1 Einleitung

Im Rückblick auf die letzten 50 Jahre wird deutlich, welch rasanten Fortschritt die Mikroelektronik seit ihrer Entstehung gemacht hat. Als ihre Entwicklung im Dezember 1947 mit der Erfindung des ersten Transistors durch Willliam B. Shockley, John Bardeen und Walter H. Brattain beginnt, war der unaufhaltsame Siegeszug der Festkörperelektronik noch nicht absehbar. Mit der Silizium-Planartechnologie und der ersten verfügbaren monolithisch integrierten Schaltung im Jahre 1961 beginnt der unaufhaltsame Trend, die Funktionalität und Komplexität mikroelektronischer Schaltungen stetig zu erhöhen. 1965 ist von Gordon E. Moore ein exponentielles Anwachsen der Integrationsdichte und eine damit einhergehende Reduzierung der Kosten pro Bauelement erkannt und in einen gesetzmäßigen Zusammenhang gebracht worden [Moe65]. Das Moore'sche Gesetz [Moe75] prognostiziert eine Verdoppelung der Integrationsdichte im 18-Monatsrhythmus, was einer Vervierfachung alle drei Jahre entspricht. In Bild 1.1 ist die tatsächliche Entwicklung der Mikroelektronik anhand von Speicherbausteinen und Mikroprozessoren seit 1970 im Vergleich zur Prognose von Moores Gesetz dargestellt. Aufgrund ihrer regelmäßigen Strukturen und der weniger komplexen Verdrahtung ergibt sich für die Speicher eine wesentlich höhere Integrationsdichte als bei den Mikroprozessoren.

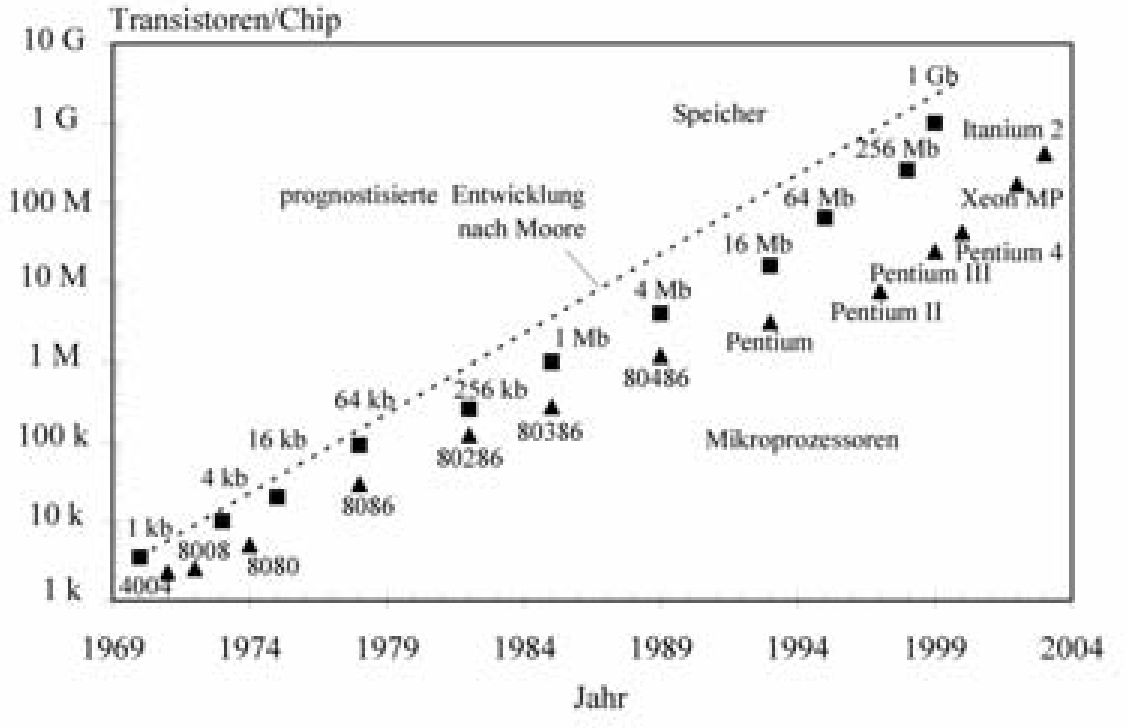

Bild 1.1: Entwicklung von Mikroprozessoren und Speichern im Vergleich zur prognostizierten Entwicklung nach dem Moore'schen Gesetz

Eine wesentliche Voraussetzung zur Fertigung immer komplexer werdender mikroelektronischer Schaltungen ist die ständige Verkleinerung derer minimalen Strukturabmessungen.

Nach der ITRS (International Technology Roadmap for Semiconductors) [SIA03] von 2003 soll sich die Entwicklung gemäß dem Moore'schen Gesetz zumindest bis zum Jahr 2018 unverändert fortsetzen. Dabei ist geplant, etwa alle drei Jahre eine neue Schaltungsgeneration einzuführen. Bild 1.2 verdeutlicht die geplante Entwicklung von minimalen Strukturgrößen und Anzahl von Transistoren bei Mikroprozessoren und Speicherbausteinen. Die minimale Strukturgröße wird von der ITRS über den DRAM half Pich definiert. Ausgehend von einer minimalen Strukturweite von 100 nm im Jahr 2003 sollen diese bis zum Jahr 2018 auf 18 nm gesenkt werden, wobei die Gatelänge der MOS-Transistoren in Mikroprozessoren sogar auf 10 nm reduziert werden soll. In dieser Zeitspanne soll sich die Anzahl der Transistoren in Mikroprozessoren von 180 Millionen auf 9,8 Milliarden erhöhen. Die Kapazität von Speicherbausteinen steigt während dessen von 1 Gb auf 32 Gb.

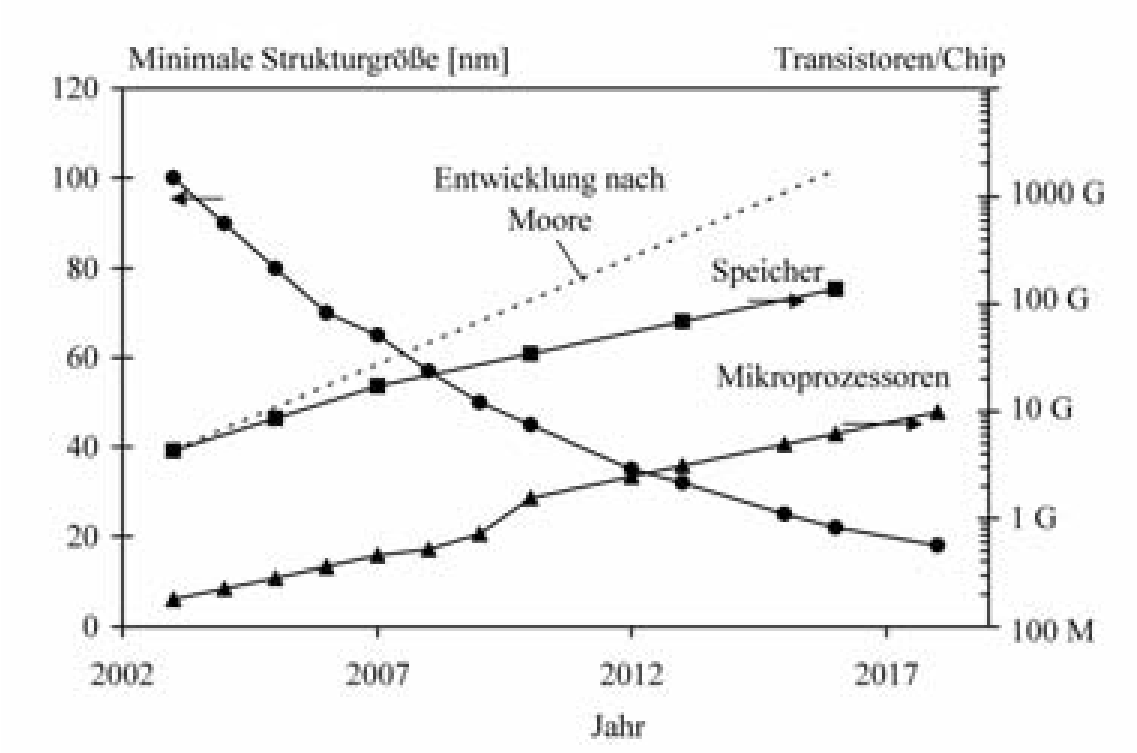

Bild 1.2: Nach der ITRS [SIA03] prognostizierte Entwicklung der minimalen Strukturgrößen und der Komplexität von Speichern und Mikroprozessoren.

Parallel zu der Verkleinerung der Transistorgeometrien soll auch die Betriebsspannung von derzeit 1,2 V auf 0,7 V (bzw. von 1 V auf 0,5 V für Anwendungen mit geringer Leistung) im Jahre 2018 gesenkt werden. Die Schwellenspannung der MOS-Transistoren muss zusammen mit der Betriebsspannung reduziert werden, um weiterhin ein brauchbares Schaltverhalten zu haben. Um die Transistorkapazitäten für hohe Taktfrequenzen schnell umladen zu können, wird im Sättigungsbetrieb ein möglichst hoher Drainstrom benötigt. Dieser hängt maßgeblich von der effektiven Gatespannung (U_{GS}-U_{Th}) ab, weshalb die Betriebsspannung mindestens den 2,3 fachen Wert der Schwellenspannung betragen sollte. Bild 1.3 zeigt die geplante Entwicklung der Betriebs- und Schwellenspannungen in Abhängigkeit der minimalen Strukturgrößen.

Die reduzierte Schwellenspannung führt aufgrund der konstanten Unterschwellenspannungsneigung (engl.: subthreshold slope) S unweigerlich zu einem erhöhten Leckstrom der Transistoren. Nach der ITRS ist ein Anstieg des maximalen, auf die Kanalweite normierten Sperrstroms von 0,03 µA/µm im Jahre 2003 auf 0,5 µA/µm vorgesehen.

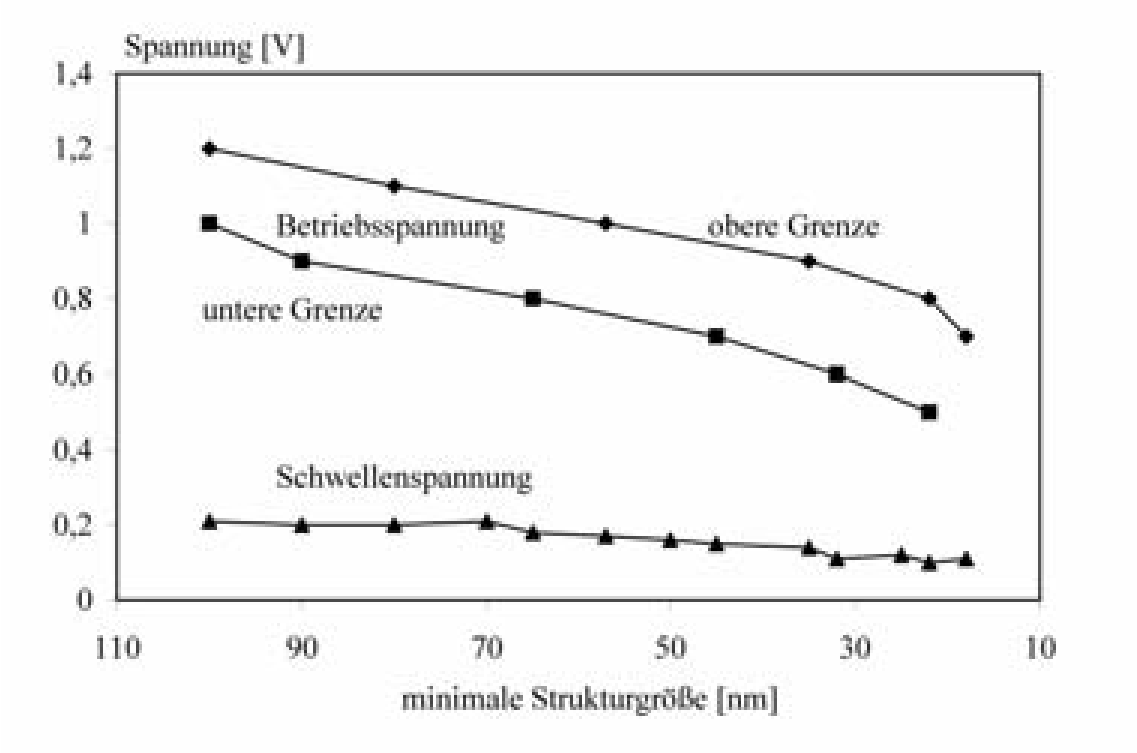

Bild 1.3: Obere und untere Grenze der Betriebsspannung sowie die geplante Schwellenspannung für MOS-Transistoren in Abhängigkeit der minimalen Strukturgröße

Die maximal zulässige Schwellenspannungsstreuung σU_{Th} muss an die geringeren Betriebs- und Schwellenspannungen angepasst werden. Einer Absenkung der Schwellenspannungsstreuung steht jedoch entgegen, dass mit abnehmender Kanalfläche die statistischen Fluktuationen der elektrischen Transistorparameter zunehmen. Diese werden durch die unvermeidbaren Schwankungen der Kanaldotierung hervorgerufen, welche nach [Mik93] und [Mik96] eine untere Grenze für die praktisch einsetzbare Kanallänge darstellen. Im Gegensatz zu älteren Versionen werden in der aktuellen ITRS keine Aussagen über die Entwicklung der Schwellenspannungsstreuung mehr getroffen.

Mit der ITRS besteht zwar eine klare Zielvorgabe bezüglich der weiter fortschreitenden Strukturverkleinerung, jedoch gibt es für die Verfahren zur Durchführung der letzten Phase unterhalb 45 nm minimaler Strukturgröße bisher allenfalls Ansätze, für deren Umsetzung noch erheblicher Forschungsaufwand betrieben werden muss. Dabei liegt das Hauptaugenmerk auf den Lithografieverfahren, mit denen die Strukturen auf der Oberfläche einer Halbleiterscheibe definiert werden. Das klassische Verfahren der optischen Lithografie soll für die nächsten zwei Technologiegenerationen weiterhin als Standard für die industrielle Serienfertigung integrierter Schaltungen dienen, bevor es durch neue, zum Teil auch innovative Verfahren abgelöst werden muss. Im zweiten Kapitel dieser Arbeit wird zunächst ein Überblick über die derzeit verwendeten optischen Lithografieverfahren mit den wichtigsten Maßnahmen

zur Auflösungsverbesserung gegeben, bevor einige entscheidende Lithografieverfahren der nächsten Generation vorgestellt und bewertet werden.

Im Mittelpunkt dieser Arbeit steht eine Prozessführung, mit der minimale Strukturweiten unterhalb von 30 nm bei ausschließlicher Verwendung einer konventionellen, optischen Lithografie reproduzierbar hergestellt werden können. Dieses Verfahren basiert auf einer modifizierten Kanten-Rückätztechnik [Fla83,Hsu92,Ari93], bei der die Linienbreite durch konforme Schichtabscheidung und anisotropes Zurückätzen homogen und reproduzierbar im Nanometerbereich definiert werden kann. In einer vorangegangenen Arbeit an der Universität Dortmund [Hor99] wurden mit Hilfe dieses Depositions- und Rückätzverfahrens MOS-Transistoren mit einer minimalen Kanallänge von L = 30 nm erfolgreich hergestellt und bezüglich ihrer elektrischen Eigenschaften untersucht.

Im Rahmen dieser Arbeit ist die Prozessführung um eine Aktivgebietsstrukturierung im Sub-100 nm-Bereich erweitert worden, so dass – zusammen mit dem bisherigen Verfahren zur Strukturierung der Gateelektrode – Transistoren gefertigt werden können, deren Kanallänge und Kanalweite im Sub-100 nm-Bereich liegen. Die Aktivgebiete werden mit Hilfe einer Technik zur lokalen Feldoxidation erzeugt, deren Oxidationsbarriere mit einer zweiten Deposition- und Rückätztechnik strukturiert wird. Unter Berücksichtigung einer möglichen Unteroxidation der Nitridbarriere können Aktivgebiete mit einer Weite im Sub-100 nm-Bereich reproduzierbar hergestellt werden.

Neben den Herausforderungen bezüglich der Auflösungserweiterung der optischen Lithografie beziehungsweise der Entwicklung völlig neuer Lithografieverfahren, stellen die physikalischen Eigenschaften von Bauelementen mit Abmessungen im Nanometerbereich eine Grenze für die weitere Strukturverkleinerung dar. Neben der minimal erreichbaren Kanallänge spielen die statistischen Schwankungen der elektrischen Parameter einzelner Bauelemente eine entscheidende Rolle für zukünftige Schaltungsgenerationen.

Mit dem erweiterten Depositions- und Rückätzverfahren können mit einer Homogenität von ± 1,5 % bei der Strukturierung ausreichend viele Bauelemente auf einem Substrat erzeugt werden, so dass neben der statischen Charakterisierung von MOS-Transistoren mit einer Länge von bis zu L = 30 nm und einer Weite bis W = 80 nm auch statistische Untersuchungen durchgeführt werden können.

2 MOS-Transistoren mit kleinen Abmessungen

Durch die stetige Strukturverkleinerung gemäß dem Moore'schen Gesetz können – mit dem damit verbundenen geringeren Flächenbedarf – die Herstellungskosten pro Bauelement deutlich gesenkt werden, auch wenn die Kosten pro Siliziumscheibe durch die fortgeschrittenen Produktionstechniken leicht steigen. Neben diesem wirtschaftlichen Vorteil verbessern sich mit der Strukturverkleinerung auch die elektrischen Eigenschaften der integrierten Schaltungen. Diese können bei erhöhter Schaltgeschwindigkeit zunehmend komplexer gestaltet werden, wodurch immer neue Anwendungsgebiete ermöglicht werden. Grundlage für die ständige Verkleinerung der Transistorgeometrien bildet die Theorie der „ähnlichen Verkleinerung" (engl. scaling theory) von MOS-Transistoren.

2.1 Das Prinzip der Skalierung – Scaling

Im Jahre 1974, etwa neun Jahre nachdem Gordon E. Moore sein Gesetz über die exponentiell anwachsenden Integrationsdichte aufgestellt hat, wurden von Robert H. Dennard et al. [Den74] erstmalig grundlegende Regeln zur Strukturverkleinerung von MOS-Transistoren veröffentlicht. Die „Constant Electric Field Scaling Theory" (CE) besagt, dass durch ein Verkleinern sämtlicher geometrischer Abmessungen und ein Absenken der Versorgungsspannung um einen Faktor $1/\alpha$ (mit $\alpha > 1$), sowie eine Anhebung der Kanaldotierung um diesen Faktor α (Bild 2.1), die elektrische Feldstärke und die Verlustleistungsdichte im Transistorkanal konstant gehalten werden können.

Die Betriebsspannungen integrierter Schaltungen konnten zum einen wegen der notwendigen Kompatibilität zu standardisierten Spannungspegeln, zum anderen wegen Schwellenspannungsstreuungen und den nicht skalierbaren Potentialdifferenzen an den pn-Übergängen nicht im gleichen Maße wie die Transistorgeometrien reduziert werden. Deshalb wurde die CE-Theorie etwa zehn Jahre nach ihrer Veröffentlichung durch neue Scaling-Regeln ergänzt. Bei dem „Constant Voltage Scaling" (CV) nach [Chi83] wird im Gegensatz zu den ursprünglichen Scaling-Regeln die Versorgungsspannung konstant gehalten, so dass sich elektrische Feldstärke und Verlustleistung um den Faktor α erhöhen. Mit dem „Quasi-Constant Voltage Scaling" (QCV) [Bac84] wurde mit dem Skalierungsfaktor $1/\alpha^{1/2}$ ein Kompromiss zwischen einer strikten Beibehaltung der Betriebsspannung und der ursprünglichen starken Skalierung mit dem Faktor α vorgestellt.

Mit der Einführung eines weiteren frei wählbaren Parameters κ ($\kappa > 1$) stellt die „Generalized Scaling Theory" (GS) von Baccarani und Dennard et al. [Bac84] eine universelle Scaling-Regel zur Verfügung, bei der die Betriebsspannung mit $1/\kappa$ unabhängig von den Transistorgeometrien ($1/\alpha$) und der Dotierung (α) skaliert werden kann. In Tabelle 2.1 sind die Auswirkungen der beiden Faktoren der GS-Theorie auf die Skalierung der einzelnen Transistorparameter dargestellt. Je nach Anforderung an eine neue Schaltungsgeneration können die Transistoren auf die Verlustleistungsdichte, Verzögerungszeit oder Steilheit hin optimiert werden.

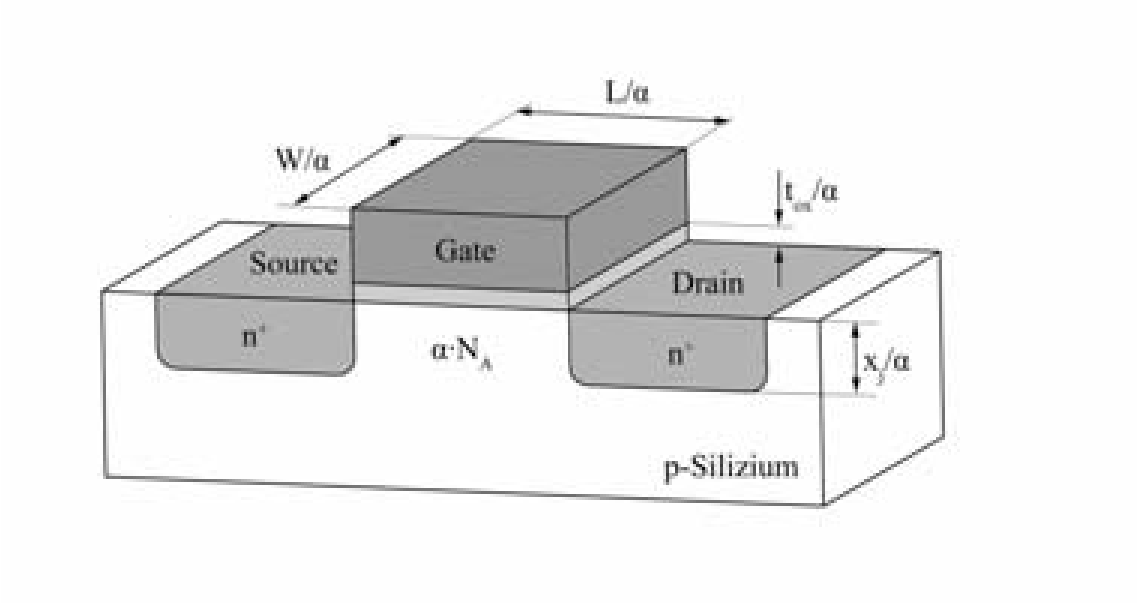

Bild 2.1: Scaling eines NMOS-Transistors

Tabelle 2.1: Skalierungsregeln nach der Generalized Scaling Theory für die Geometrien und Kenndaten von MOS-Transistoren

Parameter	Formelzeichen	Skalierungsfaktor für die GS-Theorie
Transistorabmessungen	W, L, t_{ox}, x_j	$1/\alpha$
Flächen	A	$1/\alpha^2$
normierte Kapazität	C_{ox}', C_j'	α
Kapazitäten	$A \cdot C_{ox}$, $A \cdot C_j'$	$1/\alpha$
Spannungen	U_B, U_{DS}, U_{GS}	$1/\kappa$
Drainstrom	I_D	α/κ
elektrische Feldstärke	E	α/κ
Kanaldotierung	N_A, N_D	α^2/κ
Verlustleistungsdichte	$\frac{U_{DS} \cdot I_D}{A}$	α^3/κ^3
Verlustleistung	$U_{DS} \cdot I_D$	α/κ^3
Gate Verzögerungszeit	$\frac{C_G \cdot U_{DS}}{I_D}$	κ/α^2
Zeitkonstante	$R_j \cdot C_j'$	1
mit $\kappa = \alpha$ ergeben sich die Skalierungsfaktoren für das CE-Scaling-Konzept		

Mit der weiteren Strukturverkleinerung unterhalb von 100 nm sind die vorgestellten Scaling-Konzepte durch auftretende physikalische und technologische Grenzen nur bedingt anwendbar. Zum einen schränkt der Tunneleffekt die minimale Gateoxiddicke t_{ox} auf 3 nm ein, so dass diese nicht mehr zusammen mit den weiteren Transistorgeometrien skaliert werden kann, zum anderen steht ein konsequentes Einhalten der Scaling-Regeln im Widerspruch zur Minimierung der auftretenden Kurzkanaleffekte. Für den Sub-100 nm-Bereich wurde daher eine neue Scaling-Theorie [Fie94], beziehungsweise eine überarbeitete GS-Theorie [Maa95], entworfen.

Während sich in den nächsten zehn Jahren nach den Vorstellungen der ITRS die minimalen Strukturweiten etwa um den Faktor 3,3 verkleinern, soll die Versorgungsspannung im selben Zeitraum nur um den Faktor 1,3 bis 1,8 reduziert werden. Die Schwellenspannung wird in diesem Zeitraum zwischen dem hp90 und hp22 Technolgy Node um die Hälfte reduziert. Die Betriebsspannung liegt damit weiterhin mindestens über dem 2,3fachen der Schwellenspannung.

Die Gateoxiddicke soll nach Planung der ITRS ebenfalls weiter abgesenkt werden, wobei die zunehmenden Tunnelströme unterhalb von 3 nm und die damit zusätzlich entstehende Verlustleistung toleriert werden können. Nach [Koh01] wurden MOS-Transistoren mit minimalen Gateoxiddicken von t_{ox} = 1,2 nm gefertigt, die aufgrund der Gateleckströme Fluktuationen in der Schwellenspannung zeigen, woraus eine untere Grenze für die Gateoxiddicke bei 0,8 nm gesehen wird. Für die nächsten Technologiegenerationen ist verstärkt der Einsatz alternativer Gatedielektrika vorgesehen [Tog02, Chn02]. Obwohl die physikalischen Schichtdicken dieser Materialien deutlich größer sind, können zu Siliziumoxid äquivalente Schichtdicken von bis zu einem Nanometer realisiert werden.

Zur Reduzierung der bei den geringen Betriebsspannung kritischen Schwellenspannungsstreuungen sollen in Zukunft auch neue Materialien für die Gateelektroden eingesetzt werden. Mit der Verwendung so genannter Mid-Bandgap-Materialien [Vie02], deren Austrittsarbeit in der Mitte der Siliziumbandlücke liegt, können im Vergleich zu dotiertem Polysilizium gleiche Schwellenspannungen bei einer geringeren Kanaldotierung erreicht werden. Zudem kann mit der Verwendung einer Mid-Bandgap-Gateelektrode aus Metall eine erhöhte Leitfähigkeit erzielt werden, wobei die n- und p-Dotierung einer Polysiliziumgateelektrode im CMOS-Prozess entfällt.

2.2 Kurzkanaleffekte

Die so genannten Kurzkanaleffekte stellen einen begrenzenden Faktor bezüglich der, mit der Strukturverkleinerung einhergehenden, Verbesserung der elektrischen Eigenschaften von integrierten Schaltungen dar. Diese Effekte unterliegen im allgemeinen nicht den Scaling-Regeln, so dass ihre Auswirkungen im Verhältnis zu den skalierten Abmessungen, Spannungen und Dotierungen deutlich zunehmen. Sie können die Funktion der einzelnen Bauelemente – und damit auch einer gesamten Schaltung – stören und zur Degeneration oder zum Durchbruch der Transistoren führen. Gegenmaßnahmen zur Unterdrückung dieser unerwünschten Effekte stehen dabei teilweise im Widerspruch zu den zuvor betrachteten Scaling-Regeln.

2.2.1 Kanallängenmodulation

Nach dem einfachen MOS-Transistormodell von Sah [Sah64] gilt für die Berechnung des Drainstroms im Anlauf- beziehungsweise Widerstandsbereich ($|U_{GS}-U_{Th}| > |U_{DS}|$) folgende Gleichung:

$$I_D = \frac{W}{L} \cdot \mu \cdot C_{ox} \cdot \left[\left(U_{GS} - U_{Th} \right) \cdot U_{DS} - \frac{1}{2} \cdot U_{DS}^2 \right] \tag{2.1}$$

Für den Sättigungsbereich ($|U_{GS}-U_{Th}| \leq |U_{DS}|$) gilt:

$$I_D = \frac{1}{2} \cdot \frac{W}{L} \cdot \mu \cdot C_{ox} \cdot \left(U_{GS} - U_{Th} \right)^2 \tag{2.2}$$

Nach diesem einfachen Modell fließt für eine Gate-Source-Spannung U_{GS} unterhalb der Schwellenspannung U_{Th} kein Strom durch den Transistor. Erst mit dem Überschreiten der Schwellenspannung bildet sich an der Kanaloberfläche eine Inversionsschicht, so dass der Transistorkanal leitfähig wird (Bild 2.2, links). Im Anlaufstrombereich hängt der durch den Transistor fließende Strom gemäß Gleichung 2.1 sowohl von der Gate-Source-Spannung, als auch von der Spannung zwischen Drain und Source ab.

Mit Erhöhung der angelegten Drain-Source-Spannung U_{DS} vergrößert sich zwangsläufig die Potentialdifferenz entlang des Transistorkanals, wodurch die für die Kanalinversion verantwortliche Spannungsdifferenz zwischen Gateelektrode und Kanal vom Source zum Drain kontinuierlich abnimmt. Mit dem Erreichen der Drain-Source-Sättigungsspannung U_{DSS} – der Drain-Source-Spannung, bei der am drainseitigen Kanalende die effektive Gate-Kanal-Spannung den Wert der Schwellenspannung U_{Th} annimmt – geht der Transistorarbeitspunkt in den Sättigungsbereich über. Wie in der Mitte von Bild 2.2 dargestellt ist, wird der Kanal zum Drain hin abgeschnürt.

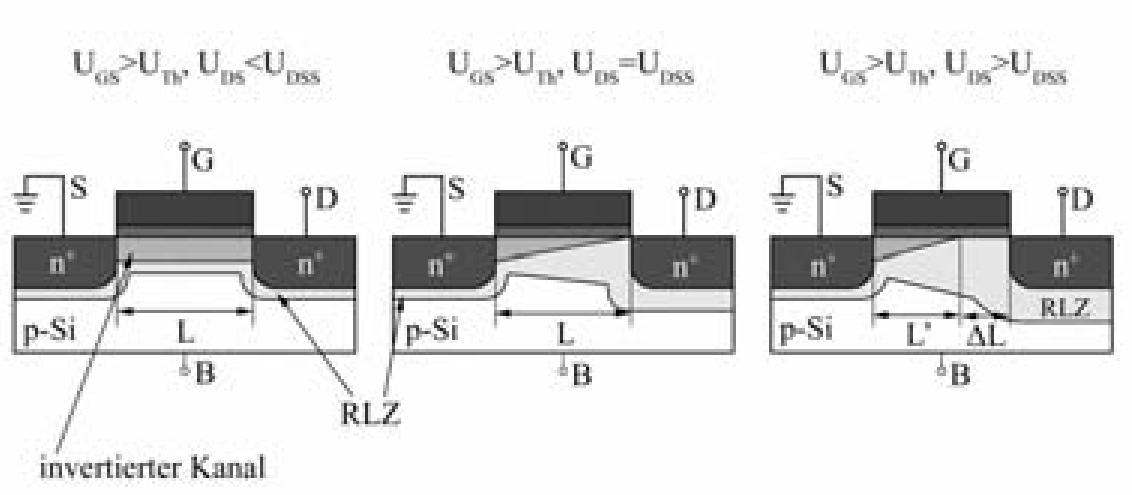

Bild 2.2: Querschnitt durch einen NMOS-Transistor im Anlaufstrombereich (links), im Übergang zum Sättigungsbereich (Mitte) und mit abgeschnürtem Kanal im Sättigungsbereich (rechts)

Nach Gleichung 2.2 bleibt der Drainstrom im Sättigungsbereich mit zunehmender Drain-Source-Spannung konstant, so dass er nur von der Gate-Source-Spannung abhängig ist. Wie in Bild 2.3 skizziert ist, weisen die Äste des Ausgangskennlinienfeldes vor allem bei Transistoren mit kurzen Kanallängen einen deutlichen Stromanstieg im Sättigungsbereich auf. Dieser Stromanstieg kann durch die mit dem Übergang vom Anlaufstrombereich in den Sättigungsbereich auftretende Kanalabschnürung erklärt werden, die sich mit zunehmender Drain-Source-Spannung weiter vergrößert. Die Raumladungszone (RLZ) des Draingebiets weitet sich dabei in die Kanalregion aus (Bild 2.2, rechts), so dass sich die aktive Kanallänge L' um den abgeschnürten Bereich (engl. Pich-Off-Bereich) ΔL verkürzt (L' = L-ΔL).

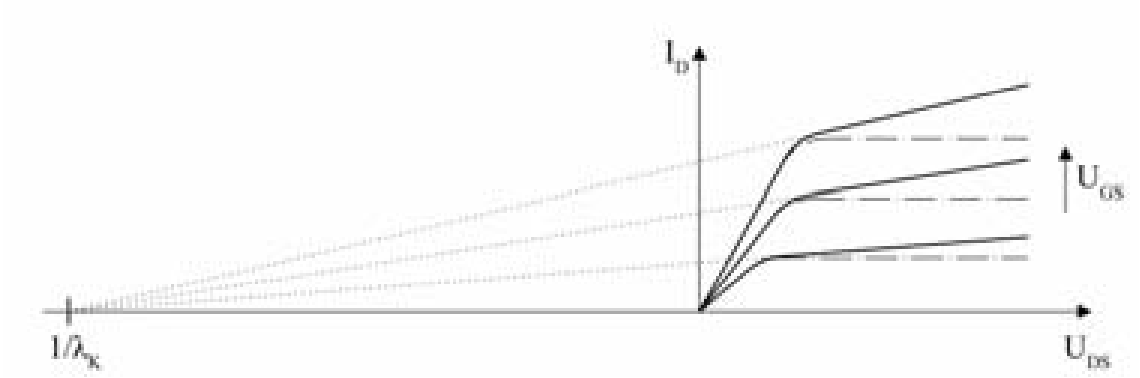

Bild 2.3: Deutlicher Anstieg des Drainstroms im Sättigungsbereich durch die Kanallängenmodulation. Im Vergleich ist der Verlauf nach Gleichung 2.2 als Strich-Punkt-Linie dargestellt

Um den als Kanallängenmodulation bekannten Effekt der ansteigenden Kennlinien berücksichtigen zu können, führt das verbesserte Shichmann-Hodges-Transistormodell [Shi68] einen Anpassungsfaktor λ_K in die Gleichungen zur Abschätzung des Drainstroms ein. Für den Anlaufstrombereich gilt demnach:

$$I_D = \frac{W}{L} \cdot \mu \cdot C_{ox} \cdot \left[\left(U_{GS} - U_{Th} \right) \cdot U_{DS} - \frac{1}{2} \cdot U_{DS}^2 \right] \cdot \left(1 + \lambda_K \cdot U_{DS} \right) \qquad (2.3)$$

Mit dem Übergang in den Sättigungsbereich ($|U_{GS}-U_{Th}| \leq |U_{DS}|$) wird nun eine Abhängigkeit des Drainstroms von der Drain-Source-Spannung berücksichtigt:

$$I_D = \frac{1}{2} \cdot \frac{W}{L} \cdot \mu \cdot C_{ox} \cdot \left(U_{GS} - U_{Th} \right)^2 \cdot \left(1 + \lambda_K \cdot U_{DS} \right) \qquad (2.4)$$

Durch Verlängerung der Kennlinien im Sättigungsbereich kann der Faktor $1/\lambda_K$ durch den Schnittpunkt mit der Spannungsachse aus dem Ausgangskennlinienfeld grafisch ermittelt werden (Bild 2.3).

Der Pinch-Off-Bereich ΔL entspricht der Ausdehnung der Raumladungszone vom Drain in den Kanalbereich des Transistors und kann allgemeingültig mit folgender Gleichung abgeschätzt werden: