Universität des Saarlandes
Lehrstuhl für Mikroelektronik

Mikroelektronik I
Wintersemester

Grundlagen der Mikroelektronischen Bauelemente und Schaltungen

Lehrstuhl für Mikroelektronik
Universität des Saarlandes
Postfach 151150
66041 Saarbrücken

Dies ist die erste modifizierte Version des Vorlesungsskriptes Mikroelektronik I. Falls Sie noch Fehler bzw. Ungereimtheiten oder Schwachstellen anderer Art entdecken, wären wir Ihnen für eine Rückmeldung darüber sehr dankbar. Für konstruktive Vorschläge zur Verbesserung des Skriptes haben wir immer ein offenes Ohr; und schließlich haben von einem guten Skript vor allem Sie, die Studierenden, die Vorteile.

Inhaltsverzeichnis

1 Einleitung

Die Mikroelektronik ist Dienstleister für die unterschiedlichsten Branchen. Von ihren Fortschritten profitieren die Computer- und die Unterhaltungsindustrie, die Luft- und Raumfahrttechnik, die Automobiltechnik, die Telekommunikationsbranche und viele damit verbundene Dienstleistungszweige sowie in starkem Maße die Medizintechnik (z.B. Herzschrittmacher, Diagnosegeräte, usw.).

Ihre rasant wachsende wirtschaftliche Bedeutung gründet sich auf den schnellen technologischen Fortschritt der Entwicklung und Fertigung integrierter Schaltungen. Zum einen wuchs der Durchmesser der Wafer (z.B. 1985: 10 cm, 1990: 20 cm, 2001: 300mm), zum anderen sank die Strukturgröße erheblich (z.B. 1985: 10 μm, 1995: 0.5 μm, 2001: 0.13μm).Damit stieg die im selben Prozeß herstellbare Chipfläche auf das Vierfache, während der einzelne Chip in neuer Technologie hergestellt nur noch $\frac{1}{400}$ der Fläche beansprucht und wegen der kleineren Strukturen auch schneller ist. Dies ermöglicht die günstigere Produktion von Chips einerseits und die Realisierung komplexerer Funktionen andererseits.

Bei der Reduzierung der Prozeßgrößen stellen die atomaren Strukturen eine untere Grenze dar, in deren Nähe dann quantenphysikalische Effekte zu berücksichtigen sind. Damit deutet sich schon an, daß die Lösung dieser Probleme wohl nicht in einer Extrapolation der aktuellen Technologie, sondern in grundlegend neuem Denken liegen muß.

Thema dieser Vorlesung sind in erster Linie monolithisch integrier analoger digitale Grundschaltungen in CMOS-Technologie.[1] Daneben werden auch andere Arten von integrierten Schaltungen betrachtet, die für deren Verständnis wichtig sind. Bei diesen monolithisch integrierten Schaltungen handelt es sich um Halbleiterschaltungen, die vollständig auf/in einer Scheibe eines Halbleitereinkristalls, dem sogenannten **Wafer** realisiert werden.

Grundätzlich beschäftigt sich die Mikroelektroink mit Dickfilm-, Dünnfilm- und integrierten Schaltungen. Bei Dickfilmschaltungen (werden für Schichtschaltungen verwendet) werden die Bauelemente auf ein Keramik-Substrat aufgeklebt, bei Dünnfilmschaltungen aufgedampft und bei integrierten Schaltungen in mehreren Prozeßschritten im Silizium Substrat selbst realisiert.

Grundlegend für das Verständnis integrierter Schaltungen sind Kenntnisse auf den Gebieten der Halbleiterphysik und der Digitaltechnik für das Verständnis der Bauelemente und der daraus zusammengesetzten Gatter sowie der Mathematik/Informatik für das Verständnis der Entwurfsverfahren. Letzteres wird in der Vorlesung Mikroelektronik II behandelt.

[1] Monolithisch leitet sich von den griechischen Worten *monos* (ein) und *lithos* (Stein) ab, bedeutet also auf einem Stein bzw. einem Einkristall. Von integrierten Schaltungen spricht man, wenn alle Komponenten und deren Verbindungen als eine Einheit hergestellt werden. CMOS bedeutet Complementary Metal-Oxide Semiconductor.

Die Halbleiterbauelemente werden hier nicht detailliert erörtert, da nur ihre Grundlagen zum Verständnis der hier verwendeten Modelle erforderlich sind. Generell gibt es drei Abstraktionsebenen, auf denen man die Schaltungen betrachten kann:

- Bauelemente-/Halbleitersicht (physikalische Gleichungen, partielle Differentialgleichungen)
- Schaltungssicht (Modelle für Bauelemente)
- Schaltungen

Neben diesen Sichten einer Schaltung gibt es noch mehrere andere, wie z.B. die physikalische Sicht, die im Rahmen dieser Vorlesung nicht betrachtet werden. Im Folgenden wird jedes Bauelement innerhalb einer integrierten Schaltung als „Black Box“ betrachtet, indem z.B. nur seine Kennlinie als charakterisierende Eigenschaft verwendet wird.

Die restlichen Sichten spielen beim Entwurf integrierter Schaltungen eine Rolle, der Thema der Vorlesung Mikroelektronik II ist.

Inhalt der Vorlesung

- Überblick und Entwicklungshistorie
- Charakteristiken und Modelle der wesentlichen Bauelemente insbes. MOS Transistoren (Vt, Gm, Sättigungsstrom... Dimensionierung)
- Grundlage der analogen IC (Inverter, Differenzstufe, Strom-Quelle und Spiegel)
- einfache Gatter und deren Layout, Übergänge und Verzögerung
- kombinatorische Logik und Sequentielle Logik
- Schiebregister, Zähler
- Tristate, Bus, I/O Schaltung
- Speicher: DRAM, SRAM, ROM, NVM
- PLA, FPGA
- Prozessor und digitaler Systementwurf

2 Der Herstellungsprozeß

Basis bei der Herstellung von integrierten Schaltungen bilden Einkristalle aus Silizium. Es muß sich um einen Einkristall handeln, da sonst Oberflächeneffekte an Kristallübergängen die Halbleitereigenschaften wesentlich stören würden. Diese Einkristalle werden in einem Zonenschmelzverfahren hergestellt, bei dem in mehreren Schritten die Verunreinigungen beim Schmelzen immer vor der Schmelze herwandern.

Der so erhaltene Einkristall wird in der Regel schwach p-vordotiert, um die Leitfähigkeit zu erhöhen. Mittels Diffusion (bei hohen Temperaturen) und Ionenimplatation (gezieltes Einschießen von Ionen in den Einkristall, wobei man Ionendichte und Eindringtiefe sehr gut einstellen kann) können danach Gebiete mit erhöhter Ladungsträgerkonzentration erzeugt werden.

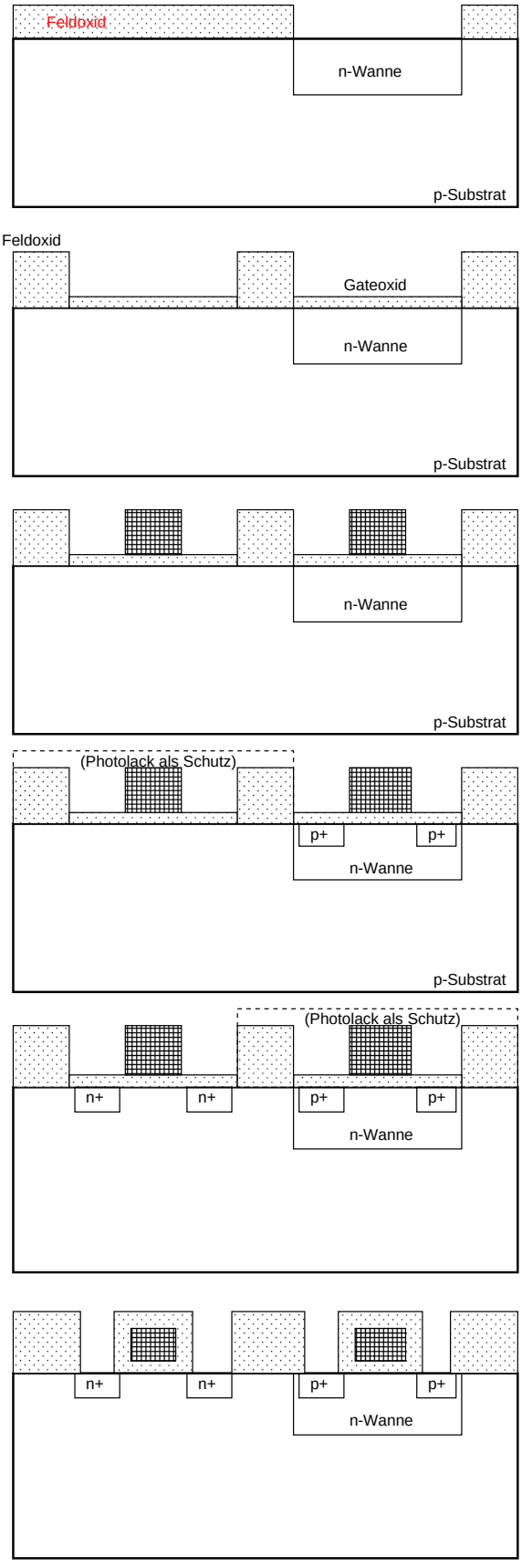
Bei Temperaturen von ca. 1000 °C und in Sauerstoffatmosphäre kann auf den Kristall eine Siliziumoxidschicht aufgewachsen werden, wobei die Dauer der Erhitzung die Dicke der Oxidschicht bestimmt.

Ebenfalls bei großer Hitze, allerdings in einer Siliziumdampfatmosphäre, kann dann auf das Oxid eine polykristalline Siliziumschicht (im folgenden kurz „Poly“ genannt) aufgewachsen werden. Später wird man sehen, daß mit diesen Schritten bereits die Grundstruktur für Transistoren herstellbar ist. In Sauerstoffatmosphäre kann dann der oberste Teil dieser Schicht ebenfalls wieder oxidiert werden.

Auf das Oxid kann dann noch ein Metallfilm aufgedampft werden, wovon Teile als Verbindungen (Leiterbahnen) verwendet werden können.

Die Auswahl, welche Gebiete eines Wafer welche Eigenschaften erhalten, wird mit sogenannten **photolithographischen Verfahren** getroffen. Dabei wird der Chip mit einem Photolack überzogen, der dann mit Hilfe von Masken teilweise belichtet bzw. strukturiert wird. Der Photolack wird dann abgelöst, wobei der belichtete Teil bestehen bleibt. Die oben genannten Prozeßschritte wirken sich dann nur auf den Teil des Wafer aus, auf dem sich kein Photolack mehr befindet.

Der Herstellungsprozeß ist in Abbildung 2.1 anhand einer kleinen integrierten Schaltung kurz dargestellt. Dort fehlt nur der Schritt der Metallisierung, in dem Metall auf der gesamten Oberfläche aufgedampft und dann nach Lithographie teilweise wieder weggeätzt wird.



Feldoxid aufwachsen
Lithographie (Lack auftragen, belichten, ätzen)
Feldoxid ätzen
schwach n dotieren (Phosphor-Implantation)
ausheilen, nachdiffundieren

Feldoxid aufwachsen
Lithographie
Feldoxid ätzen
Gateoxid aufwachsen

Polysilizium aufwachsen
Poly dotieren
Lithographie
Poly ätzen

Lithographie
stark p dotieren (Bor-Implantation)
ausheilen, nachdiffundieren

Lithographie
stark n dotieren (Arsen-Implantation)
ausheilen, nachdiffundieren

Zwischenoxid aufwachsen
Lithographie
Kontaktlöcher ätzen

Abbildung 2.1: Herstellung einer integrierten Schaltung (bis auf Metallisierung)

3 Bauelemente der Mikroelektronik

Die Vorlesung beschäftigt sich in erster Linie mit monolithisch integrierten Schaltungen. Deshalb sollen hier auch nur die Bauelemente betrachtet werden, die sich direkt in einem Halbleiter integriert realisieren lassen. Hierzu zählen:

- Diode
- Widerstand
- Kondensator
- Transistor (Bipolar und MOS)

Nur extrem kleine Induktivitäten sind realisierbar und kommen deshalb äußert selten vor. Es gibt aktuelle Bestrebungen, sie für die HF-Elektronik zu verwenden (Coil on Chip).

Integrierte Widerstände und Kondensatoren werden im Rahmen dieser Vorlesung nur mit einfachen Modellen betrachtet.

3.1 Diode

Eine Diode läßt sich als einfacher n-p-Übergang realisieren, indem man z.B. ein n-Gebiet in ein p-Substrat diffundiert oder implantiert. Auf dieselbe Art erzeugt man ein p-Gebiet, das als Anodenanschluß dient. Diese beiden Anschlüsse werden besonders hoch dotiert, damit der Metall-Halbleiter-Übergang ein möglichst ohmsches Verhalten und einen geringen Widerstand erhält (Vermeidung eines sog. Schottky-Kontaktes; dieser verhält sich selbst ähnlich einer Diode).

Auch wenn die Diode in Sperrrichtung gepolt wird, fließt ein kleiner Strom, der Sperrstrom. Die Stromgleichungen für die Diode beschreiben die Stromdichte j und die Sperrstromdichte j_s :

$$j = j_s(e^{\frac{U}{U_{\text{temp}}}} - 1) \tag{3.1}$$

$$j_s = qn_i^2 \left(\frac{D_n}{N_A L_n} + \frac{D_p}{N_D L_p} \right) \tag{3.2}$$

Dabei ist $U_{\text{temp}} = \frac{kT}{q}$ die sog. Temperaturspannung, q die Elementarladung ($\approx 1.6e^{-19}As$), k die Boltzmann-Konstante ($\approx 1.38e^{-23}J/K$), D_n, D_p die Diffusionskonstanten für Elektronen/Löcher

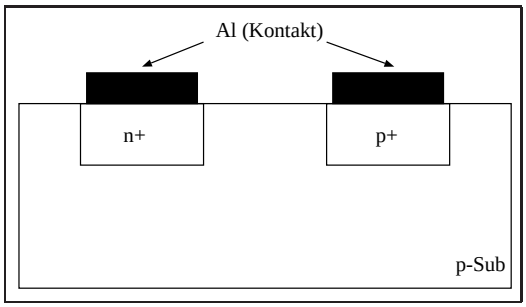


Abbildung 3.1: Realisierung einer Diode

(Einheit m^2/s), N_A, N_D die Akzeptoren- bzw. Donatorendichte (Einheit $1/m^3$), L_n, L_p die Diffusionslängen (mittlere freie Weglänge zwischen zwei Stößen mit dem Kristallgitter) von Elektronen/Löchern (Einheit m). j_s hängt damit nur vom Material und vom Herstellungsprozeß ab, ist also für den Schaltungsentwickler nicht beeinflußbar. n_i ist die Eigenleitungsladungsträgerdichte (auch intrinsische Ladungsträgerdichte genannt).

Üblicherweise ist bei den hochdotierten Gebieten $N_{A/D} \approx 10^{19...20} m^{-3}$, im Substrat dagegen $\approx 10^{14...16} m^{-3}$.

Damit ergibt sich der Verlauf für den Diodenstrom $I_D(U_D)$ bzw. $\log I_D(U_D)$ entsprechend Abbildung 3.2.

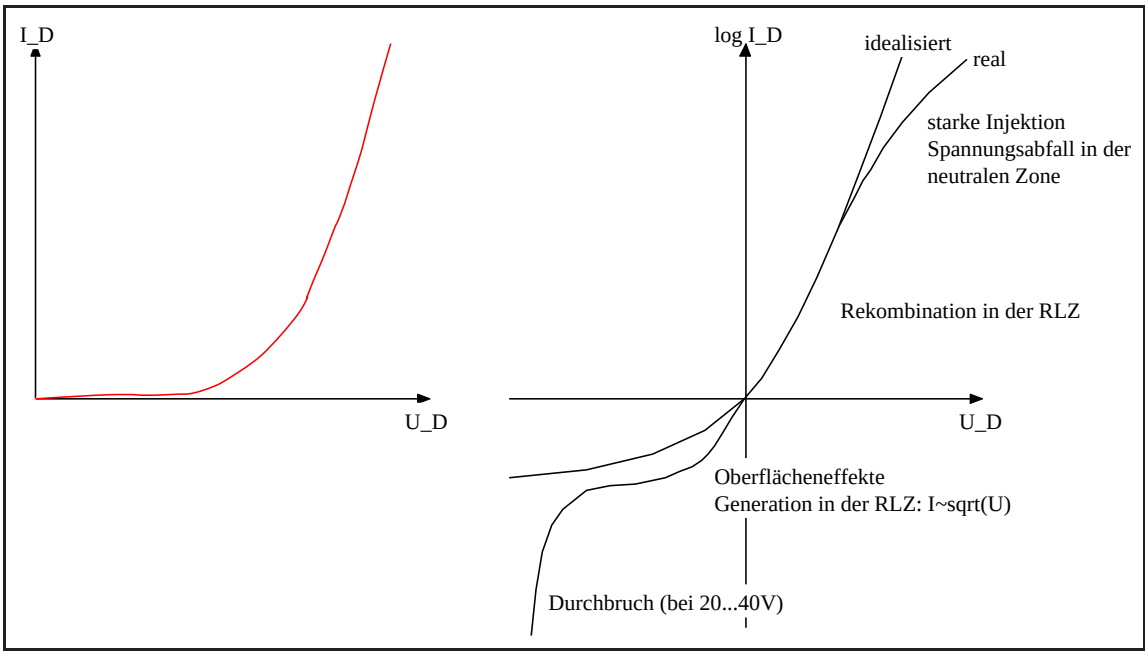


Abbildung 3.2: Verlauf des Diffusionsstroms einer Diode

Wenn die Diode sperrt, so entsteht eine Raumladungszone (RLZ), in der keine beweglichen Ladungsträger vorhanden sind. In diesem Zustand verhält sich die Diode wegen der beweglichen

Ladungsträger in n- und p-Gebiet mit dazwischenliegender „Isolationsschicht“ (der Raumladungszone) wie ein Kondensator der Größe

$$C_s = \varepsilon \cdot \frac{A}{l} \tag{3.3}$$

Dabei ist als Länge l die Länge der RLZ anzusetzen.

Dennoch kann es durch zwei Effekte zum Stromfluß kommen:

Lawineneffekt: Die Elektronen nehmen durch das von der hohen Spannung erzeugte E-Feld soviel kinetische Energie auf, daß sie die RLZ durchdringen und dabei durch Stöße mit dem Kristallgitter weitere Elektronen aus ihren Bindungen stoßen, die dann ihrerseits stark beschleunigt werden und ebenfalls neue Elektronen herausschlagen. Dieser Effekt führt zum starken Anwachsen des Stromes und damit zur thermischen Zerstörung des Bauelements.

Tunneleffekt: Wenn die p- und n-Gebiete sehr stark dotiert sind und genügend nahe beieinanderliegen, bildet sich nur ein sehr schmaler p-n-Übergang, d.h. eine kleine Raumladungszone aus. Diese können die Elektronen auf Grund quantenphysikalischer Effekte auch ohne Überwindung der Potentialbarriere mit einer gewissen Wahrscheinlichkeit überwinden (tunneln).

Der insgesamt fließende Strom ergibt sich als Produkt der Stromdichte und der Fläche des p-n-Übergangs: $I = j \cdot A$. Dabei ist zu beachten, daß es sich um eine räumliche Anordnung handelt, die in etwa die Form einer Wanne hat, so daß neben der „Bodenfläche“ auch die „Seitenflächen“ eine Rolle spielen.

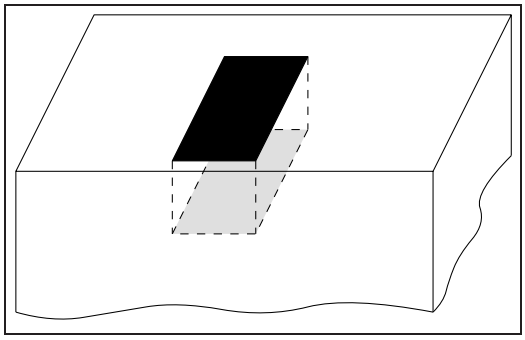


Abbildung 3.3: Perspektivische Ansicht des p-n-Übergangs

Damit ergibt sich das Ersatzschaltbild entsprechend Abb. 3.4.

3.2 Bipolartransistor

Abbildung 3.5 zeigt den Querschnitt durch einen npn-Bipolartransistor. Der dort eingezeichnete vergrabene Kollektor dient der Verringerung des Kollektorbahnwiderstandes. Im Prinzip handelt es

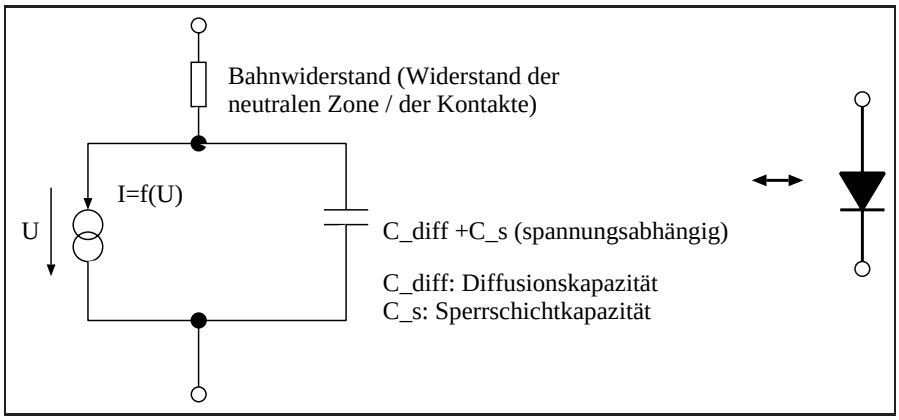


Abbildung 3.4: Ersatzschaltbild einer gesperrten Diode

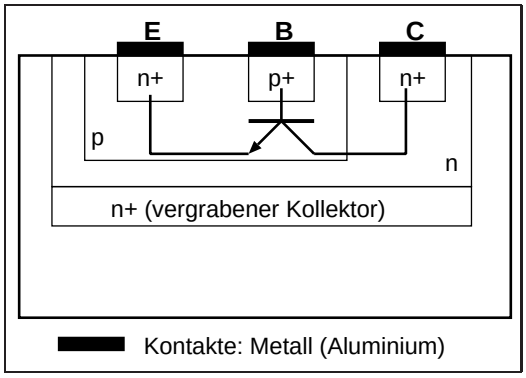


Abbildung 3.5: Querschnitt durch einen integrierten Bipolartransistor (npn)

sich bei diesem hochdotierten Gebiet um einen dem Kollektorgebiet parallelgeschalteten, kleinen Widerstand.

In Abbildung 3.6 ist das Ausgangskennlinienfeld eines Bipolartransistors dargestellt. Die Verlängerungen der Kennlinien im aktiven Bereich schneiden sich alle in einem Punkt auf der U_{CE}-Achse, der sogenannten Early-Spannung. Der Name Sättigungsbereich rührt davon her, daß in diesem Bereich die Basis mit Ladungsträgern gesättigt ist, so der Transistor auf Änderungen des Basisstromes nur sehr langsam reagiert.

3.3 MOS-Transistor

Die Grundstruktur eines MOS-Transistors beruht auf einer Leiter-Isolator-Halbleiter-Abfolge (engl.: METAL OXIDE SEMICONDUCTOR), die für sich genommen eine Kapazität darstellt (Abb. 3.7). Wir konzentrieren uns hier auf die Betrachtung des sog. **n-Kanal-Transistors**, der in einem p-Substrat realisiert werden kann. Später werden wir noch sehen, wie ein p-Kanal-Transistor realisiert werden kann.

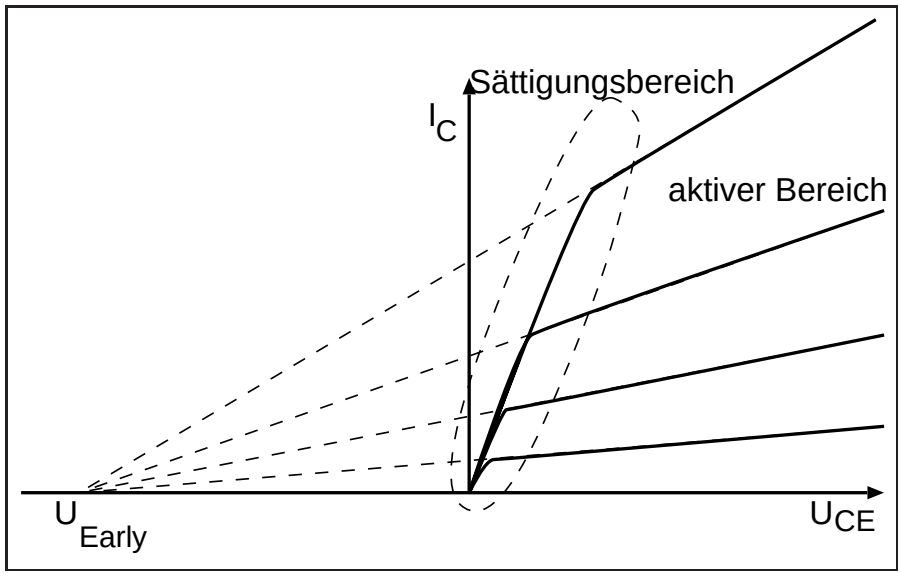


Abbildung 3.6: (Ausgangs-)Kennlinienfeld eines Bipolartransistors

Legt man an der Polysiliziumplatte[1] (Gate) eine Spannung an, so entsteht ein Feld, das je nach Polung dazu führt, daß sich unterhalb der Oxidschicht Ladungsträger ansammeln, die eine dünne n-leitende Schicht, die sogenannte **Inversionsschicht**, bilden oder aber daß durch Verdrängen der Ladungsträger unterhalb des Gate eine Raumladungszone entsteht.

Bei Inversion entsteht eine leitende Verbindung zwischen beiden n+-Gebiete. Das Gebiet unterhalb der Oxidschicht wird **Kanal** genannt. Es handelt sich hierbei — wegen der durch Inversion entstandenen n-leitenden Schicht — um einen n-Kanal-MOSFET.

3.3.1 Strom-Spannungs-Zusammenhang des MOSFET

Um überhaupt eine leitende Schicht, d.h. einen Überschuß von n-Ladungsträgern, im Kanal zu erzeugen, muß erst eine bestimmte Spannung zwischen Polysilizium und Substrat — die **Schwellenspannung** (engl. threshold voltage) U_T — überschritten werden. Die Schwellenspannung wird höher, wenn die p-Dotierung hoch ist. Bei geringerer Gatespannung existiert keine leitende Schicht und der Transistor ist gesperrt (Sperrbereich)[2].

Bei höherer Spannung als U_T entsteht ein leitender Kanal. Der Transistor ist damit leitend.

Wie groß der Strom durch den Kanal ist, hängt dann neben dem Gatepotential auch von der Spannung über dem Kanal ab. Das Gatepotential hat dabei nur Einfluß auf die Anzahl der Ladungsträger, während die Spannung U_{DS} über dem Kanal deren Geschwindigkeit bestimmt.

Die einzelnen Anschlüsse eines MOSFET werden entsprechend Abbildung 3.8 bezeichnet. Wenn U_{DS} gegenüber U_{GS} immer weiter erhöht wird, so wird vor allem am Drain die Raumladungszone

[1] Polysilizium verhält sich bei entsprechender Dotierung eher wie ein Metall denn wie ein Halbleiter
[2] Es kann dann lediglich ein Sperrstrom fließen, der jedoch unterhalb $1\mu A$ liegt

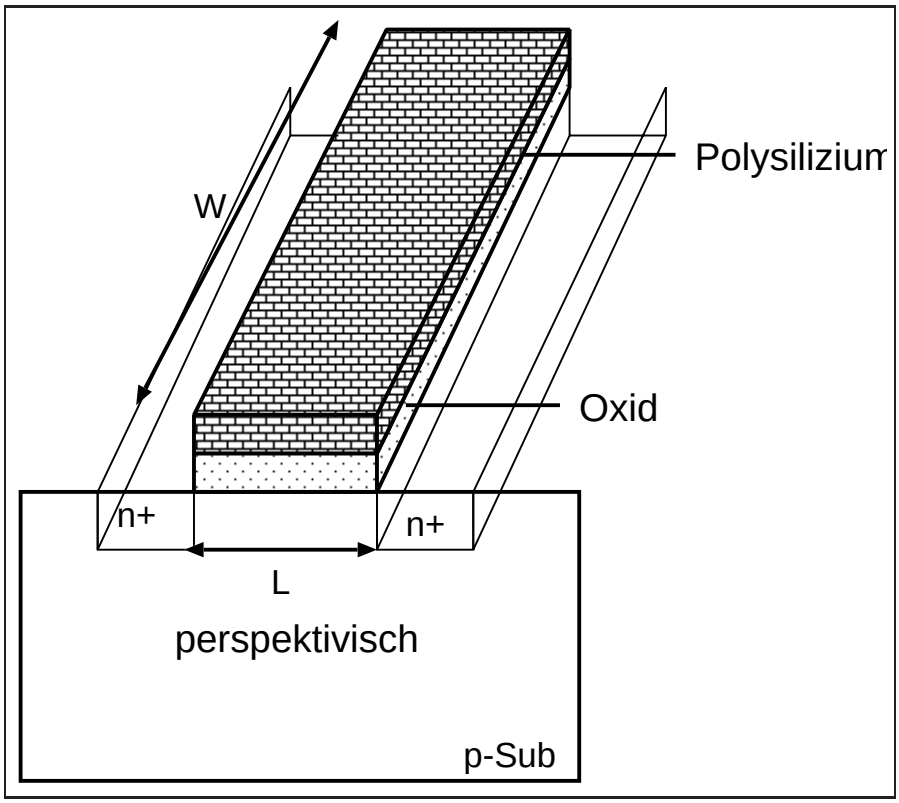


Abbildung 3.7: Grundstruktur eines n-Kanal MOSFET

immer breiter, der leitende Kanal an dieser Stelle immer dünner, bis er schließlich zu einem Punkt zusammenschrumpft. Dieser Punkt heißt **Abschnürpunkt, pinch-off**. Bei weiterer Erhöhung von U_{DS} wandert dieser Punkt näher an Source heran. Die Ladungsträger durchfliegen die restliche Strecke (in der eigentlich kein Kanal mehr vorhanden ist, da dieser ja abgeschnürt ist) mit ihrer Sättigungsdriftgeschwindigkeit. Der Spannungsbereich, in dem dies geschieht, wird deshalb auch **Sättigungsbereich** genannt.

Es ergibt sich ein Ausgangskennlinienfeld nach Abbildung 3.9.

In den drei Kennlinienbereichen gelten folgende Zusammenhänge:

Sperrbereich:

$$\begin{aligned} U_{GS} &< U_T \\ I_D &\approx 0 \end{aligned}$$

Linearer Bereich: (auch Triodenbereich genannt)

$$\begin{aligned} U_{GS} &> U_T \\ U_{DS} &< U_{GS} - U_T \\ I_D &= \beta(U_{GS} - U_T - \frac{1}{2}U_{DS})U_{DS} \end{aligned}$$

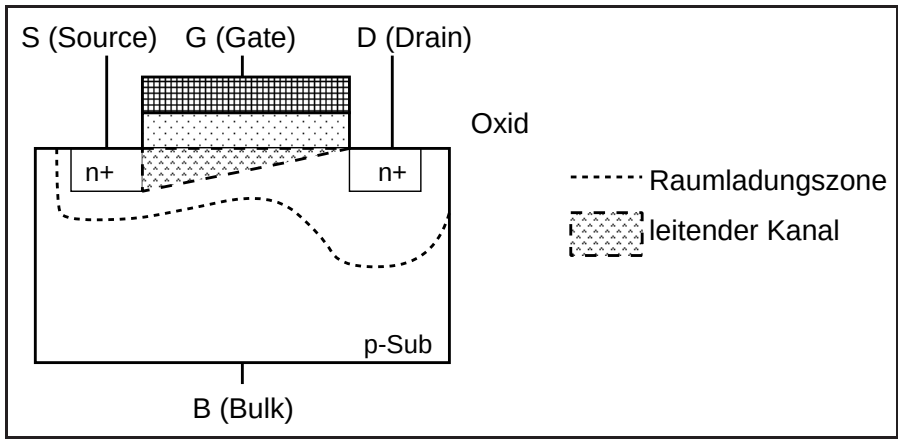


Abbildung 3.8: MOSFET bei nicht abgeschnürtem, leitendem Kanal

Sättigungsbereich:

$$\begin{aligned} U_{\mathrm{GS}} &> U_{\mathrm{T}} \\ U_{\mathrm{DS}} &\geq U_{\mathrm{GS}} - U_{\mathrm{T}} \\ I_{\mathrm{D}} &= \frac{\beta}{2}(U_{\mathrm{GS}} - U_{\mathrm{T}})^2 \end{aligned}$$

Bei allen Gleichungen ist dabei $\beta = \beta_0 \frac{W}{L}$, wobei wiederum $\beta_0 = \frac{\mu \varepsilon_{\mathrm{ox}}}{t_{\mathrm{ox}}}$ (μ: Beweglichkeit, $\varepsilon_{\mathrm{ox}}$: Oxid-Dielekrizitätszahl, t_{ox}: Dicke der Oxidschicht) ist. Es ist gut zu erkennen, daß der Transistorstrom um so größer ist, je kleiner die Kanallänge bzw. je dünner das Gateoxid ist.

Ein Faktor $\lambda \approx 0.01 \ldots 0.05 \frac{1}{V}$ kann eingeführt werden, um die dem Modell entnommenen, ideal flachen Kennlinien den leicht steigenden der Realität anzupassen, durch ihn wird der sog. Early-Effekt berücksichtigt. λ beschreibt die Kanallängenmodulation infolge der Raumladungszone, die durch U_{DS} erzeugt wird. Die Gleichung für den Sättigungsbereich lautet dann:

$$I_{\mathrm{D}} = \frac{\beta}{2}(U_{\mathrm{GS}} - U_{\mathrm{T}})^2 (1 + \lambda U_{\mathrm{DS}})$$

Für die Schwellenspannung U_{T} gilt:

$$U_{\mathrm{T}} = U_{\mathrm{T0}} + \gamma \left(\sqrt{U_{\mathrm{SB}} + 2U_{\mathrm{Diff}}} - \sqrt{U_{\mathrm{Diff}}} \right)$$

wobei U_{T0} die Schwellenspannung bei verschwindendem Substratpotential ist, γ der Substratfaktor, U_{SB} die Spannung zwischen Substrat und Source und $U_{\mathrm{Diff}} = \frac{kT}{q} \ln \frac{N_A N_D}{n_i^2}$ die Diffusionsspannung. Die Gleichung beschreibt den Effekt, daß die negative Bulkspannung positive Ladung im Kanalbereich erzeugt, so daß die Schwellenspannung erhöht wird.

Der Unterschied zwischen Drain und Source ist bei MOS-Transistoren rein interpretativ. Durch den symmetrischen Aufbau sind die beiden Anschlüsse beliebig vertauschbar. Für das tatsächliche

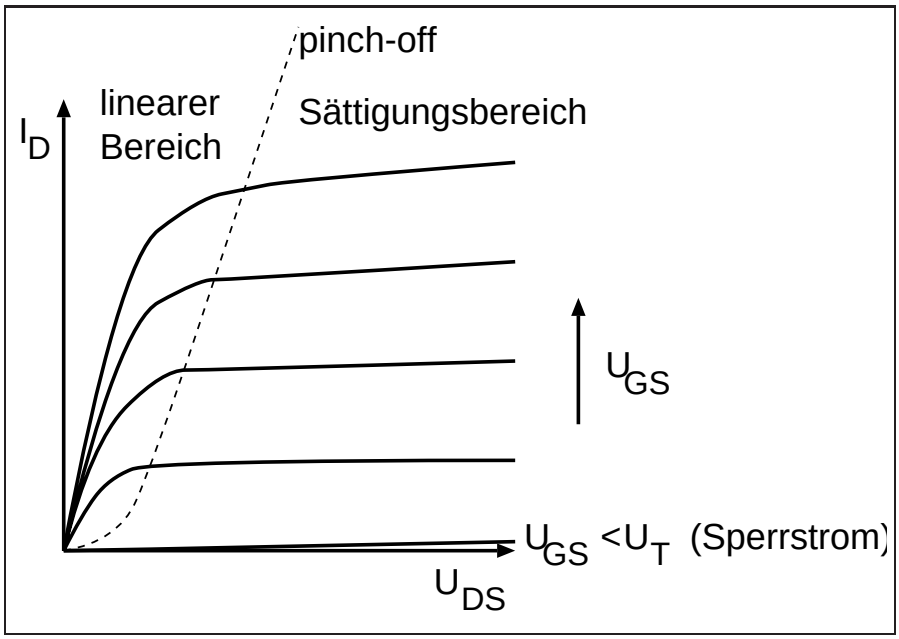


Abbildung 3.9: Ausgangskennlinienfeld eines MOSFET

Verhalten ist nur maßgeblich, welcher der beiden Anschlüsse an höherem Potential liegt und beim hier betrachteten n-Kanal-Transistor dadurch als Drain interpretiert wird.

Damit der pn-Übergang Source-Bulk niemals durchgeschaltet werden kann (Was aufgrund der bipolaren Injektion zur Aktivierung parasitärer Transistoren zur Folge hat und in der Schaltung zu Fehlfunktionen führen kann), darf U_{SB} nie kleiner als $0V$ (eigentlich $\approx -0.4 \cdots -0.7V$) werden. Dies wird dadurch gewährleistet, daß der Bulkanschluß auf das niedrigste auftretende Potential (Masse) gelegt wird, so daß an Source keine negativen Potentiale gegenüber Bulk mehr auftreten können.

Ein weiterer Vorteil dieser Vorgehensweise ist, daß alle in Sperrichtung gepolten n+-Gebiete von einer Raumladungszone umgeben sind, die diese gegeneinander isoliert. Bei Bipolartechnologie ist eine solche Isolierung durch zusätzlichen Herstellungsaufwand zu gewährleisten. Dies ist mit einer der Gründe, warum bei MOS-Technologie derart große Packungsdichten erreicht werden können.

Im linearen Bereich ist I_D linear von U_{GS} und quadratisch von U_{DS} abhängig. Die Kennlinien des Ausgangskennlinienfeldes sind in diesem Bereich äquidistant. Im Sättigungsbereich ist I_D quadratisch von U_{GS} und (schwach) linear von U_{DS} abhängig.

Im Sperrbereich ($0 \leq U_{GS} < U_T$) werden ebenfalls Elektronen in das Gebiet unterhalb des Gate gezogen, jedoch reicht deren Anzahl lediglich für eine sehr schwache Leitfähigkeit des Kanals aus (**schwache Inversion**). Damit ist ein sehr kleiner Stromfluß durch den Kanal (Größenordnung $\approx 1\mu A$) möglich, der exponentiell mit abnehmenden U_{GS} abnimmt.

Bei digitalen Schaltungen liegt U_{GS} entweder (nahe) bei $0V$ oder auf einem Wert, der wesentlich größer als U_T ist, so daß schwache Inversion vernachlässigt werden kann. Bei analogen MOS-Schaltungen kann dieser Arbeitsbereich sogar vorteilhaft eingesetzt werden (siehe Mikroelektronik 3).

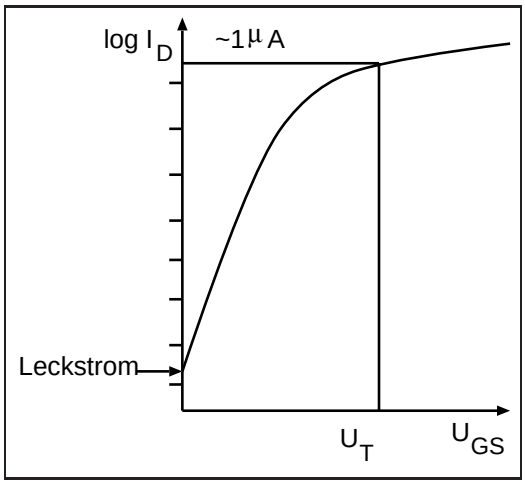


Abbildung 3.10: Strom bei schwacher Inversion

Durch die annähernd waagerechte Kennlinie im Sättigungsbereich — I_D hängt hier nur schwach von U_{DS} ab — läßt sich der Transistor als **Stromquelle** verwenden. Im linearen Bereich verhält er sich als **Widerstand** und kann auch als **Schalter** benutzt werden.

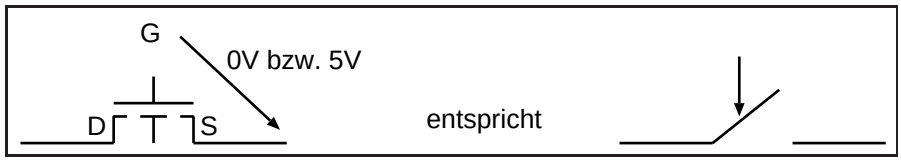


Abbildung 3.11: MOSFET als Schalter

Bei einer Verwendung als Schalter wird jedoch über dem Transistor — je nach Schaltungsgegebenheiten — eine Spannung abfallen. Der Transistor wird sich auf jeden Fall als nichtidealer Schalter verhalten.

Die Schwellenspannung U_T eines Transistors kann durch Vordotierung des Kanals in großen Bereichen eingestellt werden. Findet keine Vordotierung des Kanals statt, so spricht man vom sog. **nativen** Transistor. In der Regel gilt für in digitalen Schaltungen verwendete Transistoren (V_{DD} ist die Versorgungsspannung):

$$|U_T| \approx 0.2 V_{DD}$$

Bei (den üblichen) 5V für die Versorgungsspannung ergibt sich damit $|U_T| \approx 1V$, bei den heute bereits gebräuchlichen 3.3V ergibt sich $|U_T| \approx 0.7V$. Damit wird der „tote“ (Sperr-)Bereich für kleinere Versorgungsspannung ebenfalls verkleinert, um ein möglichst „gewohntes“ Verhalten zu gewährleisten.

Durch Ionenimplantation kann die Schwellenspannung durch Einbringen von Donatoratomen derart stark beeinflußt werden, daß selbst bei fehlender Gatespannung bereits ein leitender Kanal existiert. Damit erhält man sogenannte **depletion** oder **Verarmungs**transistoren. Der Transistortyp,

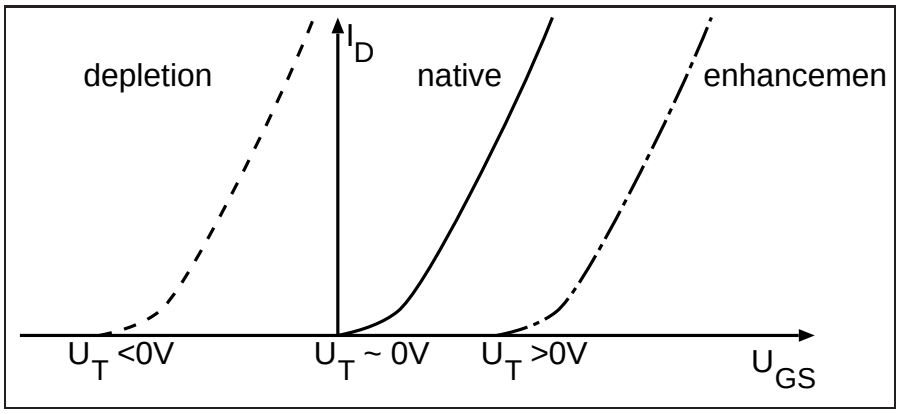


Abbildung 3.12: Eingangskennlinien von MOS-Transistoren

der bei $U_{GS} = 0$ keinen leitenden Kanal besitzt wird auch **enhancement** oder **Anreicherungs**typ genannt.

Neben den bisher betrachteten Transistoren, bei denen sich ein n-leitender Kanal im p-Diffusionsgebiet ausbilden kann gibt es noch den komplementären Typ, den **p-Kanal-Transistor**. Das Bulk besteht nun aus einem n-Gebiet. Hier kann durch Anlegen einer Gatespannung, die gegenüber Source negativ ist, ein löcherleitender Kanal erzeugt werden.

Damit erhält man im wesentlichen vier verschiedene Transistortypen, für die wir folgende Symbole verwenden werden:

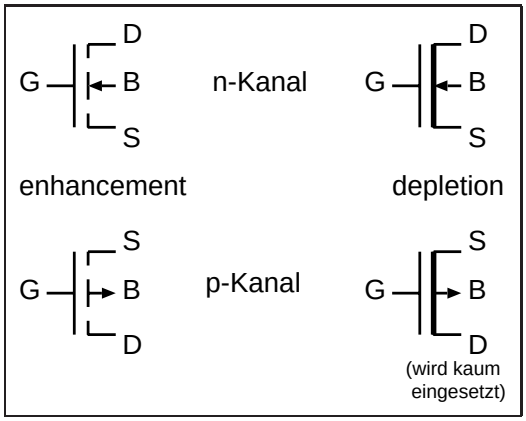


Abbildung 3.13: Schaltungssymbole für MOS-Transistoren

Daneben werden in der Literatur auch folgende Symbole verwendet:

3.3.2 Parasitäre Elemente bei MOSFET

Betrachtet werden hier nur n-Kanal-Transistoren. Die Effekte treten in entsprechender Weise selbstverständlich auch bei p-Kanal-Transistoren auf.

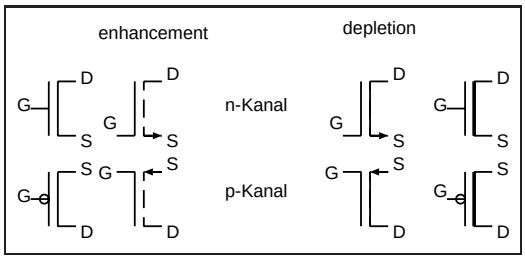


Abbildung 3.14: Schaltzeichen für MOS-Transistoren

Im Wesentlichen handelt es sich bei den parasitären Elementen um Dioden und Kapazitäten, die durch in Sperrichtung gepolte Dioden und Überlappungen verschiedener Schichten des MOS-Prozesses verursacht werden. Da die pn-Übergänge immer gesperrt sind, spielen die Dioden nur bei sehr hoher Temperatur eine Rolle, während die Kapazitäten beim dynamischen- bzw. Frequenzverhalten immer berücksichtigt werden müssen.

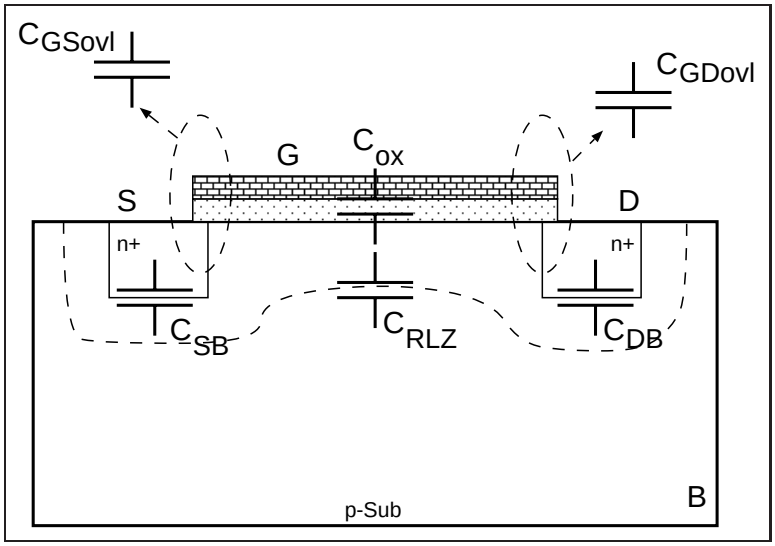


Abbildung 3.15: Parasitäre Kapazitäten beim MOS-Transistor

Die **Oxidkapazität** C_{ox} ist die Kapazität des Plattenkondensators, der aus dem Polysiliziumgate als erster Platte, dem Oxid als Dielektrikum und dem Substrat als zweiter Platte gebildet wird:

$$\begin{aligned} C_{\text{ox}} &= \varepsilon \frac{A}{d} \\ &= \varepsilon_{\text{ox}} \frac{WL}{t_{\text{ox}}} \\ &= \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}} WL \\ &= C'_{\text{ox}} WL \end{aligned}$$

Die Größe C'_{ox} hängt allein vom Herstellungsprozeß ab, das Produkt WL ist genau die Gatefläche des Transistors.

Bei negativen Gatespannungen befinden sich positive Ladungsträger unterhalb des Gate und es existiert keine Raumladungszone, so daß sich in diesem Bereich nur diese Oxidkapazität auswirkt.

In der Umgebung von $U_{GB} = 0V$ wirkt sich in Reihe zu dieser Oxidkapazität noch die Kapazität C_{RLZ} der Raumladungszone aus, die dann existiert.

Oberhalb von U_T verhält sich der Kanal durch die hohe Anzahl negativer Ladungsträger metallähnlich, so daß sich auch hier zwischen G und B nur die Kapazität C_{ox} auswirkt.

Mit den Faktoren

$$\alpha = 1 - \left(\frac{U_{GS} - U_{DS} - U_T}{2U_{GS} - 2U_T} \right)^2$$

$$\beta = 1 - \left(\frac{U_{GS} - U_T}{2U_{GS} - 2U_T - 2U_{DS}} \right)^2$$

ergibt sich dann ein Verlauf für C_{GB} in Abhängigkeit von U_{GB} bzw. U_{GS} nach Abbildung 3.16. Teilt man C_{GB} in einen Drain- (C_{GD}) und Source-Anteilt (C_{GS}) auf, so erhält man einen Verlauf dieser Größen, der ebenfalls in Abb. 3.16 dargestellt ist.

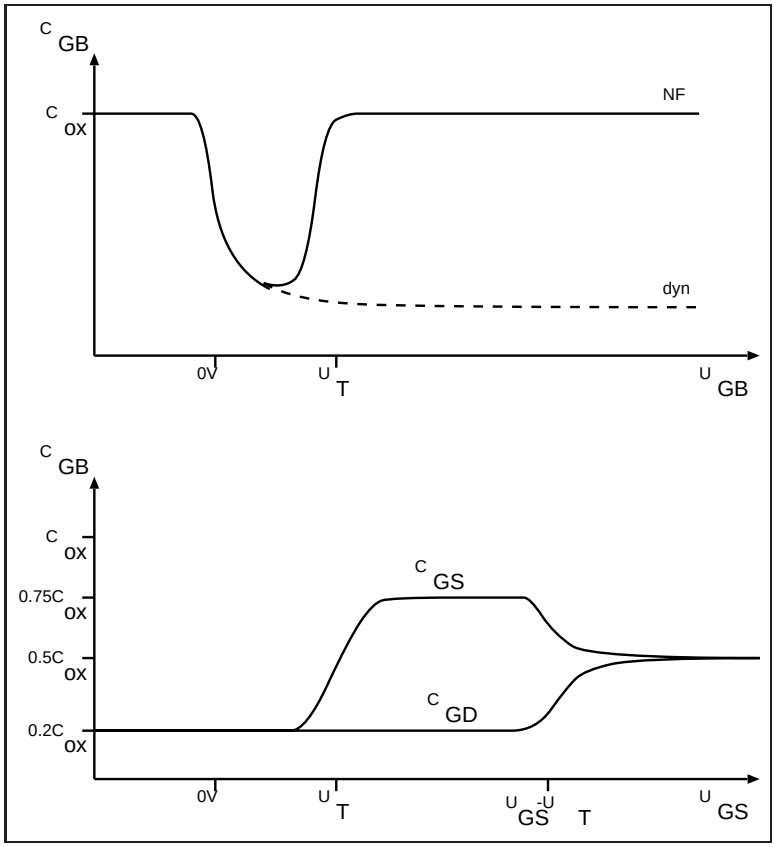


Abbildung 3.16: Gate-Substrat-Kapazität

Neben dieser Gate-Substratkapazität existieren noch die Überlappungskapazitäten von Gate und Oxid gegenüber Source bzw. Drain ($C_{\mathrm{GSovl}}, C_{\mathrm{GDovl}}$) sowie die Sperrschichtkapazitäten der Dioden am Drain- bzw. Source-Anschluß (C_{DB}, CSB).

Die Sperrschichtkapazitäten C_{SB} und C_{DB} sind nicht vernachlässigbar. Sie setzen sich jeweils aus einer Flächenkomponente C_{jF} $(\approx 0.1\frac{fF}{\mu m})$ und einer Umfangskomponente C_{jP} $(\approx 1\frac{fF}{\mu m^2})$ zusammen. Sie sind u.U. größer als die Gatekapazität selbst!

3.3.3 Effekte zweiter Ordnung

Kanallänge

Nach der Dotierung von Drain- und Sourcegebiet folgen noch einige Prozeßschritte, die hohe Temperaturen erfordern. Während dieser Schritte diffundiert ein Teil der in Drain und Source implantierten Ionen u.a. in den Bereich unterhalb des Gates, was zu einer Verkürzung des effektiven Kanals führt.

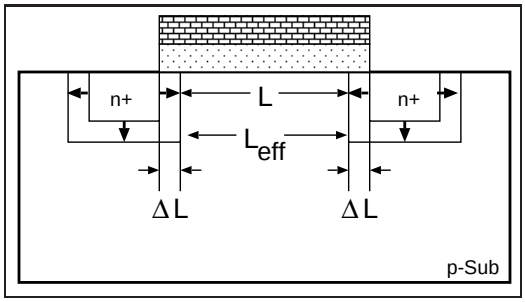


Abbildung 3.17: Kanalverkürzung

Die tatsächliche Kanallänge L_{eff} unterscheidet sich dadurch von der bei der Ionenimplantation eingehaltenen Länge L:

$$L_{\mathrm{eff}} = L - 2\Delta L$$

Kanalweite

Die Kanalweite wird durch die aktive Maske, die das Feldoxid entfernt,definiert. Der Kanal wird durch Ätzen des Feldoxides, Aufwachsen von Polysilizium und nachträglichem Aufwachsen von Oxid erzeugt. Dabei bilden sich keine ideal steilen Übergänge von Feldoxid zu Gateoxid, was zu einer Abweichung der Kanalbreite von der Maskenbreite ergibt.

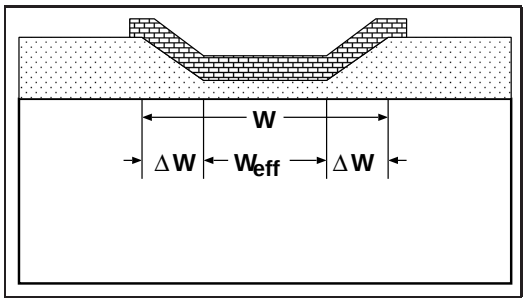


Abbildung 3.18: Kanalweitenreduktion (Seitenwände übertrieben flach dargestellt)

Hier gilt:

$$W_{\text{eff}} = W - 2\Delta W$$

Feldoxidtransistoren

Leitungen aus Metall oder Polysilizium über dem dicken Feldoxid können zusammen mit Diffusionsgebieten von Transistoren einen ungewollten Transistor bilden, da diese Bereiche mit den enhancement-Transistoren sozusagen „mitdotiert“ werden. Durch das sehr dicke Feldoxid wird dabei jedoch das Eindringen der Implantationsionen in die Halbleiterschicht sehr stark vermindert. Außerdem wird eine hohe Spannung benötigt, um eine Inversionsschicht unter dem Feldoxid zu erzeugen, da das Feldoxid sehr dick ist. Daher ist die Schwellenspannung dieser Transistoren wesentlich höher als bei den gewollten Transistoren (Faktor etwa $10 \ldots 20$) ist. Da die Schwellenspannung der gewollten Transistoren in der Größenordnung von $0.2V_{\text{DD}}$ liegt, treten jedoch nirgends Spannungen auf, die ausreichen würden, einen solchen Feldoxidtransistor ausreichend anzusteuern. Folglich können Transistoren nur in Feldoxid-Öffnungen realisiert werden, die bekanntlich durch aktive Masken erzeugt werden.

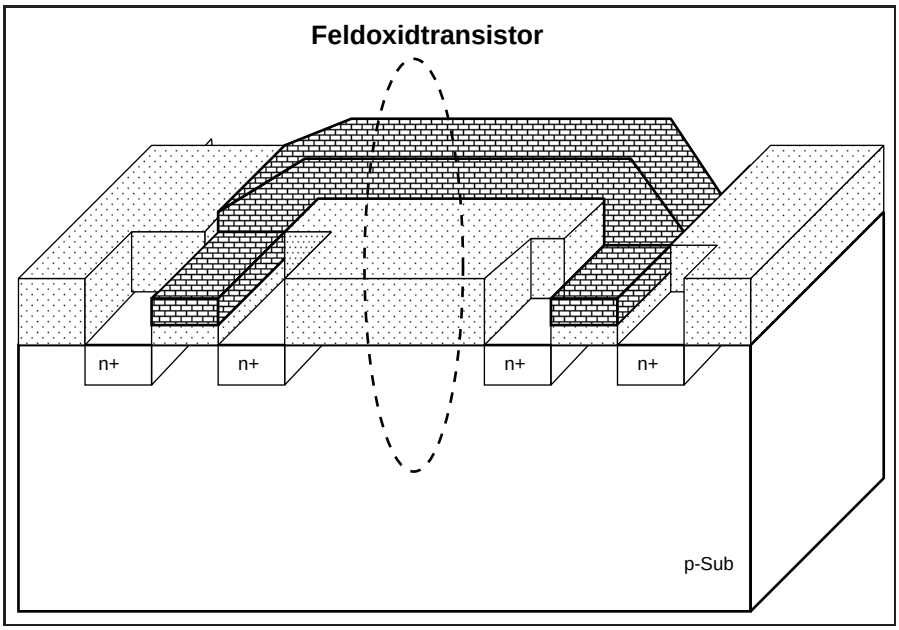


Abbildung 3.19: Feldoxidtransistor

Heiße Elektronen

Wenn das Feld über dem Gatekondensator groß genug ist und ausreichend Ladungsträger mit hoher Geschwindigkeit sich im Kanal bewegen, ist die Zahl der Ladungsträger, die ausreichend Energie besitzen, um in das Gateoxid hineinzutunneln, relativ hoch. Da die in das Gateoxid hineintunnelnden Ladungsträger dort schnell an Energie verlieren (Stöße mit dem Kristallgitter), können sie das Oxid später nicht mehr verlassen.

Diese festen Ladungen führen zu einer Verschiebung der Schwellenspannung des Transistors, so daß dieser sich nicht mehr entsprechend der Entwurfsvorgaben verhält. Solche heißen Elektronen wirken sich als verstärkte Alterung des Transistors und damit der gesamten Schaltung aus. Sie werden deswegen heiße Elektronen genannt, weil die Geschwindigkeit sehr hoch ist.

Latchup-Effekt

Bei nahe beieinanderliegenden p- und n-Kanal-Transistoren existiert immer eine laterale Zonenfolge der Art „npnp“, also zwei zusammengeschaltete Bipolartransistoren, die einen Thyristor ergeben (Abb. 3.20) können, wenn für die Stromverstärkungen der Transistoren $\beta_1\beta_2 > 1$ gilt. Dabei wirken sich die Wanne und das Substrat als Widerstände aus, die jeweils am Kollektor eines der beiden Bipolartransistoren und an der Basis des anderen angeschlossen sind.

Der kritische Schaltungsteil hat ein Aussehen entsprechend Abbildung 3.21. Schaltet einer der beiden Transistoren, so fällt auch an der Basis des anderen eine Spannung ab, die diesen ebenfalls

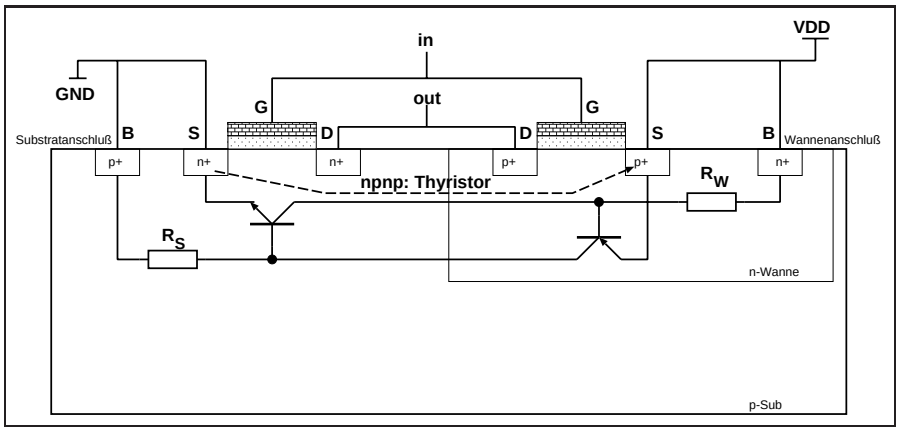


Abbildung 3.20: Latchup-Effekt

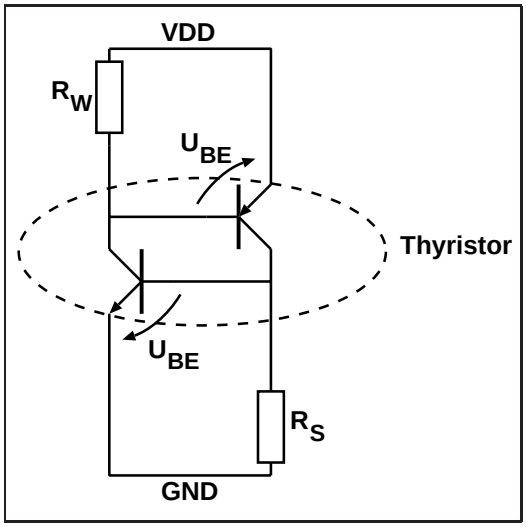


Abbildung 3.21: Ersatzschaltbild der Thyristorschaltung beim Latchup

durchschaltet. Beide Transistoren halten sich dann gegenseitig offen, es kommt zu einem dauerhaften Kurzschluß, was zur Zerstörung des Halbleiters führt.

Prinzipell kann es nicht dazu kommen, daß einer der beiden Transistoren „einfach so“ durchschaltet, eigentlich sind beide von Anfang an gesperrt und die Basis-Emitter-Dioden sind durch die Widerstände R_S und R_W kurzgeschlossen.

Es kann jedoch — z.B. bei Spannungsspitzen der Versorgungsspannungen — zu einem Stromfluß über den Wannenanschluß oder den Substratanschluß kommen, was zu einem Spannungsabfall über R_W bzw. R_S führt, so daß einer der beiden Transistoren geschaltet wird. Der dann fließende Kurzschlußstrom ($\approx 100mA$!) kann den Halbleiter zerstören.

Vermeidung des Latchup:

- Die Basis der Bipolartransistoren (laterale Ausdehnung)kann vergrößert werden, so daß höhere Spannungsabfälle zum Schalten erforderlich werden. Dadurch würden aber auch die

Abmessungen der Schaltung vergrößert, was inakzeptabel ist.

- Der kritische Wannen-Substratübergang an der Oberfläche kann durch **Schutzringe** entschärft werden. Wird im Substrat direkt neben der Wanne ein p+-Gebiet eindiffundiert und an GND angeschlossen, so kann der npn-Transistor nicht mehr schalten. Entsprechend kann innerhalb der Wanne zwischen Wannenrand und Drainanschluß des p-Kanal-Transistors ein n+-Gebiet eindiffundiert und an V_{DD} angeschlossen werden, was ein Schalten des pnp-Transistors unmöglich macht.
- Die Widerstände von Wanne und Substrat können möglichst klein gemacht werden, so daß keine Spannung abfällt, die groß genug wäre den entsprechenden Transistor zu schalten. Eine relativ einfache Möglichkeit hierzu ist möglichst **viele Kontakte** für Drain, Source, V_{DD} und GND anzubringen, was der Parallelschaltung von Widerständen entspricht. Eine weitere Möglichkeit, den Wannenwiderstand zu senken ist, einen **n+ buried layer** (ähnlich dem vergrabenen Kollektor eines Bipolartransistors) unterhalb der Wanne einzudiffundieren, der sich als parallelgeschalteter kleiner Widerstand zum Wannenwiderstand auswirkt.
- Durch einen tiefen Oxidgraben (**trench**) neben der Wanne kann die npnp-Struktur aufgetrennt werden, so daß erst gar kein Thyristor entsteht. Dies ist jedoch im Prozeß mit sehr viel Aufwand verbunden.

Durch die Vermeidungsverfahren für den Latchup wird der Prozeß komplizierter. Das Vermeiden des Latchup ist aber Voraussetzung für das Erreichen kleiner Schaltungsstrukturen und damit großer Packungsdichten.

3.3.4 Integrierte Leitungen und Kontakte

Leitungen auf dem Chip bestehen aus einer — abgesehen von den später noch betrachteten Kontaktlöchern — Metall- oder Polysiliziumschicht konstanter Dicke t. Grundsätzlich werden Leitungen aus rechteckigen Stücken zusammengesetzt.

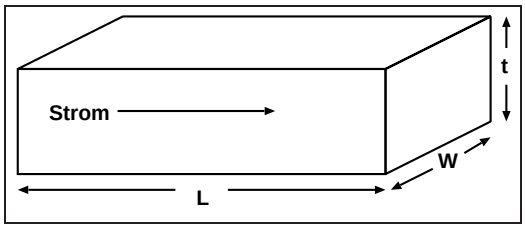


Abbildung 3.22: Leitung

Ein Leitungsstück nach Abbildung 3.22 hat einen Widerstand der Größe

$$\begin{aligned} R &= \rho\frac{L}{Wt} \\ &= \frac{\rho}{t}\frac{L}{W} \\ &= R_S\frac{L}{W} \end{aligned}$$

Sowohl die Leitungsdicke t (heutzutage $\approx 0.5 \ldots 1\mu m$) als auch der spezifische Widerstand ρ der Metallisierung sind vom Prozeß vorgegeben. Da der Widerstand eines geraden Leitungsstücks damit nur noch von der Breite (meist: $\approx 0.5 \ldots 1\mu m$) und der Länge abhängt, führt man den sogenannten **Schichtwiderstand** R_S ein.

Tabelle 3.1 zeigt einige wichtige Werte des Schichtwiderstands für unterschiedliche Materialien bei einem $1\mu m$-Prozeß. Der Name Schichtwiderstand leitet sich davon ab, daß ein (in der Draufsicht) quadratisches Leitungsstück genau den Widerstand R_S besitzt.

Material		R_S
Metall (Al)	$\approx$	$20m\Omega/\square$
Poly	$\approx$	$20\Omega/\square$
Diffusion	$\approx$	$100\Omega/\square$
Wanne	$\approx$	$1 \ldots 5k\Omega/\square$

Tabelle 3.1: Quadratwiderstände verschiedener Materialien

Für gerade Leitungsstücke gilt dann einfach die die Formel:

$$R = R_S \frac{\text{Länge}}{\text{Breite}}$$

Diese einfache Formel läßt sich jedoch nicht für Ecken in Leitungen verwenden. Für solche Fälle gibt es Näherungsformeln, die sehr gute Werte liefern, z.B. gilt als Näherung für ein quadratisches Eckstück (Abb. 3.23) die Näherung $R \approx 0.5 R_S$.

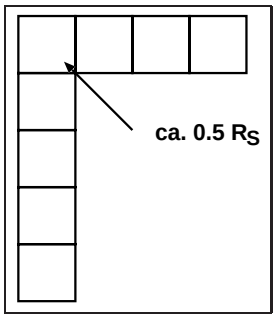


Abbildung 3.23: Draufsicht auf ein Leitungsstück mit Ecke

Metalleitungen liegen auf dem Feldoxid und besitzen daher zusammen mit dem darunterliegenden Substrat die Kapazität eines Plattenkondensators. Gleiches gilt für Polysiliziumleitungen.

$$\begin{aligned} C &= \varepsilon \frac{A}{d} \\ &= \varepsilon \frac{WL}{t} \\ &= \frac{\varepsilon}{t} WL \\ &= C_S WL \end{aligned}$$

C_S heißt auch **Kapazitätsbelag**. In Tabelle 3.2 sind die wichtigsten Kapazitätsbeläge angegeben.

Material		C_S
Poly	$\approx$	$0.05\frac{fF}{\mu m^2}$
Metall	$\approx$	$0.02\frac{fF}{\mu m^2}$

Tabelle 3.2: Kapazitätsbelag

Mehrere ausreichend nahe beieinanderliegende Leitungen besitzen auch Koppelkapazitäten zwischeneinander. Diese kann man nicht einfach vernachlässigen, da die Leitungsabstände aus Platzgründen (Packungsdichte) ebenfalls $\approx 0.5 \ldots 1\mu m$ sind und weiter abnehmen.

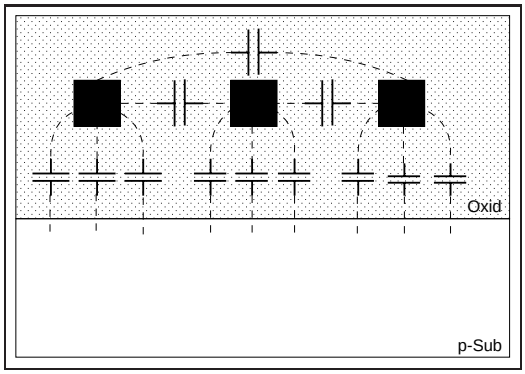


Abbildung 3.24: Kapazitäten bei integrierten Leitungen

Die Koppelkapazitäten müssen in der Regel aus den Feldgleichungen berechnet werden.

Insgesamt läßt sich die Leitung dann aus Π- oder T-Gliedern zusammengesetzt modellieren.

Integrierte Widerstände lassen sich am besten in Form von Mäandern realisieren, da diese einen relative hohen Widerstand bei geringem Platzbedarf erzeugen (Abb. 3.25).

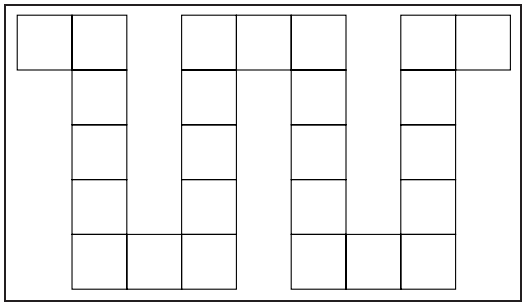


Abbildung 3.25: Mäanderwiderstand (Draufsicht)

Kontakte

Die Kontaktierung von Metalleitungen zum Substrat (z.B. Drain- und Sourceanschlüsse von Transistoren) erfordert Kontaktlöcher durch das dicke (Feld-)oxid. Diese Kontaktlöcher sind nur im Bereich von einem Mikrometer breit und lang, dagegen in der Größenordnung von 700…1500 Mikrometern tief. Dadurch besitzen auch sie einen nicht zu vernachlässigenden Widerstand.

Die Kontaktwiderstände einiger Grenzflächen sind in Tabelle 3.3 widergegeben.

Materialien	Kontaktwiderstand
Metall-Diffusion	$\approx 30\Omega$
Metall-Poly	$\approx 20\Omega$
Metall-Metall	$\approx 0.2\Omega$

Tabelle 3.3: Kontaktwiderstände

Jeder Kontakt (engl.: via) wird zumeist aus quadratischen Einheiten gebildet, wenn ein kleiner Übergangswiderstand gefordert wird, müssen dementsprechend mehrere Kontakte verwendet werden. Früher wurden Kontaktlöcher und Leitungen in einem einzigen Metallisierungsschritt hergestellt. Durch die geringeren heutzutage verwendeten Strukturgrößen wurde es jedoch immer schwieriger, die schmalen, tiefen Kontaklöcher in einem einzigen Schritt verläßlich aufzufüllen, da besonders an den Rändern Gebiete mit stark verminderter Leitfähigkeit entstehen. Darum werden diese bei modernen Verfahren in einem eigenen Schritt langsam aufgefüllt bzw. planarisiert und erst danach der gesamte Wafer mit einer homogenen Aluminiumschicht (für die Leitungen) versehen.

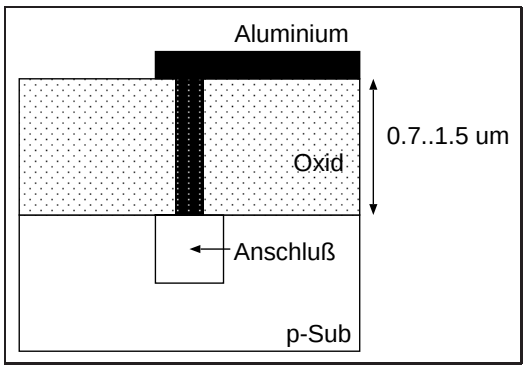


Abbildung 3.26: Kontaktloch

4 MOS-Analog Schaltungen

4.1 Inverter

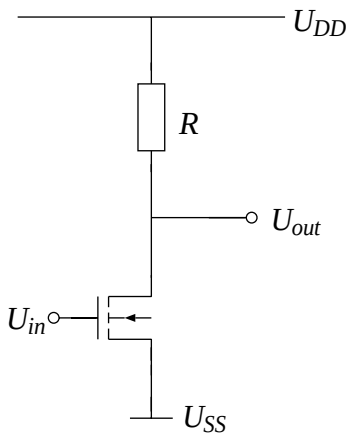


Abbildung 4.1: Inverter mit Widerstandslast

Die in Bild 4.1 dargestellte Schaltung ist in der digitalen Schaltungstechnik als Inverter mit Widerstandslast bekannt. Die Bezeichnung ”Last” kann allgemein in zwei verschiedenen Bedeutungen verwendet werden. In diesem Fall ist die Nutzlast gemeint, die für die Verstärkung nötig ist. Mit Last kann aber auch die (unerwünschte) zu treibende Last am Ausgang gemeint sein (siehe Bild 4.2).

An der Kennlinie in Bild 4.3 kann man erkennen, daß die Schaltung in Bild 4.1 auch als Verstärker verwendet werden kann: Kleine Veränderungen der Eingangsspannung U_{in} führen zu größeren Variationen der Ausgangsspannung U_{out}. Das funktioniert aber nur bei geeigneter Wahl des Arbeitspunktes: Er muß in dem steilen Bereich der Kennlinie liegen.

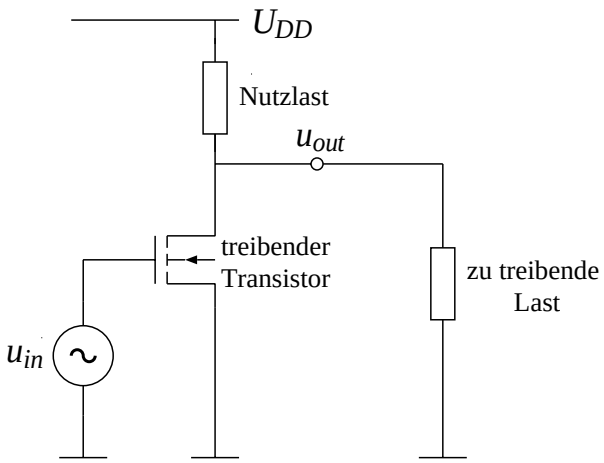


Abbildung 4.2: Zwei Bedeutungen von ”Last”

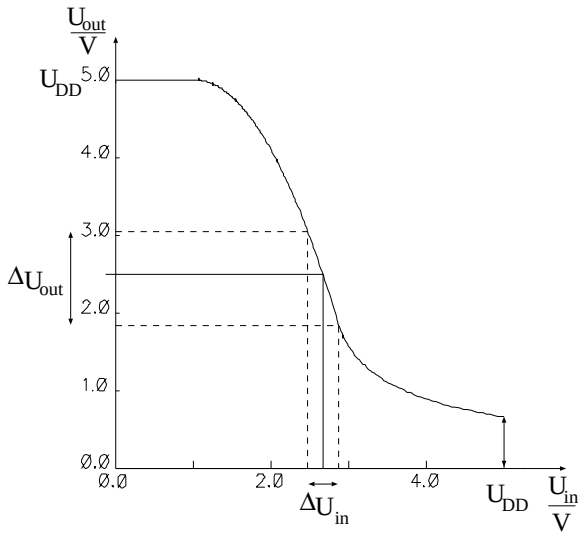


Abbildung 4.3: Kennlinie des Inverters

Mit Hilfe der Dimensionierung des Transistors kann der steile Bereich der Kennlinie verschoben werden. Damit kann die Berechnung des Arbeitspunktes auf zwei Arten erfolgen:

- Die Dimensionierung ist gegeben $\Rightarrow$ bestimme geeignetes U_{in}
- U_{in} ist gegeben $\Rightarrow$ bestimme geeignete Dimensionierung

Hier wird der zweite Fall betrachtet. Zur Berechnung des Arbeitspunktes ist das Großsignalverhalten entscheidend. Es wird angenommen, daß sich der Transistor im Sättigungsbereich befindet. Wenn der Einfluß der Kanallängenmodulation vernachlässigt wird, gilt:

$$I_D = \frac{\beta}{2} \cdot (U_{GS} - U_T)^2 \tag{4.1}$$

Folgende Werte werden zum einfachen Rechnen vorgegeben:

$$U_{DD} = 5V \tag{4.2}$$
$$U_T = 1V = 0.2 \cdot U_{DD} \tag{4.3}$$
$$U_{in} = \frac{U_{DD}}{2} = 2.5V \tag{4.4}$$

im Arbeitspunkt:

$$U_{out} = \frac{U_{DD}}{2} = 2.5V \tag{4.5}$$
$$I_{Last} = 250\mu A \tag{4.6}$$
$$I_D = I_{Last} = 250\mu A \tag{4.7}$$
$$U_{DD} = I_{Last \cdot R + U_{out}} \tag{4.8}$$
$$R = \frac{U_{DD} - U_{out}}{L_{Last}} \tag{4.9}$$
$$= 10K\Omega \tag{4.10}$$

Mit :

$$U_{GSE} = U_{in} - U_T = 1.5V < U_{DS} = U_{out} = 2.5V \tag{4.11}$$

gilt die Gleichung 4.1 für diesen Arbeitspunkt.

Mit:

$$U_{in} = U_{GS} \tag{4.11}$$
$$U_{out} = U_{DD} - I_D \cdot R \tag{4.12}$$

ergibt sich:

$$\frac{1}{2} \cdot U_{DD} = U_{DD} - \frac{\beta}{2} \cdot (U_{in} - U_T)^2 \cdot R \tag{4.12}$$
$$= U_{DD} - \frac{\beta}{2} \cdot (0.3 \cdot U_{DD})^2 \cdot R \tag{4.13}$$

nach β auflösen:

$$\Longrightarrow \beta = \frac{1}{0.09 \cdot R \cdot {U_{DD}}^2} \tag{4.14}$$
$$\tag{4.15}$$

β hängt von der Dimensionierung des Transistors ab:

$$\beta = \mu \cdot \frac{\varepsilon_{OX}}{t_{OX}} \cdot \frac{W}{L} \tag{4.16}$$

$$= \beta_0 \cdot \frac{W}{L}, \tag{4.17}$$

$$t_{OX} = 50nm$$

$$\beta_0 = \mu \cdot \frac{\varepsilon_{OX}}{t_{OX}} \approx 93.2\frac{\mu A}{V^2} \tag{4.17}$$

ergibt sich:

$$\frac{W}{L} = \frac{1}{0.09 \cdot R \cdot U_{DD} \cdot \beta_0} \tag{4.18}$$

$$= 2.38 \tag{4.19}$$

Für einen 1μ Prozeß gibt es die theoretische Lösung $L = 1\mu m$,$W = 2.38\mu m$. Die Maskenauflösung ist begrenzt bzw. höhere Auflösung ist sehr teuer. Die Lösung $L = 10\mu m$,$W = 23.8\mu m$ erfordert die 100-fache Fläche. Da der absolute Wert in einem IC-Prozeß ohnehin nicht genau zu produzieren ist, ist ein $\frac{w}{L} = 2.4$ angemessen und sinnvoll.

Die folgende Tabelle zeigt die möglichen Realisierungen.

$W(\mu m)$	$L(\mu m)$	$\frac{W}{L}$
ß.5	1.5	2.33
.4	1	2.4
.7	0.7	2.43
.2	0.5	2.4
.85	0.35	2.43

Die verwendete Technologie gibt die kleinste herstellbare Größe vor.

Schließlich ergibt sich aus Gleichung 4.1 für den Drainstrom:

$$I_D = 250\mu A \tag{4.20}$$

Damit ist die gesuchte Dimensionierung ermittelt und der Arbeitspunkt berechnet.

Jetzt soll die Verstärkung berechnet werden, die sich ergibt, wenn die Eingangsspannung leicht um den berechneten Arbeitspunkt variiert wird (Kleinsignalverhalten). Dies entspricht in Bild 4.4 einer höheren bzw. niedrigeren Kurve. An der Widerstandsgerade läßt sich die Änderung von U_{out} ablesen.

Bei einem größeren Widerstand ist die Steigung der Widerstandsgerade geringer, damit ist die Änderung von U_{out} bzw. die Verstärkung größer.

Für die Änderung der Ausgangsspannung gilt (s. Gleichung 4.12):

$$\Delta U_{out} = -R \cdot \Delta I_D \tag{4.21}$$

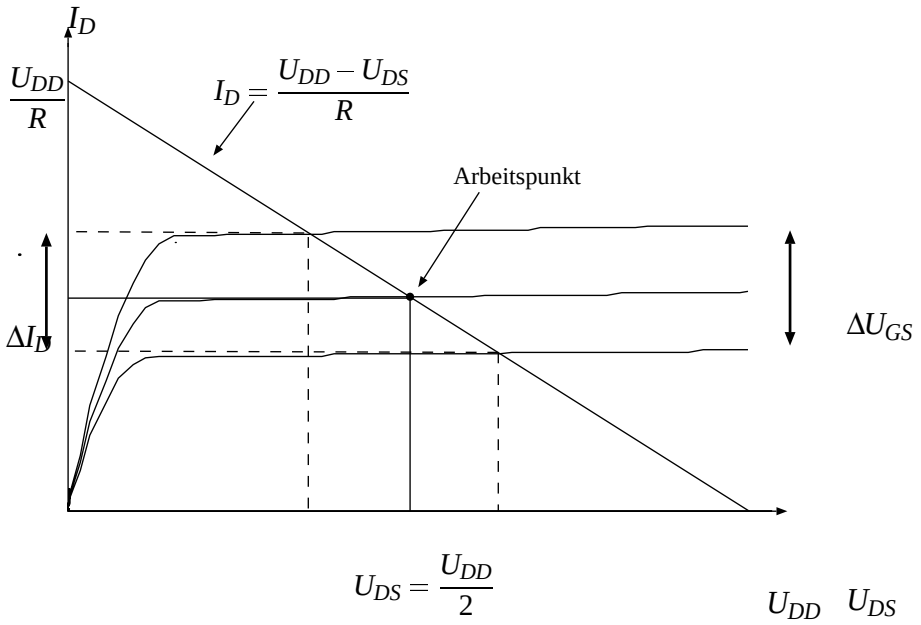


Abbildung 4.4: Kennlinienfeld mit Widerstandsgerade

Im Arbeitspunkt wird für diese kleinen Änderungen ein linearer Zusammenhang als Näherung gewählt:

$$\Delta I_D = G \cdot \Delta U_{in} \tag{4.22}$$

$$\Longrightarrow \quad G = \frac{\Delta I_D}{\Delta U_{in}} = \frac{\Delta I_D}{\Delta U_{GS}} \tag{4.23}$$

$$\Longrightarrow \Delta U_{out} = -R \cdot G \cdot \Delta U_{in} \tag{4.24}$$

Durch Übergang ins Differentielle erhält man die arbeitspunktabhängige Stromverstärkung:

$$\frac{dI_D}{dU_{GS}} = g_m \tag{4.25}$$

G_m wird auch als Übertragungssteilheit (transconductance) genannt.
Damit ergibt sich für die Verstärkung:

$$A = \frac{\Delta U_{out}}{\Delta U_{in}} = -R \cdot G \tag{4.26}$$

bzw. im Differentiellen:

$$\boxed{A = -g_m \cdot R} \tag{4.26}$$

4.2 Kleinsignalverhalten

Im vorigen Abschnitt wurde der Arbeitspunkt eines Transistors und die Verstärkung bei kleinen Änderungen der Eingangsspannung $U_{in} = U_{GS}$ bestimmt.

Bei der Betrachtung solcher kleiner Änderungen, die den Arbeitspunkt näherungsweise nicht ändern, spricht man auch vom Kleinsignalverhalten. Durch Linearisierung der Gleichungen im Arbeitspunkt können Kleinsignalersatzschaltbilder entwickelt werden, die nur diese Änderungen berücksichtigen. Durch Linearisierung ist auch eine einfache Berechnung möglich.

Im folgenden sollen für Kleinsignalgrößen Kleinbuchstaben verwendet werden. Bild 4.5 stellt ein erstes Kleinsignalersatzschaltbild eines Inverters dar.

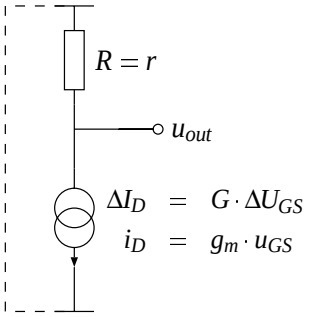


Abbildung 4.5: Kleinsignalersatzschaltbild eines Inverters

Feste Potentiale können im Kleinsignalersatzschaltbild auf Masse gelegt werden: Da sich die Spannung nicht ändert, gilt $\Delta U = 0$ bzw. $u = 0$. Entsprechend können Festspannungsquellen kurzgeschlossen werden. Feste Stromquellen können aufgetrennt werden, weil der Strom, der in sie hinein- und herausfließt, konstant ist; daher ist die Stromänderung gleich Null ($\Delta I = 0$ bzw. $i = 0$). In Bild 4.5 gilt entsprechend für die Versorgungsspannung $\Delta U_{DD} = 0$ bzw. $u_{DD} = 0$, das Potential kann auf Masse gelegt werden.

Auch der Widerstand wird nur im Arbeitspunkt betrachtet; aufgrund seiner Linearität gilt aber $r = R$, so daß Groß- oder Kleinschreibung verwendet werden kann.

Die Stromquelle ist ideal dargestellt, d.h. der Drainstrom ist unabhängig von der Drain-Source-Spannung U_{DS}. Das entspricht einer steigungslosen Geraden im Ausgangskennlinienfeld $I_D(U_{DS})$. Bild 4.6 zeigt einen Ausschnitt aus Bild 4.4, und zwar rechts im realen und links im idealen Fall. Im realen Fall führt die Kanallängenmodulation zu einer Steigung der Geraden (im Bild 4.6 übertrieben dargestellt) und damit zu einem kleineren u_{out}. Das entspricht dem Early-Effekt bei Bipolartransistoren.

Dieses Verhalten wirkt der erreichten Verstärkung entgegen, was im Kleinsignalersatzschaltbild berücksichtigt werden muß. Dazu wird der Widerstand r_{DS} eingeführt (siehe Bild 4.7); die Stromänderung bleibt also gleich, aber der Strom teilt sich jetzt auf.

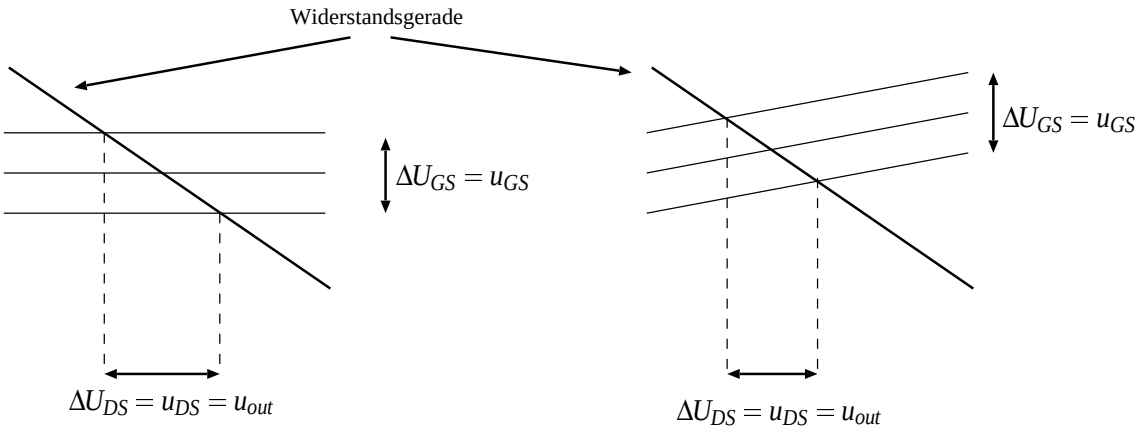


Abbildung 4.6: Ideale und nichtideale Kennlinie

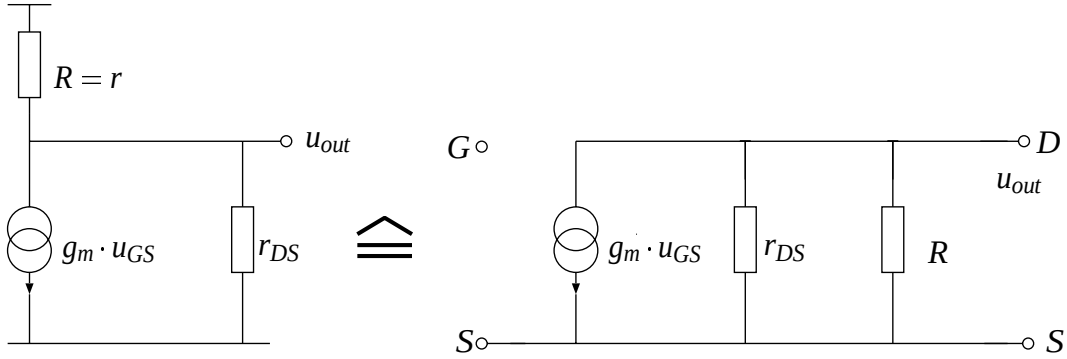


Abbildung 4.7: Erweitertes Kleinsignalersatzschaltbild

Zur Berechnung von g_m wird die Kanallängenmodulation vernachlässigt, man verwendet daher Gleichung 4.1:

$$g_m = \frac{dI_D}{dU_{GS}} \tag{4.27}$$
$$= \beta \cdot (U_{GS} - U_T) \tag{4.28}$$
$$= \beta \cdot U_{GS_{eff}} \tag{4.29}$$
$$= \frac{2 \cdot I_D}{U_{GS_{eff}}} \tag{4.30}$$
$$= \sqrt{2 \cdot I_D \cdot \beta} \tag{4.31}$$

Für β gilt dabei: $$\beta = \mu \cdot C_{OX} \cdot \frac{W}{L} \tag{4.32}$$

Damit ist β (und damit g_m) über das Verhältnis $\frac{W}{L}$ einstellbar. Es bleibt anzumerken, daß für ei-

ne größere Übertragungssteilheit ein höherer Drainstrom erforderlich ist. Diese unterschiedlichen Gleichungen können je nachdem, ob $U_{GS_{eff}}$, I_D oder β vorgegeben ist, verwendet werden.

Zur Berechnung von r_{DS} wird dagegen die um λ ergänzte Gleichung 4.33 herangezogen, da im idealen Fall die Kanallängenmodulation und damit r_{DS} nicht berücksichtigt wird. Man erhält dann:

$$I_D = \frac{\beta}{2} \cdot (U_{GS} - U_T)^2 \cdot (1 + \lambda U_{DS}) \tag{4.33}$$

$$\begin{aligned} \frac{1}{r_{DS}} = g_{DS} &= \frac{dI_D}{dU_{DS}} && (4.34)\\ &= \frac{1}{2} \cdot \beta (U_{GS} - U_T)^2 \cdot \lambda && (4.35)\\ \text{mit Gleichung 4.33:} \quad &= I_D \cdot \frac{\lambda}{1 + \lambda \cdot U_{DS}} && (4.36)\\ \text{mit } \lambda \cdot U_{DS} \ll 1 \quad &\approx I_D \cdot \lambda && (4.37) \end{aligned}$$

Eine physikalische Näherung für die Kanallängenmodulation läßt sich wie folgt herleiten.

$$\begin{aligned} I_D &= \frac{W}{L} \cdot \beta_0 \cdot (U_{SS} - UT)^2 && (4.38)\\ &= \frac{L_0}{L} \cdot \frac{W}{L_0} \cdot (U_{GS} - U_T)^2 && (4.39)\\ &= \frac{L_0}{L} \cdot I_0 && (4.40) \end{aligned}$$

L ist die aufgrung der Raumladungszone im Kanalbereich reduzierte effektive Kanallänge.L_0 ist der Abstand zwischen Drain und Source, also die unmodulierte Kanallänge. I_0 ist eine Hilfsvariable und beschreibt den Drainstrom ohne die Kanallängenmodulation.

$$I_D = \frac{L_0}{L_0 - \triangle L} \cdot I_0 \approx I_0 \cdot (1 + \frac{\triangle L}{L_0}) \tag{4.41}$$

$\triangle L$ ist die Weite der Raumladungszone. Sie ist porportional zur Wurzelfunktion der Spannung.

$$\triangle L = k_2 \cdot \sqrt[2]{U_{DS} - UDSS} \tag{4.42}$$

k_2 ist der Kanallängenmodulationsfaktor. U_{DSS} ist die Abschnürspannung $\approx U_{GSeff}$

4.2.1 Innere Verstärkung des Transistors

Unter innerer Verstärkung des Transistors versteht man den Wert der Verstärkung, der erreicht wird, wenn der Lastwiderstand beliebig groß wird, also die maximal erreichbare Verstärkung.

Dann gilt:

$$\text{mit } r_{DS} \ll R: \quad \Longrightarrow \quad A \approx -g_m \cdot r_{DS} \tag{4.43}$$

$$= -\frac{\sqrt{2 \cdot I_D \cdot \beta}}{\lambda \cdot I_D} \tag{4.44}$$

$$\text{mit Gleichung 4.37 eingesetzt} \quad \Longrightarrow \quad A \sim -\frac{1}{\sqrt{I_D}} \tag{4.45}$$

$$\text{im Beispiel:} \quad \approx -200 \tag{4.46}$$

Je kleiner der Drainstrom ist, desto größer ist also die innere Verstärkung. Ein kleiner Drainstrom bedeutet aber auch, daß die Schaltung langsamer wird.

An den Gleichungen erkennt man, daß die angegebenen AC-Parameter vom Arbeitspunkt und damit von DC-Parametern abhängig sind.

4.2.2 Admittanzparameter

Der Transistor kann auch als Vierpol angesehen werden. In diesem Abschnitt sollen die entsprechenden Parameter berechnet werden. Für einen Vierpol aus Bild 4.8 gilt ganz allgemein:

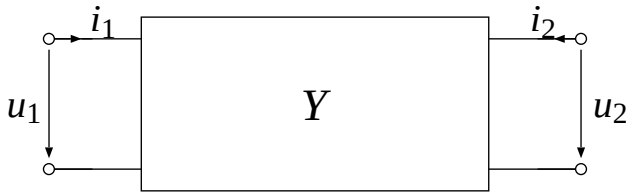


Abbildung 4.8: Vierpol

$$i_1 = y_{11} \cdot u_1 + y_{12} \cdot u_2 \tag{4.47}$$

$$i_2 = y_{21} \cdot u_1 + y_{22} \cdot u_2 \tag{4.48}$$

Für die Admittanzparameter gilt dabei:

$$y_{11} = \frac{i_1}{u_1}|_{u_2=0} = \frac{1}{r_{in}}$$ Eingangsleitwert

$$y_{12} = \frac{i_1}{u_2}|_{u_1=0}$$ bei MOS-Transistor nicht sinnvoll

$$y_{21} = \frac{i_2}{u_1}|_{u_2=0} = \frac{g \cdot u_1}{u_1} = g$$ Transkonduktanz

$$y_{22} = \frac{i_2}{u_2}|_{u_1=0} = \frac{i_2}{r_{out} \cdot i_2} = \frac{1}{r_{out}}$$ Ausgangsleitwert

Für die Verstärkung erhält man:

$$A = \frac{u_2}{u_1}|_{i_2=0} = -\frac{y_{21}}{y_{22}} = -g \cdot r_{out} \tag{4.49}$$

Bei MOS-Transistoren geht der Eingangswiderstand gegen Unendlich (bei Vernachlässigung der parasitären Kapazitäten, siehe Kapitel 3.3.2).
Im Kleinsignalbetrieb gilt für die Transkonduktanz g:

$$g = \frac{i_2}{u_1}|_{u_2=0} = \frac{dI_D}{dU_{GS}} = g_m \tag{4.50}$$

In Bild 4.9 links ist ein Lastwiderstand eingezeichnet; zur Beschreibung des Verstärkers wird er in den Vierpol hereingezogen. Damit gelangt man zu Bild 4.9 rechts und es ergibt sich für den Ausgangswiderstand:

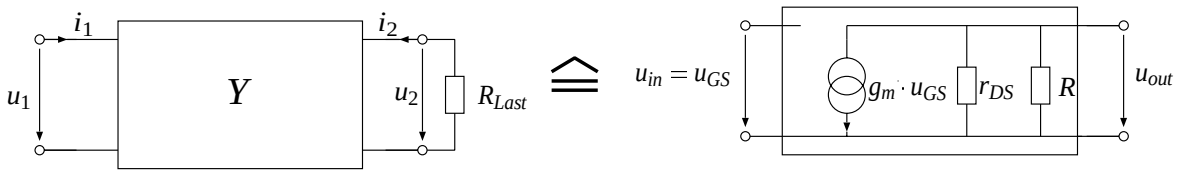


Abbildung 4.9: Vierpol mit Lastwiderstand

$$r_{out} = r_{DS}||R_{Last} = \frac{1}{\frac{1}{r_{DS}} + \frac{1}{R_{Last}}} \tag{4.51}$$

Setzt man die beiden Größen in Gleichung 4.49 ein, so erhält man:

$$\boxed{A = -g_m \cdot r_{out}} \tag{4.52}$$

Für das Verhalten einer Verstärkerstufe sind diese beiden Größen entscheidend. Dabei hängt g_m damit zusammen, was die Verstärkerstufe an Strom liefern kann, während r_{out} den Widerstand am Ausgang darstellt.

4.3 Die Differenzstufe

Bild 4.10 zeigt eine Differenzstufe. Sie besteht aus zwei Invertern, die gleich ausgelegt sind (d.h. $\beta_1 = \beta_2$, $U_{T_1} = U_{T_2}$ ...), und einer idealen Stromquelle. Es gibt zwei Eingänge U_{i_1} und U_{i_2} sowie zwei Ausgänge U_{0_1} und U_{0_2}; letztere sind allerdings nicht ganz unabhängig voneinander, da sie den gemeinsamen Knoten N_1 mit dem eingeprägten Strom I_{SS} haben. Es ist zu beachten, daß die Source-Potentiale hier nicht konstant sind; damit gilt nicht mehr unbedingt $u_{in} = u_{GS}$.

In diesem Abschnitt sollen die DC-Größen betrachtet werden. Es gilt:

$$I_{D_1} + I_{D_2} = I_{SS} \tag{4.53}$$

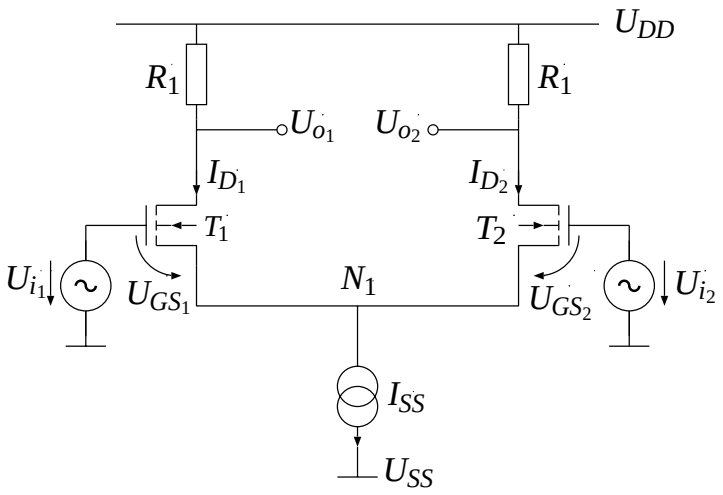


Abbildung 4.10: Differenzstufe

Wenn auch die Eingangssignale gleich sind, gilt außerdem aus Symmetriegründen:

$$\text{Bei } U_{i_1} = U_{i_2}: \qquad I_{D_1} = I_{D_2} = \frac{I_{SS}}{2} \tag{4.54}$$

Jetzt soll der Fall untersucht werden, daß ein Eingang (U_{i_2}) konstant bleibt und der andere (U_{i_1}) eine höhere Spannung erhält. Dann steigt der Strom I_{D_1}, wegen Gleichung 4.53 sinkt damit der Strom I_{D_2}.

Für die Spannungen gilt:

$$-U_{i_1} + U_{GS_1} - U_{GS_2} + U_{i_2} = 0 \tag{4.55}$$

$$\Longrightarrow U_{i_1} - U_{i_2} = U_{GS_1} - U_{GS_2} \tag{4.56}$$

Aus Gleichung 4.29 und 4.31 folgt:

$$U_{GS} = \sqrt{\frac{2 \cdot I_D}{\beta}} + U_T(U_{SB}) \tag{4.57}$$

$$\text{Mit } U_{T_1} = U_{T_2}: \qquad \Longrightarrow U_{i_1} - U_{i_2} = \sqrt{\frac{2 \cdot I_{D_1}}{\beta}} - \sqrt{\frac{2 \cdot I_{D_2}}{\beta}} \tag{4.58}$$

$$\text{und mit } \Delta U_i = U_{i_1} - U_{i_2}: \qquad \Longrightarrow \Delta U_i = \sqrt{\frac{2 \cdot I_{D_1}}{\beta}} - \sqrt{\frac{2 \cdot I_{D_2}}{\beta}} \tag{4.59}$$

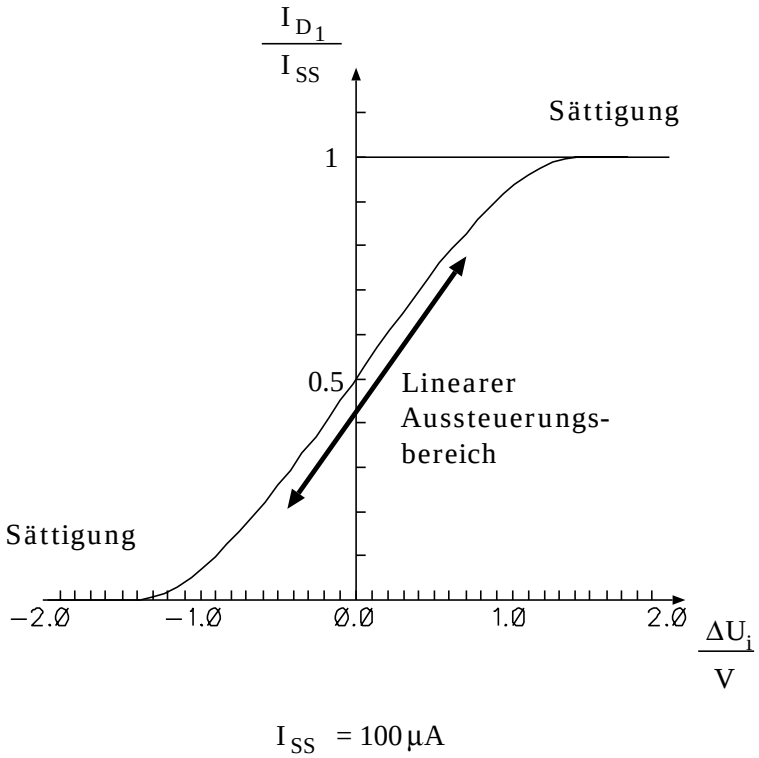
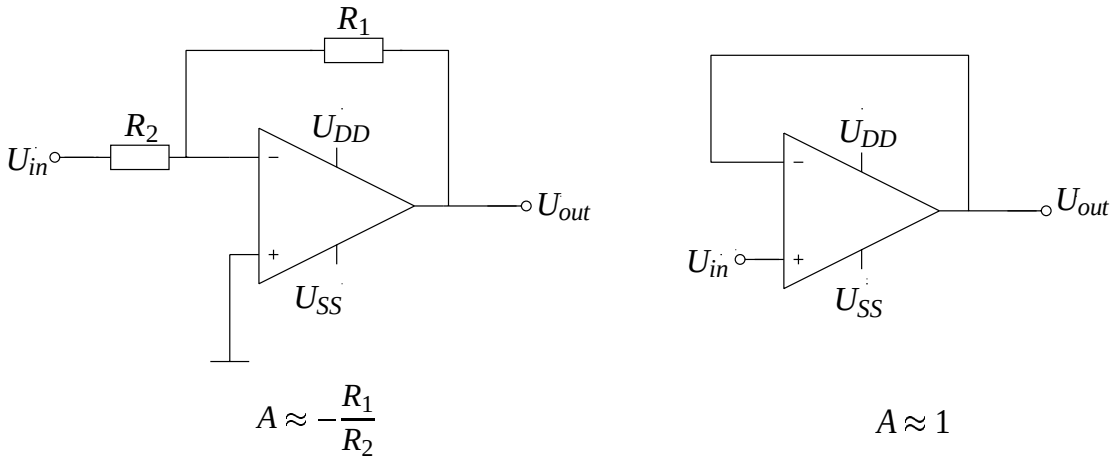


Abbildung 4.11: Übertragungsfunktion der Differenzstufe: I_{D_1} in Abhängigkeit von ΔU_i

Dabei ist zu beachten, daß wegen $U_{SB_{1,2}} > 0$ gilt: $U_{T_{1,2}} > U_{T_0}$.
Man erkennt an den Gleichungen 4.53 und 4.59, daß der Strom I_{D_1} von ΔU_i und I_{SS} abhängt. Diese Abhängigkeit soll hier nur qualitativ dargestellt werden (Bild 4.11).

Bei der oben betrachteten Differenzstufe liegt das untere Potential U_{SS} auf Masse. Bei einigen Anwendungen ist jedoch ein Massepotential nötig, das von U_{DD} und U_{SS} unabhängig ist, wie z.B. in Bild 4.12 links. Für solche Fälle werden i.Allg. U_{DD} und U_{SS} symmetrisch zur Masse gewählt, beispielsweise $\pm 5V$. Bei einem Buffer dagegen ist eine externe Masse nicht nötig (Bild 4.12 rechts).



$$A \approx -\frac{R_1}{R_2} \qquad A \approx 1$$

Abbildung 4.12: Typische Beschaltungen

4.3.1 Aussteuerungsbereich (Großsignalverhalten)

In diesem Abschnitt soll untersucht werden, in welchem Bereich die beiden Eingangsspannungen U_{i_1} und U_{i_2} in der Schaltung nach Bild 4.10 variiert werden können, ohne daß die Transistoren den Sättigungsbereich verlassen.

Differenzspannung

Aus Symmetriegründen sind beide Transistoren bei den folgenden Betrachtungen austauschbar. Wenn beide Eingänge gleich sind, gilt mit Gleichung 4.29 und 4.31:

$$U_{GS_{1,2_0}} = \sqrt{\frac{2 \cdot I_{D_{1,2_0}}}{\beta}} + U_T(U_{SB}) \tag{4.60}$$

$$\text{mit } I_{D_{1,2}} = \frac{I_{SS}}{2}: \quad = \sqrt{\frac{I_{SS}}{\beta}} + U_T(U_{SB}) \tag{4.61}$$

$$\text{bzw.} \quad U_{GS_{eff_{1,2_0}}} = \sqrt{\frac{I_{SS}}{\beta}} \tag{4.62}$$

Es sei jetzt $U_{i_2} > U_{i_1}$. Dann ist die obere Grenze erreicht, wenn der Transistor T_1 sperrt und Strom I_{SS} völlig durch den Transistor T_2 fließt. Das Potential am Knoten N_1 stellt sich entsprechend ein. Damit gilt:

$$U_{GS_{1_{min}}} \approx U_T(U_{SB}) \tag{4.63}$$

$$U_{GS_{2_{max}}} = U_{GS_{eff_{2_{max}}}} + U_T(U_{SB}) \tag{4.64}$$

$$I_{D_{2_{max}}} = I_{SS} \tag{4.65}$$

$$\Longrightarrow \Delta U_{i_{max}} = U_{GS_{eff_{2_{max}}}} \tag{4.65}$$

$$= \sqrt{\frac{2 \cdot I_{D_{2_{max}}}}{\beta}} \tag{4.66}$$

$$= \sqrt{\frac{2 \cdot I_{SS}}{\beta}} \tag{4.67}$$

$$= \sqrt{2} \cdot U_{GS_{eff_{1,2_0}}} \tag{4.68}$$

Für höhere Spannungen wird keine Verstärkung mehr erreicht, da die Differenzstufe voll ausgesteuert ist. Diese Spannung wird bei normalen Beschaltungen wie z.B. in Bild 4.12 nicht überschritten.

An Gleichung 4.67 erkennt man, daß der Aussteuerungsbereich über I_{SS} und über die β der Transistoren einstellbar ist. Ein größeres I_{SS} führt allerdings zu einer größeren Verlustleistung, ein kleineres β zu einem kleineren g_m (s. Gleichung 4.31) und damit zu einer geringeren Verstärkung.

Gleichanteil

Auch wenn $U_{i_2} = U_{i_1}$ ist (wie z.B. beim Buffer, Bild 4.12 rechts), können die Spannungen nicht beliebig variieren. In diesem Fall sind die beiden Drainströme I_{D_1} und I_{D_2} gleich. Für die Gate-Source-Spannung gilt bei beiden Transistoren gemäß Gleichung 4.61:

$$U_{GS_{1,2}} = \sqrt{\frac{I_{SS}}{\beta}} + U_T(U_{SB} > 0) \tag{4.69}$$

Dabei ist zu berücksichtigen, daß $U_{SB} > 0$ und damit $U_T > U_{T_0}$ (s. Gleichung **??**) ist.

Damit durch die Transistoren ein Drainstrom von $\frac{I_{SS}}{2}$ fließen kann, muß U_{GS} groß genug sein. Das Potential des Knotens N_1 stellt sich entsprechend ein, kann aber nicht kleiner als U_{SS} werden; im Fall $U_{N_1} = U_{SS}$ wird $U_{SB} = 0$. Somit ist eine untere Schranke für die Eingangsspannungen gegeben:

$$\begin{aligned} U_{in} &= U_{GS} + U_{N_1} && (4.70)\\ &> \sqrt{\frac{I_{SS}}{\beta}} + U_T(U_{SB} \approx 0) + U_{SS} && (4.71) \end{aligned}$$

Die obere Grenze ist dadurch gegeben, daß die Transistoren in Sättigung bleiben müssen, d.h. $U_{DS} > U_{GS} - U_T(U_{SB})$. Da die Drainströme feststehen, sind über die Widerstände die Drainpotentiale festgelegt:

$$U_{out} = U_{DD} - \frac{I_{SS}}{2} \cdot R \tag{4.72}$$

Durch den Zusammenhang

$$\begin{aligned} U_{in} &= U_{GS} - U_{DS} + U_{out} && (4.73)\\ &= U_{GS} - U_{DS} + U_{DD} - \frac{I_{SS}}{2} \cdot R && (4.74)\\ &< U_{GS} - (U_{GS} - U_T(U_{SB} > 0)) + U_{DD} - \frac{I_{SS}}{2} \cdot R && (4.75)\\ &< U_T(U_{SB} > 0) + U_{DD} - \frac{I_{SS}}{2} \cdot R && (4.76) \end{aligned}$$

ist eine Obergrenze für die Eingangsspannungen gegeben.

Insgesamt gilt also:

Differential Mode:	$\Delta U_{i_{max}} = \sqrt{2} \cdot U_{GS_{eff_{1,2_0}}}$
Common Mode:	$\sqrt{\frac{I_{SS}}{\beta}} + U_T(U_{SB\approx 0}) + U_{SS} < U_{in} < U_T(U_{SB>0}) + U_{DD} - \frac{I_{SS}}{2} \cdot R$
bzw. Common Mode Range:	$U_{DD} - U_{SS} - \frac{I_{SS}}{2} \cdot R - \sqrt{\frac{I_{SS}}{\beta}} + U_T(U_{SB>0}) - U_T(U_{SB\approx 0})$

4.3.2 Kleinsignalverhalten

Während sich die Betrachtungen im vorherigen Abschnitt auf den Arbeitspunkt bezogen, sollen jetzt kleine Änderungen der Eingangsspannungen betrachtet werden. Um die Symmetrieeigenschaften ausnutzen zu können, spaltet man dabei die Ein- und die Ausgangsspannungen folgendermaßen auf:

$$u_{i_1} = u_{i_c} + \frac{u_{i_d}}{2} \quad (4.77) \qquad u_{o_1} = u_{o_c} + \frac{u_{o_d}}{2} \quad (4.81)$$

$$u_{i_2} = u_{i_c} - \frac{u_{i_d}}{2} \quad (4.78) \qquad u_{o_2} = u_{o_c} - \frac{u_{o_d}}{2} \quad (4.82)$$

$$u_{i_d} = u_{i_1} - u_{i_2} \quad (4.79) \qquad u_{o_d} = u_{o_1} - u_{o_2} \quad (4.83)$$

$$u_{i_c} = \frac{u_{i_1} + u_{i_2}}{2} \quad (4.80) \qquad u_{o_c} = \frac{u_{o_1} + u_{o_2}}{2} \quad (4.84)$$

Dabei steht *c* für den Gleichanteil (common) und *d* für den Differenzanteil der Spannungen. Eine solche Aufspaltung ist immer möglich; die entsprechenden Ausgangsspannungen überlagern sich gemäß dem Superpositionsprinzip. Der Vorteil dieser Aufspaltung ist, daß nur ein symmetrischer und ein antisymmetrischer Anteil betrachtet werden müssen. Damit ergibt sich Bild 4.13.

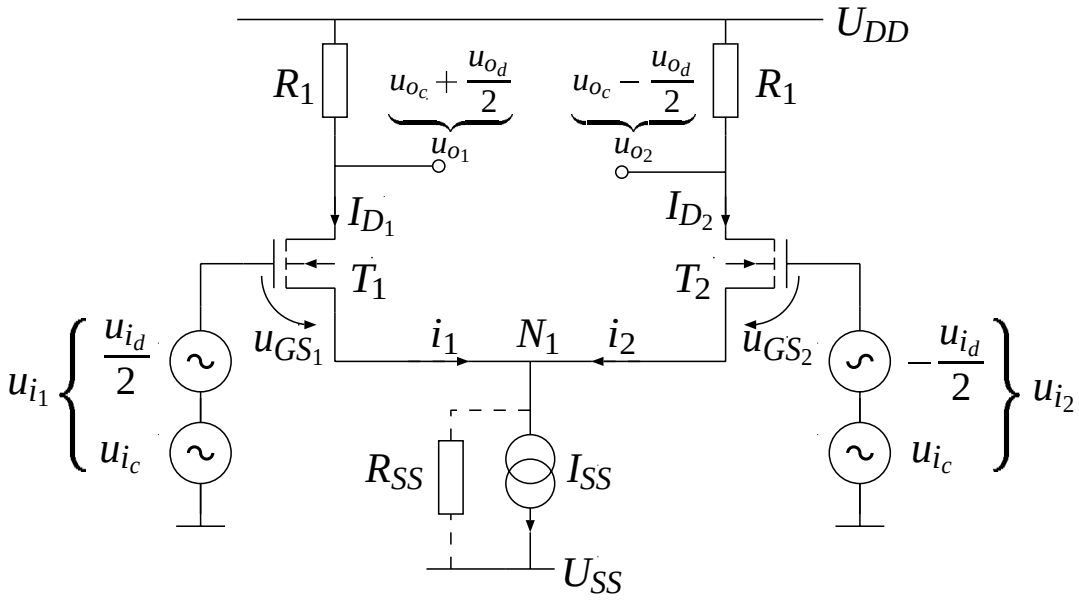


Abbildung 4.13: Differenzstufe mit aufgespaltenen Ein- und Ausgangsspannungen

Differenzmode

Zunächst soll nur der Differenzanteil betrachtet werden:

$$\text{Annahme: } u_{i_c} = 0 \implies u_{o_c} = 0 \tag{4.85}$$

$$u_{i_d} > 0 \implies i_1 > 0,\ i_2 < 0 \tag{4.86}$$

Da die beiden Eingangsspannungen betragsmäßig gleich groß sind und für das Kleinsignalverhalten die im Arbeitspunkt linearisierten Gleichungen verwendet werden, gilt:

$$i_1 = -i_2 \tag{4.87}$$

Entsprechend fließt in der Kleinsignalbetrachtung kein Strom über R_{SS}; deshalb ändert sich das Potential des Knotens N_1 nicht und er kann im Ersatzschaltbild auf Masse gelegt werden. In diesem Fall gilt also $u_{in} = u_{GS}$. Damit erhält man dann zwei voneinander entkoppelte Inverter (siehe Bild 4.14).

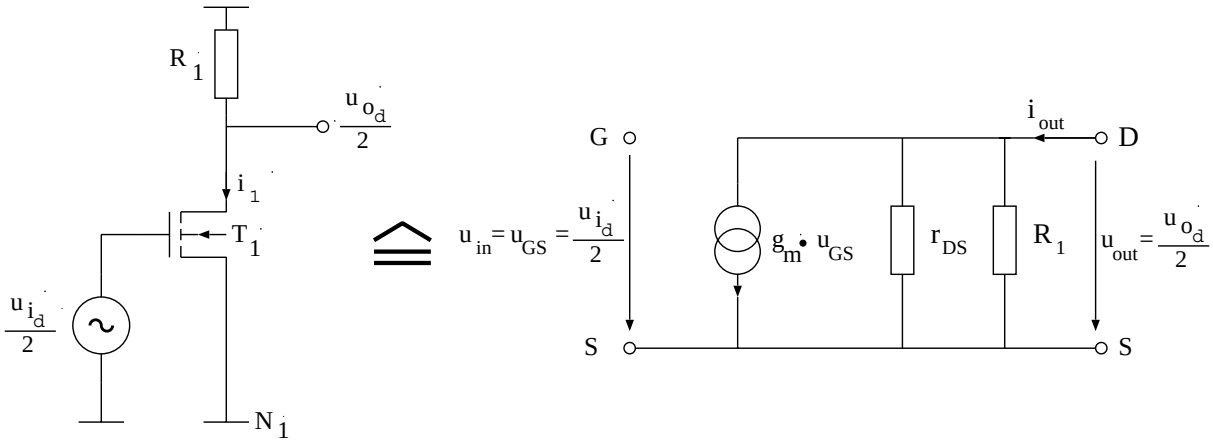


Abbildung 4.14: Kleinsignalersatzschaltbild für einen der Inverter

Die Verstärkung ergibt sich allgemein aus:

$$A_{DM} = -g \cdot r_{out} \tag{4.88}$$

g ergibt sich, wenn man den Ausgang kurzschließt und den Ausgangsstrom in Abhängigkeit von der Eingangsspannung ermittelt:

$$g = \frac{i_{out}}{u_{in}}|_{u_{out}=0} = \frac{g_m \cdot u_{GS}}{\frac{u_{id}}{2}} = \frac{g_m \cdot \frac{u_{id}}{2}}{\frac{u_{id}}{2}} = g_m \tag{4.89}$$

Um r_{out} zu bestimmen, wird der Eingang kurzgeschlossen und die Ausgangsspannung in Abhängigkeit vom Ausgangsstrom bestimmt:

$$r_{out} = \frac{u_{out}}{i_{out}}|_{u_{in}=0} \tag{4.90}$$

$$= \frac{i_{out} \cdot (R_1||r_{DS})}{i_{out}}|_{u_{GS}=0} \tag{4.91}$$

$$= R_1||r_{DS} = \frac{1}{\frac{1}{r_{DS}} + \frac{1}{R_1}} = \frac{1}{g_{DS} + \frac{1}{R_1}} \tag{4.92}$$

Bei einfachen Zusammenhängen lassen sich diese Größen auch an der Schaltung erkennen: Ein angenommener Teststrom i_{out} am Ausgang bei festgehaltenem Eingangspotential teilt sich in die beiden Zweige auf, sodaß die Parallelschaltung von R_1 und r_{DS} (wegen festgehaltenem Gatepotential) als Ausgangswiderstand erscheint. Die Drainstromänderung bei festgehaltenem Ausgangspotential und variiertem Eingangspotential hängt nur von g_m ab, sodaß $g = g_m$ ebenfalls leicht ablesbar ist.

Damit ergibt sich für die Differenzverstärkung:

$$\Longrightarrow A_{DM} = -g_m \cdot \frac{1}{g_{DS} + \frac{1}{R_1}} \tag{4.93}$$

Dies ist die gesamte Differenzverstärkung der Stufe, da sich die beiden Faktoren $\frac{1}{2}$ bei u_{i_d} und u_{o_d} aufheben. Man erkennt, daß sich diese Verstärkung nicht von der des einfachen Inverters unterscheidet.

Common Mode

Jetzt wird nur ein Gleichanteil angenommen:

$$\text{Annahme: } u_{i_d} = 0, \qquad u_{i_c} > 0 \tag{4.94}$$

$$\tag{4.95}$$

Da die beiden Eingangsspannungen gleich sind, lassen die beiden Transistoren gleichviel Strom durch:

$$I_{D_1} = I_{D_2} \Longrightarrow i_1 = i_2 \tag{4.96}$$

Zur Vereinfachung der Berechnung wird in Bild 4.15 die Stromquelle mit Innenwiderstand aufgeteilt; dabei halbiert sich der Wert der Stromquelle, weil die Ströme sich addieren, während sich der Wert des Widerstands verdoppelt, weil es sich um eine Parallelschaltung handelt.

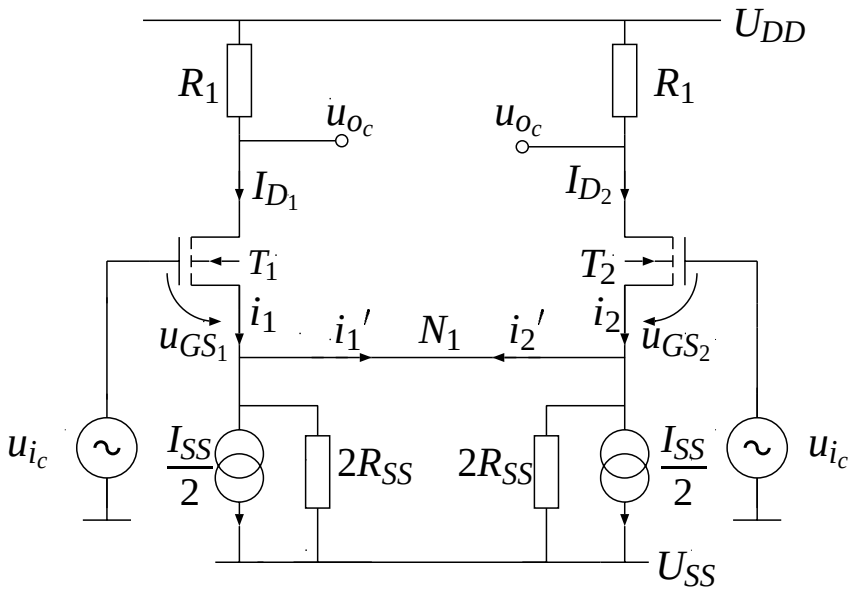


Abbildung 4.15: Differenzstufe mit aufgespaltener Stromquelle

Diese Umformung ändert nichts am Verhalten der Schaltung. Weil in diesem Fall $I_{D_1} = I_{D_2}$ ist, ist die Schaltung symmetrisch und durch den Knoten N_1 fließt kein Strom: $i_1' = i_2' = 0$. Daher läßt sich die Schaltung am Knoten N_1 auftrennen und in zwei identische Hälften teilen (siehe Bild 4.16).

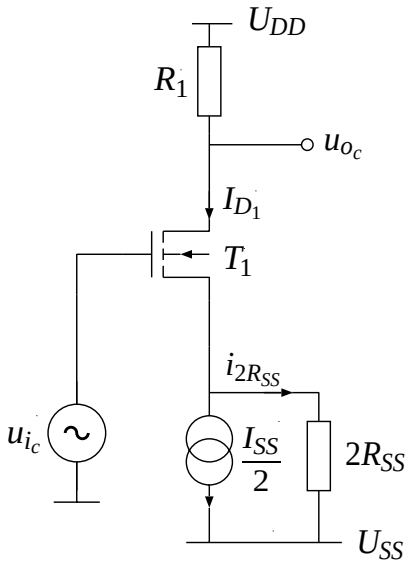


Abbildung 4.16: Teilschaltung von Bild 4.15

Für das Kleinsignalverhalten entfällt die Konstantstromquelle und kann aufgetrennt werden. In diesem Fall ist das Potential des Knotens N_1 nicht konstant, entsprechend gilt hier nicht $u_{in} = u_{GS}$. Dies wird auch im Kleinsignalersatzschaltbild 4.17 deutlich. Damit wird insbesondere die Berechnung von $r_{out} = \frac{u_{out}}{i_{out}}|_{u_{in}=0}$ komplexer.

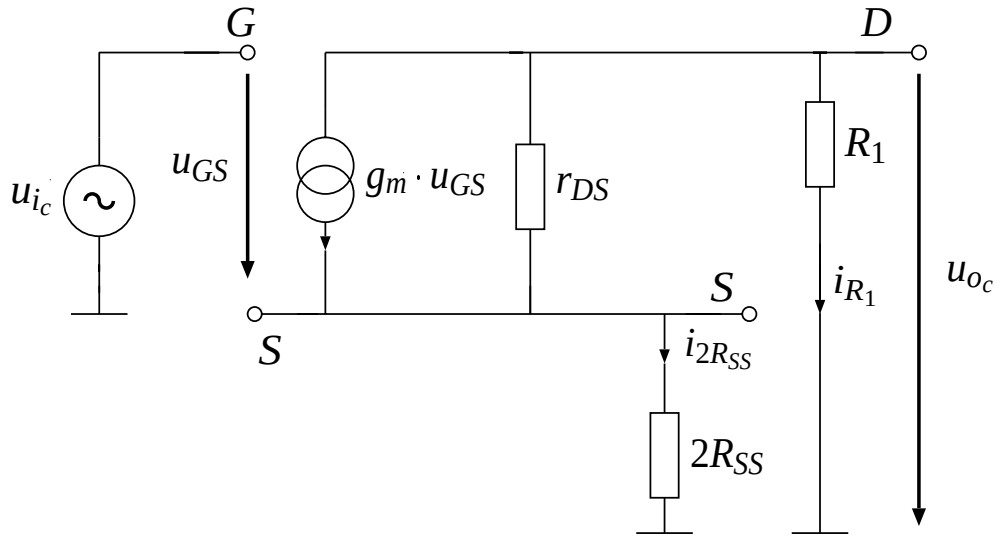


Abbildung 4.17: Kleinsignalersatzschaltbild von Bild 4.16

$i_{2R_{SS}}$ sei der Kleinsignalstrom, der durch $2R$ fließt. Damit gelten folgende Gleichungen:

$$\begin{aligned} i_{2R_{SS}} &= -i_{R_1} && (4.97)\\ u_{i_c} &= u_{GS} + i_{2R_{SS}} \cdot 2 \cdot R_{SS} && (4.98)\\ u_{o_c} &= -i_{2R_{SS}} \cdot R_1 && (4.99)\\ &= i_{2R_{SS}} \cdot 2 \cdot R_{SS} + (i_{2R_{SS}} - g_m \cdot u_{GS}) \cdot r_{DS} && (4.100) \end{aligned}$$

Für die Gleichtaktverstärkung A_{CM} (CM: Common Mode) gilt:

$$\begin{aligned} A_{CM} &= -g \cdot r_{out} && (4.101)\\ g &= \frac{i_{out}}{u_{in}}|_{u_{out}=0} && (4.102)\\ r_{out} &= \frac{u_{out}}{i_{out}}|_{u_{in}=0} && (4.103) \end{aligned}$$

Durch Einsetzen und Umformen erhält man:

$$A_{CM} = -\frac{g_m}{\frac{1}{r_{DS}} + \frac{1}{R_1} + \frac{2 \cdot R_{SS}}{R_1} \cdot (\frac{1}{r_{DS}} + g_m)} \qquad (4.104)$$

Der Widerstand R_1 kann schon aus Platzgründen nicht allzu groß sein; wie wir schon gesehen haben, gilt i.Allg. $R_1 \ll r_{DS}$. Ebenfalls kann man gewöhnlich $g_m \gg \frac{1}{r_{DS}}$ abschätzen. Der Innenwiderstand der Stromquelle R_{SS} ist möglichst groß (im Idealfall wäre $R_{SS} = \infty$); entsprechend gilt $2 \cdot R_{SS} \cdot g_m \gg 1$. Damit läßt sich die Gleichtaktverstärkung abschätzen:

$$A_{CM} = -\frac{g_m}{\frac{1}{r_{DS}} + \frac{1}{R_1} + \frac{2 \cdot R_{SS}}{R_1} \cdot (\frac{1}{r_{DS}} + g_m)} \tag{4.105}$$

$$\text{mit } R_1 \ll r_{DS}: \quad \approx -\frac{g_m}{\frac{1}{R_1} + \frac{2 \cdot R_{SS}}{R_1} \cdot (\frac{1}{r_{DS}} + g_m)} \tag{4.106}$$

$$\text{mit } g_m \gg \frac{1}{r_{DS}}: \quad \approx -\frac{g_m}{\frac{1}{R_1} + \frac{2 \cdot R_{SS}}{R_1} \cdot g_m} \tag{4.107}$$

$$= -\frac{g_m}{\frac{(1 + g_m \cdot 2 \cdot R_{SS})}{R_1}} \tag{4.108}$$

$$\text{mit } 2 \cdot R_{SS} \cdot g_m \gg 1: \quad \approx -\frac{g_m}{\frac{g_m \cdot 2 \cdot R_{SS}}{R_1}} \tag{4.109}$$

$$= -\frac{R_1}{2 \cdot R_{SS}} \tag{4.110}$$

$$\Longrightarrow \quad \boxed{A_{DM} \approx -g_m \cdot R_1 \qquad A_{CM} \approx -\frac{R_1}{2 \cdot R_{SS}}} \tag{4.110}$$

Da es sich um einen *Differenz*-Verstärker handelt, sind eine große Differenzverstärkung A_{DM} und eine kleine Gleichtaktverstärkung A_{CM} erwünscht.

Ein großer Innenwiderstand der Stromquelle R_{SS} führt zu einer kleinen Gleichtaktverstärkung A_{CM}. Dies ist auch anschaulich klar: Wenn R_{SS} groß ist, führt eine Eingangsspannungsänderung nur zu einer kleinen Drainstromänderung, weil sich das Potential von N_1 entsprechend erhöht und sich die Gate-Source-Spannungen deshalb nur wenig ändern. Damit verändert sich der Spannungsabfall an den Widerständen R_1 nur wenig und die Ausgangsspannungen bleiben annähernd gleich.

Ein anderer Weg, eine kleine Gleichtaktverstärkung zu ereichen, wären kleine Widerstände R_1. Dies würde aber auch die Differenzverstärkung vermindern.

Unter CMRR (Common Mode Rejection Ratio, Gleichtaktunterdrückung) versteht man das Verhältnis von Differenz- zu Gleichtaktverstärkung:

$$CMRR = \frac{A_{DM}}{A_{CM}} \tag{4.111}$$

$$\approx \frac{g_m \cdot R_1}{\frac{R_1}{2 \cdot R_{SS}}} \tag{4.112}$$

$$= 2 \cdot R_{SS} \cdot g_m \tag{4.113}$$

Für eine große Differenzverstärkung A_{DM} benötigt man einen großen Widerstand R_1, eine große Gleichtaktunterdrückung erhält man, wenn die Stromquelle einen großen Innenwiderstand R_{SS} hat.

4.4 Transistoren als Widerstände

Um große Verstärkungen zu erhalten, waren in den betrachteten Schaltungen große Widerstände nötig. Diese lassen sich monolithisch nur unter großem Platzbedarf realisieren. Für große Kleinsignalverstärkungen muß der Widerstand aber nur im Kleinsignalverhalten groß sein, der Absolutwert spielt dafür keine Rolle. Wie wir gesehen haben, hat ein Transistor im Sättigungsbereich zwar einen kleinen Großsignalwiderstand, aber einen hohen Kleinsignalwiderstand.

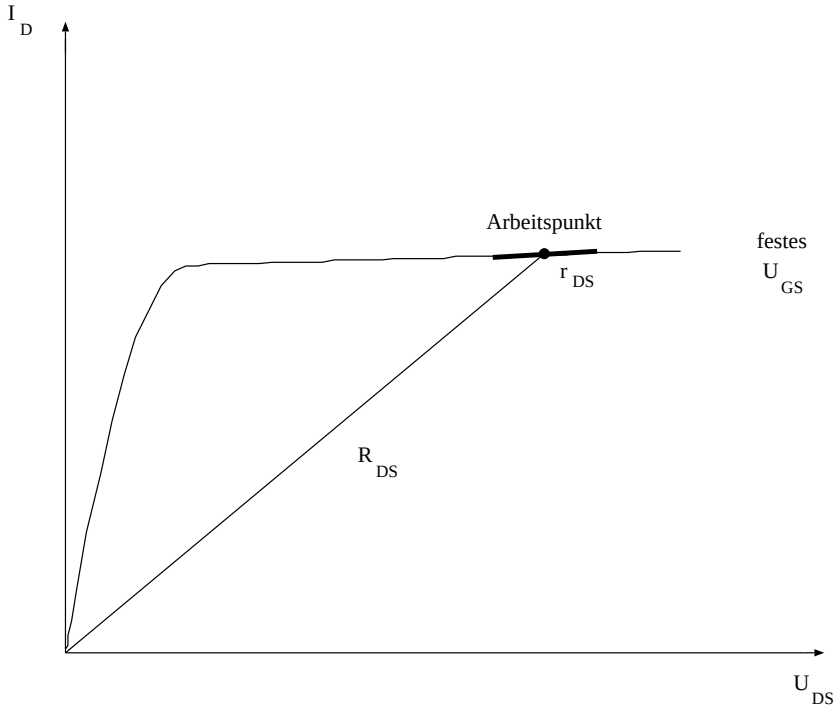


Abbildung 4.18: Differentieller Widerstand r_{DS} bei festem U_{GS}

Bild 4.18 verdeutlicht diesen Zusammenhang: Während R_{DS} (Großsignal) der Steigung der Verbindung Ursprung-Arbeitspunkt entspricht, ist r_{DS} (Kleinsignal) durch die Steigung der Tangenten im Arbeitspunkt gegeben:

$$G_{DS} = \frac{I_D}{U_{DS}} \qquad g_{DS} = \frac{dI_D}{dU_{DS}} \tag{4.114}$$

An Bild **??** erkennt man, daß ein möglichst kleines $U_{GS_{eff}}$ zwei Vorteile bietet:

- Die Steigung im Sättigungsbereich ist geringer $\Longrightarrow$ großes r_{DS}

- Der Transistor bleibt für größere Schwankungen der Spannung U_{DS} im Sättigungsbereich

Im Folgenden werden unterschiedliche Beschaltungen von Transistoren auf ihre Eignung als Kleinsignalwiderstand untersucht werden.

4.4.1 n-Kanal-Transistor als Diode

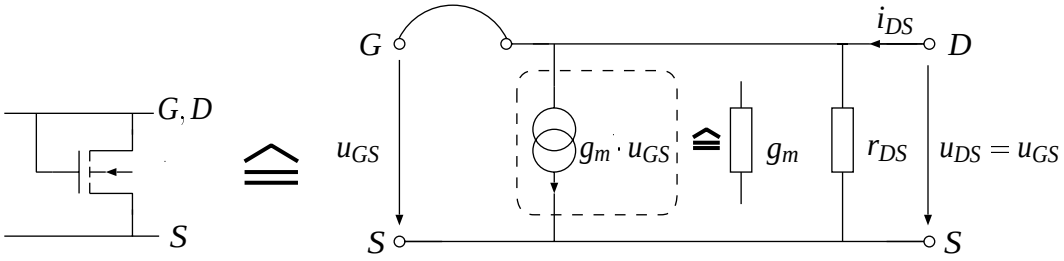


Abbildung 4.19: n-Kanal-Transistor als Diode geschaltet

In Bild 4.19 sind Gate und Drain des Transistors miteinander verbunden. Da $U_{GS} = U_{DS}$, befindet sich der Transistor in Sättigung und es gilt großsignalmäßig:

$$I_D(U_{DS}) = I_D(U_{GS}) = \frac{1}{2} \cdot \beta \cdot (U_{DS} - U_T)^2 \tag{4.115}$$

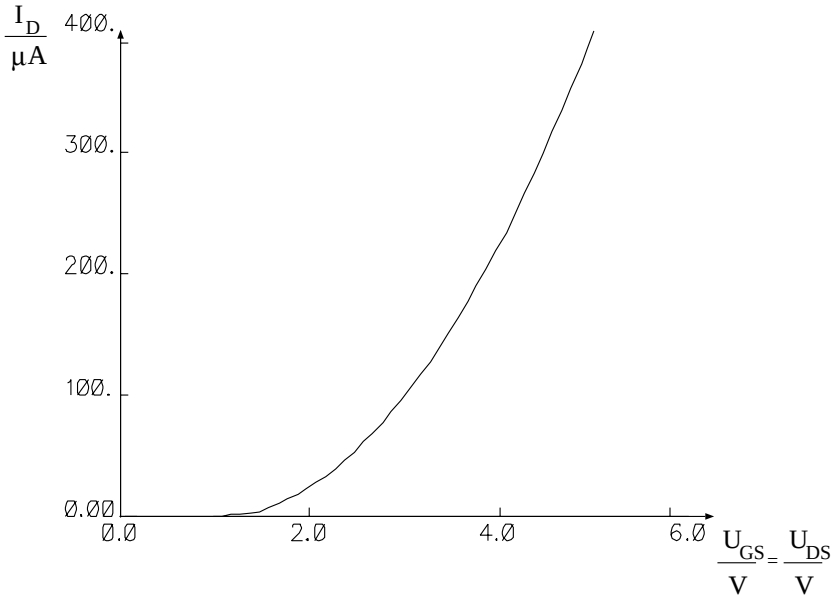


Abbildung 4.20: Kennlinie eines als Diode geschalteten n-Kanal-Transistors

Man erhält eine Kennlinie gemäß Bild 4.20. Daher bezeichnet man eine solche Beschaltung als Diode. Allerdings hat eine MOS-Diode keine Injektionladung, die unerwünschte parasitäre Effekte auslösen kann.
Der Abb.4.20 ist zu entnehmen, daß der Widerstand der MOS-Diode eher klein ist. Der Ausgangswiderstand ist $\frac{1}{g_m}$.

4.4.2 Transistor mit festem U_{GS}

Bei der MOS-Diode wurde das Gate-Potential des Widerstandstransistors auf sein Drain-Potential gelegt; damit ist die Spannung U_{GS} noch variabel. An der Ausgangskennlinie des Transistors (Bild 3.9) kann man erkennen, daß bei einem festen U_{GS} die Steigung der Geraden $I_D(U_{DS})$ fast Null ist, was einem großen differentiellen Widerstand entspricht. Deshalb wird der Transistor jetzt mit einer festen Spannungsquelle zwischen Gate und Source betrieben (siehe Bild 4.21 links). Für das Kleinsignalverhalten bedeutet das:

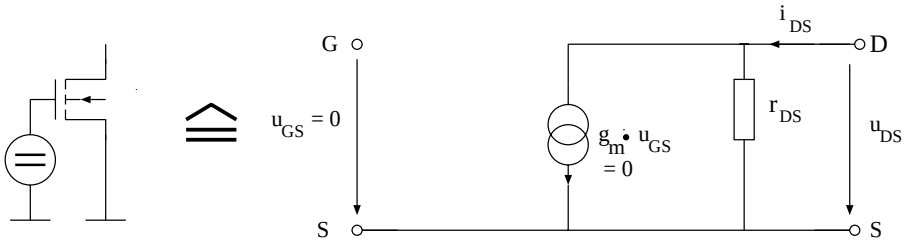


Abbildung 4.21: Kleinsignalersatzschaltbild bei festem U_{GS}

$$u_{GS} = 0 \tag{4.116}$$
$$r_{out} = r_{DS} \tag{4.117}$$

Wenn bei der Differenzstufe aus Bild 4.10 ein p-Kanal-Transistor mit fester Gate-Source-Spannung als Widerstand verwendet wird, sieht im Differenzmode eine Hälfte der Differenzstufe wie in Bild 4.22 aus:

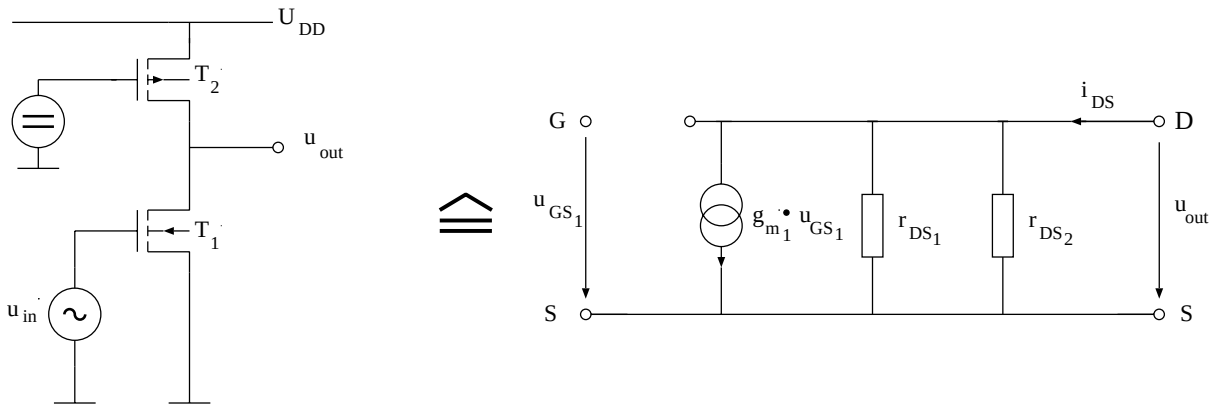


Abbildung 4.22: Linker Teil der Differenzstufe bei festem U_{GS}

$$A = -g \cdot r_{out} \tag{4.118}$$

$$g = \frac{i_{out}}{u_{in}}|_{u_{out}=0} \tag{4.119}$$

$$= \frac{g_{m_1} \cdot u_{GS_1}}{u_{in}} \tag{4.120}$$

$$= g_{m_1} \tag{4.121}$$

$$r_{out} = \frac{u_{out}}{i_{out}}|_{u_{in}=0} \tag{4.121}$$

$$= \frac{u_{out}}{u_{out} \cdot (g_{DS_1} + g_{DS_2})} \tag{4.122}$$

$$= \frac{1}{g_{DS_1} + g_{DS_2}} \tag{4.123}$$

$$\Longrightarrow \quad A = -\frac{g_{m_1}}{g_{DS_1} + g_{DS_2}} \tag{4.123}$$

$$\text{mit } g_{m_1} = 500\mu S\,, \quad = -\frac{500\mu S}{2.5\mu S + 2.5\mu S} \tag{4.124}$$

$$g_{DS_1} = g_{DS_2} = 2.5\mu S\text{:} \quad = -100 \tag{4.125}$$

Auf diese Art sind also erheblich höhere Verstärkungen erreichbar.

4.4.3 Feste Spannungsquelle

Dazu muß ein festes Potential erzeugt werden. Als Versuch kann Bild 4.23 links gelten. Der Transistor ist als Diode geschaltet. Man erhält einen Spannungsteiler, der über die Werte von R_1, W und L einstellbar ist.

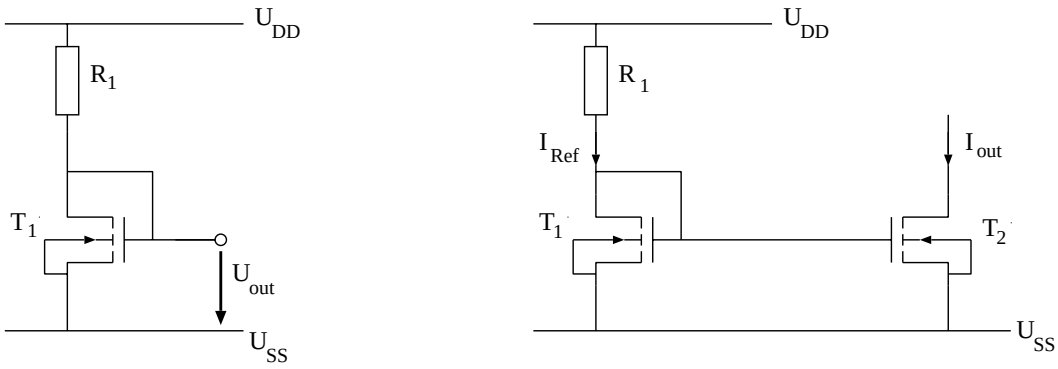


Abbildung 4.23: Spannungsteiler und Stromspiegel

4.4.4 Stromspiegel

Wird in Bild 4.23 ein zweiter Transistor an das feste Potential angeschlossen, so erhält man einen Stromspiegel (Bild 4.23 rechts). Wenn beide Transistoren gleich ($(\frac{W}{L})|_1 = (\frac{W}{L})|_2$) sind, gilt:

$$\text{mit} \qquad U_{GS_1} = U_{GS_2} = U_{GS}\text{:} \qquad I_{D_1} = I_{D_2} \tag{4.126}$$

Aus diesem Grund wird diese Schaltung als Stromspiegel bezeichnet. Sie kann als Stromquelle mit hohem Innenwiderstand $r_{out} = r_{DS}$ eingesetzt werden. Der Strom im linken Zweig wird als Referenzstrom bezeichnet. Er kann über den Transistor und den Widerstand eingestellt werden. Es gilt:

$$I_{Ref} = \frac{1}{2} \cdot \beta_1 \cdot (U_{GS} - U_T)^2 \tag{4.127}$$

$$U_{GS} = U_{DD} - U_{SS} - I_{Ref} \cdot R_1 \tag{4.128}$$

$$\Longrightarrow I_{Ref} = \frac{1}{2} \cdot \beta_1 \cdot (U_{DD} - U_{SS} - I_{Ref} \cdot R_1 - U_T)^2 \tag{4.129}$$

$$\Longrightarrow \sqrt{\frac{2 \cdot I_{Ref}}{\beta_1}} = U_{DD} - U_{SS} - I_{Ref} \cdot R_1 - U_T \tag{4.130}$$

$$\Longrightarrow I_{Ref} \cdot R_1 = U_{DD} - U_{SS} - U_T - \sqrt{\frac{2 \cdot I_{Ref}}{\beta_1}} \tag{4.131}$$

$$\Longrightarrow R_1 = \frac{U_{DD} - U_{SS} - U_T}{I_{Ref}} - \sqrt{\frac{2}{I_{Ref} \cdot \beta_1}} \tag{4.132}$$

Als Beispiel soll gelten:

$$U_{DD} - U_{SS} = 5V \tag{4.133}$$
$$U_T = 1V \tag{4.134}$$
$$\beta_1 = 50\frac{\mu A}{V^2} \tag{4.135}$$

Damit ergibt sich für den Großsignalwiderstand:

$$R_1 = \frac{4V}{I_{Ref}} - \sqrt{\frac{2}{I_{Ref} \cdot 50\frac{\mu A}{V^2}}} \tag{4.136}$$

Für einen Strom von $250\mu A$ benötigt man damit einen Widerstand der Größe $3.3k\Omega$. Um die Verlustleistung gering zu halten, soll der Strom möglichst klein sein. Für einen Strom von $20\mu A$ muß der Wert des Widerstands aber schon $340k\Omega$ betragen.

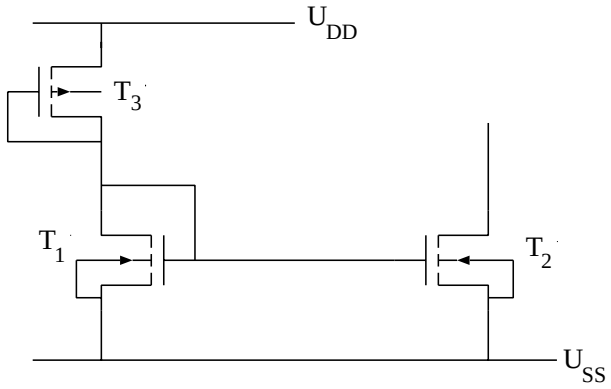


Abbildung 4.24: Stromspiegel mit p-Kanal-Transistor als Widerstand

Auch hier wird der Widerstand durch einen als Diode geschalteten Transistor realisiert, in Bild 4.24 durch einen p-Kanal-Transistor T_3. Für diesen wählt man ein kleines $\frac{W}{L}$ - Verhältnis. Damit ist der differentielle Widerstand zwar nicht sehr hoch ($\frac{1}{g_m}$), aber der Absolutwert ist hoch.

Die Schaltung in Bild 4.24 entspricht einem gegengekoppeltem Inverter: Der Ausgang des n-MOS-Inverters ist auf den Eingang (Gate) des p-MOS-Inverters geschaltet und umgekehrt. Der Strom wird durch die Spannung $U_{GS_{eff}}$ bestimmt. Wenn die beiden Transistoren T_1 und T_3 äquivalent ausgelegt sind (d.h. $\beta_n = \beta_p$), gilt für $U_{GS_{eff}}$:

$$U_{GS_{eff}} = \frac{U_{DD} - U_{SS}}{2} - U_T \tag{4.137}$$
$$= \frac{U_{DD} - U_{SS} - 2 \cdot U_T}{2} \tag{4.138}$$
$$\text{mit den Beispielwerten:} \quad = \frac{5V - 2V}{2} = 1.5V \tag{4.139}$$

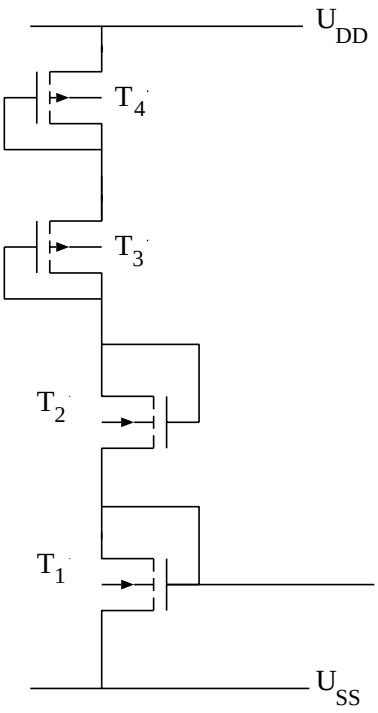


Abbildung 4.25: Stromspiegel mit 4 Transistoren

Die beiden Drainströme können daher sehr groß werden:

$$U_{DD} - U_{SS} = U_{GS_3} + U_{GS_1} \tag{4.140}$$

$$= \sqrt{\frac{2 \cdot I_{D_3}}{\beta_3}} + U_{T_3} + \sqrt{\frac{2 \cdot I_{D_1}}{\beta_1}} + U_{T_1} \tag{4.141}$$

$$= 2 \cdot (\sqrt{\frac{2 \cdot I_D}{\beta}} + U_T) \tag{4.142}$$

Um den Strom zu verringern, kann man mehrere Transistoren verwenden und so den Strom über ein kleineres U_{GS} einstellen. Beispielsweise gilt in Bild 4.25:

$$U_{GS_{eff}} = \frac{U_{DD} - U_{SS} - 4 \cdot U_T}{4} \tag{4.143}$$

$$\text{mit den Beispielwerten:} \quad = \frac{5V - 4V}{4} = 0.25V \tag{4.144}$$

Eine Stromeinstellung ist auch über die $\frac{W}{L}$ - Verhältnisse möglich, aber die Variante aus Bild 4.25 ist trotz größerer Zahl von Transistoren platzsparender. Allerdings ist die Schwankung des Referenzstromes infolge der Prozessschwankungen (V_T) auch größer.

4.4.5 Layout von Stromspiegelung

Für viele Anwendungen ist ein genaues Stromübersetzungsverhältnis erforderlich. In der Abbildung 4.26 wird der Referenzstrom mit dem Verhältnis von $T_2 : T_1$ auf das Drain von T_2 übersetzt. Für ein Übersetzungsverhältnis von z.B. 3 muss es zunächst gelten, dass $\frac{W_2}{L_2}$ dass dreifache von $\frac{W1}{L1}$ ist. Wenn jedoch L_2 und L_1 unterschiedlich ist, ist das Verhältnis zwischen den beiden effektive Kanallängen nicht identisch wie das Verhältnis zwischen den nominalen Kanallängen $\frac{L_2}{L_1}$. Der Grund liegt an der Unterdiffusion der Drain- und Sorcegebiete, die in Kapitel 3.3.3 beschrieben worden ist. Diese Unterdiffusion variiert von Waferlos zu Waferlos, von Wafer zu Wafer und auch von Chip zu Chip, sodass das Verhältnis zwischen den beiden effektiven Kanallängen nicht konstant ist.

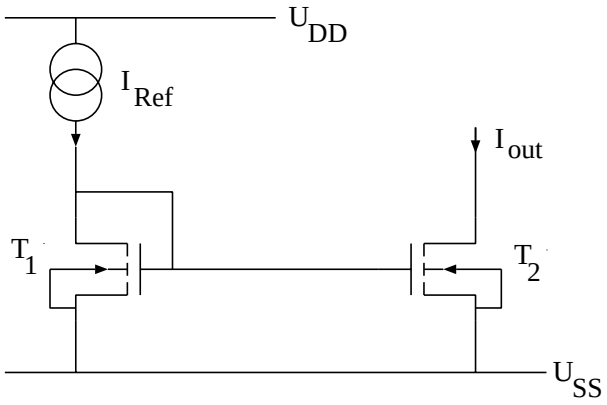


Abbildung 4.26: Stromspiegel mit idealer Stromquelle

Folglich ist es sinnvoll, die Kanallänge immer auf einen Wert für alle Transistoren im Stromspiegel festzulegen. Mit der minimalen Kanallänge ist die benötigte Fläche auch minimal. Allerdings haben Transistoren mit größerer Kanallänge eine flachere (idealere) Ausgangskennlinie. Dies würde die Qualität des Stromspiegels bzw. der Stromquelle verbessern. Die Problematik mit der unterschiedlichen Kanallänge gilt natürlich genau so für die unterschiedlichen Kanalweiten. Die effektive Kanalweite wurde auch in Kapitel 3.3.3 beschrieben. Da das Übersetzungsverhältnis nicht immer 1 betragen kann, muss das Problem im Layout gelöst werden. Beispielsweise wird der Transistor 1 mit jeweils 10μm Kanallänge und 10μm Kanalweite dimensioniert. Die Kanallänge des Transistors 2 beträgt ebenfalls 10μm. Die Kanalweite des Transistors 2 ist jedoch nicht 30μm, sondern 3x10μm. Dies bedeutet, dass anstatt 1 Transistor mit 30μm Kanalweite je 3 gleiche Transistoren mit jeweils 10μm Kanalweite verwendet werden. Alle diese 3 Transistoren sind identisch wie der Transistor 1. Damit wird ein genaues Übersetzungsverhältnis von 3 realisiert.
In der Abbildung 4.27 wird dieses Layout dargestellt. Der Transistor 1 ist ein solcher Einheitstransistor. Der Transistor 2, dessen Gate und Sorce jeweils mit dem Gate und Sorce von Transistor 1 kurz geschlossen sind, hat 3 aktive Gebiete, die die Kanalweite definieren. Es geht aber nicht nur um Einsatz von Einheitszellen. Auch die Ausrichtungen der Transistoren müssen gleich sein, da sonst aufgrund der Prozessschwankungen die effektive Kanallänge bzw. Kanalweite doch unterschiedlich sein können. Auch die Transistoren 1 und 2 müssen geometrisch nahe beieinander sein, weil der Prozessparameter auch über die Fläche variieren kann. Das dargestellt Prinzip gilt nicht nur für Stromspiegel, sondern auch für andere Strukturen, wie z.B. den Teiler, wo es auf

genaue Verhältnisse ankommt. Während die absolute Größe, wie z.B. die Schwellspannung eines Transistors oder der Widerstand stark variieren kann (z.B. 20%), kann das Verhältnis zwei gleicher Elemente sehr genau eingehalten werden. Man spricht hier auch vom Match, dass je nach verwendeten Bauelementen bis auf 0,1% genau sein kann.

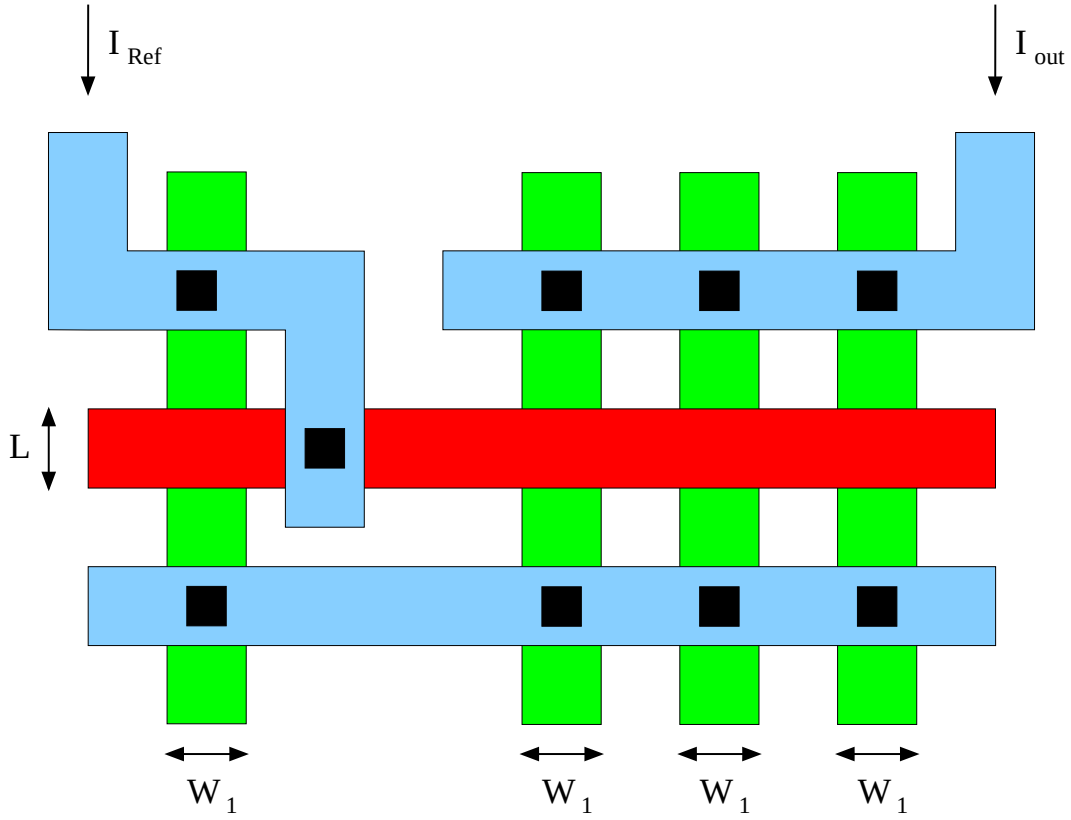


Abbildung 4.27: Layout bei $W_2 = 3 \cdot W_1$

5 Grundlagen digitaler Schaltungstechnik

Im Folgenden wird die CMOS-Schaltungstechnik im Vordergrund der Betrachtungen stehen. Für deren Verständnis ist es jedoch sinnvoll, zuerst einen Blick auf die älteren zu werfen.

5.1 Digitale MOS-Schaltungstechnik

Im Unterschied zu den Bipolartransistoren sind MOS-Transistoren spannungsgesteuert. Dadurch benötigen MOS-Gatter keine permanenten Eingangsströme. MOS-Gatter bestehen prinzipiell aus einem schaltenden Logikteil, der aus n-Kanal-Transistoren aufgebaut ist, und einem Lastelement, welches im einfachsten (aber ungebräuchlichen Fall) ein Widerstand, in der Regel ein Transistor als nichtlinearer Widerstand und in komplexeren Schaltungstechniken ein ganzes Netzwerk von Transistoren ist.

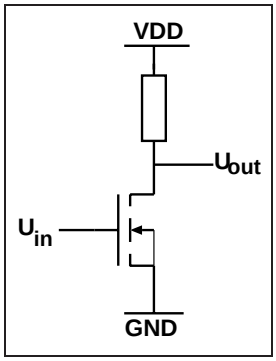


Abbildung 5.1: Inverter mit Widerstandslast

Bei niedrigem Eingang $(U_{in} - U_T)$ sperrt der Transistor in Abb. 5.1 und der Ausgangspegel V_{DD} wird voll erreicht. Da jedoch bei hohem Eingangssignal (konsequenterweise $U_{\mathrm{in}} = V_{\mathrm{DD}}$)[1] ein leitender Pfad von V_{DD} nach GND existiert, wird $U_{\mathrm{out}} = 0$ nicht erreicht. In diesem Fall hängt die Ausgangsspannung von der Kennlinie des Transistors (I_{D}) und dem Widerstand R ab: $U_{\mathrm{out}} = V_{\mathrm{DD}} - I_{\mathrm{D}}R$.

Der Widerstand in dieser Schaltung kann prinzipiell durch jede beliebige Art von Last ersetzt werden. Damit der Strom nicht zu groß wird sollte R dabei selbst möglichst groß sein, was jedoch

[1] Bei integrierten Schaltungen sind die Eingänge nahezu aller Gatter mit Ausgängen anderer Gatter verbunden

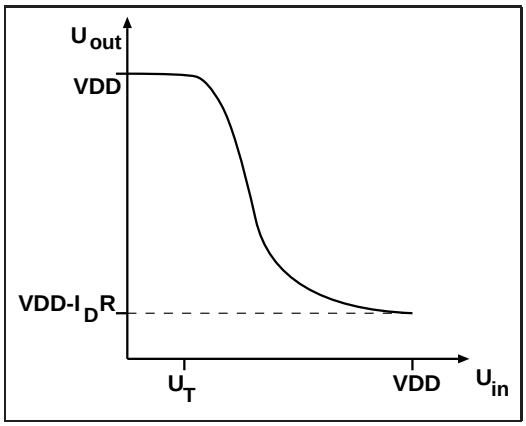


Abbildung 5.2: Übertragungskennlinie des Inverters aus Abb. 5.1

die Schaltvorgänge verlangsamt. Eine andere Möglichkeit besteht darin, Stromquellen als Last zu verwenden.

5.2 Pseudo-NMOS — Ein p-Kanal-Transistor als Last

Ähnliches Verhalten wie bei NMOS-Inverter erhält man, wenn man statt des Widerstand einen p-Kanal-Transistor einsetzt, dessen Gate an Masse angeschlossen wird (Abb. 5.3). Damit verhält sich die Last wie eine — wenn auch nichtideale — Stromquelle.

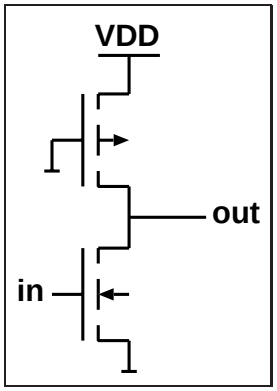


Abbildung 5.3: Pseudo-NMOS-Inverter

Bei einer solchen Schaltung gilt für den p-Kanal-Transistor $U_{GS} = -V_{DD} < U_T$, weshalb der Transistor ebenfalls immer leitend ist[2]. Dadurch ergibt sich eine Kennlinie ähnlich der entsprechenden NMOS Schaltung.

Da moderne Herstellungsverfahren in der Regel auf die später noch betrachtete CMOS-Technologie ausgerichtet sind, erfordert die Herstellung solcher p-Kanal-enhancement-Transistoren keinen zu-

[2] Bei p-Kanal-Transistoren sind die Relationszeichen umgedreht!

sätzlichen Aufwand. Solche pseudo-NMOS-Schaltungen sind damit auch in den üblichen CMOS-Prozessen günstig herstellbar. Sie haben — wie später besonders bei den Speichern gezeigt wird — auch heute noch Bedeutung.

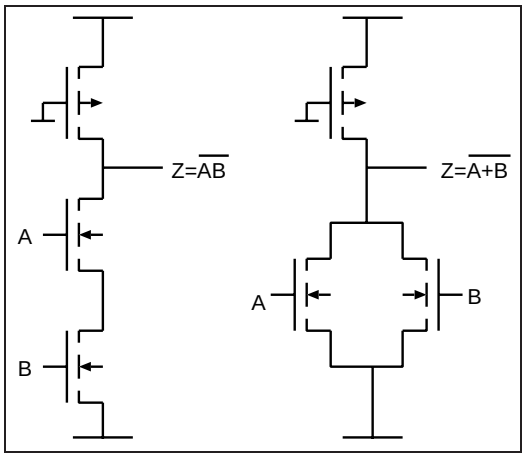


Abbildung 5.4: Pseudo-NMOS NAND und NOR

5.3 CMOS — Komplementär MOS

Wenn man nun schaltungstechnisch erreicht, daß bei hohem Eingangspotential nur der n-Transistor leitet, d.h. daß gleichzeitig die Last sperrt, wenn $U_{\text{in}} = V_{\text{DD}}$ ist, würde der Ausgang auch $0V$ voll erreichen. Dies kann einfach dadurch geschehen, daß das Gate des p-Transistors statt fest mit Masse mit dem Eingang selbst verbunden wird (Abb. 5.5).

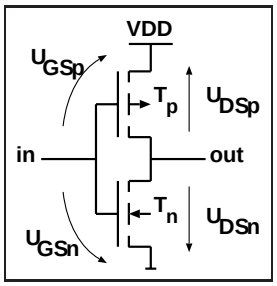


Abbildung 5.5: CMOS Inverter

Hier gilt:

- $U_{\text{GSn}} = U_{\text{in}}, U_{\text{GSp}} = U_{\text{in}} - V_{\text{DD}}$.
- Für $U_{\text{in}} =$"0"[3]: $U_{\text{GSp}} = -V_{\text{DD}}$, d.h. T_p ist leitend. $U_{\text{GSn}} = 0$, d.h. T_n sperrt. Damit wird

$$U_{\text{out}} = V_{\text{DD}}$$

[3] „0“ bedeutet hier logisch 0; elektrisch gesehen heißt das $U_{\text{in}} < U_{\text{Tn}}$.

- Für $U_{\text{in}} =$“1“[4]: $U_{\text{GSp}} = 0$, d.h. T_p sperrt. $U_{\text{GSn}} = V_{\text{DD}}$, d.h. T_n leitet. Damit wird

$$U_{\text{out}} = 0V$$

In Abbildung 5.6 ist das Schaltermodell des CMOS Inverters für die beiden Fälle $U_{\text{in}} = V_{\text{DD}}$ und $U_{\text{in}} = 0V$ dargestellt.

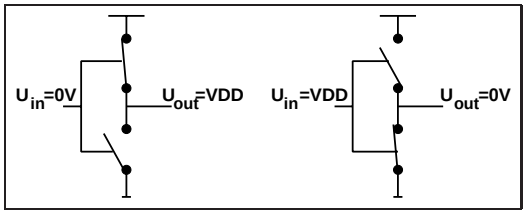


Abbildung 5.6: Schaltermodell des CMOS Inverters

Dieser geschickte komplementäre Aufbau ist — wie man später noch sehen wird — für alle CMOS Gatter gegeben. Grundsätzlich ist (außer beim Umschalten) immer einer der beiden Schaltungsteile (n-Teil, hier: n-Transistor; p-Teil, hier: p-Transistor) gesperrt, während der jeweils andere leitend ist, so daß kein statischer Strom fließen kann und die Ausgangspegel das Niveau der vollen Versorgungsspannungen erreichen.

Zur Herleitung der Kennlinie stellt man folgende Überlegungen an:

- Solange $U_{\text{in}} < U_{\text{T}}$ ist, ist der Transistor T_n gesperrt. Da T_p wegen $U_{\text{GSp}} = U_{\text{in}} - V_{\text{DD}} < U_{\text{T}}$ leitend ist, wird damit $U_{\text{out}} = V_{\text{DD}}$. Diese Spannung fällt über dem Transistor T_n ebenfalls ab. Am p-Transistor ergibt sich damit ein $U_{\text{DS}} = U_{\text{out}} - V_{\text{DD}} = 0V$. Damit befindet sich der Transistor T_p im linearen Bereich.

- Wird U_{in} geringfügig größer als U_{T}, so wird der Transistor T_n ebenfalls leitend. Noch fällt eine sehr große Spannung über diesem Transistor ab, so daß $U_{\text{GS}} < U_{\text{DS}}$ ist, d.h. der Transistor befindet sich im Sättigungsbereich. Da sich die Ausgangsspannung hier ebenfalls nur geringfügig ändert, bleibt T_p anfangs noch im linearen Bereich.

- Steigt U_{in} weiter an, so wird irgendwann der Punkt erreicht, an dem $U_{\text{GS}} - U_{\text{T}} = U_{\text{DS}}$ ist, ab dem der Sättigungsbereich verlassen und der lineare Bereich betreten wird. An genau diesem Punkt geht — wie später noch gezeigt wird — der Transistor T_p in den Sättigungsbereich über.

- Eine weitere Erhöhung von U_{in} führt irgendwann dazu, daß $U_{\text{GSp}} = U_{\text{in}} - V_{\text{DD}}$ größer wird als $U_{\text{T}}(T_p)$ (Achtung! Beides sind negative Zahlen!), so daß T_p in den Sperrbereich kommt. Dann wird $U_{\text{out}} = 0V$.

[4] „1“ bedeutet hier logisch 1; elektrisch gesehen heißt das $U_{\text{in}} > U_{\text{Tp}} + V_{\text{DD}}$.

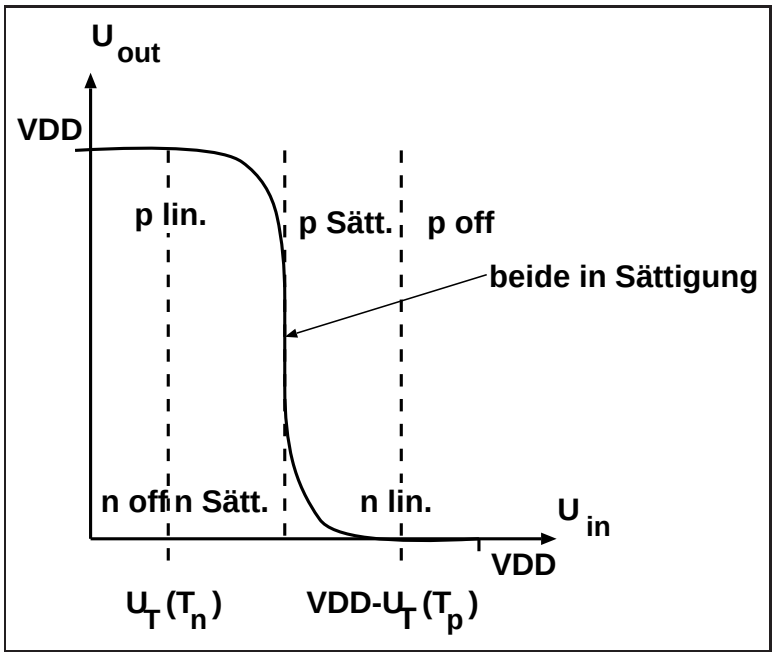


Abbildung 5.7: Übertragungskennlinie des CMOS Inverters

5.3.1 Genauere Betrachtung der Kennlinie des CMOS Inverters

Für die nun folgenden Überlegungen wird eine Bereichseinteilung der Kennlinie entsprechend Abb. 5.8 zugrunde gelegt. Gegenüber der alten NMOS-Technik Abb. 5.2 fällt auf, daß bei hohem Eingang der Ausgang Null ist. Für die folgende CMOS-Stufe bedeutet dies einen hohen Störabstand, da die Schwellenspannung von NMOS deutlich unterschritten ist. Das gilt natürlich für Ausgang=high, da der PMOS der folgenden Stufe sicher aus ist. Man spricht auch von einem hohen Noise Margin (Störabstand)

Allgemein gilt für die Schaltung nach Abb. 5.5:

$$\begin{aligned} U_{\mathrm{GS}}(T_n) &= U_{\mathrm{in}} \\ U_{\mathrm{GS}}(T_p) &= U_{\mathrm{in}} - V_{\mathrm{DD}} \\ U_{\mathrm{DS}}(T_n) &= U_{\mathrm{out}} \\ U_{\mathrm{DS}}(T_p) &= U_{\mathrm{out}} - V_{\mathrm{DD}} \\ U_{\mathrm{T}}(T_p) &= \text{Schwellenspannungp} - \text{Kanal} \\ U_{\mathrm{T}}(T_n) &= \text{Schwellenspannungn} - \text{Kanal} \end{aligned}$$

Weiter soll gelten:

$$U_{\mathrm{T}}(T_n) = -U_{\mathrm{T}}(T_p) = U_{\mathrm{T}}$$

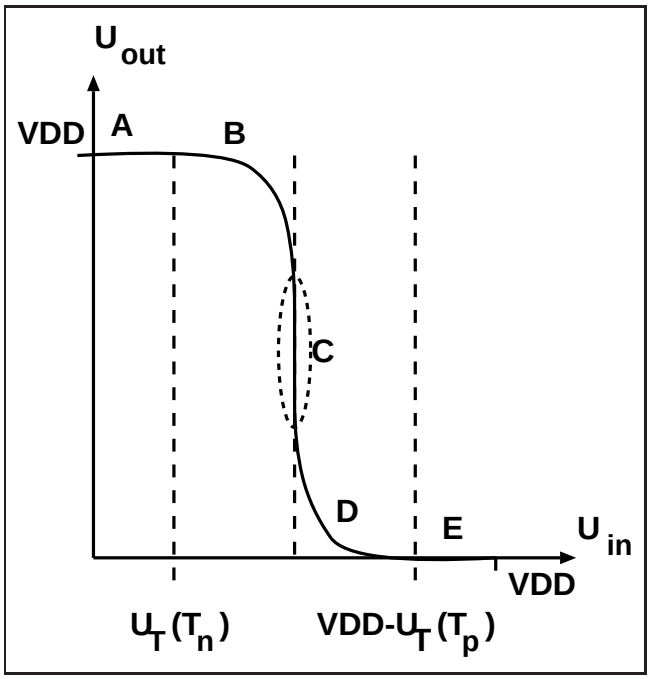


Abbildung 5.8: Bereiche der Kennlinie des CMOS Inverters

Bereich A: $0 \leq U_{\text{in}} \leq U_{\text{T}}(T_n)$

Da $U_{\text{GS}}(T_n) < U_{\text{T}}(T_n)$ ist Transistor T_n gesperrt und damit $I_{\text{D}}(T_n) = 0$. Nun ist $-V_{\text{DD}} \leq U_{\text{GS}}(T_p) = U_{\text{in}} - V_{\text{DD}} \leq U_{\text{T}}(T_n) - V_{\text{DD}}$ und damit mit Sicherheit $U_{\text{GS}}(T_p) < U_{\text{T}}(T_p)$ und T_p ist leitend. Da am Ausgang kein Strom herausfließt und auch durch T_n kein Strom fließen kann, muß $I_{\text{D}}(T_p)$ ebenfalls verschwinden. Dadurch ist klar. In welchem Bereich sich T_p befindet kann durch einsetzen der beiden (theoretisch) noch möglichen Kennlinienbereiche von T_p in die Gleichung $I_{\text{D}}(T_p) = 0$ herausgefunden werden. Da $\frac{\beta_p}{2}(U_{\text{in}} - U_{\text{T}}(T_p))^2 \neq 0$, kann sich der Transistor nicht im Sättigungsbereich befinden. Damit ist nur der lineare Bereich möglich:

$$\begin{aligned} \beta_p \left(U_{\text{in}} - V_{\text{DD}} - U_{\text{T}}(T_p) - \frac{1}{2}(U_{\text{out}} - V_{\text{DD}}) \right)(U_{\text{out}} - V_{\text{DD}}) &= 0 \\ \Rightarrow U_{\text{out}} - V_{\text{DD}} &= 0 \\ \Rightarrow U_{\text{out}} &= V_{\text{DD}} \end{aligned}$$

Bereich B: $U_{\text{T}}(T_n) < U_{\text{in}}$ **und** $U_{\text{in}} - U_{\text{T}}(T_n) < U_{\text{out}}$

Hier befindet sich T_n im Sättigungsbereich. Da $U_{\text{GS}}(T_p) = U_{\text{in}} - V_{\text{DD}}$ folgt $U_{\text{GS}}(T_p) - U_{\text{T}}(T_p) = U_{\text{in}} - U_{\text{T}}(T_p) - V_{\text{DD}} < U_{\text{out}} - V_{\text{DD}}$. Damit befindet sich T_p immer noch im linearen Bereich. Die

Gleichung der Kennlinie ergibt sich dann mit $I_D(T_n) = -I_D(T_p)$ aus den Gleichungen:

$$\begin{aligned} I_D(T_n) &= \frac{\beta_n}{2}(U_{in} - U_T(T_n))^2 \\ I_D(T_p) &= \beta_p \left[U_{in} - V_{DD} + U_T(T_p) - \frac{1}{2}(U_{out} - V_{DD})\right](U_{out} - V_{DD}) \end{aligned}$$

Bereich C: $U_{in} - U_T(T_n) = U_{out}$

T_n befindet sich gerade noch in Sättigung. An diesem Punkt gilt auch $U_{in} - V_{DD} - U_T(T_p) \leq U_{out} - V_{DD} = U_{in} - U_T(T_n)$ (da $U_T(T_p) < 0 < U_T(T_n)$), d.h. $U_{GS}(T_p) - U_T = U_{DS}(T_p)$, so daß sich auch T_p gerade in Sättigung begeben hat. Damit gelten die Stromgleichungen:

$$\begin{aligned} I_D(T_n) &= \frac{\beta_n}{2}(U_{in} - U_T(T_n))^2 \\ I_D(T_p) &= \frac{\beta_p}{2}(U_{in} - V_{DD} - U_T(T_p))^2 \end{aligned}$$

U_{out} läßt sich über diese beiden Gleichungen nicht bestimmen, da es darin gar nicht vorkommt. Dies bedeutet, daß bei solchen Spannungsverhältnissen in einem Bestimmten Bereich jeder beliebige Wert von U_{out} möglich ist. In der Kennlinie äußert sich das durch einen senkrechten Abfall in diesem Punkt, was ein Ablesen des Punktes aus der Kennlinie sehr schwierig macht. Für einen Inverter wäre es sinnvoll, diesen Punkt genau bei $U_{in} = \frac{V_{DD}}{2}$ zu erreichen. Aus obigen Gleichungen lassen sich bei gegebenem V_{DD} die Parameter $\beta_n, \beta_p, U_T(T_n)$ und $U_T(T_p)$ entsprechend bestimmen[5].

Bereich D: $U_{out} < U_{in} - U_T(T_n)$ **und** $U_{in} - V_{DD} < U_T(T_p)$

Hier ist nun T_n im linearen Bereich seiner Kennlinie angelangt. T_p bleibt im Sättigungsbereich. Dieser Bereich ist damit analog zum Bereich B zu behandeln. Die Gleichung der Kennlinie ergibt sich hier aus:

$$\begin{aligned} I_D(T_n) &= \beta_n \left(U_{in} - U_T(T_n) - \frac{1}{2}U_{out}\right)U_{out} \\ I_D(T_p) &= \frac{\beta_p}{2}(U_{in} - V_{DD} - U_T(T_p))^2 \end{aligned}$$

Bereich E: $U_{in} - V_{DD} \geq U_T(T_p)$

Da hier $U_{GS}(T_p) \geq U_T(T_p)$ ist, befindet sich T_p im Sperrbereich. Hier kann entsprechend Bereich A für T_p berechnet werden, daß dieser sich im linearen Bereich befindet. Damit ergibt sich

$$U_{out} = 0$$

[5] Wie später noch gezeigt wird, wählt man oft $\beta_n = \beta_p$. Hier erhält man dann als Ergebnis $U_T(T_p) = -U_T(T_n)$

5.3.2 Verzögerungszeiten des CMOS Inverters

In der Regel wird an den Ausgang eines Inverters der Eingang eines anderen CMOS-Gatters angeschlossen sein. Durch die Spannungssteuerung verhält sich diese Last rein kapazitiv, so daß man ein Ersatzschaltbild entsprechend Abb. 5.9 erhält.

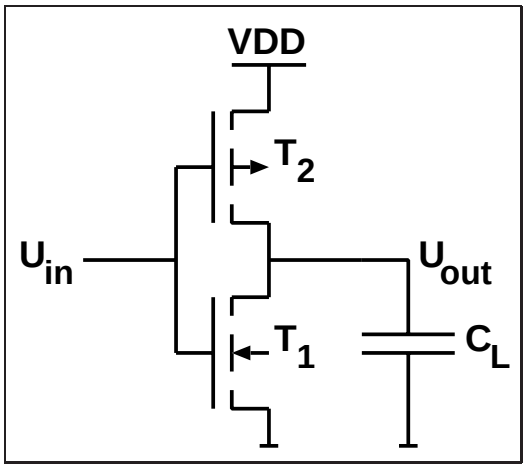


Abbildung 5.9: CMOS Inverter mit kapazitiver Last

In den folgenden Überlegungen wird ein Vorgang entsprechend Abb. 5.10 betrachtet, bei dem die — zuerst auf V_{DD} aufgeladene — Kapazität C_{L} über den Inverter ganz entladen, danach wieder voll aufgeladen wird.

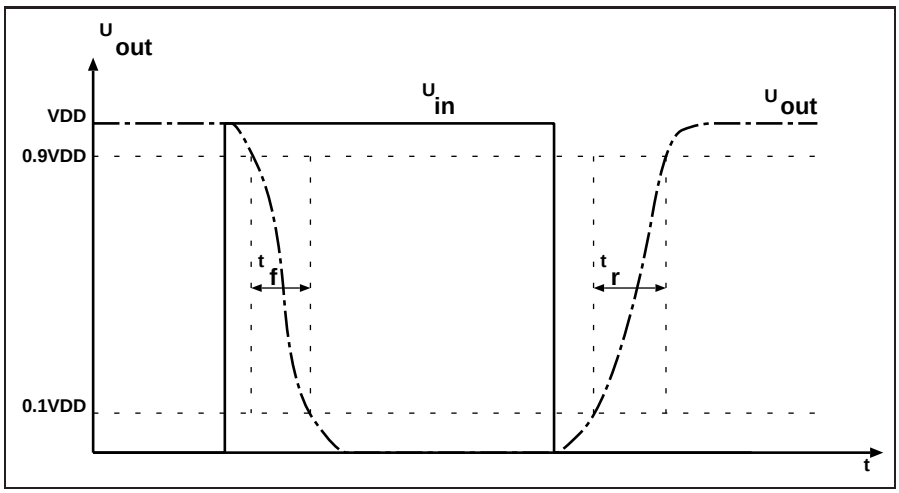


Abbildung 5.10: Umladen einer Lastkapazität über den Inverter

Da nicht alle Schaltungstechniken die Pegel V_{DD} und GND voll erreichen, begnügt man sich mit der Näherung t_f,während dieser fällt Zeit die Spannung von$0.9V_{DD}$ auf $0.1V_{DD}$. Da V_T üblicherweise $0.2V_{DD}$ ist, sind bei $0.9V_{DD}$ die PMOS immer aus und die NMOS immer ein, während bei $0.1V_{DD}$ die PMOS immer ein und die NMOS immer aus sind. t_f läßt sich hier aus Integralen über den

Sättigungsbereich (U_{out} von $0.9V_{\text{DD}}$ bis $V_{\text{DD}} - U_{\text{T}}$) und den linearen Bereich (U_{out} von $V_{\text{DD}} - U_{\text{T}}$ bis $0.1V_{\text{DD}}$) recht einfach bestimmen.

Für die Schaltung gilt: $-\frac{C_L}{I_{\text{D}}} \cdot dU_{\text{out}} = dt$.

Mit

$$\begin{aligned} & Q_C = C_L U_{\text{out}} \\ \Rightarrow \quad & \frac{d}{dt} Q_C = I = C_L \frac{d}{dt} U_{\text{out}} \\ \Rightarrow \quad & -\frac{I_D}{C_L} = \frac{d}{dt} U_{\text{out}} \end{aligned}$$

läßt sich t_f folgendermaßen bestimmen:

$$\begin{aligned} t_f &= \int_{V_{\text{DD}}-U_{\text{T}}}^{0.9 \cdot V_{\text{DD}}} \frac{2C_L}{\beta_n} \cdot \frac{dU_{\text{out}}}{(U_{\text{GS}} - U_{\text{T}})^2} + \int_{0.1V_{\text{DD}}}^{V_{\text{DD}}-U_{\text{T}}} \frac{2C_L}{\beta_n} \frac{dU_{\text{out}}}{[2(U_{\text{GS}} - U_{\text{T}}) - U_{\text{out}}] U_{\text{out}}} \\ &= \frac{2C_L}{\beta_n} \left[\frac{U_{\text{T}} - 0.1V_{\text{DD}}}{(U_{\text{GS}} - U_{\text{T}})^2} + \frac{1}{2(U_{\text{GS}} - U_{\text{T}})} ln \left(\frac{V_{\text{DD}} - U_{\text{T}}}{2(U_{\text{GS}} - U_{\text{T}}) - (U_{\text{GS}} - U_{\text{T}})} \frac{2(U_{\text{GS}} - U_{\text{T}}) - 0.1V_{\text{DD}}}{0.1V_{\text{DD}}} \right) \right] \end{aligned}$$

Mit der Angabe $U_{\text{GS}} = V_{\text{DD}}$ und der Näherung $U_{\text{T}} \approx 0.2V_{\text{DD}}$ läßt sich die Formel vereinfachen[6]:

$$\begin{aligned} t_f &\approx \frac{2C_L}{\beta_n} \left[\frac{0.1V_{\text{DD}}}{0.64V_{\text{DD}}^2} + \frac{1}{1.6V_{\text{DD}}} \ln \left(\frac{0.8V_{\text{DD}}}{1.6V_{\text{DD}} - 0.8V_{\text{DD}}} \frac{1.6V_{\text{DD}} - 0.1V_{\text{DD}}}{0.1V_{\text{DD}}} \right) \right] \\ &\approx \frac{4C_L}{\beta_n V_{\text{DD}}} \end{aligned}$$

Analog läßt sich die Anstiegszeit für das Aufladen der Kapazität von $0.1V_{DD}$ auf $0.9V_{DD}$ bestimmen zu:

$$t_r \approx \frac{4C_L}{\beta_p V_{\text{DD}}}$$

Für den Inverter kann es sinnvoll sein, daß Auf- und Entladevorgang möglichst gleich schnell vonstatten gehen, d.h. t_f und t_r sollen gleich groß sein. Dies läßt sich allgemein durch $\beta_n = \beta_p$ erfüllen, was in folgender Rechnung ausgeführt ist.

$$\begin{aligned} \beta_p &= \beta_n \\ \mu_p \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}} \frac{W_p}{L_p} &= \mu_n \frac{\varepsilon_{\text{ox}}}{t_{\text{ox}}} \frac{W_n}{L_n} \\ \text{wobei wie üblich} \quad & L_p = L_n = L \\ \frac{W_p}{W_n} &= \frac{\mu_n}{\mu_p} \approx 3 \end{aligned}$$

[6] Beim p-Kanal-Transistor würde hier bei der Berechnung der Anstiegszeit t_r stehen: $|U_{\text{T}}| \approx 0.2V_{\text{DD}}$

Macht man den p-Kanal-Transistor dreimal so weit wie den n-Kanal-Transistor, so schaltet der Inverter symmetrisch.

Im Normalfall wird die Lastkapazität selbst wieder die Eingangskapazität eines anderen Gatters sein. Um ein brauchbares Maß zu besitzen, setzt man der Einfachheit halber erst einmal als Last die Eingangskapazität eines Inverters ein, die sich aus den Gatekapazitäten des entsprechend p- Transistors bzw. n-Transistors ergibt:

$$\begin{aligned} C_L &= C_{\mathrm{Gn}} + C_{\mathrm{Gp}} \\ &= W_n L_n C'_{\mathrm{ox}} + W_p L_p C'_{\mathrm{ox}} \\ &\quad \text{mit } L_n = L_p = L, W_p = \frac{\mu_n}{\mu_p} W_n \\ C_L &= W L C'_{\mathrm{ox}} \left(\frac{\mu_n}{\mu_p} + 1 \right) \end{aligned}$$

Für die Gesamtverzögerung t_{ges} ergibt sich:

$$\begin{aligned} t_{\mathrm{ges}} &= t_f + t_r \\ &= \frac{4C_L}{\beta_n V_{\mathrm{DD}}} + \frac{4C_L}{\beta_p V_{\mathrm{DD}}} \\ &= \frac{4WLC'_{\mathrm{ox}}}{V_{\mathrm{DD}}} \left(1 + \frac{\mu_n}{\mu_p}\right) \left(\frac{1}{\beta_n} + \frac{1}{\beta_p}\right) \\ &= \frac{4W_n LC'_{\mathrm{ox}}}{V_{\mathrm{DD}}} \left(1 + \frac{\mu_n}{\mu_p}\right) \left(\frac{1}{\mu_n C'_{\mathrm{ox}} \frac{W_n}{L}} + \frac{1}{\mu_p C'_{\mathrm{ox}} \frac{W_p}{L}}\right) \\ &= \frac{4W_n LC'_{\mathrm{ox}}}{V_{\mathrm{DD}}} \left(1 + \frac{\mu_n}{\mu_p}\right) \left(\frac{1}{\mu_n C'_{\mathrm{ox}} \frac{W_n}{L}} + \frac{1}{\mu_p C'_{\mathrm{ox}} \frac{\frac{\mu_n}{\mu_p} W_n}{L}}\right) \\ &= \frac{4W_n LC'_{\mathrm{ox}}}{V_{\mathrm{DD}}} \left(1 + \frac{\mu_n}{\mu_p}\right) \left(\frac{2L}{\mu_n C'_{\mathrm{ox}} W_n}\right) \\ &= \frac{8L^2}{\mu_n V_{\mathrm{DD}}} \left(1 + \frac{\mu_n}{\mu_p}\right) \end{aligned}$$

Hieran erkennt man, daß die Festlegung der Verzögerungszeit zusammengeschalteter Gatter nur über L erfolgen kann. Um möglichst schnelle Schaltungen zu erhalten, muß L möglichst klein gehalten werden (heutzutage : $L \approx 0.90\nu m$). Hier wird die enorme Anstrengung verständlich, die Strukturgröße im Halbleiterprozeß zu verkleineren, da eine kleinere Kanallänge höhere Stromleitfähigkeit ($\approx \frac{1}{L}$) und höhere Geschwindigkeit bedeutet. Diese Vorteile stehen im Einklang mit geringeren Herstellkosten, da die notwendige Chipfläche auch kleiner wird.

Die Treiberleistung eines Gatters wird als **fan-out** f_o bezeichnet, wobei die gesamte Last als einzelner Inverter betrachtet wird. Wird an das Gatter eine andere Last C_{L1} angeschlossen, so ergibt sich die neue Verzögerungszeit als:

$$\begin{aligned} t_{\mathrm{ges}} &= f_o t_1 \text{ mit} \\ C_L &= f_o C_{L1} \end{aligned}$$

Werden mehrere Gatter an den Ausgang angeschlossen, so darf die Kapazität der Zuleitungen C_E nicht mehr vernachlässigt werden. Die Last stellt sich dann dar als (wobei wiederum die Gatekapazitäten aller Transistoren zusammengerechnet werden zu einem Inverter):

$$\begin{aligned} C_L &= C_E + f_o C_{L1} \\ &= C_E + f_o \left(W_n L_n + W_p L_p\right) C'_{\text{ox}} \end{aligned}$$

Zur Minimierung der Verzögerungszeit muß nun dasjenige Verhältnis $\frac{W_n}{W_p}$ gefunden werden, für das t_{ges} minimal wird.

$$\left.\frac{\partial t_{\text{ges}}}{\partial W_p}\right|_{W_n=\text{const}} = 0$$

ergibt nach längerer Rechnung:

$$W_n = \sqrt{\left(\frac{C_E}{2C'_{\text{ox}}L_n}\right)^2 + \frac{\mu_n}{\mu_p}W_p^2} - \frac{C_E}{2C'_{\text{ox}}L_n}$$

bzw.

$$\frac{W_p}{W_n} = \sqrt{\frac{\mu_n}{\mu_p} + \left(\frac{C_E}{2C_{\text{ox}}(p)}\right)^2} - \frac{C_E}{2C_{\text{ox}}(p)}$$

Dabei ist $C_{\text{ox}}(p) = C'_{\text{ox}}LW_p$ die Oxidkapazität des/der angeschlossenen p-Kanal-Transistoren.

Hier ist damit auch das Verhältnis der Zuleitungskapazität zur Gatekapazität von Bedeutung. Handelt es sich bei der Schaltung tatsächlich um zwei direkt hintereinandergeschaltete Inverter, so kann in der Regel die Zuleitungskapazität wieder vernachlässigt werden ($C_E = 0$) und man erhält:

$$\frac{W_p}{W_n} = \sqrt{\frac{\mu_n}{\mu_p}} \approx 1.7$$

In der Regel wird es sich jedoch um weiter auseinanderliegende Gatter handeln, so daß im Allgemeinen die Zuleitungskapazität nicht vernachlässigt werden darf.

Für die Bestimmung der Größe von C_E betrachtet man den Querschnitt und die Draufsicht auf die Anschlüsse eines Transistors (Abb. 5.11).

Entsprechend dieser Abbildung ergibt sich $C_l = W_l L_l C'_l$, wobei C'_l wegen der sehr viel größeren Dicke des Feldoxides wesentlich kleiner als C'_{ox} ist. Da die Gate- und Feldoxiddicke für jeden Prozeß fest vorgegeben sind, kann man ansetzen $C'_{\text{ox}} = KC'_l$.

Um die Leitungskapazität möglichst klein zu machen, kann W_l klein gehalten werden. Durch die Strukturgröße ist der minimale Wert von W_l gegeben als

$$W_l \approx L_n$$

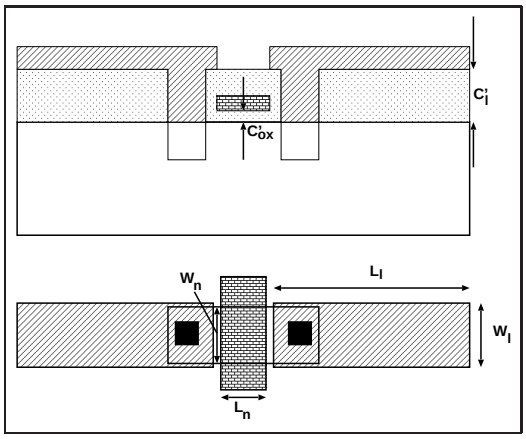


Abbildung 5.11: Zuleitung zu einem Transistor

Insgesamt erhält man für die Leitungskapazität damit

$$C_E \approx \frac{C'_{\text{ox}}}{K} L_n L_l$$

Eingesetzt ergibt das für die Dimensionierungsvorschrift:

$$\begin{aligned}
\frac{W_p}{W_n} &= \sqrt{\frac{\mu_n}{\mu_p} + \left(\frac{C_E}{2C'_{\text{ox}} L_n W_p}\right)^2} - \frac{C_E}{2C'_{\text{ox}} L_n W_p} \\
&= \sqrt{\frac{\mu_n}{\mu_p} + \left(\frac{C'_{\text{ox}} L_n L_l}{2KC'_{\text{ox}} L_n W_p}\right)^2} - \frac{C'_{\text{ox}} L_n L_l}{2KC'_{\text{ox}} L_n W_p} \\
&= \sqrt{\frac{\mu_n}{\mu_p} + \left(\frac{L_l}{2KW_p}\right)^2} - \frac{L_l}{2KW_p}
\end{aligned}$$

Es ist noch anzumerken, daß dank der immer feineren Strukturen die Chips komplexer werden und die Verdrahtung/Zuleitung überproportional komplexer werden. Die Delays, erzeugt durch Zuleitungskapazikäten und Zuleitungswiderständen, die mit kleineren Strukturen ansteigen, spielen eine immer gewichtigere Rolle, zumal die Transistoren immer schneller werden.

5.4 Verlustleistung

Ein grundsätzliches Problem besonders bei integrierten Schaltungen ist die Verlustleistung, die zur Temperaturerhöhung des Halbleiters führt. Eine solche Temperaturerhöhung muß in einem zulässigen Rahmen bleiben, damit die Bauelemente entsprechend der Spezifikation funktionieren. Eine starke Temperaturerhöhung kann dabei u.U. nicht nur zu Fehlfunktionen sondern auch zur Zerstörung des Halbleiters führen.

Verlustleistung entsteht überall dort, wo Ströme fließen. Dabei dürfen auch die Sperrströme von Dioden und Transistoren nicht vernachlässigt werden. Diese sind zwar für die logische Funktion der Schaltung meist vernachlässigbar klein, in ihrer Gesamtheit spielen sie jedoch wegen ihrer sehr großen Anzahl (mind. 2 parasitäre Dioden pro Transistor bei vielen tausend Transistoren) eine nicht zu vernachlässigende Rolle.

5.4.1 Diodensperrströme

Die Drain- und Sourcediffusionsgebiete der MOS-Transistoren ergeben zusammen mit dem jeweiligen Substrat zusammen immer eine in Sperrichtung gepolte Diode, durch die ein Sperrstrom fließt (Abb. 5.12).

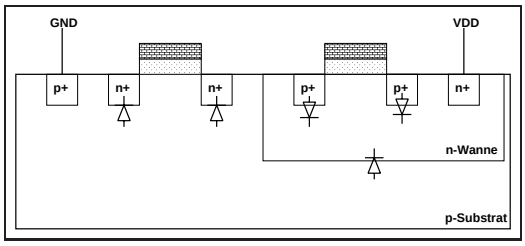


Abbildung 5.12: Parasitäre Dioden bei MOSFET

Durch die Dioden in Sperrichtung fließt ein **Leckstrom** (Sperrstrom der Diode) von:

$$I_L = I_S \left(e^{\frac{qU_D}{kT}} - 1 \right)$$

Dabei liegt I_S in der Größenordnung von $\approx 10^{-15}/\mu m^2$, d.h. im pA-Bereich. Bei heutigen Integrationsdichten können sich diese Leckströme zu vielen μA addieren, was für bestimmte Anwendungen (z.B. Medizintechnik: Implantate wie Herzschrittmacher, deren Batterien möglichst mehrere Jahre ausreichen sollen, aber auch andere Implantate, bei denen Temperaturerhöhungen gefährlich sind) nicht mehr tolerierbar ist.

Da diese Leckströme unabhängig vom Schaltzustand der Transistoren immer vorhanden sind, handelt es sich um eine **statische Verlustleistung**.

5.4.2 Ströme beim Umschalten

Während des Umschalten eines Ausgangs von 0 auf 1 (oder umgekehrt) sind für eine kurze Zeitspanne beide Transistoren leitend, so daß ein **Querstrom** von V_{DD} nach GND fließt. Solches geschieht bei jedem Umschalten jedes Gatters, so daß auch hier die Vielzahl der gleichzeitig schaltenden Gatter zu einer nicht vernachlässigbaren Verlustleistung führen. Da diese Verlustleistung nur beim Umschalten, d.h. dynamisch, auftritt, stellt sie einen Beitrag zur **dynamischen Verlustleistung** dar.

Wesentlich höher als die Querströme beim Umschalten sind die Ströme, die beim Umladen der Lastkapazitäten fließen. So wird bei jedem Umschalten des Ausgangs von „0“ auf „1“ die gesamte Lastkapazität C_L von $0V$ auf V_{DD} über (einen) p-Kanal-Transistor(en) aufgeladen. Umgekehrt wird beim Umschalten von „1“ auf „0“ die Lastkapazität über (einen) n-Kanal-Transistor(en) von V_{DD} auf $0V$ entladen. Geschieht dies im Mittel mit einer Periode t_p (d.h. mit einer mittleren Schaltfrequenz von $f_P = \frac{1}{t_p}$), so ergibt sich als Leistungsbilanz:

$$\begin{aligned} P &= \frac{\text{Gesamtenergie des Aufladens}}{\frac{t_p}{2}} \\ &= \frac{\frac{1}{2}C_L V_{\mathrm{DD}}^2}{\frac{t_p}{2}} \\ &= \frac{C_L V_{\mathrm{DD}}^2}{t_p} \\ &= C_L f_p V_{\mathrm{DD}}^2 \end{aligned}$$

Die gesamte dynamische Verlustleistung setzt sich dann aus dem (kleineren) Anteil der Querströme und dem (größeren) Anteil der Umladeströme zusammen.

Insgesamt ergibt sich damit eine Verlustleistung von:

$$P_{\text{gesamt}} = P_{\text{Umladen}} + P_{\text{Querströme}} + P_{\text{Leckströme}}$$

Dabei stellt nur der letzte Anteil einen statischen dar.

Aufgrund der Beziehung $t_p P_{\text{dyn}} \approx C_L^2 V_{\mathrm{DD}}^2$ wird das Produkt $t_p P_{\text{dyn}}$ (das power-delay-Produkt) als Gütemaß für den verwendeten Prozeß gebraucht. Je kleiner sich dieses Produkt ergibt, umso besser ist der Prozeß.

6 Layout

6.1 Standardzellenlayout

Die von den Chipherstellern zur Verfügung gestellten Standardzellbibliotheken enthalten Zellen, die besonders auf ihre Plazierbarkeit hin optimiert sind. Die Ein- und Ausgänge sowie die Versorgungsleitungen sind derart angeordnet, daß solche Zellen in einfacher Weise in Zeilen angeordnet werden können. Man nennt sie auch Standardzellen. Der verbleibende Raum zwischen den Zeilen kann dann für die Verdrahtung der Zellen genutzt werden.

Das Aussehen solcher Zellen soll hier kurz an zwei Beispielen erläutert werden, wobei noch einmal die Korrespondenzen zwischen Transistorschaltung, Layout und realisierter Schaltung dargestellt werden.

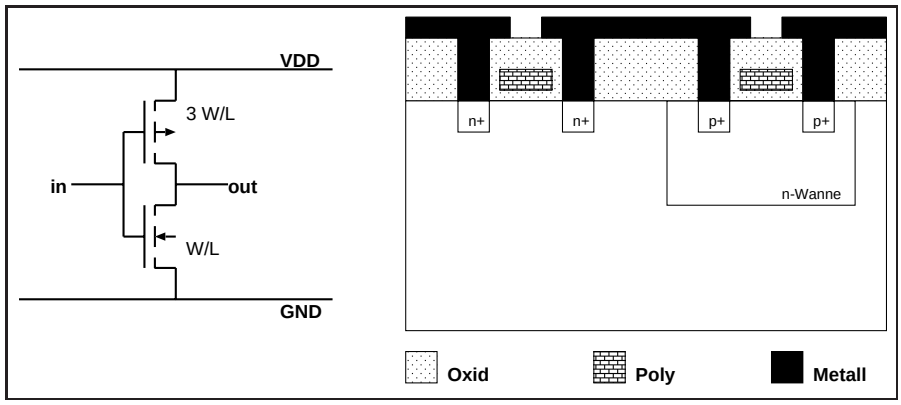


Abbildung 6.1: CMOS-Inverter in Schematic-Darstellung und als Querschnitt

Zwei mögliche Layouts dieses Inverters sind in der folgenden Abbildung dargestellt. Der Übersichtlichkeit halber ist dort der p-Transistor mit einem W/L-Verhältnis von $2:1$ gegenüber dem n-Transistor dargestellt.

Beide Layouts sind jedoch nicht sehr gut für eine Standardzellplazierung geeignet.

Gute Standardzellayouts für ein NOR-Gatter und einen Inverter ist in Abbildung 6.3 zu sehen. Dort ist auch die Verdrahtung der beiden Gatter zur in der Abbildung dargestellten Schaltung dargestellt. An diesem Layout kann man auch erkennen, daß alle Standardzellen die gleiche Höhe besitzen, so daß die Versorgungsleitungen und die n-Wannen für die PMOS Transistoren über eine Zeile einfach „durchgeschleift“ werden können. Die Breite der Zellen kann natürlich unterschiedlich sein, sie hängt im Wesentlichen von der Anzahl der Transistoren der Zelle ab.

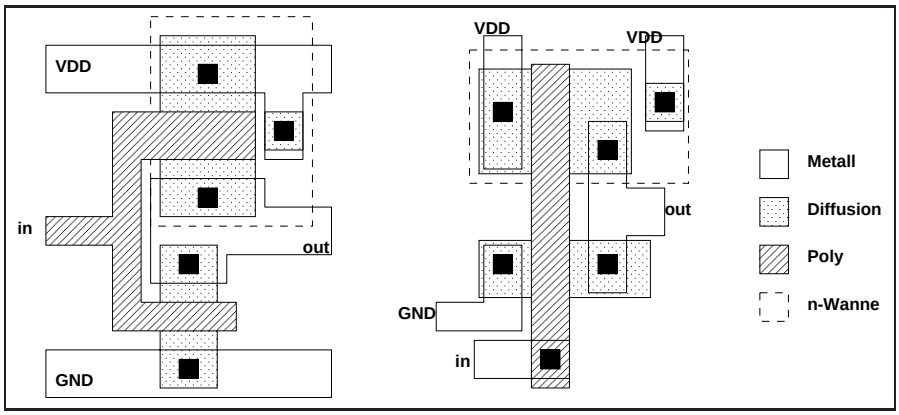


Abbildung 6.2: Mögliche Layouts des Inverters

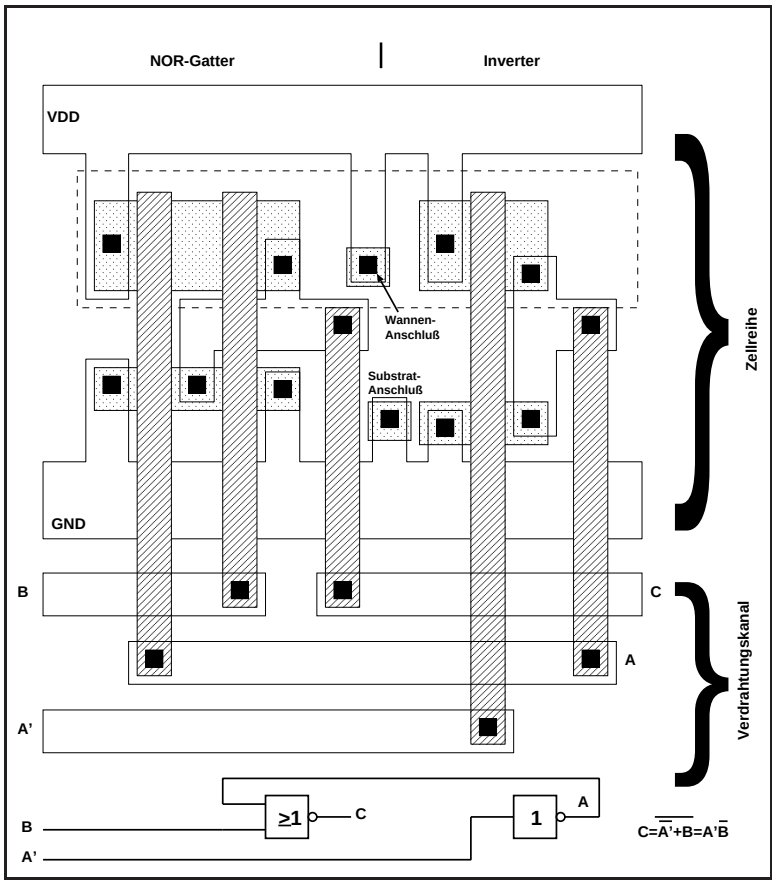


Abbildung 6.3: Standardzellayout (Inverter und NOR)

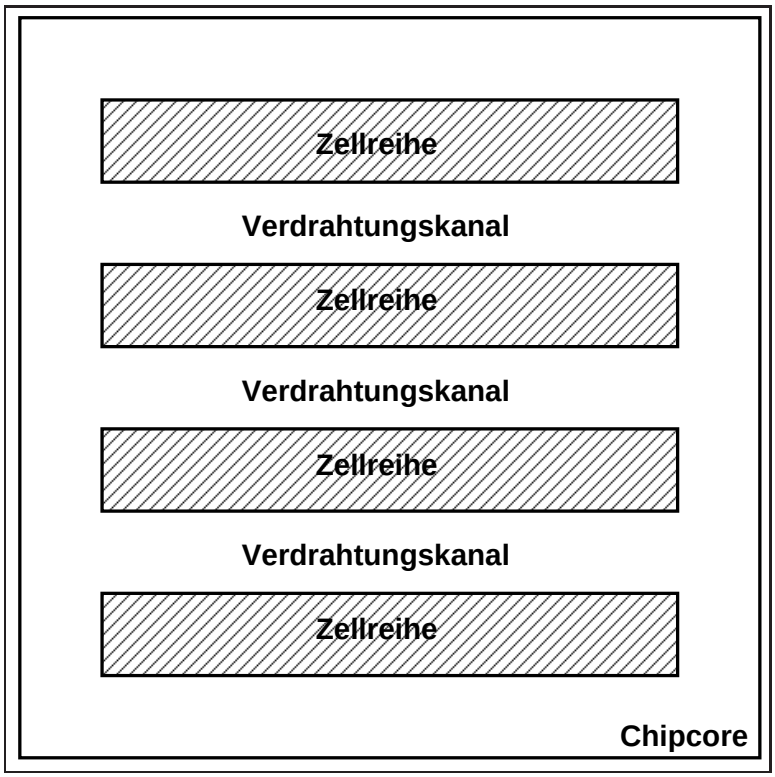


Abbildung 6.4: Gesamtansicht eines Standardzellenlayouts (ohne Verdrahtung)

6.2 Designregeln („design rules“)

Aus dem Herstellungsprozeß, insbesondere aus dessen Abbildungs- und Herstellungstoleranzen, resultieren Regeln für das Maskenlayout, die die Übereinstimmung von entworfener Funktion und hergestellter Schaltung garantieren sollen. Diese geometrische Layout-Regel werden design rules genannt.

Bei den Kontaktlöchern muß darauf geachtet werden, daß nur die hochdotierten Source- bzw. Draingebiete mit Metall in Berührung kommt, nicht aber das Substrat selbst oder gar der Kanal eines Transistors. Damit dies immer gewährleistet ist, muß berücksichtigt werden, daß die unterschiedlichem Masken für die Kontaktlöcher und Diffusionsgebiete nicht beliebig genau gefertigt und auch nicht beliebig genau „übereinandergelegt“ werden können. Abhilfe schaffen hier Mindestüberlappungen von Diffusion bzw. Polysilizium gegenüber den Kontaktlöchern, die man auch in Abbildung 6.3 erkennen kann.

Polysilizium darf nicht zu dicht an Source- und Draingkontakte herangeführt werden, da sonst durch Toleranzen der Masken Kurzschlüsse zwischen Gate und Sourcekontakt bzw. Gate und Drainkontakt entstehen können. Die entsprechende Designregel verlangt darum einen Mindestabstand der Polysiliziummaske zu Source- und Drainkontakten (siehe Abb. 6.6 unten links).

Neben diesen (und ähnlichen) Regeln, die aus den Herstellungstoleranzen und der Justierungenauigkeit der Masken resultieren gibt es noch **elektrisch begründete** Designregeln. So spielt bei den Diffusionsgebieten die Ausdehnung der Raumladungszone eine Rolle: diese dürfen sich auch bei hohen anliegenden (Sperr-)Spannungen zwischen zwei Diffusionsgebieten nicht soweit ausdehnen, daß sie einander berühren können (was einen punch through der beiden Diffusionsgebiete und damit im schlimmsten Fall zur Zerstörung des Halbleiters führen kann). Hieraus resultiert eine Designregel für den Mindestabstand zwischen Diffusionsgebieten. Aufgrund unterschiedlicher Dotierung kann dieser für n^+ und p^+ unterschiedlich sein. Die Kanallänge darf eine bestimmte Größe nicht unterschreiten, weil sonst durch Erreichen der kritischen Feldstärke der Durchschlag durch den Kanal droht.

In den Abbildungen 6.5 und 6.6 sind die Designregeln eines Beispielprozesses dargestellt.

Alle diese Designregeln müssen beim Entwurf eingehalten werden. In CAD-Entwurfssystemen bieten die entsprechenden Entwurfswerkzeuge die Möglichkeit, eine entworfene Schaltung auf Einhaltung dieser Regeln zu überprüfen und entsprechende Verletzungen anzuzeigen. Die dabei verwendeten Verfahren werden im Rahmen der Vorlesung Mikroelektronik II vorgestellt.

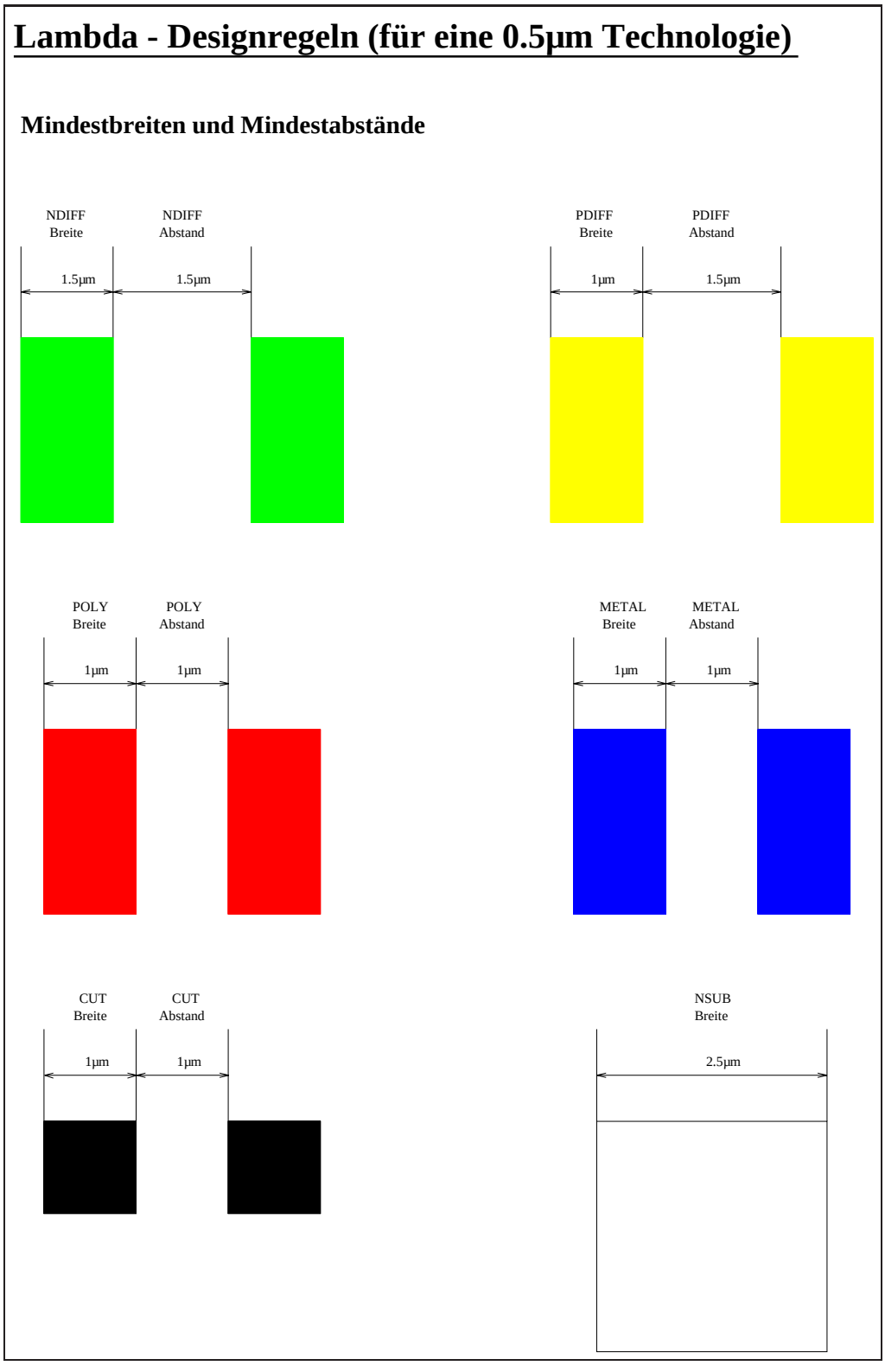
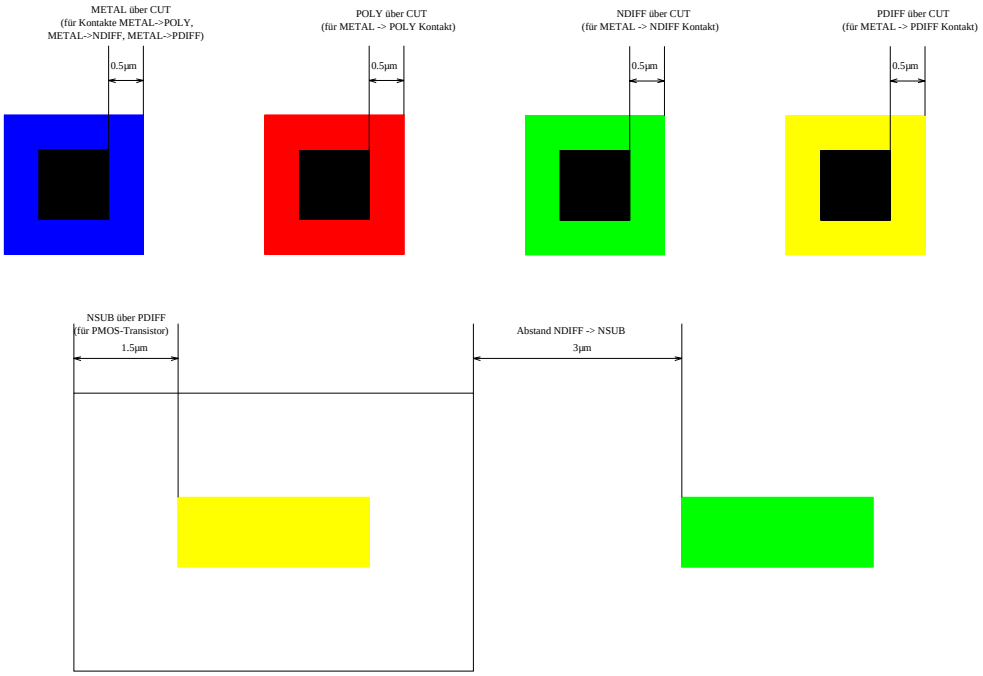
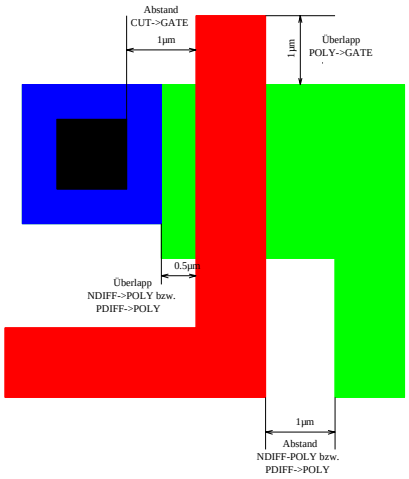


Abbildung 6.5: Abstände und Breiten

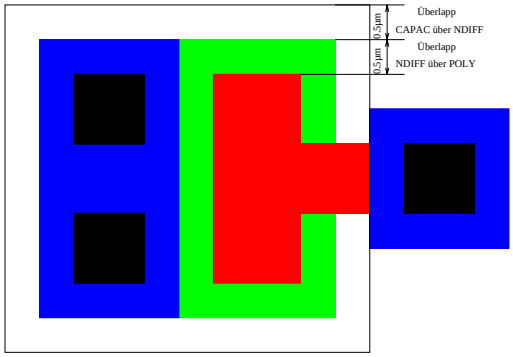
Mindestüberlappungen



Transistor-Regeln



Kondensator-Regeln



Technologie-Parameter

Lambda (L):	0.5 um
Schwellenspannung NMOS (minimale Abmessung):	+ 1 V
Schwellenspannung PMOS (minimale Abmessung):	- 1 V
Versorgungsspannung:	3-10 V
Square-/Schicht-Widerstand (NSUB):	1 kOhm/□
Square-/Schicht-Widerstand (POLY):	20 Ohm/□
Square-/Schicht-Widerstand (NDIFF):	30 Ohm/□
Square-/Schicht-Widerstand (PDIFF):	100 Ohm/□
Kapazitätsbelag (NDIFF über POLY über CAPAC):	0.5 fF/(um)^2
Kapazitätsbelag (GATE):	1 fF/(um)^2

Abbildung 6.6: Überlappungen

6.3 Skalierung

Die Verbesserung der Geräte für das photolithographische Herstellungsverfahren hat in den letzten Jahren dazu geführt, daß die Strukturen auf den Chips viel exakter und damit wesentlich kleiner hergestellt werden können, so daß die eben erwähnten, von der Photolithographie herrührenden Entwurfsregeln die Realisierung kleinerer Transistoren ermöglichen. Durch Änderung der Dotierung konnten auch die elektrisch verursachten Regeln soweit verändert werden, daß auch sie kleinere Strukturen erlauben. Das Herabsetzen der Versorgungsspannung (Feldstärke über dem Gateoxid $\approx \frac{5V}{40nm}$!) ist hierbei eine zusätzliche Methode, die Strukturen weiter zu verkleinern, um noch mehr Transistoren auf der gleichen Chipfläche realisieren zu können.

Alle geometrischen Abmessungen werden dabei um einen Faktor α verkleinert. Es ist leicht einzusehen, daß dann Dotierungen und elektrische Eigenschaften ebenfalls geändert werden müssen.

Bei einer solchen Skalierung hat man die Wahl, die Versorgungsspannung oder die Feldstärke für die neue Technologie beizubehalten (wenn man von der Möglichkeit, beides zu ändern absieht). Tabelle 6.1 gibt einen Überblick über die Änderungen physikalischer Größen, die daraus resultieren.

Skalierung bei		**konstanter Feldstärke**	**konstanter Spannung**
Kanallänge	L	$1/\alpha$	$1/\alpha$
Kanalweite	W	$1/\alpha$	$1/\alpha$
Oxiddicke	t_{ox}	$1/\alpha$	$1/\alpha$
Versorgungsspannung	V_{DD}	$1/\alpha$	1
Stromstärke	I	$1/\alpha$	α
Gatekapazität	$C_g = C_{\mathrm{ox}}\frac{W}{L}$	$1/\alpha$	$1/\alpha$
Verzögerungszeit	$t_{f/r} \sim \frac{C_L}{\beta V_{\mathrm{DD}}}$	C_{L}	C_{L}/α
dyn. Verlustleistung	$P_d = C_L f V_{\mathrm{DD}}^2$	$C_{\mathrm{L}}^2/\alpha^2$	C_{L}^2/α
Stromdichte	$J = \frac{I}{A}$	α	α^3
Leistungsdichte	$\frac{P}{A}$	C_{L}^2	$C_{\mathrm{L}}^2\alpha^2$

Tabelle 6.1: Gegenüberstellung der Skalierungsarten

Damit das der Chip, der ohnehin sehr heiß betrieben wird, nicht zu heiß wird, ist es sinnvoll, die Feldstärke konstant zu halten. Die Leistungsdichte bleibt dann annähernd konstant. Das heißt auch, daß die Versorgungsspannung in modernen Prozeßen mit immer kleineren Strukturgrößen sinkt.

7 Gatter in CMOS-Technologie

Die prinzipielle Funktionsweise und der logische Entwurf von CMOS-Gattern soll hier an einigen Beispielen erläutert werden. Betrachten wir zuerst das NAND-Gatter nach Abbildung 7.1:

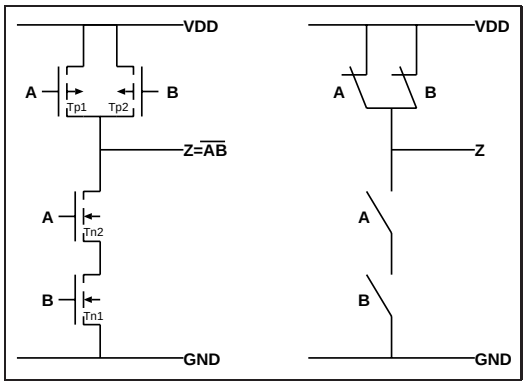


Abbildung 7.1: NAND-Gatter in CMOS-Realisierung und entsprechendes Schaltermodell

Berücksichtigt man nur die stabilen Signalzustände „0“ (high) und „1“ (low) an den Eingängen, so verhalten sich die einzelnen Transistoren der Schaltung entsprechend Tabelle 7.1. Man erkennt, daß — unabhängig von der Kombination der Eingangswerte — niemals ein leitender Pfad von VDD nach Masse (GND) existieren kann. Dies bedeutet auch, daß unabhängig vom Wert des Ausgangs Z niemals ein statischer Strom fließt. Dies ist eine Aussage, die auf alle anderen CMOS-Gatter ebenfalls zutrifft.

Betrachtet man vom Ausgang Z aus die möglichen Wege zu VDD bzw. GND, so ist zu erkennen, daß über den n-Teil der Schaltung nur das Einstellen des Wertes Z=0, über den p-Teil der Schaltung nur das Einstellen des Wertes Z=1 möglich ist. Dies ist ebenfalls eine Aussage, die allgemein für CMOS-Gatter gültig ist.

Speziell bei dieser Schaltung sieht man auch, daß für den Wert Z=0 als einziger Weg die Möglichkeit „Tn1 und Tn2 leitend“, d.h. A=1 und B=1, besteht. Betrachtet man den entsprechenden

	Tn1	Tn2	Tp1	Tp2
A=1		leitet	sperrt	
A=0		sperrt	leitet	
B=1	leitet			sperrt
B=0	sperrt			leitet

Tabelle 7.1: Zustände der Transistoren beim NAND-Gatter

Schaltungsteil im Schaltermodell, so wird dort mit den Schaltern die Funktion $A \cdot B$ realisiert. Da eine Verbindung von Z mit GND dem logischen Wert 0 enstpricht, kann man auch schreiben:

$$\begin{aligned} Z_N &= A \cdot B \\ Z_N &= \overline{Z} \end{aligned}$$

Ein Weg von Z nach VDD existiert für die Möglichkeiten „Tp1 oder Tp2 (oder beide) leitend“, was A=0 oder B=0 entspricht. Man kann deshalb auch sagen $\overline{A} + \overline{B}$. Dies kann aus dem Schaltermodell ebenfalls einfach abgelesen werden. Da eine Verbindung von Z mit VDD dem logischen Wert 1 entspricht schreiben wir auch:

$$\begin{aligned} Z_P &= \overline{A} + \overline{B} = \overline{A \cdot B} \\ Z_P &= Z \end{aligned}$$

In diesem Modell entspricht also ein p-Kanal-Transistor einem Schalter, der mit „0“ geschlossen und mit „1“ geöffnet, ein n-Kanal-Transistor dagegen einem Schalter, der mit „1“ geschlossen und mit „0“ geöffnet wird. Jedes Eingangssignal ist an jeweils einen p-Kanal-Transistor und einen n-Kanal-Transistor angeschlossen. CMOS hat damit allgemein einen Aufwand von zwei Transistoren pro Eingang eines Gatters.

Dadurch, daß über den n-Teil nur ein Festlegen des Ausgangs auf GND und damit auf den logischen Wert 0 möglich ist, ist leicht ersichtlich, daß nur logische Funktionen darstellbar sind, die als äußersten Booleschen Operator eine Negation enthalten. Ein AND-Gatter z.B. kann in dieser Weise nur durch ein NAND mit nachgeschaltetem Inverter oder durch ein NOR mit jeweils einem Inverter vor jedem Eingang realisiert werden. Dies bedeutet jedoch einen vergleichsweise hohen Aufwand für eine so elementare Funktion.

Entsprechend schwierig wird es sich gestalten, komplexe Funktionen (d.h. solche mit vielen Eingängen und/oder vielen Operatoren wie $+, \cdot, \overline{\ }, \oplus, \ldots$) direkt als ein einziges Gatter zu realisieren. Außerdem wird dies in den meisten Fällen auch zu einem sehr hohen Aufwand (Anzahl der Transistoren und Verbindungsleitungen) führen. Es ist deshalb sinnvoll, solche komplexen Funktionen durch Zusammenschalten kleinerer, einfacher Gatter zu realisieren. Dies ist in Beispiel 7.1 verdeutlicht.

Beispiel 7.1

8-fach AND

$$
\begin{aligned}
Z = \quad & ABCDEFGH & \\
= \quad & \overline{\overline{ABCDEFGH}} & \text{8-fach NAND} + \text{Inverter (18 Transistoren)} \\
= \quad & \overline{\overline{ABCD} + \overline{EFGH}} & 2 \times \text{4-fach NAND} + \text{NOR (20 Transistoren)} \\
= \quad & \overline{\overline{AB} + \overline{CD} + \overline{EF} + \overline{GH}} & 4 \times \text{NAND} + \text{4-fach NOR (24 Transistoren)} \\
= \quad & \overline{\overline{(\overline{AB} + \overline{CD})} \cdot \overline{(\overline{EF} + \overline{GH})}} & \text{NAND} + \text{Inverter} + 2 \times \text{NOR} + 4 \times \text{NAND (30 Transistoren)}
\end{aligned}
$$

Vom Flächenbedarf her ist die erste Realisierung die günstigste, leider aber sehr langsam. Die letze Realisierung benötigt am meisten Fläche, ist aber bei Vernachlässigung der Leitungskapazitäten die schnellste. In der Realität spielen jedoch auch die Leitungskapazitäten (und ihr Flächenbedarf) eine Rolle, so daß die beiden mittleren Realisierungen sowohl schneller als auch kleiner hergestellt werden können.

Mit den eben erwähnten Methoden lassen sich — neben den elementaren Gattern NAND, NOR und Inverter — auch einige komplexere Funktionen direkt als Gatter effizient realisieren; diese werden auch als **Komplexgatter** bezeichnet. Beispiel 7.2 zeigt unter anderem die Realisierung eines solchen Komplexgatters auf Transistorebene.

Beispiel 7.2

Realisierung der Funktion $Z = \overline{AB + CD}$.

Direkt zu erkennen aus der Funktionsgleichung ist eine Realisierung mit zwei AND (d.h. NAND mit nachgeschaltetem Inverter) und einem NOR. Diese Schaltung benötigt insgesamt 16 Transistoren.

Beachtet man dagegen die Eigenschaft von CMOS, daß invertierte Funktionen direkt realisierbar sind, so kann man erkennen, daß

$$Z_N = AB + CD.$$

Damit läßt sich der n-Teil des Gatters entsprechend Abbildung 7.2 bestimmen.

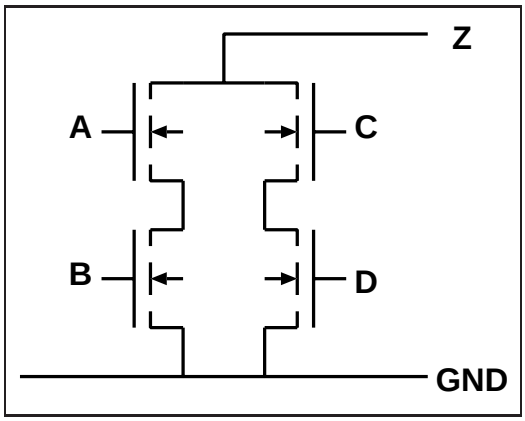


Abbildung 7.2: n-Teil des Komplexgatters

Für den p-Teil muß die Funktionsgleichung nach den deMorganschen Regeln so umgeformt werden, daß die Eingänge nur in invertierter Form benutzt werden:

$$Z = \overline{AB + CD} = \overline{AB} \cdot \overline{CD} = (\overline{A} + \overline{B}) \cdot (\overline{C} + \overline{D}) = Z_P$$

Dies entspricht der Schaltung in Abbildung 7.3.

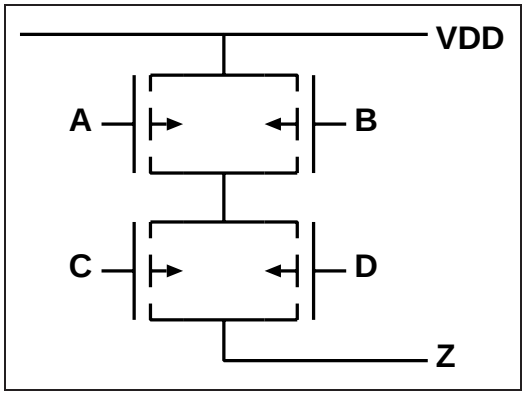


Abbildung 7.3: p-Teil des Komplexgatters

Insgesamt ergibt sich damit für das Komplexgatter die Schaltung nach Abbildung 7.4. Da es sich hier um die Realisierung in einem einzigen Gatter handelt, existiert auch ein Symbol dafür (Abbildung 7.5), sowie ein eigener Name: AOI-Gatter (von AND, OR, INVERT).

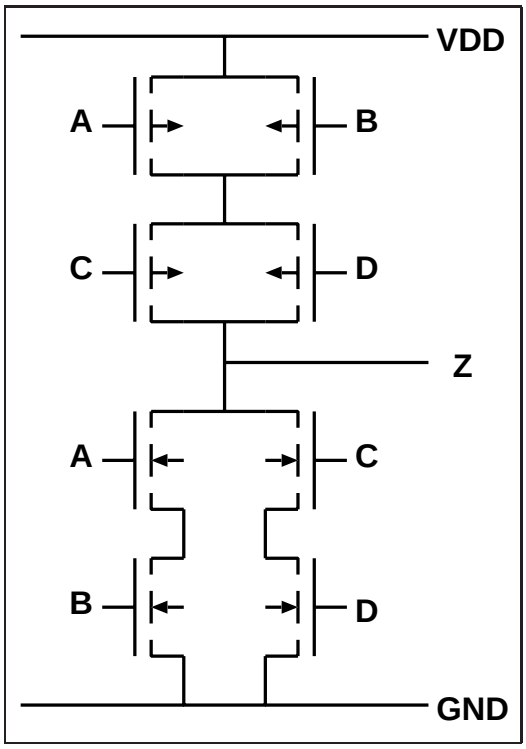


Abbildung 7.4: Komplexgatter AOI auf Transistorebene

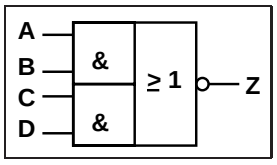


Abbildung 7.5: Symbol des Komplexgatters $Z = \overline{AB + CD}$ (AOI)

Zusammenfassung

CMOS-Gatter sind zusammengesetzt aus einem n-Teil, d.h. einem Teil, der nur aus n-Kanal-Transistoren besteht, und einem p-Teil, der nur aus p-Kanal-Transistoren besteht. Der n-Teil realisiert die Funktion Z_N, wobei $Z_N = \overline{Z}$ und nur nichtinvertierte Eingänge verwendet werden, da n-Kanal-Transistoren bei hohem Gatepotential leitend werden. Der p-Teil realisiert die Funktion $Z_P = Z$, wobei nur invertierte Eingangssignale verwendet werden, da p-Kanal-Transistoren bei niedrigem Gatepotential leitend werden.

Besonders zu beachten ist:

- möglichst wenige Transistoren sollten in Reihe geschaltet werden, da dies die Verzögerungszeit oder die Fläche vergrößert.
- Bei Zusammensetzung komplexer Funktionen aus Gattern sollten vorzugsweise NAND-Gatter verwendet werden, da sie kleiner bzw. schneller sind als NOR-Gatter ($\frac{\mu_n}{\mu_p} \approx 3$!).
- Ein Gatter besitzt für jeden Eingang genau 2 Transistoren, nämlich einen p-Kanal-Transistor und einen n-Kanal-Transistor.
- CMOS-Schaltungen nehmen keinen statischen Strom auf, da entweder der n-Teil oder der p-Teil nichtleitend ist.

8 Spezielle CMOS-Schaltungstechniken (Logik-Gatter)

8.1 Übertragungsgatter — „transmission gates“

Eben wurden die Transistoren als Schalter modelliert, mit denen eine Verbindung vom Ausgang eines Gatters zu einem der Versorgungspotentiale geschaltet werden konnte. Natürlich kann man einen solchen Schalttransistor auch verwenden, um zwei beliebigen Punkte einer Schaltung miteinander zu verbinden. Ein so verwendeter Transistor heißt auch **Übertragunsgatter** oder **transmission gate**.

Da MOS-Transistoren symmetrisch hergestellt werden, ist es eine reine Interpretationsfrage, wo Source und Drain liegen; diese beiden Anschlüsse sind beliebig austauschbar. Per Definition liegt Drain von p-Kanal-Transistoren immer am niedrigeren, von n-Transistoren immer am höheren Potential. Für die Untersuchung von transmission gates ist nur von Bedeutung, wie groß der Widerstand des Transistors im leitenden Zustand ist, d.h. wie ideal der Schalter leitet.

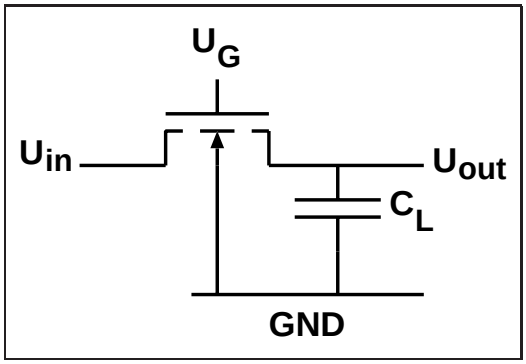


Abbildung 8.1: n-Kanal Übertragungsgatter (transmission gate)

Für das n-Kanal transmission gate in Abbildung 8.1 gelten folgende Überlegungen für $U_{GS} = V_{DD}$ (bei $U_{GS} = 0$ ist der Transistor gesperrt, was einem geöffneten Schalter entspricht):

- Wenn $U_{in} = U_{out}$, dann ist $U_{DS} = 0$, damit $I_D = 0$.

- Wenn $U_{in} > U_{out}$, dann liegt (interpretationsweise) Drain an U_{in}, Source an U_{out}. Es fließt ein Strom von Drain zur Source, solange der Transistor leitend ist, d.h. solange $U_{GS} > U_T$ bleibt. Da durch den Stromfluß das Ausgangspotential (Sourcepotential) ansteigt, sinkt $U_{GS} = U_G - U_{out}$. Ist das Ausgangspotential soweit angestiegen, daß $U_{GS} = U_T$ erreicht wird, hört der Stromfluß auf. Jetzt gilt:

  $$U_{in} = U_T + U_{out} \text{ bzw. } U_{out} = U_{in} - U_T$$

  U_{out} kann damit maximal $V_{DD} - U_T$ werden. Nun ist aber der Bulkanschluß auf Masse gelegt, d.h. nicht mit Source verbunden, so daß

  $$U_T = U_T(0) + \gamma\left(\sqrt{U_{SB} + 2\Phi_S} - \sqrt{2\Phi_S}\right).$$

  Hier gilt $U_{SB} = U_{out}$ und damit

  $$U_T \approx U_T(0) + \gamma\sqrt{U_{out}}$$

  Bei üblichen Werten — $\gamma = 0.8, V_{DD} = 5V$ — ergibt sich als maximales U_{out} ein Wert von lediglich $2.7V$. Damit ist klar, daß mit einem solchen n-Kanal transmission gate nur der untere Spannungsbereich geschaltet werden kann.

Für ein p-Kanal transmission gate nach Abbildung 8.2 gilt entsprechend, daß nur der obere Spannungsbereich geschaltet werden kann, da U_{out} minimal den Wert U_T erreichen kann.

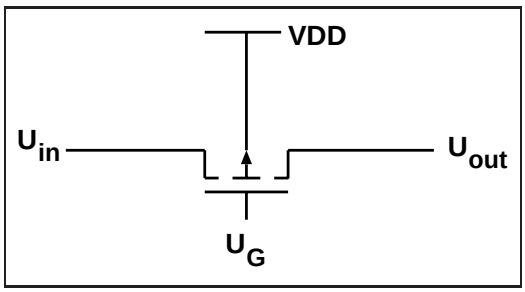


Abbildung 8.2: p-Kanal Übertragungsgatter (transmission gate)

Bei Parallelschaltung eines n-Kanal und eines p-Kanal transmission gates 8.3 ergänzen sich diese beiden so, daß das n-Kanal transmission gate den unteren, das p-Kanal transmission gate den oberen und beide zusammen den mittleren Spannungsbereich gut schalten. Nur im mittleren Spannungsbereich sind beide Transistoren leitend, in den beiden anderen leitet jeweils nur einer (Abb. 8.4). Das so erhaltene Gatter heißt auch **CMOS-Übertragungsgatter**. Es ist anzumerken, daß die beiden Gates invertiert angesteuert werden.

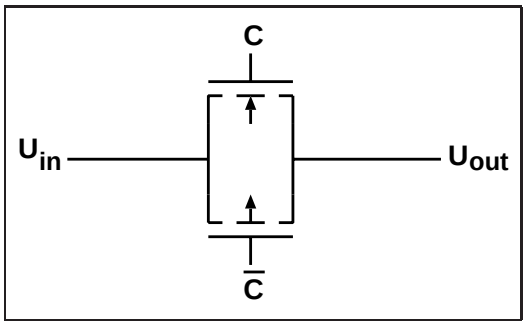


Abbildung 8.3: CMOS-Übertragungsgatter

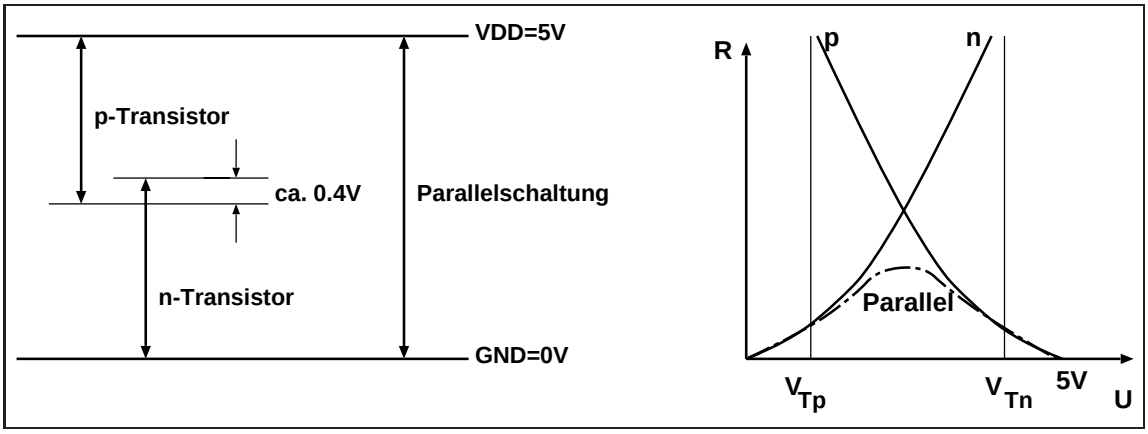


Abbildung 8.4: Spannungsbereiche und Widerstand des CMOS transmission gates

Das Verhalten des CMOS-Übertragungsgatters ist in Tabelle 8.1 zusammengefaßt.

C	T_N	T_P	out
0	sperrt	sperrt	hochohmig
1	leitet	leitet	out = in

Tabelle 8.1: Verhalten des CMOS-Übertragungsgatters

Mit solchen Übertragungsgattern lassen sich sehr einfach sogenannte **Multiplexer** realisieren. Dies sind Schaltungen, die aus mehreren Eingängen — je nach Steuersignalen — eine gewisse geringere Anzahl auf die Ausgänge durchschalten (siehe Abb. 8.5, Multiplexer mit zwei Eingängen und einem Ausgang). Der dort abgebildete Multiplexer benötigt nur vier Transistoren.

Als Komplexgatter würde diese Funktion mehr Fläche benötigen, da sie als AOI mit nachgeschaltetem Inverter realisiert werden müßte, d.h. aus 10 Transistoren bestünde.

Die CMOS transmission gates werden so häufig verwendet, daß man für sie ein eigenes Symbol eingeführt hat (das Symbol ist aus zwei ineinandergeschobenen Invertern entstanden):

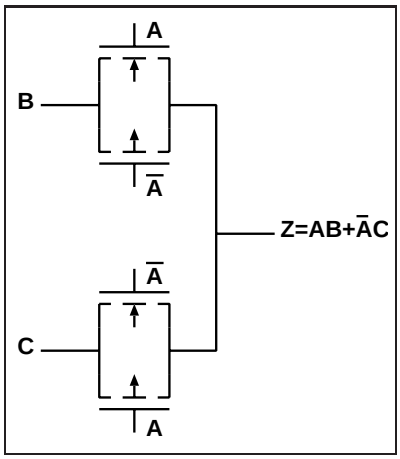


Abbildung 8.5: Multiplexer aus CMOS-Übertragungsgattern

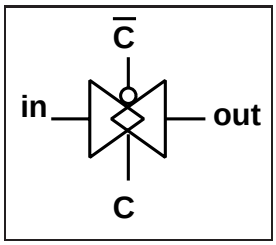


Abbildung 8.6: Symbol eines CMOS Übertragungsgatters

Die in Abbildung 8.5 dargestellte Schaltung kann man bei etwas abweichender Beschaltung zur Realisierung der Funktion $Z = AB + \overline{A}\,\overline{B}$ verwenden, der sogenannten **XNOR**-Funktion. Dazu muß nur statt C das Signal $\overline{B}$ an den Eingang des unteren Übertragunsgatters gelegt werden.

Bei Übertragungsgattern muß immer berücksichtigt werden, daß sie keine aktiven Bauelemente darstellen (sie benutzen keine Versorgungsspannungen, um das Ausgangssignal zu regenerieren) und dadurch den Rauschabstand des Ausgangs gegenüber dem Eingang verschlechtern. Diese Verschlechterung rührt — besonders bei mehreren hintereinandergeschalteten transmission gates — vom nicht vernachlässigbaren Widerstand des Kanals, d.h. vom Spannungsabfall über dem Kanal, her. Die Treibfähigkeit ist folglich schwach und die Schaltgeschwindigkeit gering. Außerdem besitzen auch die parasitären Kapazitäten (Diffusionskapazitäten der Transistoren) einen negativen Einfluß.

8.2 Pseudo-NMOS

Die bisher betrachteten CMOS-Schaltungen (abgesehen von Übertragungsgattern) gehören zur Familie der **Standard CMOS**-Schaltungen. Ihnen ist gemeinsam, daß sie als Lastelement ein komplettes (schalter-)logisches Netzwerk aus p-Kanal-Transistoren enthalten. Gegenüber NMOS hat

dies den Nachteil, daß diese Last aus wesentlich mehr Transistoren besteht, die wegen der geringeren Löcherbeweglichkeit auch noch größer dimensioniert werden müssen.

In komplementärer MOS-Technologie ist man jedoch nicht gezwungen, die Last aus mehreren Transistoren zu realisieren. Man kann sich damit begnügen, den p-Teil der Schaltung durch einen einzelnen p-Kanal-Transistor zu ersetzen, dessen Gate auf GND liegt. Diese Schaltungsart heißt dann pseudo-NMOS bzw. Entsprechende Gatter sind in Abbildung 8.7 dargestellt (siehe auch Kapitel 5.2). Die logische Funktion wird vom n-Teil realisiert. Wenn der Ausgang eins sein soll, ist der n-Teil sperrend. Der Ausgang wird vom pmos hochgezogen.

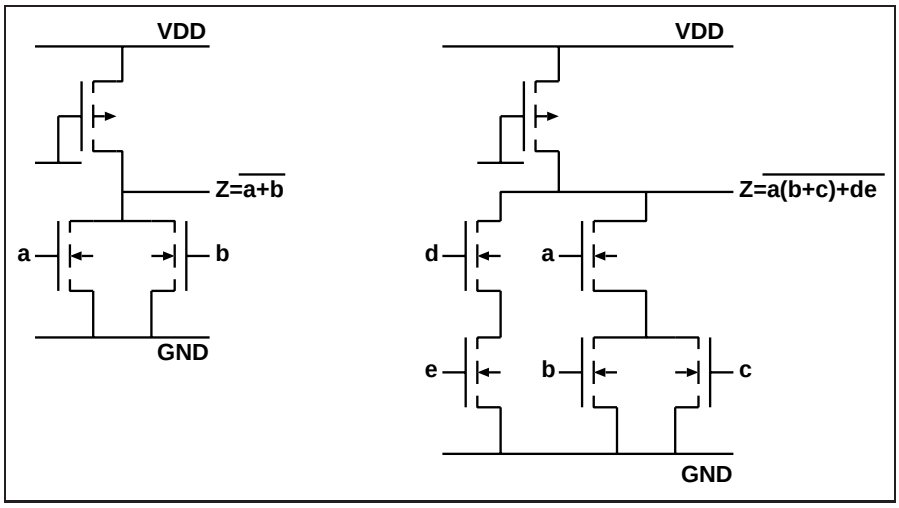


Abbildung 8.7: pseudo-NMOS-Schaltungen (NOR und Komplexgatter $Z = \overline{a(b+c)+de}$.)

Durch die geringere Lastkapazität (Last = Eingang eines anderen Gatters) sind diese Schaltungen schneller als Standard-CMOS-Gatter. Hierdurch sinkt auch die dynamische Verlustleistung ($P_d \sim C_\mathrm{L}$). Diese Vorteile werden jedoch durch den Nachteil einer **statischen Verlustleistung** erkauft, da bei $Z = 0$ mindestens ein leitender Pfad von V_DD nach GND existiert (der p-Kanal-Lasttransistor ist immer leitend!). Dieser leitende Pfad führt auch zu einer schlechteren Signalgüte für $Z = 0$, da hier die Ausgangsspannung vom **Spannungsteiler** des p-Kanal-Transistors und des leitenden n-Teils festgelegt wird (Abb. 8.8). Die Größe der Ausgangsspannung hängt in diesem Fall von den $\frac{W}{L}$-Verhältnissen der entsprechenden Transistoren, insbesondere des p-Transistors ab.

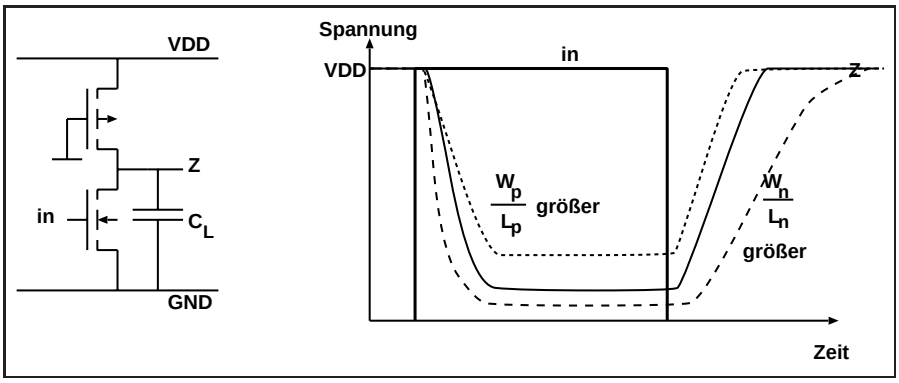


Abbildung 8.8: Übertragungskennlinie des pseudo-NMOS Inverters

Leider kann durch die Dimensionierung die Güte der Ausgangsspannung bei $Z = 0$ nicht beliebig verbessert werden, da gleichzeitig die Verzögerungszeit nachteilig beeinflußt wird. Aus Sicherheitsgründen (Rauschabstand) muß die Ausgangsspannung jedoch innerhalb der vorgesehenen Verzögerung unterhalb U_{Tn} (in der Regel sogar unterhalb $U_{\mathrm{Tn}}/2$) fallen, damit die nächste Stufe noch korrekt angesteuert wird. Diesen Effekt, daß die Ausgangsspannung von den $\frac{W}{L}$-Verhältnissen der Transistoren abhängt, bezeichnet man auch als **Verhältnislogik**.

Später werden wir noch sehen, daß diese Schaltungsart (besonders NOR-Gatter) häufig für die Realisierung von Speichern (genauer: der Adressdekodierer in Speichern) verwendet wird.

9 Zusammenfassung: CMOS-Schaltungsarten für kombinatorische Schaltungen

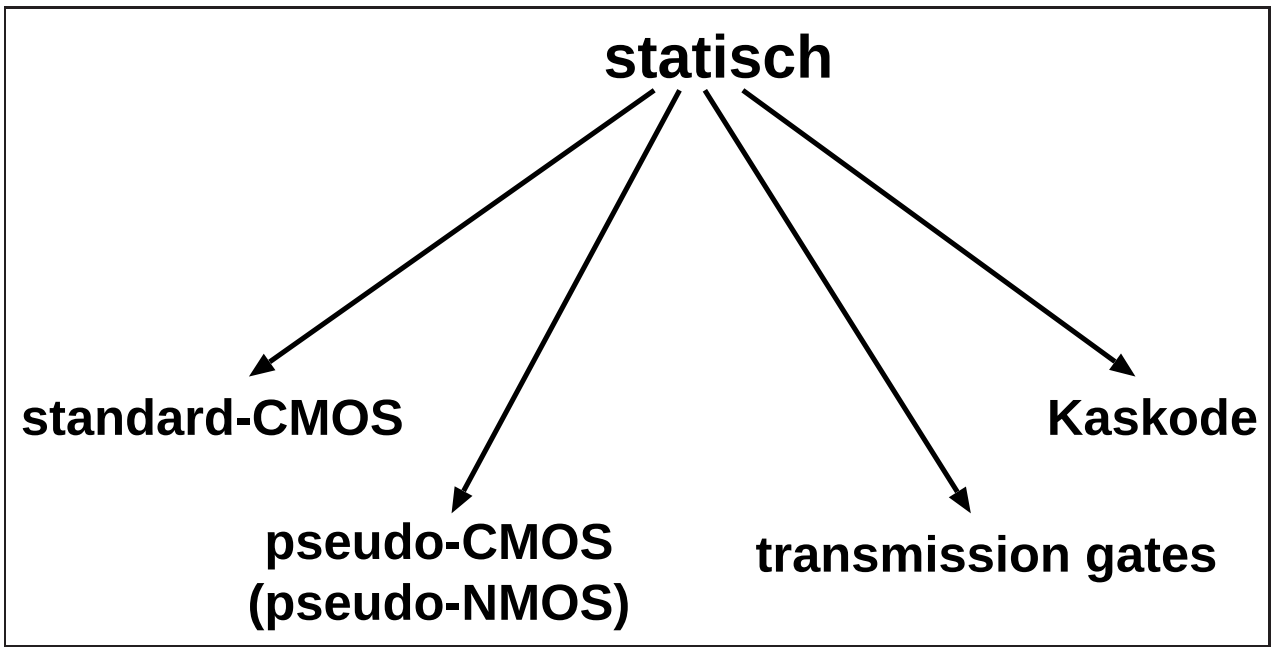


Abbildung 9.1: Übersicht über kombinatorische CMOS-Schaltungsarten

9.1 Vor- und Nachteile statischer Logik

Statische Logik hat in der Regel einen höheren Platzbedarf (bei standard-CMOS: zwei Transistoren pro Eingang). Eine Ausnahme bilden hierbei die transmission gates, die jedoch ihrerseits den Nachteil der Signalabschwächung haben. Bei standard-CMOS ist außerdem die Lastkapazität sehr groß, da sie aus mindestens einem n-Kanal- und einem p-Kanal-Transistor besteht.

Vorteile sind der hohe Störabstand (besonders bei standard-CMOS) und die gute Signalregenerierung. Bei den Familien pseudo-NMOS und transmission gates fallen diese Effekte schwächer aus. Die Verlustleistung ist bis einschließlich 0 Hz frequenzabhängig (d.h. bei 0 Hz Schaltfrequenz gibt es keine Verluste mehr).

9.2 Clocked CMOS — C^2MOS

Als Abhilfe für die kapazitive Darstellung des Logikwertes „1“ bei dynamischer Logik bietet sich noch die Möglichkeit, eine **dynamische Version von standard-CMOS** zu realisieren. Dabei wird ein standard-CMOS Gatter dynamisch betrieben. Wegen des sehr hohen Flächenbedarfs werden in dieser Schaltungsart nur sehr wenige, einfache Gatter realisiert.

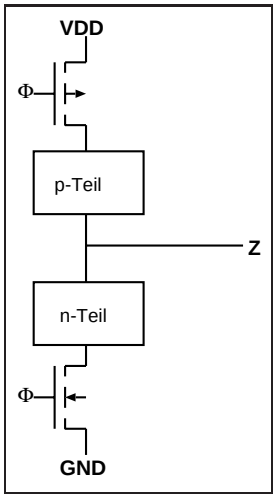


Abbildung 9.2: Aufbau eines C^2MOS Gatters

Die einfachste C^2MOS Schaltung ist der Inverter in Abbildung 9.3. In Abbildung 9.4 ist schematisch dargestellt, wie man daraus durch einige Modifikationen ein Latch gewinnt.

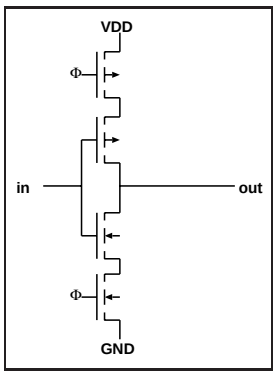


Abbildung 9.3: C^2MOS Inverter

Die so erhaltene Schaltung kann schematisch auch folgendermaßen dargestellt werden:

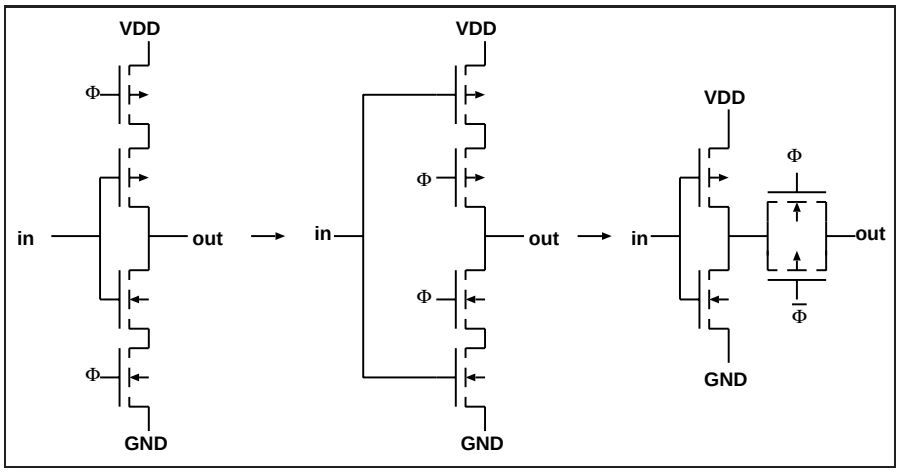


Abbildung 9.4: Vom C²MOS Inverter zum einfachen Latch

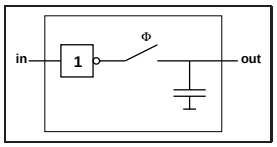


Abbildung 9.5: Invertierendes Latch

Für dieses Latch gilt $\text{out}(t) = \overline{\text{in}(t-1)}$. Solche Latches (mit nachgeschaltetem Inverter) werden dazu benutzt, um Signalzustände zu speichern. Mit ihnen können daher sequentielle Schaltungen realisiert werden.

10 Sequentielle Schaltungen

Die bisher betrachteten Schaltungen haben alle keine Möglichkeit, einen Wert zu speichern. Ihr Zustand hängt damit immer nur vom Zustand der Eingangsvariablen ab. Jetzt sollen Schaltungen betrachtet werden, deren Zustand auch von vorangehenden Zuständen, d.h. auch von den in Registern gespeicherten Werten, abhängen kann, sogenannte **sequentielle Schaltungen**.

10.1 Realisierung von Registern

Für die Realisierung der Register ist prinzipiell das Latch aus dem vorangehenden Kapitel geeignet. Es besitzt jedoch den Nachteil, daß bei geschlossenem Schalter sich die Änderungen des Eingangssignales direkt auf den Ausgang auswirken, was zu Race Problemen führen kann. Bei einem Register will man dies vermeiden, da sich der Ausgang nur zu definierten Zeitpunkten (normalerweise bei einer Taktflanke) ändern soll. Das ist das Grundelement für synchrone Logik

10.1.1 Statische Register

Eine Speicherwirkung kann durch Rückkopplung der Inverter erzielt werden (Abb. 10.1). Damit ist dann auch das Halten des gespeicherten Wertes bei geringen Taktfrequenzen gesichert.

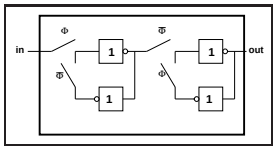


Abbildung 10.1: Statisches Flip-Flop

Die Probleme mit der Taktverschiebung bleiben allerdings. Um die Taktverzögerungen möglichst klein zu halten, ist es hier am günstigsten, das invertierte Taktsignal jeweils lokal entsprechend Abb. 10.2 zu erzeugen. Dabei werden für jede Speicherzelle insgesamt 20 Transistoren benötigt.

Statische Flip-Flops können nicht nur durch rückgekoppelte Inverter, sondern auch durch rückgekoppelte Logikgatter realisiert werden. Aus der Digitaltechnik müßte das Latch aus Abb. 10.3 bekannt sein, aus dem ein Flip-Flop leicht gewonnen werden kann.

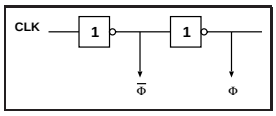


Abbildung 10.2: Erzeugung wenig verzögerter Taktsignale aus der Taktversorgung

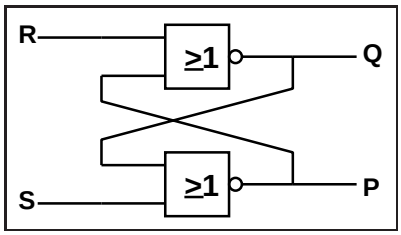


Abbildung 10.3: RS-Latch

Die Zustandstabelle des RS-Latches ist in Tabelle 10.1 zusammengefaßt. Problematisch ist der Übergang von RS = 11 auf RS = 00, da dies zu dem instabilen Zustand QP = 00 führt (Q und P müssen zueinander invertiert sein).

R	S	Q	P
0	1	1	0
1	0	0	1
1	1	0	0
0	0	Q(t-1)	P(t-1)

Tabelle 10.1: Übergangstabelle des RS-Latch

Durch einen Inverter kann verhindert werden, daß die kritische Eingangskombination auftritt und man erhält das sogenannte **D-Latch** nach Abb. 10.4. Leider geht dadurch aber auch die Eingangskombination für den speichernden Zustand mit verloren.

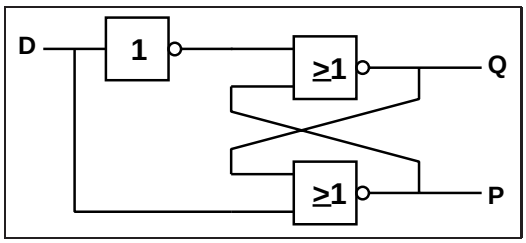


Abbildung 10.4: D-Latch

Durch Hinzufügen von zwei AND Gattern und durch Takten der Schaltung kann der speichernde Zustand wieder hinzugefügt werden, ohne daß ein kritischer Zustand auftreten kann. Man erhält so das sogenannte **getaktete D-Latch** nach Abb. 10.5.

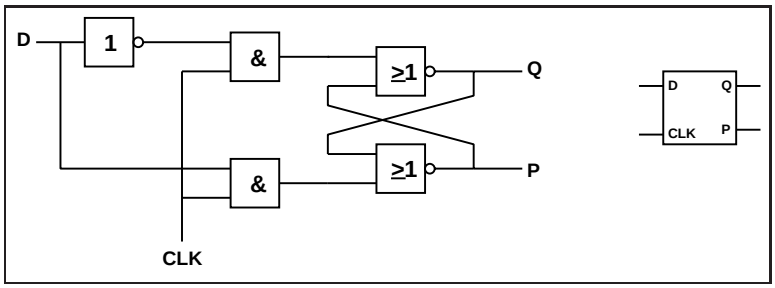


Abbildung 10.5: Getaktetes D-Latch

Aus zwei solchen getakteten D-Latches kann nun sehr einfach durch Hintereinanderschalten und komplementäres Takten ein Flip-Flop aufgebaut werden (Abb. 10.6).

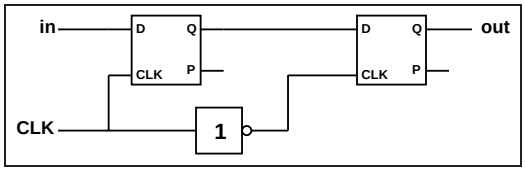


Abbildung 10.6: Flip-Flop aus zwei getakteten D-Latches

Bei CLK= 1 leitet das erste Latch den Ausgang Q weiter. Bei CLK= 0 behält es seinen Wert bei und das zweite Latch leitet diesen Zustand an den Ausgang.

Der Inverter im Innern der zweiten Latches kann eingespart werden, wenn statt des Ausgangs Q des ersten Latches P verwendet wird. Die Inversion für $\overline{\text{CLK}}$ kann ebenfalls eingespart werden, wenn nach den deMorganschen Regeln die AND Gatter durch NOR Gatter mit invertierten Eingängen ersetzt werden. Diese Inversionen können durch die Gatter „hindurchgeschoben“ werden. Dies und weitere Modifikation führen zu einer Realisierung mit Komplexgattern (Abb. 10.7).

Das so erhaltene Flip-Flop wird auch **Master-Slave-D-Flip-Flop** genannt, da das vordere Latch, der Master, den neuen Zustand bei der steigenden Taktflanke zuerst übernimmt und bei der fallenden Taktflanke diesen dem zweiten, dem Slave, „aufzwingt“. Der Ausgang des gesamten Flip-Flops kann sich dann nur bei einer fallenden Taktflanke ändern, während der Zustand am Eingang nur während der steigenden Taktflanke registriert wird.

Durch die statische Realisierung aus logischen Gattern benötigt dieses Flip-Flop allerdings insgesamt 26 Transistoren. Dafür ist es sehr sicher gegenüber Taktverschiebungen, weil es nur ein Taktsignal verwendet, und funktioniert auch bei geringen Taktfrequenzen.

Register ist ein Array (Feld) von Flipflop, was die Größe der zu verarbeitenden Daten reflektiert. Er ist wie gesagt flankengesteuert und ist der sicherste und schnellste Speicher überhaupt, leider auch der teuerste aufgrund der vielen Transistoren.

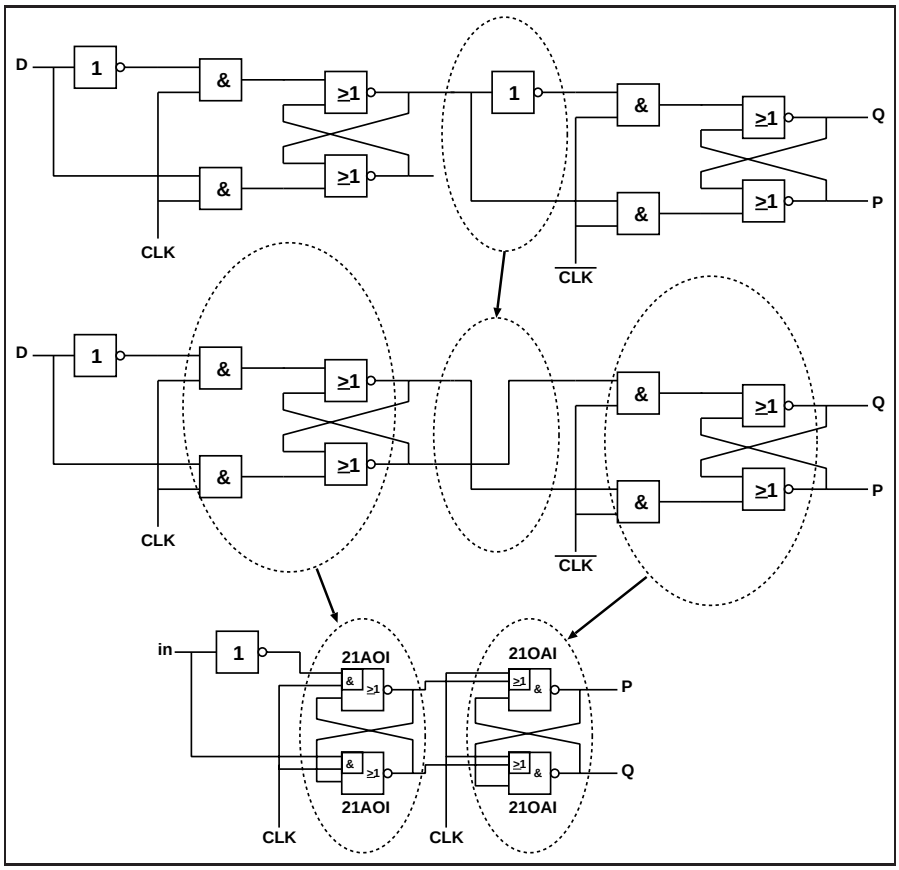


Abbildung 10.7: Von der einfachen Realisierung des D-Flip-Flop zur Realisierung mit Komplexgattern

10.2 Taktverteilung — Taktnetzwerke

Der schematische Aufbau sequentieller Schaltungen nach Art der Moore-Maschine ist abstrakt anzusehen. In Wirklichkeit zeigen die Realisierungen sequentieller Schaltungen eher folgende Struktur, die mehr der Funktionsweise folgt:

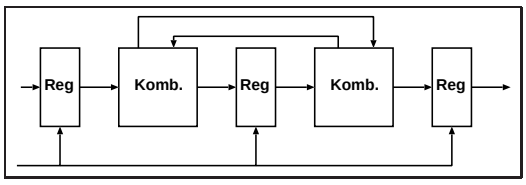


Abbildung 10.8: Realistische Struktur sequentieller Schaltungen

Im Folgenden wird davon ausgegangen, daß die Flip-Flops in den Registern ihre Eingangswerte jeweils zur steigenden Taktflanke übernehmen, d.h. daß es sich um Master-Slave-Flip-Flops handelt.

Damit eine solche Schaltung korrekt funktioniert, müssen die Eingangssignale eine kurze Zeit

τ_{setup} **vor** der steigenden Taktflanke stabil sein, damit sie korrekt übernommen werden. Ebenso müssen sie **nach** der steigenden Taktflanke noch die Zeit τ_{hold} stabil bleiben (Abb. 10.9).

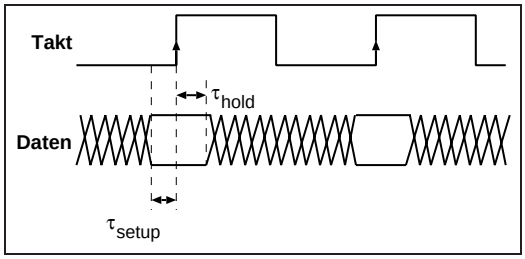


Abbildung 10.9: Setup- und Holdzeiten beim Flip-Flop

Da das Taktsignal an allen Flip-Flops von der gleichen Quelle kommt und über den gesamten Chip verteilt werden muß, kommt es immer zu Taktverzögerungen zwischen den einzelnen Flip-Flops. Eine solche Taktverzögerung kann dann dazu führen, daß die positive Taktflanke an einem oder mehreren Flip-Flops in einen Bereich fällt, in dem sich die Eingangsdaten noch ändern, so daß das Flip-Flop falsche Werte speichert.

Eine Möglichkeit, das Problem zu umgehen ist, die Taktversorgung immer „parallel“ zu den Daten und auch in der gleichen Richtung auf dem Chip zu führen, so daß Takt und Daten in gleichem Maße verzögert werden. Leider ist dies in der Regel — besonders bei komplexen Schaltungen — nicht möglich, da die Daten selbst in völlig unterschiedlichen Richtungen über den Chip fließen.

Um die Taktverzögerungen dennoch möglichst gering zu halten, muß das Taktnetzwerk von einem Treiber mit hoher Treiberleistung versorgt werden. Ein solcher Clocktreiber braucht ein sehr hohes $\frac{W}{L}$-Verhältnis, was dazu führt daß er selbst eine sehr hohe Eingangskapazität hat, die in der Regel von der eigentlichen Taktquelle (meist ein Quarzschwingkreis) nicht direkt getrieben werden kann. Das Problem würde mit einer solchen Realisierung nicht gelöst sondern nur auf den Eingang des Treibers verlagert.

Eine bessere Möglichkeit besteht darin, mit mehreren Invertern, die jeweils auf die Eingangskapazität des nachfolgenden abgestimmt sind, eine **Treiberkette** zu realisieren. Eine solche Anordnung (Abb. 10.10) besitzt ein Optimum bezüglich Anzahl und Dimensionierung der Inverter für gegebenes C_g (Treiberleistung der Quelle) und C_{L} (Kapazität des Taktnetzwerkes):

$$\begin{aligned} n = \text{Anzahl der Inverter} &= \ln\frac{C_{\text{L}}}{C_g} - 1 \\ k = \frac{\frac{W}{L}\,\text{Nachfolger}}{\frac{W}{L}\,\text{betrachteter Inverter}} &= e \qquad (\approx 2.7) \end{aligned}$$

Gute Näherungslösungen erhält man auch, wenn man bei einem Dimensionierungsfaktor von $k = 3\ldots5$ als Anzahl der Inverter den Wert $n = \log_k \frac{C_{\text{L}}}{C_g}$ wählt.

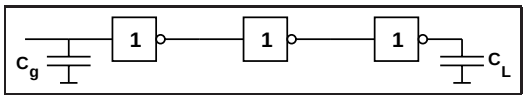


Abbildung 10.10: Treiberkette als Clocktreiber

Beispiel 10.1

Treiberkette mit $k = 4, \frac{C_L}{C_g} = 1024 \Rightarrow n = \log_4 1024 = 5$.

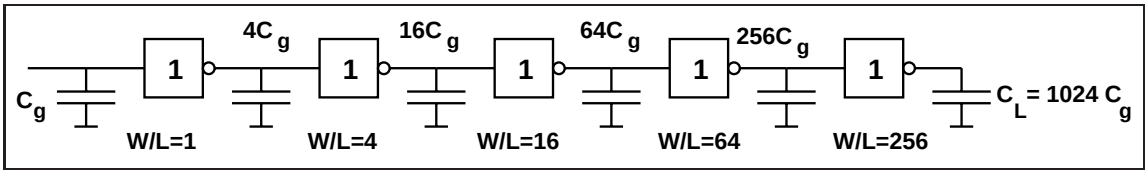


Abbildung 10.11: Fünfstufige Treiberkette für $\frac{C_L}{C_g} = 1024$

Selbstverständlich verursacht diese Art der Taktverteilung durch den hohen dynamischen Strom (Umladen des gesamten Taktnetzwerkes zweimal in jedem Takt) große dynamische Verluste. Es ist daher sinnvoll, das Taktnetzwerk in viele, kleinere Teilnetze aufzuteilen, die dann getrennt getrieben werden können.

Die in Abb. 10.12 dargestellte Anordnung stellt einen Kompromiß zwischen Anzahl der Inverter und der einfachen Kette dar. Eine solche Anordnung wird in der Regel die günstigste sein. Der Aufwand bei all diesen Anordnungen ist in etwa gleich.

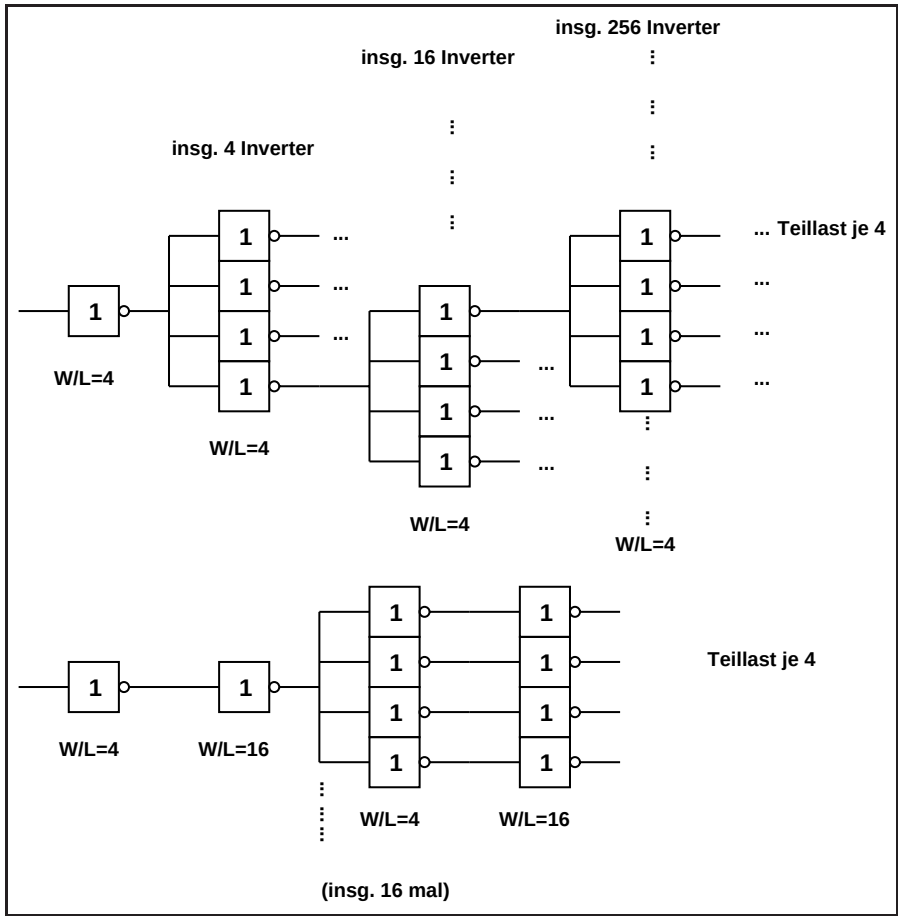


Abbildung 10.12: Treiberanordnungen für Takt-Teilnetzwerke

10.3 Takterzeugung

Die Taktsignale werden in der Regel aus Quarzoszillatoren gewonnen. Hohe Taktfrequenzen werden dabei durch Vervielfachung aus niedrigeren (bei Personal Computern in der Regel 33 MHz) erzeugt (Abb. 10.13).

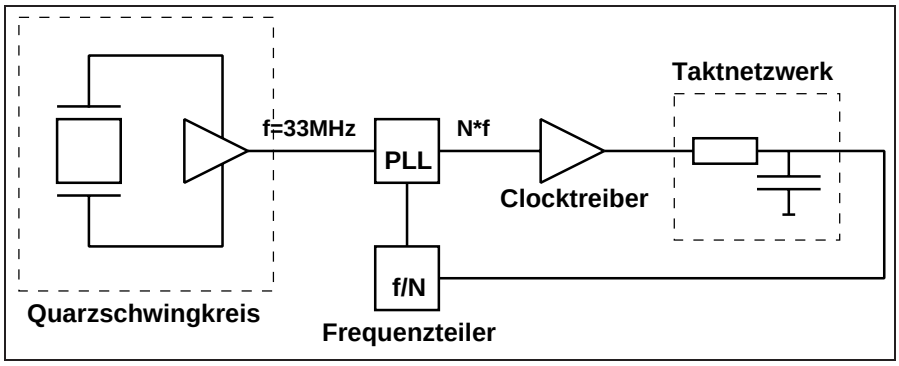


Abbildung 10.13: Erzeugung des Taktsignals

Ein Oszillator erzeugt eine hohe Frequenz, die als Taktsignal nach außen zur Schaltung geführt wird. Der Clocktreiber verstärkt dieses Signal. Dieses Signal wird zugleich mit einem Frequenzteiler, der z.B. aus eimem Zähler besteht, heruntergeteilt. Die PLL(Phase Locked Loop) Schaltung regelt den Oszillator so, daß die Phase des Quarzozillators und des Oszillators im Chip gleich sind. Da Phase das Integrals der Frequenz ist, gewährleistet die PLL- Schaltung, daß kein Taktsignal verloren geht.)

11 Ein/Ausgangstreiber — I/O-Treiber

Die Ein- und Ausgangssignale einer integrierten Schaltung werden über ca. $100 \times 100 \mu m$ große Metallkontakte, die sog. **Pads**, mit den Anschlüssen des Gehäuses verbunden. Diese Kontaktierung erfolgt mechanisch (in der Regel durch eine Maschine), was die für Chipverhältnisse sehr großen Padausdehnungen erklärt. In der Mikroelektronik II wird mehr auf das Thema Packaging eingegangen.

11.1 Eingangspads

Da in MOS-Technologie hergestellte Schaltungen sehr empfindlich gegen Überspannungen sind (Durchschlag über dem Gateoxid), werden die Eingangspads häufig noch mit Widerständen und Dioden geschützt (Abb. 10.14). Ursachen für die Uberspannung ist meistens elektro-statische Entladung ESD (electro-static Discharge).

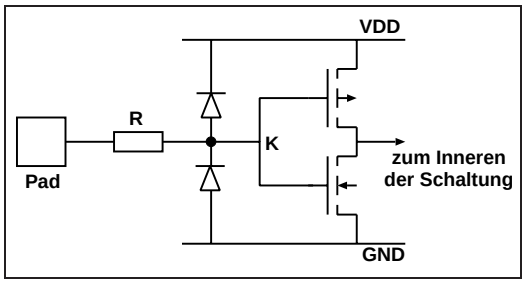


Abbildung 11.1: Eingangspad mit Schutzbeschaltung

Dabei dient der Widerstand R als Strombegrenzer für die Dioden. Überschreitet die am Pad anliegende Spannung V_{DD} oder unterschreitet sie GND, so sorgen die Dioden dafür, daß der Knoten K dennoch im Bereich von $-0.5 \ldots 5.5$ V liegt, d.h. nicht wesentlich ober/unterhalb der Versorgungsspannungen. Der Treiber ist in der Regel erforderlich, da die Leitung vom Pad bis zum eigentlichen Anschluß an der Logik relativ lange ist (siehe auch Abb. 11.1).

11.2 Ausgangspads

Einfache Ausgangspads bestehen aus einem Treiber und dem eigentlichen Pad.

Da die Ausgänge eines Chip in der Regel mit — für Chipverhältnisse — sehr langen und breiten Leitungen (z.B. den Leiterbahnen einer Platine) verbunden sind, sind hier jedoch Treiber mit hoher Treiberleistung erforderlich, auf die gleich noch eingegangen wird.

11.3 Tri-State-Treiber

Damit mehrere Chips in einem System manche Leitungen gemeinsam nutzen können, müssen die entsprechenden Ausgänge die Möglichkeit bieten, hochohmig vom Pad abgetrennt zu werden. Dies ist insbesondere bei **Bussen** erforderlich (z.B. Adressbus, Datenbus im Rechner). Solche Pads müssen daher mit sogenannten **Tri-State-Treibern** versehen werden.

Der Name Tri-State-Treiber deutet schon an, daß der Ausgang eines solchen Treibers drei definierte Zustände hat: logisch „0“, logisch „1“, und „hochohmig getrennt“. Im letzten Zustand ist die Ausgangsleitung hochohmig von beiden Versorgungspotentialen abgetrennt und damit passiv, so daß ein anderer Treiber (u.U. auf einem anderen Chip) die Leitung ungestört ansteuern kann. Um diesen zusätzlichen Zustand steuern zu können, ist ein sogenanntes „enable“-Signal erforderlich (Abb. 10.15).

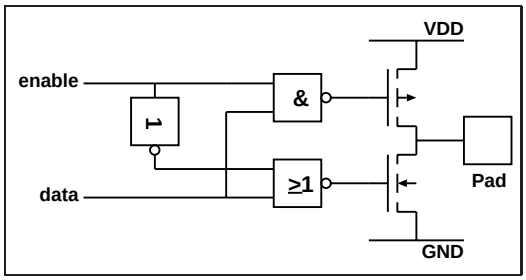


Abbildung 11.2: Ausgangsteil eines Tri-State-Pad

Solange das Signal enable= 1 ist werden die Daten auf das Pad ausgegeben. Mit enable= 0 wird das Pad hochohmig (über die beiden dann gesperrten Transistoren) von den Versorgungspotentialen abgetrennt, so daß andere Gatter (oder Chips) die Leitung verwenden können. Für einen kollisionsfreien Zugriff muß lediglich dafür gesorgt werden, daß niemals zwei unterschiedliche Pads an der gleichen Leitung gleichzeitig enable= 1 erhalten.

Da Busse in der Regel für Schreib- und Lesezugriffe ausgelegt sind, existiert zu einem solchen Pad auch ein Leseteil (Abb. 10.16).

Der Schmitt-Trigger (Kippschaltung mit Hysterese) dient zum Erkennen verzerrter Signale, die durch kapazitive und induktive Kopplungen und Einstrahlung von anderen Komponenten entstehen können. Das Lesen ist immer kollisionsfrei.

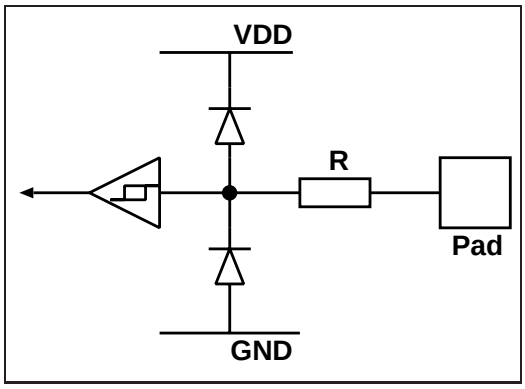


Abbildung 11.3: Eingangsteil eines Tri-State-Pad

12 Anordnung der Schaltungsteile auf einem Chip

Durch die erforderliche hohe Treiberleistung der Ausgangspads sind die Transistoren der Ausgangstreiber so groß, daß sie die Ausmaße des eigentlichen Pads erreichen und überschreiten können.

Wegen der Kontaktierung zum Chip liegen die Pads normalerweise am Rande des Chips.[1]

Damit die Spannungsabfälle über den Versorgungsleitungen möglichst gering sind, verwendet man in der Regel mehrere Anschlüsse für V_{DD} und GND.

Eine günstige Art der Realisierung eines Chip zeigt Abbildung 11.1. In dieser Realisierung enthalten die Clock-Leitungen sowie die Ein- und Ausgangsleitungen viele Brücken (bzw. Kreuzungspunkte, mehr als eine Metallisierungsebene im Prozeß möglich ist). Um die Wahrscheinlichkeit von Kurzschlüssen von Ein- und Ausgangsleitungen mit Versorgungspotential zu senken (die durch Herstellungstoleranzen möglich sind), legt man die Anschlüsse für die Versorgungsspannungen in die Ecken des Chips.

In Abbildung 11.2 sind noch einige Möglichkeiten zur Anordnung der Pads — je nach Breite und Länge des zur Verfügung stehenden Gebietes — dargestellt.

12.1 Register

Die in Kapitel 10 betrachteten Flip-Flops eignen sich nur zum Speichern jeweils eines einzigen logischen Wertes. Um mehrere Werte speichern zu können solche Flip-Flops zu sogenannten Registern zusammengeschaltet werden.

Die einfachste Art eines Register besteht darin, daß alle Flip-Flops mit dem gleichen Takt betrieben werden. Damit können gleichzeitig genausoviele Bitwerte gespeichert werden wie Flip-Flops verwendet werden. Beim Einlesen werden immer alle Flip-Flops gleichzeitig geladen, zum Auslesen stehen immer alle Flip-Flop-Ausgänge zur Verfügung.

Eine weitere naheliegende Möglichkeit ist das **Schieberegister**, auch FIFO- (FIRST IN FIRST OUT) Register genannt, bei dem die eingelesenen Werte in derselben Reihenfolge wieder am Ausgang erscheinen (Abb. 11.3).

[1] Bei den neueren Flip-Chip-Techniken muß das nicht immer der Fall sein

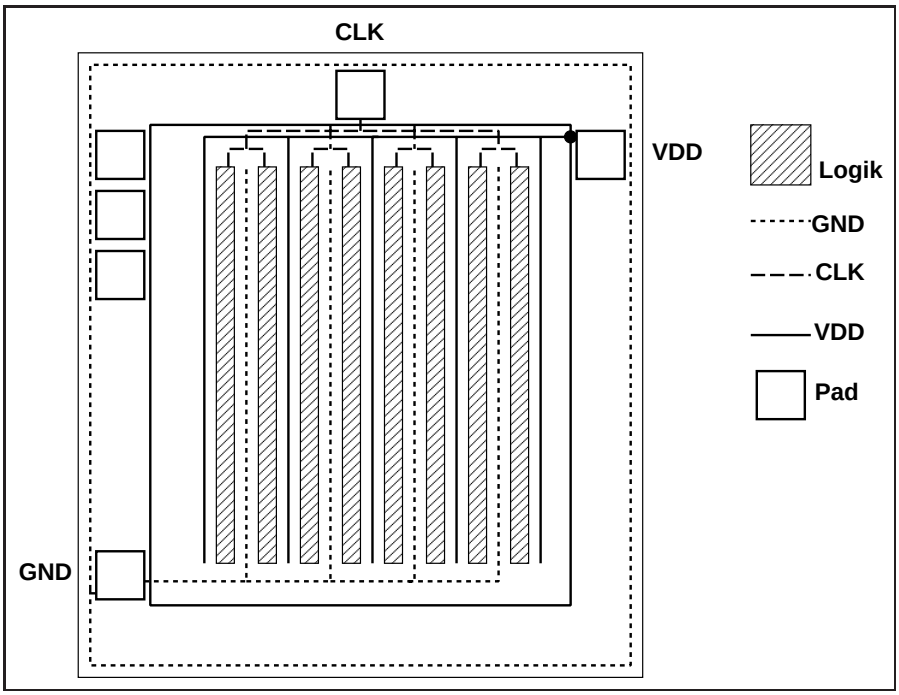


Abbildung 12.1: Anordnung von Pads, Logik, Clock- und Versorgungsleitungen

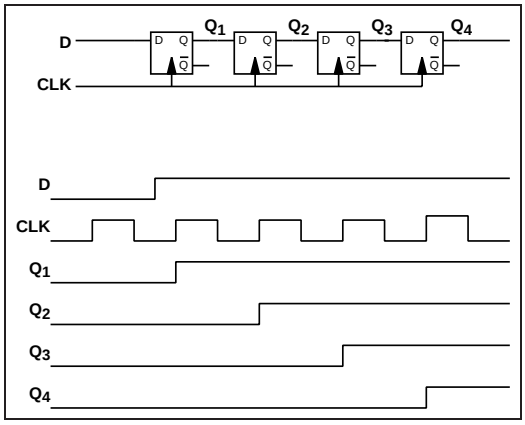


Abbildung 12.3: Schieberegister (FIFO)

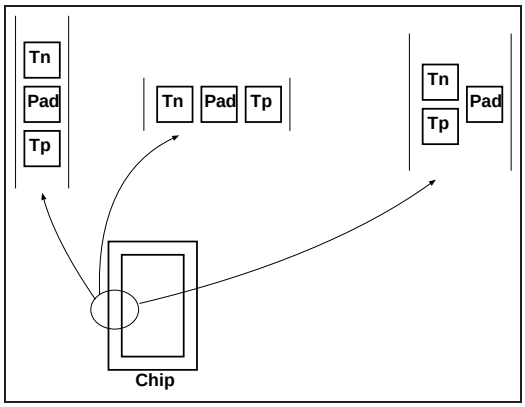


Abbildung 12.2: Mögliche Anordnungen von Pads und den zugehörigen Transistoren (Treiber)

Solche Schieberegister können zur Zwischenspeicherung von Daten über mehrere Takte verwendet werden. Macht man die Ausgänge aller Flip-Flops zugänglich, so können Daten seriell eingelesen und parallel verarbeitet werden. Solche Schaltungen finden z.B. in seriellen Schnittstellen von Rechnern Verwendung.

Daneben kann durch geeignete Beschaltung von Flip-Flops auch ein sogenanntes LIFO-Register (LAST IN FIRST OUT) realisiert werden.

Will man immer mehrere Informationseinheiten (Bit) in sogenannten Datenwörtern zusammengefaßt speichern und gemeinsam bearbeiten, so können Schieberegister in Zeilen nebeneinander angeordnet werden. Abbildung 11.4 zeigt einen FIFO Speicher mit der Wortbreite 8bit und der Tiefe (d.h. Anzahl der speicherbaren Worte) 7.

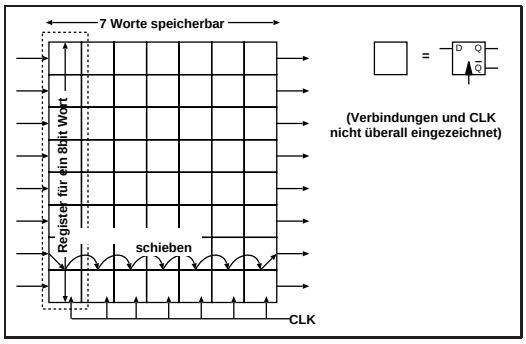


Abbildung 12.4: FIFO Speicher mit Wortbreite 8bit und Tiefe 7

12.2 RAM — Schreib-Lese-Speicher mit Wahlzugriff

Bei Registern ist die Reihenfolge von eingelesenen und ausgegebenen Werten — unabhängig davon, ob es sich um ein FIFO oder LIFO handelt — fest vorgegeben. Viele Anwendungen erfordern jedoch einen Speicher, auf den man wahlfrei zum Lesen und Schreiben zugreifen kann. Speicher mit diesen Eigenschaften heißen **RAM** (engl.: RANDOM ACCESS MEMORY). Um solche RAM zu realisieren, benötigt man an jeder Speicherzelle neben dem Dateneingang jeweils noch ein Signal für den Schreibzugriff (WR — write) und eines für den Lesezugriff (RD — read), wie in Abb. 11.5 dargestellt. Außerdem ist sinnvollerweise nur eine Leitung zum Schreiben und Lesen der Daten für jede Speicherzelle vorhanden. Für eine Zeile von Speicherzellen kommt man dann mit nur einer Leitung aus, da die entsprechenden RD und WR Signale zum Auswählen der Speicherzelle innerhalb der Zeile verwendet werden können.

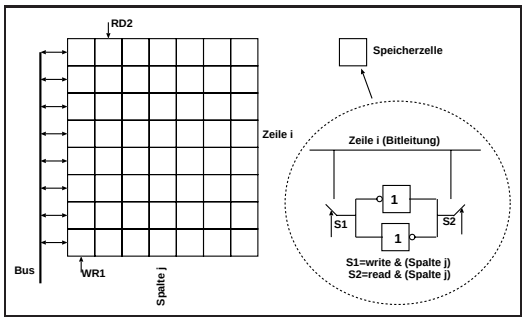


Abbildung 12.5: Schematischer Aufbau eines RAM-Speicherfeldes

Die Speicherzellen können z.B. nach Abb. 11.6 aufgebaut sein.

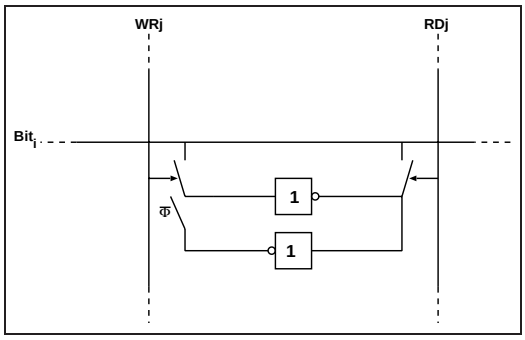


Abbildung 12.6: Realisierung einer RAM-Zelle mit Invertern

Da es sich hier um eine bistabile Zustandsspeicherung handelt muß beim Auslesen darauf geachtet werden, daß die Information nicht zerstört wird.

Um aus dem Speicherzellenfeld ein funktionierendes RAM zu erhalten, ist noch Adressierungslogik erforderlich. Insgesamt ergibt sich der Aufbau eines RAM nach Abbildung 11.7.

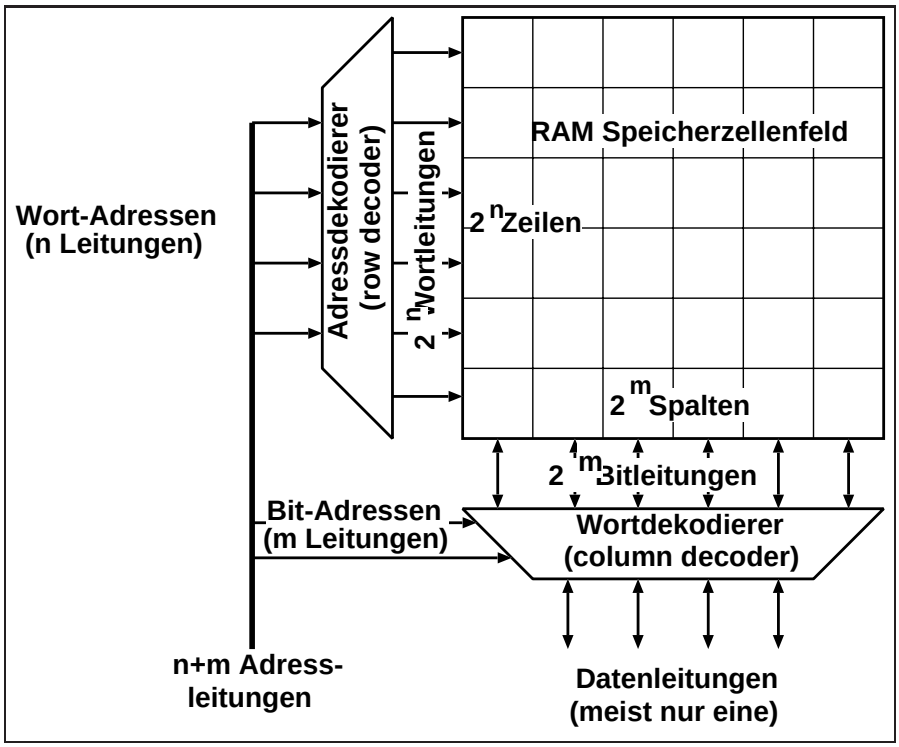


Abbildung 12.7: Schematischer Aufbau eines RAM

Der row decoder dient dazu, genau die Zeile im Speicherzellenarray anzusteuern, deren Nummer *i* in der row Adresse kodiert ist (d.h. Zeile 0 wird durch die row Adresse 0…0 angesteuert, Zeile 1 durch 0…01 usw.). Funktional gesehen ist er daher entsprechend Abbildung 11.8 aufgebaut.

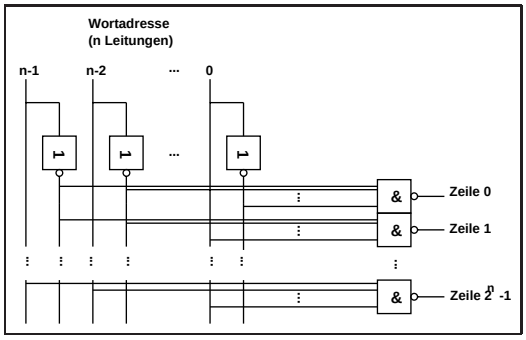


Abbildung 12.8: Funktionaler Aufbau des row decoders

In der Regel werden statt der NAND-Gatter des row decoders jedoch pseudo-NMOS NOR-Gatter verwendet, da diese flächengünstiger realisierbar sind (Bem.: es handelt sich hier um Gatter mit vielen Eingängen!). Ein solches NOR ist dann entsprechend Abbildung 11.9 aufgebaut.

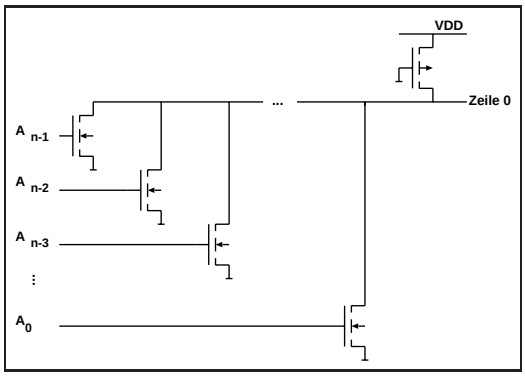


Abbildung 12.9: Aufbau eines NOR im row decoder mit Beschaltung für Zeile 0

Am column decoder muß aus 2^m Leitungen genau eine (bei einer Datenleitung, wie heute üblich) entsprechend der Bitadresse ausgewählt werden. Es handelt sich also um einen $m \times 1$-Multiplexer, der am einfachsten mit Übertragungsgattern realisiert werden kann (Abb. 11.10, column decoder mit Datenbreite 1 und $m = 3$). Dort ist auch eine Realisierung gezeigt, die auf die Hintereinanderschaltung vieler Übertragungsgatter verzichtet und deshalb längere Verzögerungszeiten durch den höheren Widerstand der Datenleitung gegenüber den Ausgängen der Speicherzellen besitzt.

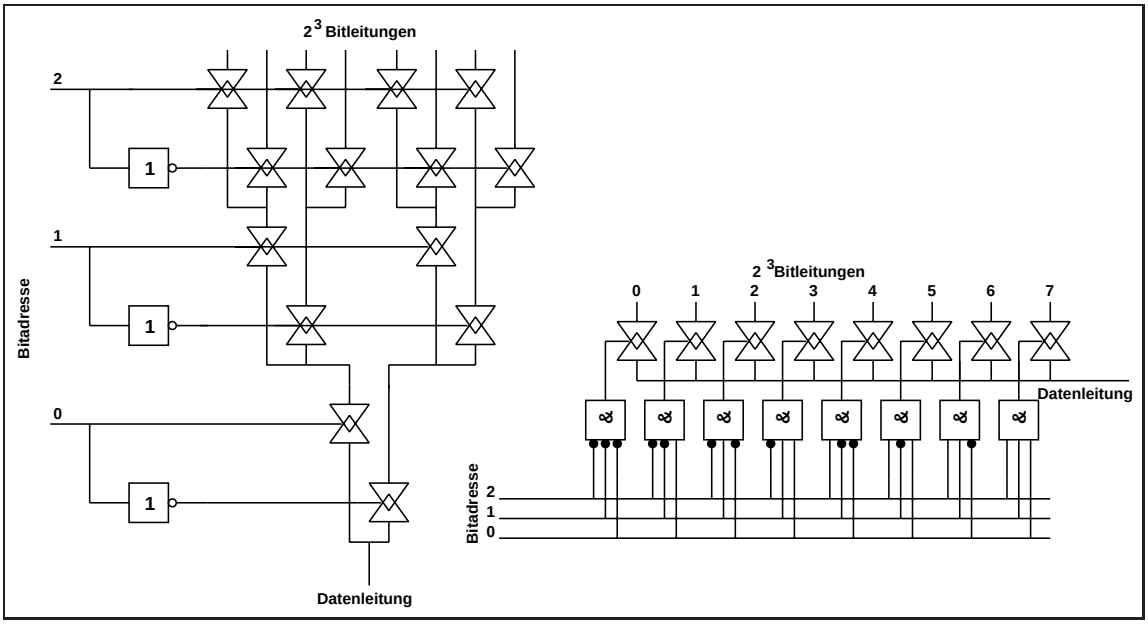


Abbildung 12.10: column decoder als Multiplexer

12.2.1 RAM-Zellen

Eine einfache Speicherzelle für ein RAM entsprechend Abbildung 11.6 (ohne synchrones Schreiben und Lesen), kann so modifiziert werden, daß sie folgendermaßen aussieht.

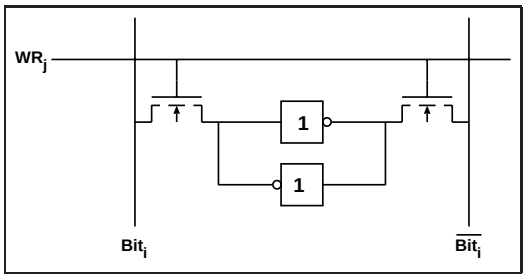


Abbildung 12.11: Statische 6-Transistor Ramzelle

Die Ansteuerung erfolgt dabei beidseitig. Beim Schreiben von Null erzwingt der linke Schalter das Kippen der beiden gekoppelten Inverter. Beim Schreiben von Eins ist der rechte Schalter wirksam. Der jeweilige andere Schalter unterstüzt das Schreiben und macht es schneller. Hierbei handelt es sich um eine statische Speicherzelle, der Zustand kann beliebig lange gehalten werden, da die beiden Inverter sich gegenseitig stabilisieren. SRAM kann gut in einem CMOS-Prozeß integriert werden (enbedded SRAM). Durch die insgesamt 6 Transistoren ist der Platzbedarf jedoch relativ groß.

Statt die Information wie hier durch den Stromfluß zu speichern könnten die Gates der beiden Speichertransistoren als dynamische Ladungsspeicher verwendet werden und die beiden PMOS wären überflüssig. Damit erhält man die dynamische 4-Transistor Ramzelle entsprechend Abb. 11.12.

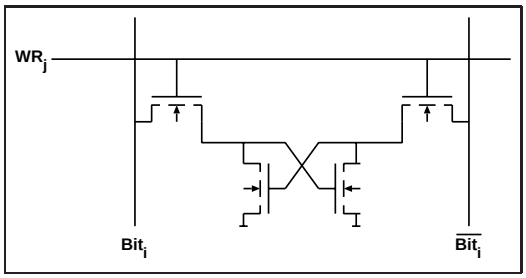


Abbildung 12.12: Dynamische 4-Transistorzelle

Eine Ansteuerungsseite kann ebenfalls noch eingespart werden, um die Zelle weiter zu verkleinern. Von den beiden so verbleibenden Transistoren wird dann eigentlich nur noch der Schalttransistor verwendet; vom Speichertransistor wird lediglich die Gatekapazität als Ladungsspeicher benötigt. Diese kann damit gleich durch einen Kondensator ersetzt werden und man erhält die dynamische 1-Transistorzelle nach Abbildung 11.13.

Eine solche 1-Transistorzelle kann dann mit folgenden Alternativen realisiert werden, von denen die zweite die modernere ist. Diese zweite Transistorzelle verwendet einen sogenannten Trench-Kondensator, dessen Platten aus dem Polysilizium-„Keil“ und dem Substrat gebildet werden. Allerdings erfordert dieser Trench-Kondensator spezielle Technologie, die von Standard CMOS abweicht.

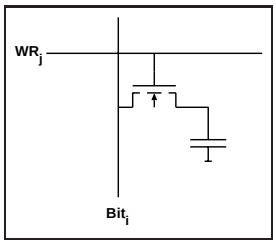


Abbildung 12.13: Dynamische 1-Transistorzelle

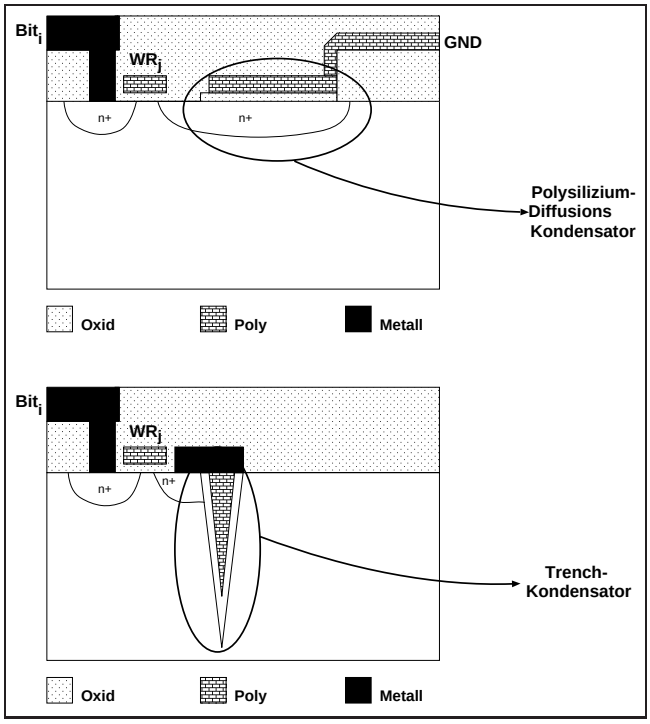


Abbildung 12.14: Mögliche Realisierungen der dynamischen 1-Transistorzelle

Der DRAM Speicher benötigt die Kleinste Fläche. Er muß aber immer refreshed werden, damit die Ladung erhalten bleibt und ist relativ lansam. Er wird meist als diskreter großer Speicher realisiert.

12.3 ROM — Festwertspeicher

Ein Festwertspeicher (ROM, engl.: read only memory) ist von der Struktur her genauso aufgebaut wie ein RAM: er besitzt einen row decoder für die Wortadressen, einen column decoder für die Bitadressen und ein Speicherzellenarray. Der Unterschied liegt darin, daß in einem ROM keine Werte mehr verändert werden können. Der Inhalt wird beim Layout einprogrammiert.

Entsprechend anders ist daher das Speicherzellenfeld des ROM aufgebaut. Im Prinzip handelt es sich bei einer Speicherzelle jeweils um einen Transistor, der so geschaltet ist, daß er bei Ansteue-

rung durch die Wortleitung die Bitleitung entsprechend auf 1 oder 0 zieht, jenachdem welchen Wert der Hersteller für diese Speicherzelle vorgesehen hat. In Abbildung 11.15 ist dargestellt, wie sich ein ROM mit pseudo-NMOS NOR-Gattern (jeweils ein Gatter pro Bitleitung) realisieren läßt.

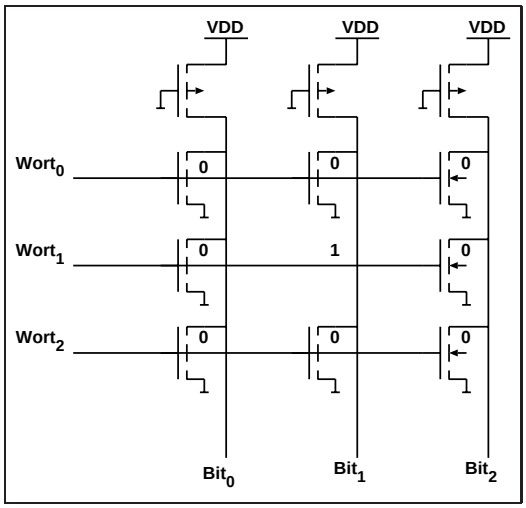


Abbildung 12.15: ROM-Speicherzellenfeld

Der in dieser Abbildung fehlende Transistor wird in allen Prozeßschritten bis auf die Metallisierung der Einfachheit halber mitrealisiert. Lediglich die Kontaktierung des Drain mit der Wortleitung wird weggelassen, so daß es für die Schaltung so „aussieht“, als sei der Transistor nicht vorhanden (siehe auch Abb. 11.16). Andere Möglichkeiten zur Realisierung sind:

- Das Diffusionsgebiet nicht zu realisieren (kompakteste Möglichkeit)
- Metallverbindung zur Bitleitung weglassen (flexibelste Möglichkeit, da diese erst in einem späten Produktionsschritt hergestellt würde)

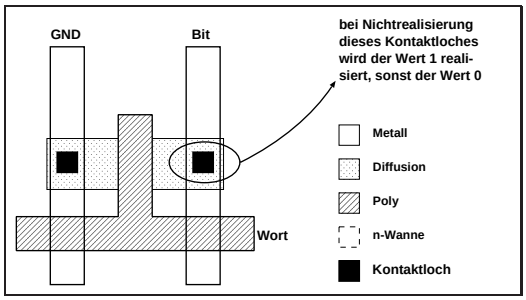


Abbildung 12.16: Speichertransistor einer ROM-Zelle

12.4 PLA — Programmierbare Logik-Arrays

Bei einem ROM sind immer sehr viele Speicherzellen realisiert und die Dekodierer (row und column decoder) sind für alle darstellbaren Adressen ausgelegt. Manche Anwendungen kommen

jedoch mit sehr kleinen Adressbereichen und einer sehr eingeschränkten Anzahl verschiedener Worte aus. Man denke da z.B. an die Zustände einer einfachen Zustandsmaschine, die ja auch in gewisser Weise gespeichert werden müssen.

Für solche Anwendungen gibt es sogenannte PLA (eng.: PROGRAMMABLE LOGIC ARRAY), mit denen mehrere boolesche Funktionen mehrerer Variablen in Form einer Tabelle („lookup table“) gemeinsam realisiert werden können.

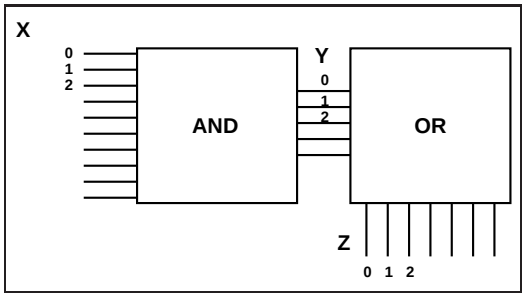


Abbildung 12.17: Blockstruktur eines PLA

Prinzipiell haben solche PLA eine AND-OR-Struktur entsprechend Abb. 11.17.

Damit lassen sich Funktionen der Art

$$\begin{aligned} y_0 &= \text{AND}(x_0, \overline{x_0}, \ldots, x_n, \overline{x_n}) \\ &\vdots \\ y_n &= \text{AND}(x_0, \overline{x_0}, \ldots, x_n, \overline{x_n}) \end{aligned}$$

$$\begin{aligned} z_0 &= \text{OR}(y_0, \ldots, y_m) \\ &\vdots \\ z_l &= \text{OR}(y_0, \ldots, y_m) \end{aligned}$$

realisieren, wobei mit AND() bzw. OR() jeweils eine Konjunktion/Disjunktion einer beliebigen Zahl der in den Klammern angegebenen Variablen handelt. Die entsprechenden Verknüpfungen können im Prozeßschritt der Metallisierung (d.h. im letzten Herstellungsschritt) durch Verbinden oder Nichtverbinden der entsprechenden Eingänge der Gatter hergestellt (sog. **Maskenprogrammierung**). Dadurch ist ein solches PLA nur genau einmal programmierbar.

Da die hier verwendeten Logikfunktionen AND und OR wegen der fehlenden Negation und der Reihenschaltung vieler Transistoren beim AND sehr viel Aufwand bedeuten nutzt man die Eigenschaft:

$$Z = \overline{A} \cdot \overline{B} \cdot \overline{C} = \overline{A + B + C}$$

Der AND-Teil kann damit durch ein NOR ersetzt werden, wenn man die negierten Eingänge verwendet. Im OR-Teil kann man ebenfalls NOR verwenden, wenn zusätzlich an die Ausgänge

noch Inverter angeschlossen werden. Damit erhält man die Realisierung eines PLA nach Abbildung 11.18. Hierbei werden pseudo-NMOS NOR-Gatter verwendet, da sie besonders platzsparend bei hoher Geschwindigkeit realisiert werden können. Sie bieten allerdings den Nachteil eines statischen Verlustes.

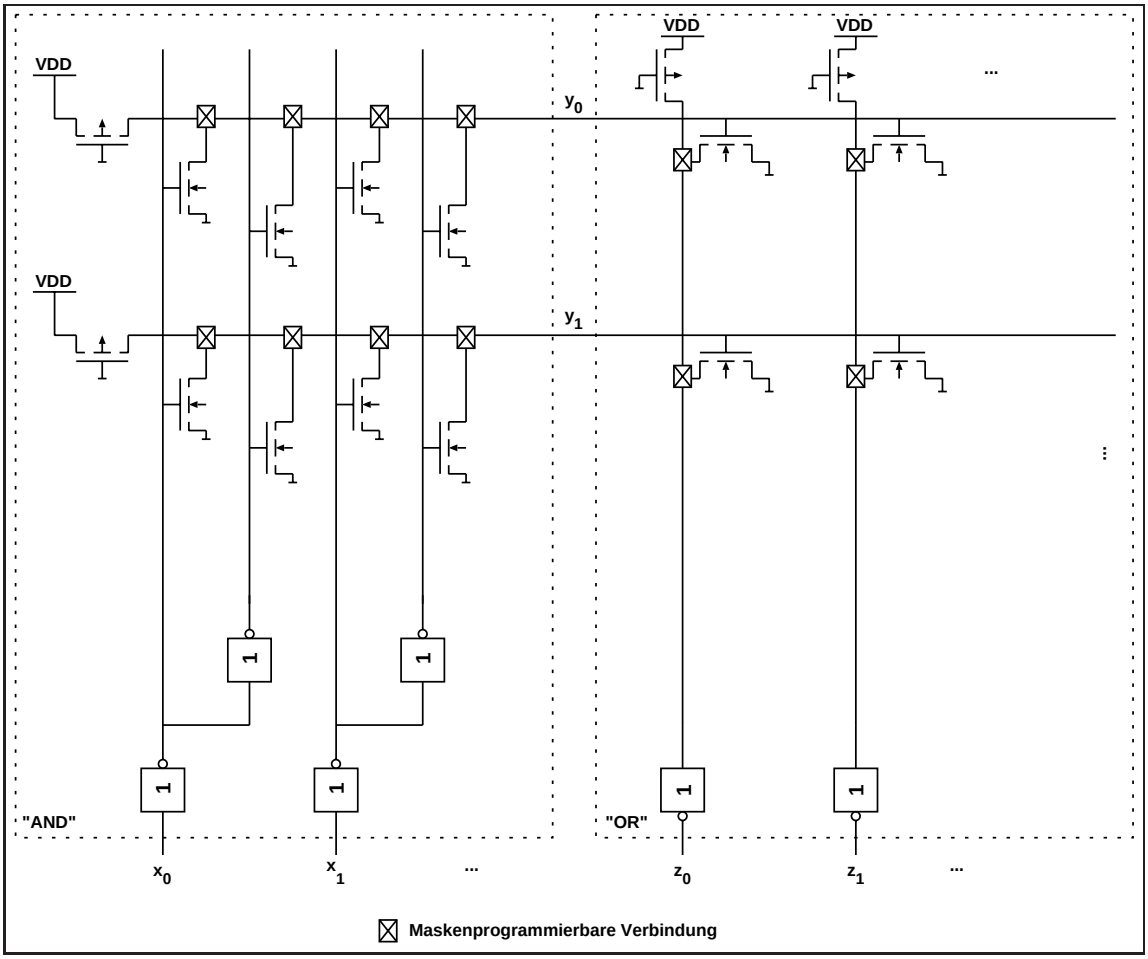


Abbildung 12.18: Aufbau eines PLA mit NOR-Gattern

12.5 NVM - Nichtflüchtiger Speicher

Die Nachteile von RAM und ROM bzw. PLA liegen auf der Hand: ein RAM „vergißt“ seine gespeicherten Werte beim Abschalten der Versorgungsspannung sofort; in einen Festwertspeicher (ROM, PLA) können keine neuen Werte mehr eingespeichert werden. Es ist jedoch für manche Anwendungen ein Speicher erforderlich, der Schreib- und Lesezugriffe zuläßt, die gespeicherten Werte beim Abschalten der Versorgungsspannung aber nicht verliert.

Um nichtflüchtigen sog. NVRAM (engl.: NON VOLATILE RAM), auch EEPROM (engl.: ELECTRICALLY ERASABLE PROGRAMMABLE ROM), zu realisieren werden **Floating Gate Transistoren** (Abb. 11.19) verwendet. Da das Poly (Floting Gate) nur vom Oxid umschlossen ist, kann

die Ladung normalerweise nicht entweichen. Diese Ladung kann auf der Siliziumoberfläche eine Inversionsschicht erzeugen und mit den Drain- und Souce-Anschlüssen abgelesen werden. Dies wird in der Abbildung Abb. 11.19 veranschaulicht.

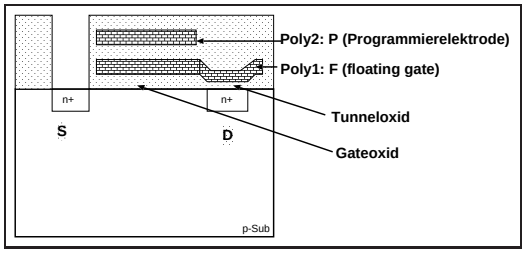


Abbildung 12.19: Floating Gate Transistor

Wenn zwischen Drain und der Programmierelektrode eine hohe positive Programmierspannung (z.B. $V_{PP} \approx 13 \ldots 15V$) angelegt wird, wird auf dem Floating Gate eine positive Ladung gespeichert (durch aus dem Floating Gate in den Kanal tunnelnde Elektronen), die nach Wegfall der Programmierspannung bestehen bleibt. Je nach Höhe dieser Spannung ist im Kanal permanent eine Inversionsschicht vorhanden, so daß der Transistor leitend bleibt.

Wird ein negatives V_{PP} als Programmierspannung angelegt, so kann die Ladung auf dem Floating Gate wieder neutralisiert werden und der Transistor wird wieder gesperrt. Entscheident bei der Programmierung ist die richtige Höhe der Spannung und auch die Pulsform. Es geht darum, daß Ladung durch das Oxid injiziert wird, ohne dabei das Oxid zu beschädigen. Natürlich muss das Oxid auch eine hohe Qualitätsanforderung erfüllen.

Damit hat man die Möglichkeit über die Programmierspannung permanent den Zustand des Transistors festzulegen. Solche Transistoren können dann in Speicherzellen entsprechend Abb. 11.20 verwendet werden, um nichtflüchtige Speicher zu realisieren.

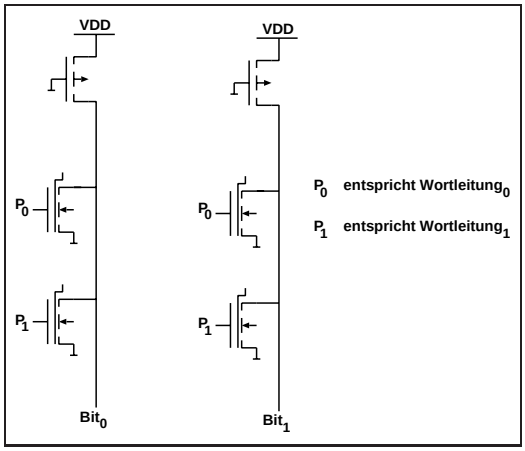


Abbildung 12.20: Aufbau eines NVRAM (EEPROM) mit Floating Gate Transistoren

13 Systemtechnik: Prozessoren

Prozessoren werden grob nach ihren Einsatzgebieten unterschieden. Für Anwendungen in der Signalverarbeitung gibt es die sog. **Signalprozessoren**, die analoge Signale verstärken, filtern und transformieren können. Diese Operationen können entweder analog oder digital (bei den digitalen Signalprozessoren, DSP) implementiert sein. Bei analogen Signalprozessoren ist der Algorithmus fest vorgegeben, er ist sozusagen fest als Hardware im Prozessor realisiert. Bei DSP gibt es neben den sehr flexiblen, programmierbaren auch die „hardwired“ Prozessoren, bei denen der Algorithmus ebenfalls fest vorgegeben ist (z.T. als Hardware, z.T. als fest eingebranntes Programm, z.B. in einem ROM). Bei solchen Prozessoren sind höchstens noch Parameter des Algorithmus von außen vorgebbar.

Hardwired DSP werden vor allem dort eingesetzt, wo ein Algorithmus auf einem sehr schnellen oder breiten Datenstrom, in der Regel in Echtzeit, arbeiten muß (Video, Audio, Grafik, Filter, Kompression, FFT).

Für die Verarbeitung allgemeiner digitaler Daten verwendet man die sogenannten **Datenprozessoren** (Mikroprozessoren, μP). Sie sind immer programmierbar und daher sehr flexibel. Man verwendet gelegentlich auch den Ausdruck „instruction set processor“ (ISP), da ihre Möglichkeiten vor allem von dem vorgegebenen Befehlssatz abhängen.

Im Folgenden werden hauptsächlich ISP betrachtet.

13.1 Befehlssatz und Maschinensprache

Jeder ISP hat durch die Hardwareimplementierung fest vorgegebene Operationen, die ausgeführt werden können. Jede dieser Operationen entspricht einem oder mehreren Datenworten, die im Programmspeicher abgelegt ist und bei seiner Ausführung das Steuerwerk in den entsprechend Zustand versetzt. Das Steuerwerk verursacht daraufhin die erforderlichen Aktivitäten wie Speicherzugriffe und Steuerung des Datenpfades.

Da solche **Maschinenworte** für Menschen schwer lesbar und auch schwierig zu merken sind, verwendet man eine besser lesbare Kurzschreibweise, sog. **mnemonics**. So könnte z.B. das mnemonic für „Lade den Inhalt von Adresse A“ lauten `LOAD A`.

Beispiel 13.1

Rechenaufgabe $C = A + B$.

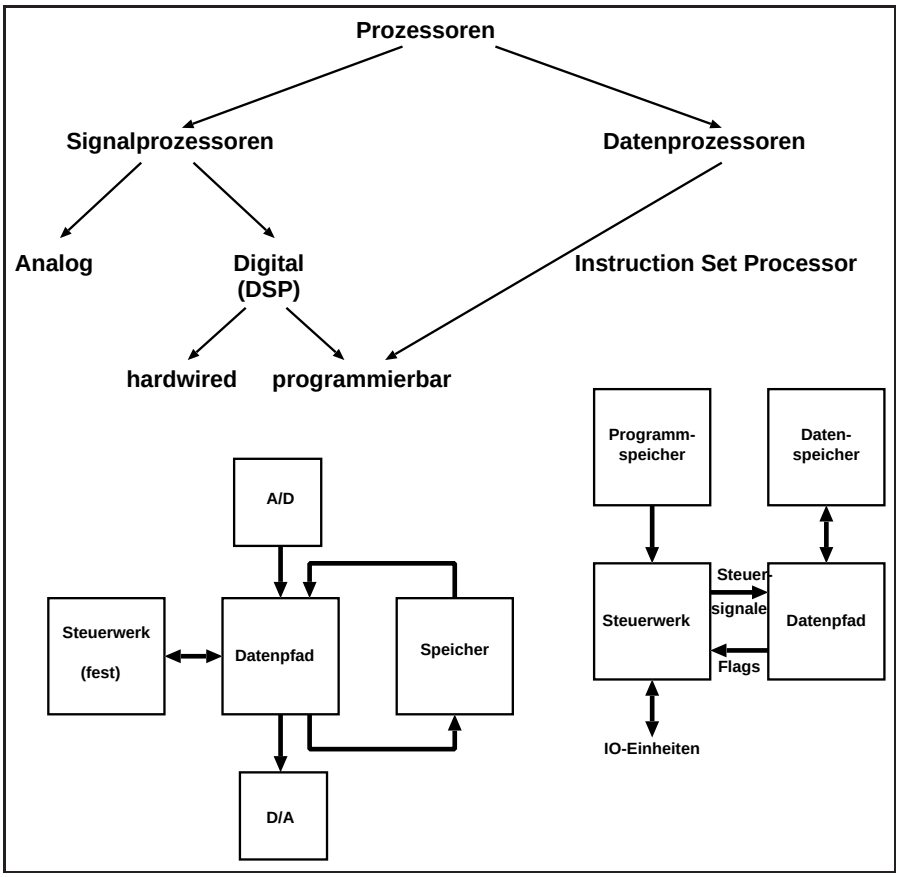


Abbildung 13.1: Übersicht über Prozessoren

```
LOAD A
ADD B
STORE C
```

Hier wird der in Adresse `A` stehende Wert in den Datenpfad geladen, dann wird der in Adresse `B` stehende Wert dort hinzuaddiert. Das Ergebnis der Addition wird dann in Adresse `C` gespeichert.

Um Programmverzweigungen und Sprünge zu ermöglichen, existiert im Prozessor ein spezielles Register, da immer auf den aktuellen Befehl im Programmspeicher zeigt. Sein Name ist **Program Counter (PC)**.

Die Speicherorganisation für das in Bsp. 12.1 beschriebene Programm könnte dann etwa folgendermaßen aussehen:

Ein Maschinenwort besteht, wie eben gesehen, aus einem Befehlsteil (`LOAD`), dem **OP-Code**, und einem (oder mehreren) Operandenteilen (`A` bzw. `B` oder `C`), hier kurz Adresse genannt. Bei einem 16bit breiten Programmspeicher könnte die Aufteilung etwa folgendermaßen sein:

Die in Abb. 12.1 dargestellte Trennung von Programm- und Datenspeicher kann eine tatsächliche, physikalische Trennung sein. In einem solchen Fall befindet sich das Programm meist in einem

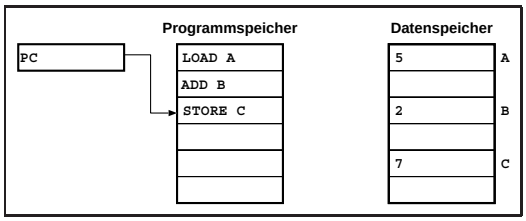


Abbildung 13.2: Speicherorganisation für Beispiel 12.1 (am Ende des Befehls `STORE C`)

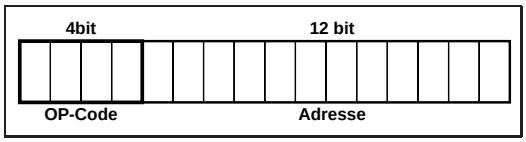


Abbildung 13.3: Aufteilung eines Maschinenwortes

ROM, die Daten in einem RAM (z.B. Steuerung einer Waschmaschine). Für sehr flexible Anwendungsgebiete liegen Programm und Daten in einem gemeinsamen RAM-Speicher (z.B. bei Personal Computern)[1]. Damit Programm und Daten nicht durcheinandergeraten sind sie in einem solchen Fall in unterschiedlichen Speicherbereichen abgelegt. Eine Maschine mit einem solchen logischen Aufbau nennt man auch eine **von Neumann Maschine**[2] (Abb. 12.4).

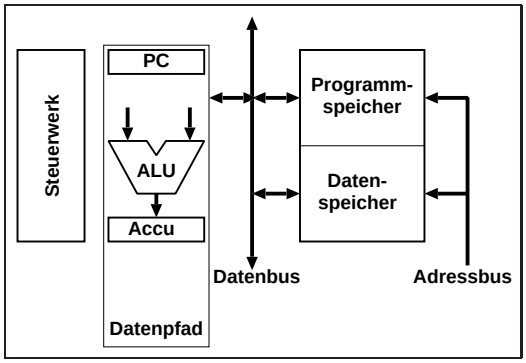


Abbildung 13.4: von Neumann Architektur

Neben dieser von Neumann Architektur existiert noch die Harvard Architektur, die für Programm und Daten getrennte Busse verwendet. Diese besitzt den Vorteil, daß gleichzeitig auf den Programmspeicher und den Datenspeicher zugegriffen werden kann. Wegen dieses Geschwindigkeitsvorteils wird sie häufig für DSP verwendet.

Der Befehlssatz eines Prozessors enthält **Datentransferbefehle** (`LOAD, STORE, ...`), **arithmetisch-logische Befehle** (`ADD, SUB, AND, SHIFT, COMP, ...`) sowie **Kontrollflußbefehle** (`JMP, BRANCH,` Funktionsaufrufe, ...). Sprungbefehle können absolut (`JMP`) sein, d.h. es wird auf die im Befehlswort angegebene Adresse gesprungen, oder bedingt (`BRANCH`), wobei als

[1] Damit so etwas funktioniert, muß beim Einschalten das Programm natürlich aus irgendeinem Festwertspeicher geladen werden, was bei z.B. beim Personal Computer in der sogenannten Bootstrap-Prozedur erfolgt

[2] nach dem Mathematiker John von Neumann (u.A. Begründer der Spieltheorie)

Anz. Variablen	Name	Beispiel
0	Stack	HP-Taschenrechner
1	Accumulator	Taschenrechner, erste μP
2		zwei Operanden, z.B. $B \leftarrow B+A$
3		zwei Operanden + Zieladresse, z.B. $< X > \leftarrow A+B$
4		wurde bisher nur von einem μP benutzt

Tabelle 13.1: Anzahl von Variablen in Maschinenbefehlen

Sprungbedingung immer der Wert eines Flags dient (z.B. „Z“ — Ergebnis der letzten Berechnung ist 0).

Für die in den Befehlsworten angegebenen Adressen gibt es mehrere Interpretationsarten, die als **Adressierungsarten** bezeichnet werden. Ein weiteres wichtiges Kriterium zur Beurteilung der Flexibilität eines Prozessors ist die Anzahl der Adressen (d.h. Variablen), die in einem Befehlswort stehen können, die auch zu unterschiedlicher Namensgebung führen.

Da man bei mehreren Operanden zu sehr breiten Datenworten kommt, ist man dazu übergegangen, eine Mehrzahl von Registern im Prozessor zu implementieren, auf denen dann (fast) alle arithmetischen Operationen ablaufen. Grundsätzlich hat man dann die Möglichkeit, folgende Arten von arithmetischen Operationen im Prozessor zu implementieren:

Memory-Memory-Operationen: Hierbei erhält man — wie eben erwähnt — sehr breite Datenworte.

Register-Register-Operationen: Dies führt zu einem sehr kompakten OP-Code, da um Größenordnungen weniger Register vorhanden als Speicheradressen möglich sind.

Register-Memory-Operationen: Eine Art Mittelweg zwischen den beiden anderen Möglichkeiten.

Beispiel 13.2

Das Programm nach Beispiel 12.1 würde für einen Prozessor, der nur Register-Register-Operationen zuläßt, etwa folgendermaßen aussehen:

```
LOAD R1, A
LOAD R2, B
ADD R1, R2
STORE R2, C
```

Falls auch Register-Memory-Operationen möglich sind, könnte man das Programm so formulieren:

```
LOAD R1, A
ADD R1, B
STORE R1, C
```

Nun seien noch die vier wichtigsten Arten der **Adressierung** erwähnt, die nicht immer alle in einem Prozessor implementiert sein müssen.

Direkt: effektive Adresse = Daten im Befehlswort

Indirekt: eff. Adr. = Inhalt der Speicheradresse, die im Befehlswort steht. Man schreibt auch [Adrs]. Diese Art der Adressierung wird z.B. für Zeiger verwendet.

Immidiate: eff. Adr. = Inhalt des Program Counter. Man schreibt auch [PC].

Indexed: eff. Adr. = Adresse im Befehlswort +[index register]. Solche Indexregister werden oft beim Auslesen automatisch erhöht. Sie können nur für solche Speicherzugriffe verwendet werden, d.h. man kann sie in der Regel nur mit einem Wert laden und dann vielleicht noch inkrementieren bzw. dekrementieren, nicht aber andere Operationen damit durchführen.

13.2 Befehlsablauf eines Prozessors

Die Ausführung eines Befehls besteht aus folgenden Schritten:

1. Programmzähler (PC) adressiert Speicherzelle
2. Befehl wird aus dem Speicher geholt
3. Befehl wird dekodiert (Aufteilung in Opcode und Operandenadresse)
4. Operandenandresse adressiert Speicherzelle
5. Operand wird aus dem Speicher geholt
6. Operation wird durchgeführt. Die Ergebnisse werden dabei in Registern oder Speicherzellen abgelegt, eventuell werden Flags gesetzt.
7. PC wird inkrementiert

Oft ist es geschickter, den PC schon nach dem 2. Schritt zu inkrementieren. Dann zeigt der Befehl immer auf die nachfolgende Speicheradresse, was besonders günstig für die immediate-Adressierung ist.

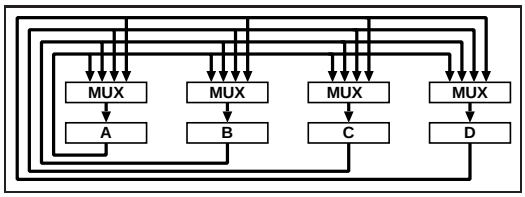


Abbildung 13.5: Registerbusnetzwerk mit paralleler Transfermöglichkeit

13.3 Datenwege

Das Laden von Registern mit Werten anderer Register oder dem Inhalt von Speicherzellen kann — je nach Aufbau des Prozessors — unterschiedlich organisiert sein.

Abb. 12.5 zeigt eine Möglichkeit, bei der jedes Register den Wert jedes anderen Register übernehmen kann. Hierbei können alle diese Transfers gleichzeitig (parallel) erfolgen. Allerdings besitzt hier jede Einheit einen eigenen Bus, was einen sehr großen Aufwand darstellt.

In Abb. 12.6 kann zu jedem Zeitpunkt nur ein Transfer durchgeführt werden. Am Multiplexer besteht immer noch ein sehr großer Verdrahtungsaufwand.

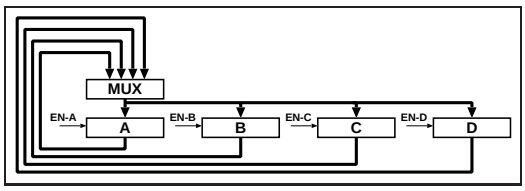


Abbildung 13.6: Registerbusnetzwerk mit Einzeltransfers

Mit Tri-State-Ausgängen an den Registern bietet sich die Möglichkeit, daß alle Register einen einzigen gemeinsamen Bus nutzen (Abb. 12.7). Hier ist jeweils auch nur ein Transfer möglich, jedoch können mehrere Register gleichzeitig mit demselben Wert geladen werden.

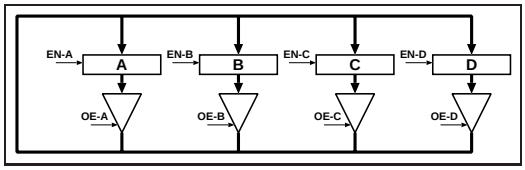


Abbildung 13.7: Register mit gemeinsam benutztem Bus

Selbstverständlich gibt es auch dazwischenliegende Möglichkeiten, die in der Regel in der Realität verwendet werden.

13.4 UKM297 — Entwurf eines Mikroprozessors

Um die Funktionsweise von Mikroprozessoren genauer zu verstehen, wird im folgenden ein (fiktiver) Mikroprozessor entworfen, der den Namen UKM297 („UNSERE KLEINE MASCHINE 2/97“) haben soll[3]. Dabei wird die Architektur nach Abbildung 12.8 vorgegeben.

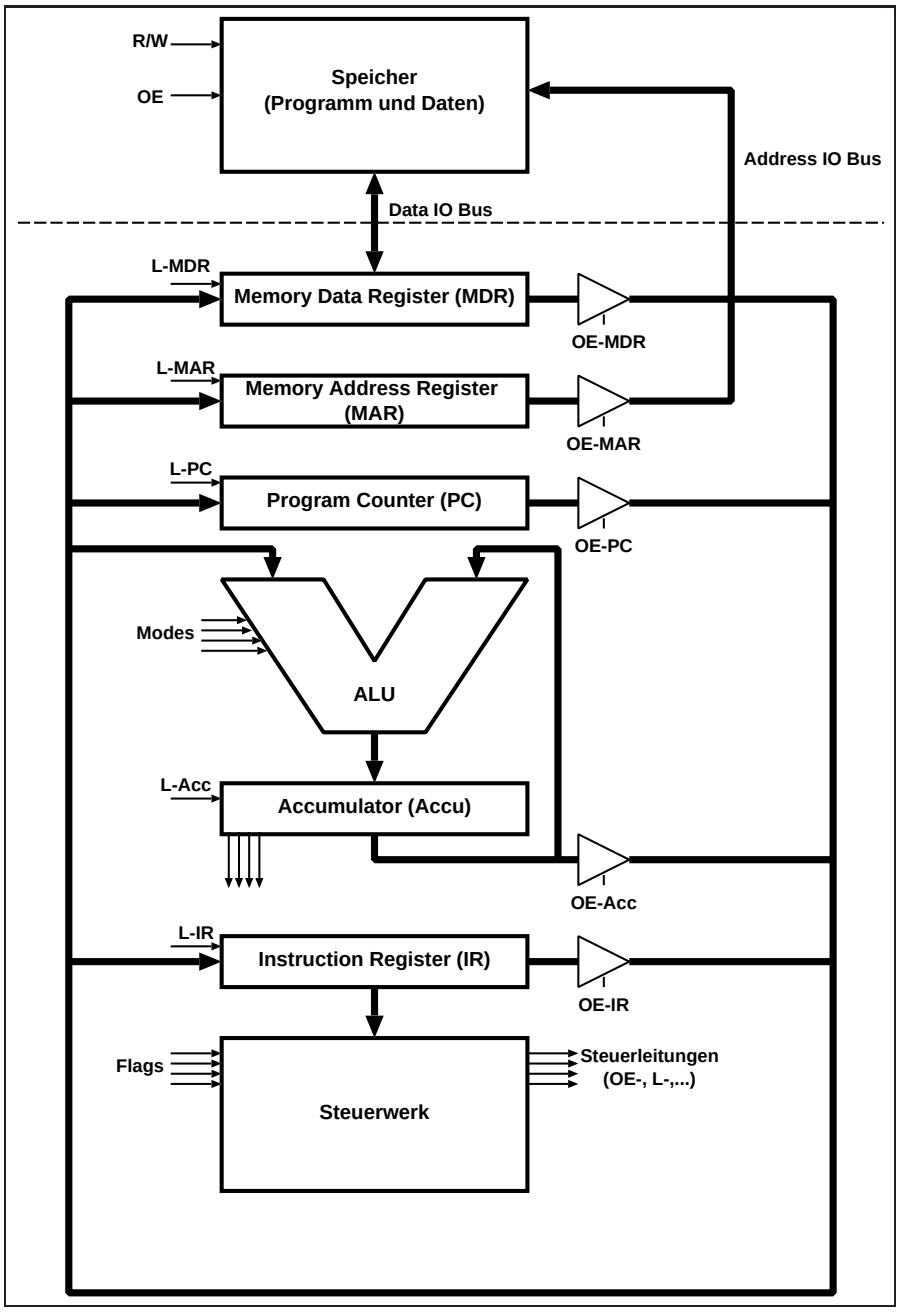


Abbildung 13.8: Architektur des UKM297

[3]Eine leicht veränderte Variante dieses Mikroprozessor wird im Rahmen des Praktikums Mikroelektronik entworfen

13.4.1 Befehlssatz des UKM297

Eigentlich gibt man sich beim Entwurf eines Prozessors erst den Befehlssatz (d.h. man bestimmt vorher, was der Prozessor „können“ soll) und entwirft dann die Architektur.

Für UKM297 ist ein Befehlswort entsprechend Abb. 12.9 aufgeteilt in OPCODE und Operanden.

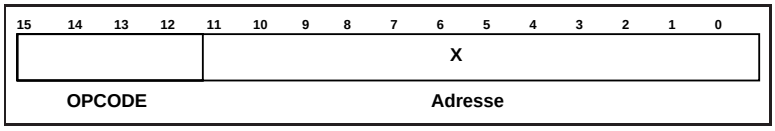


Abbildung 13.9: Aufteilung eines Befehlswortes des UKM297

UKM soll die Befehle entsprechend Tabelle 12.2 „kennen“.

Name	Funktion
LOAD	ACC $\leftarrow$ M[X]
STORE	M[X] $\leftarrow$ ACC
JUMP	PC $\leftarrow$ X
BRN	PC $\leftarrow$ X if ACC<0
COMP	ACC $\leftarrow \overline{\text{ACC}}$
SHR	ACC $\leftarrow$ ACC$/2$ (Shift Right)
ADD	ACC $\leftarrow$ ACC $+$ M[X]
AND	ACC $\leftarrow$ ACC & M[X]

Tabelle 13.2: Befehle des UKM297

Dabei bedeutet M[X] den Inhalt der Speicherzelle X.

Damit ist festgelegt, was der Mikroprozessor alles mit Daten machen können soll. Allerdings steht noch nicht fest, *wie* er das machen soll, d.h. wir die *Struktur* des Prozessors muß noch festgelegt werden.

13.4.2 Detaillierte Beschreibung des Befehlsablaufs

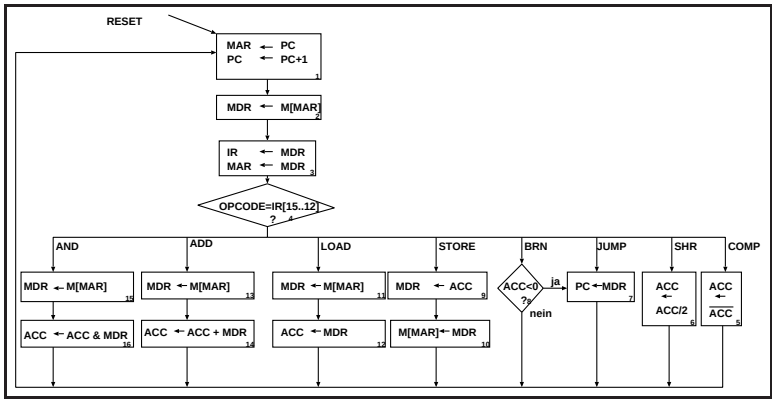


Abbildung 13.10: Befehlsablauf des UKM297

13.4.3 Datenpfad

Für den Datenpfad verwendet man in der Regel die sogenannte **bitslice**-Stuktur, eine sehr regelmäßig aufgebaute Schaltungsstruktur, die den Entwurf weitgehend vereinfacht und eine platzsparende und schnelle Realisierung ermöglicht.

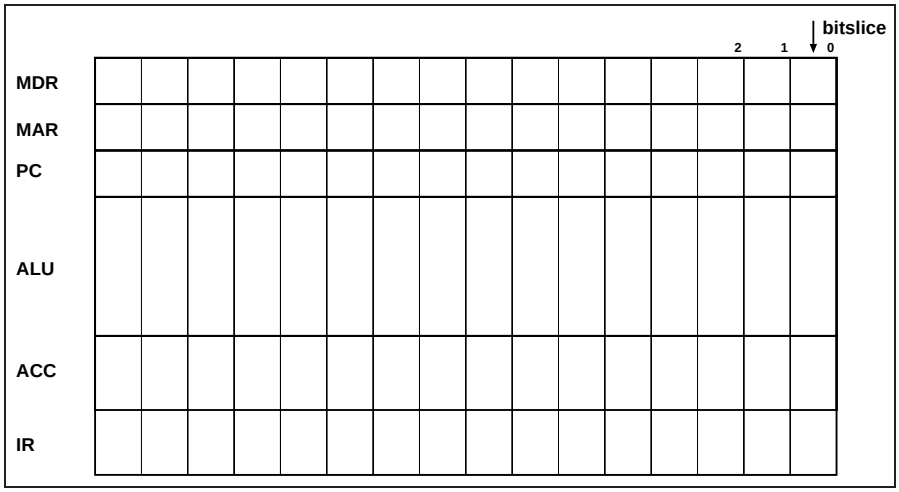


Abbildung 13.11: Anordnung der Register in der Bitslice-Struktur

Die unterschiedliche Höhe der einzelnen Elemente in der Bitslice-Anordnung soll den erwarteten höheren Flächenbedarf wegen größerer Komplexität (besonders bei der ALU) kenntlich machen.

Die Register sollen folgende Längen haben:

- MDR: 16bit

- MAR: Da UKM297 mit 12bit-Adressen arbeitet, muß das MAR mindestens 12 Bit breit sein. Um eine möglichst regelmäßige Struktur zu erhalten kann es auch als 16bit-Register realisiert werden.
- PC entsprechend MAR: 12 oder 16bit.
- ALU: Die ALU muß mindestens so breit sein wie ein Datenwort. Die Datenwortbreite wird hier sinnvollerweise zu 16bit festgelegt, da ein Befehlswort ebenfalls so breit ist und für Daten und Befehle ein gemeinsamer Speicher verwendet wird.
- ACC: 16bit (Datenwortbreite) + 1bit (Carry).
- IR: 16bit.

Bei einer vollkommen homogenen Registerbreite erhält man eine vollkommen regelmäßige Bitslice-Struktur. Es braucht dann nur ein einziges Bitslice entworfen zu werden. Durch nebeneinandersetzen erhält man dann den gesamten Datenpfad. Abb. 12.12 zeigt einen Ausschnitt aus der so erhaltenen Bitslice-Struktur.

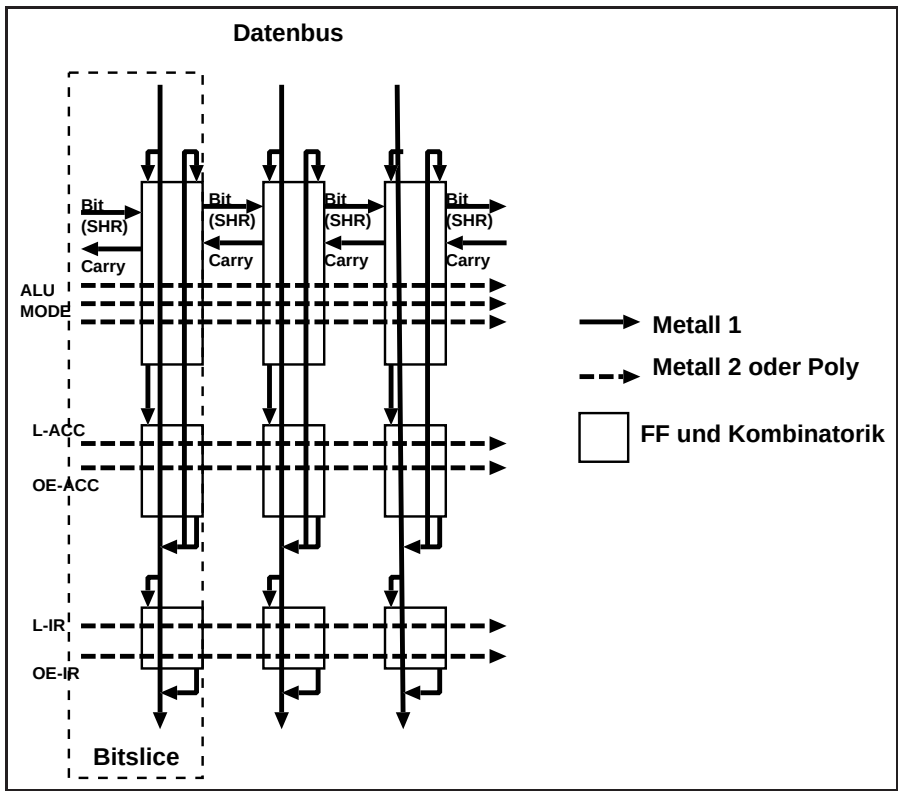


Abbildung 13.12: Ausschnitt aus dem Datenpfad des UKM297

Verbesserungsmöglichkeiten

Das Befehlsregister IR müßte eigentlich nur 4bit breit sein, um den OPCODE aufnehmen zu können. MAR würde eigentlich nur 12bit für die Adresse benötigen. Damit könnten diese beiden Re-

gister zu einem einzigen 16bit-Register zusammengefaßt werden, wodurch jedoch die regelmäßige Struktur des Datenpfades aufgebrochen würde.

Betrachtet man Abb. 12.10 genauer, so kann man erkennen, daß IR nur in einem einzigen Zustand geschrieben wird, in dem die darin stehenden Daten gleichzeitig auch in MDR gespeichert sind. Dadurch kann IR vollständig eingespart werden, wenn bei der OPCODE-Abfrage statt dessen MDR verwendet wird. Im Datenpfad müßten dann die höchstwertigen 4 Bit von MDR zum Steuerwerk herausgeführt werden, was eine kleinere Unregelmäßigkeit des Datenpfades verursachen würde.

13.4.4 Steuerwerk

Das Steuerwerk muß zu jedem Zustand entsprechend Abb. 12.10 die entsprechend erforderlichen Steuersignale (OE-ACC, …, L-IR, …, ALU-MODE) erzeugen. Insgesamt benötigt UKM297 vierzehn Steuersignale.

Eine einfache Implementierung des Steuerwerks wäre damit ein 4bit-Zustandszähler, der die Zustandsabfolge generiert (Zustände $1 \ldots 4$, Übergänge $9 \rightarrow 10, 11 \rightarrow 12, 13 \rightarrow 14, 15 \rightarrow 16$) und parallel geladen werden kann (Zuständsübergänge von 4 zu $5, 6, 7, 8, 9, 11, 13, 15$ und von den Zuständen $5, 6, 7, 8, 10, 12, 14, 16$ zum Zustand 1).

Eine entsprechende Schaltung hat dann folgenden Aufbau:

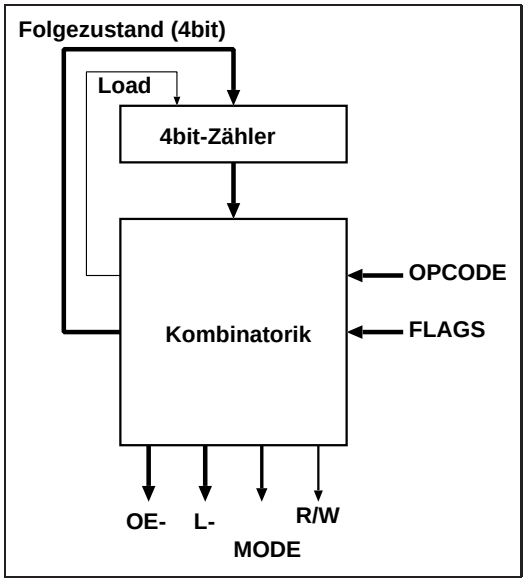


Abbildung 13.13: Steuerwerk mit Zustandszähler

Diese Art des Steuerwerks heißt auch **hardwired**, da nach dem Entwurf der Schaltung keine Änderung mehr vorgenommen werden kann.

Eine weitere Möglichkeit ist ein **mikrokodiertes Steuerwerk**. In einem ROM werden die Werte aller Steuersignale für jeden Zustand abgelegt. Jeder einzelne Zustand entspricht dann einer Adresse, so daß man insgesamt ein 16×14-ROM benötigt. Natürlich ist auch hier ein Zustandszähler und etwas Kombinatorik für die Zustandsabfolge erforderlich. Wenn statt eines ROM ein EEPROM verwendet wird, kann die Programmierung nachträglich geändert werden.

Der im ROM gespeicherte micro code beschreibt hiermit den Ablauf jedes einzelnen Maschinensprachebefehls, so daß man eine Art hierarchischen Aufbau erhält:

- Jedes Maschinenprogramm besteht aus mehreren Maschinenbefehlen.
- Jeder Maschinenbefehl besteht aus mehreren micro code Befehlen (d.h. Zuständen).

Für sehr komplexe Befehle (z.B. Multiplikation) kann unter diese Hierarchie noch eine weitere Stufe gelegt werde, so daß man auch einen nano code erhält, aus dem dann ein micro code Befehl aufgebaut sein kann. Dies erfordert dann zusätzlich einen nano code Zähler und nano code Kombinatorik, was wiederum erheblichen Aufwand verursacht.

Diese Überlegungen haben letztlich vor ca. 15 Jahren zur Entwicklung des RISC-Konzeptes geführt.

13.5 RISC und CISC

Die eben behandelte Architektur des UKM297 besitzt die Merkmale der CISC-Architektur (CISC: complex instruction set computer):

- Maschinenbefehle können sehr komplex sein und setzen sich aus micro code Operationen zusammen, die u.U. aus nano code Operationen bestehen.
- Maschinenbefehle benötigen eine unterschiedliche Anzahl von Taktzyklen (z.B. ist eine Multiplikation in der Regel aus vielen Additionen zusammengesetzt).

Der komplizierte Aufbau eines entsprechenden Steuerwerks ist sehr aufwendig. Beim RISC-Konzept (RISC: reduced instruction set computer) könnte man (als Übertreibung) sagen, daß ein Prozessor nur micro code Befehle kennt, die seine Maschinensprache darstellen.

Das RISC-Konzept besitzt folgende Eigenschaften[4]:

- Nur einfache Maschinenbefehle.

[4] Diese Eigenschaften sind heute z.T. extrem aufgeweicht. Von „reduced instruction set“ kann bei modernen RISC-Prozessoren eigentlich nicht mehr geredet werden.

- Alle Maschinenbefehle benötigen möglichst die gleiche Anzahl von Taktzyklen. In der Regel ist ein Befehl aufgebaut aus (ca.) 3 Teilen: fetch&decode (Hole und dekodiere den Befehl, FD), operand fetch&execute (Hole Operanden und führe aus, OE), memory access&write (Speicheradressierung und -zugriff, MW).
- Durch diesen gleichartigen Aufbau lassen sich mehrere Befehle versetzt gegeneinander quasi-gleichzeitig ausführen (Pipelining, Abb. 12.15). Dabei wird, während ein Befehl ausgeführt wird (OE) bereits der nächste dekodiert (FD) und im nächsten Zyklus, während der erste Befehl einen Speicherzugriff tätigt, bereits ausgeführt.

Durch Pipelining kann dann eine komplexe Berechnung, die bei RISC aus wesentlich mehr Befehlen besteht als bei CISC, in der Regel dennoch schneller ausgeführt werden.

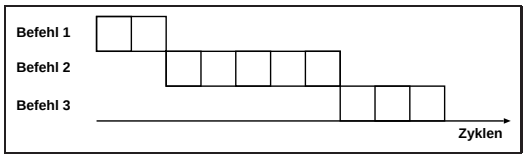


Abbildung 13.14: Abfolge von CISC-Befehlen

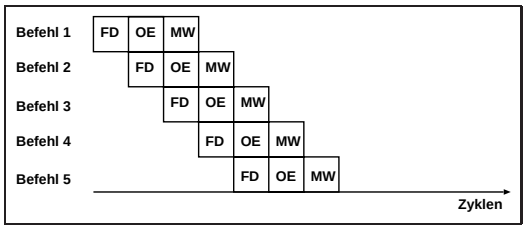


Abbildung 13.15: Pipelining bei einem RISC-Prozessor

Das Verschachteln von Befehlen ist so einfach jedoch nicht immer möglich. Oft benötigt ein Befehl das Ergebnis eines anderen, so daß Wartezustände eingelegt werden müssen, die den Befehlsablauf insgesamt verlangsamen. Moderne Compiler können solche Abhängigkeiten feststellen und entsprechend zwischen diese beiden voneinander abhängenden Befehle andere einschieben, so daß der Ablauf mit optimaler Auslastung geschehen kann. Dieses „Umsortieren“ von Befehlen heißt in der Fachsprache **instruction reordering**.

Problematisch bleiben dann noch bedingte Sprünge. Hier ist es so, daß ein Befehl ein Ergebnis liefert, wobei der Zustand der Flags geändert wird, der wiederum als Sprungbedingung für einen anderen Befehl dient. Hier muß also auch der zweite Befehl auf das Ende des ersten warten. Nun kann nicht jeder beliebige andere Befehl dazwischengeschoben werden, da dieser auch den Zustand der Flags ändern kann und damit einen falschen Sprung auslösen könnte. Sprungbefehle bei RISC-Prozessoren berechnen darum heutzutage erst beide möglichen Sprungadressen und entscheiden möglichst spät, welche verwendet wird. Oft verwenden sie auch eine Art Statistik, welche Sprung am wahrscheinlichsten ist („wird öfter die erste oder zweite mögliche Adresse angesprungen?“) und berechnen dann diese zuerst, so daß hier eventuell früher der Sprung erfolgen kann (**branch prediction**).

Moderne RISC-Prozessoren (und auch manche „moderne“ CISC-Prozessoren, die das RISC-Konzept aufgegriffen haben) besitzen mehrere Pipelines, so daß mehrere Befehle gleichzeitig bearbeitet werden können (wobei in jeder Pipeline selbst wiederum mehrere Befehle verschachtelt bearbeitet werden). Bei diesen werden dann oft beide Möglichkeiten einer Verzweigung gleichzeitig in unterschiedlichen Pipelinies ausgeführt und dann (sobald die Entscheidung gefällt werden kann, weil die Bedingung berechnet wurde) nur der korrekte Teil weiterbearbeitet und der andere verworfen.

Heutige RISC-Prozessoren besitzen ebenfalls sehr komplexe Befehlssätze mit Befehlen wie Multiplikation und Division. Außerdem haben sie oft mehrere Integer-, Memory- und Floating-Point-Einheiten, in denen die Befehle auch von der Hardware selbst umsortiert werden können (**out-of-order-execution**); sie besitzen sehr viele Register (da Registerzugriffe on-chip wesentlich schneller erfolgen können als Speicherzugriffe) und on-chip cache-Speicher für einen schnelleren Speicherzugriff, meist sogar verschiedene für Programm und Daten. In einem solchen cache-Speicher wird ein Teil des externen Speichers abgebildet (d.h. am Anfang einmal geladen), so daß ein größerer Teil des Programmes im program cache (und der Daten im data cache) während der Ausführung keine Buszugriffe sondern nur on-chip-Zugriffe erfordert. Natürlich müssen Maßnahmen getroffen werden, um die Daten im cache mit den Daten im externen Speicher konsistent zu halten, was einen weiteren Hardwareaufwand zur Folge hat. Solche RISC-Prozessoren sind in der Regel sehr groß (über 1cm Chipbreite), können dafür aber sehr schnell getaktet werden (bis ca. 400MHz beim DEC-Alpha).

Abbildungsverzeichnis

Tabellenverzeichnis

Index