-mal umprogrammierbar ist, zum Löschen keine zusätzlichen Geräte benötigt und meist auf dem Schaltkreis eingebettete Lösch- (20 V; 0,2 bis 1 s) und Programmieralgorithmen (z.B. 20 V, 0,1 ms) verwendet werden. Um die Lösch-Schreibzyklen gering zu halten, wird intern zuweilen auch zwischen gültigen und ungültigen Daten unterschieden. Eine Löschung erfolgt nur dann, wenn zuviel ungültige Daten erreicht sind.

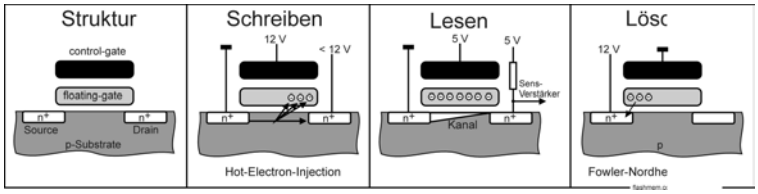
Natürlich hat der Flash-RAM auch *Nachteile*. Einmal ist er etwa 8mal so teuer wie ein Masken-ROM. Zum anderen können nur alle Speicherzellen oder zumindest eine größere Gruppe (Sektorlöschen) gemeinsam und nicht so schnell wie beim EEPROM gelöscht werden. Weiter ist auch das Schreiben deutlich langsamer. So ergibt sich in Ergänzung zu Bild 41 die folgende Aufstellung.

	UV-EPROM	Flash-RAM	EEPROM
Löschen	20 min	1 s	5 ms/Bit
Schreiben	100 ms	10 ms	5 ms

In Bild 42 sind für das Flash-RAM die einzelnen Speicherprozesse veranschaulicht.
Wegen der veränderten Dicke des Oxides werden die Zellen auch als ETOX (EPROM Tunnel Oxide) bezeichnet.
Wahrscheinlich behauptet sich von allen EPROM-Typen langfristig allein der Flash-RAM. Hierauf weist u.a. die folgende Umsatz-Tabelle von 1993 hin.

Jahr	1989	1990	1991	1992	1993	1994	1995	1996
10^7 $	1,2	3,5	13	32	68	90	110	140

Abschließend sei noch darauf hingewiesen, daß einige Firmen versuchen, in einer Speicherzelle mittels der althergebrachten Multileveltechnik mehr als ein Bit unterzubringen. Die Zelle hat dann nicht nur die Pegel 0 und 1, sondern mehrere, unterscheidbare Spannungszwischenwerte. Bei der thermischen Unsicherheit von etwa 25 mV und einem Hub von 3 bis 4 Volt dürften aber kaum jemals 16 Pegel, also 4 Bit, mit hinreichender Sicherheit zu realisieren sein.



2.8 Ergänzungen

2.8.1 Assoziativ-Speicher

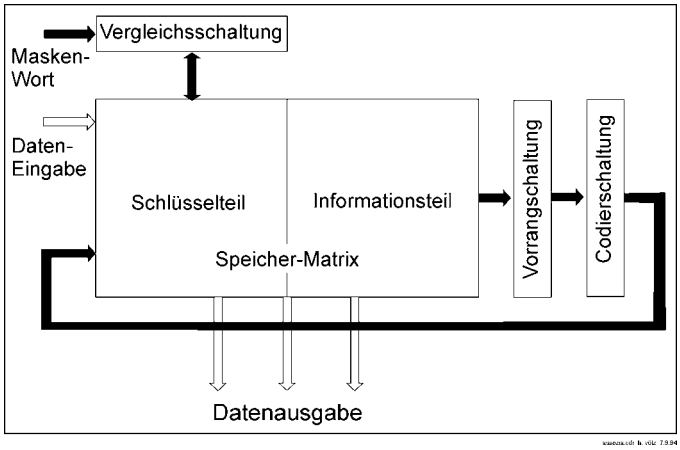
Die bisherigen elektronischen Speicher sind durch zwei typische Operationen gekennzeichnet:

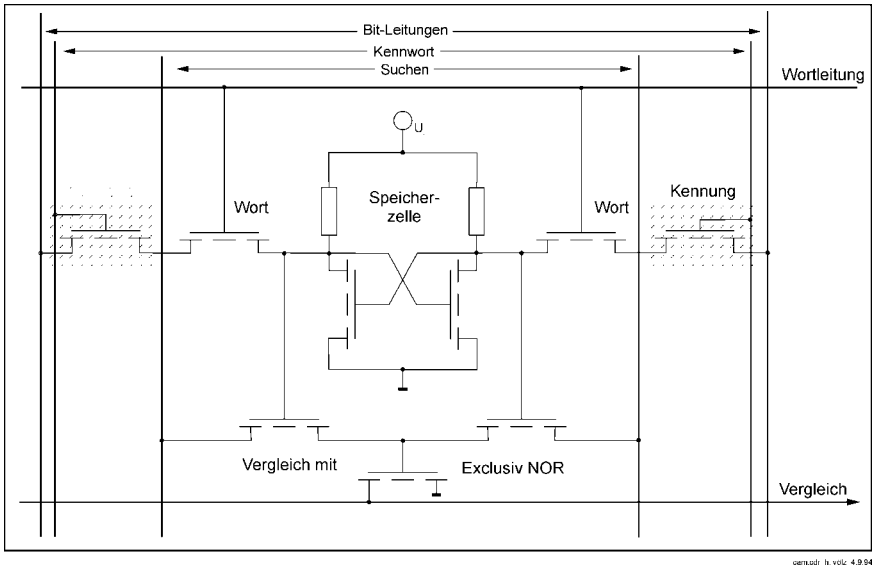
- Durch *Adressierung* wird eine Speicherzelle ausgewählt. Dies ist ein Auswahlvorgang, der zum physikalischen Ort der Speicherzelle führt.
- Neue Information wird an diesen Ort *geschrieben*; die dort befindliche Information wird *gelesen*.

Der Speicherinhalt hat also keinen Einfluß auf den Ablauf. Wenn etwas über den Speicherinhalt ausgesagt werden soll, muß er zuvor gelesen werden. Wird also eine bestimmt Information gesucht, so muß daher zuweilen der gesamte Speicher gelesen werden. Natürlich gibt es spezielle Such-Programme, die diesen Prozeß verkürzen. Dennoch sind die Such- und damit oft verbundenen Sortierprozesse sehr zeitaufwendig. Daher liegt es nahe, Speicher zu entwickeln, die nicht über ihre Adresse sondern bezüglich ihres Inhalts abgefragt werden. Das sind die CAM (content addressable memory), oder wie die „deutsche“ Bezeichnung lautet, die Assoziativ-Speicher. Ihr Prinzip ist seit langem bekannt, hat sich aber bisher praktisch nicht etablieren können. Um ihre Funktion zu erklären, sei vom einfachen (willkürlich gewählten) Auszug eines Flugplanes ausgegangen. Er möge in der folgenden Weise zeilenweise in einem Speicher stehen:

Abflug	Ort	Ziel	Marke
08.10	Berlin	Hamburg	*
08.45	Frankfurt	Berlin	
08.50	Hamburg	Köln	
09.10	Berlin	Hamburg	*
09.25	Berlin	Frankfurt	(*)
09.35	Köln	Berlin	

Möchte nun jemand wissen, wann er am Vormittag von Berlin nach Hamburg fliegen kann, dann sollte ein wirklich assoziativer Speicher, sofort die beiden Zeiten 8.10 und 9.10 ausgegeben „*“, ohne daß auch nur die anderen gelesen werden müßten. Bei der Frage wohin ich von Berlin aus am Vormittag fliegen kann, müßte die Antwort lauten: 8.10 nach Hamburg, 9.10 nach Hamburg und 9.25 nach Frankfurt „*“+ (*). Ein echt assoziativer Speicher sollte beides ohne jegliche zusätzliche Software ermöglichen. Den zugehörigen Aufbau zeigt Bild 43. Es existiert die hellgrau unterlegte Speichermatrix, deren Zellen (siehe unten) erheblich komplizierter aufgebaut sind. Die Zeile der Matrix besteht aus zwei Teilen, dem Schlüsselteil und dem Informationsteil. Der Schlüsselteil entspricht den Eingaben und der Informationsteil den Ausgaben. Nun hat aber das obige Beispiel gezeigt, daß diese Zuordnung auch verändert wird. Deshalb wird von einer Maske für die Eingaben gesprochen. Sie ist eine Speicherzeile mit 1 Bit Tiefe, aber von der Länge des Schlüsselteils oder der ganzen Speichermatrix. Als Vergleichsoperation wirkt sie bitweise parallel und gleichzeitig über alle Zeilen der Speichermatrix. Sie ist zwar binär, also auf die einzelnen Bit bezogen, kennt aber drei Zustände in der jeweiligen Spalte:





0 das Bit im Schlüsselteil des Speichers ist *nicht gesetzt*.
1 das Bit im Schlüsselteil des Speichers ist *gesetzt*.
d das Bit im Schlüsselteil Spalte des Speichers interessiert nicht, *don't care*.

Das Ergebnis der Vergleichsoperation zwischen Maske und allen Speicherzeilen wird in einem speziellen Zeilen-Bit der Vorrangschaltung gespeichert. In der obigen Tabelle steht dafür die Marke „*“. Über die Codierschaltung können dann die entsprechenden Zeileninhalte in einer wählbaren Reihenfolge ausgegeben werden. Die Dateneingabe und die Datenänderung des Speichers erfolgt nach ähnlichen Kriterien.

Die Speicherzelle eines Assoziativspeichers muß also neben der einzelnen (üblichen) Speicherzelle zusätzliche Bauelemente für die Vergleichsoperationen besitzen. Auch für sie ist ähnlich wie für die

Speicherzellen selbst, eine horizontale und vertikale Verknüpfung notwendig. So entsteht die Beispielschaltung von Bild 44.

Prinzipiell verfügen also assoziative Speicher über Möglichkeiten, die bei üblichen Speichern nur mit erheblichem Softwareaufwand zu realisieren sind und daher natürlich auch sehr viel langsamer ablaufen. Hinzu kommt, daß diese Operationen oft benötigt werden, z.B. in Datenbanken, bei der Textverarbeitung usw. Daher ist es verwunderlich, daß sich derartige Speicher bisher nicht durchsetzen konnten. Der Grund dafür ist der großen Aufwand der Speicherzelle und die sich dadurch ergebende vergleichsweise geringe Speicherkapazität. So wurden im Zeitraum von 1983 bis 1989 für spezielle LISP- und PROLOG-Rechner nur Kapazitäten bis 20 KBit realisiert. Hier könnte in den nächsten Jahren eine Änderung eintreten.

2.8.2 Geschichte

Bild 45 zeigt in einem groben Überblick die Entwicklung der typischen technischen Daten von Halbleiterspeichern. Hierzu wurden die Daten nach Asai [ASA] ergänzt und erweitert. Sie zeigen ein kontinuierliches Wachstum, welches sich nach Auffassung vieler Experten bis mindestens zum Jahre 2005 fortsetzt. Die zwischenzeitlichen Abweichungen von den Geraden sind so gering, daß sie hier problemlos vernachlässigt werden konnten. Natürlich gab es zu jedem Zeitpunkt auch Speicher mit „schlechteren" Parametern. Auch sie wurden nicht eingetragen. Eine weitere Analyse dieser Entwicklung zeigt, daß die Fortschritte der Technik zu etwa gleichen Teilen von den folgenden Einflüssen bestimmt sind:

- Verkleinerung der Dimensionen einer Zelle.
- Vergrößerung der Chipfläche.
- Neue, intelligentere (Cleverness) Prinzipien.

Die wichtigsten Zeiten zur Einführung der jeweils leistungsfähigeren Speicher zeigt die folgende Tabelle mit einer mittleren Unsicherheit von etwa einem Jahr. In mehreren Fällen konnte trotzdem kein hinreichend genaues Jahr ermittelt werden. Dann wurden die Stellen frei gelassen.

	1K	4K	16K	64K	256K	1M	4M
SRAM	1969	1977	1979	1981	1984	1987	1991
DRAM	1970	1973	1977	1979	1981	1984	1988
ROM				1977	1981		
EPROM	1975	1980	1983	1985	1987	1989	1992
EEPROM			1982		1984		

Mittels Bild 46 ist es möglich, abzuschätzen, wie sich in etwa die Preisrelationen zwischen Festplatte, DRAM und Flash-RAM verhalten. Z.Z. (1993) liegen die Kreuzungspunkte im MByte-Bereich. Die absoluten Bitpreise liegen für 1995 bei DRAMs um $4 \cdot 10^{-6}$ und bei Festplatten um 10^{-7} DM/Bit. Das weltweite Produktionsvolumen bei Speichern im Jahr 1990 zeigt die folgende Tabelle.

Speichertyp	Fest-platte	digital-optisch	Magnet-band	Diskette	DRAM	SRAM	E-PROM
Produktion in 10^9 $	26	1,1	2,8	25	30	5	3

2.8.3 Grenzen

Die Betrachtungen des vorigen Abschnittes und die Analysen auf S. 17ff legen es nahe, konkreter auf die Grenzen der Halbleiterspeicher einzugehen. Hierzu gibt es eine ausführliche Arbeit von Folberth aus dem Jahre 1983 [FOL], die auch noch heute Gültigkeit hat. Stark vereinfacht existieren drei Grenzprobleme:

- Die g*eometrische* Grenze ist durch die kleinstmöglichen Abmessungen der Transistoren bzw. Gatter bedingt.
- Bezüglich der *Signalausbreitung* wird die Größe des Chips infolge der Laufzeit der Signale zu einer Geschwindigkeitsgrenze.
- Die Grenze für die Wärmeabfuhr wird als *thermische* Grenze bezeichnet.

Ein Diagramm aus Integrationsdichte und möglicher Taktfrequenz zeigt hierzu Bild 47. Die *geometrische* Grenze ist z.B. dadurch bestimmt, daß eine Raumladungszone im Transistor kaum kleiner als 30 nm werden kann. Der Kanal eines FET läßt sich vielleicht auf 100 nm reduzieren. Geht man unter diese Werte, so macht sich das statistische Auftreten der Ladungsträger störend bemerkbar. Daher ergibt sich die mögliche Gatterdichte von einigen 10^7 pro cm^2. Die *thermische* Grenze hängt mit der Kühlung zusammen. Bei Luftkühlung ist kaum mehr als 1 W/cm^2 erreichbar. Selbst bei einer effektiven Flüssigkeitskühlung (siedende Flüssigkeiten) stellt 20 W/cm^2 die Grenze dar. Bezüglich der Laufzeit sind die Probleme komplizierter. Hier geht nämlich das Geschwindigkeits-Leistungs-Produkt ein. Es liegt grob um 10^{-12} J/Bit. Werden die Gatter so dicht gepackt, wie es die geometrische Grenze erlaubt, dann sind wegen des Wärmeumsatzes bestenfalls Verzögerungen um 50 ns möglich. Sie sind jedoch schon einige Jahre üblich. Eine Senkung ist vor allem durch kleineren Spannungshub möglich. Er muß aber immer hinreichend hoch über den thermisch bedingten 25 mV (bei Zimmertemperatur) bleiben. Folglich müssen die Gatter lockerer gepackt, wenn schnellere Bauelemente entstehen sollen. Dann sind prinzipiell Schaltzeiten bis hinab zu 25 ps möglich. Andererseits muß auch beachtet werden, daß die Ausbreitungsgeschwindigkeit auf dem Chip endlich, aber stets kleiner als Lichtgeschwindigkeit ist. So begrenzt die Größe eines Chips die Taktfrequenz um einige hundert MHz. Bei richtiger Ausnutzung von Laufzeiteffekten kann sie auch teilweise überschritten werden.