Digitale Schaltungstechniken für Sub-100 nm-CMOS-Technologien

DISSERTATION

zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften
(Dr.-Ing.)

der Technischen Fakultät
der Christian-Albrechts-Universität zu Kiel

Klaus von Arnim

Kiel, Mai 2006

1. Gutachter	Prof. Dr.-Ing. Peter Seegebrecht
2. Gutachter	Prof. Dr. Helmut Föll
3. Gutachter	Prof. Dr.-Ing. Horst Fiedler
Datum der mündlichen Prüfung	20. Juli 2006

Vorwort

Die vorliegende Arbeit entstand im Rahmen einer Anstellung als Doktorand bei der Firma Infineon Technologies im Bereich Corporate Research (CPR) in München. Viele Menschen haben zum Gelingen dieser Arbeit beigetragen, nur einigen kann ich an dieser Stelle danken.

Sehr herzlich möchte ich mich bei meinem Doktorvater Prof. Dr.-Ing. Peter Seegebrecht vom Lehrstuhl für Halbleitertechnik an der Christian-Albrechts-Universität zu Kiel für die Betreuung meiner Arbeit bedanken, sowie bei Prof. Dr. Helmut Föll vom Lehrstuhl für Materialwissenschaft und Prof. Dr.-Ing. Horst Fiedler von der Universität Dortmund für die Übernahme der Korreferate.

Mein großer Dank gebührt auch Dr. Roland Thewes, in dessen Gruppe ich stets vielfältige Unterstützung und eine inspirierende Arbeitsumgebung fand. Insbesondere bei Dr. Christian Pacha bedanke ich mich herzlich für dreieinhalb Jahre Zusammenarbeit während Diplom- und Doktorarbeit, viele hundert Stunden anregender Diskussion zu jeder Tages- und Nachtzeit sowie für die Unterstützung bei Vorträgen und Veröffentlichungen.

Darüber hinaus fand ich stets wertvolle Unterstützung bei Dr. Jörg Berthold, bei Dr. Klaus Schrüfer und dem übrigen Multi-Gate-Team bei Infineon und Texas Instruments.

Meinen geschätzten Bürokollegen Dr. Martin Jenkner und Dr. Alexander Frey wünsche ich auch in Zukunft eine so gute multidisziplinäre Büroatmosphäre und immer einen freien Blick auf die Alpen. Auch allen anderen Kolleginnen und Kollegen an dieser Stelle herzlichen Dank für fachliche und moralische Unterstützung und eine schöne Zeit bei der CPR.

Mein großer Dank gilt ebenso allen Diplomanden und Werkstudenten für den großen Einsatz und die wertvolle Arbeit: Wolfgang Penth, Martin Bach, Eduardo Borinski, Florian Bauer und Ramachandhran Balasubramanian.

Nicht zuletzt danke ich meiner Freundin Heike für ihre besondere Unterstützung und das Verständnis für die zeitintensive Erstellung dieser Arbeit, meinen Eltern, die mich nicht nur während meines Studiums immer unterstützten und natürlich meiner Schwester Henrike, die mich ins ferne München entließ.

Kiel, im Juli 2006

Inhaltsverzeichnis

1 Einleitung **1**

2 Rahmenbedingungen für Sub-100 nm-Schaltungsdesign **5**

3 Leistungsaufnahme und Schaltgeschwindigkeit digitaler CMOS-Schaltungen **17**
3.1 Leckströme und aktive Leistungsaufnahme 19
3.1.1 Unterschwellenstrom und Gate-Induced Drain Leakage 19
3.1.2 Gate-Tunnelstrom in Bulk- und Multi-Gate-Transistoren 30
3.1.3 Aktive Leistungsaufnahme 39
3.2 Schaltgeschwindigkeit digitaler CMOS-Schaltungen 41
3.2.1 Inverter-Verzögerungszeit als Funktion des On-Stroms 42
3.2.2 Transistor-Dimensionierung in CMOS-Schaltungen 46
3.2.3 Stack-Effekte in NAND- und NOR-Gattern 49
3.2.4 Einfluss der Temperatur auf die Schaltgeschwindigkeit 53

4 Body Biasing in Sub-100 nm-Technologien **57**
4.1 Leckströme und Reverse-Biasing 58
4.2 Schaltgeschwindigkeit und Forward-Biasing 62
4.3 Praktische Anwendbarkeit 66

5 Neue Ansätze für Sub-100 nm-Schaltungstechnik **68**
5.1 Skewed-CMOS-Logik 68
5.1.1 Funktionsweise 71
5.1.2 Möglichkeiten zur Reduzierung des Leckstroms 72
5.1.3 Störsicherheit in einer System-on-Chip-Umgebung 77
5.1.4 Dimensionierung und Geschwindigkeit 79
5.2 Statische Multi-V_t-Multi-t_{ox}-Logik 82
5.3 Flip-Flops und Latches 86
5.3.1 Optimiertes Sense-Amplifier-Flip-Flop 89
5.3.2 Flip-Flops mit Zustandserhaltung im Standby-Modus 94
5.3.3 Sense-Amplifier-Flip-Flop für Skewed-CMOS-Logik 96
5.3.4 Skewed-CMOS-Latch 98

6 Low-Power-Schaltungstechniken am Beispiel von 32-bit-Addierern 100
6.1 Parallel-Prefix-Algorithmus 100
6.2 Implementierung von vier 32-bit-Addierern 106
6.3 Leckstrom und Geschwindigkeit 110

7 Zusammenfassung 115

A Messergebnisse der Crosstalk-Teststruktur 118

B Ringoszillatoren in statischer Multi-V_t-Logik 124

C Transistordimensionierungen 129

Formelzeichen und Abkürzungen 131

Veröffentlichungen 134

Patentanmeldungen 135

Abbildungsverzeichnis 136

Tabellenverzeichnis 140

Literaturverzeichnis 141

Kapitel 1

Einleitung

Im Jahre 1943 schätzte Thomas J. Watson, Vorsitzender von IBM, den weltweiten Bedarf für Computer auf eine Zahl von fünf. Heute nutzt nahezu jeder Mensch in modernen Informationsgesellschaften ein vielfache Anzahl in unterschiedlichster Form. Das Anwendungsspektrum reicht von der allgegenwärtig verfügbaren Rechenleistung (*Ubiquitous Computing*) bis hin zu zentralen Systemen: vom Mobiltelefon über Automobilelektronik, tragbare und Desktop-Rechner bis zum Großrechner für die Wettervorhersage. Diese Entwicklung wurde durch die fortschreitende Skalierung der integrierten Schaltungen ermöglicht, die Gordon Moore bereits 1965 beobachtete [1]. Er prognostizierte, dass sich die Dichte der Transistoren alle 18 Monate verdoppeln würde.

Der Prozess der fortschreitenden Skalierung lässt sich zurückblickend in verschiedene Abschnitte unterteilen. Am Anfang blieb die Versorgungsspannung V_{dd} trotz kleiner werdender Abmessungen konstant bei 5 Volt (*Constant Voltage Scaling*). In Sub-Mikrometer-Technologien werden die in den Bauelementen auftretenden Felder so groß, dass auch V_{dd} skaliert wird (*Constant Field Scaling*) [2]. Die ebenfalls sinkende Schwellenspannung V_t kann jedoch nicht im gleichen Maße reduziert werden, ohne dass die Leckströme zunehmen.

Seit der 130 nm-Technologie wird die Gatelänge stärker als die übrigen Transistordimensionen skaliert, damit auch bei reduziertem Gate-Overdrive $V_{dd} - V_t$ noch eine Zunahme der Schaltgeschwindigkeit erzielt wird [3]. Diese Entwicklung, die sich voraussichtlich bis zur 45 nm-Technologie fortsetzen wird, erfolgt abgesehen von einer Nitridierung des Gateoxids weiterhin unter Verwendung von Poly-Silizium-Gates und $\mathrm{SiO_2}$ als Dielektrikum in einem planaren Transistor. Damit verbunden ist jedoch eine starke Zunahme verschiedener Leckstromkomponenten, sodass eine CMOS-Technologie heute mehrere spezialisierte Transistoren für unterschiedliche Anwendungen zur Verfügung stellen muss (Sub-100 nm-Skalierung). Transistoren in 180 nm-Technologien verfügen häufig über zwei, Transistoren in 130 nm-Technologien über bis zu drei Schwellenspannungen. In der 90 nm-Technologie kann der Schaltungsentwickler unter zwei Oxiddicken für Logikschaltungen auswählen, die wiederum mit jeweils bis zu drei Schwellenspannungen kombiniert werden können [4]. Erst in der 32 nm-Technologie erscheint aus heutiger Sicht der Einsatz neuer Gate-Materialien und Dielektrika sowie alternativer Transistorkonzepte sowohl technisch notwendig als auch ökonomisch sinnvoll [3]. Es ist offen, ob sich dabei die konventionelle Bulk-Technologie mit Metall-Gates und High-κ-Dielektrika oder

neue Konzepte wie *Fully-Depleted Silicon-on-Insulator* (FD-SOI) – als planarer oder aufrecht stehender Multi-Gate-Transistor (MuGFET) – durchsetzen wird.

Die verschiedenen Anwendungen, die bei der Spezialisierung der Transistoren in aktuellen CMOS-Technologien abgedeckt werden müssen, lassen sich prinzipiell in drei Klassen unterteilen. Auch wenn im Allgemeinen ein einzelner Transistor niemals genau einer dieser Klassen entspricht, so werden doch die meisten Fälle durch diese Einteilung gut abgedeckt.

In High-Performance-Anwendungen (HP) werden laut *International Technology Roadmap for Semiconductors* (ITRS) [3] Transistoren mit den kürzesten Gatelängen L_g, kleinen Gateoxiddicken t_{ox} und niedrigen Schwellenspannungen V_t eingesetzt, die auch bei hohen Versorgungsspannungen V_{dd} noch zuverlässig funktionieren. Mikroprozessoren stellen hierfür eine typische Anwendung dar. Deren Verlustleistung kann 100 W oder mehr betragen, sodass eine aktive Kühlung unerlässlich wird.

Im Gegensatz dazu werden Low-Operating-Power-Schaltungen (LOP) bei kleineren Spannungen betrieben, da das Absenken von V_{dd} eine sehr effiziente Methode zur Reduzierung der aktiven Leistungsaufnahme darstellt. Wichtig ist hier insbesondere eine gute Prozess-Kontrolle, damit die Schaltungen auch bei kleinem Gate-Overdrive $V_{dd} - V_t$ noch zuverlässig funktionieren. LOP-Anwendungen sind z.B. mobile Geräte mit hohen Anforderungen an die Rechenleistung, aber auch Schaltungen, die mit einfachen passiven Kühlungen oder in günstigen Kunststoffgehäusen eingesetzt werden sollen.

Low-Standby-Power-Anwendungen (LSTP) sind auf geringe Leckströme optimiert. Dies erfordert dickere Gateoxide, längere Gatelängen und höhere Schwellenspannungen. Wegen des höheren V_t kann auch V_{dd} nicht so weit abgesenkt werden wie für LOP-Schaltungen. Die elektrischen Eigenschaften der LSTP-Transistoren skalieren von der 130 nm-Technologie an nur noch sehr langsam. In den nachfolgenden Generationen können hier zwar weiterhin höhere Integrationsdichten, jedoch nur noch leichte Erhöhungen der Schaltgeschwindigkeit erreicht werden. LSTP-Schaltungen werden in mobilen Geräten eingesetzt, die im Ruhezustand wenig Leistung aufnehmen. Dies gilt z.B. für den Einsatz in Mobiltelefonen oder PDAs (*Personal Digital Assistent*).

Die verschiedenen Anwendungsklassen definieren sich in erster Linie über die Anforderungen getakteter digitaler Logikblöcke in integrierten Schaltungen. Daneben sind jedoch große Speicherblöcke, Analog-Komponenten sowie I/O-Schaltungen wesentlicher Bestandteil vieler System-on-Chip-Anwendungen. Diese Schaltungsteile greifen zwar auf die gleichen Transistoren wie die digitalen Logikblöcke zurück, stellen jedoch bezüglich ihrer technologischen und schaltungstechnischen Anforderungen vollkommen eigenständige Schaltungsteile dar und werden daher im Rahmen dieser Arbeit nicht behandelt.

Die Anforderungen in den drei Anwendungsklassen sind sehr verschieden. So können sich die Leckströme von HP- und LSTP-Transistoren um bis zu vier Dekaden voneinander unterscheiden. Aus diesem Grund führt die Optimierung aller Transistor- sowie Schaltungsparameter dazu, dass auch in HP-Schaltungen stets eine Leckstromproblematik besteht und dass LSTP-Anwendungen geschwindigkeitskritisch sind. Entsprechend gilt für LOP-Schaltungen, dass die Versorgungsspannung und die Schwellenspannung nur so weit gesenkt werden können, wie die Schaltgeschwindigkeit ausreichend groß bzw. der Leckstrom klein bleibt.

Idealerweise wird die Schaltgeschwindigkeit eines Transistors für ein bestimmtes Leckstrombudget optimiert. In der Praxis stellt sich jedoch das Problem, die Bedingungen für diese Optimierung festzulegen. Mögliche Fragestellungen können hierbei sein, welche Temperatur die Schaltung im aktiven und im Standby-Betrieb hat oder wie hoch die Versorgungsspannung ist. Außerdem müssen Parameterschwankungen eingeplant werden, deren Auswirkungen auf die Leistungsaufnahme und Schaltgeschwindigkeit berücksichtigt werden müssen [5].
Diese ohnehin hohe Komplexität des Problems erhöht sich weiter, wenn in unterschiedlichen Betriebsmodi zeitlich wechselnde Anforderungen an die Schaltgeschwindigkeit und den Leckstrom gestellt werden. Soll eine Schaltung im aktiven Betrieb schnell arbeiten können, bei geringer Aktivität aber stromsparend sein, kann dieses nicht mehr allein durch Weiterentwicklung und Spezialisierung der Technologie erreicht werden. So meldet sich ein Mobiltelefon im Standby-Betrieb einmal pro Sekunde für kurze Zeit an der Basisstation an; dieses entspricht einigen hundert Schaltvorgängen pro Sekunde. Bei rechenintensiven Funktionen soll hingegen beispielsweise eine Taktrate von 300 MHz verfügbar sein [6], sodass die Aktivität um fast sechs Größenordnungen schwankt. Es wird deutlich, dass technologische Weiterentwicklungen allein nicht zu einer Lösung des Zielkonflikts zwischen Leckstrom und Schaltgeschwindigkeit führen. Vielmehr müssen schaltungstechnische Lösungen gefunden werden, um auch in Zukunft immer kleinere und schnellere integrierte Schaltungen mit geringerer Leistungsaufnahme herstellen zu können.

Zielsetzung dieser Arbeit ist es, aufbauend auf einem grundlegenden Verständnis von Leckströmen und Schaltgeschwindigkeiten, sowohl neue Ansätze zur Verbesserung des Trade-offs zwischen Leckstrom und Schaltgeschwindigkeit zu finden als auch bekannte Schaltungstechniken auf ihre Eignung für Sub-100 nm-CMOS-Technologien zu untersuchen. Insbesondere die Möglichkeiten und Herausforderungen, die aus der Verfügbarkeit unterschiedlicher Schwellenspannungen und Oxiddicken resultieren, werden betrachtet.
Wichtig ist, dass die vorgestellten Schaltungstechniken auch in hochintegrierten System-on-Chip-Umgebungen zum Einsatz kommen können. Die Evaluation erfolgt daher nicht nur anhand von Schaltungssimulationen und Messungen an Einzeltransistoren und Ringoszillatoren, sondern auch anhand eines praxisrelevanten komplexen Schaltungsblocks. Ziel ist ein grundlegendes Verständnis der Übergänge zwischen den verschiedenen Abstraktionsebenen (Transistor–Logikgatter–Schaltung). Dies ermöglicht die Bewertung einer Schaltungstechnik bezogen auf ein Gesamtsystem und damit eine übergreifende Optimierung.

Kapitel 2 zeigt zunächst die Rahmenbedingungen für das Schaltungsdesign in Sub-100 nm-CMOS-Technologien auf. Es werden Schaltungstechniken beschrieben, die heute schon mit dem Ziel eingesetzt werden, wahlweise die Leistungsaufnahme reduzieren oder die Schaltgeschwindigkeit erhöhen zu können.
Anschließend werden in Kapitel 3 die verschiedenen Leckstrommechanismen einerseits sowie die Schaltgeschwindigkeit in digitalen Schaltungen andererseits beschrieben. Effekte wie Gateleckströme, die in Sub-100 nm-Technologien an Relevanz gewinnen, werden in aktuellen

Transistoren gemessen und physikalisch basiert modelliert. Ringoszillatoren, die trotz eines einfachen Aufbaus das digitale Schaltverhalten einer komplexen Schaltung sehr gut abbilden, dienen der Charakterisierung dynamischer Eigenschaften.

Auf dieser Grundlage wird im Kapitel 4 das *Body Biasing* untersucht. Bei dieser Technik wird die Schwellenspannung durch Variation der Bulk-Spannung unter Ausnutzung des Substrateffekts während des Betriebs angepasst. Ausgehend von Einzeltransistoren, über einfache Ringoszillatoren, bis hin zu einem komplexen Schaltungsblock werden dabei die Auswirkungen dieser Schaltungstechnik beschrieben.

In Kapitel 5 werden neue Schaltungstechniken vorgestellt, die dazu beitragen können, den Leckstrom in Sub-100 nm-Schaltungen zu reduzieren und die Schaltgeschwindigkeit zu erhöhen (Skewed-CMOS-Logik und Multi-V_t-Multi-t_{ox}-Logik). Darüber hinaus werden getaktete Speicherelemente (Flip-Flops und Latches) beschrieben, die den Datenfluss in einer Pipeline-Architektur synchronisieren und die einen effizienten Einsatz dieser Schaltungstechniken erst möglich machen.

Die vorgestellten Techniken werden anhand von Demonstratorschaltungen auf ihre Nutzbarkeit in realen Schaltungen überprüft. Ein 32-bit-Parallel-Addierer demonstriert die Funktion eines größeren und für viele Anwendungen geschwindigkeitskritischen Schaltungsblocks (Kapitel 6). Der Einsatz einer verbesserten Mikroarchitektur zeigt, dass für die Entwicklung schneller und verlustleistungsarmer Schaltungen die verwendeten Transistoren, die Schaltungstechniken sowie auch der Schaltungsentwurf aufeinander abgestimmt sein müssen.

Kapitel 2

Rahmenbedingungen für Sub-100 nm-Schaltungsdesign

In den Anfängen der Transistorskalierung konnten Schaltungslayouts noch durch einfaches Verkleinern der Strukturgrößen um einen bestimmten Faktor (*Shrink Factor*) in eine neue Technologie transferiert werden. Schon bald wurden jedoch einzelne Abmessungen schneller skaliert als andere, sodass der Transistor als Ganzes im Layout ersetzt werden musste. In Sub-100 nm-Technologien erfordert der Übergang zu einer neuen Technologie, selbst bei unveränderter Funktionalität der Schaltung, einen vollständig neuen Entwurf. Es müssen neue Technologie-Optionen wie zum Beispiel mehrere unterschiedliche Schwellenspannungen oder Oxiddicken auf einem Chip schaltungstechnisch sinnvoll genutzt werden. Außerdem spielen parasitäre Effekte, die sich aus der Miniaturisierung der Bauelemente ergeben, eine zunehmend größere Rolle. Auch diese Effekte müssen schaltungstechnisch berücksichtigt werden.

Die Entwicklung neuer schaltungstechnischer Maßnahmen kann daher heute nicht mehr sequentiell nach der Technologieentwicklung erfolgen, sondern es müssen parallel dazu Konzepte erdacht und erprobt werden, um eine neue CMOS-Technologie effektiv und schnell nutzen zu können. Insbesondere bietet sich die Chance, von schaltungstechnischer Seite frühzeitig Einfluss auf die Technologieentwicklung nehmen zu können (Co-Design von Schaltungstechnik und Technologie). Alle technologischen wie auch schaltungstechnischen Innovationen können nur dann in einer Schaltung eingesetzt werden, wenn sie in den Entwurfs-Prozess (*Design Flow*) für digitale CMOS-Schaltungen integrierbar sind. Nur für wenige, besonders geschwindigkeitskritische Schaltungsblöcke ist eine Handoptimierung möglich (*Full Custom Design*).

Dieses Kapitel gibt einen Überblick über die Rahmenbedingungen, unter denen die Entwicklung neuer Schaltungstechniken in Sub-100 nm-Technologien erfolgt.

Transistor-Optionen in Sub-100 nm-Technologien

In der ITRS-Roadmap [3] wird regelmäßig eine Prognose veröffentlicht, welche technologischen Maßnahmen erforderlich sind, um auch weiterhin dem Skalierungstrend, hin zu kleineren, schnelleren und günstigeren Schaltungen, folgen zu können. In Tabelle 2.1 sind die Anforderungen an die aktuelle 90 nm-Technologie für die drei Anwendungsszenarien *High Performance*

Technologie-Knoten	130 nm	90 nm	90 nm	90 nm	65 nm	45 nm	32 nm
Anwendung	LSTP	HP	LOP	LSTP	LSTP	LSTP	LSTP
Jahr der Einführung	2001	2004	2004	2004	2007	2010	2013
Physikalische Gatelänge (nm)	90	37	53	65	37	25	18
EOT (nm)	2.4	1.2	1.5	2.1	1.6	1.3	1.1
Versorgungsspannung V_{dd} (V)	1.2	1.2	0.9	1.2	1.1	1.0	0.9
Schwellenspannung $V_{t,sat}$ (V)	N.A.	0.20	0.26	0.50	0.5	0.39	0.34
On-Strom $I_{d,sat}$ (μA/μm)	300	1110	530	440	510	670	880
Gateleckstrom-Dichte (A/cm^2)	$\approx$0	450	1.9	0.005	0.023	0.08	0.15
Intrinsisches Delay τ (ps)	4.61	0.95	1.76	2.77	1.77	0.98	0.6
NAND2-FO3-Delay (ps)	$\approx$116	23.9	44.3	69.7	43.2	24.8	15.1

Tabelle 2.1: Anforderungen an aktuelle und zukünftige CMOS-Technologien nach ITRS von 2001 und 2004 [3, 7]. Dargestellt sind unter anderem die äquivalente Oxiddicke *EOT*, der On-Strom eines NFETs bei $V_{gs} = V_{ds} = V_{dd}$ und 25 °C sowie das intrinsische Delay eines NFETs $\tau = CV_{dd}/I_{d,sat}$.

(HP), *Low-Operating Power* (LOP) und *Low-Standby Power* (LSTP) angegeben. Selbst innerhalb eines Technologie-Knotens unterscheiden sich die Transistoren stark. So ist die Gatelänge eines 90 nm-LSTP-Transistors fast doppelt so groß wie die eines 90 nm-HP-Devices. Die größeren Abmessungen bei LSTP-Anwendungen sind nötig, um die Leckströme und damit die Verlustleistung im inaktiven Modus (*Standby Mode*) klein zu halten. So ist die Gateleckstrom-Dichte hier um fast fünf Dekaden kleiner als bei HP-Transistoren.

Die kleinere Verlustleistung kann jedoch nur auf Kosten einer reduzierten Schaltgeschwindigkeit erzielt werden, die sich im intrinsischen Delay $\tau = CV_{dd}/I_{d,sat}$ (NMOS) oder in der Verzögerungszeit eines NAND2-Gatters mit Fan-out 3 ausdrückt. Die Schaltgeschwindigkeit von LSTP- und HP-Gattern unterscheidet sich um einen Faktor 2.9. Erst mit der 45 nm-Technologie erreichen die LSTP-Schaltungen die Geschwindigkeit der HP-90 nm-Technologie.

Es fällt auf, dass die Versorgungsspannung in Sub-100 nm-Technologiegenerationen bei der Skalierung der Transistordimensionen nur noch langsam reduziert wird. Dies liegt daran, dass die Schwellenspannung aufgrund der endlichen Unterschwellenstromsteilheit auch nur noch langsam abgesenkt werden kann. Erst wenn die Unterschwellenstromsteilheit durch Einführung neuer Transistor-Konzepte wie FD-SOI erhöht wird, kann auch V_t und V_{dd} weiter sinken.

Da dieses frühestens mit der 32 nm-Technologie der Fall sein wird, müssen bis dahin schaltungstechnische Maßnahmen ergriffen werden, um mit den zur Verfügung stehenden Transistoren leckstromarme und zugleich schnelle Schaltungen zu realisieren.

Technologie	130 nm			90 nm			
Transistor	LVT	REG	HVT	LVT	REG	LL-LVT	LL
Phys. Gatelänge (nm)	90	90	90	70	70	90	90
$V_{t,sat}$ NMOS (mV)	290	370	550	270	400	460	550
Oxiddicke t_{ox} (nm)	2.2	2.2	2.2	1.6	1.6	2.2	2.2
Oxiddicke *EOT* (nm)	2.2	2.2	2.2	1.3	1.3	1.9	1.9
Triple-Well-Option	–	–	–	×	×	×	×

Tabelle 2.2: Ausgewählte Parameter von Transistoren der 90 nm- und 130 nm-Technologie [12, 13]. Die 90 nm-NMOS-Transistoren können in einer zusätzlichen Wanne implementiert werden, sodass sie vom Substrat entkoppelt sind (Triple-Well-Option).

Leckstrom und Schaltgeschwindigkeit

Grundsätzlich muss bei der Beurteilung der Schaltgeschwindigkeiten zwischen Mikroprozessor- und System-on-Chip-CMOS-Technologien unterschieden werden. In Mikroprozessor-Technologien [8,9] steht meist nur eine einzige Oxiddicke für Logikschaltungen zur Verfügung. Die Transistoren entsprechen dabei den HP-Anforderungen. Die Leistungsaufnahme aktueller Server-Prozessoren beträgt mehrere 100 W. Die dabei entstehende Wärme muss über aufwändige aktive Kühlsysteme abgeführt werden.

Im Gegensatz dazu weisen Transistoren in System-on-Chip-Technologie-Plattformen [4, 10, 11] erheblich kleinere Verlustleistungen auf. Es stehen häufig mehrere Oxiddicken und Schwellenspannungen zur Verfügung. Die Abmessungen der schnellsten Transistoren mit vergleichsweise hohem Leckstrom entsprechen etwa den LOP-Devices in Tabelle 2.1, können jedoch auch mit hohen Versorgungsspannungen V_{dd} betrieben werden. Es ist zu beobachten, dass V_{dd} in den Jahren nach einer Technologie-Einführung häufig wieder langsam angehoben wird, um die Schaltgeschwindigkeit zu erhöhen. Dieses wird durch erhöhte Zuverlässigkeit, bessere Prozess-Kontrolle und größere Erfahrung mit der Technologie möglich. Viele Untersuchungen in dieser Arbeit evaluieren daher das Schalt- und Leckstromverhalten für 90 nm-Transistoren abweichend von Tabelle 2.1 auch bei Spannungen bis zu $V_{dd} = 1.5\,\text{V}$.

Mit der in [12] beschriebenen System-on-Chip-Technologie stehen hier ebenfalls Logiktransistoren mit unterschiedlichen Oxiddicken zur Verfügung, die je nach Anforderung jeweils mit drei Schwellenspannungen kombiniert werden können. Tabelle 2.2 zeigt die Transistoren der 90 nm- [12] und 130 nm-Technologie [13]. In beiden Technologien stehen Transistoren mit unterschiedlichen Schwellenspannungen zur Verfügung: nominelle Schwellenspannung (REG), niedrige (Low-V_t, LVT) und hohe Schwellenspannung (High-V_t, HVT).

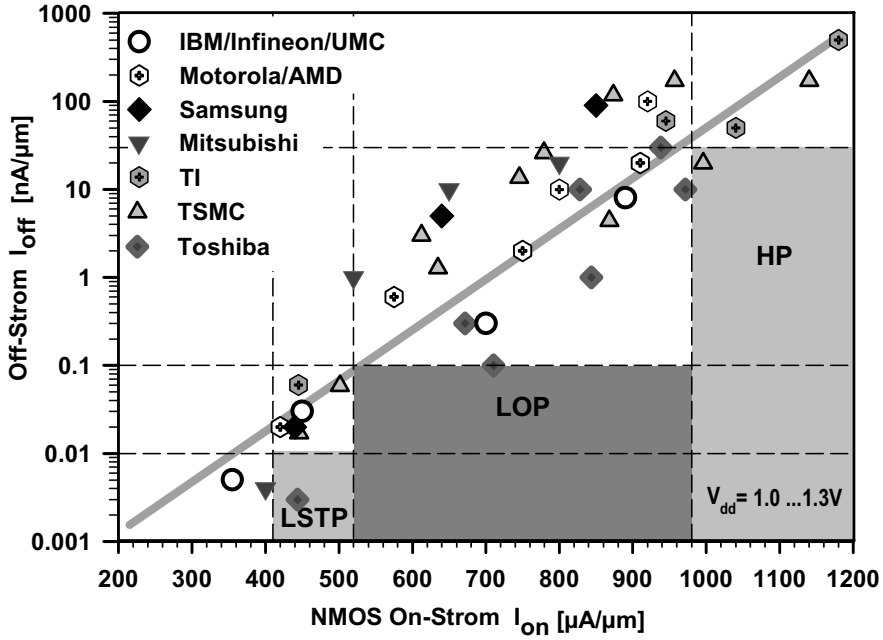


Abbildung 2.1: Leckströme und On-Ströme von NMOS-Transistoren in verschiedenen 90 nm-CMOS-Technologien [4, 8–12, 14].

In Tabelle 2.2 sind sowohl die physikalische Oxiddicke t_{ox} als auch die äquivalente Oxiddicke *EOT* angegeben. *EOT* ist kleiner als t_{ox}, wenn die Dielektrizitätskonstante des Oxids $\varepsilon_{\mathrm{SiO_2}}$ z.B. durch eine Nitridierung erhöht wird:

$$EOT = \frac{\varepsilon_{\mathrm{SiO_2}}}{\varepsilon_{\mathrm{SiON}}} \cdot t_{ox}. \tag{2.1}$$

Für die Funktion eines Transistors ist jedoch die elektrisch wirksame Oxiddicke in Inversion $t_{ox,el}$ ausschlaggebend, die sich aus der Summe der äquivalenten Oxiddicke *EOT*, der Verarmung in der Gate-Elektrode t_{depl} sowie dem mittleren quantenmechanischen Abstand des Kanals von der Oxid-Silizium-Grenzfläche t_{qm} ergibt:

$$t_{ox,el} = EOT + t_{depl} + t_{qm} \approx EOT + [0.8 \ldots 1.0]\,\mathrm{nm} \tag{2.2}$$

Die Transistoren innerhalb einer Technologiegeneration unterliegen immer einem Trade-off zwischen hohem On-Strom $I_{on} = I_{d,sat}(V_{gs} = V_{ds} = V_{dd})$ und kleinem Off-Strom $I_{off} = I_d(V_{gs} = 0, V_{ds} = V_{dd})$. In Abbildung 2.1 ist ein linearer Zusammenhang zwischen On-Strom und logarithmisch aufgetragenem Off-Strom zu erkennen. Die Trendgerade entsteht, da der On-Strom weitgehend linear von $V_{dd} - V_t$ abhängt, während der Off-Strom exponentiell mit sinkender Schwellenspannung ansteigt. In der 90 nm-Technologie können daher keine Transistoren hergestellt werden, die in Abbildung 2.1 deutlich unterhalb der Trendgeraden liegen. Obwohl sich die Leckströme entlang der Trendlinie um über fünf Dekaden und die On-Ströme um einen Faktor 5.4 unterscheiden, stellen sich für jeden Transistor je nach Anwendung die gleichen Probleme. Die Schwellenspannung, die in einem High-Performance-Mikroprozessor eingesetzt

wird, wird so lange reduziert, bis die Leckströme gerade noch eine sichere Funktionalität gewähren und die statische Verlustleistung nicht zu groß wird. Eine mobile LSTP-Anwendung ist auf minimale Leistungsaufnahme im inaktiven Betrieb optimiert, muss aber mit geringer Schaltgeschwindigkeit auskommen. Unabhängig von der Anwendung ist es also das Ziel, den Trade-off zwischen I_{on} und I_{off} auf Transistorebene bzw. den Trade-off zwischen Schaltgeschwindigkeit und passiver Leistungsaufnahme auf Schaltungsebene zu verbessern.

Die dritte kritische Größe ist die Leistungsaufnahme im aktiven Betrieb, die in Sub-100 nm-Technologien bezogen auf eine Flächeneinheit nicht mehr konstant gehalten werden kann. Ein Grund dafür sind die mit der Skalierung größer werdenden parasitären Kapazitäten. Da diese Kapazitäten einen erheblichen Anteil an der Gesamtlast eines Schaltvorgangs ausmachen, kann die während eines Schaltvorgangs umzuladende Kapazität nicht mehr effizient durch Reduzierung der Transistorweiten vermindert werden, ohne die Schaltgeschwindigkeit erheblich zu erhöhen. In einem Flip-Flop einer 90 nm-Technologie vermindern allein die lokalen Verdrahtungskapazitäten die Schaltgeschwindigkeit um bis zu 40 %.

Ein anderer Grund ist der immer größer werdende Einfluss statistischer Parameterschwankungen auf die Worst-Case-Performance einer Schaltung. Der Schaltvorgang eines minimal dimensionierten 90 nm-Transistors erfordert nur noch wenige 1000 Elektronen und die Anzahl der Dotieratome in der Kanalregion ist teilweise kleiner als 100 Atome [15–19]. Die effizienteste Methode zur Reduzierung der aktiven Leistungsaufnahme ist das Absenken der Versorgungsspannung V_{dd}. Um ein starkes Absinken der Schaltgeschwindigkeit zu verhindern, muss jedoch auch die Schwellenspannung reduziert werden. Somit ist auch für LOP-Anwendungen der Trade-off zwischen Leckstrom und Geschwindigkeit entscheidend.

Neben den elektrischen Daten müssen seitens der Technologieentwicklung und des Schaltungsdesigns stets aber auch die Kosten für die Realisierung der Schaltung beachtet werden. Dazu muss zunächst der Flächenbedarf klein sein. Außerdem erhöht jeder zusätzlich in einer Schaltung eingesetzte Transistortyp die Prozesskomplexität. Die Masken- sowie die Prozessierungskosten erhöhen sich und die Ausbeute vermindert sich wegen der erhöhten Anfälligkeit gegen Parameterschwankungen.

CMOS-Skalierung in Sub-45 nm-Technologien

Ebenso wenig wie es sich lohnt, eine technologische Innovation nur für einen einzigen Technologie-Knoten einzuführen, müssen auch neue schaltungstechnische Ideen für mehr als eine CMOS-Generation nutzbar sein, da ansonsten eine Implementierung in den automatisierten Schaltungsentwurf nicht nachhaltig in folgenden Technologieknoten genutzt werden kann. Schon heute müssen neue Konzepte daher auf ihre Eignung für die Sub-45 nm-Technologien überprüft werden.

Wie in Abbildung 2.2 dargestellt, können die Performance-Anforderungen der 45 nm-LSTP-Technologie noch allein mit Hilfe einer erhöhten Ladungsträgerbeweglichkeit im Kanal erreicht werden, z.B. durch Verspannen des Siliziums (*Strained Silicon*) oder Rotation des Substrats. Diese Techniken wirken sich unterschiedlich stark auf die Beweglichkeit von Elektronen

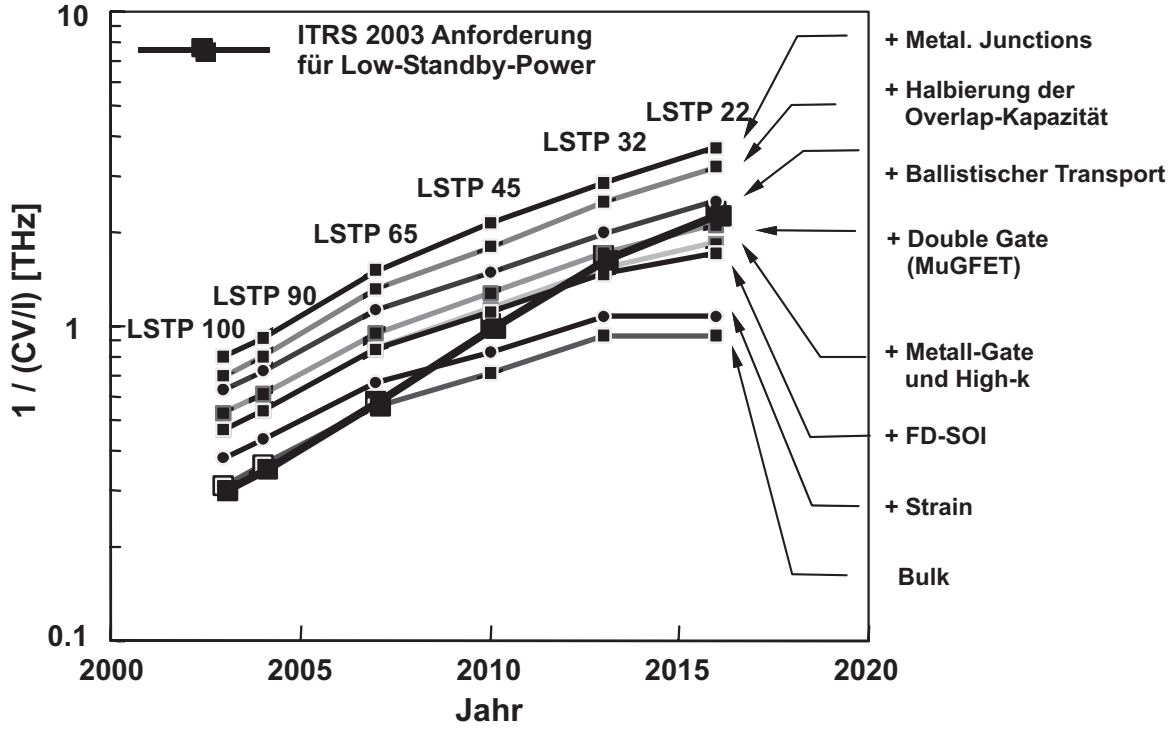


Abbildung 2.2: Intrinsische Geschwindigkeit $1/\tau$ von Low-Standby-Power-Transistoren in zukünftigen CMOS-Technologie-Generationen für verschiedene Technologieoptionen [20]. Im 45 nm-Knoten können die Performance-Anforderungen durch Einsatz von verspanntem Silizium erreicht werden. Mit der 32 nm-Technologie müssen neue Materialien und Transistor-Konzepte eingeführt werden, wie z.B. metallische Gates, High-κ-Dielektrika, Double- oder Multi-Gate-Transistoren (MuGFETs).

und Löchern aus. Dadurch verändert sich das Verhältnis der Schaltströme in N- und PMOS-Transistoren und somit auch das Schaltverhalten von digitalen CMOS-Gattern. Modelle zur Beschreibung der Schaltgeschwindigkeit in Abhängigkeit der Schaltströme von N- und PFET werden im Kapitel 3 aufgezeigt und im Folgenden angewendet.

Ab der 32 nm-Technologie werden neue Konzepte oder neue Materialien benötigt, um in Transistoren mit Gatelängen von weniger als 30 nm alle Leckstromkomponenten so weit zu kontrollieren, dass auch weiterhin in LSTP-Transistoren ein Off-Strom von weniger als 100 pA/μm erreicht werden kann.

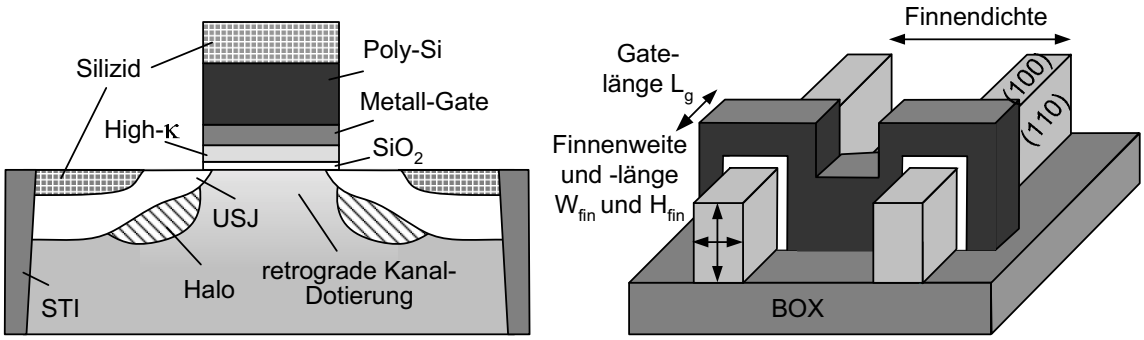
Eine planare 32 nm-CMOS-Technologie erfordert den Einsatz von metallischen Gate-Elektroden und High-κ-Dielektrika, die sich derzeit in der Entwicklung befinden. Die Verwendung einer metallischen Gate-Elektrode verhindert die weiter zunehmende Verarmung des Gates (*Poly Depletion*), während eine höhere Dielektrizitätskonstante eine Reduzierung der elektrischen Oxiddicke $t_{ox,el}$ erlaubt, ohne den Gateleckstrom zu erhöhen. Obwohl sich die elektrischen Transistor-Eigenschaften nicht grundlegend ändern, muss das Schaltungsdesign doch auf die Einführung dieser Material-Innovationen vorbereitet sein. So erfordert das Schaltungsdesign ein physikalisches Verständnis sowie eine genaue Modellierung der verschiedenen Leckstromkomponenten, insbesondere des Tunnelstroms durch das jeweilige Dielektrikum.

Darüber hinaus beeinflussen viele unterschiedliche Faktoren, die sich aus dem in Abbildung 2.3a dargestellten komplexen Aufbau von planaren Sub-100 nm-Transistoren ergeben, immer mehr das elektrische Verhalten. So ist die Schwellenspannung abhängig von der Gatelänge und der Gateweite, und der Serienwiderstand hängt von Kontaktlöchern, Silizid und den Junctions ab (*Ultra-Shallow Junction*, USJ). Der komplexe Aufbau des Gate-Stacks, bestehend aus einer dünnen Siliziumoxid-Schicht, einem Dielektrikum mit einer höheren Dielektrizitätskonstanten und einer Metall-Elektrode sowie den darin enthaltenen Ladungen und Grenzflächenzustände, erschweren die zuverlässige Realisierung solcher Transistoren.

Aus heutiger Sicht erscheint daher der alternative Einsatz senkrecht zur Silizium-Oberfläche orientierter, so genannter Multi-Gate-Transistoren (MuGFETs) als eine aussichtsreiche Möglichkeit zur Herstellung leckstromarmer 32 nm-Transistoren (Abb. 2.3b). Die Beweglichkeit der Ladungsträger ändert sich in MuGFETs nicht nur aufgrund von beabsichtigten Verspannungen, sondern auch wegen der veränderten Topologien und des Einsatzes anderer Materialien. So weist zum Beispiel jede Kristallorientierung und Kanalausrichtung unterschiedliche Beweglichkeiten für Elektronen und Löcher auf [21].

In MuGFET-Technologien kann der Schaltungsdesigner die Gateweite eines einzelnen Transistors nicht mehr frei bestimmen, sondern muss sich an die diskrete Weitenabstufung halten, die durch die effektive Weite einer einzelnen Finne vorgegeben ist. Dieser Effekt hat jedoch nur einen geringen Einfluss auf die Schaltgeschwindigkeit klassischer digitaler CMOS-Schaltungen [22]. Dennoch muss dieser Effekt für neue Schaltungstechniken, die auch in MuGFET-Technologien effizient genutzt werden sollen, beachtet werden.

In einer Serienschaltung aus Bulk-Transistoren, die an denselben Bulk-Kontakt angeschlossen sind, reduziert sich die Schaltgeschwindigkeit aufgrund des Auto-Reverse-Biasing-Effekts (Kapitel 3.2.3). Da MuGFETs keinen Bulk-Kontakt besitzen, tritt dieser Effekt hier nicht auf.



(a) Planarer Transistor **(b) Multi-Gate-Transistor**

Abbildung 2.3: Planarer und Multi-Gate-Transistor für 32 nm-CMOS-Technologien. In planaren Transistoren müssen zunehmend mehr Materialien wie High-κ-Dielektrika und metallische Gates sowie Implantationen, z.B. flache Kontakte (*Ultra-Shallow Junctions*, USJ) und Halo-Dotierung integriert und optimiert werden. Aufgrund der besseren elektrostatischen Kontrolle können MuGFETs auch ohne viele dieser Optionen skaliert werden.

Daraus resultiert eine Verbesserung der Schaltgeschwindigkeit von CMOS-Logik-Gattern mit Serienschaltungen im Vergleich zu Bulk-Technologien (Kapitel 5.1).
Wegen der geringen Finnenweite gehen die Raumladungszonen der beiden Kanalregionen in MuGFETs direkt ineinander über. Es existiert kein neutraler Bereich und die Finne ist vollkommen verarmt (*Fully Depleted*), wodurch die Kurzkanal-Eigenschaften des Transistors verbessert werden. Die parasitäre Gate-Kapazität wird reduziert, da die Raumladungszone beim Schalten des Transistors ihre Größe nicht verändert.
Der veränderte Potential-Verlauf im MuGFET im Vergleich zum Bulk-Transistor wirkt sich daneben auch auf das Tunnelstromverhalten aus. Der Gatestrom in MuGFETs wird im Kapitel 3.1.2 beschrieben.
Abbildung 2.4 zeigt eine Übersicht über die Besonderheiten im Schaltungsdesign mit MuGFETs.

Betriebsbedingung digitaler CMOS-Schaltungen

Temperatureffekte stellen für das CMOS-Schaltungsdesign stets eine besondere Herausforderung dar, weil nahezu jeder elektrische Parameter eine Temperaturabhängigkeit besitzt, die Temperatur aber direkt beeinflusst werden kann.
Obwohl die Verlustleistungsdichte von Mikroprozessoren die einer Herdplatte bereits um mehrere Größenordnungen übersteigt, müssen diese High-Performance-Schaltungen nur für einen begrenzten Temperaturbereich spezifiziert werden, da sie in einer wohl definierten, aktiv gekühlten Umgebung eingesetzt werden. Im Gegensatz dazu müssen System-on-Chip-Applikationen häufig über einen sehr viel größeren Temperaturbereich zuverlässig funktionieren, je nach

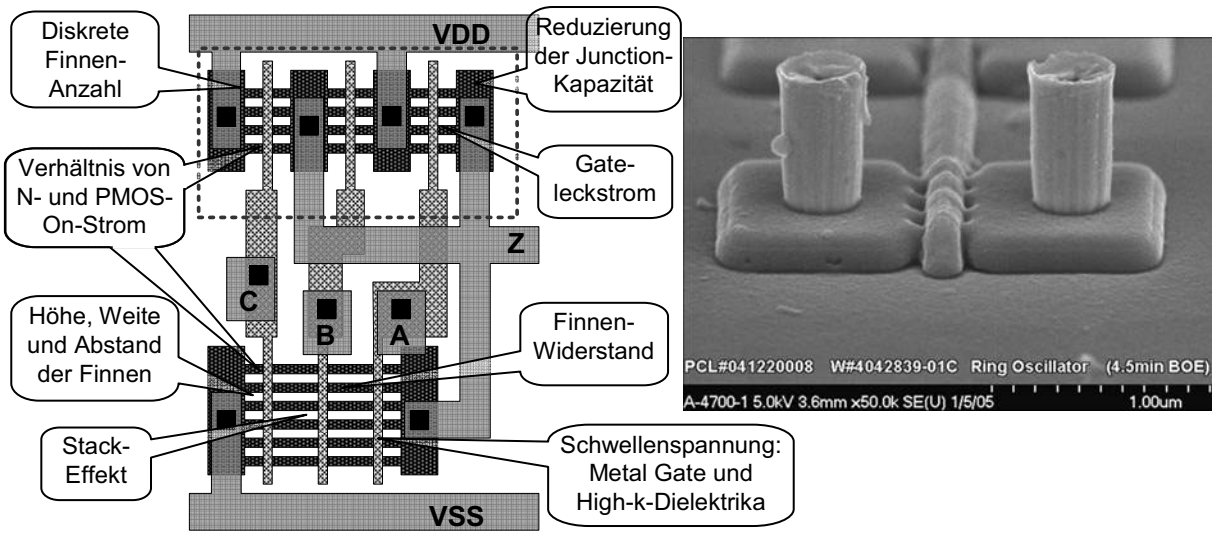


Abbildung 2.4: Besonderheiten des Schaltungsdesigns mit Multi-Gate-Transistoren und SEM-Bild eines MuGFETs.

Anwendung z.B. von $-40\,°C$ bis $125\,°C$, in der Automobilelektronik bis zu $200\,°C$ (Logik-Schaltungen). Einzelne Bauelemente arbeiten hier noch bei $300\,°C$ (Leistungshalbleiter).

Entscheidend für das elektrische Verhalten eines Transistors ist jedoch nicht diese Umgebungstemperatur, sondern die Temperatur der elektrisch aktiven Kanalregion (Junction-Temperatur). Diese ist aufgrund der elektrischen Verlustleistung, die in der Schaltung entsteht, höher als die der Chip- bzw. Gehäuse-Oberfläche. Während in Mikroprozessoren lokale heiße Stellen auftreten können (Gebiete mit hoher Schaltaktivität, *Hot Spots*) und z.B. SRAM-Blöcke kälter bleiben [23], kann die Temperatur in den meisten System-on-Chip-Anwendungen als weitgehend räumlich gleichverteilt angesehen werden.

Jedoch unterscheidet sich auch hier die Temperatur im aktiven Betrieb von der des Standby-Zustands. Es ist daher zweckmäßig, Leckströme vorzugsweise bei niedrigeren Temperaturen ($25\,°C$) und Schaltgeschwindigkeiten bei erhöhten Temperaturen ($85\,°C$) zu betrachten. Im Kapitel 3.2.4 wird gezeigt, dass die Schaltgeschwindigkeit von Sub-100 nm-Schaltungen unter typischen Betriebsbedingungen nur schwach von der Temperatur abhängt. Daher ist es meist ausreichend, Leckströme und Geschwindigkeiten bei Raumtemperatur zu vergleichen.

Dennoch dürfen andere Bedingungen nicht außer Acht gelassen werden, um eine sichere Funktion der Schaltung auch unter ungünstigen Bedingungen gewährleisten zu können (Worst-Case-Bedingungen). Alle Ränder des Spezifikationsbereichs müssen dazu untersucht werden. Die wichtigsten Parameter sind, neben der Temperatur und der Versorgungsspannung, die Gateläng en sowie die Schwellenspannungen der einzelnen Transistoren.

Die Spezifikationsbreite für die Versorgungsspannung muss sowohl verschiedene Betriebsmodi als auch Schwankungen des nominellen V_{dd} von typischerweise $10\,\%$ umfassen. In der 90 nm-Technologie werden nominelle Versorgungsspannungen von $0.85\,V$ (Low-Operating-Power-Modus) bis zu $1.35\,V$ (hohe Performance-Anforderung) eingesetzt [6, 24]. Folglich muss eine

Schaltung bzw. eine Schaltungstechnik über einen sehr weiten Spannungsbereich von 0.76 bis 1.48 V funktionsfähig sein.

Mit der Skalierung der Strukturgrößen in den Sub-100 nm-Bereich entwickeln sich Parametervariationen mehr und mehr zu einem ernsthaften Problem. Ein kritischer Parameter ist die Gatelänge, deren Abmessung (ca. 60 nm in der 90 nm-Technologie) weit unterhalb der Wellenlänge der eingesetzten Lithographie liegt (193 nm). Die minimale Gatelänge muss daher so gewählt werden, dass auch unter ungünstigen Prozess- und Betriebsbedingungen die durch die jeweilige Anwendung definierte Obergrenze für den Leckstrom eingehalten wird. Es wird ein bestimmter Offset vorgehalten, damit z.B. ein Transistor mit einer 3-σ-Abweichung vom Mittelwert maximal eine V_t-Abweichung von 80 mV besitzt.
Die Schwellenspannung kann sowohl systematisch bedingt schwanken (Weiten- und Längenabhängigkeit, Verspannungen des Siliziums, Abstand zwischen N- und PMOS-Transistoren), als auch zufällig variieren (Kanaldotierung, Gatelänge, Grenzflächenzustände [15–17]). Hinzu kommen außerdem Degradationserscheinungen über die Lebensdauer einer Schaltung. Systematische Parametervariationen können während des Schaltungsentwurfs berücksichtigt werden, indem in der Schaltungssimulation unterschiedliche Parametersätze verwendet werden. Allerdings müssen die komplexen Abhängigkeiten zwischen den physikalischen Effekten und dem Schaltungslayout verstanden werden und die Transistormodelle sowie die Layout-Vorschriften entsprechend angepasst werden. Die Weitenabhängigkeit der Schwellenspannung wird im Kapitel 3.1.1 beschrieben.
Die systematischen Parametervariationen sind häufig räumlich korreliert und beeinflussen deshalb das Schaltverhalten und den Leckstrom von Logikschaltungen signifikant. Im Gegensatz dazu sind statistische Variationen in einem Datenpfad, in dem viele Logikgatter hintereinander schalten, zu vernachlässigen, da diese sich nach wenigen Gattern herausmitteln. In SRAM-Zellen sind hingegen gerade die statistischen Parametervariationen kritisch [18], da hier das Verhältnis der Schaltströme entscheidend ist (*Matching*). Systematische Parametervariationen wirken sich dagegen in SRAMs weniger stark aus. Systematische Schwankungen können mit Hilfe von Monte-Carlo-Simulatoren untersucht werden.
Die Ursachen und Auswirkungen statistischer V_t-Variationen werden in dieser Arbeit nicht untersucht, müssen aber stets berücksichtigt werden, da sie in zukünftigen Technologien die Worst-Case-Schaltgeschwindigkeit stark beeinträchtigen können [3]. Die unterschiedlichen Parametersätze in den Transistormodellen dienen dazu, die Funktion einer Schaltung unter verschiedenen Bedingungen untersuchen zu können (hohe, nominelle und niedrige Schwellenspannungen jeweils für N- und PMOS-Transistoren).

Schaltungstechniken in System-on-Chip-Umgebungen

Neben digitalen Logikschaltungen und häufig sehr großen SRAM-Blöcken besteht in vielen Anwendungen die Herausforderung in einer zusätzlichen Integration von analogen und Hochfrequenzschaltungen auf einem Chip (Abb. 2.5). Die Komplexität der zu realisierenden Gesamtsysteme erfordert eine sorgfältige Überprüfung jeder in einem Schaltungsblock eingesetzten Schaltungstechnik auch auf eventuelle Rückwirkungen auf andere Systemkomponenten.

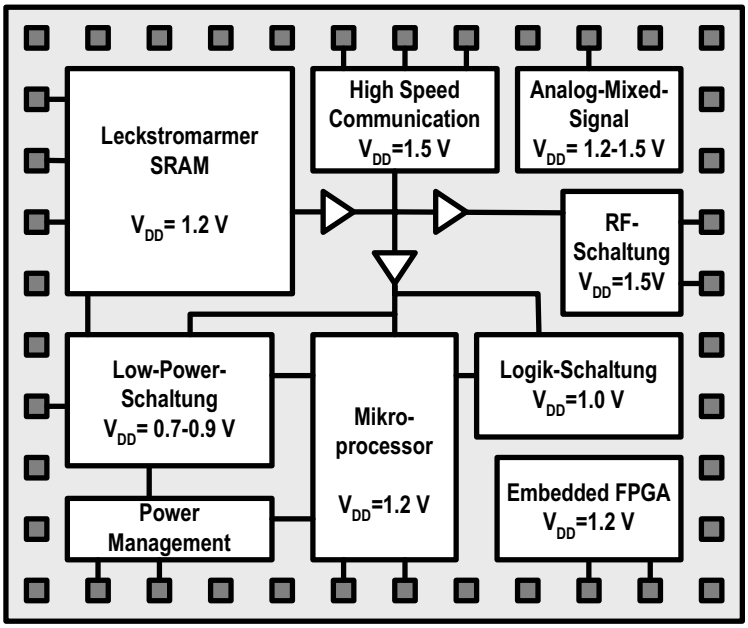


Abbildung 2.5: Schaltungsblöcke in einer System-on-Chip-Umgebung. Neben den in dieser Arbeit behandelten digitalen Schaltung müssen noch andere Blöcke integriert werden, die häufig bei unterschiedlichen Spannungen betrieben werden sowie andere Device-Typen verwenden.

Aufgrund der hohen Komplexität der Gesamtsysteme sind bis heute erst wenige Techniken zur Reduzierung des Leckstroms in automatisierte Design-Umgebungen auch tatsächlich integriert worden. In der Literatur finden sich jedoch zunehmend mehr erfolgreiche Implementierungen:

Blockabschaltung: Der Einsatz von Standby-Transistoren zur Abschaltung einzelner, vorübergehend nicht benötigter Schaltungsblöcke entwickelt sich mehr und mehr zu einer Standard-Methode. Der Leckstrom kann sehr effizient reduziert werden, jedoch erfordert die Reaktivierung der Schaltung besondere Vorsicht [25, 26].

Standby-Transistoren werden häufig nicht als große zentrale Devices ausgelegt, sondern verteilt in den Schaltungsblock integriert [6, 27, 28].

Body Biasing: Die Änderung der Substratspannung zur Anpassung der Schwellenspannungen an den Betriebszustand einer Schaltung wird vereinzelt zur Reduzierung des Leckstroms eingesetzt (*Reverse-Biasing*). Es kann aber auch die maximale Frequenz einer Schaltung durch *Forward-Biasing* erhöht werden. Auf das Body Biasing wird im Kapitel 4 eingegangen [29–33].

Multi-V_t/Multi-t_{ox}: Durch die Kombination unterschiedlicher Schwellenspannungen oder Oxiddicken in einer Schaltung lassen sich ebenfalls Leckströme reduzieren bzw. die Geschwindigkeiten erhöhen. Dazu wird der kritische Pfad einer Schaltung aus schnellen LVT-Transistoren aufgebaut, während alle anderen Transistoren ($\approx 90\,\%$) höhere Schwellenspannungen aufweisen. Nicht immer ist jedoch die Identifizierung des kritischen Pfades einfach. Auch für

die Blockabschaltung bietet sich eine Verwendung leckstromarmer Standby-Transistoren und schneller Logiktransistoren an [34–36].

Obwohl zusätzliche Transistortypen in der Logikschaltung eingesetzt werden, bleiben der prozesstechnische Aufwand bei der Herstellung und damit die Kosten konstant, wenn dieser Transistortyp in anderen Schaltungsteilen ohnehin zum Einsatz kommt (z.B. leckstromarme Devices in SRAM-Blöcken). Schaltungstechniken, die unterschiedliche Transistortypen in einer Schaltung oder sogar innerhalb einzelner logischer CMOS-Gatter kombinieren, befinden sich erst in der Entwicklung. Im Kapitel 5 werden neue Möglichkeiten hierzu vorgestellt.

Variation der Gatelänge: Da die Schwellenspannung bei sehr kurzen Gatelängen eine zunehmende Abhängigkeit von der Gatelänge besitzt, kann das V_t und damit der Leckstrom durch Variation von L_g an die Anforderungen in einer Schaltung angepasst werden [28]. Während für Multi-t_{ox}- oder Multi-V_t-Techniken zusätzliche Prozess-Schritte eingeführt werden müssen, kann die Gatelänge allein durch das Layout variiert werden.

Die Variation der Gatelänge wird im Kapitel 5 mit anderen Techniken wie Blockabschaltung, Multi-V_t oder Multi-t_{ox} kombiniert.

Variation der Taktfequenz/Multi-$\mathbf{V_{dd}}$: Bei vorübergehend reduzierten Performance-Anforderungen können die Versorgungsspannung V_{dd} und die Betriebsfrequenz f reduziert werden. Insbesondere die mehrstufige Anpassung der Taktfrequenz an die Anwendung stellt eine sehr effiziente Methode zur Reduzierung der aktiven Verlustleistung dar. Im nächsten Schritt kann zusätzlich V_{dd} abgesenkt werden, wodurch die aktive Leistungsaufnahme weiter sinkt, vor allem aber auch die Leckströme signifikant reduziert werden. Da eine gleichzeitige stufenlose Regelung von Spannung (*Variable-V_{dd}*) und Frequenz schwierig ist, werden meist nur wenige Betriebszustände mit unterschiedlichen V_{dd}/f-Kombinationen vorgesehen [6, 24, 28].

Kapitel 3

Leistungsaufnahme und Schaltgeschwindigkeit digitaler CMOS-Schaltungen

Die Unterschwellenstromsteilheit $S = 2.3 \cdot m \cdot k_B T/q$ eines MOS-Transistors kann auch mit einem Idealitätsfaktor $m = 1 + C_{depl}/C_{ox}$ von 1 bei Raumtemperatur nicht größer werden als 60 mV/Dekade. In der 130 nm-Technologie ist $S = 90$ mV/Dekade. Dies hat zur Folge, dass die Schwellenspannung bei der Transistorskalierung nicht weiter abgesenkt werden kann, ohne dass der Unterschwellenstrom ansteigt. Durch die Reduzierung der effektiven Oxiddicke kann S zwar verbessert werden, jedoch steigt der Gate-Tunnelstrom bei Oxiddicken kleiner als 2 nm sehr stark an. Außerdem trägt der *Gate-Induced Drain Leakage* (GIDL) zum Leckstrom eines ausgeschalteten Transistors bei.

Erst die Einführung neuer Materialien (High-κ-Dielektrika und Metall-Gates) und neuer Transistor-Konzepte (Fully Depleted SOI oder MuGFETs) ermöglichen eine Fortsetzung der Skalierung. Da diese Materialien und Konzepte nicht vor Einführung der 32 nm-Technologie zur Verfügung stehen, erfordern verschiedene Anwendungen eine zunehmende Spezialisierung der Transistoren auf verschiedene Anforderungen. Abbildung 3.1 zeigt die gemessenen Eingangs-Charakteristiken dreier NMOS-Transistoren aus der 90 nm-Technologie. Der Leckstrom wird bei $V_{gs} = 0$ V in jedem der abgebildeten Transistoren durch einen anderen Leckstrombeitrag dominiert (Abb. 3.2). So wird der Off-Strom I_{off} entweder durch den Unterschwellenstrom $I_{s,off}$, durch den Gateleckstrom des ausgeschalteten Transistors $I_{g,off}$ oder durch den GIDL-Strom I_{gidl} bestimmt.

Zunächst werden in diesem Kapitel die Leckstrombeiträge Unterschwellenstrom, GIDL und Gateleckstrom untersucht. Anschließend werden Modelle für die Schaltgeschwindigkeit vorgestellt. Dabei ist es nicht das Ziel, ein vollständiges Transistormodell zu entwickeln. Vielmehr sollen die Abhängigkeiten der einzelnen Effekte von z.B. Spannungen und Temperaturen physikalisch basiert und zugleich anschaulich beschrieben werden. Auf dieser Grundlage können dann in den folgenden Kapiteln schaltungstechnische Maßnahmen entwickelt werden, die dazu beitragen können, den Zielkonflikt in Sub-100 nm-Schaltungen zwischen niedrigem Leckstrom einerseits und hohen Schaltgeschwindigkeiten andererseits zu lösen.

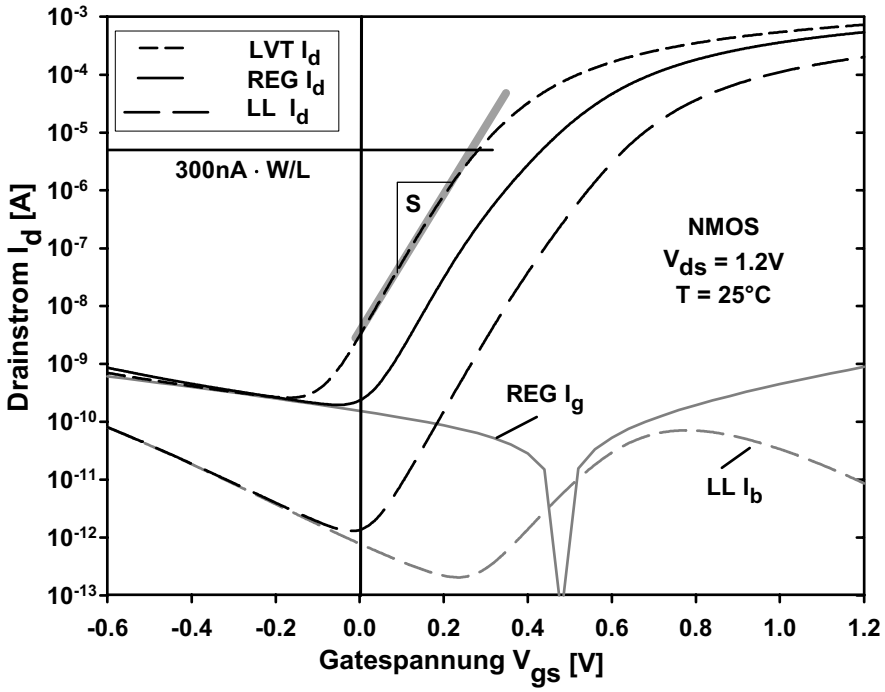


Abbildung 3.1: Eingangs-Charakteristiken von NMOS-Transistoren der 90 nm-CMOS-Technologie bei unterschiedlichen Bulk-Source-Spannungen V_{bs}. Nur der Off-Strom des LVT-Transistors wird durch den Unterschwellenstrom bestimmt. Bei dem REG-Transistor dominiert hingegen der Gate-Tunnelstrom und bei dem Low-Leakage-Transistor der GIDL-Effekt den Off-Strom.

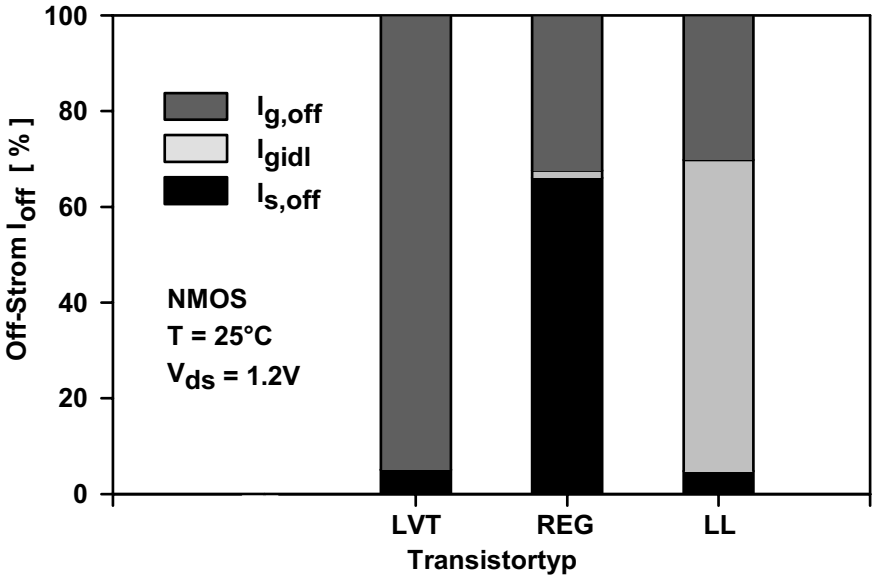


Abbildung 3.2: Relative Zusammensetzung der Leckströme von 90 nm-NMOS-Transistoren. Im Gegensatz zu den hier dargestellen NMOS-Transistoren sind PFETs häufiger durch den Unterschwellenstrom $I_{s,off}$ dominiert, da Gate-Tunnelströme und GIDL-Ströme in PMOS-Transistoren kleiner sind.

3.1 Leckströme und aktive Leistungsaufnahme

3.1.1 Unterschwellenstrom und Gate-Induced Drain Leakage

Leckströme und Schaltgeschwindigkeiten sind in hohem Maße voneinander abhängig. Wird ein Transistor mit einer hohen Schwellenspannung V_t verwendet, so ist der Unterschwellenstrom zwar klein, die Schaltgeschwindigkeit jedoch auch gering. Ein niedriges V_t erhöht die Performance, aber auch den Off-Strom. Bei allen folgenden Betrachtungen bildet die Schwellenspannung V_t daher den zentralen Bezugspunkt. Der Bereich unterhalb der Schwellenspannung $V_{gs} < V_t$ bestimmt dabei im Wesentlichen das Leckstromverhalten, der Bereich $V_{gs} > V_t$ den On-Strom des Transistors und damit die Geschwindigkeit einer Schaltung, wobei die Schwellenspannung selbst jedoch nicht unabhängig von äußeren Einflüssen wie den anliegenden Spannungen und der Temperatur ist.

Es existieren mehrere unterschiedliche Definitionen der Schwellenspannung. Nach der physikalischen Definition ist V_t die Gate-Source-Spannung V_{gs}, bei der in einem NMOS-Transistor die Elektronendichte an der Grenzfläche zwischen dem Kanal und dem Dielektrikum der Löcherdichte im Substrat entspricht [37]. Dieses ist gleichbedeutend mit dem Einsetzen der starken Inversion am source-seitigen Ende des Kanals.

Liegt die Gate-Spannung unterhalb der Schwellenspannung $V_{gs} < V_t$ und ist $V_{ds} > 4 \cdot U_T$, dann dominiert der Diffusionsstrom und der Unterschwellenstrom zwischen Source und Drain lässt sich beschreiben durch [37]

$$I_s = I_{ss} \exp \frac{V_{gs} - V_t}{m U_T} \tag{3.1}$$

$$\text{mit} \quad I_{ss} = \mu_n U_T \frac{W}{L} q N_A L_D \frac{1}{\sqrt{2\left[\frac{\psi_s - \psi_B}{U_T} - 1\right]}} \tag{3.2}$$

Dabei ist die $U_T = k_B T/q$ die Temperaturspannung, N_A die effektive Dotierung der Kanalregion und $L_D = \sqrt{\varepsilon_{\mathrm{Si}} U_T / q N_A}$ die Debye-Länge.

Da das Einsetzen der starken Inversion jedoch nicht elektrisch messbar ist, wird in der schaltungstechnischen Praxis nach dem Stromkriterium (*Constant-Current Criterion*) ein Schwellenwert für den Drain-Strom I_d^{th} definiert, bei dem $V_{gs} = V_t$ ist [38]. Im Rahmen dieser Arbeit wird für NMOS-Transistoren die Schwellenspannung durch die Gate-Source-Spannung V_{gs} definiert, bei der ein Strom $I_{dn}^{th} = 300\,\mathrm{nA} \cdot W/L$ fließt (Abb. 3.1). Für PFETs liegt die Schwelle bei $I_{dp}^{th} = 70\,\mathrm{nA} \cdot W/L$. Die unterschiedlichen Werte (300 bzw. 70 nA) gehen ursprünglich auf die unterschiedlichen Beweglichkeiten von Elektronen und Löchern zurück, die sich im Bereich der Schwellenspannung besonders stark unterscheiden, sind aber als empirisch festgelegte Werte aufzufassen.

Wird ausgehend von $V_{gs} = V_t$ ein exponentieller Abfall der Eingangskennlinie bis $V_{gs} = 0\,\mathrm{V}$ angenommen, dann ist der Unterschwellenstrom nur von der Unterschwellenstromsteilheit S abhängig

$$I_{s,off} = I_d^{th} \cdot 10^{-\frac{V_t - V_{t,offset}}{S}}, \tag{3.3}$$

	EOT	NMOS	PMOS
0.5 μm-Technologie	16 nm	1.21 mV/K	1.91 mV/K
90 nm-Technologie	3.2 nm	0.63 mV/K	0.87 mV/K
90 nm-Technologie	2.3 nm	0.53 mV/K	0.77 mV/K
MuGFET-Transistor	2.9 nm	0.33 mV/K	0.58 mV/K

Tabelle 3.1: Vergleich der Temperaturkoeffizienten k_1 für NMOS- und PMOS-Transistoren verschiedener Technologie-Generationen: simulierte Werte für die 0.5 μm-Technologie, Messergebnisse für 90 nm-, 130 nm- und MuGFET-Technologien.

wobei die Offset-Spannung $V_{t,offset}$ den nicht vollständig exponentiellen Verlauf der Kennlinie in der Nähe von $V_{gs} = V_t$ berücksichtigt. Unter anderem hat eine Weitenabhängigkeit der Schwellenspannung einen großen Einfluss auf diese Abweichung der I_d-Kennlinie vom idealen exponentiellen Verlauf (Abb. 3.1). Auf eine Bestimmung der Offset-Spannung V_{offset}, die typischerweise eine Größe von 30 bis 60 mV hat, kann verzichtet werden, wenn nur relative Vergleiche zwischen Unterschwellenströmen beabsichtigt sind. So erhöht sich z.B. für $S = 90$mV/Dekade bei einer Reduzierung der Schwellenspannung um $\Delta V_t = 100$ mV der Strom $I_{s,off}$ um den Faktor $10^{\Delta V_t/S} = 12.9$.

Temperaturabhängigkeit

Die Temperatur hat einen sehr großen Einfluss auf den Unterschwellenstrom $I_{s,off}$, der sich aus der Temperaturabhängigkeit sowohl der Unterschwellenstromsteilheit als auch der Schwellenspannung ergibt. Die Temperaturabhängigkeit der Schwellenspannung kann modelliert werden durch [39]

$$V_t(T) = V_t(T_0) - k_1 \cdot (T - T_0). \tag{3.4}$$

$V_t(T_0)$ ist die Schwellenspannung bei einer Referenztemperatur T_0. In Tabelle 3.1 werden die Temperaturkoeffizienten k_1 verschiedener Transistoren in unterschiedlichen Technologien verglichen. In den Bulk-Technologien ist k_1 weitgehend unabhängig von der Schwellenspannung $V_t(T_0)$. Bei der Skalierung von der 0.5 μm- über die 130 nm- hin zur 90 nm-Technologie fällt auf, dass der Temperaturkoeffizient mit der Oxiddicke sinkt. Dieses Verhalten korrespondiert mit der Temperaturabhängigkeit der Schwellenspannung, die sich aus den Gleichungen 3.5 bis 3.10 ergibt [37]:

$$V_t(T) = -\frac{E_g(T)}{2q} + \Phi_F(T) + \frac{qN_A}{C'_{ox}} x_{d,max}(T) \tag{3.5}$$

$$x_{d,max}(T) = \sqrt{\frac{2\varepsilon_{\mathrm{Si}}}{qN_A} \cdot 2\Phi_F(T)} \tag{3.6}$$

$$\Phi_F(T) = \frac{k_B T}{q} \ln \frac{N_A}{n_i(T)} \tag{3.7}$$

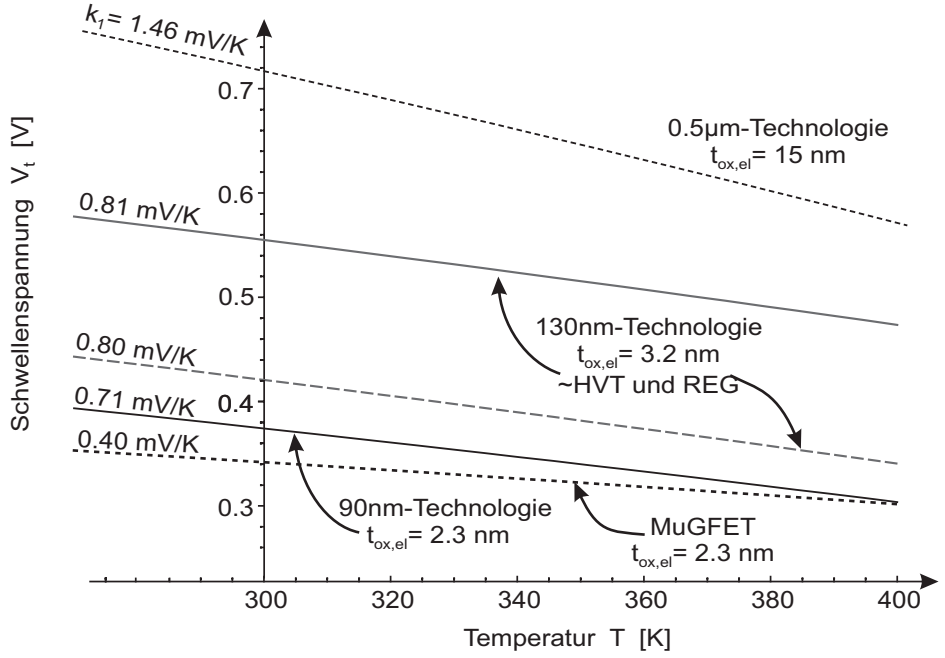


Abbildung 3.3: Schwellenspannung in Abhängigkeit von der Temperatur für verschiedene Transistoren gemäß Gleichung 3.5. Die Werte stimmen abgesehen von unterschiedlichen Offsets für N- und PFETs in ihrer Tendenz mit den messtechnisch gewonnenen Werten in Tabelle 3.1 überein.

$$n_i(T) = 2\left(m_n^* m_p^*\right)^{\frac{3}{4}} \left(\frac{k_b T}{2\pi\hbar^2}\right)^{\frac{3}{2}} e^{-\frac{E_g}{2k_B T}} \tag{3.8}$$

$$E_g(T) = \left(1.17 - 4.73 \cdot 10^{-4} \cdot \frac{T^2}{T + 636\text{K}}\right)\left[\frac{\text{eV}}{\text{K}}\right] \tag{3.9}$$

$$m_n^* = 1.1 m_0\ , \qquad m_p^* = 1.55 m_0\ , \qquad m_0 = 9.1 \cdot 10^{-31}\text{kg} \tag{3.10}$$

Dabei sind die Bandlücke E_g, das Fermi-Potential Φ_F, die intrinsische Ladungsträgerdichte n_i sowie die maximale Weite der Raumladungszone $x_{d,max}$ temperaturabhängig. Die Dotierung unterhalb des Gates N_A wird näherungsweise als konstant angesehen. Die effektiven Massen m_n^* und m_p^* ergeben sich aus der Bandstruktur des Siliziums. Die weitenbezogene Oxidkapazität ist $C'_{ox} = \varepsilon_{ox}/EOT$.

In Abbildung 3.3 ist Gleichung 3.5 für typische Bulk-Technologien und einen MuGFET in Abhängigkeit von der Temperatur aufgetragen. An der Steigung der Kennlinien kann der Temperaturkoeffizient k_1 abgelesen werden. Während die Oxiddicke den Koeffizienten stark beeinflusst, bleibt k_1 bei unterschiedlichen V_t-Implantationen nahezu konstant. Es zeigt sich eine gute Übereinstimmung mit den Messwerten in Tabelle 3.1.

In MuGFETs ist die Größe der Raumladungszone nicht durch die Dotierung, sondern durch die Weite der Finnen W_{fin} definiert. In Gleichung 3.5 wird daher $x_{d,max} = W_{fin}/2$ gesetzt. Dadurch reduziert sich k_1 bei konstanter Oxiddicke.

Die Unterschwellenstromsteilheit

$$S(T) = 2.3 \cdot \frac{k_B T}{q} \left(1 + \frac{\varepsilon_{\mathrm{Si}}}{\varepsilon_{ox}} \frac{t_{ox}}{x_{d,max}}\right) \tag{3.11}$$

besitzt ebenfalls eine Temperaturabhängigkeit. In einem MuGFET ohne Kurzkanaleffekt ist S nur von der Temperaturspannung $U_T = k_B T/q$ abhängig. In Bulk-Transistoren ist zusätzlich die Weite der Raumladungszone temperaturabhängig: $x_{d,max}$ nimmt bei Erhöhung der Temperatur ab und die Temperaturabhängigkeit der Unterschwellenstromsteilheit nimmt damit zu. Die daraus resultierende Erhöhung des Off-Stroms ist stark abhängig von der Größe der Schwellenspannung $V_t(T_0)$. Je größer V_t ist, desto höher ist die Temperaturabhängigkeit von $I_{s,off}$.

Die Effekte $V_t(T)$ und $S(T)$ führen jeweils zu einer Erhöhung des Unterschwellenstroms $I_{s,off}$ bei höheren Temperaturen. Die Zunahme fällt jedoch bei neueren Technologiegenerationen mit dünneren Oxiden und kleineren Schwellenspannungen, insbesondere aber in MuGFETs, zunehmend kleiner aus. So nimmt $I_{s,off}$ in der 0.5μm-Technologie zwischen 25 °C und 85 °C um einen Faktor 300 zu, für einen REG-Transistor in der 90 nm-Technologie um einen Faktor 13.8, in einem MuGFET mit $V_t = 0.3\,\mathrm{V}$ aber nur noch um einen Faktor 6.1. Bei kleinem V_t macht $V_t(T)$ den größeren Teil der Leckstromerhöhung aus, bei großen Schwellenspannungen überwiegt $S(T)$.

Die reduzierte Temperaturabhängigkeit des Unterschwellenstroms kann bei der Wahl der optimalen Schwellenspannung für eine Schaltung mit einem bestimmten Leckstrombudget berücksichtigt werden. So kann bei gleichem $I_{s,off}(T = 400\,\mathrm{K})$ in einer MuGFET-Technologie eine um 35 mV kleinere Schwellenspannung verwendet werden als in einer vergleichbaren Bulk-Technologie (vgl. Abb. 3.3).

In diesem Zusammenhang ist die Frage, bei welcher Temperatur der Off-Strom bewertet werden muss. Leckströme sind besonders dann relevant, wenn sich eine Schaltung im Standby-Zustand befindet und daher kälter ist. Meist jedoch wird der Leckstrom und damit z.B. die Standby-Zeit eines mobilen Gerätes lediglich bei der maximalen Temperatur spezifiziert. Dieses ist nicht immer sinnvoll, denn obwohl ein Schaltungsblock im Standby-Modus durch einen benachbarten aktiven Block erwärmt werden kann, ist in der Regel die Leistungsaufnahme des aktiven Blocks bestimmend für die Gesamtverlustleistung.

Abbildung 3.4 zeigt die Temperaturabhängigkeit von $I_{s,off}$. Verglichen mit der Zunahme des Source-Stroms um einen Faktor 13.8 ist der ebenfalls dargestellte Gateleckstrom von ein- und ausgeschaltetem Transistor nur schwach temperaturabhängig.

Längen- und Weitenabhängigkeit der Schwellenspannung

Die Schwellenspannung moderner Transistoren ist neben den anliegenden Spannungen und Temperaturen auch von den geometrischen Abmessungen des Transistors abhängig. Insbesondere ist V_t eine Funktion der Gatelänge und der Transistorweite.

Die Längenabhängigkeit wird von zwei Effekten bestimmt. Abbildung 3.5 zeigt schematisch das bekannte V_t-Verhalten in Abhängigkeit von der Gatelänge L_g. Bei kurzem L_g führt der

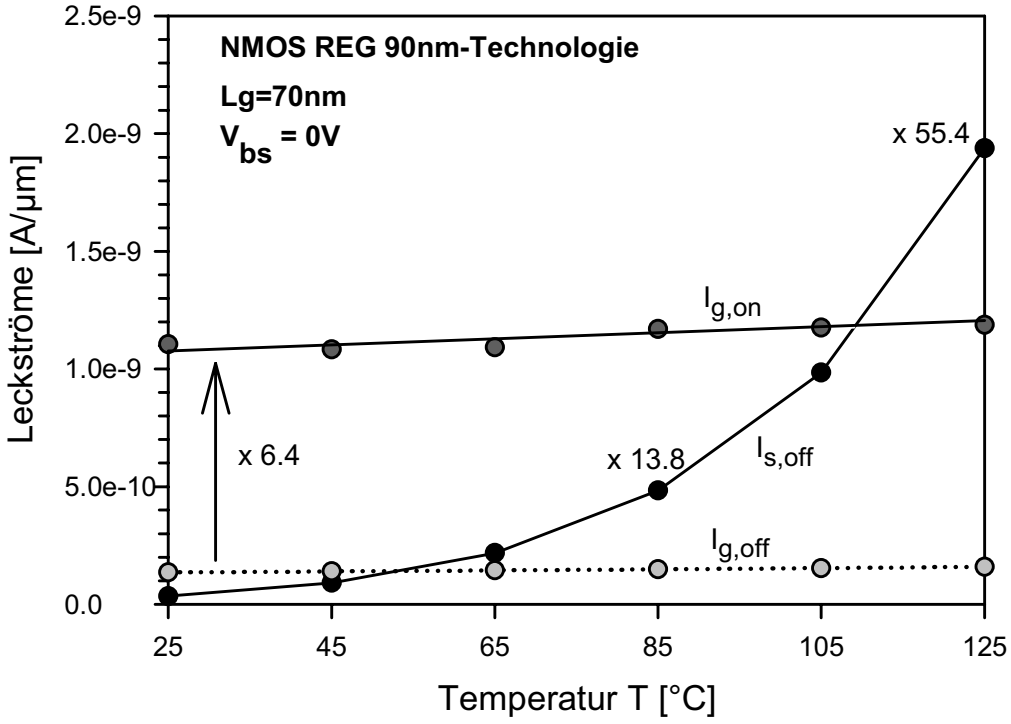


Abbildung 3.4: Temperaturabhängigkeit von Source- und Gatestrom. Im Vergleich zum Source-Strom $I_{s,off}$, der zwischen 25 und 125 °C um einen Faktor 55.4 zunimmt, sind die Gateleckströme von ein- und ausgeschaltetem Transistor $I_{g,on}$ und $I_{g,off}$ nur schwach temperaturabhängig.

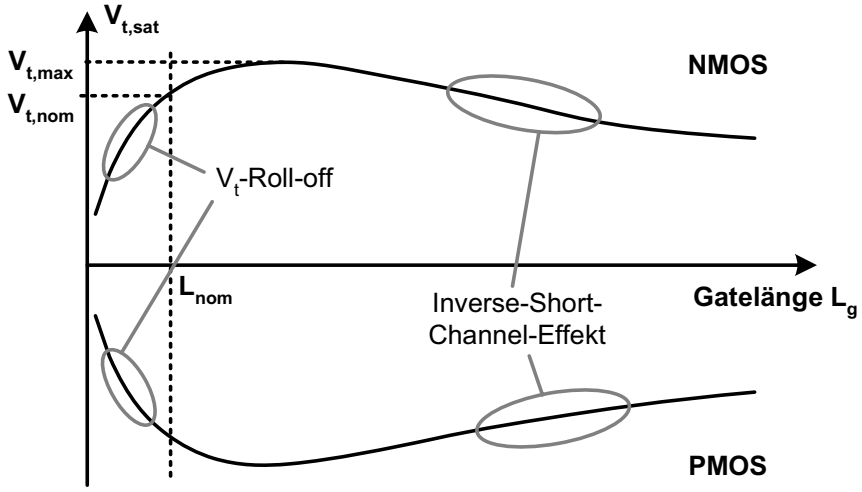


Abbildung 3.5: Schematische Darstellung der Schwellenspannung in Abhängigkeit von der Gatelänge.

V_t-Roll-off zu einem reduzierten V_t, insbesondere im Sättigungsbereich. Dieses resultiert aus einem Durchgriff des Drain-Potentials bis zum PN-Übergang des Source-Kontakts (*Drain-Induced Barrier Lowering*, DIBL).

In modernen CMOS-Technologien wird die Substratdotierung zudem in der Nähe der Source- und Drain-Inseln durch eine so genannte Halo-Implantation lokal erhöht, um einen Felddurchgriff unterhalb des Kanals zu verhindern (Abb. 2.3). Diese Implantation wirkt sich aber auch auf das Fermi-Niveau und damit auch auf die Schwellenspannung aus. Bei kurzen Gatelängen ist ein großer Teil der Kanallänge von der Halo-Dotierung beeinflusst und V_t wird angehoben. Bei langen Transistoren nimmt der Einfluss der Halos ab und nur noch die eigentliche V_t-Implantation bestimmt die Schwellenspannung. Die mittlere Dotierung in der Kanalregion ist somit bei großen Gatelängen kleiner und die Schwellenspannung sinkt. Dieser Effekt wird als *Inverse Short-Channel Effect* bezeichnet.

Aufgrund dieser beiden Effekte besitzt die Schwellenspannung bei einer bestimmten Gatelänge ein Maximum, üblicherweise etwa bei Gatelängen $L_g = L_{min} + 20 \ldots 40\,\text{nm}$. Es ist daher möglich, die Schwellenspannung zu erhöhen, indem ein nicht minimales L_g gewählt wird. In-Sub-100 nm-Technologien liegt die nominelle Gatelänge L_{nom} zunehmend weiter im V_t-Roll-off-Bereich. Bei Verwendung von Gatelängen im Roll-off-Bereich erhöht sich die Sensitivität einer Schaltung gegenüber Parameterschwankungen. Bei guter Prozess-Kontrolle lassen sich jedoch kürzere Gatelängen und damit ein schnelleres Schaltverhalten erzielen. Dieser Effekt kann dazu genutzt werden, um unterschiedliche Schwellenspannungen in einer Schaltung ohne zusätzlichen Prozessierungsaufwand zu realisieren [28]. Allerdings reduziert sich im Roll-off-Bereich vor allem $V_{t,sat}$, während die Schwellenspannung im linearen Bereich $V_{t,lin}$ weniger stark absinkt. Die Abhängigkeit der Schaltgeschwindigkeit vom DIBL-Verhalten der Transistoren wird im Abschnitt 3.2.1 beschrieben.

Die Schwellenspannung besitzt neben der Längen- auch eine Weitenabhängigkeit, die aus mehreren Effekten resultiert:

- In die Berechnung der Schwellenspannung geht u.a. das Verhältnis von Bulk-Ladung $Q_b = qN_A x_{d,max}$ zur Oxidkapazität C_{ox} ein (Gleichung 3.5). Während Q_b/C_{ox} in LOCOS-isolierten Transistoren zum Rande hin abnimmt, vergrößert sich dieser Term bei Verwendung von STI-Gebieten [40,41] (Abb. 3.6). In STI-Transistoren wird dieser Effekt als *Inverse Narrow Width Effect* bezeichnet, da die Verschiebung der Schwellenspannung im Vergleich zu den vorhergehenden LOCOS-Transistoren in die entgegengesetzte Richtung verläuft.

- An der Grenze zum STI-Oxid führen Segregationsvorgänge in PMOS-Transistoren zu erhöhten und in NMOS-Transistoren zu reduzierten Dotierstoffkonzentrationen. Die Schwellenspannung wird daher in den Randbereichen lokal ebenfalls erhöht bzw. reduziert. Während ein schmaler Transistor zu großen Teilen aus diesen Randbereichen besteht, ist dieser Effekt bei weiten Devices zu vernachlässigen. Hieraus resultiert, dass die Schwellenspannung in schmalen NFETs reduziert und in schmalen PFETs betragsmäßig angehoben wird [42, 43].

- Aufgrund von mechanischem Druck (*Stress*) kommt es an den Rändern des Transistors zu Verspannungen (*Strain*), die zu einer unterschiedlichen Verteilung und Aktivierung von Kanal-, Source/Drain- und Halo-Dotierung führt. Außerdem haben die Verspannungen Einflüsse auf die Ladungsträgerbeweglichkeit, insbesondere in PMOS-Transistoren. Dieser Effekt wird als STI-Stress-Effekt bezeichnet [44, 45].

Die beiden erstgenannten Effekte werden zusammenfassend als Narrow-Width-Effekt bezeichnet. Dieser resultiert für N- und PMOS-Transistoren mit kleinen Weiten in einer Verschiebung der Schwellenspannungen, die in entgegengesetzte Richtungen verlaufen. In NMOS-Transistoren wirken beide Effekte in dieselbe Richtung, sodass die Schwellenspannung bei minimaler Gateweite z.B. in der 130 nm-Technologie um 40 mV im Vergleich zu sehr weiten Transistoren kleiner ist (Abb. 3.7). In PMOS-Transistoren führt die Segregation der Kanaldotierung zu einer Erhöhung des Absolutwerts der Schwellenspannung im Randbereich, welche die V_t-Absenkung aufgrund des *Inverse Narrow Width Effects* teilweise ausgleichen oder umkehren kann.

Bei großen Transistorweiten bestimmt der STI-Stress-Effekt den Verlauf der Schwellenspannung. Sowohl im NFET als auch im PFET sinkt der Betrag der Schwellenspannung bei Weiten von $W > 1\,\mu$m ab. Die daraus resultierende Weitenabhängigkeit des Unterschwellenstroms sind in Abbildung 3.8 dargestellt.

In einer SoC-Schaltung kommen sehr verschiedene Transistorweiten zum Einsatz. So werden die Transistoren in SRAM-Zellen weitgehend minimal dimensioniert, während in digitalen Logikschaltungen 300 nm bis 1 μm weite Transistoren eingesetzt werden. In Analog- und RF-Schaltungen werden sehr viel weitere Transistoren benötigt, deren Gesamtweite jedoch meist auf mehrere Transistorfinger verteilt ist. Für das Schaltungsdesign in Sub-100 nm-Technologien müssen die Weitenabhängigkeiten daher stets berücksichtigt werden.

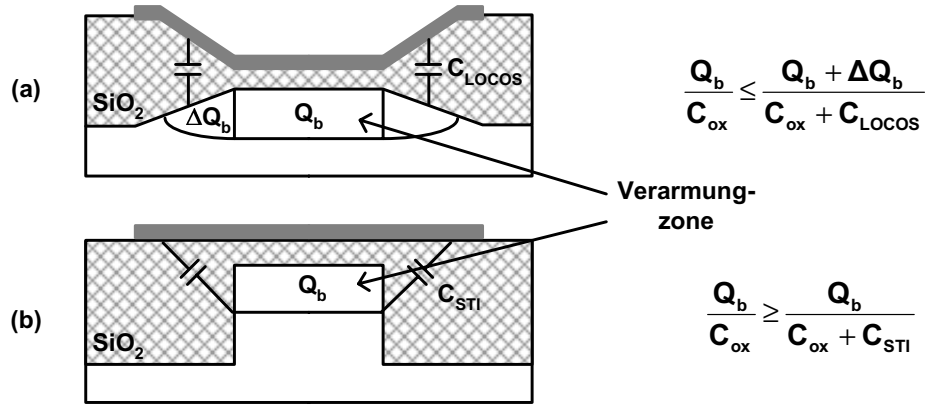


Abbildung 3.6: Querschnitte durch LOCOS- (a) und STI-isolierte (b) Transistoren. Der Term Q_b/C_{ox}, der in die Schwellenspannung unter Annahme der Depletion-Approximation $V_t = V_{fb} + \varphi_s + Q_b/C_{ox}$ eingeht, ist am Rande des LOCOS-isolierten Transistors erhöht, bzw. im STI-isolierten Transistor vermindert [40]. Die Schwellenspannung in Transistoren mit kleiner Weite ist deshalb ebenfalls erhöht (LOCOS) bzw. vermindert (STI).

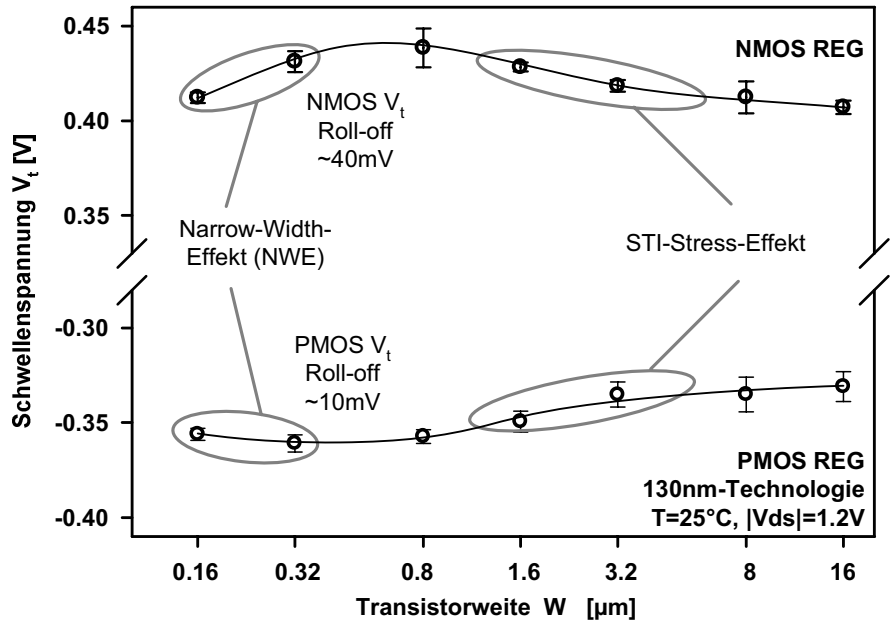


Abbildung 3.7: Gemessene Weitenabhängigkeit der Schwellenspannung für NMOS- und PMOS-Transistoren in der 130 nm-Technologie.

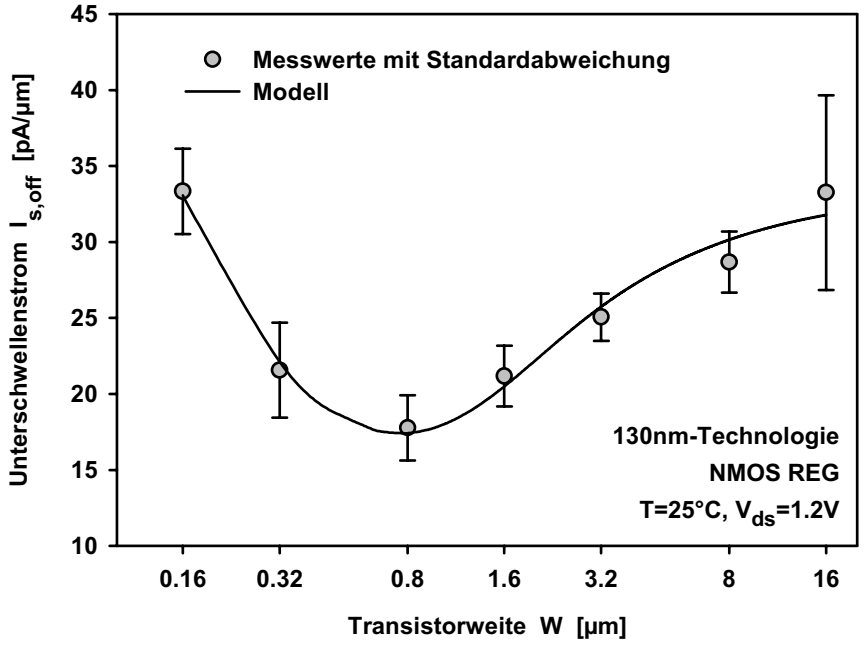


Abbildung 3.8: Weitenabhängigkeit des Unterschwellenstroms von NMOS-Transistoren in der 130 nm-Technologie. Die Messwerte lassen sich mit dem Modell in [46] beschreiben. Das Maximum und das Minimum der Ströme unterscheidet sich um einen Faktor 1.9.

Reduzierung des Leckstroms durch den Stack-Effekt

Der Unterschwellenstrom reduziert sich erheblich, wenn zwei oder mehr ausgeschaltete Transistoren in Serie geschaltet werden, z.B. in einem NAND-Gatter, dessen Eingänge alle mit 0 V beschaltet sind (Abb. 3.9). Dieser Effekt wird als *Stack Effect* bezeichnet.

Der Stack-Effekt lässt sich durch Betrachtung des Knoten-Potentials V_M erklären, das in der Serienschaltung auf 100 bis 150 mV angehoben wird (Abb. 3.9). Dadurch wird $V_{gs,1}$ negativ und der Transistor N1 wird stärker ausgeschaltet. Weiterhin wird $V_{bs,1}$ negativ und die Schwellenspannung des Transistors steigt aufgrund des Substrateffekts. Außerdem reduziert sich die Drain-Source-Spannung des unteren Transistors N2 auf $V_{ds,2} = V_M$, sodass sich der DIBL-Effekt vermindert. Die Kombination der drei Effekte führt zu dem in Abbildung 3.9b dargestellten Arbeitspunkt. Trotz einer Reduzierung des Substrateffekts erhöht sich der Stack-Effekt mit fortschreitender Technologie-Skalierung (Abb. 3.9), da der DIBL-Effekt größer wird [47, 48]. Der Unterschwellenstrom reduziert sich für einen Zwei-Transistor-Stack in einer aktuellen CMOS-Technologie etwa um eine Dekade.

Andere Leckstromkomponenten bleiben vom Stack-Effekt jedoch weitgehend unbeeinflusst. Der GIDL-Effekt im oberen NMOS-Transistor N1 reduziert sich nur sehr wenig, da Drain-Gate-

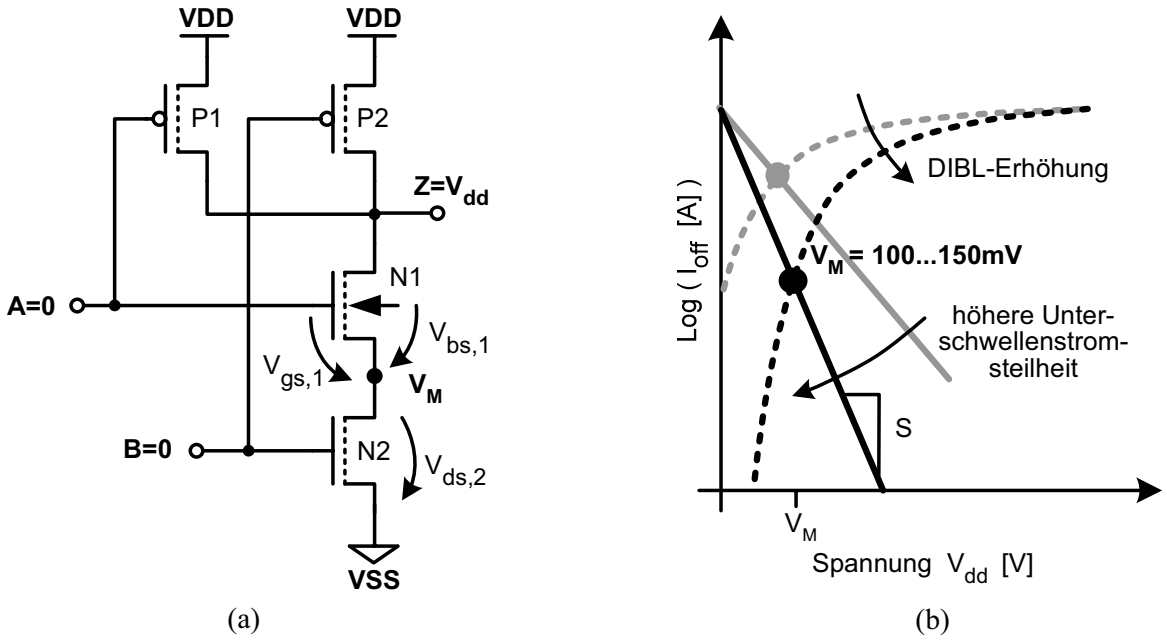


Abbildung 3.9: Stack-Effekt in einem NAND2-Gatter. Das Potential des internen Knotens V_M wird in der Serienschaltung zweier ausgeschalteter Transistoren auf etwa 100 bis 150 mV angehoben. Dadurch wird die Gate-Source-Spannung des oberen Transistors negativ und der Unterschwellenstrom sinkt. Gleichzeitig reduziert sich die Drain-Source-Spannung des unteren Transistors auf V_M und der Unterschwellenstrom sinkt wegen des reduzierten DIBL-Effekts. Mit fortschreitender Skalierung erhöht sich die Unterschwellenstromsteilheit S und auch der DIBL-Effekt ist in der Regel größer. Die Reduzierung des Unterschwellenstroms durch den Stack-Effekt fällt deshalb in neueren Technologie-Generationen (schwarz) stärker aus als in älteren (grau).

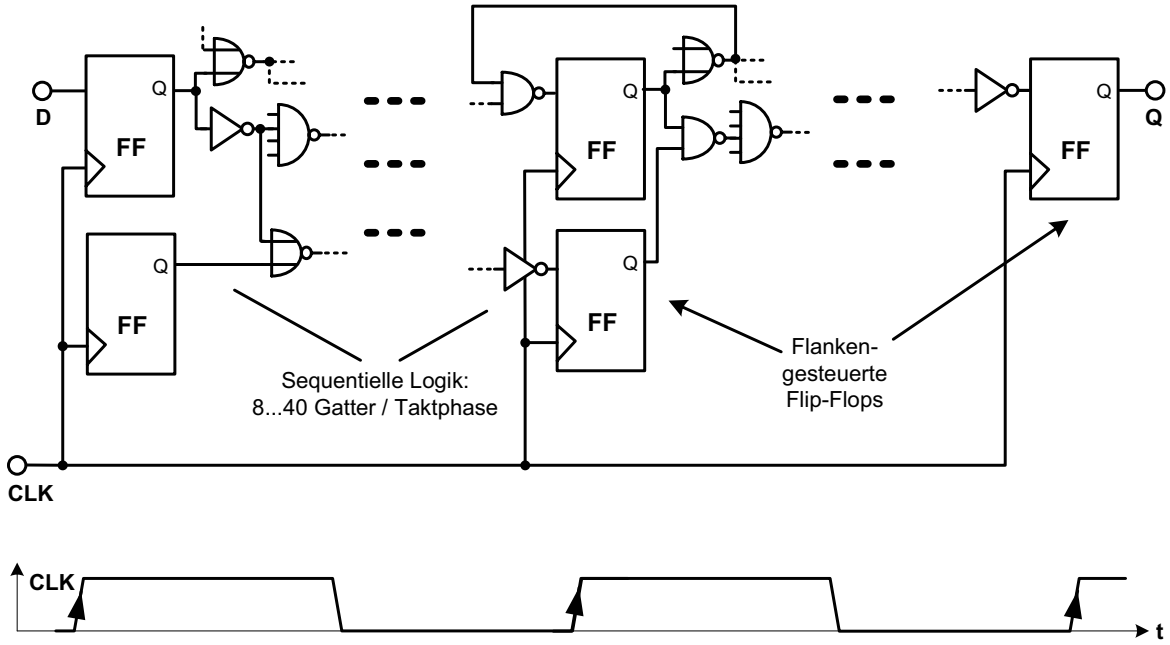


Abbildung 3.10: Aufbau einer getakteten CMOS-Schaltung. Dargestellt ist eine getaktete CMOS-Schaltung mit Flip-Flops, die mit der steigenden CLK-Flanke die Daten an die jeweils nächste Taktphase weitergeben. Die Anzahl der Logikgatter zwischen zwei Flip-Flops kann zwischen 8 in schnell getakteten Datenpfaden von High-Performance-Schaltungen und bis zu 50 für Low-Power-Anwendungen liegen.

und Drain-Bulk-Spannung unverändert hoch sind. Auch die Gateleckströme in den Transistoren P1, P2 und N1 sind unverändert.

Darüber hinaus hat die Stärke des Stack-Effekts selbst ohne signifikante Gateleckströme im Normalfall nur einen geringen Einfluss auf den Leckstrom in komplexen Schaltungen. Dies wird am Beispiel einer Schaltung aus NAND2- und NOR2-Gattern deutlich: Die Wahrscheinlichkeit, dass in einem Gatter der Stack-Effekt wirksam wird, beträgt jeweils 1/4 (beide NAND-Eingänge A=B=0 bzw. NOR-Eingänge A=B=1). Das bedeutet, dass in drei von vier Gattern kein Stack-Effekt wirksam ist. Ein verstärkter Stack-Effekt, z.B. eine Verstärkung der Leckstromreduzierung von 1/10 auf 1/20 des ungestackten Stroms $I_{s,off}$, würde den Leckstrom der Gesamtschaltung nur unwesentlich verringern.

Gelingt es hingegen, die Anzahl der Stacks in der Schaltung zu erhöhen, kann der Leckstrom signifikant verringert werden. Die schaltungstechnischen Maßnahmen *Minimum Leakage Vector* [49] und *Forced Stack* [50] versuchen die passive Leistungsaufnahme auf diese Art zu reduzieren:

In jeder Schaltung weist der Leckstrom eine Abhängigkeit von den am Eingang anliegenden Daten (Eingangsvektor) auf. Als Schaltung wird in diesem Zusammenhang eine kombinatorische Logik-Schaltung in einem Datenpfad zwischen zwei Flip-Flops bezeichnet (Abb. 3.10). Zwar können die Eingangspotentiale der einzelnen Gatter ab der zweiten Logikstufe nicht mehr direkt eingestellt werden, jedoch kann durch eine geschickte Wahl des Eingangsvektors ein minimaler Leckstrom eingestellt werden. Da die Anzahl der möglichen Eingangsvektoren 2^n

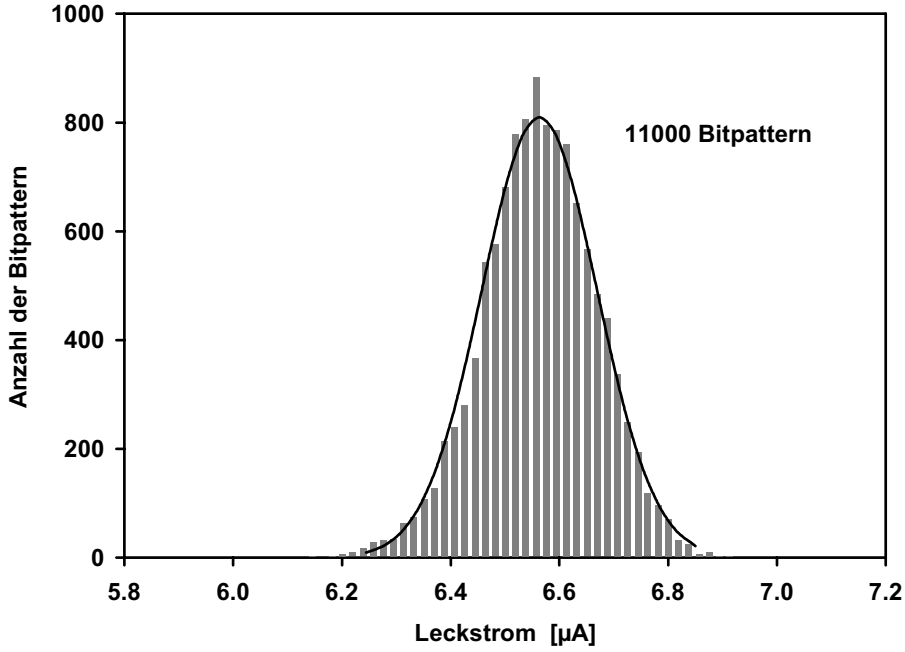


Abbildung 3.11: Leckstrom-Histogramm eines 32-bit-Multiplizierers für 11 000 zufällig ausgewählte Eingangsvektoren (Messwerte) [52]. Der Leckstrom hängt nur in sehr geringem Maße vom Eingangsvektor der Schaltung ab.

mit der Anzahl der Eingänge n schnell ansteigt, wurden Algorithmen entwickelt, die einen Minimum-Leakage-Vektor (MLV) annähernd auswählen können [34]. Während in Abbildung 3.10 maximal zwei Flip-Flops am Anfang einer Taktphase dargestellt sind, benötigt ein 32-bit-Addierer bereits 64 Eingangs-Flip-Flops. Der MLV kann im Standby-Modus mit Hilfe spezieller Flip-Flops eingeprägt werden oder durch serielles Beschreiben der Flip-Flops mit Hilfe des Scan-Pfads hergestellt werden [51]. Untersuchungen an einem Multiplizierer in einer 130 nm-Technologie [26, 52] zeigen jedoch nur eine geringe Leckstrom-Schwankungsbreite von 5–7 % bei 11 000 zufällig ausgewählten Eingangsvektoren (Abb. 3.11). MLV-Techniken werden mit zunehmender Datenpfadlänge weniger effizient, da sich die Eingangssignale von Gattern, die weit vom Eingangs-Flip-Flop entfernt sind, nur sehr indirekt beeinflussen lassen. Auch wenn der in [26] untersuchte Multiplizierer aufgrund seiner hohen Regularität möglicherweise nicht repräsentativ für alle Anwendungen ist, so deutet sich jedoch an, dass diese Technik trotz hohem design- und schaltungstechnischen Aufwand nur einen geringen Nutzen bringt. Dies gilt insbesondere für Low-Power-Schaltungen, die eine höhere Anzahl von Logikgattern pro Taktphase aufweisen.

Eine Möglichkeit, den Stack-Effekt besser nutzen zu können, stellt die Forced-Stack-Technik dar [50]. Hierbei werden Serientransistoren gezielt an Stellen hinzugefügt, an denen ansonsten nur ein ungestackter, ausgeschalteter Transistor einen hohen Leckstrom generieren würde, bevorzugt außerhalb des kritischen Pfades. Die zusätzlichen Transistoren werden im Standby-Modus über ein Standby-Signal abgeschaltet oder werden ebenfalls an das Eingangssignal des

zweiten Stack-Transistors angeschlossen. Insbesondere in Verbindung mit MLV-Techniken lässt sich in einigen Fällen eine signifikante Leckstromreduzierung erzielen. In [50] wird der Leckstrom um 35 bis 90 % reduziert.

Alle Techniken, die auf eine Reduzierung des Leckstroms durch Transistor-Serienschaltungen abzielen, verlieren jedoch mit der Zunahme von Gateleckströmen in Sub-100 nm-CMOS-Technologien stark an Effizienz. Es müssen daher Möglichkeiten gefunden werden, den Leckstrom auch in Schaltungen mit dünnen Oxiden zu reduzieren. Dies kann dadurch geschehen, dass ein regelmäßiger Schaltungsaufbau erzwungen wird, oder indem dickere und dünnere Gateoxide in einer Schaltung kombiniert werden. Zwei Möglichkeiten hierzu werden im Kapitel 5 vorgestellt.

Gate-Induced Drain Leakage

Gate-Induced Drain Leakage (GIDL) entsteht an der Drain-Junction in der Nähe der Silizium-Grenzfläche zum Dielektrikum. Aufgrund starker Bandverbiegungen kommt es zu Tunnelvorgängen, sowohl zum Tunneln über Störstellen als auch Band-zu-Band-Tunneln [53]. Der GIDL-Effekt wird durch größere Oxid-Feldstärken und höhere Dotierungen verstärkt. Besonders die lokale Erhöhung der Dotierung in der Nähe der Drain-Junction durch Halo-Dotierungen trägt dazu bei, dass der GIDL-Effekt in Sub-100 nm-Technologien kritisch ist.

Bei der Entwicklung neuer CMOS-Technologien wird stets darauf Wert gelegt, dass der GIDL-Effekt andere Leckstromkomponenten nicht signifikant übersteigt. Dieser Effekt stellt somit eher eine technologische Herausforderung bezüglich der Optimierung der Dotierprofile dar. Für die Entwicklung neuer Schaltungstechniken ist es jedoch erforderlich, die grundlegenden Abhängigkeiten des GIDL-Stroms zu kennen.

Aufgrund der komplexen Dotierstoffprofile sowie der Überlagerung von vertikalen und lateralen elektrischen Feldern an der Drain-Junction lässt sich der GIDL-Effekt nur empirisch beschreiben. In [54, 55] wird folgender Ansatz verwendet:

$$I_{gidl} = A_{gidl} \cdot V_{db} \cdot V_{tov}^2 \cdot \exp\left(-\frac{B_{gidl}}{V_{tov}}\right) \tag{3.12}$$

$$V_{tov} = \sqrt{V_{ov}^2 + (C_{GIDL} \cdot V_{db})^2} \tag{3.13}$$

Dabei sind A_{gidl}, B_{gidl} und C_{gidl} empirische Parameter, V_{ov} ist die Oxidspannung in der Overlap-Region und V_{db} die Drain-Bulk-Spannung. Besonders kritisch ist der GIDL in MuGFETs, da die Oxidspannung in der Drain-Overlap-Region aufgrund der von null verschiedenen Flachbandspannung höher ist.

3.1.2 Gate-Tunnelstrom in Bulk- und Multi-Gate-Transistoren

Mit Hilfe der in aktuellen Transistormodellen eingesetzten Beschreibungen des Tunnelstroms [55, 56] lassen sich Gateströme in Sub-100 nm-Transistoren zwar hinreichend gut beschreiben, im BSIM-4-Modell [56] sind dazu aber zum Beispiel mehr als 40 Gate-Tunnelstrom-Parameter

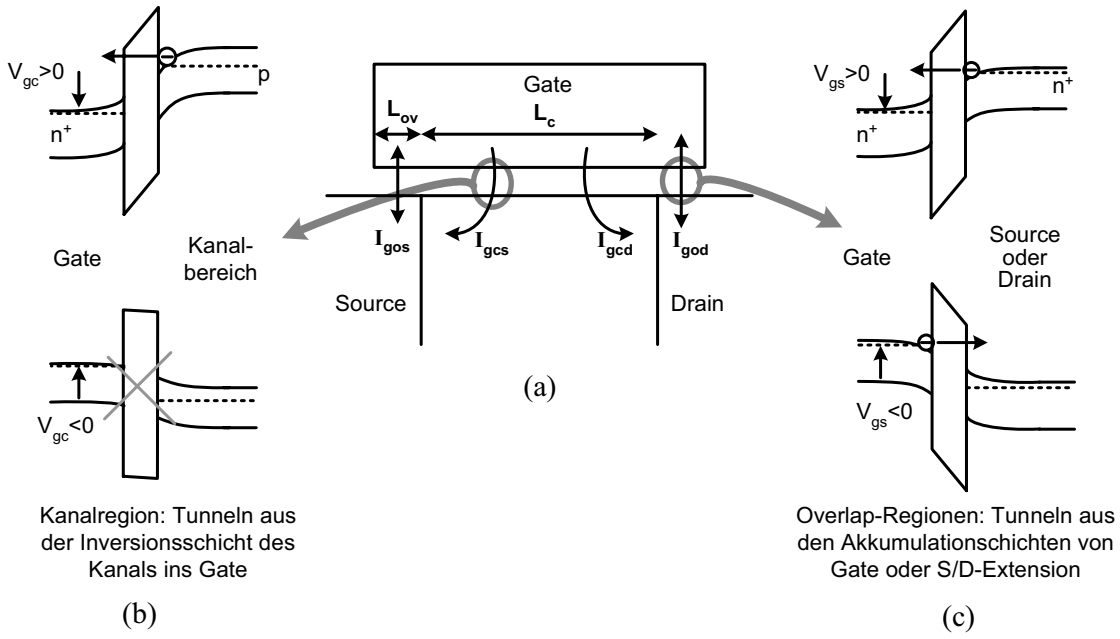


Abbildung 3.12: Tunnelstromkomponenten in MOSFETs (a). Im Kanalgebiet können Ladungsträger nur aus dem Kanalgebiet in Richtung des Gates tunneln (b). In den Overlap-Regionen können je nach Spannungsrichtung Ladungsträger in beide Richtungen tunneln (c).

erforderlich, deren ursprüngliche, physikalische Bedeutung nicht mehr erkennbar ist. In diesem Abschnitt wird ein Kompaktmodell zur Beschreibung des Tunnelstroms aus der grundlegenden Transmissionswahrscheinlichkeit hergeleitet (vgl. [57]).

Grundsätzlich wird bei der Tunnelstrommodellierung zwischen den Überlappungsbereichen (*Overlap*) zwischen Gate und Source/Drain sowie dem Kanalgebiet unterschieden (Abb. 3.12 und 3.13). Im Kanalgebiet können Ladungsträger nur aus dem Kanalgebiet in Richtung des Gates tunneln. Bei entgegengesetzter Spannung setzt in einem NFET erst ab einer Gate-Source-Spannung von etwa $V_{gs} = -1.5\,\mathrm{V}$ Akkumulation ein und es können Ladungsträger auch aus dem Gate in die Kanalregion tunneln. Da in Sub-100 nm-CMOS-Schaltungen so hohe Spannungen nicht auftreten, wird auf eine Modellierung dieses Tunnelprozesses verzichtet. Die Overlap-Regionen von Bulk-Transistoren sind bezüglich der Dotierungen in Gate und Source/Drain weitgehend symmetrisch. Daher können je nach Spannungsrichtung hier Ladungsträger in beide Richtungen tunneln.

Mit Hilfe der Transmissionsmatrixmethode [58] kann die Wahrscheinlichkeit berechnet werden, mit der ein Elektron durch eine rechteckige Barriere der Höhe φ_0 und der Dicke t_{ox} tunnelt. In Abhängigkeit von der Energie der Ladungsträger E ist die Transmissionswahrscheinlichkeit

$$\begin{aligned} T(E) &= \frac{1}{1+\frac{\varphi_0^2}{4E(\varphi_0-E)}\sinh^2\left(\sqrt{\frac{2m_{ox}}{\hbar^2}(\varphi_0-E)}t_{ox}\right)} \\ &\approx \frac{16E(\varphi_0-E)}{\varphi_0^2}\exp\left(-\gamma\sqrt{\varphi_0}\right)\cdot\exp\left(\frac{\gamma E}{2\sqrt{\varphi_0}}\right), \quad \gamma=\frac{4\pi t_{ox}\sqrt{2m_{ox}}}{h}. \end{aligned} \tag{3.14}$$

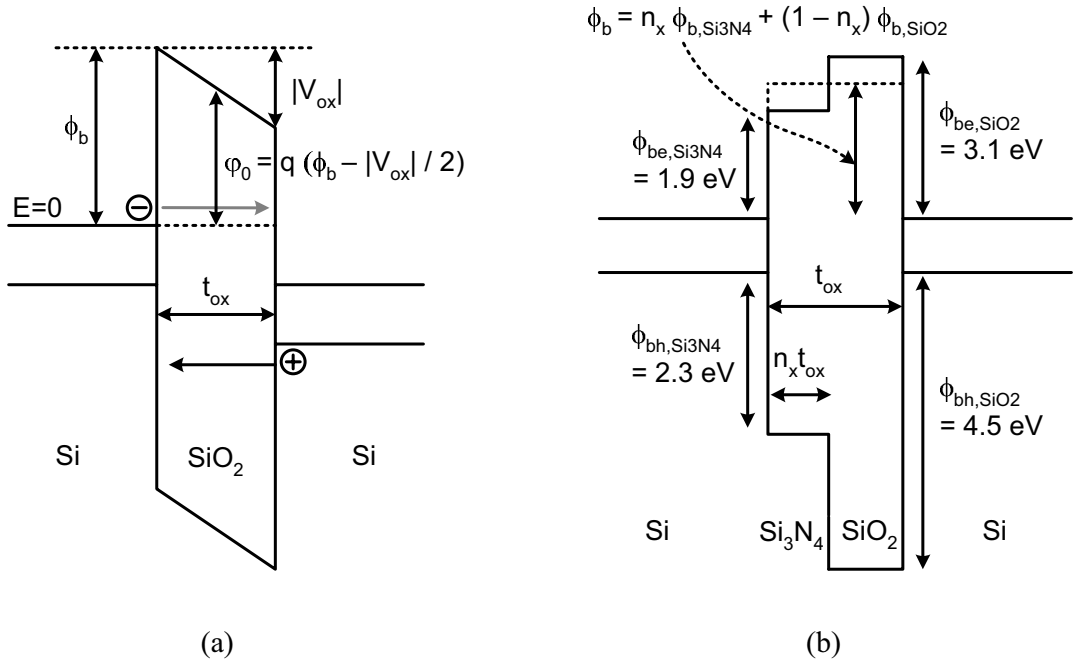


Abbildung 3.13: Bestimmung der durchschnittlichen Barrierenhöhe einer trapezförmigen Barriere bei Anliegen einer Oxidspannung $V_{ox} \neq 0$ (a) und in einem nitridierten Oxid (b).

Im Fall eines Feldeffekttransistors bildet das Gate-Dielektrikum diese Barriere, über dem im Allgemeinen eine Spannung V_{ox} anliegt, sodass sich eine trapezförmige Barriere ausbildet. Um Gleichung 3.14 auf diese trapezförmige Barriere anwenden zu können, wird eine mittlere Höhe berechnet (Abb. 3.12b, [59]):

$$\varphi_0 = q\left(\phi_b - \frac{|V_{ox}|}{2}\right). \tag{3.15}$$

Ein Vorteil der Annahme einer mittleren Barrierenhöhe φ_0 ist, dass sich diese nicht nur auf ein homogenes Oxid, sondern auch auf geschichtete Dielektrika anwendbar ist, z.B. ein High-κ-Dielektrikum auf einer dünnen $\mathrm{SiO_2}$-Schicht. Im Falle eines nitridierten Oxids (Oxinitrid, SiON) werden die physikalischen Parameter des Dielektrikums durch lineare Interpolation zwischen den Werten für $\mathrm{SiO_2}$ und $\mathrm{Si_3N_4}$ ermittelt (Abb. 3.12c):

$$\phi_b = n_x\phi_{b,\mathrm{Si_3N_4}} + (1 - n_x)\phi_{b,\mathrm{SiO_2}}. \tag{3.16}$$

Die Barrierenhöhen für $\mathrm{SiO_2}$ und $\mathrm{Si_3N_4}$ $\phi_{b,\mathrm{SiO_2}}$ und $\phi_{b,\mathrm{Si_3N_4}}$ sind dabei mit dem jeweiligen relativen Anteil am Dielektrikum $1 - n_x$ und n_x gewichtet. In vergleichbarer Weise wird eine mittlere Dielektrizitätskonstante berechnen:

$$\varepsilon_{\mathrm{SiON}} = n_x\varepsilon_{\mathrm{Si_3N_4}} + (1 - n_x)\varepsilon_{\mathrm{SiO_2}}. \tag{3.17}$$

Mit Hilfe der Transmissionswahrscheinlichkeit 3.14 lässt sich jetzt die Tunnelstromdichte basierend auf dem Ansatz von Esaki und Tsu [60] analytisch berechnen:

$$J_{tun} = \frac{4\pi q m^*}{h^3} \int_0^{\varphi_0} \left\{ \int_0^{\infty} f_1(E)\, dE_\perp \right\} T(E_z)\, dE_z. \tag{3.18}$$

Die Energie E ist die Summe der Energie in Tunnelrichtung E_z und der Energie senkrecht dazu $E_\perp$ [59]. $f_1(E)$ beschreibt die Verteilung der Ladungsträger, die durch die Fermi-Dirac-Statistik gegeben ist. Bei Temperaturen $T \gg 0$K kann diese durch die Boltzmann-Verteilung approximiert werden:

$$f_1(E) = \frac{1}{1+\exp\left(\frac{E-E_{F,1}}{k_B T}\right)} \approx \exp\left(-\frac{E-E_{F,1}}{k_B T}\right) = \exp\left(-\frac{E_z+E_\perp-E_{F,1}}{k_B T}\right). \tag{3.19}$$

Die obere Integrationsgrenze φ_0 lässt sich durch $\varphi_0 \to \infty$ ersetzen, da die Boltzmann-Verteilung für große Energien sehr klein wird. Damit ergibt sich für die Tunnelstromdichte in Abhängigkeit von der Oxidspannung [59]:

$$\begin{aligned} J_{tun} &= \frac{4\pi q m^*}{h^3} \int_0^{\varphi_0} \left\{ \int_0^{\infty} \exp\left(-\frac{E_\perp}{k_B T}\right) dE_\perp \right\} \exp\left(\frac{-E_z+E_{F,1}}{k_B T}\right) T(E_z)\, dE_z \\ &\approx \frac{4\pi q m^*}{h^3} k_B T \exp\left(\frac{E_{F,1}}{k_B T}\right) \frac{16}{\varphi_0^2} \exp\left(-\gamma\sqrt{\varphi_0}\right) \frac{\varphi_0}{b^2} \\ \text{mit} \quad b &= \frac{1}{k_B T}\left(1-\frac{\gamma k_B T}{2\sqrt{\varphi_0}}\right), \quad 1-\frac{2}{\varphi_0 b} \approx 1, \quad E_{F,1} = q\phi_s - q\phi_F - \frac{E_g}{2} \end{aligned} \tag{3.20}$$

In Gleichung 3.20 wird jetzt das Grenzflächen-Ferminiveau $E_{F,1} = q\phi_s - q\phi_F - \frac{E_g}{2} < 0$ berechnet. Dazu wird zunächst die elektrische Feldstärke im Halbleiter an der Grenzfläche zum Dielektrikum E_s in Akkumulation oder Inversion bestimmt:

$$E_s^2 = \frac{2 k_B T n_0}{\varepsilon_s} e^{\frac{q\phi_s}{k_B T}}. \tag{3.21}$$

Dabei ist die Ladungsträgerdichte im Substrat

$$n_0 = n_i e^{-\frac{q\phi_F}{k_B T}} \tag{3.22}$$

und die intrinsische Ladungsträgerdichte

$$n_i = 2\left(\frac{2\pi k_B T}{h^2}\right)^{\frac{3}{2}} (m_n^* m_p^*)^{\frac{3}{4}} e^{-\frac{E_g}{2k_B T}}. \tag{3.23}$$

Für die Feldstärke im Oxid $E_{ox} = V_{ox}/t_{ox}$ folgt damit

$$E_{ox}^2 = \frac{\varepsilon_s^2}{\varepsilon_{ox}^2} E_s^2 = \frac{2 k_B T \varepsilon_s}{\varepsilon_{ox}^2} \cdot 2\left(\frac{2\pi k_B T}{h^2}\right)^{\frac{3}{2}} (m_n^* m_p^*)^{\frac{3}{4}} \underbrace{\exp\left(\frac{q\phi_s - q\phi_F - \frac{E_g}{2}}{k_B T}\right)}_{=\exp(E_{F,1}/k_B T)}. \tag{3.24}$$

Daraus ergibt sich der in Gleichung 3.20 gesuchte Term

$$\exp\left(\frac{E_{F,1}}{k_B T}\right) = \frac{V_{ox}^2}{t_{ox}^2 \cdot \frac{2k_B T \varepsilon_s}{\varepsilon_{ox}^2} \cdot 2\left(\frac{2\pi k_B T}{h^2}\right)^{\frac{3}{2}} (m_n^* m_p^*)^{\frac{3}{4}}} \tag{3.25}$$

Somit folgt für die Tunnelstromdichte

$$J_{tun}(V_{ox}) = \frac{\sqrt{k_B T}}{\left(1 - \frac{\gamma k_B T}{2\sqrt{\varphi_0}}\right)^2} \cdot \frac{8 m^* q \varepsilon_{ox}^2}{\sqrt{2\pi}\varepsilon_s \varphi_0 (m_n^* m_p^*)^{\frac{3}{4}}} \frac{V_{ox} \cdot |V_{ox}|}{t_{ox}^2} e^{-\gamma\sqrt{\varphi_0}} \tag{3.26}$$

$$\text{mit} \quad \gamma = \frac{4\pi t_{ox}\sqrt{2 m_{ox}}}{h} \qquad \text{und} \quad \varphi_0 = q\left(\phi_B - \frac{|V_{ox}|}{q}\right).$$

Dabei sind m_n^* und m_p^* die mittleren effektiven Massen für Elektronen und Löcher im Leitungsband und im Valenzband, m^* die transversale Masse der Elektronen in Tunnelrichtung (bzw. die der Löcher für Löchertunneln), m_{ox} die Masse der tunnelnden Ladungsträger im Oxid sowie ε_s und ε_{ox} die Dielektrizitätskonstanten des Halbleiters und des Dielektrikums.

Zur Berechnung der Tunnelstromdichte muss jetzt die Oxidspannung V_{ox} in den Überlappungsgebieten und im Kanal bestimmt werden. In den Überlappungsgebieten wird die jeweilige, von außen anliegende Spannung (V_{gs} am Source-Kontakt und V_{gd} am Drain-Kontakt) um den Spannungsabfall in den Verarmungszonen korrigiert. In MuGFETs ist zudem die Flachbandspannung in der Overlap-Region zu berücksichtigen, die im Gegensatz zu Bulk-Transistoren nicht null ist. Die Verarmung tritt je nach Spannungsrichtung im Gate oder im Source/Drain-Gebiet auf. In den Source/Drain-Gebieten wird dazu eine durchschnittliche Dotierung angenommen [55].

In der Kanalregion ist die Oxidspannung ortsabhängig. Da auch die anderen Leckstromkomponenten in Abhängigkeit von der Schwellenspannung V_t modelliert werden, wird auch hier ein V_t-basiertes Modell verwendet, um die Oxidspannung zu bestimmen. Der in Abbildung 3.14 dargestellte Verlauf der Oxidspannung im Kanalbereich wird beschrieben durch

$$V_{gc}(x) = V_{gs} - \frac{2}{3} \cdot V_t - \frac{x}{L_c} \cdot V_{ds} \tag{3.27}$$

Der empirische Faktor $2/3$ berücksichtigt quantenmechanische Effekte im Kanal. Wie in den Überlappungsgebieten wird auch hier der Spannungsabfall in der Verarmungszone des Gates abgezogen (*Gate Depletion*):

$$V_{ox}^c(x) = \frac{n\, k_B T}{q} \ln\left(1 + \exp\left(\frac{q V_{gc}(x)}{n\, k_B T}\right)\right). \tag{3.28}$$

Zur Berechnung des weitenbezogenen Gatestroms muss die ortsabhängige Tunnelstromdichte $J_{tun}(V_{ox}^c(x))$ über die Kanallänge integriert werden. Die Integration ist analytisch jedoch nicht möglich. Deshalb wird hier eine einfache numerische Integration durchgeführt. Dazu wird die

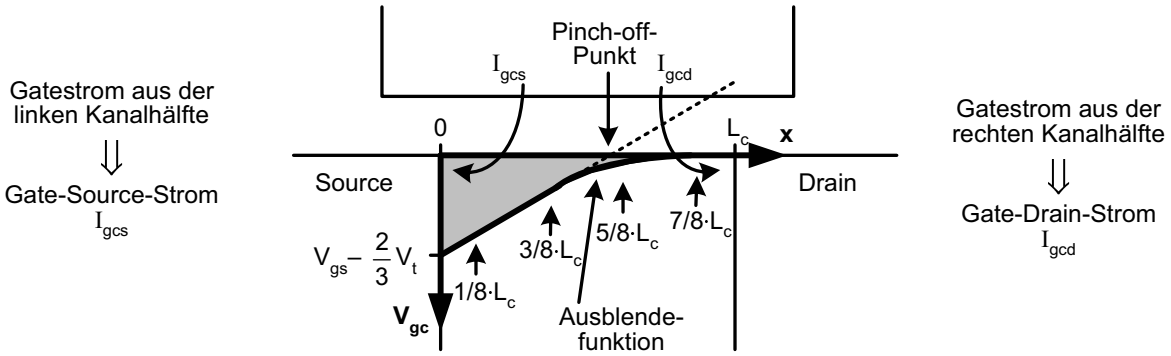


Abbildung 3.14: Verlauf des Kanalpotentials.

Tunnelstromdichte an vier Punkten im Kanal ausgewertet. Der Strom in der linken Hälfte des Kanals wird als Gate-Source-Strom betrachtet, der Strom in der rechten Hälfte als Gate-Drain-Strom:

$$I_{gcs} = L_c/4\left[J_{tun}\left(x = 1/8 \cdot L_c\right) + J_{tun}\left(x = 3/8 \cdot L_c\right)\right] \tag{3.29}$$

$$I_{gcd} = L_c/4\left[J_{tun}\left(x = 5/8 \cdot L_c\right) + J_{tun}\left(x = 7/8 \cdot L_c\right)\right].$$

Auch in den Überlappungsgebieten wird der weitenbezogene Strom durch Multiplikation der Tunnelstromdichte mit einer effektiven Überlappungslänge L_{ov} berechnet. Die Abbildungen 3.15 und 3.16 zeigen die gute Übereinstimmung des Modells mit Messungen an Bulk-Transistoren und MuGFETs.

Anders als in [56, 61] oder [55] ist die Tunnelstromdichte 3.26 abhängig von der Temperatur. Diese Temperaturabhängigkeit wird durch Messungen bestätigt (Abb. 3.17). Der Anstieg des Tunnelstroms, der je nach Transistor bei Anhebung der Temperatur von 300 K auf 400 K zwischen 20 % und 40 % beträgt, erscheint gegen die Erhöhung des Unterschwellenstroms vernachlässigbar klein (Abb. 3.4). Im Abschnitt 3.1.1 wird jedoch gezeigt, dass die Temperaturabhängigkeit mit der fortschreitenden Transistorskalierung abnimmt.

In Abbildung 3.18 sind alle Gateleckstrom-Komponenten eines Inverterpaars dargestellt. Neben den Gateleckströmen in den Drain-Overlap-Regionen der ausgeschalteten NMOS- und PMOS-Transistoren $I_{g,off}$ treten stets auch Tunnelströme in den eingeschalteten Transistoren $I_{g,on}$ auf. Da in diesen Transistoren ein Kanal ausgebildet ist, tunneln Ladungsträger nicht nur aus den beiden Overlap-Regionen, sondern auch aus der Kanalregion. Unterschwellenströme und GIDL-Ströme treten hingegen nur in den ausgeschalteten Transistoren auf. Abhängig von der Gatelänge ist $I_{g,on}$ in der Regel um einen Faktor 5 bis 10 größer als $I_{g,off}$. Aufgrund der für die meisten Dielektrika höheren Barriere für Löcher als für Elektronen ist der Gateleckstrom in den PMOS-Transistoren kleiner als in NMOS-Transistoren. Die PMOS-Transistoren werden jedoch wegen des kleineren On-Stroms stets größer ausgelegt.

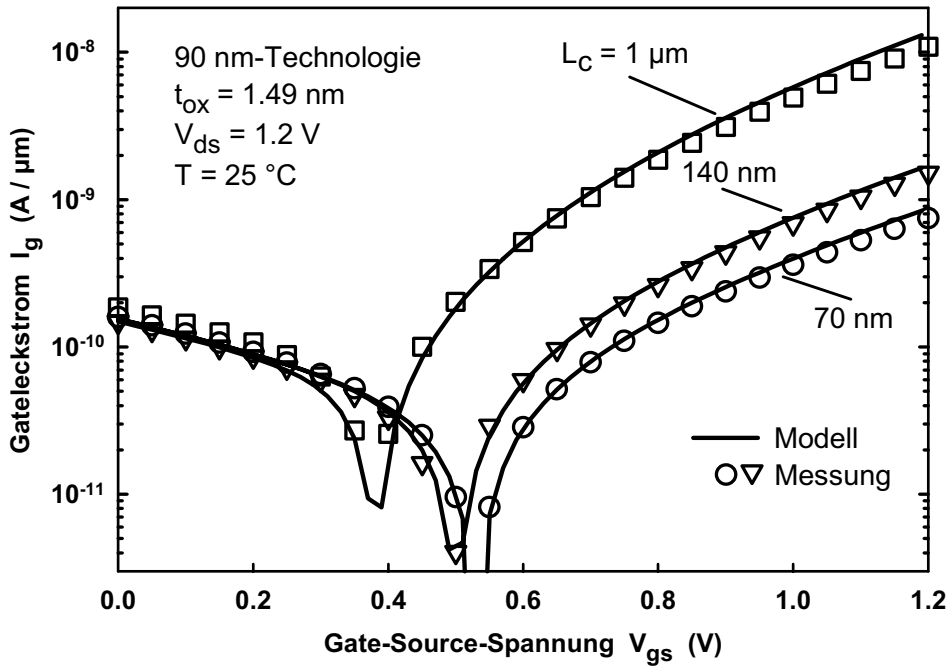


Abbildung 3.15: Modellierter und gemessener Gateleckstrom in Bulk-Transistoren bei Variation der Gatelänge.

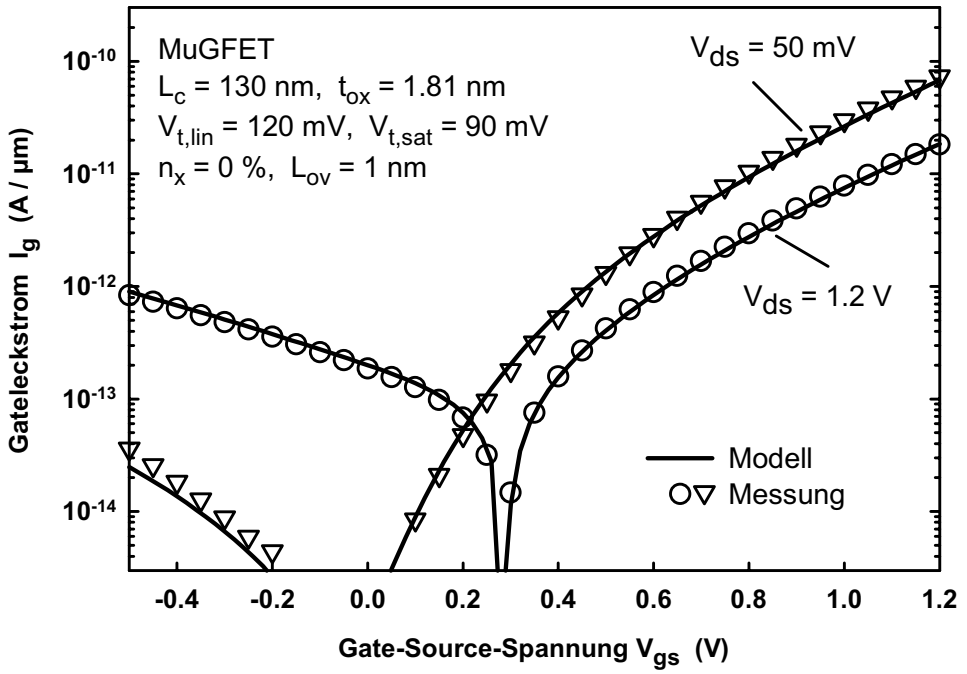


Abbildung 3.16: Modellierter und gemessener Gateleckstrom in Bulk-Transistoren in einem Multi-Gate-Transistor.

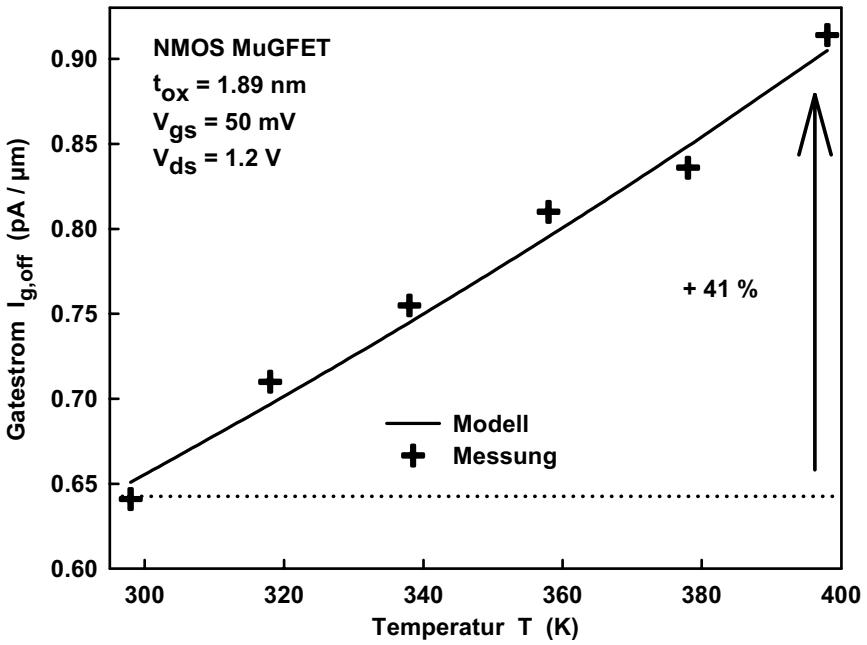


Abbildung 3.17: Einfluss der Temperatur auf den Gateleckstrom in einem MuGFET.

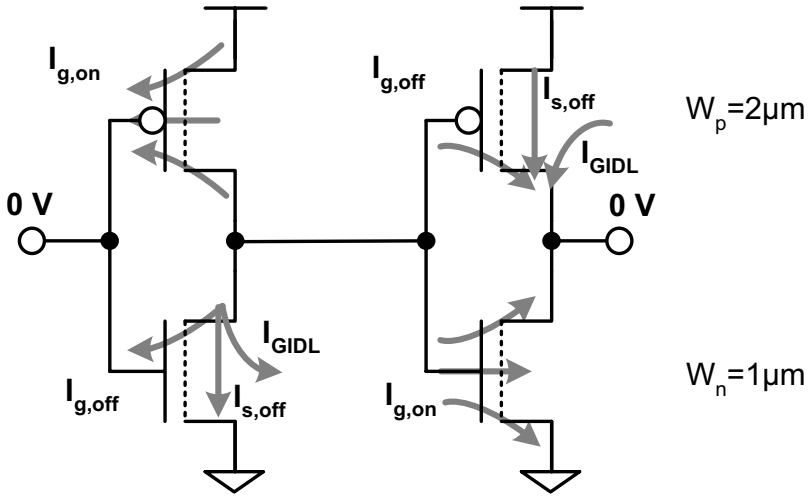


Abbildung 3.18: Leckströme in einem Inverterpaar.

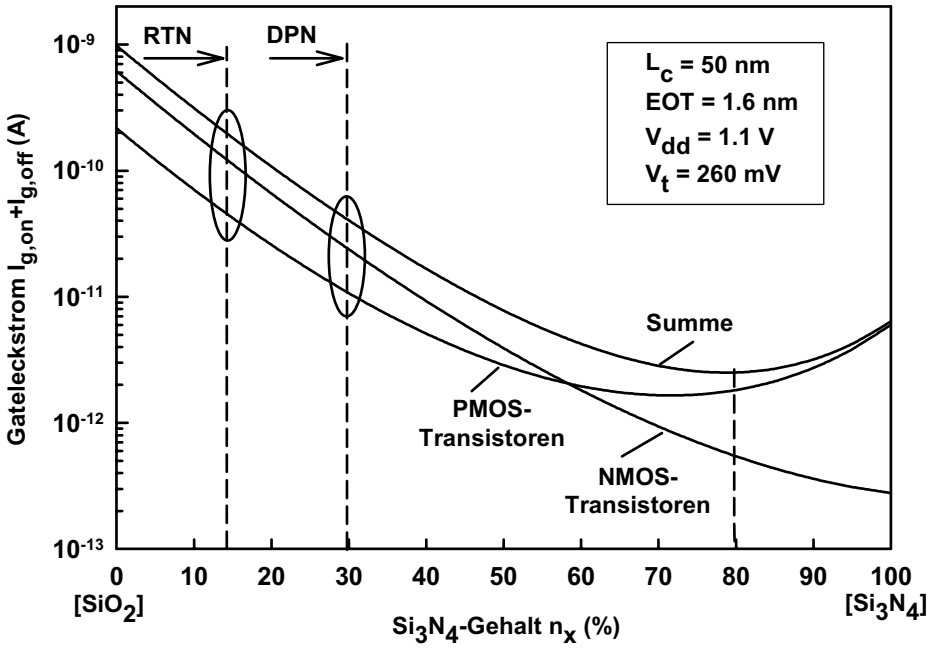


Abbildung 3.19: Einfluss der Nitridation auf den Gateleckstrom eines Inverterpaars in einer typischen 65 nm-Technologie. Die äquivalente Oxiddicke $EOT = \varepsilon_{\mathrm{SiO_2}}/\varepsilon_{\mathrm{SiON}} \cdot t_{ox}$ wird durch Anpassung von t_{ox} konstant gehalten. Der Gatestrom erreicht erst bei $n_x = 80\,\%$ ein Minimum. Zur Zeit werden nicht mehr als $n_x = 30\,\%$ erreicht (Rapid Thermal Nitridation, RTN oder Decoupled Plasma Nitridation, DPN).

Schon in der 90 nm-CMOS-Technologie kann die Gateoxiddicke nicht mehr skaliert werden, ohne dass der Tunnelstrom stark ansteigt. Da auch in den nächsten Jahren keine High-κ-Dielektrika einsatzfähig sein werden, kann die Dielektrizitätskonstante lediglich durch Nitridierung des Gateoxids leicht erhöht werden. Mit Hilfe von *Rapid Thermal Nitridation* (RTN) lassen sich bis zu 14 % Nitridanteil im Oxid erzielen, ohne die Beweglichkeit der Ladungsträger im Kanal herabzusetzen. Mit *Decoupled Plasma Nitridation* (DPN) lassen sich bis zu 30 % erreichen [62]. 100 % Nitridation ($\mathrm{Si_3N_4}$) entspricht einer atomaren Stickstoff-Konzentration von 57 at-%.

Die Nitridierung hat zwei Einflüsse auf den Tunnelstrom: Zum einen erhöht sich die Dielektrizitätskonstante, sodass das Dielektrikum bei gleicher elektrischer Dicke $EOT = \varepsilon_{\mathrm{SiO_2}}/\varepsilon_{\mathrm{SiON}} \cdot t_{ox}$ dicker ausgelegt werden kann und der Tunnelstrom sinkt. Zum anderen wird jedoch auch die Barrierenhöhe reduziert (Abb. 3.13b), sodass der Tunnelstrom wieder steigt. Dies gilt insbesondere für PMOS-Transistoren, da die Barriere für Löcher ϕ_{bh} stärker degradiert als die für Elektronen ϕ_{be}. Abbildung 3.19 zeigt den Tunnelstrom zweier Inverter in 65 nm-Technologie gemäß ITRS-Roadmap [3]. Es ist zu erkennen, dass eine fortschreitende Erhöhung des Nitridgehalts weiterhin dazu geeignet ist, den Gatestrom zu reduzieren.

3.1.3 Aktive Leistungsaufnahme

Im Gegensatz zu den Leckströmen nimmt die aktive Leistungsaufnahme je Gatter mit der Skalierung der Transistordimensionen ab. Da sich jedoch die Anzahl der Logikgatter je Flächeneinheit mit jeder Technologiegeneration verdoppelt, steigt die Leistungsdichte der aktiven Schaltströme dennoch an. Der Skalierungstrend ist in Abbildung 3.20 dargestellt. Der Anstieg der aktiven Leistungsdichte ist gemessen an der Zunahme der Unterschwellenströme relativ klein.

Im Gegensatz zu Leckströmen, die zwar auch eine Zustandsabhängigkeit aufweisen, im Mittel aber über die Zeit konstant sind, tritt die aktive Leistungsaufnahme nur während der sehr kurzen Zeit des Schaltvorgangs auf. Deshalb ist die Schaltaktivität des einzelnen Logikgatters von großer Bedeutung. Diese richtet sich erstens nach der Frequenz der gesamten Schaltung, zweitens nach der Wahrscheinlichkeit, dass eines der Eingangssignale eines Gatters seinen Zustand ändert und drittens nach der Wahrscheinlichkeit, dass ein Gatter bei dem Wechsel eines Eingangssignals auch seinen Ausgangszustand ändert. So schaltet z.B. das Ausgangssignal eines NAND4-Gatters selbst bei statistisch gleichverteilten Eingangssignalen nur in einem von acht Fällen.

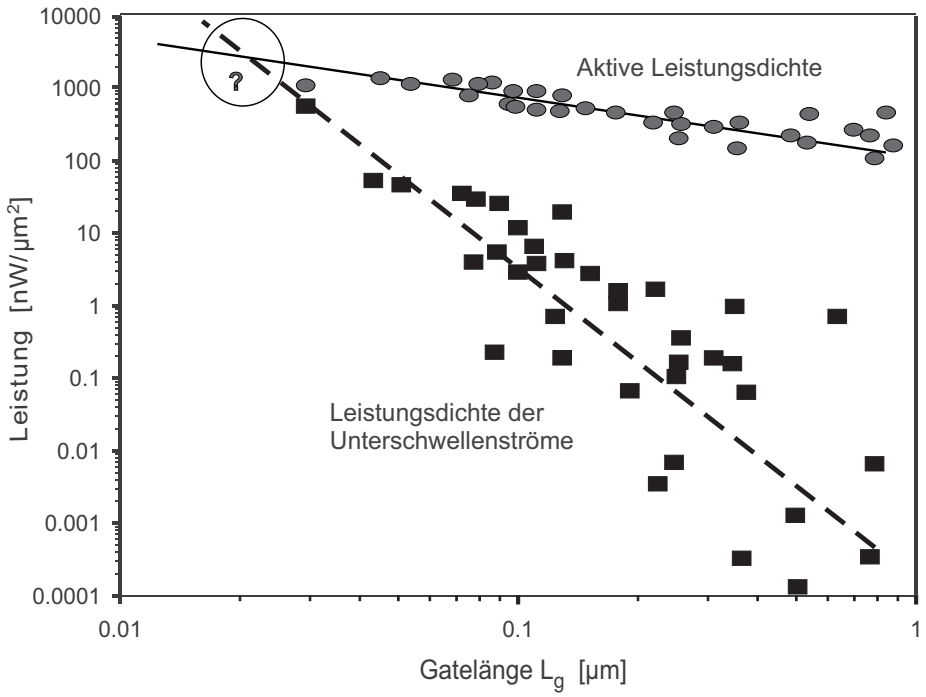


Abbildung 3.20: Extrapolationen des industriellen Trends der Leistungsaufnahme durch aktive Schaltvorgänge und Unterschwellenströme von Hochgeschwindigkeits-CMOS-Anwendungen bei 25 °C [2]. Bei erhöhter Temperatur und unter Einbeziehung von Gateleckströmen erreichen die Leckströme die aktive Leistungsaufnahme schon sehr viel früher.

Besonders häufig sind aktive Schaltereignisse im Taktverteilungsbaum sowie in den Flip-Flops einer getakteten Logikschaltung, da hier in jedem Taktzyklus eine große Anzahl an Transistoren geschaltet wird. So macht die aktive Leistungsaufnahme der Flip-Flops in vielen Anwendungen einen signifikanten Anteil an der Gesamtleistung aus, sodass insbesondere bei Latches und Flip-Flops auf die Schaltenergie geachtet werden muss (Kapitel 5.3). Im Gegensatz dazu sind Schaltereignisse in großen SRAM-Blöcken weitaus seltener. Hier spielt die aktive Leistungsaufnahme meist nur eine untergeordnete Rolle.

Es wird deutlich, dass die Aktivität eines Logikgatters sehr stark von der Anwendung abhängt. Während hoch getaktete Datenpfade mit weniger als 10 Logikgattern pro Taktphase Aktivitäten von 10 % bei Frequenzen über 3 GHz erreichen (entsprechend 300 Millionen Schaltereignissen pro Sekunde), liegt die Aktivität einer SRAM-Zelle oder die eines Logikgatters in einer Low-Power-Schaltung bei reduzierter Taktfrequenz um viele Größenordnungen niedriger. Dieser Unterschied in der Aktivität und damit in der pro Gatter benötigten durchschnittlichen Schaltenergie begründet die Existenz unterschiedlicher Transistortypen mit Leckströmen, die sich ebenfalls um viele Größenordnungen unterscheiden.

Die aktive Leistungsaufnahme lässt sich in das Umladen von Gate- und Parasitär-Kapazitäten sowie in Querströme aufteilen. Das Umladen der Gate-Kapazitäten stellt dabei die wichtigste Komponente dar. Die Eingangskapazität eines Logikgatters lässt sich verhältnismäßig einfach bestimmen und ist ein wichtiges Charakterisierungsmerkmal eines Gatters für die Schaltungssynthese. Die Kapazität eines Transistors ist jedoch für das Schaltungsdesign als technologisch gegebener Parameter anzusehen, der durch das Layout nur sehr begrenzt beeinflusst werden kann. Selbst die Verwendung unterschiedlicher Gatelängen und Oxiddicken hat nur begrenzte Auswirkungen, da Junction- und Parasitär-Kapazitäten einen immer größeren Anteil ausmachen. Lediglich die Reduzierung der Transistorweiten in nicht-kritischen Pfaden – zum Beispiel die internen Rückkopplungen in einem Flip-Flop – kann signifikant zur Energieeinsparung beitragen.

Der Anteil der Leistungsaufnahme, der auf das Umladen der Leitungskapazitäten entfällt, ist schwer einzuschätzen, mit fortschreitender Skalierung aber zunehmend wichtig. Der minimale Leitungsabstand sinkt mit jeder Technologie-Generation, wodurch sich die Kapazitäten zwischen den Leitungen trotz Einführung von Low-κ-Intermetalldielektrika erhöhen.

Querströme fließen, wenn während des Schaltvorgangs kurzzeitig sowohl der NMOS- als auch der PMOS-Pfad eines Logikgatters leitend sind. Dazu müssen die Gatespannungen von N- und PFET gleichzeitig $|V_{gs}| > |V_t|$ sein. Da mit fortschreitender Technologie-Skalierung der Bereich, in dem Querströme auftreten ($V_{dd} - 2 \cdot V_t$), immer kleiner wird, spielt diese Komponente der Leistungsaufnahme in Sub-100 nm-Schaltungen nur noch eine untergeordnete Rolle. Nur wenn der Eingang eines Logikgatters sehr langsam schaltet, ist eine signifikante Erhöhung der Leistungsaufnahme festzustellen, etwa wenn die Eingangsflanke 5 bis 10 mal langsamer als die nominelle Schaltgeschwindigkeit ist. Dieser Zustand sollte in einer gut dimensionierten Digitalschaltung jedoch nur selten auftreten. Abhängig von den gewählten Versorgungs- und Schwellenspannungen sowie der Temperatur liegt der Anteil des Querstroms in Sub-100 nm-Technologien nicht über 10 %.

3.2 Schaltgeschwindigkeit digitaler CMOS-Schaltungen

Die Entwicklung und Bewertung von Schaltungstechniken zur Erhöhung der Schaltgeschwindigkeit oder Reduzierung des Leckstroms erfordern ein grundlegendes Verständnis des digitalen Schaltverhaltens. In diesem Abschnitt werden daher Methoden vorgestellt, die Geschwindigkeit einer Schaltung, insbesondere mit gestackten Device-Anordnungen, aus den Charakteristiken der einzelnen Transistoren abzuleiten.

Beim Betrieb statischer CMOS-Schaltungen werden alle internen Knoten stets wechselweise auf die vollen Betriebspotentiale V_{dd} oder V_{ss} aufgeladen oder entladen. Im Gegensatz zu Schaltungstechniken mit reduziertem Spannungshub (*Reduced Swing Logic*, z.B. [63, 64]), erlaubt die hohe Störsicherheit statischer CMOS-Logik einen Einsatz in System-on-Chip-Umgebungen und in automatisierten Designumgebungen. Die Schaltgeschwindigkeit hängt von zwei Faktoren ab. Dabei ist der On-Strom, den der einschaltende Transistor in seinem jeweiligen Betriebspunkt liefern kann, ins Verhältnis zu der Kapazität zu setzen, die umgeladen werden muss. Diese Kapazität setzt sich aus den Kapazitäten des treibenden Gatters (Gate- und Junction-Kapazitäten), den Parasitärkapazitäten der Metall-Leitungen sowie den Kapazitäten der zu treibenden Transistoren (Gate-, Junction- und Parasitärkapazitäten, Abb. 3.21) zusammen. Alle diese Parameter unterliegen Prozess-Schwankungen und sind zudem in jeder Schaltung unterschiedlich gewichtet.

Die Analyse der Geschwindigkeit einer Schaltung kann auf verschiedenen Ebenen erfolgen. Auf der einen Seite geben die elektrischen Parameter eines Einzeltransistors einen messtechnisch einfachen Einblick in die physikalischen Ursachen der Schaltgeschwindigkeit. Auf der anderen Seite bildet nur die aufwändige Messung an einer komplexen Schaltung das Geschwindigkeitspotential eines bestimmten Dies oder Wafers richtig ab, ohne jedoch einen Einblick in die Ursachen zu erlauben. Dazwischen bieten Ringoszillatoren die Möglichkeit, ein komplexes Schaltverhalten schnell und präzise durch Messung einer einfachen Testschaltung zu charakterisieren.

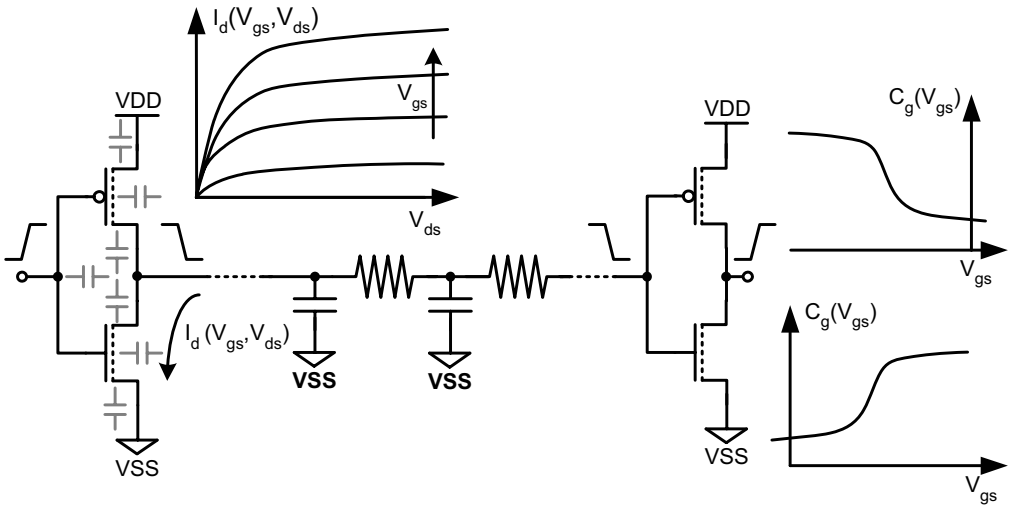


Abbildung 3.21: Schematische Darstellung eines Schaltvorgangs in statischer CMOS-Logik.

3.2.1 Inverter-Verzögerungszeit als Funktion des On-Stroms

Allgemein wird bei einem Schaltvorgang in einer CMOS-Schaltung eine Kapazität über einen Transistor umgeladen. Der einfachste Ansatz zur Abschätzung der mittleren Schaltdauer τ ist daher, das Verhältnis zwischen On-Strom und Kapazität zu bestimmen:

$$\tau \propto \frac{V_{dd}C}{I_d(V_{gs} = V_{ds} = V_{dd})}. \tag{3.30}$$

Der Strom I_d eines Transistors in Sättigung lässt sich beschreiben durch:

$$I_d \propto \mu_n C'_{ox} \frac{W}{L}(V_{gs} - V_t)^{\alpha}. \tag{3.31}$$

Daraus folgt das bekannte Alpha-Exponential-Gesetz (*Alpha Power Law*) [65]:

$$\tau_\alpha \propto \frac{V_{dd}C}{(V_{dd} - V_t)^{\alpha}}. \tag{3.32}$$

Der Exponent α beschreibt die Sättigung der Ladungsträgergeschwindigkeit. Nach dem ursprünglichen Shockley-Modell ist $\alpha = 2$. In den Technologiegenerationen des *Constant Voltage Scalings* von der 0.5 μm- bis zur 130 nm-Technologie bleiben die maximal in den Transistoren auftretenden Felder bei gleichzeitig skalierten Abmessungen und Spannungen konstant. α ist hier weitgehend konstant $\alpha \approx 1.3$ [66]. In Sub-100 nm-Technologien werden die elektrischen Felder jedoch wieder größer, sodass $\alpha \approx 1$ ist und der Strom fast nur noch linear von der Spannung $V_{dd} - V_t$ abhängt.

In Gleichung 3.32 wird angenommen, dass während des gesamten Schaltvorgangs $V_{gs} = V_{ds} = V_{dd}$ gilt. In einer statischen CMOS-Schaltung hat das Eingangs-Signal jedoch immer eine endliche Steigung. Da die Steigungen von Ein- und Ausgangssignal im Mittel gleich groß sind, wird zunächst dieser Fall untersucht.

Sind die Steigungen von Eingangs- und Ausgangssignal gleich, so durchläuft der schaltende Transistor die in Abbildung 3.22 dargestellte Trajektorie, die bei den nachfolgenden Betrachtungen ein zentrales Instrument zur Beurteilung des Schaltverhaltens digitaler Logikgatter bildet. Es ist zu erkennen, dass schon während $V_{gs} < V_{dd}$ ein signifikanter Strom fließt. Die Trajektorie erreicht die Kennlinie mit maximaler Gate-Spannung ($V_{gs} = V_{dd}$) erst bei etwa $V_{dd}/2$. Daher wird in [67] ein effektiver On-Strom I_{eff} berechnet, der als proportional zur Schaltgeschwindigkeit eines CMOS-Gatters angesehen werden kann:

$$I_{eff} = \frac{I_L + I_H}{2} \tag{3.33}$$

$$\text{mit } I_L = I_{ds}\left(V_{gs} = \frac{V_{dd}}{2}, V_{ds} = V_{dd}\right) \text{ und } I_H = I_{ds}\left(V_{gs} = V_{dd}, V_{ds} = \frac{V_{dd}}{2}\right). \tag{3.34}$$

Es wird deutlich, dass der maximale On-Strom des Transistors, der bei $V_{ds} = V_{gs} = V_{dd}$ fließt, in einem realen CMOS-Gatter nie erreicht wird. Dieses hat verschiedene Auswirkungen auf die Schaltgeschwindigkeit. Insbesondere resultiert der DIBL-Effekt, der bei $V_{ds} = V_{dd}$ maximal ist

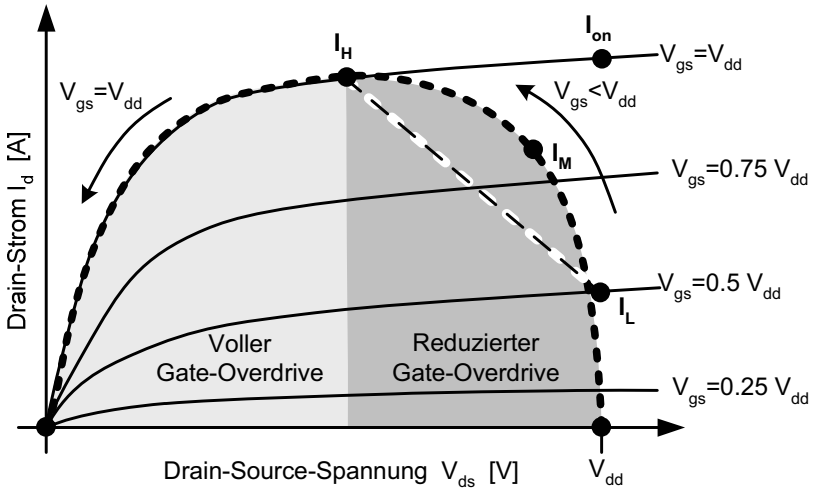


Abbildung 3.22: Zustand eines NMOS-Transistors während eines Ausschaltvorgangs, entsprechend einem logischen 1–0-Übergang am Ausgang eines CMOS-Inverters. Der PMOS-Transistor, der von dem leitenden in den sperrenden Zustand übergeht, stellt während des 1-0-Übergangs lediglich eine kapazitive Last dar und hat damit keinen Einfluss auf die dargestellte Schalttrajektorie.

und den On-Strom in diesem Bereich stark erhöht, nur in einer relativ schwachen Geschwindigkeitserhöhung, da der maximale Strom bei $V_{ds} \approx V_{dd}/2$ fließt.
Gleichung 3.33 kann noch weiter vereinfacht werden, indem anstelle der beiden Ströme I_L und I_H nur ein einziger repräsentativer Strom I_M betrachtet wird, der gemessen wird bei

$$I_M = I_d(V_{ds} = V_{gs} = 0.8 \cdot V_{dd}). \tag{3.35}$$

Die Schaltgeschwindigkeit eines Inverters hängt neben dem On-Strom im Wesentlichen von den umzuladenen Kapazitäten ab. Neben der Last am Inverter-Ausgang (Fan-out und parasitäre Leitungskapazitäten) sind dies die Gate- und Junction-Kapazitäten sowohl des einschaltenden als auch des ausschaltenden Transistors. Außerdem muss der schaltende Transistor die Verschiebungsladung aufnehmen, die sich beim Schalten des Eingangs ergibt, wiederum für den einschaltenden und den ausschaltenden Transistor.
Der auf externe Lasten zurückzuführende Anteil der Verzögerungszeit wird als extrinsisches Delay t_{ext} bezeichnet. Dieser ist weitgehend proportional zur Last. In Abbildung 3.21 ist die Leitung zwischen den Logikgattern als verteiltes RC-Netzwerk dargestellt. In allen hier betrachteten Fällen ist es jedoch ausreichend, nur eine Kapazität sowie nur einen Widerstand als Ersatzschaltbild anzunehmen. Erst bei Verdrahtungslängen von etwa 1 mm, die in Bus-Leitungen, aber nicht innerhalb eines Schaltungsblocks auftreten, ist es erforderlich, mehrere RC-Glieder zu betrachten.
Die Kapazitäten des schaltenden Inverters selbst führen zu einer intrinsischen Verzögerungszeit t_{int}, die bei konstantem Weitenverhältnis zwischen N- und PFET prinzipiell unabhängig von den verwendeten Transistorweiten ist. Nur bei kleinen Weiten nahe W_{min} führt der erhöhte Anteil der Parasitärkapazitäten zu einer erhöhten Verzögerungszeit. Die intrinsischen Verzögerungszeiten können innerhalb einer Technologie für einen Transistortyp als konstant angesehen

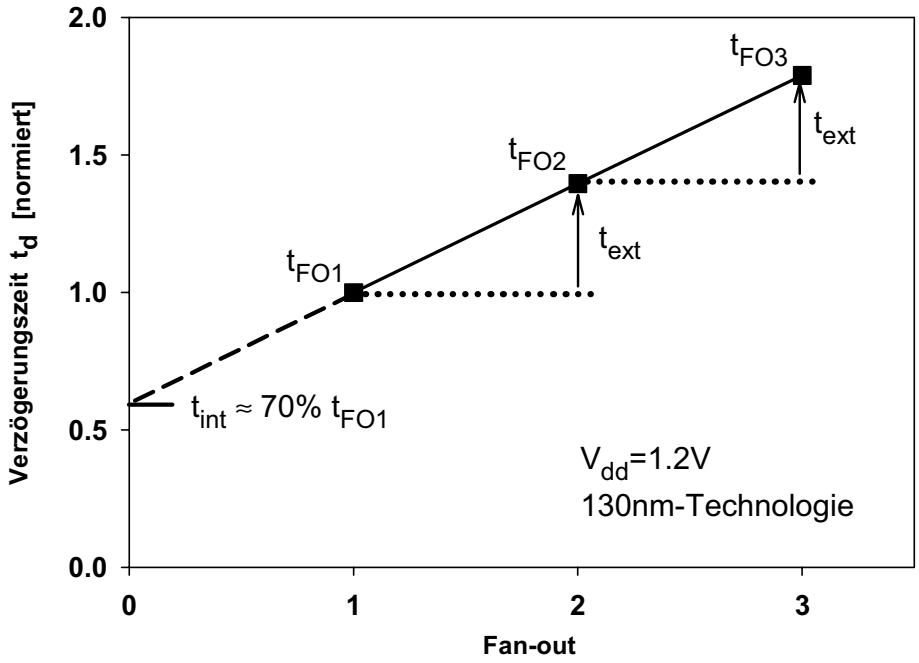


Abbildung 3.23: Gemessene mittlere Verzögerungszeit für 1–0- und 0–1-Übergang in Abhängigkeit vom Fan-out in 130 nm-CMOS-Ringoszillatoren. Die Erhöhung des Fan-outs um 1 führt jeweils zu einer Delay-Erhöhung von t_{ext}. Die Extrapolation der Geraden auf einen theoretischen Fan-out-0-Wert entspricht dem intrinsischen Delay t_{int}. Die Messungen stammen aus der im Anhang B beschriebenen Teststruktur.

werden, da sie sich aus dem Verhältnis von On-Strom und Kapazitäten des Transistors selbst bestimmen. Wird die Verzögerungszeit in Abhängigkeit vom Fan-out aufgetragen, kann der intrinsische Anteil durch Extrapolation der Kennlinie bestimmt werden (Abb. 3.23). Dieser Wert für einen theoretischen Fan-out von 0 beträgt in der 130 nm-Technologie ca. 70 % des Fan-out-1-Wertes t_{FO1}.

In Abbildung 3.24 sind die Ein- und Ausgangssignale eines Inverters für unterschiedliche Eingangs- und Ausgangsbelastungen (Fan-in und Fan-out) gegenüber gestellt. Der Fan-in gibt an, wie viele Gatter sich das Eingangssignal des betrachteten Gatters teilen; ein hoher Fan-in bedeutet eine langsame Eingangsflanke (diese Definition unterscheidet sich von des des gatterbezogenen Fan-ins, bei dem z.B. ein Fan-in von 4 ein NAND4- oder NOR4-Gatter bezeichnet). Der Fan-out bezeichnet die Anzahl der Gatter, die an den Ausgang des Gatters angeschlossen sind. Ein hoher Fan-out bedeutet eine hohe Ausgangskapazität und damit eine langsame Ausgangsflanke. Insbesondere die Fälle, in denen Fan-in und Fan-out gleich groß sind (Abb. 3.24a, c und e), werden gut durch Gleichung 3.33 beschrieben.

Im Falle einer steilen Eingangsflanke und einer großen Ausgangslast (Abb. 3.24d) verläuft die Trajektorie näher an der $V_{gs} = V_{dd}$-Kennlinie. Der maximale Schaltstrom liegt näher an I_{on}. Im Gegensatz dazu verläuft die Trajektorie bei hohem Fan-in und kleiner Ausgangslast sehr flach (Abb. 3.24b). Der Schaltvorgang ist beim Erreichen von $V_{gs} = V_{dd}$ bereits weitgehend abgeschlossen.

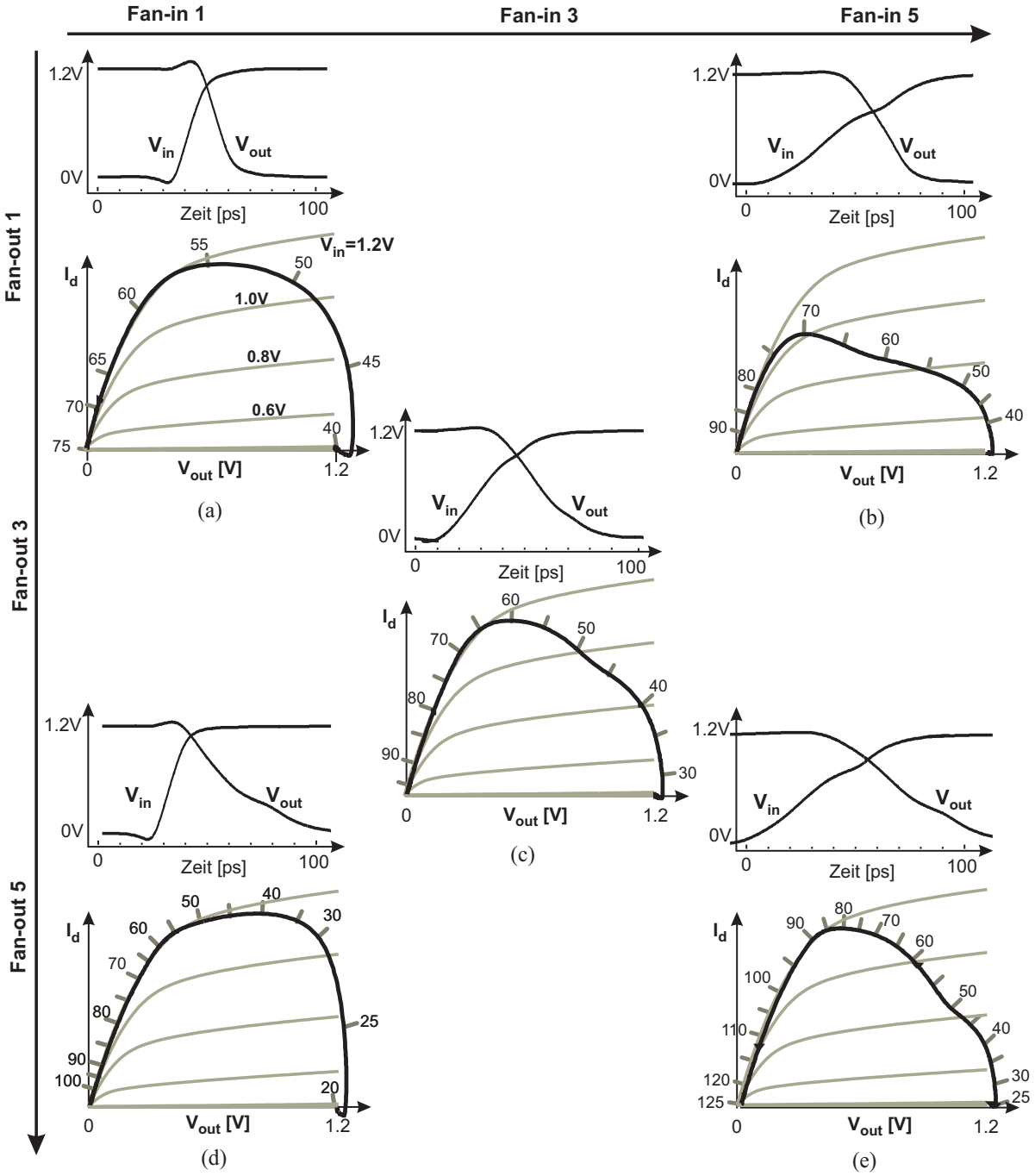


Abbildung 3.24: NMOS-Schalttrajektorien und Zeitdiagramme in einem symmetrischen Inverter bei unterschiedlichen Eingangs- und Ausgangsbelastungen (Schaltungssimulation einer 90 nm-Technologie, REG, 27 °C). Die Zahlen an den Trajektorien geben die parametrische Zeit in Picosekunden an und beziehen sich jeweils auf die Zeitskala im dazu gehörigen Timing-Diagramm. Die absoluten Werte der Zeiten lassen sich nicht direkt untereinander vergleichen, da der Zeitpunkt des betrachteten Schaltereignisses von den Verzögerungszeiten der vorgeschalteten Gatter abhängt. Die Trajektorien in (a), (c) und (e) erreichen die Kennlinie für $V_{in} = 1.2\,\mathrm{V}$ bei gleichem V_{out}, da jeweils Fan-in und Fan-out gleich groß sind. Lediglich die Geschwindigkeit, mit der die Trajektorie durchlaufen wird, nimmt mit wachsenden Belastungen ab (Fall (a) 75–40=35 ps und Fall (d) 125–25=100 ps).

In einer realen Logik treten die beiden Fälle (b) und (d) immer mit gleicher Häufigkeit auf, da die Flankensteilheit nicht von Stufe zu Stufe kontinuierlich kleiner werden kann. Zudem gleichen die in automatisierten Design-Umgebungen verwendeten Tools die Flankensteilheit weitgehend durch die Wahl von Gattern mit geeigneter Treiberstärke an. Im Mittel ergibt sich damit ein Schaltverhalten nach Trajektorie (c).

In einer gut dimensionierten Schaltung wird das Schaltverhalten daher stets durch die Trajektorie in Abbildung (c) beschrieben. Zur Veranschaulichung einzelner Effekte kann es jedoch zweckmäßig sein, auch andere Fälle zu untersuchen. So wird im Kapitel 4 die Trajektorie (d) bei verschiedenen Substrat-Spannungen betrachtet.

Im Gegensatz zu der schematische Trajektorie in Abbildung 3.22 zeigen die simulierten Trajektorien in Abbildung 3.24, insbesondere (b), (c) und (e), einen nicht gleichmäßig runden Verlauf. Dieser ergibt sich aus der Spannungsabhängigkeit der Eingangskapazitäten der zu treibenden Gatter (Abb. 3.21). Bei $V_{dd}/2$ ist die Eingangskapazität der Fan-out-Gatter maximal, da hier die Summe der CV-Kurven von NFET und PFET ein Maximum ausbildet [68].

3.2.2 Transistor-Dimensionierung in CMOS-Schaltungen

Für die Berechnung der Schaltzeiten für einen 1–0-Übergang t_n und für einen 0–1-Übergang t_p wird ein einfaches Modell verwendet, welches das Schaltverhalten digitaler Schaltungen anschaulich und hinreichend genau beschreibt. Bei konstanter Versorgungsspannung ergeben sich aus Gleichung 3.30 die Verzögerungszeiten

$$t_n \propto \frac{C_l + C_i + C_p}{W_n \cdot I_n} \tag{3.36}$$

$$t_p \propto \frac{C_l + C_i + C_p}{W_p \cdot I_p}. \tag{3.37}$$

Dabei setzt sich die intrinsische Kapazität C_i aus den Gate- und Parasitärkapazitäten des schaltenden Gatters zusammen, während die extrinsischen Kapazitäten (Last-Kapazität C_l und parasitäre Leitungskapazitäten C_p) von der Ausgangslast und der Leitungslänge abhängen. Im kritischen Pfad einer typischen CMOS-Schaltung kann der Anteil von C_p an der Gesamtlast $C_{ges} = C_l + C_i + C_p$ selbst bei lokalen Verdrahtungen 50 % oder mehr betragen. Anstelle von $I_{d,sat}$ werden die effektiven Schaltströme der NFETs und PFETs I_n und I_p nach Gleichung 3.33 oder 3.35 eingesetzt.

Bei der Untersuchung unterschiedlicher Strategien zur Transistor-Dimensionierung wird stets eine Kette von Invertern angenommen, die jeweils gleich dimensioniert sind. Dies führt dazu, dass sich durch Variation einer Gateweite nicht nur der Strom des schaltenden Gatters, sondern auch die Kapazität des nachgeschalteten Logikgatters ändert. Wie im vorhergehenden Abschnitt beschrieben, gelten die Untersuchungen für unterschiedliche Fan-outs. Eine Erweiterung auf NAND- und NOR-Gatter erfolgt im Abschnitt 3.2.3.

Bei der klassischen CMOS-Skalierung werden die NMOS- und PMOS-Transistoren so dimensioniert, dass sich der statische Schaltpunkt der Logikgatter $V_{out} = V_{dd}/2$ bei $V_{in} = V_{dd}/2$

befindet. Diese Dimensionierung führt dazu, dass Anstiegs- und Abfallzeiten der Gatter etwa gleich lang sind: $t_p = t_n$ [69]. Diese so dimensionierten Gatter werden als symmetrische Gatter bezeichnet. Der kleinere effektive Schaltstrom des PMOS-Transistors I_p muss dafür durch eine größere PMOS-Weite W_p kompensiert werden. Daraus ergibt sich ein Weitenverhältnis

$$\beta_W^{sym} = \frac{W_p}{W_n} = \frac{I_n}{I_p} = \beta_I. \tag{3.38}$$

Dabei ist β_I das Verhältnis der effektiven Schaltströme nach Gleichung 3.33.

Diese symmetrische Dimensionierung ist jedoch nicht geschwindigkeitsoptimal. Das Weitenverhältnis, mit dem in einem Inverter die höchste mittlere Schaltgeschwindigkeit erzielt wird, kann anhand einer einfachen Rechnung bestimmt werden. Dazu wird die Verzögerungszeit zweier aufeinander folgender Schaltvorgänge t_{np} in einer Inverterkette minimiert (Abb. 3.25). Intrinsische und parasitäre Lasten werden zunächst nicht berücksichtigt:

$$t_{np} = t_n + t_p \quad \propto \quad \frac{W_p + W_n}{W_n \cdot I_n} + \frac{W_p + W_n}{W_p \cdot I_p} = \frac{1}{I_n}(\frac{W_p}{W_n} + 1) + \frac{1}{I_p}(\frac{W_n}{W_p} + 1) \tag{3.39}$$

$$t_{np}(\beta_W) \quad \propto \quad \frac{1}{I_n}(\beta_W + 1) + \frac{1}{I_p}(\frac{1}{\beta_W} + 1) \tag{3.40}$$

Diese Paar-Verzögerungszeit wird für ein optimales Weitenverhältnis β_W^{opt} minimal:

$$t'_{np}(\beta_W^{opt}) \quad = \quad \frac{1}{I_n} - \frac{1}{I_p(\beta_W^{opt})^2} = 0 \tag{3.41}$$

$$\Rightarrow \beta_W^{opt} \quad = \quad \frac{W_p}{W_n} = \sqrt{\frac{I_n}{I_p}} = \sqrt{\beta_I} \tag{3.42}$$

$$\Rightarrow t_{np}^{opt} \quad \propto \quad \left(\frac{1}{\sqrt{I_n}} + \frac{1}{\sqrt{I_p}}\right)^2 \tag{3.43}$$

Bei einem angenommenen Verhältnis der Ströme von $\beta_I = 2$ muss der PMOS-Transistor um einen Faktor $\beta_W^{opt} = \sqrt{2}$ größer dimensioniert werden. Dieses entspricht dem Ergebnis in [70]. Die hier vorgestellte Herleitung erfolgt jedoch aus dem einfachen Ansatz in den Gleichungen 3.36 und 3.37.

Im Vergleich zum symmetrisch schaltenden Inverter ist der PMOS-Transistor schwächer. Die Schaltzeit für einen 0–1-Übergang ist somit länger als für einen 1–0-Übergang. Für das Verhältnis der Schaltzeiten gilt:

$$\beta_t^{opt} = \frac{t_p}{t_n} = \frac{W_n \cdot I_n}{W_p \cdot I_p} = \frac{\beta_I}{\beta_W} = \sqrt{\beta_I} \tag{3.44}$$

Der Schaltvorgang über den NMOS-Transistor entspricht bei dieser Dimensionierung eher der Trajektorie (b), der des PMOS-Transistors der Trajektorie (d). Vorteil ist, dass der PMOS-Transistor, der einen höheren Anteil an der Schaltzeit hat, bei höheren Spannungen $V_{dd} - V_t$ betrieben wird und somit weniger anfällig gegen Schwankungen der Schwellenspannung ist.

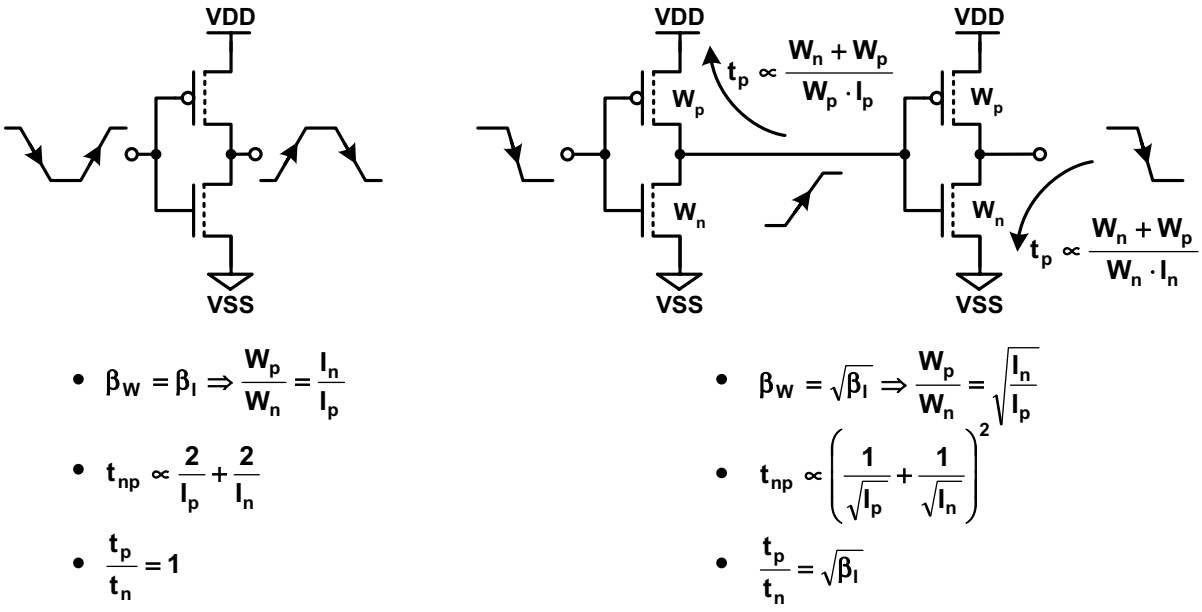


Abbildung 3.25: Dimensionierungen für symmetrisches und geschwindigkeitsoptimiertes Schalten.

Es ist zu beachten, dass der statische Schaltpunkt eines Gatters, bei dem $V_{in} = V_{out}$ ist, mit geschwindigkeitsoptimierter Dimensionierung nicht bei $V_{dd}/2$, sondern aufgrund des stärkeren NFETs bei kleineren Spannungen liegt. Die Schaltzeiten t_n und t_p müssen daher auch bei kleineren Spannungen gemessen werden (Abb. 3.26).

Bisher wurde in Gleichung 3.39 nur ein Fan-out von 1 ohne parasitäre und intrinsische Lasten berücksichtigt (Abb. 3.25). Bei höheren Ausgangsbelastungen vergrößern sich die Schaltzeiten. Da jedoch die Schaltzeiten für NMOS- und PMOS-Übergang jeweils proportional zum Fan-out ansteigen, gilt Gleichung 3.39 auch für größere Lasten. Auch die intrinsischen Lasten C_i werden durch die Gleichung berücksichtigt, da diese ebenfalls mit den NMOS- und PMOS-Weiten skalieren.

Bei Betrachtung einer großen Schaltung verhalten sich außerdem die parasitären Leitungskapazitäten C_p weitgehend proportional zu der Summe der Transistorweiten $W_n + W_p$. Größere Transistoren bedeuten eine größere Gesamtfläche und damit längere Leitungen und höhere Leitungskapazitäten. Die Gleichungen 3.39–3.44 sind somit sehr allgemein gültig.

Ein Nachteil bei der geschwindigkeitsoptimierten Dimensionierung ist, dass sich bei unterschiedlichen Belastungen der Gatter die Zeitabstände zwischen steigender und fallender Flanke verändern. Ist der PMOS-Schaltvorgang z.B. doppelt so lang wie der NMOS-Schaltvorgang, so wird die steigende Flanke bei einer hohen Belastung stärker verzögert als die fallende Flanke. Ist die folgende Stufe einer Inverterkette schwächer belastet, so wird nur ein Teil dieses Zeitunterschieds kompensiert. Insbesondere muss der Taktbaum mit symmetrischen Gattern ausgelegt werden, da sich sonst das Tastverhältnis des Taktsignals (*Duty Cycle*) verändert. In

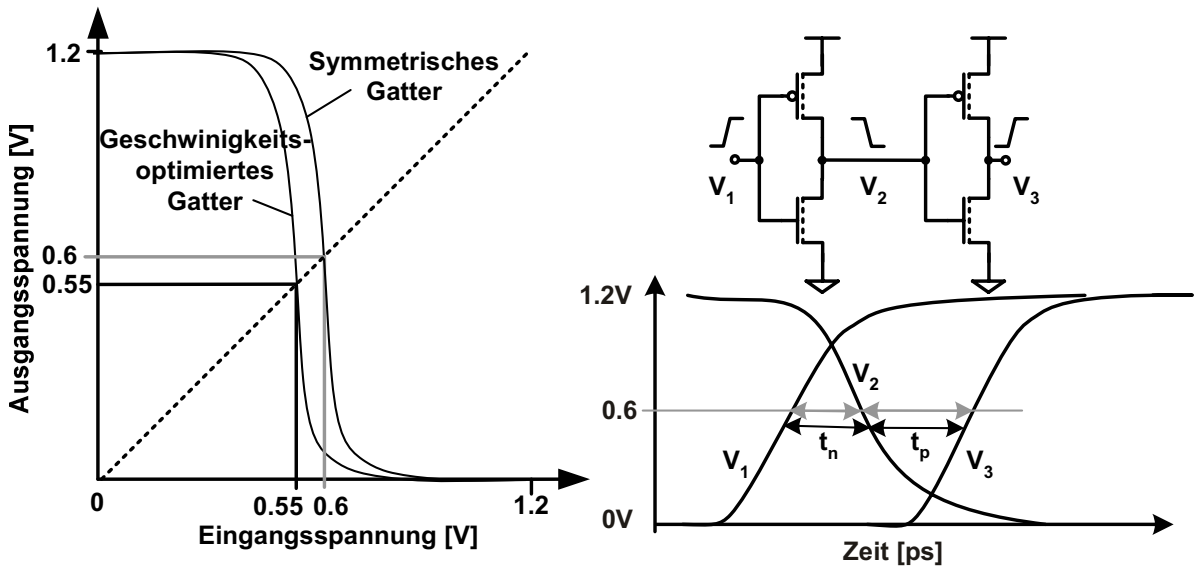


Abbildung 3.26: Statische Schaltpunkte von symmetrischen und geschwindigkeitsoptimierten Logikgattern. Im Gegensatz zum symmetrischen Gatter liegt der Schaltpunkt bei dem geschwindigkeitsoptimierten Gatter nicht bei $V_{dd}/2$.

dieser Arbeit wird meist die symmetrische Dimensionierung bei der Untersuchung verschiedener Effekte diskutiert, sodass nicht zwischen NMOS- und PMOS-Schaltvorgang differenziert werden muss. Die im Kapitel 6 beschriebenen Demonstratorschaltungen verwenden dagegen die geschwindigkeitsoptimierte Dimensionierung.

3.2.3 Stack-Effekte in NAND- und NOR-Gattern

Im Unterschied zu einem einfachen Inverter erfolgen Schaltvorgänge in digitalen Schaltungen auch über Transistor-Serienschaltungen, so genannte *Stacks*. Insbesondere beinhaltet der kritische Pfad einer Logikschaltung häufig überproportional viele Stacks, da diese einen logischen Pfad langsamer werden lassen und dieser Pfad dadurch zum kritischen Pfad wird.

Die Schaltreihenfolge der Transistoren in einem Stack hat wesentliche Auswirkungen auf die Schaltgeschwindigkeit. Insbesondere bei einer kleinen Ausgangslast ist die intrinsische Verzögerungszeit deutlich größer, wenn der langsamere Eingang (z.B. der untere Transistor in einem NMOS-Stack) schaltet, da hier zusätzlich die Kapazität des oberen Transistors entladen werden muss. Für große Lasten gleichen sich die Verzögerungszeiten an, da hier der extrinsische Anteil überwiegt.

Zwei Effekte vermindern die Schaltgeschwindigkeit eines Transistorstacks gegenüber einem einfachen Inverter, welche beide aus der Anhebung des Potentials des internen Knotens des Stacks resultieren. Zum einen wird die Gate-Source-Spannung des oberen Transistors reduziert. Zum anderen liegt am oberen Transistor effektiv eine negative Bulk-Source-Spannung an (*Auto-Reverse-Biasing*).

Zur Untersuchung der Strom- und Spannungsverhältnisse in gestackten Transistoren während des Schaltvorgangs wird zunächst der Fall betrachtet, bei dem eine große Kapazität mit einer schnellen Eingangsflanke geschaltet wird (Abb. 3.27). Insbesondere das Schalten des unteren Transistors N2 in einer Serienschaltung ist geschwindigkeitskritisch und stellt damit den in erster Linie zu untersuchenden ungünstigsten Fall dar. Der Transistor N1 ist bereits geöffnet. Um die Betriebspunkte der Transistoren zu erhalten, wird das dynamische Verhalten der Serienschaltung mit Hilfe von Schaltungssimulationen untersucht. Für die verschiedenen Transistortypen sowie über einen großen V_{dd}-Bereich erhält man dabei für das Potential des inneren Knotens

$$V_x \approx \frac{V_{dd} - V_t}{\beta}. \tag{3.45}$$

Diese empirische Formel mit $\beta \approx 4.2$ gilt für das innere Knotenpotential in einem NMOS-Stack ebenso wie für einen PMOS-Stack, z.B. in einem NOR2-Gatter. Aus dem erhöhten Potential am Knoten V_x folgt u.a. eine reduzierte Gate-Source-Spannung des oberen Transistors N1 ($V_{gs,1} < V_{dd}$, Abb. 3.27b). Die Gate-Source-Spannung des oberen Transistors $V_{gs,1}$ erreicht damit zu keinem Zeitpunkt des Schaltvorgangs die volle Betriebsspannung, während V_{gs} in einem Inverter nur während der ersten Hälfte des Schaltvorgangs kleiner als V_{dd} ist (Abb. 3.22).

Im nächsten Schritt wird nun das dynamische Schaltverhalten bei gleich schnellen Ein- und Ausgangsflanken untersucht. Beim Einschalten des Transistors N2 ist dessen Gate-Source-Spannung zunächst kleiner als die des Transistors N1, da das Eingangssignal noch nicht V_{dd} erreicht hat: $V_{gs,2} < V_{gs,1}$. N2 hat daher eine geringere Leitfähigkeit als N1 und bestimmt zunächst den Stromfluss durch die gesamte Anordnung (Abb. 3.28).

Ist das Eingangssignal B so weit angestiegen, dass nun $V_{gs,2} > V_{gs,1}$ ist, dann geht das kritische Schaltverhalten auf den Transistor N1 über, der als Stromquelle den Schaltvorgang bestimmt. Die Kombination der beiden kritischen Trajektorien-Abschnitte von N1 und N2 in Abbildung 3.28 liefert eine Stack-Trajektorie, die in ihrer Form ähnlich der Inverter-Trajektorie ist, jedoch insgesamt bei kleineren Spannungen verläuft.

In Abbildung 3.29 werden die Schalttrajektorien von einem ungestackten Transistor sowie Zweier- und Dreier-Stacks, entsprechend einem Inverter, einem NAND2- und einem NAND3-Gatter, verglichen. Der mittlere Schaltstrom I_M nach Gleichung 3.35 ist danach für den ungestackten Fall bei 80 %, für den Zweier-Stack bei 65 % und für den Dreier-Stack bei 55 % V_{dd} zu betrachten. Um die Leistungsfähigkeit eines Transistors in einer realen digitalen Schaltung richtig einzuschätzen, sollte der Drain-Strom daher nicht nur bei $V_{gs} = V_{ds} = V_{dd}$, sondern besonders bei kleineren Spannungen mit $V_{gs} = V_{ds}$ gemessen werden. Beispielsweise besitzen Transistoren in Sub-100 nm-Technologien ein relativ hohes *Drain-Induced Barrier Lowering* (DIBL-Effekt). Dieser Effekt erhöht zwar den Strom I_{on} bei 100 % V_{dd} deutlich, hat aber bei 55 % V_{dd} einen erheblich reduzierten Einfluss (effektive Spannung eines NAND3-Gatters, Abb. 3.29).

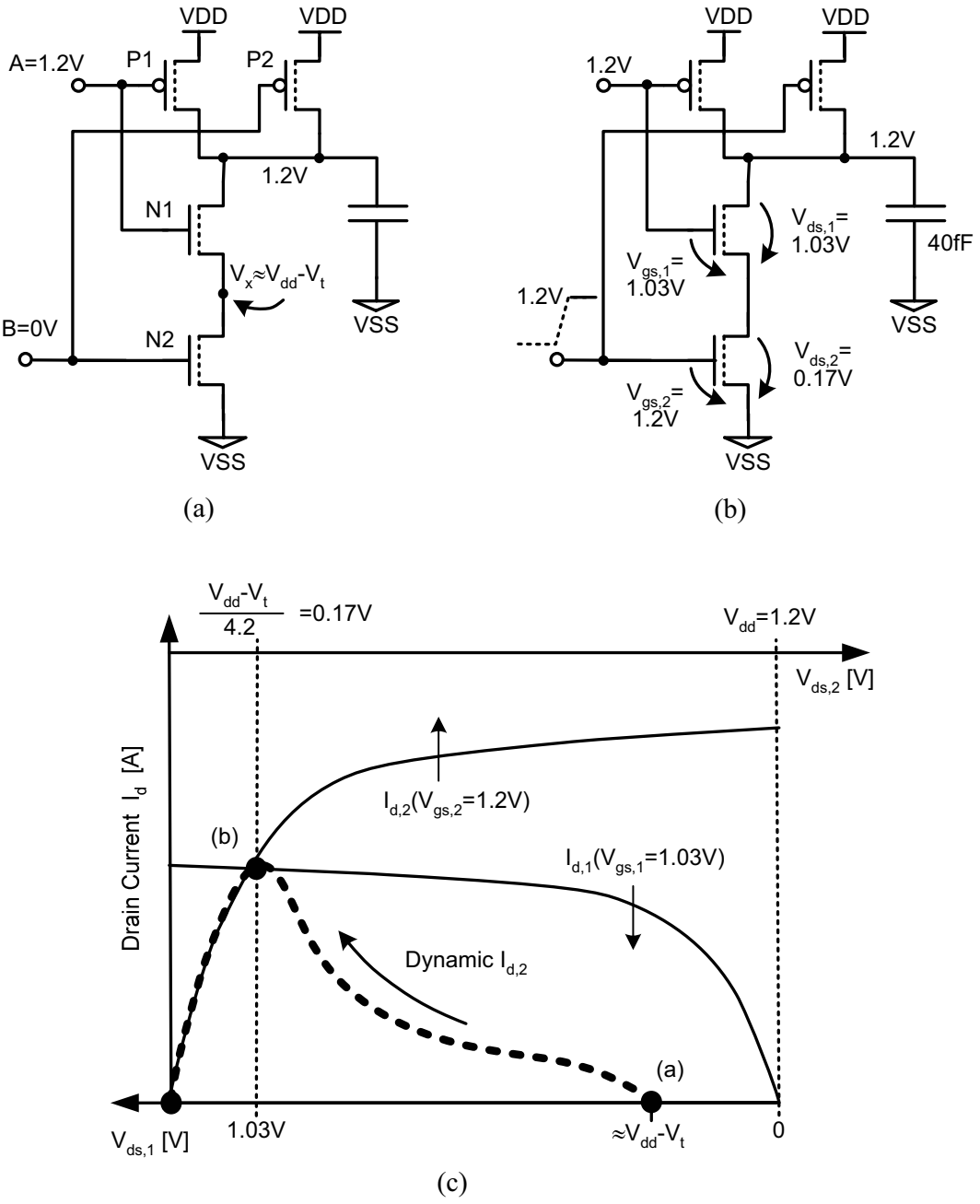


Abbildung 3.27: Schematische Darstellung des Entladevorgangs einer großen Kapazität über eine NMOS-Serienschaltung. Bevor der untere Eingang B von 0 V auf V_{dd} schaltet, liegt am internen Knoten das Potential $V_x \approx V_{dd} - V_t$ an (a). Kurz nach dem Öffnen des unteren Transistors ist $V_x \approx (V_{dd} - V_t)/4.2$ (b). Mit der Entladung der Ausgangskapazität verschiebt sich die Kennlinie $I_{d,1}$ in Abbildung (c) nach links. Da sich N1 in Sättigung befindet, hängt $I_{d,1}$ nur schwach von der Drain-Spannung $V_{ds,1}$ ab. Der Strom bleibt damit zunächst weitgehend konstant und V_x sinkt nur sehr langsam ab. Erst wenn sowohl N1 als auch N2 im linearen Bereich sind, reduziert sich V_x signifikant.

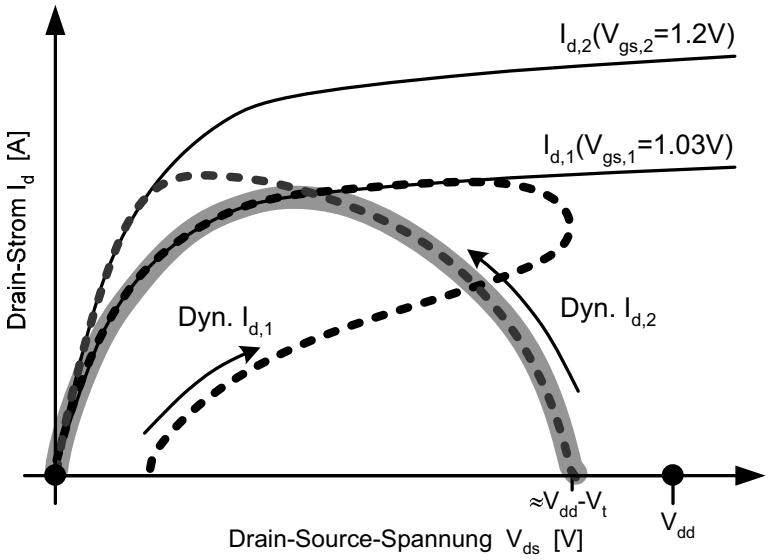


Abbildung 3.28: Dynamische Schalttrajektorien zweier gestackter NFETs. Die grau hinterlegten Abschnitte der Trajektorien kennzeichnen den kritischen Verlauf des gesamten Schaltvorgangs.

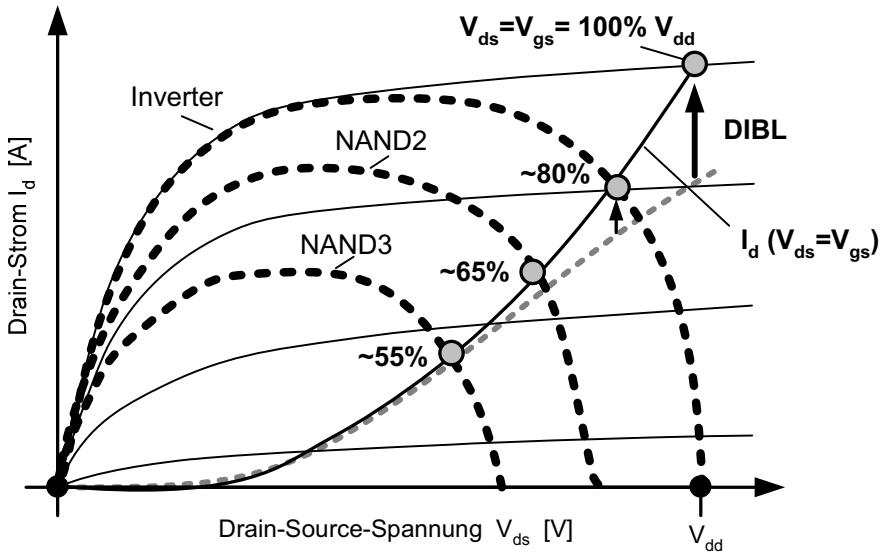


Abbildung 3.29: Schalttrajektorien von Inverter, NAND2- und NAND3-Gattern.

3.2.4 Einfluss der Temperatur auf die Schaltgeschwindigkeit

Während des aktiven Betriebs einer Schaltung steigt die Temperatur im Kanalbereich der Transistoren abhängig von der umgesetzten Leistung sowie der Wärmeableitung an. Eine typische Junction-Temperatur für Logikschaltungen ist 85 °C, aber auch bei 125 °C soll die Schaltung in der Regel noch funktionsfähig sein. Bei Erhöhung der Temperaturen beeinflussen zwei Effekte den Drainstrom der Transistoren, die in entgegengesetzte Richtungen wirken. Zum einen reduziert sich die Beweglichkeit der Ladungsträger im Kanalbereich durch verstärkte Phononen- und Oberflächenstreuung. Zum anderen reduziert sich die Schwellenspannung wie im Abschnitt 3.1.1 beschrieben wegen des geringeren Fermi-Potentials. Für kleine Gate-Source-Spannungen V_{gs} ist die Reduzierung der Schwellenspannung der bestimmende Faktor und der Drainstrom I_d erhöht sich in diesem Bereich. Bei höheren Spannungen wirkt sich hingegen die geringere Beweglichkeit stärker aus. Deshalb reduziert sich der Drainstrom hier (Abb. 3.30).

Der statische *Zero Temperature Coefficient Point* (ZTC-Punkt) ist definiert als der Wert der Gate-Source-Spannung, bei dem der Drainstrom unabhängig von der Temperatur ist. In älteren Technologiegenerationen liegt der statische ZTC-Punkt im Verhältnis zur Versorgungsspannung bei deutlich kleineren Gate-Source-Spannungen. In der 0.5 μm-Technologie überkreuzen sich die Kennlinien bei $V_{gs} = 1.13\,\mathrm{V}$ (NMOS) bzw. 1.66 V (PMOS), während die Versorgungsspannung 5 V beträgt. In [71] wird daraus geschlossen, dass in Technologiegenerationen, deren Versorgungsspannung unterhalb von 1.66 V bzw. 1.13 V liegt, der Strom für höhere Temperaturen zunimmt. Die dabei getroffene Annahme, dass der ZTC-Punkt bei der Skalierung der Transistoren konstant bleibt, ist jedoch nicht richtig. Vielmehr sinkt die Temperaturabhängig-

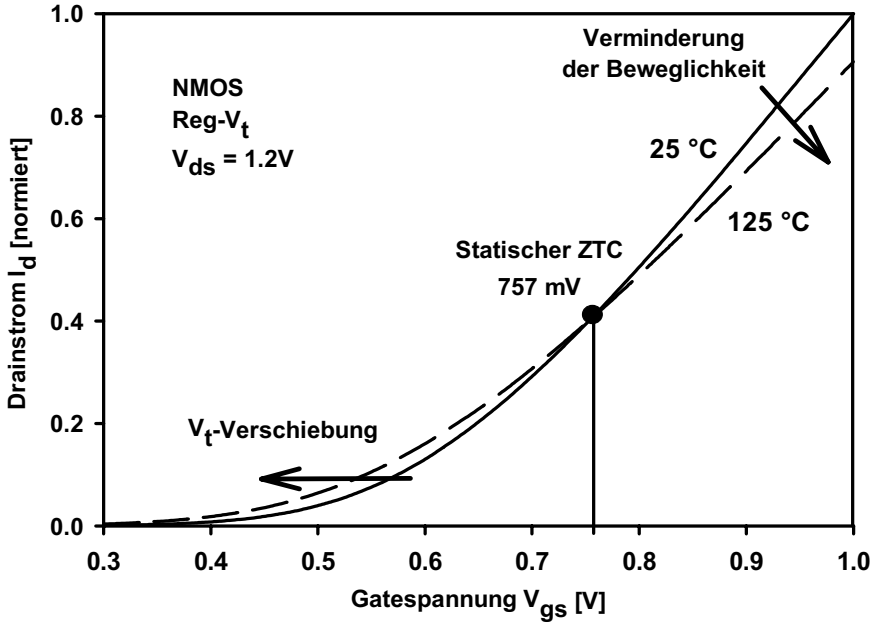


Abbildung 3.30: Statischer Zero-Temperature-Coefficient-Punkt eines REG-Transistors. Dargestellt ist die Überkreuzung der Kennlinien bei 25 °C und 125 °C, jedoch auch die Kennlinien anderer Temperaturen kreuzen sich in dem selben Punkt.

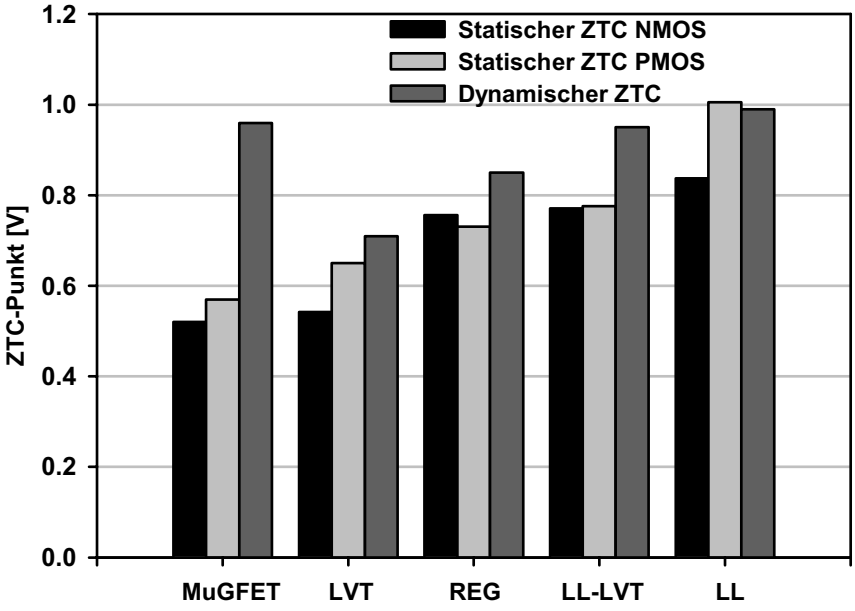


Abbildung 3.31: Vergleich statischer und dynamischer ZTC-Punkte in einem Multi-Gate-Transistor sowie verschiedenen Transistoren der 90 nm-Technologie.

keit der Schwellenspannung, wenn das Kanalpotential stärker an das Gate-Potential gekoppelt ist, z.B. beim Einsatz dünnerer Gateoxide.

Die Temperaturabhängigkeit der Beweglichkeit wird semi-empirisch durch

$$\mu(T) = \mu(T_0) \cdot \left(\frac{T}{T_0}\right)^{k_2} \tag{3.46}$$

beschrieben [39]. Die genaue Messung der Ladungsträger-Beweglichkeit erfordert dynamische CV-Messungen, die nicht zur Verfügung standen. Jedoch bereits die Analyse der Transistor-Charakteristiken zeigt kein eindeutiges Bild der Veränderung der Eingangskennlinien-Steigungen bei Erhöhung der Temperatur.

In Abbildung 3.31 sind die ZTC-Punkte von Transistoren aus der 90 nm-Technologie und für MuGFETs dargestellt. In den meisten Fällen ist der PMOS-ZTC-Punkt wegen der größeren Temperaturabhängigkeit der Schwellenspannung etwa 100 mV größer als der NMOS-ZTC-Punkt.

Teilweise sind die statischen ZTC-Punkte in Abbildung 3.31 größer als 1 V. Das bedeutet, dass die Versorgungsspannung insbesondere bei Low-Power-Anwendungen unterhalb des statischen ZTC-Punktes liegen kann. Daher ist es wichtig, das dynamische Schaltverhalten digitaler CMOS-Schaltungen in Abhängigkeit von der Temperatur zu untersuchen [72]. Es werden Ringoszillatoren in der 90 nm-Technologie untersucht. Die Ringoszillatoren bestehen jeweils aus 17 Invertern mit einem Fan-out von 1. Die Frequenz wird über einen zehnstufigen Frequenzteiler gemessen. Die Messung der Frequenz bei unterschiedlichen Versorgungsspannungen und Tem-

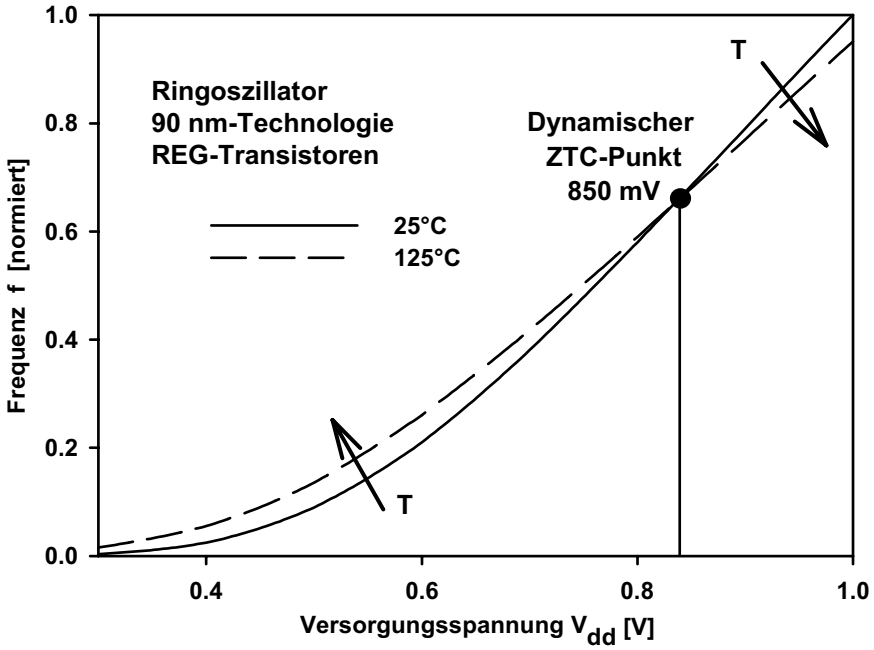


Abbildung 3.32: Dynamischer Zero-Temperature-Coefficient-Punkt eines Ringoszillators. Dargestellt ist die Überkreuzung der Frequenz-Kennlinien in Abhängigkeit von V_{dd} bei 25 °C und 125 °C. Diese Darstellung besitzt Ähnlichkeit mit der Abbildung 3.30, in der der Strom eines einzelnen Transistors in Abhängigkeit von der Gate-Source-Spannung V_{gs} dargestellt ist.

peraturen liefert wie die Messung der Transistorströme charakteristische Kennlinien, die sich in einem Punkt des Frequenz/V_{dd}-Diagramms schneiden (Abb. 3.32).

Der *dynamische* ZTC-Punkt ist deshalb als der Wert der Versorgungsspannung definiert, bei dem sich die Frequenz-Kennlinien schneiden. In diesem Punkt ist die Frequenz des Ringoszillators von der Temperatur unabhängig. Beim Vergleich der dynamischen mit den statischen ZTC-Punkten der 90 nm-Technologie fällt auf, dass die dynamischen Punkte etwa 100 mV über dem Durchschnitt der beiden jeweiligen statischen Punkte liegen. Diese Abhängigkeit lässt sich anhand der Abbildung 3.22 erklären. Schon bevor der Eingang eines schaltenden Transistors auf das volle V_{dd} aufgeladen ist, fließt ein signifikanter Strom. In Abbildung 3.30 bedeutet das, dass die Kennlinie von links nach rechts durchlaufen wird. Solange noch $V_{gs} < ZTC_{stat}$ ist, fließt bei 125 °C mehr Strom als bei 25 °C. Erst oberhalb des statischen ZTC-Punktes führt die reduzierte Beweglichkeit zu einem kleineren Drainstrom. Folglich ist der dynamische ZTC-Punkt stets höher als die statischen ZTC-Punkte.

In Schaltungen, die auch bei niedrigen Spannungen noch korrekt arbeiten sollen, muss der dynamische ZTC-Punkt bei der Geschwindigkeits-Charakterisierung berücksichtigt werden. So werden Schaltungen in der Simulation oder messtechnisch unter den ungünstigsten Bedingungen getestet, üblicherweise bei einer um 10 % verminderten Versorgungsspannung und bei der höchsten spezifizierten Temperatur. Ist $V_{dd} - 10\,\%$ kleiner als der dynamische ZTC-Punkt, so wird die Schaltung für hohe T schneller und es muss daher bei der kleinsten Temperatur getestet werden. Die geringere Temperaturabhängigkeit in der Nähe des dynamischen ZTC-Punkts

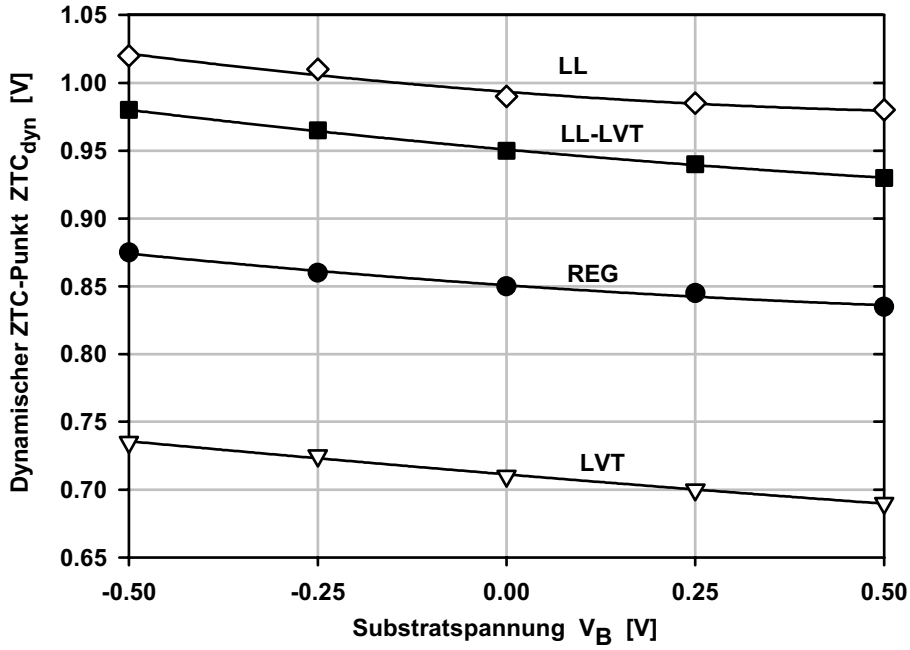


Abbildung 3.33: Dynamischer ZTC-Punkt in der 90 nm-Technologie in Abhängigkeit von der Substratspannung.

reduziert insgesamt die Schwankungsbreite der Schaltgeschwindigkeit. Leckströme müssen jedoch in jedem Fall bei der höchsten Temperatur abgeschätzt werden.

Der ZTC-Punkt kann dazu genutzt werden, um Anwendungen mit einer temperaturunabhängigen Schaltgeschwindigkeit zu betreiben ($V_{dd} = ZTC_{dyn}$). Allerdings liegen die dynamischen ZTC-Punkte in der 90 nm-Technologie (0.72 bis 0.99 V) noch im unteren Bereich der V_{dd}-Spezifikation (0.85 bis 1.45 V), sodass möglicherweise die Schaltgeschwindigkeit nicht ausreicht oder Parameterschwankungen einen zu großen Einfluss haben. In diesem Fall kann die Schaltgeschwindigkeit durch *Forward-Biasing* erhöht werden (vgl. Kapitel 4). Der dynamische ZTC-Punkt wird durch das Anlegen einer Substratspannung nur leicht verschoben (Abb. 3.33).

Kapitel 4

Body Biasing in Sub-100 nm-Technologien

Die Abhängigkeit der Schwellenspannung von der Bulk-Source-Spannung V_{bs} wird als Substrateffekt bezeichnet. Mit Hilfe dieses Effekts ist es möglich, die Schwellenspannungen der Transistoren an die jeweiligen Anforderungen anzupassen [73, 74]. So lässt sich V_t im aktiven Betrieb senken und damit der On-Strom und die Schaltgeschwindigkeit erhöhen. Alternativ kann im Standby-Modus der Leckstrom reduziert werden, indem V_t erhöht wird. Bei NMOS-Transistoren vermindert sich die Schwellenspannung durch Anlegen einer positiven Bulk-Source-Spannung $V_{bs} > 0$V (Forward-Biasing, FB), während sie für $V_{bs} < 0$V angehoben wird (Reverse-Biasing, RB). Für PMOS-Transistoren gilt Entsprechendes mit umgekehrten Vorzeichen.

Der Substrateffekt, d.h. die Verschiebung der Schwellenspannung, lässt sich beschreiben durch [29]

$$\Delta V_t = \gamma \cdot \left(\sqrt{2\phi_F - V_{bs}} - \sqrt{2\phi_F} \right) \tag{4.1}$$

$$\gamma = \frac{\sqrt{2\varepsilon_{Si} q N_A}}{C'_{ox}} = \sqrt{2\varepsilon_{Si} q N_A} \frac{t_{ox}}{\varepsilon_{ox}}, \tag{4.2}$$

γ ist die Substratkonstante, ϕ_F das Fermi-Potential, N_A die Kanaldotierung, t_{ox} die Oxiddicke und ε_{Si} und ε_{ox} die Dielektrizitätskonstanten von Silizium und des Gate-Dielektrikums. Die Verschiebung der Schwellenspannung ist in Abbildung 4.1 für einen NMOS-LL-Transistor dargestellt.

Um in einer Schaltung Forward- oder Reverse-Biasing sowohl auf NMOS- als auch auf PMOS-Transistoren gleichzeitig anzuwenden zu können, müssen an die beiden Transistortypen entgegengesetzte Spannungen angelegt werden. Im Folgenden bezeichnet eine positive Substratspannung V_B ein Forward-Biasing von NFET ($V_{bs,n} > 0$) und PFET ($V_{bs,p} < 0$), eine negative Substratspannung V_B ein Reverse-Biasing. Es wird dabei angenommen, dass ein Drei-Wannen-Prozess zur Verfügung steht, in dem die P-Wannen-Spannung der NMOS-Transistoren unabhängig von der Substratspannung des Chips eingestellt werden kann. Body Biasing ist auch in einem einfachen Zwei-Wannen-Prozess realisierbar. Allerdings können dann nur die PMOS-Transistoren vorgespannt werden [75]. Die NMOS-Transistoren und damit das gesamte Substrat eines Chips vorzuspannen, würde andere Schaltungsteile in einer System-on-Chip-Umgebung

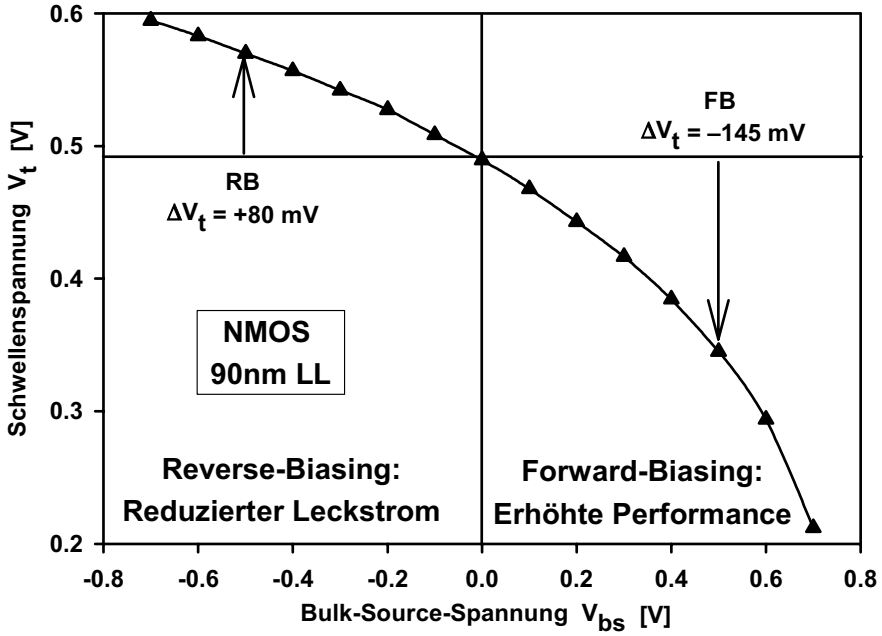


Abbildung 4.1: Messung der Schwellenspannung in Abhängigkeit von der Bulk-Source-Spannung.

stören. Da die PMOS-Transistoren in stärkerem Maße als die NMOS-Transistoren zur Gesamtschaltzeit beitragen (Abschnitt 3.2), ist diese Technik sehr effizient. In diesem Kapitel werden jedoch NFETs und PFETs immer symmetrisch vorgespannt, da in der 90 nm-Technologie eine Drei-Wannen-Option zur Verfügung steht [29].

Durch die Skalierung der Transistoren hin zu dünneren Oxiden und kleineren Kanaldotierungen (respektive kleineren Schwellenspannungen) wird die Substratkonstante und damit der Substrateffekt kleiner. Dennoch ist der Effekt in der 90 nm-Technologie noch so groß, dass die Schwellenspannung eines NMOS-LL-Transistors bei Anlegen von $V_{bs} = 0.5$V um 145 mV sinkt und sich der On-Strom um 30 % erhöht. Wegen der nichtlinearen Abhängigkeit der Schwellenspannung von V_{bs} erhöht sich V_t beim Reverse-Biasing ($V_{bs} = -0.5V$) nur um 80 mV und der On-Strom vermindert sich nur um 20 % (Abb. 4.1).

4.1 Leckströme und Reverse-Biasing

Für eine umfassende Bewertung der Effizienz von Body Biasing in Sub-100 nm-Technologien sind neben den Einflüssen auf das einzelne Device auch Schaltungsaspekte zu berücksichtigen. So können zusätzliche Leckströme auftreten oder Kapazitäten ihren Wert ändern.

Wie im Abschnitt 3.1 beschrieben, setzt sich der Off-Strom eines Transistors aus dem Unterschwellenstrom, dem GIDL-Strom und dem Gateleckstrom des ausgeschalteten Transistors $I_{g,off}$ zusammen. Zusätzlich trägt in CMOS-Schaltungen noch der Gateleckstrom der angeschalteten Transistoren $I_{g,on}$ zum gesamten Leckstrom bei (Abb. 4.2). Abbildung 4.3 zeigt die Abhängigkeit der NMOS-Leckströme von V_{bs} bei unterschiedlichen Temperaturen.

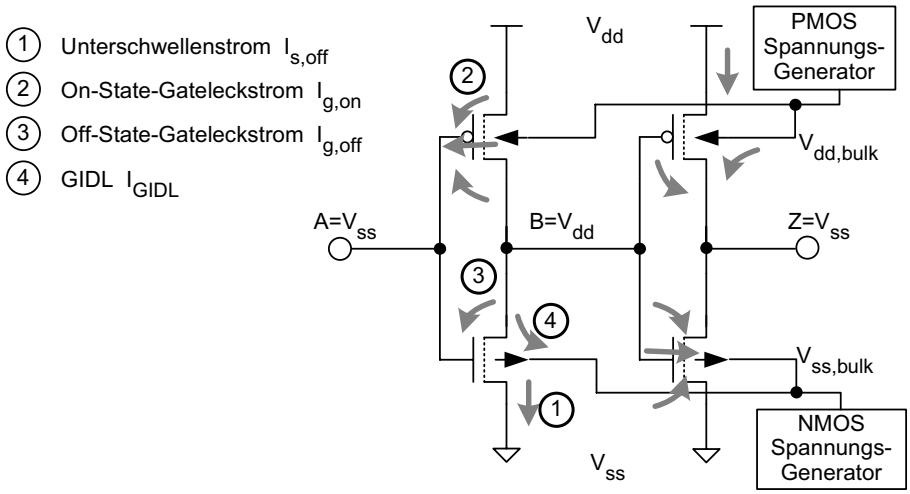


Abbildung 4.2: Leckströme und Body Biasing in einem CMOS-Inverter-Paar.

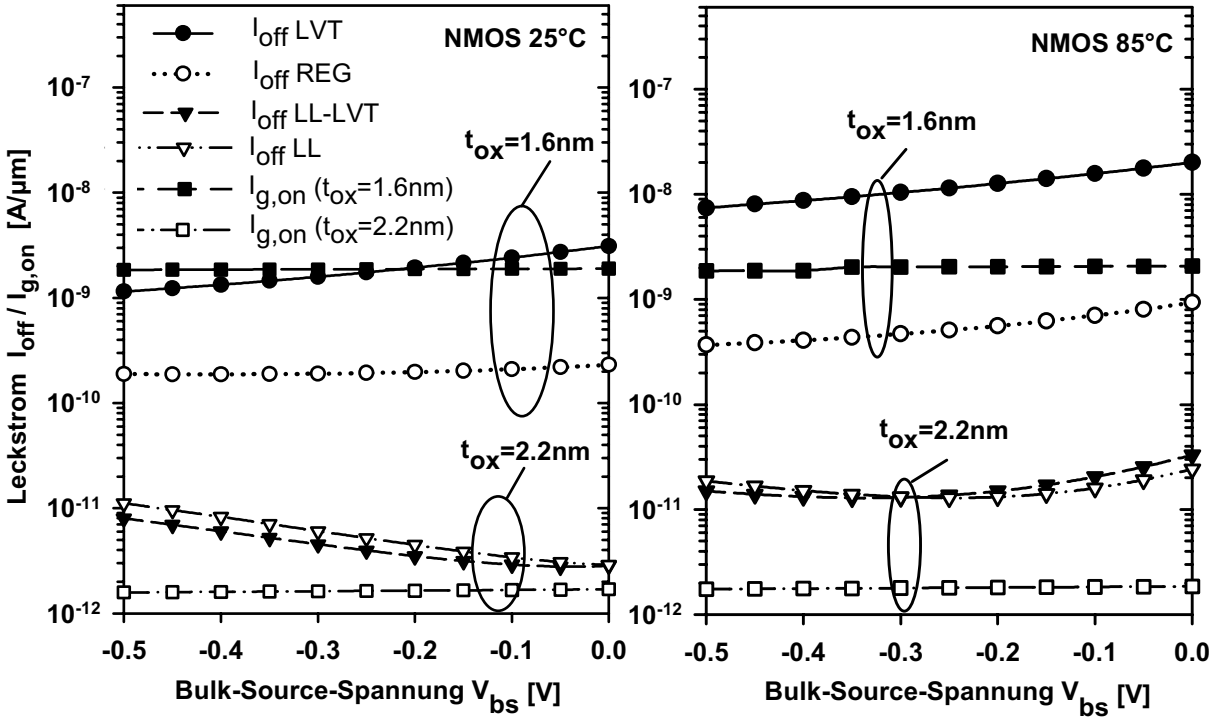


Abbildung 4.3: NMOS-Leckströme bei 25 °C und 85 °C und variabler negativer Bulk-Source-Spannung.

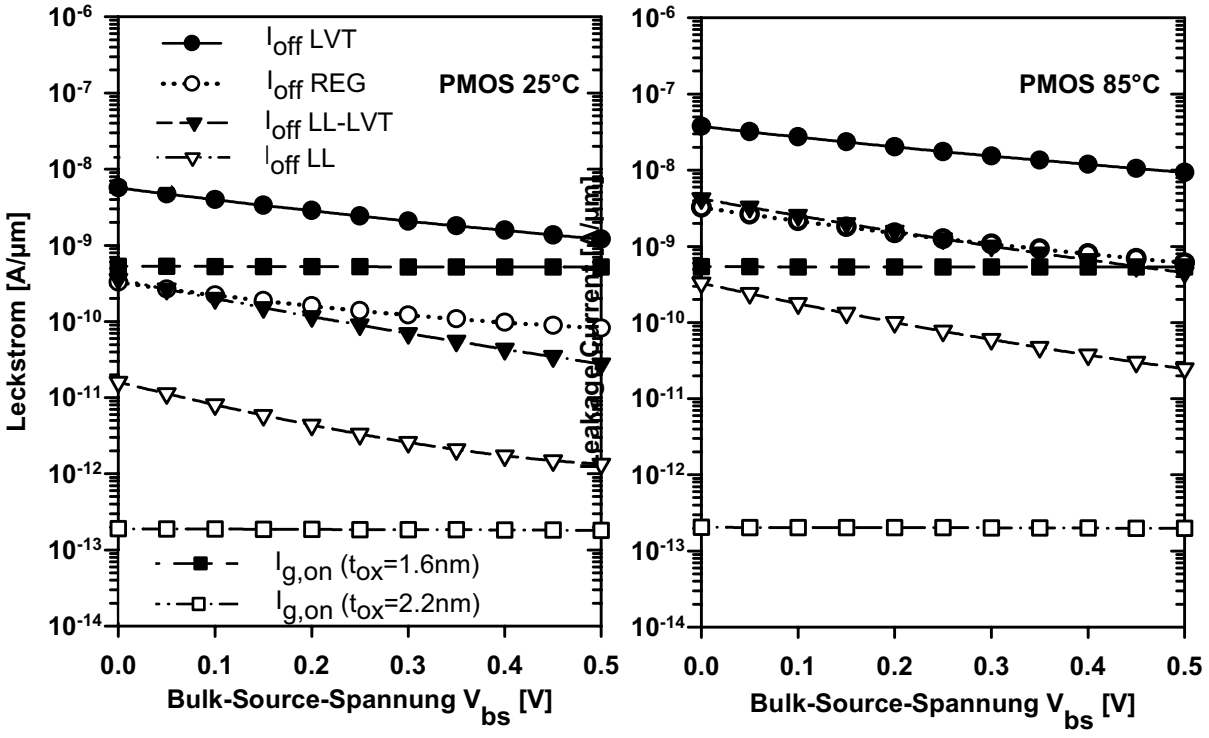


Abbildung 4.4: PMOS-Leckströme bei 25 °C und 85 °C und variabler negativer Bulk-Source-Spannung.

Bei Raumtemperatur und $V_{bs} = 0\text{V}$ liegt $I_{g,on}$ für alle Transistortypen mindestens in der gleichen Größenordnung wie der Off-Strom des jeweiligen Transistors, sogar für die Dickoxid-Transistoren. Hier ist Reverse-Biasing nicht dazu geeignet, den Leckstrom einer Schaltung signifikant zu reduzieren. Für den REG-Transistor ist $I_{g,on}$ auch noch bei 85 °C die bestimmende Leckstromkomponente.

Bei NMOS-LL-Transistoren führt Reverse-Biasing von mehr als $-0.25\,\text{mV}$ darüber hinaus sogar zu einer Erhöhung des Leckstroms, da der GIDL-Strom aufgrund der erhöhten Drain-Bulk-Spannung zunimmt (Gleichung 3.12). Da jedoch bei PMOS-Transistoren (Abb. 4.4) der GIDL-Anteil kleiner ist, führt Reverse-Biasing dennoch zu einer Reduzierung des Leckstroms, wenn zusätzlich die PMOS-Transistoren einbezogen werden. Dies gilt insbesondere für LL-LVT-Transistoren. Der gesamte Leckstrom des Inverterpaars in Abb. 4.2 lässt sich als Summe der gemessenen Leckstromkomponenten bestimmen (Abb. 4.5).

In Tabelle 4.1 werden die Veränderungen der Leckströme in Einzeltransistoren, Inverterpaaren und komplexen Schaltungen (32-bit-Addierer) verglichen. Bei den Einzeltransistoren sind nur die Leckströme der ausgeschalteten Transistoren angegeben, während bei den Inverterpaaren und den Addierern zusätzlich $I_{g,on}$ zum Leckstrom beiträgt. Dieser Anteil, der nicht durch Reverse-Biasing vermindert wird, erklärt die geringere Abnahme bei den Inverterpaaren. Der 32-bit-Addierer wird im Kapitel 6 beschrieben. Auch wenn in dem Addierer teilweise neue Schaltungstechniken verwendet werden, kann dieser trotzdem als repräsentatives Beispiel für eine komplexe CMOS-Schaltung dienen. Anders als bei einem Inverterpaar treten hier auch

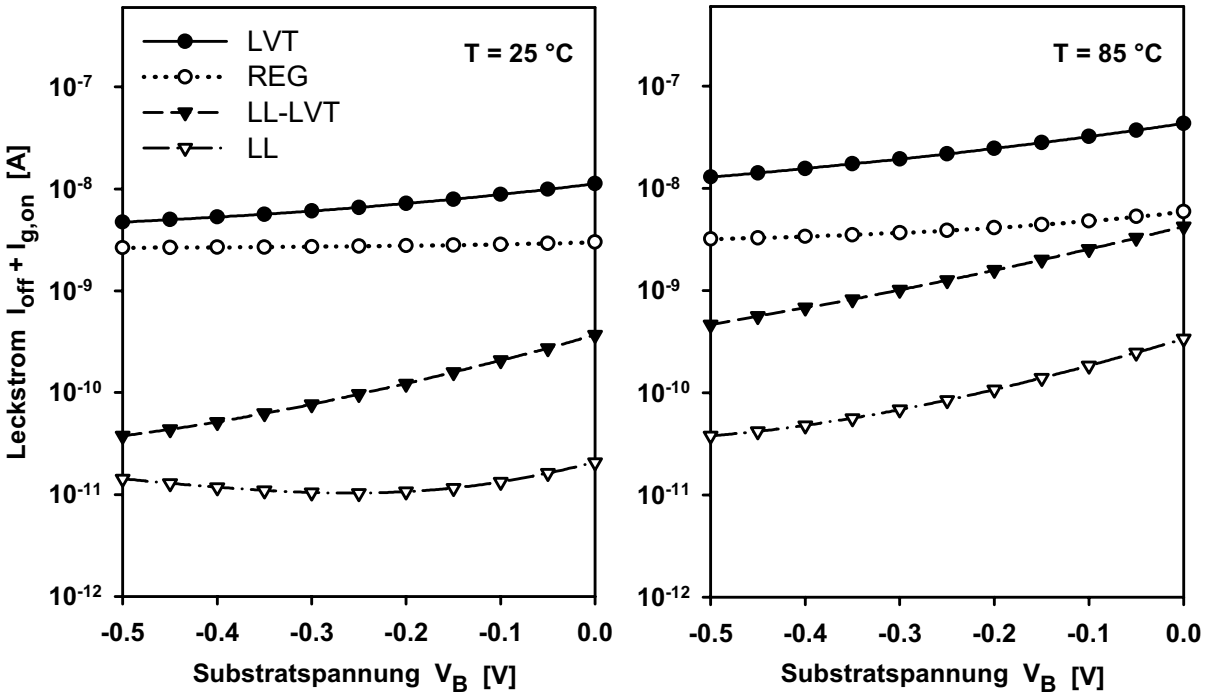


Abbildung 4.5: Summe der Leckströme eines Inverterpaars bei 25 °C und 85 °C und Reverse-Biasing. Die Transistorweiten von PFETs und NFETs betragen 2 und 1 μm.

	LVT	REG	LL-LVT	LL
NFET/PFET	0.37 / 0.21	0.81 / 0.25	**2.81** / 0.08	**3.85** / 0.08
Inverterpaar	0.42	0.88	0.10	0.69
32-bit-Addierer	0.47	0.90	—	0.48

Tabelle 4.1: Veränderung des Leckstroms durch Reverse-Biasing ($V_B = -0.5$ V) im Vergleich zu $V_B = 0$ V ($= 1$). Im Gegensatz zu den PFETs erhöht sich beim Reverse-Biasing der NFETs der Leckstrom wegen des GIDL-Effekts um einen Faktor 3 bis 4 (T=25 °C). Da der Anteil des PFETs am Leckstrom des Inverters jedoch höher ist als der Anteil der NFETs, sinkt der Inverter-Leckstrom insgesamt dennoch.

Transistor-Stacks auf. Der Vergleich der Leckstromveränderungen zwischen den Inverterpaaren und den Addierern ergibt nur kleine Abweichungen, die sich zum einen daraus ergeben, dass die beiden Testschaltungen auf unterschiedlichen Testchips realisiert wurden. Zum anderen führen Stack-Effekte zu kleinen Abweichungen. Allgemein lässt sich jedoch feststellen, dass die Betrachtung eines Inverterpaars für Leckstrom-Abschätzungen ausreichend ist.

Nur wenn der Unterschwellenstrom größer ist als andere Komponenten, ist es möglich, den Leckstrom signifikant durch Reverse-Biasing zu reduzieren. Bei den hier untersuchten Transistortypen erfordert der Einsatz von Reverse-Biasing eine sehr genaue Analyse des Einsatzbereiches der Schaltung (Temperaturen und Timing-Anforderungen). Der Einsatz von Reverse-Biasing kann effizient werden, wenn Transistoren mit kleineren Schwellenspannungen verwendet werden, sodass der Leckstrom einer Schaltung bei Raumtemperatur durch die Unterschwellenströme bestimmt wird (LL-LVT-Transistor, Tabelle 4.1). Allerdings nimmt der Substrateffekt für kleinere Schwellenspannungen ab. Eine signifikante V_t-Verschiebung ist dann nur noch durch stärkeres Reverse-Biasing zu erreichen. Hier wiederum sinkt die Effizienz noch weiter, da die Schwellenspannung hier schwächer von V_{bs} abhängt. (Abb. 4.1).

Daneben erfordert die Spannungsgenerierung und -verteilung von $V_{ss,bulk}$ und $V_{dd,bulk}$ zusätzliche Fläche und Energie. Für die Spannungsgenerierung auf dem Chip müssen Ladungspumpen implementiert werden, gerade im Standby-Modus wirkt sich die hierzu benötigte Energie negativ aus. Bei der Spannungsverteilung ergibt sich das Problem, dass Spannungen geschaltet werden müssen, die um V_B größer als V_{dd} sind. Falls V_{dd} ohnehin schon am Rande der Spezifikation liegt, müssen hier andere Transistoren, z.B. mit dickeren Oxiden verwendet werden. Dies erhöht den Design-, Prozess- und Flächenaufwand noch weiter.

4.2 Schaltgeschwindigkeit und Forward-Biasing

Wie im Abschnitt 3.2.1 beschrieben, hängt die Schaltgeschwindigkeit eines Logikgatters nicht nur von der Schwellenspannung und dem On-Strom, sondern auch von den Kapazitäten in den Transistoren ab. Da eine Substratspannung nicht nur V_t (und damit den On-Strom), sondern auch die Gate- und Parasitär-Kapazitäten sowie den Stack-Effekt beeinflusst, wirkt Body Biasing in vielfältiger Weise auf die Schaltgeschwindigkeit in CMOS-Schaltungen.

Der primäre Effekt des Body Biasing ist die Veränderung der Schwellenspannung und damit des On-Stroms (Abb. 4.6). Wegen des wurzelförmigen Verlaufs in Abbildung 4.1 ist die On-Strom-Zunahme für Forward-Biasing (+ 30 %) größer als die On-Strom-Abnahme für Reverse-Biasing ($-$ 20 %). Allerdings erhöhen sich im Gegenzug auch die parasitären Junction-Kapazitäten, da sich die Weite der Raumladungszone bei Forward-Biasing reduziert.

Für die Evaluation der Schaltgeschwindigkeiten werden in der 90 nm-Technologie Ringoszillatoren für jeden Transistortyp untersucht. Abbildung 4.7 zeigt die relativen Schaltgeschwindigkeiten des LL-Ringoszillators für unterschiedliche Substratspannungen. Unter Forward-Biasing erhöht sich die Frequenz bei $V_{dd} = 1.2\,\mathrm{V}$ um 34 %. Alternativ kann die Versorgungsspannung bei konstanter Frequenz um 180 mV abgesenkt werden. Die Kennlinie verschiebt sich in erster Näherung parallel nach links.

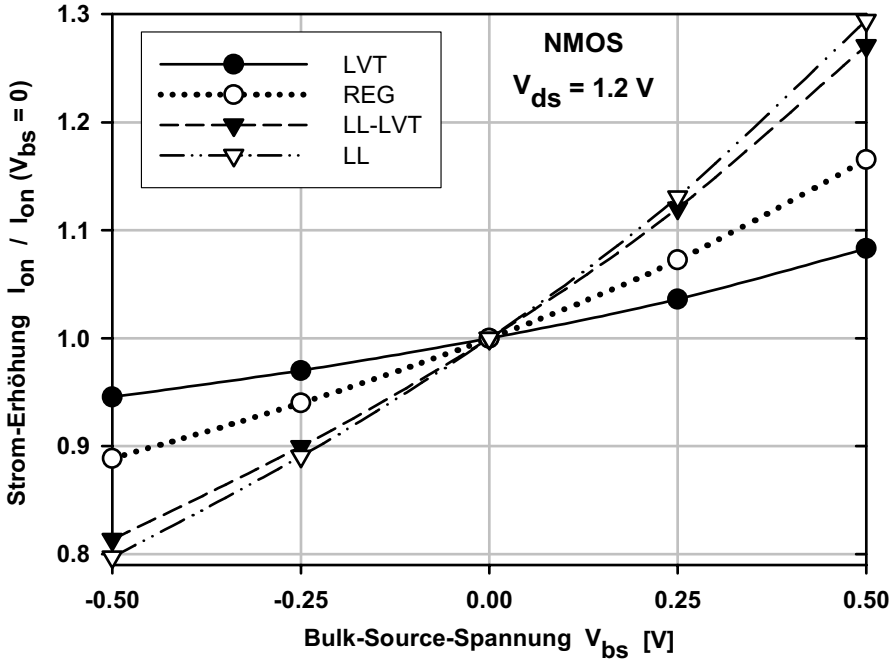


Abbildung 4.6: Abhängigkeit des On-Stroms von der Bulk-Source-Spannung.

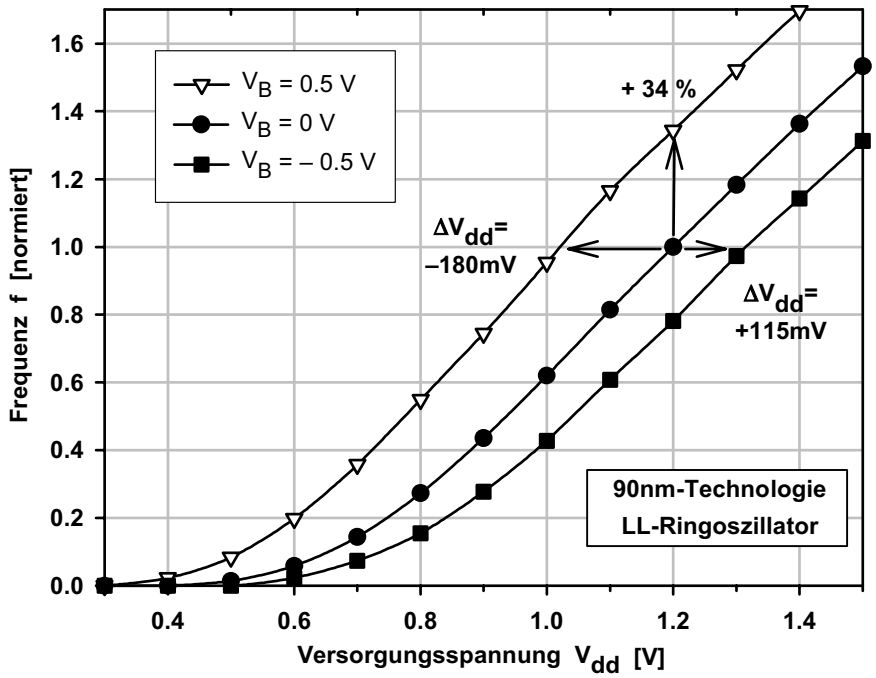


Abbildung 4.7: Frequenz eines LL-Ringoszillators in Abhängigkeit von der Versorgungsspannung unter bei Forward- und Reverse-Biasing.

	LVT	REG	LL-LVT	LL
NFET/PFET	8 % / 12 %	17 % / 19 %	27 % / 20 %	30 % / 27 %
Ringoszillator	10 %	18 %	28 %	34 %
32-bit-Addierer	16 %	28 %	—	37 %

Tabelle 4.2: Vergleich der relativen Erhöhungen der On-Ströme von Einzeltransistoren sowie der Frequenzen der Ringoszillatoren und Addierer durch Forward-Biasing von $V_B = 0.5\,\mathrm{V}$ ($T = 25\,^{\circ}\mathrm{C}$, $V_{gs} = V_{ds} = 1.2\,\mathrm{V}$ bzw. $V_{dd} = 1.2\,\mathrm{V}$).

Der Vergleich der Geschwindigkeitszunahme der Ringoszillatoren (34 %) mit der On-Strom-Zunahme (30 %) in Abbildung 4.6 bzw. der Vergleich der Verschiebung der Kennlinie (180 mV) mit der Verschiebung der Schwellenspannung in Abbildung 4.1 (145 mV) zeigt, dass sich Forward-Biasing stärker auf das dynamische Verhalten einer Schaltung auswirkt, als anhand des statischen Verhaltens zu erwarten ist.

Hierfür verantwortlich ist die höhere Effizienz von Forward-Biasing bei kleinerer Drain-Spannung $V_{ds} < V_{dd}$. Im Abschnitt 3.2.1 wurde die Relevanz dieses Bereiches für einen digitalen digitalen Schaltvorgang aufgezeigt (Abb. 3.22). Dieser Effekt wirkt sich so stark aus, dass die Geschwindigkeitsreduzierung, die sich aus den größeren Kapazitäten unter Forward-Biasing-Bedingungen ergeben, mehr als ausgeglichen wird.

Da die Geschwindigkeit einer digitalen CMOS-Schaltung jedoch in verstärktem Maße von der Verzögerungszeit komplexer Gatter mit höherem Fan-out abhängt, müssen die Ergebnisse für die Inverter-Ringoszillatoren auch an komplexen Schaltungen verifiziert werden. Tabelle 4.2 vergleicht die relativen Geschwindigkeitserhöhungen durch Forward-Biasing für Einzeltransistoren, Ringoszillatoren sowie die 32-bit-Addierer.

Es fällt auf, dass die Geschwindigkeitserhöhungen bei den komplexen Addierer-Schaltungen noch einmal höher sind als bei den entsprechenden Inverter-Ringoszillatoren. Dieser Effekt kann auf die im Abschnitt 3.2.3 beschriebene kleinere effektive Spannung in Gattern mit gestackten Transistoren zurückgeführt werden (vgl. Abb. 3.29). Die Gate-Source-Spannung des oberen Transistors ist während des gesamten Schaltvorgangs reduziert ($V_{gs} < V_{dd}$). Bei kleineren Spannungen wirkt sich daher die reduzierte Schwellenspannung prozentual stärker auf den Drain-Strom aus. Folglich ist Forward-Biasing in komplexen Schaltungen effizienter als in Inverter-Ringoszillatoren. Daraus resultiert für die 32-bit-Addierer eine deutlich höhere Effizienz von Forward-Biasing als für die Ringoszillatoren, insbesondere für die LVT- und REG-Addierer (von 10 % auf 16 %, bzw. von 18 % auf 28 %, Tabelle 4.2).

Für den LL-Addierer ist der Zugewinn jedoch vergleichsweise klein (von 34 % auf 37 %). Dieses lässt sich anhand Abbildung 4.8 erklären. Aufgrund des nicht-linearen Verlaufs der $V_t(V_{bs})$-Kennlinie in Abbildung 4.1 ist der 'Self-Reverse-Biasing-Effekt' in einer Serienschaltung unter Forward-Biasing stärker ausgeprägt als ohne Body Biasing. Die Schwellenspannung des oberen Transistors ist bei Forward-Biasing um 71 mV höher als die des unteren Transistors, für $V_B = 0$ beträgt die Differenz nur 32 mV. Die Nicht-Linearität der V_t-Kennlinie ist bei den

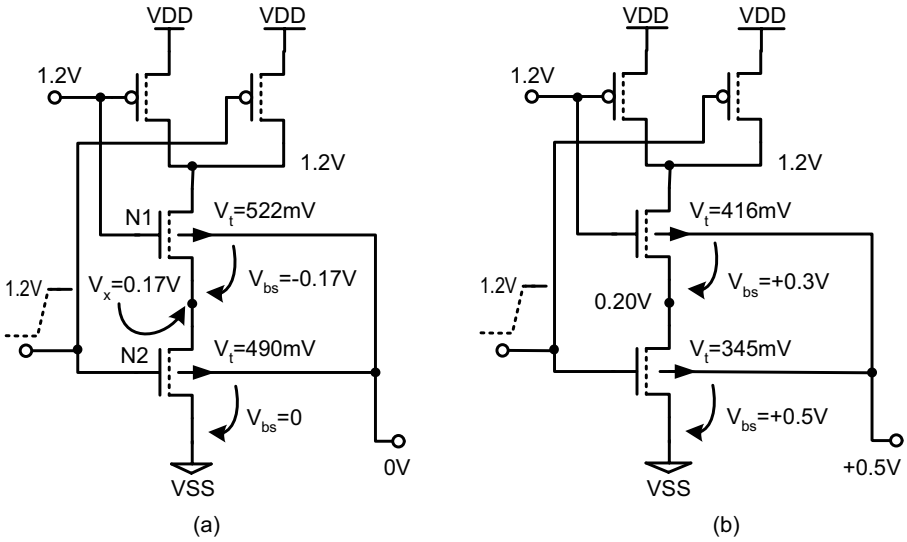


Abbildung 4.8: Schwellenspannung von Transistoren in Serienschaltung unter verschiedenen Body-Biasing-Verhältnissen. Ohne Body Biasing (a) besitzt der obere Transistor N1 im Verhältnis zu N2 ein um 522 mV – 490 mV = 32 mV höheres V_t. Liegen an dem Gatter hingegen 0.5 V Forward-Biasing an, so erhöht sich die Differenz auf 416 mV – 345 mV = 71 mV. Dieser so genannte Self-Reverse-Biasing-Effekt vermindert die Effizienz von Forward-Biasing für Schaltungen mit Serienschaltungen und wirkt damit dem Effekt in Abbildung 3.27 entgegen.

LL-Transistoren wegen des großen Substrateffekts sehr ausgeprägt, sodass sich der Effekt bei diesem Transistortyp am stärksten auswirkt. Der Self-Reverse-Biasing-Effekt wirkt dem Effekt der erhöhten Effizienz von Forward-Biasing in Gattern mit Serienschaltungen (Abb. 3.27) entgegen, insbesondere bei dem LL-Addierer. Daher ist die Geschwindigkeitserhöhung bei dem LL-Addierer mit 37 % nur geringfügig höher als bei dem LL-Ringoszillator mit 34 %. Für die LVT- und REG-Transistoren ist der Effekt weniger ausgeprägt. Hier erhöht Forward-Biasing die maximale Frequenz der Addierer stärker als die der Ringoszillatoren.

Auch für das Forward-Biasing ergeben sich verschiedene Probleme, die bezüglich eines Einsatzes in einer System-on-Chip-Umgebung gelöst werden müssen. Zwar sind die Forward-Biasing-Potentiale einfacher zu generieren, da diese zwischen V_{dd} und V_{ss} liegen. Allerdings müssen die Potentiale, die in der Nähe von $V_{dd}/2$ liegen, über Transmission-Gates geschaltet werden, deren Widerstand gerade bei $V_{dd}/2$ maximal ist. Ein niederohmiger Anschluss ist jedoch erforderlich, um Latch-up-Bedingungen zu verhindern. Die zur Spannungsgenerierung benötigte Energie wirkt sich bei Forward-Biasing weniger störend aus, da diese nur im aktiven Modus der Schaltung zur Verfügung stehen muss. Allerdings treten bei Forward-Biasing erhöhte Diodenleckströme auf, die von der Spannungsquelle aufgenommen werden müssen. Prinzipiell ist Forward-Biasing zudem auf maximal 0.5 V beschränkt, da bei höheren Substratspannungen die Substrat-Source-Dioden in Vorwärtsrichtung gespannt werden und gerade bei erhöhten Temperaturen leitend werden. Insbesondere bei höheren Temperaturen muss daher das Forward-Biasing unter Umständen noch weiter beschränkt werden.

4.3 Praktische Anwendbarkeit

In Tabelle 4.3 sind die Herausforderungen für Forward und Reverse-Biasing gegenübergestellt, die sich bei einer praktischen Implementierung in einer System-on-Chip-Umgebung stellen.

Insgesamt beschränken die Probleme beim Reverse-Biasing eine effiziente Anwendbarkeit auf wenige Spezialfälle. So wird in [76] ein Dickoxid-Transistor für den Einsatz in SRAMs so modifiziert, dass der Off-Strom nicht mehr durch den GIDL-Anteil dominiert ist.

Im Gegensatz hierzu kann Forward-Biasing eine praktikable Möglichkeit darstellen, um das Verhältnis zwischen Leckstrom und Schaltgeschwindigkeit zu verbessern. Jedoch auch für die praktische Anwendbarkeit von Forward-Biasing stellen sich einige in Tabelle 4.3 zusammengestellte Herausforderungen.

Der zusätzliche Flächenbedarf ist, anders als im SRAM-Teil, im Logikteil eines Chips weniger kritisch. Eine Begrenzung auf $V_B = 0.3\,\mathrm{V}$ kann jedoch erforderlich sein, da zum einen über die vorwärts gespannten Dioden zwischen den Source-/Drainkontakten und den Wannen bei hohen Temperaturen (125 °C) ein großer Leckstrom fließt. Zum anderen kann hierdurch eine höhere Latch-up-Immunität gewährleistet werden. Selbst bei einem reduzierten Forward-Biasing kann z.B. die maximale Frequenz des REG-Addierers bei $V_{dd} = 1\,\mathrm{V}$ um 20 % erhöht werden.

Latch-up ist ebenfalls bei hohen Temperaturen kritisch. Der Widerstand der Wanne wird wegen der kleineren Ladungsträgerbeweglichkeit größer und es kann schneller zu einem latch-up-kritischen Zustand kommen, insbesondere da die vier Versorgungspotentiale (V_{dd}, V_{ss}, $V_{dd,bulk}$ und $V_{ss,bulk}$) jeweils getrennt über große Distanzen auf dem Chip zugeführt werden. Die Zuleitungen sind damit unabhängig voneinander unterschiedlichen Störeinflüssen ausgesetzt, die im schlechtesten Fall in entgegengesetzter Richtung wirken (z.B. Crosstalk oder Schwankungen der Versorgungsspannungen). Der Effekt verstärkt sich noch, wenn die Spannungen außerhalb des Chips erzeugt werden.

Body Biasing kann durch Trennung der Wannen auf einzelne geschwindigkeitskritische Schaltungsblöcke begrenzt werden und auch I/O-Blöcke sowie SRAM-Blöcke können ausgespart werden. Insbesondere in SRAM-Blöcken besteht die Gefahr, dass der Speicherinhalt einzelner Zellen durch das Umschalten der Wannen-Potentiale verloren gehen kann.

In diesem Kapitel wurden die Auswirkungen von Reverse- und Forward-Biasing auf die Leckströme und Schaltgeschwindigkeiten von Einzeltransistoren, Ringoszillatoren und einen Parallel-Addierer aufgezeigt. Trotz kleiner werdendem Substrateffekt in Sub-100 nm-CMOS-Technologien stellt vor allem Forward-Biasing eine aussichtsreiche Möglichkeit dar, die Schaltgeschwindigkeit im aktiven Modus zu erhöhen.

Ein effizienter Einsatz in einer System-on-Chip-Anwendung erfordert darüber hinausgehende Bewertungen, die jedoch von der zu implementierenden Funktionalität der Schaltung abhängen und damit nicht mehr Bestandteil dieser Arbeit sind. Zu untersuchen sind hierbei unter anderem die Zuverlässigkeit einer Schaltung, die Anfälligkeit gegenüber Schwankungen der Versorgungsspannung und der Substratspannungen, die Möglichkeiten zur Spannungsgenerierung sowie die Höhe des zusätzlichen Designaufwands.

Forward-Biasing	Reverse-Biasing
• Niederomige Spannungsgenerierung zur Aufnahme erhöhter Stoßionisations- und Leckströme • Schalten von $\approx V_{dd}/2$ (maximaler Transmission-Gate-Widerstand) • Vermeidung von Latch-Up • Erhöhter Flächenbedarf (Routing, Generierung, drei Wannen) • Beschränkt auf maximal 0.5 V, für praktische Anwendung ≈ 0.3 V • Leckstrombegrenzung bei erhöhten Temperaturen und unter Einfluss von Prozess-Schwankungen • Zusätzliche Spannungsschwankungen auf $V_{ss,bulk}$ und $V_{dd,bulk}$	• Andere Leckstromkomponenten können Leckstrom dominieren • Leistungsarme Spannungsgenerierung im Standby-Modus • Schalten hoher Spannungen • Zuverlässigkeit der Transistoren • Erhöhter Flächenbedarf • Abnehmende Effizienz für stärkeres Reverse-Biasing • Abnehmende Effizienz für kleinere V_t und dünnere Oxide • Zusätzliche Spannungsschwankungen auf $V_{ss,bulk}$ und $V_{dd,bulk}$

Tabelle 4.3: Herausforderungen beim Einsatz von Forward-Biasing und Reverse-Biasing.

Kapitel 5

Neue Ansätze für Sub-100 nm-Schaltungstechnik

In neueren CMOS-Technologien eröffnet die Verfügbarkeit unterschiedlicher Schwellenspannungen und Oxiddicken neue Möglichkeiten, die Schaltfrequenz zu verbessern, ohne den Leckstrom einer Schaltung zu erhöhen. Hierzu werden in weniger geschwindigkeitskritischen Teilen, wie z.B. SRAM-Blöcken, höhere Schwellenspannungen eingesetzt. Weiterhin ist es möglich, innerhalb eines Schaltungsblocks geringere V_t's nur in Logikgattern einzusetzen, die in kritischen Pfaden liegen [48]. Zudem können in Standby-Transistoren größere Schwellenspannungen eingesetzt werden [77].

Als nächster Schritt ist denkbar, unterschiedliche Transistortypen innerhalb der einzelnen Logikgatter zu kombinieren [78]. In diesem Kapitel werden hierfür verschiedene Möglichkeiten mit ihren Vor- und Nachteilen vorgestellt.

5.1 Skewed-CMOS-Logik

In dem bei weitem größten Teil aller Logikschaltungen wird heute die klassische statische CMOS-Technik eingesetzt. Vorteile sind die geringe statische Verlustleistung, eine hohe Störsicherheit, ein einfach zu automatisierendes Design und verhältnismäßig kleine Schaltzeiten. Nur da, wo sehr hohe Schaltgeschwindigkeiten notwendig sind und sich der erhöhte Designaufwand lohnt, werden z.B. Schaltungstechniken eingesetzt, die nach einem Aufladen der Logikgatter (*Precharge*) eine höhere Schaltgeschwindigkeit erlauben.

Eine solche dynamische Domino-Logik ist in Abbildung 5.1 dargestellt [79]. Die erhöhte Schaltgeschwindigkeit resultiert aus dem fehlenden komplementären PMOS-Pfad in der ersten Stufe. Stattdessen wird der Knoten $\overline{Z}$ vor jeder aktiven Phase des Gatters über den Transistor P1 in einer Precharge-Phase auf V_{dd} aufgeladen. Nach diesem Aufladen wird P1 geschlossen und der logische Zustand des Gatters wird nur noch durch die auf dem Knoten $\overline{Z}$ gespeicherte Ladung repräsentiert (dynamische Logik). Danach kann das Gatter über die geschwindigkeitsoptimierten NMOS-Transistoren N2–4 entladen werden. Allerdings kann ein Domino-Gatter in jeder aktiven Phase nur einmal entladen werden. Wenn der Ausgang $\overline{Z}$ des AND-NOR-Gatters

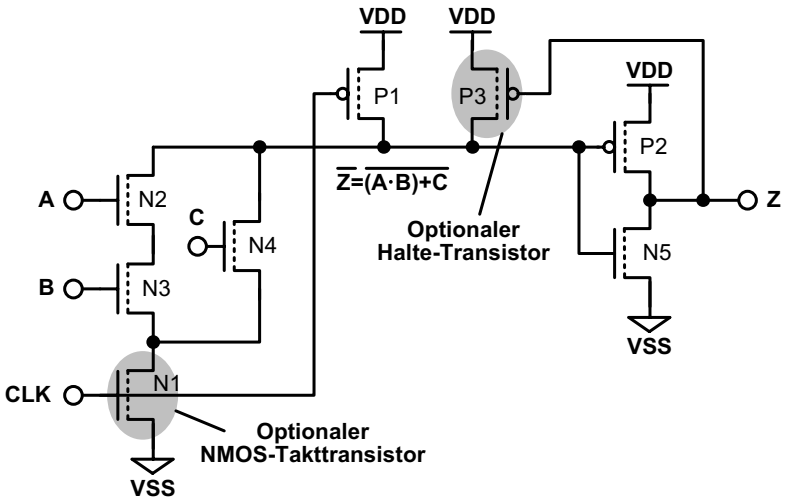


Abbildung 5.1: Domino-Logik-Gatter mit der Funktionalität $Z = (A \cdot B) + C$. Diese Funktionalität ist hier und in den folgenden Abbildungen beispielhaft dargestellt, da sie sowohl Serien- als auch Parallelschaltungen in NMOS- und, wenn vorhanden, PMOS-Pfad enthalten. Außerdem wird die AND-NOR-Funktion in Parallel-Prefix-Addierern häufig verwendet (Kapitel 6).

in Abbildung 5.1 einmal entladen ist, so bleibt dieser Zustand erhalten, selbst wenn die Eingänge zurückschalten. Dieses wird als monotones Schaltverhalten bezeichnet. Da die Eingänge eines Domino-Gatters wiederum von einem vorhergehenden Domino-Gatter stammen, das ebenfalls ein monotones Schaltverhalten besitzt, ist dieses zunächst kein Nachteil.

In der Domino-Logik ist es nicht ohne weiteres möglich, ein invertiertes Signal zu erzeugen (Nicht-invertierende Logik). Nur an der Grenze zwischen zwei Taktphasen kann eine Inversion realisiert werden. Hiervon ausgehend kann an jeder anderen Stelle in der Schaltung ebenfalls ein invertiertes Signal zur Verfügung gestellt werden, indem die Logikschaltung in invertierter Form ein zweites Mal aufgebaut wird (*Dual Rail*). Alternativ können differentielle Logiken eingesetzt werden, bei denen in jedem Gatter nicht-invertierte und invertierte Signale gleichzeitig vorhanden sind [80].

Der damit verbundene zusätzliche Flächen- und Energiebedarf kann durch Umformung der zu berechnenden Logikfunktion in einigen Fällen erheblich reduziert werden. In der Regel ist dieses bei einer komplexen Kontroll-Logik schwieriger als in regulären Datenpfad-Strukturen wie Multiplizierern, Filterschaltungen oder einem Parallel-Addierer. Ein solcher Addierer wird im Abschnitt 6 beschrieben.

In [30] wird in die zweite Stufe des Domino-Gatters zusätzlich eine statische Logikfunktion integriert. Dadurch werden die Schaltzeiten sowie der Flächen- und Energieverbrauch reduziert. Am monotonen Schaltverhalten ändert sich jedoch nichts.

Die Funktionssicherheit von Domino-Logik-Schaltungen wird durch Leckströme, kapazitive Kopplungen, Parametervariationen und Schwankungen in der Versorgungsspannung stärker beeinträchtigt als die von statischen CMOS-Schaltungen. Es existieren viele Vorschläge, dyna-

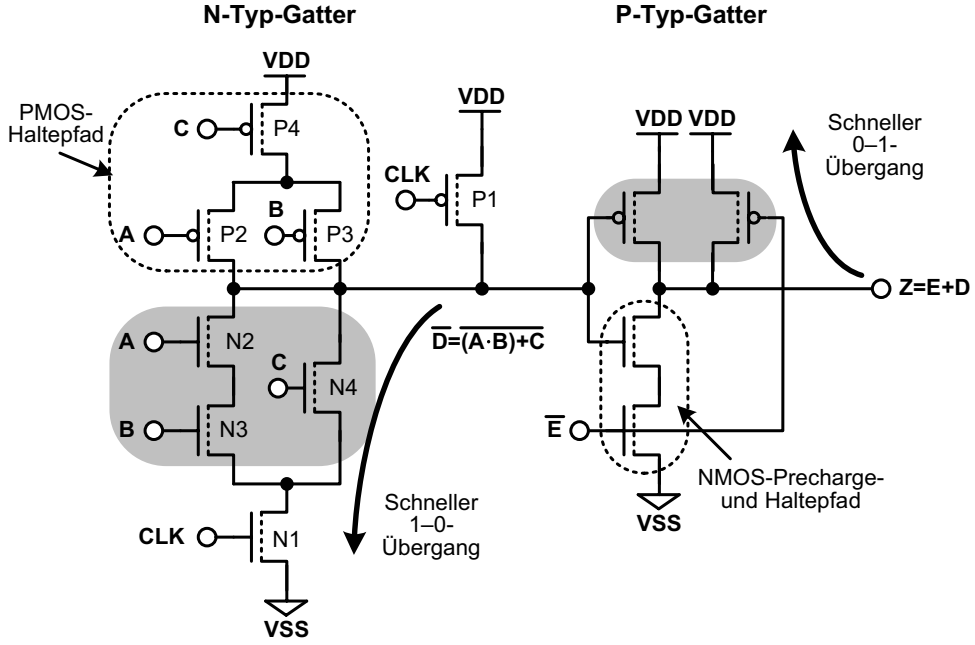


Abbildung 5.2: Skewed-CMOS-Gatter.

mische Logiken robuster gegen solche Einflüsse zu machen [81, 82]. Zum Beispiel wird ein Halte-Transistor (*Keeper* P3, Abb. 5.1) hinzugefügt, der Ladungsverluste durch die Leckströme ausgleichen soll. Dieser Keeper reduziert jedoch auch die Schaltgeschwindigkeit, da die NMOS-Logiktransistoren den Knoten $\overline{Z}$ entladen müssen, während über den Keeper eine leitende Verbindung zu V_{dd} besteht. Kommen Parametervariationen hinzu, kann das Gatter unter ungünstigen Bedingungen nor noch sehr langsam oder gar nicht schalten (starker Keeper und schwache gestackte NMOS-Transistoren).

Dieser Zustand wird bei Schaltungen in Skewed-CMOS-Logik verhindert [83]. Anstelle eines Haltetransistors wird ein zum NMOS-Logikpfad komplementärer PMOS-Pfad wie in statischen CMOS-Schaltungen hinzugefügt (Abb. 5.2). Verglichen mit der Domino-Logik steigt die Eingangslast der Gatter und die Schaltgeschwindigkeit wird reduziert. Die PMOS-Transistoren werden jedoch im Vergleich zu statischer CMOS-Logik kleiner ausgelegt, da sie nicht zum geschwindigkeitskritischen Schalten, sondern nur zum Ausgleich von Leckströmen eingesetzt werden. Während des Schaltvorgangs arbeiten die NMOS-Transistoren des N-Typ-Gatters nicht mehr gegen einen Keeper, wodurch die Dimensionierung vereinfacht wird und Querströme vermieden werden.

Das Gatter besitzt einen schnellen und einen langsamen Pfad, die Asymmetrie führt zu der Bezeichnung *Skewed CMOS*. Neben unterschiedlichen Transistorweiten kann der Skew-Effekt auch durch unterschiedliche Schwellenspannungen (Multi-V_t) und Oxiddicken (Multi-t_{ox}) erreicht oder verstärkt werden. Die Möglichkeit, Logikransistoren mit unterschiedlichen Oxiddicken zu kombinieren, bietet sich erst seit Einführung der 90 nm-Technologie. Die neuen schaltungstechnischen Möglichkeit werden bisher erst wenig genutzt.

5.1.1 Funktionsweise

Wie im vorhergehenden Abschnitt beschrieben, besitzt Skewed-CMOS-Logik ein monotones Schaltverhalten. Gatter mit schnellem NMOS- und PMOS-Pfad (N- und P-Typ-Gatter) müssen sich daher stets abwechseln. Die erste Stufe der in Abbildung 5.2 dargestellten Schaltung ist ein AND-NOR-N-Typ-Gatter mit weiten NMOS-Transistoren, die in der aktiven Phase eine schnelle Auswertung erlauben. Im Precharge-Modus wird der NMOS-Takttransistor N1 geschlossen, sodass der Knoten $\overline{D}$ über den PMOS-Takttransistor P1 aufgeladen werden kann. Die PMOS-Logiktransistoren P2–4 dienen lediglich dazu, den während des Precharge-Vorgangs aufgeladenen Knoten $\overline{D}$ nach Beendigung des Precharge-Vorgangs auf 1 zu halten und können daher kleiner dimensioniert werden. Zusätzlich kann der PMOS-Pfad des N-Typ-Gatters den Precharge-Vorgang unterstützen, der hauptsächlich über den Transistor P1 erfolgt.
Auch in Skewed-CMOS-Logik kann in der zweiten Stufe eine zusätzliche Logik-Funktion integriert werden. In Abbildung 5.2 ist dieses z.B. ein NAND-Gatter. Der Eingang $\overline{E}$ muss dabei ebenfalls von einem N-Typ-Gatter stammen, damit während des Precharge-Vorgangs alle Eingänge der P-Typ-Gatter auf 1 liegen. Die zweite Stufe arbeitet invers zur ersten Stufe. Während des Precharge-Vorgangs entläd sich der Knoten Z auf 0. Dieses geschieht im Gegensatz zur ersten Stufe nicht über einen Precharge-Transistor, sondern über den NMOS-Pfad des Gatters. Der NMOS-Pfad ist geöffnet, sobald alle Eingänge (= Ausgänge der N-Typ-Gatter) auf 1 aufgeladen sind. Die vergleichsweise hohen On-Ströme der NMOS-Transistoren gewährleisten eine relativ schnelle Entladung, auch wenn kleine Transistorweiten verwendet werden.
Der Precharge-Vorgang ist abgeschlossen, wenn alle Ausgänge der P-Typ-Gatter auf 0 entladen sind. Die NMOS-Takttransistoren der N-Typ-Gatter können geöffnet werden ohne die Ausgänge zu entladen, da die NMOS-Logiktransistoren der N-Typ-Gatter geschlossen sind.

Die Erhöhung der Schaltgeschwindigkeit beruht zum einen darauf, dass die Eingangskapazität der Logikgatter durch die Reduzierung der Transistorweiten in den Haltepfaden verkleinert wird. Zum anderen schalten die Gatter früher, weil die statischen Schaltpunkte im Vergleich zu statischen CMOS-Gattern verschoben sind. Abbildung 5.3 zeigt die Transferkennlinien von statischer und Skewed-CMOS-Logik. Sowohl die statischen als auch die dynamischen Kennlinien für den 1–0- und den 0–1-Übergang sind im Vergleich zur statischen CMOS-Logik zu früheren Schaltpunkten verschoben.
Skewed-CMOS-Schaltungen benötigen ein zweiphasiges Taktschema, um die Precharge-Phase ohne eine Halbierung der Datenrate zu realisieren (Abb. 5.4). Während $CLK = 1$ ist, befindet sich die erste Hälfte einer Schaltung im aktiven Modus. Die zweite Hälfte, die mit einem inversen Taktsignal $\overline{CLK} = 0$ versorgt wird, befindet sich im Precharge-Modus. Für $CLK = 0$ und $\overline{CLK} = 1$ ist dann die zweite Hälfte der Schaltung aktiv und die erste Hälfte läd auf.
Wegen der notwendigen Precharge-Phase unterscheiden sich die Signale in statischen und Skewed-CMOS-Schaltungen grundlegend. Es werden daher spezielle Flip-Flops und Latches benötigt, sowohl um die beiden Taktphasen voneinander zu trennen, als auch um Skewed-CMOS-Schaltungen mit statischen Schaltungen kombinieren zu können (Abb. 5.4). Die Möglichkeit dieser Kombinierbarkeit stellt eine wichtige Voraussetzung für die praktische Nutzbarkeit der Skewed-CMOS-Logik dar. Denn aufgrund des monotonen Schaltverhaltens eignet sich

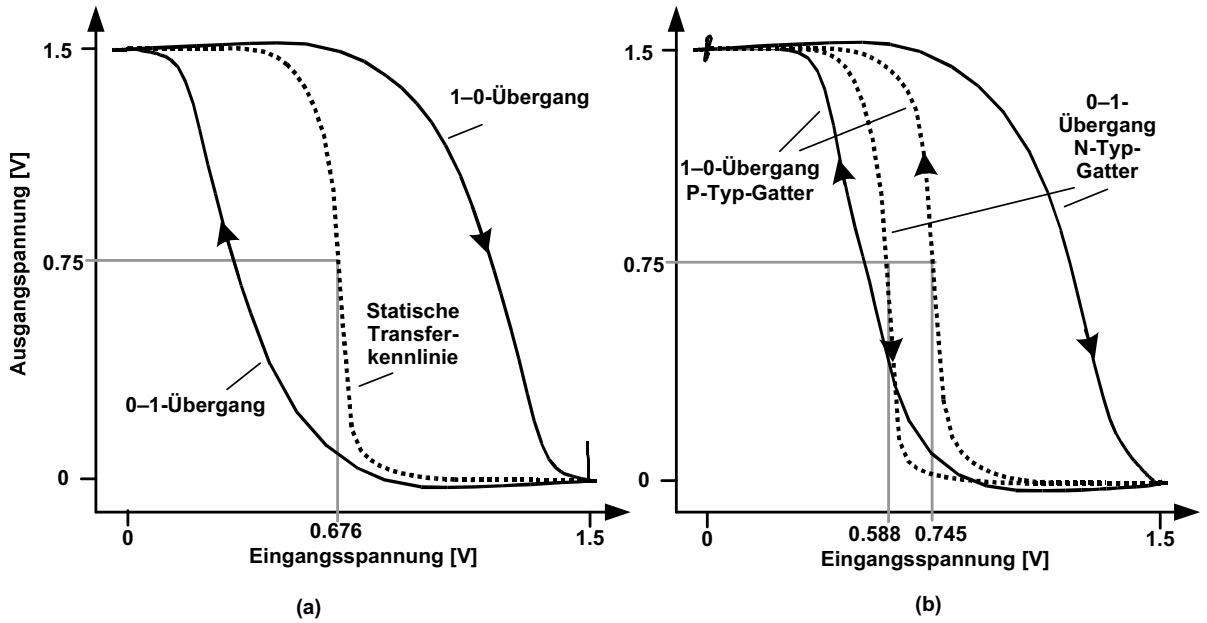


Abbildung 5.3: Statische und dynamische Transferkennlinien in statischer (a) und Skewed-CMOS-Logik (b) mit Fan-out 3. Der statische Schaltpunkt des statischen CMOS-Gatters liegt nicht genau bei $V_{dd}/2$, da dieses auf optimale Geschwindigkeit dimensioniert ist (Kapitel 3.2.1). Die Transistorweiten der Skewed-CMOS-Gatter ergeben sich aus den Dimensionierungsregeln im Kapitel 5.1.4.

diese Schaltungstechnik nicht für automatisierte Design-Umgebungen. Nur in einem Full- oder Semi-Custom-Design besteht die Möglichkeit, einen geschwindigkeitskritischen Schaltungsteil in Skewed-CMOS-Logik zu realisieren und diesen in die Gesamtschaltung zu integrieren. Flip-Flops und Latches, die speziell für Skewed-CMOS-Logik ausgelegt sind, werden im Abschnitt 5.3 vorgestellt.

Das monotone Schaltverhalten hat jedoch nicht nur Nachteile, sondern bietet auch neue Möglichkeiten zur Leckstromreduzierung, da nach Beendigung des Precharge-Vorgangs alle Zustände in der Schaltung wohl definiert und bekannt sind.

5.1.2 Möglichkeiten zur Reduzierung des Leckstroms

Minimum Leakage Vector

Wie im Abschnitt 3.1.1 gezeigt, führt das Anlegen eines *Minimum Leakage Vector* (MLV) nur zu einer geringen Reduzierung des Leckstroms, wenn, wie in statischen CMOS-Schaltungen üblich, eine unregelmäßige Schaltungsstruktur vorliegt. Skewed-CMOS-Schaltungen sind hingegen sehr regelmäßig aufgebaut, da ein strikter Wechsel zwischen N- und P-Typ-Gattern eingehalten werden muss. Es ist daher möglich, die geschwindigkeitskritischen Pfade der Gatter auf hohe Schaltgeschwindigkeit und die nicht-kritischen Pfade unabhängig davon auf einen kleinen Leckstrom zu optimieren.

Abbildung 5.5 zeigt den MLV-Modus für Skewed-CMOS-Schaltungen. Alle Precharge- und

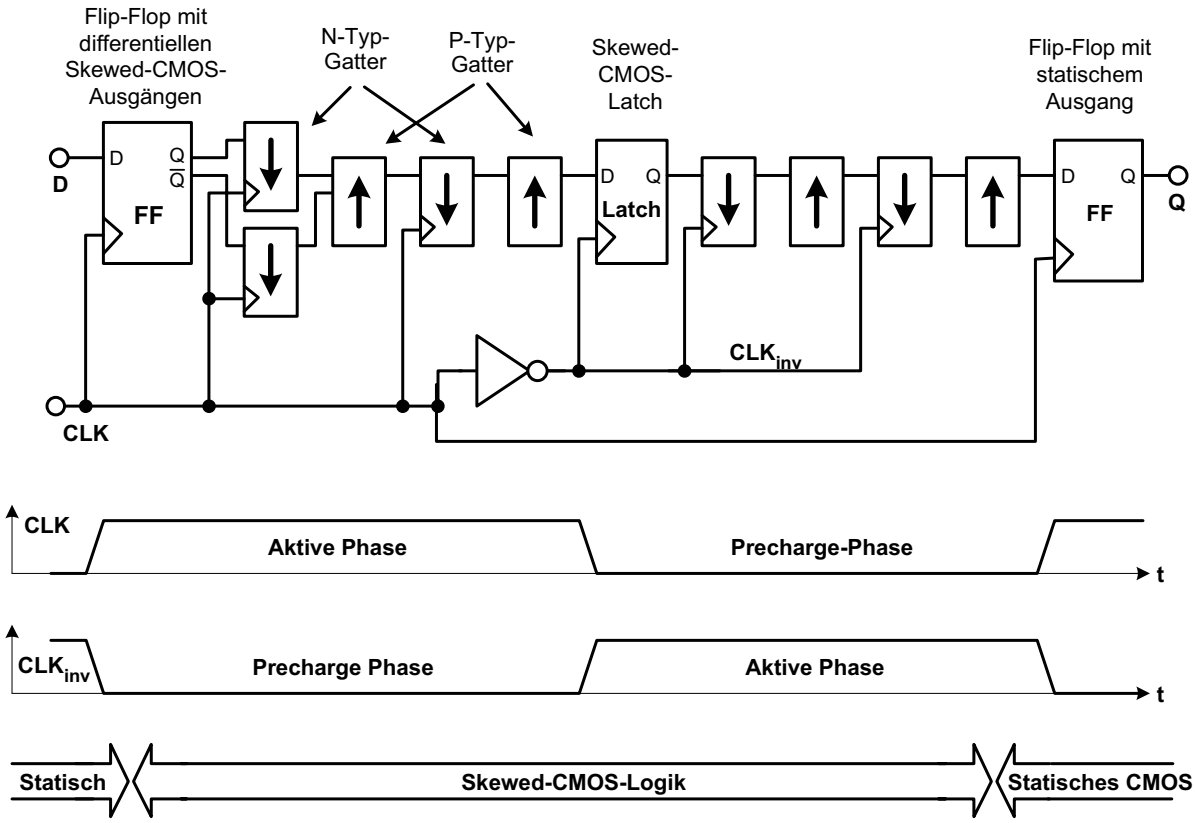


Abbildung 5.4: Zweiphasiges Taktschema für Skewed-CMOS-Logik. Die dargestellten Skewed-CMOS-Gatter sind mit Hilfe spezieller Flip-Flops in eine statische CMOS-Logik eingebettet. Es kann sich jedoch alternativ auch eine weitere Skewed-CMOS-Schaltung in der nächsten Taktphase anschließen. In diesem Fall müsste das Ausgangs-Flip-Flop mit statischen Ausgängen durch ein Latch oder ein Flip-Flop mit vorgeladenen Ausgängen ersetzt werden.

Halte-Transistoren verwenden eine höhere Schwellenspannung, während die schnellen Evaluationspfade kleinere Schwellenspannungen besitzen. In einem MLV-Standby-Modus werden an die Eingänge des ersten N-Typ-Gatters in einem Datenpfad V_{dd}-Potentiale angelegt. Außerdem ist $CLK = V_{dd}$. Flip-Flops, die diese Aufgabe übernehmen können, werden im Abschnitt 5.3 vorgestellt. Die High-V_t-Transistoren sind geschlossen und die Ausgänge der N-Typ-Gatter liegen auf V_{ss}-Potential. Dadurch wiederum sperren die High-V_t-NMOS-Transistoren der P-Typ-Gatter. Dieser leckstromarme Zustand propagiert durch die gesamte Schaltung, da die Ausgänge der P-Typ-Gatter (=V_{dd}) wiederum die Eingänge für die nächsten N-Typ-Gatter bilden. Anders als bei der Festlegung eines MLV für eine statische CMOS-Schaltung (Abb. 3.11) ist in diesem Fall kein aufwendiges Suchverfahren notwendig, da alle Eingänge auf V_{dd}-Potential gelegt werden. Der Designaufwand wird dadurch reduziert.

Der MLV-Standby-Modus führt zu einer signifikanten Reduzierung des Leckstroms, wenn ausschließlich Transistoren verwendet werden, deren Gateleckstrom klein ist. Werden dünne Oxide verwendet, trägt der Gateleckstrom der eingeschalteten Transistoren $I_{g,on}$ erheblich zur Verlust-

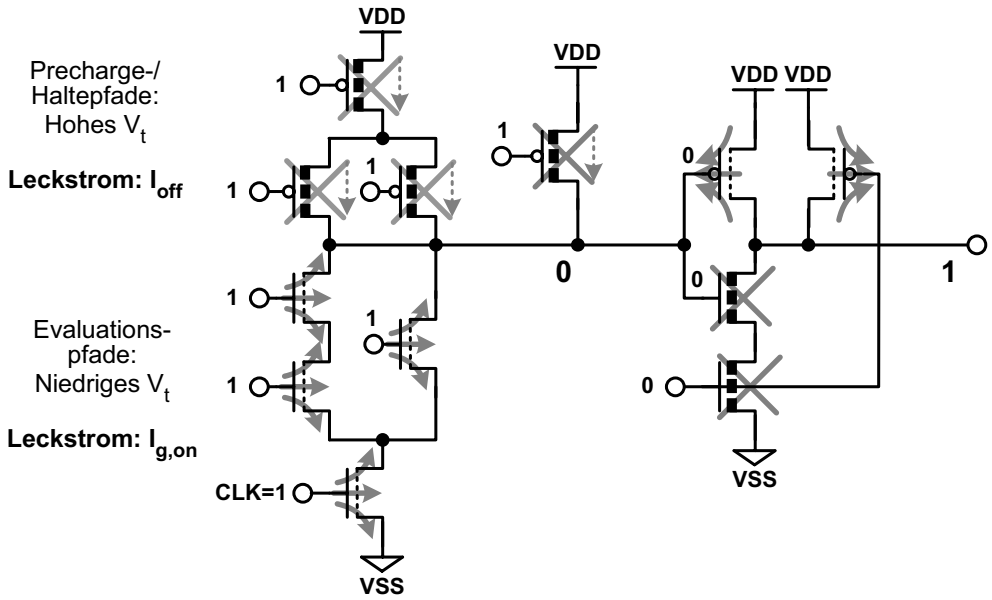


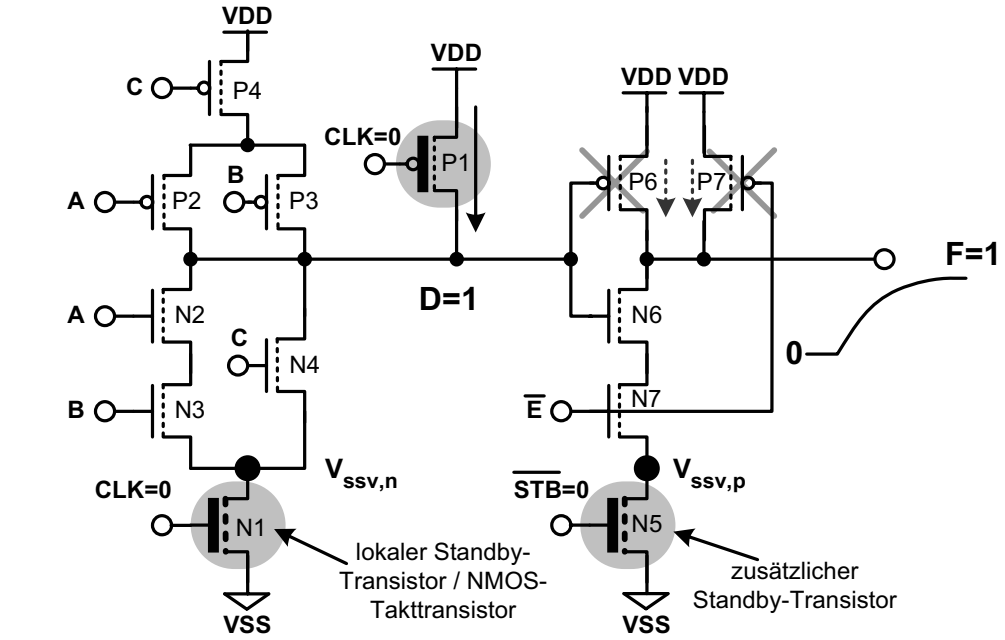
Abbildung 5.5: Minimum-Leakage-Vector-Modus in Skewed-CMOS-Multi-t_{ox}-Logik.

leistung bei. Wenn der On-Gateleckstrom die bestimmende Leckstromkomponente ist, wird die Verlustleistung durch das Anlegen von V_{dd} an die Eingänge der N-Typ-Gatter sogar erhöht, da die weiten Evaluations-Transistoren einen besonders hohen Gateleckstrom aufweisen. Meistens stehen jedoch auch Transistoren zur Verfügung, die zwar hohe Unterschwellenströme, aber einen kleinen Gateleckstrom aufweisen, sodass der MLV-Modus sinnvoll einsetzbar ist, insbesondere wenn High-κ-Dielektrika verfügbar sind oder Multi-Gate-Transistoren verwendet werden.

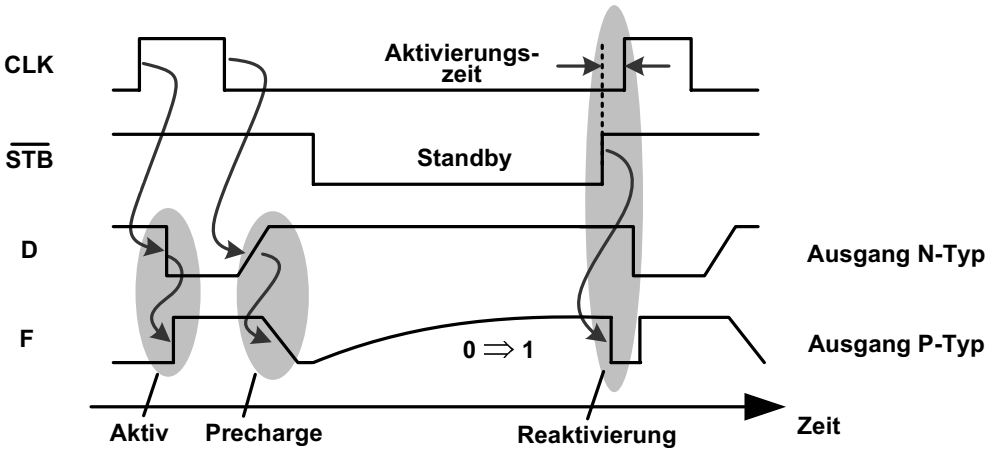
Skewed-CMOS-Logik mit lokalen Standby-Transistoren

Ein alternativer Ansatz zur Reduzierung des Leckstroms in Skewed-CMOS-Schaltungen besteht darin, den NMOS-Takttransistor N1 in Abbildung 5.6 als lokalen Standby-Transistor für die N-Typ-Gatter zu nutzen. In den P-Typ-Gattern wird ein zusätzlicher Standby-Transistor hinzugefügt, sodass die gesamte Logikschaltung in diesem Standby-Modus von V_{ss} getrennt werden kann (Skewed-CMOS-Standby-Modus). Beide Transistoren besitzen einen niedrigen Off-Strom, in Abbildung 5.6 besitzen sie ein dickeres Oxid und eine höhere Schwellenspannung. Der PMOS-Precharge-Transistor P1 besitzt ebenfalls ein dickes Oxid, da hier $CLK = 0$ anliegt. Andernfalls würden hier Gateleckströme ($I_{g,on}$) auftreten.

Der N-Typ-Standby-Transistor befindet sich während der aktiven Phase im kritischen Pfad. Es besteht daher die Gefahr, dass die Schaltgeschwindigkeit signifikant reduziert wird. Da das Taktsignal den Transistor N1 jedoch schon öffnet, bevor der Knoten D entladen wird, kann der NMOS-Stack schon bis zu den Source-Kontakten der Transistoren N3 und N4 entladen werden. Außerdem kann N1 größer als die übrigen NMOS-Transistoren dimensioniert werden, da der Schaltvorgang nicht zeitkritisch ist. Die Degradation der Schaltgeschwindigkeit durch Verwen-



(a)



(b)

Abbildung 5.6: Skewed-CMOS-Multi-t_{ox}-Logik mit lokalen Standby-Transistoren (a) und Timing-Diagramm für den Standby-Modus (b). Alternativ zu der hier gezeigten Implementierung mit lokalen Standby-Transistoren können jeweils mehrere N-Typ- und P-Typ-Gatter an zentrale Standby-Transistoren über zwei virtuelle V_{ss}-Leitungen ($V_{ssv,n}$ und $V_{ssv,p}$) angeschlossen werden.

dung eines leckstromarmen Transistors kann damit zwar auf null reduziert werden, dieses geht jedoch zu Lasten einer erhöhten aktiven Taktlast.

Wird zum Beispiel ein LL- anstelle eines REG-Transistors verwendet, so muss die Gateweite und damit die kapazitive Last basierend auf Simulationen um 50 % erhöht werden, um den geringeren On-Strom auszugleichen. Die gesamte Leistungsaufnahme der Logikschaltung wird um 10 % erhöht, da die NMOS-Takttransistoren einen Anteil von etwa 20 % an der Schaltaktivität der Gesamtschaltung haben. Diese Berechnung beruht einerseits darauf, dass jedes zweite Gatter einen NMOS-Takttransistor besitzt und andererseits darauf, dass ein mittleres Logikgatter aus vier Transistoren besteht sowie eine Schaltwahrscheinlichkeit von 50 % aufweist. In Abbildung 5.2 z.B. ist einer von zwölf Transistoren betroffen.

Der Standby-Transistor des P-Typ-Gatters N5 liegt außerhalb des kritischen Pfades. Lediglich während der Precharge-Phase reduziert die zusätzliche Serienschaltung die Leitfähigkeit des NMOS-Pfades. Um den Flächenbedarf klein zu halten, sollten größere Transistorweiten hier nur verwendet werden, wenn der Precharge-Vorgang signifikant behindert wird.

Der Standby-Modus $\overline{STB} = 0$ der Skewed-CMOS-Logik mit lokalen Standby-Transistoren wird während der Precharge-Phase $CLK = 0$ eingeleitet (Abb. 5.6b). Der Transistor P1 leitet, sodass der Ausgang des N-Typ-Gatters 1 bleibt. Der Ausgang des P-Typ-Gatters ist isoliert und nähert sich langsam V_{dd}, da die Logiktransistoren P6 und P7 einen höheren Leckstrom aufweisen als der lokale Standby-Transistor N5. Für $F = 1$ ist ein leckstromarmer Zustand erreicht. Die statische Verlustleistung wird nur noch durch die Off-Ströme der lokalen Standby-Transistoren bestimmt.

Damit die Schaltung wieder für die nächste Auswertung bereitsteht, muss lediglich der P-Typ-Standby-Transistor N5 geöffnet werden. Der Ausgang des Gatters F entlädt sich sofort, da die Eingänge D und E durch PMOS-Takttransistoren auf 1 gehalten werden. Auf diese Weise kann verhindert werden, dass die Schaltung während der Reaktivierung unkontrollierte Zustände annimmt, die dann wiederum zu schnellen und unkontrollierten Schaltereignissen führen. Diese so genannten *Glitches* verhindern in Schaltungen mit zentralen Standby-Transistoren eine schnelle Rückkehr in den aktiven Zustand [84]. Glitches können hier nur durch langsames Einschalten vermieden werden, sodass die globale Spannungsversorgung nicht einbricht und umliegende Schaltungsteile in ihren Funktionen nicht beeinträchtigt werden.

Alternativ zu den lokalen Standby-Transistoren können sich mehrere N- und P-Typ Gatter jeweils einen Standby-Transistor teilen. Insbesondere reduziert sich dadurch der Spannungsabfall über dem Standby-Transistor, da die N-Typ-Gatter in der aktiven Phase nacheinander schalten. Die P-Typ-Gatter werden in der Precharge-Phase hingegen alle gleichzeitig über ihren Standby-Transistor entladen. Der Verdrahtungsaufwand steigt, da zwei virtuelle V_{ss}-Potentiale zu den Gattern geführt werden müssen. Gleichzeitig wird jedoch die Anzahl der Transistoren mit dickem Oxid in den Logikgattern reduziert und die Gesamtfläche der Standby-Transistoren kann ohne Geschwindigkeitsreduzierung vermindert werden. In [84] wird eine vergleichbare Technik für statische CMOS-Schaltungen vorgestellt.

5.1.3 Störsicherheit in einer System-on-Chip-Umgebung

Eine Schaltung in Skewed-CMOS-Logik besitzt wie eine statische CMOS-Schaltung komplementäre N- und PMOS-Pfade. Der Zustand eines Gatters wird aktiv durch Haltetransistoren aufrecht erhalten. Die Störsicherheit gegenüber Leckströmen, Parametervariationen und kapazitiven Kopplungen (Übersprechen, *Crosstalk*) wird dadurch im Vergleich zur Domino-Logik deutlich erhöht.

Leckströme: Im Gegensatz zur Domino-Logik treten in Skewed-CMOS-Schaltungen keine isolierten Knoten auf, deren logische Zustände sich durch Leckströme ändern können. Leckströme beeinträchtigen daher den Betrieb von Skewed-CMOS-Schaltungen nicht, sodass die Betriebssicherheit vergleichbar ist mit der statischer CMOS-Schaltungen.

Parametervariationen: Zum Ausgleich von Leckströmen werden Domino-Logik-Schaltungen häufig um einen PMOS-Haltetransistor erweitert. Während der Auswertung eines Gatters arbeiten die NMOS-Logiktransistoren jedoch gegen diesen Transistor. Liegt durch Parametervariationen ein starker PMOS-Transistor vor, kann die Auswertung verzögert werden und es tritt ein Querstrom auf. Im Extremfall schaltet das Gatter überhaupt nicht. Dieser Fall kann in Skewed-CMOS-Logik nicht auftreten.
Bei dem Vergleich der Auswirkungen von Parametervariationen auf die Schaltgeschwindigkeit in statischen und Skewed-CMOS-Schaltungen sind zwei gegenläufige Effekte zu beobachten. Zum einen ist die Serienschaltung der NMOS-Transistoren im N-Typ-Gatter länger geworden. Dadurch reduzieren sich die effektiven Gate-Source-Spannungen der Logiktransistoren und die Abhängigkeit gegenüber V_t-Schwankungen wird erhöht. Zum anderen sinkt die Schwankungsbreite der Verzögerungszeiten, da die Ausgangslast in Skewed-CMOS-Schaltungen aufgrund kleinerer Haltetransistoren reduziert ist. In Abbildung 5.7 werden die Verzögerungszeiten in statischen und Skewed-CMOS-Schaltungen für langsame, nominelle und schnelle Parametersätze verglichen. Nicht nur die absolute, sondern auch die relative Geschwindigkeitsveränderung ist in Skewed-CMOS-Schaltungen kleiner als bei statischer CMOS-Logik. Damit ist Skewed-CMOS weniger anfällig gegen Parametervariationen.

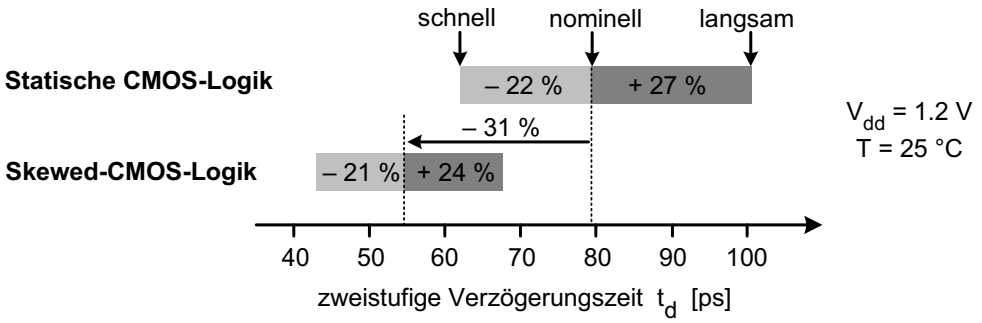


Abbildung 5.7: Vergleich der Variationen der Verzögerungszeiten in statischer und Skewed-CMOS-Logik, Verzögerungszeiten von zwei in Serie geschalteten NAND-Gattern mit Fan-out 2 für langsame, nominelle und schnelle Parametersätze.

Crosstalk: Während in einer dynamischen Logik ein Zustand nur durch eine kapazitiv gespeicherte Ladungsmenge repräsentiert wird, wird ein logischer Zustand in Skewed-CMOS-Gattern stets aktiv durch einen leitenden Transistorpfad zu V_{dd} oder V_{ss} gestützt. Dennoch ist Skewed-CMOS keine vollständig statische Logik. Ein einmal z.B. durch Crosstalk unbeabsichtigt entladener Ausgang eines N-Typ-Gatters wird zwar durch den PMOS-Haltepfad langsam wieder auf V_{dd} aufgeladen. Der 1–0-Übergang kann jedoch schon weitere schnelle Schaltereignisse in den nachgeschalteten Gattern ausgelöst haben, die über die Precharge-Pfade nur langsam rückgängig gemacht werden können. Die Skewed-CMOS-Schaltung vergrößert somit von Stufe zu Stufe die Länge des Störimpulses. Es kommt zu einer Fehlfunktion, wenn ein Flip-Flop am Ende der aktiven Taktphase das Datum während dieses vergrößerten Impulses übernimmt.

Crosstalk stellt somit ein ernst zu nehmendes Risiko für den Einsatz von Skewed-CMOS-Logik in einer System-on-Chip-Umgebung dar. Mittels einer Teststruktur auf einem 130-nm-Testchip werden Skewed-CMOS- und statische CMOS-Gatter unter Einfluss kapazitiver Störimpulse verglichen.

Auf dem Testchip wird nur die Signal-Integrität untersucht (vergleichbar mit Simulationen in [85]), hingegen nicht die Veränderung der Verzögerungszeiten [86]. Abbildung 5.8 zeigt eine von insgesamt 128 Teststrukturen. Ein Aggressor wirkt über zwei Inverter auf eine Metallleitung, die kapazitiv an eine Victim-Leitung gekoppelt ist. Für maximale Kapazität umschließt die Aggressor-Leitung (Metallisierungslagen M1, M2 und M3 und Kontaktlöcher Via1 und Via2) die Victim-Leitung (M2) mit minimalen Abständen. Durch die fallende Flanke an AGG wird auch das Victim-Potential abgesenkt. Der Impuls erreicht ein nachgeschaltetes NAND-Gatter. Bei einer ausreichenden Dauer und Höhe des Eingangspulses wird der Puls invertiert an den Ausgang weitergegeben. Dabei kann dieser entweder größer oder kleiner werden. Eine Pulsvergrößerung tritt insbesondere dann auf, wenn zum einen der Eingangspuls hinreichend groß

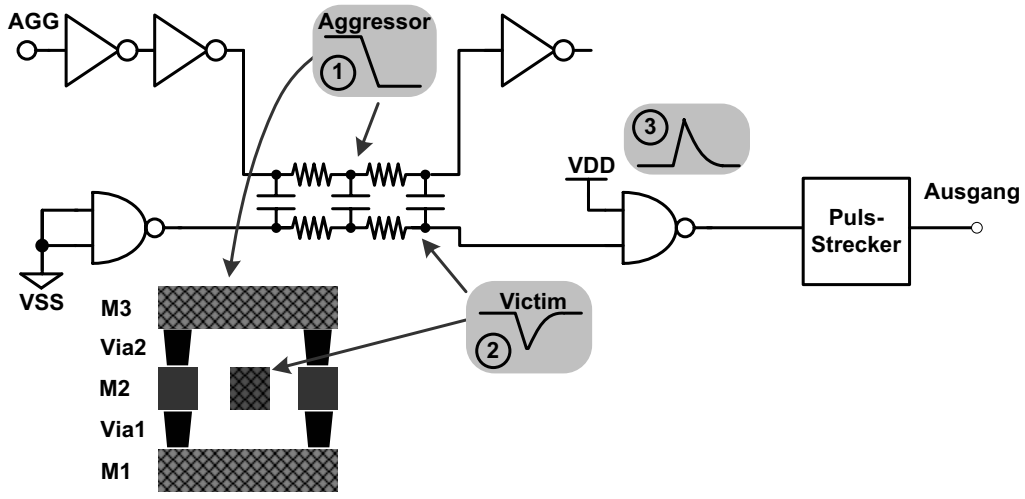


Abbildung 5.8: Testschaltung zur Crosstalk-Detektion. Der Querschnitt unten links zeigt die vom Aggressor umschlossene Victim-Leitung.

ist und das Gatter zum anderen asymmetrisch dimensioniert ist, wies es bei Skewed-CMOS-Gattern der Fall ist.

Ab einer bestimmten Höhe und Länge wird der Puls detektiert und in einen statischen Zustand umgewandelt. Dies ist vergleichbar mit dem Verhalten einer realen Schaltung, in der ein Flip-Flop am Ende des Datenpfades während der Zeitdauer eines Störimpulses ein falsches Datum übernimmt. Die Teststrukturen sind so gewählt, dass einzelne Effekte unabhängig voneinander untersucht werden können. Im Anhang A werden einige Messergebnisse dargestellt und analysiert.

Es zeigt sich, dass es in Schaltungen mit asymmetrischen Skewed-CMOS-Gattern unter ungünstigen Bedingungen zu einer Zustandsänderung durch Crosstalk und damit zu einer Fehlfunktion kommen kann. Wird die Victim-Leitung von allen Seiten von dem Aggressor umschlossen, so kommt es ab einer Leitungslänge von 120 μm bei ungünstiger Dimensionierung zu einem Crosstalk-Ereignis. Es ist jedoch möglich, durch geeignete Dimensionierung diese Bedingungen zu vermeiden. Diese Randbedingung wird im folgenden Abschnitt berücksichtigt.

5.1.4 Dimensionierung und Geschwindigkeit

In statischen CMOS-Schaltungen ist es ausreichend, die Transistoren so zu dimensionieren, dass die Summe der NMOS- und PMOS-Verzögerungszeiten, z.B. eines Inverterpaars, minimiert wird. Wie die Crosstalk-Messergebnisse zeigen, ist es nahezu unmöglich, hier ein fehlerhaftes Datum durch Übersprechen zu generieren. In einer Skewed-CMOS-Schaltung müssen hingegen die Evaluationstransistoren, die Halte-Transistoren, die Takttransistoren sowie die Precharge-Transistoren nach unterschiedlichen Kriterien dimensioniert werden. Zunächst werden hier Dimensionierungsregeln für Skewed-CMOS-Schaltungen allgemein aufgestellt. Im Folgenden kann dann auf die Besonderheiten von leckstromoptimierten Varianten eingegangen werden.

Die Transistorweiten des Evaluationspfades W_p^{eval} und W_n^{eval} werden so dimensioniert, dass die Summe der Verzögerungszeiten von N- und P-Typ-Gattern minimal werden. Vergleichbar mit der im Abschnitt 3.2 für statische Schaltungen gezeigten Rechnung ergibt sich auch hier in erster Näherung ein Weitenverhältnis $W_p^{eval}/W_n^{eval} = (I_n/I_p)^{1/2}$ (Gleichung 3.42). Lediglich die zusätzliche Serienschaltung von NMOS-Takttransistor und NMOS-Logiktransistoren im N-Typ-Gatter muss berücksichtigt werden.

Der NMOS-Takttransistor kann eine andere Weite als die NMOS-Logiktransistoren besitzen, da er bereits vor der Auswertung des Gatters geöffnet wird und daher nicht geschwindigkeitskritisch ist. Je größer W_n^{CLK} ist, desto schneller wird das Gatter. Allerdings erhöht sich bei Verwendung großer Takttransistoren die in Logiken mit einem Precharge-Vorgang ohnehin schon große Taktlast noch weiter. Die Weite des NMOS-Takttransistors wird basierend auf Schaltungssimulationen auf $W_n^{CLK} = 1.2 \cdot W_n^{eval}$ begrenzt, da der darüber hinausgehende Performance-Gewinn nur noch minimal ist.

Während des Precharge-Vorgangs werden zunächst alle Ausgänge der N-Typ-Gatter über die PMOS-Takttransistoren auf 1 aufgeladen. Daraufhin beginnt der Precharge-Vorgang der P-Typ-Gatter. Der PMOS-Takttransistor und der NMOS-Precharge-Pfad des P-Typ-Gatters werden daher als Precharge-Pfad bezeichnet. Diese beiden Schaltvorgänge müssen während einer halben Taktperiode abgeschlossen sein. Die für den Precharge-Vorgang zur Verfügung stehende Zeit ergibt sich damit direkt aus der Taktfrequenz des Schaltungsblocks. Die Weite der Transistoren im Precharge-Pfad sollte in jedem Fall hinreichend groß gewählt werden, damit die Geschwindigkeit der Schaltung stets nur von der aktiven Evaluation und nicht vom Precharge-Vorgang abhängt. Insbesondere muss die Ausgangslast bei der Dimensionierung des Precharge-Pfades berücksichtigt werden. Im Falle eines sehr kurzen Datenpfades mit nur vier Gattern je Taktphase (Abb. 5.4) muss der Precharge-Vorgang fast halb so schnell wie der Evaluationsvorgang ablaufen.

Der NMOS-Precharge-Pfad des P-Typ-Gatters ist zugleich ein Haltepfad, der die Schaltung im aktiven Modus robuster gegen Leckströme und Crosstalk machen soll. Wie auch für den PMOS-Haltepfad des N-Typ-Gatters muss hier ein Kriterium festgelegt werden, wie die Transistoren zu dimensionieren sind.
Die Ergebnisse der Crosstalk-Messungen im Anhang A zeigen, dass der NMOS-Haltepfad nicht kritisch ist. Zudem ist dieser schon aufgrund des oben genannten Precharge-Kriteriums ausreichend groß dimensioniert. Der PMOS-Haltepfad ist jedoch anfällig gegen Crosstalk. Insbesondere vermindern Serienschaltungen die Stabilität (Abb. A.2a). Da nicht bekannt ist, wie groß die Ausgangslast des Gatters ist, kann im Allgemeinen nicht von einer Stabilisierung durch diese parasitären Lasten ausgegangen werden (Abb. A.8). Es ergibt sich z.B. eine Schwelle für einen sicheren Betrieb der Schaltung von 1.5 V, wenn zu einer Versorgungsspannung von 1.2 V ein Sicherheitsabstand von 0.3 V hinzu addiert wird. In Abbildung A.2b ist zu erkennen, dass auch für eine PMOS-Serienschaltung eine hinreichende Stabilität durch Verwendung von Transistoren mit größerer Weite erreicht werden kann.

Abbildung 5.9 zeigt die Leckströme und Schaltverzögerungen von repräsentativen Skewed-CMOS- und statischen CMOS-Schaltungen. Während die statischen CMOS-Schaltungen einer Trendlinie folgen, die mit der I_{on}-I_{off}-Trendlinie für Einzeltransistoren vergleichbar ist, werden mit Skewed-CMOS-Schaltungen Punkte links unterhalb der Trendlinie erreicht. Die Skewed-CMOS-Variante mit LVT-Transistoren in den Evaluationspfaden und REG-Transistoren in den Haltepfaden (LVT–REG) reduziert die Schaltverzögerung um 26 % im Vergleich zur statischen LVT-Schaltung. Der Leckstrom reduziert sich nur geringfügig, da die Schwellenspannungen von LVT und REG sehr ähnlich sind und große Gateleckströme fließen. Ähnliches gilt für die entsprechenden Dickoxid-Varianten (31 % niedrigere Verzögerungszeit).
Die größte Verbesserung kann mit der Leckstrom-optimierten Skewed-CMOS-Logik erzielt werden. Evaluations- und Haltepfade sind aus LVT-Transistoren aufgebaut. Die lokalen Standby-Transistoren sind LL-LVT-Transistoren. Im Vergleich zur LVT–REG-Skewed-CMOS-Schaltung wird der Leckstrom um etwa zweieinhalb Dekaden reduziert, während die Schaltgeschwindigkeit nur um 5 % abnimmt.

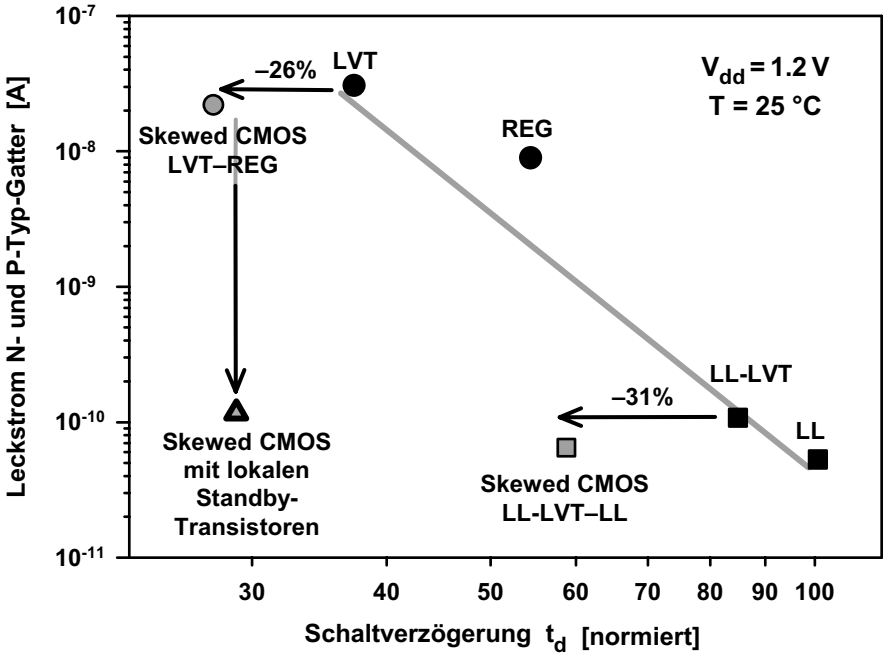


Abbildung 5.9: Simulierte Leckströme und Schaltverzögerungen von NAND2-Ketten mit Fan-out 3 in Skewed-CMOS- und statischen CMOS-Schaltungen.

Die simulierten Werte in Abbildung 5.9 beziehen sich auf einzelne Gatter. In einer realen Schaltung können jedoch verschiedene Effekte wieder zur einer Erhöhung des Leckstroms oder zu einer Reduzierung der Schaltgeschwindigkeit führen (z.B. Narrow-Width-Effekt, Well-Proximity-Effekt oder stärkere Auswirkungen von Parametervariationen). Im Kapitel 6 wird daher die Effizienz von Skewed-CMOS-Logik für reale Schaltungen anhand mehrerer 32-bit-Addierer demonstriert.

Zusammenfassend ist festzustellen, das sich die Skewed-CMOS-Logik im Gegensatz zur statischen CMOS-Logik nicht für den Einsatz in einer voll automatisierten Designumgebung eignet. Das monotone Schaltverhalten sowie die höhere Anzahl getakteter Speicherelemente und Logikgatter erfordert einen manuellen Schaltungsentwurf, der nur für besonders geschwindigkeitskritische Schaltungsblöcke vertretbar sein kann. Jedoch gerade in diesen Schaltungsteilen, in denen bevorzugt Transistoren mit niedrigen Schwellenspannungen und dünnen Oxiden eingesetzt werden, ist eine Reduzierung des Leckstroms im Standby-Modus sowie eine schnelle Rückkehr in den aktiven Betrieb wichtig. Die leckstromoptimierte Skewed-CMOS-Logik stellt hierzu eine effiziente Lösung bereit.

5.2 Statische Multi-V_t-Multi-t_{ox}-Logik

Im vorhergehenden Abschnitt wurde eine leckstromoptimierte Skewed-CMOS-Logik vorgestellt, in der die einzelnen NMOS- und PMOS-Pfade getrennt auf hohe Schaltgeschwindigkeit und geringen Leckstrom optimiert sind. Ein Nachteil von Skewed-CMOS-Logik ist jedoch das monotone Schaltverhalten dieser Logik-Familie. In [34] wird eine Schaltungstechnik vorgestellt, die zwar weniger Geschwindigkeits- und Leckstromvorteile als Skewed-CMOS-Logik bringt, dafür aber invertierend und damit frei kombinierbar ist.

Die grundsätzliche Idee ist, die Logik-Gatter im Hinblick auf die Schaltgeschwindigkeit weiterhin symmetrisch zu dimensionieren. Bezüglich des Leckstroms besitzen die Logikgatter jedoch einen bevorzugten leckstromarmen Zustand. In [34] wird dieses durch die Verwendung unterschiedlicher V_t erreicht (Multi-V_t-Logik). In Abbildung 5.10a ist eine einfache Inverterkette dargestellt. Die Gatter besitzen wie in der leckstromoptimierten Skewed-CMOS-Logik abwechselnd leckstromarme NMOS- und PMOS-Pfade, die im Standby-Modus den Verluststrom bestimmen. Die Weiten der Transistoren werden so angepasst, dass das symmetrische Schaltverhalten erhalten bleibt. Ein Transistor mit hoher Schwellenspannung (HVT) ist dementsprechend weiter als ein LVT-Transistor. In Abbildung 5.10a sind relative Weiten für die unterschiedlichen Transistortypen angegeben.

Um die Effizienz dieser Schaltungstechnik experimentell zu verifizieren, werden Teststrukturen auf einem 130 nm-Testchip integriert und gemessen (Anhang B). Abbildung 5.11 zeigt die Verbesserung von Leckstrom und Schaltgeschwindigkeit. Die klassischen CMOS-Ringoszillatoren mit unterschiedlichen Schwellenspannungen (HVT, REG und LVT) zeigen in diesem Leckstrom-Geschwindigkeits-Diagramm ein ähnliches Verhalten wie die Einzeltransistoren im I_{on}-I_{off}-Diagramm (Abb. 2.1. Davon abweichend kann mit Multi-V_t-Logik bei Aktivierung des leckstromarmen Zustands ($SEL = 0$) ein besserer Trade-off erreicht werden.

In der Abbildung wird zudem deutlich, dass die Schaltgeschwindigkeit bei konstant niedrigem Leckstrom gesteigert werden kann, indem sub-nominale Gatelängen eingesetzt werden. Die Kombination unterschiedlicher Gatelängen wird in statischen CMOS-Schaltungen bereits angewendet [28].

In einer komplexeren Schaltung wird zunächst eine beliebige Kombination von Eingangssignalen als Minimum-Leakage-Vektor definiert. Im einfachsten Fall werden dazu alle Ausgänge der Flip-Flops, die am Eingang des nicht benötigten Blocks liegen, auf 0 gesetzt. Wie in Abbildung 3.11 gezeigt, ist eine darüber hinausgehende Wahl eines bestimmten Eingangsvektors nicht effizient. Der logische Ausgangszustand eines jeden Gatters beim Anliegen des MLV bestimmt dann die Anordnung der unterschiedlichen Schwellenspannungen (Abb. 5.10b).

Eine darüber hinausgehende Optimierung der Gatter ist möglich, indem z.B. in einem NAND-Gatter mit Ausgang 1 nur einer der Serien-NMOS-Transistoren mit hohem V_t ausgelegt wird (Abb. 5.10c). In einem realen Layout entsteht jedoch das Problem, dass der Abstand von Transistoren unterschiedlicher Schwellenspannung mehr als doppelt so groß sein muss wie der Abstand bei gleicher V_t-Implantation. Außerdem würde jede Kombination von Eingangssignalen

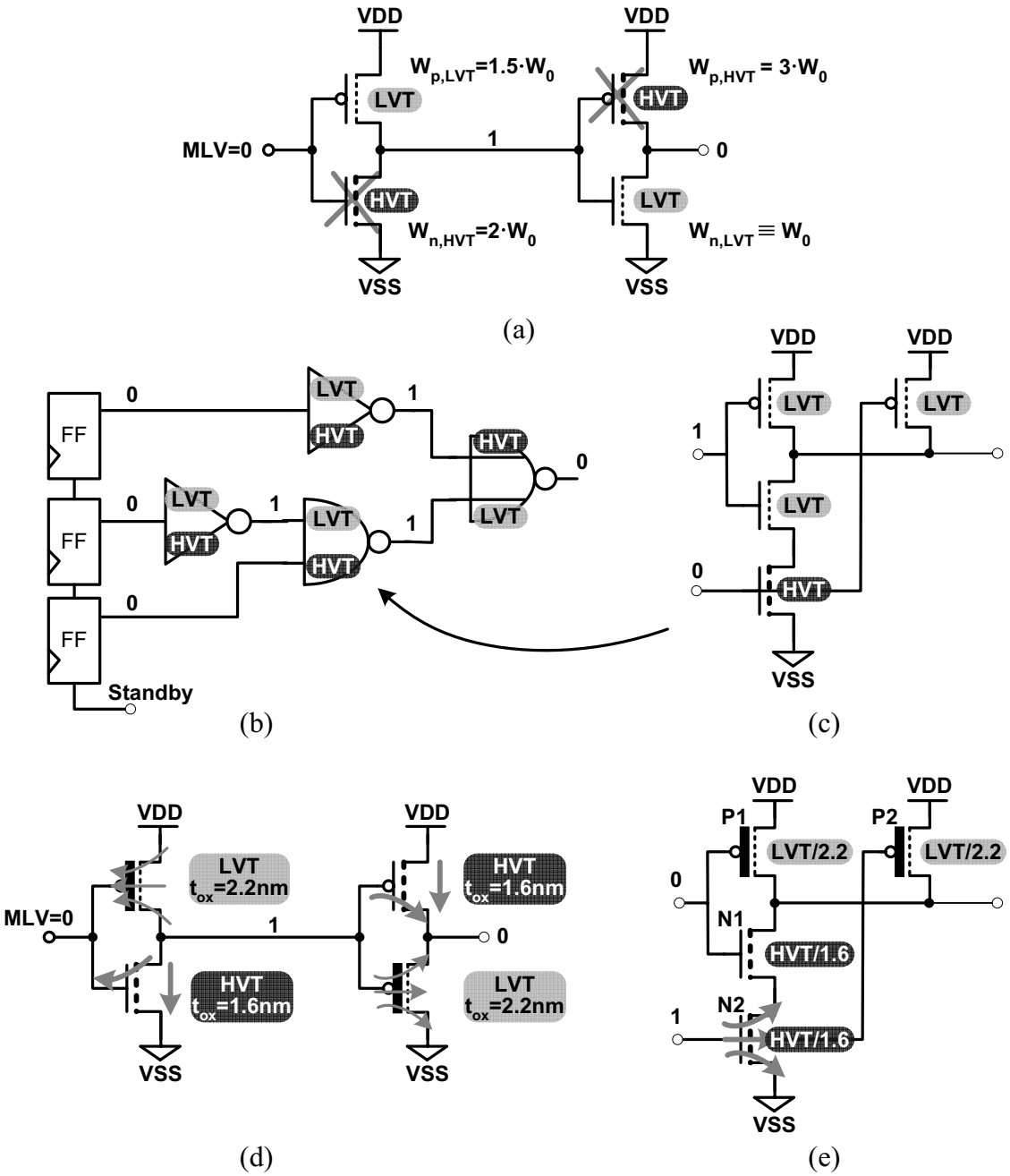


Abbildung 5.10: Einfache Inverterkette (a), komplexere Logikschaltung (b) und ein NAND-Gatter (c) in Multi-V_t-CMOS-Logik sowie Multi-V_t-Multi-t_{ox}-Logik (d). In (e) ist ein NAND-Gatter in Multi-V_t-Multi-t_{ox}-Logik dargestellt, bei dem ein hoher Gateleckstrom auftritt, der durch Vertauschen der Eingänge jedoch vermieden werden kann.

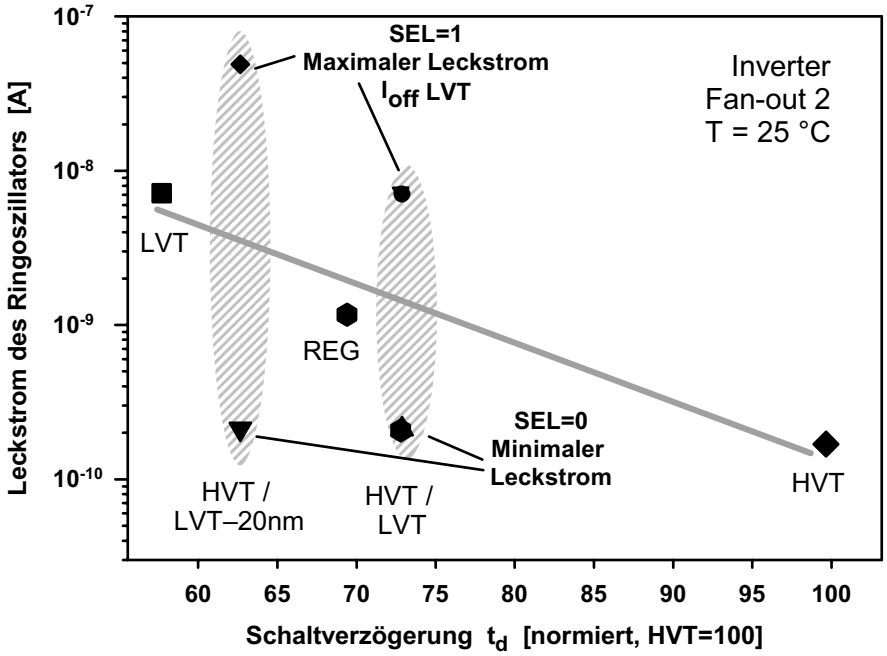


Abbildung 5.11: Gemessene Schaltgeschwindigkeiten und Leckströme von Ringoszillatoren in klassischer CMOS-Logik (HVT, REG,LVT) und Multi-V_t-Logik in der 130 nm-Technologie. Wenn bei der Multi-V_t-Logik der leckstromarme Zustand aktiviert wird ($SEL = 0$), kann ein besserer Trade-off erreicht werden.

eine eigene Zelle mit individuell angepassten Transistoren erfordern. Die Anzahl der Zellen in einer Standardzellen-Bibliothek würde sich erheblich vergrößern.

Der Vorteil dieser Technik ist, dass der leckstromarme Zustand schnell hergestellt werden kann und die Schaltung danach auch sofort wieder betriebsbereit ist. Der zusätzliche Flächenbedarf ist begrenzt auf die MLV-Erweiterung in den Flip-Flops sowie auf einen eventuell größeren Abstand, der zwischen Transistoren unterschiedlicher Schwellenspannung oder Oxiddicke eingehalten werden muss.

Wie in der Skewed-CMOS-Logik reduzieren auch hier hohe Gateleckströme, die in der 90 nm-Technologie auftreten, die Effizienz dieser Schaltungstechnik, da nur die Unterschwellenströme reduziert werden. In Abbildung 5.10d ist daher im Unterschied zu [34] ein Konzept zur Reduzierung des Leckstroms in einer CMOS-Schaltung dargestellt, das auch den Gateleckstrom reduziert. Dabei haben die im Standby-Modus ausgeschalteten Transistoren weiterhin ein hohes V_t. Da hier jedoch nur im Überlappungsbereich zwischen Source und Gate ein Tunnelstrom fließen kann, wird ein dünnes Oxid eingesetzt. Der komplementäre Pfad besitzt eine niedrige Schwellenspannung, aber ein dickeres Oxid, sodass der On-Gatestrom unterdrückt werden kann (Statische Multi-V_t-Multi-t_{ox}-Logik). Da der On-Gatestrom üblicherweise 5 bis 10 mal größer als der Off-Gatestrom ist, kann der Gateleckstrom der Gesamtschaltung um bis zu 90 % reduziert werden. Auch hier kann die Gatelänge der HVT-Transistoren größer als die der LVT-Transistoren gewählt werden. Der Trade-off mit der Summe aller Leckströme in einer Multi-

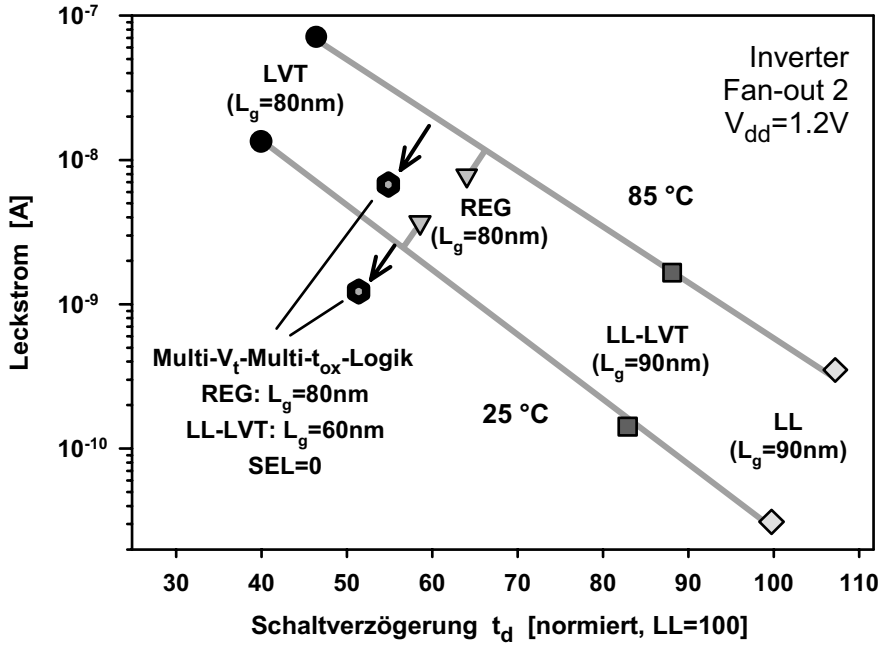


Abbildung 5.12: Leckstrom-Performance-Tradeoff für Multi-V_t-Multi-t_{ox}-Logik in der 90 nm-Technologie (simulierte Daten). Wie in der 130 nm-Technologie liegen die klassischen CMOS-Schaltungen (LVT, REG, LL-LVT und LL) auf der Trendlinie. Die Multi-V_t-Multi-t_{ox}-Logik stellt bei beiden Temperaturen eine Verbesserung dar.

V_t-Multi-t_{ox}-Schaltung ist in Abbildung 5.12 dargestellt. Es wird eine deutliche Verbesserung erzielt.

Bei einer Serienschaltung, in der nur ein Transistor ausgeschaltet ist, ist darauf zu achten, dass kein zusätzlicher Gateleckstrom auftritt. Die Eingänge des in Abbildung 5.10e dargestellten NAND-Gatters müssen daher getauscht werden, sodass der 0-Eingang im MLV-Modus unterhalb des 1-Eingangs liegt.
Auch hier ist es möglich, jedes Gatter noch weiter zu optimieren. Anstelle einer Umordnung der Eingänge könnte im NAND-Gatter in Abbildung 5.10e der Transistor N2 ein dickes Oxid und ein niedriges V_t erhalten und der Transistor P2 ein dünnes Oxid. Insgesamt würden in einer Schaltung dann aber acht verschiedene Transistortypen verwendet (jeweils NMOS/PMOS, HVT/LVT, dickes/dünnes Oxid). Die Anzahl der Standardzellen in einer Bibliothek sowie deren Flächenbedarf würde weiter zunehmen. Schon in statischen CMOS-Schaltungen werden häufig unterschiedliche Standardzellen entworfen z.B. für unterschiedliche Schwellenspannungen, flächen- bzw. geschwindigkeitsoptimierte Zellen sowie für verschiedene Versorgungsspannungen optimierte Zellen.
In Abbildung 5.10a sind die relativen Transistorweiten für ein geschwindigkeitsoptimiertes Design angegeben. Im Gegensatz zu den im Abschnitt 3.2 beschriebenen Skalierungsregeln ist es hier nicht ausreichend, die Verzögerung eines Inverterpaars zu minimieren (Gleichung 3.43). Dort erfolgte eine Signalpropagation stets abwechselnd über die verschiedenen Transistortypen N- und PFET. Hier kann es hingegen vorkommen, dass eine Folge von Schaltvorgän-

gen zum Beispiel nur über HVT-Transistoren erfolgt. Die Verzögerungszeit für Multi-V_t-Multi-t_{ox}-Schaltungen oder Multi-V_t-Schaltungen wird daher nur unter der Bedingung

$$t_{np} = \max[t_n^{HVT} + t_p^{HVT}\,,\, t_n^{LVT} + t_p^{LVT}] \tag{5.1}$$

minimal. Die auf dem Testchip integrierten Inverterketten (Abb. 5.11) wurden mit einem Optimierungstool in Cadence für diese Bedingung optimiert. Die daraus resultierenden Weitenverhältnisse entsprechen etwa denen in Abbildung 5.10a. Das Weitenverhältnis von N- und PFET eines Transistortyps entspricht dem Optimum aus β_W^{opt} aus Gleichung 3.42. Im Gegensatz dazu ist z.B. für die beiden NFETs unterschiedlichen Transistortyps ein symmetrisches Weitenverhältnis zu wählen ($W_{LVT}/W_{HVT} = I_{HVT}/I_{LVT}$). Die exakten Dimensionierungen sind im Anhang in Abbildung B.2 angegeben.

Aus Gleichung 5.1 ist ersichtlich, dass die gesamte Verzögerungszeit von Multi-V_t-Multi-t_{ox}-Schaltungen das Maximum einer HVT- und einer LVT-Verzögerungszeit ist. Dies wirkt sich negativ auf die Anfälligkeit gegenüber Parameterschwankungen aus. Fallen auf einem Chip z.B. die LVT-Transistoren prozessbedingt langsam aus, so kann die daraus resultierende Verlangsamung der Schaltung nicht durch schnellere HVT-Transistoren ausgeglichen werden.

Darüber hinaus lässt sich die Technik mit bekannten Multi-V_t-Techniken auf Gatter-Ebene kombinieren, indem nur geschwindigkeitskritische Pfade mit Multi-V_t-Multi-t_{ox}-Logik-Gattern ausgestattet werden, während alle anderen Pfade leckstromarme Transistoren mit höheren Schwellenspannungen und größeren Gatelängen verwenden.

Insgesamt stellen statische Multi-V_t-Multi-t_{ox}-Schaltungen eine gute Möglichkeit dar, den Trade-off zwischen Leckstrom und Schaltgeschwindigkeit zu verbessern. Zwar sind die Verbesserungen nicht so hoch wie bei der leckstromoptimierten Skewed-CMOS-Logik, jedoch kann jetzt eine beliebige logische Funktion ohne Einschränkung realisiert werden. Lediglich die Aktivierung des Standby-Zustands mit Hilfe spezieller Flip-Flops erfordert einen zusätzlichen schaltungstechnischen Aufwand.

5.3 Flip-Flops und Latches

Flip-Flops sind elementare Bestandteile jeder getakteten CMOS-Digitalschaltung [69]. Sie dienen der Unterteilung des Datenpfades, wobei z.B. in statischer CMOS-Technik die Logikgatter zwischen zwei Flip-Flops in einer Taktperiode ausgewertet werden (Abb. 3.10).

Damit die in diesem Kapitel vorgestellten Schaltungstechniken effizient zur Reduzierung des Leckstroms bei hoher Schaltgeschwindigkeit eingesetzt werden können, sind spezielle Flip-Flops erforderlich oder aber es können besonders einfache Latches verwendet werden. So werden in Skewed-CMOS-Schaltungen Flip-Flops mit differentiellen Ausgangssignalen benötigt, oder es ist erforderlich, dass ein Flip-Flop auch im Standby-Modus seinen Zustand erhält.

Zudem machen Flip-Flops häufig einen sehr großen Anteil an der Gesamtschaltung aus. Sie können bis zu 50 % der Fläche in Anspruch nehmen. Da Flip-Flops stets direkt getaktet sind,

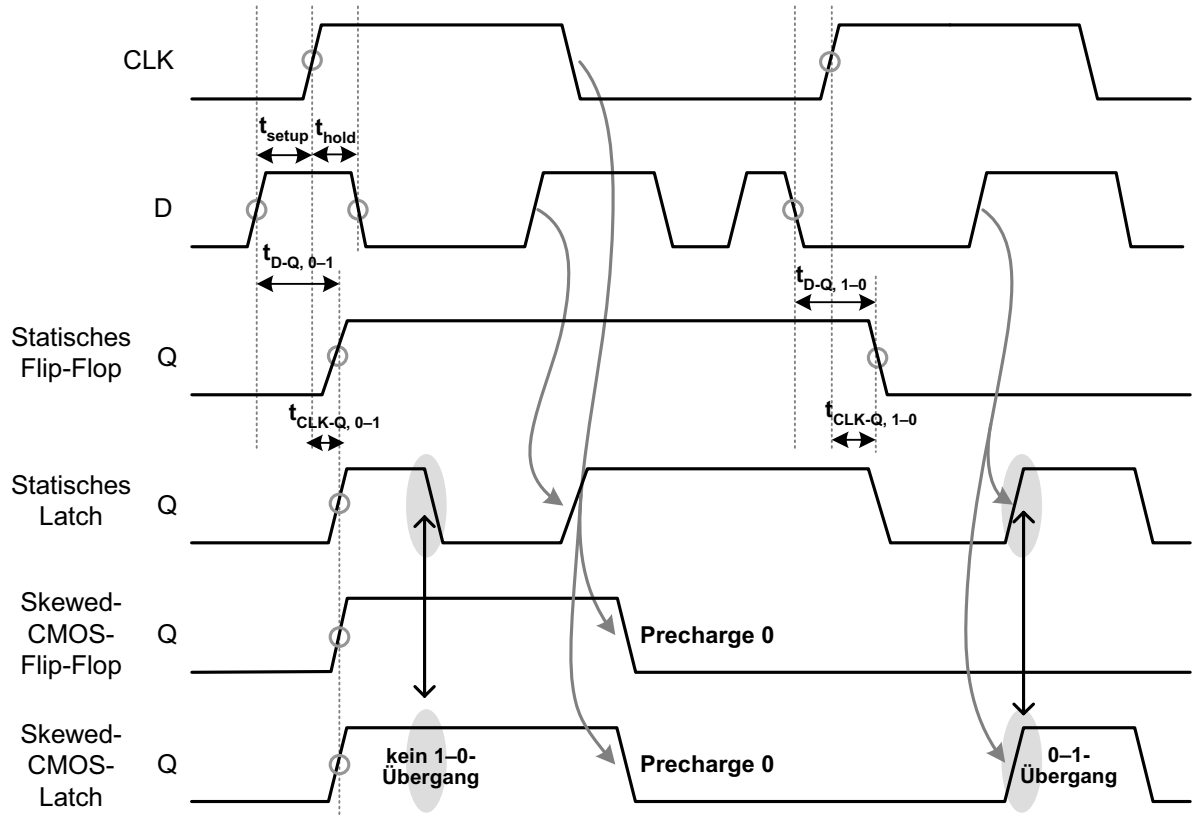


Abbildung 5.13: Timing-Diagramme für statische und Skewed-CMOS-Flip-Flops und Latches.

weisen sie zudem eine hohe Aktivität auf. Aus diesem Grund ist eine gleichzeitige Optimierung von aktiver Leistungsaufnahme und Schaltgeschwindigkeit wichtig.

Ein Flip-Flop – auch flankengesteuertes Flip-Flop – übernimmt das am Eingang D anliegende Datum mit der steigenden Flanke des Taktsignals und stellt dieses Datum bis zur nächsten steigenden CLK-Flanke am Ausgang Q bereit. Daneben gibt es auch Varianten, die mit der fallende Taktflanke schalten. Dieses kann durch einfache Inversion des Taktsignals erfolgen und wird daher nicht weiter diskutiert.

Im zeitlichen Ablauf der Eingangssignale müssen Setup- und Hold-Zeiten eingehalten werden. Diese Zeiten geben an, wie lange D vor und nach der steigenden Taktflanke anliegen muss, um eine sichere Funktion zu gewährleisten (Abb. 5.13).

Normalerweise umschließt das durch Setup- und Hold-Zeit definierte Zeitfenster die CLK-Flanke (positive Setup- und Hold-Zeiten), kann aber auch hinter (negative Setup-Zeit) oder vor der Flanke liegen (negative Hold-Zeit). Prinzipiell hat die Lage des Setup-Hold-Fensters keinen Einfluss auf die gesamte Verzögerungszeit eines Datenpfades. Werden jedoch Flip-Flops mit unterschiedlichen Setup-Zeiten kombiniert, erhöht sich die Gefahr von *Race Conditions*. Diese können auftreten, wenn der Ausgang eines Flip-Flops direkt mit dem Eingang des nächsten Flip-Flops verbunden ist. Das bereits vorliegende Ausgangsdatum des ersten Flip-Flops wird einen Taktzyklus zu früh durch das zweite Flip-Flop übernommen, wenn eine direkte Verbin-

dung zwischen zwei aufeinander folgenden Flip-Flops besteht und wenn das zweite Flip-Flop eine negative Setup-Zeit besitzt. Kurze Schaltzeiten sowie ungleichmäßige Taktsignalversorgung (*Clock Skew*) erhöhen die Gefahr.

Im Gegensatz zu flankengesteuerten Flip-Flops sind Latches während der gesamten CLK=1-Phase transparent (Abb. 5.13). Sie werden daher auch als zustandsgesteuerte Flip-Flops bezeichnet. Wird ein Latch mit einem Takt-Puls anstelle eines normalen Taktsignals mit 50 % Tastverhältnis betrieben, so entsteht wiederum ein Flip-Flop. Diese so genannten pulsgesteuerten Flip-Flops sind jedoch in einer SoC-Umgebung weniger robust.

Die Datenausgänge von Flip-Flops und Latches in Skewed-CMOS-Schaltungen müssen während der Precharge-Phase auf 0 gesetzt werden (Abb. 5.13). Wie in der statischen Variante übernimmt dabei ein Skewed-CMOS-Flip-Flop das Datum mit der steigenden Taktflanke und hält den Ausgang dann bis zum Ende der aktiven Phase konstant. Ein Skewed-CMOS-Latch ist zwar auch wie ein statisches Latch während $CLK = 1$ transparent, darf aber nur einen einzigen 0–1-Übergang übernehmen, da die Eingänge von Skewed-CMOS-Schaltungen nur einmal in jeder aktiven Phase schalten können (Abb. 5.13).

Die wichtigste Größe für die Beurteilung der Flip-Flop-Geschwindigkeit ist das D-Q-Delay t_{D-Q}. Dieses setzt sich aus der Setup-Zeit t_{setup} und dem CLK-Q-Delay t_{CLK-Q} zusammen. Die Schaltzeiten des Flip-Flops unterscheiden sich allgemein für verschiedene Schaltrichtungen an Ein- und Ausgang. Folglich müssen beide Schaltrichtungen oder deren Maximum betrachtet werden.

Größtenteils werden in statischen CMOS-Schaltungen Master-Slave-Flip-Flops (MS-FF, Abb. 5.14) eingesetzt [87]. Diese zeichnen sich durch ein sehr robustes Verhalten aus. Allerdings sind die Schaltgeschwindigkeiten niedrig und die aktive Leistungsaufnahme hoch. Mit abnehmender Anzahl der Logikgatter zwischen zwei Flip-Flops erhöht sich der Anteil von Flip-Flops an der Gesamtschaltung. Zwischen acht Logikstufen in Mikroprozessoren [30] und etwa 40 Stufen in ASIC-Designs [6] sind üblich. Da mit der Verkürzung der logischen Tiefe der Anteil von Flip-Flops an der Gesamtschaltung steigt, stellt die Optimierung von Flip-Flops eine effiziente Methode zur Erhöhung der Geschwindigkeit sowie zur Reduzierung der aktiven und passiven Verlustleistung dar.

In [88] wird ein so genanntes Sense-Amplifier-Flip-Flop (SAFF, Abb. 5.15a) beschrieben, das in den Mikroprozessoren DEC Alpha 21264 und StrongARM 110 verwendet wird. Die obere Sense-Amplifier-Stufe muss im Betrieb zunächst über die PMOS-Takttransistoren P3 und P4 aufgeladen werden. Mit der steigenden Taktflanke wird eine Seite über D oder $\overline{D}$ entladen. $\overline{D}$ kann durch lokale Inversion erzeugt werden. Für $D = 1$ wird das inverse Set-Signal $\overline{S} = 0$ und die rückgekoppelten NAND-Gatter ändern ihren Zustand, wenn zuvor $Q = 0$ war. In der verbleibenden Zeit des Taktzyklusses muss im Sense-Amplifier zum einen gewährleistet sein, dass $\overline{R}$ nicht ebenfalls entladen wird, wenn am Eingang D ein Zustandswechsel auftritt und $\overline{D} = 1$ wird. Dieses wird durch den gesperrten Transistor N2 verhindert. Zum anderen muss eine leitende Verbindung von $\overline{S}$ zu V_{ss} erhalten bleiben, wenn N3 sperrt. Dieses gewährleistet der Quertransistor N6. Im Zeitpunkt der Auswertung ist die Verbindung über N6 hochohmig, da V_{gs} am Beginn der aktiven Phase klein ist. Zudem wird N6 minimal dimensioniert.

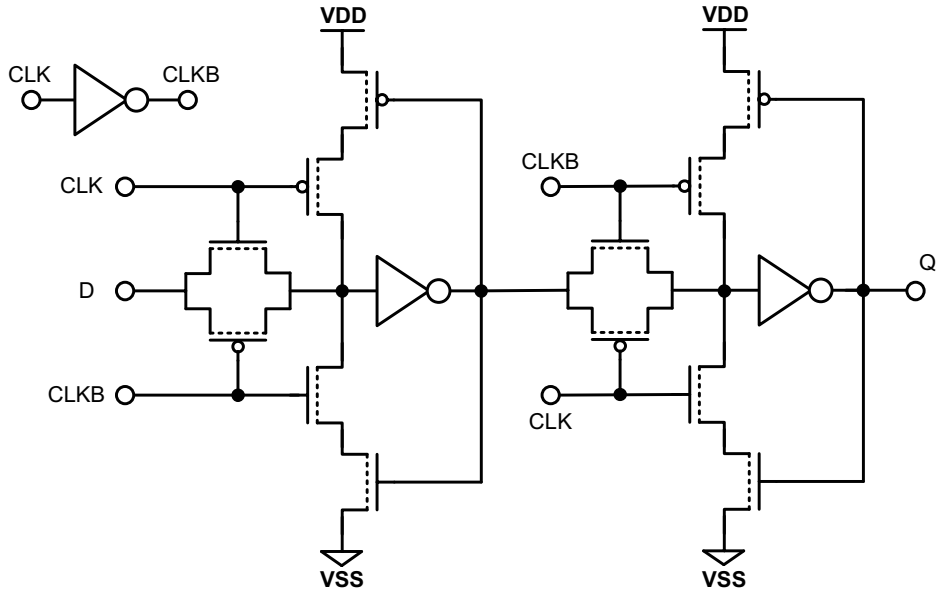


Abbildung 5.14: Tranmissiongate-Master-Slave-Flip-Flop [87]. Dargestellt ist ein MS-FF mit minimaler Anzahl an Transistoren. In einer Standardzellen-Ausführung werden alle Ein- und Ausgänge zusätzlich durch Inverter entkoppelt. Außerdem müssen häufig noch zusätzliche Funktionen wie Set, Reset, Enable, Scan oder differentielle Ausgänge implementiert werden.

In der Ausgangsstufe werden zunächst beide NAND-Gatter durchlaufen, bis das Datum an beiden Ausgängen Q und $\overline{Q}$ anliegt. Diese Verzögerungszeit wird durch die veränderte Ausgangsstufe in Abbildung 5.15b reduziert (modifiziertes SAFF [89]). Durch Hinzufügen zweier Inverter sowie der Transistoren P9, P10, N11 und N12 werden die NAND-Gatter zu rückgekoppelten Invertern (N7/N8, P7/P8, vergleichbar mit einer SRAM-Zelle), deren Rückkopplung unterbrochen werden kann (N9/N10, P9/P10).
Durch geeignete Dimensionierung kann ein schnelles, gleichzeitiges Schalten der Ausgänge erreicht werden. Zwar beginnt der Entladevorgang über N12 um eine Inverter-Verzögerungszeit später als der Aufladevorgang über P11. Wegen der höheren Schaltgeschwindigkeit des NMOS-Transistors holt der Entladevorgang jedoch auf und die Ausgänge schalten gleichzeitig. Das modifizierte SAFF wird zum Beispiel in [90] eingesetzt. In einer anderen Ausführung [91] entfallen die Transistoren zur Auftrennung der Rückkopplung P/N 9/10. In diesem Fall muss durch geeignete Dimensionierung wie in einer SRAM-Zelle darauf geachtet werden, dass auch bei Parameter-Schwankungen die rückgekoppelten Inverter sicher umgeschaltet werden können.

5.3.1 Optimiertes Sense-Amplifier-Flip-Flop

Bei der Optimierung von Flip-Flops können verschiedene Ziele definiert werden. Die wichtigsten Kriterien sind die aktive Verlustleistung und die Verzögerungszeit. Daneben müssen aber auch stets Setup- und Hold-Zeiten, Flächenbedarf, Leckströme, Störanfälligkeit und Möglichkeiten zur Implementierung von zusätzlicher Funktionalität, wie Set und Reset, Standby oder Scan-Möglichkeit, berücksichtigt werden.

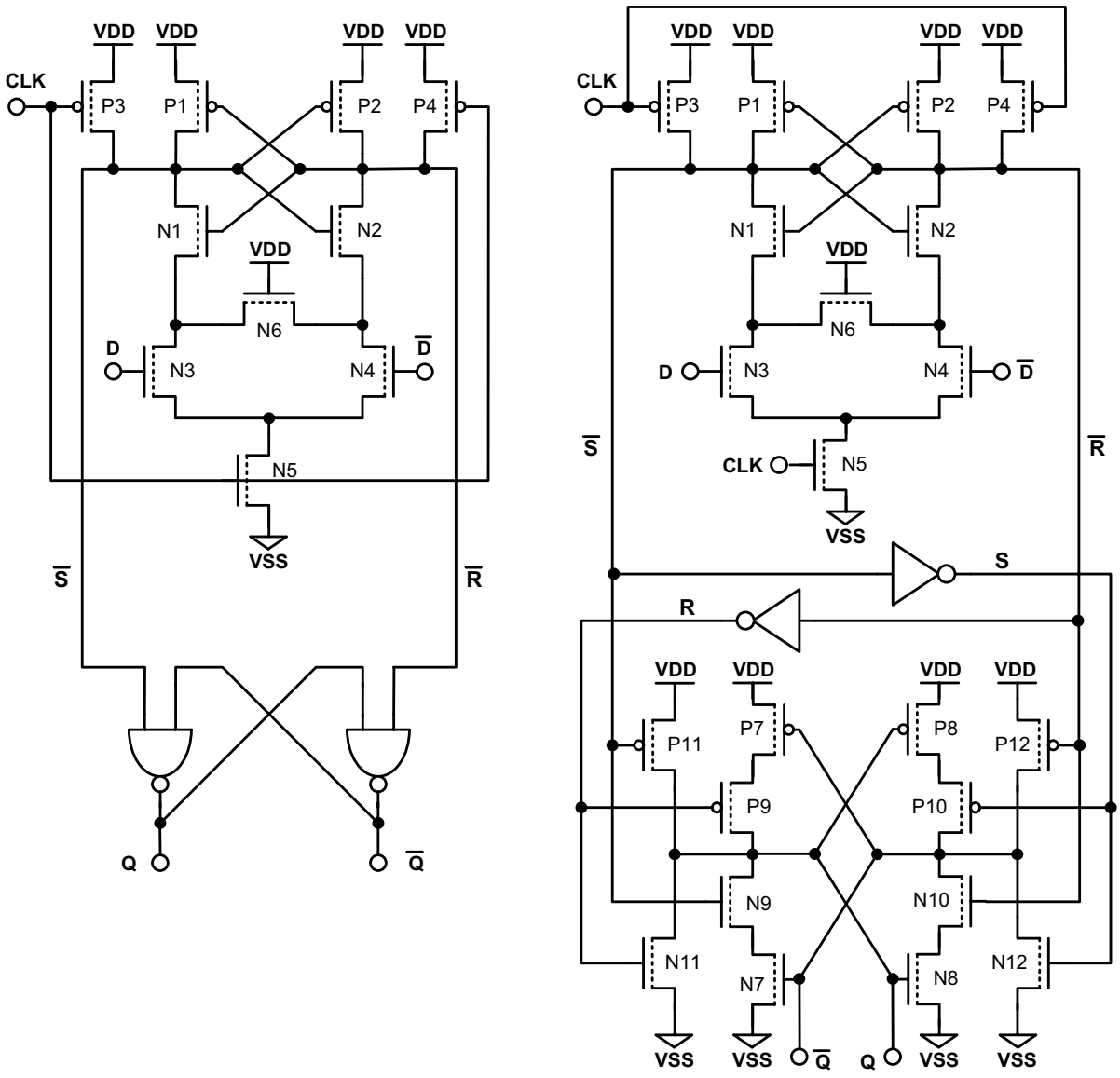


Abbildung 5.15: Sense-Amplifier-Flip-Flop (a) und modifiziertes Sense-Amplifier-Flip-Flop (b).

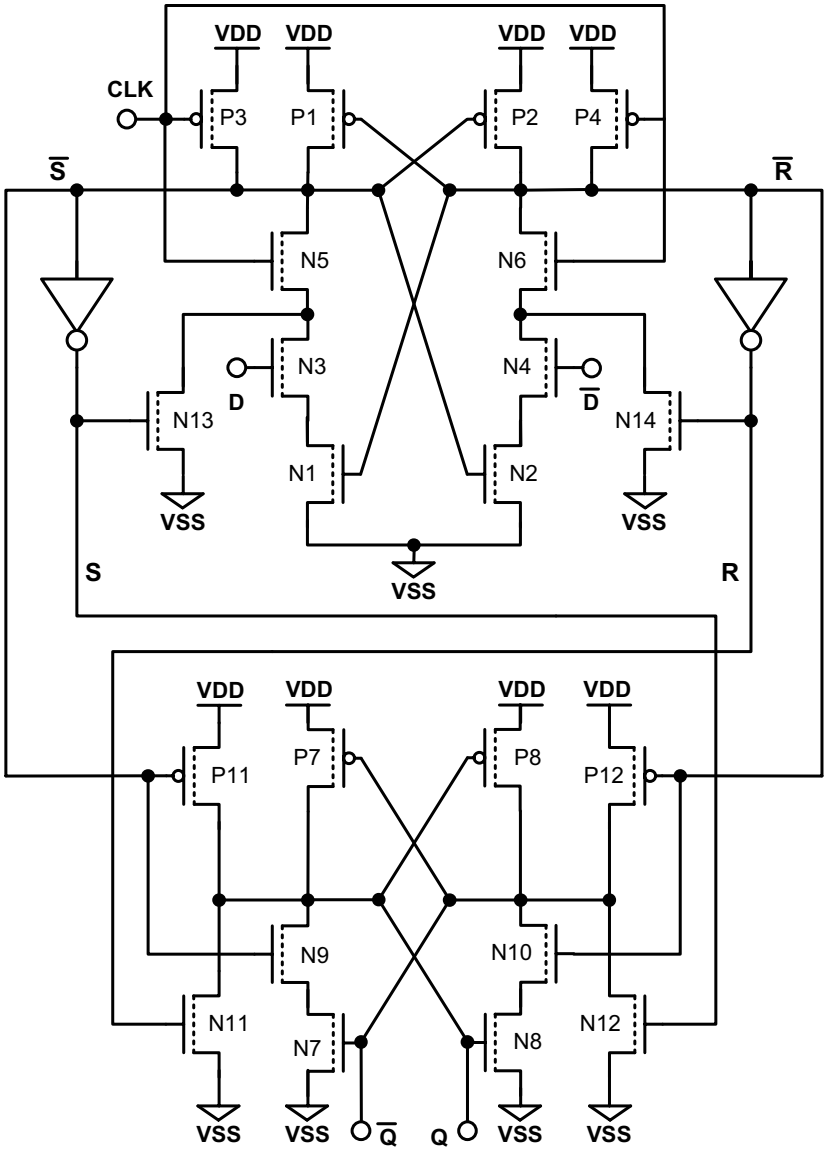


Abbildung 5.16: Optimiertes Sense-Amplifier-Flip-Flop.

Das in Abbildung 5.16 dargestellte Flip-Flop weist im Vergleich zum Flip-Flop in Abbildung 5.15a eine kleinere Verzögerungszeit sowie eine geringere aktive Leistungsaufnahme auf. Dies wird durch Modifikation sowohl der Eingangs- als auch der Ausgangsstufe erreicht. In der Sense-Amplifier-Eingangsstufe sind die Transistoren mit den zeitkritischen Eingangssignalen D, $\overline{D}$ und CLK im NMOS-Stack nach oben verschoben. Zum Beispiel wird für $D = 1$ die NMOS-Serienschaltung der linken Seite N1/N3 schon bis zum Source-Kontakt des CLK-Transistors N5 auf das V_{ss}-Potential entladen. Die Ladung, die bei Eintreffen der ansteigenden Taktflanke durch N5 fließen muss, wird dadurch deutlich reduziert. Der Schaltvorgang folgt damit am Anfang der Trajektorie des ungestackten Transistors (vgl. Abb. 3.29) und geht erst im Laufe des Schaltvorgangs auf den dreifach gestackten Fall über.

Wie in dem Sense-Amplifier-Flip-Flop in Abbildung 5.15 verhindert der Transistor N2 eine Entladung des Knotens $\overline{R}$ für den Fall eines Zustandswechsels $D = 1 \Rightarrow 0$. Allerdings wird ein zweiter Takttransistor N6 benötigt. Da sich die Takttransistoren jetzt an oberster Stelle der Serienschaltung befinden, können diese jeweils kleiner dimensioniert werden als bisher, ohne die Schaltgeschwindigkeit signifikant zu erhöhen.

Ein Quertransistor N6 wie in Abbildung 5.15 kann in dem optimierten SAFF nicht mehr eingesetzt werden, um eine leitende Verbindung zwischen $\overline{S}$ und V_{ss} aufrecht zu erhalten. Stattdessen überbrückt N13 den Transistor N3. Da ohnehin die invertierten Signale S und R zur Verfügung stehen und N13/N14 als Halte-Transistoren klein ausgelegt werden können, ist der zusätzliche Energie- und Flächenbedarf gering.

Die Ausgangsstufe wird ebenfalls modifiziert. Die Transistoren P9 und P10 können entfallen, da z.B. für $D = 1$ der PMOS-Transistor P11 den Knoten Q schon so weit aufgeladen hat, dass P8 bereits sperrt. Beim Einschalten des NMOS-Transistors N12 ist P8 schon weitgehend geschlossen, sodass die beiden Transistoren zu keinem Zeitpunkt gleichzeitig geöffnet sind und kein Querstrom von V_{dd} nach V_{ss} fließt.

Beide Modifikationen tragen dazu bei, die aktive Verlustleistung zu reduzieren. In der Ausgangsstufe wirkt sich die kleinere Anzahl an Transistoren positiv aus. Zudem wird eine PMOS-Serienschaltung vermieden, wodurch P7 und P8 kleiner ausgelegt werden können. Eine größere Einsparung wird im Sense-Amplifier erreicht, da die Anzahl der Transistoren, die in jedem Taktzyklus auf- und wieder entladen werden, kleiner geworden ist. Dies resultiert ebenfalls aus der geänderten Transistorfolge in der NMOS-Serienschaltung.

Die Taktlast von SAFFs ist im Vergleich zu Master-Slave-Flip-Flops nicht höher. Obwohl MS-FFs nicht vor jeder Auswertung aufgeladen werden müssen, werden hier dennoch in jedem Zyklus zehn Takttransistoren geschaltet, um die Transmissiongates und die Rückkopplungen zu öffnen und zu schließen.

Leistungsaufnahme und Geschwindigkeit einer Schaltung können nicht unabhängig voneinander beurteilt werden. Werden zum Beispiel alle Transistoren minimal dimensioniert, ist zwar im Allgemeinen die Leistungsaufnahme klein, aber auch die Schaltverzögerung groß. Deshalb werden die beiden Größen E und τ in unterschiedlicher Gewichtung $E \cdot \tau^n$ miteinander multipliziert, um eine unabhängige Größe für den Vergleich verschiedener Schaltungen bei unterschiedlichen Spannungen zu erhalten [92, 93]. Während für den Vergleich bei variabler V_{dd} ein Wert von $n = 1.5 \ldots 2.0$ sinnvoll ist [93], führt hier die Verwendung des Energie-Delay-Produkts mit $n = 1$ zu sinnvollen Bewertungen

$$EDP = E \cdot \tau. \tag{5.2}$$

Im Gegensatz dazu liefert ein an der Universität Dortmund entwickelter Algorithmus [94] eine große Anzahl unterschiedlicher Designs, die eine pareto-optimale Menge von Transistorweiten in Bezug auf Leistungsaufnahme und Schaltgeschwindigkeit bilden. Ist eine Schaltung pareto-optimal dimensioniert, dann existiert sowohl keine Schaltung, die bei gleicher Leistungsaufnahme schneller ist, als auch keine Schaltung, die bei gleicher Verzögerungszeit eine niedrigere Leistungsaufnahme besitzt.

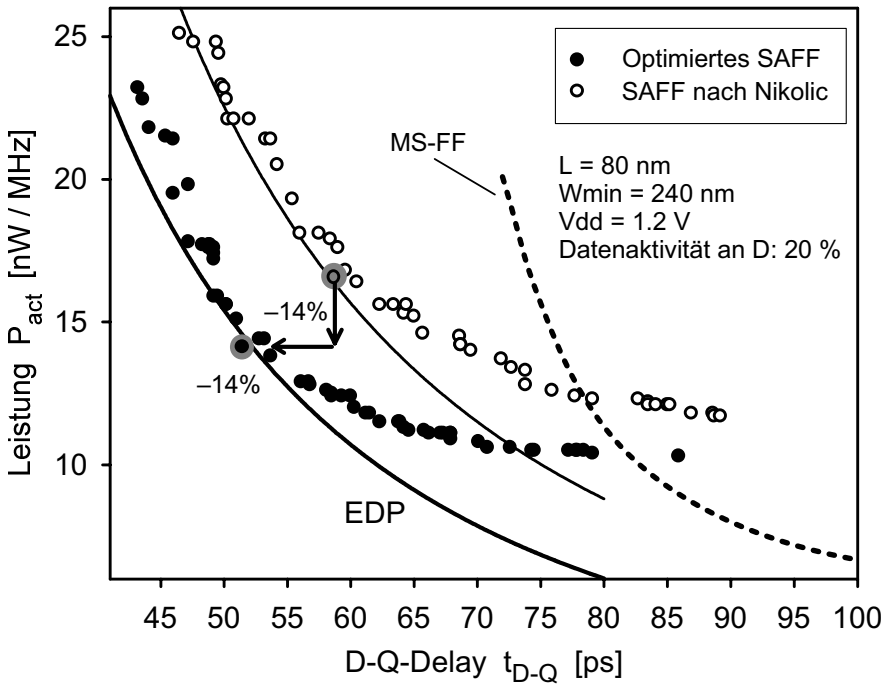


Abbildung 5.17: Ergebnisse der SAFF-Optimierung mit evolutionärem Algorithmus [95]. Für jedes Flip-Flop liefert der Algorithmus verschiedene Designs, indem die Transistorweiten optimiert werden. Für keines der Designs (repräsentiert durch einen Punkt im Energie-Delay-Diagramm) ist eine andere Lösung bekannt, die gleichzeitig sowohl sparsamer im Energieverbrauch als auch schneller ist (pareto-optimal).

Abbildung 5.17 vergleicht das optimierte SAFF mit dem Flip-Flop aus Abbildung 5.15b. Es wird deutlich, dass das optimierte Flip-Flop z.B. für jede vorgegebene Leistungsaufnahme schneller bzw. für jede Verzögerungszeit sparsamer ist. Zusätzlich ist die Trendlinie für optimierte Designs des Master-Slave-Flip-Flop (MS-FF) nach Abbildung 5.14 eingetragen. Dieses Flip-Flop erreicht eine geringere Leistungsaufnahme, allerdings nur bei langsamen Schaltgeschwindigkeiten von über 85 ps. Wichtig ist, den Vergleich unter realistischen Randbedingungen vorzunehmen. So ist die minimale Transistorweite mit 240 nm größer gewählt als prozesstechnisch erlaubt, um die Auswirkungen von Parametervariationen klein zu halten. Es wird eine Datenaktivität von 20 % am Eingang D angenommen. Die Aktivität beeinflusst das Ergebnis, da die Leistungsaufnahme eines Flip-Flops höher ist, wenn sich das Eingangs- und damit auch das Ausgangsdatum in jedem Taktzyklus ändert. Die Verzögerungszeit ist das minimale D-Q-Delay, dass sich bei optimaler Setup-Zeit t_{setup} einstellt (Abb. 5.18).

In Abbildung 5.17 ist eine Linie mit konstantem Energie-Delay-Produkt eingezeichnet. Der Vergleich zweier Designs mit gutem *EDP* zeigt für das optimierte Flip-Flop eine um 14 % höhere Schaltgeschwindigkeit bei einer um 14 % reduzierten Leistungsaufnahme. Das Energie-Delay-Produkt wird um 23 % vermindert. Die Transistorweiten sind im Anhang in Abbildung C.2 angegeben.

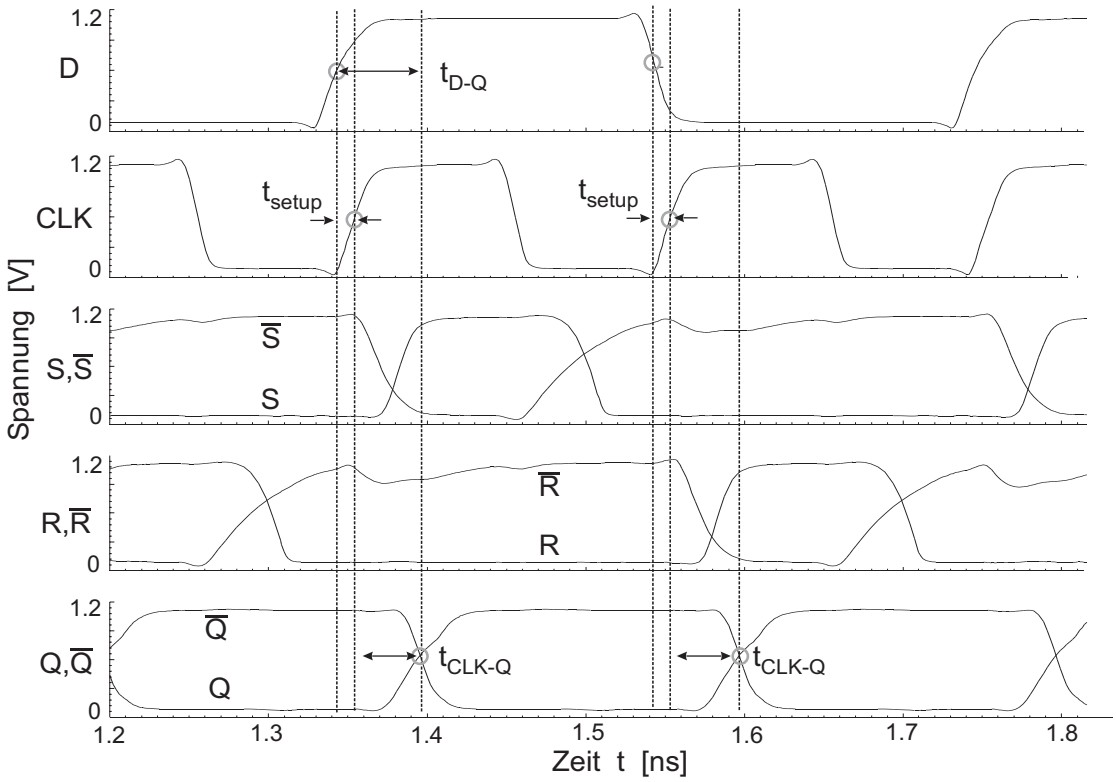


Abbildung 5.18: Timing-Diagramm eines Sense-Amplifier-Flip-Flops. Durch geeignete Transistordimensionierungen schalten die Ausgänge Q und $\overline{Q}$ genau gleichzeitig.

5.3.2 Flip-Flops mit Zustandserhaltung im Standby-Modus

Die Verlustleistung eines inaktiven Schaltungsblocks kann signifikant reduziert werden, wenn dieser durch Power-Switches von der Versorgungsspannung getrennt wird. Im Allgemeinen gehen dabei jedoch die in den Flip-Flops des betreffenden Schaltungsblocks gespeicherten Daten durch Leckströme verloren. Die Daten müssen daher aus den Flip-Flops ausgelesen werden und außerhalb des abzuschaltenden Blocks gespeichert werden. Nach dem Wiedereinschalten muss die Schaltung dann zunächst zurückgesetzt werden. Anschließend werden die Daten wieder zurück in die Flip-Flops geschrieben. Die Schaltung kann jetzt in demselben Zustand weiter arbeiten, der vor der Abschaltphase vorlag.

Mit Hilfe von Flip-Flops, die ihren Zustand auch im Standby-Modus nicht verlieren (*State Retention Flip-Flop*), lässt sich dagegen eine schnellere Rückkehr in den aktiven Betrieb erreichen. Die Schwierigkeit besteht darin, alle Leckstrompfade auszuschalten, insbesondere wenn auch Transistoren mit dünnen Oxiden eingesetzt werden. Gleichzeitig soll die Schaltgeschwindigkeit im aktiven Betrieb so wenig wie möglich beeinträchtigt werden.

Das State-Retention-Flip-Flop in [78] verwendet Transistoren mit unterschiedlichen Schwellenspannungen in einem Master-Slave-Flip-Flop, um einen leckstromarmen Zustand zu realisieren. Gateleckströme können jedoch nicht unterdrückt werden, ohne dickere Oxide im kritischen Pfad einzusetzen.

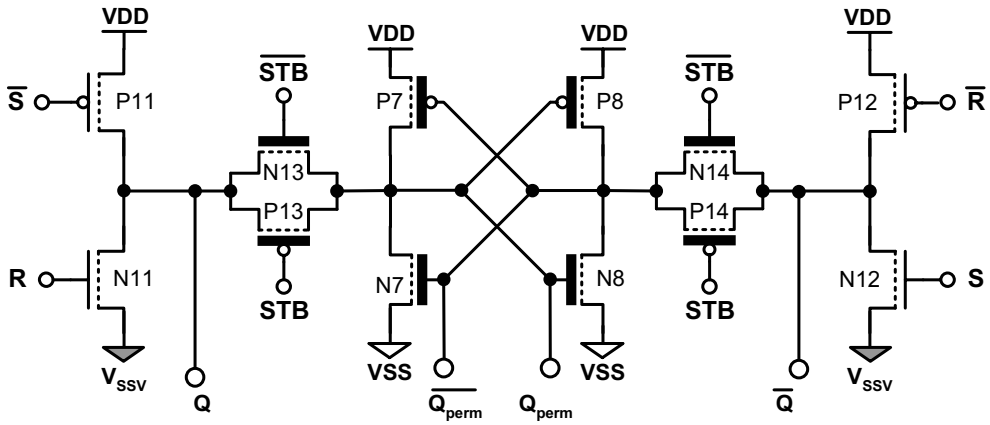


Abbildung 5.19: Ausgangsstufe eines Sense-Amplifier-Flip-Flops mit Zustandserhaltung. Die Ausgangsstufe kann mit einer der Eingangsstufen in den Abbildungen 5.15b oder 5.16 kombiniert werden.

Das so genannte Ballon-Flip-Flop [96] setzt ebenfalls eine Master-Slave-Architektur ein. Während der Zustand im vorhergehenden Flip-Flop in der zweiten Latch-Stufe gehalten wird, besitzt das Ballon-Flip-Flop einen zusätzlichen Speicher in Form rückgekoppelter Inverter. Diese werden im Standby-Modus durch leckstromarme Transistoren vollständig vom abgeschalteten Flip-Flop getrennt, sodass Gateleckströme nicht auftreten.

Nachteil dieser Ausführung ist, dass vier globale Kontroll-Signale erzeugt und zu jedem State-Retention-Flip-Flop geleitet werden müssen. Außerdem müssen diese Kontrollsignale noch mit dem Taktsignal synchronisiert werden, da der im leckstromarmen Inverterpaar gespeicherte Zustand nur für $CLK = 0$ zurück gelesen werden kann.

In [97] wird ein Sense-Amplifier-Flip-Flop mit einer Zustandserhaltung kombiniert. Das klassische SAFF aus Abbildung 5.15a wird dabei um eine Scan- und um eine Retention-Funktionalität erweitert. Die Erweiterung ist hier wie beim Ballon-Flip-Flop ein zusätzlicher, eigenständiger Zustandsspeicher. Auch hier können Gateleckströme unterdrückt werden.

Der in Abbildung 5.19 dargestellte Teil eines neuen State-Retention-Flip-Flops benötigt hingegen kein zusätzliches rückgekoppeltes Inverterpaar zur Zustandserhaltung [6]. Stattdessen ist die Ausgangsstufe des SAFFs in Abbildung 5.15b so modifiziert, dass der Zustand auch im Standby-Modus gehalten werden kann. Die rückgekoppelten Inverter N7/P7 und N8/P8 können mit Hilfe der beiden Transmissiongates N13/P13 und N14/P14 vom übrigen Flip-Flop-Teil getrennt werden.

Die rückgekoppelten Inverter und die Transmissiongates können in einem Multi-t_{ox}-Design mit einem dickeren Oxid ausgeführt werden. Nur hier wird eine permanente Spannungsversorgung angelegt. Der übrige Dünnoxid-Teil des Flip-Flops wird an das virtuelle Null-Potential V_{ssv} angeschlossen. Gateleckströme werden so vollständig unterdrückt. Gleichzeitig verringert sich die Schaltgeschwindigkeit des Flip-Flops nur minimal, da die Ausgänge Q und $\overline{Q}$ außerhalb der Transmissiongates liegen. Alternativ können diese Ausgänge auch innerhalb der Transmissiongates abgeleitet werden. In diesem Fall liegen die Ausgangsdaten auch während des Standby-

Modus an. Dieses kann zum Beispiel dann erforderlich sein, wenn sich das Flip-Flop am Rande eines abgeschalteten Blocks befindet und die Ausgangsdaten in einem anderen Block, der weiterhin aktiv ist, benötigt werden.

Um einen sicheren Betrieb des Flip-Flops im aktiven und im Standby-Modus zu gewährleisten, muss zum einen ein sicheres Schalten der rückgekoppelten Inverter möglich sein. Dazu müssen die Transmissiongates im Verhältnis zu den Invertern ausreichend groß dimensioniert werden. Zum anderen dürfen Ladungen, die möglicherweise auf den Knoten Q und $\overline{Q}$ am Ende des Standby-Modus gepeichert sind, den gespeicherten Zustand nicht verändern. Diese Bedingungen sind vergleichbar mit der Schreib- bzw. Lesestabilität von SRAM-Zellen. Anders als in SRAM-Zellen, in denen besonders kleine Transistoren eingesetzt werden, ist hier ein erhöhter Flächenbedarf weniger kritisch. Es können daher größere Transistorweiten verwendet werden, um den Einfluss von Parametervariationen klein zu halten. Außerdem vereinfachen die PFETs P13 und P14 den Betrieb im Vergleich zu einer SRAM-Zelle. Die Dimensionierung ist dennoch sorgfältig durchzuführen.

Kritisch ist der zusätzliche Flächenbedarf des Gatters zu beurteilen, insbesondere wenn eine größere Oxiddicke in der Ausgangsstufe verwendet werden soll. Dieser Effekt lässt sich jedoch minimieren, wenn mehrere State-Retention-Flip-Flops im Layout so kombiniert werden, dass die Dickoxid-Anteile zusammen gelegt werden können.

Da die Ausgangsstufe in Abbildung 5.19 keine anderen Schaltungsteile treibt und sich somit außerhalb des kritischen Pfades befindet, reduziert die zusätzliche State-Retention-Funktionalität die Schaltgeschwindigkeit des Sense-Amplifier-Flip-Flops nicht. Gleichzeitig lässt sich der Leckstrom des Gatters abhängig von den Schwellenspannungen und Oxiddicken der verwendeten Transistoren um mehrere Größenordnungen reduzieren.

5.3.3 Sense-Amplifier-Flip-Flop für Skewed-CMOS-Logik

Skewed-CMOS-Schaltungen besitzen ein monotones Schaltverhalten, die daraus resultierenden Beschränkungen können jedoch umgangen werden. Wird an einer bestimmten Stelle in einer Schaltung ein inverses Signal benötigt, so wird die gesamte Logik zur Erzeugung dieses Signals in invertierter Form ein zweites Mal aufgebaut und am Eingang des Datenpfades mit jeweils inversen Eingangssignalen versorgt. In Skewed-CMOS-Logik können diese Signale jedoch nicht mit Hilfe eines einfachen Inverters erzeugt werden, da während der Precharge-Phase an allen N-Typ-Gatter eine 0 anliegen muss. Beide Flip-Flop-Ausgänge Q und $\overline{Q}$ müssen daher in der Precharge-Phase 0 sein. Mit der steigenden Taktflanke schaltet nur einer der Ausgänge auf 1.

Genau diese Eigenschaft weisen die beiden internen Knoten S und R in einem Sense-Amplifier-Flip-Flop auf. Die Sense-Amplifier der beiden SAFFs in den Abbildungen 5.15 und 5.16 können daher als Eingangs-Flip-Flop in einer Skewed-CMOS-Schaltung eingesetzt werden [98, 99]. Dabei kann das Eingangssignal D von einer statischen CMOS-Schaltung stammen. $\overline{D}$ entsteht durch lokale Inversion. Die Sense-Amplifier können somit eine Konversion von statischen zu Skewed-CMOS-Signalen durchführen (Abb. 5.18) und stellen in einer Skewed-CMOS-Logik ein vollwertiges Flip-Flop dar.

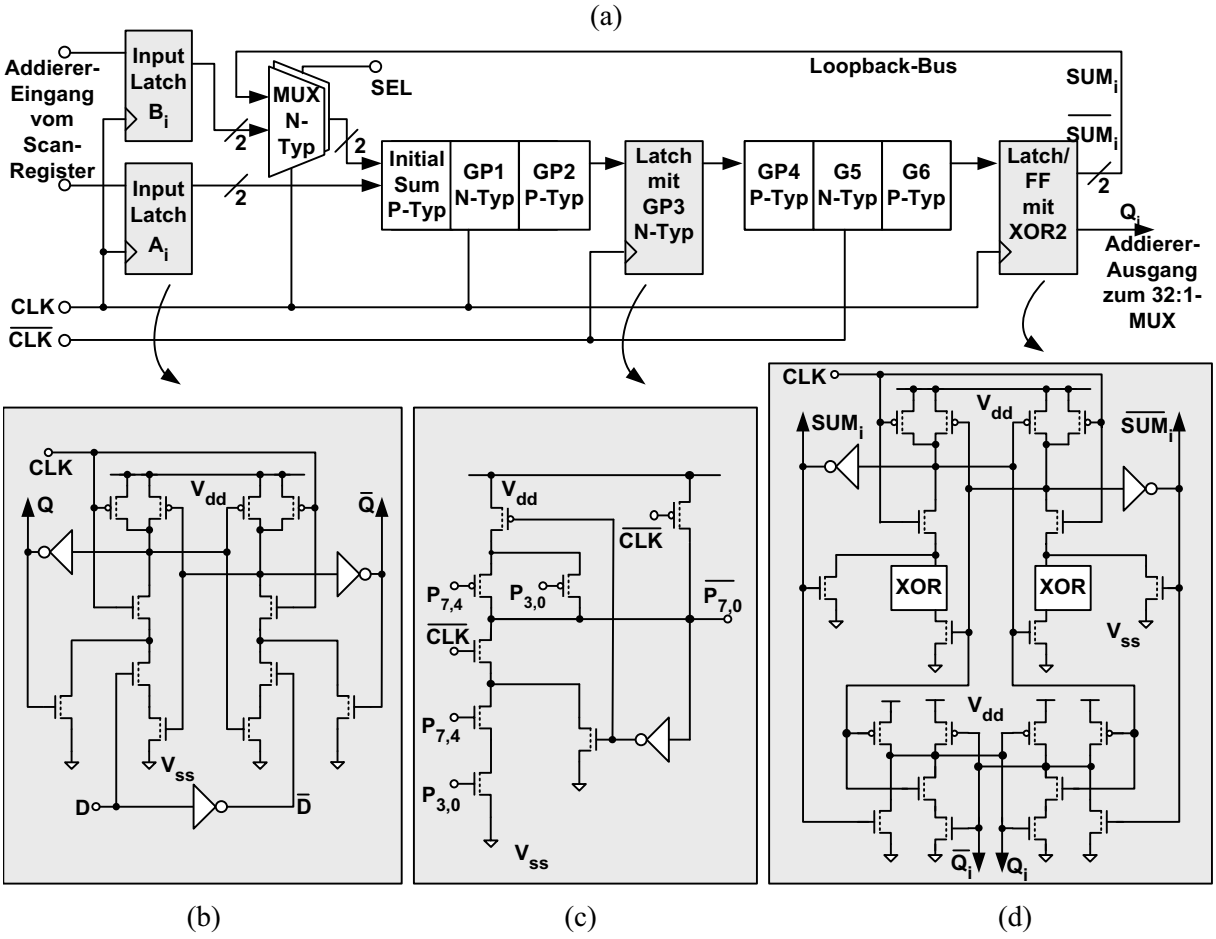


Abbildung 5.20: Flip-Flops und Latches im Datenpfad eines 32-bit-Parallel-Addierers (vgl. Kapitel 6).

Abbildung 5.20 zeigt einen Skewed-CMOS-Datenpfad, dessen logische Funktionalität im Kapitel 6 beschrieben wird. Am Eingang wird das beschriebene Skewed-CMOS-Flip-Flop dazu eingesetzt, ein statisches CMOS-Signal D in differentielle Skewed-CMOS-Signale Q und $\overline{Q}$ umzuwandeln.

Am Ausgang eines Skewed-CMOS-Datenpfades müssen die Skewed-CMOS-Signale wieder in Statisch-CMOS-Signale zurück konvertiert werden. Auch hier können SAFFs eingesetzt werden. Der Inverter, der zur Erzeugung des Flip-Flop-Eingangs $\overline{D}$ notwendig ist, benötigt dabei keinen Precharge-Transistor, selbst wenn es sich um ein Gatter vom N-Typ handelt. Das Aufladen kann über den langsamen PMOS-Pfad erfolgen, da sich dieser Inverter ganz am Ende der Taktphase befindet und somit nicht nur die Precharge-Phase, sondern auch der größte Teil der aktiven Phase zum Aufladen zur Verfügung steht.

Die Ausgänge des Sense-Amplifiers S und R können entweder wiederum als Eingangssignale einer weiteren Skewed-CMOS-Schaltung dienen (*SUM* und $\overline{SUM}$ in Abbildung 5.20a), oder sie stehen als statische Ausgänge Q und $\overline{Q}$ zur Verfügung (Abbildung 5.20d).

Mit Hilfe dieser Sense-Amplifier-Flip-Flops kann somit ein Skewed-CMOS-Schaltungsblock in eine statische CMOS-Schaltung eingebettet werden.

5.3.4 Skewed-CMOS-Latch

Das im letzten Abschnitt beschriebene Ausgangs-Flip-Flop kann als Interface zwischen einer Skewed-CMOS-Schaltung und einer statischen CMOS-Schaltung dienen. Zusätzlich werden skewed-CMOS-kompatible Signale zur Verfügung gestellt. Somit könnte die Sense-Amplifier-Stufe auch innerhalb eines Skewed-CMOS-Blocks dazu eingesetzt werden, die zwei Taktphasen voneinander zu trennen. Es besteht jedoch die Möglichkeit, hier ein sehr viel einfacheres Latch einzusetzen.

Das Skewed-CMOS-Latch in Abbildung 5.21 wird anstelle eines Skewed-CMOS-N-Typ-Gatters im Datenpfad eingesetzt. Es bildet das erste Gatter einer neuen Taktphase, das Eingangssignal D stammt von dem letzten P-Typ-Gatter der vorhergehenden Taktphase.

Während der Precharge-Phase wird der Ausgang Q auf 1 aufgeladen. Am Beginn der aktiven Phase wird $CLK = 1$ und N2 leitet. Hat das vorhergehende P-Typ-Gatter geschaltet, so ist $D = 1$ und der Ausgang des Latches wird $Q = 0$. Der Inverter in Verbindung mit den Transistoren N3 und P3 verhindert, dass der Ausgang Q wieder zurückschaltet, wenn das vorhergehende Gatter wieder auf 0 vorgeladen wird.

Es kann vorkommen, dass bei Beginn der aktiven Phase die vorhergehende Taktphase noch nicht vollständig durchlaufen wurde, z.B. durch Parameter-Schwankungen oder ungünstig platzierte Taktgrenzen. Erreicht ein Datum den Eingang D noch bevor der Precharge-Vorgang der vorhergehenden Taktphase beginnt, so wird dieses noch an den Ausgang Q invertiert weitergegeben. Die Schaltung wird dadurch auch gegen Schwankungen in der Taktversorgung weniger anfällig. Es muss daher keine weitere Sicherheit im Timing der Schaltung vorgehalten werden.

Zusätzlich ist es möglich, eine Logik-Funktionalität in das Latch zu integrieren. Anstelle der Transistoren N1 und P1 kann jede beliebige CMOS-Schaltung eingesetzt werden. Dadurch wird

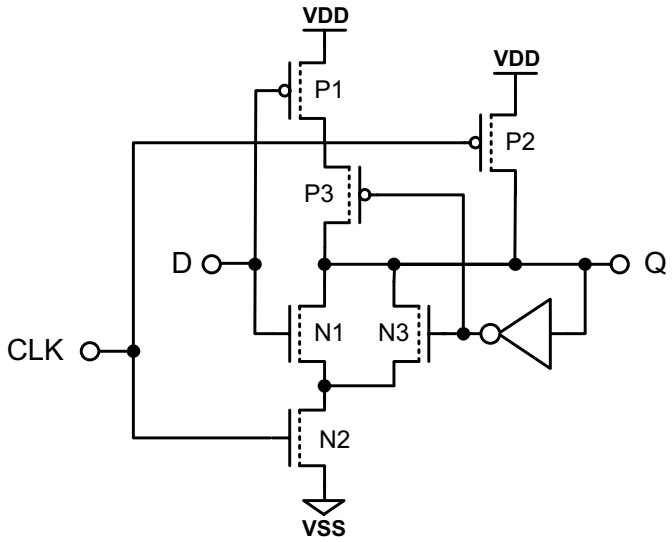


Abbildung 5.21: Skewed-CMOS-Latch. Ein spät eintreffender 0–1-Übergang an D wird noch nach Q propagiert, während der nachfolgende 0–1-Precharge-Übergang nicht weitergegeben wird.

die durch das Latch verursachte zusätzliche Verzögerungszeit auf ein Minimum reduziert. Von dieser Möglichkeit wird in Abbildung 5.20 in der Mitte des Datenpfads Gebrauch gemacht. Dargestellt ist eine NAND-Funktionalität, die anstelle der Transistoren N1 und P1 eingefügt wurde.

Mit Hilfe der in diesem Kapitel vorgestellten getakteten Speicherelemente lassen sich schnelle geschwindigkeitskritischen Schaltungsblöcke in Skewed-CMOS-Logik effizient realisieren. Das Sense-Amplifier-Flip-Flop eignet sich dazu, eine Skewed-CMOS-Schaltung in eine statische CMOS-Logik einzubetten. Das Skewed-CMOS-Latch verhindert, dass die Geschwindigkeitsvorteile dieser Logikfamilie durch das erforderliche zweiphasige Taktschema reduziert werden.

Die höhere Schaltgeschwindigkeit des optimierten Sense-Amplifier-Flip-Flops kann auch in klassischen CMOS-Schaltungen genutzt werden. Gleichzeitig kann eine Reduzierung der aktiven Leistungsaufnahme erreicht werden.

Die Zustandserhaltung von Flip-Flops im Standby-Modus stellt ein Schlüsselelement für die Realisierung von Low-Standby-Power-Anwendungen in Sub-100 nm-Technologien dar. Die vorgestellte Retention-Funktionalität für Sense-Amplifier-Flip-Flops ist ebenfalls in statischen CMOS-Schaltungen einsetzbar [6].

Kapitel 6

Low-Power-Schaltungstechniken am Beispiel von 32-bit-Addierern

Zwar können durch Simulationen und einfache Testschaltungen bereits weit gehende Erkenntnisse über die Wirksamkeit einer Schaltungstechnik gewonnen werden. Eine abschließende Beurteilung ist aber erst am Beispiel einer konkreten Anwendung möglich. In einer komplexen Schaltung können z.B. zusätzliche Leckstrompfade auftreten, oder es wird erkennbar, dass sich bestimmte Anforderungen an das Layout nicht effizient realisieren lassen.
Als Demonstrator wird ein 32-bit-Parallel-Prefix-Addierer gewählt, da dieser in vielen Anwendungen eine geschwindigkeitskritische Schaltungskomponente ist, z.B. in schnellen DSP-Anwendungen oder Mikroprozessoren. Zudem existieren zahlreiche vergleichbare Veröffentlichungen, die eine Einordnung der Ergebnisse erlauben [30, 100]. Häufig werden daneben auch Carry-Select-Addierer eingesetzt [69], die sich aufgrund ihres weniger regulären Aufbaus jedoch schlechter für ein manuelles Layout eignen. Zudem ist ihre Implementierung mit einer monoton schaltenden CMOS-Logik wie der Skewed-CMOS-Logik schwieriger.
Am Beispiel des im Folgenden beschriebenen Addierers werden verschiedene Techniken zur Leckstromreduzierung und zur Erhöhung der Schaltgeschwindigkeit verglichen: Die Verwendung von statischer und Skewed-CMOS-Logik, Forward- und Reverse-Biasing, der Einsatz von lokalen und zentralen Standby-Transistoren sowie die Minimum-Leakage-Vektor-Technik. Zudem werden dabei die im vorhergehenden Kapitel vorgestellten Flip-Flops und Latches verwendet. Darüber hinaus wird eine Mikroarchitektur für Parallel-Prefix-Addierer vorgestellt, die gleichzeitig eine hohe Schaltgeschwindigkeit, eine niedrige Leistungsaufnahme sowie ein flächensparendes Layout ermöglicht.

6.1 Parallel-Prefix-Algorithmus

Neben der Verbesserung des Transistors auf technologischer Ebene und der Anwendung neuer Techniken auf elementarer Schaltungsebene, ermöglichen Optimierungen auf Systemebene teilweise eine erhebliche Erhöhung der Schaltgeschwindigkeit oder eine Reduzierung der Leistungsaufnahme und der Fläche. Dabei kann z.B. für zwei unterschiedliche Schaltungstechniken jeweils ein anderer Algorithmus zu einem optimalen Ergebnis führen. Hier wird der

Parallel-Prefix-Addierer-Algorithmus verwendet und optimiert, der eine einfache Implementierung in einer Schaltungstechnik mit monotonem Schaltverhalten wie der Skewed-CMOS-Logik erlaubt.

Ein Ripple-Carry-Addierer (RCA) stellt die einfachste Realisierung eines n-bit-Addierers dar. In jeder Bit-Position i addieren Volladdierer die jeweiligen Daten a_i und b_i zu dem Übertrag des vorhergehenden Volladdierers c_{i-1}:

$$c_i = a_i \cdot b_i + (a_i + b_i) \cdot c_{i-1} \tag{6.1}$$

$$SUM_i = a_i \oplus b_i \oplus c_{i-1} \tag{6.2}$$

Die Verarbeitung erfolgt fast vollständig seriell. Lediglich dadurch, dass die Überträge schon im ersten Halbaddierer erzeugt werden, kann eine minimale Parallelität erreicht werden. Der Flächenbedarf ist zwar klein, jedoch steigt die Verarbeitungszeit proportional mit n an und wird daher z.B. für einen 32-bit-Addierer sehr groß ($t_d \propto n$) [69].

Der Parallel-Prefix-Algorithmus bietet eine Möglichkeit, die Verzögerungszeit eines n-bit-Addierers auf $t_d \propto 3 + \log_2 n$ zu reduzieren. Auf diese Weise kann eine 32-bit-Addition in einer Taktperiode erfolgen. In [101] wird ein solcher Addierer beschrieben, [102] liefert einen verallgemeinerten Ansatz zur parallelen Berechnung arithmetischer Operationen. Abbildung 6.1a zeigt die schematische Darstellung eines Kogge-Stone-Addierers. Jeder schwarz gefüllte so genannte Punkt-Operator besteht dabei aus zwei Logik-Gattern zur Berechnung des Generations-Übertrags $G_{i,j}$ und des Propagations-Übertrags $P_{i,j}$ eines Blocks von Bit-Position j (LSB) bis Bit-Position i (MSB):

$$G_{i,j} = \begin{cases} a_i \cdot b_i & \text{für} \quad i = j \\ G_{i,k} + P_{i,k} \cdot G_{k-1,j} & \text{für} \quad i \geq k > j \end{cases} \tag{6.3}$$

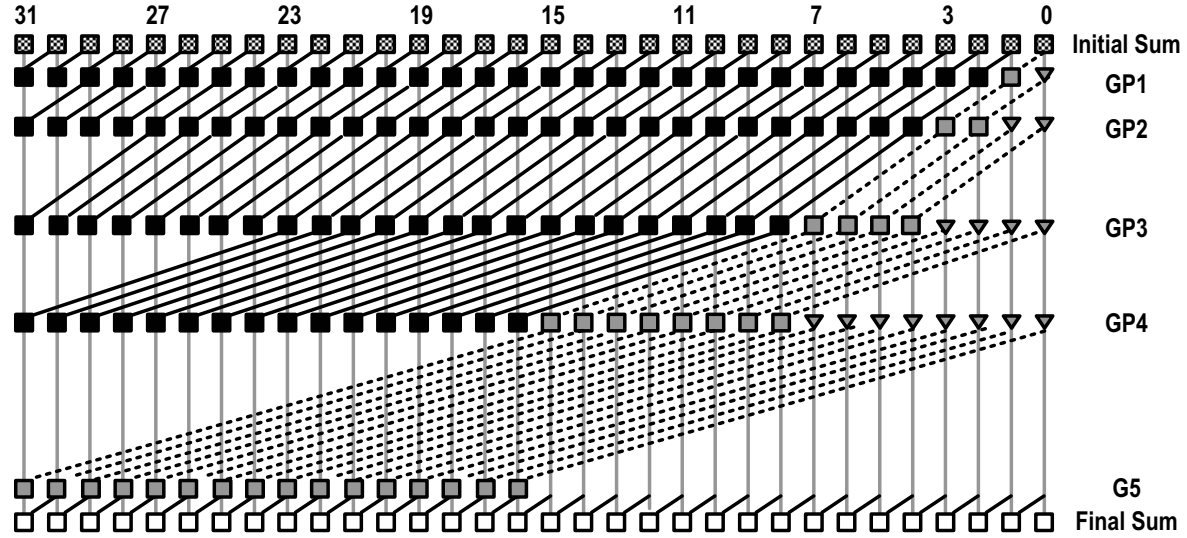
$$P_{i,j} = \begin{cases} a_i \oplus b_i & \text{für} \quad i = j \\ P_{i,k} + P_{k-1,j} & \text{für} \quad i \geq k > j. \end{cases} \tag{6.4}$$

$$SUM_i = P_{i,i} \oplus G_{i-1,0} \tag{6.5}$$

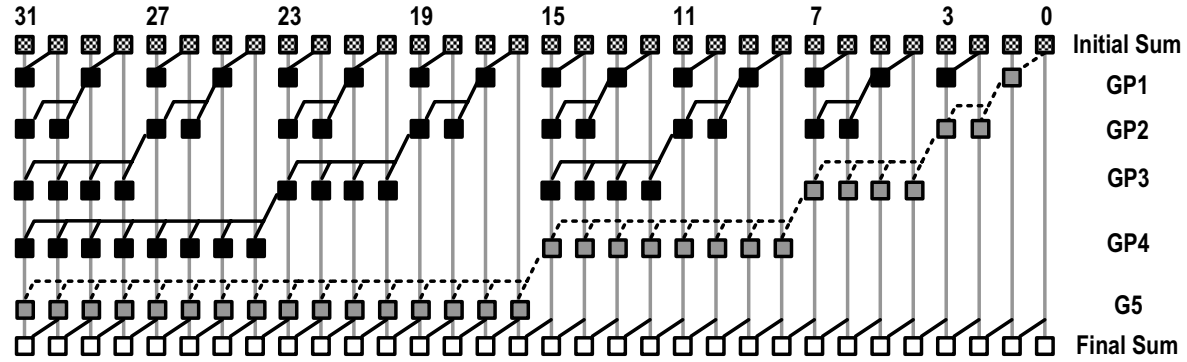
Das Gleichungssystem ist rekursiv und assoziativ. Dieses ermöglicht viele unterschiedliche Anordnungen der Punkt-Operatoren. Durch geeignete Umgruppierungen der Operatoren für VLSI-Implementierungen entstehen flächen- oder geschwindigkeitsoptimierte Addierer. Der jeweils letzte Punkt-Operator einer Bit-Position (grau) kann auf die Berechnung von $G_{i,j}$ reduziert werden, da $P_{i,j}$ hier nicht benötigt wird. Jede durchgezogene Linie entspricht den beiden Signalen G und P, die unterbrochenen Linien repräsentieren das einfache Signal G.

In einer effiziente CMOS-Realisierung wird jede zweite Stufe des Gleichungssystems 6.3–6.5 invertiert aufgebaut. Die Serienschaltung zweier G-Operatoren ist in Abbildung C.1 dargestellt.

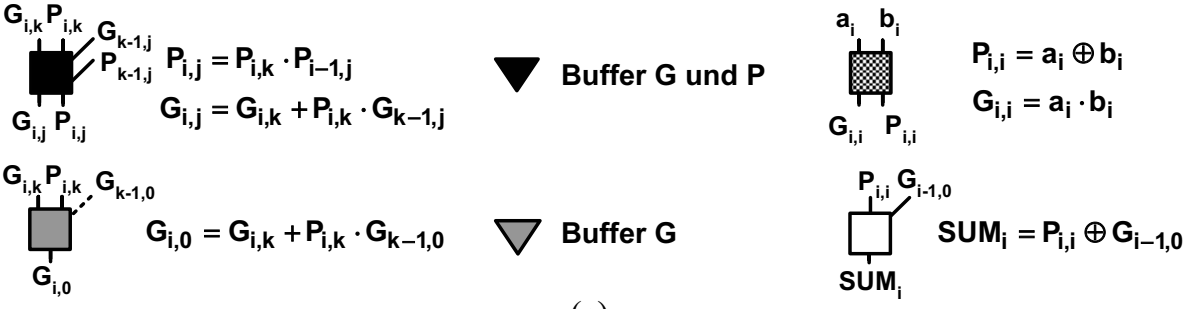
Der Kogge-Stone-Addierer benötigt die minimal mögliche Anzahl an Carry-Merge-Stufen $\log_2 32 = 5$. Zudem treibt jeder Punkt-Operator maximal 2 Operatoren, wodurch die Ausgangslast auf 3 begrenzt wird: Der Propagate-Operator P_i wird zweifach in derselben Bit-Position benötigt (P_{i+1} und G_{i+1}) und zudem einfach in einer höheren Bit-Position (G_{i+m}). Allerdings



(a)



(b)



(c)

Abbildung 6.1: Kogge-Stone- (a) und Ladner-Fischer-Addierer (b).

Addierer	GP-Stufen	Maximaler FO	Operatoren	Max. Verdrahtungsdichte	Verdrahtungslänge
Ladner-Fischer	5	16 + 1 Inv.	80	9	129
Kogge-Stone	5	3	129	46	961
Brent-Kung	9	3	57	8	178
Han-Carlson	6	3	81	32	495
diese Arbeit	6	3 + 1 Inv.	77	15	293

Tabelle 6.1: Vergleich von Kenngrößen verschiedener Addierer-Architekturen. Verglichen werden die Anzahl der Generate-Propagate-Stufen, der maximale Fan-out, die Anzahl der Punkt-Operatoren (GP und G), die maximale Anzahl an Verdrahtungsleitungen zwischen zwei Bit-Positionen sowie die Summe der Verdrahtungslängen in x-Richtung (Länge in Bit-Positionen).

ist auch die Anzahl der Operatoren mit 129 sehr groß, sodass die aktive Leistungsaufnahme sowie der Flächenbedarf maximal werden.

Prinzipiell gibt es zwei Möglichkeiten, die Anzahl der Dot-Operatoren zu reduzieren. Entweder wird die Anzahl der Carry-Merge-Stufen oder die maximal zulässige Ausgangslast erhöht.

Eine vollständig serielle Ausführung in 32 Stufen reduziert die Anzahl der Operatoren auf 32, allerdings bei maximaler logischer Tiefe.

Der Ladner-Fischer-Addierer [103] besitzt wie der Kogge-Stone-Addierer minimale logische Tiefe (Abb. 6.1b). Durch eine höhere Ausgangslast kann die Anzahl der aktiven Operatoren reduziert werden. Allerdings wird die Last in einem 32-bit-Addierer in der letzten Stufe so groß, dass hier zusätzliche Treiber eingesetzt werden müssen, sodass sich die Verzögerungszeit des Addierers trotz minimaler Tiefe erhöht.

In Tabelle 6.1 werden verschiedene Addierer nach den Kriterien logische Tiefe, maximale Ausgangslast, Anzahl der Operatoren, Verdrahtungsdichte und -länge verglichen. Die Verdrahtungsdichte wird berechnet, indem die Längen aller Carry-Leitungen (in Bit) addiert werden. Kürzere Verdrahtungslängen und -dichten reduzieren die parasitären Verluste, die beim Umladen der Leitungskapazitäten entstehen. Zudem wird ein kompakteres Layout möglich. Die drei ersten Addierer in der Tabelle weisen bezüglich einzelner Kriterien weitgehend ideale Eigenschaften auf. So besitzen der Ladner-Fischer- sowie der Kogge-Stone-Addierer eine minimale logische Tiefe von $\log_2 32 = 5$. Brent-Kung- und Ladner-Fischer-Addierer besitzen niedrige Verdrahtungsdichten und -längen. Dieses kann jedoch jeweils nur auf Kosten anderer Kriterien erreicht werden, sodass Implementierungen als 32-bit-Addierer ineffizient werden: im Ladner-Fischer-Addierer mit großem Fan-out, im Kogge-Stone-Addierer mit vielen aktiven Gattern und hohem Verdrahtungsaufwand und im Brunt-Kung-Addierer mit großer logischer Tiefe.

In einer Vielzahl von Veröffentlichungen werden verschiedene Addierer zwischen diesen extremen Lösungen vorgeschlagen, um je nach Anforderung eine unterschiedliche Kombination zwischen logischer Tiefe, Ausgangslast und Anzahl der Operatoren zu finden [104, 105]. Am häufigsten wird heute die Han-Carlson-Architektur verwendet, die einen guten Kompromiss

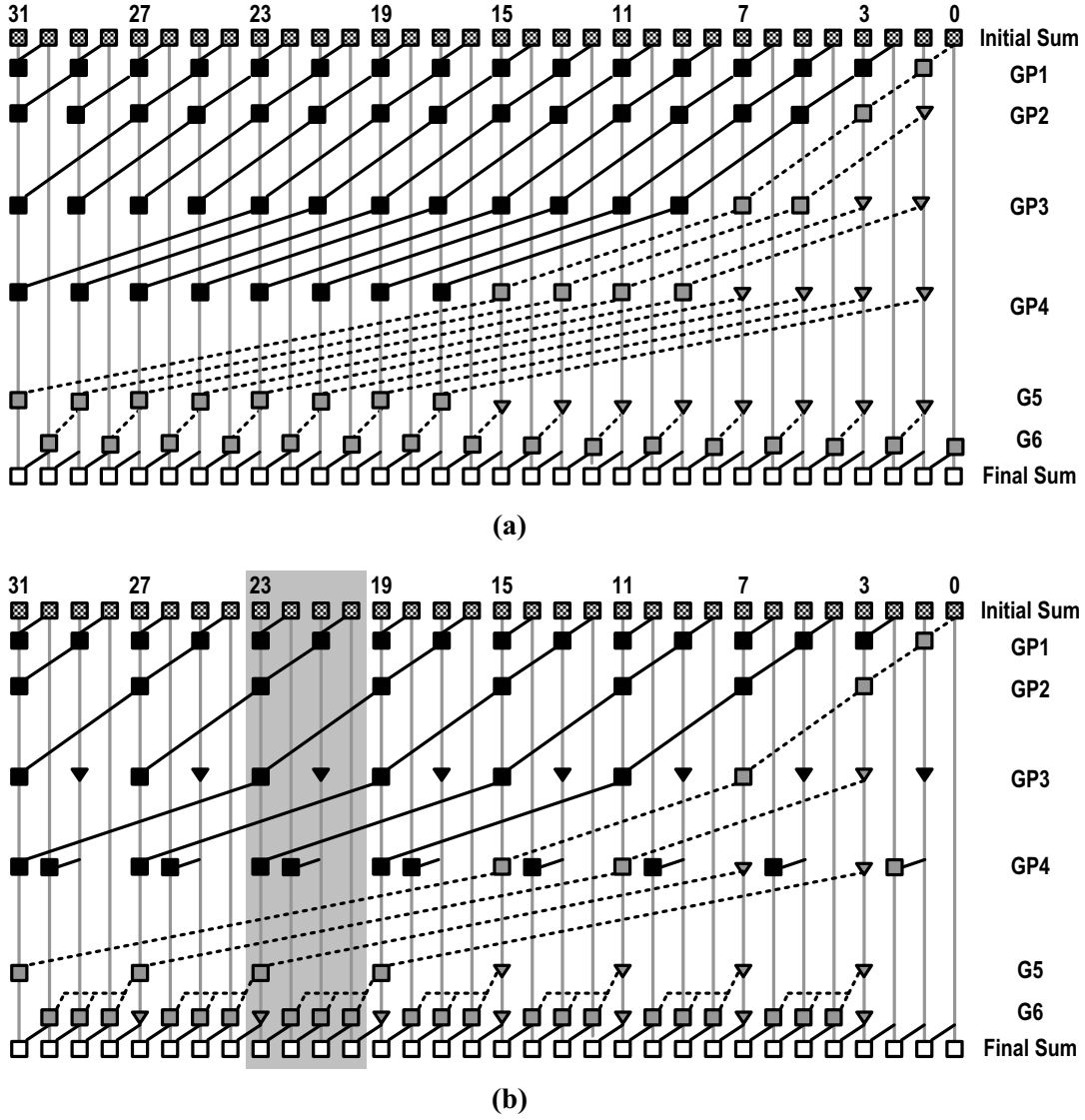


Abbildung 6.2: Han-Carlson-Addierer (a) und neuer Addierer (b).

zwischen den Kriterien in Tabelle 6.1 darstellt [26, 30, 100, 106]. Bei einer logischen Tiefe von 6 und kleinem Fan-out bleiben die Anzahl der Operatoren und die Verdrahtungsdichte relativ klein (Abb. 6.2a).

Abbildung 6.2b zeigt einen Addierer, der wie der Han-Carlson-Addierer sechs Stufen, aber weniger Operatoren benötigt und eine deutlich geringere Verdrahtungsdichte aufweist. Es werden nur vier 16-bit-Carry-Leitungen benötigt, wodurch die Verdrahtungsdichte reduziert wird. Die Ausgangslast in der letzten Stufe ist nur leicht erhöht, obwohl drei Operatoren und ein Inverter mit dem Ausgang verbunden sind. Dieser Inverter kann minimal dimensioniert werden und jeder der drei Operatoren hat nur einen Fan-out von 1, da kein P-Gatter in den nachgeschalteten

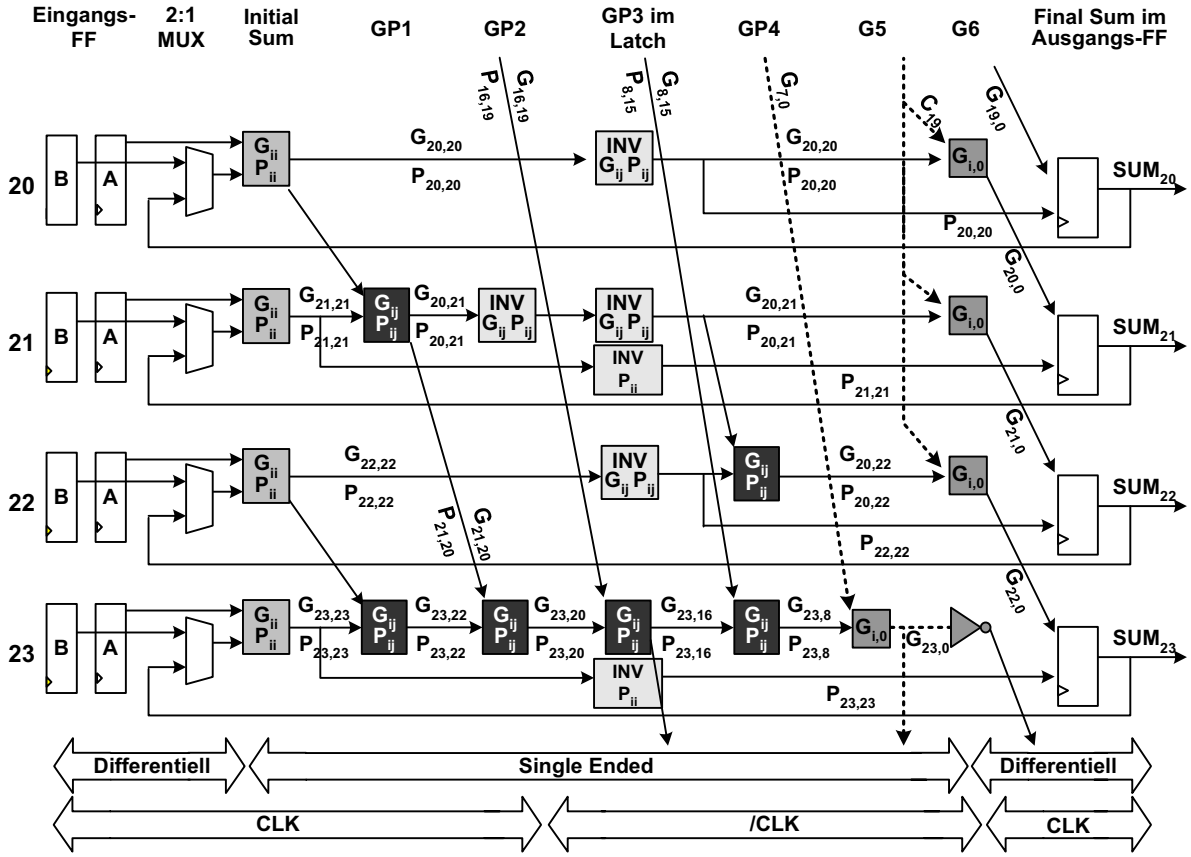


Abbildung 6.3: 4 Bit des Blockschaltbildes eines Skewed-CMOS-Addierers (vgl. grau hinterlegten Bereich in Abb. 6.2b).

Operatoren berechnet werden muss. Die Erhöhung der maximalen Ausgangslast um einen Inverter ist vertretbar, da die kurzen 3-bit-Leitungen nur geringe parasitäre Kapazitäten besitzen.

Erst die endgültige Implementierung einer realen Schaltung wie einem Skewed-CMOS-Addierer macht die Unterschiede zu einer klassischen statischen CMOS-Schaltung deutlich. Ein 4-bit-Block des vorgestellten Addierers ist in Abbildung 6.3 dargestellt.

Die Unterschiede resultieren hauptsächlich aus dem monotonen Schaltverhalten. So muss in der Mitte des Datenpfades ein Latch eingesetzt werden, welches jedoch auch eine Logikfunktion ausführen kann (GP3 in Abbildung 6.3). Alle Signale, die aus der ersten in die zweite Hälfte führen, werden mit dem Latch aus Abbildung 5.21 versehen. Darüber hinaus muss auch im übrigen Logikteil ein strikter Wechsel zwischen N- und P-Typ-Gattern eingehalten werden. Dazu wird z.B. an Position GP2 in der Bitposition 21 ein Inverter hinzugefügt.

Anstelle einer automatisierten Design-Umgebung erfolgt die Layout-Erstellung vollständig manuell. Dieses verbessert die Vergleichbarkeit der Messergebnisse, da direkt Einfluss auf die Verdrahtung genommen werden kann. Die reguläre Struktur der einzelnen 4-Bit-Blöcke ermöglicht ein rekursives Design.

6.2 Implementierung von vier 32-bit-Addierern

Das Ziel der Implementierung von vier verschiedenen Addierern auf einem 90-nm-Testchip ist, die verwendeten Techniken zur Reduzierung der Leistungsaufnahme und Erhöhung der Schaltgeschwindigkeit unter realistischen Bedingungen zu testen.
Insgesamt werden drei Skewed-CMOS- und ein statischer Addierer vorgestellt (Tabelle 6.2). Der erste Skewed-CMOS-Addierer verwendet schnelle LVT-Transistoren sowohl in den Evaluationspfaden als auch in den Precharge- und Haltepfaden. Er soll die mit Skewed-CMOS-Logik erreichbare hohe Schaltgeschwindigkeit demonstrieren. Der Skew-Effekt wird ausschließlich über unterschiedliche Transistorweiten für Evaluations- und Haltepfade erreicht. Ein zweiter ansonsten identischer Addierer verwendet statt der LVT-Transistoren REG-Transistoren. Beide Addierer, die in einem Modul auf dem Testchip zusammen gefasst sind (Abb. 6.4), können jeweils mit Hilfe eines zentralen NMOS-Standby-Transistors (LL-LVT) in einen leckstromarmen Zustand versetzt werden. Das Blockschaltbild eines Addiererkerns mit Peripherieschaltungen zur Takterzeugung und Taktaufbereitung sowie Eingangsdaten-Register, Auslese-Multiplexer und Standby-Transistor ist in Abbildung 6.5 dargestellt.
Der Standby-Transistor ist so dimensioniert, dass der maximale Spannungseinbruch auf der V_{ss}-Leitung, der am Anfang der Taktperiode beim Schalten des Flip-Flops auftritt, in der Simulation nicht mehr als 30 mV beträgt. In der realen Schaltung stabilisieren parasitäre Kapazitäten die Spannung der Versorgungsleitungen zusätzlich. Der Standby-Transistor hat eine Weite von 350 μm. Der zusätzliche Flächenbedarf beträgt 225 $\mu\mathrm{m}^2$ oder 2.5 % der Fläche des Addiererkerns und stellt somit nur eine geringe Vergrößerung der Schaltung dar.
Die Gatelänge des Standby-Transistors ist um 20 nm größer als die minimale Gatelänge, da der Off-Strom aufgrund V_t-Roll-offs in Abbildung 3.5 dadurch kleiner wird. Die Gesamtweite von 350 μm setzt sich aus mehreren parallel geschalteten Transistorfingern zusammen, deren Weite 3.5 μm beträgt.
Der Off-Strom ließe sich weiter reduzieren, indem die Weite der einzelnen Finger mit 0.8 μm so gewählt wird, dass der Off-Strom minimal wird (Abb. 3.8). Alternativ kann der Off-Strom des LL-LVT-Transistors dadurch reduziert werden, dass die Gatespannung unter $V_{gs} = 0\,\mathrm{V}$ abgesenkt wird. Diese in [107] als *Super Cut-off* bezeichnete Technik ist jedoch nur effizient, wenn der Off-Strom des Standby-Transistors wie hier durch den Unterschwellenstrom dominiert wird. Würde an dieser Stelle der GIDL-limitierte LL-Transistor verwendet, könnte die Super-Cut-off-Technik nicht sinnvoll eingesetzt werden.
Der dritte Addierer ist eine Skewed-CMOS-Variante, welche beide im Abschnitt 5.1.2 vorgestellten Techniken zur Leckstromreduzierung einsetzt: Minimum-Leakage-Vektor-Technik und die Verwendung lokaler Standby-Transistoren. Die Evaluationspfade bestehen aus LVT-Transistoren, die Precharge-Pfade sowie die Takttransistoren aus LL-LVT-Transistoren. Der Skew-Effekt entsteht hier nicht nur durch die Transistorweite, sondern zusätzlich durch die Verwendung unterschiedlicher Schwellenspannungen und Oxiddicken.
Der statische Addierer verwendet ausschließlich Gatter aus einer Standardzellen-Bibliothek. Dieser Addierer dient als Referenz. Er ist wegen der hohen Schwellenspannung und des dicken Oxides der verwendeten LL-Transistoren relativ langsam und leckstromarm.

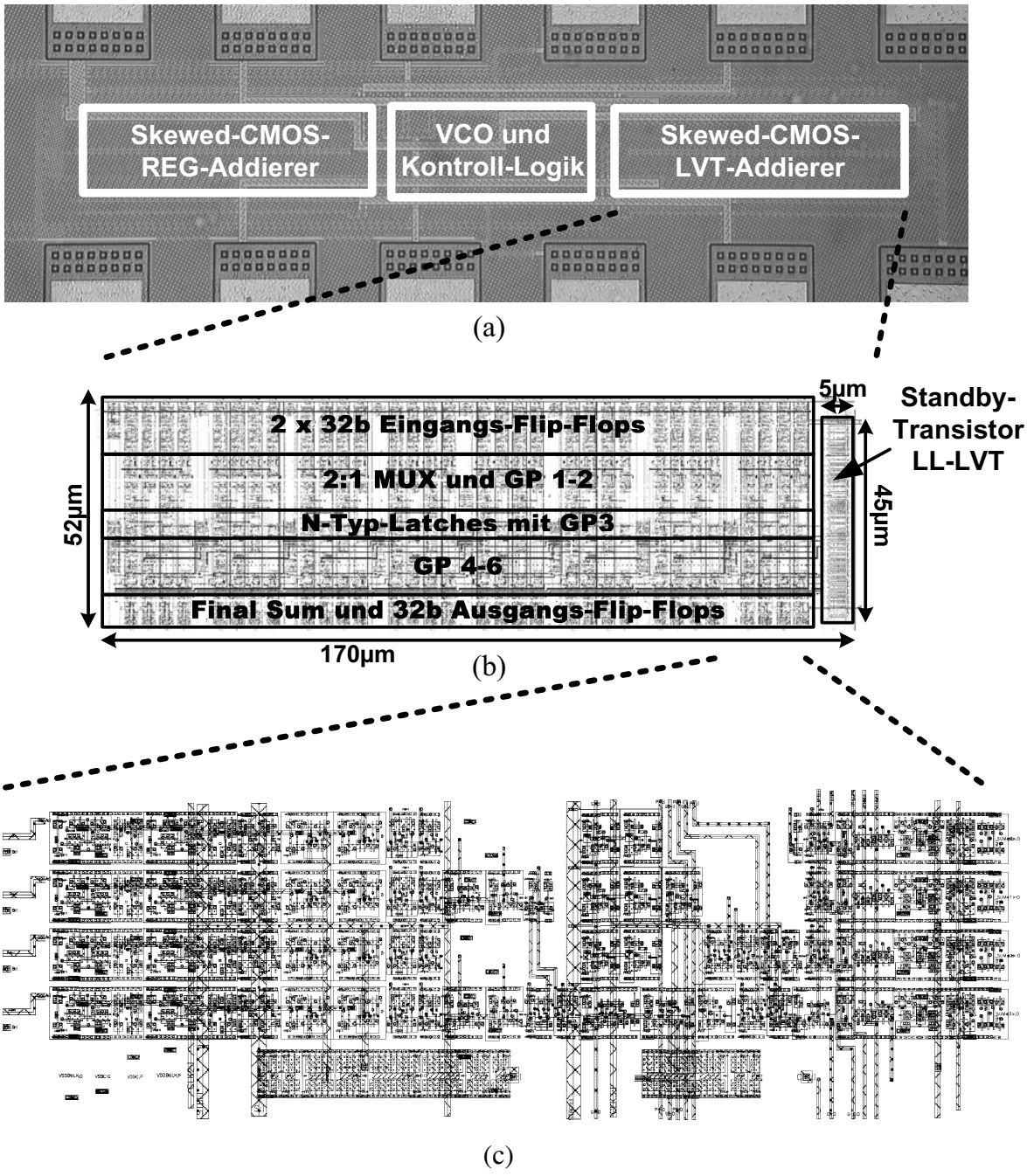


Abbildung 6.4: Foto eines Testchip-Moduls mit zwei Addierern (a), Schematic eines Addiererkerns (b) und Schematic ein 4-bit-Blocks (c). In dem 4-bit-Block sind die lokalen Takt-Treiber unterhalb der vier Bit-Streifen zu erkennen.

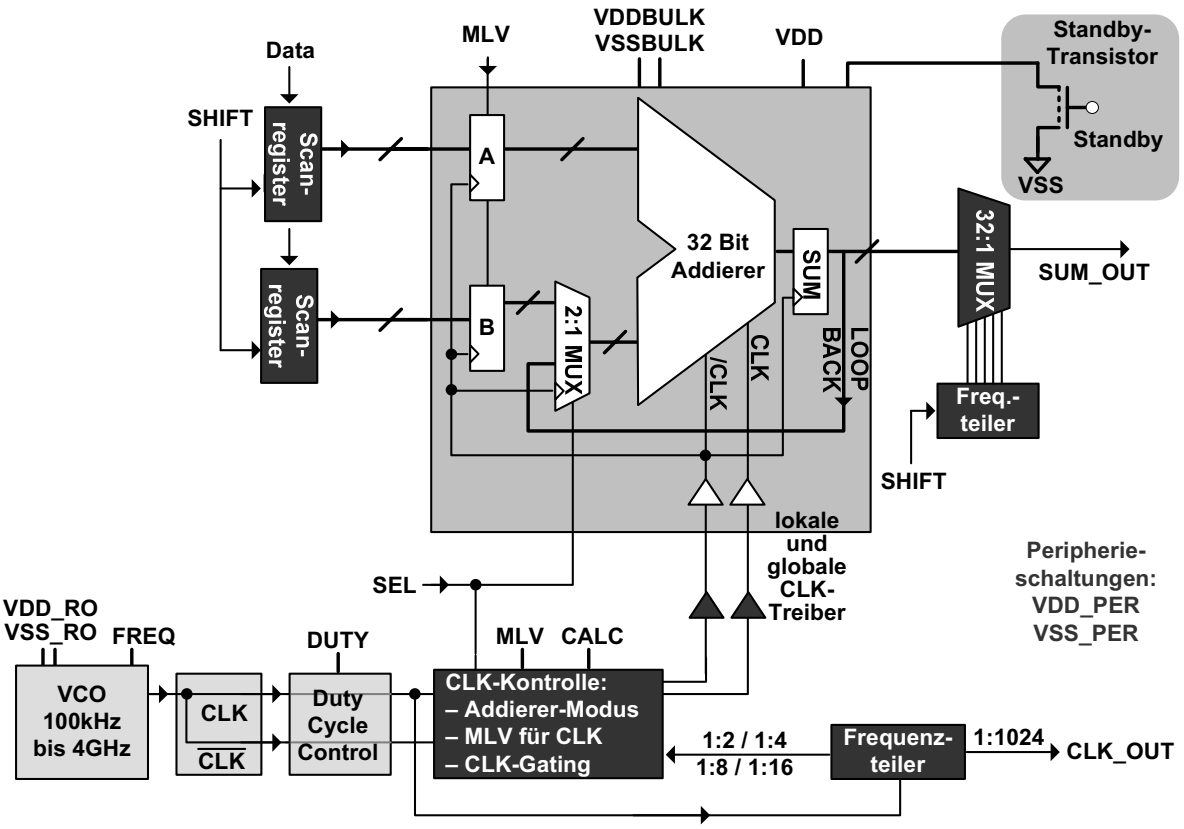


Abbildung 6.5: Blockschaltbild eines Addierers mit Peripherieschaltungen.

Bezeichnung	Transistoren schneller Pfad	Transistoren langsamer Pfad	Techniken zur Leckstromreduzierung	Skew-Effekt
Skewed-CMOS LVT	LVT	LVT	zentrale Blockabschaltung	W
Skewed-CMOS REG	REG	REG	zentrale Blockabschaltung	W
Sk.-CMOS Multi-t_{ox}	LVT	LL	lokale Stb.-Trans. und MLV	W, V_t und t_{ox}
Statisches CMOS	LL		nur Body Biasing	–

Tabelle 6.2: Übersicht über die vier Addierer-Varianten auf dem Testchip.

Die Wannenkontakte aller vier Addierer werden direkt mit Ausgangs-Pads des Chips verbunden. Dadurch können Body-Biasing-Potentiale direkt von außen an die Addiererkerne angelegt werden. Die Effizienz von Body-Biasing kann so für komplexe Logikschaltungen bestimmt werden.

Bei der Implementierung wurde darauf Wert gelegt, dass alle Signale, die auf den Chip oder von dem Chip herunter geführt werden, nicht geschwindigkeitskritisch sind und keiner externen Synchronisation bedürfen (Abb. 6.5).
Mit Hilfe eines abstimmbaren Ringoszillators (*Voltage-Controlled Oscillator*, VCO) können Frequenzen zwischen 100 kHz und 4 GHz erzeugt werden. Die Regelung erfolgt über zuschaltbare Lastkapazitäten (Faktor 4) und über die, von der übrigen Schaltung getrennte, Spannungsversorgung des Ringoszillators. Die aktuelle Frequenz kann über einen Frequenzteiler ausgelesen werden. Anschließend wird das inverse Taktsignal $\overline{CLK}$ erzeugt und mit CLK synchronisiert. In der nachfolgenden Schaltung *Duty Cycle Control* kann das Tastverhältnis der Taktsignale eingestellt werden, da die Betriebssicherheit von Skewed-CMOS-Schaltungen erhöht wird, wenn sich die $CLK = 1$- und $\overline{CLK} = 1$-Taktphasen überlappen.
Zu Testzwecken wird ein Multiplexer und eine Rückführung der Summe (*Loopback Bus*) zum Addiereingang in den Datenpfad aufgenommen, sodass zwischen zwei Betriebsmodi gewählt werden kann:

$$SUM_i = A_i + (B_i \cdot SEL) + (SUM_{i-1} \cdot \overline{SEL}). \tag{6.6}$$

Im ersten Modus wird eine begrenzte Anzahl von Additionen durchgeführt, sodass die logische Funktionalität des Addierers getestet werden kann. Dabei werden im ersten Taktzyklus die beiden Eingangsregister A und B addiert ($SEL = 1$). Für die drei folgenden Zyklen schaltet der 2:1-Multiplexer das Loopback-Signal auf den zweiten Addierer-Eingang, sodass jetzt A zum Ergebnis der ersten Addition hinzu addiert wird. Am Ende liegt das Ergebnis $SUM = 4A + B$ in den Ausgangs-Flip-Flops und kann über einen 32:1-Multiplexer ausgelesen werden. Der zweite Testmodus (kontinuierlicher Modus) dient der Bestimmung der maximalen Geschwindigkeit. Der Addierer wird dazu als Akkumulator betrieben. Der Multiplexer schaltet mit $SEL = 0$ das *SUM*-Signal fest auf den zweiten Eingang $SUM_i = A_i + SUM_{i-1}$. Wenn jetzt z.B. $A = 1$ gesetzt wird, zählt der Addierer am Ausgang langsam hoch, sodass im Laufe von $2^{32} = 4\,294\,267\,296$ Taktzyklen alle Ausgangszustände durchlaufen werden. Über den 32:1-Multiplexer kann jedes beliebige Ausgangsbit während des Betriebs beobachtet werden. Selbst bei einer Betriebsfrequenz von 4 GHz würde das höchstwertige Bit (*Most Significant Bit*, MSB) eine Frequenz von weniger als 1 Hz besitzen. Indem die Frequenz des VCOs variiert wird, kann die maximale Betriebsfrequenz einfach und schnell bestimmt werden.
Der Datenpfad des Skewed-CMOS-LVT-Addierers sowie die verwendeten Flip-Flops sind in Abbildung 5.20 dargestellt. In der Skewed-CMOS-Ausführung wird die *Final Sum* erst im Ausgangs-Flip-Flop berechnet, da die Realisierung der XOR-Funktion nicht in Skewed-CMOS-Logik möglich ist. Es müsste schon vor dem Ausgangs-Flip-Flop statische CMOS-Logik eingesetzt werden, wodurch sich die Schaltgeschwindigkeit reduzieren würde. Das Ausgangssignal und die Rückführung der Summe (Loopback-Bus) sind erst mit Beginn der steigenden Taktflanke verfügbar. Im Unterschied zu statischer CMOS-Logik ist der Multiplexer (2:1 MUX) daher

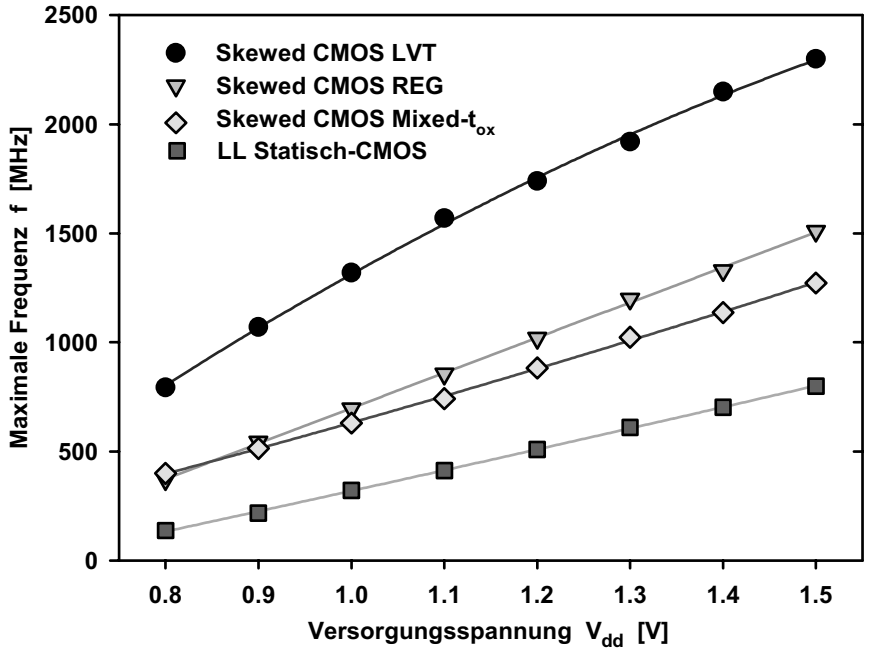


Abbildung 6.6: Gemessene maximale Betriebsfrequenzen bei Variation der Versorgungsspannung.

auch hinter den Eingangs-Flip-Flops anzuordnen. Zusätzlich muss das Auswahlsignal für den Multiplexer während der gesamten CLK=1-Phase gültig sein. Der Loopback-Bus sowie die Ausgänge der Flip-Flops A und B sind Skewed-CMOS-Signale, die jeweils differentiell zur Berechnung der *Initial Sum* im ersten Logikgatter geführt werden.

In Skewed-CMOS-Schaltungen müssen nicht nur die Flip-Flops, sondern auch die Latches sowie jedes N-Typ-Gatter mit einem Taktsignal versorgt werden. Auf dem Testchip blieb die Metall-Lage 4 der Takt-Versorgung vorbehalten. Während in einer realen Anwendung das inverse Taktsignal $\overline{CLK}$ nach Möglichkeit durch lokale Inversion erzeugt wird, werden CLK und $\overline{CLK}$ hier getrennt generiert, aufbereitet und über die Schaltung verteilt, um zu Testzwecken das Tastverhältnis (*Duty Cycle*) der beiden Signale sowie deren Überlappung individuell einstellen zu können (Abb. 6.5). Bei der Messung zeigt sich jedoch, dass der Addierer durch die Verwendung des Skewed-CMOS-Latches in der Mitte des Datenpfades tolerant gegen Schwankungen in der Taktversorgung wird (*Clock Jitter* von CLK und $\overline{CLK}$).

6.3 Leckstrom und Geschwindigkeit

Die maximale Betriebsfrequenz der einzelnen Addierer wird im kontinuierlichen Modus bestimmt (Abb. 6.6). Der Skewed-CMOS-LVT-Addierer erreicht bei einer nominalen Versorgungsspannung von $V_{dd} = 1.2\,\mathrm{V}$ eine Frequenz von 1.73 GHz. Bei erhöhter Spannung $V_{dd} = 1.5\,\mathrm{V}$ und 0.5 V Forward-Biasing werden bis zu 2.5 GHz erreicht (Abb. 6.7). Die Geschwindigkeit des Skewed-CMOS-REG-Addierers ist wegen der höhere Schwellenspannung der REG-Transistoren bei 1.2 V um 42 % kleiner (Abb. 6.6).

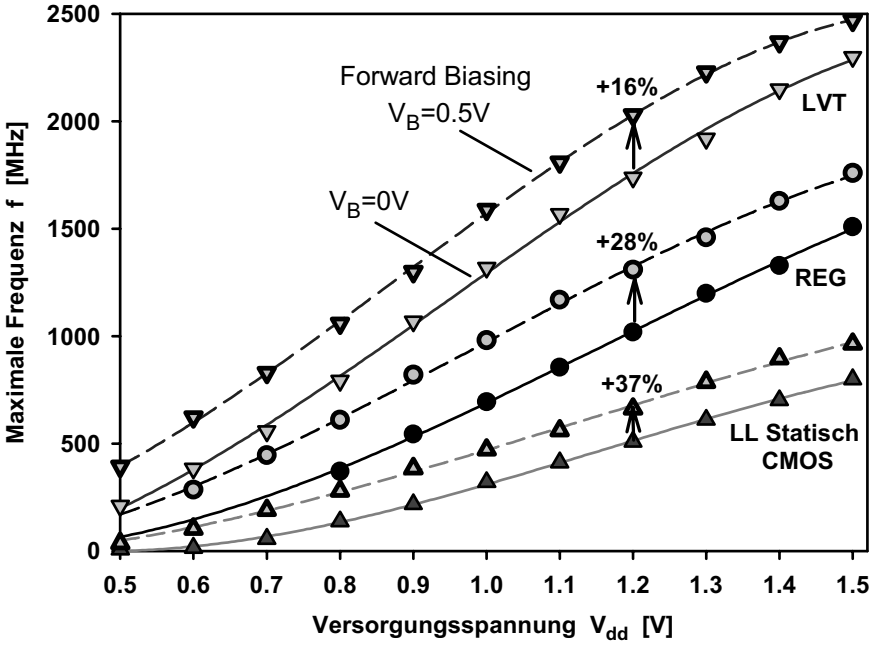


Abbildung 6.7: Erhöhung der Addierergeschwindigkeit durch Forward-Biasing (Messung). Die Geschwindigkeitszunahme des Mixed-t_{ox}-Addierers entspricht der des LVT-Addierers (jeweils 16 %), da beide ausschließlich LVT-Transistoren im kritischen Pfad einsetzen.

Der statische LL-Addierer benötigt die 3.4-fache Zeit für eine Addition. Diese Zunahme ergibt sich zum einen aus der höheren Schwellenspannung (Faktor 2.4), zum anderen aus der unterschiedlichen Schaltungstechnik und der Verwendung der optimierten Sense-Amplifier-Flip-Flops (Faktor 1.4). Die Erhöhung der Schaltgeschwindigkeit um 40 % demonstriert die Fähigkeit der Skewed-CMOS-Logik, eine signifikante Performance-Steigerung zu erzielen.

Sowohl der Skewed-CMOS-LVT- als auch der Skewed-CMOS-Mixed-t_{ox}-Addierer verwenden LVT-Transistoren im kritischen Pfad. Die maximale Frequenz des Mixed-t_{ox}-Addierers ist jedoch deutlich kleiner. Dieses hat mehrere Gründe. Zunächst mussten beim Schaltungsentwurf und beim Layout Kompromisse eingegangen werden, um mehrere unterschiedliche Standby-Modi in einem einzigen Addierer zu implementieren. Auch die Flip-Flops mussten für den MLV-Modus modifiziert werden. Außerdem wurden die Transistorweiten des kritischen Pfades auf eine kleine aktive Leistungsaufnahme hin dimensioniert, wodurch der Einfluss parasitärer Leitungskapazitäten größer wird und sich Weiteneffekte stärker auswirken. Darüber hinaus waren auf dem experimentellen Testchip nicht alle der acht Schwellenspannungen (LVT, REG, LL-LVT und LL jeweils für NMOS- und PMOS-Transistor) vollständig aufeinander abgestimmt.

Zusätzlich zur Versorgungsspannung können in den Addierern die Bulk-Potentiale $V_{ss,bulk}$ und $V_{dd,bulk}$ variiert werden (Abb. 6.7). Wie im Kapitel 4 beschrieben, ist die Geschwindigkeitszunahme durch Forward-Biasing in dieser komplexen Schaltung aufgrund der häufig verwendeten Gatter mit Transistorstacks größer als in einfachen Ringoszillatoren (vgl. Tabelle 4.2).

Leckströme sowie aktive Schaltströme werden in Abbildung 6.8 den Schaltgeschwindigkeiten der Addiererkerne gegenübergestellt. Die Leckströme beziehen sich auf den gesamten Addie-

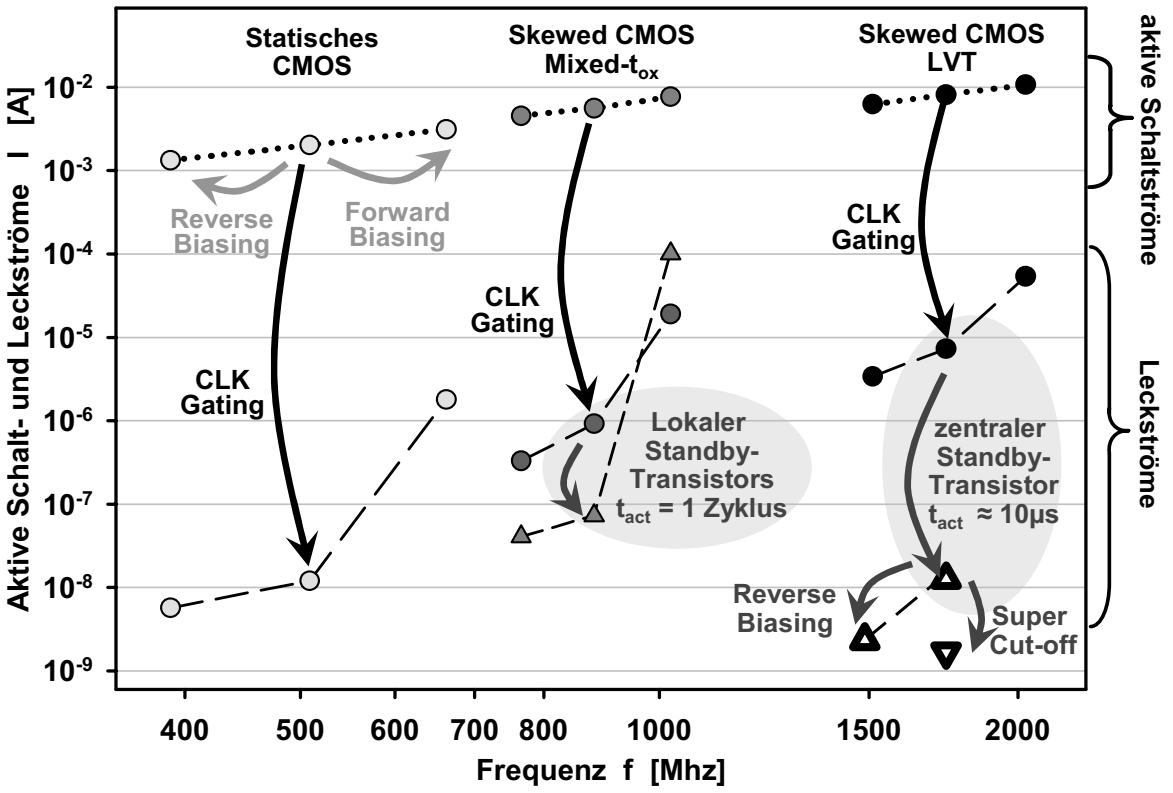


Abbildung 6.8: Leckströme und Schaltgeschwindigkeit von drei Addierern bei $V_{dd} = 1.2\,\mathrm{V}$ und 25 °C.

rerblock, der inklusive Flip-Flops und Latches aus ca. 3000 Transistoren besteht. Wie schon für Einzeltransistoren und Ringoszillatoren existiert auch für komplexe Schaltungsblöcke ein Trade-off. Daraus resultiert auch hier eine Trendlinie. Neben technologischen und schaltungstechnischen Optionen hat hier außerdem die Art der Implementierung des Addierers einen entscheidenden Einfluss.

Die Schaltströme des statischen CMOS-Addierers betragen bei einer Betriebsfrequenz von 510 MHz und bei hoher Datenaktivität an den Eingängen des Addierers 1.9 mA. Durch den hohen Anteil an getakteten Master-Slave-Flip-Flops unterscheidet sich die Stromaufnahme bei konstanten Eingangssignalen hiervon nur wenig (1.5 mA, Abb. 6.8).

Beim Übergang vom statischen zu den Skewed-CMOS-Addierern erhöht sich die aktive Stromaufnahme zum einen wegen der höheren Taktfrequenz und der damit verbundenen höheren Aktivität. Zum anderen besitzen Skewed-CMOS-Schaltungen aber auch eine größere Anzahl getakteter Gatter. Darüber hinaus werden einige interne Knoten in Logiken mit einem Precharge-Vorgang in jedem Taktzyklus aufgeladen und wieder entladen, auch ohne dass sich der Eingangsvektor ändert. Die aktive Stromaufnahme des Skewed-CMOS-LVT-Addierers ist bezogen auf einen Taktzyklus je nach Eingangsvektor um 30 bis 60 % höher als der des statischen Addierers.

Es ist zu erkennen, dass Forward-Biasing im Low-Leakage-Addierer am effizientesten ist. Die erhöhte Leistungsaufnahme resultiert zum einen aus der höheren Taktfrequenz, zum anderen aber auch aus den erhöhten Gate- und Junction-Kapazitäten.

Bevor andere Maßnahmen zur Leckstromreduzierung greifen können, muss zunächst die hohe aktive Stromaufnahme eines vorübergehend nicht benötigten Schaltungsblocks ausgeschaltet werden. Dieses geschieht durch eine Unterbrechung der Taktversorgung (*Clock Gating*). Allein durch diese Maßnahme lässt sich der Leckstrom des statischen LL-Addierers um über fünf Dekaden auf 12 nA reduzieren (Stromaufnahme für den gesamten 32-bit-Addierer). Reverse-Biasing reduziert den Leckstrom nur um 50 %, da der GIDL-Effekt eine weitere Reduzierung von I_{off} verhindert.

Die Reduzierung der Stromaufnahme fällt bei dem LVT-Addierer sehr viel geringer aus (6.7 mA auf 7.3 μA). Dennoch beträgt der Abstand zwischen den Strömen in diesem Block immer noch fast drei Dekaden. Diese Beobachtung scheint nicht mit der Aussage übereinzustimmen, wonach Leckströme in 90 nm-CMOS-Technologien teilweise schon die bestimmende Quelle der gesamten Leistungsaufnahme sind (Kapitel 3). Der Unterschied lässt sich dadurch erklären, dass die Aktivität in dieser Addierer-Schaltung wegen des kurzen Datenpfads und der maximalen Parallelität sehr hoch ist. Insbesondere wenn die Frequenz während einer Phase mit geringeren Performance-Anforderungen abgesenkt wird, erhöht sich der relative Anteil der Leckströme. Zudem machen die Flip-Flops in dieser Implementierung einen sehr großen Anteil an der Gesamtschaltung aus.

Durch das Abschalten des Standby-Transistors lässt sich der Leckstrom jedoch auch hier auf 13 nA reduzieren. Da es sich um einen LL-LVT-Transistor mit einer relativ niedrigen Schwellenspannung handelt, ist das Minimum des Off-Stroms bei $V_{gs} = 0\,\mathrm{V}$ noch nicht erreicht. Der Leckstrom lässt sich auf bis zu 1.7 nA reduzieren, indem ein negatives Potential von 0.3 V an das Gate des Standby-Transistors angelegt wird (*Super Cut-off*). Alternativ kann die Schwellenspannung über Reverse-Biasing gesenkt werden, womit ein minimaler Off-Strom von 2.4 nA bei $V_{bs} = -0.5\,\mathrm{V}$ erzielt werden kann. Für beide Techniken sind jedoch Potentiale kleiner als V_{ss} erforderlich, deren Generierung eine zusätzliche Leistungsaufnahme des Gesamtsystems bedeutet. Es ist zu erwarten, dass dieselbe Leckstromreduzierung, die mit Hilfe der Super-Cut-off-Technik erreicht wird, sich auch durch Einsatz eines LL-Standby-Transistors erzielen ließe.

Zur Abschätzung der Verminderung der Schaltgeschwindigkeit durch den Standby-Transistor wird die Gate-Spannung des Standby-Transistors im aktiven Betrieb auf $V_{gs} = V_{dd} - 200\,\mathrm{mV}$ abgesenkt. Es kann jedoch keine Reduzierung der Performance festgestellt werden. Daraus kann geschlossen werden, dass das Absinken der Schaltgeschwindigkeit bei $V_{gs} = V_{dd}$ aufgrund der sicheren Dimensionierung des Standby-Transistors mit einer Weite von 350 μm weitgehend zu vernachlässigen ist.

Verglichen mit anderen Implementierungen von 32- und 64-bit-Addierern fällt auf, dass die maximale Betriebsfrequenz selbst des Skewed-CMOS-LVT-Addierers um einen Faktor drei kleiner ist als in [108]. Allerdings werden in [108] High-Performance-Transistoren mit hohen On-Strömen eingesetzt. Der Leckstrom des Addierers ist mit 4.9 mA mehr als zwei Dekaden höher als im Skewed-CMOS-LVT-Addierer, verglichen mit dem niedrigsten Leckstrom im Super-Cut-off-Modus sogar mehr als sechs Dekaden. Auch der in [30] beschriebene Addierer weist eine ähnliche Mikroarchitektur wie der hier beschriebene Addierer auf und erreicht ebenfalls wesentlich höhere Frequenzen. Allerdings werden keine Angaben über den Leckstrom gemacht, sodass ein Vergleich im Sinne des Trade-offs zwischen Leckstrom und Geschwindigkeit nicht möglich ist. Mit Hilfe einer für Mikroprozessoren optimierten CMOS-Technologie [30]

mit kürzeren Gatelängen, kleineren Schwellenspannungen und dünneren Oxiden lassen sich jedoch keine Low-Power-Schaltungen realisieren. Ebenso ist die hier eingesetzte System-on-Chip-Technologie nicht dazu geeignet, höchste Schaltgeschwindigkeiten über 4 GHz zu erzielen.

Alle internen Knoten einer Schaltung, die mittels eines zentralen NMOS-Transistors von der V_{ss}-Leitung abgetrennt sind, werden aufgrund von Leckströmen fast vollständig langsam auf V_{dd}-Potential aufgeladen. Wird der Standby-Transistor anschließend schlagartig geöffnet, so kommt es auf dem globalen V_{ss}-Netzwerk zu einem Spannungseinbruch, der zu einer fehlerhaften Funktion in anderen, noch aktiven Schaltungsblöcken führen kann. Zudem treten während der Einschaltphase zufällige Schaltereignisse, so genannte *Glitches* auf. Diese führen ebenfalls zu einem vorübergehenden Absinken der Versorgungsspannung. Die Reaktivierung aus dem Standby-Modus muss daher langsam erfolgen [28], oder es müssen Techniken zur Vermeidung von Glitches eingesetzt werden [26, 109, 110].

Im Gegensatz dazu kann der Mixed-t_{ox}-Addierer innerhalb eines Taktzyklusses aus dem Standby-Modus mit lokalen Standby-Transistoren reaktiviert werden. Im Vergleich zum Clock-Gating-Leckstrom (920 nA) reduzieren die lokalen Standby-Transistoren I_{off} auf 73 nA. Der Unterschied zum abgeschalteten LVT-Addierer (13 nA) resultiert daraus, dass die lokalen Standby-Transistoren nicht unter allen Gattern geteilt werden können und sie daher in der Summe eine größere Gesamtweite besitzen. Außerdem haben die lokalen NMOS-Standby-Transistoren minimale Gatelängen. Die Schwellenspannung ist daher wegen des Short-Channel-Effekts kleiner als bei 120 nm Länge (Abb. 3.5).

Wegen der hohen Gateleckströme erhöht sich der Leckstrom bei Anlegen des Minimum-Leakage-Vektors (hier eigentlich Maximum-Leakage-Vektor). Alle Dünnoxid-Transistoren sind eingeschaltet, deren On-Gateleckströme insbesondere bei Raumtemperatur einen hohen Anteil am gesamten Leckstrom ausmachen.

Je nach Anteil des Gateleckstroms am gesamten Off-Strom des Addierers ergibt sich eine Reduzierung des Leckstroms durch Reverse-Biasing. Bei Clock Gating und wenn der Anteil des Unterschwellenstroms groß ist, reduziert sich der Leckstrom um einen Faktor 3. Im MLV-Modus, bei dem der Gateleckstromanteil hoch ist, lässt sich durch Reverse-Biasing lediglich ein Faktor 1.1 erzielen. Dieses erklärt die Überkreuzung der Kennlinien für den Skewed-CMOS-Mixed-t_{ox}-Addierer beim Anlegen von Forward-Biasing in Abbildung 6.8.

Der untersuchte Parallel-Addierer stellt für viele Anwendungen eine geschwindigkeitskritische Schaltungskomponente dar. Jedoch auch andere Schaltungskomponenten, die auch im Standby-Modus aktiv sind, weisen selbst bei reduzierten Betriebsfrequenzen und -spannungen eine Leistungsaufnahme auf, die eine Leckstromreduzierung der inaktiven Blöcke um mehrere Dekaden nur bis zu einem bestimmten Punkt effizient macht.

Zur Reduzierung von Leckströmen in Sub-100 nm-Technologien müssen daher die vorgestellten Schaltungstechniken sinnvoll in die Gesamtschaltung integriert werden. Für die Zukunft scheinen aus heutiger Sicht der Einsatz von Standby-Transistoren und Clock-Gating in Kombination mit Multi-V_t- und Multi-t_{ox}-Techniken als besonders aussichtsreich.

Kapitel 7

Zusammenfassung

Mit Unterschreiten der minimalen Strukturgröße von 100 nm stellen sich für die Realisierung integrierter Digitalschaltungen neue Herausforderungen. Diese resultieren aus der endlichen Unterschwellenstromsteilheit und dem Auftreten von Tunnelströmen bei Oxiddicken von weniger als 2 nm. Selbst die zukünftige Einführung neuer Transistorkonzepte wie Multi-Gate-Transistoren oder alternativer Gate-Materialien und Dielektrika kann nicht verhindern, dass Sub-100 nm-CMOS-Technologien bezüglich der Skalierbarkeit bei konstantem Leckstrom zunehmend an physikalische Grenzen stoßen.
Dennoch wird sich auch in Zukunft der Trend hin zu größeren Integrationsdichten, höherer Schaltgeschwindigkeit und geringerer Leistungsaufnahme fortsetzen, wenn es gelingt, durch geeignete schaltungstechnische Maßnahmen die Schaltgeschwindigkeit zu erhöhen und höhere Transistor-Leckströme in der Gesamtschaltung zu unterdrücken.

In der vorliegenden Arbeit wird der Trade-off zwischen Leckstrom und Schaltgeschwindigkeit in digitalen CMOS-Schaltungen übergreifend vom Einzeltransistor, über elementare Logikgatter und Ringoszillatoren bis hin zu einem 32-bit-Addierer hardwarebasiert untersucht. Ziel ist es, die Übergänge zwischen den verschiedenen Abstraktionsebenen grundlegend zu verstehen und auf diese Weise eine ganzheitliche Optimierung zu ermöglichen.

Zunächst werden dazu die physikalischen Grundlagen von Leckströmen und Schaltgeschwindigkeiten aufgezeigt. Insbesondere werden die Gateleckströme, die Weitenabhängigkeit der Schwellenspannung sowie die Temperaturabhängigkeit verschiedener Effekte physikalisch basiert beschrieben. Dabei zeigt sich unter anderem, dass die in bisherigen Modellen nicht enthaltene Temperaturabhängigkeit des Tunnelstroms gerade vor dem Hintergrund einer abnehmenden Temperaturabhängigkeit des Unterschwellenstroms signifikant wird.
An der Schnittstelle zwischen Einzeltransistor und elementarem Logikgatter wird die Schaltgeschwindigkeit von Invertern sowie NAND- und NOR-Gattern untersucht. Es wird deutlich, dass insbesondere für die Schaltgeschwindigkeit von NAND- und NOR-Gattern der Drain-Strom bei reduzierten Gate-Source- und Drain-Source-Spannungen $V_{gs} < V_{dd}$ und $V_{ds} < V_{dd}$ entscheidend ist. Die Untersuchung der Temperaturabhängigkeiten von Schaltströmen und Schaltgeschwindigkeiten verdeutlicht, dass in Sub-100 nm-Technologien der dynamische Zero-Temperature-Coefficient-Punkt bei kleinen Versorgungsspannungen unterschritten

werden kann, wodurch sich die maximale Betriebsfrequenz entgegen dem bekannten Verhalten mit der Temperatur erhöht.

Anschließend wird die Effizienz von Body Biasing zur Erhöhung der Schaltgeschwindigkeit (Forward-Biasing) und zur Reduzierung des Leckstroms (Reverse-Biasing) für Einzeltransistoren, Ringoszillatoren und komplexere Schaltungen experimentell untersucht. Es ist festzustellen, dass die Einsatzfähigkeit von Reverse-Biasing auf wenige Spezialfälle begrenzt bleibt, da nur der Unterschwellenstrom, jedoch nicht GIDL und der Gateleckstrom reduziert werden.

Trotz kleiner werdenden Substrateffekts in Sub-100 nm-CMOS-Technologien stellt Forward-Biasing in aktuellen Low-Power-Technologien eine viel versprechende Möglichkeit dar, auch die Geschwindigkeit komplexer Schaltungsblöcke zu erhöhen. Die Performance einer Schaltung, in der Dickoxid-Transistoren eingesetzt werden, lässt sich um bis zu 37 % steigern. Aber auch in Schaltungen mit reduzierter Oxiddicke (1.6 nm) werden noch um 28 % höhere Frequenzen gemessen.

Bevor Forward-Biasing in einer Schaltung eingesetzt werden kann, ist noch das Latchup-Verhalten zu untersuchen. Je nach Spezifikation des Temperaturbereichs muss die maximale Spannung möglicherweise auf 0.3 oder 0.4 V begrenzt werden. Außerdem ist zu untersuchen, wie schnell die Schaltung in den Forward-Biasing-Zustand versetzt werden, ohne logische Zustände in der Schaltung zu verändern.

Im Kapitel 5 werden neue Schaltungstechniken präsentiert, die die Möglichkeit ausnutzen, Transistoren unterschiedlicher Oxiddicke und Schwellenspannung in einer Schaltung zu kombinieren. In der statischen Multi-V_t-Multi-t_{ox}-Logik werden Transistoren unterschiedlicher Schwellenspannung und Oxiddicke so miteinander kombiniert, dass im Standby-Modus sowohl Gateleckströme als auch Unterschwellenströme reduziert werden. Die Effizienz dieser Technik wird exemplarisch an Ringoszillatoren demonstriert.

Eine weitere Möglichkeit zur Erhöhung der Schaltgeschwindigkeit stellt die Skewed-CMOS-Logik dar [83]. Im Vergleich zu statischen CMOS-Gattern können 45 % höhere Geschwindigkeiten erzielt werden. Darüber hinaus kann der Leckstrom im Standby-Modus mit Hilfe der vorgestellten leckstromoptimierten Skewed-CMOS-Logik oder durch lokale Standby-Transistoren um teilweise mehr als zwei Dekaden reduziert werden. Die Schaltungen können jeweils innerhalb eines Taktzyklusses wieder in den aktiven Modus versetzt werden kann, während die Reaktivierung eines Schaltungsblocks mit zentralem Standby-Transistor mehrere Mikrosekunden dauert.

Der Reduzierung von Schaltverzögerung und Leckstrom steht, bedingt durch das monotone Schaltverhalten, ein erhöhter Entwurfsaufwand entgegen. Dieser ist jedoch vertretbar, wenn die Skewed-CMOS-Logik nur in geschwindigkeitskritischen Blöcken eingesetzt wird. Im Vergleich zu der Domino-Logik kann die Störsicherheit gegen Leckströme, Parametervariationen und Störungen durch kapazitive Kopplung (Crosstalk) experimentell verifiziert deutlich erhöht werden. Dennoch muss vor allem in einer System-on-Chip-Umgebung auf eine ausreichende Störsicherheit des Schaltungsblocks geachtet werden.

Neben den Logik-Gattern sind getaktete Speicherelemente wie Flip-Flops und Latches ein elementarer Bestandteil jeder digitalen CMOS-Schaltung mit Pipeline-Architektur. Insbesondere Sense-Amplifier-Flip-Flops (SAFF) eignen sich für den Einsatz in Skewed-CMOS-Schaltungen [99]. Das vorgestellte optimierte SAFF weist im Vergleich zu [89] bei kleinerer Verzögerungszeit eine geringere aktive Leistungsaufnahme auf und eignet sich zudem auch für den Einsatz in statischen CMOS-Schaltungen.

Das außerdem präsentierte Skewed-CMOS-Latch ist speziell auf das monotone Schaltverhalten dieser Logikfamilie abgestimmt und vermindert die Anfälligkeit der Schaltung gegen ungleichmäßige Taktsignalversorgung.

Abschließend wird ein 32-bit-Parallel-Addierer mit einer für CMOS-Logik optimierten Mikroarchitektur vorgestellt. Die verwendete Mikroarchitektur stellt im Hinblick auf die Anzahl der Operatoren sowie auf die Verdrahtungsdichte und -länge eine Verbesserung gegenüber der weit verbreiteten Han-Carlson-Architektur dar.

Sowohl statische CMOS- als auch Skewed-CMOS-Addierervarianten wurden in einer 90 nm-CMOS-Technologie hergestellt. Sie demonstrieren die Fähigkeiten der verschiedener Schaltungstechniken, den Trade-off zwischen Leckstrom und Geschwindigkeit zu verbessern.

Aufgrund der verbesserten Mikroarchitektur und der sorgfältigen Implementierung erzielt bereits ein Addierer in statischer CMOS-Logik unter Verwendung leckstromarmer Transistoren eine Betriebsfrequenz von 510 MHz ($V_{dd} = 1.2$ V). Der Leckstrom der gesamten Addiererschaltung beträgt 12 nA. Drei weitere Addierer-Implementierungen in Skewed-CMOS-Logik setzen die vorgestellten Schaltungstechniken sowie Flip-Flops und Latches effizient ein, um die Schaltgeschwindigkeit weiter zu erhöhen, den Leckstrom klein zu halten sowie den Flächenbedarf und die Leistungsaufnahme zu minimieren. Bei Einsatz schneller Transistoren mit niedriger Schwellenspannung (Low-V_t, LVT) wird eine Taktfrequenz von 1.73 GHz erreicht, die durch V_{dd}-Erhöhung auf 1.5 V und Forward-Biasing auf 2.5 GHz gesteigert werden kann.

Da der Leckstrom im Vergleich zum leckstromarmen Addierer um drei Dekaden höher ist, müssen Maßnahmen zur Leckstromreduzierung ergriffen werden. Die Leckströme sowohl des LVT-Addierers mit zentralem Standby-Transistor als auch die des Skewed-CMOS-Addierers mit lokalen Standby-Transistoren können so ebenfalls auf weniger als 100 nA reduziert werden.

Es wird erkennbar, wie dem Trade-off zwischen niedriger Leistungsaufnahme und hoher Schaltgeschwindigkeit auf Systemebene durch geeignete schaltungstechnische Maßnahmen begegnet werden kann. Auch in Zukunft lassen sich so immer höher integrierte und schnellere Schaltungen mit geringerer Leistungsaufnahme realisieren.

Anhang A

Messergebnisse der Crosstalk-Teststruktur

Im Abschnitt 5.1.3 wird auf die erhöhte Störanfälligkeit von Skewed-CMOS-Schaltungen hingewiesen. Anhand einer Teststruktur, die ein Crosstalk-Ereignis in einen statischen Zustand umwandelt (Abb. A.1), soll die Crosstalk-Anfälligkeit von statischen und Skewed-CMOS-Schaltungen verglichen werden. Das Aggressor-Signal wird von einem Pulsgenerator erzeugt und über zwei Inverter an eine Metall-Leitung (Signal 1 in Abb. A.1) angelegt, an deren Ausgang eine Last geschaltet ist. Der Spannungshub des Aggressor-Inverter V_{agg} kann unabhängig von der übrigen Schaltung eingestellt werden. Die Aggressor-Leitung, die unterschiedliche Formen und Längen L_{agg} besitzen kann, ist kapazitiv über das Inter-Metall-Dielektrikum an eine Victim-Leitung gekoppelt und ruft dort eine vorübergehende Absenkung des Potentials hervor (Signal 2).

Die Höhe des Störimpulses auf der Victim-Leitung wird durch mehrere Faktoren bestimmt. Geringere Transistorweiten in Verbindung mit Serienschaltungen im Eingangsgatter *IN* erhöhen dessen Größe, parasitäre Kapazitäten im Eingangs- und Ausgangsgatter *IN* und *OUT* reduzieren sie. Darüber hinaus hat die Steilheit der Aggressorflanke Auswirkungen auf die Form des Störimpulses.

Abhängig von der Dimensionierung und der logischen Funktionalität des Ausgangsgatters *OUT* wird der Impuls bei der Propagation zum Signal 3 größer oder kleiner. Ziel des nachfolgenden Schaltkreises ist es, die Größe dieses Impulses zu bewerten und ab einer bestimmten Länge und Höhe als statisches digitales Signal zu speichern. Dazu vergrößern 16 asymmetrische Inverter mit wechselweise großen NMOS- und PMOS-Transistoren den Impuls. Am Ausgang des Puls-Streckers ist der Impuls (Signal 5) ausreichend lang, um als Takt-Signal ein Flip-Flop anzusteuern. Dieses Flip-Flop, das zuvor über einen Reset-Impuls auf 0 gesetzt wurde, übernimmt dadurch eine fest am Eingang anliegende 1.

Der Aggressor wirkt in der Testschaltung auf alle 128 Strukturen gleichzeitig. Nach Anlegen einer einzelnen Aggressor-Flanke können alle Flip-Flops der Reihe nach zeitlich unkritisch über einen 128:1-Multiplexer ausgelesen werden.

Auf dem Testchip sind sowohl Strukturen vorhanden, die auf einen 1–0-Übergang, als auch Strukturen, die auf einen 0–1-Übergang sensitiv reagieren. Hierzu muss der Puls-Strecker in die andere Richtung arbeiten als in Abbildung A.1 dargestellt. Allerdings zeigen die Teststrukturen, die auf den 0–1-Übergang reagieren sollten, selbst bei hohem V_{agg} und niedrigem V_{vic}

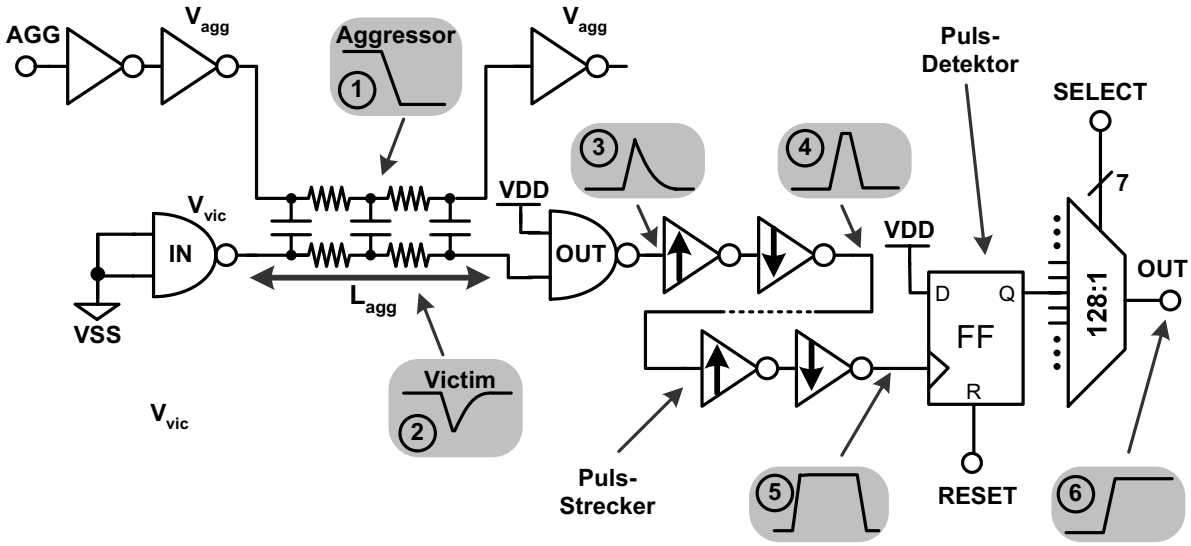


Abbildung A.1: Testschaltung zur Detektion von Crosstalk.

kein Crosstalk-Ereignis an. Dieses Verhalten lässt sich durch den hohen Leitwert der leitenden NMOS-Transistoren im Gatter *IN* im linearen Betrieb erklären.

Eine Teststruktur wird für ein bestimmtes V_{vic} sensitiver, je höher V_{agg} ist. In einem relativ schmalen Übergangsbereich zeigt sich eine statistische Verteilung. Bei konstantem V_{vic} wird V_{agg} in 50 mV-Schritten erhöht und jeweils zehnmal gemessen (Abb. A.2). Die Erhöhung der Aggressor-Spannung ist in diesem Zusammenhang als Testmethodik zu verstehen, um das Crosstalk-Verhalten der Gatter untersuchen zu können. Die Aggressor-Spannung, bei der die Crosstalk-Wahrscheinlichkeit 50 % beträgt, stellt ein Maß für die Sensitivität der untersuchten Gatter dar. In den Abbildungen A.3 bis A.8 wird nur noch der 50 %-Wert der Crosstalk-Wahrscheinlichkeit als Schwelle zwischen Crosstalk und Nicht-Crosstalk angegeben.

Jedoch auch in einer realen Schaltung kann es vorkommen, dass Signale mit einem größeren Spannungshub über einen Schaltungsblock mit reduzierter Versorgungsspannung geleitet werden. Darüber hinaus ist zu beachten, dass auch schon bei Aggressor-Spannungen, die noch nicht zu einem vollständigen Umwerfen des logischen Zustands führen, eine signifikante Erhöhung der Schaltzeiten auftreten kann [86].

Die Aggressor-Längen L_{agg} variieren zwischen 10 und 200 μm. Die 10 μm-Strukturen zeigen jedoch in keinem Fall ein Crosstalk-Ereignis, da die Koppel-Kapazität C_c hier im Verhältnis zu den übrigen parasitären Leitungskapazitäten klein ist.

Die Teststruktur eignet sich auch zur Bewertung alternativer Dielektrika (Low-κ-Dielektrika), da sie Ergebnisse liefert, die von der Schwellenspannung der verwendeten Gatter unabhängig sind (Abb. A.6).

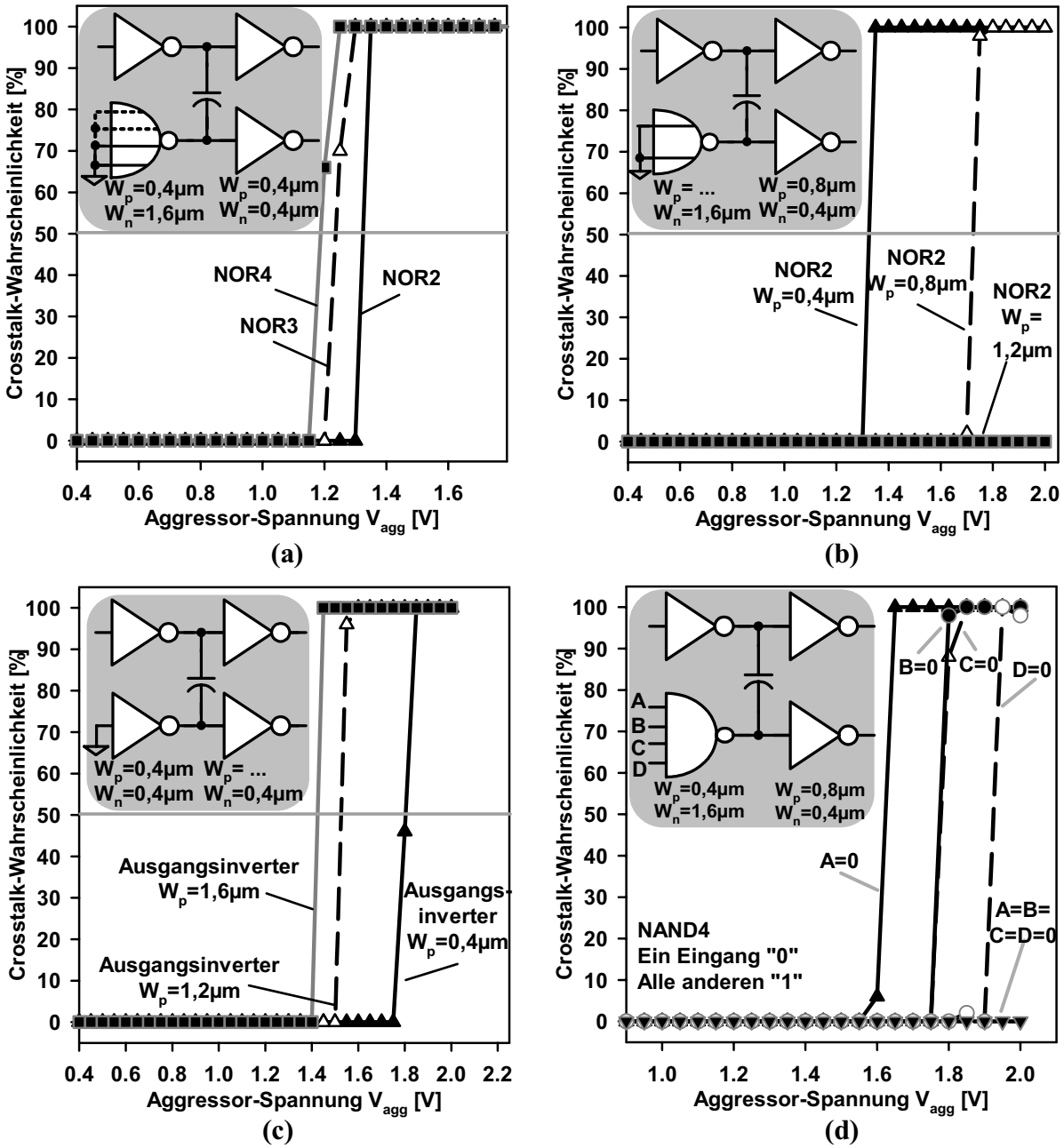


Abbildung A.2: Gemessene Häufigkeiten von Crosstalk, $L_{agg} = 40\,\mu$m.

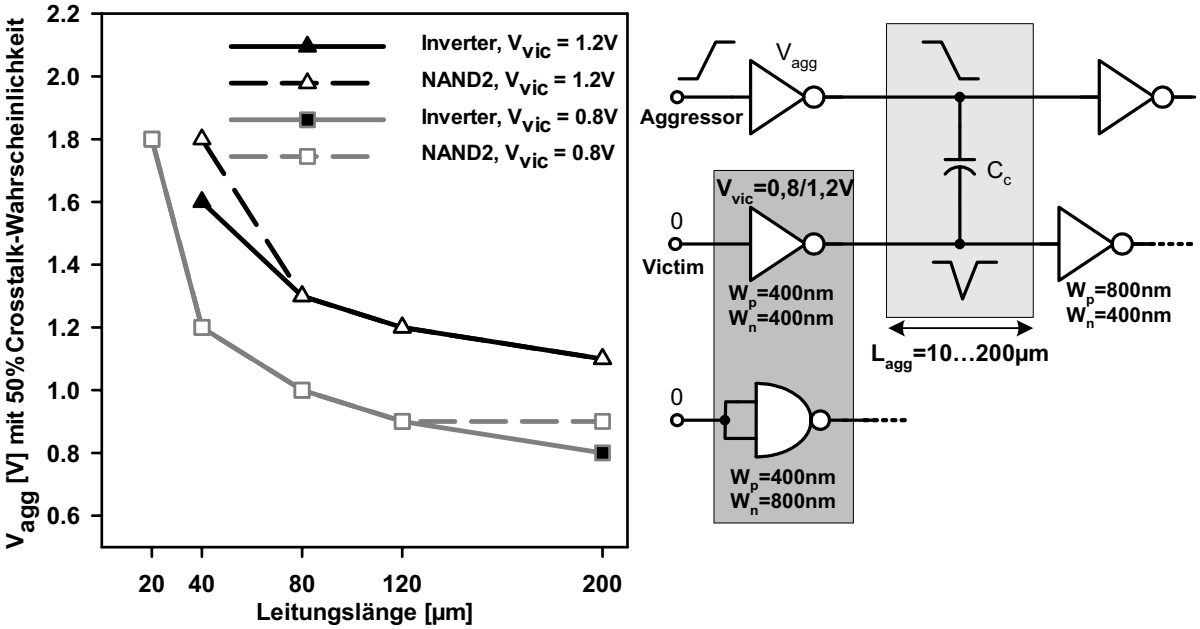


Abbildung A.3: Aggressor-Spannung V_{agg}, bei der eine 50 %-ige Wahrscheinlichkeit für Crosstalk besteht, in Abhängigkeit von der Leitungslänge L_{agg} und von der logischen Funktion des Eingangsgatters. Da die leitenden NMOS-Transistoren in der NAND-Serienschaltung doppelt so weit sind, zeigen NAND2-Gatter und Inverter sowohl bei $V_{vic} = 0.8\,\mathrm{V}$ als auch bei 1.2 V ein annähernd gleiches Verhalten. Die zusätzlichen parasitären Kapazitäten im NAND-Gatter wirken sich nun sehr schwach aus.

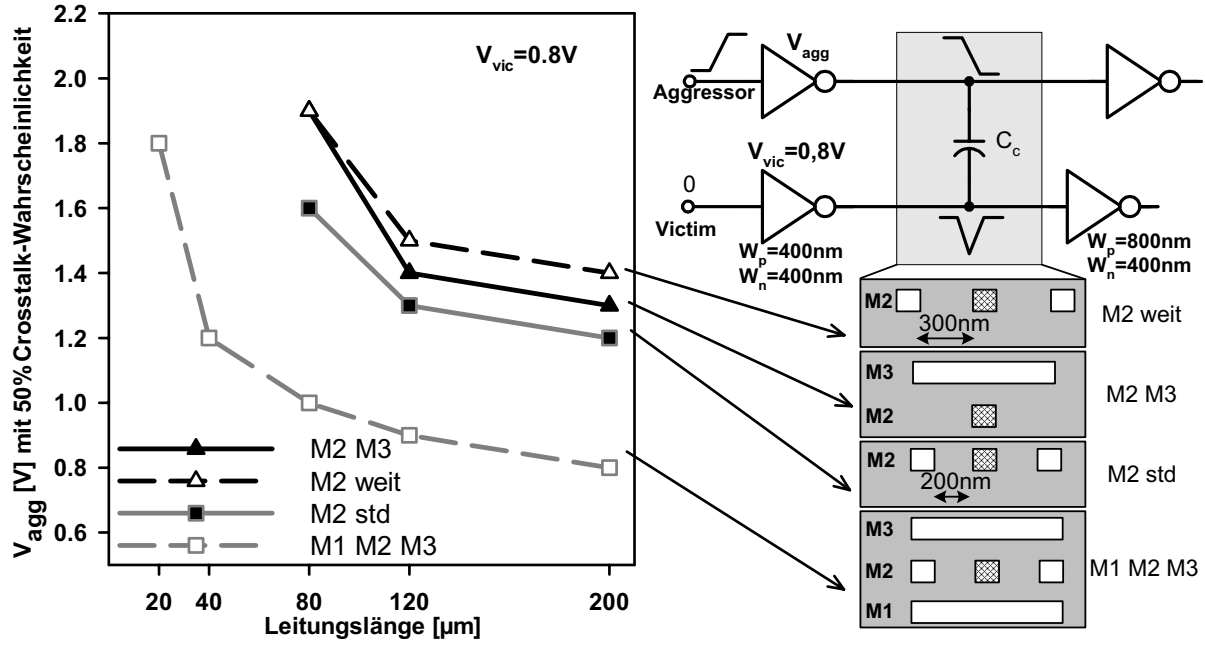


Abbildung A.4: Crosstalk bei Variation der Aggressor-Form. Die Stärke der Kopplung zwischen Aggressor und Victim hängt von der Form der Leitung ab. In der Standard-Konfiguration M1 M2 M3, die auch bei allen anderen Strukturen verwendet wird, ergibt sich die mit Abstand höchste Empfindlichkeit. Bei minimalem Abstand zwischen zwei Metall-2-Leitungen ist die Empfindlichkeit nur geringfügig höher als bei Inter-Metall-Crosstalk (M2 M3).

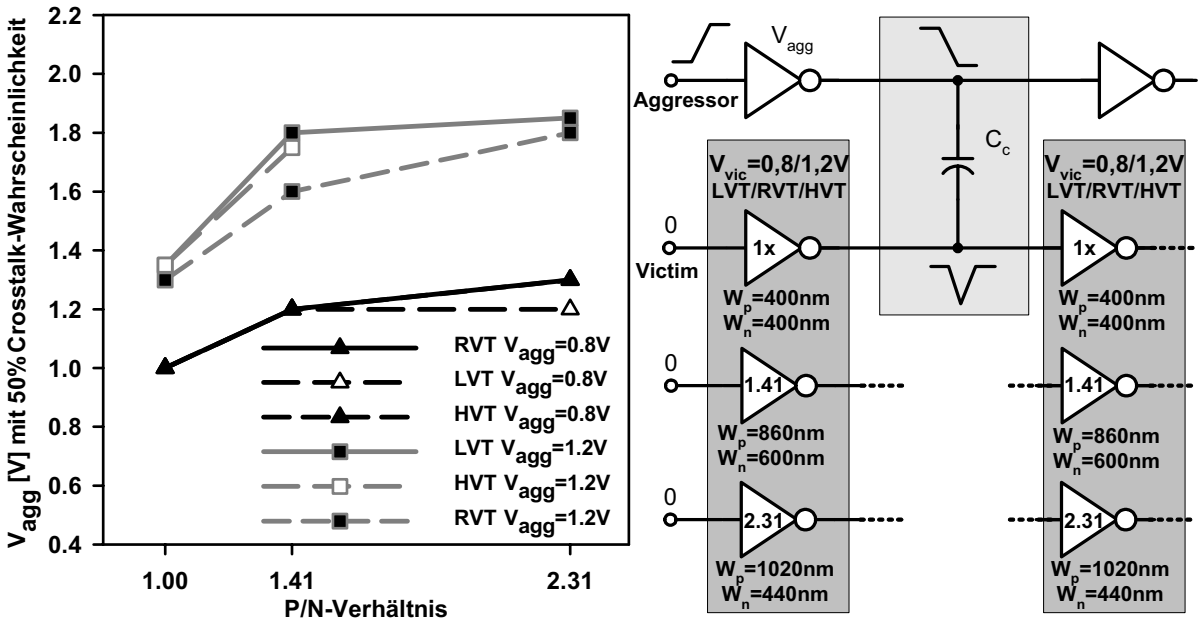


Abbildung A.5: Variation des P/N-Verhältnisses in der Victim-Leitung. Die geschwindigkeitsoptimierte Skalierung zeigt, verglichen mit der symmetrischen Skalierung (P/N-Verhältnis 2.31x und 1.41x), ein nur geringfügig schlechteres Verhalten. Hingegen nimmt die Empfindlichkeit bei der 1x-Skalierung deutlich zu. Außerdem ist zu erkennen, dass die Schwellenspannung keinen Einfluss hat.

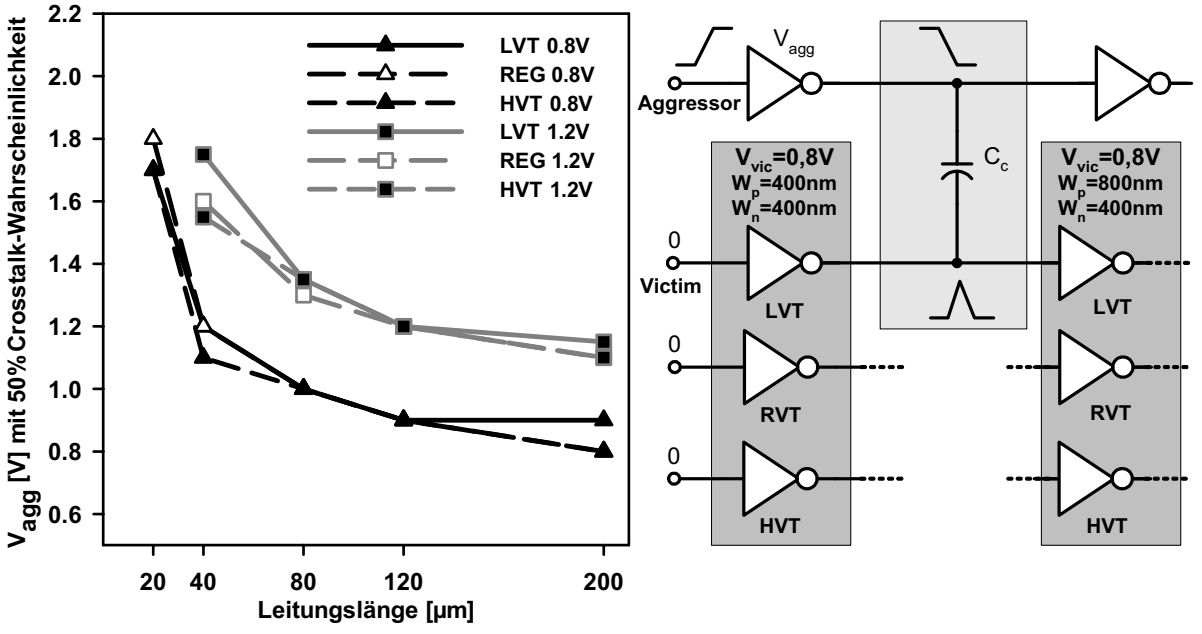


Abbildung A.6: Crosstalk bei unterschiedlichen Schwellenspannungen. Wie schon in Abbildung A.5 zu erkennen war, ist Crosstalk nicht von der in der Victim-Leitung verwendeten Schwellenspannung abhängig. Allerdings stammt hier das Aggressor-Signal immer von einem REG-Inverter. In einer realen Schaltung würde das Aggressor-Signal bei Verwendung eines LVT- bzw. HVT-Inverters steiler bzw. weniger steil sein (vergleiche Abb. A.7).

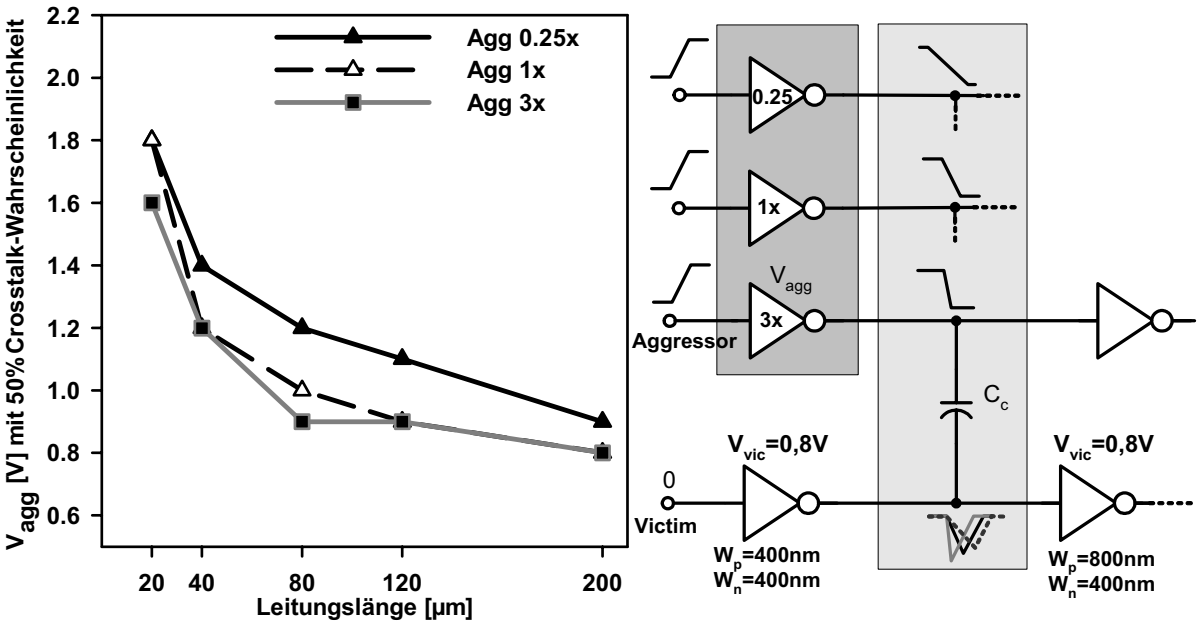


Abbildung A.7: Variation der Steilheit des Aggressorsignals durch Verwendung unterschiedlicher Treiberstärken. Vom schwächsten bis zum mittleren Aggressor-Treiber verstärkt sich der Crosstalk signifikant. Danach tritt eine Sättigung ein. Das bedeutet, dass der Zustand der Victim-Leitung bei schnellen Störsignalen nur sehr schwach durch den (sehr kleinen) leitenden Transistor im Eingangsgatter IN stabilisiert wird.

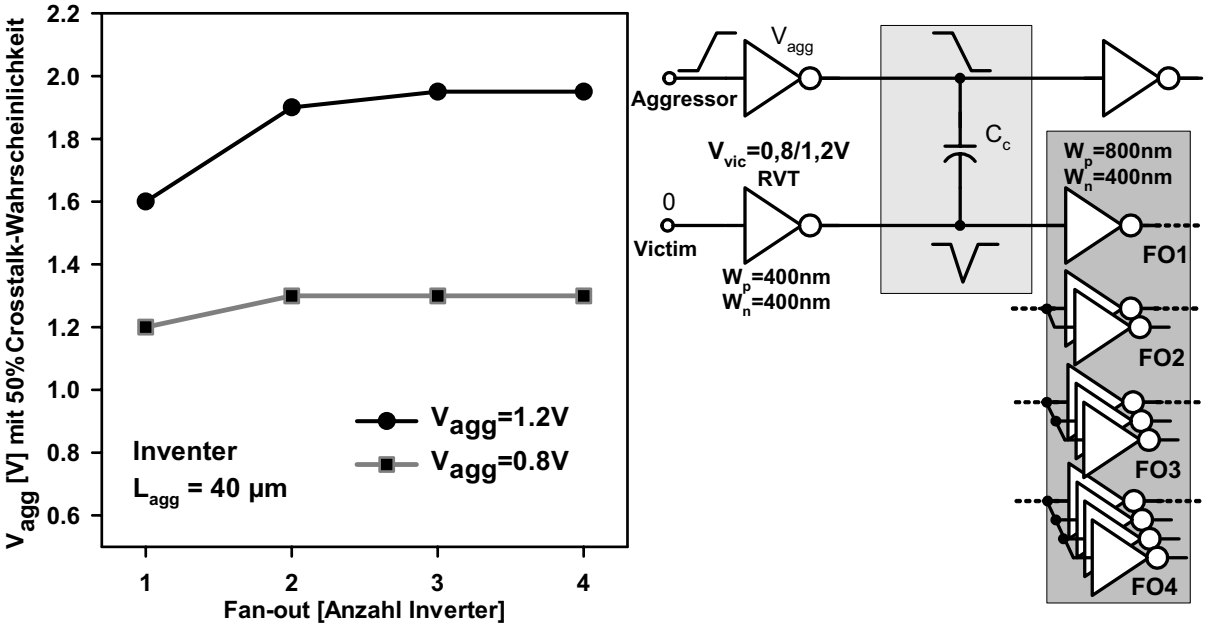


Abbildung A.8: Crosstalk in Abhängigkeit vom Fan-out der Victim-Leitung. Zunächst nimmt die Empfindlichkeit von FO1 zu FO2 wie erwartet wegen der höheren, stabilisierenden Kapazität ab. Bei FO3 und FO4 tritt eine Sättigung ein.

Anhang B

Ringoszillatoren in statischer Multi-V_t-Logik

Auf einem 130 nm-Testchip wurden verschiedene Standard- und Multi-V_t-Ringoszillatoren implementiert. Multi-t_{ox}-Gatter werden nicht eingesetzt, da in der verwendeten 130 nm-Technologie nur eine Oxiddicke für Core-Devices verfügbar ist. Mit Hilfe einer Auswahllogik lassen sich mehrere Ringoszillatoren in einem Modul zusammenfassen. Auf diese Weise wird der Flächenbedarf klein gehalten und die Messung vereinfacht. Außerdem vermindert sich der Abstand der verschiedenen Testschaltungen, wodurch die Auswirkungen von räumlichen Parameterschwankungen auf ein Minimum reduziert werden.

Ein Nachteil der Auswahllogik ist, dass der Leckstrom eines Ringoszillators nicht unabhängig von den anderen Teststrukturen gemessen werden kann, wenn deren Leckströme nicht durch einen sehr leckstromarmen Standby-Transistor, z.B. einen HVT-Device mit geringer Weite, abgeschaltet werden. Der Spannungsabfall über dem Standby-Transistor würde jedoch im aktiven Betrieb das Ergebnis verfälschen. Aus diesem Grund wird jede der 16 zu untersuchenden Testschaltungen jeweils einmal zur Bestimmung der Schaltgeschwindigkeit ohne Standby-Transistor und einmal zur Leckstrom-Messung mit Standby-Transistor in dem Testmodul implementiert.

Abbildung B.1 zeigt ein Blockschaltbild der Schaltung. Insgesamt 32 Flip-Flops bilden am Eingang ein Schieberegister, dessen erstes Flip-Flop mit Hilfe des *Reset*-Signals auf 1 gesetzt wird, während alle anderen Flip-Flops auf 0 gesetzt werden. Damit ist der unterste aktive Ringoszillator aktiviert. Dessen Frequenz wird zunächst mit einem zweistufigen Frequenzteiler um den Faktor 4 reduziert, um die Frequenz auf der nachfolgenden Bus-Leitung nicht zu groß werden zu lassen. Das Signal wird über einen Tri-State-Buffer zu einem zentralen zehnstufigen Frequenzteiler geführt, der die Frequenz am Ende auf 1/4096 der Ringoszillator-Frequenz teilt und verstärkt an den Ausgang weiterleitet. Die positive Flanke des *Shift*-Signals schiebt die im ersten Flip-Flop gespeicherte 1 um eine Struktur nach oben. Auf diese Weise lassen sich der Reihe nach die Frequenzen aller 16 aktiven Strukturen messen. Die aktive Leistungsaufnahme kann an *VDD-ACTIVE* gemessen werden, da immer nur ein Ringoszillator aktiv ist.

Nach dem letzten aktiven Ringoszillator wird die erste Teststruktur zur Messung des Leckstroms aktiviert. Die Logik-Gatter in der ersten Leckstrom-Struktur entsprechen denen

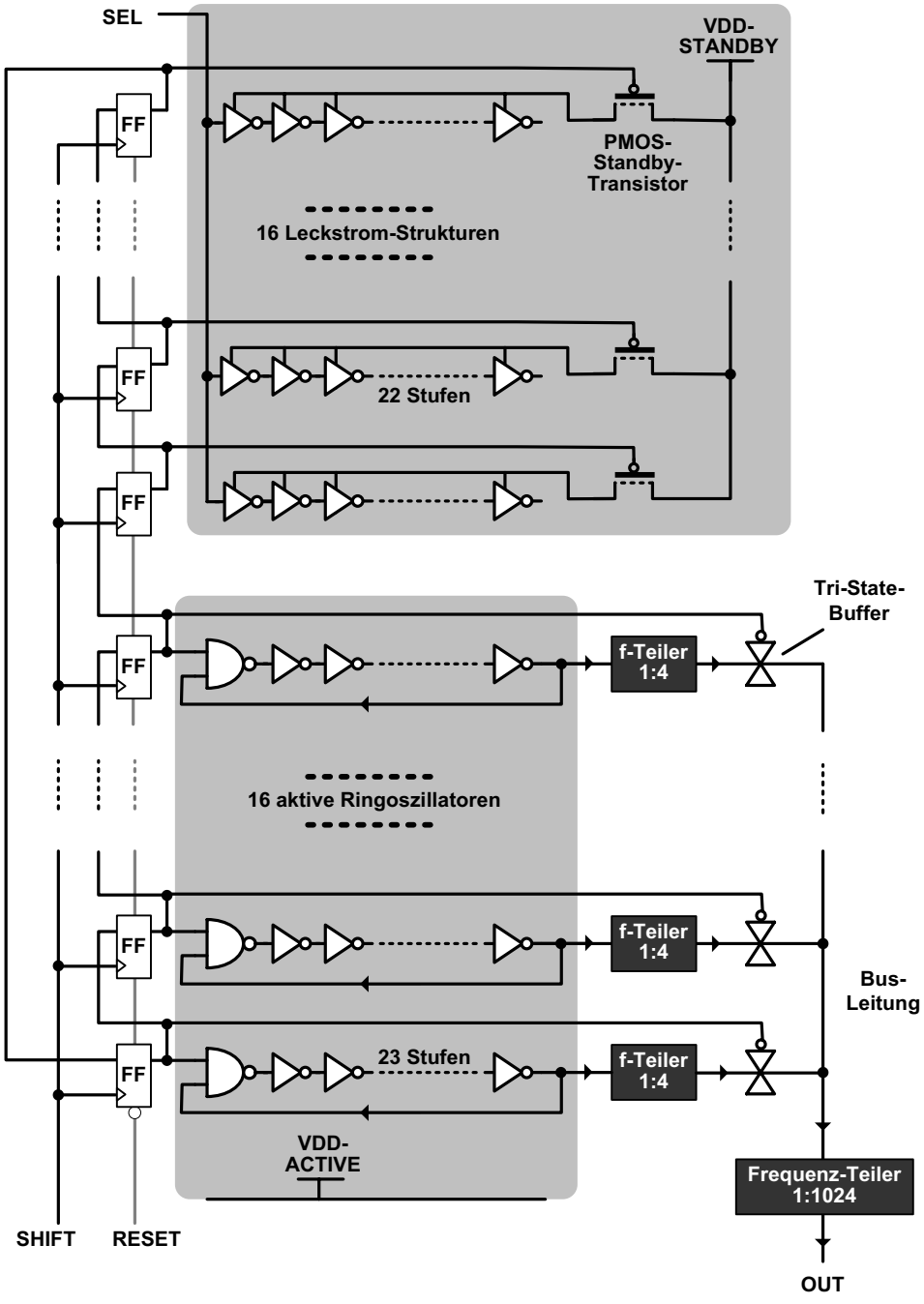


Abbildung B.1: Blockschaltbild der Ringoszillator-Auswahlschaltung.

der ersten aktiven Struktur. Der Eingang des jeweils ersten Gatters kann mit Hilfe von *SEL* auf 1 oder 0 gesetzt werden, da sich die Leckströme statischer Multi-V_t-Gatter je nach Eingangssignal unterscheiden. *SEL*=0 entspricht dem Anlegen des Minimum-Leakage-Vektors (MLV) in einer realen Schaltung. Nur die jeweils aktive Leckstrom-Struktur wird an *VDD-STANDBY* angeschlossen, um den Leckstrom zu messen. Es kommt ein PMOS-Standby-Transistor zum Einsatz, da in der verwendeten 130 nm-Technologie keine Triple-Well-Option zur Verfügung steht und somit ein NMOS-Standby-Transistor nicht auf einfache Weise eingesetzt werden kann.

Nach Erreichen der letzten Leckstrom-Struktur schaltet das nächste *SHIFT*-Signal wieder zurück zum ersten Ringoszillator und die Messung kann mit einer anderen Versorgungsspannung wiederholt werden. Alle Peripherie-Schaltungen werden mit einer dritten Versorgungsspannung betrieben, um die Messung der Stromaufnahme nicht zu stören.

Allerdings wird mit der beschriebenen Teststruktur nicht die eigentlich relevante Verzögerungszeit nach Gleichung 5.1 gemessen, sondern die durchschnittliche Verzögerungszeit $t_{np,avg} = 1/4 \cdot (t_n^{HVT} + t_p^{HVT} + t_n^{LVT} + t_p^{LVT})$. Zur korrekten Messung von t_{np} müssten spezielle Ringoszillatoren oder Delay-Lines eingesetzt werden, deren Messung aufwendiger ist. Da die Gatter jedoch mit Hilfe eines Algorithmus auf $t_{np} = t_{np,avg}$ optimiert wurden, kann die gemessene Verzögerungszeit im Rahmen der Simulationsgenauigkeit als gute Näherung von t_{np} angesehen werden.

Abbildung B.2 zeigt verschiedene Realisierungen von Multi-V_t-Gattern mit LVT- und HVT-Transistoren, kombiniert mit variablen Gatelängen. Alle Ringoszillatoren in statischer Multi-V_t-Logik haben einen Fan-out von 2. Jedes Gatter besitzt abhängig vom Eingangssignal zwei verschiedene Leckstrom-Werte. Lediglich die als Referenz dienenden statischen CMOS-Gatter sind davon unabhängig.

Abbildung B.3 vergleicht klassische statische CMOS-Schaltungen bei unterschiedlichen Schwellenspannungen und Variationen der Gatelängen. Zwar lässt sich die Schaltgeschwindigkeit durch die Reduzierung von L_g signifikant erhöhen, jedoch nur auf Kosten eines erhöhten Leckstroms. Alle Punkte in Abbildung B.3 liegen auf der Trendgeraden.

Im Unterschied dazu erreichen statische Multi-V_t-Logik-Gatter Punkte unterhalb der Trendgeraden. Abhängig vom Eingangssignal der Inverterkette *SEL* gibt es in diesen Gattern zwei unterschiedliche Leckstrom-Zustände, wobei sich für *SEL*=1 der Leckstrom des LVT-Transistors einstellt. Der Vergleich der Strukturen 7 und 10 macht deutlich, dass eine Reduzierung der Gatelänge der LVT-Transistoren zu einer signifikanten Erhöhung der Schaltgeschwindigkeit führt. Im leckstromarmen Zustand erhöht sich der Off-Strom nicht, da dieser von den HVT-Transistoren bestimmt wird. Im Gegensatz dazu kann der Leckstrom durch Erhöhung der HVT-Gatelänge (Vergleich der Strukturen 9 und 10) nicht weiter abgesenkt werden. Dieses stimmt mit der Beobachtung in Abbildung B.3 überein, in der die Strukturen 13 und 14 keine Reduzierung des Leckstroms im Vergleich zu Struktur 6 zeigen.

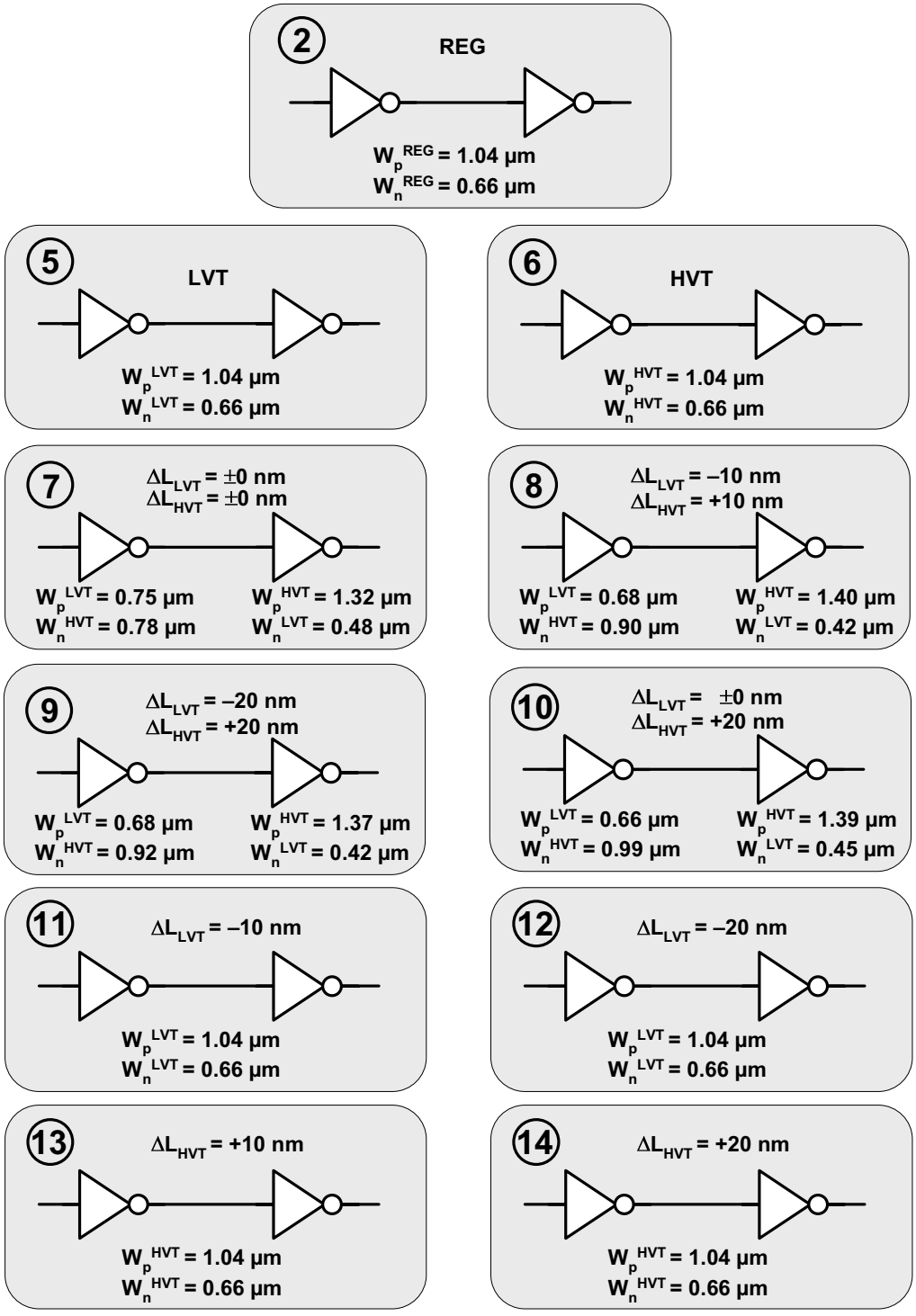


Abbildung B.2: Statische Multi-V_t-Gatter in Ringoszillatoren. Die Strukturen 2, 5 und 6 sind klassische CMOS-Schaltungen mit nominellen Gatelängen, die Strukturen 11–14 sind klassische CMOS-Schaltungen mit Variationen von L_g, die Strukturen 7–11 entsprechen statischen Multi-V_t-Schaltungen. Die nicht dargestellten Strukturen 1, 3 und 4 enthalten Fan-out-Variationen (vgl. Abb. 3.23). Die Strukturen 15 und 16 sind nicht funktional, da die Gatelänge hier zu kurz gewählt wurde.

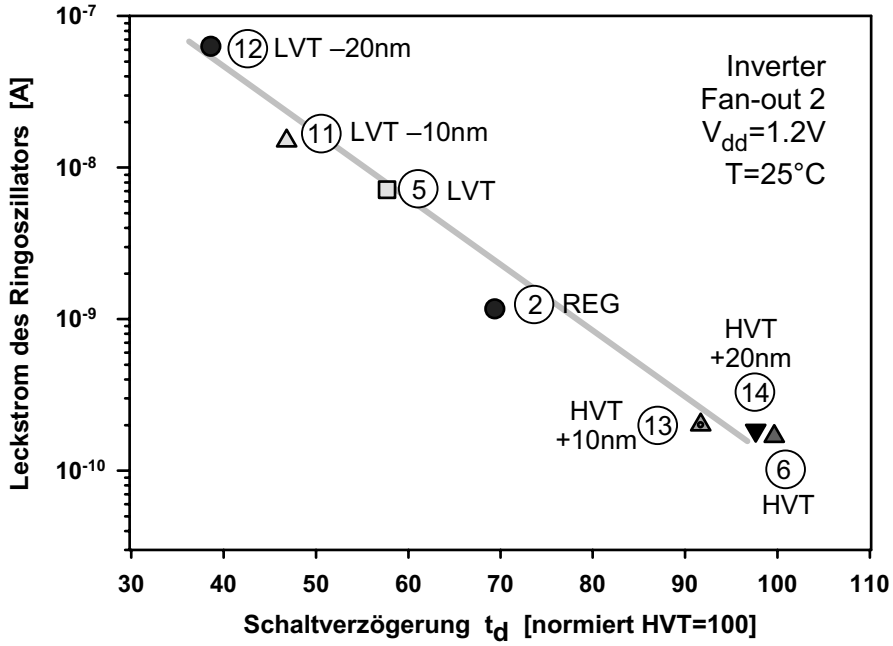


Abbildung B.3: Variation der Gatelängen in statischen CMOS-Schaltungen. Die Reduzierung der Gatelänge erhöht die Schaltgeschwindigkeit in LVT-ROs signifikant. Im Gegensatz dazu kann der Leckstrom von HVT-ROs durch größere Gatelängen nicht reduziert werden, da der Transistor auf die nominelle Gatelänge optimiert ist.

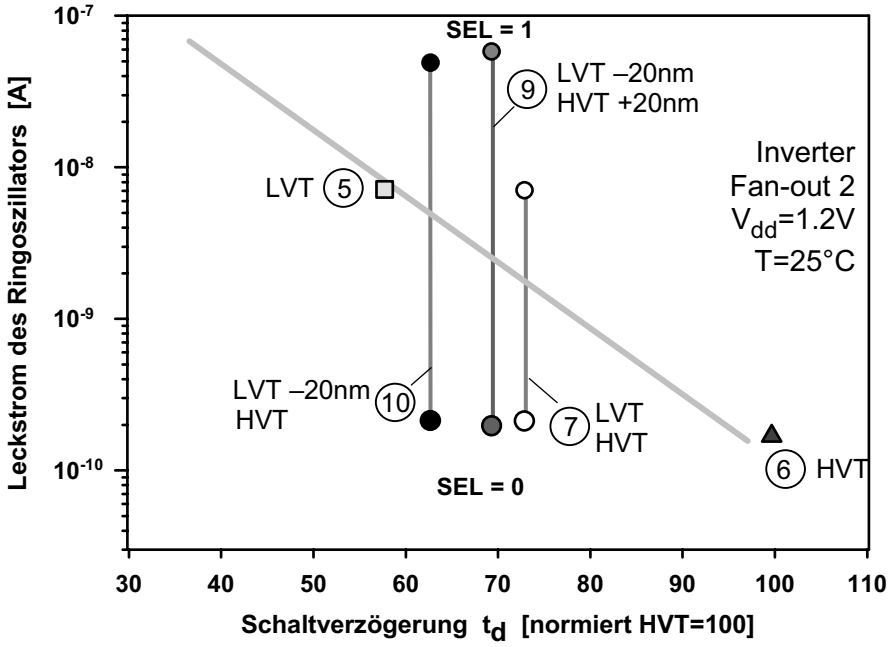


Abbildung B.4: Variation der Gatelängen in Multi-V_t-Schaltungen. Die Erhöhung der Schaltgeschwindigkeit zwischen den Strukturen 7 und 10 entspricht etwa der Hälfte der Verbesserung zwischen den Strukturen 5 und 12 in Abbildung B.3. Nicht dargestellt ist Struktur 8, die sich weitgehend identisch mit Struktur 7 verhält.

Anhang C

Transistordimensionierungen

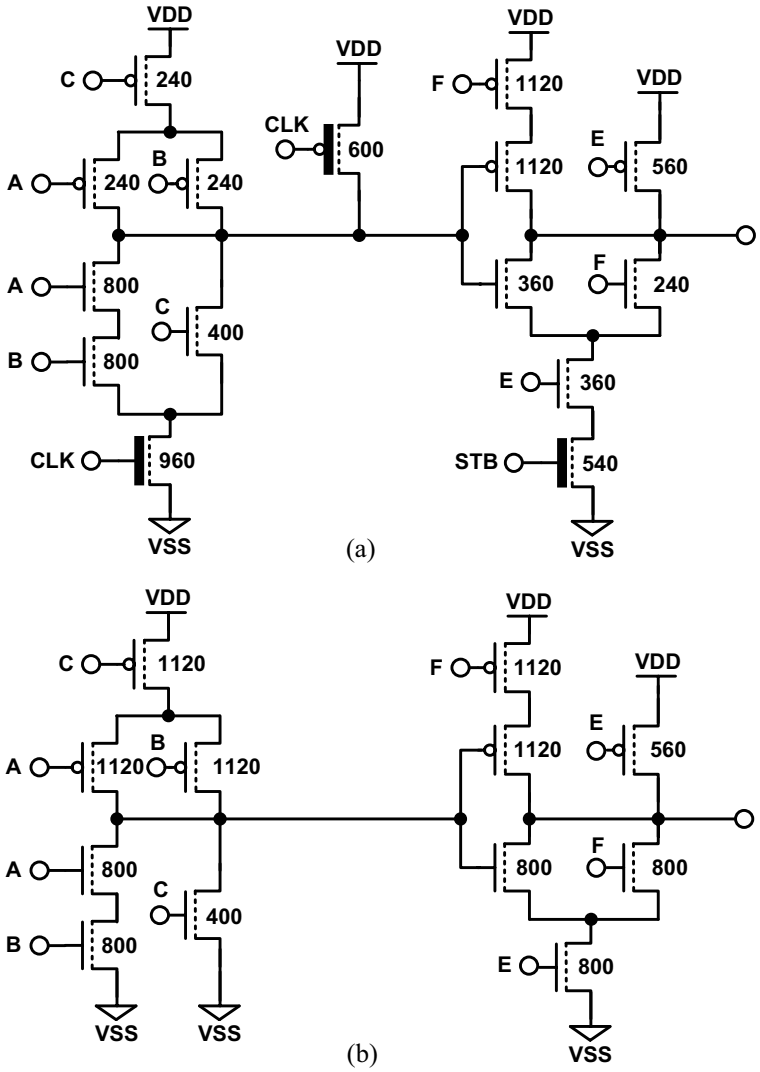


Abbildung C.1: Transistordimensionierungen (Gate-Weiten in nm) in Skewed-CMOS-Logik mit lokalen Standby-Transistoren (a) und statischer CMOS-Logik (b).

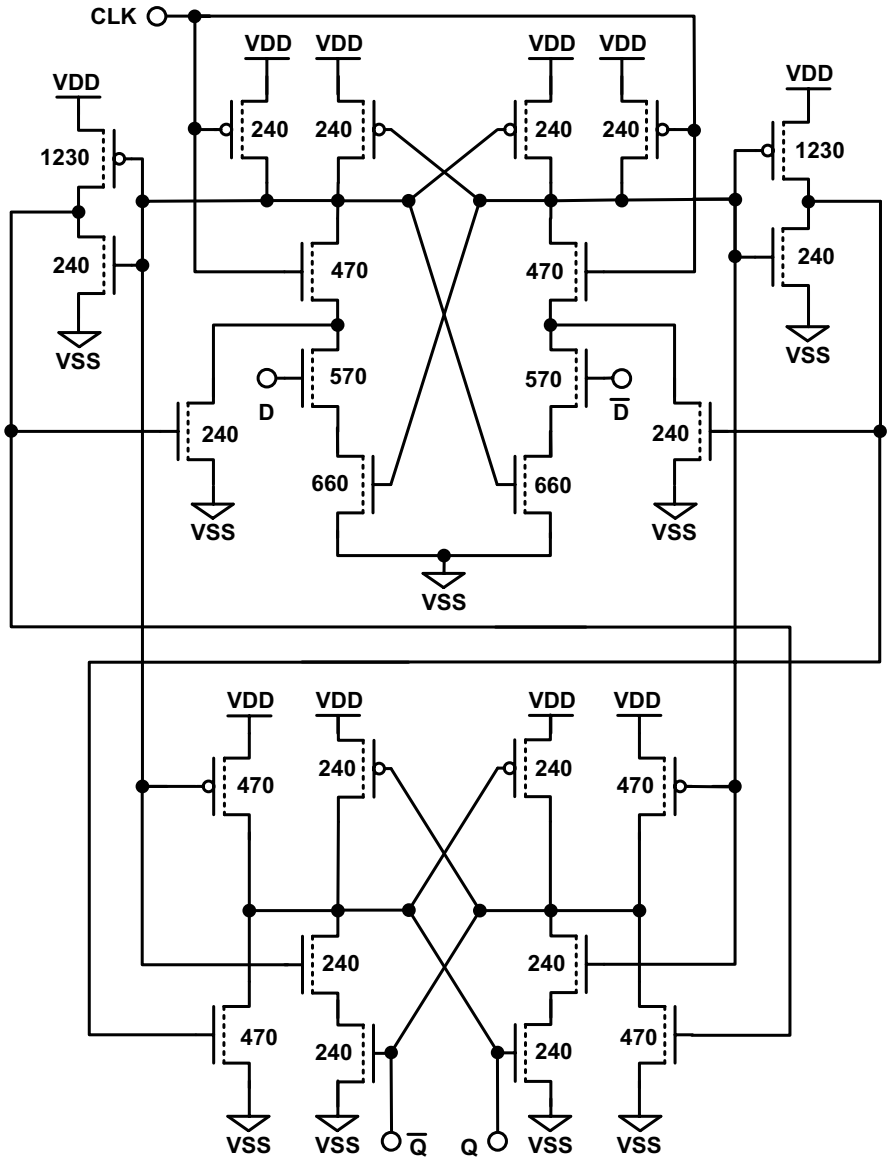


Abbildung C.2: Transistordimensionierungen im optimierten Sense-Amplifier-Flip-Flop (Gate-Weiten in nm, vgl. Abb. 5.17). Die minimale Transistorweite beträgt 240 nm. Zur Reduzierung der aktiven Leistungsaufnahme besitzen viele Transistoren geringe Weiten. In einer realen Schaltung sind die Weiten größer gewählt, um die Auswirkungen von Parametervariationen und zusätzlichen Verdrahtungskapazitäten zu reduzieren.

Formelzeichen und Abkürzungen

β_I	Verhältnis der On-Ströme von NMOS- und PMOS-Transistoren
β_t	Verhältnis der NMOS- und PMOS-Verzögerungszeiten
β_W	Weitenverhältnis von NMOS- und PMOS-Transistoren
C_g	Gate-Kapazität
C_{ox}	Oxid-Kapazität
CLK	Taktsignal (*Clock*)
CLK_{inv}	Taktsignal der zweiten Taktphase in Skewed-CMOS-Logik
DIBL	Drain-Induced Barrier Lowering
DPN	Decoupled Plasma Nitridation
DSP	Digitaler Signal-Prozessor
EDP	Energy Delay Product
EOT	Equivalent Oxide Thickness
f	Frequenz
FB	Forward-Biasing
FD-SOI	Fully-Depleted Silicon-on-Insulator
FF	Flip-Flop
FO	Fan-out
$G_{i,j}$	Generate-Operator
$G_{i,j}$	Propagate-Operator
GIDL	Gate-Induced Drain Leakage
h	Plank'sches Wirkungsquantum
H_{fin}	Finnenhöhe eines Multi-Gate-Transistors
HP	High-Performance, nach Definition der ITRS-Roadmap
HVT	High-V_t-Transistor der 130 nm oder 90 nm-Technologie
I_d	Drain-Strom
I_{eff}	Effektiver Schaltstrom eines Transistors in einer digitalen Schaltung
$I_{g,off}$	Gateleckstrom des ausgeschalteten Transistors ($V_{gs} = 0$ V, $V_{ds} = V_{dd}$)

$I_{g,on}$	Gateleckstrom des eingeschalteten Transistors ($V_{gs} = V_{dd}$, $V_{ds} = 50\,\mathrm{mV}$)
I_n	Effektiver Schaltstrom eines NMOS-Transistors
I_{off}	Source-Strom des ausgeschalteten Transistors ($V_{gs} = 0\,\mathrm{V}$, $V_{ds} = V_{dd}$)
I_{on}	On-Strom, Drain-Strom bei $V_{gs} = V_{ds} = V_{dd}$
I_p	Effektiver Schaltstrom eines PMOS-Transistors
$I_{s,off}$	Source-Strom des ausgeschalteten Transistors ($V_{gs} = 0\,\mathrm{V}$, $V_{ds} = V_{dd}$)
J_{tun}	Tunnelstromdichte
k_1	Temperaturkoeffizient der Schwellenspannung
L_c	Kanallänge
L_g	Physikalische Gatelänge
L_{nom}	Nominelle Gatelänge eines Transistors
L_{ov}	Effektive Länge der Überlappung zwischen Gate und Source/Drain
LL	Low-Leakage-Transistor der 90 nm-Technologie
LL-LVT	Low-Leakage-Low-V_t-Transistor der 90 nm-Technologie
LOCOS	Local Oxidation of Silicon
LOP	Low Operating Power, nach Definition der ITRS-Roadmap
LSTP	Low Standby Power, nach Definition der ITRS-Roadmap
LVT	Low-V_t-Transistor der 130 nm oder 90 nm-Technologie
m	Idealitätsfaktor der Unterschwellenstromsteilheit
m_n^*, m_p^*	Mittlere effektive Massen von Elektronen und Löchern im Silizium
μ_n,μ_p	Beweglichkeit von Elektronen und Löchern
MLV	Minimum Leakage Vector
MS-FF	Master-Slave-Flip-Flop
MuGFET	Multi-Gate-Transistor
N_A	Effektive Dotierung in der Kanalregion
n_i	Intrinsische Ladungsträgerdichte
NWE	Narrow-Width Effect
PD-SOI	Partially-Depleted Silicon-on-Insulator
q	Elementarladung
Q_b	Bulk-Ladung
RB	Reverse-Biasing
REG	Transistor der 130 nm oder 90 nm-Technologie mit regulärem V_t
RTN	Rapid Thermal Nitridation
S	Unterschwellenstromsteilheit
SAFF	Sense-Amplifier-Flip-Flop

SCE	Short-Channel Effect
SoC	System-on-Chip
STI	Shallow Trench Isolation
T	Temperatur
t_{act}	Benötigte Zeit für die Aktivierung einer Schaltung
$T(E)$	Transmissionswahrscheinlichkeit
t_n, t_p	NMOS- und PMOS-Verzögerungszeit eines CMOS-Gatters
$t_{ox,el}$	Elektrisch wirksame Oxiddicke
t_{ox}	Physikalische Oxiddicke
t_{setup}	Setup-Zeit in einem Flip-Flop
τ	Intrinsische Verzögerungszeit
U_T	Thermische Spannung
V_B	Substratspannung bei Body Biasing
V_{bs}	Bulk-Source-Spannung
V_{dd}	Versorgungsspannung
V_{ds}	Drain-Source-Spannung
V_{fb}	Flachbandspannung
V_{gc}	Spannung zwischen Gate-Elektrode und Kanalregion
V_{gs}	Gate-Source-Spannung
V_{in}	Eingangssignal eines CMOS-Gatters
V_M	Potential zwischen den Transistoren eines Stacks
V_{out}	Ausgangssignal eines CMOS-Gatters
V_{ox}	Oxid-Spannung
V_{ssv}	Virtuelle V_{ss}-Leitung bei Verwendung eines Standby-Transistors
V_t	Schwellenspannung
$V_{t,lin}$	Schwellenspannung bei $V_{ds} = 50\,\text{mV}$
$V_{t,sat}$	Schwellenspannung bei $V_{ds} = V_{dd}$
VCO	Voltage-Controlled Oscillator
W	Gate-Weite
W_{fin}	Finnenweite in einem Multi-Gate-Transistor
W_{min}	Minimale Transistorweite
$x_{d,max}$	Maximale Weite der Raumladungszone
ZTC	Zero Temperature Coefficient

Veröffentlichungen im Rahmen dieser Arbeit

VON ARNIM, K., E. BORINSKI, P. SEEGEBRECHT, H. FIEDLER, R. BREDERLOW, R. THEWES, J. BERTHOLD und C. PACHA: *Efficiency of body biasing in 90 nm CMOS for low power digital circuit*. In *Proceedings of the 30th European Solid-State Circuits Conference (ESSCIRC)*:175–178, Leuven, September 2004.

PACHA, C., M. BACH, K. VON ARNIM, R. BREDERLOW, D. SCHMITT-LANDSIEDEL, P. SEEGEBRECHT, J. BERTHOLD und R. THEWES: *Impact of STI-induced stress, inverse narrow width effect, and statistical VTH variations on leakage currents in 120 nm CMOS*. In *Proceedings of the 34th European Solid-State Device Research Conference (ESSDERC)*:397–440, Leuven, September 2004.

VON ARNIM, K., P. SEEGEBRECHT, R. THEWES und C. PACHA: *A low-leakage 2.5GHz skewed CMOS 32b adder for nanometer CMOS technologies*. In *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*:380–381, San Francisco, Februar 2005.

VON ARNIM, K., E. BORINSKI, P. SEEGEBRECHT, H. FIEDLER, R. BREDERLOW, R. THEWES, J. BERTHOLD und C. PACHA: *Efficiency of body biasing in 90 nm CMOS for low power digital circuit*. IEEE Journal of Solid-State Circuits, 39(7):1549–1556, Juli 2005.

PACHA, C., K. VON ARNIM, T. SCHULZ, W. XIONG, M. GOSTKOWSKI, G. KNOBLINGER, A. MARSHALL, T. NIRSCHL, J. BERTHOLD, C. RUSS, H. GOSSNER, C. DUVVURY, P. PATRUNO, R. CLEAVELIN und K. SCHRUEFER: *Circuit design issues in multi-gate FET CMOS technologies*. In *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*:420–421, San Francisco, Februar 2006.

MARSHALL, A., M. KULKARNI, M. CAMPISE, R. CLEAVELIN, C. DUVVURY, H. GOSSNER, M. GOSTKOWSKI, G. KNOBLINGER, C. PACHA, C. RUSS, K. SCHRÜFER, T. SCHULZ, K. VON ARNIM, B. WILKS und W. XIONG: *FinFET current mirror design and evaluation*. In *IEEE Dallas Circuits and Systems Workshop*, Dallas, Januar 2006.

Patentanmeldungen im Rahmen dieser Arbeit

CHRISTIAN PACHA, KLAUS VON ARNIM, RALF BREDERLOW und JÖRG BERTHOLD: *CMOS-Schaltkreis-Anordnung*. Skewed-CMOS mit Multi-t_{ox} und Multi-V_t-Erweiterung, angemeldet am Deutschen Patentamt (Anmeldekennzeichen 10348018.8) und am US-Patentamt.

KLAUS VON ARNIM, CHRISTIAN PACHA und ROLAND THEWES: *Schaltkreis-Anordnung*. Flip-Flop mit Speicherfunktionalität im Stand-By Modus, eingereicht am Deutschen Patentamt (Anmeldekennzeichen 10255636.9) und am US-Patentamt (Anmeldekennzeichen 10/723,309).

KLAUS VON ARNIM und CHRISTIAN PACHA: *Pulsgenerator-Schaltkreis und Schaltkreis-Anordnung*. Modifizierte Eingangs- und Ausgangsstufe für Sense-Amplifier-Flip-Flops, eingereicht beim Deutschen Patentamt.

Abbildungsverzeichnis

2.1 Leckströme und On-Ströme in verschiedenen 90 nm-CMOS-Technologien . . 8
2.2 Geschwindigkeit von Low-Standby-Power-Transistoren in zukünftigen CMOS-Technologie-Generationen 10
2.3 Planarer und Multi-Gate-Transistor 12
2.4 Schaltungsdesign mit Multi-Gate-Transistoren 13
2.5 Schaltungsblöcke in einer System-on-Chip-Umgebung 15

3.1 Eingangs-Charakteristiken von 90 nm-NMOS-Transistoren 18
3.2 Relative Zusammensetzung der Leckströme von 90 nm-NMOS-Transistoren . 18
3.3 Schwellenspannung in Abhängigkeit von der Temperatur 21
3.4 Temperaturabhängigkeit von Source- und Gatestrom 23
3.5 Schwellenspannung in Abhängigkeit von der Gatelänge 23
3.6 Querschnitte durch LOCOS- und STI-isolierte Transistoren 25
3.7 Weitenabhängigkeit der Schwellenspannung 26
3.8 Weitenabhängigkeit des Unterschwellenstroms 26
3.9 Stack-Effekt in einem NAND2-Gatter 27
3.10 Aufbau einer getakteten CMOS-Schaltung 28
3.11 Leckstrom-Histogramm eines 32-bit-Multiplizierers 29
3.12 Tunnelstromkomponenten in MOSFETs 31
3.13 Bestimmung der Oxiddicke 32
3.14 Verlauf des Kanalpotentials 35
3.15 Gateleckstrom in Bulk-Transistoren, Modell und Messung 36
3.16 Gateleckstrom in einem MuGFET, Modell und Messung 36
3.17 Einfluss der Temperatur auf den Gateleckstrom 37
3.18 Leckströme in einem Inverterpaar 37
3.19 Einfluss der Nitridation auf den Gateleckstrom 38

3.20 Leistungsdichte von aktiven Schaltströmen und Unterschwellenströmen 39
3.21 Schaltvorgänge in statischer CMOS-Logik 41
3.22 Schaltverhalten eines NMOS-Transistors 43
3.23 Verzögerungszeit in Abhängigkeit vom Fan-out 44
3.24 Schalttrajektorien und Zeitdiagramme eines Inverters 45
3.25 Dimensionierungen für symmetrisches und geschwindigkeitsoptimiertes Schalten 48
3.26 Statische Schaltpunkte von symmetrischen und geschwindigkeitsoptimierten Gattern 49
3.27 Stack-Effekt unter Forward-Biasing 51
3.28 Dynamische Schalttrajektorien zweier gestackter NFETs 52
3.29 Schalttrajektorien von Inverter, NAND2 und NAND3 52
3.30 Statischer Zero-Temperature-Coefficient-Point 53
3.31 Vergleich statischer und dynamischer ZTC-Punkte 54
3.32 Dynamischer Zero-Temperature-Coefficient-Punkt 55
3.33 Dynamischer ZTC-Punkt in Abhängigkeit von der Substratspannung 56

4.1 Schwellenspannung in Abhängigkeit von der Bulk-Source-Spannung 58
4.2 Leckströme und Body Biasing 59
4.3 NMOS-Leckströme bei 25 °C und 85 °C. 59
4.4 PMOS-Leckströme bei 25 °C und 85 °C. 60
4.5 Summe der Leckströme bei 25 °C und 85 °C 61
4.6 Abhängigkeit des On-Stroms von der Bulk-Source-Spannung 63
4.7 Frequenz eines LL-Ringoszillators bei Forward- und Reverse-Biasing 63
4.8 Schwellenspannung von Transistoren in Serienschaltung 65

5.1 Domino-Logik-Gatter 69
5.2 Skewed-CMOS-Gatter 70
5.3 Statische und dynamische Inverter-Transferkennlinien in statischer und Skewed-CMOS-Logik 72
5.4 Zweiphasiges Taktschema für Skewed-CMOS-Logik 73
5.5 Minimum-Leakage-Vector-Modus in Skewed-CMOS-Multi-t_{ox}-Logik 74
5.6 Skewed-CMOS-t_{ox}-Logik mit lokalen Standby-Transistoren 75
5.7 Variation der Verzögerungszeit in statischer und Skewed-CMOS-Logik 77

5.8 Testschaltung zur Crosstalk-Detektion 78
5.9 Leckströme und Schaltverzögerungen Skewed-CMOS- und statischen CMOS-Schaltungen 81
5.10 Multi-V_t-Multi-t_{ox}-Logik 83
5.11 Leckstrom-Performance-Tradeoff für Multi-V_t-Logik in der 130 nm-Technologie 84
5.12 Leckstrom-Performance-Tradeoff für Multi-V_t-Multi-t_{ox}-Logik in der 90 nm-Technologie 85
5.13 Timing-Diagramme für statische und Skewed-CMOS-Flip-Flops und Latches 87
5.14 Transmissiongate-Master-Slave-Flip-Flop 89
5.15 Sense-Amplifier-Flip-Flop und modifiziertes Sense-Amplifier-Flip-Flop . . . 90
5.16 Optimiertes Sense-Amplifier-Flip-Flop 91
5.17 Ergebnisse der SAFF-Optimierung mit evolutionärem Algorithmus 93
5.18 Timing-Diagramm eines Sense-Amplifier-Flip-Flops 94
5.19 Ausgangsstufe eines Sense-Amplifier-Flip-Flops mit Zustandserhaltung 95
5.20 Datenpfad des Skewed-CMOS-LVT-Addierers 97
5.21 Skewed-CMOS-Latch 98

6.1 Kogge-Stone- und Ladner-Fischer-Addierer 102
6.2 Han-Carlson- und neuer Addierer 104
6.3 4 Bit des Blockschaltbildes eines Skewed-CMOS-Addierers 105
6.4 Foto eines Testchip-Moduls mit zwei Addierern 107
6.5 Blockschaltbild eines Addierers mit Peripherieschaltungen 108
6.6 Maximale Betriebsfrequenzen bei Variation der Versorgungsspannung 110
6.7 Erhöhung der Addierergeschwindigkeit durch Forward-Biasing 111
6.8 Leckströme und Schaltgeschwindigkeit von drei Addierern 112

A.1 Testschaltung zur Detektion von Crosstalk 119
A.2 Gemessene Häufigkeiten von Crosstalk 120
A.3 Crosstalk: Variation der logischen Funktion des Eingangsgatters 121
A.4 Crosstalk: Variation der Aggressor-Form 121
A.5 Crosstalk: Variation des P/N-Verhältnisses in der Victim-Leitung 122
A.6 Crosstalk bei unterschiedlichen Schwellenspannungen 122
A.7 Crosstalk: Variation der Steilheit des Aggressorsignals 123

A.8 Crosstalk in Abhängigkeit vom Fan-out der Victim-Leitung 123

B.1 Blockschaltbild der Ringoszillator-Auswahlschaltung 125
B.2 Statische Multi-V_t-Gatter in Ringoszillatoren 127
B.3 Variation der Gatelängen in statischen CMOS-Schaltungen 128
B.4 Variation der Gatelängen in Multi-V_t-Schaltungen 128

C.1 Transistordimensionierungen in Skewed-CMOS- und statischer CMOS-Logik 129
C.2 Transistordimensionierungen im optimierten SAFF 130

Tabellenverzeichnis

2.1 Anforderungen an aktuelle und zukünftige CMOS-Technologien 6
2.2 Transistoren der 90 nm- und 130 nm-Technologie 7

3.1 Vergleich der Temperaturkoeffizienten k_1 20

4.1 Veränderung des Leckstroms durch Reverse-Biasing 61
4.2 On-Ströme und Schaltgeschwindigkeiten bei Forward-Biasing 64
4.3 Herausforderungen beim Einsatz von Forward-Biasing und Reverse-Biasing . 67

6.1 Vergleich von Kenngrößen verschiedener Addierer-Architekturen 103
6.2 Addierer-Varianten auf Testchip 108

Literaturverzeichnis

[1] MOORE, G.: *Cramming more components onto integrated circuit*. Electronics, 38(8), April 1965.

[2] NOWAK, E. J.: *Maintaining the benefits of CMOS scaling when scaling bogs down*. IBM Journal of Research & Development, (2/3):169–180, März/Mai 2003.

[3] ASSOCIATION, SEMICONDUCTOR INDUSTRY: *International technology roadmap for semiconductors, Process integration, devices, and structures*. 2004.

[4] WU, C. C., Y. K. LEUNG, C. S. CHANG, Y.C. SUN et al.: *A 90-nm CMOS device technology with high-speed, general-purpose, and low-leakage transistors for system on chip applications, 7 layers of Cu interconnects, low k ILD, and 1 um2 SRAM cell*. In: *International Electron Device Meeting Technical Digest, IEDM*, Seiten 65–68, Dezember 2002.

[5] DEVGAN, A.: *Leakage issues in IC design*. Tutorial im Rahmen der IEEE/ACM International Conference on Computer-Aided Design, ICCAD, November 2003.

[6] LÜFTNER, T., J. BERTHOLD, C. PACHA, G. GEORGAKOS et al.: *A 90 nm CMOS low-power GSM/EDGE multimedia-enhanced baseband processor with 380MHz ARM9 and mixed-signal extensions*. In: *Int. Solid-State Circuits Conf. Dig. Tech. Papers, ISSCC*, Seiten 252–253, Februar 2006.

[7] ASSOCIATION, SEMICONDUCTOR INDUSTRY: *International technology roadmap for semiconductors, Process integration, devices, and structures*. 2001.

[8] THOMPSON, S., N. ANAND, M. ARMSTRONG, M. BOHR et al.: *90 nm logic technology featuring 50 nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and 1 um2 SRAM cell*. In: *International Electron Device Meeting Technical Digest, IEDM*, Seiten 61–64, Dezember 2002.

[9] XIANG, Q., B. YU, H. WANG und M.-R. LIN: *High performance sub-50 nm CMOS with advanced gate stack*. In: *Symposium on VLSI Technology, Digest of Technical Papers*, Juni 2001.

[10] KIM, Y. W., C. B. OH, Y. G. KO, K. P. SUH et al.: *50 nm gate length logic technology with 9-layer Cu interconnects for 90 nm node SoC applications*. In: *International Electron Device Meeting Technical Digest, IEDM*, Seiten 69–72, Dezember 2002.

[11] PARIHAR, S., M. ANGYAL, B. BOECK, C. LAGE et al.: *A high density 0.10μm CMOS technology using low k dielectric and copper interconnect*. In: *International Electron Device Meeting Technical Digest, IEDM*, Seiten 11.4.1–4, Dezember 2001.

[12] HUANG, S.-F. et al.: *High performance 50 nm CMOS devices for microprocessor and embedded processor core applications*. In: *International Electron Device Meeting Technical Digest, IEDM*, Seiten 237–240, Dezember 2001.

[13] SCHIML, T., S. BIESEMANS, G. BRASE, L. BURRELLAND E. CRABBÉ et al.: *A 0.13μm CMOS platform with Cu/ low-k interconnects for system on chip applications*. In: *Symposium on VLSI Technology, Digest of Technical Papers*, Juni 2001.

[14] INOHARA, M., I. TAMURA, T.YAMAGUCHI, H. KOIKE, K. SUNOUCHI et al.: *High performance copper and low-k interconnect technology fully compatible to 90 nm-node SOC application (CMOS4)*. In: *International Electron Device Meeting Technical Digest, IEDM*, Seiten 77–80, Dezember 2002.

[15] TUINHOUT, H.: *Impact of parametric mismatch and fluctuations on performance and yield of deep-submicron CMOS technologies*. In: *Proceedings of the 32nd European Solid-State Circuits Conference, ESSDERC*, September 2002.

[16] GYVEZ, J. P. DE und H. P. TUINHOUT: *Threshold voltage mismatch and intra-die leakage current in digital CMOS circuits*. IEEE Journal of Solid-State Circuits, 39(1):157–168, Januar 2004.

[17] EISELE, M., J. BERTHOLD, D. SCHMITT-LANDSIEDEL und R. MAHNKOPF: *The impact of intra-die device parameter variations on path delays and on the design for yield of low voltage digital circuits*. IEEE Transactions On Very Large Scale Integration (VLSI) Systems, 5(4):360–368, Dezember 1997.

[18] ASENOV, A., SAVAS KAYA und J. H. DAVIES: *Intrinsic threshold voltage fluctuations in decanano MOSFETs due to local oxide thickness variations*. IEEE Transactions on Electron Devices, 49(1):112–119, Januar 2002.

[19] CHENG, B., S. ROY und A. ASENOV: *The impact of random doping effects on CMOS SRAM cell*. In: *Proceedings of the 30rd European Solid-State Circuits Conference, ESSCIRC*, Seiten 219–220, September 2004.

[20] ASSOCIATION, SEMICONDUCTOR INDUSTRY: *International technology roadmap for semiconductors, Emerging research devices*. 2004.

[21] DOYLE, B., B. BOYANOV, S. DATTA, M. DOCZY, S. HARELAND, B. JIN, J. KAVALIEROS, T. LINTON, R. RIOS und R. CHAU: *Scalability and biasing strategy for CMOS with active well bias*. In: *Symposium on VLSI Technology, Digest of Technical Papers*, Juni 2003.

[22] WATANABE, S.: *Impact of three-dimensional transistor on the pattern area reduction for ULSI*. IEEE Transactions on Electron Devices, 50(10):2073–2080, Oktober 2003.

[23] GENOSSAR, D. und N. SHAMIR: *Intel Pentium M processor power estimation, budgeting, optimization, and validation*. Intel Technology Journal, (2):45–49, Mai 2003.

[24] TORII, S., S. SUZUKI, H. TOMONAGA, T. TOKUE, N. NISHI et al.: *A 600MIPS 120mW 70μA leakage triple-CPU mobile application processor chip*. In: *Int. Solid-State Circuits Conf. Dig. Tech. Papers, ISSCC*, Seiten 136–137, Februar 2005.

[25] HART, J., S. Y. CHOE, L. CHENG, C. CHOU, A. DIXIT, K. HO, J. HSU, K. LEE und J. WU: *Implementation of a 4th-generation 1.8GHz dual-core SPARC V9 microprocessor*. In: *Int. Solid-State Circuits Conf. Dig. Tech. Papers, ISSCC*, Seiten 186–187, Februar 2005.

[26] HENZLER, S., T. NIRSCHL, D. SCHMITT-LANDSIEDEL et al.: *Sleep transistor circuits for fine-grained power switch-off with short power-down times*. In: *Int. Solid-State Circuits Conf. Dig. Tech. Papers, ISSCC*, Seiten 302–303, Februar 2005.

[27] LONG, C. und L. HE: *Distributed sleep transistor network for power reduction*. In: *Proceedings of the Conference on Design Automation, DAC*, Seiten 181–186, Juni 2003.

[28] ROYANNEZ, P., H. MAIR, F. DAHAN et al.: *90 nm low leakage SoC design techniques for wireless applications*. In: *Int. Solid-State Circuits Conf. Dig. Tech. Papers, ISSCC*, Seiten 138–139, Februar 2005.

[29] HUANG, S.-F. et al.: *Scalability and biasing strategy for CMOS with active well bias*. In: *Symposium on VLSI Technology, Digest of Technical Papers*, Seiten 107–108, Juni 2001.

[30] TSCHANZ, J. W., S. G. NARENDRA, Y. YE, B. A. BLOECHEL, S. BORKAR und V. DE: *Dynamic-sleep transistor and body bias for active leakage power control of microprocessors*. IEEE Journal of Solid-State Circuits, 38(11):1838–1845, November 1995.

[31] TSCHANZ, J. W., J. T. KAO, S. G. NARENDRA, R. NAIR, D. A. ANTONIADIS, A. P. CHANDRAKASAN und V. DE: *Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage*. IEEE Journal of Solid-State Circuits, 37(11):1396–1402, November 2002.

[32] NARENDRA, S., A. KESHAVARZI, B. A. BLOECHEL, S. BORKAR und V. DE: *Forward body bias for microprocessors in 130-nm technology generation and beyond*. IEEE Journal of Solid-State Circuits, 38(5):696–701, Mai 2003.

[33] OOWAKI, Y., M. NOGUCHI, S. TAKAGI, D. TAKASHIMA et al.: *A sub-0.1μm circuit design with substrate-over-biasing*. In: *Int. Solid-State Circuits Conf. Dig. Tech. Papers, ISSCC*, Seiten 6.2–1–6.2–15, Februar 1998.

[34] LEE, D. und D. BLAAUW: *Static leakage reduction through sumultaneous threshold voltage and state assignment*. In: *Proceedings of the Conference on Design Automation, DAC*, Seiten 191–194, Juni 2003.

[35] KAO, J., A. CHANDRAKASAN und D. ANTONIADIS: *Transistor sizing issues and tool for multi-threshold CMOS technology*. In: *Proceedings of the Conference on Design Automation, DAC*, Seiten 409–414, Juni 1997.

[36] WEI, L., Z. CHEN, K. ROY, M. C. JOHNSON, Y. YE und V. K. DE: *Design and optimization of dual-threshold circuits for low-voltage low-power applications*. IEEE Transactions On Very Large Scale Integration (VLSI) Systems, 7(1):16–24, März 1999.

[37] SEEGEBRECHT, P.: *Skript zur Vorlesung Bauelemente und Schaltungen*. Script zur Vorlesung, Lehrstuhl für Halbleitertechnik, Christian-Albrechts-Universität zu Kiel, 2001.

[38] SZE, S. M.: *Physics of semiconductor devices*. Wiley-Interscience publication, 1981.

[39] TSIVIDIS, T.: *Operation and modeling of the MOS transistor*. WCB/McGraw-Hill, 2. Auflage, 1999.

[40] CHUNG, S. S.-S. und T. C. LI: *An analytical threshold–voltage model of trench isolated mos devices with nonuniformly doped substrates*. IEEE Transactions on Electron Devices, 39(3):614–622, März 1992.

[41] SUGINO, M., L. A. AKERS und J.M. FORD: *Optimum p-channel isolation structure for CMOS*. IEEE Transactions on Electron Devices, 31(12):1823–1828, Dezember 1984.

[42] OHE, K., S. ODANAKA, K. MORIYAMA, T. HORI und G. FUSE: *Narrow-width effects of shallow trench-isolated CMOS with n+-polysilicon gate*. IEEE Transactions on Electron Devices, 36(6):1110–1116, Juni 1989.

[43] NOURI, F., G. SCOTT, M. RUBIN, M. MANLEY und P. STOLK: *Narrow device issues in deep-submicron technologies–the influence of stress, TED and segregation on device performance*. In: *Proceedings of the 30th European Solid-State Device Research Conference*, Seiten 112–115, September 2000.

[44] WANG, Y. G., D. B. SCOTT, J. WU, J. L. WALLER, J. HU, K. LIU und V. UKRAINTSEV: *Effects of uniaxial mechanical stress on drive current of 0.13 μm MOSFETs*. IEEE Transactions on Electron Devices, 50(2):529–531, Februar 2003.

[45] SCOTT, G., J. LUTZE, M. RUBIN, F. NOURI und M. MANLEY: *NMOS Drive current reduction caused by transistor layout and trench isolation induced stress*. In: *International Electron Device Meeting Technical Digest, IEDM*, Seiten 827–830, Dezember 1999.

[46] PACHA, C., M. BACH, K. VON ARNIM, R. BREDERLOW, D. SCHMITT-LANDSIEDEL, P. SEEGEBRECHT, J. BERTHOLD und R. THEWES: *Impact of STI-induced stress, inverse narrow width effect, and statistical VTH variations on leakage currents in 120 nm CMOS*. In: *Proceedings of the 34th European Device Research Conference, ESSDERC*, Seiten 397–440, September 2004.

[47] CHEN, Z., M. JOHNSON, L. WEI und K. ROY: *Estimation of standby leakage power in CMOS circuits considering accurate modeling of transistor stacks*. In: *Proceedings of the International Symposium on Low Power Electronics and Design, ISLPED*, Seiten 239–244, August 1998.

[48] ROY, K., S. MUKHOPADHYAY und H. MAHMOODI-MEIMAND: *Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits*. Prodeedings of the IEEE, 91(2):305–327, Februar 2003.

[49] YE, Y., S. BORKAR und V. DE: *A new technique for standby leakage reduction in high-performance circuits*. In: *Symposium on VLSI Circuits, Digest of Technical Papers*, Seiten 40–41, Juni 1998.

[50] JOHNSON, M. C., D. SOMASEKHAR und K. ROY: *Leakage control with efficient use of transistor stacks in single threshold CMOS*. In: *Proceedings of the Conference on Design Automation, DAC*, Seiten 442–445, Juni 1999.

[51] ABDOLLAHI, A., FARZAN F. und M. PEDRAM: *Leakage current reduction in sequential circuits by modifying the scan chains*. In: *Proceedings of the International Symposium on Quality Electronic Design, ISQED*, Seiten 49–54, März 2003.

[52] BAUER, F.: *Charakterisierung von leckstromarmen CMOS Schaltungskonzepten mit Sleep Transistoren*. Diplomarbeit, Lehrstuhl für Technische Elektronik, Technische Universität München, Dezember 2004.

[53] SCOTT, D., S. TANG, S. ZHAO und M. NANDAKUMAR: *Device physics impact on low leakage, high speed DSP design techniques*. In: *Proceedings of the International Symposium on Quality Electronic Design, ISQED*, Seiten 349–354, März 2002.

[54] LANGEVELDE, R. VAN, A. J. SCHOLTEN und D. B. M. KLAASSEN: *Physical Background of MOS Model 11, Level 1101*. Report, Koninklijke Philips Electronics N. V., April 2003.

[55] LANGEVELDE, R. VAN: *MOS Model 11, Level 1102*. Report, Koninklijke Philips Electronics N. V., Oktober 2004.

[56] XI, X., K. M. CAO, H. WAN, M. CHAN und C. HU: *BSIM 4.2.1 MOSFET Model*. Manual, Department of Electrical Engineering and Computer Sciences, Univ. of California, Berkeley, 2001. www-device.eecs.berkeley.edu/∼bsim3/bsim4_get.html.

[57] ARNIM, K. V.: *Modellierung von Sub-50-nm-Dünnschicht-SOI-Transistoren*. Diplomarbeit, Lehrstuhl für Halbleitertechnik, Christian-Albrechts-Universität zu Kiel, Mai 2002.

[58] DAVIES, J. H.: *The physics of low-dimensional semiconductors*, Kapitel 5: Tunneling Transport. Cambridge University Press, 1997.

[59] BOWMAN, K. A., X. TANG, J. D. MEINDL et al.: *A curcuit-level perspective of the optimum gate oxide thickness*. IEEE Transactions on Electron Devices, 48(8):1800–1810, August 2001.

[60] ESAKI, L. und P. J. STILES: *New type of negative resistance in barrier tunneling*. Physical Review Letters, 16(24):1108–1111, Juni 1966.

[61] CAO, K. M., X. XI, C. HU et al.: *BSIM 4 gate leakage model including source-drain partition*. In: *International Electron Device Meeting Technical Digest, IEDM*, Dezember 2000.

[62] VELOSO, A., F. N. CUBAYNES, A. ROTHSCHILD, S. MERTENS, R. DEGRAEVE, R. O'CONNOR, C. OLSEN, L. DATE, M. SCHAEKERS, C. DACHS und M. JURCZAK: *Ultra-thin oxynitride gate dielectrics by pulsed-RF DPN for 65 nm general purpose CMOS applications*. In: *Proceedings of the 33rd European Device Research Conference, ESSDERC*, 2003.

[63] RJOUB, A., O. KOUFOPAVLOU und S. NIKOLAIDIS: *Low-power/low-swing domino CMOS logic*. In: *Proceedings of the 1998 IEEE International Circuits and Systems, ISCAS*, Seiten 13–16, Mai 1998.

[64] WANG, J.-S. und H.-Y. LI: *0.9-V sense-amplifier-based reduced-clock-swing MTCMOS flip-flops*. In: *Proceedings of the Conference on Design Automation, DAC*, Juni 2003.

[65] SAKURAI, T. und A. R. NEWTON: *Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas*. IEEE Journal of Solid-State Circuits, 25(2):584–594, April 1990.

[66] SAKURAI, T.: *Alpha power-law MOS model*. IEEE Solid-State Circuits Society Newsletter, 9(4):4–5, Oktober 2004.

[67] NA, M. H., E. J. NOWAK, W. HAENSCH und J. CAI: *The effective drive current in CMOS inverters*. In: *International Electron Device Meeting Technical Digest, IEDM*, Seiten 121–124, Dezember 2002.

[68] NOSE, K., S.-I. CHAE und T. SAKURAI: *Voltage dependent gate capacitance and its impact in estimating power and delay of CMOS digital circuits with low supply voltage*. In: *Proc. IEEE Intl. Symp. Low Power Electronics and Design, ISLPED*, Seiten 228–230, August 2000.

[69] RABAEY, J. M.: *Digital integrated circuits: a design perspective*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

[70] KUNG, D. S. und R. PURI: *Optimal P/N width ratio selection for standard cell libraries*. In: *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, ICCAD*, Seiten 178–184, Piscataway, NJ, USA, 1999. IEEE Press.

[71] KANDA, K., K. NOSE, H. KAWAGUCHI und T. SAKURAI: *Design impact of positive temperature dependence on drain current in sub-1V CMOS VLSI's*. IEEE Journal of Solid-State Circuits, 36(10):1559–1564, Oktober 2001.

[72] BITTLESTONE, C., A. HILL, V. SINGHAL und ARVIND N. V.: *Architecting ASIC libraries and flows in nanometer era*. In: *Proceedings of the Conference on Design Automation, DAC*, Seiten 776–781, Juni 2003.

[73] MARTIN, S., K. FLAUTNER, T. MUDGE und D. BLAAUW: *Combined dynamic voltage scaling and adaptive body biasing for lower power microprocessors under dynamic workloads*. In: *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, ICCAD*, November 2002.

[74] CHEN, T. W. und J. GREGG: *A low cost individual-well adaptive body bias (IWABB) scheme for leakage power reduction and performance enhancement in the presence of intra-die variations*. In: *Design, Automation and Test in Europe Conference and Exhibition Volume I, DATE*, Februar 2004.

[75] NARENDRA, S. et al.: *Ultra-low voltage circuits and processor in 180 nm to 90 nm technologies with a swapped-body biasing technique*. In: *Int. Solid-State Circuits Conf. Dig. Tech. Papers, ISSCC*, Seiten 156–157, Februar 2004.

[76] ZHAO, S., A. CHATTERJEE, S. TANG et al.: *Transistor optimization for leakage power management in a 65 nm CMOS technology for wireless and mobile applications*. In: *Symposium on VLSI Technology, Digest of Technical Papers*, Seiten 14–15, Juli 2004.

[77] INUKAI, T. und T. HIRAMOTO: *Suppression of stand-by tunnel current in ultra-thin oxide MOSFETs by dual oxide thickness-multible threshold voltage CMOS (DOT-MTCMOS*. Japanese Journal of Applied Physics, (4B):2287–2290, April 2000.

[78] MUTOH, S. et al.: *1-V power supply high-speed digital circuit technology with multithreshold-voltage CMOS*. IEEE Journal of Solid-State Circuits, 30(8):847–854, August 1995.

[79] KRAMBECK, R. H., C. M. LEE und H.-F. S. LAW: *High-speed compact circuits with CMOS*. IEEE Journal of Solid-State Circuits, SC-17(3):614–619, Juni 1982.

[80] NG, P., P. T. BALSARA und D. STEISS: *Performance of CMOS differential circuits*. IEEE Journal of Solid-State Circuits, 31(6):841–846, Juni 1996.

[81] CHANG, S.-C., C.-H. CHENG, W.-B. JONE, S.-D. LEE und J.-S. WANG: *Charge-sharing alleviation and detection for CMOS domino circuits*. IEEE Transactions On Computer-Aided Design Of Integrated Circuits And Systems, 20(2):266–280, Februar 2001.

[82] WANG, L. und N. R. SHANBHAG: *An energy-efficient noise-tolerant dynamic circuit technique*. IEEE Transactions On Circuits And Systems—II: Analog And Digital Signal Processing, 47(11):1300–1306, November 2000.

[83] SOLOMATNIKOV, A. et al.: *Skewed CMOS: noise-immune high-performance low-power static circuit family*. In: *Intertational Conference on Computer Design, ICCD*, Seiten 241–246, September 2000.

[84] HENZLER, S., G. GEORGAKOS, J. BERTHOLD und D. SCHMITT-LANDSIEDEL: *Fast power-efficient circuit-block switch-off scheme*. Electronic Letters, 40(2):103–104, Januar 2004.

[85] WERNER, C., R. GÖTTSCHE, A. WERNER und U. RAMACHER: *Crosstalk noise in future digital CMOS circuits*. In: *Proceedings of the Conference on Design, Automation and Test in Europe*, Seiten 331–335, 2001.

[86] MOLL, F., M. ROCA und A. RUBIO: *Measurement of crosstalk-induced delay errors in digital circuits*. Electronic Letters, 33(19):1623–1624, September 1997.

[87] GEROSA, G., S. GARY, C. DIETZ et al.: *A 2.2 W, 80 MHz superscalar RISC microprocessor*. IEEE Journal of Solid-State Circuits, 29(12):1440–1154, Dezember 1994.

[88] MATSUI, M. et al.: *A 200 MHz 13 mm^2 2-D DCT macrocell using sense-amplifying pipeline flip-flop scheme*. IEEE Journal of Solid-State Circuits, 29(12):1482–1490, Dezember 1994.

[89] NIKOLIC, B., V. G. OKLOBDZIJA et al.: *Improved sense-amplifier-based flip-flops: design and measurements*. IEEE Journal of Solid-State Circuits, 35(6):876–884, Juni 2000.

[90] CHINNERY, D. G., B. NIKOLIC und K. KEUTZER: *Achieving 550 MHz in an ASIC Methodology*. In: *Proceedings of the Conference on Design Automation, DAC*, Seiten 420–425, Juni 2001.

[91] OKLOBDZIJA, V. G. und V. STOJANOVIC: *Flip-Flop*. U.S. Patent 6232810 B1, 15. Mai 2001.

[92] PAPADANTONAKIS, K.: *A theory of constant $E\tau^2$ CMOS circuits*. Technischer Bericht, California Institute of Technology, 2001.

[93] MARTIN, A. J., M. NYSTRÖM und P. I. PÉNZES: *Series In Computer Science, Power aware computing*, Kapitel ET2: A metric for time and energy efficiency of computation, Seiten 293–315. Cambridge University Press, 2002.

[94] BEIELSTEIN, T. et al.: *Circuit design using evolutionary algorithms*. In: *Proc. IEEE Congress Evolutionary Computation*, Band 2, Seiten 1904–1909, Mai 2002.

[95] DIENSTUHL, J.: *Optimierung und Modellierung von sub-100 nm CMOS-Speicherelementen mit Methoden der Computational Intelligence*. Doktorarbeit, Fakultät für Elektrotechnik und Informationstechnik, Universität Dortmund, August 2005.

[96] SHIGEMATSU, S., S. MUTOH, Y. MATSUYA, Y. TANABE und J. YAMADA: *A 1-V high-speed MTCMOS circuit scheme for power-down application circuits*. IEEE Journal of Solid-State Circuits, 32(6):861–869, Juni 1997.

[97] ZYUBAN, V. und S. KOSONOCKY: *Low power integrated scan-retention mechanism*. In: *Proc. IEEE Intl. Symp. Low Power Electronics and Design, ISLPED*, Seiten 98–105, August 2002.

[98] ZYUBAN, V. und D. MELTZER: *Clocking strategies and scannable latches for low power applications*. In: *Proc. IEEE Intl. Symp. Low Power Electronics and Design, ISLPED*, Seiten 346–351, August 2001.

[99] KOSONOCKY, S. V. et al.: *Low-power circuits and technology for wireless digital systems*. IBM Journal of Research & Development, (2/3):283–298, März 2003.

[100] PESSOLANO, F. und R. I. M. P. MEIJER: *A sub-260ps quasi-static 32-bit ALU*. In: *14th International Workshop on Power and Timing Modeling, Optimization and Simulation, PATMOS*, Seiten 372–380, September 2004.

[101] KOGGE, P. M. und H. S. STONE: *A parallel algorithm for the efficient solution of a general class of recurrence equations*. IEEE Transactions on Computers, C-22(8), August 1973.

[102] KOGGE, P. M.: *Parallel solution of recurrence problems*. IBM Journal of Research & Development, Seiten 138–148, März 1974.

[103] FISHER, M. J. und R. E. LADNER: *Propositional dynamic logic of regular programs*. Journal of Computer and System Sciences, 18(2):194–211, 1979.

[104] ZIMMERMANN, R.: *Binary adder architectures for cell-based VLSI and their synthesis*. Doktorarbeit, Swiss Federal Institute of Technology (ETH), Zürich, 1998.

[105] KNOWLES, S.: *A family of adders*. In: *14th IEEE Symposium on Computer Arithmetic*, Seiten 30–37, April 1999.

[106] HAN, T. und D. A. CARLSON: *Fast area-efficient VLSI adders*. In: *Proceedings of the 8th Symposium on Computer Arithmetic*, Seiten 49–55, Mai 1987.

[107] KAWAGUCHI, H., K. NOSE und T. SAKURAI: *A super cut-off CMOS (SCCMOS) scheme for 0.5-V supply voltage withpicoampere stand-by current*. IEEE Journal of Solid-State Circuits, 35(10):1498–1501, Oktober 2000.

[108] KAO, S., R. ZLATANOVICI und B. NIKOLIC: *A 240ps 64b carry-lookahead adder in 90 nm CMOS*. In: *Int. Solid-State Circuits Conf. Dig. Tech. Papers, ISSCC*, Seiten 438–440, Februar 2006.

[109] MIN, K.-S., H. KAWAGUCHI und T. SAKURAI: *Zigzag super cut-off CMOS (ZSC-CMOS) block activation with self-adaptive voltage level controller: an alternative to clock-gating scheme in leakage dominant era*. In: *Int. Solid-State Circuits Conf. Dig. Tech. Papers, ISSCC*, Seiten 400–503, Februar 2003.

[110] HENZLER, S., M. KOBAN, D. SCHMITT-LANDSIEDEL, J. BERTHOLD und G. GEORGAKOS: *Design aspects and technological scaling limits of zigzag circuit block switch-off schemes*. In: *IFIP Inertational Conference on Very Large Scale Integration, VLSI-SOC*, Seiten 246–251, Dezember 2003.