TECHNISCHE UNIVERSITÄT
MÜNCHEN
Lehrstuhl für Technische Elektronik

Variationen und ihre Kompensation in CMOS Digitalschaltungen

Thomas Baumann

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. techn. Josef A. Nossek
Prüfer der Dissertation:	
1.	Univ.-Prof. Dr. rer. nat. Doris Schmitt-Landsiedel
2.	Univ.-Prof. Dr.-Ing. Tobias G. Noll, Rheinisch-Westfälische Technische Hochschule Aachen

Die Dissertation wurde am 25.03.2010 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 14.09.2010 angenommen.

Danksagung

„**Leider läßt sich eine wahrhafte Dankbarkeit mit Worten nicht ausdrücken.**“

(Johann Wolfgang von Goethe)

Obwohl ich J. W. von Goethe zustimme, möchte ich dennoch versuchen, meiner Dankbarkeit Ausdruck zu verleihen.

Ich danke meiner Doktormutter, Prof. Doris Schmitt-Landsiedel für die Betreuung meiner Industrie-Promotion und die warmherzige Aufnahme am Lehrstuhl für technische Elektronik. Herrn Prof. Noll danke ich für die Begutachtung meiner Arbeit.

Ein großer Dank geht an Dr. Matthias Schoebinger, Infineon Abteilungsleiter von Advanced Systems and Circuits (ASC). Ich freue mich sehr darüber, dass Sie mir die Möglichkeit gegeben haben, meine Arbeit im freundlichen Umfeld der ASC Kollegen durchzuführen.

Besonderer Dank gilt auch meinen Kollegen Dr. Jörg Berthold und Dr. Karl Hofmann, die mir stets für Diskussionen zur Verfügung standen und mit vielen fruchtbaren Gesprächen täglich zu meiner Motivation beigetragen haben.

Dr. Christian Pacha - mein Ansprechpartner und Gruppenleiter von ASC TOC - gilt in diesem Zusammenhang mein größter Dank. Lieber Christian, ich möchte mich für die zahlreichen Gespräche und Diskussionen bei Dir bedanken, die mich ermuntert haben, die Ergebnisse meiner Arbeit immer wieder kritisch zu hinterfragen. Ohne deine tägliche Motivation und die ehrliche und aufrichtige Art und Weise technische Probleme zu diskutieren, hätte ich nicht soviel gelernt, wie ich es in den letzten Jahren getan habe. Vielen Dank dafür!

Doch mein größter Dank gilt meinen Eltern, Helmut und Theresia Baumann. Ich danke Euch, dass Ihr meine Affinität zur Technik schon früh gefördert habt. Besonders dankbar bin ich auch für den Aufwand, den Ihr getrieben habt, damit ich schon zur Jugendzeit meinen Lieblingssport ausüben konnte. Auch heute noch ist dieser Sport das beste Mittel für mich um vom beruflichen Alltag abschalten zu können. Insbesondere in stressigen Zeiten konnte ich dabei wieder Kraft sammeln und neue Energien freisetzen.
Dank Euch konnte ich meine Schulausbildung und mein Studium frei von finanziellen Sorgen absolvieren. Dafür bin ich Euch sehr dankbar! Doch viel wichtiger war mir Euer Rückhalt, der mir in schweren Zeiten Kraft gab.
Liebes Schwesterherz auch Dir danke ich für immer offene Ohren, gute Gespräche und motivierende Worte. Ich bin froh, dass es Dich gibt.
Liebe Karin auch Dir gilt besonderer Dank! Danke für Dein Verständnis, dass viele Abende und Wochenenden vorwiegend mit Arbeit belegt waren, danke, dass Du immer ein offenes Ohr für mich hattest und danke fürs „Glücklichmachen“ .

Zu guter Letzt noch ein Gruß an meine Oma! Gerne hätte ich dieses Ereignis noch mit Dir gefeiert. Leider blieb mir dieser Wunsch verwehrt. Aber ich bin mir sicher, Du freust Dich mit mir.

Für meine Eltern, aus wahrhafter und tiefer Dankbarkeit!

Inhaltsverzeichnis

1 Einleitung **13**

2 Vorgehensweise zur abstraktionsebenenübergreifenden Variationsanalyse **17**

3 Variationen in modernen sub-100nm low-power CMOS Technologien **21**
3.1 Räumliche und zeitliche Klassifizierung von Variationen 21
3.1.1 Prozessvariationen 23
3.1.2 Umgebungsvariationen 32
3.1.3 Alterungseffekte 36
3.1.4 Zeitliche Klassifizierung von Variationen 37
3.2 Sensitivitätsanalyse und technologiebasierte Trendaussagen 39
3.2.1 Analyse der Laufzeitsensitivität 39
3.2.2 Schwankungen bei fortschreitender Technologieskalierung 43
3.2.3 Schaltungstechnische Aspekte der Laufzeitsensitivität 49

4 Mikroprozessormodell zur Bestimmung technologischer und mikroarchitektonischer Einflussgrößen **53**
4.1 Strukturanalyse eines ARM926 Mikroprozessor Produktdesigns 53
4.1.1 Setup-Zeit kritische Pfade 54
Beschaffenheit des Logikteils 54
Beschaffenheit des Taktbaums 62
4.1.2 Hold-Zeit kritische Pfade 63
Beschaffenheit des Logikteils 64
Beschaffenheit des Taktbaums 65
4.2 Aufbau des Mikroprozessormodells 66
4.2.1 Modellierung der Registeranzahl 67
4.2.2 Modellierung des Logikblocks 69
4.2.3 Modellierung des Taktverteilungsnetzes 76
4.2.4 Auswirkungen auf das Timing Verhalten 82
4.3 Ergebnisse für die ARM Mikroprozessor-Familie 83
4.4 Bemerkungen zu den Ergebnissen 91

5 Topologieanalysen und Robustheit **95**
5.1 Pfadübergreifende Topologieanalyse 96
5.2 Definition von topologischen und strukturellen Bewertungskenngrößen . . . 102
5.2.1 Topologische Korrelationen in kritischen Pfaden 103
5.2.2 Struktur- und topologieabhängige Bewertung der Schaltungssensitivität 111

6 Schaltungstechnische Ansätze zur Kompensation von Laufzeitschwankungen 117
6.1 Globale Post-Fabrikation Adaptionstechniken 118
6.1.1 Process und Adaptive Voltage Scaling 118
6.1.2 Adaptive Body Biasing 122
6.1.3 On-Chip Monitorschaltungen 129
6.1.4 Vergleich der Techniken 133
6.2 Präventive Kompensationstechniken 134
6.2.1 Long-L_{Poly} Design 134
6.2.2 Selektiver Einsatz von low-V_T Zellen im Taktbaum 137
6.2.3 Selektiver Einsatz von low-V_T Zellen in geschwindigkeitskritischen Pfaden 140
6.2.4 Einsatz von gepulsten Flip Flops (P-FF) / Latches (P-L) 145
Selektiver Einsatz von P-FFs in geschwindigkeitskritischen Pfaden . 146
Globaler Einsatz von gepulsten Latches (Pulsed Latch Design) . . . 154
6.2.5 Einfluss der Techniken auf die Schaltungssensitivität 161
6.2.6 Validierung des Sensitivitätsfaktors als Robustheitsmaß 164

7 Zusammenfassung und Schlussfolgerung 171
7.1 Zusammenfassung 171
7.2 Schlussfolgerung 174

Publikationsliste 193

Abbildungsverzeichnis 194

Tabellenverzeichnis 200

Abkürzungen und Formelzeichen

α_{dyn}	Dynamisches α zur Modellierung des spannungsabhängigen Laufzeitverhaltens
α_{FF}	Exponent zur Modellierung des super-linearen Anstiegs der Register/Flip Flop Anzahl mit zunehmender Pipelinetiefe
α_{Schalt}	Schaltaktivität
A_{VT0}	Mismatch-Konstante der Transistoreinsatzspannung
β	Glitch-bedingte Schaltaktivität
BEOL	Back-End Of Line, d.h. Prozessierung der Verdrahtung
BR	Branching im Taktbaum, d.h. Aufspaltung eines Netzes im Taktbaum
BR_{Log}	Branching in der Logik, d.h. Aufspalung eines Netzes im Logikpfad
c	Konstante in der gatterspektrenabhängigen Gewichtung des Schaltungssensitivitätsfaktors
C_{eff}	Effektive Lastkapazität
C_K	Koppelkapazität zwischen benachbarten Leitungen
$C_{K_{eff}}$	Effektive, d.h. schaltende Koppelkapazität benachbarter Leitungen
C_{Ltg}	Leitungskapazität
C_{MOS}	Eingangskapazität von MOS Transistoren
C_{Off}	Offset Kapazität zur Berechnung zellinterner Laufzeitbeiträge
C_{ox}	Oxidkapazität
C_{stat}	Statische Haltekapazität bei der Crosstalkberechnung
D2D	Die-to-Die Variationen
DF	Dämpfungsfaktor statistischer
DIBL	Drain Inducded Barrier Lowering
D-TB	Dynamisches Time-Borrowing
DUT	Device under Test
E_{FF}	Empfangendes Flip Flop am Ende eines Logikpfades
E_{tot}	Gesamtenergie
ϵ_{ox}	Dielektrizitätskonstante des Gateoxids
F	Flächenfaktor (μP-Modell)
f_{Clk}	Taktfrequenz
ϕ_F	Fermipotential
F_{rel}^{Clk}	Relative Laufzeitänderung eines Taktpfads aufgrund von IR-Drop
F_{rel}^{Log}	Relative Laufzeitänderung eines Logikpfads aufgrund von IR-Drop
FEOL	Front-End Of Line, d.h. Prozessierung aktiver Schaltelemente
γ_{V_T}	Temperaturkoeffizient der Einsatzspannung

H(x)	Häufigkeitsverteilung des Parameters x
H_{Ltg}	Vertikaler Leitungsabstand
I_D	Transistor Drainstrom (allgemein)
I_{eff}	Effektiver Transistorstrom
I_H	Transistor Drainstrom bei $V_{GS} = V_{DD}$ & $V_{DS} = V_{DD}/2$
I_L	Transistor Drainstrom bei $V_{GS} = V_{DD}/2$ & $V_{DS} = V_{DD}$
I_{leak}	Leckstrom (allgemein)
I_{lin}	Transistror Drainstrom bei $V_{GS} = V_{DD}$ & $V_{DS} = 50mV$
I_{sc}	Kurzschlussstrom eines CMOS Gatters während des Schaltvorgangs
k_μ	Temperaturkoeffizient der Beweglichkeit
κ_{Top}	Topologischer Korrelations Faktor (TKF)
L	Transistor-Gatelänge (allgemein)
L2L	Los-zu-Los Variationen
L_{eff}	Effektive Transistor-Gatelänge
λ_{XT}^{WC}	Skalierungsfaktor des von der STA angegebenen Crosstalkbeitrags
$\sigma_{L_{eff}}$	Standardabweichung der statistischen Gatelängenschwankung
$\delta_{L_{eff}}$	Lokale statistische Gatelängenschwankung
$\Delta L_{eff,glo}$	Globale systematische Gatelängenschwankung
$\Delta L_{eff,lok}$	Lokale systematische Gatelängenschwankung
L_{Ltg}	Leitungslänge
L_{nom}	Nominelle Transistor-Gatelänge
μ	Ladungsträgerbeweglichkeit
σ_μ	Standardabweichung der statistischen Beweglichkeitsschwankung
δ_μ	Lokale statistische Beweglichkeitsschwankung
$\Delta\mu_{glo}$	Globale systematische Beweglichkeitsschwankung
$\Delta\mu_{lok}$	Lokale systematische Beweglichkeitsschwankung
μ_{nom}	Nominelle Ladungsträgerbeweglichkeit
N_σ	Anzahl der zu berücksichtigenden Standardabweichungen der stat. Variationen
N_{Aggr}	Aggressoranzahl
n_{Arch}	Architekturbedingte zusätzliche Logikstufenanzahl
n_{CB}	Anzahl von Buffer-Zellen im Taktverteilungsnetz (Clock Buffer)
N_{Dot}	Dotierstoffkonzentration
N_{FF}	Flip Flop Anzahl
$N_{FF/LCB}$	Anzahl der von einem LCB versorgten Flip Flop Zellen
$N_{G/P}$	Mittlere Anzahl von Gattern pro Pfad
n_{Gatter}	Gatteranzahl
N_{Gatter}	Gesamtgatteranzahl einer Schaltung
n_{Log}	Logiktiefe eines Pfades
n_{Pfade}	Pfadanzahl
$N_{Pipeline}$	Pipelinestufenanzahl
R_{Ltg}	Leitungswiderstand
ρ_{Ltg}	Spezifischer Widerstand des Leitungsmaterials

R_{Tr}	Effektiver Transistorwiderstand
RSAT	Relative Signal Arrival Time, d.h. zeitlicher Abstand zweier Signale
s_1, s_2	Schaltungssensitivität bzw. Schaltungssensitivitätsfaktor
s_{C_k}	Skalierungsfaktor der lateralen Koppelkapazität zweier Leitungen
S_i	Laufzeitsensitivität gegenüber Einflussgröße i
$S_{i,rel}$	Relative Laufzeitsensitivität gegenüber Einflussgröße i
S_{Ltg}	Horizontaler Leitungsabstand
S-FF	Sendendes Flip Flop am Anfang eines Logikpfades
$\sigma_{HD-Pfad}$	Statistische Laufzeitschwankung von $t_{Clk-Q} + t_{Log} + t_{HD}$
SP	Splitting Point im Taktbaum (Zählung beginnend von der Takteinspeisung)
SSTA	Statistische Statische Timing Analyse
STA	Statische Timing Analyse
S-TB	Statisches Time-Borrowing
T	Temperatur
TKF	Topologischer Korrelations-Faktor
t_{CB}	Laufzeit einer Clock Buffer Zelle
$\sigma_{t_{CB},rel}$	Relative statistische Laufzeitschwankung einer Clock Buffer Zelle
T_{Clk}	Taktperiode
$t_{Clk,E}$	Laufzeit des Taktpfades zum empfangenden Flip Flop
$t_{Clk,S}$	Laufzeit des Taktpfades zum sendenden Flip Flop
t_{Clk-Q}	Clock-Q Laufzeit von Flip Flop/Latch
t_{Comb}	Laufzeit der Kombinatorik (t_{Clk-Q}, t_{SU}, t_{Log})
t_{Comb}^{IR}	IR-Drop induzierte Laufzeitschwankung
Δt_{Comb}^{IR}	Korrekturterm bei gleichzeitiger Berücksichtigung von Clock Jitter und IR-Drop induzierter Laufzeitschwankung im Mikroprozessormodell
t_d	Laufzeit (allgemein)
$\sigma_{t_{d,rel}}(k)$	Statistische Laufzeitschwankung aufgrund von Strom-Mismatch
t_d^{nom}	Nominelle Laufzeit
t_{D-Clk}	Data-Clk Laufzeit von Flip Flop/Latch
t_{D-Q}	Data-Q Laufzeit von Flip Flop/Latch
t_{Gatter}	Gatterlaufzeit
$\sigma_{t_{Gatter}}$	Statistische Laufzeitschwankung eines Einzelgatters (allgemein)
T_{HD}	Pfad Timing von Hold-Zeit kritischen Pfaden
t_{HD}	Hold-Zeit
t_{Jitter}	Clock Jitter induzierte Laufzeitschwankung
ΔT_{krit}	Kritischer Timing Bereich innerhalb ΔT_{Var}
t_{Log}	Logiklaufzeit
T_{Ltg}	Leitungshöhe
t_{ox}	Gateoxiddicke
t_{Pfad}	Pfadlaufzeit
T_{Pfad}	Gesamtlaufzeit eines Pfades
$\sigma_{t_{Pfad}}$	Statistische Laufzeitschwankung eines Pfades

t_{Pfad}^{max}	Maximale Pfadlaufzeit
$t_{Pipeline}$	Laufzeit einer Pipelinestufe
t_{RC}	RC Laufzeit der Leitungen
t_{RW}	Laufzeitanteil des Leitungswiderstandes
t_{Sig}^{Agr}	Signalflanke des Aggressornetzes
t_{Skew}	Zeitlicher Unterschied im Taktverteilungsnetz (Clock Skew)
t_{Skew}^{Design}	Clock Skew aufgrund von Designunsicherheiten
t_{Skew}^{IR}	IR-Drop induzierter Clock Skew in Hold-Zeit kritischen Pfaden
t_{Skew}^{WID}	Clock Skew aufgrund systematischer WID Prozessvariationen
T_{SU}	Pfad Timing von Setup-Zeit kritischen Pfaden
t_{SU}	Setup-Zeit
ΔT_{Var}	Zeitliche Spanne der Laufzeitvariation
t_{Var}^{HD}	WID Timing Unsicherheit Hold-Zeit kritischer Pfade (allgemein)
t_{Var}^{SU}	WID Timing Unsicherheit Setup-Zeit kritischer Pfade (allgemein)
$\Delta t_{WID,rel}$	Systematische relative WID Laufzeitschwankung
t_{XT}	Crosstalk induzierte Laufzeit
Δt_{XT}^{WC}	Worst-case Crosstalk-Beitrag eines Pfades der von der STA berechnet wird
TB	Time-Borrowing
V_{DD}	Versorgungsspannung
$\Delta V_{DD,glo}$	Globale systematische Versorgungsspannungsschwankung
$\Delta V_{DD,lok}$	Lokale systematische Versorgungsspannungsschwankung
$V_{DD,nom}$	Nominelle Versorgungsspannung
V_{DIBL}	Einsatzspannungsabsenkung durch DIBL
V_{DS}	Drain-Source Spannung eines MOS-Transistors
V_{GS}	Gate-Source Spannung eines MOS-Transistors
V_T	Transistoreinsatzspannung
σ_{V_T}	Standardabweichung der statistischen Einsatzspannungsschwankung
δ_{V_T}	Lokale statistische Einsatzspannungsschwankung
$\Delta V_{T,glo}$	Globale systematische Einsatzspannungsschwankung
$\Delta V_{T,lok}$	Lokale systematische Einsatzspannungsschwankung
$V_{T,nom}$	Nominelle Transistoreinsatzspannung
$V_{T_{eff}}$	Effektive Transistoreinsatzspannung
$V_{T_{sat}}$	Einsatzspannung des Transistors im Sättigungsbereich
W	Transistorweite
$w(t_d)$	Pfad Timing abhängiger Gewichtungsfaktor des Gatterspektrums bei Bestimmung der Schaltungssensitivität
W_{Ltg}	Leitungsweite
W_{min}	Minimale Transistorweite einer Standardzellenbibliothek
W2W	Wafer-zu-Wafer Variation
WID	Within-Die Variation (Variation innerhalb eines Dies/Chips)
X_{Schalt}	Mikroarchitekturabhängiger Gewichtungsfaktor des Crosstalk-induzierten Laufzeitbeitrags

Z_{Log}	Vorfaktor zur Berücksichtigung der stärkeren Spannungsabhängigkeit von Logik- gegenüber Taktpfaden

1 Einleitung

Die Reduzierung der Energieaufnahme ist wesentlicher Faktor der fortschreitenden CMOS Technologieskalierung. Doch neben der Absenkung der Versorgungsspannung wird beim Übergang zu einer neuen Technologiegeneration auch die für die Schaltung benötigte Fläche deutlich reduziert. Dieser Flächengewinn ist Voraussetzung, um zusätzliche Funktionalität wie z.B. superskalare und Multi-Core Mikroprozessoren, Multimedia Erweiterungen etc. kosteneffizient implementieren zu können. Insbesondere mobile Anwendungen wie z.B. Mobiltelefone und Handheld PCs erfordern eine stete Reduzierung der aufgenommen Energie, um die Batterielaufzeiten zu erhöhen, während die Geschwindigkeitsanforderungen aufgrund neuer Standards wie z.B. HSxPA weiter ansteigen. Deshalb werden vermehrt Schaltungsblöcke verwendet, die für dedizierte Anwendungen optimiert sind (Hardware-Beschleuniger). All diese Veränderungen führen zu immer komplexeren, heterogenen Systemen [1].
Zusätzlich erhöht sich mit fortschreitender Technologieskalierung und der einhergehenden Absenkung der Versorgungsspannung unter 1V der Einfluss von prozess- und betriebsbedingten Variationen auf die Geschwindigkeit und Energieaufnahme von Digitalschaltungen. Zusammen mit der steigenden Schaltungskomplexität erschweren die erhöhten Laufzeitsensitivitäten den Entwurf von Digitalschaltungen in variationsbehafteter Umgebung.

Daher behandelt diese Arbeit den Einfluss von Prozess- und Umgebungsvariationen auf die Laufzeitschwankung in digitalen sub-100nm CMOS Logikschaltungen. Der Schwerpunkt der Arbeit liegt dabei auf der Analyse eingebetteter Mikroprozessoren in low-power CMOS Technologien. Dazu wird ein auf alle Digitalschaltungen verallgemeinerbarer Ansatz zur Identifikation, Klassifizierung, Quantifizierung und Bewertung von sensitiven Schaltungsstrukturen sowie der Quantifizierung der einzelnen Variationsbeiträge zur Laufzeitschwankung erarbeitet. Die einzelnen Variationseffekte werden in dieser Arbeit nicht isoliert behandelt, d.h. bei der Bewertung der Effekte werden die Interaktion und Randbedingungen der verschiedenen Effekte sowie der im Schaltungsentwurf durchlaufenen Abstraktionsebenen berücksichtigt (siehe Bild 1.1). Die vorgenommenen Untersuchungen beschränken sich auf den digitalen Logikteil der jeweiligen Mikroprozessoren. Cache- und on-chip Kommunikationsstrukturen werden nicht behandelt, da der Einfluss von Variationen auf diese Schaltungsteile im Gegensatz zu Mikroprozessorkerne bereits ausführlich untersucht wurden [2, 3, 4, 5, 6, 7]. Unter Berücksichtigung der gewonnenen Kenntnisse werden Kosten und Nutzen bekannter schaltungstechnischer Maßnahmen und daraus abgeleitete Erweiterungen zur Kompensation von Laufzeitschwankungen in Semi-Custom CMOS Digitalschaltungen bestimmt. Zur Bewertung der Robustheit einer Schaltung gegenüber Within-Die Prozess-, und on-chip Umgebungsvariationen erfolgt die erstmalige Definition mehrerer Kenngrößen, die als Metrik für die Verwundbarkeit einer Schaltung dienen. Diese Kenngrößen werden auch zur Bewertung der angewandten Kompensationstechniken verwendet.

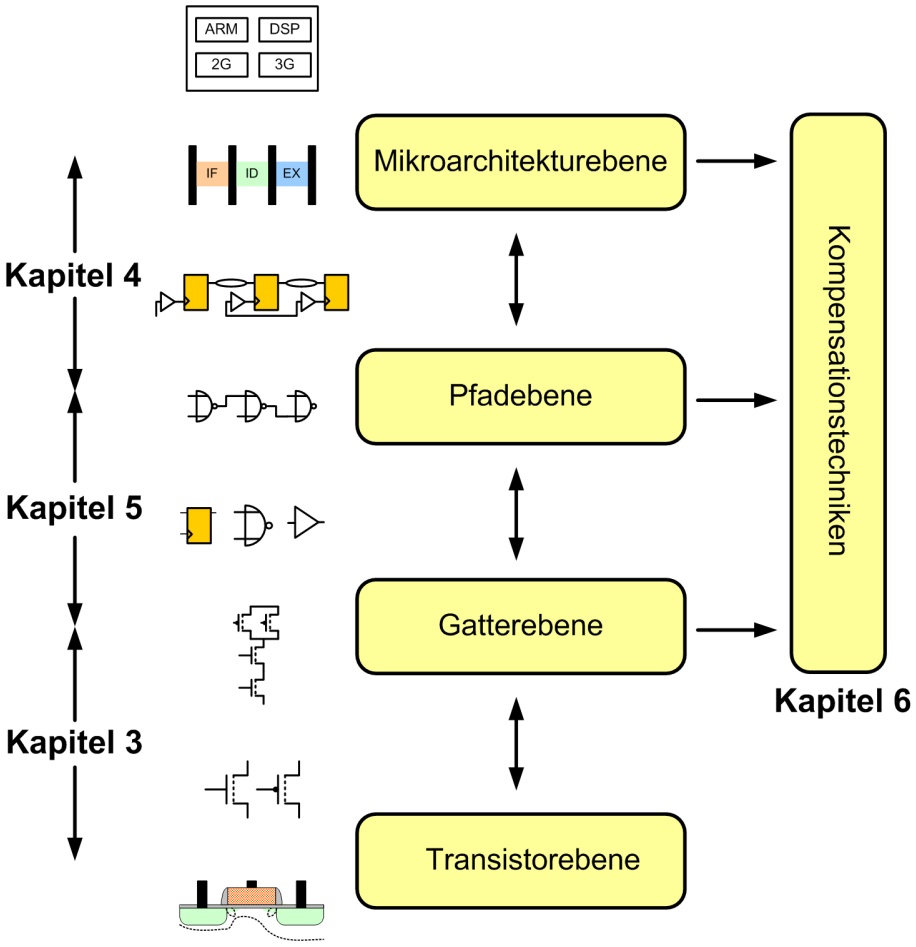


Bild 1.1: Gliederung der Arbeit nach verschiedenen Abstraktionsebenen.

Kapitel 2 beschreibt die entwickelte Vorgehensweise zur Bewertung des Einflusses von Variationen auf die Geschwindigkeit einer getakteten Digitalschaltung. Dabei wird das Zusammenspiel der in Kapitel 3, 4, 5 und 6 erarbeiteten Erkenntnisse hervorgehoben und die verwendeten Simulatoren bzw. Tools aufgeführt.

In Kapitel 3 werden alle relevanten Variationsquellen nach ihrem Auftreten klassifiziert und ihr Einfluss auf die Laufzeitschwankung analysiert. Dazu werden die physikalischen Einflussgrößen betrachtet und deren Beitrag zur Laufzeitschwankung quantifiziert. Um technologische Trendaussagen treffen zu können, werden die Laufzeitsensitivitäten von 130nm bis 40nm CMOS Technologien gegenüber den wichtigsten Variationsquellen untersucht und verglichen. Ergänzend werden auch die für low-power Schaltungen charakteristischen wechselnden Betriebsbereiche (Dynamic Voltage Scaling) berücksichtigt.

Kapitel 4 beinhaltet eine detaillierte strukturelle Analyse der kritischen Strukturen eines ARM926 Produktdesigns in 90nm low-power CMOS Technologie. Die Untersuchung des - in der Praxis relevantesten - Vertreters eines fünfstufigen RISC Prozessors ermöglicht die repräsentative Analyse struktureller und topologischer Eigenschaften wie z.B. Abfolge kritischer Pfade, die Aufspaltung im Taktverteilungsnetz etc. sowie der einzelnen Laufzeitbeiträge. Die Analyse erfolgt unter der Berücksichtigung von Randbedingungen und

Designkriterien eines state-of-the-art Design Flows, der alle Anforderungen an ein konkurrenzfähiges low-power Produktdesign erfüllt.
Auf Basis der strukturellen Analyse wird ein Mikroprozessormodell vorgestellt, das für die Abschätzung der einzelnen Laufzeitbeiträge sowohl technologische als auch strukturelle, topologische und mikroarchitektonische Aspekte berücksichtigt. Der wesentliche Kern des Mikroprozessormodells beruht auf der Beschreibung von Pipelinestrukturen mittels eines generischen 'kritischen Pfad' Modells, das neben Register zu Register Pfaden auch die Registerelemente (Flip Flops) und das Taktverteilungsnetz beinhaltet. Aufgrund des generischen Charakters ist das hier vorgestellte Modell auf beliebige getaktete Digitalschaltungen anwendbar. Ziel des Modells ist die Quantifizierung aller strukturabhängiger Beiträge von WID Prozess- und on-chip Umgebungsvariationen zur Laufzeitschwankung und die Bewertung von schaltungstechnischen Maßnahmen zur Kompensation von variationsbedingten Einflüssen.
Um den Einfluss von tieferem Pipelining auf das Laufzeitverhalten in variationsbehafteter Umgebung zu zeigen, werden auf Basis des Mikroprozessormodells stellvertretend für alle RISC Prozessoren der ARM926, ARM1176 und ARM Cortex A8 der ARM Mikroprozessorfamilie analysiert.

Kapitel 5 behandelt die Beschreibung und Quantifizierung des Einflusses schaltungsspezifischer Eigenschaften auf die Sensitivität einer Schaltung gegenüber Variationen. Zum einen wird die Auswirkung verschiedener Pfadtopologien auf das Laufzeitverhalten einer Schaltung analysiert. Die Unterscheidung der einzelnen Topologien hinsichtlich ihres Einflusses auf die Funktionalität der Schaltung erfolgt auf Basis neu definierter Pfadtopologien. Zum anderen wird erstmals der Einfluss von Pfad- bzw. Gatterspektrum und Pipeline-interner topologischer Korrelationen auf das Laufzeitverhalten digitaler Schaltungen untersucht. Mit der Definition des topologischen Korrelationsfaktors wird die Verflechtung der Gatter in geschwindigkeitskritischen Strukturen und somit der Einfluss lokaler Variationen auf die Gesamtschaltung beschrieben. Aus der Analyse der 'kritischen Hardware' erfolgen zwei Vorschläge zur Definition der Schaltungssensitivität gegenüber Variationen.

In Kapitel 6 werden globale Adaptionstechniken und präventive Designtechniken zur Kompensation bzw. Vermeidung variationsbedingter Laufzeitschwankungen bewertet. Als präventive Kompensationstechniken von Within-Die (WID) Laufzeitschwankungen, d.h. die Berücksichtigung von WID Laufzeitvariationen während des Schaltungsentwurfs, werden der selektive Einsatz von low-V_T Gattern und gepulsten Flip Flops in geschwindigkeitskritischen Pfaden sowie die globale Ersetzung der Standard Master-Slave Flip Flops durch gepulste Latches und Pulsgenerator diskutiert. Die Evaluation von Kosten und Nutzen erfolgt auf Basis der in Kapitel 4 erarbeiteten Erkenntnisse. Zur Bewertung der einzelnen präventiven Maßnahmen werden neben den herkömmlichen Größen wie Geschwindigkeit, Leistungsaufnahme und Flächenbedarf die in Kapitel 5 definierten Bewertungskenngrößen für die Robustheit einer Schaltung verwendet.
Zur Reduzierung der Laufzeitsensitivitäten findet eine Untersuchung des globalen Einsatzes von Long-L_{Poly} Gattern und des selektiven Einsatzes von low-V_T Gattern im Taktverteilungsnetz statt.
Als globale Adaptionstechniken werden das aus der Literatur bekannte Process/Adaptive Voltage Scaling (PVS/AVS) sowie Adaptive Body Biasing (ABB) behandelt und der für die Implementierung erforderliche Aufwand gegenübergestellt. Hierzu werden auch expe-

rimentelle Ergebnisse in 90nm, 65nm und 45nm CMOS verwendet. Als Teststrukturen dienen repräsentative Schaltungsstrukturen und -topologien, die aus den Erkenntnissen der Mikroprozessoranalyse in Kapitel 4 abgeleitet wurden.

Das letzte Kapitel der Arbeit fasst die wichtigsten Ergebnisse, Erkenntnisse und Wertungen zusammen und gibt einen Ausblick auf die Bedeutung der gewonnenen Erkenntnisse für kommende Generationen digitaler CMOS Schaltungen.

2 Vorgehensweise zur abstraktionsebenenübergreifenden Variationsanalyse

In diesem Abschnitt wird die erarbeitete Vorgehensweise zur Bewertung des Einflusses von Variationen auf die Geschwindigkeit und Robustheit von Digitalschaltungen vorgestellt. Die Bewertung von Variationseffekten wird durch zunehmende Schaltungskomplexität und verschiedene Betriebsbedingungen von low-power Schaltungen erschwert. Neben über 300 Parametern eines aktuellen BSIM Transistormodells, mehreren 100 Gattern einer Standardzellenbibliothek und mehreren 100.000 Gattern eines eingebetteten Mikroprozessors, die mehrere 100.000 kritischen Pfade bilden, müssen bei der Bewertung von Variationen auch die unterschiedlichen Versorgungsspannungs- und Temperaturbereiche der zu entwerfenden Schaltung berücksichtigt werden. Diese Beispiele zeigen, dass eine Aussage zum Einfluss von Variationen auf die Geschwindigkeit einer Schaltung nicht isoliert auf einer der in Bild 1.1 gezeigten Abstraktionsebenen getroffen werden kann, sondern die Berücksichtigung aller Ebenen vom Transistor bis hin zur Mikroarchitektur erfordert. Dies bedarf einer umfassenden Vorgehensweise zur Abstraktion und Bewertung aller verfügbaren Daten, wie sie im Rahmen dieser Dissertation erarbeitet wurde.

Bild 2.1 gibt einen Überblick über die wichtigsten Komponenten der Vorgehensweise. Ausgangspunkte sind Technologie-Datenbasen, Standarzellenbibliotheken, Betriebsparameter sowie Sign-Off Daten von eingebetteten Mikroprozessoren, hier eines ARM926 in 90nm CMOS.
Unter Sign-Off Status versteht man den Status eines Schaltungsentwurfs zum Zeitpunkt der Freigabe zur Produktion, d.h. nach der erfolgreichen Verifikation der Spezifikationen mittels EDA Tools. Diese Daten beinhalten auch layoutspezifische Informationen, die durch Extraktion des Layouts gewonnen wurden. Durch die Verwendung von Timing-Tools wie z.B. der Statischen Timing Analyse (STA) können zeitkritische Strukturen identifiziert werden. Während des herkömmlichen Timing Sign-Offs werden nur dedizierte Informationen ausgegeben, um zum einen geringen Rechenaufwand und damit verbundene geringe Rechnerlaufzeiten zu gewährleisten, zum anderen die zu bewertende Datenmenge gering zu halten. So wird z.B. während des Timing Sign-Offs im Wesentlichen nur der Slack des zeitkritischsten aller Pfade zur Verifikation der in der Spezifikation angegebenen Taktfrequenz herangezogen. Mit Slack wird in diesem Zusammenhang die zeitliche Marge eines Pfades bezeichnet, die vorhanden ist, bevor eine Setup-Zeit Verletzung hervorgerufen wird. Ist der Slack positiv, so wird davon ausgegangen, dass die spezifizierte Geschwindigkeit gewährleistet werden kann. Im vorliegenden Fall eines ARM926 ergab eine herkömmliche STA Analyse (PrimeTime SI) der oberen 10% des Timings Daten im Umfang von ca. 100MB, um den Slack des kritischsten Pfades bestimmen zu können. Der Dateninhalt der in dieser Arbeit verwendeten Timing-Reports ist für den gleichen Timing-Bereich um einen Faktor von ca. 100 erhöht. Diese erhöhte Datenmenge resultiert aus der Aufhebung

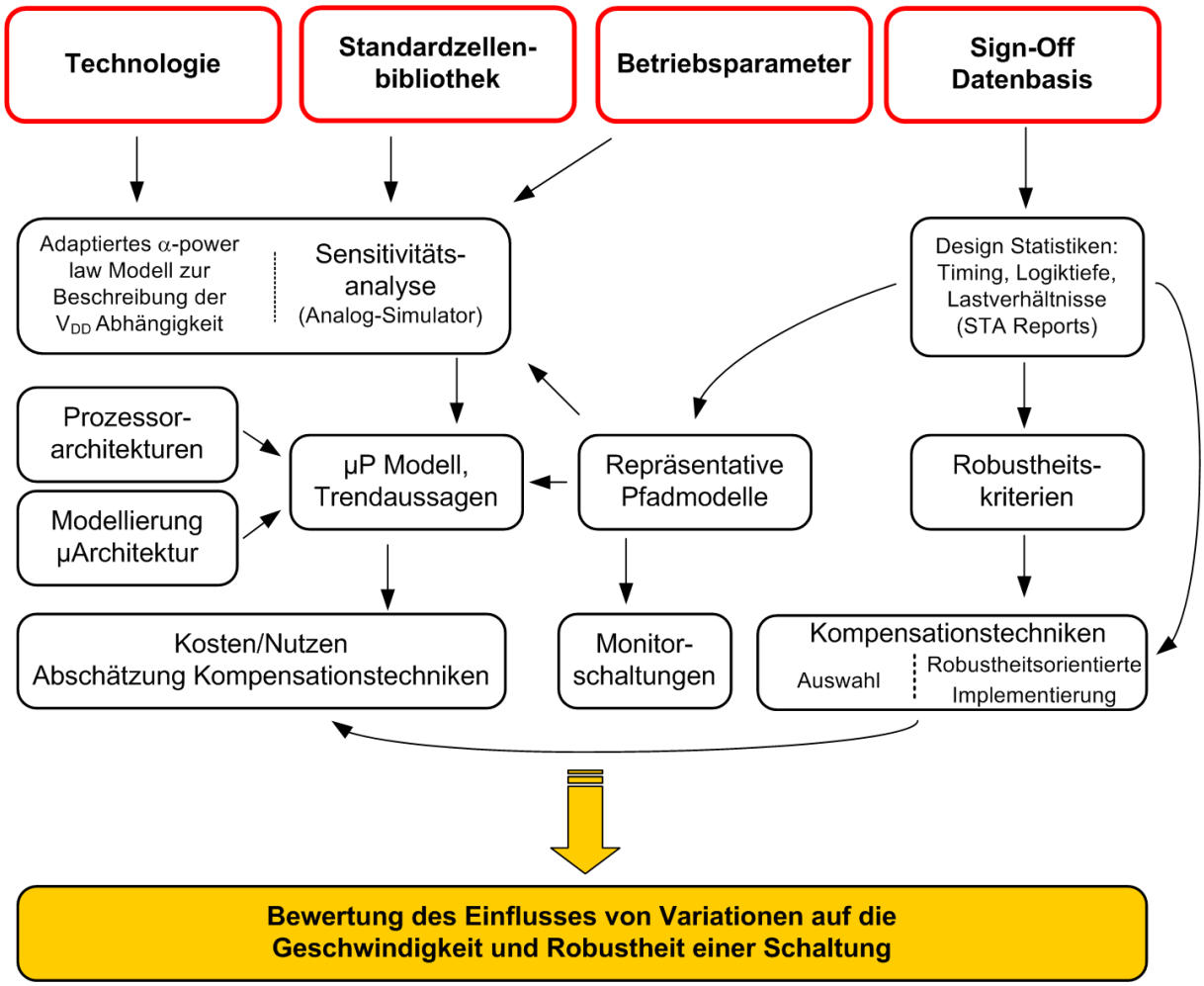


Bild 2.1: Überblick über die einzelnen Komponenten der konzipierten Vorgehensweise zur Bewertung des Einflusses von Variationen auf die Geschwindigkeit und Robustheit von integrierten Schaltungen.

struktureller Beschränkungen wie z.B. die beschränkte Anzahl an Pfaden pro empfangendem Register. Zusammen mit weiteren Informationen aus Sign-Off Reports z.B. zur Verteilung und Größe von Kapazitäten, beträgt die erzeugte Datenmenge eines untersuchten Mikroprozessors zwischen 20 und 25GB, was in etwa der Größenordnung von 10.000.000 DIN-A4 Text-Seiten entspricht (DIN-A4: ca. 2kB). Hier wird die Notwendigkeit einer geeigneten Abstraktion der zur Verfügung stehenden Informationen besonders deutlich.
Um repräsentative Aussagen generieren zu können, wurden Statistiken über topologische Kenngrößen wie z.B. die verwendeten Transistorgrößen und Gattertypen in den kritischen Pfaden, Treiber-Last Verhältnisse, die pipeline-übergreifende Abfolge von kritischen Pfaden etc. und strukturellen Kenngrößen wie z.B. die Lage des Aufspaltungspunktes von sendendem und empfangendem Taktpfad, Logiktiefe der kritischen Pfade etc. erzeugt. Dazu wurden die Daten der einzelnen Sign-Off Reports mittels neu erstellter Add-On Software verknüpft und topologische bzw. strukturelle Eigenschaften extrahiert.
Die Häufigkeitsverteilungen der jeweiligen Kenngrößen wurden analysiert und repräsentative Eigenschaften durch Abstraktion der Daten identifiziert. Diese Ergebnisse dienen als Basis für repräsentative Pfadmodelle, neu definierte Robustheitskriterien und zur Be-

Tabelle 2.1: Übersicht über die technologischen Kernparameter von 180nm bis 45nm low-power CMOS Technologien (reg-V_T Transistoren).

	180nm	130nm	90nm	65nm	45nm
$V_{DD,nom}$ [V]	1.8	1.5	1.2	1.2	1.1
T_{ox} [nm]	3.5	2.2	1.6	1.6	1.8
V_{TN} [mV]	430	385	370	380	410
V_{TP} [mV]	380	310	290	340	380
$I_{on,N}$ [$\mu A/\mu m$]	600	935	890	600	650
$I_{on,P}$ [$\mu A/\mu m$]	260	450	390	275	320

wertung von Kosten und Nutzen verschiedener Kompensationstechniken.

Die Überwachung des Schaltungszustandes nach der Produktion bzw. während des Betriebs der Schaltung (Monitoring) gewinnt weiter an Bedeutung, wie an der vermehrten Anzahl von Publikationen zum Thema Monitorschaltungen in den letzten Jahren zu erkennen ist. Hier spielt die Auswahl geeigneter, repräsentativer Testschaltungen CUT (Circuits Under Test) eine wichtige Rolle. In High-Performance Mikroprozessoren kommen häufig Monitorkonzepte zum Einsatz, die durch einen aufwendigen Off-Chip Test konfiguriert und kalibriert werden. Im Vergleich zu diesen standalone Prozessoren, die für mehrere 100 Dollar verkauft werden, sind für günstige low-power Schaltungen, wie sie z.B. in Mobiltelefonen eingesetzt werden, geringe Testkosten besonders wichtig, da diese einen großen Anteil am Produktpreis haben. Der Anspruch an die Implementierung von Monitoring-Konzepten ist daher, keine zusätzlichen Testkosten zu verursachen. Deshalb ist es insbesondere hier wichtig eine präzise Auswahl von geeigneten Testschaltungen bereits während des Schaltungsentwurfs zu treffen. Die Abstraktion von strukturellen und topologischen Schaltungseigenschaften zur Auswahl von repräsentativen Testschaltungen gewinnt daher an Bedeutung.

Eine ähnliche Abstraktion findet auf der Technologieebene incl. der zur Verfügung stehenden Standardzellenbibliothek statt. Zur Modellierung von Prozessvariationen werden ca. 20 Transistormodellparameter verwendet, die sich verschiedenartig auf die Geschwindigkeit von CMOS Logikschaltungen auswirken. Hier wurden Analysen vorgenommen, um die wichtigsten Parameter, d.h. die Parameter mit größtem Einfluss auf die Schaltgeschwindigkeit eines CMOS Gatters, zu identifizieren. Die wichtigsten nominellen Transistoreigenschaften der in dieser Arbeit verwendeten CMOS Technologien sind in Tabelle 2.1 zusammengefasst [8, 9, 10, 11, 12].
Unter Verwendung des Infineon internen Analog-Simulators Titan wurde die Sensitivität von repräsentativen Pfaden gegenüber den wichtigsten Prozess- und Betriebsparametern analysiert.
Um die Eigenschaften der repräsentativen kritischen Pfade nachbilden zu können, wurde die Laufzeitsensitivität von Standardzellen untersucht. Dazu wurden Gatter ausgewählt, die die wichtigsten schaltungstechnischen Unterschiede der Standardzellenbibliothek repräsentieren. Ein Vergleich der Laufzeitsensitivitäten von einzelnen Gattertypen und kritischen Pfaden ermöglicht somit die Nachbildung des Verhaltens kritischer Pfade gegenüber Variationen durch vereinfachte generische Pfadmodelle.
Diese repräsentativen Pfadmodelle und die für verschiedene Technologiegenerationen ermittelten Laufzeitsensitivitäten bilden zwei der drei Säulen des entwickelten Mikroprozes-

sormodells.
Dritte Säule des Modells bildet die Modellierung von Schaltungsstruktur und -topologie in Abhängigkeit mikroarchitektonischer Eigenschaften wie z.B. erhöhter Parallelität, tieferem Pipelining etc.. Hierzu wurden aus den Strukturanalysen eines ARM926 strukturelle und topologische Kenngrößen definiert, die eine Modellierung der Mikroarchitektur hinsichtlich des Einflusses von Variationen auf die Geschwindigkeit einer Schaltung ermöglichen.
Neben den Laufzeitsensitivitäten erfordert das Mikroprozessormodell auch die Eingabe von Schwankungsbreiten der einzelnen Prozess- und Umgebungsvariationen um den Beitrag variationsbedingter Laufzeitschwankungen zur Gesamtlaufzeit abschätzen zu können.

Zur Bewertung der Robustheit einer Schaltung werden Kenngrößen definiert. Diese basieren auf den Erkenntnissen der detaillierten Strukturanalyse der Schaltungen. Bei der Untersuchung der im Kapitel 5 eingeführten Robustheitskriterien wurden sowohl Titan-Simulationen als auch Matlab Modelle verwendet. Matlab Modelle wurden eingesetzt um Laufzeitverteilungen kleinerer generischer Gatternetzlisten für die strukturelle Extrapolation auf Schaltungen größerer Gatteranzahl verwenden zu können.

In dieser Arbeit wird besonders auf die Repräsentativität der gewonnen Ergebnisse geachtet. Aus diesem Grund wurde mit der ARM Mikroprozessor-Familie die repräsentativsten Vertreter eines RISC Prozessors zur Analyse und Bewertung von Variationseffekten auf die Geschwindigkeit einer Schaltung ausgewählt. Die untersuchten ARM Mikroprozessor-Designs wurden unter Verwendung von state-of-the-art Synthese und Place & Route Tools sowie unter industriellen Randbedingungen und Anforderungen an ein konkurrenzfähiges low-power Schaltungsdesign entworfen.

Zur Kosten- und Nutzenanalyse von verschiedenen Kompensationstechniken werden die ursprünglichen Schaltungsdesigns, die Ergebnisse des Mikroprozessormodells und die erstmals in dieser Form definierten Robustheitskriterien herangezogen. So werden im einen Fall die bestehenden Schaltungsdesigns abgeändert und das Timing mittels STA neu berechnet, im anderen Fall werden veränderte Laufzeitsensitivitäten und strukturelle Kenngrößen im Mikroprozessormodell berücksichtigt, um den Einfluss der Kompensationstechnik auf die Geschwindigkeit der jeweiligen Schaltung abzuschätzen.

Somit ergibt sich ein ganzheitlicher Blick auf die Wirkung von Variationen auf allen Abstraktionsebenen und die Möglichkeit zur Bewertung des Einflusses von Variationen aller Art auf die Geschwindigkeit und Robustheit einer Schaltung.
Obwohl die hier vorgestellte Vorgehensweise anhand detaillierter Analysen von eingebetteten Mikroprozessoren in sub-100nm CMOS Technologien konzipiert wurde, ist sie auf alle getakteten Digitalschaltungen übertragbar. Die detaillierten Teil- und Gesamtergebnisse der hier vorgestellten Vorgehensweise werden in den folgenden Kapiteln diskutiert.

3 Variationen in modernen sub-100nm low-power CMOS Technologien

Die Abweichung eines Prozess- bzw. Betriebsparameters vom nominellen Wert wird in der Mikroelektronik als Variation bezeichnet [13]. Variationen lassen sich nach der Art ihres Auftretens, der Zeitskala ihres Wirkens und ihrer örtlichen Ausdehnung klassifizieren (Bild 3.1).

Die Art des Auftretens unterscheidet man nach deterministischem, pseudo-statistischem und statistischem Verhalten. Deterministische Schwankungen treten nach einer bestimmten Systematik auf. Deshalb werden diese Variationen oftmals auch als systematische Variationen bezeichnet. Beispiel hierfür sind z.B. unterschiedliche Ätzzeiten und Aberrationen der Projektionslinsen [14] oder topologische Abhängigkeiten auf Layoutebene (STI stress, n-well proximity).

Statistische Variationen hingegen sind unkorrelierte Variationen, die statistisch unabhängig von Transistor zu Transistor schwanken. Beispiel hierfür sind statistische Dotierstoffatom-Schwankungen (Random Dopant Fluctuations RDF) und die Genauigkeit bei der Abbildung des Transistor-Gates (Line Edge Roughness LER).

Mit pseudo-statistischen Variationen bezeichnet man betriebsbedingte Schwankungen, die im Schaltungsentwurf meist über worst-case Annahmen berücksichtigt werden, da deren Größenordnung und Auswirkungen auf die Schaltung z.B. von Betriebsparametern wie der Schaltaktivität und Temperatur abhängig sind [15]. Als Beispiel hierfür dient der dynamische Einbruch der Versorgungsspannung (IR-Drop), der aufgrund betriebs- bzw. benutzerspezifischer Lastwechsel durch die Aktivierung bzw. Deaktivierung verschiedener Schaltungsblöcke hervorgerufen wird. Die Komplexität der Schaltung ist jedoch häufig zu groß um z.B. Schaltaktivitäten vorherzusagen bzw. zu bestimmen, so dass man trotz systematischen Ursprungs von einem pseudo-statistischen Verhalten sprechen kann.

Die weitere Differenzierung von Variationen nach Variationsquelle, räumliche Ausdehnung und den Zeitkonstanten ihres Wirkens erfolgt in den folgenden Abschnitten.

3.1 Räumliche und zeitliche Klassifizierung von Variationen

Der Fokus dieser Arbeit liegt auf dem Einfluss von Variationen auf die Geschwindigkeit digitaler Schaltungen. Im Schaltungsentwurf muss sichergestellt werden, dass die Schaltung selbst unter worst-case Bedingungen die spezifizierte Geschwindigkeit bzw. Taktfrequenz erreicht. Für die Modellierung der Laufzeit in Abhängigkeit von Transistor- und Betriebsparametern existieren in Genauigkeit und Komplexität verschiedenste Ansätze [16, 17, 18]. Für die Modellierung der Laufzeit unter nominellen Betriebsbedingungen liefert die Methode des effektiven Schaltstroms ausreichend genaue Ergebnisse [19] um

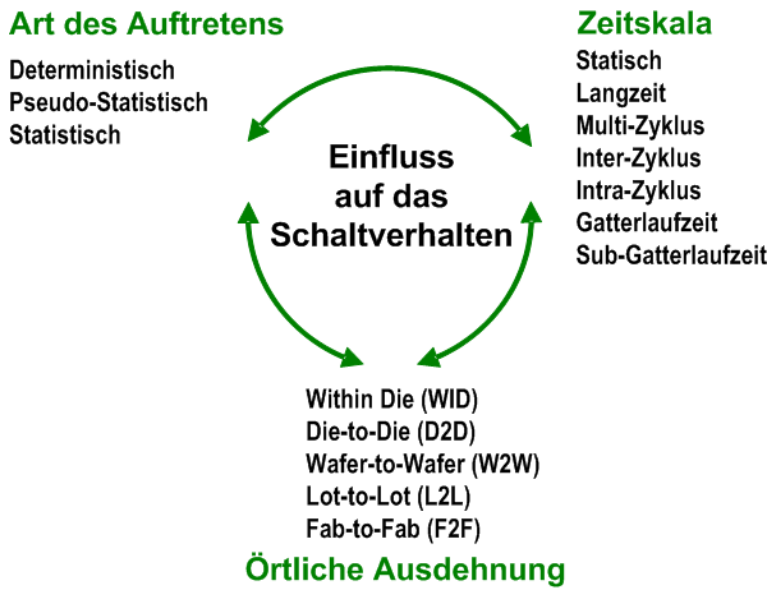


Bild 3.1: Kriterien zur Klassifizierung von Variationen.

generelle Trends zur Geschwindigkeit von CMOS Technologien zu untersuchen. Die Laufzeit eines Inverters berechnet sich wie folgt:

$$I_H = I_{DS}\,|_{(V_{GS}=V_{DD},V_{DS}=V_{DD}/2)} \tag{3.1}$$

$$I_L = I_{DS}\,|_{(V_{GS}=V_{DD}/2,V_{DS}=V_{DD})} \tag{3.2}$$

$$I_{eff} = \frac{I_H + I_L}{2} \tag{3.3}$$

$$t_d = \frac{C_{Last} \cdot V_{DD}}{2 \cdot I_{eff}} \tag{3.4}$$

Eine Erweiterung des effektiven Stroms zur Berechnung der Laufzeit komplexerer Gatter wurde in [17] vorgestellt. Der Effektivstrom I_{eff} beinhaltet hier gewichtete Anteile von I_H, I_L sowie $I_{lin} = I_{DS}\,|_{(V_{GS}=V_{DD},V_{DS}=0.05\cdot V_{DD})}$. Die Gewichtung dieser Anteile hängt von der jeweiligen Gattertopologie ab, d.h. für NAND, NOR und Inverter Gatter müssen verschiedene Gewichtungsfaktoren bestimmt werden. Die Gewichtungsfaktoren ergeben sich aus der Schalttrajektorie des jeweiligen Gatters, die wiederum von der zu treibenden Lastkapazität C_{Last} und der Signalflanke am Gattereingang abhängt.
Bild 3.2 zeigt die Schalttrajektorie eines Inverters und eines 2-fach NAND Gatters. Im Hintergrund ist die Schwankung des Drainstroms I_D aufgrund von Prozessvariationen gezeigt. Es ist zu erkennen, dass beide Gatter unterschiedlich sensitive Bereiche während des Schaltvorgangs durchlaufen. Die Berücksichtigung dreier verschiedener Basisströme im erweiterten Modell ermöglicht eine frühe Abschätzung der Geschwindigkeit einer Schaltung für künftige Technologiegenerationen. Der Einfluss der Eingangs- und Ausgangsflanken auf die Laufzeit wird in diesem Modell nicht berücksichtigt. Auch die Schwankung der Laufzeit aufgrund von Prozess- und Umgebungsvariationen kann nicht mit ausreichender Genauigkeit modelliert werden. Versuche, die Schwankungen der drei Stromkomponenten zur Bestimmung der Laufzeitschwankung zu verwenden, zeigen signifikante Unterschiede zur Simulation extrahierter Netzlisten, so dass sich dieser Ansatz nicht zur Untersuchung des Einflusses von Variationseffekten auf die Laufzeit einer Schaltung eignet.
Im Allgemeinen führen Variationen, steigende Kurzkanaleffekte sowie dynamische Ef-

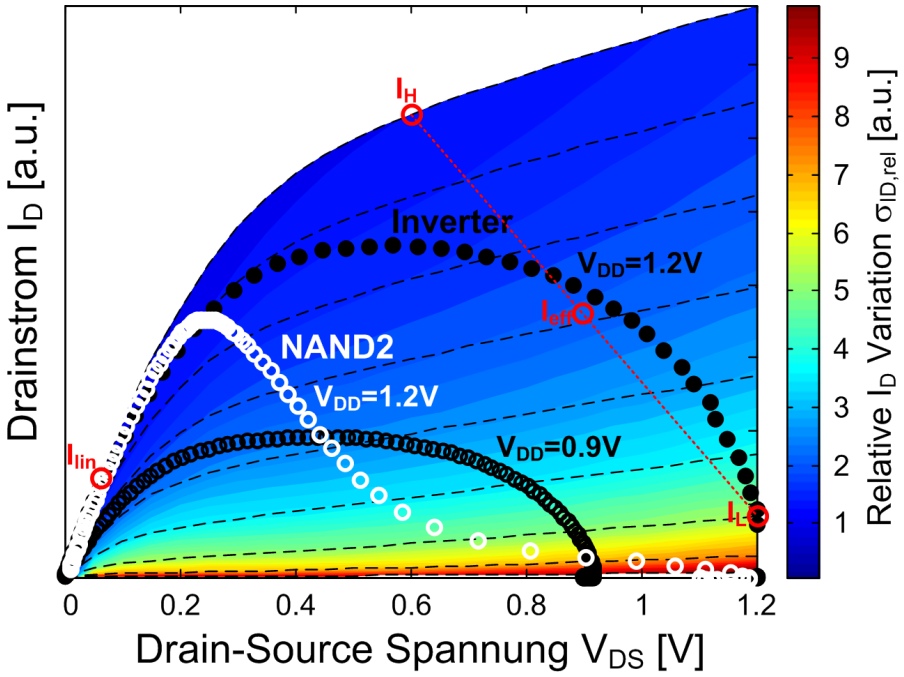


Bild 3.2: Schalttrajektorien von Inverter und 2-fach NAND in 65nm. Im Hintergrund ist die globale, prozessbedingte 1σ Schwankung des Drainstroms I_D eines NMOS Transistors gezeigt.

fekte während des Schaltens wie z.B. (Ent-)laden von internen Kapazitäten, kapazitive Kopplungen (Miller-Effekt) etc. zu erhöhtem Aufwand und erhöhter Komplexität bei der Modellierung von Transistorströmen und daraus resultierenden Laufzeiten. Da die zu bestimmenden variationsbedingten Schwankungsbreiten in der Größenordnung der Modellierungsgenauigkeit vereinfachter Laufzeitmodelle (wenige Modellparameter) liegen, ist für die Bestimmung des Einflusses von Prozess- und Umgebungsvariationen ein genaues Kompakt-Modell erforderlich. Deshalb werden in dieser Arbeit, bis auf vereinzelt gekennzeichnete Ausnahmen, nur Ergebnisse aus SPICE Simulationen mit BSIMv4.5 Modellen verwendet (312 Modellparameter).
Im Folgenden wird die Variation der Transistorparameter diskutiert, die den größten Einfluss auf die Schwankung von Gatter- und Pfadlaufzeiten haben.

3.1.1 Prozessvariationen

Prozessvariationen sind Abweichungen von Transistor- und Leitungsstrukturen vom nominellen Wert (Target-Value), die aufgrund der begrenzt möglichen Prozesskontrolle während der einzelnen Herstellungsschritte auftreten. Bild 3.3 zeigt die räumliche Klassifizierung von Prozessvariationen. Man unterscheidet Prozessschwankungen zwischen einzelnen Losen (L2L), zwischen Wafern des gleichen Loses (W2W), zwischen einzelnen Dies (Chips) eines Wafers (D2D), sowie zwischen identischen Strukturen auf einem einzelnen Die (WID). Auf Schaltungsebene, d.h. auf einem einzelnen Die, werden L2L, W2W und D2D Variationen als globale, WID Variationen als lokale Schwankungen bezeichnet.
Da alle Variationen außer den WID Variationen global wirken, findet man in der Literatur auch eine vereinfachte Klassifizierung, die nicht zwischen L2L, W2W und D2D Variationen unterscheidet, sondern diese Parameterschwankungen unter dem Begriff D2D Variationen zusammenfasst, wie in Bild 3.3 angedeutet ist.

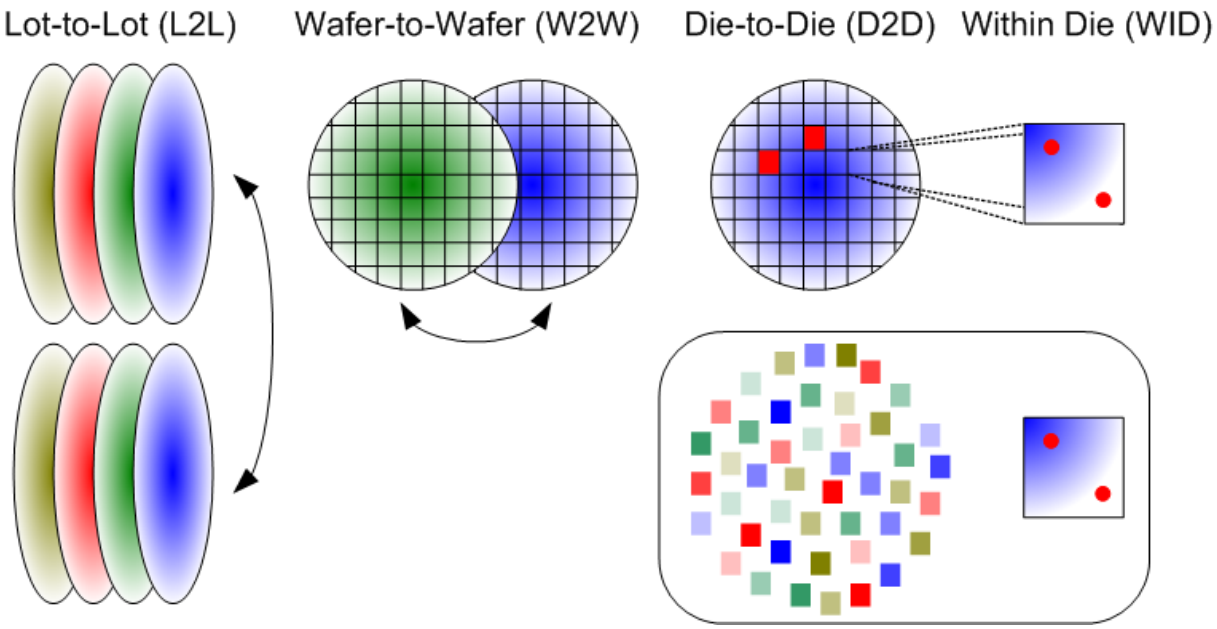


Bild 3.3: Räumliche Klassifizierung von Prozessvariationen.

Im Schaltungsentwurf werden die Schwankungsbreiten aller systematischen und statistischen Prozessvariationen einer Technologie als Monte-Carlo Parameter in Simulationsdateien abgelegt. Diese Parameter repräsentieren die Schwankungsbreiten aller Modellparameter dieser Technologie. Unabhängig davon, ob die Quelle der Variation systematischen oder statistischen Ursprungs ist, wird für alle Schwankungen als Wahrscheinlichkeitsverteilung die Gauß'sche Normalverteilung gewählt. Der Mittelwert μ_i^{MC} liegt beim Target-Wert des jeweiligen Parameters i, die Standardabweichung des Parameters i bei σ_i^{MC}. Im digitalen Schaltungsdesign wird der $\mu + 3\sigma$ Wert der Laufzeitschwankung oftmals als worst-case Szenario bezeichnet. Da aufgrund von Modellierungsproblemen selbst systematische Variationen als statistische Schwankungen modelliert werden [20, 21], wird in dieser Arbeit auch im Zusammenhang systematischer Schwankungen von Mittelwert und Standardabweichung gesprochen.
Im folgenden Abschnitt werden die Schwankungsquellen und -beiträge der wichtigsten Transistorparameter diskutiert.

- **Gatelänge:** $L_{eff} = L_{nom} + \underbrace{\Delta L_{eff,glo} + \Delta L_{eff,lok} + \delta_{L_{eff}}}_{\sigma_L^{MC}}$

  Die Gatelänge eines Transistors verändert sich z.B. aufgrund von systematischen Unterschieden in der Belichtung (Zeit und Dosis), unregelmäßiger Belackung, sowie Schwankungen im Ätzprozess. Die Gatelängenschwankung setzt sich aus einem globalen Anteil, bestehend aus L2L, W2W und D2D Komponenten, sowie einem lokalen Anteil zusammen.
  Die globale, systematische Schwankungsbreite $\Delta L_{eff,glo}$ hängt dabei vorwiegend von der Prozesskontrolle ab. Inhomogenitäten im Layout, basierend auf unterschiedlicher Anordnung und Dichte der Transistor-Gates, führen zu inhomogener Belichtung, Belackung und Ätzraten [22]. Dies resultiert in einem systematischen, lokalen Anteil der Gatelängenschwankung $\Delta L_{eff,lok}$. Systematische Variationsquellen können durch verbessertes OPC (Optical Proximity Correction) [23], auflösungsverbessern-

der Maßnahmen (Resolution Enhancement Techniques RET) [24] und regulären Layoutstrukturen wie z.B. PLA oder Logic Bricks Designs [25, 26] verringert werden. Der statistische Anteil der Gatelängenschwankung $\delta_{L_{eff}}$, der sich aufgrund von Mittelungseffekten mit zunehmender Weite des Transistors verringert, beruht unter anderem auf Quanteneffekten während der Belichtung. Beispiel hierfür ist die diskrete Anzahl von Photonen, die bei der Belichtung vom Photolack absorbiert werden. Ein zusätzlicher Anteil an der statistischen Schwankung der Gatelänge wird beim Ätzen des Poly-Gates hervorgerufen. Hier wird die Gatelänge durch statistische Schwankungen der Reaktanden beim Ätzvorgang beeinflusst. Die Bedeutung von LER nimmt aufgrund eines steigenden relativen Anteils an der Gatelänge für künftige Technologieknoten zu [27, 28]. In den aktuell verwendeten Technologien bis zu 45nm ist dieser Effekt in der digitalen Logik jedoch noch nicht sichtbar. Aufgrund der im Vergleich zur Logik kleinen Transistorweiten in SRAM Zellen wird dieser Effekt erstmals dort zu sehen sein.

- **Einsatzspannung:** $V_T = V_{T,nom} + \underbrace{\Delta V_{T,glo} + \Delta V_{T,lok} + \delta_{V_T}}_{\sigma_{V_T}^{MC}}$

  In modernen CMOS Technologien hängt die Einsatzspannung von verschiedenen Dotierungsschritten ab. Das effektive Dotierprofil nach Implantation der Substratdotierung und Halo-Dotierung bestimmt die Einsatzspannung des Transistors. Da die Einsatzspannung nicht nur vom Dotierprofil selbst sondern auch von der Aktivierung der Dotieratome abhängt, ist es wichtig, bei der Aktivierung der Dotierstoffe (Rapid Thermal Anneal RTA) einen Temperaturgang auf dem Wafer bzw. den einzelnen Dies zu vermeiden. Schwankungen während der Dotierstoff-Implantation und Temperaturunterschiede während des RTA stellen globale, systematische Quellen der Einsatzspannungsschwankung $\Delta V_{T,glo}$ dar.
  Als lokale systematische Quellen gelten layoutabhängige Schwankungen der Einsatzspannung $\Delta V_{T,lok}$[29]. Dazu zählen WID Temperaturgradienten während des RTA [30], z.B. durch Irregularitäten der Transistorgate-Dichte, Umgebungseffekte wie z.B. die Lage des Transistors zur n-Wanne (n-well proximity) [31] sowie der Einfluss der Transistorisolierung auf die Gitterstruktur des Transistor-Kanalgebiets (STI stress) [32].
  Die Einsatzspannung hängt neben der Temperatur auch von der Gatelänge des Transistors ab (V_T roll-off) [33]. Je kürzer die Gatelänge desto kleiner die Einsatzspannung, d.h. Gatelängen- und Einsatzspannungsschwankungen korrelieren partiell.
  Die statistische Einsatzspannungsschwankung δ_{V_T} basiert vorwiegend auf der statistischen Verteilung (Random Dopant Fluctuations RDF) der Dotierstoffatome. Da sich die zu dotierende Fläche bzw. das zu dotierende Volumen für Bulk-Transistoren von einer Technologiegeneration zur nächsten stark verringert, nimmt die Anzahl der Dotieratome, die die Einsatzspannung bestimmen, ab. Die relative Schwankung, die durch ein einzelnes gestreutes Dotieratom verursacht wird, nimmt daher für fortschreitende Technologieskalierung zu. Dabei zeigt die statistische Schwankung der Einsatzspannung folgende Abhängigkeit [34, 35]:

$$\sigma_{V_T} = \frac{A_{VT0}}{\sqrt{WL}} \sim \frac{N_{Dot}^{\frac{1}{4}}}{C_{ox}} \cdot \frac{1}{\sqrt{WL}} = \frac{t_{ox} \cdot N_{Dot}^{\frac{1}{4}}}{\varepsilon_{ox}} \cdot \frac{1}{\sqrt{WL}} \tag{3.5}$$

  A_{VT0} ist die Mismatch-Konstante der Einsatzspannung, t_{ox} die Dicke des Gateoxids,

N_{Dot} die Dotierstoffkonzentration im Kanal, ε_{ox} die Dielektrizitätskonstante des Gateoxids und $W \cdot L$ die Transistorgeometrie.
Die Mismatch-Konstante A_{VT0} zeigt eine lineare Abhängigkeit von t_{ox}. Ein weiteres Skalieren der Oxiddicke ist mittelfristig zu erwarten [36]. Für low-power Technologien kann eine Oxiddicke von 1.5nm als untere Grenze gesehen werden, um einen weiteren Anstieg der Gate-Leckströme zu verhindern [37]. Die Erhöhung der Kanaldotierung, die sich für moderne Bulk-Transistoren in der Größenordnung von $10^{18} - 10^{19} \frac{1}{cm^3}$ befindet [37], hat zwar eine geringere Auswirkung auf die Einsatzspannungsschwankung, trägt aber dennoch zu erhöhten V_T-Variationen mit fortschreitender Technologieskalierung bei. Die Größe der Einsatzspannungsschwankung in Bulk-CMOS Technologien wird vorwiegend durch die Transistorgeometrie beeinflusst, was eine weitere Zunahme der statistischen Einsatzspannungsschwankung für künftige Technologien zur Folge hat. Eine verringerte Schwankung der Einsatzspannung kann über eine erhöhte Dielektrizitätskonstante des Gateoxids erfolgen (high-k), oder bedarf einer fundamentalen Änderung des Transistors, z.B. die Einstellung der Einsatzspannung über das Gate-Material (metal gate) anstatt über die Dotierstoffkonzentration im Kanal. So kann die Dotierstoffkonzentration N_{Dot} signifikant reduziert werden bzw. wegfallen, und somit die Einsatzspannungsschwankung deutlich verringert werden.

- **Beweglichkeit:** $\mu = \mu_{nom} + \underbrace{\Delta\mu_{glo} + \Delta\mu_{lok} + \delta_\mu}_{\sigma_\mu^{MC}}$

  Die Beweglichkeit der Ladungsträger im Kanal wird durch die freie Weglänge der Ladungsträger und damit durch die Häufigkeit deren Streuung an Gitteratomen bestimmt. Diese ist abhängig von der Beschaffenheit der Oxid-Kanal-Grenzfläche, sowie von der Dotierstoffkonzentration und -verteilung im Kanal. Die Beweglichkeit nimmt mit geringerer Oxiddicke ab, da das vertikale elektrische Feld ansteigt, und die Ladungsträger stärker in Richtung Oxid-Kanal-Grenzfläche beschleunigt werden, was zu erhöhter Streuung der Ladungsträger an der Grenzfläche führt. So hängt die Beweglichkeitsschwankung auch von der Oxiddickenschwankung ab. Während die durch Oxiddickenschwankung hervorgerufene Beweglichkeitsschwankung vorwiegend global $\Delta\mu_{glo}$ auftritt, trägt die Schwankung der Dotierstoffatome sowohl zur globalen als auch lokalen ($\Delta\mu_{lok}$) Beweglichkeitsschwankung bei.
  Zusätzliche systematische Komponenten resultieren aus Verspannungen der Gitterstruktur, die z.B. aufgrund von STI und Ätz-Stop Schichten hervorgerufen werden. Kompressiver Stress bewirkt eine deutlich erhöhte Beweglichkeit von positiven Ladungsträgern (Löchern), die Beweglichkeit der negativen Ladungsträger (Elektronen) wird reduziert [32, 38]. In neuen Technologien werden zur Verbesserung der Beweglichkeit spezielle Schichten auf PMOS und NMOS Transistoren aufgebracht, um kompressive Verspannung im PMOS Kanal sowie Zugspannung im NMOS Kanal zu induzieren. Die von der Transistorstruktur und -anordnung abhängige Verspannung ist ebenfalls eine weitere Quelle systematischer Beweglichkeitsschwankung.
  Neben den systematischen Quellen existieren auch statistische Schwankungen in der Gitterstruktur, der Oxid-Kanal-Grenzfläche etc., die in dem statistisch schwankenden Anteil der Beweglichkeit δ_μ zusammengefasst werden.
  In der Praxis werden neben statistischen Einsatzspannungsschwankungen durch RDF auch statistische Drainstrom-Schwankungen (Strom-Mismatch) gemessen die

als statistische Schwankung der Beweglichkeit modelliert wird. Der Strom-Mismatch wird auch mit σ_k bezeichnet und entspricht im Wesentlichen δ_μ.

- **Oxiddicke:** t_{ox}

  Variationen bei der thermischen Oxidation zur Herstellung des Gateoxids haben eine unterschiedliche Oxiddicke zur Folge. Eine relativ hohe Fertigungsgenauigkeit bei der thermischen Oxidation und Nitridierung des Oxids ermöglicht eine sehr gute Reproduzierbarkeit und nahezu konstante Oxiddicken. Eine starke Verringerung der physikalischen Gateoxiddicke ist für zukünftige low-power CMOS Technologien nicht zu erwarten, da der damit verbundene Anstieg der Gateleckströme nicht hinnehmbar ist [37, 39, 36]. Die Schwankung der Oxiddicke wird sich daher für künftige Technologien nicht wesentlich verändern. Gegenüber den Schwankungen von L und V_T ist die Variation des Gateoxids gering [14].

- **Verdrahtung:** R_{Ltg}, C_{Ltg}

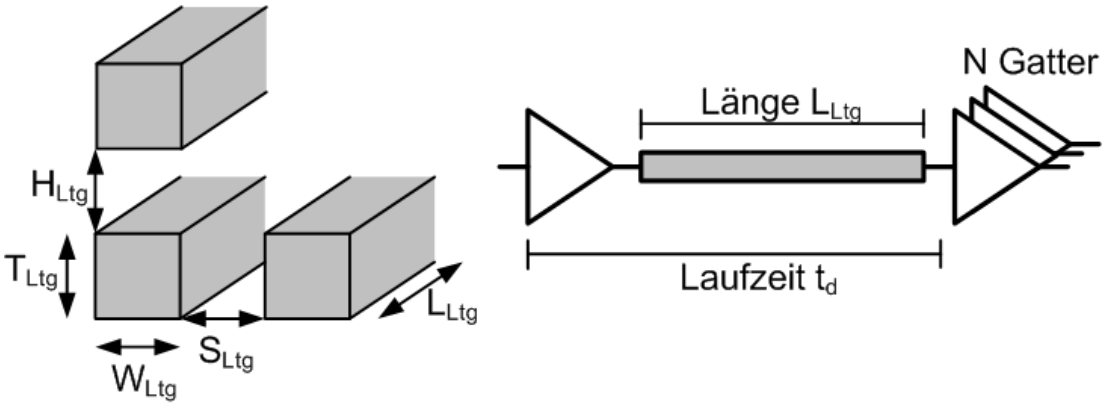
  Neben den Schwankungen in den FEOL Prozessen (Front End of Line FEOL) treten auch Variationen in den BEOL Prozessschritten (Back End of Line BEOL) auf, die sich sowohl auf den Widerstands- als auch den Kapazitätsbelag der einzelnen Metallebenen auswirken. Im Allgemeinen liegt die Schwankungsbreite des Widerstands deutlich über der der Leitungskapazität [40]. Der globale Beitrag zur Widerstandsschwankung liegt bei über 90%, der in globalen Prozess-Cornern abgedeckt wird. Der Within-Die Anteil resultiert aus Schwankungen der Leitungsgeometrie (Lithographie) und CMP-induzierten Schwankungen der Leitungshöhe [40, 41]. Diese lokalen, umgebungsabhängigen Schwankungen sind auch für die Schwankung der Leitungskapazität verantwortlich.
  Kapazität C_{Ltg} und Widerstand der Leitung R_{Ltg}, sowie die Laufzeit vom Gattereingang des Treibergatters bis zum Ende der Leitung können wie folgt modelliert werden [42]:

$$\begin{aligned} C_{Ltg} =& \varepsilon L_{Ltg}\left[1.15(\frac{W_{Ltg}}{H_{Ltg}}) + 2.8(\frac{T_{Ltg}}{H_{Ltg}})^{0.222}\right] \\ &+ 2\varepsilon L_{Ltg}(\frac{S_{Ltg}}{H_{Ltg}})^{-1.34}\left(0.03(\frac{W_{Ltg}}{H_{Ltg}}) + 0.83(\frac{T_{Ltg}}{H_{Ltg}}) - 0.07(\frac{T_{Ltg}}{H_{Ltg}})^{0.222}\right) \end{aligned} \tag{3.6}$$

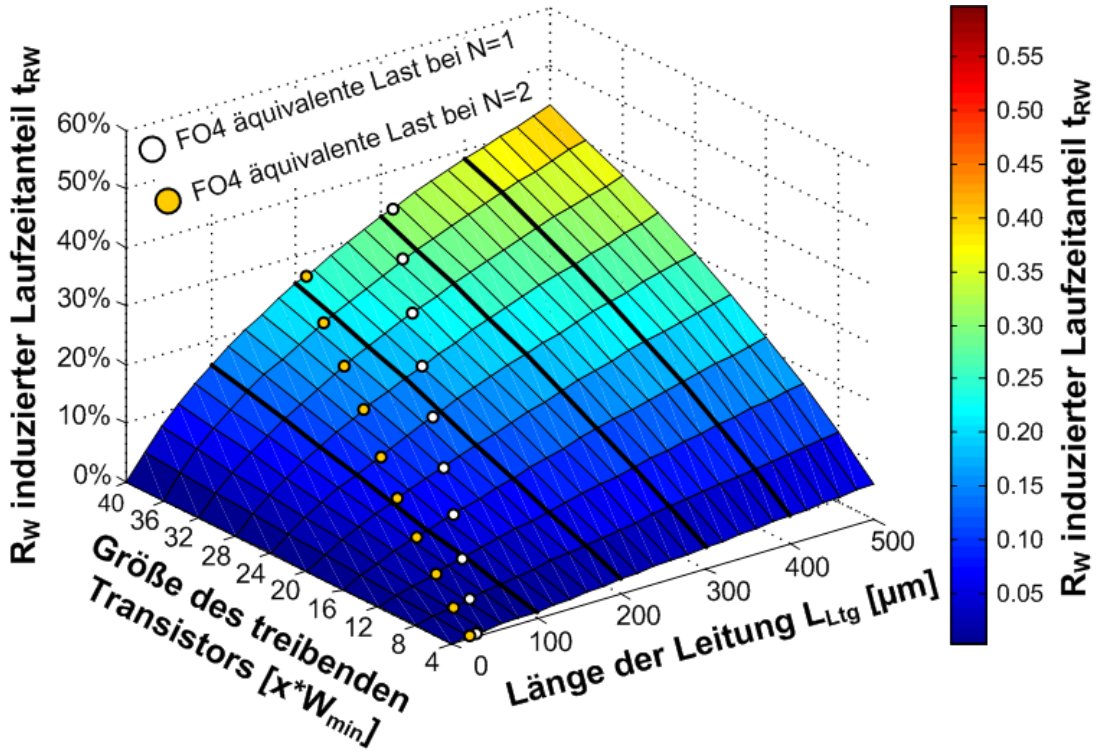
$$R_{Ltg} = \rho_{Ltg} \cdot \frac{L_{Ltg}}{W_{Ltg} \cdot T_{Ltg}} \tag{3.7}$$

$$t_d = 0.69 \cdot R_{Tr}\left(C_{Off} + C_{MOS} + C_{Ltg}\right) + R_{Ltg}\left(0.38 \cdot C_{Ltg} + 0.69 \cdot C_{MOS}\right) \tag{3.8}$$

  Die Größen $W_{Ltg}, T_{Ltg}, S_{Ltg}, H_{Ltg}$ beschreiben wie in Bild 3.4(a) gezeigt die Geometrie der Leitung. R_{Tr} ist der effektive Widerstand des Treibergatters, C_{Off} repräsentiert die zellinternen intrinsischen Kapazitäten, C_{MOS} ist die Kapazität der Eingangspins am Ende der Leitung, C_{Ltg} und R_{Ltg} sind Widerstand und Kapazität der Leitung. ρ_{Ltg} ist der spezifische Widerstand des Metalls z.B. Kupfer, Aluminium.
  Für die folgende Untersuchung wird der effektive Widerstand des Treibergatters



(a) Schematische Abbildung der Teststruktur.



(b) Beitrag des Leitungswiderstandes zur Gesamtlaufzeit in 65nm für verschiedene Treibertransistoren und Leitungslängen (nomineller Prozess.

Bild 3.4: Modellierung zur Bestimmung des Beitrags des Leitungswiderstands zur Laufzeit eines Pfades.

über den Effektivstrom I_{eff} bei der jeweiligen Versorgungsspannung V_{DD} berechnet. Wie bereits in diesem Kapitel diskutiert wurde, eignet sich der Effektivstrom vorwiegend zur Modellierung der Laufzeit unter nominellen Transistorparametern. Da in diesem Fall der Einfluss des Leitungswiderstandes bestimmt werden soll, werden die nominellen Transistoreigenschaften zur Bestimmung von I_{eff} herangezogen. Bild 3.4(b) zeigt den Laufzeitanteil der Gesamtlaufzeit t_d, die durch den Leitungswiderstand der in Bild 3.4(a) gezeigten Leitungsstruktur hervorgerufen wird. Die Ergebnisse wurden für verschiedene Transistorgrößen und Leitungslängen in 65nm CMOS berechnet. Als Leitungsparameter werden die Eigenschaften von Minimalleitungen der Metallebenen M2-M4 verwendet. Für jede Transistorgröße wird die äquivalente Leitungslänge berechnet, die zusammen mit der Anzahl N an Gattern am Ende der Leitung die vierfache kapazitive Last der Eingangslast des Treibergatters ergibt (Fanout-4 Last, FO4). Dieses Verhältnis von kapazitiver Last am Gatterausgang und Gattereingang wird von vielen Tools im Schaltungsentwurf angestrebt, um eine annähernd minimale Pfadlaufzeit zu erzielen [43, 44]. Neben der Auswirkung auf die Laufzeit werden ferner steile Signalflanken gewährleistet, was zum einen geringe Verluste durch Kurzschlussströme, zum anderen einen verminderten Einfluss von Crosstalkeffekten auf die Propagation des Signals zur Folge hat. In Bild 3.4(b) ist deutlich zu erkennen, dass für Gatter, bestehend aus Transistoren mit der Treiberstärke eines Transistors der Weite $W = 16W_{min}$, der resistive Laufzeitbeitrag einer solchen Leitung bei nur 7% liegt. Selbst eine Widerstandsschwankung von 10% führt zu einer Laufzeitschwankung von weniger als 1%. Für Treibertransistoren mit 40facher Minimalgröße liegt der Beitrag bei 34%. Derartige Strukturen sind hauptsächlich im Taktbaum und in Bus-Strukturen zu finden. Während im Taktbaum der Leitungsanteil an der kapazitiven Last aufgrund einer erhöhten Anzahl $N \geq 2$ an Empfangsgattern geringer ist, stellt eine Bus-Struktur eine Punkt-zu-Punkt Verbindung dar. In diesen Strukturen werden Leitungen verwendet, die deutlich breiter sind als die Minimalgröße, so dass sich der Widerstand der Leitung deutlich reduziert, während der Kapazitätsbelag aufgrund der höheren Kopplung auf die untere und obere Metallebene ansteigt. Die FO4-äquivalente Leitungslänge nimmt somit ab. Geometriebedingte R_{Ltg} und C_{Ltg} Schwankungen kompensieren sich hinsichtlich ihres Einflusses auf die Laufzeit jedoch teilweise gegenseitig, da beide stark negativ korrelieren, so dass die Schwankungsbreite der $R_{Ltg} \cdot C_{Ltg}$ Laufzeit deutlich geringer ist als die Schwankungsbreiten von Leitungskapazität und -widerstand. Bild 3.5 zeigt die nach Gleichung 3.6 und 3.7 berechneten globalen, geometrieabhängigen Schwankungen von R_{Ltg}, C_{Ltg} und $R_{Ltg} \cdot C_{Ltg}$ für 65nm CMOS unter der Annahme normalverteilter Geometrieschwankungen. Es ist deutlich zu erkennen, dass die negative Korrelation von R_{Ltg} und C_{Ltg} zu geringeren Schwankungsbreiten der $R_{Ltg} \cdot C_{Ltg}$ Laufzeit führen.

Im geschwindigkeitskritischen worst-case Fall, d.h. bei langsamem Prozess, verringert sich zudem der Einfluss von Widerstandsschwankungen der Verdrahtung auf die Laufzeit, da der effektive Transistorwiderstand und somit der Laufzeitanteil des Treibergatters an der Gesamtlaufzeit ansteigt.

Im Gegensatz zur Schwankung des Leitungswiderstandes führt eine veränderte Kapazitätsschwankung bereits bei kürzeren Leitungen zu Laufzeitschwankungen. Die vorwiegend systematischen Variationen der Leitungsstrukturen wirken sich jedoch wesentlich geringer auf die Geschwindigkeit der Schaltung aus, als die Schwankungen der Transistorparameter [45, 46].

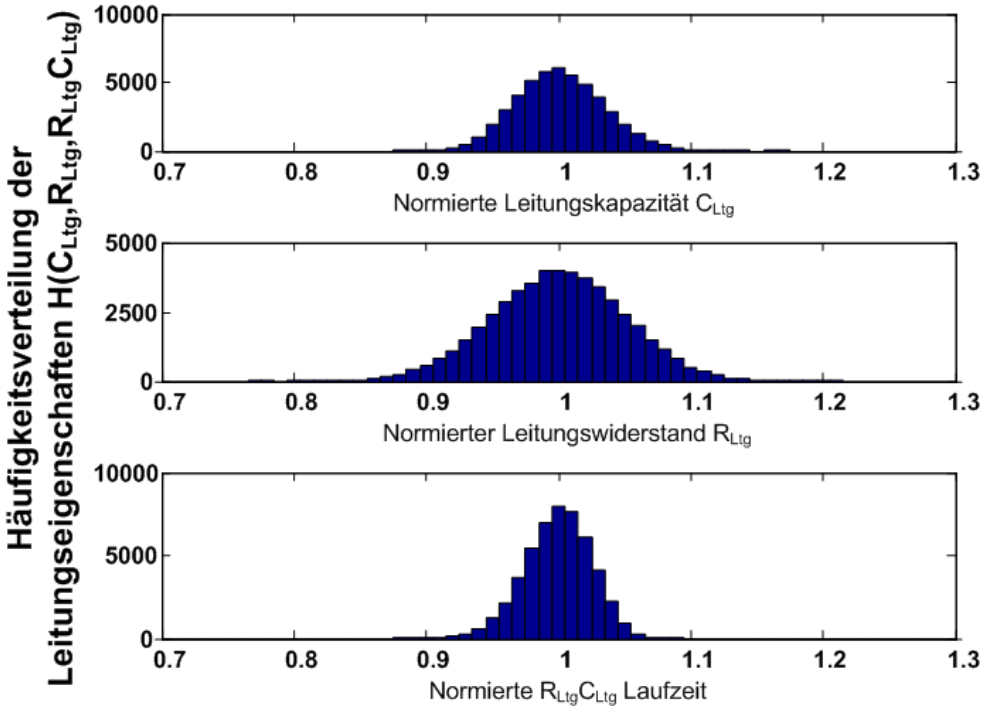


Bild 3.5: Häufigkeitsverteilung von Leitungskapazität, -widerstand und RC Laufzeit der Verdrahtung (nach Glg. 3.6, 3.7) in 65nm CMOS unter der Annahme globaler, normalverteilter Geometrieschwankungen.

Die geometrieabhängige Schwankung von R_{Ltg} und C_{Ltg} wird neben Lithographie-Effekten hauptsächlich durch Dishing und Erosion verursacht. Diese wirken sich auf die Dicke der Leiterbahn und somit auf die Kapazität und den Widerstand der Leitung aus. Sie treten beim chemisch-mechanischen Polierschritt (Chemical Mechanical Polishing CMP) auf und sind abhängig von der Dichte und Breite der Leiterbahnen. Somit ergibt sich eine vorwiegend lokale, systematische Schwankung in der Verdrahtung.
Statistische Effekte wie LER in der Verdrahtung nehmen aufgrund von Mittelungseffekten mit der Länge der Leitung ab. Schwankungen in kurzen Leitungen bewirken vernachlässigbar kleine Laufzeitschwankungen, da deren Kapazität und Widerstand nur einen geringen Anteil zur Gatter- bzw. Pfadlaufzeit beitragen.

Die Ergebnisse der in dieser Arbeit durchgeführten Untersuchungen zeigen, dass die Schwankungen von Gatelänge L und Einsatzspannung V_T für ca. 90% der Laufzeitschwankungen im gesamten produktrelevanten Betriebsbereich ($V_{DD} \in [V_{DD_{nom}} - 300mV; V_{DD_{nom}}]$) verantwortlich sind. Ähnliche Erkenntnisse werden in [40, 47, 48] berichtet. Bild 3.6 zeigt die Ergebnisse einer Monte-Carlo Simulation eines extrahierten NAND2-NOR2 Pfades in 65nm CMOS Technologie für den originalen sowie einen reduzierten Parametersatz bei nomineller Betriebsspannung. Anstelle von 17 global schwankenden Parametern beinhaltet der reduzierte Parametersatz lediglich die globale Schwankung der Gatelänge L und der Einsatzspannung V_T. Der Parametersatz der lokal schwankenden Größen bleibt unverändert und beinhaltet die schwankenden Größen V_T und μ. Es ist deutlich erkennbar, dass sich die Schwankungsbreite der Pfadlaufzeit für den reduzierten Parametersatz nur geringfügig gegenüber der Verwendung des vollständigen Parametersatzes ändert.
Die weiteren Untersuchungen beschränken sich daher im Folgenden auf die Betrachtung

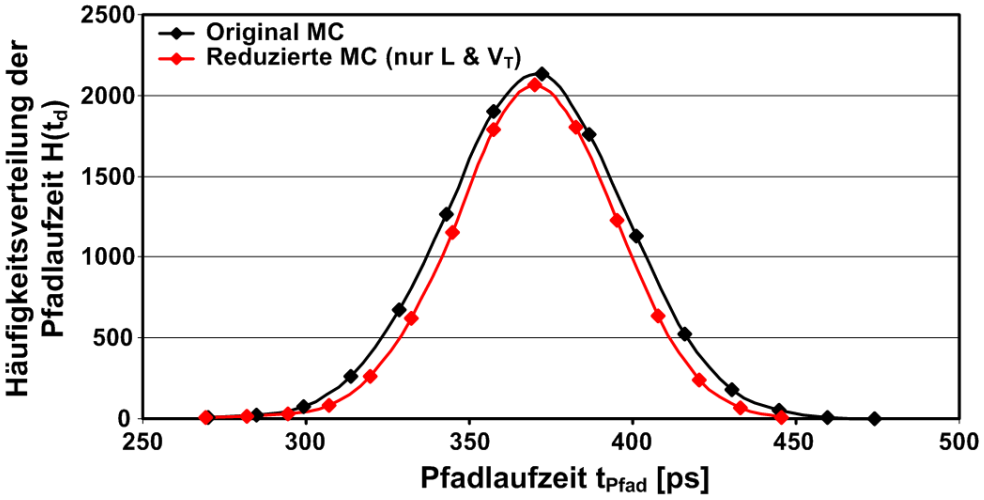


Bild 3.6: Vergleich von Monte Carlo Simulationen mit standardmäßigem und reduziertem Parametersatz ($V_{DD} = V_{DD}^{nom}$, T=27°C).

der drei wichtigsten Transistorparameter L, V_T und μ.

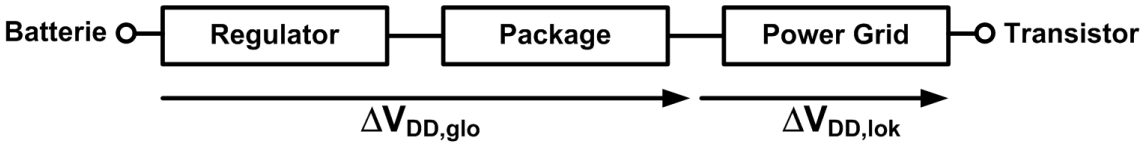


Bild 3.7: Schematische Darstellung der einzelnen IR-Drop Komponenten.

3.1.2 Umgebungsvariationen

Neben Prozessvariationen beeinflussen auch Umgebungsvariationen das Schaltverhalten von Transistoren. Diese Variationen, die während des Betriebs der Schaltung auftreten, sind im Wesentlichen von Betriebsparametern (z.B. V_{DD}, T) und der Schaltungstopologie abhängig. Während des Schaltungsentwurfs kann z.B. über die Dimensionierung der Leitungen Einfluss auf die Sensitivität der Schaltung gegenüber Umgebungsvariationen genommen werden. Das Schaltverhalten während des Betriebs der Schaltung, d.h. Schalttaktivität, Schaltprofile (switching pattern) etc. ist allerdings nicht bekannt. Die Emulation bzw. Modellierung all dieser Parameter während des Designs ist aus Komplexitätsgründen nicht möglich. Aus diesem Grund spricht man von einem pseudo-statistischen Charakter der Umgebungsvariationen. Im Folgenden werden die wichtigsten Quellen von Umgebungsvariationen diskutiert:

- **IR-Drop:** $V_{DD} = V_{DD,nom} + \Delta V_{DD,glo} + \Delta V_{DD,lok}$

  Der IR-Drop, d.h. der Spannungsabfall ΔV an einem Widerstand bei Stromfluss, entlang von Versorgungsleitungen zum einzelnen Transistor, reduziert die effektive Betriebsspannung des Transistors und damit den Gate-Overdrive $V_{DD} - V_T$. Der Spannungsabfall ist abhängig vom Schaltverhalten der gesamten Schaltung (Aktivität der Schaltung), sowie von Unterschieden in der lokalen Leistungsdichte [49]. Die effektive Betriebsspannung des Transistors setzt sich aus der externen Versorgungsspannung (Batterie), den Pegelverlusten am externen Spannungsregler und an den Widerständen des 'Package' $\Delta V_{DD,glo}$, sowie an den Leitungswiderständen $\Delta V_{DD,lok}$ zusammen. Bild 3.7 zeigt schematisch alle Komponenten des Spannungseinbruchs bis zum einzelnen Transistor der Schaltung.

  Die Höhe des Spannungseinbruchs ist abhängig von der Positionierung der Schaltung relativ zur Package-Chip Schnittstelle [50, 51]. Die am Transistor anliegende effektive Versorgungsspannung wird durch diesen Spannungsabfall verringert und beeinflusst aufgrund des reduzierten Gate-Overdrives $V_{DD} - V_T$ das Schaltverhalten des Transistors und somit auch die Laufzeit von Pfaden. Dabei dient der Zyklusmittelwert des Spannungseinbruchs als Hauptindikator bezüglich des Einflusses auf die Pfadlaufzeit [51].
  Der Spannungsabfall in der Versorgungsspannung resultiert im Taktverteilungsnetz (Clock Tree) in zeitabhängigen, zeitlichen Unausgeglichenheiten des Clock Trees. Dies führt neben einem Intra-Zyklus Anteil des Clock Skews zu einem von Zyklus zu Zyklus unterschiedlichen Spannungsabfall an den Clock-Buffern. Dieser Clock Jitter hat eine Vergrößerung bzw. Verkleinerung der effektiven Taktperiode zur Folge.
  Neben diesem resistiven Spannungsabfall ist es möglich, dass über starke Lastwechsel und damit verbundenen Sprüngen im Stromverlauf, das Versorgungsnetz zu einer Schwingung angeregt wird. Man spricht vom $\frac{dI}{dt}$ Effekt. Die Dämpfung der angereg-

Tabelle 3.1: Relevanter Temperaturbereich von CMOS Digitalschaltungen.

Anwendung	Minimale Temperatur	Maximale Temperatur
Automobilelektronik	-40°C	125°C
Mobiltelefon	-20°C	85°C

ten Schwingung ist dabei vom Widerstand, der Kapazität sowie der Induktivität der Schaltung abhängig. Für die Betrachtung der Laufzeitschwankung kann dieser Effekt im Allgemeinen jedoch vernachlässigt werden [51].

- **Betriebstemperatur: T:**

  Sowohl die Einsatzspannung der Transistoren als auch die Beweglichkeit der Ladungsträger zeigt eine deutliche Temperaturabhängigkeit [52, 53]:

$$\mu(T) = \mu(T_0) \cdot \left(\frac{T}{T_0}\right)^{-k_\mu} \tag{3.9}$$

$$V_T(T) = V_T(T_0) - \gamma_{V_T} \cdot (T - T_0) \tag{3.10}$$

  T bezeichnet die Temperatur [°K], T_0 die Temperatur zum Zeitpunt t=0. k_μ ist der Beweglichkeits-Temperaturkoeffizient und liegt zwischen 1.2 und 2.0. γ_{V_T} bezeichnet den Temperaturkoeffizienten der Einsatzspannung.
  Mit zunehmender Temperatur nimmt die Beweglichkeit der Ladungsträger im Kanal aufgrund erhöhter Streuungen an Phononen ab und verringert somit den Transistorstrom. Gleichzeitig führt die Zunahme der thermischen Energie zu einem reduzierten Fermi-Potential ϕ_F, das eine verringerte Einsatzspannung zur Folge hat und der beweglichkeitsbedingten Abnahme des Transistorstroms entgegenwirkt. Je nach Versorgungsspannung V_{DD} hat eine der beiden Größen einen dominierenden Einfluss auf die Laufzeit. Für hohe Versorgungsspannungen dominiert das Temperaturverhalten der Beweglichkeit, für niedriges V_{DD} das Temperaturverhalten der Einsatzspannung. Bei einer bestimmten Versorgungsspannung V_{DD} kompensieren sich beide Beiträge, und die Laufzeit bleibt unverändert. Dieser Punkt wird als Zero Temperature Coefficient Point (ZTCP) bezeichnet und liegt für 90nm CMOS bei ca. 0.85V [53].
  Die Temperatur kann als globaler Betriebsparameter betrachtet werden. Signifikante Temperaturunterschiede treten vorwiegend zwischen Prozessorkern und Speicher bzw. Peripherie, sowie zwischen einzelnen Mikroprozessorkernen (Multi-Core Design) auf. Selbst für diese räumlich getrennten Komponenten werden in den häufigsten Fällen Temperaturgradienten von lediglich 10°C gemessen [54, 55]. Die meisten zeitkritischen Strukturen befinden sich im Prozessorkern und weisen keine großen räumlichen Distanzen auf [56], so dass insbesondere in kleinen Schaltungen keine signifikanten lokalen Temperaturgradienten zu erwarten sind.
  Neben dem direkten Einfluss der Temperatur auf die Laufzeit von Gattern und Pfaden führt eine erhöhte Temperatur zur verstärkter Alterung durch NBTI (Negative Bias Temperature Instability, siehe 3.1.3). Tabelle 3.1.2 zeigt für verschiedene Anwendungsbereiche die für den Betrieb der Schaltung relevanten Temperaturbereiche.

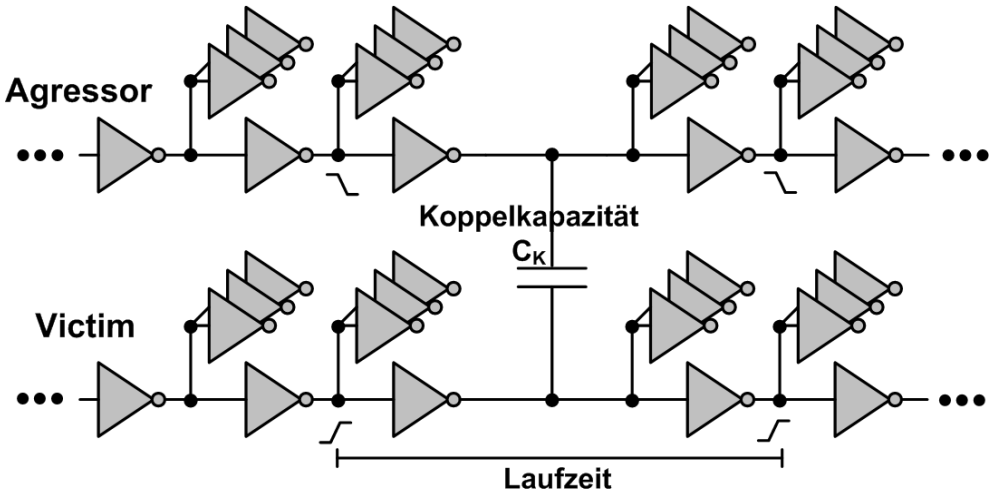


Bild 3.8: Schematische Darstellung eines Crosstalk relevanten Netzes in einem Fanout-4 Inverter Pfad.

- **Übersprechen/Crosstalk:**

  Die kapazitive Kopplung zwischen zwei Netzen führt beim Schaltvorgang zur gegenseitigen Beeinflussung des Potentials durch Ladungsverschiebung, d.h. der Signalwechsel an einem der Netze (Aggressor) beeinflusst den Spannungspegel am ruhenden bzw. ebenfalls schaltenden anderen Netz (Victim). Bild 3.8 zeigt eine schematische Anordnung zweier kapazitiv gekoppelter Netze innerhalb eines Fanout-4 Inverter Pfades.

  Liegt das schaltende Victim-Netz in einem kritischen Pfad, so beeinflusst der Crosstalkeffekt die minimal mögliche Taktperiode der Schaltung. Für gleichgerichtete Schaltvorgänge am Victim- und Aggressor-Netz verringert sich die Laufzeit, bei entgegengesetzten Transitionen erhöht sich die Laufzeit des Pfades im Vergleich zu einem ruhenden Aggressor-Netz. Der Einfluss auf die veränderte Laufzeit des Victim Netzes lässt sich unter Annahme eines gleichzeitigen Schaltvorgangs von Victim und Aggressor wie folgt beschreiben [43]:

$$\Delta t_D \sim \frac{C_{K_{eff}}}{C_{K_{eff}} + C_{stat}} \cdot \frac{1}{t_{Sig}^{Agr}} \tag{3.11}$$

  $C_{K_{eff}}$ bezeichnet die effektive Koppelkapazität, C_{stat} die statische Kapazität des Victim Netzes, die sich aus der Leitungskapazität C_{Ltg} zu nicht schaltenden Netzen sowie den Diffusionskapazitäten von VP1, VN1 und den Gatekapazitäten von VP2, VN2 zusammensetzt. Die Koppelkapazität beinhaltet den erhöhten Wert der im statischen Fall wirkenden Kapazität C_K, deren effektiver Wert durch den Miller-Effekt bei schaltendem Aggressor steigt [57]. Dabei hängt der Miller-Effekt von der Steilheit der Signalflanken ab, d.h. je steiler die Signalflanke des Aggressornetzes t_{Sig}^{Agr}, desto höher die Laufzeiterhöhung [58]. Zusätzlich hängt die Größe der Laufzeitänderung vor allem von der zeitlichen Synchronität der Signalwechsel auf Aggressor und Victim-Netz ab (Relative Signal Ankunftszeit RSAT). Generell gilt, je größer der zeitliche Abstand zwischen den jeweiligen Signalwechseln, desto kleiner die Laufzeitänderung [59].

  Bild 3.9 zeigt die in 65nm simulierte Laufzeiterhöhung bei entgegengesetzt schaltenden Aggressor- und Victim-Netzen in Abhängigkeit der zeitlichen Synchronität

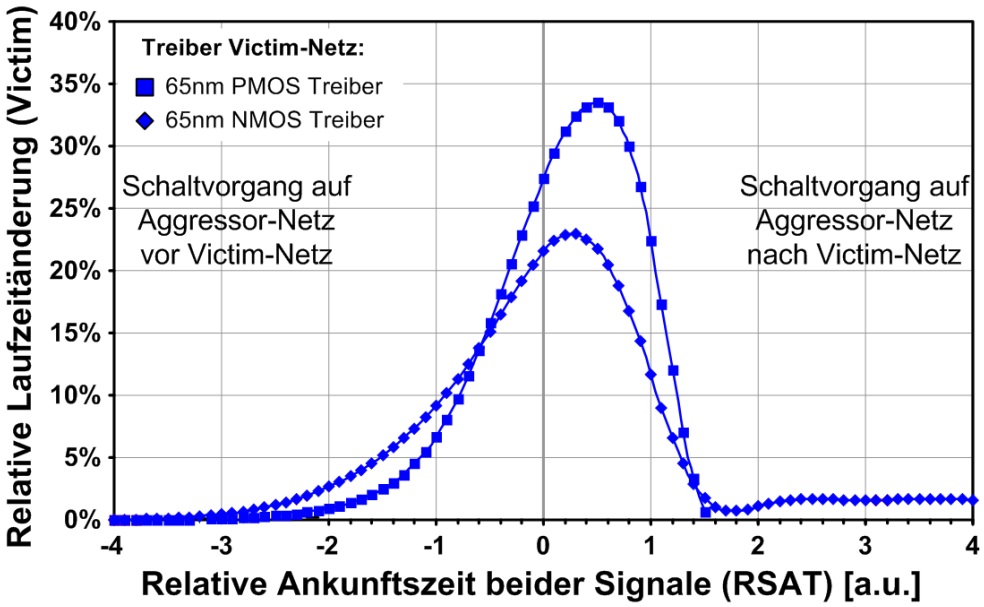


Bild 3.9: Simulierte crosstalk-bedingte Laufzeiterhöhung der in Bild 3.8 gezeigten Testschaltung in 65nm CMOS ($V_{DD} = V_{DD,nom} + 10\%$, T=27°C).

beider Signalwechsel. Bei schwacher Flanke am Victim-Netz (PMOS-Treiber) und steiler Flanke am Aggressor-Netz (NMOS Treiber) ist bei zeitlicher Synchronität der Signalwechsel ein deutlich erhöhter Einfluss auf die Laufzeit zu erkennen als im umgekehrten Fall. Dies verdeutlicht die Abhängigkeit des Effektes von der Flankensteilheit an den gekoppelten Netzen.
Bild 3.10 zeigt die gemessene Frequenz eines Ringoszillators mit Crosstalk-Struktur, d.h. kapazitiv gekoppelten Leitungsstrukturen als Lastelemente, in 45nm CMOS Technologie bei deaktiviertem und aktiviertem Aggressor. Bei steigender Versorgungsspannung erhöht sich die Steilheit der Signalflanken und die effektive Last am Crosstalk-behafteten Netz erhöht sich aufgrund des stärker wirkenden Miller-Effektes. Somit nimmt die relative Frequenzabnahme mit steigendem V_{DD} zu.

Insbesondere Netze mit starker Kopplung zu benachbarten Signalleitungen und einem geringen Anteil der Gatterlast (Fanout) sind potentiell anfällig für Crosstalk. Derartige Netzstrukturen sind vor allem in Bussen und im Taktbaum zu finden, die Signale über lange Strecken auf dem Chip verteilen. Für Leitungen, die das Taktsignal führen, werden deshalb alle benachbarten Leitungen oftmals mit zweifachem Minimalabstand (double spacing) positioniert, um den Anteil der Koppelkapazität an der Gesamtkapazität des Netzes zu verringern [60]. Somit wird ein Crosstalk-induzierter Beitrag zum Clock Jitter deutlich reduziert. Für kritische Strukturen im Taktbaum und in Busstrukturen werden zahlreiche Maßnahmen wie z.B. Gate-Sizing (Anpassung der Gattertreiberstärken an Victim- und Aggressor-Netz) [61], relative Signalverzögerung [62], Abschirmung (Shielding) [63], Kodiertechniken für Busstrukturen [64] usw. angewandt, um den Beitrag der Crosstalk induzierten Laufzeitschwankung gering zu halten. In [65] werden diese und weitere Techniken zusammengefasst und einzeln diskutiert.

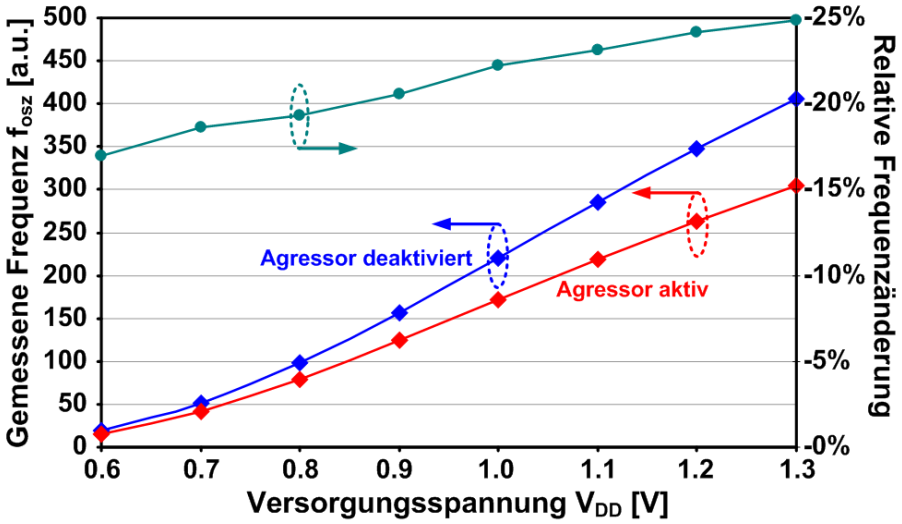


Bild 3.10: Gemessene Frequenz einer Crosstalk-Struktur in 45nm low-power CMOS Technologie bei T=27°C.

3.1.3 Alterungseffekte

Neben statischen Prozessvariationen verändern Alterungseffekte die Transistoreigenschaften und beeinflussen somit Gatter- und Pfadlaufzeiten. Die für die Geschwindigkeit der Schaltung relevantesten Alterungseffekte werden im Folgenden diskutiert:

- **Negative/Positive Bias Temperature Instability (NBTI/PBTI):**

  NBTI ist die betragsmäßige Erhöhung der Einsatzspannung von PMOS-Transistoren in Inversion. Insbesondere bei erhöhten Temperaturen brechen Si-H Bindungen an der Oxid-Kanal Schnittstelle auf und es entstehen freie kovalente Bindungen (Interface Traps) sowie feste Oxidladungen. Diese führen zur Verschiebung der Einsatzspannung. Haupteinflussgrößen sind die elektrische Feldstärke an der Oxid-Kanal Grenzfläche und somit die Versorgungsspannung V_{DD}, sowie die Betriebstemperatur T und die Betriebszeit t. Der Effekt verstärkt sich, wenn V_{DD} und T gleichzeitig hohe Werte annehmen, was vorwiegend im High-Performance Modus von Digitalschaltungen der Fall ist [66]. Je größer V_{DD} und T, desto stärker die Zunahme von V_T. Die Einsatzspannungserhöhung setzt sich aus einem irreversiblen und einem reversiblen Anteil zusammen. Wird der Transistor vom Stress ($|V_{GS}| > 0$) befreit, d.h. es liegt keine Spannung zwischen Gate- und Source an ($V_{GS} = 0$), so startet der Ausheilprozess des reversiblen Anteils. Auch hier gilt: Je höher die Temperatur, desto schneller reduziert sich der reversible Anteil der Einsatzspannungserhöhung [67]. Für neue high-k Technologien und dem damit verbundenen Anstieg der elektrischen Feldstärke im Gate-Dielektrikum stellt neben NBTI, auch PBTI bei NMOS Transistoren eine weitere Variationsquelle dar. Für existierende SiON Technologien ist PBTI jedoch vernachlässigbar [68].

- **Channel Hot Carrier (CHC):**

  Ladungsträger im Kanal, die über das horizontale Drain-Source Feld beschleunigt werden, nehmen kinetische Energie auf, bis sie durch Streuung an Phononen wieder an Energie verlieren. Gleichzeitig lenkt das vertikale elektrische Feld zwischen Gate und Kanal die Ladungsträger zur Oxid-Kanal Grenzfläche ab. Bleibt eine Streuung

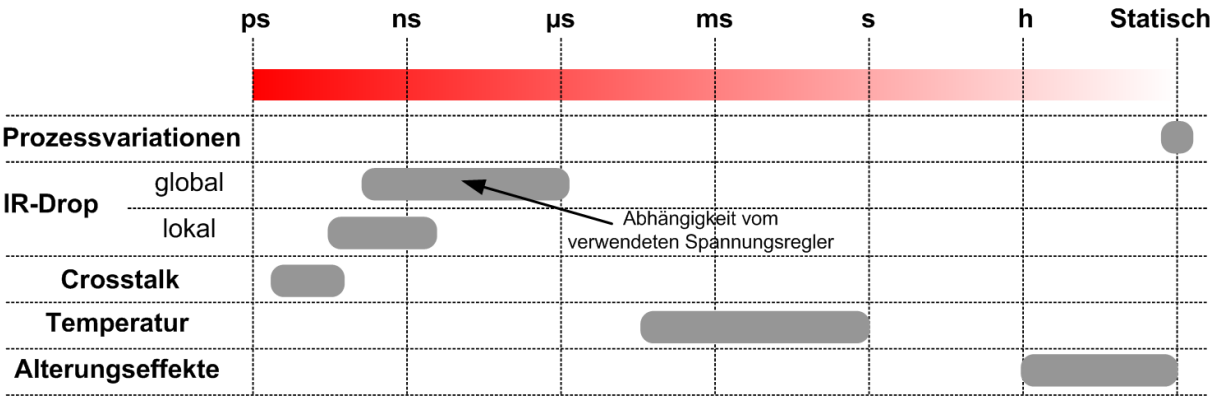


Bild 3.11: Zeitkonstanten von Prozess-, Umgebungsvariationen und Alterungseffekten.

an Phononen aus, so ist es möglich, dass Ladungsträger ausreichend Energie aufnehmen, so dass ein Überwinden der Potentialbarriere zwischen Kanal und Gate möglich ist. Diese Ladungsträger schädigen die Grenzfläche zwischen Oxid und Kanal, was zur Minderung der Beweglichkeit μ führt. Zum anderen werden im Oxid Ladungen eingeschlossen, was eine Verschiebung der Einsatzspannung zur Folge hat. Je höher die lateralen elektrischen Felder und je dicker das Gateoxid, desto höher die Schädigung des Transistors.

3.1.4 Zeitliche Klassifizierung von Variationen

Bild 3.11 zeigt den zeitlichen Wirkungsbereich von Prozess-, Umgebungsvariationen und Alterungseffekten. Während Prozessvariationen gänzlich statisch wirken, bewegen sich die Umgebungsvariationen sowohl im Intra-Zyklus (Crosstalk, IR-Drop) als auch im Multi-Zyklen (Temperatur) Bereich. Für die zeitliche Klassifizierung von IR-Drop Effekten muss zwischen globalen und lokalen Spannungsabfällen unterschieden werden. Während die Schwankungen am Spannungsregler höhere Zeitkonstanten aufweisen können, zeichnen sich lokale Spannungsschwankungen im Versorgungsspannungsnetz durch kurze Zeitkonstanten aus. Alterungseffekte, die mit der Betriebszeit der Schaltung zunehmen, reichen vom langzeitlichen Bereich (>Stunden) bis in den statischen Bereich [69, 70]. Diese unterschiedlichen Zeitkonstanten müssen insbesondere bei der Implementierung adaptiver Techniken zur Kompensation von Variationseffekten berücksichtigt werden, da kurzzeitige Variationen wie Crosstalk und IR-Drop hohe Anforderungen an die Geschwindigkeit adaptiver Maßnahmen stellen. Die Problematik der Kompensation von kurzzeitigen und lokalen Variationen wird in Kapitel 6.1.1 diskutiert.

Um neben den Zeitkonstanten einen Überblick über die Größenordnung der Schwankungsbreite der wichtigsten Einflussparameter zu geben, fasst Tabelle 3.2 die in der Literatur zu findenden Aussagen zur Gatelängen-, Transistoreinsatzspannungs-, Versorgungsspannungs- und Temperaturschwankung zusammen. Eigene Analysen ergeben für die systematische 3σ-Schwankungsbreite von L und V_T eine Schwankungsbreite zwischen 10% und 20%. Für eingebettete Mikroprozessoren, wie die der ARM und MIPS Familie, ist eine on-chip Temperaturschwankung von ca. 10°C zu erwarten. Grund hierfür sind die im Vergleich zu High-Speed Mikroprozessoren wie z.B. von Intel, AMD, IBM, relativ geringen Abmessungen der Schaltung. Eine vergleichbare Aussage über die Größenordnung von Crosstalk und Alterungseffekten ist kaum möglich, da sich die in der Literatur verwende-

Tabelle 3.2: In der Literatur zu findende Aussagen zu relativen Schwankungsbreiten der wichtigsten Einflussparameter.

Parameter	**130nm**	**90nm**	**65nm**	**45nm**
L (3σ)	17% [71] 32% [73]	10% [72] 40% [73]	10% [72] 15% [74] 47% [73]	10% [72]
V_T (3σ)	30% [71] 10% [73]	30% [14] 12% [73]	33% [72] 13% [73]	40% [72]
Max. ΔV_{DD}	10% [72, 14, 71, 73]	10% [72, 14, 71, 73]	10% [72, 71]	10% [71, 75, 73]
Max. On-Chip ΔT	35°C[76], 35°C[75], 10°C[54], 7°C[55]			
Allg. Betriebsbereich T	-40°C bis 130°C			

ten Anwendungsbeispiele und Stressbedingungen signifikant unterscheiden. Somit ist die Vergleichbarkeit nicht gewährleistet und eine den Schwankungsbreiten von Prozessvariationen ähnliche Aussage nicht sinnvoll.

3.2 Sensitivitätsanalyse und technologiebasierte Trendaussagen

3.2.1 Analyse der Laufzeitsensitivität

Sensitivitäten bei nomineller Versorgungsspannung

In zahlreichen Veröffentlichungen der letzten Jahre werden Variationen, insbesondere Prozessvariationen, als Barriere auf dem Weg der CMOS Roadmap gesehen. Variationen sind jedoch keine für sub-100nm Technologien charakteristischen, neuartigen Effekte, sondern sind bereits aus Zeiten von Langkanaltransistoren bekannt [77]. Im Gegensatz zu diesen Technologien haben sich die Herausforderungen bei der Herstellung moderner CMOS Technologien geändert. Insbesondere die Belichtung von Strukturgrößen kleiner als die Wellenlänge des für die Lithographie verwendeten Lichts erschwert die sichere und reproduzierbare Darstellung der kritischen Minimalstrukturen. Die Erkenntnisse dieser Arbeit und neue Veröffentlichungen zeigen jedoch, dass ein besseres Verständnis der Variationsquellen und verbesserte Prozesstechniken zu erheblichen Fortschritten in der Prozesskontrolle führen. Die Schwankungsbreite der Gatelänge skaliert daher im gleichen Verhältnis wie die nominellen Strukturgrößen. Die relative Schwankung der Gatelänge bleibt somit konstant [30].
Trotz dieser Verbesserungen nehmen die Laufzeitschwankungen moderner Technologien zu. Zur weiteren Untersuchung dieses Zusammenhangs wird im Folgenden das kanonische Laufzeitmodell verwendet. Da für dieses Modell statistisch unabhängige Parameter x_i erforderlich sind, wird die Korrelation der einzelnen Parameter über eine Hauptkomponentenanalyse (Principal Component Analysis PCA) entfernt [78], so dass die Laufzeit wie folgt modelliert werden kann:

$$t_d = t_d^{nom} + \sum_i^N S_i \cdot \Delta x_i = t_d^{nom} + \sum_i^N \frac{\partial t_d^{nom}}{\partial x_i} \cdot \Delta x_i \tag{3.12}$$

Dabei bezeichnet t_d^{nom} die nominelle Laufzeit des Pfades, S_i die Sensitivität der Laufzeit gegenüber dem Parameter x_i (Gatelänge L, Einsatzspannung V_T etc.) und Δx_i die Auslenkung des Parameters x_i aus dem nominellen Wert. N ist die Anzahl der zu berücksichtigenden Parameter.
Die Laufzeit wird über die Taylorentwicklung approximiert, wobei die Sensitivität S_i dem ersten Glied der Taylorapproximation entspricht. Tabelle 3.3 zeigt den relativen Fehler einer linearen Taylorapproximation für eine globale 3σ Auslenkung der Transistorparameter L, V_T und μ in 90nm, 65nm und 40nm CMOS Technologie. Eine Vernachlässigung der Terme höherer Ordnung ist aufgrund der nur sehr kleinen Abweichungen vom simulierten Wert gerechtfertigt. Eine Berücksichtigung der quadratischen Terme ist erst ab Schwankungsbreiten von ±30% notwendig [79]. Eigene Untersuchungen und die in der Literatur zu findenden Angaben zu Schwankungsbreiten, wie z.B. ±10% Gatelängenschwankung zeigen, dass derartige Abweichungen vom nominellen Wert nicht erreicht werden. Im Gegensatz zu L, V_T und μ reicht für die Laufzeitsensitivität gegenüber V_{DD} Schwankungen der lineare Anteil nicht aus, da sich der Betriebsbereich moderner low-power Schaltungen über mehrere 100mV erstreckt. Deshalb werden im Folgenden die Laufzeitsensitivitäten bei nomineller und um 300mV reduzierter Versorgungsspannungen untersucht.
Um einen Vergleich der Sensitivitäten verschiedener Technologieknoten zu ermöglichen,

Tabelle 3.3: Relativer Fehler der linearen Approximation für eine 3σ Auslenkung der wichtigsten Transistorparameter.

Technologie	L	$V_{T,N}$	$V_{T,P}$	μ_N	μ_P
90nm	0.2%	1.2%	0.9%	0.0%	0.2%
65nm	0.9%	0.4%	0.5%	0.2%	0.2%
40nm	0.2%	0.6%	1.0%	0.0%	0.0%

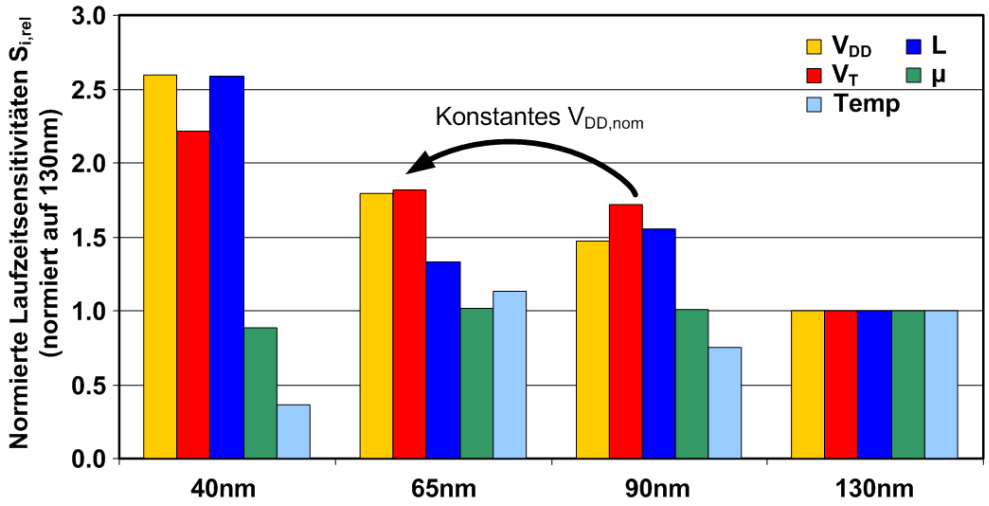


Bild 3.12: Laufzeitsensitivitäten einer NAND2-NOR2 Kette gegenüber L, V_T, μ und V_{DD} Schwankungen.

werden im Folgenden die relativen Laufzeitsensitivitäten betrachtet.

$$t_d = t_d^{nom} + \sum_i^N S_i \cdot \Delta x_i = t_d^{nom} \cdot \left(1 + \sum_i^N \frac{1}{t_d^{nom}} \cdot \frac{\partial t_d^{nom}}{\partial x_i} \cdot \Delta x_i\right) = t_d^{nom} \cdot \left(1 + \sum_i^N S_{i,rel} \cdot \Delta x_i\right) \tag{3.13}$$

Damit ergibt sich für $S_{i,rel}$:

$$S_{i,rel} = \frac{S_i}{t_d^{nom}} \tag{3.14}$$

Die Laufzeitschwankung hängt sowohl von der Auslenkung der Parameter Δx_i aus dem nominellen Fall als auch von der Laufzeitsensitivität des Pfades gegenüber Parameterschwankungen ab. So kann trotz skalierender Parameterschwankungen aufgrund von erhöhten Sensitivitäten eine stärkere Schwankung der Laufzeit erfolgen.
Bild 3.12 zeigt die Sensitivität der Pfadlaufzeit einer NAND2-NOR2 Kette für CMOS Technologien von 130nm bis 40nm. Es ist deutlich erkennbar, dass sich die Sensitivität der Laufzeit gegenüber Schwankungen der Transistor- und Betriebsparameter mit fortschreitender Skalierung signifikant erhöht. Der kleine Anstieg der Sensitivitäten von 90nm auf 65nm ist auf eine konstante nominelle Versorgungsspannung und somit nahezu gleichbleibenden Gate-Overdrive $V_{DD} - V_T$ zurückzuführen.
Die Sensitivität gegenüber der Gatelängenschwankung hängt stark von den Kurzkanaleffekten in der jeweiligen Technologie ab. Ein starker V_T roll-off und ein großer DIBL verstärken den Einfluss der Gatelänge auf die Laufzeit. Die Laufzeitsensitivität gegenüber

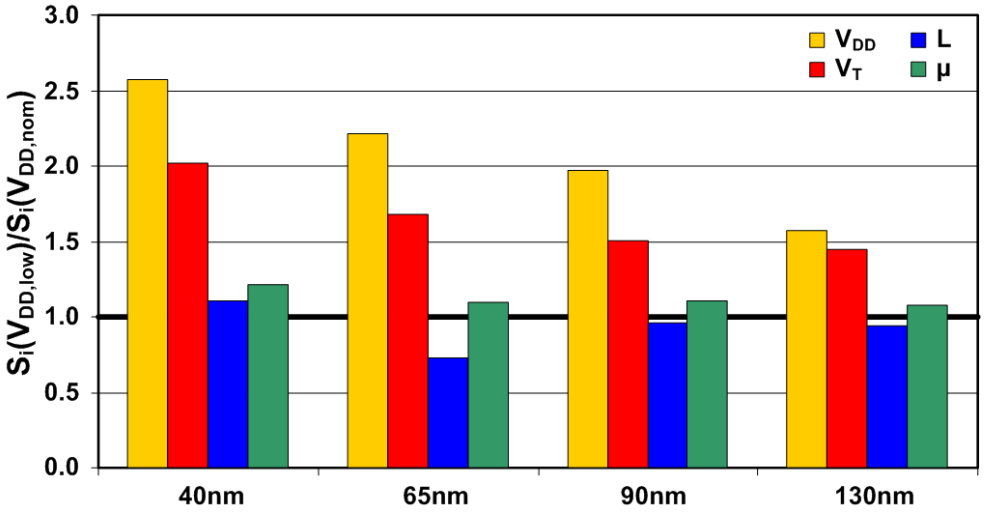


Bild 3.13: Änderung der Laufzeitsensitivitäten einer NAND2-NOR2 Kette bei reduzierter Versorgungsspannung.

der Beweglichkeit μ bleibt über alle untersuchten Technologiegenerationen hinweg nahezu konstant und kann somit als technologieunabhängig angesehen werden. Die Sensitivität gegenüber der Temperatur schwankt in dieser Darstellung deutlich von einem Technologieknoten zum nächsten. Da der Einfluss der Temperatur abhängig von der Versorgungsspannung ist, ist ein Trend hinsichtlich der Laufzeitsensitivität gegenüber Temperaturschwankungen aus dieser Darstellung nicht ersichtlich. Hierzu bedarf es einer erweiterten Betrachtung bei variierender Versorgungsspannung, wie in Kapitel 3.2.2 zu sehen ist.

Sensitivitäten bei Spannungsskalierung

Low-power Schaltungen, die stringente Spezifikationen hinsichtlich der Energieaufnahme zu erfüllen haben, nutzen zur Energieersparnis die dynamische Anpassung der Versorgungsspannung (Dynamic Voltage Scaling DVS) an die sich zeitlich ändernden Geschwindigkeitsanforderungen. Daraus ergibt sich ein Betriebsbereich für die Versorgungsspannung, der sich bis ca. 300mV unter der nominellen Versorgungsspannung erstreckt. Auch in Bereichen niedrigerer Versorgungsspannungen müssen festgelegte Taktfrequenzen gewährleistet werden. Aus diesem Grund ist es wichtig, den Einfluss von Variationen auch für geringere Versorgungsspannungen zu untersuchen.

Bild 3.13 zeigt das Verhältnis der relativen Laufzeitsensitivitäten für nominelle und eine um 300mV reduzierte Versorgungsspannung. Der Einfluss der Gatelängenschwankung nimmt im Vergleich zur nominellen Versorgungsspannung ab, d.h. die relative Laufzeitschwankung aufgrund variierender Gatelängen nimmt für verringertes V_{DD} ab. Dies ist durch einen geringeren V_T roll-off und reduzierten DIBL bei niedrigeren Versorgungsspannungen zu erklären. Die Sensitivität gegenüber der Beweglichkeit zeigt über alle Technologieknoten hinweg keine signifikante Abhängigkeit von der Versorgungsspannung. Im Gegensatz dazu ist die Laufzeitsensitivität gegenüber V_{DD} und V_T Schwankungen bei niedriger Versorgungsspannung deutlich erhöht.
Betrachtet man alle Prozessvariationen in Kombination, so führt die deutliche Erhöhung von $S_{V_T,rel}$ trotz reduziertem $S_{L,rel}$ zu einer insgesamt steigenden Laufzeitschwankung bei reduzierten Versorgungsspannungen. Bei fortschreitender Technologieskalierung er-

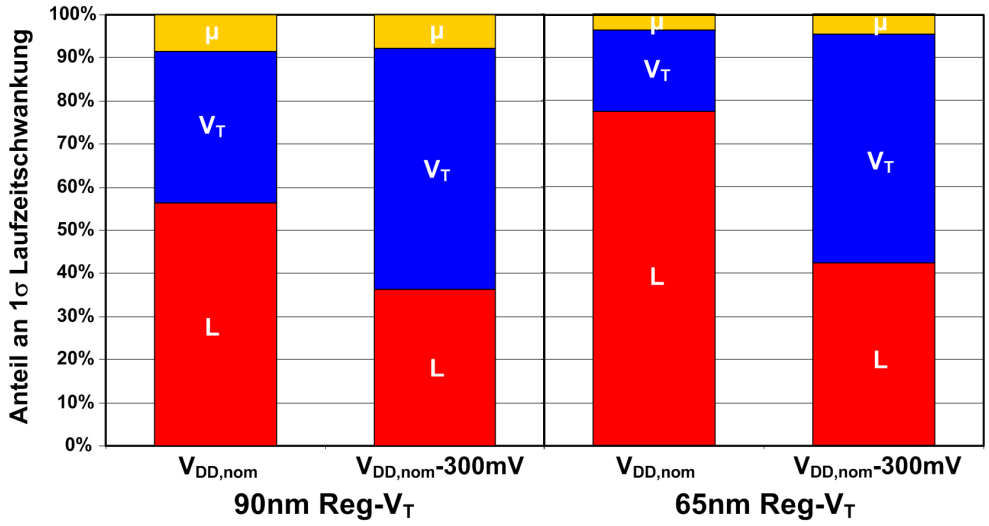


Bild 3.14: Anteile an der von L, V_T und μ Variationen induzierten Laufzeitschwankung.

höht sich der Unterschied der Laufzeitsensitivitäten zwischen nomineller und reduzierter Versorgungsspannung weiter.

Bild 3.14 zeigt die Anteile der durch globale L, V_T und μ Variationen bedingten relativen Laufzeitschwankung für einen ND2-NR2 Pfad in 65nm und 90nm reg-V_T CMOS. Die Dominanz der Gatelängenschwankung bei nomineller Versorgungsspannung ist deutlich erkennbar. Die gleichzeitige Prozessierung von PMOS und NMOS Gate führt zu einer korrelierten Schwankung der PMOS und NMOS Gatelänge. Im Gegensatz dazu sind die Implantationen von PMOS und NMOS Transistor zwei verschiedene Prozessschritte, so dass beide Einsatzspannungen unabhängig voneinander schwanken. Für verringerte Versorgungsspannungen erhöht sich aufgrund des reduzierten Gate-Overdrives $V_{DD} - V_T$ die V_T bedingte Laufzeitschwankung und die V_T Variation stellt den größten Anteil. In 65nm CMOS steigt der Gatelängen-bedingte Anteil an der Laufzeitschwankung auf fast 80%. Dieser große Anteil resultiert nicht aus einer höheren Gatelängenschwankung, sondern aus einer deutlich reduzierten globalen Schwankungsbreite der Einsatzspannung aufgrund verbesserter Prozesskontrolle. Bei nomineller Versorgungsspannung trägt die globale V_T Variation nur noch zu 19% zur Laufzeitschwankung bei. Aufgrund der im Vergleich zu 90nm erhöhten Spannungssensitivität in 65nm stellt die V_T Variation bei reduziertem V_{DD} erneut den größten Anteil an der Laufzeitschwankung.
Je kleiner der Gate-Overdrive $V_{DD} - V_T$, desto größer ist sowohl die Laufzeit t_d als auch die Laufzeitschwankung aufgrund gestiegener Sensitivitäten gegenüber V_{DD} und V_T Variationen [40, 45, 80].

Im Folgenden wird der Einfluss von Prozess-, Umgebungsvariationen und Alterungseffekten auf die Geschwindigkeit von digitalen Schaltungen bei fortschreitender Technologieskalierung diskutiert.

3.2.2 Schwankungen bei fortschreitender Technologieskalierung

Prozessvariationen:

- **Transistorparameter (FEOL):**

  Wie bereits gezeigt wurde, führen Verbesserungen der Prozesstechnik und die bessere Kenntnis über die Ursachen von Prozessvariationen zu mit der Technologie skalierenden Gatelängenschwankungen und verbesserten V_T Schwankungen [30, 81, 82]. In [30] werden systematische Schwankungen der Einsatzspannung durch das Einfügen von Dummy-Gates verringert. Die daraus resultierende homogene Verteilung der Transistor-Gates im Layout reduzieren die Temperaturgradienten während des RTA, so dass eine relativ homogene V_T-Verteilung erzielt wird. Für SRAM Zellen, bestehend aus den kleinsten Transistoren im Layout, wurde durch verbessertes Zell-Layout ein erhöhter SNM Wert bei voller Flächenskalierung erreicht. Durch Verbesserungen beim CMP Prozessschritt wurde eine signifikante Verkleinerung der Leitungswiderstandsschwankung erzielt. In [81] werden spezielle Füllstrukturen ins Layout eingebracht, um die Temperaturhomogenität während des RTA Schritts zu erhöhen. Gemessene Ringoszillatorfrequenzen zeigen eine um 30% reduzierte WID Laufzeitschwankung. Diese Beispiele zeigen, dass die stetige Reduzierung systematischer Effekte ein variationsbedingtes Ende der CMOS Roadmap aus heutiger Sicht noch nicht erkennbar machen. Vielmehr besteht bei der Produktion der Anspruch, Grenzen der maximalen Schwankungsbreite zu gewährleisten, um eine Technologie überhaupt als produkttauglich qualifizieren zu können.

  Neben systematischen (kontrollierbaren) Schwankungen nimmt der Einfluss statistischer Schwankungen mit fortschreitender Technologieskalierung zu. Aufgrund der kontinuierlichen Verkleinerung der Transistorgeometrien nähert sich die CMOS Technologie atomaren Grenzen. So nimmt bei der Implantation des V_T Dotierprofils der Einfluss jedes einzelnen Dotierstoffatoms zu, da die Gesamtanzahl der Dotierstoffatome stetig abnimmt. Bild 3.15 zeigt für verschiedene Technologieknoten die relative Zunahme der gemessenen statistischen Einsatzspannungsschwankung aufgrund von RDF normiert auf 180nm Bulk CMOS.

  In 65nm und 45nm Technologien tragen die statistischen Variationen zu mehr als 50% der gesamten Einsatzspannungsschwankung eines einzelnen Transistors bei [30]. Die Größenordnung der statistischen Schwankung hat jedoch neben der in Gleichung 3.5 aufgeführten Parameter noch andere systematische Einflussgrößen. In [83] konnte die statistische V_T Schwankung um 40% reduziert werden, indem der Implantationswinkel der Halo-Dotierung um 15 Grad reduziert wurde. In [84] wurde durch ein spezielles Implantationsverfahren eine um 15% reduzierte statistische V_T Schwankung für 20nm NMOS Transistoren erzielt. Diese Beispiele zeigen, dass systematische Effekte existieren, die Einfluss auf die Schwankungsbreite statistischer Variationen haben, d.h. auch hinsichtlich statistischer Variationen führt besseres Verständnis und erhöhte Prozesskontrolle zur Verringerung der Schwankungsbreiten, wenngleich insgesamt eine Erhöhung der statistischen Variationen zu erwarten ist.
  Aber auch fundamentale Eingriffe in die Transistorstruktur tragen zur Abnahme von statistischen Variationen bei. Die bekannteste Maßnahme zur Reduzierung von Kurzkanaleffekten und Verbesserung der Transistoreigenschaften ist der Einsatz von high-k Materialien [85]. Durch die Erhöhung der Gatekapazität wird die Mismatchkonstante reduziert, so dass statistische Variationen abgeschwächt werden.

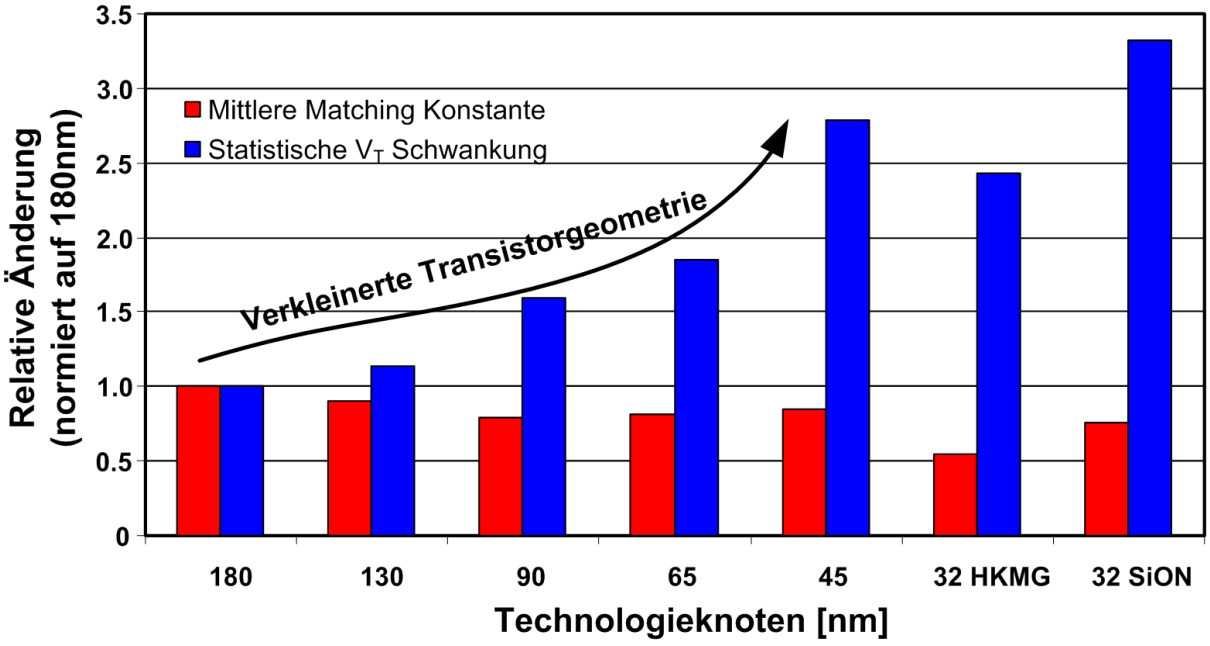


Bild 3.15: Skalierungsverhalten der statistischen Einsatzspannungsschwankung.

Der Einsatz von Multi-Gate Transistoren (MUGFET) stellt eine weitere Möglichkeit dar, statistische Variationen zu reduzieren [30]. Die Einsatzspannung der Multi-Gate Transistoren wird vorwiegend über die Austrittsenergie des Gate-Materials eingestellt. Das Substrat wird im Vergleich zum Bulk-CMOS geringer dotiert bzw. bleibt undotiert. Zusätzlich werden durch die verbesserte Kanalkontrolle die Kurzkanaleffekte reduziert [86, 87]. Neben offenen Fragen hinsichtlich parasitärer Widerstände und Kapazitäten konnte im Gegensatz zu high-k Metal-Gate Transistoren die produkttaugliche Skalierbarkeit dieses Transistortyps, d.h. die Skalierbarkeit bei Bulk CMOS ähnlicher Flächeneffizienz bisher nicht gezeigt werden.
Während eine erhöhte statistische Einsatzspannungsschwankung für die Minimaltransistoren von SRAM Speicherzellen hohe Relevanz hat, ist der Einfluss von statistischen Schwankungen auf die Geschwindigkeit von kritischen Pfaden moderner Mikroprozessoren stark vermindert. Dafür verantwortlich ist eine starke Mittelung der statistischen Schwankungen über die Logikstufenanzahl n_{Log} des kritischen Pfades, sowie über die Größe der Transistoren in den Standardzellen, wie in Kapitel 4.1 näher diskutiert wird.

- **Leitungsstrukturen/Verdrahtung (BEOL):**

  Um das Skalierungsverhalten von Leitungsstrukturen qualitativ zu bewerten werden die aus den Gleichungen 3.6 unf 3.7 bekannten Zusammenhänge vereinfacht dargestellt:

$$C_{Ltg} \sim \varepsilon_{ILD} \frac{T_{Ltg} \cdot L_{Ltg} \cdot W_{Ltg}}{S_{Ltg} \cdot H_{Ltg}} \qquad R_{Ltg} \sim \frac{\rho_{Ltg} \cdot L_{Ltg}}{W_{Ltg} \cdot T_{Ltg}} \tag{3.15}$$

  ρ_{Ltg} ist der spezifische Widerstand des verwendeten Metalls. Bei idealer Skalierung aller Dimensionen, d.h. für lokale und regionale Leitungen, skaliert auch die Kapazität der Leitung. Die Kapazität globaler Leitungen, deren Längen proportional zu

den Abmessungen des Chips sind, bleibt jedoch konstant und skaliert nicht. Bild 3.4(a) veranschaulicht die Leitungsgeometrien.
Die Skalierung der Leitungsgeometrien führt sowohl für lokale und regionale als auch globale Leitungen zu deutlich erhöhten Leitungswiderständen. Zusätzlich nimmt die spezifische Leitfähigkeit der Leitungsstrukturen ab, da mit weiterer Skalierung die Streuung der Ladungsträger am Rand der Leitung einen erhöhten Einfluss auf den Stromfluss hat.
Aufgrund des stark zunehmenden Leitungswiderstandes skaliert die Höhe T_{Ltg} der Leitung weniger als alle anderen Dimensionen. Somit wird die Zunahme von R_{Ltg} abgeschwächt, was jedoch zu erhöhter kapazitiver Kopplung in horizontaler Ebene führt. Um diese Kapazitäten zu verringern, werden als Isolatoren innerhalb eine Metallebene (Intra Layer Dielectric ILD) Dielektrika mit geringerer Dielektrizitätskonstante ε_{ILD} verwendet.
Durch dieses, für jeden Prozess individuell festgelegte Skalierungsverhalten, ist eine genaue quantitative Aussage zur Skalierung von R_{Ltg} und C_{Ltg} nur bedingt möglich. Aktuelle Trends zeigen jedoch, dass der Widerstandsbelag weiter ansteigt, während der Kapazitätsbelag nahezu konstant bleibt. Die starke Zunahme von R_{Ltg} hat insbesondere Einfluss auf die Laufzeit globaler Leitungen und das RC Verhalten der Spannungsversorgung.
Wie in Kapitel 3.1 bereits beschrieben, wird der größte Anteil der Leitungsschwankungen durch die globale Prozesskontrolle bestimmt. WID Variationen basieren vorwiegend auf der Dichte der Leitungsstruktur und umgebungsbedingten Lithographieeffekten. So ist für die Schwankung der Leitungsbreite W_{Ltg} und somit auch für den Abstand zweier benachbarter Leitungen S_{Ltg} eine der Gatelängenschwankung entsprechende konstante relative Schwankungsbreite zu erwarten. Die Höhe der Leitung T_{Ltg} wird durch die Effektivität von Fill-Strukturen zur Reduzierung von Dishing und Erosion während des CMP Schritts dominiert.
Neben diesen Aspekten muss jedoch berücksichtigt werden, dass der Absolutwert der Gatekapazität eines MOS-Transistors bei fortschreitender Skalierung abnimmt ($C_{ox} \cdot W \cdot L$), so dass sich die Länge der Leitung für eine äquivalente kapazitive FO4-Last (Design Kriterium) weiter reduziert. Zusammen mit erhöhten effektiven Schaltwiderständen der leitungstreibenden Transistoren wird somit auch der Einfluss der dominanten globalen und lokalen Schwankung des Leitungswiderstandes abgeschwächt.
Dementsprechend ist zu erwarten, dass vorwiegend kapazitive Schwankungen zur Laufzeitvariation von kritischen Pfaden beitragen. Als generell sensitive Strukturen können Anordnungen aus treiberstarken Gattern (kleines R_{Tr}) und hohem Anteil der lateralen Leitungskapazität an der Gesamtlast genannt werden. Bus-Strukturen, globale Verdrahtungen (u.a. Network on Chip NOC), sowie Taktverteilungsnetze erfüllen diese Bedingungen und gelten daher als besonders sensitiv gegenüber Schwankungen der Leitungstopologien.

Umgebungsvariationen:

- **IR-Drop:**

  Das Skalierungsverhalten von IR-Drop ist von vielen Faktoren abhängig. Neben der Stromaufnahme ist der Zuleitungswiderstand eine entscheidende Größe. Um den Flächengewinn der Technologieskalierung nicht zu reduzieren, muss auch die Leitungs-

geometrie angepasst werden. Durch diese Skalierung erhöht sich mit jeder Technologiegeneration der Widerstand der Leitung. Vergleicht man die Leistungsaufnahme eines flächenoptimierten ARM926 Mikroprozessorkerns in 130nm und 90nm CMOS [88], so zeigt sich eine um 40% erhöhte Stromdichte des Designs in 90nm gegenüber 130nm. Zusammen mit erhöhten Leitungswiderständen resultiert dies in erhöhten Spannungsabfällen.
Diese theoretische Aussage setzt jedoch voraus, dass die Randbedingungen beim Schaltungsentwurf nicht geändert wurden, d.h. gleiches Standardzellenlayout, Geometrieanpassung nach idealem Scaling etc.. Eine qualitative Aussage zum Skalierungsverhalten von IR-Drop ist somit nur unter stringenter Vorgabe von Randbedingungen möglich.
Unabhängig davon erhöht sich für moderne CMOS Technologien die Laufzeitsensitivität gegenüber Versorgungsspannungsschwankungen, d.h. selbst ein konstanter IR-Drop führt zu deutlich erhöhten Laufzeitschwankungen.
Da der IR-Drop von vielen Faktoren im Schaltungsentwurf abhängt, ist eine Pauschalaussage hinsichtlich des Skalierungsverhaltens nicht möglich. Indizien, wie erhöhte Laufzeitsensitivitäten, Stromdichten und Verdrahtungswiderstände deuten jedoch auf eine zunehmende Bedeutung von IR-Drop Effekten hin.

- **Crosstalk:**

  Der Einfluss von Crosstalk auf die Laufzeit von Pfaden hängt vorwiegend von der Koppelkapazität zweier benachbarter Netze und deren Schaltaktivität ab. Für ideale Technologieskalierung skaliert auch die Koppelkapazität eines Netzes ideal um den Faktor 0.7. Aufgrund der deutlichen Zunahme des Leitungswiderstandes mit der Skalierung und dem damit verbundenen IR-Drop Problem werden die Querschnittsflächen der Leitungen angepasst, so dass bei idealem Flächengewinn durch die Technologieskalierung der Anstieg der Leitungswiderstände abgeschwächt wird. Daher wird die Höhe T_{Ltg} der Leitung nur geringfügig reduziert, so dass die Verringerung der Querschnittsfläche moderater ausfällt. Diese Maßnahme verändert die Skalierung der Koppelkapazitäten zweier benachbarter Leitungen. Der Skalierungsfaktor s_{C_K} für Koppelkapazitäten bei ideal skalierendem L_{Ltg} und S_{Ltg} lässt sich demnach wie folgt bestimmen (Technologie 1 $\longrightarrow$ Technologie 2):

  $$s_{C_K} = \frac{C_{K_2}}{C_{K_1}} = \frac{\varepsilon_{ILD_2} \frac{L_2 \cdot T_2}{S_2}}{\varepsilon_{ILD_1} \frac{L_1 \cdot T_1}{S_1}} = \frac{\varepsilon_{ILD_2} L_2 S_1 T_2}{\varepsilon_{ILD_1} L_1 S_2 T_1} = \frac{\varepsilon_{ILD_2} L_1 S_1 T_2 0.7}{\varepsilon_{ILD_1} L_1 S_1 T_1 0.7} = \frac{\varepsilon_{ILD_2} \cdot T_2}{\varepsilon_{ILD_1} \cdot T_1} \quad (3.16)$$

  Dabei bezeichnet C_K die Koppelkapazität zweier Leitungen und ε_{ILD} die Dielektrizitätskonstante des Intra Layer Dielektrikums (ILD). Bei gleichem ILD bestimmt somit die Skalierung der Leitungshöhe die Veränderung der Koppelkapazitäten.
  Neben der Koppelkapazität zu benachbarten Leitungen sind die Schaltaktivität, die Richtung des Signalwechsels sowie die zeitliche Synchronität (RSAT) des Signalwechsels auf den Nachbarnetzen wesentliche Einflussfaktoren. Eine Vorhersage des Crosstalkverhaltens ist daher für komplexere Schaltungen nicht möglich. Im Schaltungsentwurf werden Crosstalk-Effekte deshalb durch sehr pessimistische worst-case Analysen abgedeckt. Insbesondere unter Berücksichtigung einer variationsbehafteten Umgebung ist die Bestimmung der zeitlichen Synchronität der Schaltvorgänge von Aggressor und Victim nicht möglich. Eine qualitative Aussage zum Skalierungsverhalten von Crosstalk-Effekten ist daher nur auf Basis von Koppelkapazitätswerten und Schaltaktivität möglich. Auch hier ist für komplexe Schaltungen wie z.B.

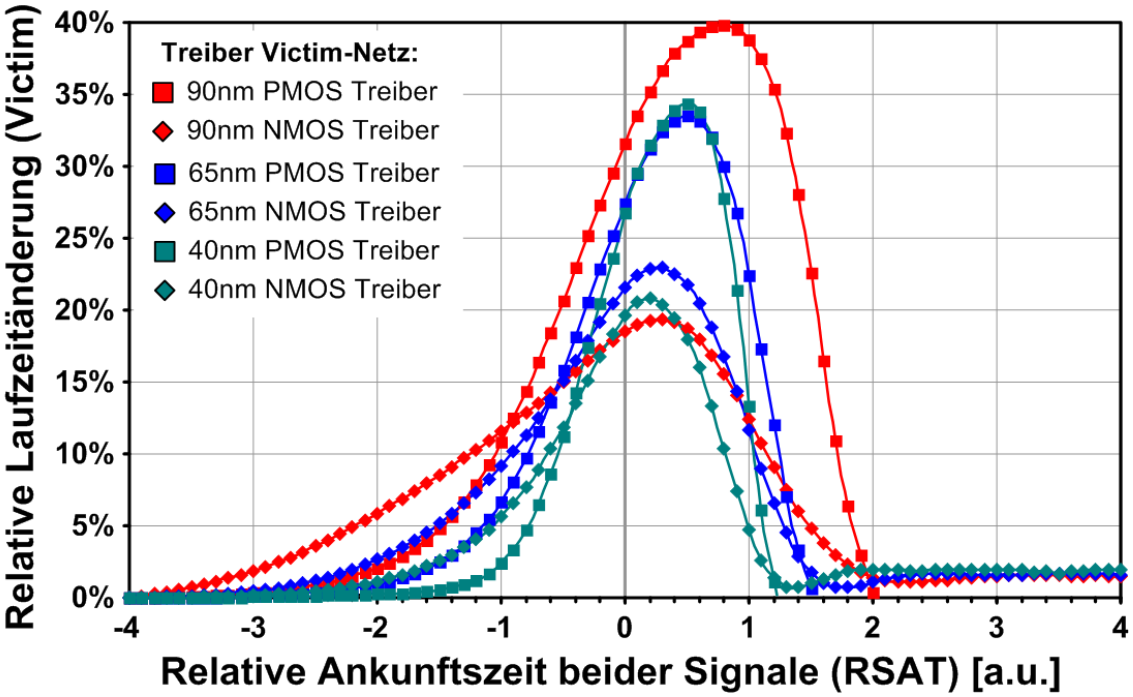


Bild 3.16: Simulierte Crosstalk-induzierte Laufzeitänderung für 90nm, 65nm und 40nm CMOS auf Basis extrahierter Netzlisten ($V_{DD} = V_{DD,nom} + 10\%$, T=27°C).

Mikroprozessoren keine Pauschalaussage möglich, da dieser dynamische Effekt neben strukturellen Schaltungseigenschaften auch von benutzerprofilabhängigen Schaltvorgängen abhängt.
Für die folgende Untersuchung wird eine Kette aus Invertern mit Fanout-4 Last und zusätzlicher Koppelkapazität untersucht. Bild 3.8 zeigt die verwendete Simulationsschaltung. Zwei Netze benachbarter Fanout-4 Pfade sind mit einer FO-3.3 Last äquivalenten Kapazität gekoppelt, d.h. die Koppelkapazität hat einen Anteil von ca. 45% an der Gesamtlast. Somit ist jedes der beiden treibenden Gatter - unabhängig vom jeweiligen Prozess - mit einem effektiven Fanout von 7.3 belastet. Da schwache Signalflanken am Victim-Netz einen höheren Einfluss des Aggressornetzes auf die Laufzeitänderung hervorrufen, kann die verwendete Struktur als Crosstalk-sensitiv angesehen werden. Bild 3.16 zeigt die Simulationsergebnisse für 90nm, 65nm und 40nm CMOS Technologien in Abhängigkeit der relativen Signalankunftszeit an der Koppelkapazität (extrahierte Netzlisten).

Durch die mit der Technologieskalierung einhergehende Reduzierung der Gatterlaufzeit verkleinert sich das zeitliche Fenster, in dem Signale auf Nachbarleitungen die Propagation von Signalen im Victim-Pfad beeinflussen. Betrachtet man, wie hier, Crosstalkeffekte unter Berücksichtigung von Designkriterien für kritische Pfade, wie sie von Synthese und Place & Route Tools angewandt werden, so ist zu erkennen, dass trotz großer Koppelkapazitäten kein Anstieg der Crosstalk-induzierten Laufzeitschwankung mit fortschreitender Technologieskalierung zu erwarten ist. Dieser Zusammenhang basiert auf der Tatsache, dass z.B. bei erhöhten Leitungskapazitäten zusätzliche Treiberstufen eingefügt und somit konstante relative Lastverhältnisse beibehalten werden. So werden Technologieeffekte auf höherer Ebene (hier: Pfadebene) kompensiert. Trotzdem ist es beim Schaltungsentwurf weiterhin wichtig, hohe Anteile der Koppelkapazitäten an der Gesamtlast zu vermeiden, um schwerwiegende

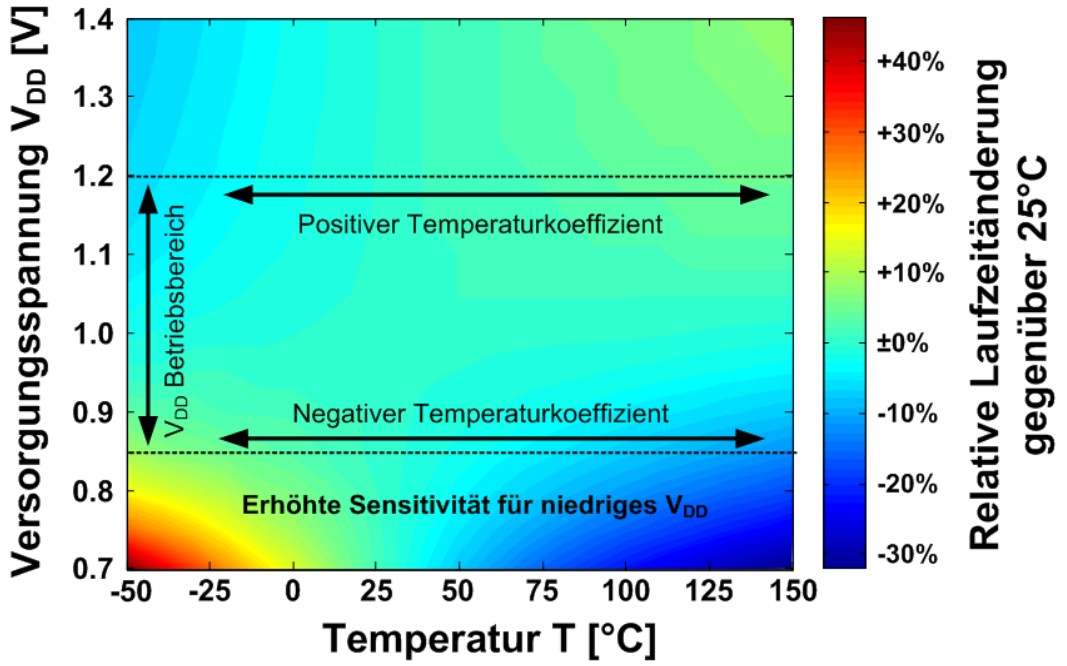


Bild 3.17: Simulierte relative Laufzeitänderung aufgrund von Temperaturschwankungen in 40nm CMOS auf Basis extrahierter Netzlisten.

Laufzeitschwankungen durch Crosstalk zu verhindern.

- **Temperaturschwankung:**

  Die Temperatur T hat sowohl Einfluss auf die Beweglichkeit μ der Ladungsträger als auch die Einsatzspannung V_T des Transistors. Bei sinkender Versorgungsspannung V_{DD} nimmt auch der Gate-Overdrive $V_{DD} - V_T$ ab und die Sensitivität der Laufzeit gegenüber V_T steigt. Dies hat einen erhöhten Einfluss der Temperatur zur Folge. Im Gegensatz zu den vorhergehenden CMOS Technologien, deren Versorgungsspannungsbereich oberhalb des Zero Temperature Coefficient Point (ZTCP) lag, liegt der ZTCP in 40nm/45nm CMOS innerhalb des Betriebsbereichs von V_{DD}, so dass das Vorzeichen der Laufzeitsensitivität gegenüber T je nach Betriebspunkt wechseln kann. Der steigende Einfluss von V_T hat somit insbesondere bei niedrigem V_{DD} auch einen steigenden Einfluss der Temperatur auf die Laufzeit zur Folge. Obwohl die Laufzeit beim Betrieb am ZTCP temperaturunabhängig ist, steigt die Sensitivität bei weiterer Skalierung der Versorgungsspannung an. Bild 3.17 zeigt für eine NAND2-NOR2 Kette in 40nm die relative Laufzeitänderung in Abhängigkeit von Versorgungsspannung und Temperatur gegenüber dem Betrieb bei Raumtemperatur. Bereits bei $V_{DD} = 0.85V$ ist eine deutlich höhere Laufzeitschwankung über den gesamten Temperaturbereich zu erkennen als im gesamten Spannungsbereich $V_{DD} > 0.85V$.

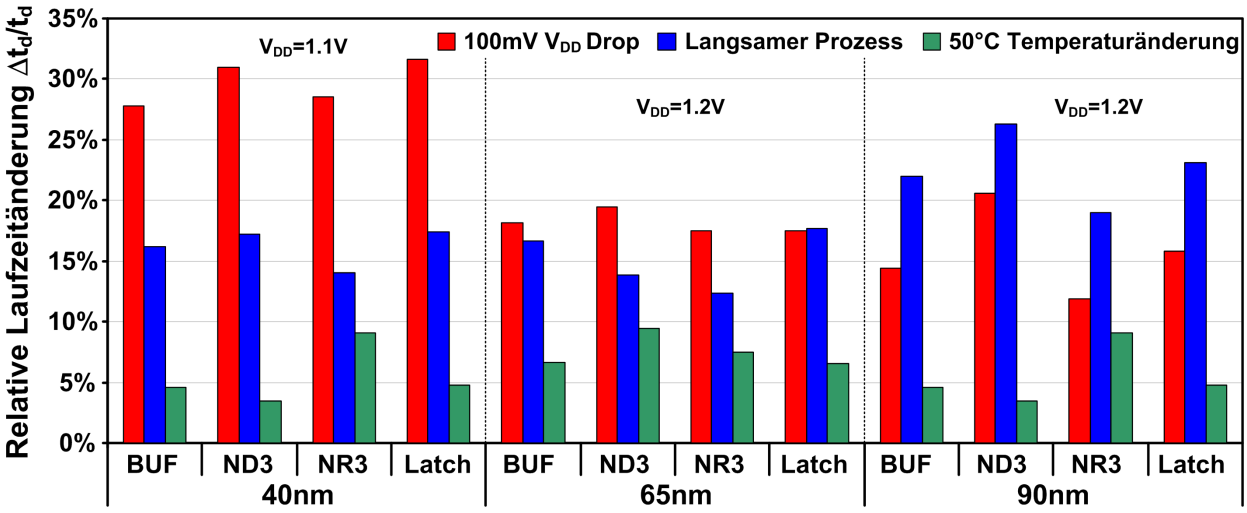


Bild 3.18: Simulierte technologie- und schaltungstechnikabhängige Unterschiede der Laufzeitschwankung auf Basis extrahierter Netzlisten (Nomineller Betrieb bei: $V_{DD} = V_{DD}^{nom}$, T=27°C).

3.2.3 Schaltungstechnische Aspekte der Laufzeitsensitivität

Bisher wurde nur der generelle Technologietrend der Laufzeitsensitivität gegenüber Prozess-, Spannungs- und Temperaturschwankungen (PVT) untersucht. Schaltungstechnische Aspekte wurden bisher vernachlässigt. In diesem Abschnitt werden die simulierten Laufzeitvariationen verschiedener Gattertypen untersucht und für verschiedene Technologieknoten gegenübergestellt. Für die folgenden Untersuchungen erhalten alle Gatter die gleiche Eingangsflanke und sind mit einer Fanout-4 Last belastet.
Neben einer Buffer-Zelle, die das Verhalten eines einzelnen schaltenden Transistors repräsentiert, werden ein NAND3 und NOR3 Gatter gewählt, um den Einfluss von schaltenden Serientransistoren zu untersuchen. Das untersuchte Latch repräsentiert das Verhalten von C^2MOS Invertern. Diese werden neben Latches auch in Multiplexern und Flip Flops verwendet. Somit werden die wichtigsten schaltungstechnischen Eigenschaften einer Standardzellenbibliothek abgedeckt. Die Gatterauswahl ist daher geeignet, den Einfluss schaltungstechnischer Eigenschaften auf die Laufzeitsensitivität zu untersuchen.
Bild 3.18 zeigt die Laufzeitänderung der Gatterauswahl gegenüber 100mV IR-Drop, der $+3\sigma$ entsprechenden globalen worst-case Corner der Prozessschwankung sowie einer Temperaturänderung von 50°C jeweils relativ zum nominellen Betriebspunkt.
Die Laufzeitschwankungen der verschiedenen Gattertypen unterscheiden sich deutlich von einander und variieren abhängig vom jeweiligen Technologieknoten. Die von 65nm auf 40nm trotz erhöhter Laufzeitsensitivität konstant bleibende prozessbedingte Laufzeitänderung ist Indiz für die Skalierung der globalen Schwankungsbreiten aufgrund verbesserter Prozesskontrolle [30]. Unabhängig vom Gattertyp steigt mit fortschreitender Technologieskalierung die Laufzeitsensitivität aller Gatter gegenüber Versorgungsspannungsschwankungen. Dies zeigt sich besonders beim Übergang von 65nm auf 40nm und der damit verbundenen Reduktion der nominellen Versorgungsspannung um 100mV.

Die unterschiedlichen Laufzeitsensitivitäten ergeben sich aus verschiedenen Betriebspunkten der Transistoren beim Schaltvorgang. Je nach Gattertopologie durchlaufen die Transistoren Betriebspunkte verschiedener Gate-Source V_{GS} und Drain-Source V_{DS} Spannungen,

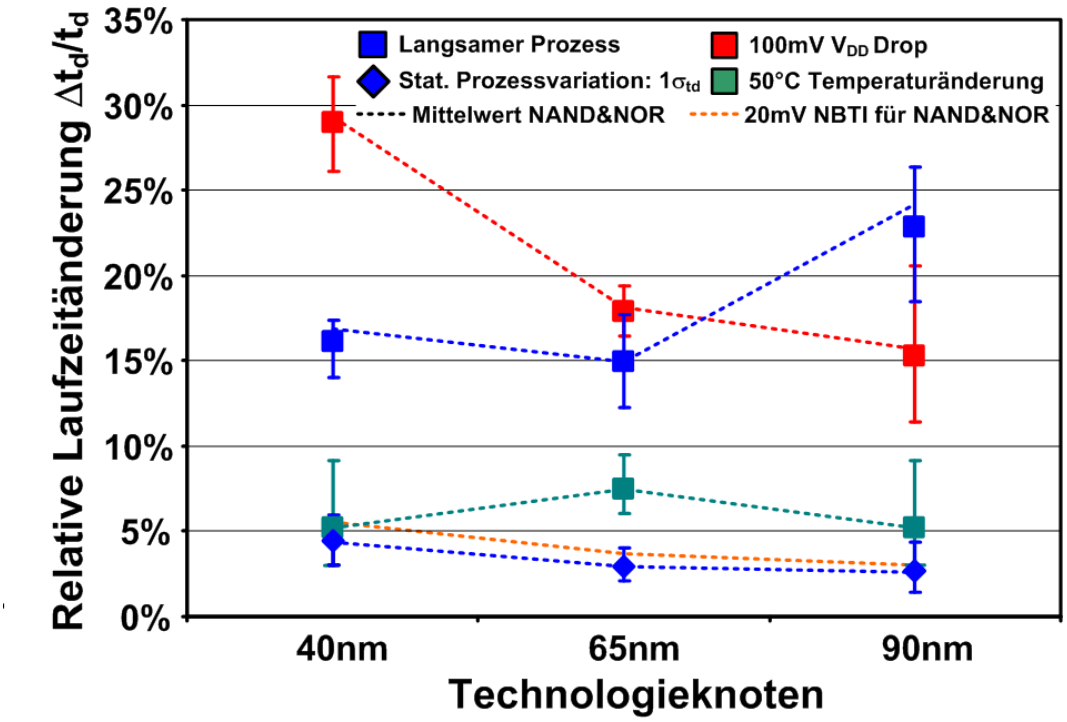


Bild 3.19: Vergleich der Laufzeitschwankung von Gattermix und ND2-NR2 Äquivalent für verschiedene Technologieknoten auf Gatterebene (Einzelgatter). Simulationsergebnisse auf Basis extrahierter Netzlisten (Nomineller Betrieb bei: $V_{DD} = V_{DD}^{nom}$, T=27°C).

so dass aufgrund des sich ändernden Gate-Overdrives auch die Laufzeitsensitivität betroffen ist. Zur Veranschaulichung zeigt Bild 3.2 das Ausgangskennlinienfeld eines 65nm NMOS Transistors mit der jeweiligen 1σ Schwankung (globale Prozessschwankung) des Drainstroms sowie die Trajektorien schaltender Transistoren eines NAND2-Gatters und Inverters. Bei nomineller Versorgungsspannung von 1.2V durchläuft der schaltende Serientransistor des NAND2-Gatters deutlich sensitivere Betriebsbereiche als der NMOS Transistor des schaltenden Inverters. Da sich der schaltende Transistor des NAND Gatters nur kurz in den deutlich sensitiveren Bereichen befindet, ist der Effekt auf die Laufzeitsensitivität jedoch geringer als Bild 3.2 vermuten lässt. Die weitere Absenkung der Versorgungsspannung hat die Verschiebung der durchlaufenen Betriebspunkte in deutlich sensitivere Betriebsbereiche zur Folge. Unabhängig vom Gattertyp ergibt sich somit generell eine höhere Laufzeitsensitivität gegenüber Prozess- und Umgebungsvariationen.
Um einen generellen Trend der Laufzeitsensitivität abzuleiten wird im Folgenden der Mittelwert der gatterspezifischen Laufzeitänderungen von 11 unterschiedlichen Gattertopologien gebildet, wie in Bild 3.19 für den Betrieb bei nomineller Versorgungsspannung dargestellt ist.
Für alle Gattertopologien wurde der Einfluss von globalen Prozessschwankungen, lokalen statistischen Variationen, 100mV IR-Drop, 50°C Temperaturunterschied und 20mV NBTI auf die Signallaufzeit des Gatters untersucht. Es sind sowohl die Mittelwerte als auch die Spannbreite (Maximum und Minimum) der Laufzeitänderung gezeigt. Mit fortschreitender Technologieskalierung steigt insbesondere der Einfluss von IR-Drop deutlich an. Bessere Prozesstechnik und besseres Verständnis von Variationen führen zu nur geringfügigem Anstieg des Einflusses globaler Prozessschwankungen von 65nm auf 40nm

CMOS. Wie im vorangegangen Abschnitt bereits diskutiert wurde, muss der Einfluss von Temperaturschwankungen auf die Laufzeit in Abhängigkeit der Versorgungsspannung erfolgen. Um jedoch einen Vergleich aller Variationen zu ermöglichen ist hier der Einfluss einer 50°C Temperaturänderung bei nomineller Versorgungsspannung gezeigt. Ein Trend hinsichtlich Temperatureffekte sollte aus dieser Grafik jedoch nicht abgeleitet werden. Der Einfluss von lokalen, statistischen Prozessvariationen wird durch die Darstellung der 1σ Gatterlaufzeitschwankung gezeigt. Der Mittelwert liegt für alle Technologien bei unter 5% Laufzeitschwankung eines Einzelgatters. Um auch den Einfluss von NBTI zu berücksichtigen, wird die Laufzeitänderung eines schaltenden NAND2/NOR2 Paars aufgrund einer um 20mV veränderten Einsatzspannung des PMOS Transistors gezeigt. Auch dieser Effekt liegt bei etwa 5% in 40nm CMOS. Für alle Variationen ist auch der Mittelwert aus NAND2 und NOR2 Gatter eingezeichnet, der nahezu immer dem Mittelwert aller 11 Gattertopologien entspricht.
Bild 3.19 zeigt den Einfluss von Variationen auf die Laufzeit eines Gattermixes in sub-100nm CMOS Technologien. In digitalen Schaltungen bestehen die geschwindigkeitskritischen Pfade jedoch aus einer Kombination verschiedener Gattertypen. Deshalb unterliegt die Sensitivität der Pfadlaufzeit schaltungstechnischen Mittelungseffekten, so dass die aus der Simulation von Einzelgattern erhaltenen maximalen Unterschiede auf Pfadebene nur in Einzelfällen auftreten. Als Beispiel hierfür kann das Taktverteilungsnetz genannt werden, das meist nur aus wenigen verschiedenen Gattertypen aufgebaut ist. Dieser Schaltungsteil wird im Schaltungsentwurf gesondert behandelt, so dass derartige Unterschiede ebenfalls Berücksichtigung finden. Zusätzlich wird auf Gatterebene der Einfluss der propagierenden Ausgangsflanke auf die Laufzeit vernachlässigt.
Das folgende Kapitel berücksichtigt bei der Analyse von Variationseffekten den schaltungstechnischen Einfluss in kritischen Pfaden durch die Auswahl repräsentativer Pfadstrukturen. Diese orientieren sich an den Eigenschaften geschwindigkeitskritischer Pfade und erfüllen wesentliche Designkriterien wie z.B. korrekte Lastverhältnisse, Flankensteilheit etc., wie sie in state-of-the-art Design-Tools verwendet werden.

4 Mikroprozessormodell zur Bestimmung technologischer und mikroarchitektonischer Einflussgrößen

4.1 Strukturanalyse eines ARM926 Mikroprozessor Produktdesigns

Im Folgenden werden die Ergebnisse einer detaillierten Strukturanalyse eines ARM926 Mikroprozessors der ARM Familie vorgestellt. Der ARM926 ist ein 32-bit RISC Prozessor mit einer klassischen fünfstufigen Pipeline. Er ist der am meisten verkaufte eingebettete Mikroprozessor im Bereich eingebetteter Mikroprozessoren und dient somit als repräsentative Schaltung für die folgende Fallstudie.
Die Ergebnisse basieren auf einer Produktimplementierung des ARM926 unter Verwendung von state-of-the-art Entwurfsprogrammen (EDA Tools) [89]. Alle laufzeit-relevanten Daten wurden durch Verwendung des Sign-Off Tools PrimeTime SI von Synopsys generiert [90].
Ziel dieser Untersuchung ist es, die Beschaffenheit von Setup- und Hold-Zeit kritischen Pfaden zu bestimmen und strukturelle Einflussgrößen auf das Laufzeitverhalten der Schaltung in variationsbehafteter Umgebung zu extrahieren.
Bild 4.1 zeigt die Verteilung des Pfad-Timings im Setup-Zeit und Hold-Zeit kritischen Bereich des ARM926 Designs. Es gilt:

$$\text{Pfad-Timing: } T_{SU} = t_{Clk-Q} + t_{Pfad} + t_{SU} + t_{var}^{SU} \overset{!}{\leq} T_{Clk} \tag{4.1}$$

$$\text{Pfad-Timing: } T_{HD} = t_{Clk-Q} + t_{Pfad} - t_{HD} - t_{var}^{HD} \overset{!}{>} 0 \tag{4.2}$$

Dabei bezeichnet T_{Clk} die Taktperiode, t_{Clk-Q} die Clock-Q Laufzeit des sendenden Flip Flops, t_{SU} und t_{HD} Setup- und Hold-Zeit des empfangenden Flip Flops, t_{Pfad} die Laufzeit des Logikpfades und t_{var}^{SU} bzw. t_{var}^{HD} die variationsbedingten Laufzeitschwankungen in Setup-Zeit und Hold-Zeit kritischen Pfaden. Während Setup-Zeit kritische Pfade die maximale Taktfrequenz limitieren führen Hold-Zeit Verletzungen zu einem funktionellen Ausfall unabhängig von der Taktfrequenz. Setup-Zeit Verletzungen können daher als Zyklus-zu-Zyklus Effekte, Hold-Zeit Verletzungen als Intra-Zyklus Effekte betrachtet werden. Da Setup-Zeit kritische und Hold-Zeit kritische Pfade gemeinsame Logikgatter besitzen und am gleichen Register (Flip Flop) enden können, muss bei allen Designmaßnahmen die Wirkung auf beide Timing-Bereiche überprüft werden.
Im Folgenden werden die Eigenschaften von Setup-Zeit und Hold-Zeit kritischen Pfaden untersucht.

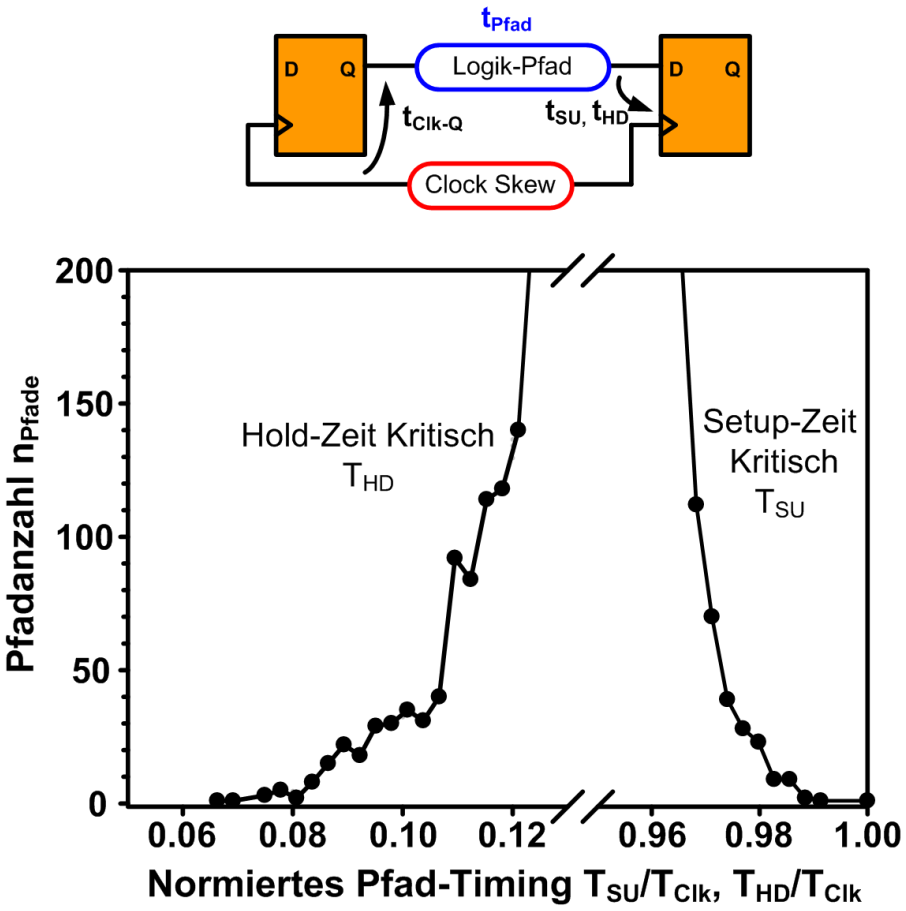


Bild 4.1: Pfadverteilung des untersuchten ARM926 Designs in 90nm CMOS.

4.1.1 Setup-Zeit kritische Pfade

Die Geschwindigkeit einer Schaltung wird durch den zeitlich längsten aller Pfade limitiert. Dieser Pfad wird 'kritischster Pfad' genannt. Im Folgenden wird die Bezeichnung 'kritischer Timing-Bereich' für alle Pfade mit einem Pfad-Timing von mindestens 90% des kritischsten Pfad verwendet. Als 'sub-kritische Pfade' werden Pfade bezeichnet, die sich ebenfalls im kritischen Timing-Bereich befinden, deren Pfad-Timing aber einen deutlichen zeitlichen Abstand zum kritischsten Pfad aufweist.
Die detaillierte Untersuchung des variationsrelevanten Timing-Bereichs wird durch die große Anzahl von Gatterkombinationen und der mit abnehmender Laufzeit nahezu exponentiell ansteigenden Anzahl an Pfaden limitiert. Für die strukturelle Untersuchung der oberen 10% des Timing-Bereichs wurden Daten in der Größenordnung von ca. 20GB generiert und verarbeitet. Aus diesem Grund beschränken sich die weiteren strukturellen Untersuchungen auf die oberen 10% der maximalen Pfadlaufzeit.

Beschaffenheit des Logikteils

Dieser Abschnitt beschäftigt sich mit der Beschaffenheit von Logikpfaden zwischen sendendem und empfangendem Flip Flop. Eine wesentliche Eigenschaft der Schaltung ist durch die Pfadverteilung im kritischen Timing-Bereich gegeben. Bild 4.2 zeigt die Pfadverteilung der geschwindigkeitskritischen Pfade.
Es ist deutlich zu erkennen, dass innerhalb der oberen 2.5% des Pfad-Timings eine relativ

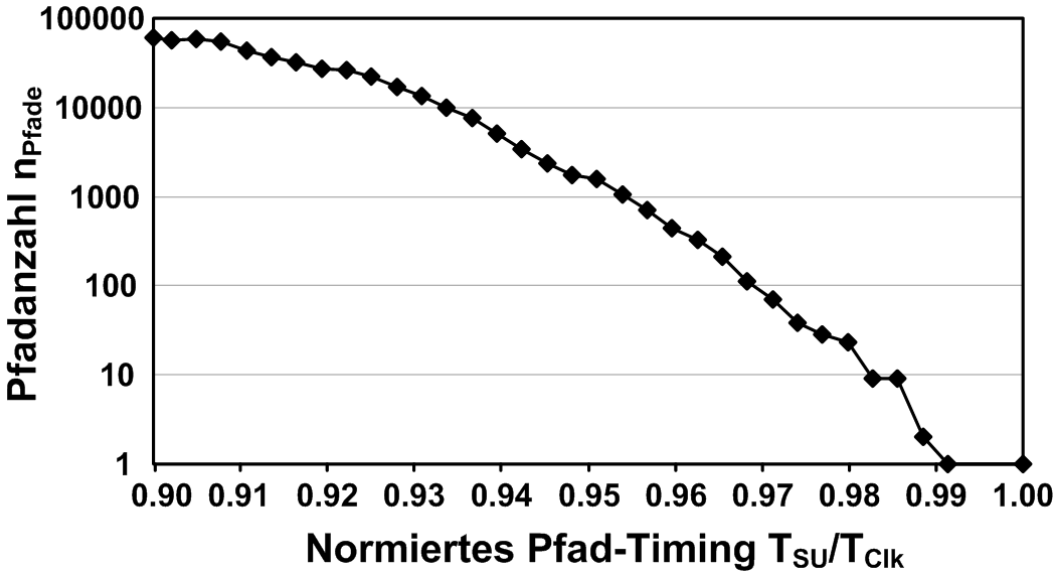


Bild 4.2: Pfadspektrum der geschwindigkeitskritischen Pfade des ARM926 Designs.

geringe Anzahl von Pfaden liegt. Anschließend steigt die Zahl der Pfade exponentiell an, da sich die Anzahl an Gattern, die in den folgenden Timing-Bereichen liegen, erhöht, und sich somit eine Vielzahl von neuen möglichen Gatterkombinationen (Pfade) ergibt. Gegen Ende des untersuchten Timing-Bereichs läuft die Pfadanzahl in eine Sättigung. Dies ist dadurch zu erklären, dass vom Timing-Tool keine zusätzlichen Gatter über den untersuchten Bereich hinaus berücksichtigt werden und somit die Anzahl an Kombinationen verschiedener Gatter sättigt. Diese Sättigung hat demnach keinen schaltungstechnischen Ursprung, sondern kann als Artefakt des Tool-Setups betrachtet werden. Da sich im untersuchten Bereich nur ca. 12.7% aller Gatter befinden, und somit die Anzahl an Gatterkombinationen auch über die Grenzen des untersuchten Bereichs hinaus zunehmen kann, ist eine weitere exponentielle Zunahme der Pfadanzahl zu erwarten.
Die im Pfadspektrum gezeigten Pfade des ARM926 sind Bestandteil verschiedenster Schaltungsblöcke innerhalb des Mikroprozessors, d.h. es liegt kein dominanter kritischer Schaltungsblock wie z.B. die arithmetische Einheit (ALU) vor, der die Geschwindigkeit der Schaltung limitiert.

Um den Einfluss von Variationen auf die Pfadlaufzeiten zu bestimmen, ist es notwendig, die einzelnen Anteile von kombinatorischen Logikgattern, RC Laufzeit etc. an der Pfadlaufzeit zu kennen. Bild 4.3 zeigt die Verteilung von Logikgatter, RC-Laufzeit und Crosstalk Effekten für alle kritischen Pfade.
Es ist deutlich zu erkennen, dass der Laufzeitanteil der Logik gegenüber allen anderen Beiträgen dominiert. Selbst für eine worst-case Crosstalk Berücksichtigung, d.h. die Koppelkapazitäten aller Aggressoren, die in das Schaltfenster der Victim-Leitung fallen, werden für die Berechnung des Crosstalk Laufzeitbeitrags herangezogen, weisen 90% aller Pfade einen Laufzeitbeitrag der Logik von über 85% auf. Die top-kritischen Pfade zeichnen sich durch einen überdurchschnittlich hohen Beitrag der Logik zur Gesamtlaufzeit des Pfades aus.
Es ist ersichtlich, dass bei derartig großen Logikbeiträgen das Verhalten der Schaltung gegenüber Variationen insbesondere durch die Variation der Transistorparameter und den Sensitivitäten der Logikgatter gegenüber schwankenden Betriebsbedingungen dominiert

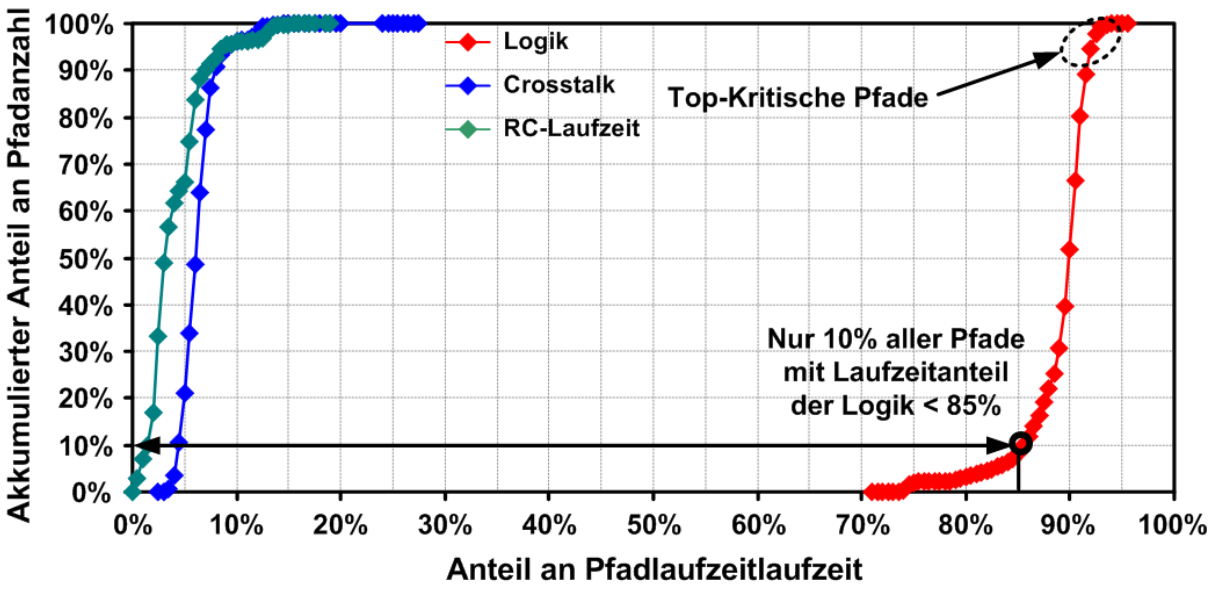


Bild 4.3: Akkumulierte Verteilung der Laufzeitbeiträge aller kritischen Logikpfade.

wird. Aus diesem Grund wird im Folgenden der Aufbau des Logikpfades durch die Analyse der für kritische Pfade typischen Gatter, deren Aufbau sowie deren Belastung näher untersucht.

Laufzeitbeiträge aus der Logik werden durch die für CMOS Schaltungen typische Invertierung des Signals und der damit verbundenen Umladung der intrinsischen Gatterkapazitäten, sowie der als Last zu betrachtenden Gate- und Leitungskapazitäten am Ausgang des treibenden Gatters verursacht. Je größer die Anzahl an invertierenden Stufen (Logikstufen), desto größer die Laufzeit. Bild 4.4 zeigt die Verteilung der Gatteranzahl bzw. der Anzahl an Logikstufen (Logiktiefe) der kritischen Pfade.
Ungefähr 95% aller Pfade zeichnen sich demnach durch eine Logiktiefe zwischen 26 und 41 aus. Der Mittelwert liegt bei 33 Logikstufen bei durchschnittlich 26 Gattern pro Pfad. Im Durchschnitt sind ca. 30% der Gatter im Logikpfad zwei- bzw. mehrstufig. Kritische Pfade mit hoher Logikstufenanzahl zeigen einen großen Anteil an Komplexgattern, d.h. mehrfach kombinatorische Gatter wie z.B. ein dreifach-Input AND-OR Gatter. Pfade mit geringer Logiktiefe liegen vorwiegend in sub-kritischen Timing Bereichen und weisen im Vergleich zu anderen kritischen Pfaden ein höheres Verhältnis zwischen Ausgangslast und Treiberstärke auf. Diese Systematik ist wie folgt zu erklären. Das Synthese Tool erkennt, dass diese Pfade im sub-kritischen Pfad liegen und erachtet aufgrund der fehlenden unmittelbaren Gefahr einer Timing-Verletzung die vorliegende Treiberstärke als ausreichend. Die kritischsten Pfade zeichnen sich durch dem Mittelwert ähnliche Werte von Logiktiefe und Gatteranzahl aus.
Diese Analyse ermöglicht die Untersuchung des Einflusses von statistischen Gatterlaufzeitschwankungen auf das Timing des Logikpfades. Im Wesentlichen wird der Einfluss von zwei verschiedenen strukturellen Eigenschaften bestimmt, der Logiktiefe n_{Log}, d.h. der Anzahl invertierender CMOS Stufen, sowie der Transistorgröße $W \cdot L$.
Mit zunehmender Logiktiefe bewirken Mittelungseffekte eine Abnahme der relativen Pfad-

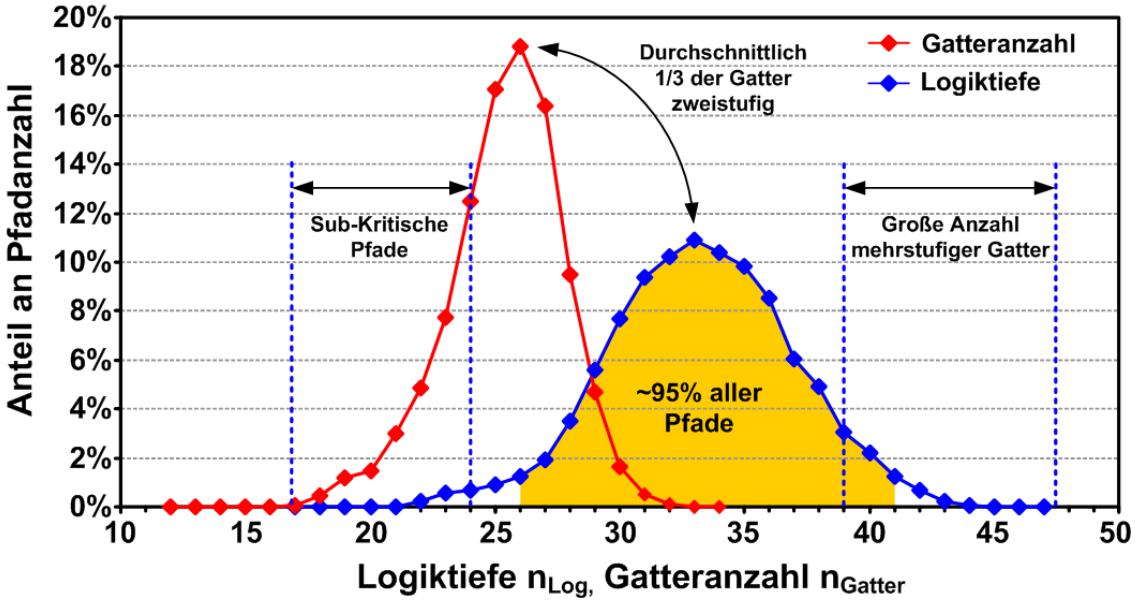


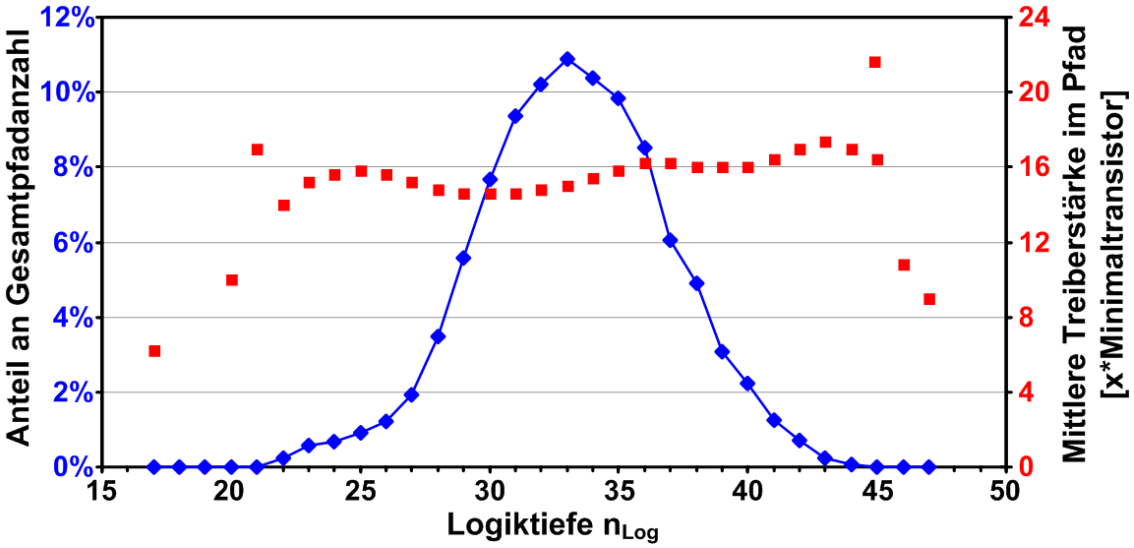
Bild 4.4: Verteilung der Logiktiefe n_{Log} bzw. Gatteranzahl n_{Gatter} in den kritischen Pfaden.

laufzeitschwankung [75, 22, 91]. Es gilt:

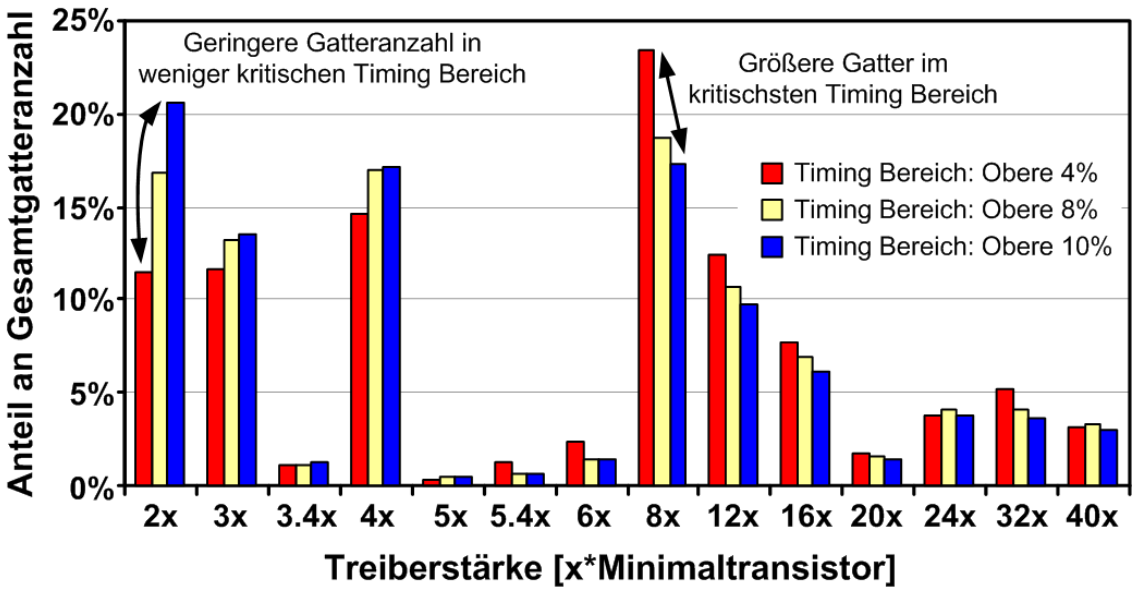
$$\frac{\sigma_{t_{Pfad}}}{\mu_{t_{Pfad}}} \sim \frac{1}{\sqrt{n_{Log}}} \frac{\sigma_{t_{Gatter}}}{\mu_{t_{Gatter}}} \tag{4.3}$$

Im vorliegenden Fall werden statistische Laufzeitschwankungen durchschnittlich um den Faktor 0.17 aufgrund der hohen Anzahl an Logikstufen abgeschwächt, d.h. eine Gatterlaufzeitschwankung von z.B. $\sigma_{t_{Gatter}} = 5\%$ führt zu einer Pfadlaufzeitschwankung von $\sigma_{t_{Pfad}} = 0.87\%$.
Zusätzlich gilt, je größer die Transistoren eines Gatters, desto geringer deren Laufzeitschwankung. Die Größe der verwendeten Gatter in den kritischen Pfaden ist somit entscheidend für den Einfluss statistischer Laufzeitschwankungen.
Bild 4.5(a) zeigt die Verteilung der Logiktiefe und die mittlerer Treiberstärke in den kritischen Pfaden, Bild 4.5(b) zum anderen die Verteilung der Gattertreiberstärke für drei unterschiedliche Timing Bereiche. Unabhängig von der Logiktiefe liegt die mittlere Treiberstärke bzw. Transistorgröße in den kritischen Pfaden beim 16fachen der in der Standardzellen verfügbaren Minimaltransistoren. Die statistische Transistoreinsatzspannung und damit auch die statistische Laufzeitschwankung wird durch die Verwendung derart großer Transistoren zusätzlich abgeschwächt. Eine Monte Carlo Simulation des kritischsten Pfades auf Basis extrahierter Netzlisten zeigt für 90nm CMOS Technologie bei nomineller Spannung eine Standardabweichung der Pfadlaufzeit von 0.5%. Die Schwankungsbreite liegt damit in der Größenordnung der Modell-Genauigkeit, d.h. der Einfluss von statistischen Schwankungen auf die Laufzeit des Logikpfades ist für diese Schaltungsimplementierung vernachlässigbar gering.
In Bild 4.5 wird insbesondere sichtbar, dass in Pfaden im top-kritischen Timing-Bereich im Vergleich zu sub-kritischen Bereichen eine größere Anzahl treiberstarker Gatter verwendet werden. Die Ursache für diese systematische Auffälligkeit muss anhand der Lastverteilung in den kritischen Pfaden diskutiert werden, wie sie im folgenden Abschnitt erfolgt.

Die Verwendung von treiberstarken Gattern in den top-kritischen Pfaden beruht nicht



(a) Verteilung der Logiktiefe mit zugehöriger mittlerer Transistortreiberstärke.



(b) Treiberstärkenverteilung in unterschiedlich kritischen Timing Bereichen.

Bild 4.5: Logiktiefe und Treiberstärken im untersuchten ARM926.

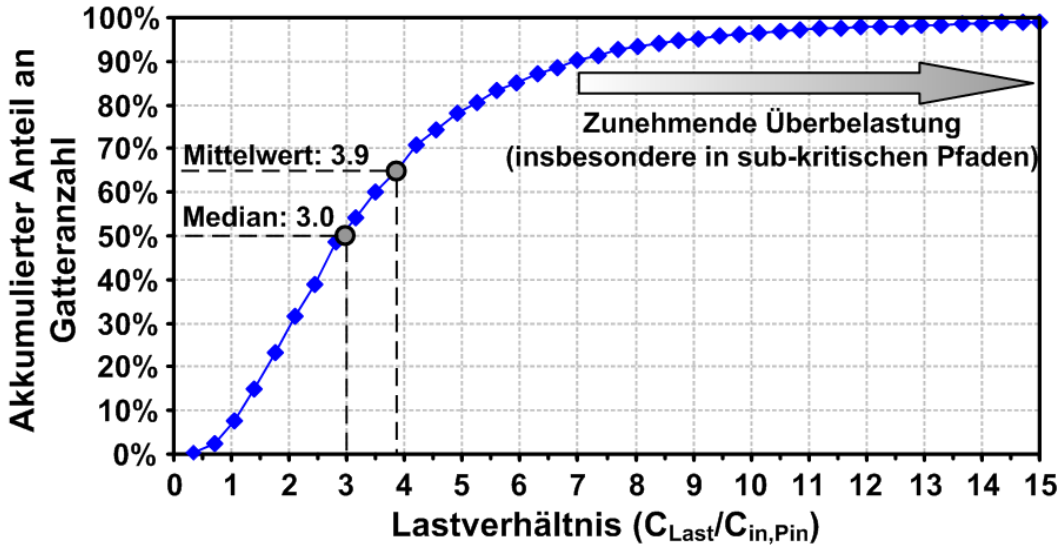


Bild 4.6: Treiber zu Last Verhältnis für alle Gatter im kritischen Timing-Bereich.

auf erhöhten kapazitiven Lasten. Vielmehr ist zu erkennen, dass das Verhältnis aus Eingangskapazität $C_{in,Pin}$ und Lastkapazität C_{Last} einen sehr geringen Wert aufweist. Bild 4.6 zeigt die Verteilung des Lastverhältnisses in den kritischen Pfaden. Etwa 50% aller Gatter weisen ein Lastverhältnis (Fanout) kleiner 3 auf. Das für eine minimale Laufzeit erforderliche Verhältnis von ca. 2,7 wird somit von fast der Hälfte aller Gatter eingehalten [44]. In den sub-kritischen Pfaden sind höhere Lastverhältnisse zu erkennen, da vom Synthese-Tool hier noch zeitliche Margen erkannt werden, die eine weitere Optimierung des Timings auf Kosten zusätzlicher Verlustleistung und erhöhten Flächenbedarfs nicht erforderlich machen.
Der Ausgangsknoten jedes Gatters ist mit mindestens einem Eingangs-Pin eines Folgegatters verbunden. Unter der Annahme gleicher Transistorgeometrien ergibt sich daraus eine Belastung von FO1. Dementsprechend verbleibt eine weitere Belastung von FO2, die sich aus der Verdrahtungskapazität, sowie evtl. weiteren Pin-Kapazitäten (falls die Aufspaltung des Netzes größer 1 ist; $BR_{Log} \geq 1$) zusammensetzt. Ein Lastverhältnis von FO1 kann daher unter obiger Annahme als strukturell-bedingtes, mittleres Minimum angesehen werden.
Die hohen Treiberstärken im top-kritischen Timing-Bereich sind durch das Bemühen des Synthese-Tools zu erklären, die Laufzeiten durch das Einfügen treiberstarker Gatter zu reduzieren. Eine weitere Erhöhung der Treiberstärken wäre jedoch aufgrund der gleichzeitig erhöhten Pin-Kapazitäten der zu treibenden Gatter nicht sinnvoll. Diese geringen Lastverhältnisse bei gleichzeitig kleinen RC-Laufzeit-Beiträgen in den kritischen Pfaden weisen demnach auf eine strukturelle Begrenzung der Pfadlaufzeit hin. Untersuchungen eines ARM1176 Designs zeigen ähnlich geringe Lastverhältnisse in den kritischen Pfaden. Diese Pfade können daher als intrinsisch-kritisch betrachtet werden, d.h. künstliche Effekte wie z.B. Überbelastung der Gatter oder unsichere Modellierung von Crosstalkbeiträgen sind nicht für die vergleichbar hohen Laufzeiten verantwortlich.

Um den Einfluss von Verdrahtung und Crosstalk auf die Pfadlaufzeit besser abschätzen zu können, werden die einzelnen Lastbeiträge näher analysiert. Bild 4.7 zeigt die Verteilung der Anteile von Verdrahtungskapazität, sowie der vom STA Tool berechneten, effektiv

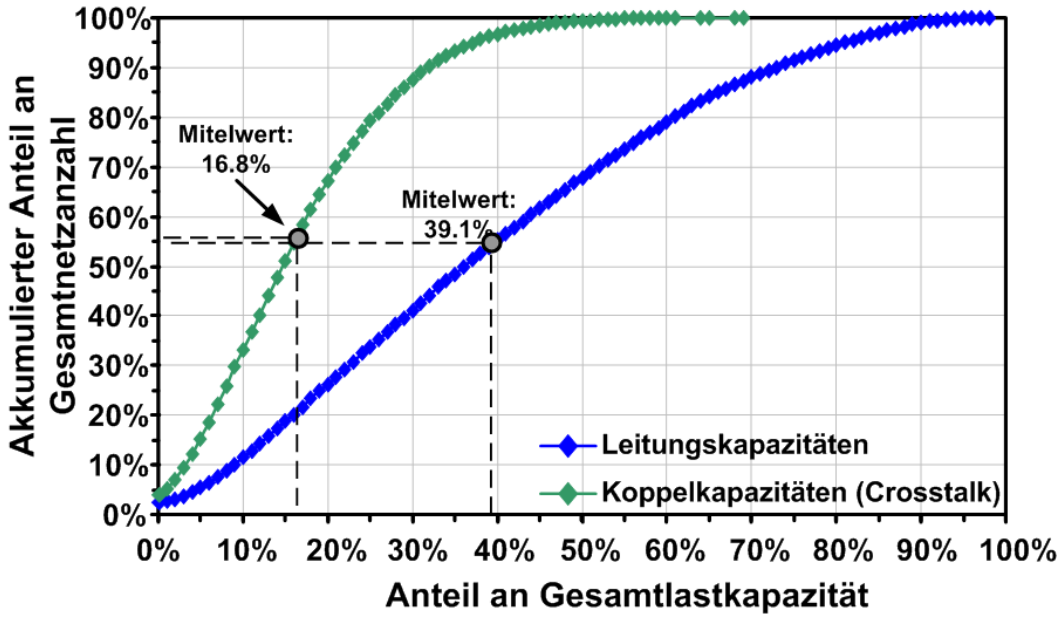


Bild 4.7: Verteilung von Verdrahtungs- und Koppelkapazitätsanteilen an der Gesamtkapazität der Netze innerhalb des kritischen Timing Bereichs.

wirkenden Koppelkapazität an der Gesamtkapazität des jeweiligen Netzes für alle Netze innerhalb des kritischen Timing-Bereichs.
Im Mittel beträgt der Verdrahtungsanteil an der Gesamtkapazität ca. 40% der Gesamtlast, d.h. im Durchschnitt ca. den Faktor 1.6 der Eingangskapazität der Treiberstufe. Dieser relativ geringe Wert zeigt, dass innerhalb der kritischen Pfade nur wenige Netze mit langen Leitungen vorhanden sind, so dass Widerstandseffekte kaum auftreten, wie bereits in Bild 4.3 zu sehen war.
Der Anteil der für Crosstalkeffekte verantwortlichen Koppelkapazität ist deutlich geringer und liegt im Mittel bei ca. 17%. Für die Bewertung von Crosstalkeffekten ist die Größe des Koppelkapazitätsanteils alleine nicht ausreichend. Da sich die Koppelkapazität aus einzelnen, kleineren Koppelkapazitäten verschiedener Aggressorleitungen zusammensetzt, wäre es entscheidend, den Teil der Kapazität zu bestimmen, der von schaltenden Aggressorleitungen stammt. Da das Schaltverhalten von vielen, unvorhersagbaren Faktoren wie z.B. benutzerspezifischen Anwendungen abhängt, ist die Bestimmung der effektiven Koppelkapazität daher nicht möglich. Ein Indiz für den Einfluss von Crosstalkeffekten kann jedoch die Anzahl der Aggressorleitungen N_{Aggr} sowie eine Abschätzung der maximalen Schaltaktivität liefern. Eigene Untersuchungen zeigen, dass die vom STA Tool in einer worst-case Berechnung bestimmten Crosstalkbeiträge in den kritischen Pfaden durchschnittlich von 254 ebenfalls schaltenden Aggressoren verursacht wird. Das zur Victim-Leitung entgegengesetzte, gleichzeitige Schalten aller Aggressorleitungen führt somit zu den in Bild 4.3 gezeigten, durch Crosstalk verursachten Laufzeitbeiträgen Δt_{XT}^{WC} von durchschnittlich ca. 7.5%. Dabei liegt das Minimum der Aggressoranzahl pro kritischem Pfad bei 105, das Maximum bei 428. Der worst-case Crosstalk Beitrag zur Gesamtlaufzeit des Pfades tritt nur auf, wenn alle Aggressoren gleichzeitig in die dem Victim-Netz entgegengesetzte Richtung schalten.
Setzt man gleichverteilte Schaltvorgänge und eine vergleichsweise hohe Schaltaktivität α_{Schalt} von 0.2 voraus (IBM Power6: $\alpha_{high} = 0.14$ [92]), so reduziert sich die effektive Aggressoranzahl auf durchschnittlich 51 Aggressoren. Für einen negativen Einfluss auf

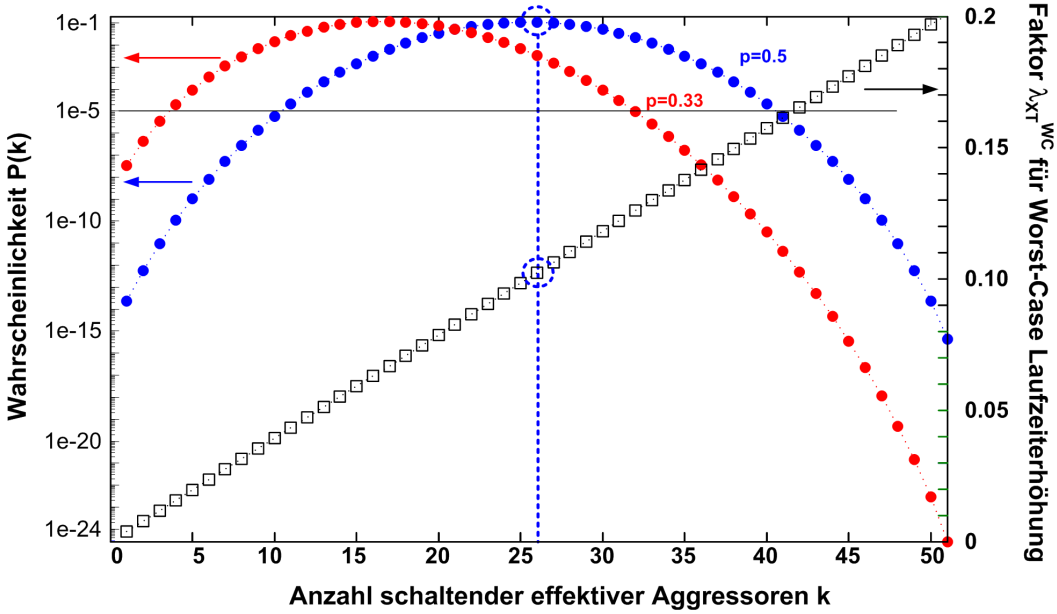


Bild 4.8: Wahrscheinlichkeitsverteilung für das Eintreten einer um den Faktor x reduzierten crosstalk-bedingten worst-case Laufzeiterhöhung.

die Laufzeit geschwindigkeitskritischer Pfade ist ein zur Victim-Leitung entgegengesetztes Schalten der Aggressoren notwendig. Zusätzlich beschleunigt ein mögliches, gleichgerichtetes Schalten der anderen Aggressoren den Schaltvorgangs auf der Victim-Leitung.
Nimmt man an, dass auf einer Aggressor-Leitung ein der Victim-Leitung entgegengesetzter, zeitgleicher Schaltvorgang mit der Wahrscheinlichkeit von p=0.5 auftritt und die restlichen Nachbarleitungen keinen Signalwechsel vollziehen, so kann die Wahrscheinlichkeit für eine bestimmte Anzahl k an entgegengesetzt und statistisch unabhängig schaltenden Aggressoren P(k) aus der binomialen Dichtefunktion bestimmt werden:

$$P(k) = \frac{n!}{k!\,(n-k)!} p^k q^{n-k} = \left(\begin{array}{c} n \\ k \end{array} \right) p^k q^{n-k} \tag{4.4}$$

$$\lambda_{XT}^{WC} = \frac{k}{N_{Aggr}} \tag{4.5}$$

Bild 4.8 zeigt die Wahrscheinlichkeit, dass eine um den Faktor λ_{XT}^{WC} reduzierte worst-case Laufzeiterhöhung Δt_{xTalk}^{WC} eintritt, wie sie vom STA Tool angegeben wird. Dabei wird angenommen, dass alle Aggressorkapazitäten gleich groß sind und die Anzahl der effektiv schaltenden Aggressoren bereits um den Faktor α_{Schalt} reduziert wurde ($254 \cdot 0.2 = 51$). Gleichzeitig reduziert sich die maximal schaltende Koppelkapazität ebenfalls um den Faktor 0.2, so dass die worst-case Laufzeiterhöhung unter obigen Annahmen ebenfalls um den Faktor 0.2 reduziert wird. So tritt bei p=0.5 eine 11% ige Wahrscheinlichkeit auf, dass ein einzelner Pfad eine um den Faktor 0.1 reduzierte crosstalk-bedingte worst-case Laufzeiterhöhung aufweist, d.h. im Mittel 0.75% der Taktperiode. Der starke Abfall der Wahrscheinlichkeit unter $P(k) < 10^{-10}$ zeigt, dass der worst-case Fall, d.h. alle als mögliche Aggressor identifizierten Netze schalten in die entgegengesetzte Richtung, unter den getroffenen Annahmen keine Praxisrelevanz besitzt. Die worst-case Wahrscheinlichkeit für eine bestimmte Anzahl an Aggressoren N_{Aggr} ist in Tabelle 4.1 gezeigt. Bei einer ursprünglichen minimalen Aggressoranzahl von 105 - der worst-case Fall tritt nach Gleichung 4.5

Tabelle 4.1: Abschätzung der Wahrscheinlichkeit des worst-case Crosstalk Szenarios in Abhängigkeit der Aggressoranzahl bei p=0.5.

Aggressoranzahl N_{Agr}	Wahrscheinlichkeit des worst-case
10	0.976e-3
50	0.888e-15
105	0.025e-30

für eine geringere Aggressoranzahl häufiger auf - ergibt sich für $\alpha_{Schalt} = 0.2$ eine effektive Aggressoranzahl von 21. Nimmt man an, dass die einzelnen Koppelkapazitäten je Aggressor gleich groß sind so ergeben sich folgende Szenarien. Schalten alle Aggressoren in die der Victim-Leitung entgegengesetzte Richtung, so tritt eine Crosstalk-induzierte Laufzeiterhöhung von 20% des worst-case mit einer Wahrscheinlichkeit von ca. 0.5e-6 auf. Berücksichtigt man nun, dass nur die Hälfte der Netze entgegengesetzt schalten während der Rest stabile Signale hält, so ist nach Gleichung 4.5 eine Laufzeiterhöhung von 10% des worst-case mit einer Wahrscheinlichkeit von ca. 16.8% zu erwarten.
Diese einfache Abschätzung auf Basis unabhängig voneinander schaltender Aggressoren zeigt, dass in komplexen Systemen wie dem untersuchten Mikroprozessor Design eine worst-case Abschätzung von Crosstalkeffekten einen großen Pessimismus aufweist. Ein gleichgerichtetes Schalten der Aggressorleitung, das die Laufzeiterhöhung durch entgegengesetztes Schalten kompensieren kann, wird in diesem Zusammenhang nicht berücksichtigt.
Im Gegensatz zu kleinen Designs, wie z.B. Time-to-Digital Converter (TDC), für die der Einfluss von Crosstalkeffekten auf die Zeitauflösung wesentlich einfacher zu bestimmen ist, stellen unvorhersagbare, benutzerprofilabhängige Faktoren wie Bit-Pattern Abhängigkeiten, das Auftreten von Glitches, sowie die hohe Komplexität der Koppelkapazitätsanteile und die zeitlichen Relationen der Schaltvorgänge für komplexe Schaltungen ein nahezu unüberwindbares Problem für die Abschätzung von Crosstalkeffekten dar.
Für eine genauere Bestimmung dieser dynamischen Laufzeiteffekte, müssen kombinatorischer Abhängigkeiten zwischen den einzelnen Aggressoren sowie zwischen Aggressor- und Victim-Leitung berücksichtigt werden. Zusätzlich kann auf Basis aller möglichen Bit-Pattern die theoretische Schaltwahrscheinlichkeit der verschiedenen Netze berechnet werden und somit eine Gewichtung der einzelnen Koppelkapazitäten erfolgen.
Die Ergebnisse der Untersuchungen von Koppelkapazitäten und Aggressoranzahl im ARM926 Mikroprozessor sind jedoch deutliche Indikatoren, dass die Wahrscheinlichkeiten für worst-case Crosstalkeffekte - wie sie noch immer von aktuellen Sign-Off Tools bestimmt werden - sehr gering sind. Die nach obigen Annahmen realitätsnahen Laufzeiterhöhungen liegen bei unter 2% der Taktperiode und haben deshalb in dieser Schaltung einen vernachlässigbar geringen Einfluss.

Beschaffenheit des Taktbaums

Das Pfad-Timing aus 4.1 und 4.2 beinhaltet mit t_{Skew} auch einen abstrahierten Beitrag des Taktbaums, so dass die Designqualität des Taktbaums wesentlich zur Vermeidung von Setup- und Hold-Zeit Fehlern beiträgt. Im Idealfall erreicht das für die Synchronisation erforderliche Taktsignal zur exakt gleichen Zeit die Takt-Eingänge aller Synchronisationselemente (Flip Flops, Latches). Jede Abweichung vom Idealfall führt zu einer verkürzten oder verlängerten effektiven Taktperiode.

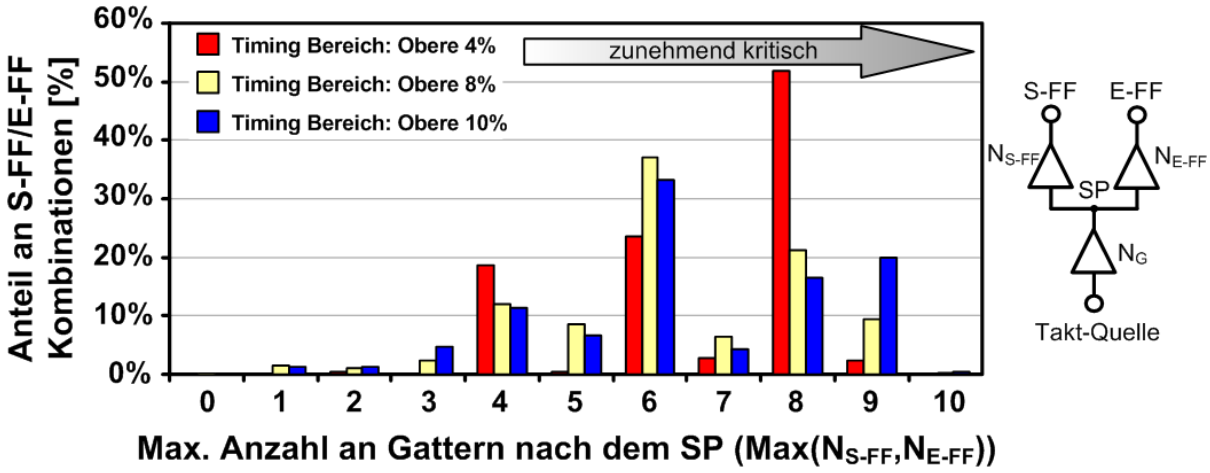


Bild 4.9: Verteilung der maximalen Anzahl an Clock Buffern nach dem Aufspalten im Taktbaum.

Im Vergleich zur komplexen Logik ist der Taktbaum strukturell einfach zu beschreiben. Jeder Taktpfad beginnt am Einspeisepunkt des Taktsignals und führt über eine bestimmte Anzahl an Clock Buffern zu einem Flip Flop bzw. Latch. Dabei werden einzelne Clock Buffer von mehreren Taktpfaden gleichzeitig genutzt, z.B. von den Taktpfaden zum sendenden und empfangenden Flip Flop eines kritischen Pfades, bis schließlich eine Aufspaltung beider Taktpfade erfolgt.
Bild 4.9 zeigt die Verteilung der maximalen Anzahl an Clock Buffern nach der Aufspaltung des Taktbaums (Splitting Point SP). Je größer die Anzahl der Clock Buffer nach dem Splitting Point, desto schwieriger ist es unter prozess- und betriebsbedingten Variationen einen zeitlich ausgeglichenen Taktbaum zu erzielen.
In Bild 4.9 ist zu erkennen, dass für mehr als die Hälfte aller kritischer Pfade in den oberen 4% des Timing Bereichs das Taktsignal nach dem Aufspalten noch über weitere acht für sendenden und empfangenden Taktpfad unterschiedliche Clock Buffer geführt wird, bis es an den entsprechenden Registerelementen ankommt.
Im Allgemeinen werden im Clock Buffer große Transistoren eingesetzt, um die große Gesamtlast der Taktsignal-Eingänge aller Flip Flops bzw. Latches treiben zu können. Im vorliegenden Fall werden nur Clock Buffer mit einer Mindesttreiberstärke von 16 Minimaltransistoren eingesetzt. Dabei besteht ein Taktpfad aus durchschnittlich 22 Logikstufen. Deshalb ist im Taktverteilungsnetz eine hohe Mittelung von statistischen Laufzeitschwankungen zu erwarten.

4.1.2 Hold-Zeit kritische Pfade

Während Setup-Zeit Verletzungen die Geschwindigkeit einer getakteten Schaltung limitieren, führen Hold-Zeit Verletzungen zu funktionalen Ausfällen einer Schaltung. Setup-Zeit Verletzungen können grundsätzlich durch Verwendung niedrigerer Taktfrequenzen verhindert werden, sofern vorhandene Geschwindigkeitsmargen genutzt werden können. Hold-Zeit Verletzungen treten jedoch unabhängig von der Betriebsfrequenz auf. Aus diesem Grund ist es sehr wichtig unter allen Umständen Hold-Zeit Verletzungen bereits während des Schaltungsentwurfs insbesondere unter Berücksichtigung von variierenden Prozessparametern und Betriebsbedingungen ausschließen zu können.

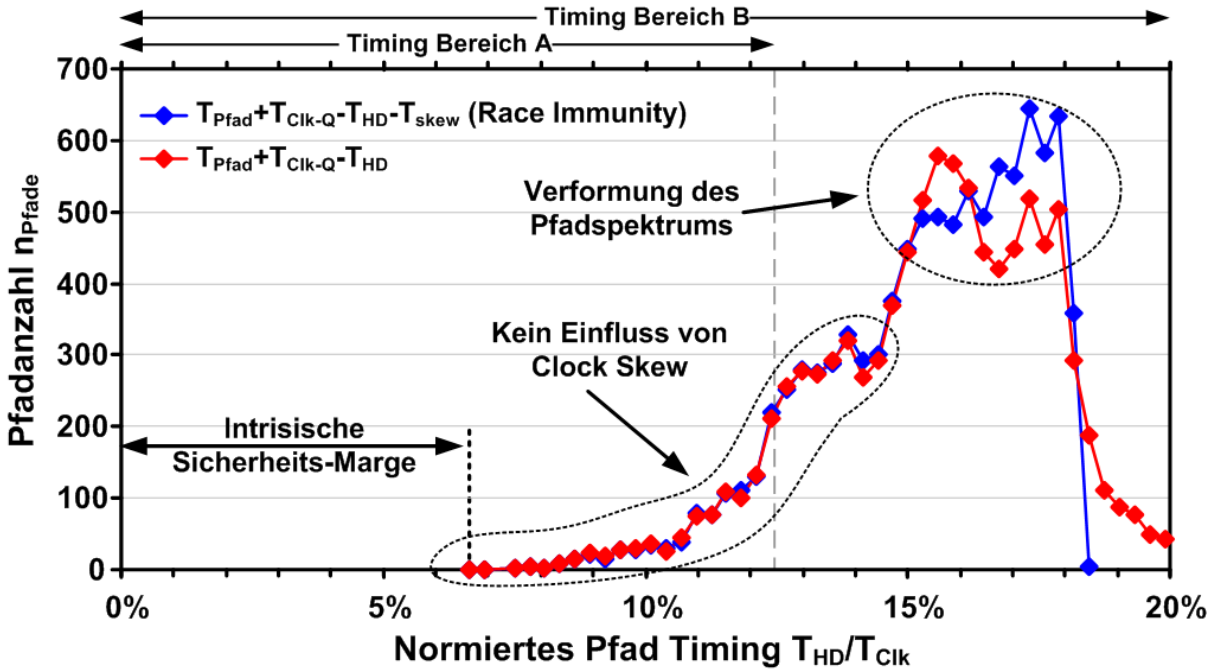


Bild 4.10: Pfadspektrum der Hold-Zeit kritischen Pfade des ARM926 Designs.

Beschaffenheit des Logikteils

Gleichung 4.2 beschreibt die Laufzeitbedingung, die alle Pfade zu erfüllen haben, damit Hold-Zeit Verletzungen ausgeschlossen werden können. Bild 4.10 zeigt die Pfadverteilung der Hold-Zeit kritischen Pfade. Das Pfad-Timing ist auf die durch die Setup-Zeit kritischen Pfade limitierte minimale Taktperiode normiert. Eine große Sicherheits-Marge von ca. 7% der Taktperiode ist zu erkennen, d.h. eine unmittelbare Gefahr von Hold-Zeit Verletzungen ist nicht gegeben. Obwohl die 100 kritischsten Pfade zu 97% aus direkten Flip Flop zu Flip-Flop-Verbindungen bestehen, d.h. $t_{Pfad} \approx 0$, ist eine deutlich positive Race-Immunity erkennbar. Da das vorliegende Design ausschließlich Master-Slave Flip Flops zur Synchronisation der Schaltung verwendet, gilt für die Hold-Zeit $t_{HD} \approx 0$ [93, 94]. Die Race-Immunity ergibt sich somit aus den Größen t_{Clk-Q} und t_{Skew}. Wie in Bild 4.10 sichtbar ist, ist zwischen Berücksichtigung des Clock Skews und dessen Vernachlässigung kein Unterschied in der Race-Immunity erkennbar, d.h. die Race-Immunity wird in diesem Design vorwiegend durch die Clock-Q Laufzeit des sendenden Flip Flops bestimmt.
Da es sich bei Hold-Zeit kritischen Pfaden immer um kurze Pfade mit geringer Laufzeit handelt, ist die Untersuchung der Logiktiefe sowie der Transistorgrößen in diesen Pfaden hinsichtlich der Auswirkung statistischer Variationen besonders wichtig. Bild 4.11 zeigt die Verteilung der Logiktiefe in den kritischen Pfaden und die mittlere Treiberstärke der verwendeten Gatter im Logikpfad.
Im Vergleich zu den Setup-Zeit kritischen Pfaden ist die mittlere Treiberstärke um den Faktor 0.2 geringer. Die minimale Logiktiefe, die für dieses Design bei 3 liegt, wird durch die verwendeten Flip Flops festgelegt. Die aus den Setup-Zeit kritischen Pfaden bekannten starken Mittelungseffekte sind hier nicht erkennbar. Dennoch muss festgehalten werden, dass die intrinsische Sicherheits-Marge von ca. 7% der Taktperiode, die ohne Logikgatter zwischen sendendem und empfangendem Flip Flop vorhanden ist, daher nur durch Unsicherheiten aus dem Taktbaum verringert werden kann. Zusätzliche Logik, wenn auch schwankend, würde die Race-Immunity stets verbessern. Da Hold-Zeit kritische Pfade

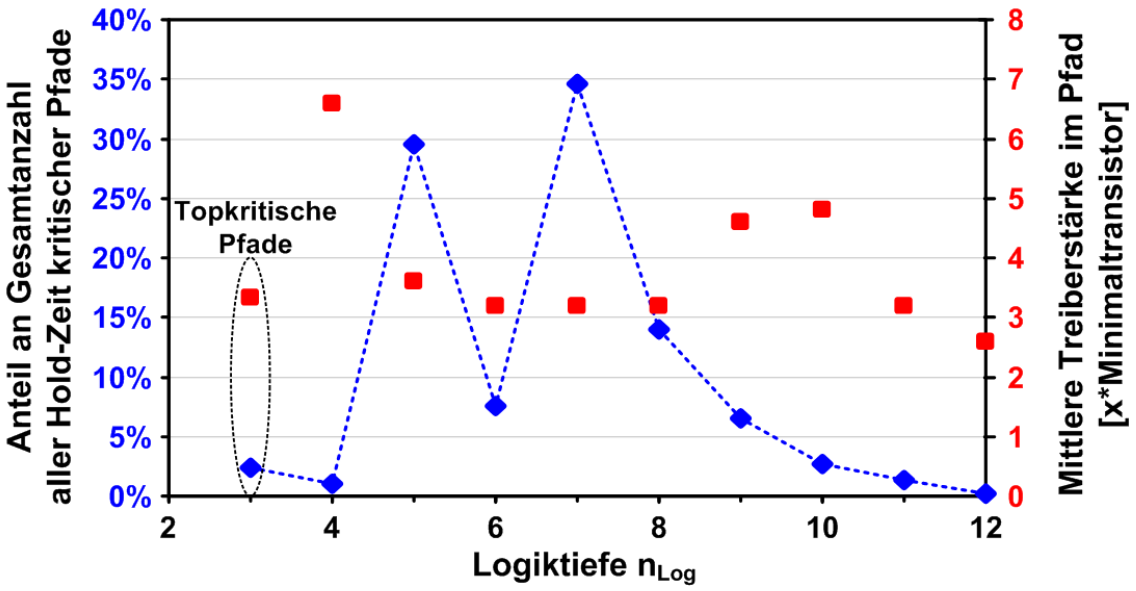


Bild 4.11: Verteilung der Logiktiefe und der mittleren Treiberstärke in den Hold-Zeit kritischen Pfaden.

über den selben Taktbaum mit dem Taktsignal versorgt werden wie Setup-Zeit kritischen Pfade, findet auch hier eine starke Mittelung statistischer Schwankungen über die großen Transistorgeometrien in den verwendeten Clock Buffern statt.

Beschaffenheit des Taktbaums

Da im vorliegenden Fall Hold-Zeit Verletzungen nur durch Unsicherheiten im Taktbaum verursacht werden können, folgt eine nähere Untersuchung der Aufspaltung im Taktbaum. In Bild 4.12 ist die maximale Anzahl an Clock Buffern nach dem Aufspalten für die in Bild 4.10 definierten Timing Bereiche dargestellt. Die top-kritischen Pfade im Timing Bereich A werden zu 30% vom selben lokalen Clock Buffer (LCB) mit dem Taktsignal versorgt, weitere 60% der Taktpfade zu sendendem und empfangendem Flip Flop unterscheiden sich maximal durch einen einzigen Clock Buffer. Diese strukturelle Eigenschaft limitiert den möglichen Clock Skew und somit auch das Risiko einer Hold-Zeit Verletzung.

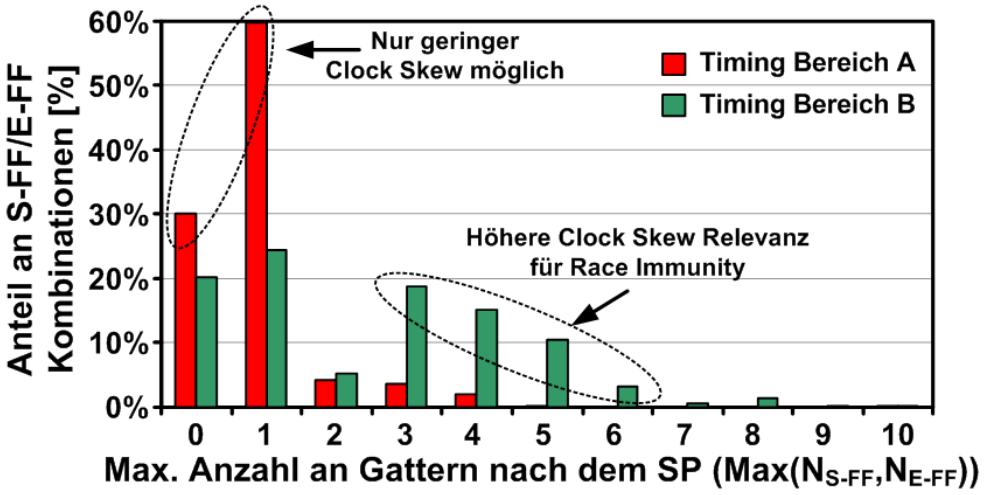


Bild 4.12: Verteilung der Anzahl an Clock Buffern nach dem Aufspalten im Taktbaum für Hold-Zeit kritische Pfade.

4.2 Aufbau des Mikroprozessormodells

In diesem Abschnitt wird ein neuartiges Mikroprozessormodell vorgestellt, das der Analyse von strukturellen Einflussgrößen auf die Geschwindigkeit von getakteten Digitalschaltungen, sowie der Identifikation von sensiblen, sensitiven Schaltungsteilen unter Berücksichtigung von technologischen und betriebsbedingten Variationen dient. Das Modell basiert auf der Beschreibung von Pipelinestrukturen mittels generischer 'kritischer Pfad' Modelle und ermöglicht neben der Abschätzung von variationsbedingten Geschwindigkeitsschwankungen auch Trendaussagen hinsichtlich Veränderungen in der Mikroprozessorarchitektur.

Während L2L, W2W und D2D Variationen alle Transistoren auf dem Chip in gleicher Weise beeinflussen [40, 47], wirken sich WID Variationen, die mit fortschreitender Technologieskalierung einen zunehmend größeren Anteil an der Gesamtvariation stellen [95, 96], je nach Schaltung unterschiedlich auf die Geschwindigkeit aus [45]. Aus diesem Grund werden L2L, W2W und D2D Variationen durch die Verwendung von Prozess-Cornern berücksichtigt. Der Einfluss von WID und betriebsbedingten on-chip Umgebungsvariationen hingegen wird vorwiegend durch OCV Faktoren (On Chip Variation) berücksichtigt, die unabhängig von Mikroarchitektur und Schaltungstopologie zeitliche Margen generieren [15]. Das hier vorgestellte Mikroprozessormodell berücksichtigt schaltungstechnische, topologische und mikroarchitektonische Eigenschaften einer Schaltung bei der Bestimmung der variationsbedingten Laufzeitschwankung.
Als Zeiteinheit wird die aus der Literatur bekannte Fanout-4 Laufzeit verwendet. Fanout-4 bezeichnet die Laufzeit eines Inverters, der mit vier gleichartigen Invertern belastet wird [97]. Ziel dieser Metrik ist es eine technologieunabhängige Zeiteinheit zu verwenden, die einen rein strukturellen Vergleich verschiedener Schaltungen wie z.B. verschiedener Mikroprozessoren ermöglicht. Die FO4 Laufzeit wird deshalb beim Design von Mikroprozessoren verwendet um zeitliche Zielgrößen technologieunabhängig festzulegen [98, 99, 100, 101, 102].
Bild 4.13 zeigt die Simulationsergebnisse auf Basis extrahierter Netzlisten zum Skalierungsverhalten der FO4-Laufzeit im Vergleich zum Gatter-Mix aus NAND4, NOR3, NAND2, NOR2, Buffer, Inverter und Multiplexer Zellen. Dieser Gatter-Mix repräsentiert alle wich-

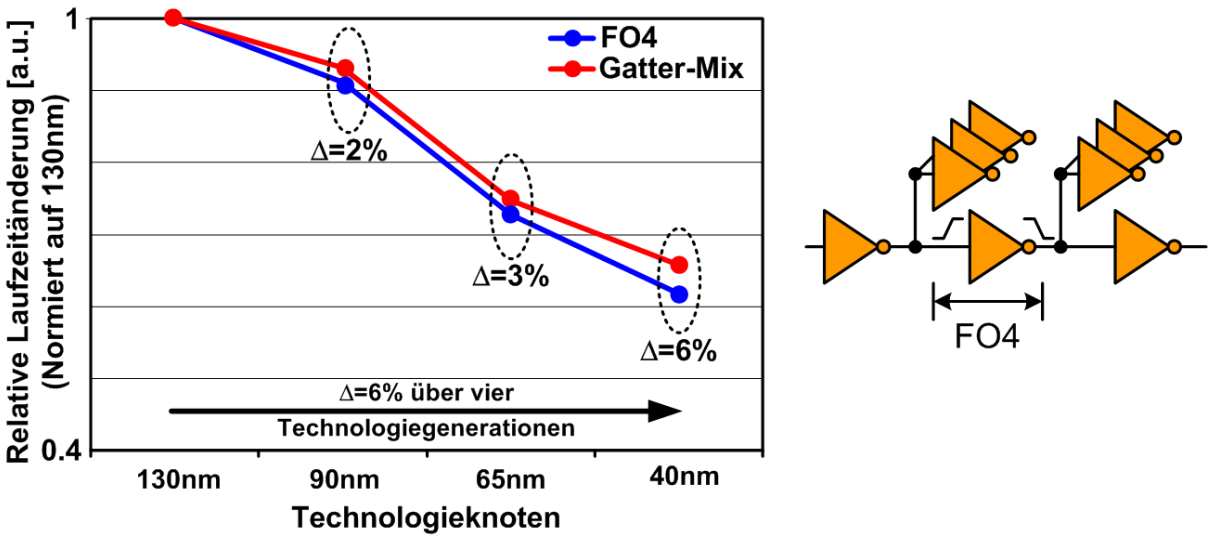


Bild 4.13: Skalierungsverhalten der Fanout-4 Laufzeit im Vergleich zu einem repräsentativen Gattermix.

tigen schaltungstechnischen Unterschiede einer Standardzellen-Bibliothek und ist somit für den Vergleich des Skalierungsverhaltens geeignet.
Während in 90nm und 65nm CMOS Technologien die mittlere Laufzeitskalierung des repräsentativen Gattermixes nur geringfügig um ca. 2% von der FO4-Laufzeit abweicht, ist für 40nm ein Unterschied von ca. 6% zu erkennen. Neben der Technologieskalierung beeinflussen auch das Design der Standardzellen sowie topologieabhängige Größen wie z.B. das Treiber zu Lastverhältnis die Skalierung der Pfadlaufzeiten. So ist es schwer, ein pauschales Skalierungsverhalten der Laufzeit von Standardzellenbibliotheken mit mehr als 300 Logikgattern anzugeben. Die kleinen relativen Abweichungen der skalierten Laufzeiten von maximal nur 6% über vier Technologieknoten hinweg erlauben auch für sub-100nm CMOS Schaltungen die Verwendung der bewährten FO4-Laufzeit als technologieunabhängige Zeitbasis. Eine abstrahierte Betrachtung und Untersuchungen struktureller und mikroarchitektonischer Einflussgrößen auf die von Variationen beeinflusste Geschwindigkeit von getakteten Digitalschaltungen ist somit möglich.

Das Mikroprozessormodell gliedert sich in drei funktionale Einheiten: Den Logikteil, den Taktbaum und die Registerelemente. Bild 4.14 zeigt eine generische Pipelinestufe mit Aufteilung der drei funktionalen Einheiten, die im Folgenden diskutiert werden.

4.2.1 Modellierung der Registeranzahl

Registerelemente sind wesentliche Bestandteile eines Mikroprozessors. Zur synchronen Verarbeitung der verschiedenen Aufgaben sind diese Elemente für die Speicherung der Ergebnisse zuständig. In der Mikroprozessorpipeline werden in den Pipeline-Registern die Zwischenergebnisse gespeichert und für die Verarbeitung in der nächsten Pipelinestufe zur Verfügung gestellt. Neben diesen Datenregistern müssen auch die Kontrollsignale im Prozessor zwischengespeichert und an die jeweiligen Pipelinestufen weitergegeben werden. Daten- und Kontroll-Pipelineregister tragen zur Gesamtanzahl der Registerelemente im Prozessor bei. Diese ändert sich mit der Pipelinetiefe des Prozessors und kann wie folgt

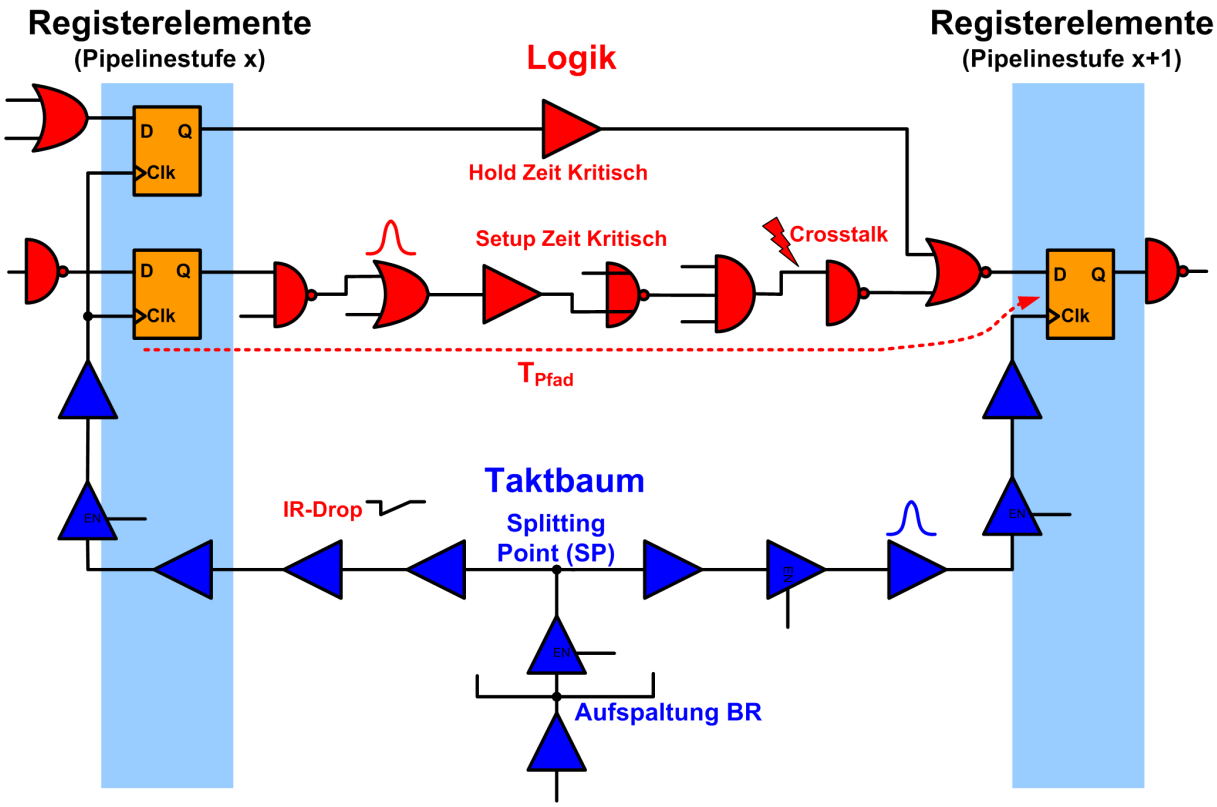


Bild 4.14: Generische Pipelinestruktur: Registerelemente, Logik und Taktbaum.

modelliert werden:

$$t_{Pipeline} \sim \frac{1}{N_{Pipeline}}$$
$$N_{FF}^{B} = N_{FF}^{A} \cdot \left(\frac{t_{Pipeline}^{A}}{t_{Pipeline}^{B}} \right)^{\alpha_{FF}} \sim N_{FF}^{A} \cdot \left(\frac{N_{Pipeline}^{B}}{N_{Pipeline}^{A}} \right)^{\alpha_{FF}} \tag{4.6}$$

Dabei bezeichnet $t_{Pipeline}$ die Laufzeit der Pipelinestufe, $N_{Pipeline}$ die Pipelinestufenanzahl und N_{FF}^{B} die Registeranzahl des Mikroprozessors B mit der im Vergleich zu Mikroprozessor A veränderten Pipelinestufenanzahl. Der Exponent $\alpha_{FF} > 1$ bestimmt den superlinearen Zuwachs an Registerelementen mit steigender Pipelinetiefe, der auf zusätzlich benötigter Hardware z.B. erweiterte Kontrolllogik, Multiplexing etc. zurückzuführen ist. Die aus der Literatur zur Leistungsabschätzung bekannte Modellierung der Registeranzahl zeigt für α_{FF} Werte zwischen 1.1 und 1.3 [99, 103, 104]. Der Vergleich eines 5-stufigen ARM926 und eines 8-stufigen ARM1176 Mikroprozessors ergibt einen Wert von 1.17, so dass als Standardwert 1.2 verwendet wird. Die Anzahl der Registerelemente ist insbesondere für die Verteilung des Taktes entscheidend, da sie die Laufzeit des Taktpfades, sowie die Anzahl der Bufferzellen im Taktbaum beeinflusst.

Im Semicustom-Schaltungsentwurf hat sich als Registerelement das Master-Slave Flip Flop (MS-FF) [105] bewährt. Insbesondere in low-power Schaltungen wird das MS-FF aufgrund der relativ geringen Energieaufnahme bevorzugt verwendet [106]. Der Aufbau aus zwei komplementär getakteten Latches in Serie resultiert jedoch in einer im Vergleich zu anderen Flip Flop Typen relativ hohen Laufzeit, bestehend aus Setup-Zeit t_{SU} und der

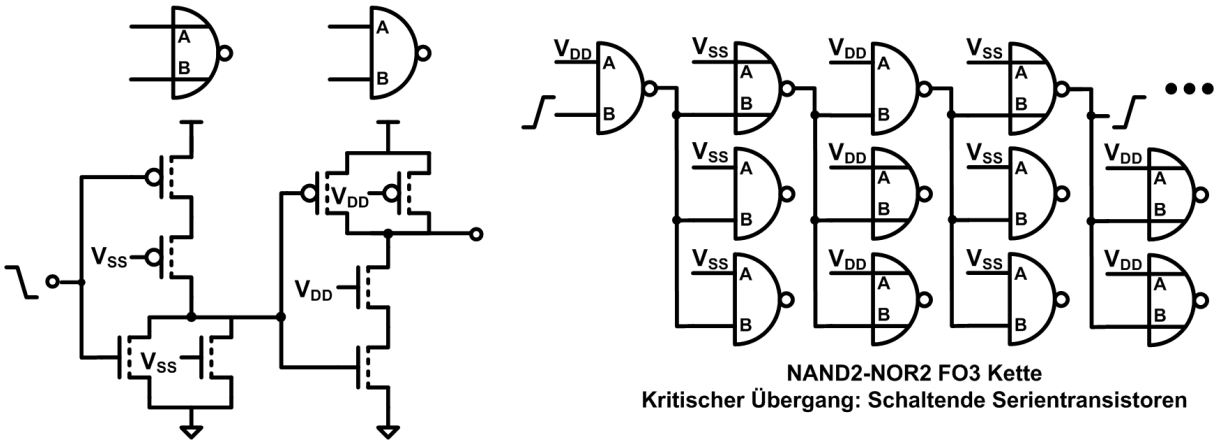


Bild 4.15: NAND2-NOR2 Kette: Generische Struktur zur Nachbildung des Laufzeitverhaltens kritischer Pfade.

Tabelle 4.2: Prozessschwankungsbedingte, relative 1σ Laufzeitschwankung für NAND2-NOR2 Kette und Mittelwert eines für kritische Pfade repräsentativen Gattermixes.

Typ	65nm			90nm		
	0.9V	1.18V	1.32V	0.9V	1.18V	1.32V
Gattermix	8.7%	7.7%	7.6%	9.2%	7.3%	6.9%
ND2-NR2	8.7%	7.7%	7.6%	9.0%	7.3%	6.9%

Clock-zu-Q Laufzeit t_{Clk-Q}. Die Fallstudie eines ARM926 und ARM1176 Mikroprozessors zeigt einen mittleren zeitlichen Beitrag von 7 Inverter-FO4 Laufzeiten zur Gesamtlaufzeit. Studien zum Vergleich von Full-Custom High Performance und Semi-Custom Schaltungen in [107, 108] zeigen Register Beiträge in ASIC Designs von bis zu 10 FO4 Laufzeiten.

4.2.2 Modellierung des Logikblocks

Die maximale Taktfrequenz f_{Clk} einer getakteten Digitalschaltung wird vorwiegend von der Logiklaufzeit zwischen den einzelnen Pipelinestufen bestimmt. Unter Annahme einer idealen Taktverteilung limitiert somit der Pfad mit der maximalen Laufzeit die maximale Taktfrequenz. Da das Mikroprozessormodell nicht alle kritischen Pfade einer Schaltung berücksichtigen kann, ist es wichtig eine repräsentative Darstellung der Schaltungseigenschaften zu finden, die von Schaltung zu Schaltung variieren kann. Daher ist ein gleichartiges Verhalten unter Einfluss von Prozess- und Umgebungsvariationen wichtig. Die Fallstudie des ARM926 zeigt, dass sich die Laufzeit kritischer Pfade eines Mikroprozessors sehr ähnlich dem kritischen Übergang einer NAND2-NOR2 Kette (schaltende Serientransistoren), mit einer Last von jeweils 2 NAND2 und 1 NOR2 bzw. 1 NAND2 und 2 NOR2 Gattern verhält. Bild 4.15 zeigt das schematische Bild einer Fanout-3 belasteten NAND2-NOR2 Kette auf Gatterebene und die zugehörige Implementierung auf Transistorebene.

Tabelle 4.2 zeigt die relative Laufzeitschwankung ($1\sigma_{t_d}$) einer NAND2-NOR2 Kette und den Mittelwert eines für kritische Pfade repräsentativen Gattermixes. Die Ergebnisse der Monte-Carlo Simulation auf Basis extrahierter Netzlisten zeigt nahezu gleiche Schwan-

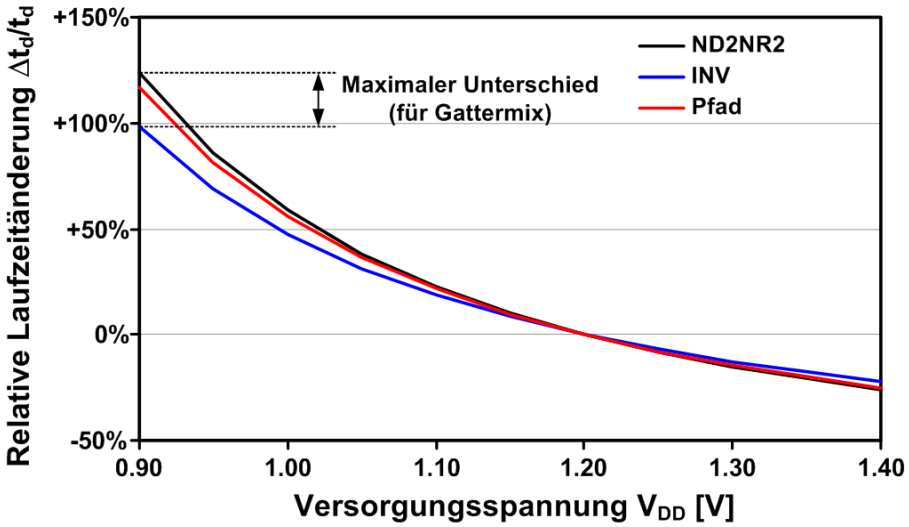


Bild 4.16: Spannungsabhängigkeit von Inverter-Kette, NAND2-NOR2 Kette und repräsentativem kritischen Pfad des ARM926 Mikroprozessors in 65nm CMOS (T=27°C).

kungsbreiten von Gattermix und NAND2-NOR2 Kette für drei verschiedene Spannungen über den gesamten Betriebsbereich einer low-power Schaltung mit Dynamic Voltage Scaling. Die Simulation zweier kritischer Pfade des ARM926 zeigt eine Differenz der Schwankungsbreite zur NAND2-NOR2 Kette von nur 0.2-0.4%. Diese nur geringen Abweichungen zeigen, dass hinsichtlich Prozessvariationen das Verhalten kritischer Pfade durch die NAND2-NOR2 Kette gut nachgebildet werden kann.
Um den Einfluss von betriebsbedingten Variationen auf die Laufzeit zu berücksichtigen, spielt insbesondere die Abhängigkeit von der Versorgungsspannung V_{DD} eine wesentliche Rolle. Bild 4.16 zeigt die Spannungsabhängigkeit der Laufzeit für eine Inverter-Kette, den kritischen Übergang der NAND2-NOR2 Kette sowie einen repräsentativen kritischen Logikpfad des ARM926, d.h einen Logikpfad frei von Design Fehlern, wie z.B. Überbelastungen oder hohem Clock Skew .
Die Untersuchung der Spannungsabhängigkeit zeigt, dass der kritische Übergang einer NAND2-NOR2 Kette, d.h. schaltende Serientransistoren, im Vergleich zu allen Gattern des repräsentativen Gattermixes die stärkste Abhängigkeit von der Versorgungsspannung aufweist. Sie dient als obere Grenze und bildet das Spannungsverhalten kritischer Pfade im geschwindigkeitskritischen Bereich mit guter Genauigkeit - wenngleich etwas pessimistisch - nach. Die FO4 Inverter-Kette zeigt zusammen mit Buffer-Zellen, die in der Regel aus zwei FO4 Invertern aufgebaut sind, die geringste Laufzeitsensitivität gegenüber der Versorgungsspannung V_{DD}, was insbesondere bei der Modellierung des Taktverteilungsnetzes berücksichtigt werden muss.

Die absolute Laufzeit eines Pfades setzt sich aus der Laufzeit der sich im Pfad befindlichen Logikgatter, den RC-Elementen der Signalleitungen sowie dem Crosstalk-Beitrag aller Netze im Pfad zusammen. Da, wie die Ergebnisse der Strukturanalysen zeigen, die Laufzeit kritischer Pfade logikdominiert ist, wird die NAND-NOR Kette als repräsentative Anordnung für kritische Logikpfade verwendet. Tabelle 4.3 zeigt die für ARM926 und ARM1176 repräsentativen Laufzeitbeiträge für drei verschiedene Pfadtypen, die sich durch verschiedene Laufzeitanteile aus Logik, Verdrahtung und Crosstalk unterscheiden.

Tabelle 4.3: Laufzeitbeiträge im Logikpfad für verschiedene Pfadtypen.

Pfadtyp	Logik-Anteil	Crosstalk-Anteil	RC-Anteil
Logikdominiert	94%	4%	2%
Crosstalk-Sensitiv	85%	10%	5%
RC-Sensitiv	86%	6%	8%

Der Crosstalk-sensitive Pfadtyp erhält als Crosstalk-Beitrag den Wert des 90% Perzentils, der sich aus der Verteilung der Crosstalk-Beiträge aller kritischen Pfade ergibt. Ebenso werden die Beiträge von logikdominiertem Pfad und RC-sensitivem Pfad gewonnen. Im Laufe der Untersuchungen wurde jedoch festgestellt, dass für die in Tabelle 4.3 aufgeführten Werte die Laufzeitschwankung von logikdominierten Pfaden stets größer ist als die Schwankung der anderen Pfadtypen, so dass im Folgenden nur logikdominierte Pfade verwendet werden.

Die Laufzeit des Logikblocks wird für den nominellen Fall wie folgt berechnet:

$$T_{Pfad} = \underbrace{t_{Clk-Q} + t_{Log} + t_{SU}}_{t_{Comb}} + t_{XT} + t_{RC} \tag{4.7}$$

Die Pfadlaufzeit setzt sich aus den Beiträgen von sendendem Flip Flop (S-FF) und empfangendem Flip Flop (E-FF) t_{Clk-Q} und t_{SU}, sowie den Beiträgen aus Logiklaufzeit t_{Log}, Crosstalk t_{XT} und Leitungsverzögerung t_{RC} zusammen.
Im folgenden Abschnitt wird die Modellierung des strukturabhängigen Einflusses von Variationen gezeigt.

Berücksichtigung von Prozessvariationen:

Da sich kritische Pfade gegenüber globalen L2L, W2W und D2D Variationen, deren Einfluss unabhängig von der jeweiligen Schaltungsstruktur ist, gleich verhalten, werden diese Variationen im Modell über globale Worst-Case Prozess-Corner berücksichtigt [40, 47].
Lokale, statistische Variationen gehen durch die Beiträge von Random Dopant Fluctuations zur Einsatzspannungsschwankung (σ_{V_T}) und Strom-Mismatch (σ_k) in die Laufzeitschwankung ein. Die Schwankungsbreiten der jeweiligen Technologie sind bekannt und werden im Modell standardmäßig für eine Transistorgröße von einem vollen Transistorfinger verwendet (abhängig von der jeweiligen Standardzellenbibliothek). Der Einfluss auf die Laufzeit des Logikteils wird wie folgt berechnet:

$$\sigma_{t_{Comb}} = t_{Comb} \cdot \sqrt{\frac{1}{F} \cdot \frac{1}{n_{Log}} \left(\left(S_{V_T,rel} \cdot \sigma_{V_T} \right)^2 + \left(S_{k,rel} \cdot \sigma_k \right)^2 \right)} \tag{4.8}$$

Dabei bezeichnet F die mittlere Anzahl (Mittelwert über den Pfad) an Paralleltransistoren (Finger) je Gatter , n_{Log} die Anzahl der Logikstufen im Pfad, $S_{V_T,rel}$ und $S_{k,rel}$ die relative Laufzeitsensitivität gegenüber V_T Schwankungen und Strom-Mismatch.
Über den Faktor F wird die Größe der verwendeten Transistoren modelliert. Je größer ein Transistor, desto geringer der V_T Schwankungsanteil aufgrund von RDF, sowie der Strom-Mismatch Anteil. Da die Laufzeit ein lineares Verhalten gegenüber V_T Variationen aufzeigt, kann die Schwankungsbreite für einen Standardtransistor σ_{V_T} verwendet und die Laufzeit somit in direktes Verhältnis zum Flächenfaktor F gestellt werden.

Über die Anzahl der Logikstufen n_{Log} wird die Größe der Mittelungseffekte von statistischen Variationen festgelegt. Je höher die Logikstufenanzahl, desto kleiner die relative Laufzeitschwankung [75, 22, 91]. Die Logikstufenanzahl n_{Log} eines Pfades verringert sich mit tieferem Pipelining und kann wie folgt dargestellt werden:

$$n_{Log}^{B} = n_{Log}^{A} \cdot \frac{N_{Pipeline}^{A}}{N_{Pipeline}^{B}} + n_{Arch} \tag{4.9}$$

$N_{Pipeline}^{A}$ ist die Pipelinestufenanzahl eines Mikroprozessor vor und $N_{Pipeline}^{B}$ nach mikroarchitektonischer Veränderung der Pipeline. Über n_{Arch} werden mikroarchitekturspezifische Veränderungen berücksichtigt, die z.B. in den einzelnen Pipelinestufen zusätzliche Multiplexer Zellen erfordern, um zwischen verschiedener Hardware wie z.B. verschiedenen Execution Units zu wählen.
Zur Veranschaulichung der Mittelungseffekte wird ein Dämpfungsfaktor DF gegenüber der statistischen Laufzeitschwankung eines Minimalinverters $\sigma_{t_{Inv}}$ wie folgt definiert:

$$DF = \frac{1}{\sqrt{\frac{W}{W_{min}} \cdot n_{Log}}} \tag{4.10}$$

W bezeichnet die mittlere Transistorweite, W_{min} die Transistorweite eines Minimalinverters der jeweiligen Standardzellenbibliothek. Bild 4.17 zeigt schematisch den Dämpfungsfaktor statistischer Laufzeitvariationen durch statistische Mittelung als Funktion von Logikstufenanzahl n_{Log} und dem Vielfachen x der minimalen Transistorgröße der verwendeten Standardzellenbibliothek. Weiße und gelbe Kreuze markieren die 95% und 5% Perzentile sowie den Mittelwert der Logikstufenverteilung in den geschwindigkeitskritischen Pfaden der untersuchten Mikroprozessoren. In den kritischen Pfaden des ARM926 und ARM1176 werden statistische Variationen um den Faktor 0.04-0.06 gegenüber der Laufzeitschwankung eines Minimalinverters abgeschwächt, d.h. eine Laufzeitschwankung des Minimalinverters von $\sigma_{t_{Inv}} = 10\%$ führt zu statistischen Laufzeitschwankungen des Pfades von $\sigma_{t_{Pfad}} < 1\%$. Im Vergleich zu den geschwindigkeitskritischen Strukturen in eingebetteten Mikroprozessoren ist die Dämpfung für High-Speed Addierer Architekturen deutlich geringer. Bei einer Logikstufenanzahl von 5-10 Stufen (je nach Implementierung), liegt der Dämpfungsfaktor bei unter 20%. Es ist jedoch zu erkennen, dass für Transistorgrößen ab $5W_{min}$ und einer Logiktiefe des Pfades von 10 Stufen eine deutliche Abschwächung der statistischen Laufzeitschwankung stattfindet, so dass selbst für tieferes Pipelining (z.B. 13-stufiger ARM Cortex A8) kein signifikanter Anstieg der statistischen Laufzeitschwankung im Logikpfad zu erwarten ist.
Hold-Zeit kritische Pfade unterliegen einem deutlich geringerem Mittelungseffekt der relativen Schwankungsbreite. Es ist jedoch zu beachten, dass die absolute Schwankungsbreite mit $\sqrt{n_{Log}}$ zunimmt, so dass Pfade mit geringer Logikstufenanzahl zwar eine hohe relative Schwankungsbreite, absolut gesehen jedoch geringe Schwankungsbreiten aufweisen.

Berücksichtigung von Umgebungsvariationen:

Aufgrund der stark ansteigenden Laufzeitsensitivität gegenüber Versorgungsspannungsschwankungen trägt IR-Drop signifikant zur Laufzeitschwankung bei. Zur Bestimmung von IR-Drop Beiträgen zur Gesamtlaufzeit wird im Mikroprozessormodell die Spannungsabhängigkeit der Laufzeit wie folgt modelliert.

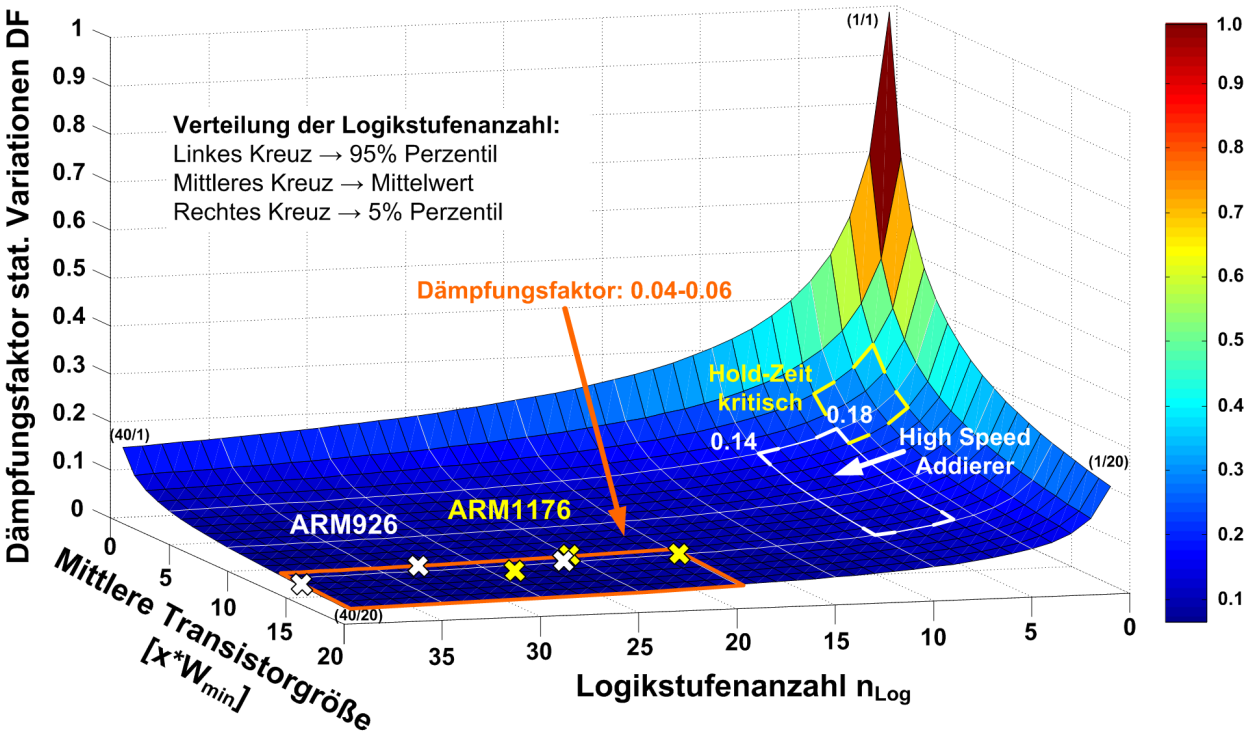


Bild 4.17: Gatter- und pfadtopologieabhängige Dämpfung von statistischen Laufzeitvariationen.

In Bild 3.2 ist die Schalttrajektorie eines Inverters und eines NAND2 Gatters mit schaltendem Serientransistor gezeigt. Es ist zu erkennen, dass die Form der Trajektorie und somit die durchlaufenen Betriebspunkte (V_{GS}, V_{DS}) maßgeblich durch die Gattertopologie bestimmt werden. Dieser Einfluss der Gattertopologie wird für ein spannungsabhängige Laufzeitmodell in den Parametern $V_{T_{eff}}$ und α_{dyn} berücksichtigt, so dass sich für die Laufzeit folgender Zusammenhang ergibt [109]:

$$t_d \sim \frac{C_{eff} \cdot V_{DD}}{\left(V_{DD} - V_{T_{eff}}\right)^{\alpha_{dyn}}} \tag{4.11}$$

Die effektive Einsatzspannung $V_{T_{eff}}$ für schaltende Inverter-Ketten und kombinatorische Logik wird wie folgt bestimmt:

$$V_{T_{eff}} = V_{T_{sat}} + x \cdot V_{DIBL} \tag{4.12}$$

$V_{T_{sat}}$ ist die Einsatzspannung des Transistor in Sättigung bei nomineller Versorgungsspannung, V_{DIBL} die Einsatzspannungsänderung durch Drain Induced Barrier Lowering (DIBL) bei nomineller Versorgungsspannung und x ein schaltungsspezifischer Faktor. Der Faktor x beträgt $\frac{2}{3}$ für Inverter-Ketten und $\frac{9}{10}$ für kombinatorische Logik und berücksichtigt somit die spezifische Abhängigkeit von V_T gegenüber V_{DS}. Der Exponent α_{dyn} ist nahezu technologieunabhängig und liegt bei ca. 2. Er darf nicht mit dem in [16] verwendetem α verwechselt werden!
Ein Vergleich von Simulationen auf Basis extrahierter Netzlisten von FO4 Inverter-Kette und NAND2-NOR2 Kette mit den Ergebnissen des Laufzeitmodells ergibt für Technologien von 130nm bis 65nm einen relativen Fehler von unter ±2% für den produktrelevanten Betriebsbereich der Versorgungsspannung. Bild 4.18 zeigt einen Vergleich von Simulation

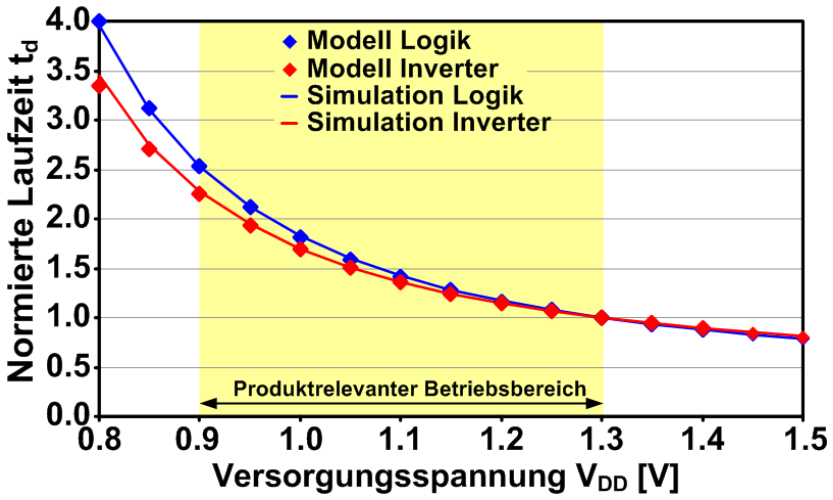


Bild 4.18: Spannungsabhängiges Laufzeit-Modell für kombinatorische Logik und Inverter Kette in 65nm CMOS. Die Simulation basiert auf extrahierten Netzlisten (T=27°C).

und Laufzeitmodell für 65nm CMOS normiert auf 1.3V (High-V_{DD} Overdrive).
Die Spannungsabhängigkeit der Laufzeit t_d kann somit mit ausreichender Genauigkeit modelliert werden.

Im Modell wird zwischen zwei Komponenten von IR-Drop unterschieden: Dem lokalen Spannungseinbruch $\Delta V_{DD,lok}$, d.h. der unterschiedliche Spannungseinbruch zwischen zwei verschiedenen Positionen auf dem Die, sowie dem globalen IR-Drop, der alle Komponenten der Schaltung in gleichem Maße beeinflusst. Die relative Laufzeitänderung aufgrund von IR-Drop wird wie folgt bestimmt:

$$t_{Comb}^{IR} = t_{Comb}^{nom} \cdot \left(\frac{(V_{DD,nom} - \Delta V_{DD,glo} - \Delta V_{DD,lok}) \cdot (V_{DD,nom} - V_{T,eff})^{\alpha_{dyn}}}{V_{DD,nom} \cdot (V_{DD,nom} - \Delta V_{DD,glo} - \Delta V_{DD,lok} - V_{T,eff})^{\alpha_{dyn}}} - 1 \right) \quad (4.13)$$

t_{Comb}^{nom} ist die nominelle Laufzeit des Logikanteils, $V_{DD,nom}$ die nominelle Versorgungsspannung, $\Delta V_{DD,glo}$ der globale, mittlere Spannungsabfall (z.B. am Spannungsregler) und $\Delta V_{DD,lok}$ der lokale, mittlere Abfall der Versorgungsspannung am Zuleitungswiderstand. Da im Modell mit verschiedenen Pfadtypen gerechnet werden kann, ist es erforderlich zu überprüfen, ob sich das Spannungsverhalten von logikdominierten und leitungsdominierten Pfaden signifikant unterscheidet. Dazu wurde eine Testschaltung in 45nm CMOS konzipiert, die verschiedene Ringoszillatoren beinhaltet, die als Lastelement unterschiedlich lange Leitungen mit beidseitigen Aggressoren treiben. Eine crosstalkbedingte Laufzeiterhöhung von über 30%, wie in Bild 3.10 über die relative Frequenzänderung dargestellt ist, zeigt das crosstalkdominierte Variationsverhalten der gewählten Struktur und den großen Einfluss der Leitungen auf das Laufzeitverhalten der Testschaltung.
Bild 4.19 zeigt die in 45nm CMOS gemessene normierte Oszillationsfrequenz eines logik-, leitungs- und crosstalkdominierten Pfades in Abhängigkeit der Versorgungsspannung.
Die relative Abweichung der Laufzeitsensitivität von logik- und leitungsdominiertem Pfad im IR-Drop relevanten Bereich liegt hier für einen Bereich von 200mV bei unter 1.5%. Unter Berücksichtigung der Messgenauigkeit bleibt somit ein vernachlässigbar geringer Unterschied. Im Vergleich dazu zeigt der crosstalkdominierte Pfad einen größeren Unterschied zum Spannungsverhalten des logikdominierten Pfades. Da der Crosstalkbeitrag zur

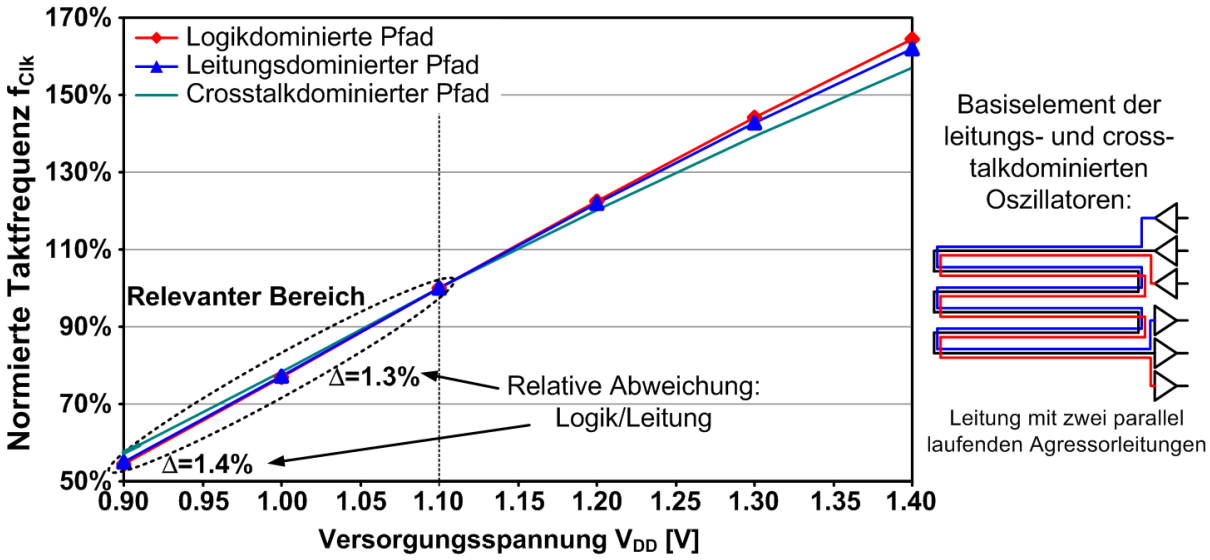


Bild 4.19: Gemessenes Spannungsverhalten logik-, leitungs- und crosstalkdominierter Pfade in 45nm CMOS (T=27°C).

Gesamtlaufzeit im vorliegenden Fall immer deutlich geringer ist als der Laufzeitbeitrag der Logik, tritt kein crosstalkdominiertes Verhalten auf, so dass das Spannungsverhalten des logikdominierten Pfads auch für die Modellierung der Laufzeitsensitivität von crosstalksensitiven Pfaden gegenüber V_{DD} verwendet wird. Aus diesem Grund wird im Folgenden nicht zwischen den einzelnen Laufzeitsensitivitäten gegenüber V_{DD} unterschieden.

Im Modell werden verschiedene Pfadtypen verwendet, um unter anderem auch den Laufzeitbeitrag aus Crosstalk Ereignissen zu quantifizieren. Wie bereits in Kapitel 3.1 festgestellt, ist der Crosstalk-induzierte Laufzeitbeitrag von vielen verschiedenen topologischen sowie zeitlichen Einflussgrößen abhängig. Nachdem Crosstalkeffekte durch Transitionen auf kapazitiv gekoppelten Nachbarleitungen hervorgerufen werden, lässt sich der Crosstalkbeitrag unabhängig von implementierungsabhängigen Zusammenhängen über die Anzahl der Signaltransitionen gewichten. Für die Modellierung ist die architekturabhängige Veränderung der Transitionsanzahl entscheidend. Diese ist abhängig von der Schaltaktivität α_{Schalt} und der durch Glitches erzeugten zusätzlichen Anzahl von Transitionen. Der Gewichtungsfaktor des Crosstalkbeitrags X_{Schalt} lässt sich aus der Glitch-Power Modellierung in [99] wie folgt bestimmen:

$$\text{Anzahl an Transitionen:} \qquad \sim \alpha_{Schalt} + \beta \tag{4.14}$$

$$\text{Architekturabhängigkeit:} \qquad X_{Schalt} \sim \frac{\alpha_{Schalt,B} + \beta_B}{\alpha_{Schalt,A} + \beta_A} \sim \frac{1 + \beta_A \cdot \frac{N_{Pipeline,A}}{N_{Pipeline,B}}}{1 + \beta_A} \quad \text{für}_{\alpha_{Schalt,A}=\alpha_{Schalt,B}} \tag{4.15}$$

Mit β_A wird der aufgrund von Glitches verursachte Anteil an der aktiven Verlustleistung bezeichnet, der mit tieferem Pipelining, d.h. $N_{Pipeline,B} > N_{Pipeline,A}$, abnimmt. Die Schaltaktivität α_{Schalt} ist unabhängig von der Mikroarchitektur, da die eigentliche Befehlsverarbeitung nicht verändert wird. Für die gleiche Schaltaktivität ergibt sich dadurch eine um den Faktor X_{Schalt} reduzierte Anzahl von Transitionen bei erhöhter Pipelinestufenanzahl

$N_{Pipeline,B}$. Der Einfluss des mikroarchitekturabhängigen Crosstalkanteils ist dabei stark vom Anteil der durch Glitches verursachten Verlustleistung an der Gesamtverlustleistung abhängig. Da diese mit einem Anteil von 10%-60% sehr stark schwankt [110], ist die Modellierung des Crosstalkanteils an der Laufzeitschwankung nur bei genauer Kenntnis von β_A möglich.

4.2.3 Modellierung des Taktverteilungsnetzes

Der Taktbaum besitzt im Vergleich zur Logik eine deutlich homogenere Struktur. Da der Taktbaum im Wesentlichen aus Bufferzellen besteht, die aus zwei Inverterstufen aufgebaut sind, kann das Verhalten des Taktbaums gegenüber Prozess- und Umgebungsvariationen durch eine Inverterkette mit FO4 Belastung nachgebildet werden.
Im Gegensatz zum Logikteil kann im Taktbaum auch eine verkürzte Laufzeit zu Setup-Zeit Fehlern und somit zum funktionalen Ausfall der Logik führen. Grund hierfür ist die differenzielle Anordnung zweier Taktpfade. Ein Taktpfad führt zum Startpunkt des Logikpfades, dem sendenden Register. Seine Laufzeit wird mit $t_{Clk,S}$ bezeichnet. Der andere Taktpfad mit der Laufzeit $t_{Clk,E}$ führt zum empfangenden Register. Bild 4.20(a) zeigt die generische Struktur zweier Taktpfade, die das sendende und empfangende Register eines Logikpfades synchronisieren. Beide Pfade beinhalten die Clock Buffer (CB) 1 und 2 und jeweils drei weitere Clock Buffer. Der Punkt der Aufspaltung wird als Splitting Point SP bezeichnet. Im vorliegenden Fall gilt $SP = 2$.

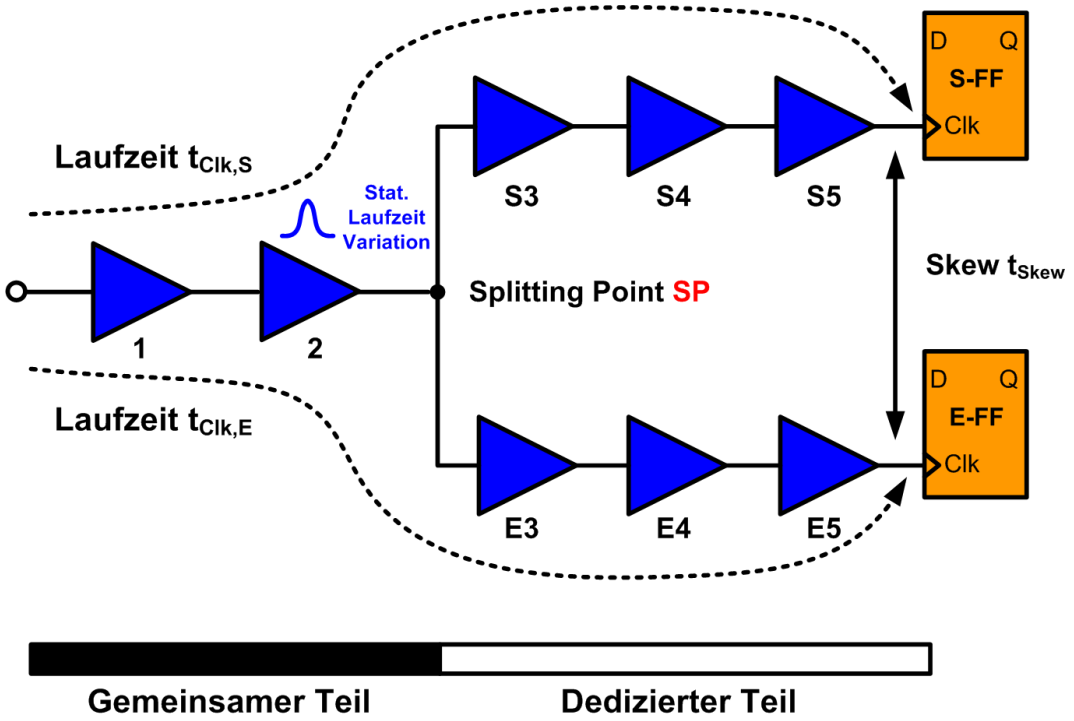
Die Struktur des Taktbaumes ist von verschiedenen Faktoren abhängig. Im Modell wird der Taktbaum ausgehend von der Registeranzahl über die Parameter SP, BR, N_{FF} und $N_{FF/LCB}$ bestimmt. Die Registeranzahl, die entweder bekannt ist oder mittels Gleichung 4.6 bestimmt wird, dient als Ausgangspunkt. Um das Taktsignal an alle Register der Schaltung zu verteilen, ist es notwendig, dass vom Einspeisepunkt des Taktsignals aus eine Aufspaltung erfolgt. Der mittlere Faktor der Aufspaltung, die an jedem Clock Buffer Ausgang erfolgt, wird mit BR (engl.: Branching) bezeichnet. Die aus der Literatur bekannten idealen Taktbaumstrukturen „Binary Clock-Tree“ und „H-Tree“ zeichnen sich durch Aufspaltungsfaktoren von $BR = 2$ bzw. $BR = 4$ aus [111, 112].
Dabei lässt sich die Anzahl der verwendeten lokalen Clock Buffer N_{LCB} wie folgt bestimmen:

$$N_{LCB} = BR^{n_{CB}-1} = N_{FF}/N_{FF/LCB} \tag{4.16}$$

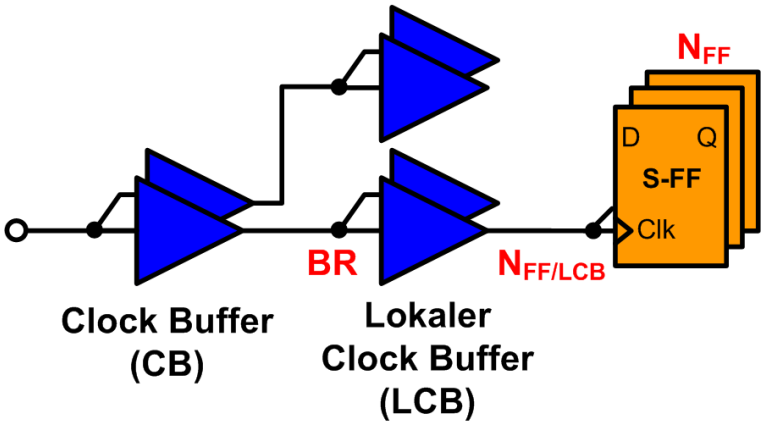
Je nach Platzierung der Registerzellen im Design versorgt ein Lokaler Clock Buffer (LCB) eine unterschiedliche Anzahl von Registern $N_{FF/LCB}$. Je mehr Register vom selben LCB synchronisiert werden, desto geringer die Anzahl der Clock Buffer n_{CB} in einem Taktpfad, die benötigt werden, um das Taktsignal an die Register zu verteilen.

$$n_{CB} = \log_{BR}\left(\frac{N_{FF}}{N_{FF/LCB}}\right) + 1 \tag{4.17}$$

Bild 4.20(b) veranschaulicht die Bedeutung der einzelnen Modellparameter.
Neben der Anzahl der Clock Buffer Zellen, die das Taktsignal vom Einspeisepunkt bis zum jeweiligen Register verteilen, ist es notwendig die mittlere Laufzeit dieser Zelle zu bestimmen. Basierend auf der Analyse des ARM926 Mikroprozessor wird als Standardwert eine Clock Buffer Laufzeit t_{CB} von 2.7 FO4 festgelegt.



(a) Generische Taktbaum-Struktur mit den Pfaden zu sendendem und empfangendem Register.



(b) Kernparameter zur Modellierung des Taktbaums.

Bild 4.20: Modellierung des Taktverteilungsnetzes.

Berücksichtigung von Prozessvariationen:

Lokale statistische Schwankungen, die die Laufzeit der Taktpfade beeinflussen, führen zu einer schwankenden Ankunftszeit des Taktsignals am Register. Im Idealfall erreichen die Taktsignale das sendende Register und das empfangende Register zur gleichen Zeit. Es gilt:

$$t_{Skew} = t_{Clk,E} - t_{Clk,S} = 0 \tag{4.18}$$

Mit t_{Skew} wird der zeitliche Unterschied der Pfadlaufzeiten zum sendenden $t_{Clk,S}$ und empfangenden Register $t_{Clk,E}$ bezeichnet. Die Varianz von t_{Skew} wird wie folgt berechnet:

$$\sigma_{t_{Skew}} = (n_{CB} - SP) \cdot t_{CB} \cdot \frac{\sqrt{2} \cdot \sigma_{t_{CB,rel}}}{\sqrt{n_{CB} - SP}} = \sqrt{2} \cdot t_{CB} \cdot \sigma_{t_{CB,rel}} \cdot \sqrt{n_{CB} - SP} \tag{4.19}$$

Die beiden Taktpfade, die sich am Splitting Point aufspalten, werden dabei als gleichartig betrachtet. Die Schwankungsbreite der Laufzeitdifferenz von $t_{Clk,S}$ und $t_{Clk,E}$ wird daher über die mit $\sqrt{2}$ multiplizierte relative Laufzeitschwankung eines einzelnen Clock Pfades nach dem Aufspalten berücksichtigt.

Systematische, layoutunabhängige WID Prozessschwankungen werden im Modell durch den Faktor $\delta t_{WID,rel}$ berücksichtigt, der dem maximalen Wert der relativen Laufzeitunterschiede auf einem Die entspricht. Dieser Beitrag resultiert aus globalen Schwankungen über den Wafer, die zu einem globalen, monotonen Die-Trend führen. In [113] wird experimentell gezeigt, dass sich diese langreichweitigen Variationen über den gesamten Die monoton verhalten. Je nach Position des Dies auf dem Wafer verändert sich die Orientierung des Die-Trends und somit auch der Einfluss auf die on-chip Laufzeitschwankung [13]. Der Beitrag dieser WID Schwankung auf den Clock Skew t_{Skew}^{WID} wird im Modell wie folgt berücksichtigt:

$$t_{Skew}^{WID} = \left(\frac{1}{BR}\right)^{\lfloor \frac{SP}{2} \rfloor} \cdot (n_{CB} - SP) \cdot t_{CB} \cdot \delta t_{WID,rel} \tag{4.20}$$

Beim Entwurf des Taktbaums ist die Balancierung der Pfadlaufzeiten ein wesentliches Designkriterium, das hauptsächlich über eine gleichartige Belastung der Clock Buffer realisiert wird. Bild 4.21 veranschaulicht diesen Zusammenhang am Beispiel einer generischen Taktbaumstruktur mit monotonem Die-Trend der Laufzeitschwankung im Hintergrund. Wählt man als Splitting Point SP=2, so ergibt sich aus 4.20 eine um den Faktor 0.25 reduzierte Schwankung gegenüber $\delta t_{WID,rel}$. Der in den Chip eingezeichnete Sektor, der durch den Clock Buffer am Punkt 2 versorgt wird, erstreckt sich über ein Chipgebiet, das nur noch von 25% der ursprünglichen WID Schwankungsbreite betroffen ist, d.h. die differenzielle Laufzeitschwankung wird auf 25% absinken.
Im Falle eines horizontalen Die-Trends wäre bereits für SP=1 nur die Hälfte der Laufzeitschwankung zu erwarten. Da die Orientierung des Die-Trends von Die zu Die unterschiedlich ist [13], wird im Modell erst nach jedem zweiten Orientierungswechsel die Schwankungsbreite um den Branching-Faktor BR reduziert.

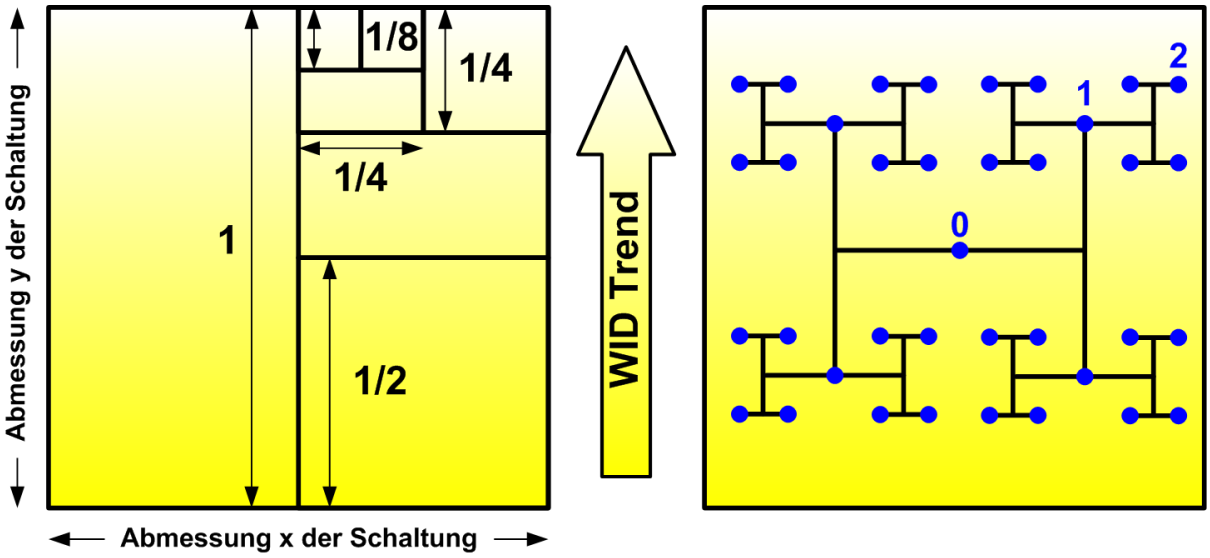


Bild 4.21: Berücksichtigung von WID Laufzeitschwankungen im Taktverteilungsnetz.

Berücksichtigung von Umgebungsvariationen:

Neben den statisch wirkenden Prozessvariationen führen betriebsbedingte Schwankungen der Versorgungsspannung zu veränderten Laufzeiten im Taktbaum. Verkürzt sich die Laufzeit des Taktpfades zum Empfangsregister des kritischen Pfades im Vergleich zum Taktpfad des sendenden Registers, so verkürzt sich die effektive Taktperiode. Die Wahrscheinlichkeit für eine Setup-Zeit Verletzung aufgrund der verkürzten Taktperiode steigt. Ein derartiges Szenario tritt auf, wenn der Spannungsabfall der Versorgungsspannung am Zuleitungswiderstand des Empfangs-Taktpfades geringer ist, als der Spannungsabfall an der Versorgungsleitung des Taktpfades zum sendenden Register einen Taktzyklus zuvor. Diese Spannungsdifferenz kann sowohl aus einem lokalen Spannungsunterschied $\Delta V_{DD,loc}$ als auch aus einer schwankenden, globalen Versorgungsspannung $V_{DD,nom}(t)$ erfolgen. Der Clock Jitter lässt sich wie folgt bestimmen:

$$t_{Jitter} = -F_{rel}^{Clk} \cdot n_{CB} \cdot t_{CB} \tag{4.21}$$

Der Faktor F entspricht der relativen Laufzeitänderung für eine um ΔV_{DD} veränderten Versorgungsspannung im Logik- und Taktpfad und wird wie folgt berechnet:

$$F_{rel}^{Clk} = \frac{\left(V_{DD,nom} - \Delta V_{DD}^{Clk-E}\right) \cdot \left(V_{DD,nom} - \Delta V_{DD}^{Clk-S} - V_{T,eff}^{Clk}\right)^{\alpha_{dyn}}}{\left(V_{DD,nom} - \Delta V_{DD}^{Clk-S}\right) \cdot \left(V_{DD,nom} - \Delta V_{DD}^{Clk-E} - V_{T,eff}^{Clk}\right)^{\alpha_{dyn}}} - 1 \tag{4.22}$$

$$F_{rel}^{Log} = \frac{\left(V_{DD,nom} - \Delta V_{DD}^{Log(Z_2)}\right) \cdot \left(V_{DD,nom} - \Delta V_{DD}^{Log(Z_1)} - V_{T,eff}^{Log}\right)^{\alpha_{dyn}}}{\left(V_{DD,nom} - \Delta V_{DD}^{Log(Z_1)}\right) \cdot \left(V_{DD,nom} - \Delta V_{DD}^{Log(Z_2)} - V_{T,eff}^{Log}\right)^{\alpha_{dyn}}} - 1 \tag{4.23}$$

Die veränderte Versorgungsspannung ΔV_{DD} resultiert aus zeitabhängigen lokalen und globalen Spannungsschwankungen im sendenden und empfangenden Takt- sowie im Logikpfad während des Taktzyklus Z_1 und Z_2. Da die Laufzeitänderung aufgrund von Versorgungsspannungsschwankungen vom Mittelwert des Spannungseinbruchs abhängig ist [51, 114], wird der zeitabhängige Spannungseinbruch für die Jitter-Berechnung wie folgt

bestimmt:

$$\Delta V_{DD}^{Clk-S} = \frac{1}{n_{CB} t_{CB}} \int_{T_0+T_{Clk}-n_{CB}t_{CB}}^{T_0+T_{Clk}} (V_{DD}^{nom} - V_{DD}^{S-Clk}(t))\, dt \tag{4.24}$$

$$\Delta V_{DD}^{Clk-E} = \frac{1}{n_{CB} t_{CB}} \int_{T_0+2\cdot T_{Clk}-n_{CB}t_{CB}}^{T_0+2\cdot T_{Clk}} (V_{DD}^{nom} - V_{DD}^{E-Clk}(t))\, dt \tag{4.25}$$

$$\Delta V_{DD}^{Log(Z_1)} = \frac{1}{T_{Clk}} \int_{T_0}^{T_0+T_{Clk}} (V_{DD}^{nom} - V_{DD}^{Log}(t))\, dt \tag{4.26}$$

$$\Delta V_{DD}^{Log(Z_2)} = \frac{1}{T_{Clk}} \int_{T_0+T_{Clk}}^{T_0+2\cdot T_{Clk}} (V_{DD}^{nom} - V_{DD}^{Log}(t))\, dt \tag{4.27}$$

$$\tag{4.28}$$

Bild 4.22 zeigt die vereinfachte schematische Darstellung der zeitabhängigen Versorgungsspannungen an sendendem und empfangendem Takt- sowie am kritischen Logikpfad. Hier werden die mittleren Spannungseinbrüche während des Taktzyklus Z_1 in der Logik $\Delta V_{DD}^{Log(Z_1)}$ und im Taktpfad zum sendenden Register ΔV_{DD}^{Clk-S}, sowie während des Taktzyklus Z_2 in der Logik $\Delta V_{DD}^{Log(t_2)}$ und im Taktpfad zum empfangenden Register ΔV_{DD}^{Clk-E} berechnet. Im Modell wird ein aus [115, 114] abgeleitetes, sägezahnartiges Spannungsprofil für die an den Gattern wirkende Versorgungsspannung verwendet. Die lokalen Spitzen des Spannungseinbruchs liegen bei steigender und fallender Flanke des verteilten Taktes. Der Spannungseinbruch bei fallender Taktflanke wird nach den Ergebnissen von [115, 114] auf 75% des maximalen Spannungseinbruchs geschätzt. Die IR-Drop Spitzen werden im Modell so angepasst, dass der Mittelwert über einen Taktzyklus z.B. $T_0 \rightarrow T_0 + T_{Clk}$ dem als mittleren IR-Drop bekannten bzw. angenommenen Wert entspricht.
Die Spannungseinbrüche im Logikpfad müssen ebenfalls berücksichtigt werden, da im Falle einer Taktzyklen-Verkürzung durch Clock Jitter gleichzeitig eine Beschleunigung des Logikpfades aufgrund der im Vergleich zum vorherigen Zyklus möglicherweise erhöhten globalen Versorgungsspannung auftreten kann. Deshalb wird ein Korrekturterm eingeführt, der die Beschleunigung der Logik gegenüber dem worst-case IR-Drop berücksichtigt:

$$\Delta t_{Comb}^{IR} = F_{rel}^{Log} \cdot t_{Log} \tag{4.29}$$

Im Taktzyklus $[T_0; T_0 + T_{Clk}]$ werden sowohl sendender und empfangender Taktpfad als auch der Logikpfad mit der Spannung $V_{DD} - \Delta V_{DD,glo} - \Delta V_{DD,lok}$ betrieben. Im Taktzyklus $[T_0 + T_{Clk}; T_0 + 2 \cdot T_{Clk}]$ reduziert sich der globale IR-Drop um $\Delta V_{DD,glo}$ und der lokale IR-Drop des empfangenden Taktpfades um $\Delta V_{DD,lok}$. Somit verringert sich in diesem Taktzyklus die Laufzeit des Taktpfades zum E-FF im Vergleich zur Laufzeit des S-FF im vorherigen Taktzyklus (Gatterlaufzeit $t_{Bi} < t_{Ai}$), und die effektive Taktperiode wird kleiner, wie in Bild 4.22 zu sehen ist. Gleichzeitig muss jedoch berücksichtigt werden, dass die Laufzeit der Logik durch den Anstieg der Versorgungsspannung nach Gleichung 4.29 sinkt.
Wie in Bild 4.22 zu erkennen ist, ist das Verhältnis der Laufzeiten von Logik- und Taktpfaden entscheidend für die Bestimmung des effektiven, mittleren IR-Drops und somit für die Höhe des IR-Drop induzierten Clock Jitters. Die Untersuchungen der verschiedenen ARM Mikroprozessoren zeigen, dass das Verhältnis zwischen Laufzeit im Taktbaum und Taktperiode weiter ansteigt, d.h. der zeitliche Unterschied zwischen Laufzeit im Taktbaum und Laufzeit der Logik nimmt weiter ab. Für die untersuchten Mikroprozessoren liegt das

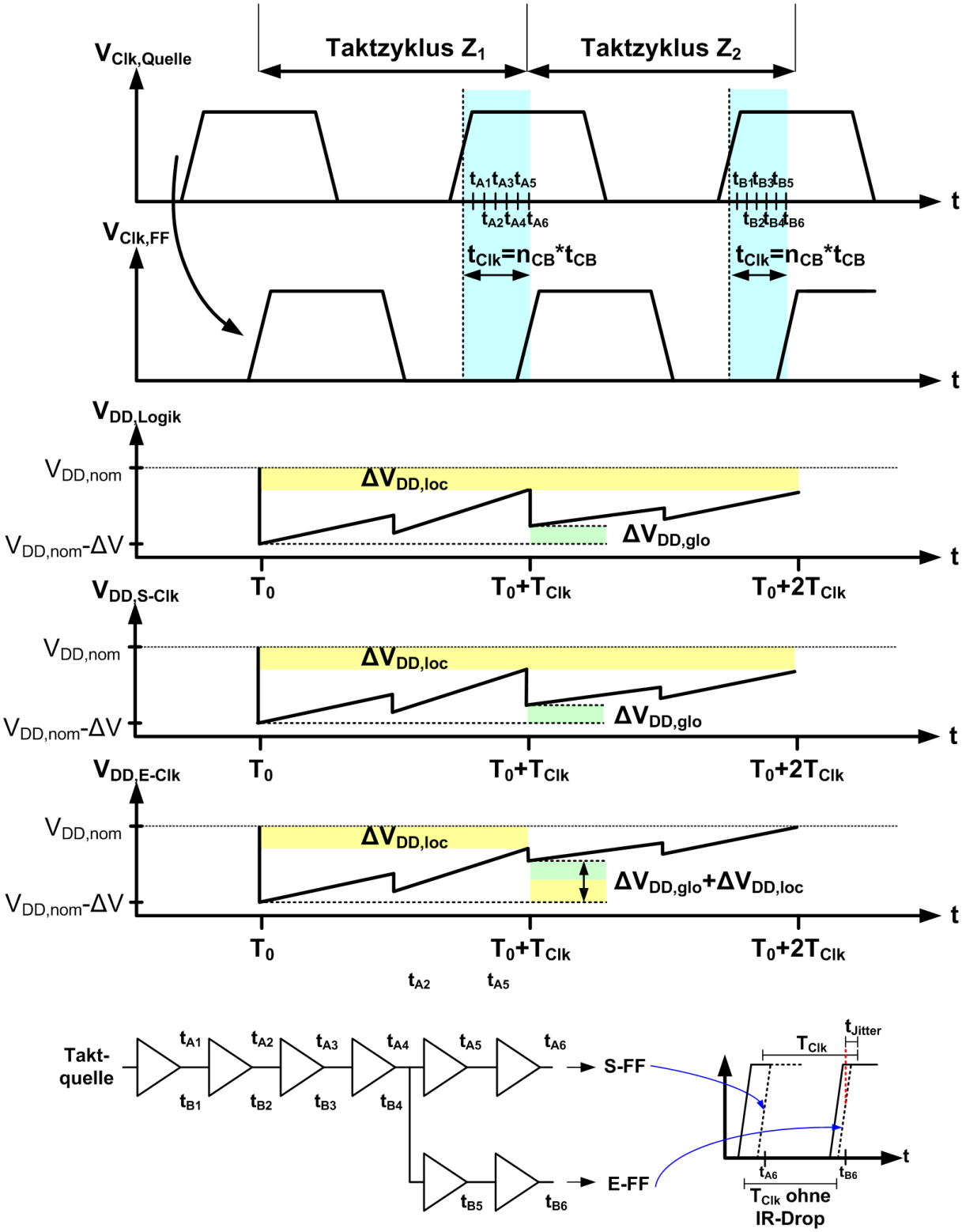


Bild 4.22: Schematische beispielhafte Darstellung der Versorgungsspannungsschwankungen in den verschiedenen Schaltungsteilen, die zur Clock Jitter Abschätzung berücksichtigt werden müssen.

Verhältnis von Laufzeit im Taktpfad und im Logikpfad zwischen 60% und 75%. Damit nähert sich der im Taktpfad gesehene effektive IR-Drop dem über die gesamte Taktperiode gemittelten Spannungseinbruch in der Logik weiter an, so dass die teilweise Kompensation des Clock Jitters durch die beschleunigte Signalpropagation im Logikpfad weiter abnimmt.

Für Hold-Zeit kritische Pfade ist nicht der von Zyklus zu Zyklus wirkende Clock Jitter entscheidend, sondern der innerhalb eines Zyklus durch IR-Drop verursachte Clock Skew. Dieser lässt sich wie folgt modellieren:

$$t_{Skew}^{IR} = (n_{CB} - SP) \cdot t_{CB} \cdot \left(\frac{(V_{DD,nom} - \Delta V_{DD,glo} - \Delta V_{DD,lok})}{(V_{DD,nom} - \Delta V_{DD,glo})} \cdot \frac{(V_{DD,nom} - \Delta V_{DD,glo} - V_{T,eff})^{\alpha_{dyn}}}{(V_{DD,nom} - \Delta V_{DD,glo} - \Delta V_{DD,lok} - V_{T,eff})^{\alpha_{dyn}}} - 1 \right) \tag{4.30}$$

Für den Clock Skew von Hold-Zeit kritischen Pfaden ist daher nur der lokale Spannungseinbruch $V_{DD,lok}$ zwischen beiden Taktpfaden entscheidend. Der globale Spannungseinbruch muss dennoch berücksichtigt werden, da sich die Sensitivität der Laufzeit gegenüber Versorgungsspannungsschwankungen je nach Betriebspunkt $V_{DD,nom} - \Delta V_{DD,glo}$ verändert.

Der Einfluss von Crosstalk auf Clock Skew und Clock Jitter wird im Modell nicht berücksichtigt, da beim Entwurf des Taktbaums generell besondere Maßnahmen ergriffen werden, um den Einfluss von Crosstalk auf die Laufzeit der Taktsignale zu minimieren. Unter anderem werden die taktsignalführenden Leitungen geschützt, indem in direkter Nachbarschaft Leitungen mit festem Potential geführt werden (V_{DD} bzw. V_{SS}), um die kapazitive Kopplung zu reduzieren. Nachteil dieser Methode sind die hohen Lasten, die zu erhöhter Leistungsaufnahme und längeren Laufzeiten des Taktsignals führen. Daher werden die zum Taktbaum benachbarten Signalleitungen in der Regel mit erhöhtem Abstand positioniert (Double Spacing/Double Track). Somit verringert sich die Lastkapazität, so dass die Laufzeit des Taktsignals bei deutlich reduziertem Crosstalkeinfluss reduziert wird. Des Weiteren werden im Taktbaum starke Treiberstufen verwendet, um das Taktsignal effizient über den gesamten Chip zu verteilen. Dies führt zu steilen Signalflanken, was den Einfluss von Signalwechseln auf Aggressorleitungen zusätzlich verringert [116, 117, 118]. Diese Maßnahmen führen im Vergleich zu IR-Drop induzierten Versorgungsspannungsschwankungen zu einem vernachlässigbar geringen Einfluss von Crosstalk auf Clock Skew und Clock Jitter.

4.2.4 Auswirkungen auf das Timing Verhalten

Die Modellierung der einzelnen Beiträge zur Laufzeitschwankung ermöglicht die Bestimmung des Einflusses von technologischen und strukturellen, schaltungsabhängigen Eigenschaften auf das Timing-Verhalten von synchronen Digitalschaltungen. Es erlaubt die Identifikation sensitiver Schaltungsbereiche und Trendaussagen hinsichtlich mikroarchitektonischer Veränderungen im Taktbaum sowie in der Logik. Die einzelnen Laufzeitbeiträge werden wie folgt kombiniert, um die maximal zu erwartende Geschwindigkeit der

Schaltung zu bestimmen:

$$T_{Clk} \geq \underbrace{Z_{Log} \cdot \left(t_{Comb} + t_{XT} + t^{IR}_{Comb} + \Delta t^{IR}_{Comb}\right) + t_{RC}}_{Logik} + \underbrace{t^{Design}_{Skew} + t^{WID}_{Skew} + t_{Jitter}}_{Taktbaum} + \underbrace{N_\sigma \cdot \sqrt{\left(Z_{Log} \cdot \sigma_{t_{Comb}}\right)^2 + \sigma^2_{t_{Skew}}}}_{Stat. Variationen\ \sigma_{t_D}} \tag{4.31}$$

Die systematischen Anteile aus Logik und Taktbaum werden addiert. Im schlimmsten Fall (worst case) führen die statistischen Variationen zu einem negativen Clock Skew und einer gleichzeitig verzögerten Logiklaufzeit. Da diese Schwankungen statistisch unabhängig erfolgen, wird die Standardabweichung für diesen worst-case über die Wurzel aus der Summe der einzelnen Varianzen bestimmt. Der Vorfaktor N_σ bestimmt die Anzahl der zu berücksichtigenden Standardabweichungen. Als Standardwert ist $N_\sigma = 3$ gesetzt.
Da sich das Spannungsverhalten von Logik und Taktbaum unterscheidet (siehe Bild 4.16), wird bei der Verwendung der FO4 Laufzeit als Zeitbasis der Vorfaktor Z_{Log} eingeführt, der für niedrigere Spannungen eine stärkere Gewichtung des Logikanteils gewährleistet. Bei nomineller Versorgungsspannung $V_{DD,nom}$ gilt $Z_{Log} = 1$.

Im Gegensatz zu Setup-Zeit kritischen Pfaden verursachen Hold-Zeit kritische Pfade unabhängig von der Taktfrequenz funktionale Fehler. Um Hold-Zeit Verletzungen zu verhindern muss unter Berücksichtigung von Variationen folgende Bedingung erfüllt werden:

$$\begin{aligned} t_{HD} \leq & Z_{Log} \cdot (t_{CQ} + t_{Log} + t_{XT}) + t_{RC} \\ & - N_\sigma \cdot \sqrt{\left(Z_{Log} \cdot \sigma_{Pfad}\right)^2 + \sigma^2_{t_{Skew}}} - t^{Design}_{Skew} - t^{WID}_{Skew} - t^{IR}_{Skew} \end{aligned} \tag{4.32}$$

Dabei bezeichnet σ_{Pfad} die Schwankungsbreite von Logik, Clock-Q Laufzeit und Hold-Zeit aufgrund von statistischen Variationen. Die Berechnung der einzelnen Komponenten muss unter Verwendung der für Hold-Zeit kritischen Pfade repräsentativen strukturellen Eigenschaften erfolgen (z.B. veränderte Lage des Splitting Points (SP) im Taktbaum, Transistorgröße und Logiktiefe im Logikpfad etc.).

4.3 Ergebnisse für die ARM Mikroprozessor-Familie

Die erhöhte Geschwindigkeitsanforderung an moderne Systems on Chip (SoC) Designs kann in sub-100nm Technologien nicht durch die Technologieskalierung allein erreicht werden. Veränderungen in der Mikroarchitektur wie z.B. tieferes Pipelining und Parallisierung ermöglichen höhere Taktfrequenzen der Mikroprozessoren bzw. einen erhöhten Datendurchsatz. Die Architekturtrends eingebetteter Mikroprozessoren zeigen primär ein tiefes Pipelining bis zu einer Tiefe von 15 Pipelinestufen. Insbesondere Mikroprozessoren mit erhöhter Pipelinestufenanzahl weisen beginnend mit SIMD-Erweiterungen eine zunehmend parallele Befehlsverarbeitung auf. Tabelle 4.4 zeigt eine Übersicht synthetisierbarer RISC Mikroprozessoren von ARM und MIPS. Die Mikroprozessortypen der beiden Marktführer im Bereich eingebetteter Mikroprozessoren zeigt eine deutliche Erhöhung der Pipelinetiefe sowie eine aufkommend erhöhte Parallelität in der Daten- bzw. Befehlsverarbeitung.

Tabelle 4.4: Pipelining und Parallelität eingebetteter Mikroprozessoren.

Prozessor	Pipelinetiefe	Parallelität
ARM7	3	-
ARM9	5	-
MIPS 4k	5	-
ARM11	8	SIMD Erweiterung
MIPS 24k	8	-
MIPS 34k	9	SIMD Erweiterung
ARM Cortex-A8	11	Dual-Issue, In-Order
MIPS 74k	15	Dual-Issue, Out-of-Order
Intel Atom Silverthorne	16	Dual-Issue, In-Order

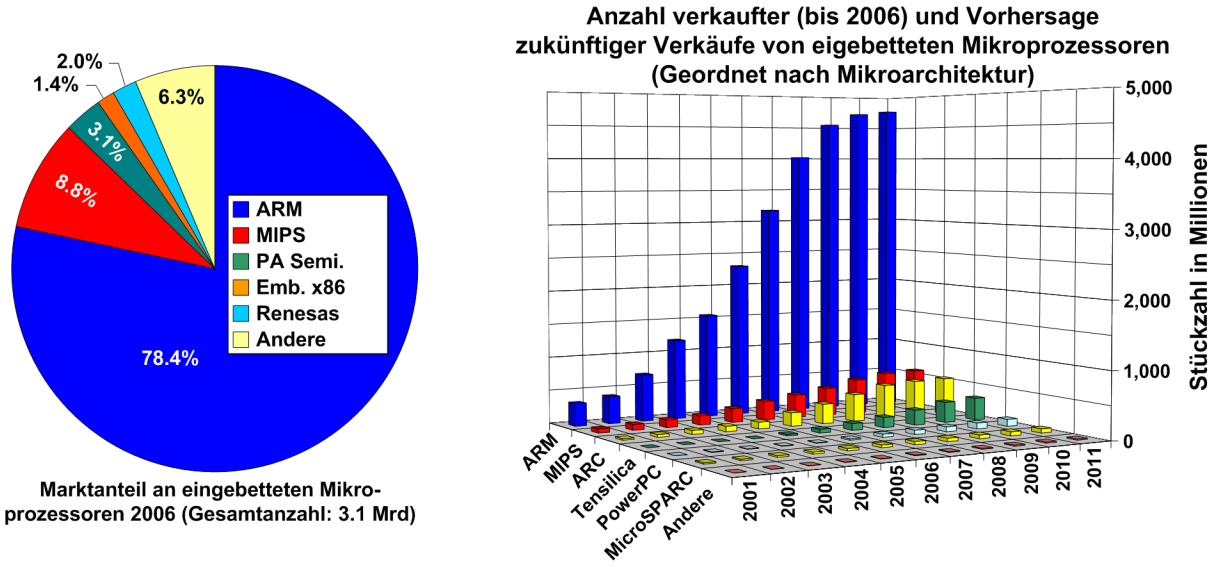
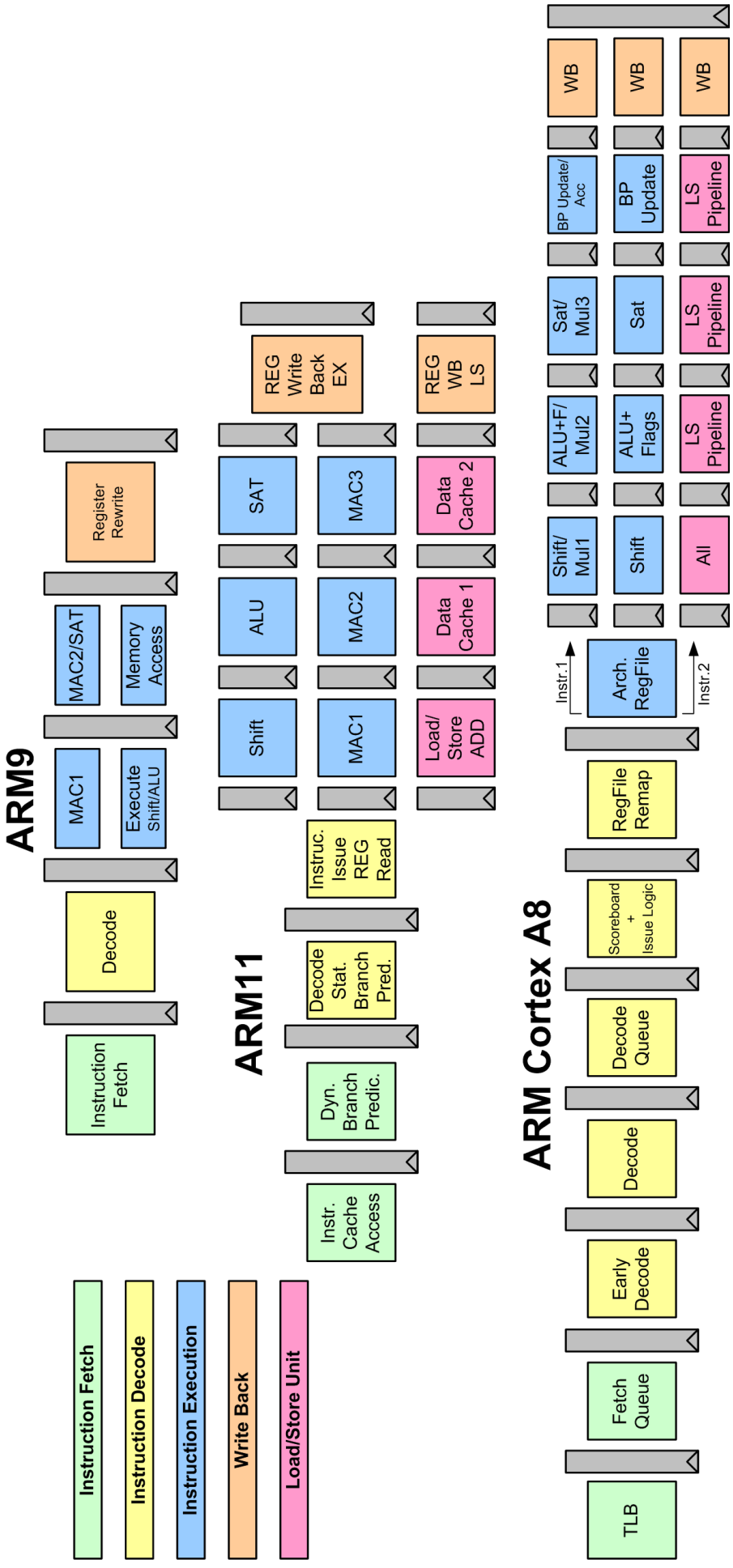


Bild 4.23: Marktdaten eingebetteter Mikroprozessoren: Hersteller und Architekturen [119].

Die Entwicklung des neuen Atom 'Silverthorne' Mikroprozessors, der den Markt für Mobile Internet Devices (MID) adressiert, zeigt, dass trotz vieler Gemeinsamkeiten in der Mikroarchitektur von High-Performance und low-power Prozessoren unterschiedliche Randbedingungen im Schaltungsentwurf vorherrschen. Der große Unterschied hinsichtlich Energieaufnahme und Fläche beeinflusst daher auch signifikant die Geschwindigkeit und Robustheit der Schaltung [1].
Bild 4.23 zeigt den Marktanteil verschiedener Hersteller von eingebetteten Mikroprozessoren sowie die verkaufte Stückzahl verschiedener eingebetteter Mikroprozessorarchitekturen seit 2001 mit Vorhersagen bis zum Jahre 2011.
Es ist deutlich zu erkennen, dass ARM mit 78.4% Marktanteil bei insgesamt 3.1 Mrd. weltweit verkauften Mikroprozessoren im Jahre 2006 der mit Abstand wichtigste Hersteller von eingebetteten Mikroprozessoren ist. Zusammen mit MIPS Technologies deckt ARM im Jahre 2006 87.2% des weltweiten Marktes ab.
In der Sparte Mobilfunk ist ARM in 75% aller Mobiltelefone weltweit vertreten. Der ARM9 Prozessor, der die klassische 5-stufige Pipeline eines RISC Prozessors aufweist, ist mit 249 verkauften Lizenzen der meist verkaufte Mikroprozessor von ARM [120].

Die in dieser Arbeit beschriebenen Analysen und Untersuchungen, basierend auf der ARM Mikroprozessorfamilie, sind somit sehr repräsentativ und weisen eine hohe Praxisrelevanz auf.
Im Folgenden werden die Ergebnisse des Mikroprozessormodells für ARM9, ARM11 und ARM Cortex A8 diskutiert. Die der ARM Familie ähnliche Entwicklung der MIPS Mikroprozessor Architektur zeigt, dass diese Ergebnisse stellvertretend für die allgemeinen Architekturtrends von RISC und low-power Prozessoren verwendet werden können. Bild 4.24 veranschaulicht die Änderungen der Mikroarchitektur für ARM926, ARM1176 und ARM Cortex A8.

Die in Tabelle 4.4 aufgeführten Veränderungen der Mikroarchitektur greifen stark in die Schaltungstopologie ein. Die Sensitivität der Schaltung gegenüber Variationen verändert sich und führt zu einem erhöhten Anteil der Laufzeitschwankung an der Gesamtlaufzeit.
Bild 4.25 zeigt die Ergebnisse des Mikroprozessormodells für die ARM Familie [121]. Der ARM9 erhält dabei die Prozessparameter für 90nm, der ARM11 die für 65nm und der ARM Cortex A8 für 40nm low-power CMOS Technologie. Dabei werden alle Prozessoren für nominelle Versorgungsspannungen $V_{DD,nom}$ untersucht.
Die fortschreitende Verkürzung der Logik und der Bedarf an einer höheren Anzahl an Pipelineregistern führt zu einem sinkenden Laufzeitbeitrag der kombinatorischen Logik zur Gesamtlaufzeit. Die Effizienz von tieferem Pipelining zur Erhöhung der Geschwindigkeit sinkt signifikant. Während bei Verwendung von Standard Master-Slave Flip Flops der absolute Beitrag der Pipelineregister unabhängig von der Architektur konstant bleibt, nimmt der absolute Beitrag variationsbedingter Timing Unsicherheit zu. Dies führt im Falle des ARM Cortex A8 zu einem effektiven Laufzeitanteil der Logik von nur noch ca. 56%. Der steigende Anteil der Pipelineregister sowie der nahezu gleich große Anteil der Timing Unsicherheit wächst auf 44% an. Der Anteil der Timing Unsicherheit, die sich aus statistischen Prozessvariationen, systematischen WID Prozessvariationen, on-chip IR-Drop, Crosstalkeffekten in der Logik sowie Clock Skew und Clock Jitter zusammensetzt, wächst von 13% für den ARM9 in 90nm auf über 23% für den ARM Cortex A8 in 40nm. Grund dafür sind sowohl strukturelle als auch technologische Sensitivitäten, die sich gegenseitig verstärken.
In Bild 4.26 sind für alle untersuchten Mikroprozessoren die Anteile an der WID Laufzeitschwankung geordnet nach der entsprechenden Variationsquelle dargestellt. Der deutlich ansteigende Beitrag der Clock Uncertainty zeigt den steigenden Einfluss des Taktbaums auf die Taktperiode einer Schaltung. Die mit tieferem Pipelining einhergehende Verkürzung der Logik lässt das Verhältnis zwischen den Propagationszeiten von Taktpfad und Logikpfad ansteigen. Selbst absolut gleichbleibend große Variationen im Taktbaum führen somit in relativem Maßstab zu steigenden Schwankungen. Der Schwankungsbeitrag statistischer Variationen beinhaltet auch den Anteil des Taktbaums, d.h. den statistisch schwankenden Clock Skew Anteil. Für den ARM11 Mikroprozessor liegt der Anteil der statistischen Laufzeitschwankung im Taktbaum bereits über dem Anteil der Logik. Im Vergleich zu den Laufzeitbeiträgen durch lokal schwankende Betriebsbedingungen ist der Anteil der WID Prozessschwankung (statistisch & systematisch) mit maximal 25% für den ARM Cortex A8 gering. IR-Drop induzierte Schwankung im Taktbaum und in der Logik trägt zu über 50% zur Gesamtschwankung bei und ist somit die mit Abstand einflussreichste Variationsquelle. Da während der Taktbaum-Synthese effektive Maßnahmen zur Vermeidung von Crosstalkeffekten ergriffen werden (z.B. Double-Spacing, Shielding usw.) ist in Bild 4.26 nur der Crosstalkbeitrag der Logik berücksichtigt. Mit sinkenden



86 Bild 4.24: Struktur der Integer-Pipeline aller untersuchten ARM Mikroprozessoren.

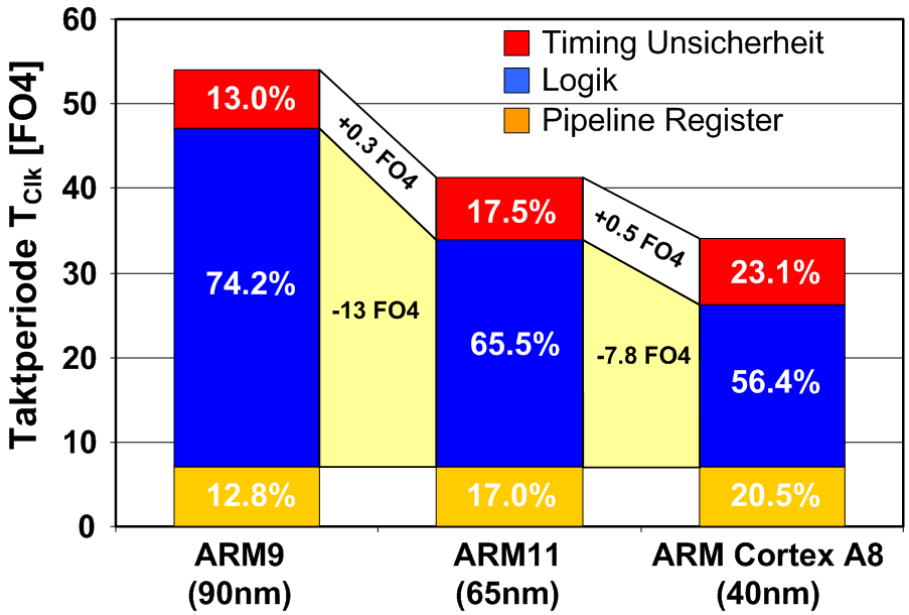


Bild 4.25: Laufzeitbeiträge von Pipelineregistern, Logik und Timing Unsicherheit für ARM9, ARM11 und ARM Cortex A8.

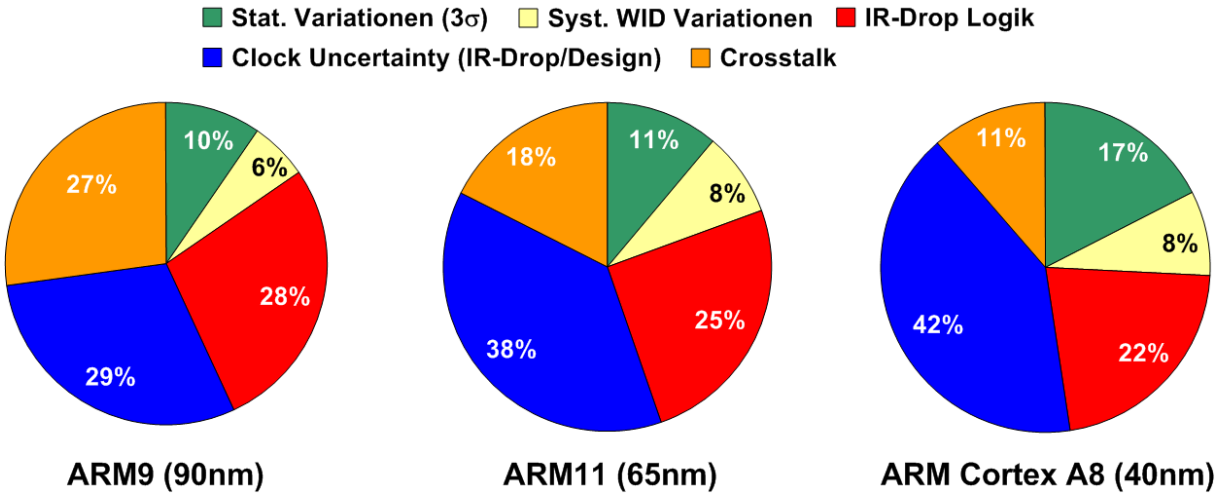


Bild 4.26: Beiträge zur Timing Unsicherheit.

Laufzeiten der Logikpfade nimmt auch der Einfluss von Crosstalk auf die maximale Taktperiode ab.

Betrachtet man den Betrieb in niedrigeren Spannungsbereichen ($V_{DD,nom} - 300mV$) wie sie z.B. durch Dynamic Voltage Scaling (DVS) in low-power Produkten auftreten, so ist Folgendes festzustellen. Bei niedrigem V_{DD} nimmt zum einen die Sensitivität gegenüber Versorgungsspannungsschwankungen deutlich zu, zum anderen aber verringert sich die absolute Auslenkung ΔV_{DD} aufgrund des verringerten Stromflusses. Dieses entgegengesetzte Verhalten dämpft den Einfluss von betriebsbedingten Variationen in niedrigeren Spannungsbereichen. Im Gegensatz dazu nimmt der von Prozessschwankungen verursachte Beitrag für sinkende Versorgungsspannungen zu. Die Abweichungen vom nominellen Prozess bleiben konstant, während die Laufzeitsensitivität gegenüber Prozessschwankungen ansteigt. Dies führt zu einem steigenden Anteil prozessbedingter Laufzeitschwankungen an der gesamten Laufzeitschwankung. Für den ARM Cortex A8 beträgt dieser Anteil für $V_{DD} = V_{DD,nom} - 300mV$ ca. 40%. Bei Spannungsabsenkung führt dies jedoch nicht zu essentiell zeitkritischen Problemen, da die Versorgungsspannung und somit auch die Geschwindigkeit wieder erhöht werden kann, sofern die variationsbedingte Laufzeiterhöhung detektiert wird. Wird die Schaltung während des Betriebs nicht durch Monitor-Schaltungen überwacht, so müssen die veränderten Sensitivitäten gegenüber Prozess- und Umgebungsvariationen jedoch bei der Festlegung der verschiedenen, betriebsmodusabhängigen Versorgungsspannungen berücksichtigt werden.

Um zwischen den beiden Einflussgrößen Technologie und Schaltungsstruktur unterscheiden zu können werden alle Prozessoren in 65nm untersucht.
Bild 4.27 zeigt die Laufzeitbeiträge von ARM9, ARM11 und ARM Cortex A8 für 65nm low-power CMOS Technologie-Parameter [122]. Während sich der Beitrag der Timing Unsicherheit für den untersuchten ARM9 Prozessor beim Technologiewechsel von 90nm auf 65nm von 12.8% um 1.5% auf 14.3% nur geringfügig erhöht, hat ein Architekturwechsel von ARM9 auf ARM11 eine Erhöhung des Timing Unsicherheitsbeitrags von 14.3% auf 17.5% zur Folge. Es zeigt sich, dass hier strukturelle Aspekte einen größeren Einfluss haben als technologische Sensitivitäten. Je sensitiver die Schaltungsstruktur, desto stärker ist auch die Wirkung prozessspezifischer Schwankungen.
Für den ARM Cortex A8 bedeutet die Wahl der 65nm Technologie im Vergleich zu 40nm Technologie einen um 3.0% reduzierten Timing Unsicherheitsbeitrag. Der Architekturwechsel von Cortex A8 zu ARM11 bewirkt eine Verringerung von 2.6%. Die Anteile der Laufzeitschwankungsbeiträge bleiben weitgehend unverändert, d.h. die Schwankungsbeiträge werden vorwiegend von strukturellen Aspekten bestimmt. Der größte Unterschied zeigt sich beim ARM Cortex A8. Hier verringert sich der Anteil der statistischen Variationen auf 12%. Statistische Variationen, deren Einfluss sowohl mit steigenden statistischen Schwankungen aus der Technologie als auch sinkenden Logiktiefen zunimmt, zeigen hier den größten Einfluss auf die Verteilung der Schwankungsanteile.
Auch hier wird die verstärkende Interaktion von strukturellen und technologiespezifischen Sensitivitäten deutlich. Die Ergebnisse des Mikroprozessormodells zeigen, dass die fortschreitende Technologieskalierung zusammen mit mikroarchitektonischen Veränderungen zu deutlich erhöhten Beiträgen von Timing Unsicherheiten führt. Bei der Diskussion von Prozess- und Umgebungsvariationen und deren Einfluss auf die Geschwindigkeit von digitalen Schaltungen ist es nicht ausreichend nur technologiebasierte Laufzeitschwankungen zu untersuchen. Wie die Ergebnisse des Mikroprozessormodells deutlich zeigen, müssen

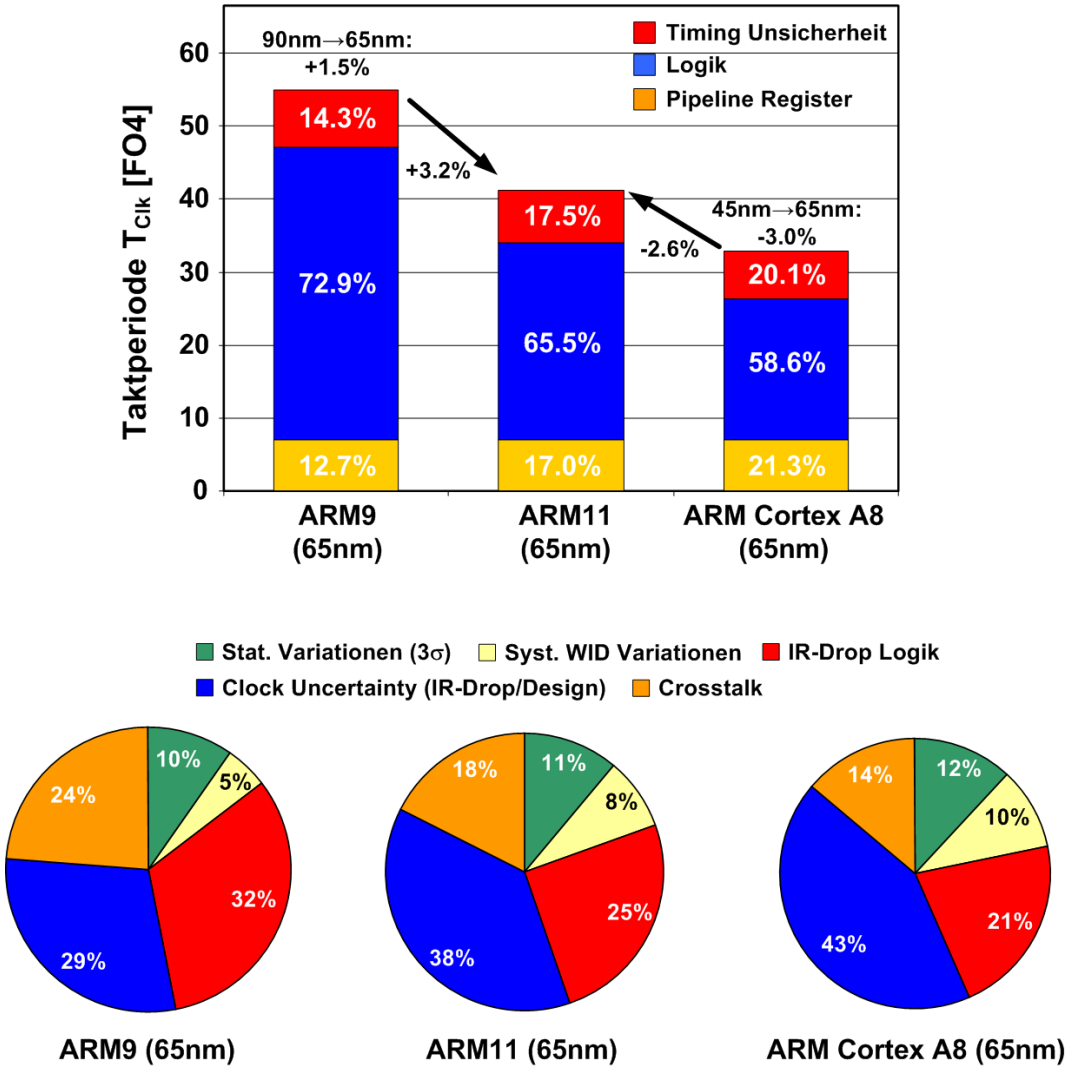


Bild 4.27: Ergebnisse des Mikroprozessormodells für ARM9, ARM11 und ARM Cortex A8 in 65nm low-power CMOS.

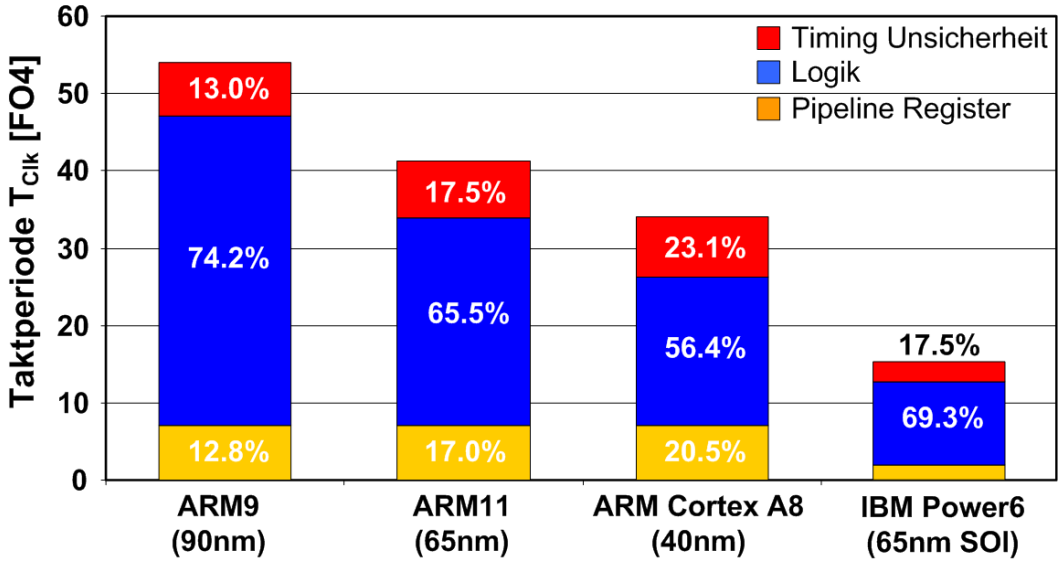


Bild 4.28: Semi-Custom low-power und Full-Custom high-speed Prozessoren im Vergleich.

neben Technologieaspekten auch die schaltungsspezifischen strukturellen Eigenschaften einer Schaltung für die Bewertung des Einflusses von Variationen auf die Geschwindigkeit einer Schaltung herangezogen werden. Zur Bestimmung von Trendaussagen kann daher nur die Kombination beider Einflussgrößen dienen. Dies zeigt auch Bild 4.28, das um den Full-Custom High-Speed Prozessor Power6 von IBM ergänzte Bild 4.25.
Der Power6 Prozessor [100, 123, 124] der mit seiner 15-stufigen Integer-Pipeline eine Mindestfrequenz von 4GHz erreicht, zeigt, dass selbst für sehr tiefe Pipelines der Anteil variationsbedingter Laufzeitschwankungen durch schaltungstechnische Maßnahmen reduziert werden kann. Selbst bei 8 bis 10-fach erhöhter Taktfrequenz des Power6 in 65nm SOI gegenüber dem untersuchten ARM11 in 65nm Bulk-CMOS, liegt der Beitrag der Timing Unsicherheit des Power6 mit 17.5% auf dem gleichen Niveau des langsameren ARM11.
Dies wird vor allem durch ein mit hohem Aufwand balanciertes Taktverteilungsnetz, bestehend aus Taktbaum und Takt-Grid, mit mehreren Laufzeitadaptionsstufen erreicht. Die zum low-power Design deutlich unterschiedlichen Randbedingungen ermöglichen den Einsatz solch kostenintensiver Designtechniken. Auch dieses Beispiel zeigt den wesentlich, teilweise sogar dominanten Einfluss der Schaltungsstruktur bzw. Mikroarchitektur auf den Einfluss von Variationen auf die Geschwindigkeit getakteter Schaltungen.

Die Eichung des Mikroprozessormodells bzw. die Bestimmung des Fehlers ist aufgrund fehlender Informationen über die Verteilung der maximalen Taktfrequenz nach Produktion der Schaltungen sehr schwierig, da kein Vergleich zwischen Designpunkt und realer Taktfrequenz gezogen werden kann. Im Vergleich zu der in [89] veröffentlichten maximalen Taktfrequenz von 342MHz (langsamer Die, eines Split-Loses), liefert das Mikroprozessormodell eine maximale Taktfrequenz von 378MHz und somit eine um ca. 10% erhöhte Geschwindigkeit. Es ist jedoch zu beachten, dass die Messungen in [89] auf der Aktivierung des im Sign-Off bestimmten kritischsten Pfades basiert. Im Gegensatz dazu repräsentiert das Mikroprozessormodell die Eigenschaften einer Vielzahl von kritischen Pfaden, so dass die Wahrscheinlichkeit groß ist dass auch die Eigenschaften des tatsächlichen kritischen Pfades repräsentiert werden (siehe „false-path“-Problematik in Kapitel 5). Ob der während des Sign-Offs identifizierte kritischste Pfad im normalen Betrieb der Schaltung sensibili-

siert wird und somit die Geschwindigkeit der Schaltung limitiert, ist fraglich.
High-Speed Prozessoren werden nach der Produktion je nach maximal zu erzielender Frequenz in Frequenz-Cluster einsortiert und unter verschiedenen Spezifikationen individuell verkauft. Aufgrund der geringeren Anzahl an Test-Cases eines General Purpose Prozessors im Vergleich zu modernen System-on-Chips ist die Bestimmung der maximal möglichen Taktfrequenz einfacher, erfordert jedoch zeitintensive Tests. Die Minimalfrequenz des Power6 Prozessors wird nach Angaben von [100] auf 4GHz festgesetzt. Die vorhergesagte worst-case Frequenz des Mikroprozessormodells beträgt 3.96GHz.
Aufgrund der o.g. Probleme und fehlender Detailinformationen bleibt ein Vergleich zwischen Mikroprozessormodell und publizierten, gemessenen Taktfrequenzen schwierig.
Auch erste strukturelle Untersuchungen eines ARM11 Mikroprozessor Designs zeigen zahlreiche Übereinstimmungen mit den strukturellen Vorhersagen des Mikroprozessormodells (siehe z.B. α_{FF}, Abschnitt 4.2.1), so dass sich das Modell für qualitative und quantitative Trendaussagen eignet.
Neben prädiktiven qualitativen Trendaussagen ermöglicht das Modell bei Eingabe der exakten Basisparameter (siehe Kapitel 4.2) die Bestimmung der schaltungsspezifischen, variationsbedingten Laufzeitbeiträge. Somit ist es möglich, die worst-case Laufzeit für ein nominelles Design in Abhängigkeit der strukturellen Schaltungseigenschaften abzuschätzen. Es ist klar, dass insbesondere bei der Abstraktion über viele technische Ebenen hinweg (Transistor-, Gatter-, Pfad- und Architekturebene) genaue Modell-Eingangsgrößen erforderlich sind, um eine gute Genauigkeit der Ausgangsgröße (Taktperiode) zu erhalten.
Zusätzlich liefert das Modell auch strukturelle Aussagen hinsichtlich sensitiver Bereiche einer Schaltung. Die Ergebnisse des Mikroprozessormodells zeigen deutlich, dass der Einfluss des Taktverteilungsnetzes signifikant an Bedeutung gewinnt. Zusammen mit der Timing Unsicherheit tragen die Flip Flop Zellen bis zu 45% (Cortex A8) an der Gesamtlaufzeit bei. Um den Geschwindigkeitsgewinn durch tieferes Pipelining aufrecht erhalten zu können sind daher schaltungstechnische Maßnahmen wie z.B. der Einsatz von gepulsten Flip Flops zu ergreifen. In Kapitel 6.2.4 werden daher auf Basis des Mikroprozessormodells Kosten und Nutzen des Einsatzes von gepulsten Latches und Flip Flops näher untersucht.

4.4 Bemerkungen zu den Ergebnissen

Die Ergebnisse in diesem Kapitel zeigen das Verhalten von eingebetteten RISC Mikroprozessoren unter Einfluss von WID Prozess- und on-chip Umgebungsvariationen. Obwohl das Modell für qualitative und quantitative Trendaussagen geeignet ist, dürfen die Ergebnisse der Analysen nicht allgemeingültig für die Bewertung des Einflusses von Variationen auf die Geschwindigkeit von digitalen Schaltungen gesehen werden.
Der Einfluss von Variationen hängt neben technologischen und strukturellen Einflussgrößen von weiteren Faktoren wie z.B. dem Betriebsbereich der Schaltung (vgl. z.B. Sub-Threshold Logik), der Komplexität der Schaltung (vgl. z.B. Mikroprozessor↔Addierer), der Homogenität der Schaltung etc. ab. Diese Zusammenhänge werden im Folgenden näher diskutiert:

- **Betriebsbereich der Versorgungsspannung**

  Während die Bedeutung von Prozessvariationen, die statisch wirken, bei niedrigen Versorgungsspannungen zunimmt, kann die Bedeutung von betriebsbedingten Variationen aufgrund von reduzierten Schwankungsbreiten (z.B. IR-Drop) abnehmen.

Für Schaltungen, die bei nomineller Spannung der jeweiligen Technologie betrieben werden, sind vorwiegend Schwankungen der Betriebsparameter geschwindigkeitskritisch. Die in diesem Kapitel durchgeführten Untersuchungen zeigen, dass für alle untersuchten Mikroprozessoren statistische Variationen eine untergeordnete Rolle spielen und keinen signifikanten Einfluss auf die maximale Taktfrequenz haben. Für extrem tiefes Pipelining, d.h. sehr kurze Pfade, wie z.B. von High-Performance Arithmetik Einheiten nimmt die Bedeutung dieser Variationen ebenso wie für Schaltungen bei extrem niedriger Versorgungsspannung (z.B. Sub-Threshold Logik) zu.

- **Schaltungskomplexität**

  Vergleicht man z.B. einen Time-to-Digital Converter (TDC) mit einer hochkomplexen Schaltung wie z.B. einem ARM926 Mikroprozessor, so zeigt sich eine signifikant unterschiedliche Bedeutung einzelner Variationen für die Kenngrößen der Schaltungen. Während statistische Variationen und Crosstalkeffekte im TDC die zeitliche Auflösung drastisch verändern können, spielen derartige Variationen in den komplexen geschwindigkeitskritischen Pfaden eine untergeordnete Rolle und können als geschwindigkeitsunkritisch bezeichnet werden. Im Vergleich zum Mikroprozessor ist für einen TDC die Bewertung des Einflusses von Variationen auf die Schaltungskenngrößen einfach. Die sehr homogene Struktur des TDC, bestehend aus der wiederholten Anordnung von Verzögerungselementen und Speichergliedern (Flip Flop), erlaubt eine einfache Basisuntersuchung der Hauptelemente, was in den komplexen, unterschiedlichen und stark verzweigten Strukturen der geschwindigkeitskritischen Pfade nicht möglich ist. Allgemein gilt, je geringer die Signalanzahl bzw. Signalkombinationen und je höher die Regularität bzw. Homogenität einer Schaltung, desto einfacher ist die Bewertung des Einflusses von Variationen auf die Kenngrößen einer Schaltung. Deshalb können Bewertungen, wie sie in sehr homogenen Strukturen wie z.B. TDC oder SRAM Blöcken angewandt werden, nur schwer bzw. gar nicht auf komplexere Systeme übertragen werden.

- **Schaltungsimplementierung**

  Zur Bewertung des Einflusses von Variationen sind neben fundamentalen strukturellen und topologischen Eigenschaften (z.B. Mikroarchitektur) auch implementierungsbedingte Einflussgrößen zu berücksichtigen. Dazu zählen unter anderem die Belastung und Größe der gewählten Standardzellen, die Implementierung von Taktverteilungs- und Versorgungsspannungsnetz, diverse Designkriterien und Optimierungsgrößen, Multi-V_T Design etc.. Aus diesem Grund wurden für die Analysen in dieser Arbeit Produktdesigns repräsentativer Mikroprozessoren verwendet, die mittels modernem industriellem Designflow unter Berücksichtigung aller Design- und Optimierungskriterien von low-power Schaltungen implementiert wurden.

Für die untersuchten Mikroprozessoren und alle strukturähnlichen Digitalschaltungen können folgende Erkenntnisse zum Einfluss von WID Prozess- und on-chip Umgebungsvariationen zusammengefasst werden:

Die Ergebnisse des Mikroprozessormodells zeigen, dass bei nomineller Versorgungsspannung, d.h. im geschwindigkeitsrelevanten Bereich der Technologie, IR-Drop mit Abstand die schwerwiegendste Variation darstellt. Da sowohl der Taktbaum als auch der Logikteil von IR-Drop betroffen sind, stellen erhöhte Logiklaufzeiten und Clock Jitter die größten Laufzeitvariationen dar. Crosstalkeffekte tragen selbst im worst-case nur geringfügig zur

gesamten Laufzeitvariation bei. Die Analysen des Crosstalkbeitrags in einem ARM926 und ARM1176 Design ergeben sehr geringe Wahrscheinlichkeiten für den worst-case Laufzeitbeitrag durch Crosstalk. Im ARM926 müssen pro Netz durchschnittlich acht Aggressoren, im ARM11 durchschnittlich ca. 40 Aggressoren pro Netz in die entgegengesetzte Richtung des Victim-Netzes schalten, um den worst-case Einfluss auf die Laufzeit zu erzielen. Basierend auf diesen Untersuchungen können Crosstalkeffekte in der kombinatorischen Logik komplexer getakteter Digitalschaltungen im Allgemeinen als geschwindigkeitsunkritisch angesehen werden, da die Wahrscheinlichkeit für das mehrheitlich entgegengesetzte Schalten der Aggressoren verschwindend gering ist.
Statistische Laufzeitvariationen spielen hier ein untergeordnete Rolle und tragen im Vergleich zu Umgebungsvariationen geringfügig zur gesamten Laufzeitschwankung bei. Selbst bei weiter steigenden statistischen Schwankungen ist ein signifikanter Einfluss dieser Variationen aufgrund starker Mittelungseffekte nicht zu erwarten. Systematische WID Prozessvariationen haben ebenfalls geringfügigen Einfluss auf die Laufzeitschwankung. Da detaillierte Untersuchungen von 45nm und 90nm CMOS Technologien in [24, 125] zeigen, dass die systematischen Laufzeitvariationen zurückgehen, ist mit einer Zunahme des Anteils systematischer WID Prozessvariationen zur gesamten Laufzeitschwankung nicht zu rechnen. WID Temperatureffekte spielen aufgrund der geringen Temperaturgradienten ebenfalls eine untergeordnete Rolle. Die globale Temperatur hingegen ist beim Schaltungsentwurf eine wichtige Größe und gewinnt, wie in Kapitel 2 gezeigt wird, hinsichtlich Laufzeitvariationen in künftigen CMOS Technologien weiter an Bedeutung.

5 Topologieanalysen und Robustheit

Neben technologieabhängigen Sensitivitäten spielen die Strukturen sowie die Implementierung einer Schaltung hinsichtlich ihrer Sensitivität gegenüber Prozess- und Umgebungsvariationen eine entscheidende Rolle. In synchronen Digitalschaltungen muss jeder einzelne Pfad, egal ob Setup-Zeit oder Hold-Zeit kritisch, die jeweiligen Timingbedingungen bezüglich der Ankunftszeit von Daten- und Taktsignal am empfangenden Flip Flop erfüllen, um die Funktionalität der Schaltung zu gewährleisten. Die Einhaltung dieser Bedingungen wird im Schaltungsentwurf mittels deterministischer Statischer Timing Analyse (STA/D-STA) überprüft, die das Timing aller kritischen Pfade berechnet.
Berücksichtigt man jedoch mögliche Variationen während der Herstellung und des Betriebs der Schaltung, so funktioniert diese Methodik nur für globale Effekte mit ausreichender Genauigkeit. Hierfür werden globale Prozess-Corner verwendet, die dem 3-Sigma Punkt einer globalen Monte-Carlo Analyse entsprechen. Für WID Prozess- und on-chip Umgebungsvariationen werden Sicherheitsmargen verwendet (OCV Methodik), die in einer sehr pessimistischen Berechnung enden und somit die Kosten (Fläche, Leistungsaufnahme) für die Gewährleistung der spezifizierten Taktfrequenz erhöhen [126, 15].
Um den Pessimismus der herkömmlichen Methodik zu reduzieren, werden zunehmend statistische Methoden für die Timing Analyse entwickelt. Die in der Literatur und bei den EDA Herstellern am weitesten verbreiteten Ansätze sind die Statistische Statische Timing Analyse (SSTA) sowie die Advanced OCV Methodik (AOCV). Beide Methodiken berücksichtigen schaltungsstrukturabhängige Mittelungseffekte und modellieren im Vergleich zur herkömmlichen Methodik statistische Variationen genauer.
Im Zusammenhang mit SSTA wird oft die Bestimmung der Pfadlaufzeit-Verteilung einer Schaltung als Vorteil der SSTA gegenüber der deterministischen STA (D-STA) diskutiert. Bei der Bestimmung der Ausbeute im Allgemeinen und für die SSTA im Speziellen existieren jedoch unterschiedlichste Hürden, die bisher nicht überwunden werden konnten. Ein grundsätzliches Problem, das die herkömmliche deterministische STA und die statistische STA gemeinsam haben, sind sogenannte 'falsche Pfade' (false paths) [127, 128]. Diese Pfade stellen bei isolierter Betrachtung des Pfades einen möglichen Propagationsweg für ein Signal dar. In Realität verhindert die den Pfad umgebende Beschaltung durch die Belegung der jeweiligen Eingangs-Pins der Einzelgatter die Propagation des Signals und somit die Sensibilisierung des Pfades. Dabei gibt es sowohl statisch als auch dynamisch falsche Pfade. Das Auffinden dynamischer Pfade erfordert eine genaue Berechnung aller Signalpropagationen. Da dazu wiederum statische und dynamische falsche Pfade ausgeschlossen werden müssen, endet die Problemstellung in der eigentlichen Ausgangslage (Henne-Ei Problem).
Insbesondere in komplexen Schaltungen ist die Identifikation von falschen Pfaden schwierig, und die Existenz nicht identifizierter falscher Pfade kann nicht ausgeschlossen werden. Aus diesem Grund kann nicht geklärt werden, inwiefern die Ergebnisse aus STA bzw. SSTA mit der realen Schaltung korrelieren. Eine Bestimmung der Korrelation würde die Sensibilisierung und Messung jedes einzelnen Pfades erfordern, was bei großer Pin-Anzahl zu immensen Messzeiten führt und daher nicht praxistauglich ist. Daher ist die Bestimmung

der realen Pfadlaufzeit-Verteilung einer Schaltung und ein Vergleich zwischen vorhergesagten Laufzeiten durch D-STA bzw. SSTA und realen Laufzeiten in der Praxis nahezu unmöglich.
Während die Corner-basierte STA gute Ergebnisse für globale Prozessschwankungen liefert, werden statistische Laufzeitschwankungen bei der Analyse mit SSTA und AOCV exakter auf die Pfadlaufzeiten abgebildet. Systematische Variationen wie z.B. layoutabhängige WID Variationen werden in der SSTA als statistisch schwankende Variationen modelliert. Die Crosstalk Analyse wird im Vergleich zur D-STA kompliziert, da die zur Bestimmung des Crosstalkbeitrags erforderlichen Timing-Fenster nicht mehr existieren. Ferner können pseudo-statistische Schwankungen der Versorgungsspannungen bisher nicht modelliert werden. Um die Verteilung der Pfadlaufzeit zu bestimmen müssen alle topologischen Korrelationen berücksichtigt werden (siehe 5.2), was nur mit der im Vergleich zur pfadbasierten SSTA (PB-SSTA) ungenaueren blockbasierten SSTA (BB-SSTA) möglich ist [129, 130]. Zur genauen Analyse der Pfadlaufzeiten wird jedoch die PB-SSTA benötigt, die topologische Korrelationen nicht berücksichtigen kann.
Die Advanced OCV Methodik bedient sich der Eigenschaft, dass die relative Laufzeitschwankung eines Pfades mit der Anzahl seiner Logikstufen abnimmt. Diese Mittelungseffekte werden berücksichtigt und ein für jeden einzelnen Pfad berechneter effektiver Derating-Faktor verwendet. Die pauschale Erhöhung der Pfadlaufzeit um einen sogenannten Derating-Faktor und der damit verbundene, vorwiegende Pessimismus werden reduziert [131].
Damit verbessert sich im Vergleich zur herkömmlichen D-STA die Berücksichtigung statistischer Prozessschwankungen in SSTA und AOCV. Da, wie in Kapitel 4.3 gezeigt wird, statistische Variationen nur einen geringen Anteil an der Laufzeitschwankung von eingebetteten Mikroprozessoren haben, adressieren AOCV und SSTA die bessere Modellierung eines nur geringen Anteils der gesamten WID Laufzeitschwankung. Der Einfluss großer betriebsbedingter Variationsbeiträge von IR-Drop und Clock Jitter sowie Crosstalk-Effekten ist nach wie vor nur durch corner-Methodik und der Implementierung von zeitlichen Sicherheitsmargen möglich. Somit bleiben für SSTA und D-STA viele Hürden zur Bestimmung der realen Laufzeitverteilung unter Berücksichtigung aller Variationen bestehen [129, 132, 133].
Im Gegensatz zu diesen statistischen Modellierungsansätzen werden in diesem Kapitel schaltungstechnische Bewertungskenngrößen vorgeschlagen, um die Verwundbarkeit einer Schaltung gegenüber Variationen und Alterungseffekten analysieren zu können. Ferner erfolgt die Definition eines Schaltungssensitivitätsfaktors, der als Robustheitskriterium einer Schaltung dient und somit die Implementierung verschiedener Schaltungen hinsichtlich Robustheit vergleichbar macht.

5.1 Pfadübergreifende Topologieanalyse

Um variationsbedingte, teilweise unmodellierbare (pseudo-statistische) Laufzeiteffekte auf Schaltungsebene zu kompensieren bzw. zu verringern, werden verschiedenste Techniken eingesetzt, die auf dynamischem und statischem Time-Borrowing (D-TB, S-TB) basieren [134, 135, 136, 137, 138, 139]. Der Einsatz dieser Methodiken erfordert jedoch die genaue Kenntnis der Pfadtopologien in den kritischen Bereichen, d.h. die serielle Anordnung kritischer und unkritischer Pfade, um zum einen die Anwendbarkeit, zum anderen den resultierenden Vorteil der Maßnahme zu bestimmen.

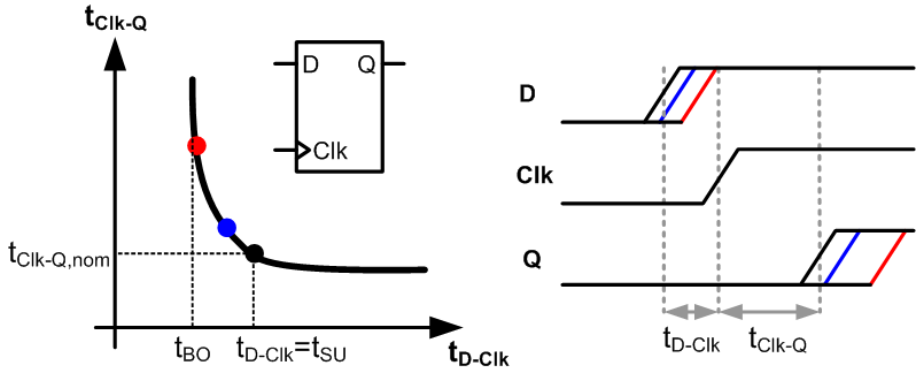


Bild 5.1: Schematische Darstellung des Schaltverhaltens von Flip Flop Zellen.

In Schaltungen mit Master-Slave Flip Flops (MS-FF) wird auch ohne explizite Techniken Time-Borrowing betrieben [140]. Im Folgenden wird daher das Schaltverhalten von taktflankengesteuerten Flip Flops näher analysiert.

In der Statischen Timing Analyse wird die Setup-Zeit Verletzung als diskretes Ereignis interpretiert, d.h. das Datensignal wird im Flip Flop gespeichert, sobald die zeitliche Differenz zur speichernden Taktflanke größer gleich der Setup-Zeit t_{SU} des empfangenden Flip Flops ist.
In der Realität findet jedoch ein Wettlauf zwischen Datensignal und Taktsignal statt, der wie folgt erklärt werden kann. Das Datensignal muss den internen Speicherknoten des Flip Flops umladen, um bei der speichernden Taktflanke, die die Verbindung zwischen Speicherknoten und Dateneingang sperrt, sicher im Flip Flop gespeichert zu werden. Erreicht das Datensignal den internen Speicherknoten zeitgleich zur speichernden Taktflanke, so ist es fraglich, ob das Datum einen Potentialwechsel im Speicherknoten erzielen kann, bevor das Taktsignal den Ladungstransport vom Dateneingang unterbricht. Aus dieser Konstellation ergibt sich somit eine Transferkennlinie in Abhängigkeit des zeitlichen Unterschieds von Daten- und Taktflanke t_{D-Clk}. Eine ins Unendliche steigende Laufzeit t_{Clk-Q} entspricht dabei einem zu spät ankommenden Datensignal, so dass das Speichern des Datums im Flip Flop nicht mehr möglich ist.

Bild 5.1 stellt die Abhängigkeit der Clock-Q Laufzeit t_{Clk-Q} vom zeitlichen Abstand von Daten- und Taktsignal t_{D-Clk} schematisch dar. Mit abnehmendem zeitlichen Abstand von Daten- und Taktflanke erhöht sich die Clock-Q Laufzeit, bis für einen bestimmten Wert von t_{D-Clk} ein Speichern des Datums nicht mehr möglich ist. Dieser Zeitpunkt t_{D-Clk} wird als Blackout-Zeit t_{BO} bezeichnet [141]. Der zeitliche Unterschied zwischen t_{SU} und t_{BO} ist typischerweise kleiner gleich einer Fanout-4 Laufzeit [140] und liegt somit in der Größenordnung statistischer Laufzeitschwankungen in modernen Mikroprozessoren.
Unabhängig davon, welche Variation eine Schwankung von t_{D-Clk} hervorruft, erhöht sich für $t_{BO} < t_{D-Clk} < t_{SU}$ die Pfadlaufzeit des folgenden Pfades, ohne dass im untersuchten Pfad ein fehlerhaftes Datum im Flip Flop gespeichert ist. Bild 5.2 veranschaulicht diesen Zusammenhang. Ist der darauf folgende Pfad (Logik 2) ebenfalls zeitlich kritisch, so kann die starke Erhöhung der Clock-Q Laufzeit (z.B. bis $2t_{Clk-Q,nom}$) zu einem deutlich reduzierten t_{D-Clk} an FF3 führen, so dass an diesem Flip Flop das Datensignal zu spät ankommt, um im Flip Flop gespeichert zu werden. Die eigentlich kritische Situation an FF2 hat sich somit auf FF3 übertragen.

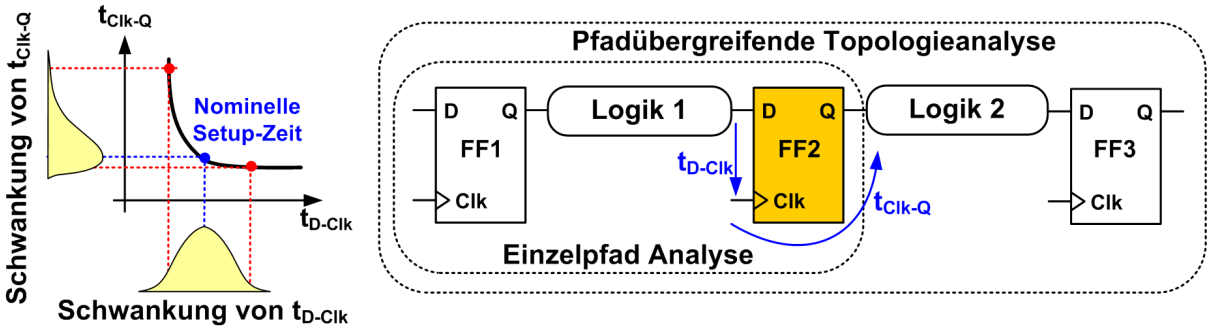


Bild 5.2: Übertragung von Laufzeitschwankungen auf nachfolgende Pipelinestufen.

Die Robustheit der Schaltung gegenüber Prozessschwankungen hängt somit auch von der Umgebung der kritischen Pfade ab. Aus diesem Grund wird eine pfadübergreifende Topologieanalyse und die Klassifizierung verschiedener Pfadtopologien hinsichtlich ihres kritischen Zustands eingeführt.

- **Der isoliert-kritische Pfad (Pfadtyp 1)**

  Der in Bild 5.3 dargestellte Pfadtyp 1 ist ein geschwindigkeitskritischer Pfad einer bestimmten Pipelinestufe. Alle Pfade der vorhergehenden Pipelinestufe, die am sendenden Flip Flop des kritischen Pfades ankommen, sind zeitlich unkritisch. Ebenso sind auch alle Pfade der folgenden Pipelinestufe, die vom empfangenden Flip Flop des untersuchten Pfades abgehen, unkritisch. Dieser Pfad liegt somit isoliert von anderen kritischen Pfaden.
  Eine variationsbedingte Laufzeiterhöhung des vorangegangenen Pfades hat keine negative Beeinträchtigung des untersuchten Pfades zur Folge, da das sendende Flip Flop noch immer im nominellen Bereich und somit ohne zusätzliche Erhöhung der Clock-Q Laufzeit t_{Clk-Q} betrieben wird.
  Der hier betrachtete Pfad bleibt demnach unbeeinflusst von Laufzeitschwankungen anderer Pfade und wird daher als unkritischste Pfadtopologie mit Pfadtyp 1 bezeichnet.

- **Der seriell-kritische Pfad (Pfadtyp 2)**

  Im Gegensatz zum isoliert-kritischen Pfad wird der seriell-kritische Pfad nach drei Unterklassen unterschieden.

  Pfadtyp 2A:

  Die in Bild 5.3 gezeigte Pfadtopologie zeigt in der 1. Pipelinestufe einen zeitlich unkritischen Pfad, gefolgt von zwei kritischen Pfaden in den folgenden Pipelinestufen 2 und 3. Dieser Pfad wird wie der isoliert-kritische Pfad von keinem anderen Pfad durch dessen Laufzeitschwankung beeinträchtigt. Im Gegensatz zum Pfadtyp 1 beeinflusst der hier betrachtete Pfadtyp 2A bei erhöhter Laufzeitschwankung jedoch das Timing des kritischen Pfades in der nächsten Pipelinestufe. Eine Laufzeiterhöhung in diesem Pfad ist daher im Vergleich zum Pfadtyp 1 kritischer.

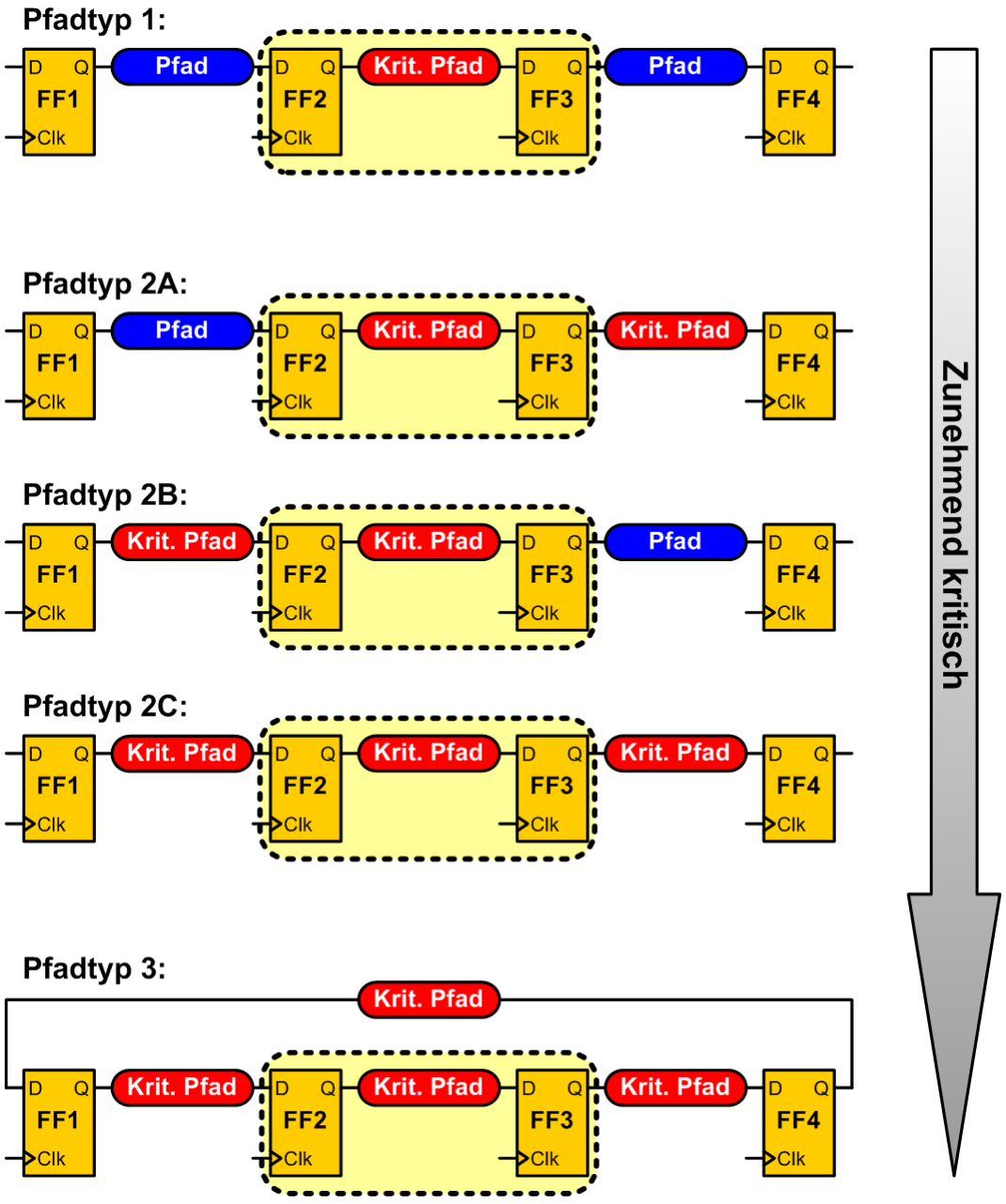


Bild 5.3: Klassifizierung von Pfadtopologien nach Lage und Umgebung von kritischen Pfaden.

Pfadtyp 2B:

In diesem Fall liegt der betrachtete kritische Pfad zwischen einem kritischen Pfad in der vorangegangenen und einem unkritischen Pfad in der nachfolgenden Pipelinestufe. In dieser Anordnung führt eine Laufzeitschwankung in der 1. Pipelinestufe mit erhöhter Wahrscheinlichkeit zu einem fehlerhaften Speichern des Datums. Entweder verursacht die Laufzeiterhöhung von Pfad 1 bereits in FF2 einen Fehler, oder die deutliche Erhöhung der Clock-Q Laufzeit von FF2 führt zum Speichern eines fehlerhaften Datums in FF3. Die korrekte Übernahme des Datums in FF3 ist somit stark abhängig von der Laufzeit der vorangegangenen Pipelinestufe.

Pfadtyp 2C:

Der dargestellte Pfadtyp zeigt den Mix aus Pfadtyp 2A und 2B. Er wird vom vorausgegangenen kritischen Pfad beeinflusst und beeinträchtigt selbst den ihm folgenden kritischen Pfad der nächsten Pipelinestufe. Dieser Pfadtyp wird als kritischste Struktur des Pfadtyps 2 klassifiziert. Eine weitere Unterscheidung wird bezüglich der Lage des Pfades in einer Kette von kritischen Pfaden vorgenommen. Liegt der Pfad z.B. in einer Kette von vier kritischen Pfaden an dritter Stelle, so wird dieser als kritischer gegenüber dem Pfad an zweiter Stelle eingestuft. Der erste Pfad entspricht der Kategorie 2A, der letzte, vierte Pfad der Kategorie 2B.

- **Der 'Loop-interne' Pfad (Pfadtyp 3)**

  Die Darstellung in Bild 5.3 zeigt eine aus kritischen Pfaden bestehende Schleife (Loop), entsprechend einem Ringschluss des Pfadtyps 2C mit einem weiteren kritischen Pfad. Im Falle von variationsbedingten Laufzeiterhöhungen beeinflusst jeder Pfad das Timing aller anderen Pfade. Es existiert in dieser Schleife kein unkritischer Pfad, der die Timing-Bedingungen entspannt, so dass ständig ein kritischer Zustand vorherrscht. Da es nicht sicher ist, jedoch davon ausgegangen werden muss, dass alle kritischen Pfade der Loop nacheinander schalten, werden diese Pfadanordnungen hinsichtlich ihres kritischen Zustands noch weiter unterschieden. Je größer die Loop, d.h. je mehr kritische Pfade in verschiedenen Pipelinestufen, desto unwahrscheinlicher ist es, dass alle kritischen Pfade der Loop nacheinander schalten. Deshalb werden kurze Loops als kritischer gegenüber langen Loops betrachtet, so dass Pfade deren sendendes und empfangendes Flip Flop identisch sind als kritischste Pfadtopologie klassifiziert werden. Schalten nicht alle Pfade der Loop, so ergibt sich automatisch eine Pfadtopologie gemäß Pfadtyp 1 und 2. Als kritischstes Beispiel kann ein Flip Flop gesehen werden, das sowohl sendendes als auch empfangendes Flip Flop des gleichen Pfades ist, wie es bei Finite State Machines (FSM) vorkommen kann.

Die Anzahl der einzelnen Pfadstrukturen in den kritischen Timing Bereichen hängt neben der Schaltungsimplementierung auch von der Mikroarchitektur der jeweiligen Schaltung ab. Die maximal mögliche Anzahl von kritischen Pfaden in Serie steigt mit der Anzahl der Pipelinestufen.
Die strukturbedingte Anzahl von 'Pipelined Loops' nimmt ebenfalls mit der Anzahl an Pipelinestufen zu [142]. Insbesondere Forwarding-Pfade und die zusätzliche Kontrolllogik für Branch Prediction, das für Mikroprozessoren mit tiefen Pipelines erforderlich ist, um einen Einbruch des Datendurchsatzes aufgrund von zusätzlichen Pipeline-Stalls zu verhindern, erhöhen die mögliche Anzahl an Pipelined Loops [143, 144].

Tabelle 5.1: Die mittels Multi-Stage STA identifizierten Pfadtypen des untersuchten ARM926 in den oberen 5% des Pfad Timings.

Isoliert-kritisch	Seriell-kritisch	Loop-intern
24.7%	55.9%	19.4%

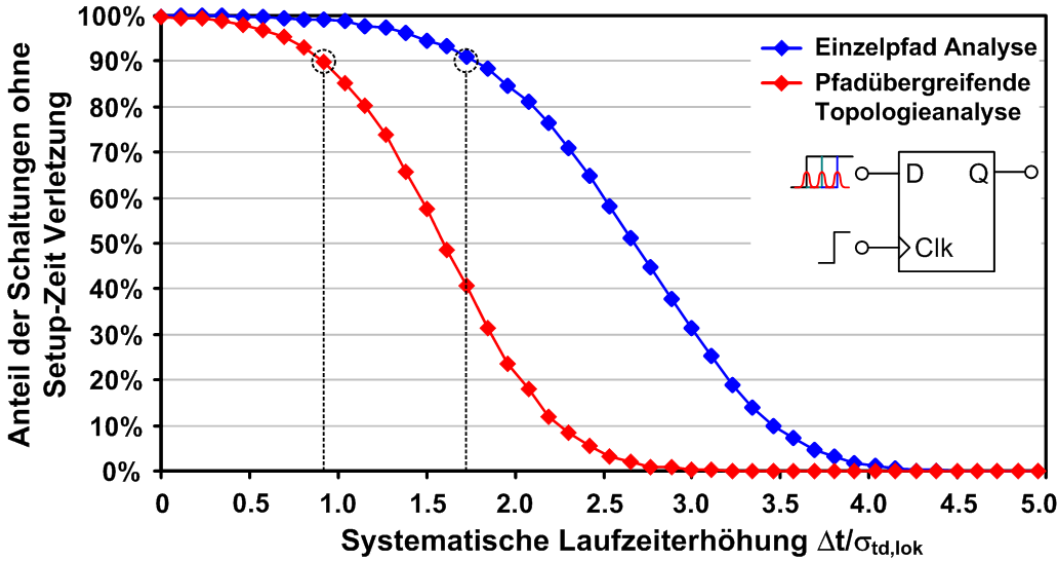


Bild 5.4: Einfluss der Pfadtopologie auf die Verteilung von Setup-Zeit Verletzungen bei Überlagerung globaler systematischer und lokaler statistischer Laufzeitvariationen.

Tabelle 5.1 zeigt für die oberen 5% des Pfad Timings des ARM926 die Zuordnung aller kritischer Pfade gemäß den oben definierten Pfadtypen. Des Weiteren führt tieferes Pipelining zu größeren Timing-Beiträgen der Flip Flop Zellen zum Gesamt-Timing, und die Bedeutung des Flip Flop Schaltverhaltens bei Laufzeitänderungen steigt.
Die Existenz von Loop-internen kritischen Pfaden limitiert den Vorteil von Time-Borrowing Techniken bzw. macht im Extremfall den Einsatz solcher Techniken unmöglich. Ist die 'geliehene' Gesamtzeit, die über die verschiedenen Pfade der 'Loop' angesammelt wird, größer als die für TB zur Verfügung stehende Zeit am Ausgangs-FF, so findet an diesem Flip Flop eine Setup-Zeit Verletzung statt, und das gespeicherte Datensignal im FF ist inkorrekt.
Aus diesem Grund wird in dieser Arbeit eine erweiterte Timing Analyse vorgeschlagen (Multi-Stage STA), die mittels selbst entwickelter perl-basierter Add-On Software durchgeführt wird, um die Anwendbarkeit und den zu erwartenden zeitlichen Gewinn von Time-Borrowing Techniken zu überprüfen. Insbesondere die Adaption des Clock Skews durch programmierbare Laufzeitelemente im Taktverteilungsnetz [139] erfordert die genaue Kenntnis der Pfadtopologien im geschwindigkeitskritischen Bereich, um künstlich erzeugte Setup-Zeit Verletzungen zu verhindern.
Unabhängig von Time-Borrowing Techniken ist die unterschiedliche Robustheit der Pfadtopologien gegenüber Variationen zu sehen.
So bestehen Unterschiede zwischen der Einzelpfadanalyse und der Analyse von Pfadtopologien, wie die Ergebnisse eines an Titan-MC Simulationen geeichten Matlab Modells in Bild 5.4 zeigen. Hier wird der Vergleich zwischen der Einzelpfad und 'Multi-Stage' Analyse von fünf gleich kritischen Pfaden in Serienschaltung gezeigt. Die Einzelpfadanalyse

vernachlässigt im Gegensatz zur Analyse der Pfadtopologie die Kopplung der Pfade über die Schaltcharakteristik der Flip Flop Zellen. Bild 5.4 zeigt den Anteil der Schaltungen ohne Setup-Zeit Verletzung (Survivor Function) bei systematischer, globaler Auslenkung aus dem nominellen Betriebspunkt (z.B. IR-Drop, Temperatur) und gleichzeitiger Überlagerung von lokalen, statistischen Prozessvariationen, wie im Piktogramm von Bild 5.4 dargestellt ist. Die systematische Auslenkung ist normiert auf die Standardabweichung der statistischen Laufzeitschwankung eines ARM926 Pfades. Es wird deutlich, dass die Pfadtopologie die Robustheit der Schaltung beeinflusst, was bei herkömmlicher Einzelpfadanalyse vernachlässigt wird.

Eine weitere, zusätzliche Bewertung des kritischen Zustands ist über die Anzahl an Pfaden möglich, die auf einen kritischen Pfad folgen. Je mehr kritische Pfade am empfangenden Flip Flop des untersuchten Pfades starten, desto größer ist die Wahrscheinlichkeit, dass einer dieser Pfade schaltet. Gleiches gilt für die Anzahl an kritischen Pfaden die am gleichen empfangenden Flip Flop enden. Je höher deren Anzahl, desto höher die Wahrscheinlichkeit, dass den folgenden Pfaden ein zusätzlicher Laufzeitbeitrag übertragen wird. Dementsprechend werden Registerelemente (Flip Flops, Latches) mit einer großen Anzahl von endenden und startenden kritischen Pfaden als besonders kritisch bewertet.
Es ist festzuhalten, dass unabhängig davon, ob Time-Borrowing Techniken eingesetzt oder andere Maßnahmen zur Kompensation von Variationseffekten getroffen werden, jeder Setup-Zeit Verletzung das Durchlaufen der Flip Flop Schaltcharakteristik vorausgeht, d.h. die Pfadtopologie entscheidet stets über die letzten Zeitreserven. Die Klassifizierung von Pfadtopologien ist daher wichtig, um zum einen den Einsatz von Designtechniken über den Bereich des Flip Flop Schaltverhaltens hinaus auf seine Wirkung prüfen zu können und einen effektiven Einsatz solcher Maßnahmen zu ermöglichen, zum anderen den Einfluss von variationsbedingten Laufzeitschwankungen besser abschätzen zu können. Dadurch unterstützt die Klassifizierung von Pfadtopologien eine geeignete Auswahl von Designtechniken bzw. die bevorzugte Behandlung einzelner Pfadelemente, um eine bestmögliche Verbesserung der Schaltung hinsichtlich Geschwindigkeit und Robustheit zu erzielen.

5.2 Definition von topologischen und strukturellen Bewertungskenngrößen

Die im vorherigen Abschnitt 5.1 durchgeführten Untersuchungen basieren auf pfadübergreifenden Strukturen und klassifizieren diese hinsichtlich ihres kritischen Zustands. In diesem Abschnitt werden pfadbasierte Eigenschaften, d.h. strukturelle Eigenschaften von Pfaden unabhängig von der Position und Umgebung der jeweiligen Pipelinestufe untersucht.

Das Pfad- bzw. Gatterspektrum zeigt die Verteilung von Pfaden bzw. Logikgattern in zuvor ausgewählten, kritischen Timing bzw. Laufzeit Bereichen. Als Logikpfade werden alle Gatterkombinationen von sendendem zu empfangendem Register bezeichnet, deren Timing in den ausgewählten Timing Bereich fällt.
Da ein Gatter in verschiedenen Pfaden vorkommen kann, wird für die Darstellung des

Gatterspektrums die Gatteranzahl wie folgt bestimmt. Jedes Gatter wird dem jeweils kritischsten Pfad in dem es vorkommt und somit dessen Pfad-Timing zugeordnet. Doppelzählungen werden somit vermieden.
Die Untersuchungen eines ARM926 Produktdesigns zeigen, dass pro empfangendem Flip Flop im Mittel mehrere 100 kritische Pfade enden. Daher ist es für die Analyse von Pfad- und Gatterspektren erforderlich, nicht nur den kritischsten Pfad pro empfangendem Flip Flop (Standard Analyse) sondern alle Pfade im kritischen Bereich zu berücksichtigen. Dies führt im Vergleich zu Standard Timing-Reports zu einer um den Faktor 100 erhöhten Datenmenge.
Die schaltungstechnische Bewertung einer Schaltung erfordert daher eine abstrahierte Darstellung der hohen Schaltungskomplexität. Die Verwendung von Gatter- und Pfadspektrum hilft bei der Analyse der kritischen Schaltungsteile:

- **Das Pfadspektrum ist Indikator für die Anzahl kritischer Strukturen und Signalkombinationen**

  Ein perfektes 'Path Balancing' führt zu der aus der Literatur bekannten Timing-Wall, d.h. eine Anhäufung von Pfaden mit maximaler Pfadlaufzeit. Ein flacher Anstieg im Pfadspektrum zeigt, dass einzelne Pfade die Geschwindigkeit der gesamten Schaltung limitieren und somit die kritischsten Strukturen darstellen. Da die Anzahl der Pfade auch von den Kombinationsmöglichkeiten der einzelnen Gatter abhängt, reicht das Pfadspektrum alleine nicht als Bewertungsinstrument aus.

- **Das Gatterspektrum gibt Auskunft über die kritische Hardware einer Schaltung**

  Das Gatterspektrum zeigt die Anzahl der Gatter im kritischen Timing Bereich und somit die Anzahl kritischer Netze. So gibt das Gatterspektrum auch Auskunft über den Einfluss lokal auftretender Umgebungsvariationen.

- **Die Kombination aus Gatter- und Pfadspektrum gibt Auskunft über die Verzahnung der Einzelgatter und Pfade**

  Die gemeinsame Analyse von Gatter- und Pfadspektrum ermöglicht es, die Verzahnung bzw. Vernetzung der kritischen Hardware in den geschwindigkeitskritischen Timing Bereichen zu beschreiben und somit den Einfluss von Variationen auf das Schaltverhalten besser abzuschätzen. Dies ermöglicht die Definition von Robustheitskriterien.

Bild 5.5 zeigt das Gatter- und Pfadspektrum der geschwindigkeitskritischen Pfade des untersuchten ARM926 Mikroprozessors. Am deutlich geringeren Anstieg des Gatterspektrums erkennt man, dass viele Pfade topologisch korreliert sind, d.h. dass diese Pfade zum Teil aus den gleichen Gattern bestehen. Am Kreuzungspunkt von Gatter- und Pfadspektrum wird im Mittel jedem neuen Pfad nur ein einziges individuelles Gatter zugeordnet. Diese Beschaffenheit der Pfadkombinationen wird im nächsten Abschnitt näher diskutiert.

5.2.1 Topologische Korrelationen in kritischen Pfaden

Wie im Pfadspektrum des ARM926 bereits zu sehen war, bestimmen topologische Korrelationen den Verlauf von Gatter- und Pfadspektrum. Je größer die topologischen Kor-

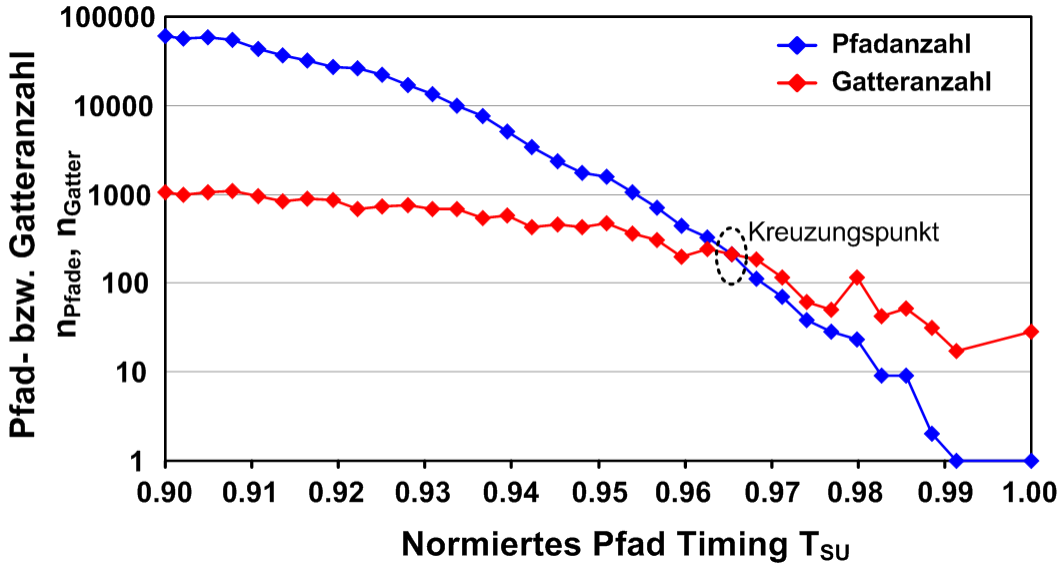


Bild 5.5: Gatter- und Pfadspektrum des ARM926 Produktdesigns in 90nm CMOS Technologie.

relationen, desto enger die Verflechtung der verschiedenen Gatter im kritischen Timing Bereich. Diese Schaltungseigenschaft blieb bei der Betrachtung des Einflusses von Variationen bisher unberücksichtigt.

Bild 5.6 zeigt schematisch die Verflechtung von kombinatorischen Gattern zu verschiedenen Pfaden. Die Pfade 2, 3 und 4 sind stark topologisch korreliert, da die Gatter 7, 8, 9 und FF4 Bestandteil aller dieser Pfade sind. Die Laufzeiteigenschaften dieser Gatter sind somit für alle drei Pfade identisch, so dass sich eine variationsbedingte Laufzeitänderung dieser Gatter auf alle drei Pfade gleichartig auswirkt. Die Vielzahl an Kombinationsmöglichkeiten ist abhängig vom mittleren Fan-In der Gatter, d.h. der Anzahl an Eingangs-Pins und dem mittleren Branching am Ausgang, d.h. der Aufspaltung des zu treibenden Netzes. Diese Größen sind von der jeweiligen Schaltung vorgegeben. Zum anderen hängt die Anzahl an Kombinationen auch von der Pfadlänge ab [99]. Je weniger Gatter zwischen sendendem und empfangendem Flip Flop liegen, desto geringer ist die Kombinationsmöglichkeit. Tieferes Pipelining reduziert daher im Allgemeinen die Kombinationsmöglichkeiten. Die Anzahl zu berücksichtigender Pfade hängt jedoch davon ab, wie viele der Pfade sich im kritischen Bereich befinden. Verschiedene Implementierungsoptionen beeinflussen somit die topologische Korrelation im kritischen Bereich, so dass für jede Implementierung einer vorgegebenen Schaltung eine eigenständige Analyse erforderlich ist.
Zur Bewertung der topologischen Korrelation wird der Topologische Korrelationsfaktor (TKF) κ_{Top} eingeführt, der wie folgt definiert wird:

$$\kappa_{Top}(\Delta T) = \left(1 - \frac{\sum_{i=1}^{Z} \frac{n^i_{Gatter}}{N^i_{G/P}}}{\sum_{i=1}^{Z} n^i_{Pfade}} \right) \tag{5.1}$$

ΔT bezeichnet dabei das zeitliche Intervall im kritischen Bereich, das aus Z Diskretisierungslevel der Größe Δt besteht, beginnend beim kritischsten Pfad. Die Anzahl der Gatter und Pfade im Diskretisierungsintervall i werden mit n^i_{Gatter} und n^i_{Pfade}, die mitt-

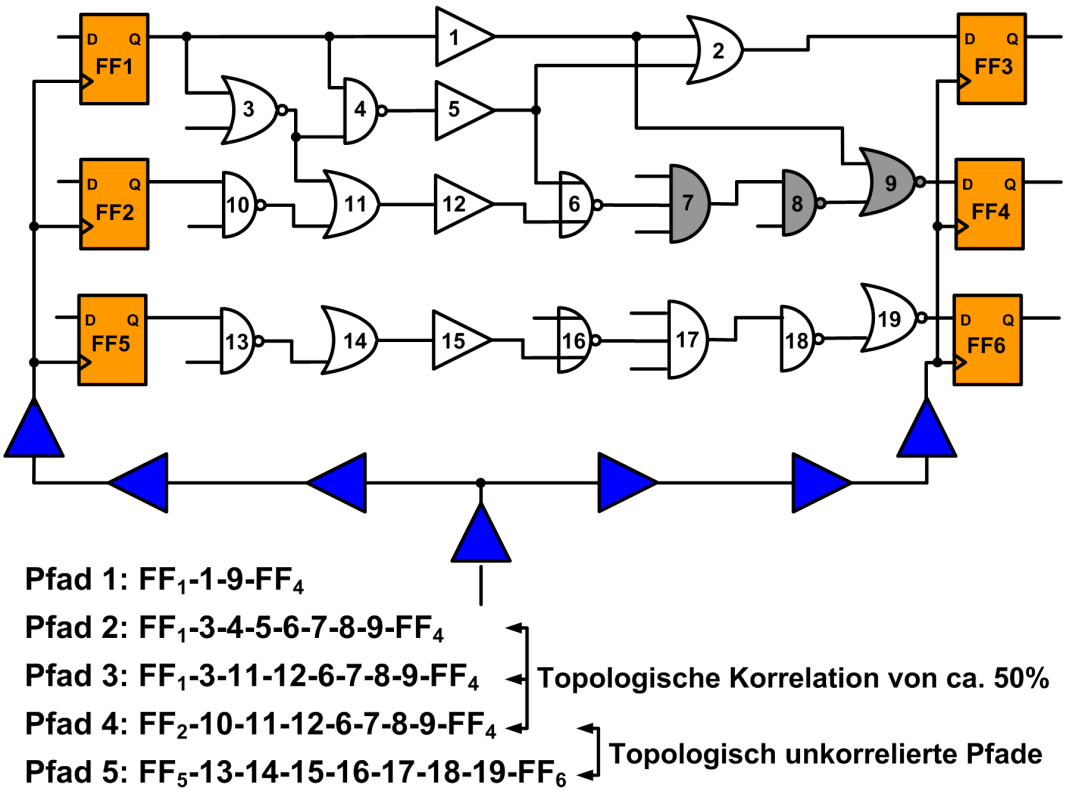


Bild 5.6: Schematische Darstellung der topologischen Korrelation von Pfaden durch die gemeinsame Nutzung von Gattern.

lere Anzahl an Gattern pro Pfad im Diskretisierungsintervall i mit $N^i_{G/P}$ bezeichnet. Ziel ist es, Pfad- und Gatterspektrum zu verknüpfen, indem die mittlere Anzahl an Gattern pro Pfad bestimmt wird. Je geringer diese Anzahl, desto weniger Gatter können einem einzigen Pfad zugeordnet werden, was wiederum eine hohe topologische Korrelation vermuten lässt. Am Beispiel in Bild 5.6 ist zu sehen, dass für kleinere Schaltungen kein topologischer Korrelationsfaktor benötigt wird. Für komplexe Schaltungen mit zehntausenden von Gattern, mehreren hunderttausenden Pfaden und mehreren Pipelinestufen ist jedoch eine abstrakte Darstellung erforderlich, um die schaltungstechnische Komplexität zu beschreiben.
In zahlreichen Publikationen wird der Begriff des unkorrelierten kritischen Pfades verwendet, um die Auswirkungen von Variationen auf synchrone Digitalschaltungen abzuschätzen. Die Verwendung der Anzahl an unkorrelierten Pfaden basiert auf der in [145, 146] beschriebenen Methodik zur Bestimmung des Einflusses von normalverteilten D2D und WID Laufzeitschwankungen auf die maximale Taktfrequenz einer Digitalschaltung. Bild 5.7a illustriert anhand der zugehörigen diskreten Pfad- und Gatterverteilung die für diese Methode getroffenen Annahmen.
Die Fokussierung auf die top-kritischen Pfade gleicher, maximaler Laufzeit vernachlässigt den signifikanten Einfluss subkritischer Pfade innerhalb der n-fachen Standardabweichung der Pfadlaufzeit. Bild 5.7b zeigt schematisch die Erweiterung dieses Ansatzes mit entsprechendem Pfad- und Gatterspektrum unter Annahme einer ansteigenden Pfadanzahl. Die Gatteranzahl liegt in diesem Fall stets über der Pfadanzahl, da mit jedem neuen Pfad zusätzliche Gatter im kritischen Timing Bereich liegen. In realen Schaltungen zeigen sich jedoch hohe topologische Korrelationen, wie in Tabelle 5.2 beispielhaft für ein ARM926

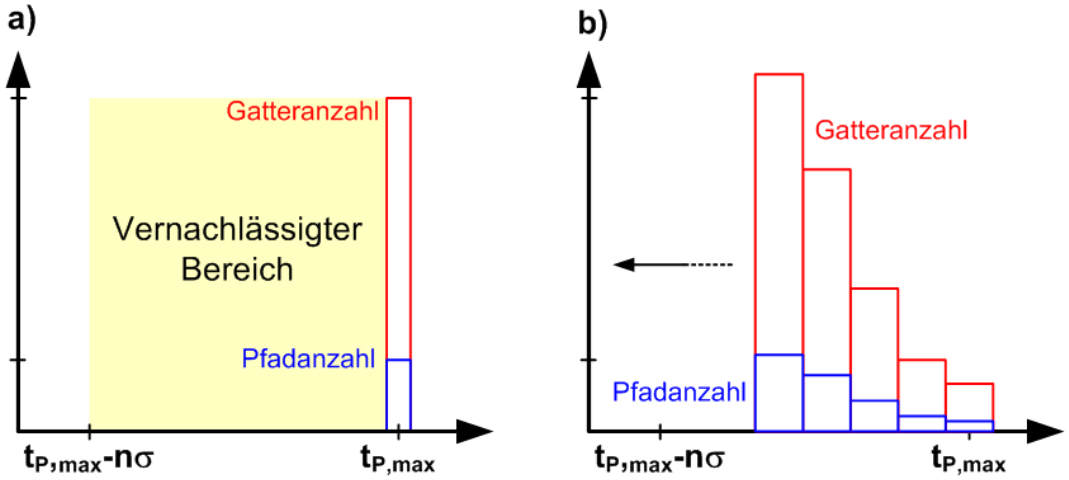


Bild 5.7: Schematische Darstellung der in [145, 146] getroffenen Annahmen und die mögliche Erweiterung dieses Ansatzes.

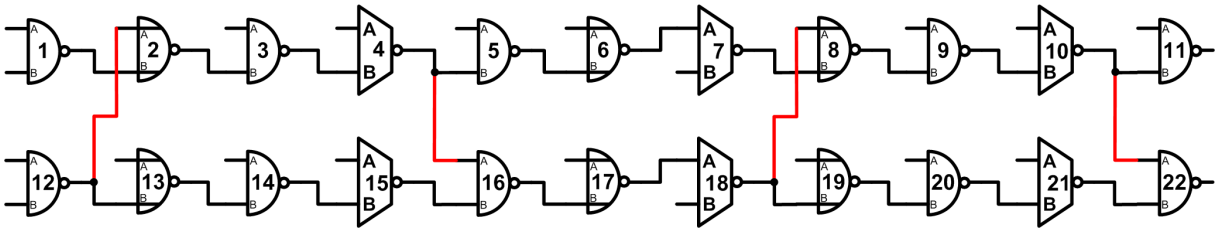
Tabelle 5.2: Topologischer Korrelationsfaktor in den geschwindigkeitskritischen Pfaden des untersuchten ARM926 Produktdesigns.

Zeitliches Intervall ΔT	Obere 3%	Obere 5%	Obere 7%
TKF	88%	96%	98%
Anzahl unabhängiger Pfade	8	31	80
Gesamtanzahl an Pfaden	182	4612	47906

Produktdesign in 90nm CMOS gezeigt wird. Für die Analyse des Einflusses topologischer Korrelationen wird das Pfadspektrum, d.h. die Verteilung sub-kritischer Pfade vorerst vernachlässigt.

Das mit abnehmender Laufzeit zunehmende Verhältnis aus Gesamtpfadanzahl und der Anzahl unabhängiger, unkorrelierter Pfade zeigt, dass einzelne Gatter von einer Vielzahl von Pfaden genutzt werden, so dass sich große topologische Korrelationsfaktoren ergeben. Eine Modellierung der Schaltung nach [145, 146] führt demnach zur Vernachlässigung von wesentlichen strukturellen Eigenschaften der Schaltung. Welche Auswirkung die Vernachlässigung der topologischen Korrelation auf die Laufzeit hat, zeigt Bild 5.8(b) für die Untersuchung der in Bild 5.8(a) gezeigten Testschaltung.

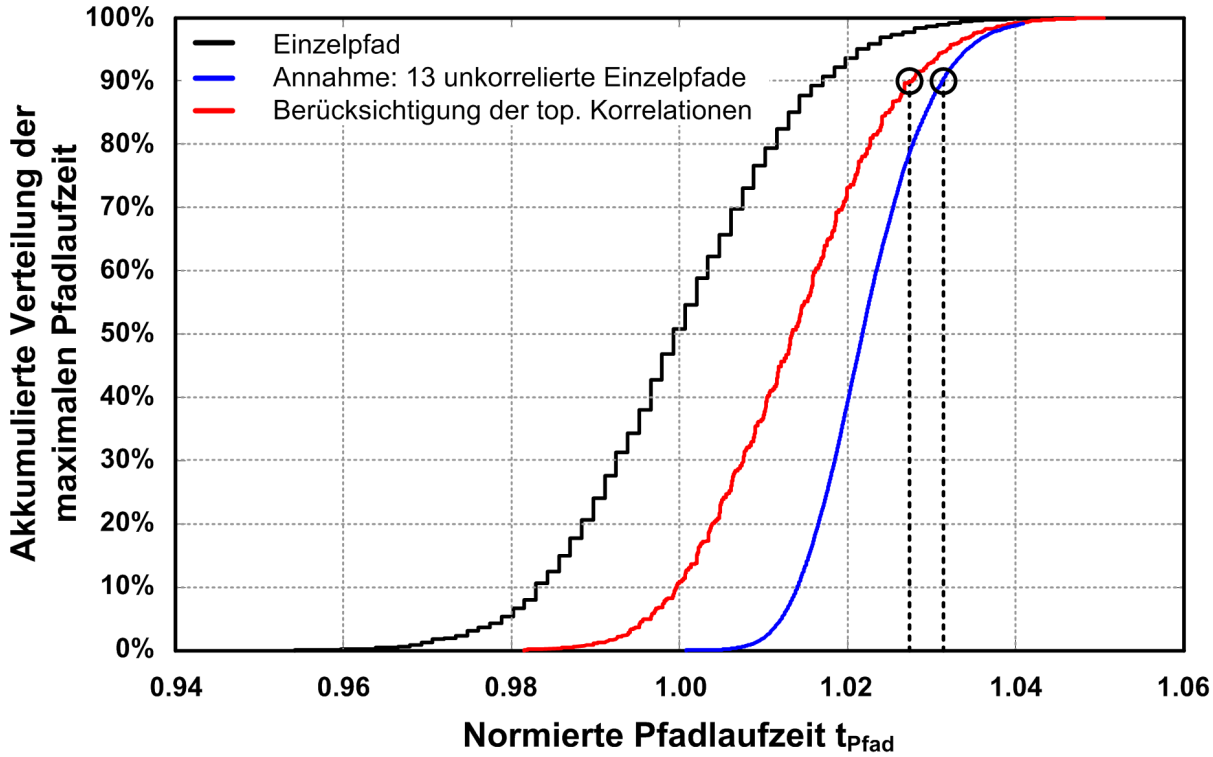
Die Anordnung aus NAND2, NOR2 und Multiplexern wurde so konzipiert, dass die Laufzeiten der einzelnen Pfade auf wenige Pikosekunden übereinstimmen. Somit kann der Einfluss von topologischen Korrelationen im Vergleich zu unkorrelierten Pfaden auf die Verteilung der Pfadlaufzeit analysiert werden. In diesem Szenario existieren im Gegensatz zu realen Schaltungen keine sub-kritischen Pfade, so dass das Ergebnis nur den Unterschied beider Methodiken zeigt. Die Testschaltung wird so konzipiert, dass der topologische Korrelationsfaktor annähernd dem des ARM926 für die oberen 3% der Pfadlaufzeit entspricht.
Bild 5.8(b) zeigt die simulierte Verteilung der maximalen Pfadlaufzeiten bei lokalen statistischen Schwankungen der Transistorparameter. Nimmt man wie in [145, 146] an, dass alle 13 Pfade der Testschaltung unkorreliert sind, so ergibt sich die in Blau dargestellte



Pfad 1: 1-2-3-4-5-6-7-8-9-10-11
Pfad 2: 1-2-3-4-5-6-7-8-9-10-22
Pfad 3: 1-2-3-4-16-17-18-19-20-21-22
Pfad 4: 1-2-3-4-16-17-18-8-9-10-11
Pfad 5: 1-2-3-4-16-17-18-8-9-10-22
Pfad 6: 12-13-14-15-16-17-18-19-20-21-22
Pfad 7: 12-13-14-15-16-17-18-8-9-10-11
Pfad 8: 12-13-14-15-16-17-18-8-9-10-22
Pfad 9: 12-2-3-4-5-6-7-8-9-10-11
Pfad 10: 12-2-3-4-5-6-7-8-9-10-22
Pfad 11: 12-2-3-4-16-17-18-19-20-21-22
Pfad 12: 12-2-3-4-16-17-18-8-9-10-11
Pfad 13: 12-2-3-4-16-17-18-8-9-10-22

$$\kappa_{Top} = \left(1 - \frac{\frac{22}{11}}{13}\right) = 84.6\%$$

(a) Testschaltung aus NAND2, NOR2 und MUX2 Gattern mit 2 unkorrelierten und 13 korrelierten Pfaden



(b) Ergebnisse der Monte-Carlo Simulationen in 65nm CMOS ($V_{DD} = V_{DD}^{nom}$, T=27°C).

Bild 5.8: Untersuchung des Einflusses topologischer Korrelationen auf die statistische Laufzeitschwankung unter der Annahme zeitlich gleich langer kritischer Pfade und der Vernachlässigung sub-kritischer Pfade.

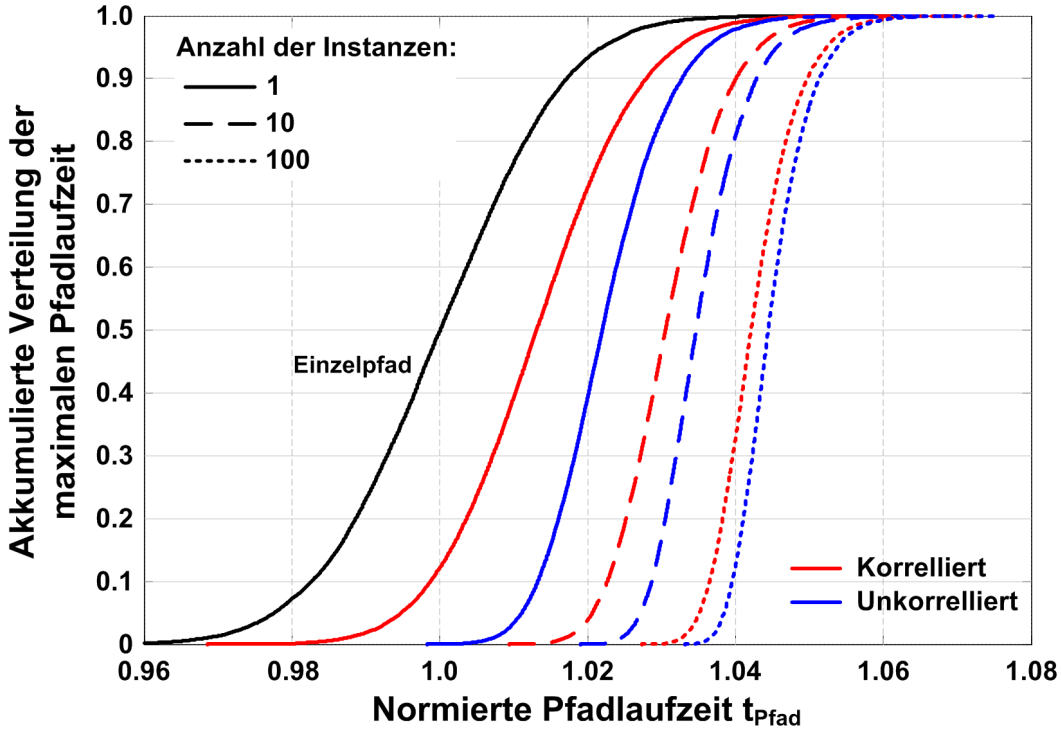


Bild 5.9: Ergebnisse eines erweiterten Matlab Modells zur Analyse des Einflusses der absoluten Anzahl unkorrelierter und topologisch korrelierter Pfade auf die akkumulierte Verteilung der maximalen Pfadlaufzeit.

Verteilung. Da die Pfade jedoch topologisch miteinander korreliert sind, entspricht die für diesen Fall notwendige Testschaltung 13 Pfaden mit jeweils 11 Gattern, so dass die Gesamtgatteranzahl 143 beträgt. Berücksichtigt man die topologischen Korrelationen der Pfade wie sie in Bild 5.8(a) gezeigt sind, so ergibt sich für die Verteilung der maximalen Pfadlaufzeit die rote Kurve. Zum Vergleich ist ebenfalls die Laufzeitverteilung eines Einzelpfades gezeigt.

Die Anzahl der Pfade wird nun erhöht, indem die Testschaltung als Instanz mehrfach verwendet wird. Hierzu werden die simulierten statistischen Schwankungen der Gatterlaufzeiten auf ein Matlab Modell der Testschaltung übertragen. Somit ist es möglich, die nicht-gaußsche und analytisch nicht beschreibbare Laufzeitverteilung der Testschaltung bei Verwendung mehrerer Instanzen und der Nachbildung des Pfadspektrums zu berücksichtigen. In Bild 5.9 sind die Ergebnisse einer Matlab Simulation zu sehen. Das Matlab Modell ermöglicht die Parametrisierung der Instanzanzahl und errechnet die Verteilung der maximalen Pfadlaufzeiten. Dieses Modell liefert leicht optimistische Ergebnisse, da die einzelnen Instanzen keine Querverbindungen enthalten. In Realität wären einzelne Gatter verschiedener Instanzen miteinander verbunden und führten zu einer erhöhten Pfadanzahl.

Es zeigt sich, dass sich der Unterschied zwischen Berücksichtigung der topologischen Korrelation und der Annahme unkorrelierter Pfade weiter verringert und für eine Anzahl von größer als 100 Modulen (d.h. 1300 Pfaden) nahezu verschwindet. Somit hat die topologische Korrelation auf den ersten Blick bei Berücksichtigung statistischer Prozessvariationen keinen Einfluss auf die Verteilung der Pfadlaufzeiten.

Es ist jedoch zu beachten, dass bei Berücksichtigung der topologischen Korrelation im Vergleich zur Annahme unkorrelierter Pfade bereits 14% der Gatteranzahl ausreicht, um

für den 90% Punkt der akkumulierten Verteilung der maximalen Laufzeit die annähernd gleiche Laufzeit zu erhalten. Das heißt eine deutlich geringere Anzahl von Gattern ist aufgrund ihrer Kopplung untereinander hinsichtlich lokaler Variationen gleich unrobust wie eine Schaltung, die aus einer Vielzahl von Gattern besteht. Dieser Aspekt ist insbesondere für die Bewertung weiterer Variationsquellen wie z.B. lokalem IR-Drop, Crosstalk etc. wichtig.
Ein weiterer Unterschied in den Ergebnissen zeigt sich bei der Berücksichtigung des Pfadspektrums. Im Folgenden wird gemäß Bild 5.7b das Pfadspektrum mithilfe von Mehrfach-Instanzen der in Bild 5.8(a) gezeigten Testschaltungen durch Skalierung der Pfadlaufzeiten nachgebildet, um die Auswirkungen des Pfadspektrums auf die Wahrscheinlichkeitsverteilung der Pfadlaufzeiten generisch zu untersuchen.
In Bild 5.10 sind verschiedene Wahrscheinlichkeitsverteilungen der maximalen Pfadlaufzeit unter Berücksichtigung von lokalen statistischen Prozessvariationen gezeigt. Der Vergleich von Mehrfach-Instanz A (117 Pfade) mit den beiden anderen Mehrfach-Instanzen B&C (je 793 Pfade) zeigt, dass die Vernachlässigung des Pfadspektrums große Unterschiede in der Verteilung der maximalen Pfadlaufzeit hervorruft. Selbst bei gleicher Pfadanzahl zeigen sich deutliche Unterschiede je nach zeitlicher Anordnung der Pfade im Spektrum (Mehrfach-Instanzen B&C). Verwendet man die Methodik unkorrelierter kritischer Pfade, indem alle Pfade im gezeigten Pfadspektrum als unkorreliert und mit maximaler nomineller Laufzeit betrachtet werden, so ist zur genaueren Nachbildung durch Berücksichtigung von topologischen Korrelationen und unterschiedlichen nominellen Pfadlaufzeiten ein nicht zu vernachlässigender Unterschied zu erkennen. Weitere topologische Korrelationen, die zwischen den einzelnen Pfaden sowohl gleicher als auch unterschiedlicher nomineller Laufzeit bestehen können, werden in diesem Beispiel nicht berücksichtigt. Die durch zusätzliche Querverbindungen der einzelnen Instanzen zu erwartende deutlich erhöhte Pfadlaufzeit würde zu weiter zunehmenden Unterschied der Wahrscheinlichkeitsverteilungen führen.
Die Darstellung der Verteilungen im Wahrscheinlichkeits-Plot lässt den nicht-gaußschen Charakter der Verteilungen erkennen. Da das Maximum zweier normalverteilter Größen nicht normalverteilt ist, treten insbesondere bei Berücksichtigung vieler teilkorrelierter Pfade Verschiebungen und Verformungen der ursprünglichen Gaußverteilung auf, und eine analytische Darstellung der Wahrscheinlichkeitsverteilung der maximalen Pfadlaufzeit ist nicht mehr möglich [147]. Selbst numerische Lösungen stoßen aufgrund der hohen Schaltungskomplexität moderner integrierter Schaltungen und der Korrelationen zwischen verschiedenen Pfaden an die Grenzen der Durchführbarkeit [133].

Auch wenn in diesen Beispielen die absoluten Unterschiede zwischen den einzelnen Ansätzen aufgrund geringer statistischer Laufzeitschwankungen klein sind ($\leq 1\sigma_{Einzelpfad}$), ist bei Berücksichtigung lokaler Umgebungsvariationen wie z.B. lokalem IR-Drop und Crosstalkeffekten ein deutlich erhöhter Wert zu erwarten. Diese Untersuchungen zeigen, dass ein hoher topologischer Korrelationsfaktor demnach als Indikator für einen starken Einfluss lokaler Variationseffekte auf das Laufzeitverhalten der Gesamtschaltung zu sehen ist.

Bild 5.11 veranschaulicht den erhöhten Einfluss von lokalen Variationen bei topologischer Korrelation im Vergleich zu unkorrelierten Strukturen. Im vorliegenden Fall werden zwei Szenarien betrachtet, die eine gleiche Gesamtgatteranzahl aufweisen. So wirkt sich im Falle topologischer Korrelationen eine lokale Laufzeitschwankung auf alle Pfade aus, die die markierten Zellen beinhalten. Damit ist die Wahrscheinlichkeit höher, dass einer der

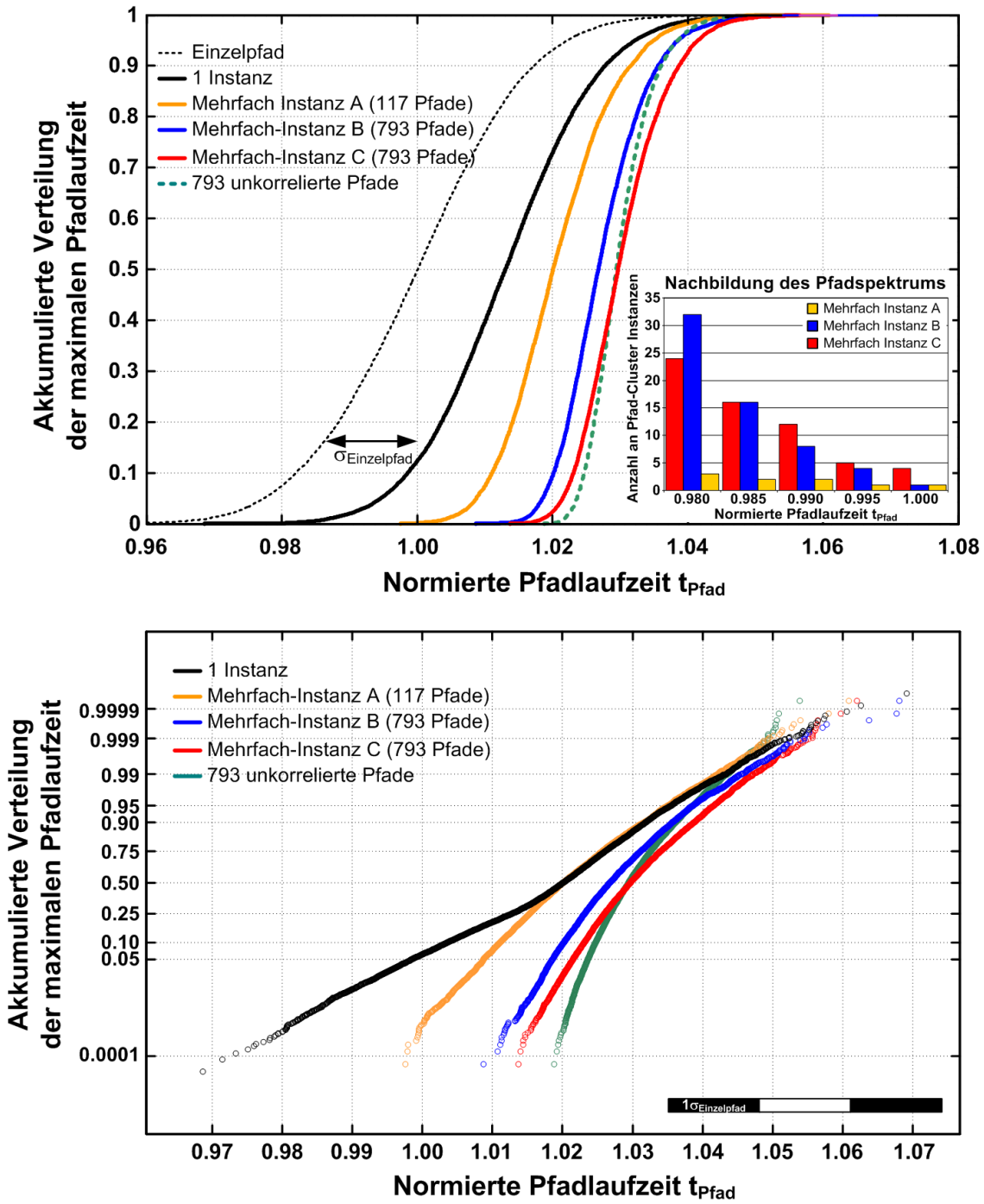


Bild 5.10: Wahrscheinlichkeitsverteilung der maximalen Pfadlaufzeit für die Nachbildung verschiedener Pfadspektren und Vergleich mit der Methodik unkorrelierter Pfade mit maximaler nomineller Laufzeit.

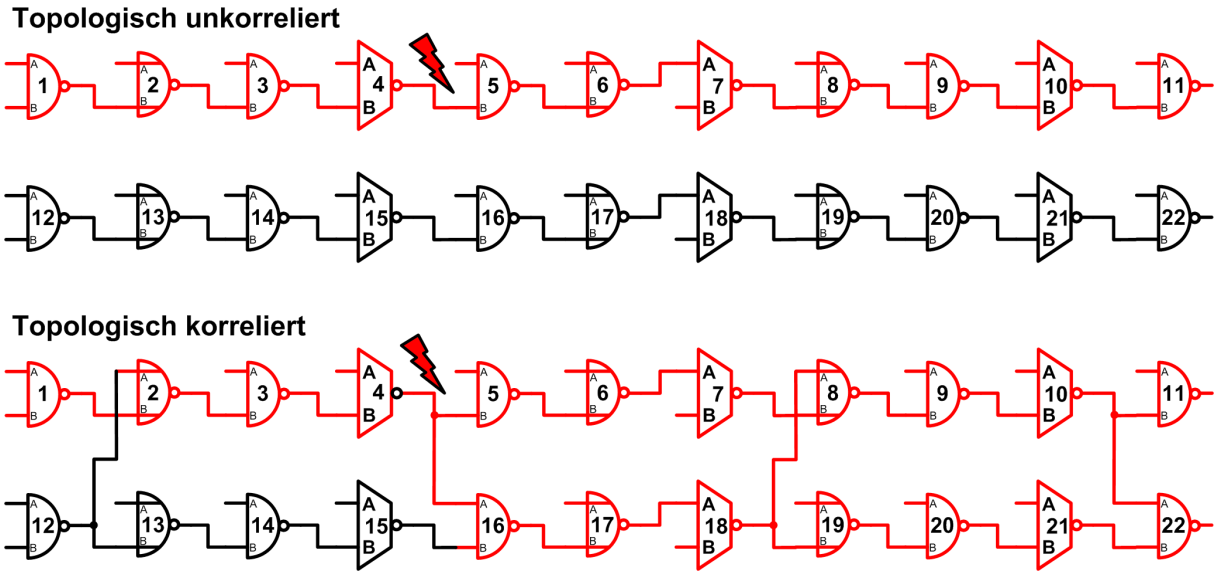


Bild 5.11: Einfluss lokaler Laufzeitschwankungen bei topologisch korrelierten und unkorrelierten Pfaden.

kritischen Pfade sensibilisiert wird und sich die lokale Laufzeitschwankung somit auf das Schaltverhalten der Gesamtschaltung auswirken kann.
Im folgenden Abschnitt werden auf Basis der gewonnen Erkenntnisse Robustheitsdefinitionen vorgeschlagen, die auf der Kombination von Gatter- und Pfadspektren sowie topologischer Korrelationen beruhen.

5.2.2 Struktur- und topologieabhängige Bewertung der Schaltungssensitivität

Prozess- und Umgebungsvariationen führen dazu, dass sich das Schaltverhalten der einzelnen produzierten Schaltungen voneinander unterscheidet. Die Bestimmung, wie groß der Unterschied zwischen den einzelnen Schaltungen sein darf, ohne dass die Funktionalität der Schaltung leidet, ist Ziel eines Robustheitsmaßes. In der Praxis ist die exakte Bestimmung dieses Unterschieds nahezu unmöglich. So ist z.B. die im vorherigen Abschnitt diskutierte Verteilung der maximalen Pfadlaufzeit einer Schaltung nur bedingt als Maß für die Robustheit einer Schaltung anwendbar. Viele nicht modellierbare systematische und pseudo-statistische Variationen machen die Bestimmung der maximalen Pfadlaufzeiten nahezu unmöglich. Da die letztendliche Laufzeitverteilung der Schaltung nicht bestimmt werden kann, ist es notwendig andere Bewertungskriterien zur qualitativen und quantitativen Bewertung von Robustheit zu definieren. Dabei ist es sinnvoll eine Metrik einzuführen, die alle auftretenden WID Prozess- und Umgebungsvariationen berücksichtigt.
Im Folgenden werden zwei topologie- und strukturbasierte Definitionen für das Maß der Schaltungssensitivität gegenüber Variationen vorgeschlagen und diskutiert. Je sensitiver eine Schaltung, desto geringer ist deren Robustheit gegenüber Variationen und Alterungseffekten.
Um die Sensitivität einer Schaltung in Abhängigkeit der Schaltungsstruktur und -topologie zu bestimmen ist es notwendig zwischen lokalen und globalen Schwankungsgrößen zu unterscheiden. In vielen low-power Schaltungen wird die Versorgungsspannung an die globale

Prozessschwankung angepasst, um bei schnellen Schaltungen die dynamischen Verluste zu verringern (Process Voltage Scaling PVS). Dadurch laufen alle Schaltungen - unabhängig von der maximal möglichen Taktfrequenz - bei gleicher Taktfrequenz und sind hinsichtlich weiterer WID Prozess- und on-chip Umgebungsvariationen nahezu gleich verwundbar bzw. kritisch. Deshalb werden im Folgenden für die Bewertung der Schaltungsrobustheit globale Variationen ausgeschlossen.
Der Anteil der geschwindigkeitskritischen Pfade wird über die maximal zu erwartende Auslenkung der Laufzeit aus dem nominellen Fall bestimmt. Der geschwindigkeitskritische Bereich erstreckt sich daher von der Laufzeit des zeitlich längsten Pfades t_{Pfad}^{max} bis zur unteren Timing Grenze $t_{Pfad}^{max} - \Delta T_{Var}$. Mit ΔT_{Var} wird die zu betrachtende Spanne der möglichen Laufzeitvariation bezeichnet.

In der Literatur werden bei der Analyse des Einflusses von Variationen auf die Laufzeit einer Schaltung schaltungsspezifische strukturelle und mikroarchitektonische Eigenschaften meist vernachlässigt. In einzelnen Fällen werden kritische Pfade modelliert und Pfadspektren gezeigt, die Eigenschaften von Pfad- und Gatterspektrum werden jedoch vernachlässigt. Auch das Pfadspektrum bzw. die Anzahl kritischer Pfade alleine reicht zur Analyse nicht aus, da die in Abschnitt 5.2.1 diskutierte topologische Korrelation dazu führt, dass die Anzahl an kritischen Pfaden je nach topologischer Korrelation stark variieren kann. Des Weiteren ist die Bestimmung der Gesamtpfadanzahl einer Schaltung nur mit unverhältnismäßig hohem Aufwand möglich (15GB für die oberen 10% des Timing Bereichs), so dass eine Normierung der Bewertungsgröße und somit ein Vergleich z.B. verschiedener Synthesen einer Schaltung kaum möglich ist.
Aus diesem Grund basiert die folgende Definition der Schaltungssensitivität auf der Verteilung der Gatter im kritischen Bereich (kritische Hardware). Dabei werden die Gatter, wie im vorherigen Abschnitt beschrieben, den jeweiligen Timing Bereichen zugeordnet. Die Gatteranzahl einer Schaltung ist bekannt und kann daher als Normierung verwendet werden. Unabhängig von der jeweiligen Variationsquelle gilt unter der Annahme einer konstanten topologischen Korrelation, je geringer die Anzahl der im kritischen Bereich liegenden Gatter, desto geringer die Wahrscheinlichkeit, dass lokal wirkende Variationen die kritische Hardware beeinflussen und somit die Schaltungsfunktionalität stören. Dies gilt sowohl für Prozessvariationen, wie z.B. den WID Trend als auch für Umgebungsvariationen wie lokalen IR-Drop und Crosstalk.
Eine erste Abschätzung der Schaltungssensitivität s_1 kann somit über die Bestimmung der kritischen Hardware erfolgen [122]:

$$s_1 = \frac{1}{N_{Gatter}} \sum_{i=1}^{Z} n_{Gatter}(i) \tag{5.2}$$

Z bezeichnet die Anzahl der Zeitintervalle Δt_i innerhalb des kritischen Timing Bereichs $\Delta T_{krit} = Z \cdot \Delta t_i$ und N_{Gatter} die Gesamtanzahl der Gatter einer Schaltung. Je mehr Zellen sich im kritischen Bereich befinden, desto unrobuster das Schaltungsdesign. Diese Definition der Schaltungssensitivität ermöglicht erstmals eine quantitative Bewertung der Schaltungsrobustheit und eröffnet die Möglichkeit zwischen Geschwindigkeitsmarge und Robustheit zu unterscheiden [122].
Bild 5.12 veranschaulicht den Vergleich der Schaltungssensitivität z.B. vor bzw. nach dem Einsatz von präventiven Designtechniken zur Implementierung einer zeitlichen Sicherheitsmarge. Die Pfadverteilung und die akkumulierte Gatterverteilung der Schaltung sind vor

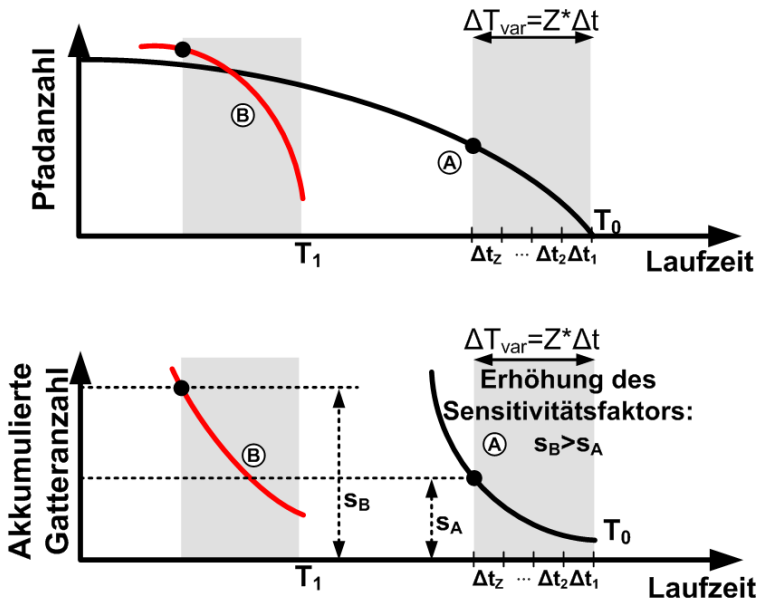


Bild 5.12: Veranschaulichte Definition der Schaltungssensitivität.

(A) und nach Ergreifen der Maßnahme (B) gezeigt.
Als Beispiel für den Einfluss der kritischen Hardware auf den Einfluss von Variationen auf die Robustheit einer Schaltung dient die Wahrscheinlichkeitsverteilung der maximalen Pfadlaufzeit von Mehrfach-Instanz A&B in Bild 5.10. Der Sensitivitätsfaktor nach Gleichung 5.2 ist für Mehrfach-Instanz B demnach um den Faktor 6.8 höher als für Mehrfach-Instanz B. Da der Sensitivitätsfaktor auf Gatterebene definiert wird, ist dessen Anwendung nicht auf statistische Prozessvariationen beschränkt sondern deckt alle lokalen Laufzeitschwankungen ab. Eine hohe Schaltungssensitivität s_1 ist daher Indikator für einen starken Einfluss von WID Prozess- und Umgebungsvariationen auf die Robustheit einer Schaltung.

Bei dieser Definition der Schaltungssensitivität werden alle Gatter im kritischen Bereich gleich gewichtet, so dass die Veränderungen im Gatterspektrum direkt am Sensitivitätsfaktor abzulesen sind. Um die Anordnung der jeweiligen Zellen im kritischen Bereich besser zu berücksichtigen wird für eine zweite Definition der Schaltungssensitivität eine Gewichtungsfunktion der jeweiligen Timing Bereiche vorgeschlagen. Mit zunehmendem zeitlichem Abstand zum kritischsten Pfad nimmt der Einfluss der einzelnen Zellen (siehe Bild 5.10) und somit auch der Gewichtungsfaktor der Zellen im jeweiligen Timing Bereich ab. Als Gewichtungsfunktion wird eine modifizierte Tangens Hyperbolicus Funktion verwendet:

$$w(i) = 0.5 \cdot tanh\left(\frac{5.3}{\Delta T_{krit}} \cdot (-i \cdot \Delta t) + 2.65\right) + 0.5 \tag{5.3}$$

Diese Funktion stellt sicher, dass der Gewichtungsfaktor $w(i)$ aller Zellen im kritischsten Pfad bei 0.995 und der aller Zellen am Ende des kritischen Bereiches ΔT_{krit} bei 0.005 liegt. $w(i)$ bezeichnet dabei den Gewichtungsfaktor aller im zeitlichen Intervall i liegenden Gatter. Bild 5.13 zeigt die Gewichtungsfaktoren für einen Bereich von $\Delta T_{krit} = 0.1$.
Zum Vergleich ist eine normierte (einseitige) Gaußverteilung gezeigt ($\mu = 1, \sigma = 1/30$), wie sie im Schaltungsentwurf als Verteilung von Gatter- und Pfadlaufzeiten verwendet

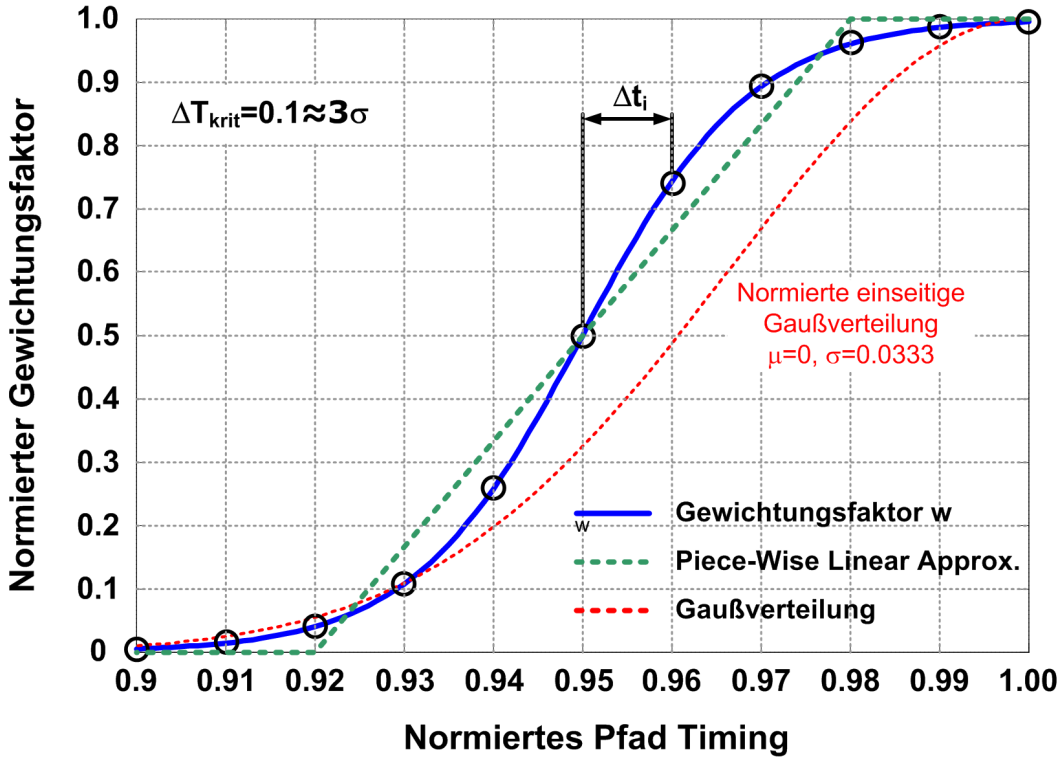


Bild 5.13: Darstellung des Gewichtungsfaktors für einen kritischen Timingbereich von 10%.

wird. Im Vergleich zur Gaußverteilung werden die Zellen mit geringem und mittlerem Abstand stärker gewichtet. Da Umgebungsvariationen den größten Einfluss auf die Laufzeitschwankungen haben, können selbst einzelne Ereignisse zu größeren Laufzeitsprüngen führen, ohne dass dafür mehrere Variationen positiv korreliert sein müssen. Für weite Abstände zum kritischsten Pfad ist eine Kombination mehrerer Ereignisse bzw. Effekte notwendig, um geschwindigkeitskritisch zu sein. Deshalb werden die Zellen mit kurzem und mittlerem Abstand stärker gewichtet, da für die Überbrückung dieses Abstandes nicht zwangsläufig mehrere Variationseffekte gleichzeitig auftreten müssen.
Die Wahl des Gewichtungsfaktors basiert daher auf den im Kapitel 3 und 4 gewonnen Erkenntnissen zur Größenordnung der einzelnen Variationen und ist somit ein empirischer, heuristischer Ansatz um die Position der kritischen Hardware im Gatterspektrum zu berücksichtigen. Die Gewichtung erfolgt anhand des Abstandes zum kritischsten aller Pfade. Im Fall des ARM926 ist es auch möglich den prinzipiellen Verlauf des Gewichtungsfaktors aus Einfachheitsgründen z.B. wie in Bild 5.13 gezeigt durch eine Piece-Wise Linear Approximation nachzubilden. Wichtig dabei ist jedoch, die Größenordnung der einzelnen WID Laufzeitschwankungen bei der Wahl des Gewichtungsfaktors zu berücksichtigen. Existiert z.B. eine dominante Laufzeitschwankung mit z.B. 75% Anteil an der Gesamtlaufzeitschwankung, so ist es ratsam, den Gewichtungsfaktor anzupassen, z.B. durch die Verschiebung der vorgeschlagenen Gewichtungsfunktion in Richtung kleinerer normierter Pfadlaufzeiten.
Die Optimierung der Verlustleistung z.B. durch Mixed-V_T Design, ist für energiearme CMOS Digitalschalschaltungen essentiell und führt zur Verlangsamung der sub-kritischen Pfade und somit zu einer erhöhten Anzahl von top-kritischen Pfaden. Im Pfadspektrum ist dies am Aufbau einer Timing-Wall zu erkennen. Eine ideale Timing-Wall existiert,

wenn sich alle Pfade durch das gleiche Pfad-Timing auszeichnen. Ein Gewichtungsfaktor für das Gatterspektrum wäre in diesem Fall nicht notwendig. Für komplexe Semi-Custom Schaltungen wie z.B. eingebetteten Mikroprozessoren wird eine ideale Timing-Wall jedoch nicht erreicht. Ausgeglichene Laufzeiten der Logikpfade werden nur durch Verlangsamung und Beschleunigung der Gatterlaufzeiten erzielt. Eine Verlangsamung der Pfade ist nur bedingt möglich, da neben den Gatterlaufzeiten auch Designkriterien wie z.B. die minimale Steilheit der Signalflanken betroffen sind, was eine beliebige Reduzierung des Treiber-Last Verhältnisses und somit die Balancierung der Laufzeiten verhindert. Aus diesem Grund wird in der Realität ein Pfadspektrum generiert, das sich stets von einer idealen Timing-Wall unterscheidet. Aus diesem Grund ist es wichtig und sinvoll, eine Gewichtung der Gatter je nach ihrer Position im Gatterspektrum vorzunehmen.

Kombiniert man alle strukturellen Eigenschaften zur Bestimmung der Schaltungssensitivität so ergibt sich folgender Zusammenhang:

$$s_2 = \frac{1}{N_{Gatter}} \sum_{i=1}^{Z} (c + \kappa_{Top}(\Delta t \cdot i)) \cdot w(i) \cdot n_{Gatter}(i) \tag{5.4}$$

Die Konstante c wird bei der Analyse einer Schaltung so gewählt, dass eine angemessene Gewichtung der topologischen Korrelation erfolgt. Für die Untersuchungen in Kapitel 6 wurde c=2 gewählt, da der Beitrag aller umittelbar lokal auftretenden Variationen (stat. Prozessschwankung, Crosstalk) zur Timing Unsicherheit des ARM926 mittels Mikroprozessormodell auf 37% bestimmt wurde. Bei einer idealen topologischen Korrelation $\kappa_{Top} = 1$ erfolgt für $c = 2$ eine Gewichtung der lokalen Effekte mit 0.33.
Da die hohe Komplexität der Schaltung auf topologischen und strukturellen Eigenschaften beruht, basieren die hier vorgeschlagenen Definitionen der Schaltungsrobustheit genau auf diesen Aspekten. Durch die Beschreibung struktureller und topologischer Eigenschaften ist es möglich, die Komplexität der Schaltung auf vereinfachte Art und Weise phänomenologisch zu beschreiben. Eine Erweiterung der bisherigen Definitionen von Schaltungsrobustheit ist durch die Hinzunahme weiterer struktureller und topologischer Kenngrößen, wie z.B. die Lage des Splitting Points im Taktbaum, Transistorgröße, Gattertopologie etc. möglich und sinnvoll.
Ziel dieser Definitionen ist es, robustheitsbeeinflussende Eigenschaften der Schaltungsstruktur und -topologie geeignet zu gewichten, um somit unabhängig von der jeweiligen Variationsquelle ein Maß für die Verwundbarkeit der Schaltung bestimmen zu können.

Die Definition der Schaltungssensitivität ermöglicht es erstmals, zwischen zeitlichen Sicherheitsmargen einer Schaltung und Robustheitsaspekten zu unterscheiden. Hintergrund dieses Vergleichs ist die Tatsache, dass im Schaltungsentwurf oftmals zeitliche Sicherheitsmargen abgebaut werden um die maximale Taktfrequenz zu erhöhen bzw. abzusenken. Eine Analyse der Schaltungsrobustheit vor und nach dem Abbau der Sicherheitsmarge wird durch die Definition der Schaltungssensitivität möglich. Kompensationstechniken, wie sie in Kapitel 6 diskutiert werden, können somit hinsichtlich ihres Einflusses auf die Schaltungssensitivität bewertet und somit verglichen werden. Zusätzlich zu den traditionellen Kenngrößen Leistungsaufnahme, Flächenbedarf und Geschwindigkeit für die Bewertung von Kosten und Nutzen verschiedener Maßnahmen eröffnet die Kenngröße des Schaltungssensitivitätsfaktors neue Entscheidungsgrundlagen beim Schaltungsentwurf.
Ein weiterer Vorteil der in diesem Kapitel definierten strukturellen Kenngrößen ist die

einfache Einbindung als Optimierungsgröße in die Schaltungssynthese, wozu z.B. die Verwendung von SSTA nicht geeignet ist [129, 125]. Da alle Kenngrößen auf Standard Design-Flow Daten beruhen, ist die Integration in den Design-Flow durch geeignete add-on Software problemlos möglich.
Die Schaltungssensitivität dient als Kriterium für die Robustheit einer Schaltung. Der Zusammenhang zwischen Schaltungssensitivität und der Verteilung der maximalen Pfadlaufzeiten kann aufgrund der Komplexität moderner Schaltungen und der in der Praxis nicht-modellierbaren Einflussgrößen nicht evaluiert werden. Dieser Ansatz zur Bewertung der Schaltungsrobustheit muss demnach als heuristische Methode zum Umgang mit Variationen in komplexen Schaltungen auf Basis erstmals definierter struktureller und topologischer Kenngrößen gesehen werden. Die Anwendbarkeit der Bewertungsgrößen beschränkt sich nicht auf Mikroprozessoren sondern ist für alle Digitalschaltungen möglich.

6 Schaltungstechnische Ansätze zur Kompensation von Laufzeitschwankungen

In diesem Kapitel werden Designtechniken diskutiert, die zur Kompensation von variationsbedingten Laufzeitschwankungen dienen. Dabei wird zwischem aktiver und inhärenter Kompensation unterschieden. Mit aktiver Kompensation wird die Detektion von zu kompensierenden Laufzeitschwankungen und die entsprechende Anpassung z.B. der Betriebsparameter bezeichnet. Inhärente Kompensation ist die eigenständige Kompensation des Systems bzw. der Schaltung z.B. durch bestehende zeitliche Sicherheitsmargen oder durch den Betrieb in insensitiven Bereichen. Man unterscheidet dabei zwischen verschiedenen schaltungstechnischen Kompensationstechniken (Bild 6.1):

- **Adaption nach der Herstellung (Post-Fabrikation)**

  Derartige Gegenmaßnahmen basieren auf der Bestimmung des Chip-Zustands nach der Herstellung und der anschließenden Adaption von Betriebsparametern zur Sicherung der spezifizierten Kenngrößen, wie z.B. Taktfrequenz, Leckstrom etc.. Die jeweiligen Kenngrößen werden durch Testläufe vor der Auslieferung oder regelmäßige Überwachung (Monitoring) durch on-chip Teststrukturen während des Betriebs bestimmt. Es lassen sich sowohl statische (Process Voltage Scaling PVS, De-/Skewing im Taktbaum) als auch dynamische Adaptionsmaßnahmen (Adaptive Voltage Scaling AVS) ableiten.

- **Präventive Techniken**

  Präventive Techniken werden während des Schaltungsentwurfs vorgenommen. Die Implementierung von zeitlichen Sicherheitsmargen während des Schaltungsentwurfs kann sowohl global als auch lokal erfolgen. Eine Möglichkeit ist, die Laufzeitsensitivität gegenüber PVT Variationen und somit auch die Breite der Laufzeitschwankung zu reduzieren. Eine weitere Option liegt in der Verringerung der Laufzeit und dem damit verbundenen Geschwindigkeitsgewinn bzw. der Reduzierung der variationsbedingten, absoluten Laufzeitschwankungen.
  Ein anderer Ansatz ist die Implementierung von Time-Borrowing (TB) Maßnahmen [134], z.B. durch den Einsatz von gepulsten Flip Flops (P-FF) bzw. gepulsten Latches (P-L). Neben der durch Time-Borrowing integrierten zeitlichen Elastizität führt die verkürzte Setup-Zeit von P-FFs und P-Ls zu verringerten Pfadlaufzeiten und somit auch zu zeitlichen Sicherheits-Margen.

- **Fehlerdetektion und -korrektur**

  Im Gegensatz zu den o.g. Maßnahmen, die der Vermeidung von Fehlern dienen, basiert dieser Ansatz auf der Fehlerdetektion. Durch die Überwachung der kritischen Schaltungsteile und die einhergehende Detektion eines Fehlers kann der funktionale

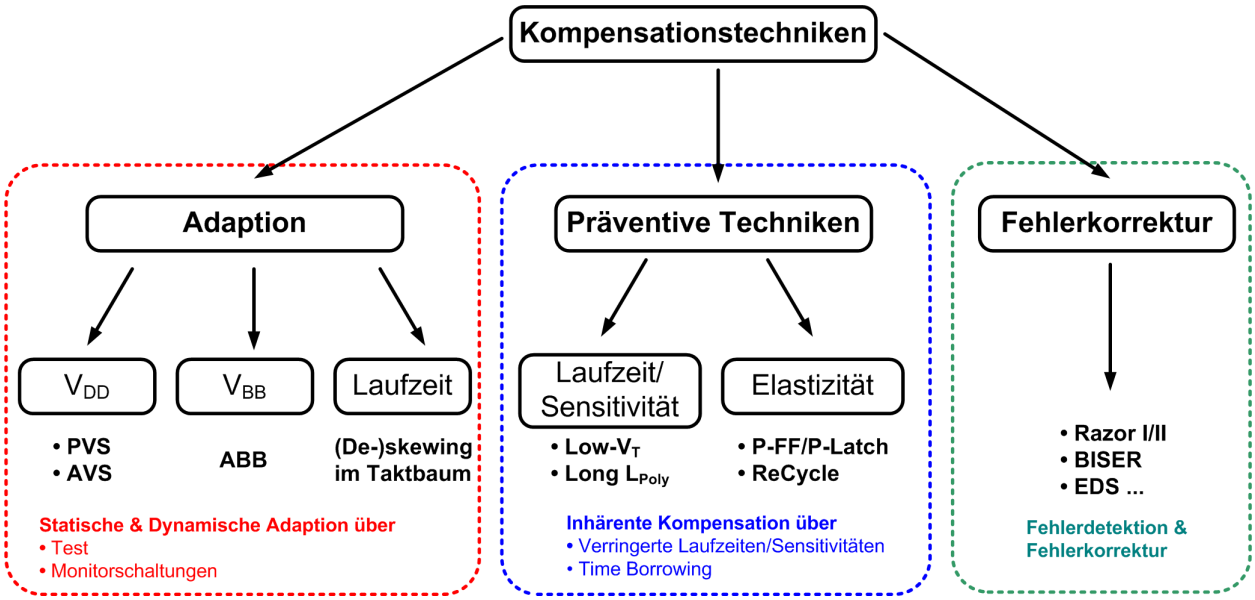


Bild 6.1: Klassifizierung verschiedener Techniken zur Kompensation von Laufzeitvariationen.

Ausfallpunkt der Schaltung bestimmt werden. Aus der veränderten Fehlerrate nach Adaption der Betriebsparameter lässt sich die Größenordnung der Anpassung z.B. von V_{DD} bestimmen.

Der Einfluss dieser Maßnahmen auf die Transistor- und Zelltopologie, Synthese und Mikroarchitektur ist je nach Ansatz stark unterschiedlich. Bild 6.2 zeigt eine Übersicht über die Anforderungen bzw. Voraussetzungen bei Implementierung der Maßnahmen.

6.1 Globale Post-Fabrikation Adaptionstechniken

In diesem Abschnitt wird der Umgang mit globalen Schwankungen bzw. deren Kompensation diskutiert. Global wirkende Variationen können als Offset zwischen verschiedenen Dies betrachtet werden. Alle Elemente eines Dies sind von globalen Schwankungen in gleicher Weise betroffen [40, 47], so dass es keinen signifikanten schaltungsspezifischen Einfluss der strukturellen und topologischen Eigenschaften einer Schaltung auf das Verhalten gegenüber solchen Variationen gibt.

6.1.1 Process und Adaptive Voltage Scaling

Der gegensätzliche Anspruch, immer höhere Taktfrequenzen und Datenraten bei gleichzeitig sinkender Energieaufnahme zu ermöglichen, führte zur Implementierung von Dynamic Voltage Scaling (DVS) in low-power Digitalschaltungen. Die Absenkung der Versorgungsspannung für Anwendungen mit geringeren Geschwindigkeitsanforderungen ermöglicht dabei die deutliche Reduzierung der dynamischen Verlustleistung [148]. Insbesondere für Technologien vor dem 130nm CMOS Knoten war das Potential zur Energieersparnis aufgrund der vergleichsweise geringen Laufzeitsensitivität gegenüber veränderten Versorgungsspannungen hoch.
Die nahezu standardmäßige Implementierung von DVS [149] in modernen low-power Schaltungen eröffnet die Möglichkeit variationsbedingte Laufzeitschwankungen durch die

Process Voltage Scaling	---	---	• V_{DD} abhängige Zellcharakterisierung	• Adaptionsfähige Spannungsregler • Prozessmonitor mit Programmiereinheit	• Referenztest $t=t_0$ zur Bestimmung der Prozessklasse
Adaptive Voltage Scaling [75,148,149,150]	---	---	• V_{DD} abhängige Zellcharakterisierung	• Adaptionsfähige Spannungsregler • Laufzeit-Monitor mit Interface zu on-chip Power Management	• Referenztest $t=t_0$ • On-line Test und Auswertung des Laufzeit Monitors
Adaptive Body Biasing [36,151-159]	• Triple Well Prozess	• Body-Kontakte • Metal-Track für Verteilung von V_{BB}	• ABB abhängige Zellcharakterisierung	• 2 zusätzliche, adaptionsfähige Spannungsregler (LDO) • Laufzeit-Monitor mit Interface zu on-chip Power Management	• Referenztest $t=t_0$ • On-line Test und Auswertung des Laufzeit Monitors
Adaptive Skewing [139]	---	• Programmierbare Laufzeitelemente	• Multi-Stage STA • Hold-Time Fixing	• Programmiereinheit	• Referenztest $t=t_0$ • Testalgorithmus für statisches Time Borrowing
(De-) Skewing [137,138]	---	---	• Multi-Stage STA • Hold-Time Fixing	---	---
Low-V_T [160-162]	• Zusätzliche Implants	• Standardzellenbibliothek	• Zellcharakterisierung	---	---
Long L_{Poly} [163-165]	• Long L_{Poly} Maske	• Standardzellenbibliothek	• Zellcharakterisierung	---	---
P-FF/P-Latch [135,136,166-170]		• Erweiterte Standardzellenbibliothek	• FF Ersetzungsstrategie • Multi-Stage STA • Pulse Propagation • Hold-Time Fixing	---	---
ReCycle [144]	---	---	• Hold-Time Fixing	• Zusätzliche Pipelinestufe(n)	---
Fehlerdetektion am Flip Flop	**Allgemein:** • Zweifache Abtastung des Datensignals		• Hold-Time Fixing		• In-situ on-line Test
Beispiel 1: Razor I [171]	---	---	• Zusätzliches Latch • Hold-Time Fixing • Clock-Gating	• Fehlerdetektion & Single Cycle Verarbeitung • Clock-Gating Kontrolle • IPC Reduktion	• In-situ on-line Test und Auswertung der Razor FF Zustände
Beispiel 2: Razor II [172]	---	• Latch mit Detektion von Signal übergängen	• Hold-Time Fixing • Clock-Gating	• Fehlerdetektion & Single Cycle Verarbeitung • Clock-Gating Kontrolle • Pipelinekontrolle (Pipeline Flush) • IPC Reduktion • Gefahr einer Systemblockade	• In-situ on-line Test und Auswertung der Razor FF Zustände
Beispiel 3: BISER [173]	---	• Flip Flop mit , redundantem Latch und Muller-C Element • Flip Flop mit Stabilitäts-Check	• FF Ersetzungsstrategie • Hold-Time Fixing		• In-situ on-line Test mit automatischer Fehlerkorrektur

Bild 6.2: Aufwand zur Implementierung verschiedener Kompensationstechniken.

Adaption der Versorgungsspannung zu kompensieren, ohne dabei erhebliche Zusatzkosten hinsichtlich der zusätzlich benötigten Hardware zu verursachen.
Um den Einfluss von Process und Adaptive Voltage Scaling auf die Energieaufnahme E_{tot} und die Laufzeit t_d abzuschätzen, können folgende Gleichungen verwendet werden:

$$E_{tot} = E_{dyn} + E_{stat} = \alpha_{Schalt} \cdot C \cdot V_{DD}^2 + \alpha_{Schalt} \cdot V_{DD} \cdot I_{sc} + V_{DD} \cdot I_{leak} \cdot T_{Clk} \quad (6.1)$$

$$t_d \sim \frac{C_{eff} \cdot V_{DD}}{\left(V_{DD} - V_{T_{eff}}\right)^{\alpha_{dyn}}} \approx t_d^{nom} \cdot \left(1 + S_{rel}^{V_{DD}} \cdot \Delta V_{DD}\right) \quad (6.2)$$

E_{tot} bezeichnet die mittlere Gesamtenergie pro Taktzyklus, I_{sc} den Kurzschlussstrom während des Schaltvorgangs der Gatter, I_{leak} den statischen Leckstrom der Schaltung und $S_{rel}^{V_{DD}}$ die relative Sensitivität der Laufzeit gegenüber V_{DD} Schwankungen.

Bei der Anpassung der Versorgungsspannung an die Beschaffenheit des einzelnen Dies unterscheidet man zwischen Process- und Adaptive Voltage Scaling.

- **Process Voltage Scaling (PVS)**

  Statische Prozessvariationen können durch statische Maßnahmen zur Laufzeitadaption kompensiert werden. Je nach Schaltung werden verschiedene Maßnahmen zum Umgang mit globalen Variationen angewandt. Während bei modernen standalone CPUs z.B. von Intel, AMD und IBM langsame Chips, die nicht mit der nominellen Taktfrequenz betrieben werden können, für niedrigere Frequenzen spezifiziert werden (Frequency-Binning) [126], ist dies für System on Chip (SoC) Designs aufgrund spezifizierter Standards wie z.B. HSDPA/HSUPA in Mobilfunk-Chips nicht möglich. Um diese Standards zu erfüllen, müssen Mindestanforderungen an Frequenzen bzw. Datenraten eingehalten werden. Daher werden Schaltungen stets für den worst-case Fall entworfen, so dass der Großteil aller produzierten Chips diese Geschwindigkeitsanforderungen deutlich übertreffen. Hier kann die Versorgungsspannung gesenkt und Energie eingespart werden. Da neben den global wirkenden Variationen, die im Design durch verschiedene Prozess- und Spannungs-Corner berücksichtigt werden, auch WID Variationen auftreten, deren Einfluss auf die Schaltgeschwindigkeit bedeutend schwieriger zu modellieren ist, ist es möglich, dass die daraus resultierenden Laufzeiterhöhungen eine Erhöhung der Versorgungsspannung erfordern. Die für die jeweiligen Betriebsmodi festgelegten Versorgungsspannungswerte werden nach dem Test für jeden Chip individuell angepasst und in einer Lookup-Table hinterlegt.

- **Adaptive Volatge Scaling (AVS)**

  Adaptive Voltage Scaling ist vom Grundgedanken dem Process Voltage Scaling sehr ähnlich. Auch hier werden die festgelegten Versorgungsspannungen der einzelnen Betriebsmodi je nach Beschaffenheit des Dies angepasst [75, 148, 149, 150]. Im Gegensatz zur statischen Anpassung bei PVS besteht das AVS-System aus Sensorik und Aktorik und bildet einen Regelkreis. Der Sensor besteht dabei aus einem/mehreren Device under Test (DUT) z.B. Ringoszillatoren oder Replika von kritischen Pfaden, deren Geschwindigkeit in zuvor festgelegten zeitlichen Abständen on-chip gemessen wird.
  Da Prozessvariationen statisch wirken und sich mit der Betriebszeit nicht ändern, ist die Anwendung von Adaptive Voltage Scaling nur sinnvoll, um Laufzeitschwankungen aufgrund von dynamischen bzw. zeitlich abhängigen Variationen zu kompensieren. IR-Drop bedingte Spannungsschwankungen und Crosstalk-Effekte zeichnen sich

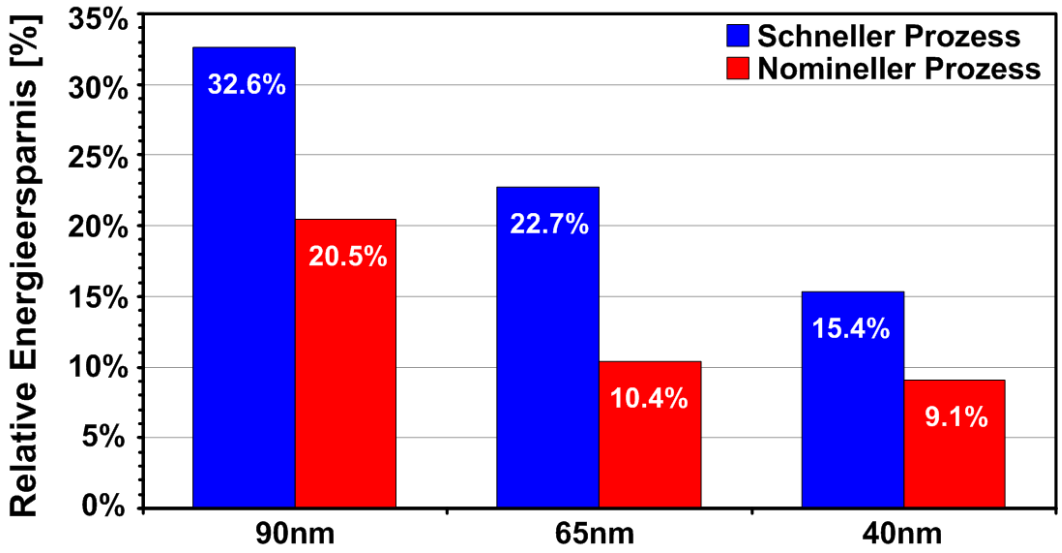


Bild 6.3: Abschätzung der dynamischen Energieersparnis durch PVS bei T=27°C.

durch ihre geringen Zeitkonstanten aus, so dass eine Anpassung der Versorgungsspannung äußert schwierig ist. In low-power Schaltungen werden als Spannungsregler oftmals die sehr energieeffizienten Buck-Converter verwendet [174, 175, 176, 177]. Dabei wird darauf geachtet, dass die Taktrate des Spannungsreglers möglichst klein gewählt wird, so dass die dynamische Verlustleitung im Regler gering gehalten wird. Bei einer Taktrate von wenigen MHz, d.h. mehreren 10-100 Taktzyklen des Systemtakts, ist eine Anpassung der Versorgungsspannung innerhalb weniger Nanosekunden nicht möglich. Erhöhte Frequenzen des Spannungsreglers führen insbesondere bei hohen Idle bzw. Standby Zeiten zu großen Power-Overheads [175]. Die stringenten Vorgaben hinsichtlich der Leistungsaufnahme verhindern somit die schnelle Taktung des Spannungsreglers und somit die Anpassung der Versorgungsspannung an IR-Drop und Crosstalk-Effekte.
Temperaturschwankungen weisen unter allen dynamischen Effekten die größten Zeitkonstanten auf. Die Ausbreitung der Temperatur auf einem Chip erfolgt im Bereich von Millisekunden. Bei transienten Response-Zeiten eines herkömmlichen Buck-Converters von ungefähr 10-20μs ist die Anpassung der Versorgungsspannung an die globale Betriebstemperatur möglich [177].
Auch Alterungseffekte wie NBTI und Hot Carrier Injection, die die Laufzeit von Gattern und Pfaden erhöhen, können durch AVS kompensiert werden.
Die Adaption von Betriebsparametern an die Beschaffenheit, bzw. den Zustand des Chips erfordert neben der Aktorik eine geeignete Sensorik. Aus diesem Grund gewann das 'on-chip Monitoring' in den letzten Jahren an Bedeutung. Kern der diesbezüglichen Aktivitäten ist die Suche nach adäquaten Monitor-Schaltungen und deren Implementierung [174, 150]. In Abschnitt 6.1.3 werden zwei unterschiedliche Konzepte zur Chip-Überwachung vorgestellt und diskutiert.

Aus oben stehenden Gleichungen lässt sich der Einfluss der ansteigenden Laufzeitsensitivität auf die dynamische Energie E_{dyn} wie folgt abschätzen:

$$\Delta V_{DD} = \frac{\Delta t_d^{rel}}{S_{rel}^{V_{DD}}} \tag{6.3}$$

Δt_d^{rel} bezeichnet den Prozentsatz, um den die Geschwindigkeit mittels reduziertem V_{DD} verringert werden kann, so dass die Laufzeiten der Schaltung auf schnellem und langsamen Die gleich sind. $S_{rel}^{V_{DD}}[\%/V]$ ist die relative Laufzeitsensitivität gegenüber V_{DD} Schwankungen. Da schnelle und langsame Dies nun bei gleicher Frequenz laufen ergibt sich unter Vernachlässigung der Kurzschlussströme für den schnellsten Die eine relative Energiereduktion von

$$\Delta E_{dyn}^{rel} \approx 1 - \left(\frac{V_{DD} - |\Delta V_{DD}|}{V_{DD}}\right)^2 = 1 - \left(1 - \frac{\Delta t_d^{rel}}{S_{rel}^{V_{DD}} \cdot V_{DD}}\right)^2 \tag{6.4}$$

Diese Zusammenhänge gelten nur unter der Annahme, dass die Schaltungskapazität unabhängig von der Versorgungsspannung ist und der Kurzschlussstrom nur geringfügig zur Gesamtenergieaufnahme beiträgt. Sie zeigen jedoch den generellen Trend hinsichtlich PVS/AVS. Grundsätzlich führt die mit fortschreitender Technologieskalierung zunehmende Laufzeitsensitivität gegenüber V_{DD} bei konstanten relativen Prozessvariationen (siehe Parameter Δt_d^{rel}) zu verringerter Energieersparnis durch PVS bzw. AVS. Bild 6.3 zeigt die mit obigen Gleichungen abgeschätzte Ersparnis der dynamischen Energiekomponente für die V_{DD} Adaption von schnellem und nominellem Prozess an die Geschwindigkeit des langsamen Prozesses. Die eingesparte Energie durch PVS reduziert sich vom 90nm auf den 40nm Technologieknoten um ca. 50% und beträgt im nominellen Fall ca. 10%.

Im Gegensatz dazu ermöglicht die erhöhte Sensitivität gegenüber V_{DD} die Kompensation von Laufzeiterhöhungen zu geringeren Kosten. Degradationseffekte wie Time-Dependent Dielectric Breakdown (TDDB), NBTI etc. limitieren jedoch aus Zuverlässigkeitsgründen die beliebige Erhöhung der Versorgungsspannung [151]. In 40nm CMOS ist die Kompensation einer 5% igen Laufzeiterhöhung durch eine Versorgungsspannungserhöhung von nur 20mV ($< 2\%$ von V_{DD}, nom) möglich, so dass Zuverlässigkeitsaspekte bei der Kompensation von Laufzeitschwankungen an Bedeutung verlieren.
Bild 6.4 zeigt die Messergebnisse eines ARM926 kritischen Pfades in 65nm und 45nm CMOS Technologie. Für einen nominellen Die wird die normierte Energieaufnahme über der normierten Taktfrequenz dargestellt. Für 45nm nimmt sowohl die durch AVS bzw. PVS gesparte Energieaufnahme als auch die Energieerhöhung für eine zur Kompensation von WID Laufzeitschwankungen benötigte V_{DD} Erhöhung ab. Bei weiter steigenden Laufzeitsensitivitäten ist ein Fortlauf dieses Trends zu erwarten.

6.1.2 Adaptive Body Biasing

Neben der Versorgungsspannung, die zur Adaption von Laufzeitschwankungen verwendet werden kann, ist es möglich, über den Body-Effekt die Transistoreinsatzspannung und somit die Treiberfähigkeit (I_{eff}) und den Unterschwellstrom I_{Sub} des Transistors zu verändern. Durch die kontrollierte Veränderung des Potentialunterschieds zwischen Bulk und Source Kontakt kann die Einsatzspannung V_T des Transistors sowohl erhöht als auch reduziert werden. Die folgenden Gleichungen zur Bestimmung der Einsatzspannung zeigen den Zusammenhang von Einsatzspannung, Body- Effekt und der Spannung an der

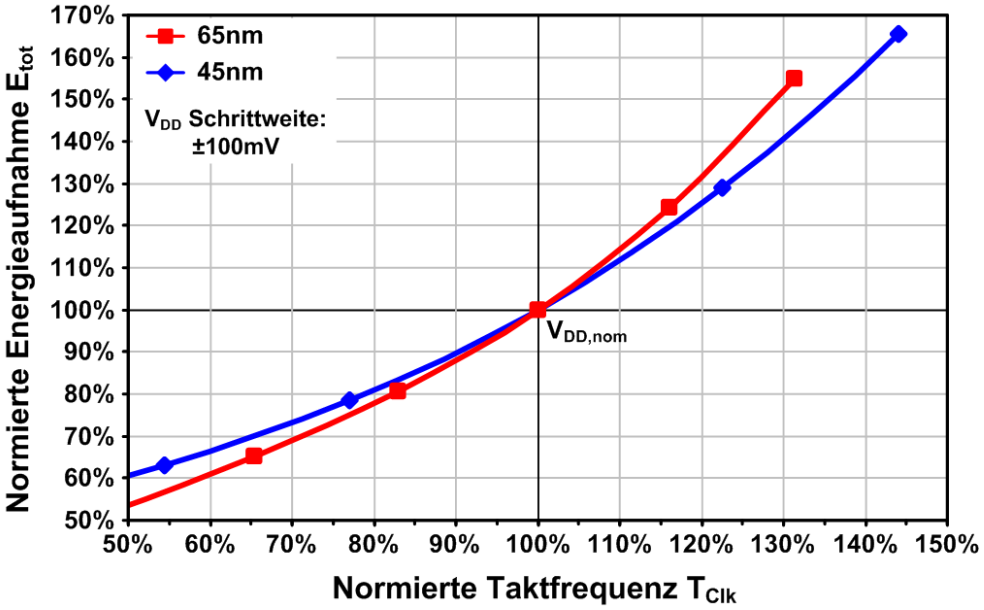


Bild 6.4: Gemessene normierte Taktfrequenz und normierte Energieaufnahme eines ARM926 kritischen Pfades in 65nm und 45nm low-power CMOS bei T=27°C ($\alpha_{Schalt} = 0.1$).

Source-Bulk Diode [178]:

$$V_T = V_{T0} + \Delta V_T \tag{6.5}$$

$$V_{T0} = V_{FB} + 2\phi_F + \gamma\sqrt{2\phi_F} - \frac{Q_S}{C_{ox}} \tag{6.6}$$

$$\Delta V_T = \gamma\left(\sqrt{2\phi_F - V_{BS}} - \sqrt{2\phi_F}\right) \tag{6.7}$$

$$\gamma = \frac{t_{ox}}{\varepsilon_{ox}}\sqrt{2 \cdot e \cdot N_A \varepsilon_{Si}} \tag{6.8}$$

V_{FB} ist die Flachbandspannung, die sich im Wesentlichen aus dem Unterschied der Austrittsarbeit zwischen Gate und Bulk-Material ergibt, γ ist der Body-Koeffizient, Q_S die Ladungen an der Gate-Oxid Grenzfläche C_{ox} die Oxid-Kapazität. Mit ϕ_F wird das Fermi-Potential, mit t_{ox} die Oxid-Dicke, ε_{ox} die Dielektrizitätskonstante des Gate-Oxids und mit N_A die Dotierdichte des Bulk-Materials bezeichnet.
Formel 6.7 zeigt die Abhängigkeit der Einsatzspannungsänderung von der Bulk-Source Spannung V_{BS}. Diese Gleichungen gelten für klassische MOSFET Transistoren und zeigen den generellen Trend des Body-Effektes. Mit fortschreitender Technologieskalierung nimmt der Body-Effekt ab, wie an Gleichung 6.8 zu sehen ist. In modernen sub-100nm Technologien hängt die Größe des Body-Effektes zudem von den verschiedenen Implantationsprofilen (z.B. LDD, Halo, Retrograde Wannendotierung) ab.
Bild 6.5 zeigt die Abnahme des Body-Effekts mit fortschreitender Technologieskalierung. Beim Übergang von 65nm auf 45nm CMOS bleibt der Body-Effekt konstant, so dass bei geringerer Versorgungsspannung und höherer Laufzeitsensitivität gegenüber verändertem V_T mit einem erhöhten Einfluss der Bulk-Source Spannung auf die Laufzeit zu rechnen ist.
Bei Adaptive Body Biasing wird die Bulk-Source Spannung je nach Chip-Zustand angepasst, so dass V_T verändert und die Laufzeit adaptiert wird. Man unterscheidet zwei

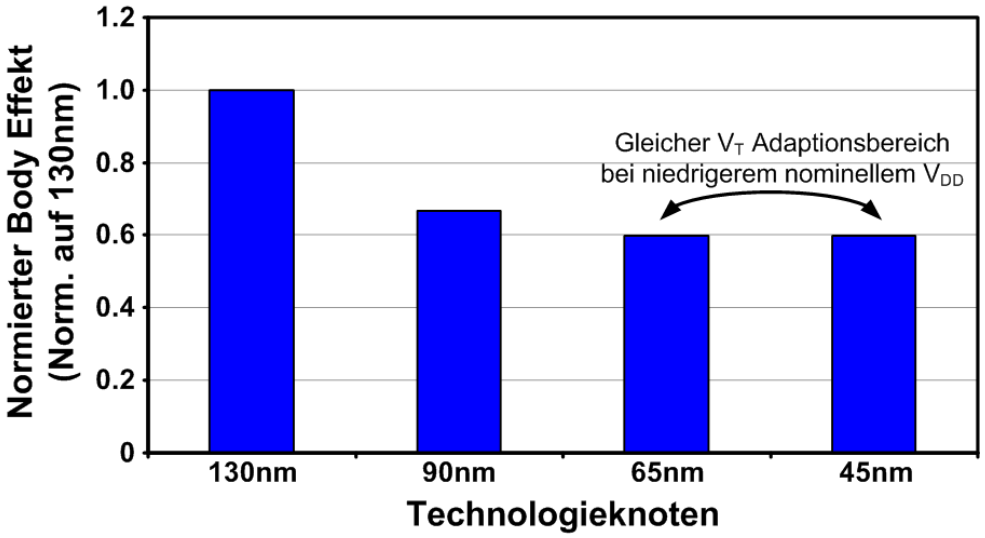


Bild 6.5: Technologietrend: Body-Effekt-bedingtes ΔV_T bei nominellem V_{DD} normiert auf 130nm low-power CMOS.

unterschiedliche Betriebsmodi, die im Folgenden am Beispiel eines NMOS Transistors diskutiert werden:

- **Reverse Body-Biasing (RBB)**

  Der Unterschwellstrom I_{sub} trägt im Vergleich zu Diodenströmen (Sperrströme) und Tunnelströmen am stärksten zum Gesamtleckstrom bei [39]. Wird die Diode zwischen Source und Bulk Kontakt in Sperrrichtung betrieben, d.h für den NMOS Transistor gilt $V_{BS} < 0$, so erhöht sich die Einsatzspannung des Transistors. Sowohl der Unterschwellstrom als auch der effektive Strom I_{eff} des Transistors nehmen ab. RBB führt somit zur Reduzierung von Leckströmen bei gleichzeitiger Verringerung der Transistor-Treiberstärke. Ein im Vergleich zum nominellen Fall erhöhter Leckstrom resultiert vorwiegend aus globalen Prozessschwankungen, wie z.B. kürzere Gateläng en (V_T roll-off) und kleine Einsatzspannungen, so dass die Geschwindigkeit einer Schaltung mit global erhöhtem Leckstrom a priori sehr hoch ist. Aus diesem Grund stellt die Laufzeiterhöhung aufgrund des reduzierten Drainstroms kein wesentliches Problem dar.
  Obwohl die Source-Bulk Diode in Sperrrichtung betrieben wird, darf V_{BS} nicht beliebig klein gewählt werden. Mit sinkendem V_{BS} erhöhen sich Bulk- (Dioden-Sperrstrom) und Tunnelströme [153, 155], so dass der Gesamtleckstrom für ein bestimmtes V_{BS} ein Minimum aufweist. Daraus resultiert eine optimale Bulk-Source Spannung V_{BS} zur Kompensation von Leckstromschwankungen.
  Die mit der Technologieskalierung verstärkt auftretenden Kurzkanaleffekte wie V_T roll-off und DIBL reduzieren den Body-Effekt [154] und führen zu erhöhten V_T-Schwankungen [45]. Zur Reduzierung der Kurzkanaleffekte werden in der Nähe der Source-/Drain-Bulk Dioden schärfere Dotierprofile implantiert, so dass der Beitrag der Tunnelströme zum Gesamtleckstrom ansteigt. Die Leckstromreduzierung durch Reverse Body Biasing nimmt daher mit fortschreitender Technologieskalierung ab [75].
  Während NBTI mit zunehmendem RBB zunimmt, bleibt der Hot Carrier Effekt durch RBB unbeeinträchtigt [152, 156].

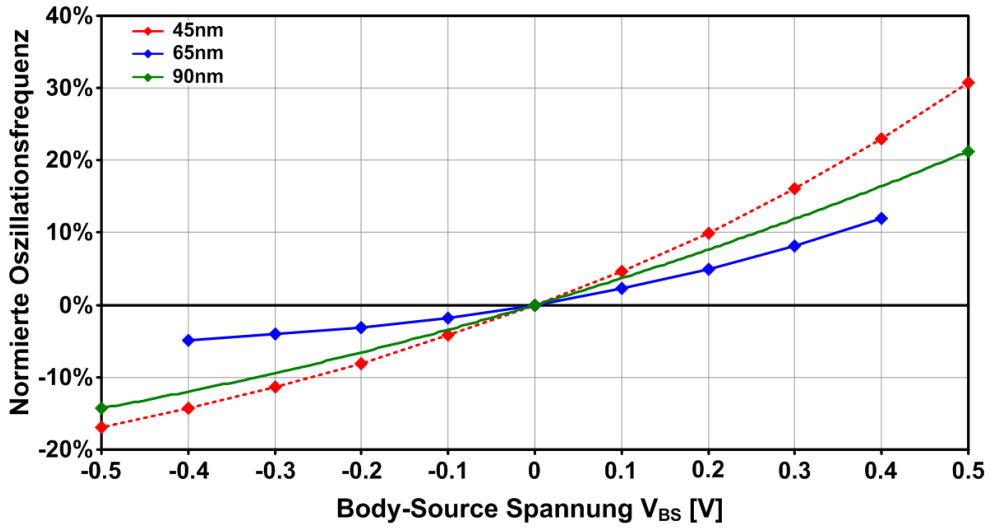


Bild 6.6: Technologietrend: Einfluss der Body-Source Spannung auf die Laufzeit von CMOS Digitalschaltungen ($V_{DD} = 1.2V$, T=27°C).

- **Forward Body-Biasing (FBB)**

  Im Gegensatz zum RBB wird beim FBB die Bulk-Source Diode in Vorwärtsrichtung betrieben, d.h. $V_{BS} > 0$. Die Einsatzspannung wird gesenkt und der Effektivstrom I_{eff} des Transistors erhöht. Eine Erhöhung der Bulk-Source Spannung ist nur bis zu Spannungen von ca. 100-200mV unter der Diodenspannung sinnvoll, da eine Spannungserhöhung über diesen Punkt hinaus zu hohen Strömen vom Bulk zum Source Kontakt führt. Der Betriebsbereich für Forward Body-Biasing liegt daher im Allgemeinen bei $0V < V_{BS} \leq 500mV$. Neben der Geschwindigkeitserhöhung durch FBB reduziert sich der Einfluss von Kurzkanaleffekten sowie die Laufzeitsensitivität gegenüber V_T, L und V_{DD} [75, 155]. Die steigende Laufzeitsensitivität gegenüber diesen Prozess- und Betriebsparametern nimmt mit fortschreitender Technologieskalierung zu, so dass trotz eines abnehmenden Body-Effekts weiterhin deutliche Geschwindigkeitsverbesserungen zur Kompensation von Laufzeiterhöhungen möglich sind.
  Zusätzlich zur Leckstromerhöhung durch den Vorwärtsbetrieb der Bulk-Source Diode werden auch die aktiven Verluste durch Forward Body Biasing beeinflusst. Zum einen führt die Verschmälerung der Raumladungszonen an den Source-/Drain-Kontakten zu erhöhten PN-Kapazitäten, zum anderen erhöht sich gleichzeitig die effektive Gate-Kapazität [155].
  Im Gegensatz zu RBB verbessert FBB die Hot Carrier Zuverlässigkeit des Transistors [156].

Bild 6.6 zeigt für drei verschiedene CMOS Technologien den gemessenen Einfluss der Body-Source Spannung V_{BS} auf die Laufzeit von CMOS Schaltungen. Die Messstruktur besteht in 65nm und 45nm aus einem als Oszillator verschaltenen kritischen Pfad, in 90nm aus einem einfachen Ringoszillator, bestehend aus Inverter Zellen. Während der Einfluss von Body Biasing auf die Pfadlaufzeit für Technologien bis 65nm CMOS abnimmt, zeigen Messungen in 45nm einen Anstieg des Effekts. Die steigende Laufzeitsensitivität gegenüber V_T Schwankungen beeinflusst die Auswirkung von Body Biasing auf die Laufzeit, so dass Body Effekt und Laufzeitsensitivität bei der Analyse von Body Biasing nicht isoliert voneinander betrachtet werden dürfen. Da neben den nominellen Zielwerten bei der Technologieentwicklung weitere Optimierungen z.B. hinsichtlich steigender statistischer

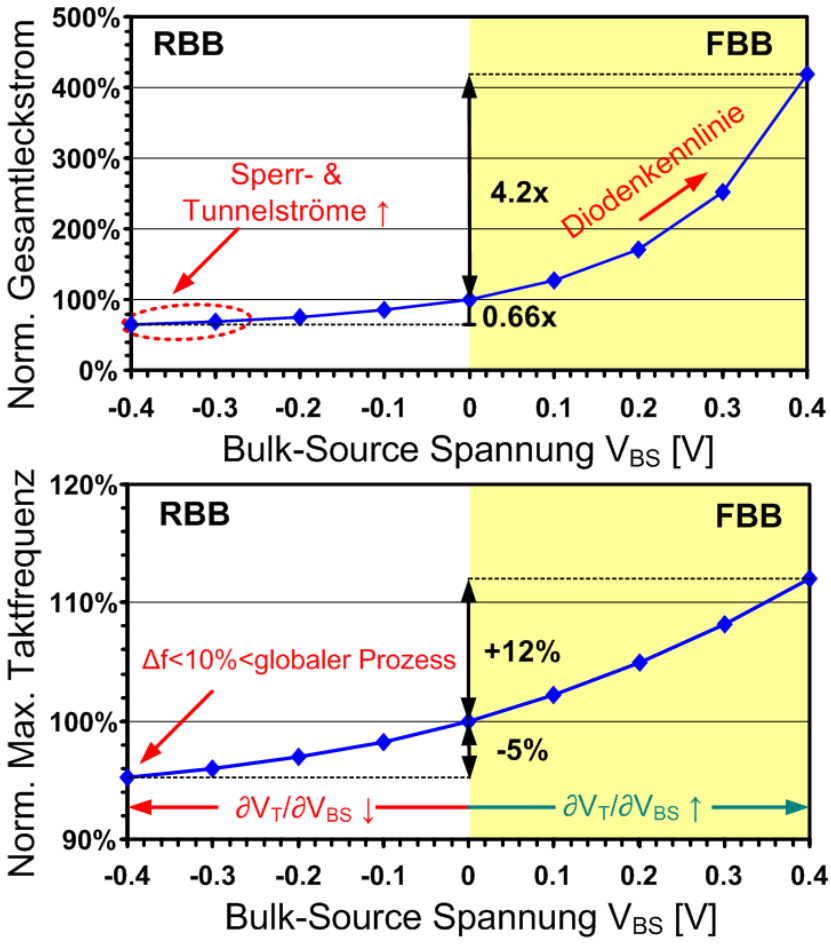


Bild 6.7: Messergebnisse eines ARM926 kritischen Pfades in 65nm low-power CMOS bei symmetrischem Reverse und Forward Body Biasing ($V_{DD} = V_{DD,nom}$, T=27°C)

.

Schwankungen der Dotierstoffatome vorgenommen werden [83, 84], ist eine Pauschalaussage zum Einfluss von Body Biasing auf die Laufzeit moderner CMOS Technologien nicht möglich. Im Folgenden wird der Einfluss von Body Biasing auf Schaltungen in der 65nm low-power CMOS Technologie gezeigt.

Bild 6.7 zeigt Messergebnisse eines geschwindigkeitskritischen ARM926 Pfad in 65nm low-power CMOS. Gemessen wurde der Einfluss von symmetrischem Body Biasing auf den Gesamtleckstrom und die Geschwindigkeit der Schaltung bei nomineller Versorgungsspannung.

Reverse Body Biasing reduziert für $V_{BS} = -0.4V$ den Leckstrom um ein Drittel, bei gleichzeitiger Laufzeiterhöhung von nur ca. 5%. Der Laufzeitunterschied zwischen schnellen und langsamen Chips liegt deutlich über 10%, so dass die Leckstromreduktion durch RBB geschwindigkeitsunkritisch ist. Die Reduktion des Gesamtleckstroms mit zunehmendem RBB nimmt ab, da die Beiträge aus Sperr- und Tunnelströmen zunehmen.

Während die Sensitivität von V_T gegenüber V_{BS} im RBB Betrieb abnimmt, wird sie im FBB Betrieb erhöht. Bei betragsmäßig gleichem Body Biasing von 400mV zeigt FBB mit 12% höherer Taktfrequenz im Vergleich zum RBB einen um den Faktor 2.4 höheren Einfluss auf die Pfad-Laufzeit. Bei 12% Geschwindigkeitsgewinn erhöht sich der Gesamtleckstrom um den Faktor 4.2. Je nach Verhältnis von Leckstrom und Aktivstrom ergibt sich der Einfluss von Body Biasing auf die Energieaufnahme der Schaltung.

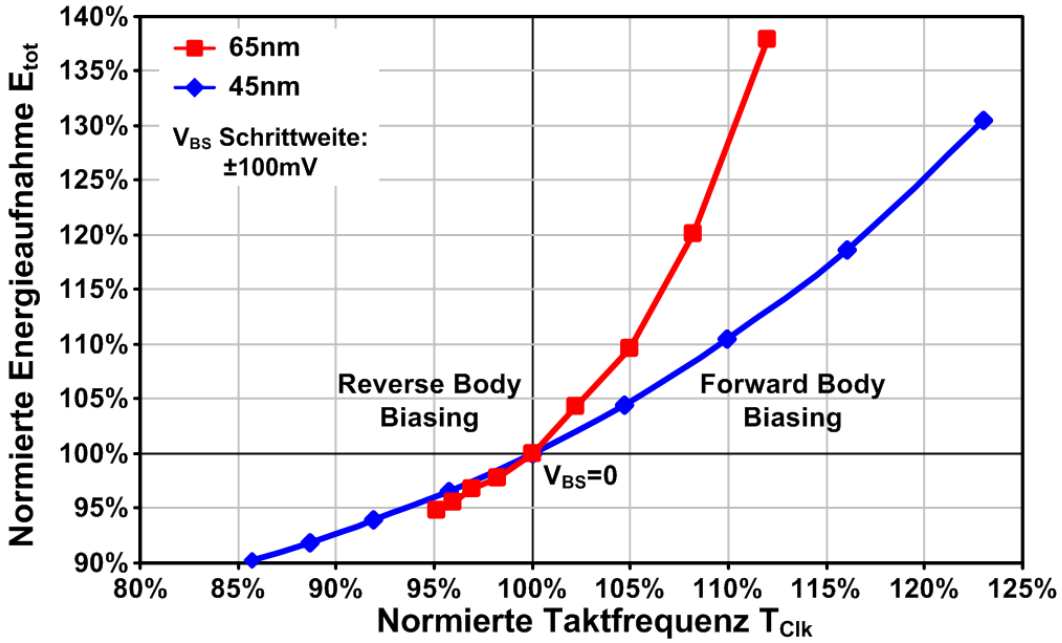


Bild 6.8: Normierte Energieaufnahme und Maximalfrequenz bei ABB eines ARM926 kritischen Pfads auf nominellem Die ($\alpha_{Schalt} = 0.1$).

Je nach Schaltung und Benutzerprofil ermöglicht gleichzeitiges Adaptive Voltage Scaling und Adaptive Body Biasing [157, 158] die Anpassung der Geschwindigkeit zu optimalen Kosten hinsichtlich Leistungsaufnahme, da sowohl dynamische Verluste als auch Verluste durch Leckströme angepasst werden können.
Um den Adaptionsbereich für Laufzeitschwankungen und den dafür erhöhten Energieaufwand quantifizieren zu können, zeigt Bild 6.8 die Messergebnisse eines ARM926 geschwindigkeitskritischen Pfads.
Im Vergleich zu 65nm nimmt in 45nm die Energieersparnis durch RBB zu, während die zusätzlich benötigte Energie zur Kompensation von Laufzeiterhöhungen abnimmt. Für die Adaption einer Laufzeiterhöhung von 10% zeigt sich für 45nm ein nahezu gleich großer Energiebedarf für PVS/AVS und ABB.
Bild 6.9 zeigt für 45nm CMOS den Einfluss von PVS/AVS und symmetrischem ABB, d.h. die Bulk-Source Spannung V_{BS} von PMOS und NMOS Transistoren sind betragsmäßig gleich groß, auf den Energiebedarf der Schaltung. Dazu sind in beiden Grafiken die Bereiche ähnlicher Laufzeitänderung (ca. ±10%) eingezeichnet. Es zeigt sich, dass bei niedrigem Beitrag der Leckströme zur Gesamtenergie kaum ein Unterschied zwischen Voltage Scaling und Reverse Body Biasing hinsichtlich Energieaufnahme zu sehen ist. Für einen höheren Leckstromanteil zeigt Reverse Body Biasing erhöhtes Einsparpotential. Bei größerer Prozessschwankung verringert sich dieses Einsparpotential gegenüber Voltage Scaling, da die Reduktion des Leckstroms mit weiterem RBB abnimmt.
Betrachtet man nun die Beschleunigung der Propagationszeiten durch PVS/AVS bzw. ABB zur Kompensation von variationsbedingten Laufzeiterhöhungen, so zeigt sich ein deutlicherer Unterschied beider Konzepte. Die erforderliche Energieerhöhung ist bei PVS/AVS stets geringer als für ABB. Mit steigendem Leckstromanteil nimmt dieser Unterschied weiter zu, da mit erhöhtem Leckstromanteil für ABB der starke Anstieg der Leckströme die Gesamtenergieaufnahme drastisch erhöht.
Auch wenn in diesem Beispiel Messungen von Schaltungen auf nominellem Die als Da-

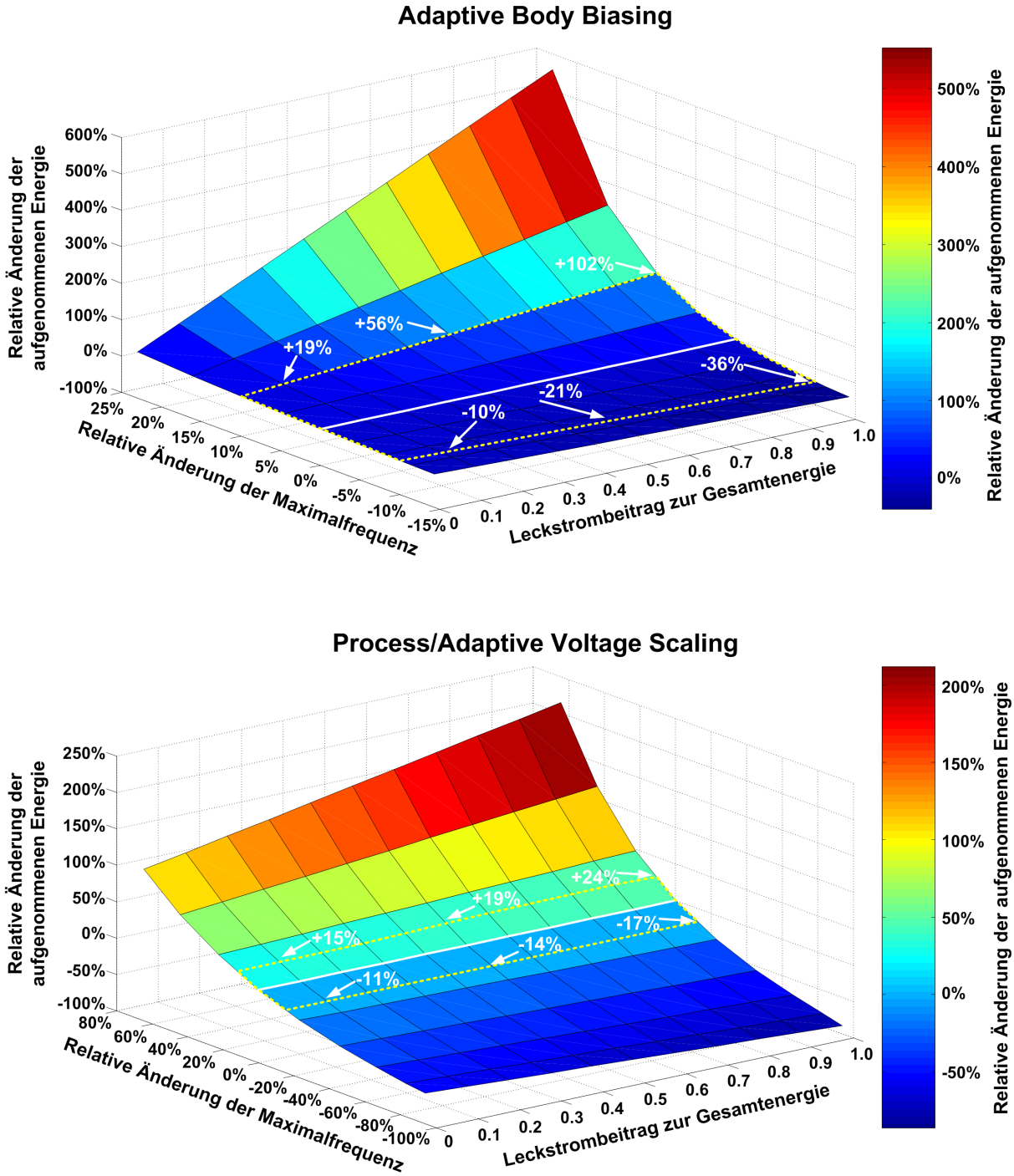


Bild 6.9: Vergleich der Messergebnisse von PVS/AVS und ABB hinsichtlich Energieerhöhung zur Adaption der Pfadlaufzeiten in 45nm CMOS (nomineller Die, 27°C, $V_{DD} = V_{DD}^{nom}$).

tenbasis dienen, können diese Ergebnisse zur Bewertung von PVS/AVS bzw. ABB herangezogen werden. Da schnelle Schaltungen geringere Einsatzspannungen aufweisen, nimmt der Body Effekt ab. Dadurch kann die hier getroffene Annahme als optimistisch hinsichtlich Energiereduzierung gesehen werden. Da der Leckstromanteil schneller Schaltungen an der Gesamtenergieaufnahme steigt, muss lediglich der X-Wert (Leckstrombeitrag) in der Grafik gewechselt werden, um die zu erwartende Änderung in der Energieaufnahme ablesen zu können. Der leichte Optimismus wird dadurch jedoch nicht korrigiert.
Die Implementierung von Adaptive Body Biasing erfordert zusätzliche Spannungspegel, um die Body-Potentiale von NMOS und PMOS Transistoren zu kontrollieren [159]. Zur Verteilung dieser Spannungen sind eigene Versorgungsleitungen notwendig, was im Standardzellen-Layout die Anzahl der Routing-Tracks reduziert. Im Allgemeinen führt dies zu einer erhöhten Anzahl an Filler-Zellen, um zusätzlichen Platz für die Verdrahtung zur Verfügung zu stellen. Der für Adaptive Body Biasing Techniken erforderliche zusätzliche Flächenaufwand ergibt sich somit aus zusätzlichem Bedarf an Verdrahtungsflächen, Kontrolllogik sowie den Spannungsgeneratoren für die Body-Potentiale [45, 157].

6.1.3 On-Chip Monitorschaltungen

Mit zunehmender Technologieskalierung und tieferem Pipelining von Mikroprozessoren steigt der Beitrag von WID Prozess- und Umgebungsvariationen zur Laufzeitschwankung einer Schaltung [80, 96, 122]. Da der große Anteil der Umgebungsvariationen von den Betriebsbedingungen und somit vom Benutzerprofil abhängt, ist die Modellierung dieser Effekte kaum möglich. Aus diesem Grund werden Monitorschaltungen eingesetzt, die den Zustand der Schaltung kontrollieren und daher als Sensor für adaptive Techniken dienen. Da die Adaption an Laufzeitschwankungen nur für statische Prozessvariationen und betriebsabhängige Temperaturschwankungen sowie Alterungseffekte möglich ist, beschränkt sich die folgende Übersicht über bestehende Monitorkonzepte zur Bestimmung der on-chip Variabilität auf diese Effekte.

Replika basierte Monitorschaltungen:

Die meisten on-chip Monitorkonzepte haben zum Ziel, mögliche funktionale Fehler zu erkennen bzw. vorherzusagen bevor diese eintreten. Um dies zu gewährleisten wird für eine Monitorschaltung die Einhaltung der Timing Bedingungen während des Betriebs überwacht.

Während für hohe Spannungen einfache Ringoszillatoren (Bild 6.10a) aus Inverterzellen nur geringe Abweichungen gegenüber anderen Gatter- und Pfadstrukturen hinsichtlich ihres Spannungsverhaltens zeigen, sind für größere Spannungs- und Temperaturbereiche komplexere Monitorschaltungen notwendig, um das Verhalten von kritischen Pfaden nachzubilden [179]. Hier gibt es verschiedene Ansätze wie z.B. die Verwendung von kritischen Pfaden oder die generische Nachbildung des Verhaltens von kritischen Pfaden durch Verwendung spezifischer Gattertopologien [180].
Einige dieser Methoden messen direkt die Laufzeit der Signale mittels Time-to-Digital Convertern (TDC) und adaptieren je nach gemessenem Wert die Betriebsparameter der Schaltung [181, 182]. Das Grundprinzip ist in Bild 6.10b dargestellt. Da in diesen Konzepten keine Fehlerkorrektur vorgesehen ist, müssen Fehler vorhergesagt werden, bevor sie auftreten. In [179] werden Setup-Zeit Fehler der Pfad-Replika mittels Error Detection Se-

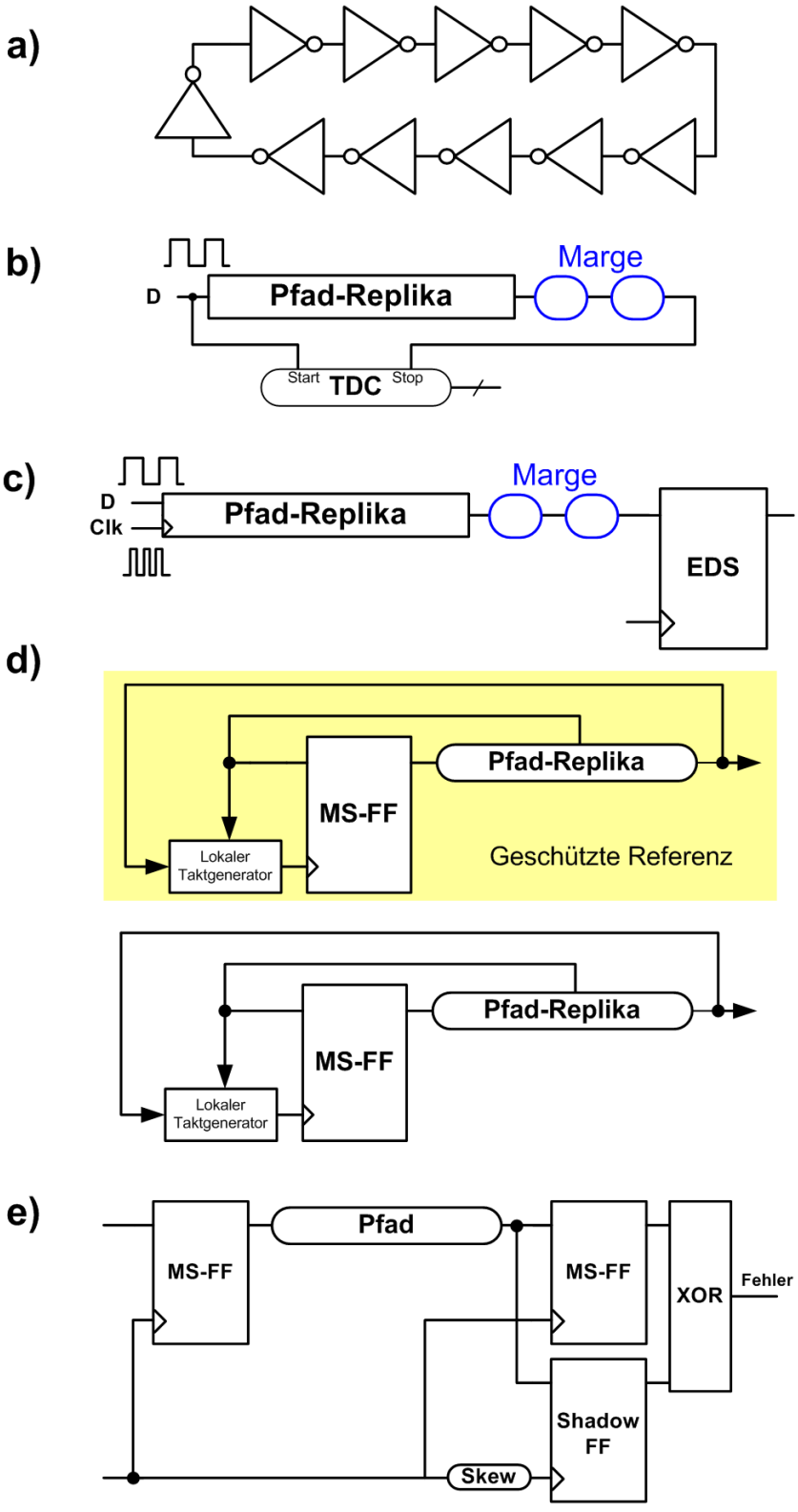


Bild 6.10: Grundprinzip verschiedener Monitorkonzepte zur Überwachung des Schaltungszustands.

quentials (EDS) detektiert (Bild 6.10c). Um das Auftreten von Fehlern in der eigentlichen Schaltung zu vermeiden werden zusätzliche Laufzeit-Blöcke in den Monitor eingefügt, die eine zeitliche Marge zwischen der zu überwachenden Struktur und den realen kritischen Pfaden der Schaltung bilden.
Der Bestimmung der absoluten Laufzeit eines kritischen Pfades stehen Konzepte gegenüber, die relative Laufzeitänderungen erfassen [183]. Das Grundprinzip ist in Bild 6.10d dargestellt. Diese Konzepte haben vergleichsweise hohe Messzeiten, ermöglichen jedoch die Messung geringer Laufzeitunterschiede zu deutlich geringerem Aufwand. Die Messzeiten sind zu hoch, um Zyklus-zu-Zyklus Variationen wie IR-Drop und Clock-Jitter zu detektieren, so dass diese Konzepte nur für Variationseffekte mit relativ großen Zeitkonstanten wie Temperatur und Alterungseffekten geeignet sind. Durch die Verteilung mehrerer Oszillatoren, bestehend aus einfachen Invertern oder komplexeren kritischen Pfaden über die gesamte Schaltung, wird die Messung von WID Variationen ermöglicht. Binär-Zähler bestimmen die Anzahl der Oszillationen für ein global vorgegebenes Zeitintervall. Die Anzahl der zu durchlaufenden Zyklen n_{Zyk} und die daraus resultierende Breite des Zeitintervalls ergibt sich für einen maximalen relativen LSB-Fehler e_{LSB}, wie folgt:

$$n_{Zyk} = \frac{1}{e_{LSB}} \tag{6.9}$$

Für einen LSB-Fehler von 0.25% sind somit 400 Zyklen Messzeit erforderlich. Bei einer nominellen Zykluszeit von z.B. 4ns ergibt sich somit eine Messzeit von $1.6\mu s$.
Die Messung von Alterungseffekten bei Verwendung einzelner Teststrukturen erfordert zusätzliche, nicht flüchtige Speicher, bzw. den Zugriff auf bestehende Speicher, um die Messwerte früherer Messungen protokollieren zu können. Eine Kombination aus geschützter Referenzstruktur und alternder Messstruktur ermöglicht die Messung von Alterungseffekten ohne Speicherelemente [183]. Dazu wird die Referenzstruktur nur im Messbetrieb ans Versorgungsnetz angeschlossen, so dass während des Normalbetriebs keine Alterung stattfindet. Die zu überwachende Teststruktur wird lokal ans Versorgungsnetz und ans Taktsignal angeschlossen, so dass diese Struktur ähnlich wie die realen kritischen Pfadstrukturen altert.
Für diese Anordnung ergeben sich im Vergleich zur Messung nur einer Teststruktur folgende Probleme. Lokale, statistische Variationen verursachen Offsetfehler zwischen zu messender Test- und Referenzstruktur, so dass sich die Anzahl der Oszillationen im vorgegebenen Zeitintervall von Die zu Die unterscheiden und die Differenz sowohl positiv als negativ sein kann. Dieser Offsetfehler variiert je nach Implementierung der Teststruktur (Anzahl und Größe der Gatter im Pfad), kann bei Bedarf jedoch über eine Offsetkorrektur justiert werden. Durch die Differenzmessung erhöht sich zusätzlich die Messzeit, da sich bei Vorgabe der Messgenauigkeit a_{Mess} und einem maximalen LSB-Fehler e_{LSB} folgende Bedingung für die Anzahl zu zählender Oszillationen n ergibt:

$$\begin{aligned} n_{Ref} - n_{Stress} &\geq \frac{1}{e_{LSB}} \\ (1 + a_{Mess}) \cdot n_{Stress} - n_{Stress} &\geq \frac{1}{e_{LSB}} \\ n_{Stress} &\geq \frac{1}{a_{Mess} \cdot e_{LSB}} \end{aligned} \tag{6.10}$$

Mit n_{Stress} bzw. n_{Ref} wird die Anzahl der jeweiligen Oszillationen von gestresster und Referenzstruktur bezeichnet. Für eine nominelle Zykluszeit von 4ns, einen LSB-Fehler von

0.25% und eine Messgenauigkeit von 1% ergibt sich eine Messzeit von 160μs.

Vor- und Nachteile Replika basierter Monitorschaltungen:

+ Kein Eingriff in die Mikroarchitektur, da Replika von kritischen Strukturen verwendet werden
+ Fehlervermeidung anstatt Fehlerkorrektur
+ Einfaches Konzept, das schnell auf andere Schaltungen übertragen werden kann
+ 'Kontinuierliche' Bestimmung anstatt binärer Bewertung (pass/fail) des Chip-Zustands
- Gute Übereinstimmung von realen kritischen Pfaden und im Schaltungsentwurf identifizierten kritischen Pfaden erforderlich
- Berücksichtigung von Variationen im Taktnetz nur über Margen möglich
- Prädiktives Konzept, daher keine Information über reale Fehlerrate

Fehlerdetektion: Beispiel Razor

Razor ist eine AVS-Strategie, die auf Basis der zu bestimmenden Fehlerrate die Versorgungsspannung einer Schaltung adaptiert [171, 184]. Als Fehler wird dabei jede Setup-Zeit Verletzung in den kritischen Pfaden bezeichnet, die über ein zusätzliches Latch mit verzögertem Taktsignal detektiert wird. Bild 6.10e zeigt das Grundprinzip der Fehlerdetektion, wie sie auch in einem Replika-basierten Monitorkonzept verwendet wird (siehe 6.10c). Der Vergleich der gespeicherten Logikwerte in regulärem Flip Flop und zusätzlichem Shadow Latch erzeugt für unterschiedliche Logikwerte ein Fehlersignal. Tritt ein Fehler auf, so wird die Verteilung des Taktsignals für die gesamte Schaltung unterbrochen, zum anderen der fehlerhafte Wert im regulären Flip Flop durch den Wert des Shadow Latchs ersetzt. Es werden ein bis mehrere Taktzyklen gewartet, bevor das Taktsignal wieder freigegeben wird. Eine andere Möglichkeit ist es, nach Fehlerdetektion die gesamte Pipeline zu löschen und eine erneute Berechnung zu starten. Die Fehlerkorrektur muss nun nicht mehr im Flip Flop realisiert werden, so dass sich der Flächenbedarf reduziert [172].

Vor- und Nachteile des Razor-Konzepts:

+ Die Messung erfasst alle Variationseffekte. Neben den global wirkenden Variationen werden auch lokale Variationen berücksichtigt
+ Es handelt sich um eine in-situ Messung, d.h. die zu überwachende Schaltung dient gleichzeitig als Sensor, so dass auch Variationen aus dem Taktnetz wie z.B. Clock Jitter bei der Bewertung miteinfließen
+ Schaltungsspezifische Optimierung der dynamischen Verlustleistung
+ Information über reale Fehlerraten
- Es müssen alle Fehler-Signale miteinander verknüpft werden, so dass innerhalb eines Zyklus die globale Verteilung des Taktsignals gestoppt werden kann. Mit steigender Anzahl an kritischen Pfaden mit tieferem Pipelining erhöht sich der Aufwand bei gleichzeitig erhöhten Geschwindigkeitsanforderungen (geringere Taktzyklen). In

[184] ist das Razor Prinzip für eine relativ geringe Taktfrequenz von nur 120MHz gezeigt.

- Deutlicher, zusätzlicher Flächenbedarf für steigende Anzahl von kritischen Pfaden.

- Für geringen zusätzlichen Flächenbedarf kann nur ein kleiner Teil der Schaltung überwacht werden, so dass erneut eine gute Übereinstimmung von realen Pfaden und den im Schaltungsentwurf identifizierten kritischen Pfaden erforderlich ist

- Fehlerkorrektur anstatt Fehlervermeidung

- Bei Deaktivierung des Taktsignals an der Einspeisequelle kann es bei der erneuten Freigabe des Taktsignals durch stark erhöhte Stromflüsse zu Versorgungsspannungsschwankungen kommen. Eine dezentrale Abschaltung des Taktsignals ist mit deutlich mehr Aufwand verbunden und daher geschwindigkeitskritischer.

6.1.4 Vergleich der Techniken

Globale Designtechniken wie Adaptive Voltage Scaling und Adaptive Body Biasing ermöglichen es D2D Laufzeitvariationen zu kompensieren. Die Implementierung von AVS ist im Vergleich zu ABB ohne erheblichen zusätzlichen Aufwand möglich, da in den meisten low-power Produkten Dynamic Voltage Scaling als standardmäßige low-power Option integriert ist. Für ABB hingegen ist neben mindestens einer zusätzlichen Versorgungsspannung weiterer Flächenbedarf für die Verteilung der Body-Potentiale und zusätzlichen Verdrahtungsmöglichkeiten notwendig, so dass eine deutliche Erhöhung der Chipfläche zu erwarten ist. Da die Leckstromreduktion durch Reverse Body Biasing mit fortschreitender Technologieskalierung im Allgemeinen abnimmt, ist die Implementierung von Adaptive Voltage Scaling für low-power Schaltungen mit DVS und Power-Switch Konzept attraktiver als Adaptive Body Biasing. Untersuchungen in 45nm low-power CMOS zeigen, dass der Energieaufwand für die Adaption von Laufzeitschwankungen für PVS/AVS und ABB annähernd gleich groß sind, so dass der für ABB erforderliche zusätzliche Aufwand nicht kosteneffizient ist. Für geschwindigkeitsoptimierte Schaltungen, die sehr stringente Geschwindigkeitsanforderungen aufweisen, können durch Forward Body Biasing zusätzliche Geschwindigkeitsmargen generiert werden, ohne die Zuverlässigkeitsanforderungen der Schaltung zu verschärfen. Hier steigt die zusätzliche Energieaufnahme im Vergleich zu AVS jedoch signifikant an. Im Gegensatz zu AVS birgt ABB die Gefahr, während des Betriebs der Schaltung zusätzliche Timing Unsicherheit zu generieren. Versorgungsspannungsschwankungen am Source-Knoten der Transistoren verursachen Schwankungen der Body-Source Spannung V_{BS}. Dabei gilt, je höher der Body-Effekt und je höher die Laufzeitsensitivität gegenüber V_T, desto höher auch die potentiellen Laufzeitschwankungen durch Body-Biasing. In sub-100nm CMOS Technologien ist aufgrund verschiedener Optimierungskriterien eine Pauschalaussage zum weiteren Trend des Body Effektes nicht möglich. Auch, wenn Messungen in 45nm CMOS einen gegenüber 65nm deutlich erhöhten Einfluss der Body-Spannung auf die Laufzeit kritischer Pfade zeigen, kann für 32nm keine verlässliche Vorhersage getroffen werden.
Die Kompensation von WID Variationen durch AVS und ABB ist, wie in [159] gezeigt wird, möglich, jedoch für Semicustom Schaltungen und Systems on Chip nur bedingt anwendbar. Zum einen muss die Gesamtschaltung in mehrere Einheiten unterteilt werden, die eine eigene Versorgungsspannung bzw. eigene Body-Potentiale erhalten. Für jede

dieser Einheiten wird neben den eigenen Spannungsquellen auch ein eigener Monitor benötigt, um die adäquaten Spannungspegel zu bestimmen. Die Unterteilung der Schaltung in kritische und nicht kritische Bereiche ist nur für Schaltungen möglich, die eine dominante, geschwindigkeitskritische Funktionseinheit besitzen, die sich von allen anderen Schaltungsblöcken absetzt. In Multi-Core Prozessoren können die einzelnen Prozessorkerne als Funktionseinheit gesehen werden.
Die durchgeführten Strukturanalysen der verschiedenen ARM Mikroprozessoren zeigen jedoch, dass kritische Strukturen über alle Pipelinestufen hinweg sowohl im Datenpfad als auch in der Kontrolllogik zu finden sind. Eine a-priori Selektion von kritischen und unkritischen Bereichen ist daher nur in Einzelfällen möglich.
Der zusätzliche Aufwand für die Adaption steigt erheblich mit der gewählten Granularität für die Aufteilung in verschiedene Schaltungsbereiche, da jeder Teilbereich eine eigene Versorgung mit V_{DD} bzw. V_T und eine eigene Monitorschaltung erfordert. Globale nicht-adaptive Konzepte (ohne Monitorschaltungen) zur Kompensation von WID Laufzeitschwankungen wie in [185] über die lokale Erzeugung der V_T Pegel sind zwar flächeneffizienter, benötigen jedoch einen zusätzlichen Test der Schaltung und die Fixierung der V_T-Pegel. Die Aufteilung der Schaltung in verschiedene Voltage Islands (VI) ist aufgrund der komplexen Schaltungsstrukturen nur für eine geringe Anzahl von VIs zu vertretbaren Kosten möglich [186]. Bei kleiner Anzahl von VIs ist jedoch die klare Aufteilung von kritischen und sub-kritischen Schaltungsbereichen Voraussetzung, so dass dieses Konzept auch nur bedingt angewandt werden kann. Auch hier ist ein komplexes Testkonzept notwendig, das neben zusätzlichen Testressourcen weiteren Flächenaufwand generiert.

Daher eignen sich adaptive Maßnahmen unter Berücksichtigung von Kosten und Nutzen vorwiegend zur Kompensation von D2D Prozessvariationen und dynamischen Umgebungsvariationen bzw. Alterungseffekten mit großen Zeitkonstanten.
Neben der Steuerung der zu adaptierenden Größen ist die zuverlässige Überwachung des Chips durch repräsentative Monitorschaltungen die größte Herausforderung um Timing Unsicherheiten zu kompensieren, ohne zusätzlich Unsicherheiten zu generieren. Unabhängig davon, ob Replika von kritischen Pfaden oder Fehlerdetektion, wie z.B. Razor, als Monitor genutzt wird, ist eine zuverlässige Identifikation der kritischen Strukturen erforderlich, um den zusätzlichen Flächenaufwand für die Überwachung der Schaltung gering zu halten.

6.2 Präventive Kompensationstechniken

6.2.1 Long-L_{Poly} Design

Die erhöhte Laufzeitsensitivität moderner CMOS Technologien gegenüber Versorgungsspannungsschwankungen ermöglicht es, kleine Laufzeitunterschiede durch vergleichsweise geringe Erhöhung der Versorgungsspannung zu kompensieren. Dies ist nur möglich, wenn keine Zuverlässigkeitsaspekte wie z.B. Gate-Oxid Breakdown, NBTI und Hot Carrier Injection einer Erhöhung entgegenstehen.
Im Folgenden wird der globale Einsatz von Standardzellen diskutiert, die aus Transistoren mit erhöhter Gatelänge $L_{nom} + \Delta L$ bestehen (Long-L_{Poly}). Die Verwendung von erhöhten Gatelängen wurden in [163] zum ersten Mal vorgeschlagen, um die Leistungsaufnahme einer Schaltung zu optimieren. Im Gegensatz dazu werden im Folgenden nur geringe Er-

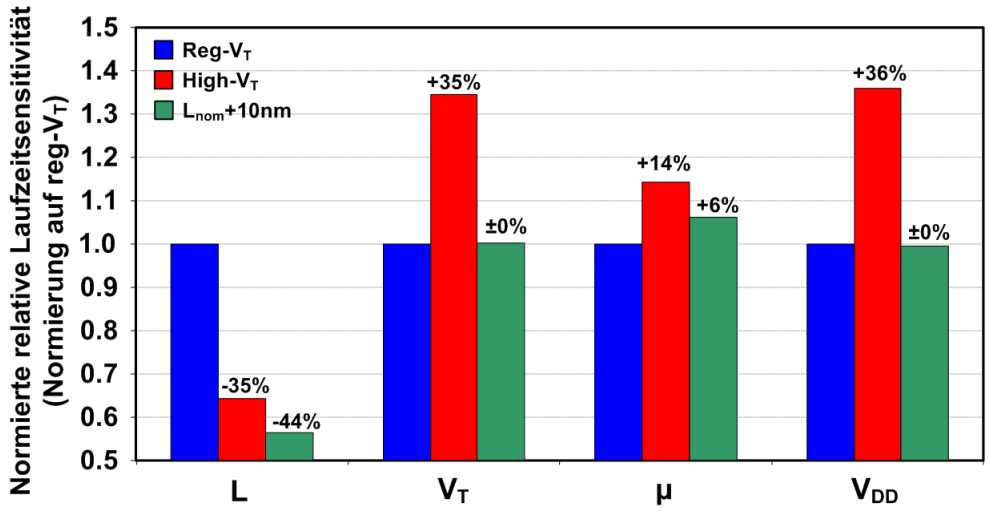


Bild 6.11: Normierte Laufzeitsensitivitäten von Reg-V_T, High-V_T und Long-L_{Poly} NAND2-NOR2 Pfaden in 40nm CMOS Technologie ($V_{DD} = V_{DD,nom}, T = 27°C$).

höhungen der Gatelänge betrachtet, so dass die Größe der Standardzellen unverändert und Pin-kompatibel bleibt. Bisher wurden erhöhte Gatelängen nur selektiv zur Leckstromreduktion eingesetzt [164, 165], da high-V_T Zellen aufgrund des mit fortschreitender Technologieskalierung weiter sinkenden Gate-Overdrives $V_{DD} - V_T$ deutlich sensitiver gegenüber PVT Variationen reagieren. Die folgenden Untersuchungen, basierend auf Simulationen extrahierter Netzlisten, zeigen den kosteneffizienten Einsatz von Transistoren mit erhöhter Gatelänge in 40 nm low-power CMOS. Die Gatelängen werden um $\Delta L = 10nm$ erhöht. Eine Vergrößerung der Standardzellen ist nicht erforderlich. Ziel ist es, alle Zellen einer Schaltung durch Zellen mit erhöhter Gatelänge zu ersetzen, um die Schwankungsbreite der PVT induzierten Laufzeitvariation ohne Einbußen in Geschwindigkeit, Fläche und Leistungsaufnahme zu ermöglichen. Als Testschaltung wird eine NAND2-NOR2 Kette verwendet, die wie in Kapitel 4.2 gezeigt, das Verhalten von geschwindigkeitskritischen Pfaden gegenüber PVT Variationen sehr gut nachbildet.

Bild 6.11 zeigt die veränderten Laufzeitsensitivitäten gegenüber Gatelängen-, Einsatzspannungs-, Beweglichkeits- und Versorgungsspannungsschwankungen einer NAND2-NOR2 Kette für die Implementierung mit nominellen reg-V_T Zellen, high-V_T Zellen und Zellen mit erhöhter Gatelänge. Eine deutliche Reduzierung der Sensitivität gegenüber Gatelängenschwankungen um 44% resultiert aus dem geringeren V_T roll-off bei erhöhter Gatelänge. Die Sensitivität gegenüber der Transistoreinsatzspannung und V_{DD} bleibt unverändert. Obwohl die Implementierung mit high-V_T Zellen eine deutliche Reduzierung der Laufzeitsensitivität gegenüber Gatelängenschwankungen zeigt, stellen die erhöhten Sensitivitäten gegenüber V_{DD} und V_T insbesondere für low-power Schaltungen mit DVS ein großes Problem dar, so dass ein Einsatz von high-V_T Zellen über den 40nm Knoten hinaus fraglich ist. Die Erhöhung der Gatelänge hingegen führt zu deutlich reduzierter Sensitivität gegenüber L bei gleichzeitig unveränderten Sensitivitäten gegenüber V_{DD} und V_T. Dies hat eine deutlich geringere Laufzeitschwankung aufgrund globaler Prozessvariationen zur Folge.
Nimmt man die aus Kapitel 3.2 bekannten Anteile an der Laufzeitschwankung in 65nm (nominelle Versorgungsspannung) als Basis für die folgende Bewertung, so führt der Ein-

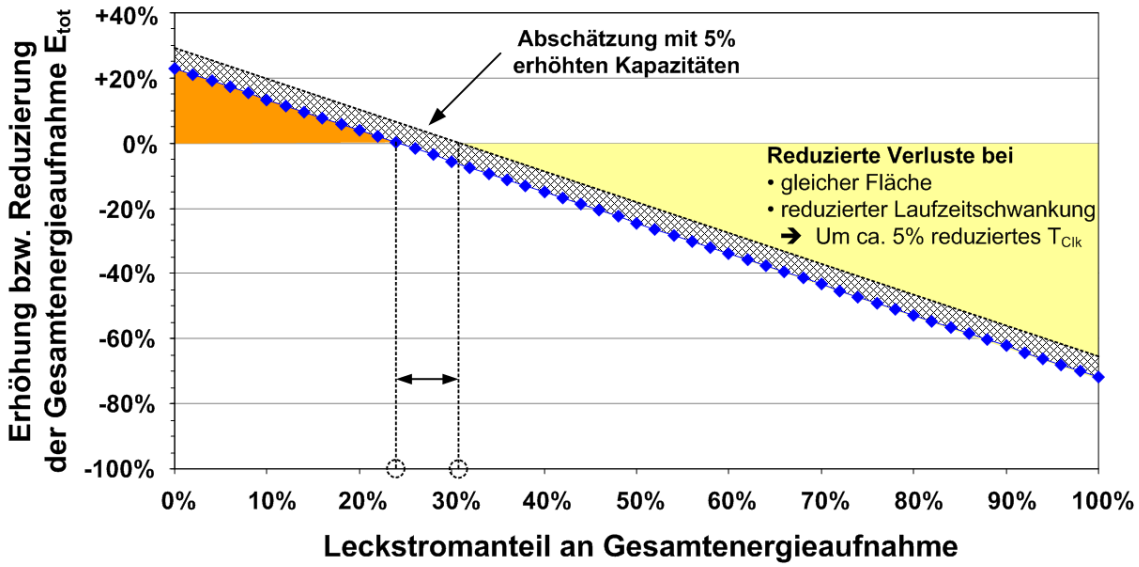


Bild 6.12: Simulierte Erhöhung bzw. Abnahme der aufgenommen Energie in Abhängigkeit des Leckstromanteils an der Gesamtenergieaufnahme in 40nm CMOS bei gleicher Geschwindigkeit($V_{DD} = V_{DD}^{nom} + 60mV$, T=85°C).

satz von erhöhten Gatelängen in 40nm CMOS zu einer Reduzierung der globalen Laufzeitschwankung aufgrund von Prozessvariationen um 30%, was in etwa 5% der Taktperiode entspricht.

Die Verwendung von Transistoren mit erhöhter Gatelänge (Long L_{Poly}) zur Reduzierung der Laufzeitsensitivitäten digitaler Schaltungen ist neu. Es muss jedoch berücksichtigt werden, dass die Verwendung von Long L_{Poly} Transistoren zu erhöhten Laufzeiten führt. Zur Kompensation der erhöhten Laufzeit wird die Versorgungsspannung leicht erhöht. Gegenüber der Implementierung mit nomineller Gatelänge ist beim Betrieb mit nomineller Versorgungsspannung bei einem langsamen Prozess und 85°C eine um 60mV erhöhte Versorgungsspannung für die Implementierung mit Long L_{Poly} Transistoren erforderlich. Neben den erhöhten Gatekapazitäten führt die erhöhte Versorgungsspannung zu erhöhten dynamischen Energieverlusten. Gleichzeitig reduziert sich der Leckstrom der Schaltung durch den verringerten Beitrag des Unterschwellstroms zum Gesamtleckstrom, so dass bei einem bestimmten Verhältnis der Beiträge von Leckstrom- und dynamischen Verlusten weder Energie eingespart noch zusätzliche Energie verbraucht wird. Bild 6.12 zeigt die veränderte Energieaufnahme für verschiedene Leckstrombeiträge zur Gesamtenergieaufnahme.

Für einen Leckstromanteil von 24% ist die Energieaufnahme beider Implementierungen gleich, d.h. es wird weder Energie eingespart noch zusätzliche Energie benötigt. Für höhere Leckstromanteile ist bei gleichbleibender Geschwindigkeit und Fläche neben einer deutlich reduzierten Laufzeitschwankung auch die Reduzierung der Energieaufnahme möglich. Zusätzlich ergibt sich bei langsamem Prozess und 60mV erhöhter Versorgungsspannung eine leicht reduzierte Laufzeitschwankung gegenüber V_{DD} und V_T Schwankungen, so dass sich die in Bild 6.11 gezeigte Reduzierung der Sensitivität weiter erhöht. Neben den Simulationsergebnissen ist auch eine Abschätzung der Energieaufnahme für eine um 5% erhöhte Lastkapazität, die aufgrund der erhöhten kapazitiven Kopplung nach der Gatelängenerhöhung ansteigt, eingezeichnet. In diesem Fall ist eine reduzierte Energieaufnahme für einen Leckstrombeitrag größer als 30% zu erwarten.

Ist die Erhöhung der Versorgungsspannung um 60mV aus dem nominellen Betriebspunkt möglich, so bietet der globale Einsatz von Long L_{Poly} Transistoren die Möglichkeit, gleichzeitig prozessbedingte Laufzeitschwankungen und die Energieaufnahme zu reduzieren.
Die Simulationsergebnisse zeigen vielversprechende Ergebnisse, die jedoch noch auf Schaltungsebene unter Verwendung einer Long L_{Poly} Standardzellenbibliothek verifiziert werden müssen.

6.2.2 Selektiver Einsatz von low-V_T Zellen im Taktbaum

Wie im vorherigen Abschnitt gezeigt wurde, ermöglicht die Verwendung von Long-L_{Poly} Zellen die Reduzierung der Laufzeitsensitivität gegenüber einzelnen Prozess- und Betriebsparametern. Da die Laufzeitsensitivität mit sinkendem Gate-Overdrive $V_{DD} - V_T$ zunimmt, führt die Erhöhung von V_{DD} bzw. die Verringerung von V_T im Allgemeinen zu verringerten Laufzeitsensitivitäten. Da zur Implementierung von Design-Margen nur die Sensitivität bzw. die Laufzeit geschwindigkeitskritischer Schaltungsteile verbessert werden müssen, ist eine globale Erhöhung der Versorgungsspannung nicht sinnvoll. Der selektive Einsatz von Gattern mit reduzierter Transistoreinsatzspannung (low-V_T Gatter) ermöglicht es, punktuell Laufzeiten und Sensitivitäten anzupassen.
Mixed-V_T Design, d.h. die gleichzeitige Verwendung von Transistoren mit unterschiedlichen V_Ts, ist eine seit langem bekannte Technik, die es ermöglicht, die Verluste durch Leckströme zu reduzieren und gleichzeitig die Geschwindigkeit der Schaltung zu erhöhen [160, 162]. Der Einsatz dieser Gatter beschränkt sich bisher auf die geschwindigkeitskritischen Logikpfade. Hier wird ein großer Teil aller Logikgatter von ca. 30-40% der High-V_T bzw. Reg-V_T Transistoren im Design durch low-V_T Transistoren ersetzt [161] und somit die Laufzeit des zeitlich längsten Logikpfades reduziert [187].

Im Gegensatz zu bisherigen Ansätzen wird in diesem Abschnitt der Einsatz von low-V_T Zellen zur Reduzierung der Laufzeitsensitivität einer getakteten Digitalschaltung diskutiert.
Bild 6.13 zeigt für 90nm, 65nm und 40nm low-power CMOS Technologien die auf reg-V_T Gatter normierte Laufzeitsensitivität einer FO4 Inverter-Kette gegenüber Prozess- und Betriebsparametern sowie die prozentuale Abnahme der Laufzeit.
Der Einsatz von low-V_T Gattern hat eine deutliche Reduzierung der Laufzeitsensitivitäten gegenüber den wichtigsten Prozess- und Betriebsparametern zur Folge. Die Ergebnisse für 40nm basieren auf ersten, noch unvollständigen Transistormodellen. Dies muss bei der Bewertung der Daten berücksichtigt werden. Die verringerten Sensitivitäten werden im Folgenden genutzt, um das Taktverteilungsnetz von synchronen Digitalschaltungen robuster gegenüber Prozess- und Umgebungsvariationen zu machen. Wie die Ergebnisse des Mikroprozessormodells in Kapitel 4.3 zeigen, ist der Beitrag des Taktbaums zur gesamten WID Timing-Unsicherheit sehr groß und nimmt mit tieferem Pipelining zu. Aus diesem Grund werden für die folgende Untersuchung alle Clock Buffer und Clock Gating Zellen durch ihre entsprechenden low-V_T Gatter ersetzt. Da das Gatterlayout unabhängig von der Wahl der Transistoreinsatzspannung ist, erfolgt der Austausch der reg-V_T Zellen pin-kompatibel und flächenneutral. Als Beispiel zur Abschätzung von Kosten und Nutzen dient das Produktdesign des ARM926 in 90nm CMOS Technologie [89].
Die erhöhten effektiven Schaltströme der low-V_T Transistoren haben eine signifikante Reduzierung der Laufzeiten zur Folge. Da die Laufzeitschwankung im Taktverteilungsnetz auch von der Propagationszeit des Taktsignals von der Signalquelle bis zu den empfan-

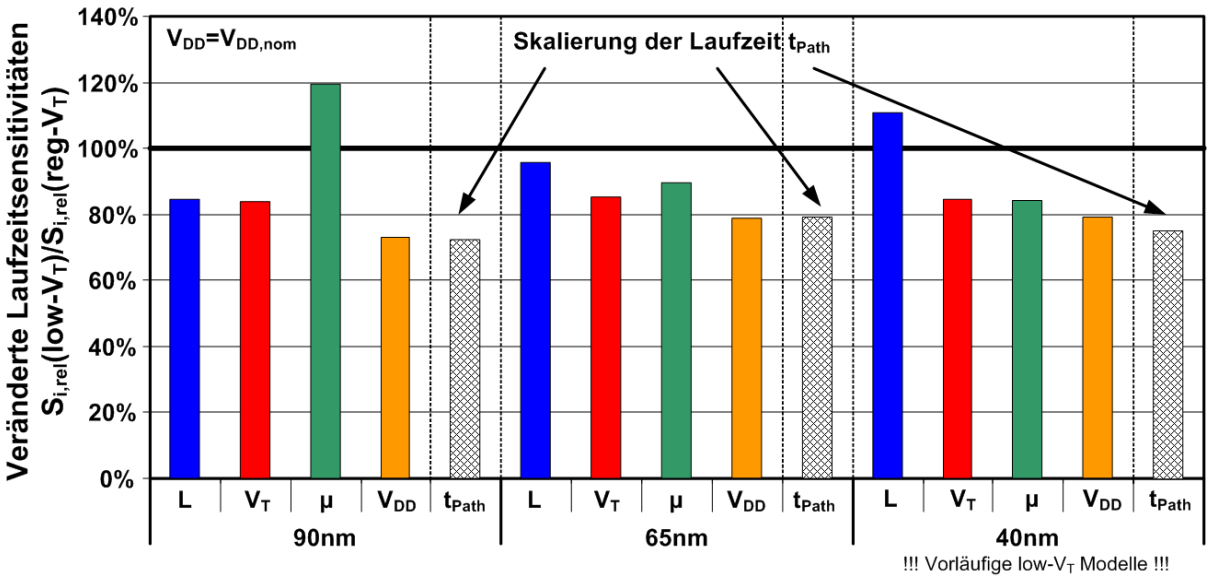


Bild 6.13: Laufzeitsensitivitäten einer low-V_T FO4 Inverter-Kette normiert auf eine reg-V_T Implementierung ($V_{DD} = V_{DD}^{nom}$, T=27°C).

Tabelle 6.1: Relative Laufzeitänderung eines low-V_T Taktpfades im Vergleich zur reg-V_T Implementierung.

90nm		65nm		40nm	
$V_{DD,nom}$	$V_{DD,low}$	$9V_{DD,nom}$	$V_{DD,low}$	$V_{DD,nom}$	$V_{DD,low}$
-27.8%	-38.3%	-20.7%	-33.1%	-24.0%	-41.6%

genden Flip Flop Zellen abhängt, verringert sich die absolute Laufzeitschwankung linear zur Laufzeit des Taktsignals. Tabelle 6.1 zeigt die relative Laufzeitänderung eines low-V_T Taktpfades gegenüber der ursprünglichen Implementierung mit reg-V_T Gattern. Für den untersuchten ARM926 in 90nm CMOS Technologie hat dies einen um 17% reduzierten Beitrag der Timing Unsicherheit zur Taktperiode zur Folge. Auf die Taktperiode umgerechnet ergibt sich daraus eine zeitliche Marge von 2-3%. Für einen ARM1176 in 65nm ist eine Marge von 3-4%, für einen Cortex A8 in 40nm eine Marge von 5-6% zu erwarten. Die Abschätzung für den ARM1176 basiert auf einer Test-Implementierung in 65nm, die Abschätzung für den Cortex A8 in 40nm auf den Ergebnissen des Mikroprozessormodells (Kapitel 4.3).
Bild 6.14 zeigt die aus dem Mikroprozessormodell stammenden Beiträge des Taktverteilungsnetzes zur Laufzeitschwankung für die Implementierung des Taktbaums mit reg-V_T und low-V_T Gattern.
Clock Skew und Jitter stellen den stärksten Anteil. Die deutlich reduzierte Sensitivität gegenüber Versorgungsspannungsschwankungen sowie die reduzierte nominelle Laufzeit führen zu signifikanter Verbesserung dieses Beitrags. Dies ist insbesondere für moderne System on Chips (SoC) sehr wertvoll, da das Taktsignal vom Taktgenerator (PLL) bis zum Einspeisepunkt eines bestimmten Schaltungsteils z.B. Applikationsprozessor etc. häufig eine nicht zu vernachlässigende Propagationszeit zurücklegt. Diese externe Laufzeit hat Einfluss auf die Größe des Clock Jitters. Eine Reduzierung dieser 'externen' Laufzeit durch den Einsatz von low-V_T Gattern bei gleichzeitiger Reduzierung der Laufzeitsensitivität gegenüber Versorgungsspannungsschwankungen kann hier die Laufzeitschwankung

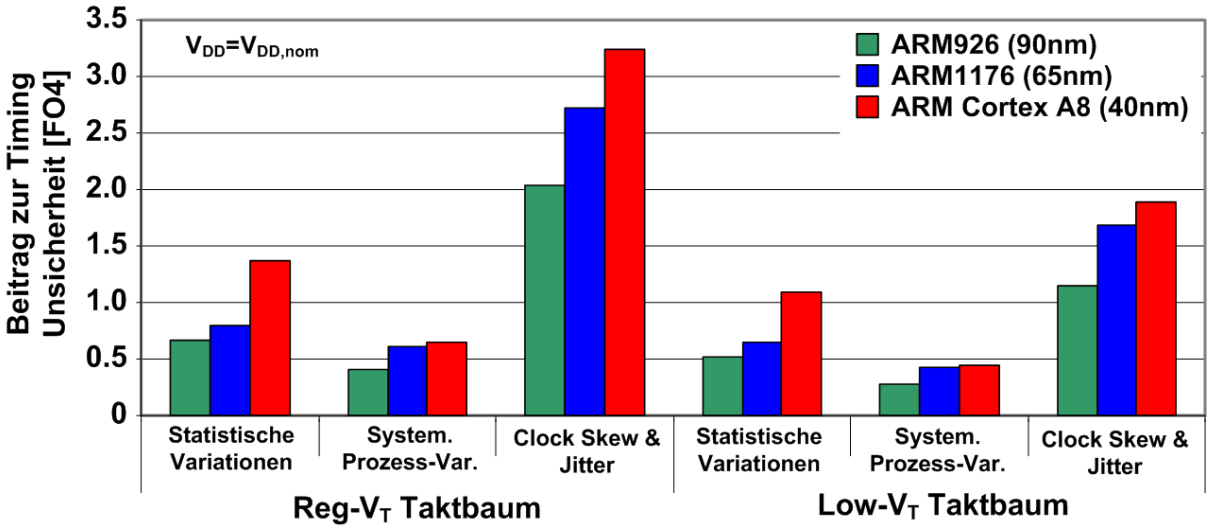


Bild 6.14: Mikroprozessormodell - Vergleich der Beiträge des Taktverteilungsnetzes zur Laufzeitschwankung für reg-V_T und low-V_T Taktverteilungsnetze.

Tabelle 6.2: Relative Leckstromerhöhung für low-V_T Taktverteilungsnetze in ARM926, ARM1176 und ARM Cortex A8.

	ARM926 (90nm)	ARM1176 (65nm)	ARM Cortex A8 (40nm)
100% low-V_T	26.7%	ca. 47% [1]	ca. 57% [1]
low-V_T, außer LCB	14.1%	ca. 24% [1]	ca. 29% [1]
[1] Abschätzungen anhand Gatternetzlisten und Mikroprozessormodell			

weiter verringern.

Mit tieferem Pipelining erhöht sich der Timing-Unsicherheits-Beitrag des Taktbaums. Somit steigt gleichzeitig der Vorteil von low-V_T Zellen im Taktbaum. Da sich die low-V_T Gatter in ihrer Fläche nicht von den ursprünglichen Gattern unterscheiden, ist für einen Kosten-Nutzen Vergleich neben der Größe der gewonnenen zeitlichen Marge lediglich die zusätzliche Leistungsaufnahme zu beachten.
Auf Gatterebene führt der Einsatz von low-V_T Gattern zu ca. 8-10 fachem Leckstrom. Tabelle 6.2 zeigt für die verschiedenen Mikroprozessoren den zu erwartenden Anstieg der Leckströme. Für die Abschätzung wurde eine gatterspezifische Leckstromcharakterisierung unter statistischer Berücksichtigung der Eingangs-Pin Belegung vorgenommen. Die Verwendung großer, treiberstarker Gatter im Taktbaum führt trotz einer relativ geringen Gatteranzahl zu einem deutlichen Anstieg des Leckstroms von über 50% für den Cortex A8. Da der Taktbaum den Schaltungsteil mit höchster Datenaktivität darstellt, ist die Leckstromerhöhung kein wesentlicher Kostenfaktor. Neben den zusätzlichen Leckstromverlusten führt die Verwendung von low-V_T Gattern zu einer Erhöhung der effektiven Gatekapazität und somit zu erhöhten dynamischen Verlusten. Simulationen einzelner ARM926 und ARM1176 Taktbaumsegmente auf Basis extrahierter Netzlisten zeigen je nach Prozess einen Anstieg der dynamischen Verlustleistung von 7-10% in 90nm und 6-9% in 65nm.
Da die Implementierung eines Binärbaums als Taktverteilungsnetz dazu führt, dass die Anzahl der Lokalen Clock Buffer (LCB) ca. 50% der Gesamtanzahl aller im Taktbaum verwendeten Gattern entspricht, kann die Verlustleistung durch den Ausschluss der LCB

während der reg-V_T Ersetzung nahezu halbiert werden. Da der Beitrag des Taktbaums zur Timing-Unsicherheit vorwiegend durch dessen Laufzeit bestimmt wird, hat der Ausschluss einzelner Zellen während der reg-V_T Ersetzung eine verringerte Abnahme der Laufzeitschwankung zur Folge. Im Fall des ARM926 und ARM1176, deren Taktpfade durchschnittlich 10-11 Gatter (ohne Flip Flops) beinhalten, liegt der Einfluss der LCB auf die Propagationszeit im Bereich von ca. 10%. Somit bleibt die Laufzeit und damit auch die reduzierte Laufzeitschwankung nahezu unverändert, während sich die ursprüngliche Zunahme der aktiven Verlustleistung halbiert. Ein stärkerer Eingriff in das Taktverteilungsnetz, d.h. eine inhomogenere Verteilung von mixed-V_T Gattern ist nicht sinnvoll, da das zeitliche Balancieren des Taktverteilungsnetzes unter verschiedenen Prozess- und Betriebsbereichen aufgrund unterschiedlicher Laufzeitsensitivitäten deutlich erschwert wird. Neben dem Einsatz von low-V_T Gattern können Laufzeiten auch durch erhöhte Treiberstärken (Gate-Sizing) reduziert werden. Simulationen einzelner Segmente aus den Taktbäumen des ARM926 und ARM1176 zeigen jedoch, dass die Aufdopplung der verwendeten Treiberstärken zu lediglich 5-7% reduzierten Laufzeiten führen, bei gleichzeitiger Erhöhung der dynamischen Verluste von 46-52% und einem Flächenzuwachs von 1.2% für den ARM926 bzw. 2.3% für den ARM1176. Die Laufzeitsensitivitäten bleiben jedoch unverändert. Die doppelte Transistorfläche führt zu um ca. 30% verringerten statistischen Schwankungen der Einsatzspannungen. Da diese statistischen Schwankungen nur einen geringen Anteil an der Laufzeitschwankung im Taktbaum haben, trägt dieser Effekt unwesentlich zur Verbesserung der Laufzeitschwankung bei. Grund für die geringe Laufzeitreduktion sind die relativ hohen Treiberstärken der verwendeten Gatter, um steile Flanken im Taktbaum sicherzustellen. Dies wird als Design-Kriterium von den Synthese-Tools überwacht. Gate-Sizing kann somit nicht als Alternative zum Einsatz von low-V_T Zellen im Taktbaum gesehen werden.

6.2.3 Selektiver Einsatz von low-V_T Zellen in geschwindigkeitskritischen Pfaden

Wie im Abschnitt 6.2.2 gezeigt wurde, reduziert sich beim Einsatz von low-V_T Gattern im Vergleich zu entsprechenden reg-V_T und high-V_T Implementierungen sowohl die Laufzeitsensitivität gegenüber Prozess- und Umgebungsvariationen als auch die Laufzeit selbst. Beim Einsatz der low-V_T Gatter im Taktbaum wurden globale zeitliche Sicherheitsmargen durch verringerte Sensitivitäten im Taktbaum erzielt.
Mixed-V_T Schaltungsdesign ist eine wohl bekannte Maßnahme um den Leckstrom durch den Einsatz von high-V_T Gattern in geschwindigkeitsunkritischen Pfaden zu reduzieren, während in den geschwindigkeitskritischen Pfaden Gatter mit niedrigerer Transistoreinsatzspannung verwendet werden, um die Geschwindigkeitsanforderungen erfüllen zu können. Der Anteil von Gattern mit niedrigem V_T liegt in den meisten Implementierungen von mixed-V_T Designs bei ca. 30-40% [161].
Im Gegensatz dazu wird in diesem Abschnitt ein Ansatz diskutiert, der die Anzahl der zu ersetzenden Zellen gering hält. Die robustheitsorientierte Ersetzung hat das Ziel, bereits während des Entwurfs eine zeitliche Sicherheitsmarge zu erzeugen. Da der Einsatz von low-V_T Zellen pin-kompatibel und flächenneutral erfolgt, ist unter bestimmten Voraussetzungen auch ein selektiver Austausch von Gattern nach dem Timing Sign-Off möglich.

Basis der Ersetzungsstrategie sind verschiedene, robustheitsspezifische Bewertungskenngrößen, die eine Priorisierung der einzelnen Gatter während der Ersetzung ermöglichen.

Im Folgenden werden die verschiedenen Bewertungskenngrößen einzeln aufgeführt.

1. **Pfadlaufzeit t_{Pfad}:**

   Im Allgemeinen gilt, dass Pfade mit geringem 'Slack', d.h. Pfade, die eine geringere zeitliche Marge zum funktionalen Ausfallpunkt haben, eine hohe Pfadlaufzeit aufweisen. Dennoch können Pfade mit gleichem Slack unterschiedliche Laufzeiten haben, da der eine Pfad z.B. einem negativen, während der andere Pfad einem positiven Clock Skew unterliegt. Da kombinatorische Logik höhere Laufzeitsensitivität als einfache Inverter oder Buffer-Zellen aufweist, wie sie im Taktbaum vorkommen, werden Gatter in Logikpfaden stärker gewichtet.
   Je höher die Pfadlaufzeit desto höher die Priorisierung des Gatters.

2. **Gatterlaufzeit t_{Gatter}:**

   Neben der Pfadlaufzeit ist die absolute Laufzeit eines Gatters entscheidend. Je höher der Anteil an der Pfadlaufzeit desto wirksamer die Beschleunigung eines solchen Gatters.
   Je höher die Gatterlaufzeit desto höher die Priorisierung des Gatters.

3. **Anzahl der Pfade pro Gatter $N_{P/G}$:**

   Je höher die Anzahl der Pfade, die ein bestimmtes Gatter beinhalten, desto größer der Einfluss auf das Gesamt-Timing der Schaltung. Diese Größe wird in Kombination mit der Pfadlaufzeit der einzelnen Pfade betrachtet. Dieser Aspekt kann durch die topologische Korrelation, wie in Abschnitt 5.2 definiert, makroskopisch beschrieben werden.
   Je größer die Anzahl von Pfaden pro Gatter desto höher die Priorisierung des Gatters.

4. **Lage des Aufspaltungspunktes im Taktbaum:**

   Die Lage des Aufspaltungspunktes beider Taktpfade zum sendenden und empfangenden Register des das Gatter beinhaltenden Logikpfades bestimmt den Einfluss von lokalen Variationen auf den Clock Skew. Je näher der Aufspaltungspunkt an der Einspeisequelle, desto höher die Priorisierung des Gatters.

5. **Gattertopologie:**

   Die Höhe des Transistor-Stacks, d.h. die Anzahl in Serie verschaltener Transistoren zwischen Ausgangsknoten und V_{DD} bzw. V_{SS} beeinflusst die Priorität eines Gatters. So wird z.B. ein NAND Gatter stärker gewichtet als ein Inverter oder eine Buffer Zelle.
   Je höher der Transistor-Stack desto höher die Priorisierung des Gatters.

6. **Transistorgröße:**

   Die Schwankungsbreite statistischer Prozessvariationen nimmt mit der Größe des Transistors ab. Somit werden Gatter mit geringer Treiberstärke bevorzugt ersetzt.
   Je kleiner die Treiberstärke, desto höher die Priorisierung des Gatters.

7. **Pfadtopologie:**

   In Kapitel 5.1 werden verschiedene Pfadtopologien gezeigt und hinsichtlich ihrer

Tabelle 6.3: Normierte Sensitivitätsfaktoren für den Einsatz von low-V_T Gattern.

$\Delta t^{Soll}_{Path,max}$	$\Delta t_{Path,max}$	S_1	S_2	$\Delta t_{Path,max}$ für $S_1 = const.$	$\Delta t_{Path,max}$ für $S_2 = const.$
low-V_T: -5%	-5.2%	5.1x	11.8x	-1.6%	-2.7%
low-V_T: -7%	-7.2%	7.7x	19.9x	-3.1%	-4.4%

Umgebung in verschiedene Pfadgruppen eingeteilt. So werden Gatter, die in 'Loop-internen' kritischen Pfaden vorkommen, bevorzugt gegenüber isoliert kritischen Pfaden ersetzt.
Je kritischer die Pfadtopologie, desto höher die Priorisierung des Gatters.

Im Gegensatz zu herkömmlichen Ersetzungsstrategien basiert der hier vorgestellte Ansatz neben Geschwindigkeitsaspekten auch auf zuvor definierten, topologie- und architekturabhängigen Eigenschaften der Schaltung. Bild 6.15 zeigt ein schematisches Ablaufdiagramm des verwendeten, robustheitsorientierten Ersetzungsalgorithmus.
Nach der Bestimmung der Gewichtungsfaktoren erfolgt ein Vergleich, der für ähnliche Gewichtungsfaktoren, d.h. Gewichtungsfaktoren, die sich relativ gesehen nicht mehr als um einen zuvor festgelegten prozentualer Wert (hier: 5%) unterscheiden, eine neue Priorisierung vornimmt. Dabei gehen die in der Darstellung gezeigten Kriterien ein.
Der Einfluss der Ersetzungsstrategie auf die Verteilung der maximalen Pfadlaufzeit in variationsbehafteter Umgebung kann in dieser Arbeit aufgrund der Schaltungskomplexität, wie in Kapitel 5 diskutiert, nicht quantifiziert werden. Aus diesem Grund wird nochmals auf Kapitel 2, 3 und 4 verwiesen, die zahlreiche generische Untersuchungen zu den o.g. Bewertungskriterien beinhalten. Im Folgenden wird daher der Einfluss der Ersetzungsstrategie anhand des in Kapitel 5.2 definierten Faktors der strukturellen Schaltungssensitivität und den Veränderungen in Gatter- und Pfadspektren diskutiert.

Bei Bestimmung der Kosten und Nutzen von lokalem low-V_T Einsatz wurde für jedes Gatter ein Gewichtungsfaktor bestimmt, der die Punkte 1 bis 4 berücksichtigt. Bei ähnlichem Gewichtungsfaktor wurden die Punkte 5 bis 7 als direktes Entscheidungskriterium herangezogen.
Bild 6.16 zeigt Gatter- und Pfadspektrum des ARM926 vor und nach dem selektiven Einsatz von low-V_T Gattern für eine zeitliche Sicherheitsmarge von $\Delta t^{Soll}_{Path,max} = 5\%$ bzw. $\Delta t^{Soll}_{Path,max} = 7\%$.
Das Pfadspektrum zeigt, dass die kritischen Logikpfade durch den Einsatz von low-V_T Gattern beschleunigt werden und sich hinter der neuen maximalen Pfadlaufzeit aufreihen. Dieser zeitliche Ausgleich resultiert im Aufbau einer Timing-Wall, die mit zunehmender zeitlicher Marge an Steilheit gewinnt. Diese Verformung des Pfadspektrums ist auch am Gatterspektrum zu erkennen. Die akkumulierte Gatteranzahl an der maximalen Pfadlaufzeit nimmt mit steigender zeitlicher Marge exponentiell zu. Diese Zunahme wirkt sich auf die in Kapitel 5.2 definierten Sensitivitätsfaktoren der Schaltung aus. Normiert man den Sensitivitätsfaktor auf das Originaldesign, so ergeben sich für die Definition des Sensitivitätsfaktor S_1 nach Gleichung 5.2 und für die Definition S_2 nach Gleichung 5.4 die in Tabelle 6.3 aufgeführten Werte. Für eine zeitliche Sicherheitsmarge von 5% erhöht sich der Sensitivitätsfaktor S_1 um den Faktor 5.1, für eine zeitliche Marge von 7% um den Faktor 7.7. Aufgrund der höheren Gewichtung der Gatter an der Geschwindigkeitsgrenze erhöht sich der Sensitivitätsfaktor S_2 stärker als S_1.

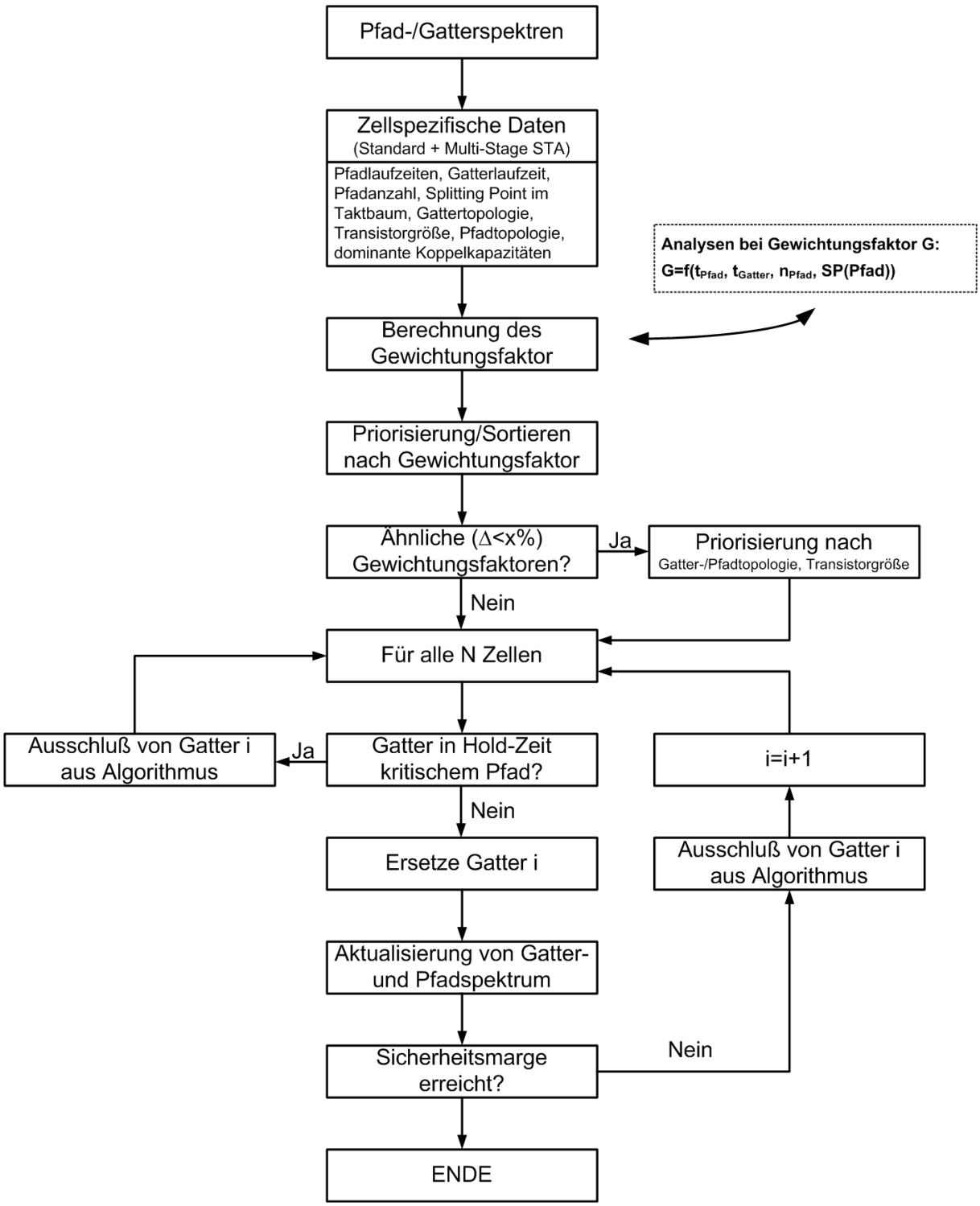


Bild 6.15: Schematisches Ablaufdiagramm des robustheitsorientierten Ersetzungsalgorithmus.

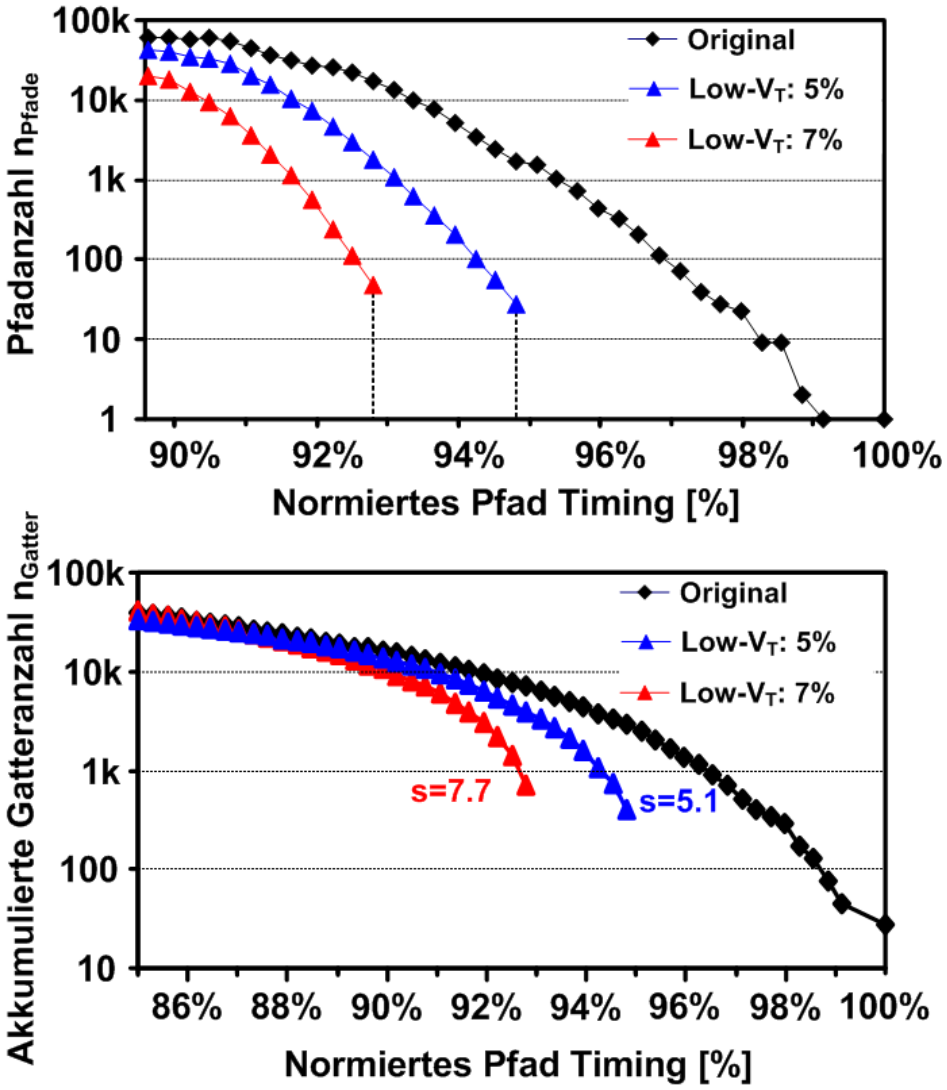


Bild 6.16: Gatter- und Pfadspektrum des ARM926 vor und nach selektivem Einsatz von low-V_T Gattern.

Tabelle 6.4: Zusammenfassung zum low-V_T Gatter Einsatz in einem 90nm ARM926.

$\Delta t^{Soll}_{Path,max}$	$\Delta t_{Path,max}$ für $S_2 = const.$	S_2	P_{Stat}	Fläche
-5%	-2.7%	11.8x	+0.8%	0%
-7%	-4.4%	19.9x	+2.6%	0%

Es zeigt sich, dass für die Randbedingung einer konstanten Schaltungssensitivität, für beide Definitionen des Schaltungssensitivitätsfaktors eine deutlich reduzierte Geschwindigkeitsmarge $\Delta t_{Path,max}$ erzielt wird (siehe Tabelle 6.3, Spalten 5,6). Die Anhäufung der Gatter sowie eine hohe topologische Korrelation führt zu einem erhöhten Risiko, dass ein kritischer Pfad sensibilisiert wird und dabei eine hohe Laufzeitschwankung aufweist.

Durch den Einsatz von low-V_T Zellen erhöht sich der Leckstrom der Zellen, so dass die statische Verlustleistung ansteigt. Im Gegensatz zu herkömmlichen mixed-V_T Designs liegt der Anteil der durch low-V_T Gatter ersetzten Zellen bei maximal 1%, so dass sich der Leckstrom nur unwesentlich erhöht. Der Ersetzungsalgorithmus arbeitet im Gegensatz zur STA Analyse zellbasiert und berücksichtigt bei der Geschwindigkeitserhöhung Setup-Zeit kritischer Pfade gleichzeitig Hold-Zeit kritische Strukturen, d.h. die Ersetzung von Gattern, die sowohl in Setup-Zeit kritischen Pfaden als auch in Hold-Zeit kritischen Pfaden liegen, wird ausgeschlossen, so dass keine Veränderung der Hold-Zeit kritischen Pfade vorgenommen wird (siehe Bild 6.15). Somit ist kein zusätzliches Hold-Zeit Fixing erforderlich und das Layout der Schaltung bleibt unverändert. Dies ermöglicht eine Post-Sign-Off Geschwindigkeits- und Robustheitserhöhung zu geringen Kosten. Tabelle 6.4 fasst den Einfluss von selektivem low-V_T Gatter Einsatz auf Geschwindigkeit, Sensitivitätsfaktor, statische Verlustleistung und Fläche eines ARM926 Mikroprozessors in 90nm CMOS Technologie zusammen.

6.2.4 Einsatz von gepulsten Flip Flops (P-FF) / Latches (P-L)

In diesem Abschnitt wird der Einsatz von gepulsten Flip Flops (P-FF) und gepulsten Latches (P-L) im Semicustom Schaltungsentwurf diskutiert. In low-power Schaltungen werden standardmäßig Master-Slave Flip Flops (MS-FF) verwendet, da sie sich durch geringe Hold-Zeiten und vergleichsweise geringe Energieverluste auszeichnen [105, 188]. Aufgrund des ansteigenden Laufzeitbeitrags von Flip Flop Zellen zur Taktperiode ist der Einsatz schneller Flip Flops bzw. Latches auch für low-power Schaltungen attraktiv.
Die wesentlichen Laufzeitgrößen von Flip Flop Zellen sind der zeitliche Abstand zwischen Daten- und Taktflanke am Eingang t_{D-Clk}, Taktflanke am Eingang und Datenflanke am Ausgang t_{Clk-Q} sowie Datenflanke am Eingang und Datenflanke am Ausgang $t_{D-Q} = t_{D-Clk} + t_{Clk-Q}$. Gepulste Flip Flops bzw. Latches zeichnen sich insbesondere durch sehr niedrige oder sogar negative Setup-Zeiten aus, d.h. ein Datum, das den Dateneingang kurz nach der Taktflanke erreicht, kann noch immer korrekt gespeichert werden. Dazu ist eine Transparenzphase notwendig, die durch die Breite des Pulses festgelegt wird. Während dieser Zeit bleibt t_{D-Q} konstant, so dass die Laufzeit von empfangendem Pfad und sendendem Pfad ausbalanciert werden kann. So ermöglichen P-FF und P-L sowohl statisches als auch dynamisches Time-Borrowing [134]. Der Einsatz von P-FF/P-L bringt somit eine von der Pulsbreite abhängige zeitliche Elastizität in die Schaltung. Diese Elastizität kann genutzt werden, um Laufzeitschwankungen aus Prozess- und Umgebungs-

variationen zu kompensieren.
Aus diesem Grund werden P-FFs unter anderem in Multi-Zyklen Interconnects verwendet. Diese Verbindungen sind für den Einsatz von P-FFs besonders attraktiv, da diese eine Punkt-zu-Punkt Verbindung darstellen. Ist diese Verbindung geschwindigkeitskritisch, so kann eine große Pulsbreite gewählt werden, da keine kurzen Pfade an diesen Flip Flops enden, und daher keine Hold-Zeit Verletzungen zu befürchten sind [135].
In [136] werden erstmals gepulste Flip Flops in synthetisierbaren Semicustom Schaltungen verwendet. Hier werden in den kritischen Pfaden P-FFs mit verschiedenen Pulsbreiten für statisches Time-Borrowing eingesetzt. Dieser Einsatz von gepulsten Flip Flops führt in loop-internen kritischen Pfaden jedoch zu Setup-Zeit Verletzungen [166]. In [167, 168, 169] werden gepulste Latches in High Performance Mikroprozessoren verwendet, um zum einen Fläche und Leistungsaufnahme zu reduzieren, zum anderen durch verringerte Laufzeiten und die Möglichkeit von Time-Borrowing zeitliche Margen bzw. Elastizität im Hinblick auf variationsbedingte Laufzeitschwankungen zu schaffen. In [170] wird der Einsatz von gepulsten Flip Flops vorgeschlagen, deren Pulsbreite mit fortschreitender Zeit und damit verbundenen, erhöhten Alterungseffekten angepasst werden kann.
Sowohl gepulste Latches als auch gepulste Flip Flops ermöglichen es, auch ohne Kenntnis des einzelnen Effekts, durch dynamisches Time-Borrowing Laufzeitvariationen bis zu einer bestimmten Größe (Pulsbreite) selbständig und ohne äußeren Einfluss zu kompensieren [134]. Dies ermöglicht auch die Kompensation von kurzzeitigen Umgebungsvariationen, deren Detektion zwar möglich aber eine Kompensation mittels Regelkreis nicht implementierbar ist.

Selektiver Einsatz von P-FFs in geschwindigkeitskritischen Pfaden

Im Folgenden wird der Einsatz von gepulsten Flip Flops mit integriertem Pulsgenerator in geschwindigkeitskritischen Pfaden diskutiert. Da selbst in zeitlich stark ausbalancierten Pipelinestufen neben sukzessiv kritischen Pfaden, für die die verbesserte Daten-zu-Ausgang Laufzeit t_{D-Q} als Bewertungskriterium herangezogen wird [189], auch isoliert kritische Pfade vorkommen (siehe Kapitel 5.1), muss beim Austausch von Flip Flop Zellen zwischen den Änderungen von t_{D-Clk} und t_{Clk-Q} unterschieden werden. Bild 6.17 zeigt den Schaltplan des gepulsten Flip Flops, das für die weiteren Untersuchungen in diesem Abschnitt verwendet wird [166].
Die Kernidee dieses Flip Flops ist es, die Propagation einer steigenden und fallenden Datenflanke individuell zu beschleunigen. Aus diesem Grund wird der Propagationspfad der beiden Flanken nach dem Dateneingang aufgespaltet. Gemeinsam mit dem generierten Puls wird das Datensignal an ein 2-fach NAND bzw. NOR geführt. Während der Transparenzphase schaltet je nach Datensignal einer der beiden Transistor-Stacks (P6/P7 bzw. N8/N9). Hier werden treiberstarke Transistoren eingesetzt um die Propagation zu beschleunigen und gleichzeitig eine hohe Flankensteilheit zu gewährleisten. Die ebenfalls treiberstarken Transistoren P1 bzw. N1 laden nun den internen Speicherknoten sowie die Eingangskapazität der Ausgangsstufe um.
Im Semicustom Design müssen weitere Randbedingungen wie z.B. die Testbarkeit (Scan) und ein asynchroner Reset gewährleistet werden. Während der Scan-Eingang ebenso wie im Master-Slave Flip Flop vor dem Dateneingang liegt und somit die Propagationszeit des Datums erhöht, liegt der asynchrone Reset außerhalb des kritischen Propagationspfads.
Diese Anpassungen ermöglichen es, dass das gepulste FF im Vergleich zum Standard Master-Slave Flip Flop neben der für P-FF charakteristischen negativen Setup-Zeit auch

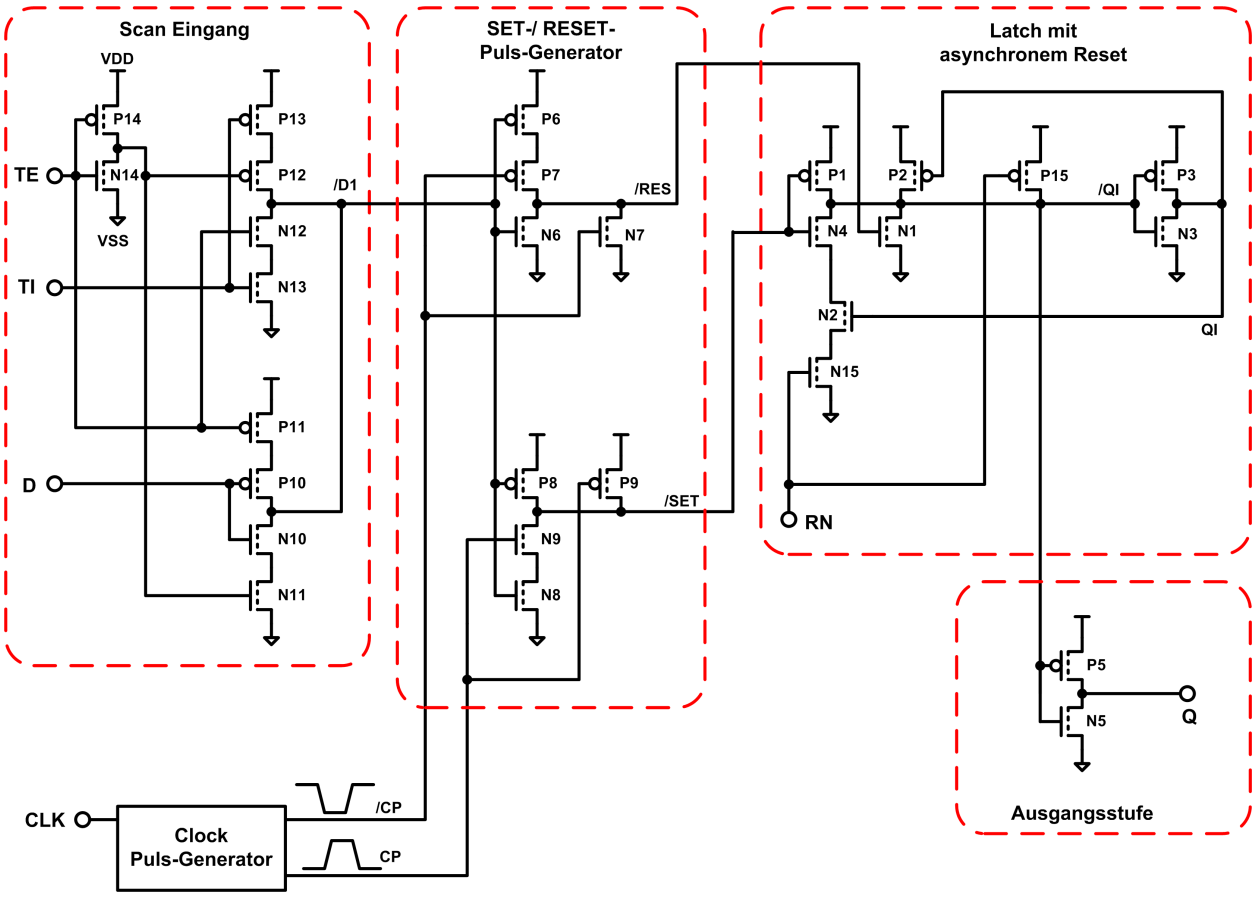


Bild 6.17: P-FF mit aufgespaltetem Propagationspfad zur Beschleunigung der Clock-Q Laufzeit.

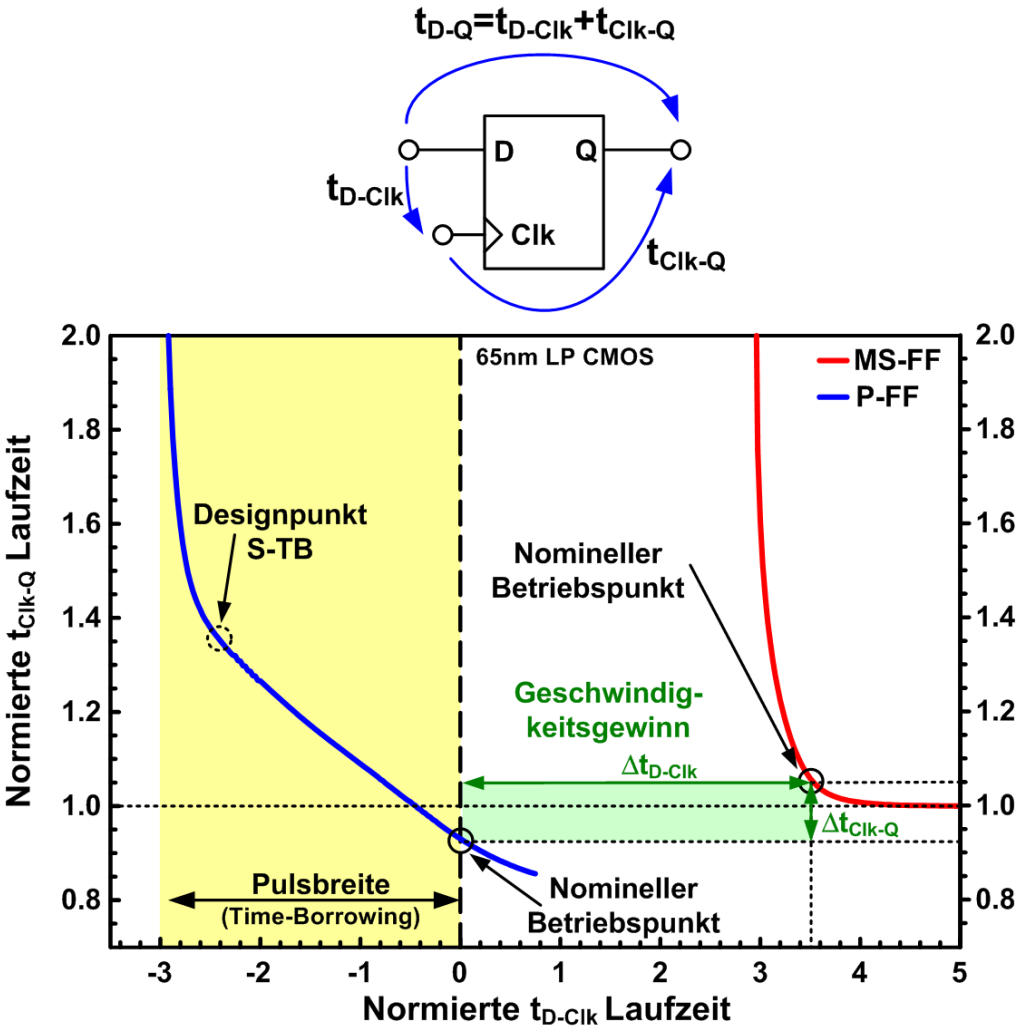


Bild 6.18: Simulierte Clock-Q Laufzeit von MS-FF und P-FF in Abhängigkeit der Data-Clock Laufzeit auf Basis extrahierter Netzlisten in 65nm CMOS ($V_{DD} = V_{DD,nom}$, T=27°C).

eine deutlich reduzierte t_{Clk-Q} Laufzeit aufweist. Diese Eigenschaft ist insbesondere in sukzessiv kritischen Pfaden und loop-internen kritischen Pfaden von Vorteil, da sowohl der am FF ankommende als auch der vom FF abgehende kritische Pfad beschleunigt wird.

Bild 6.18 zeigt die Clock-Q Laufzeit von gepulstem Flip Flop und Master-Slave Flip Flop. Der lineare Anstieg von t_{Clk-Q} der P-FF Kennlinie kennzeichnet die Transparenzphase des P-FF während des Pulses. Zu diesem Zeitpunkt, d.h. während der Transparenzphase, ist die Data-Q Laufzeit minimal. Eine eindeutige Setup-Zeit, wie Sie für Master-Slave Flip Flops definiert ist, kann gepulsten Flip Flops daher nicht zugeordnet werden. Aus diesem Grund ist es notwendig den nominellen Design-Punkt von gepulsten Flip Flops festzulegen, um den Einfluss auf die Geschwindigkeit der Schaltung abschätzen zu können. Für die folgenden Untersuchungen wird die Setup-Zeit des gepulsten Flip Flops auf $t_{SU}^{P-FF} = 0ps$ gesetzt. Im Vergleich zum Master-Slave Flip Flop ist in diesem Betriebspunkt auch die Clock-Q Laufzeit t_{Clk-Q} geringer.
Die Transparenzphase von gepulsten Flip Flops und Latches ermöglicht es zeitliche Variationen des ankommenden Pfades auf den nächsten Pfad zu übertragen, so dass Laufzeiten zwischen verschiedenen Pipelinestufen ausgeglichen werden können. Dies kann sowohl statisch (Static Time-Borrowing S-TB) als auch dynamisch geschehen (Dynamic Time-Borrowing D-TB). S-TB entspricht dabei der Wahl des nominellen Betriebspunkts nahe des Ausfallpunktes, d.h. $\frac{\partial t_{Clk-Q}}{\partial t_{D-Clk}} \mapsto \infty$. Dieser Betriebspunkt ist dem nominellem Betriebspunkt des Master-Slave Flip Flops sehr ähnlich, da mit betragsmäßig zunehmendem zeitlichen Abstand t_{D-Clk} eine Setup-Zeit Verletzung generiert wird.
Die Transparenzphase des P-FF erhöht gleichzeitig die Hold-Zeit, so dass die Vermeidung von Hold-Zeit Verletzungen im Vergleich zu Schaltungen mit MS-FFs im Allgemeinen höheren Aufwand erfordert. Die erhöhte Hold-Zeit liegt dabei in der Größenordnung der Pulsbreite [134].

Im Folgenden wird der Einsatz des o.g. P-FF in den kritischen Pfaden diskutiert, d.h. alle Pfade, deren Pfad Timing in den oberen 10% des Pfad Timings des kritischsten Pfades liegt. Dazu werden alle MS-FFs in den kritischen Pfaden durch P-FFs ersetzt. Bild 6.19 zeigt schematisch die Pipelinestruktur vor (links) und nach (rechts) der Ersetzung. Die blau gezeichneten Schaltungsteile werden ersetzt.

Einfluss auf die Pfadlaufzeiten

Für die Bestimmung der Pfadlaufzeiten wird für die Setup-Zeit des gepulsten Flip Flops $t_{SU}^{P-FF} = 0ps$ gewählt. In diesem Betriebspunkt erfolgt kein statisches Time-Borrowing. Bild 6.20 zeigt das Pfad- und Gatterspektrum vor und nach der selektiven Ersetzung von MS-FFs in den kritischen Pfaden.
Durch die Verringerung der Setup-Zeit t_{SU} auf 0ps sowie der verbesserten Clock-Q Laufzeit t_{Clk-Q} reduziert sich das Pfad Timing des kritischsten Pfads auf ca. 93% des ursprünglichen Wertes. Somit stehen nach der Ersetzung ca. 7-8% zeitliche Marge zur Kompensation von Laufzeitschwankungen zur Verfügung. Da auch sub-kritische Pfade beschleunigt werden, sofern diese an einem der ersetzten Flip Flops enden bzw. beginnen, ist die Akkumulation von kritischen Pfaden an der neuen maximalen Pfadlaufzeit gering. Daher ist im kritischen Bereich eher eine Verschiebung als eine Verformung des Pfad- und Gatterspektrums zu erwarten. Zusätzlich zur Beschleunigung der Logikpfade ermöglicht die Transparenzphase der gepulsten Flip Flops die Kompensation von dynamischen Variationen in Logik und Taktbaum. Der Einfluss dieser Elastizität auf das Schaltverhalten kann jedoch nicht quantifiziert werden, da dazu die genaue Kenntnis über alle dynamischen Laufzeitschwankungen in zeitlicher Reihenfolge benötigt wird.
Bei der dynamischen Adaption von Betriebsparametern ist eine zeitliche Elastizität in den kritischen Pfaden von Vorteil, da nicht gewährleistet werden kann, dass die Stell-

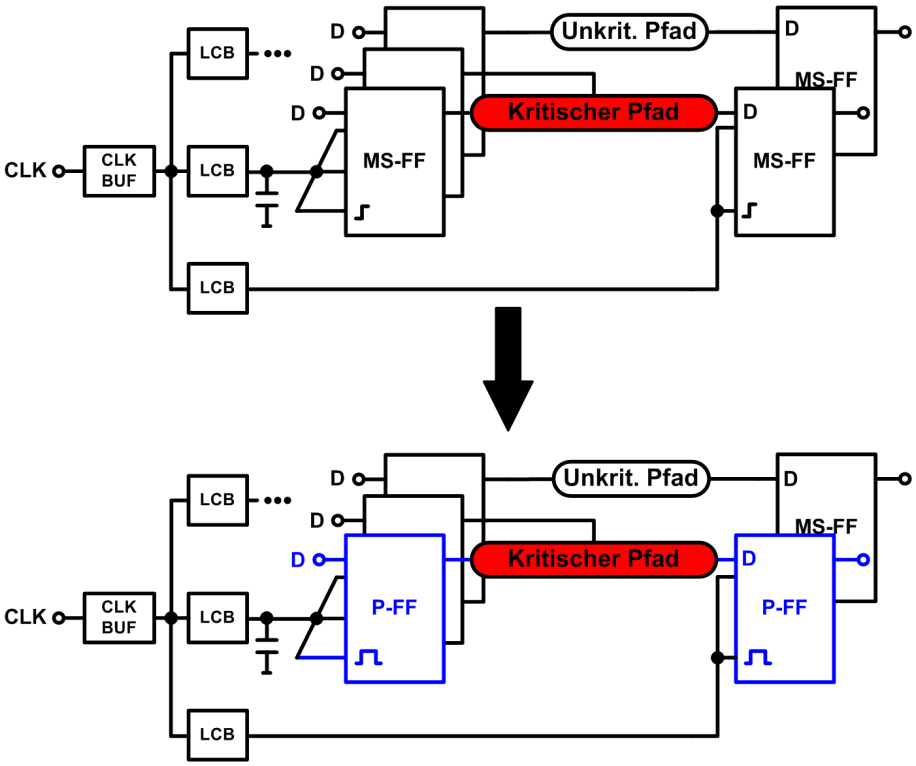


Bild 6.19: Schematische Darstellung der Ersetzungsstrategie.

größe zu jeder Zeit an jedem Ort gleich ist. Aus diesem Grund kann es zu künstlich erzeugten Variationen z.B. durch die Adaption der Versorgungsspannung oder der Body-Spannung (ABB) kommen. Selbst eine Erhöhung der Versorgungsspannung und die damit verbundene Verringerung der Laufzeit kann im Taktbaum zur Verkürzung der effektiven Taktperiode führen. Da die Clock-Q Laufzeit des Flip Flops vom zeitlichen Abstand von Daten- und Taktflanke t_{D-Clk} abhängt, kann das Flip Flop als Schnittstelle zwischen Logik und Taktbaum gesehen werden. Die zeitliche Elastizität des gepulsten Flip Flops kann daher sowohl Schwankungen der Laufzeit aus dem Logikteil als auch aus dem Taktbaum kompensieren.
Um die maximale zeitliche Marge zu bestimmen, die durch den Einsatz von gepulsten Flip Flops gegenüber einer Implementierung mit MS-FFs zu erzielen ist, wird als Testschaltung eine 'pipelined loop' aus vier identischen kritischen Pfaden verwendet, die in Bild 6.21 gezeigt ist [166]. Im Gegensatz zum obigen Beispiel (STA Ergebnisse eines ARM926 in 90nm), zeigen die in Bild 6.21 dargestellten Messungen in 65nm LP CMOS den Betrieb von P-FFs bei minimalem t_{D-Q} anstatt bei $t_{SU} = 0ps$.

Als Logikpfad zwischen den einzelnen Flip Flops wird der in die 65nm Technologie übertragene kritischste Pfad des ARM926 verwendet. Zu Beginn werden alle Flip Flops auf logisch '0' gesetzt. Im ersten Zyklus wird eine logische '1' im ersten Flip Flop gespeichert. In den nachfolgenden Taktzyklen propagiert nun die logische '1', so dass am Ausgang eine Oszillation mit einem Duty Cycle von 25% zu sehen ist. Wird nun die Taktfrequenz weiter erhöht, so wird zu einem bestimmten Zeitpunkt eines der Flip Flops aufgrund eines zu geringen t_{D-Clk} nicht mehr in der Lage sein, die propagierende '1' zu speichern. Die '1' geht verloren, und die Oszillation ist beendet.
Das Verhältnis der maximal zu erzielenden Taktfrequenz von P-FF gegenüber MS-FF liegt

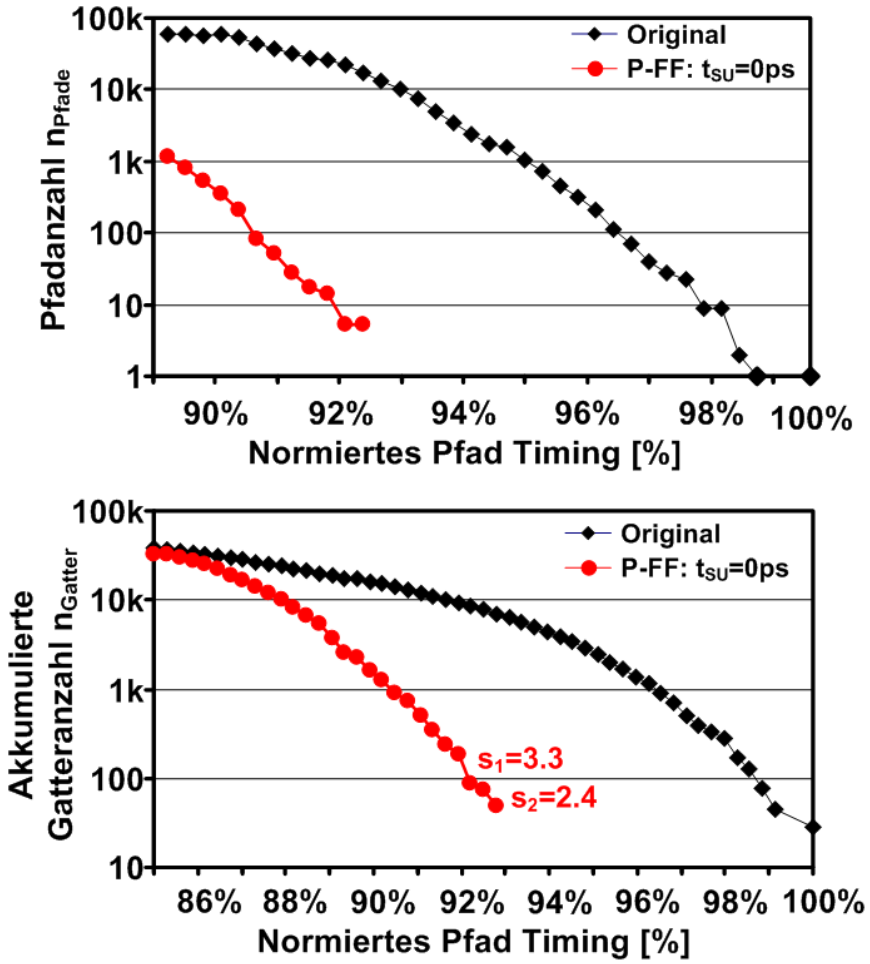


Bild 6.20: Pfad- und akkumulierte Gatterverteilung des ARM926 in 90nm CMOS vor und nach dem Ersetzen von MS-FFs in den kritischen Pfaden.

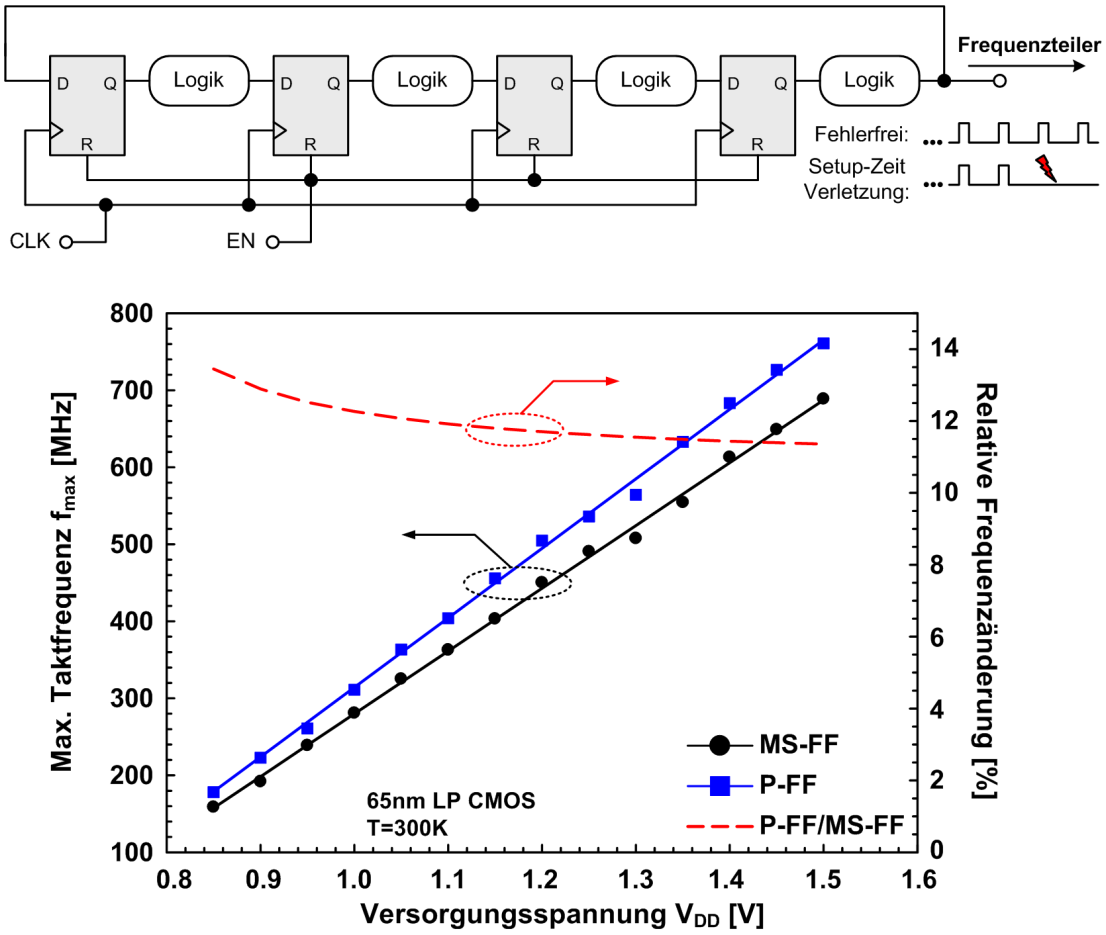


Bild 6.21: Gemessene maximale Taktfrequenz der obigen Anordnung für den Einsatz von MS-FFs und P-FFs in loop-internen kritischen Pfaden am Beispiel eines 65nm CMOS Testchips bei T=27°C.

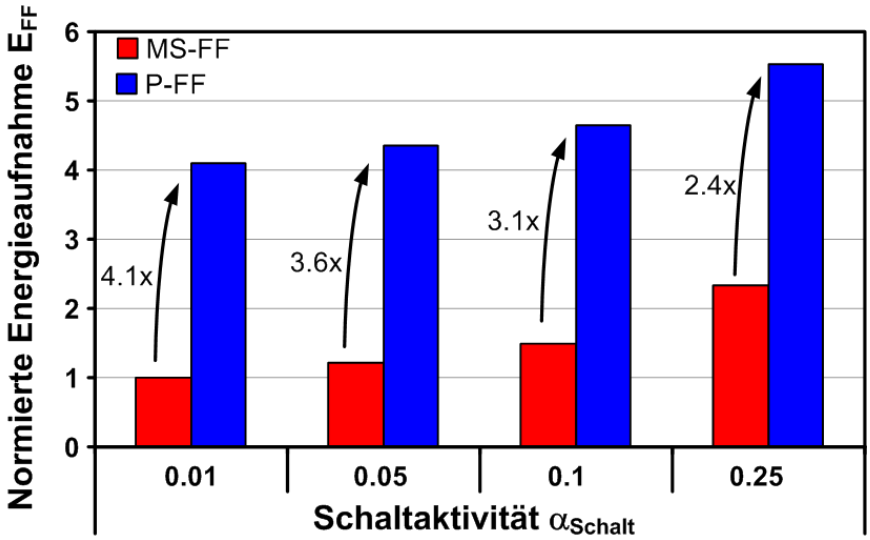


Bild 6.22: Simulation der Energieaufnahme von P-FF und MS-FF in 65nm CMOS auf Basis extrahierter Netzlisten ($V_{DD} = V_{DD}^{nom}$, T=27°C).

bei nomineller Versorgungsspannung bei ca. 11%. Für niedrigere Versorgungsspannungen ist ein geringfügig erhöhter Geschwindigkeitsgewinn zu erkennen, der grundsätzlich durch die - wie Simulationen zeigen - im Vergleich zum P-FF erhöhte Sensitivität der Daten-zu-Ausgang Laufzeit t_{D-Q} des Master-Slave Flip Flops gegenüber V_{DD} erklärt werden kann. Durch die Anpassung der Taktfrequenz (VCO) ist insbesondere für geringe Frequenzen ein klarer Ausfallpunkt nur mit einer Genauigkeit von ca. 1.5% zu bestimmen, so dass die Zunahme des gemessenen Geschwindigkeitsgewinns im Bereich der Messgenauigkeit liegt. Daher darf dieser Effekt nicht als systematischer Unterschied zwischen P-FF und MS-FF betrachtet werden.

Einfluss auf die Energieaufnahme

Neben den Geschwindigkeitsvorteilen und der Möglichkeit zu dynamischem Time-Borrowing erhöht sich durch den Einsatz von gepulsten Flip Flops sowohl der Flächenbedarf als auch die statische und dynamische Leistungsaufnahme. In Bild 6.22 wird die Energieaufnahme von MS-FF und P-FF in Abhängigkeit der Schaltaktivität gezeigt.
Für niedrige Schaltaktivitäten ist der Anteil der internen Taktverteilung sehr hoch, so dass das P-FF eine signifikant erhöhte Energieaufnahme aufweist. Grund hierfür ist die Generation des Pulses, der durch das Umladen von Kapazitäten erzeugt wird. Ist die Schaltaktivität hoch, so trägt neben den Anteilen aus der Taktverteilung auch das Umladen der internen sowie der externen Lastkapazität bei. Da die externe Lastkapazität im Vergleich zu allen anderen Kapazitäten deutlich größer ist, verringert sich das Verhältnis der Energieaufnahme von MS-FF und P-FF. Selbst für hohe Schaltaktivitäten bleibt die Energieaufnahme des P-FFs gegenüber dem Standard MS-FF deutlich erhöht, da die internen Kapazitäten des P-FF durch die Verwendung von treiberstarken Transistoren deutlich höher sind als die des MS-FFs.
Im Fall des untersuchten ARM926 ergibt eine Abschätzung unter Berücksichtigung aller ersetzten FF Typen eine Zunahme der von allen Flip Flop Zellen verbrauchten Energie um ca. 7-8%. Nimmt man an, dass 20% der Gesamtenergie von den Flip Flop Zellen in der Schaltung verbraucht wird, so ist eine Gesamtenergieerhöhung von weniger als 2% zu erwarten.

Tabelle 6.5: Zusammenfassung für den Einsatz von P-FFs in geschwindigkeitskritischen Pfaden eines 90nm ARM926.

$\Delta t_{Path,max}$	$\Delta t_{Path,max}$ $S_2 = const.$	S_2	E_{Dyn}	Fläche
-7.2%	-6.3%	2.4x	+2%	+1%

Einfluss auf den Flächenbedarf

Der für mehr als 20 verschiedene Flip Flops (Funktion und Treiberstärke) berechnete Mittelwert für den Zuwachs der Einzelgatterfläche liegt bei 28%. Auch hier trägt vorwiegend der interne Pulsgenerator zum Flächenzuwachs bei. Da jedoch nur MS-FFs in den geschwindigkeitskritischen Pfaden ersetzt werden, verringert sich der Flächenzuwachs auf Schaltungsebene auf ca. 1%. Durch die um die Pulsbreite erhöhte Hold-Zeit ist zusätzliches Hold-Zeit Fixing durch Einfügen von Hold-Zeit Buffer Zellen erforderlich. Unter Berücksichtigung der Verteilung aller Hold-Zeit kritischen Pfade (siehe Kapitel 4.1) steigt der Flächenbedarf um vernachlässigbare 0.1-0.2%, so dass der Gesamtflächenzuwachs bei ungefähr 1% liegt. Die Verwendung breiterer Pulse führt zu einem exponentiellen Anstieg der für Hold-Zeit Fixing benötigten Fläche, so dass die Anzahl Hold-Zeit kritischer Pfade die maximale Pulsbreite für einen kosteneffizienten Einsatz von P-FF in geschwindigkeitskritischen Pfaden limitiert.
Ebenso wie beim Einsatz von low-V_T Gattern in geschwindigkeitskritischen Pfaden erhöht auch der selektive Einsatz von P-FFs den Sensitivitätsfaktor. In diesem Fall beträgt der auf die ursprüngliche Schaltung normierte Sensitivitätsfaktor $S_2 = 2.4$ bei einer maximalen Laufzeitreduzierung auf Entwurfsebene von $\Delta t_{Path,max} = 7.2\%$. Setzt man die Randbedingung einer konstanten Schaltungssensitivität, so kann als zusätzliche zeitliche Sicherheitsmarge nur 6.3% anstatt $\Delta t_{Path,max} = 7.2\%$ genutzt werden.

Tabelle 6.5 fasst die Ergebnisse der selektiven Ersetzung von MS-FFs durch P-FFs in geschwindigkeitskritischen Pfaden des ARM926 zusammen.

Globaler Einsatz von gepulsten Latches (Pulsed Latch Design)

Im vorherigen Abschnitt wird der Einsatz von gepulsten Flip Flops in geschwindigkeitskritischen Pfaden diskutiert. Die Ersetzung aller MS-FFs durch P-FFs hingegen ist aufgrund eines signifikanten Flächen- und Energiezuwachses nicht kosteneffizient.
Eine Möglichkeit die Kosten zu senken ist es, den Pulsgenerator aus der Zelle zu entfernen, indem ein externer Pulsgenerator verwendet wird, der als lokaler Clock Buffer (LCB) gleichzeitig mehrere Flip Flops/Latches mit einem Puls versorgt. Da gepulste Latches weniger Fläche benötigen als gepulste Flip Flops, wird im Folgenden der Einsatz gepulster Latches mit externem Pulsgenerator in eingebetteten Mikroprozessoren untersucht.

Bild 6.23 veranschaulicht die Ersetzungsstrategie beim Pulsed Latch Design. Es werden sowohl alle Flip Flop Zellen durch P-Latches als auch alle LCB durch Pulsgeneratoren ersetzt. Bild 6.24 zeigt die Schaltpläne des verwendeten Latches und Pulsgenerators.
Im Gegensatz zum gepulsten Flip Flop ist der Propagationspfad für einen $0 \rightarrow 1$ und $1 \rightarrow 0$ gleich. Während des Pulses öffnen die Transistoren N8/P9 den Propagationspfad zum Speicherknoten und die Transistoren N11/P12 brechen die Rückkopplung des Speicherknotens auf. Im Vergleich zum Standard Master-Slave Flip Flop erhöht sich die Clock-Q Laufzeit um bis zu 10%, da mit dem C^2MOS Inverter (P8,P9,N8,N9) eine zu-

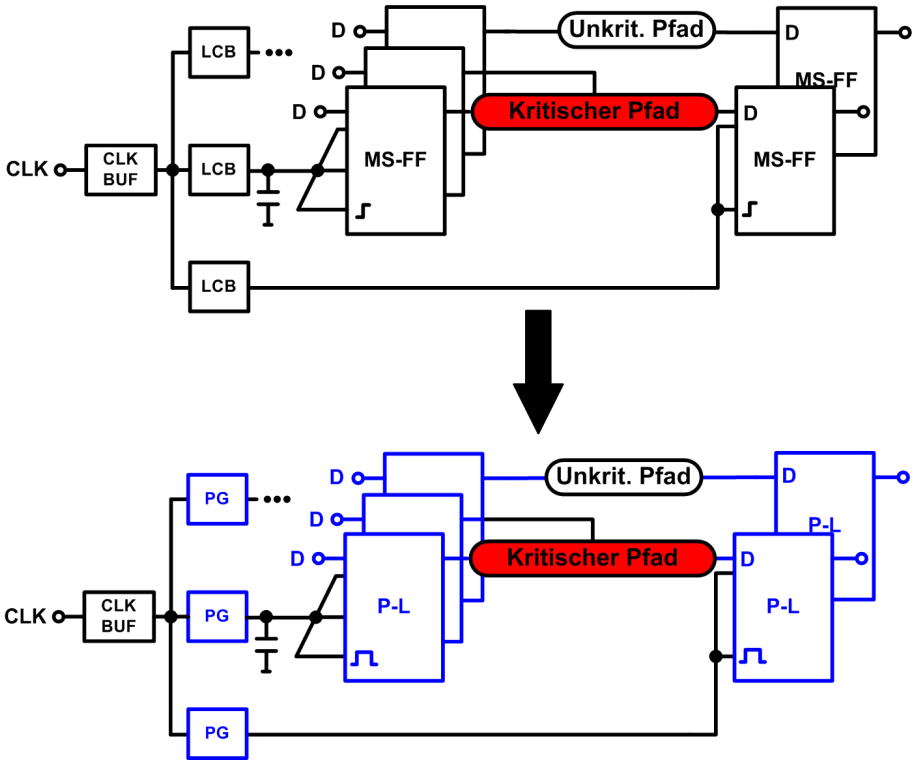
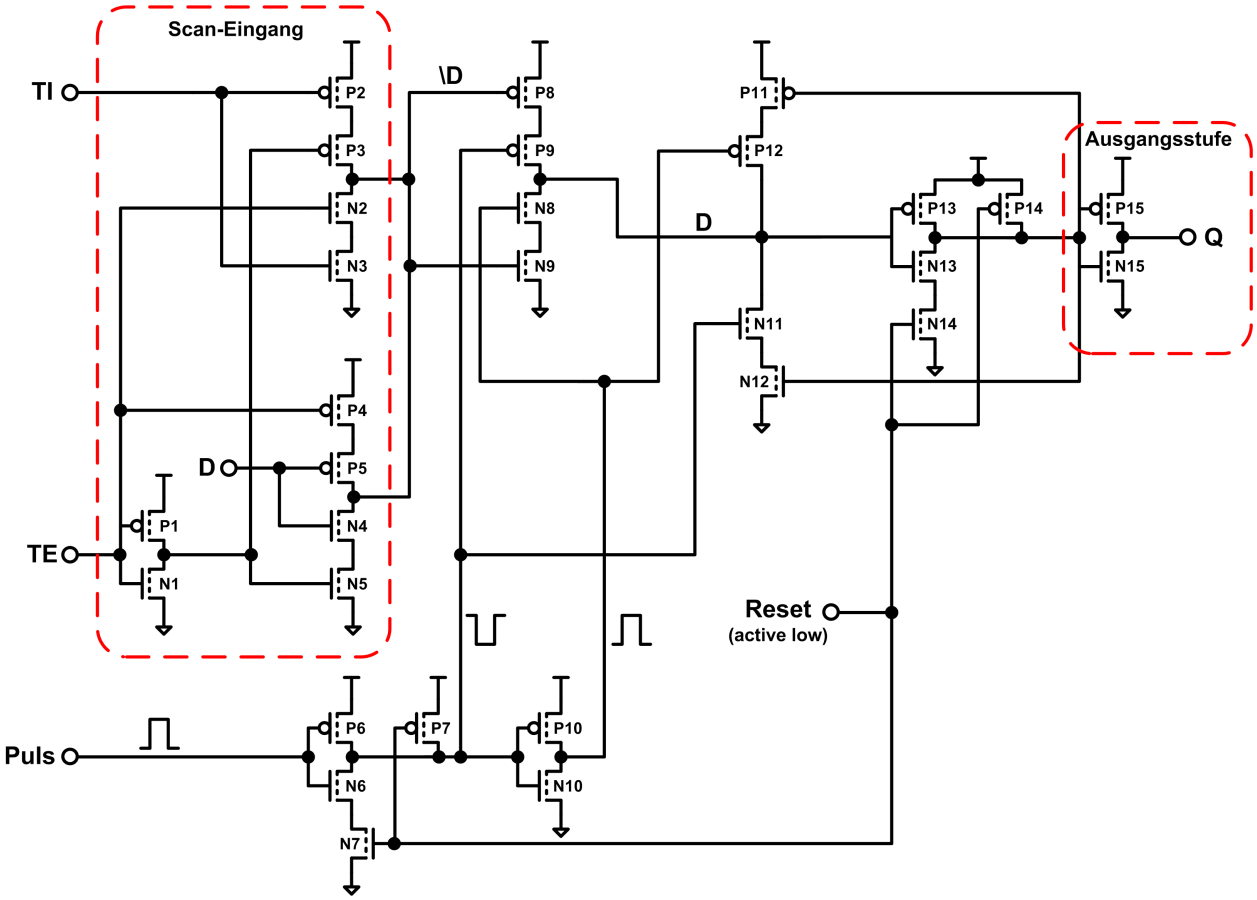


Bild 6.23: Schematische Darstellung der Ersetzungsstrategie.

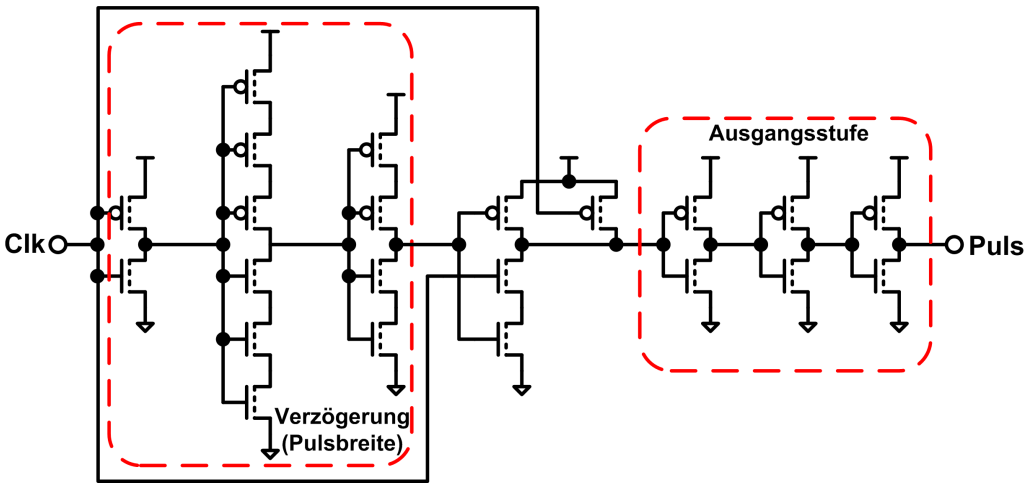
sätzliche Logikstufe im Propagationspfad liegt. Dieser Anstieg stellt kein Problem dar, da alle Flip Flop Zellen der Schaltung ersetzt werden und somit die Setup-Zeit aller FFs deutlich reduziert wird. Berücksichtigt man beide Effekte, so ist stets eine verringerte Gesamtlaufzeit zu erwarten. Für die kommenden Untersuchungen wird die Setup-Zeit des gepulsten Latches auf $t_{SU} = 0ps$ gesetzt.

Im Gegensatz zu gatterinternen Pulsgeneratoren, die eine feste Ausgangslast treiben, variiert die Belastung des externen Pulsgenerators aufgrund von variierenden Leitungslängen und unterschiedlicher Anzahl zu versorgender Latch-Zellen. Aus diesem Grund ist es insbesondere für synthetisierbare Semicustom Designs sehr wichtig, einen zuverlässigen Full-Swing Puls unter allen möglichen Prozess- und Betriebsbedingungen zu gewährleisten.
Die untere Grenze der Pulsbreite wird vorwiegend vom Tiefpassverhalten der Logik bestimmt. Je schmaler der Puls, desto stärker die Pulsverformung. Für Pulsbreiten unter 3 FO4 Laufzeiten konnte nicht für alle Betriebs- und Prozessbedingungen ein Full-Swing Puls sichergestellt werden, so dass für eine Pulsverteilung diese Pulsbreite als untere Grenze gesehen werden kann. Um diesen Effekt auch in den Standardzellen zu berücksichtigen, wird im Vergleich zu herkömmlichen LCBs die Treiberstärke der Ausgangsstufe um 30% erhöht, d.h. für gleiche Lasten ist die Ausgangsstufe des Pulsgenerators um 30% größer als die der ursprünglichen LCBs.
Die obere Grenze der Pulsbreite ist vom Aufwand für zusätzliches Hold-Zeit Fixing abhängig. Je größer die Pulsbreite desto mehr Pfade werden Hold-Zeit kritisch. Die maximale Pulsbreite für einen kosteneffizienten Einsatz von gepulsten Latches wird durch den zusätzlichen Flächen- und Energiebedarf bestimmt und ist daher abhängig von der zu entwerfenden Schaltung. Für die Untersuchungen des ARM926 wird die Verzögerungskette des Pulsgenerators so gewählt, dass ein Puls der Breite von 5 FO4 Laufzeiten verteilt

Bild 6.24: Gepulstes Latch mit externem Pulsgenerator.



(a) Schaltplan des gepulsten Latches mit Scan und asynchronem Reset.



(b) Pulsgenerator zur lokalen Verteilung des Pulses.

wird. Die effektive Pulsbreite, die für dynamisches Time-Borrowing zur Verfügung steht, beträgt ca. 3 FO4, da ca. 2 FO4 Laufzeiten für die Propagation des Datensignals vom Dateneingang bis zum Speicherknoten benötigt werden. Damit liegt die effektive Pulsbreite im Bereich der Pulsbreite des im vorherigen Abschnitt besprochenen P-FFs.

Einfluss auf die Laufzeitschwankung

Da die Pulsbreite durch ein Verzögerungselement bestimmt wird, ist die Anzahl an Logikstufen bei geringen Pulsbreiten klein. Um Energie und Fläche zu sparen werden kleine Transistoren verwendet, so dass der Mittelungseffekt statistischer Schwankungen für das Verzögerungselement gering ist. Monte Carlo Simulationen mit globaler und lokaler Parameterschwankung zeigen bei nomineller Versorgungsspannung eine relative Schwankungsbreite des Pulses von $\sigma_{t_{puls}}$=9.6%. Die entsprechende Clock-Q Laufzeit des Latches schwankt nur um $\sigma_{t_{Clk-Q}}$=8%. Die Schwankung der Pulsbreite korreliert jedoch aufgrund der globalen Prozessschwankung stark mit der Clock-Q Laufzeitschwankung. Eine genaue Untersuchung der Monte Carlo Simulationen zeigt einen unkorrelierten Schwankungsanteil beider Schwankungen von ρ =20%. Nach folgender Formel [190] ergibt sich für die Varianz der Differenz zweier Zufallsvariablen X und Y

$$\sigma^2[X-Y] = \sigma^2[X] + \sigma^2[Y] - 2\rho\sqrt{\sigma^2[X]\cdot\sigma^2[Y]} \tag{6.11}$$

Für den 1-Sigma Fall der Differenz aus t_{Clk-Q} und t_{puls} bei einer Pulsbreite von 5 FO4 Laufzeiten ergibt sich eine Laufzeit von ca. 0.3 FO4. Bei Clock-Q Laufzeiten von 3-4 FO4 Laufzeiten stellt dieser Schwankungsanteil keine wesentliche Reduzierung der inhärenten Race-Immunity dar. Zur Vermeidung von Hold-Zeit Fehlern ist daher die genaue Bestimmung der maximalen Pulsbreite unter globalen Prozessschwankungen ausschlaggebend.
Für den hier gewählten nominellen Betriebspunkt von $t_{SU} = 0ps$ stellt die Pulsbreitenschwankung kein Problem dar. Die Schwankung der Pulsbreite limitiert jedoch die maximale zeitliche Größe für Time-Borrowing. Da im gewählten Betriebspunkt nur dynamisches Time-Borrowing durch eine Pulsverkürzung beeinflusst wird, schlägt sich die Schwankung der Pulsbreite nicht auf die nominell zu erzielende zeitliche Marge nieder.

Im Vergleich zum LCB erhöht sich die Laufzeit des Pulsgenerators um ca. 1.5 FO4 Laufzeiten. Diese Erhöhung hat keinen signifikanten Einfluss auf den Clock Jitter des Taktverteilungsnetzes und kann daher vernachlässigt werden. Die relative Laufzeitschwankung von LCB und PG zeigen keine zu berücksichtigenden Unterschiede, so dass der lokale Clock Skew durch die Verwendung von Pulsgeneratoren als LCB-Zelle nur unwesentlich verändert wird.

Einfluss auf die Energieaufnahme

Um den Anstieg der Clock-Q Laufzeit gegenüber dem ursprünglich verwendeten Master-Slave Flip Flops gering zu halten, werden im gepulsten Latch die maximalen Transistorweiten, die durch die Standardzellenhöhe beschränkt wird, bestmöglich ausgenutzt. Es ergibt sich trotz signifikantem Flächengewinn eine geringe Ersparnis der Energieaufnahme. In Bild 6.25 ist die simulierte Energieaufnahme von Master-Slave FF und gepulstem Latch in Abhängigkeit der Schaltaktivität gezeigt. Bei hoher Aktivität ist eine Energieersparnis von ca. 10% zu erwarten.
Berücksichtigt man neben den Flip Flop Zellen auch die Energieaufnahme von LCB und Pulsgenerator, so ergibt sich für eine Anzahl von ca. acht gepulsten Latches pro Pulsgenerator eine zur Anordnung aus LCB und MS-FF identische Energieaufnahme. Werden mehr als acht P-L von einem Pulsgenerator versorgt, so sinkt die Energieaufnahme im

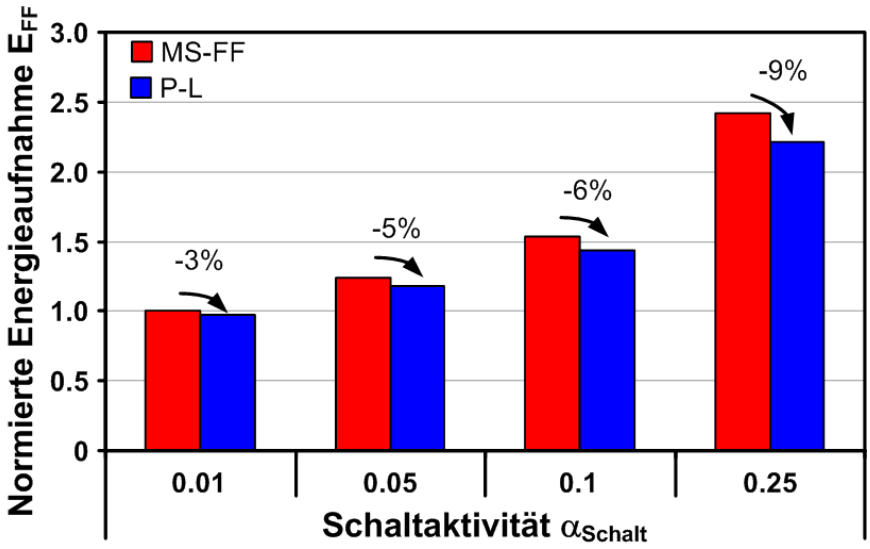


Bild 6.25: Simulierte Energieaufnahme von MS-FF und P-L in 65nm CMOS in Abhängigkeit der Schaltaktivität ($V_{DD} = V_{DD}^{nom}$, T=27°C).

Vergleich zum ursprünglichen Design. Da im analysierten ARM926 Produktdesign durchschnittlich 27 MS-FFs von einem LCB versorgt werden, ist ein zwischen 5-10% verringerter Energiebeitrag für die Kombination aus von PG und P-L zu erwarten. Um eine sichere Verteilung eines Full-Swing Pulses gewährleisten zu können, ist die räumliche Anordnung von LCB und den zu treibenden FF/Latch Zellen (Relative Placement) empfehlenswert. Gleichzeitig erhöht sich somit die Anzahl an Latch Zellen pro PG.

Einfluss auf den Flächenbedarf

Auf Gatterebene nimmt die Transistoranzahl eines gepulsten Latches im Vergleich zum MS-FF signifikant ab. Da im Semicustom Schaltungsentwurf Flip Flop Zellen einen Scan-Eingang zum Test der Schaltung erhalten, müssen auch die gepulsten Latches mit Scan-Eingängen versehen werden. Dies reduziert den ursprünglichen relativen Flächengewinn. Mittelt man den Flächenunterschied von über 20 verschiedenen Flip Flop Typen und gepulsten Latches, so resultiert ein mittlerer Flächengewinn von ca. 11%. Der zellspezifische Austausch von insgesamt ca. 20000 MS-FFs des untersuchten ARM926 durch P-Ls führt zu einer relativen Flächenersparnis von 16%. Da die ursprünglich für alle MS-FFs benötigte Fläche ca. 49% beträgt, liegt der Flächengewinn auf Mikroprozessorkernebene bei 8%.
Die 768 LCB Zellen des Taktbaums müssen durch größere Pulsgeneratoren ersetzt werden. Da der Flächenbeitrag der LCB zur Gesamtfläche gering ist, führt der Einsatz von Pulsgeneratoren zu einer Flächenerhöhung um 1%.
Zur Kompensation der erhöhten Hold-Zeiten werden Hold-Zeit Buffer eingefügt, so dass für alle Hold-Zeit kritischen Pfade vergleichbar große zeitliche Hold-Zeit Margen verfügbar sind wie im ursprünglichen Design. Im Gegensatz zum Einsatz von P-FF in geschwindigkeitskritischen Pfaden ist der Aufwand für Hold-Time Fixing beim globalen Einsatz gepulster Latches höher. Für die verwendete Pulsbreite von 5 FO4 Laufzeiten liegt der zusätzliche Flächenaufwand für Hold-Time Buffer bei 1.4% der Gesamtfläche.
Es ergibt sich für den untersuchten ARM926 ein Flächengewinn (ohne Cache) von 5.6%.

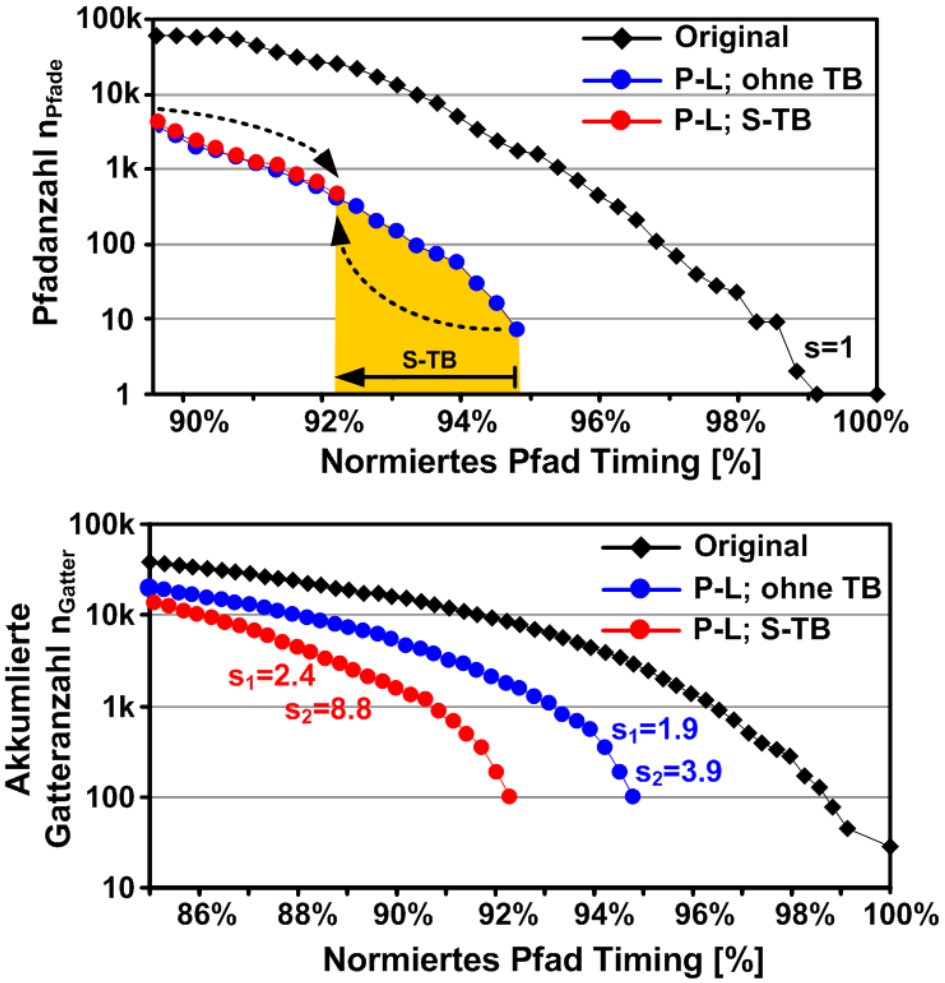


Bild 6.26: Pfad- und Gatterspektrum des ARM926 für das ursprüngliche MS-FF Design und nach globalem Einsatz von P-Ls.

Einfluss auf Pfadlaufzeiten

Bild 6.26 zeigt Gatter- und Pfadverteilung des ARM926 vor und nach dem globalen Einsatz von gepulsten Latches.

Für den bevorzugten Betriebspunkt (ohne Time-Borrowing) ist eine zeitliche Marge von 5% erkennbar. Durch die 'Multi-Stage Analyse', die für die Untersuchung der Pfadtopologien in Kapitel 5.1 erforderlich ist, kann für jeden Pfad die maximal mögliche Zeit für statisches Time-Borrowing bestimmt werden. Während die Laufzeit des kritischen Pfades reduziert wird, erhöht sich die Laufzeit eines bisher sub-kritischen Pfades. In Bild 6.26 ist der Bereich für statisches Time-Borrowing markiert. Beispielhaft für alle anderen Pfade ist für den kritischsten aller Pfade der entsprechende sub-kritische Pfad gezeigt, dessen Laufzeit mit zunehmendem Time-Borrowing ansteigt. Dieser Pfad limitiert somit das Potential von statischem Time-Borrowing hinsichtlich der maximal zu erreichenden Geschwindigkeitsmarge.

Alle anderen Pfade im markierten Bereich haben ebenfalls die Möglichkeit zum statischen Time-Borrowing, wobei die Größenordnung von 0.5% bis zu ca. 6% variiert. Im vorliegenden Fall bleibt der ursprünglich kritischste Pfad auch nach statischem Time-Borrowing der geschwindigkeitslimitierende Pfad.

Bild 6.27 veranschaulicht den Aufbau von Pulsgenerator und gepulsten Latches und die

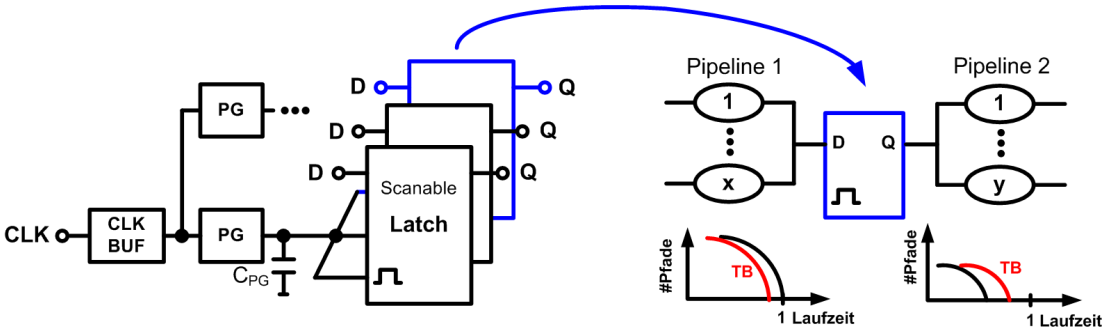


Bild 6.27: Veranschaulichung von Time-Borrowing in Schaltungen mit gepulsten Latches.

Tabelle 6.6: Übersicht über die Ergebnisse des globalen Einsatzes von P-Latches in einem 90nm ARM926.

	$\Delta t_{Path,max}$	$\Delta t_{Path,max}$ für $S_2 = const.$	S_2	E_{Dyn}	Fläche
Ohne TB	-5.2%	-3.8%	3.9x	ca. -5%	-5.6%
S-TB	-7.7%	-5.8%	8.8x	ca. -5%	-5.6%

von der Pfadtopologie abhängige Anwendung von Time-Borrowing. Da lediglich 2.5% statisches Time-Borrowing erforderlich ist, um die maximale Geschwindigkeitsmarge zu erreichen, sind nur geringfügige Änderungen des ARM926 Pfad- und Gatterspektrums zu erkennen. Für homogener verteilte Pfadlaufzeiten ist eine stärkere Veränderung der Spektren und somit auch ein deutlich erhöhter Sensitivitätsfaktor s zu erwarten. Hier ist die große Bedeutung der Multi-Stage Analyse bei der Implementierung von Time-Borrowing Techniken erkennbar.

Der Ansatz, MS-FFs global durch gepulste Latches zu ersetzen, zeigt für einen ARM926 mit einer Geschwindigkeitsmarge von 5%, einer um ca. 5.5% reduzierten Fläche, geringfügiger Energieersparnis und einer im Vergleich zu selektivem low-V_T Einsatz moderaten Erhöhung des Sensitivitätsfaktors s gute Ergebnisse. Um Kosten und Nutzen für kommende Generationen eingebetteter Mikroprozessoren abzuschätzen werden die Ergebnisse des ARM926 Produktdesigns mit Hilfe des Mikroprozessormodells aus Kapitel 4.2 für die ARM Prozessoren ARM1176 und Cortex A8 skaliert.
Für tieferes Pipelining ist eine deutlich erhöhte zeitliche Sicherheitsmarge von bis zu 8% zu erwarten. Aufgrund der erhöhten Anzahl an Flip Flops steigt der Flächenanteil des Taktverteilungsnetzes und somit auch der zusätzliche Flächenaufwand für Pulsgeneratoren und Hold-Zeit Buffer. Bild 6.28 zeigt die auf das Mikroprozessormodell basierenden Ergebnisse für ARM1176 und ARM Cortex A8. Zusätzlich ist das zeitliche Budget für die Verwendung des gepulsten Latches aus Bild 6.24 gezeigt, das bei der Wahl des Betriebspunkts $t_{SU} = 0ps$ grundsätzlich für statisches und dynamisches Time-Borrowing zur Verfügung steht. Im Falle des ARM Cortex A8 kann bis zu 40% der Timing Unsicherheit durch dynamisches Time-Borrowing kompensiert werden. Tabelle 6.7 fasst die wichtigsten Ergebnisse für den Einsatz von gepulsten Latches in eingebetteten Mikroprozessoren zusammen.

Tabelle 6.7: Mikroprozessormodellbasierte Kosten-Nutzen Analyse für den globalen Einsatz gepulster Latches in eingebetteten Mikroprozessoren.

		ARM926 Produktdesign	**ARM1176**	**ARM Cortex A8**
			μP-**Modell**	
Laufzeit	Ohne TB	-4.7%	-7%	-8%
	S-TB	-7.5%	-9%	-11%
Gesamtfläche		-5.6%	-4%	-1%
	Pulsgenerator	+1.0%	+2%	+3%
	HD-Buffer	+1.4%	+2%	+4%
TB Budget		4.8%	ca. 7%	ca. 8%

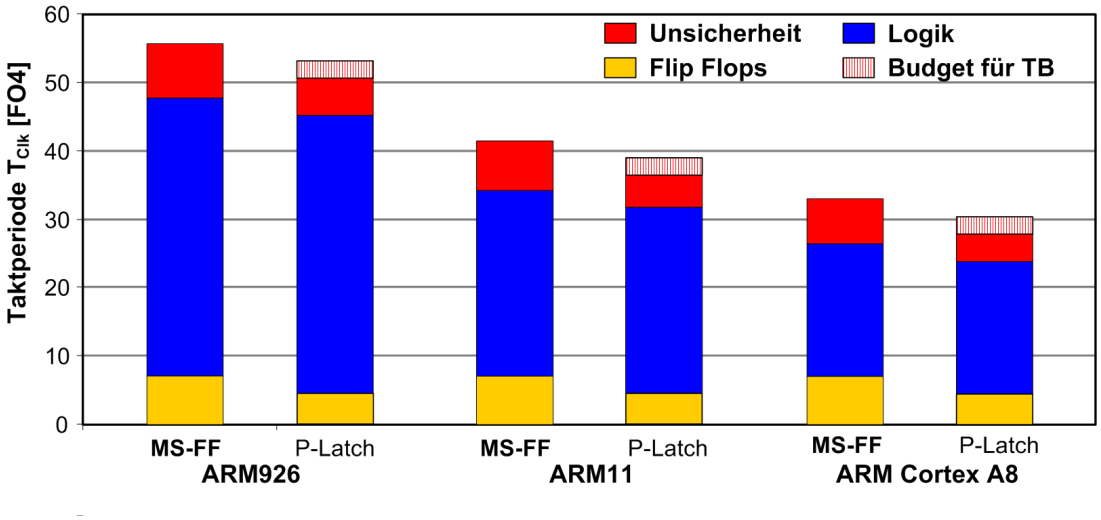


Bild 6.28: Ergebnisse des Mikroprozessormodells für den globalen Einsatz von gepulsten Latches in ARM926, ARM1176 und ARM Cortex A8.

6.2.5 Einfluss der Techniken auf die Schaltungssensitivität

In den vorangegangenen Abschnitten werden verschiedene Maßnahmen diskutiert, die es ermöglichen, präventive, zeitliche Sicherheitsmargen zur Kompensation von Laufzeitschwankungen zu erzeugen. Neben den Kosten hinsichtlich des Energiebedarfs und Flächenaufwands werden die erzielte zeitliche Sicherheitsmarge und die in Kapitel 5.2 definierte Schaltungssensitivität als Bewertungsgrößen herangezogen.
Um einen besseren Überblick über die verschiedenen Maßnahmen hinsichtlich ihres Einflusses auf das Laufzeitverhalten der Schaltung zu geben, werden in Bild 6.29 die zeitliche Sicherheitsmarge sowie die Erhöhung der Sensitivitätsfaktoren s_1, s_2 [122] gleichzeitig dargestellt.
Um den Anstieg der kritischen Hardware im geschwindigkeitskritischen Bereich zu zeigen wird in der oberen Darstellung der normierte Schaltungssensitivitätsfaktor s_1 aus Gleichung 5.2 verwendet. Es zeigt sich, dass die Verwendung von gepulsten Latches und Flip Flops im Vergleich zum Einsatz von low-V_T Gattern zu geringerem Anstieg des Sensitivitätsfaktors bei gleichzeitig erhöhtem Geschwindigkeitsgewinn führt. Die low-V_T Optionen zeigen daher im Vergleich zu P-FF und P-L einen negativen Einfluss auf die Sensitivität der Schaltung gegenüber WID Prozess- und Umgebungsvariationen. Es muss jedoch berücksichtigt werden, dass die hohe topologische Korrelation zu einer deutlichen Ge-

schwindigkeitserhöhung für vernachlässigbar geringe Erhöhung der Leckströme führt. Für Schaltungen, die bereits ausreichende zeitliche Sicherheitmargen besitzen, d.h. eine Erhöhung der Schaltungssensitivität vertretbar ist, können die low-V_T Optionen zur weiteren Optimierung der Schaltung hinsichtlich Geschwindigkeit und Energieaufnahme eingesetzt werden.
Die Darstellung des normierten Schaltungssensitivitätsfaktors s_2 nach Gleichung 5.4 unterscheidet sich zum Sensitivitätsfaktor s_1. Da dieser Faktor auch die Zuordnung der Gatter zum jeweiligen Pfad-Timing gewichtet, ist für den selektiven Einsatz von low-V_T Gattern und gepulsten Latches mit statischem Time-Borrowing, die zum Aufbau einer Timing-Wall führen, ein starker Anstieg des Sensitivitätsfaktors im Vergleich zum Original-Design zu sehen. Nach wie vor zeigt sich für alle low-V_T Optionen bei gleicher zeitlicher Sicherheitsmarge ein gegenüber P-FF und P-L Design deutlich erhöhter Sensitivitätsfaktor. Die im Vergleich zu s_1 veränderte Anordnung der Sensitivitätsfaktoren s_2 von P-FF und P-L trotz gleicher Setup-Zeiten $t_{SU} = 0ps$ resultieren aus den veränderten Gatterspektren, die sich aufgrund unterschiedlicher Clock-Q Laufzeiten von P-FF und P-L ergeben.
Neben einem relativ geringen Anstieg der Schaltungssensitivitätsfaktoren beim Einsatz von P-FFs und P-Ls eröffnen diese Techniken eine zusätzliche Kompensationsmöglichkeit von dynamischen Laufzeitschwankungen. Durch dynamisches Time-Borrowing (D-TB), d.h. zusätzliche zeitliche Elastizität zwischen verschiedenen Pipelinestufen, können Laufzeitschwankungen z.B. aus Clock Jitter, lokalem IR-Drop etc. kompensiert werden. Die maximale Größenordnung dieser zeitlichen Elastizität ist für P-L und P-FF durch gelbe, horizontale Balken dargestellt. Der genaue Einfluss von D-TB auf die maximale zeitliche Sicherheitsmarge erfordert die Kenntnisse aller Schaltabhängigkeiten und Bit-Pattern Folgen für jede einzelne dynamische Laufzeitänderung. Somit ist eine Quantifizierung der effektiv wirkenden zeitlichen Sicherheitsmarge aufgrund unvorhersagbarer Einflussgrößen nicht möglich. Diese zusätzliche Elastizität schafft nicht nur dynamisch veränderbare zeitliche Sicherheitsmargen sondern vereinfacht auch die Implementierung von Adaptionstechniken, da z.B. die Adaption der Versorgungsspannung durch die unterschiedliche Ausbreitungsgeschwindigkeit des angepassten Spannungspegels Timing Unsicherheit wie z.B. Clock Jitter verursachen kann.

Da mit tieferem Pipelining die Anzahl der kritischen Pfade steigt und aufgrund der erhöhten Anzahl an Pipelinestufen ein verringerter topologischer Korrelationsfaktor zu erwarten ist, wird mit einem Anstieg der Kosten für den selektiven Einsatz von low-V_T Gattern gerechnet. Im Gegensatz dazu zeigt sich, dass die zu erreichende zeitliche Sicherheitsmarge beim Einsatz von gepulsten Flip Flops und Latches ansteigt. Zusammen mit niedrigerem Schaltungssensitivitätsfaktor und zusätzlicher zeitlicher Elastizität sind selektiver Einsatz von P-FFs in geschwindigkeitskritischen Pfaden und die globale Verwendung von gepulsten Latches attraktive Maßnahmen zur Kompensation von Laufzeitschwankungen.

Die Gegenüberstellung von zeitlicher Sicherheitsmarge und normiertem Schaltungssensitivitätsfaktor erleichtert die Auswahl geeigneter präventiver Kompensationstechniken. Ferner ermöglicht die Bewertung der Schaltungssensitivität eine Optimierung hinsichtlich Robustheitsaspekten bei der Implementierung verschiedener Kompensationstechniken wie z.B. des selektiven low-V_T Einsatzes. Hier kann durch zusätzliche low-V_T Gatter das Gatterspektrum weiter verformt und der Sensitivitätsfaktor reduziert werden.
Die Untersuchungen präventiver Kompensationstechniken zeigen, dass der Schaltungssen-

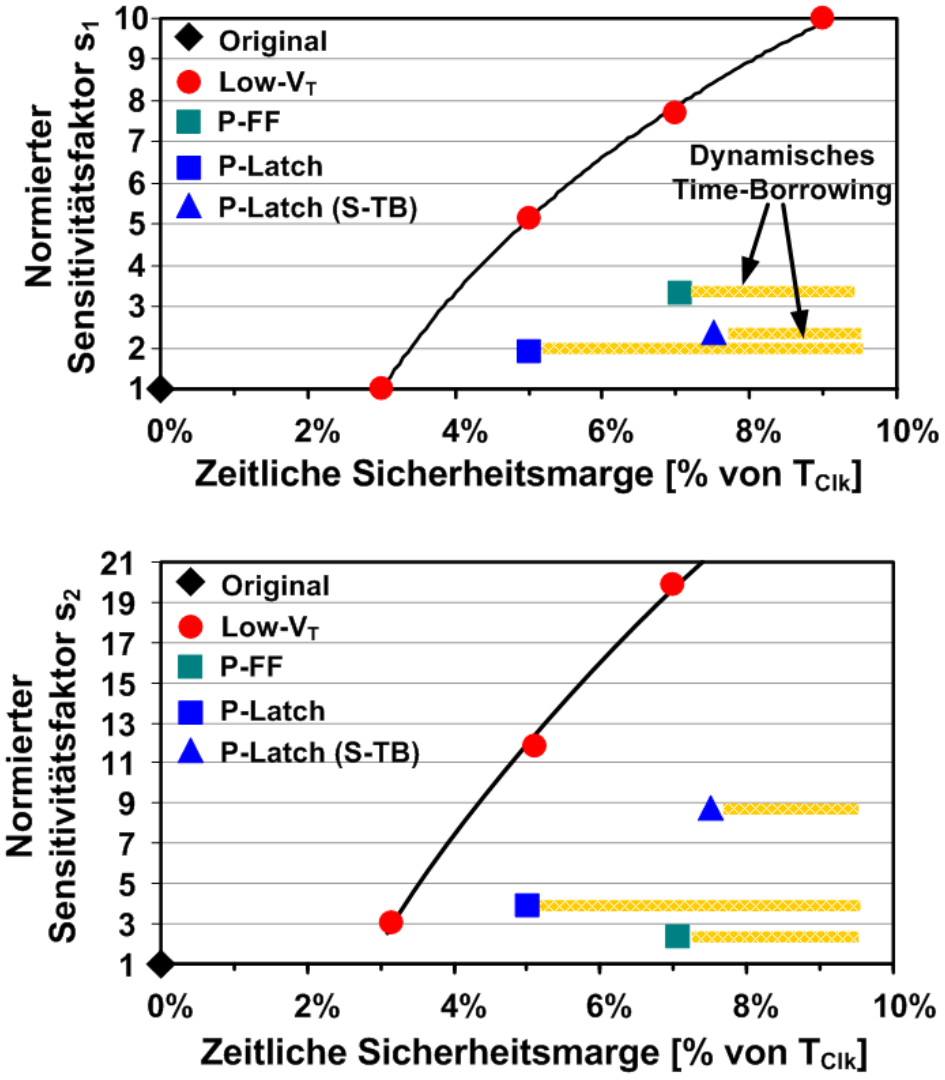


Bild 6.29: Zeitliche Sicherheitsmarge und normierter Schaltungssensitivitätsfaktor der einzelnen präventiven Kompensationstechniken für den ARM926 in 90nm CMOS.

sitivitätsfaktor ein geeignetes Maß ist, um die strukturellen und topologischen Veränderungen durch die Implementierung schaltungstechnischer Maßnahmen und deren Einfluss auf die Robustheit der Schaltung gegenüber Variations- und Alterungseffekten zu quantifizieren. Im folgenden Abschnitt wird die Schwierigkeit der Validierung von Robustheitskenngrößen diskutiert, sowie die Korrelation von Laufzeitschwankung und Schaltungssensitivitätsfaktor unter der Annahme von rein statistischen Laufzeitschwankungen gezeigt.

6.2.6 Validierung des Sensitivitätsfaktors als Robustheitsmaß

In diesem Abschnitt werden die Probleme der Validierung von Robustheitskenngrößen diskutiert und die Gültigkeit des gewichteten Schaltungssensitivitätsfaktors unter der Annahme rein statistischer Laufzeitschwankungen gezeigt.
Eine Validierung von Robustheitskenngrößen ist generell nicht ohne bestimmte Randbedingungen und Annahmen möglich, da realistische Verteilungen von variationsbedingten Laufzeitschwankungen - insbesondere für komplexe Schaltungen, wie die untersuchten ARM Mikroprozessoren - nicht bestimmt werden können. Daher kann nicht gezeigt werden, dass sich unter allen Kombinationsmöglichkeiten der einzelnen Variationsquellen das Robustheitsmaß gemäß den maximalen Pfadlaufzeiten verändert und somit als gültig zu betrachten ist. Die folgende Auflisting fasst die entscheidenden Einflussfaktoren, die eine Validierung von Robustheitskriterien erschweren bzw. unmöglich machen, zusammen.

- Die hohe Anzahl an systematischen und pseudo-statistischen Variationen, die den größten Beitrag zur variationsbedingten Laufzeitschwankung stellen, machen die Bestimmung einer realistischen Laufzeit z.B. durch Simulation in der Praxis unmöglich. Insbesondere die Abhängigkeit der einzelnen Variationsquellen von einer hohen Zahl an nur schwer oder sogar unbestimmbarer Einflussfaktoren, wie in folgender Auflistung zusammengefasst, stellt hier das größte Problem dar.
  - Prozessvariationen: z.B. Belichtungszeit und -dosis, Dotierstoffkonzentration, Chemical Mechanical Polishing, Temperaturhomogenität während des Rapid Thermal Anneal, Lage und Typ benachbarter Zellen, Zell-Layout, ...
  - IR-Drop: Versorgungsspannung, Taktfrequenz, Schaltaktivität, Zuleitungswiderstand, Lastkapazität am „Device Under Test“ sowie die Versorgungsspannung, Taktfrequenz, Schaltaktivität, Lastkapazität aller Subkomponenten eines Chips und die Antwortzeit (Response-Time) des Spannungsreglers
  - Crosstalk: Rise- und Fall-Time von Aggressor- und Victimleitung, Anzahl an Agressornetzen sowie deren Koppelkapzität zum Victimnetz, die statische Haltekapazität am Victimnetz und die Relative Signal Ankunftszeit (RSAT) für alle Netze der Schaltung, sowie die Versorgungsspannung
  - Temperatur: Umgebungstemperatur sowie Taktfrequenz, Schaltaktivität, Lastkapazität, Versorgungsspannung aller Subkomponenten eines Chips
  - Aging: Versorgungsspannung, Taktfrequenz, Temperatur, Betriebszeit
- Technologieabhängige Laufzeitsensitivität
- Topologische Korrelation innerhalb geschwindigkeitskritischer Pfade
- Die hohe Anzahl an kritischer Hardware und die damit verbundene hohe Anzahl an kritischen Pfaden (ca. 1e5-1e7), deren Laufzeit unter allen obigen Gesichtspunkten bestimmt werden muss

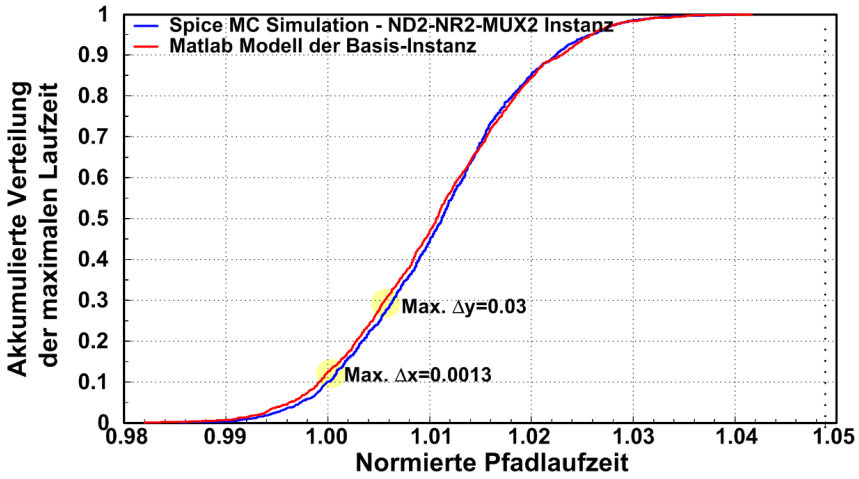
- False-Path Problematik (siehe Kapitel 5)
- Anzahl und Art der laufenden Anwendungen bzw. Prozesse

Es wird deutlich, dass eine Validierung des Schaltungssensitivitätsfaktors nur unter bestimmten Annahmen vorgenommen werden kann, da die Kenntnis über alle aufgelisteten Einflussgrößen sowie deren Kombination fehlt. Aus diesem Grund werden im Folgenden die Ergebnisse einer vereinfachten Schaltungssensitivitätsfaktor-Validierung unter der Annahme rein statistischer Laufzeitschwankungen gezeigt.

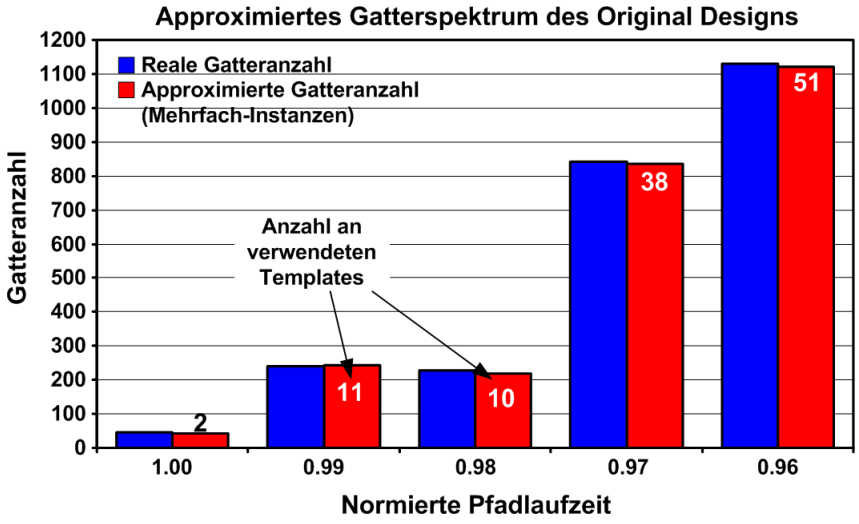
Stand der Technik bei der Beurteilung variationsbedingter Laufzeitschwankungen stellt die Monte-Carlo (MC) Analyse dar, die unter der Annahme rein statistischer Variationen als Referenz technische Akzeptanz findet. Aufgrund der Komplexität der Schaltung, d.h. die hohe Anzahl an Pfaden, die aus der starken Vernetzung aller Gatter resultiert, ergeben sich folgende Probleme für die Verwendung der MC-Analyse als Referenz zur Validierung des Schaltungssensitivitätsfaktors.

- Um ein Gatterspektrum zu erstellen muss jedes Gatter dem kritischsten aller Pfade zugeordnet werden. Dies erfordert die Sensibilisierung jedes einzelnen möglichen Pfades und eine individuelle Monte-Carlo Simulation zur Bestimmung der Laufzeit. Um das gleiche Gatterspektrum wie durch eine STA Analyse zu erhalten, ist es daher erforderlich, alle möglichen Signalpropagationspfade ausgehend von einem Registerelement zu betrachten, d.h. pro verändertem Bit-Pattern der Eingangsregister können 1 bis mehrere Tausend Pfade resultieren, deren Laufzeit analysiert werden muss, um das Gatterspektrum zu bestimmen. Dies erfordert offensichtlich einen untragbar hohen Rechenaufwand.
- Eine individuelle Simulation der Einzelpfade verhindert die Berücksichtigung topologischer Korrelationen.
- Der systematische Laufzeitunterschied von Titan Monte-Carlo Simulation und Statischer Timing Analyse liegt in der Größenordnung einzelner Variationsbeiträge (ca. 5% Laufzeitunterschied), was einen Vergleich beider Methodiken verhindert.
- Die Statistische Statische Timing Analyse (SSTA) wäre unter der Annahme rein statistischer Schwankungen prinzipiell als Referenz geeignet. Die in Kapitel 5 dargelegten Probleme der SSTA haben jedoch zur Folge, dass eine funktionierende SSTA Lösung für die Verifikation komplexer Schaltungen mit der erforderlichen Genauigkeit nicht zur Verfügung steht.

Aus den oben genannten Gründen ist es für die Validierung des Schaltungssensitivitätsfaktors erforderlich, zum einen den Rechenaufwand zu reduzieren, zum anderen die Vergleichbarkeit von Monte-Carlo Simulation und STA Ergebnissen zu gewährleisten. Deshalb wird für die Validierung des Schaltungssensitivitätsfaktors die folgende, hier erarbeitete Methodik angewandt.
Für die aus Bild 5.8(a) bekannte Testschaltung, die die wesentlichen Eigenschaften kritischer Pfade hinsichtlich ihres Verhaltens gegenüber Variationen nachbildet und einen realistischen Korrelationsfaktor aufweist, wird mit dem Analogsimulator Titan eine Monte-Carlo Analyse durchgeführt, um die maximale Laufzeit durch diese Gatteranordnung zu bestimmen. Parallel dazu wird die Gatteranordnung in Matlab als Summe unterschiedlicher, statistisch verteilter Basis-Laufzeiten modelliert. Die Genauigkeit dieses Ansatzes



(a) Vergleich von Titan-Simulation und Matlab Modell für eine Basisschaltung (Template).

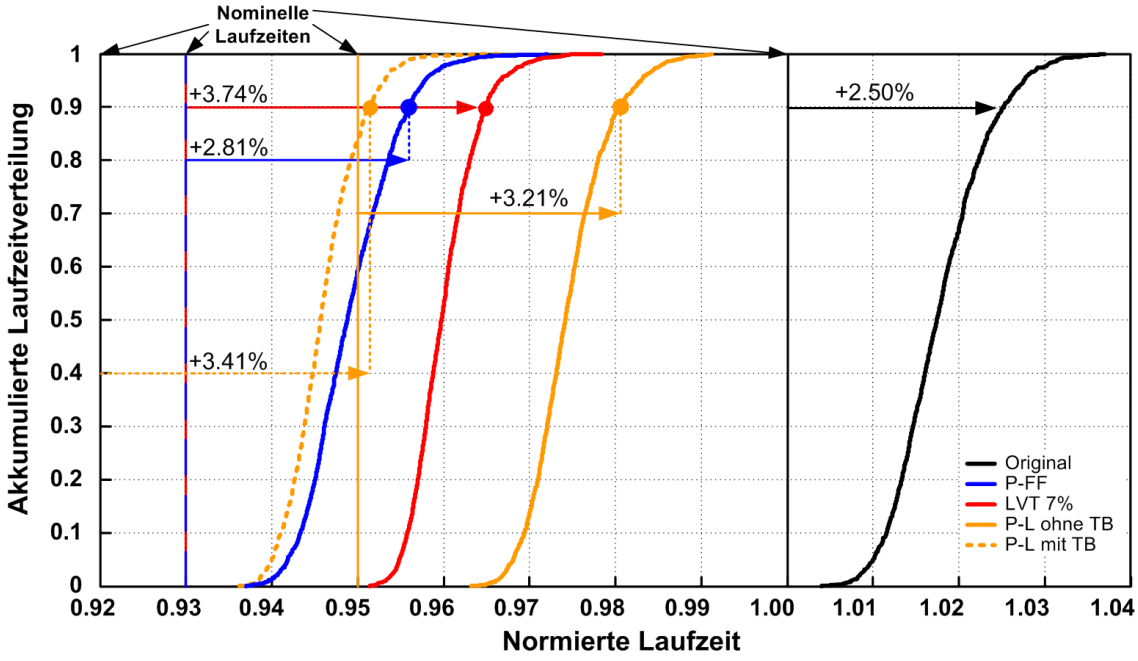


(b) Nachbildung der originalen ARM926 Gatterspektrums durch die Verwendung mehrere Templates (Mehrfach-Insatnzen).

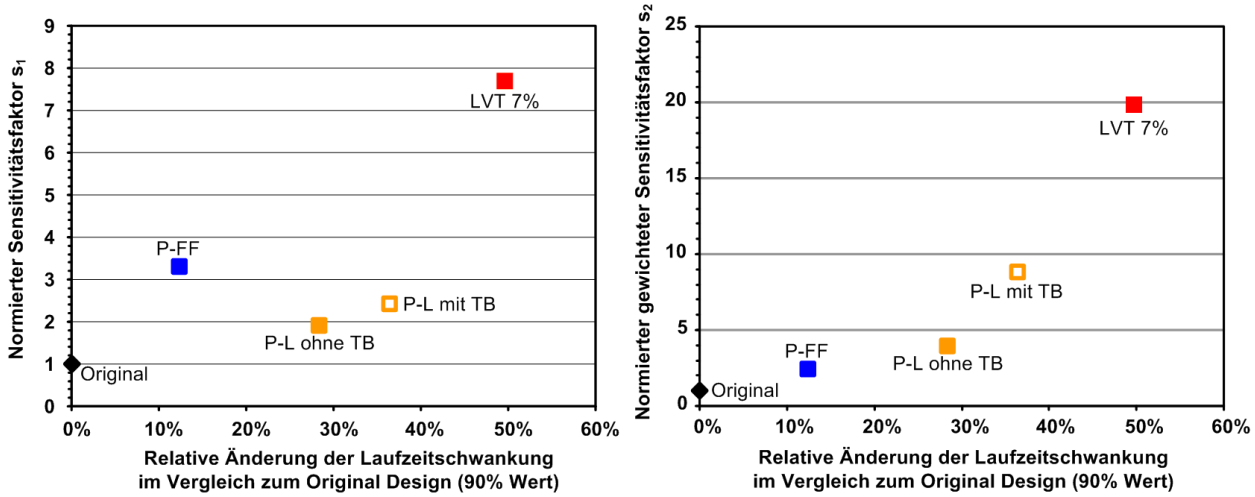
Bild 6.30: Modell-Genauigkeit und Nachbildung des Gatterspektrums durch die Verwendung von Templates als Mehrfach-Instanzen.

(Matlab Modells) wird durch die geringen Unterschiede der bestimmten maximalen Laufzeiten im Vergleich zur Analogsimulation in Bild 6.30(a) gezeigt. Diese Basisschaltung wird im Folgenden als Template bezeichnet und in einem erweiterten Matlab-Modell mehrfach verwendet (Mehrfach-Instanz), um das STA basierte Gatterspektrum nachzubilden. Dazu wird jeder Pfadlaufzeit eine bestimmte Anzahl an Templates zugewiesen, bis die reale Gatteranzahl für dieses Pfad-Timing möglichst genau getroffen wird. Die Verwendung der Templates als Basiseinheit resultiert in einem geringen Diskretisierungsfehler, der jedoch aufgrund der hohen absoluten Anzahl an Gattern vernachlässigt werden kann, wie in Bild 6.30(b) zu sehen ist.
Die nominelle Laufzeit der Templates wird gemäß den zugeordneten Pfadlaufzeiten angepasst. Die Laufzeiten der im Matlab-Modell verwendeten Basis-Lauffzeiten werden aus dem nominellen Punkt ausgelenkt (Normalverteilung), so dass sich für jedes Template eine maximale Pfadlaufzeit ergibt. Nun wird für alle Templates, die zur Nachbildung des Gat-

terspektrums verwendet werden, die maximale aller Laufzeiten bestimmt. Dieser Vorgang wiederholt sich mehrmals (MC-Analyse), so dass sich nach mehreren Tausend Durchläufen die Verteilung der maximalen Laufzeit für die gesamte Anordnung aus Mehrfach-Instanzen ergibt.
Dieser Ansatz ermöglicht es, eine statistische Methode mit dem STA Gatterspektrum einer Schaltung zu koppeln, so dass ein Vergleich zwischen Laufzeitverteilung und Schaltungssensitivitätsfaktor auf Basis der gleichen Ausgangslage, d.h. des gleichen Gatterspektrums erfolgen kann. So können einige der o.g. Schwierigkeiten bei der Validierung des Schaltungssensitivitätsfaktors umgangen werden.



(a) Simulierte Laufzeitverteilung (Matlab) auf Basis von Mehrfach-Instanzen.



(b) Vergleich von Schaltungssensitivitätsfaktor und simulierter Laufzeitschwankung (Matlab-Simulation)

Bild 6.31: Validierung des Schaltungssensitivitätsfaktors.

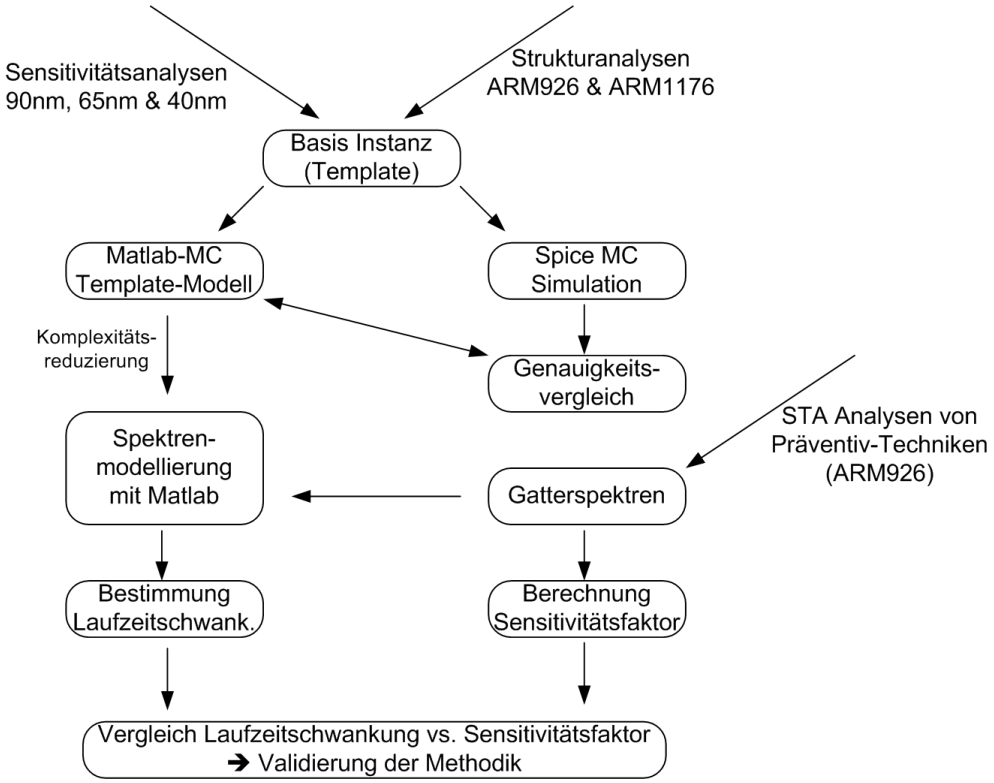


Bild 6.32: Methodik zur Validierung des Schaltungssensitivitätsfaktors.

Im Folgenden werden nun die Ergebnisse der Validierung des Schaltungssensitivitätsfaktor auf Basis der hier dargelegten Methodik diskutiert. Dazu werden für selektiven low-V_T Zellen und P-FF Einsatz in geschwindigkeitskritischen Pfaden, sowie den globalen Einsatz von gepulsten Latches die jeweiligen Gatterspektren durch die o.g. Mehrfach-Instanzen nachgebildet und die relative Zunahme des 90% Punktes der akkumulierten Laufzeitverteilung im Vergleich zum Original Design mit dem Schaltungssensitivitätsfaktors verglichen. Bild 6.31(a) zeigt die Ergebnisse der mit Matlab bestimmten Laufzeitverteilung basierend auf der Nachbildung der jeweiligen Gatterspektren.
In Bild 6.31(b) sind der ungewichtete Schaltungssensitivitätsfaktor s_1 nach Gleichung 5.2 und der gewichtete Schaltungssensitivitätsfaktor s_2 nach Gleichung 5.4 im Vergleich zur Erhöhung der Laufzeitschwankung dargestellt. Es zeigt sich, dass s_1 nur bedingt als Indikator für den Einfluss von Variationen auf die Laufzeitschwankung herangezogen werden kann, da im Falle des selektiven Einsatzes von gepulsten Flip Flops im Vergleich zu anderen Maßnahmen mit größerer Laufzeitschwankung ein erhöhter Sensitivitätsfaktor resultiert. Daher kann für Schaltungssensitivitätsfaktor s_1, der lediglich auf der Anzahl kritischer Hardware beruht, der Zusammenhang zwischen erhöhtem Sensitivitätsfaktor und erhöhten Laufzeitschwankungen nicht für alle ARM926 Design-Optionen verifiziert werden. Im Vergleich zu s_1 berücksichtigt s_2 durch individuelle Gewichtung der einzelnen Gatter weitere Eigenschaften der Schaltung. Der gewichtete Sensitivitätsfaktor s_2 bezieht bei der Bewertung der Schaltungsrobustheit neben der topologischen Korrelation in den kritischen Pfaden auch den zeitlichen Abstand eines Gatters zum kritischsten aller Pfade durch individuelle Gewichtung mit ein. In diesem Fall kann für alle untersuchten Designs des ARM926 ein klarer Zusammenhang zwischen Schaltungssensitivitätsfaktor s_2 und Laufzeitschwankung und damit deren positive Korrelation nachgewiesen werden. Bild

6.32 veranschaulicht die einzelnen Schritte der angewandten Methodik zur Validierung des Schaltungssensitivitätsfaktors.
Unter den vorherrschenden Randbedingungen und den getroffenen Annahmen kann festgestellt werden, dass der gewichtete Schaltungssensitivitätsfaktor s_2 eine geeignete Kenngröße zur Bewertung der Robustheit einer Schaltung gegenüber Variationen ist. Über die getroffenen Annahmen hinaus dient der Schaltungssensitivitätsfaktor zusätzlich als Indikator für den Einfluss unmodellierbarer systematischer und pseudo-statistischer Variationen auf die Geschwindigkeit einer Schaltung, wie in Kapitel 5.2 gezeigt werden konnte.

7 Zusammenfassung und Schlussfolgerung

7.1 Zusammenfassung

Die Analysen und Ergebnisse dieser Arbeit zeigen, dass entgegengesetzt vieler in der Literatur zu findenden Aussagen die Schwankungsbreiten der wichtigsten systematischen Prozess- und Betriebsparameter moderner sub-100nm CMOS Schaltungen bis 40nm nicht zunehmen. Aufgrund eines besseren Verständnisses der Variationsquellen und besserer Prozesskontrolle skalieren die Variationen mit der Technologie.
Trotz gleichbleibender relativer Parameterschwankungen erhöht sich dennoch die Laufzeitschwankung moderner CMOS Schaltungen. Dieser Zusammenhang lässt sich anhand steigender Laufzeitsensitivitäten gegenüber Prozess- und Betriebsparametern erklären. Dabei spielen vorwiegend der reduzierte Gate-Overdrive $V_{DD} - V_T$ sowie stärker werdende Kurzkanaleffekte eine entscheidende Rolle. Es zeigt sich, dass die Schwankungen der Gatelänge und der Transistoreinsatzspannung die dominierenden Einflussgrößen darstellen. Während Gatelängenschwankungen bei nomineller Versorgungsspannung den größten Einfluss auf die Laufzeit haben, nimmt die Bedeutung von Einsatzspannungsschwankungen bei verringerter Versorgungsspannung signifikant zu. Bei der Analyse des Einflusses von Variationen auf das Laufzeitverhalten ist es daher wichtig, den Betriebsbereich der jeweiligen Schaltung bei der Variationsanalyse zu berücksichtigen, da sich nicht nur der Einfluss von Prozess- sondern auch Umgebungsvariationen in verschiedenen Betriebsbereichen deutlich unterscheidet.
Globale Schwankungen wirken sich auf alle Transistoren gleich aus und können als globaler Offset angesehen werden. Within-Die Variationen, die statistischer, pseudo-statistischer und systematischer Natur sind, wirken sich abhängig von Gatter- und Pfadtopologie unterschiedlich auf das Laufzeitverhalten der Gesamtschaltung aus. Aus diesem Grund ist es wichtig, die mikroarchitektonischen, topologischen und strukturellen Eigenschaften einer Schaltung bei der Analyse von Variationseffekten miteinzubeziehen. Das in dieser Arbeit vorgestellte Mikroprozessormodell ermöglicht es, technologische und schaltungstechnische Aspekte in Kombination zu analysieren und deren Auswirkung auf das Laufzeitverhalten zu bestimmen. Dies erfordert zum einen die Modellierung struktureller und topologischer Eigenschaften geschwindigkeitskritischer Schaltungsteile, zum anderen die Bestimmung der technologieabhängigen Laufzeitsensitivitäten. Dazu wurde ein generisches Pfadmodell erarbeitet, das die Eigenschaften kritischer Pfade hinsichtlich ihrer Laufzeitsensitivität gegenüber Prozess- und Umgebungsvariationen über mehrere Technologiegenerationen hinweg sehr gut nachbildet. Dies ermöglicht die frühe Abschätzung des Einflusses von Variationen auf das Laufzeitverhalten kritischer Pfade in eingebetteten Mikroprozessoren. Die Analyse der ARM Mikroprozessorfamilie (ARM926, ARM1176, ARM Cortex A8), die stellvertretend für die mikroarchitektonische Entwicklung des klassischen RISC Prozessors untersucht wurde, zeigt, dass der Großteil der on-chip Laufzeitschwankungen aus Umge-

bungsvariationen wie IR-Drop und Clock Jitter stammt. Dabei nimmt die Bedeutung des Taktverteilungsnetzes zu, da das Verhältnis von Laufzeit im Taktpfad zur Laufzeit im Logikpfad künftig weiter ansteigt. Damit tragen auch die Laufzeitschwankungen aus dem Taktverteilungsnetz künftig stärker zur Gesamtlaufzeitschwankung bei. Lokale statistische Variationen gewinnen zwar an Bedeutung, wirken sich jedoch im Vergleich zu anderen Variationsquellen deutlich geringer auf die Laufzeitschwankungen eingebetteter Mikroprozessoren aus.
Detaillierte Crosstalk-Untersuchungen anhand von Aggressor-Netzen in einem ARM926 Produktdesign zeigen, dass mittlere bis hohe Crosstalk-bedingte Laufzeitvariationen nur mit vernachlässigbar kleiner Wahrscheinlichkeit auftreten, so dass in eingebetteten Mikroprozessoren kein signifikanter Einfluss dieses Effekts auf die Laufzeit kombinatorischer Logik zu erwarten ist.
Die Ergebnisse des Mikroprozessormodells ergeben ferner, dass bei tiefem Pipelining eines ARM Cortex A8 der Laufzeitanteil der Flip Flop Zellen und der Timing Unsicherheit gemeinsam knapp 45% der Taktperiode beanspruchen. Dabei erhöht sich der Beitrag der Timing Unsicherheit von einem ARM926 in 90nm auf einen Cortex A8 in 40nm um ca. 78% und die Effizienz von tieferem Pipelining nimmt signifikant ab. Da in sub-100nm CMOS Technologien der Geschwindigkeitsgewinn beim Wechsel in eine neue Technologiegeneration abnimmt, sind weitergehende schaltungstechnische Maßnahmen erforderlich, um die Effizienz von tieferem Pipelining aufrecht zu erhalten und somit auch weiterhin höhere Taktfrequenzen zu ermöglichen.
Im Vergleich zu den untersuchten Mikroprozessoren kann der Einfluss insbesondere von lokal wirkenden Variationen wie z.B. Crosstalk und statistischen Variationen in anderen Schaltungen signifikant abweichen. Daher ist eine genaue Kenntnis der Schaltungsstruktur, des Betriebsbereichs und der während des Schaltungsentwurfs vorherrschenden Randbedingungen zur Evaluation des Einflusses von Variationen entscheidend.

Zur Kompensation von Laufzeitschwankungen sind verschiedene Techniken bekannt. Globale Laufzeitschwankungen können durch die Adaption von Versorgungsspannung V_{DD} und Body-Source Spannung V_{BS} global kompensiert werden. Messungen in 45nm low-power CMOS Technologien zeigen, dass hinsichtlich der Energieaufnahme Reverse Body Biasing (RBB) nur für hohe Energiebeiträge des Leckstroms zur Gesamtenergieaufnahme einen deutlichen Vorteil gegenüber Process/Adaptive Voltage Scaling (PVS/AVS) aufweisen. Da in low-power Produkten häufig Power-Switches verwendet werden, die inaktive Schaltungsteile vom Versorgungsnetz trennen, sind keine derart hohen Leckstrombeiträge zum Gesamtstrom zu erwarten. Für die Kompensation von Laufzeiterhöhungen zeigt sich jedoch unabhängig vom Leckstromanteil an der Gesamtenergieaufnahme ein deutlicher Vorteil von PVS/AVS gegenüber Forward Body Biasing (FBB). Aufgrund der steigenden Laufzeitsensitivität gegenüber veränderten Versorgungsspannungen sinkt für moderne CMOS Technologien die Energieersparnis durch PVS/AVS, während gleichzeitig der zusätzliche Energieaufwand für die Kompensation von Laufzeiterhöhungen ebenfalls sinkt. Da in low-power Schaltungen häufig Dynamic Voltage Scaling implementiert ist, stellt PVS/AVS keinen großen zusätzlichen Aufwand dar.

Da WID Variationen durch adaptive Techniken nicht kosteneffizient kompensiert werden können und sich abhängig von Schaltungsstruktur und -topologie unterschiedlich auf die Laufzeitschwankung auswirken, ist eine spezielle Behandlung von WID Variationen während des Schaltungsentwurfs erforderlich. Hierzu wurden verschiedene Bewertungs-

kenngrößen zur Beschreibung der komplexen Schaltungsstruktur sowie zur Quantifizierung der Robustheit gegenüber Variations- und Alterungseffekten eingeführt. Es wurde gezeigt, dass die topologische Korrelation, d.h. die Verknüpfung der einzelnen Gatter untereinander, einen wesentlichen Einfluss auf die variationsbedingte Laufzeitschwankung der Schaltung hat. Die Definitionen des Schaltungssensitivitäts-Faktors basieren auf strukturellen und topologischen Eigenschaften der Schaltung. Sie ermöglichen es über die Analyse der 'kritischen Hardware' die Robustheit einer Schaltung gegenüber allen auftretenden Variationen, d.h. sowohl Prozess- als auch Umgebungsvariationen, zu quantifizieren und vergleichbar zu machen. Die Beschreibung struktureller und topologischer Eigenschaften durch diese Bewertungskenngröße ermöglicht somit die Quantifizierung der 'Verwundbarkeit bzw. Anfälligkeit' einer Schaltung gegenüber Variationen. Somit werden erstmals strukturelle Aspekte wie die topologische Korrelation von Gattern und der Verlauf von Pfad- und Gatterspektren bei der Bewertung der Robustheit einer Schaltung einbezogen. Diese heuristische Herangehensweise eröffnet neben der um den Robustheitsfaktor erweiterten Kosten-Nutzen Analyse die Möglichkeit zur Optimierung der Schaltung hinsichtlich des Einflusses von WID Prozess- und on-chip Umgebungsvariationen. Im Gegensatz zur vorgeschlagen Metrik ist eine Optimierung der Robustheit durch die oft zitierte Statistische Statische Timing Analyse nicht möglich.
Zur Umsetzung kosteneffizienter präventiver Kompensationstechniken wie dem selektiven Einsatz von gepulsten Flip Flops in geschwindigkeitskritischen Pfaden bzw. der globalen Ersetzung von MS-FFs durch gepulste Latches, wurde eine Multi-Stage Erweiterung der herkömmlichen STA erarbeitet, die den funktionalen Einsatz und den Gewinn von vielfach verwendeten Time-Borrowing Techniken überprüft. Diese ermöglicht die Identifikation von Pfadtopologien, die den Einsatz solcher Techniken limitieren bzw. bei Anwendung sogar zu funktionalen Fehlern führen können. Als kritischste Struktur wurde die 'Pipelined Loop' identifiziert, deren Anzahl mit tieferem Pipelining wächst. Im ARM926 konnte jeder fünfte kritische Pfad dieser Pfadtopologie zugeordnet werden.
Basierend auf den neuen Bewertungskenngrößen wurden präventive Designtechniken zur Kompensation von WID Variationen ausgewählt und näher analysiert. Aufgrund der hohen topologischen Korrelation der geschwindigkeitskritischen Pfade im ARM926, die Indikator für einen großen Hebel lokaler, selektiver Maßnahmen ist, wurde der selektive Einsatz von low-V_T Gattern in geschwindigkeitskritischen Pfaden untersucht. Unter Berücksichtigung der erarbeiteten Bewertungskenngrößen wurde ein robustheitsorientierter Ersetzungsalgorithmus aufgestellt und angewandt. Dieser ermöglicht für den untersuchten ARM926 die Implementierung einer zeitlichen Sicherheitsmarge von 7% der Taktperiode für weniger als 3% Leckstromerhöhung. Der Flächenbedarf bleibt aufgrund der zum reg-V_T Gatter pin- und flächenkompatiblen low-V_T Gatter und dem Ausschluss Hold-Time relevanter Zellen während der Ersetzung unverändert. Der mit der Ersetzung verbundene Aufbau einer Timing-Wall erhöht jedoch den Sensitivitätsfaktor der Schaltung um den Faktor 7.7.
Im Vergleich zum lokalen Einsatz der low-V_T Gatter erzielt der globale Einsatz von gepulsten Latches eine zeitliche Sicherheitsmarge von 5% bei einer vergleichsweise geringen Erhöhung des Sensitivitätsfaktors um den Faktor 1.9. Die Untersuchung der Pfadtopologien mittels Multi-Stage Erweiterung der STA zeigt, dass statisches Time-Borrowing in der Größenordnung von nur 2.5% der Taktperiode möglich ist. 'Pipelined Loops' und der zeitliche Ausgleich (Path Delay Balancing) zwischen sukzessiven Pfaden führen bei erhöhtem statischen Time-Borrowing zu Setup-Zeit Verletzungen. Für den untersuchten ARM926 ergibt sich aufgrund der im Vergleich zum Master-Slave Flip Flop kleineren ge-

pulsten Latches zusätzlich eine Flächenreduktion von ca. 5.6%.
Die Ergebnisse des Mikroprozessormodells liefern für den globalen Einsatz von gepulsten Latches im Cortex A8 eine zeitliche Sicherheitsmarge von ca. 8% der Taktperiode bei ca. 1% Flächengewinn. Zusätzlich ermöglicht die zeitliche Elastizität des gepulsten Flip Flops dynamisches Time-Borrowing in der Größenordnung von 40% der WID Timing Unsicherheit.
Der Einsatz von gepulsten Flip Flops und Latches ist daher eine skalierbare präventive Kompensationstechnik, da die zu erzielende relative zeitliche Sicherheitsmarge mit tieferem Pipelining weiter zunimmt. Zusätzlich erleichtert die zeitliche Elastizität die fehlerfreie Implementierung von on-chip Adaptionstechniken, da es aufgrund adaptionsbedingter Schwankungen der Betriebsparameter zu kurzzeitigen Laufzeitschwankungen kommen kann, die durch dynamisches Time-Borrowing kompensiert werden können.
Die Evaluation der einzelnen Kompensationstechniken zeigt, dass die eingeführten strukturellen Bewertungsgrößen zum Vergleich der Schaltungssensitivität geeignet sind und die Auswahl adäquater Kompensationstechniken sowie deren Bewertung erleichtern.

Im Rahmen dieser Arbeit entstand somit eine abstraktionsebenenübergreifende Vorgehensweise zur Bewertung des Einflusses von Variationen auf die Geschwindigkeit und Robustheit einer Schaltung. Dies ermöglicht sowohl die frühe Abschätzung von Variationseffekten während des Schaltungsentwurfs und unterstützt bei der Auswahl und Implementierung von Monitor-Konzepten & Kompensationstechniken und den damit verbundenen Trade-Offs zwischen Energieaufnahme, Flächenbedarf, Geschwindigkeit und erstmals Schaltungsrobustheit.

7.2 Schlussfolgerung

Die Ergebnisse dieser Arbeit zeigen, dass eine Bewertung des Einflusses von Variationen auf die Geschwindigkeit von getakteten CMOS Digitalschaltungen nur durch eine umfassende Analyse von Variationen auf allen Abstraktionsebenen vom Transistor bis zur Mikroarchitektur erfolgen kann.
Insbesondere low-power Techniken, wie z.B. Dynamic Voltage Scaling, Clock Gating etc. führen dazu, dass die Schaltung zum einen bei höheren Laufzeitsensitivitäten betrieben wird, zum anderen höhere Umgebungsvariationen z.B. durch dynamische Lastwechsel erfährt. Deshalb ist es insbesondere bei fortschreitender Technologieskalierung wichtig, dass low-power Maßnahmen nicht nur anhand von eingesparter Energie, sondern auch unter Berücksichtigung der veränderten Laufzeitsensitivitäten und Umgebungsvariationen bewertet werden.
Zusätzlich erschwert die steigende Schaltungskomplexität zusammen mit steigenden Laufzeitschwankungen die Bewertung des Einflusses von Prozess- und Umgebungsvariationen auf die Geschwindigkeit der Schaltung. Da hierfür bisher weder analytische noch numerische Lösungen existieren, gewinnen heuristische Ansätze zur Beschreibung, Quantifizierung und Bewertung von Schaltungsrobustheit wie die in dieser Arbeit vorgestellten Vorgehensweise, an Bedeutung. Dabei müssen vor allem strukturelle Aspekte, wie z.B. topologische Korrelation, Pfad- und Gatterspektren etc. berücksichtigt werden, die bisher vernachlässigt wurden. Dies erfordert eine detaillierte Strukturanalyse, die die Identifikation von schaltungsspezifischen Einflussgrößen ermöglicht.
Eine sinnvolle Bewertung des Einflusses von Variationen und die Auswahl kosteneffizienter Gegenmaßnahmen in getakteten Digitalschaltungen ist daher nur auf Basis eines abstrak-

tionsebenenübergreifenden Ansatzes möglich, der sowohl technologische (Laufzeitsensitivität) und strukturelle, schaltungsspezifische Eigenschaften als auch schaltungstechnische und betriebsbedingte Randbedingungen bei der Implementierung der Schaltung berücksichtigt.

Literaturverzeichnis

[1] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, K. Bernstein: *Scaling, Power, and the Future of CMOS*; In: IEEE International Electronic Device Meeting, Dezember 2005, pp. 9-15.

[2] A. J. Bhavnagarwala, X. Tang, J. D. Meindl: *The Impact of Intrinsic Device Fluctuations on CMOS SRAM Cell Stability*; In: IEEE Journal of Solid-State Circuits, Vol. 36, No. 4, April 2001, pp. 658-665.

[3] R.Venkatraman, R.Castagnetti, S.Ramesh: *The Statistics of Device Variations and its Impact on SRAM Bitcell Performance, Leakage and Stability*; In: International Symposium on Quality Electronic Design, März 2006 2001, pp. 190-195.

[4] A. Agarwal, B. C. Paul, S. Mukhopadyhyay, K. Roy: *Process Variation in Embedded Memories: Failure Analysis and Variation Aware Architecture*; In: IEEE Journal of Solid-State Circuits, Vol. 40, No. 9, September 2005, pp. 1804-1814.

[5] B. E. Stine, D. S. Boning, J. E. Chung, D. J. Ciplickas, J. K. Kibarian: *Simulating the Impact of Pattern-Dependent Poly-CD Variation on Circuit Performance*; In: IEEE Transactions on Semiconductor Manufacturing, Vol. 11, No. 4, November 1998, pp. 552-556.

[6] R. H. Voelkler: *Transposing Conductors in Signal Buses to Reduce Nearest-Neighbor Crosstalk*; In: IEEE Transactions on Microwave Theory and Techniques, Vol. 43, No. 5, Mai 1995, pp. 1095-1099.

[7] T. Numi, S. Tuuna, J. Isoaho: *Evaluating the Relative Effect of Process Variations and Switching Patterns on BUS Performance towards Nano-Scale Interconnects*; In: International Symposium on Signals, Circuits and Systems, Juli 2005, pp. 59-62.

[8] R. Mahnkopf, K.-H. Allers, M. Armacost, A. Augustin, J. Barth, G. Brase, R. Busch, E. Demm, G. Dietz, B. Flietner, G. Friese, F. Grellner, K. Han, R. Hannon, H. Ho, M. Hoinkis, K. Holloway, T. Hook, S. Iyer, P. Kim, G. Knoblinger, B. Lemaitre, C. Lin, R. Mih, W. Neumueller, J. Pape, O. Prigge, N. Robson, N. Rovedo, T. Schafbauer, T. Schiml, K. Schruefer, S. Srinivasan, M. Stetter, F. Towler, P. Wensley, C. Wann, R. Wong, R. Zoeller, B. Chen: *'System on a chip' Technology Platform for 0.18 μm Digital, Mixed Signal and eDRAM Applications*; In: IEEE International Electron Devices Meeting, 1999, pp. 849-852.

[9] L. K. Han, S. Biesemans, J. Heidenreich, K. Houlihan, V. McGahay, T. Schiml, A. Schmidt, U. P. Schroeder, M. Stetter, C. Wann, D. Warner, R. Mahnkopf, B. Chen: *A Modular 0.13 μm Bulk CMOS Technology for High*

Performance and Low Power Applications; In: Symposium on VLSI Technology, 2000, pp. 12-13.

[10] T. Schafbauer, J. Brighten, Y.-C. Chen, L. Clevenger, M. Commons, A. Cowley, K. Esmark, A. Grassmann, U. Hodel, H.-J. Huang, S.-F. Huang, Y. Huang, E. Kaltalioglu, G. Knoblinger, M.-T. Lee, A. Leslie, P. Leung, B. Li, C. Lin, Y.-H. Lin, W. Nissl, P. Nguyen, A. Olbrich, P. Riess, N. Rovedo, S. Sportouch, A. Thomas, D. Vietzke, M. Wendel, R. Wong, Q. Ye, K.-C. Lin, T. Schiml, C. Wann: *Integration of High-performance, Low-leakage and Mixed Signal Features into a 100nm CMOS Technology*; In: Symposium on VLSI Technology, 2002, pp. 62-63.

[11] Z. Luo, A. Steegen, M. Eller, R. Mann, C. Baiocco, P. Nguyen, L. Kim, M. Hoinkis, V. Ku, V. Klee, F. Jamin, P. Wrschka, P. Shafer, W. Lin, S. Fang, A. Ajmera, W. Tan, D. Park, R. Mo, J. Lian, D. Vietzke, C. Coppock, A. Vayshenker, T. Hook, V. Chan, K. Kim, A. Cowley, S. Kim, E. Kaltalioglu, B. Zhang, S. Marokkey, Y. Lin, K. Lee, H. Zhu, M. Weybright, R. Rengarajan, J. Ku, T. Schiml, J. Sudijono, I. Yang, C. Wann: *High Performance and Low Power Transistors Integrated in 65nm Bulk CMOS Technology*; In: IEEE International Electron Devices Meeting, 2004, pp. 661-664.

[12] J. Yuan, V. Chan, M. Eller, N. Rovedo, H. K. Lee, Y. Gao, V. Sardesai, N. Kanike, V. Vidya, O. Kwon, O. S. Kwon, J. Yan, S. Fang, W. Wille, H. Wang, Y. T. Chow, R. Booth, T. Kebede, W. Clark, H. Mo, C. Ryou, J. Liang, J. H. Yang, C.W. Lai, S.S. Naragad, O. Gluschenkov, M. R. Visokay, C. Radens, S. Deshpande, H. Shang, Y. Li, N. Cave, J. Sudijono, J. Ku, R. Divakaruni: *A 45nm Low Power Bulk Technology Featuring Carbon Co-implantation and Laser Anneal on 45°-rotated Substrate*; In: International Conference on Solid-State and Integrated-Circuit Technology, 2008, pp. 1130-1133.

[13] D. S. Boning, K. Balakrishnan, H. Cai, N. Drego, A. Farahanchi: *Variation*; In: Transactions on Semiconductor Manufacturing, Vol. 21, No. 1, Februar 2008, pp. 63-71.

[14] K. Bernstein et al.: *High Performance CMOS Variability in the 65nm Regime and Beyond*; In: IBM Journal of Research and Developement, Vol. 50, No. 4/5, Juli/September 2006, pp. 433-449.

[15] B.-S. Kim, B.-H. Lee, H.-B. Choi, S.-I. Heo, J.-R. Lee, Y.-C. Kim, C. Rim, K.-M. Choi: *Parametric Yield-Aware Sign-Off Flow in 65/45nm*; In: International SoC Design Conference, November 2008, pp. 74-77.

[16] T. Sakurai, R. Newton: *Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and Other Formulas*; In: IEEE Journal of Solid-State Circuits SSC, Vol. 25, No. 2, April 1990, pp. 584-594.

[17] K. von Arnim, C. Pacha, K. Hofmann: *An Effective Switching Current Methodology to Predict the Performance of Complex Digital Circuits*; In: IEEE International Electronic Devices Meeting, Dezember 2007, pp. 483-486.

[18] Y. Cao, L. T. Clark: *Mapping Statistical Process Variations Toward Circuit Performance Variability: An Analytical Modeling Approach*; In: IEEE/ACM Design Automation Conference, Juni 2005, pp. 658-663.

[19] M. H. Na, E. J. Nowak, W. Haensch, J. Cai: *The Effective Drive Current in CMOS Inverters*; In: IEEE International Electron Devices Meeting, Dezember 2002, pp. 121-124.

[20] K. Cao, S. Dobre, J. Hu: *Standard Cell Characterization Considering Lithography Induced Variations*; In: IEEE/ACM Design Automation Conference, Juli 2006, pp. 801-804.

[21] K. L. Shepard, D. N. Maynard: *Variability and Yield Improvement: Rules, Models, and Characterization*; In: IEEE/ACM International Conference on Computer Aided Design, November 2006, pp. 834-835.

[22] L.T. Pang, B. Nikolic: *Impact of Layout on 90nm CMOS Process Parameter Fluctuations*; In: Symposium on VLSI Circuits, 2006, pp. 69-70.

[23] S. Banerjee, P. Elakkumanan, L. W. Liebmann, M. Orshansky: *Electrically Driven Optical Proximity Correction Based on Linear Programming*; In: IEEE/ACM International Conference on Computer-Aided Design, November 2008, pp. 473-479.

[24] L. T. Pang, B. Nikolic: *Measurement and Analysis of Variability in 45nm Strained-Si CMOS Technology*; In: IEEE Custom Integrated Circuits Conference, September 2008, pp. 129-132.

[25] F. Mo, R. Brayton: *Design Methodology of Regular Logic Bricks for Robust Integrated Circuits*; In: International Conference on Computer Design, Oktober 2007, pp. 162-167.

[26] K. Y. Tong, V. Rovner, L. T. Pileggi, V. Kheterpal: *PLA-Based Regular Structures and Their Synthesis*; In: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 22, No. 6, Juni 2003, pp. 723-729.

[27] A. Asenov, A. Brown, J. Davies, S. Kaya, and G. Slavcheva: *Simulation of Intrinsic Parameter Fluctuations in Decananometer and Nanometer-scale MOSFETs*; In: IEEE Transactions on Electron Devices, Vol. 50, No. 9, September 2003.

[28] D. Frank, W. Haensch, G. Shahidi, O. Dokumaci: *Optimizing CMOS Technology for Maximum Speed*; In: IBM Journal of Research and Developement, Vol. 50, No. 4/5, Juli/September 2006, pp. 419-431.

[29] D. Pramanik, V. Moroz, X. W. Lin: *Process Induced Layout Variability for Sub 90nm Technologies*; In: International Conference on Solid-State and Integrated Circuit Technology, 2006, pp. 1849-1852.

[30] K. Kuhn: *Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS*; In: IEEE International Electronic Devices Meeting 2007, pp. 471-474.

[31] T. Kanamoto, Y. Ogasaharat, K. Natsume, K. Yamaguchi, H. Amishiro, T. Watanabe, M, Hashimoto: *Impact of Well Edge Proximity Effect on Timing*; In: European Solid State Device Research Conference, September 2007, pp. 115-118.

[32] C. Pacha, B. Martin, K. von Arnim, R. Brederlow, D. Schmitt-Landsiedel, P. Seegebrecht, J. Berthold, R. Thewes: *Impact of STI-Induced Stress, Inverse Narrow Width Effect, and Statistical V_T Variations On Leakage Current in 120nm CMOS*; In: European Solid-State Device Research Conference, September 2004, pp. 397-400.

[33] Y. Taur: *CMOS Design Near the Limit of Scaling*; In: IBM Journal of Research and Developement, Vol. 46, No. 2/3, März/Mai 2002, pp. 213-222.

[34] M. Pelgrom, A. Duinmaijer, A. Welbers: *Matching Properties of MOS Transistors*; In: IEEE Journal of Solid-State Circuits, Vol. 24, No. 5, October 1989, pp. 1433-1440.

[35] P. Stolk, F. Widdershoven, D. Klaassen: *Modeling Statistical Dopant Fluctuations in MOS Transistors*; In: IEEE Transactions on Electron Devices, Vol. 45, No. 9, Sep. 1998, pp. 1960-1971.

[36] A. Hokazono, S. Balasubramanian, K. Ishimaru, H. Ishhiuchi, C. Hu, T. Liu: *Forward-Body Biasing as a Bulk-Si CMOS Technology Scaling Strategy*; In: IEEE Transactions on Electron Devices, Vol. 55, No. 10, Oktober 2008, pp. 2657-2664.

[37] *www.ITRS.net*;

[38] P. Fantini, G. Giuga, S. Schippers, A. Marmiroli, G. Ferrari: *Modeling of STI-induced Stress Phenomena in CMOS 90nm Flash Technology*; In: European Solid-State Device Research Conference, September 2004, pp. 401-404.

[39] R. Kaushik: *Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits*; In: Proceedings of the IEEE, Vol. 91, No. 2, Februar 2003, pp. 305-327.

[40] P. McGuinness: *Variations, Margins, and Statistics*; In: International Symposium on Physical Design, April 2008, pp. 60-67.

[41] P. McGuinness: *Vortragsfolien: Variations, Margins, and Statistics*; In: International Symposium on Physical Design, April 2008, (www.ispd.cc).

[42] V. Mehrotra, D. Boning: *IEEE Technology Scaling Impact of Variation on Clock Skew and Interconnect Delay*; In: IEEE International Interconnect Technology Conference, 2001, pp. 122-124.

[43] Rabaey, J., Chandrakasan, Nikolic B.: *Digital Integrated Circuits - A Design Perspective*; Electronics and VLSI Series; Prentice Hall, 2. Auflage, 2003; ISBN 0-13-120764-4.

[44] I. E. Sutherland, B. F. Sproull, D. L. Harris: *Logical Effort Designing Fast CMOS Circuits* ; In: Morgan Kaufman Publications, 1999, ISBN: 1-55860-557-6.

[45] M. H. Abu-Rahma, M, Anis: *Variability in VLSI Circuits: Sources and Design Considerations*; In: IEEE International Symposium on Circuits and Systems, May 2007, pp. 3215-3218.

[46] X. Liang, D. Brooks: *Mitigating the Impact of Process Variations on Processor Register Files and Execution Units*; In: IEEE/ACM International Symposium on Microarchitecture 2006, Dezember. 2006, pp. 504-515.

[47] C. Amin, N. Menezes, K. Killpack, F. Dartu, U. Choudhury, N. Hakim, Y. I. Ismail: *Statistical Static Timing Analysis: How simple can we get?*; In: ACM/IEEE Design Automation Conference (DAC), Juni 2005, pp. 652-657.

[48] D. Sylvester, K. Agarwal, S. Shah: *Variability in Nanometer CMOS: Impact, Analysis, and Minimization*; In: Integration-The VLSI Journal, Vol. 41, No. 3, Mai 2008, pp. 319-339.

[49] S. Kirolos, Y. Massoud, Y. Ismail: *Power-Supply-Variation-Aware Timing Analysis of Synchronous Systems*; In: IEEE International Symposium on Circuits and Systems, Mai 2008, pp. 2418-2421.

[50] A. Dubey: *P/G Pad Placement Optimization: Problem Formulation for Best IR Drop*; In: International Symposium on Quality Electronic Design, Vol. 6, März 2005, pp. 340-345.

[51] M. Eireiner, S. Henzler, X. Zhang, J. Berthold, D. Schmitt-Landsiedel: *Impact of On-chip Inductance on Power Supply Integrity*; In: Advances in Radio Science, URSI Kleinheubacher Berichte, 2008, pp. 227-232.

[52] R. Kumar, V. Kursun: *Voltage Optimization for Temperature Variation Insensitive CMOS Circuits*; In: Midwest Symposium on Circuits, August 2005, pp. 476-479.

[53] K. von Arnim, E. Borinski, P. Seegebrecht, H. Fiedler, R. Brederlow, R. Thewes, J. Berthold, C. Pacha: *Efficiency of Body Biasing in 90-nm CMOS for Low-Power Digital Circuits*; In: IEEE Journal of Solid-State Circuits, Vol. 40, No. 7, Juli 2005, pp. 1549-1556.

[54] E. Kursun, C.-Y. Cher: *Temperature Variation Characterization and Thermal Management of Multicore Architectures*; In: IEEE Micro, Vol. 29, No. 1, Januar/Februar 2009, pp. 116-126.

[55] A. S. Leon, B. Langley, J. L. Shin: *The UltraSPARC T1 Processor: CMT Reliability*; In: IEEE Custom Integrated Circuits Conference, September 2006, pp. 555-562.

[56] J. Friedrich, B. McCredie, N. James et al.: *Design of the Power6 Microprocessor*; In: IEEE International Solid-State Circuits Conference, Februar 2007, pp. 96-97.

[57] P. Chen, D. A. Kirkpatrick, K. Keutzer: *Miller Factor for Gate-Level Coupling Delay Calculation*; In: IEEE/ACM International Conference on Computer Aided Design, November 2000, pp. 68-74.

[58] J.-S. Yang, A. R. Neureuther: *Crosstalk Noise Variation Assessment and Analysis for the Worst Process Corner*; In: International Symposium on Quality in Electronic Design, 2008, pp. 352-356.

[59] S. Nazarian, M. Pedram, E. Tuncer: *An Empirical Study of Crosstalk in VDSM Technologies*; In: ACM Great Lake Symposium on VLSI, April 2005, pp. 317-322.

[60] H. Tseng, L. Scheffer, C. Sechen: *Timing- and Crosstalk-Driven Area Routing*; In: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 20, No. 4, April 2001, pp. 528-544.

[61] T. Xiao, M. Sadowska: *Crosstalk Reduction by Transistor Sizing*; In: Asia and South Pacific Design Automation Conference, Januar 1999, pp. 137-140.

[62] K. Hirose, H. Yasuura: *A Bus Delay Reduction Technique Considering Crosstalk*; In: Design, Automation and Test in Europe, März 2000, pp. 441-445.

[63] R. Arunachalam, E. Acar, S. R. Nassif: *Optimal Shielding/Spacing Metrics for Low Power Design*; In: IEEE Computer Society Annual Symposium on VLSI, Februar 2003, pp. 167-172.

[64] B. Victor, K. Keutzer: *Bus Encoding to Prevent Crosstalk Delay*; In: IEEE/ACM International Conference on Computer Aided Design, November 2001, pp. 57-63.

[65] S. Pasricha, N. Dutt: *On-Chip Communication Architectures - System on Chip Interconnect*; In: Elsevier, 2008, ISBN: 978-0-12-373892-9.

[66] D. K. Schroder: *Semiconductor Material and Device Characterization*; In: John Wiley & Sons Interscience, 2006, ISBN: 978-0-471-73906-7.

[67] D. K. Schroder, J. A. Babcock: *Negative Bias Temperature Instability: Road to Cross in Deep Submicron Silicon Semiconductor Manufacturing*; In: Journal of Applied Physics, Vol. 94, No. 1, Juli 2003, pp. 1-18.

[68] J. Hicks, D. Bergstrom, M. Hattendorf, J. Jopling, J. Maiz, S. Pae, C. Prasad, J. Wiedemer: *45nm Transistor Reliability*; In: Intel Tech. J., Vol. 12, Juli 2008, pp. 131-144.

[69] N. J. Rohrer: *Introduction to Statistical Variation and Techniques for Design Optimization*; In: IEEE International Solid-State Circuits Conference (Tutorial), 2006.

[70] S. R. Nassif, A. J. Strojwas, K. Osada, J. Tschanz, M. Clinton, S. Ohshima: *2007 VLSI Circuits Short Course Program 'Design for Variability in Logic, Memory and Microprocessors'*; In: IEEE International Symposium on VLSI Circuits, Juni 2007.

[71] B. P. Wong, A. Mittal, Y. Cao, G. Starr: *Nano-CMOS Circuit and Physical Design*; In: Wiley-Interscience, 2005, ISBN: 0-471-46610-7.

[72] A. de Carvalho, E. Flores, N. Hakim, J. Hemmet, S. Idgunji, K. Kalafala, P. McGuinness, A. Mutlu, M. Sharma, O. M. Siguenza, A. Srivastava, A. Tetelbaum: *Statistical Methods For Semiconductor Chip Design*; In: Publiziert von Silicon Integration Initiative, Inc., Dezember 2008, pp. 1-53.

[73] S. R. Nassif: *Design For Variability in DSM Technologies*; In: International Symposium on Quality of Electronic Design, März 2000, pp. 451-454.

[74] W. Zhao, Y. Cao, F. Liu: *Rigorous Extraction of Process Variations for 65nm CMOS Design*; In: European Sloid State Device Research Conference, September 2007, pp. 89-92.

[75] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, V. De: *Parameter Variations and Impact on Circuits and Michroarchitecture*; In: IEEE/ACM Design Automation Conference, Juni 2003, pp. 338-342.

[76] J. Jaffari, M. Anis: *Statistical Thermal Profile Considering Process Variations: Analysis and Applications*; In: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 27, No. 6, Juni 2008, pp. 1027-1040.

[77] W. Schemmert, G. Zimmer: *Threshold-voltage Sensitivity of Ion-Implanted M.O.S. Transistors Due to Process Variations*; In: Electronics Letters, Vol. 10, No. 9, Mai 1974, pp. 151-152.

[78] W. Härdle, L. Simar: *Applied Multivariate Statistical Analysis*; In: 2. Auflage, Springer Verlag 2007, ISBN: 978-3-540-72243-4.

[79] Y. Zhan, A. J. Strojwas, X. Li, L. T. Pileggi, D. Newmark, M. Sharma: *Correlation-Aware Statistical Timing Analysis with Non-Gaussian Delay Distributions*; In: IEEE/ACM Designa Automation Conference, Juni 2005, pp. 77-82.

[80] B. Romanescu, S. Ozev, D. J. Sorin: *Quantifying the Impact of Process Variability on Microprocessor Behavior*; In: 2nd Workshop on Architectural Reliability (WAR), Dezember 2006.

[81] J.C. Scott, O- Gluschenkov, B. Goplen, H. Landis, E. Nowak, F. Clougherty, A. Mocuta, T. Hook, N. Zamdmer, C. W. Lai, M. Eller, D. Chidambarrao, J. Yu, P. Chang, J. Ferris, S. Despande, Y. Li, H. Shang, G. Hefferon, R. Divakaruni, E. Crabbe, X. Chen: *Reduction of RTA-Driven Intra-Die Variation via Model-Based Layout Optimization*; In: Symposium on VLSI Technology, Juni 2009, pp. 152-153.

[82] L.-T. Pang, K. Qian , C. J. Spanos, B. Nikolic : *Measurement and Analysis of Variability in 45nm Strained-Si CMOS Technology*; In: IEEE Journal of Solid-State Circuits, Vol. 44, No. 8, August 2009, pp. 2233-2243.

[83] S. Ekbote, K. Benaissa, B. Obradovic, S. Liu, H. Shichijo, F. Hou, T. Blythe, T. W. Houston, S. Martin, R. Taylor, A. Singh, H. Yang, G. Baldwin: *45nm Low-Power CMOS SoC Technology with Aggressive Reduction of Random Variation for SRAM and Analog Transistors*; In: International Symposium on VLSI Technology, Juni 2008, pp. 160-161.

[84] H. Fukutome, Y. Hori, L. Sponton, K. Hosaka, Y. Momiyama, S. Satoh, R. Gull, W. fichtner, T. Sugii: *Comprehensive Design Methodology of Dopant Profile to Suppress Gate-LER-induced Threshold Voltage Variability in 20nm NMOSFETs*; In: Symposium on VLSI Technology, Juni 2009, pp. 146-147.

[85] H. Iwai: *Roadmap for 22 nm and Beyond*; In: Microelectronic Engineering Vol. 86, No. 7-9, Juli/September 2009, pp. 1520-1528.

[86] C. Pacha, K. von Arnim, F. Bauer, T. Schulz, W. Xiong, K.T. San, A. Marshall, T. Baumann, C.-R. Cleavelin, K. Schruefer, J. Berthold: *Efficiency of Low-Power Design Techniques in Multi-Gate FET CMOS Circuits*; In: European Solid State Circuits Conference, September 2007, pp. 111-114.

[87] K. Kuhn: *Moore's Law Past 32nm: Future Challenges in Device Scaling*; In: International Workshop on Computational Electronics, Mai 2009, pp. 1-6.

[88] *http://www.arm.com/products/CPUs/ARM926EJ-S.html*.

[89] T. Lueftner, J. Berthold, C. Pacha et al.: *A 90-nm CMOS Low-Power GSM-EDGE Multimedia-Enhanced Baseband Processor With 380-MHz ARM926 Core and Mixed-Signal*; In: IEEE Journal of Solid-State Circuits, Vol. 42, No. 1, Januar 2007, pp. 1-12.

[90] *www.synopsys.com*.

[91] M. Eisele, J. Berthold, D. Schmitt-Landsiedel, R. Mahnkopf: *The Impact of Intra-Die Device Parameter Variations on Path Delays and on the Design for Yield of Low Voltage Digital Circuits*; In: IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 5, No. 4, Dezember 1997, pp. 360-368.

[92] N. James, P. Restle, J. Friedrich, B. Huott, B. McCredie: *Comparison of Split- Versus Connected-Core Supplies in the POWER6T Microprocessor*; In: IEEE International Solid State Circuits Confernece, Februar 2007, pp. 298-299.

[93] V. Stojanovic, V. G. Oklobdzija: *Comparative Analysis of Master-Slave Latches and Flip-Flops for High-Performance and Low-Power Systems*; In: IEEE Journal of Solid-State Circuits, Vol. 34, No. 4, April 1999, pp. 536-548.

[94] H. Abrishami, S. Hatami, M. Pedram: *Characterization and Design of Sequential Circuit Elements to Combat Soft Error*; In: IEEE International Conference on Computer Design, Oktober 2008, pp. 194-199.

[95] P. Min, C. Chu, Z. Hai: *Timing Yield Estimation Using Statistical Static Timing Analysis*; In: IEEE International Symposium on Circuits and Systems, Mai 2005, pp. 2461-2464.

[96] R. Srinivasa, J. Srivatsava, N. Tondamuthuru: *Process Variability Analysis In DSM Through Statistical Simulations And Its Implications To Design Methodologies*; In: International Symposium on Quality Electronic Design, März 2008, pp. 325-329.

[97] D. Harris, R. Ho, G.-Y. Wei, M. Horowitz: *The Fanout-of-4 Inverter Delay Metric*; In: Unveröffentlichtes Manuskript: http://odin.ac.hmc.edu/ harris/research/FO4.pdf.

[98] X. Liang, D. Brooks: *Microarchitecture Parameter Selection to Optimize System Performance Under Process Variation*; In: IEEE/ACM Conference on Computer Aided Design, November 2006, pp. 429-436.

[99] D. Brooks, P. Bose, V. Srinivasan, M. K. Gschwind, P.G. Emma, M.G. Rosenfield: *New Methodology for Early-stage, Microarchitectural-level Power-performance Analysis of Microprocessors*; In: IBM Journal for Research and Developement Vol.47, No.5/6, September/November 2003, pp. 585-598.

[100] R. Berridge, R. Averil, A. Barish et al.: *IBM POWER6 Microprocessor Physical Design and Design mMethodology*; In: IBM Journal of Research and Developement, Vol. 51, No. 6, November 2007, pp. 685-714.

[101] M.S. Hrishikesh, N. P. Jouppi, K. I. Farkas, D. Burger, S. W. Keckler, P. Shivakumary: *The Optimal Logic Depth Per Pipeline Stage is 6 to 8 FO4 Inverter Delays*; In: International Symposium on Computer Architecture, 2002, pp. 14-24.

[102] S. Naffziger: *Microprocessors of the Future: Commodity or Engine of Growth?*; In: IEEE Solid-State Magazine, 2009, pp. 76-82.

[103] V. Srinivasan, D. Brooks, M. Gschwind, P. Bose, V. Zyuban, P. N. Strenski, P. G. Emma: *Optimizing Pipelines for Power and Performance*; In: IEEE/ACM International Symposium on Microarchitecture, November 2002, pp. 333-344.

[104] A. Hartstein, T. R. Puzak: *Optimum Power/Performance Pipeline Depth*; In: IEEE/ACM International Symposium on Microarchitecture, Dezember 2003, pp. 117-125.

[105] G. Gerosa, S. Gary, C. Dietz et al.: *A 2.2W, 80MHz Superscalar RISC Microprocessor*; In: IEEE Journal of Solid-State Circuits, Vol. 29, No. 12, Dezember 1994, pp. 1440-1454.

[106] V. Stojanovic, V. G. Oklobdzija: *Comparative Analysis of Master-Slave Latches and Flip-Flops for High-Performance and Low-Power Systems*; In: IEEE Journal of Solid-State Circuits, Vol. 34, No. 4, April 1999, pp. 536-548.

[107] D. G. Chinnery, K. Keutzer: *Closing the gap between ASIC and custom: an ASIC perspective*; In: IEEE/ACM Design Automation Conference, 2000, pp. 637-642.

[108] D. G. Chinnery: *Low Power Design Automation*; In: Dissertation an der University of California, Berkeley, 2006.

[109] K. von Arnim, K. Schruefer, T. Baumann, K. Hofmann, T. Schulz, C. Pacha, J. Berthold: *A Voltage Scaling Model for Performance Evaluation in Digital CMOS Circuits*; In: Accepted Paper, IEEE International Electron Devices Meeting, Dezember 2009.

[110] D. Rabe, W. Nebel: *Short Circuit Power Consumption of Glitches*; In: International Symposium on Low Power Electronics and Design, August 1996, pp. 125-128.

[111] C. Yeh, G. Wilke, H. Chen, S. Reddy, H. Nguyen, T. Miyoshi, W. Walker, R. Murgai: *Clock Distribution Architectures: A Comparative Study*; In: International Symposium on Quality of Electronic Design, März 2006, pp. 85-91.

[112] A. Rajaram, H. Jiang, R. Mahapatra: *Reducing Clock Skew Variability via Crosslinks*; In: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 25, No. 6, Juni 2006, pp. 1176-1182.

[113] M. Orshansky, L. Milor, P. Chen, K. Keutzer, C. Hu: *Impact of Spatial Intrachip Gate Length Variability on the Performance of High-Speed Digital Circuits*; In: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 21, No. 5, Mai 2002, pp. 544-553.

[114] K. Arabi, R. Saleh, X. Meng: *Power Supply Noise in SOCs: Metrics, Management, and Measurement*; In: IEEE Design and Test of Computers, Vol. 24, No. 3, Juni 2007, pp. 236-244.

[115] E. Alon, M. Horowitz: *On-Chip Power Supply Noise Measurement and Regulation Techniques for PLLs*; In: IBM Academy of Technology, PLL Best Practices Conference, September 2005.

[116] L. Yuyun, G. Mehta, R. Abdel Karim, V. Le, J. Gandhi: *An Improved ASIC/SOC Design Methodology for Quick Design Convergence*; In: Internationsl Conference on Solid-State and Integrated Circuit Technology, Oktober 2006, pp. 1883-1885.

[117] H. Menager, R. Kadiyala: *An Efficient Approach to the Challenges of a True Multi-chip Integration into a Single SOC*; In: Sophia Antipolis Micro Electronics, Oktober 2004.

[118] M. R. Becer, D. Blaauw, V. Zolotov, R. Panda, I. N. Hajj: *Analysis of Noise Avoidance Techniques in DSM Interconnects Using a Complete Crosstalk Noise Model*; In: Design, Automation and Test in Europe Conference, 2002, pp. 456-463.

[119] MIPS Technologies: Marktanalysen von MIPS Technologies; Kontaktperson: Karin Neubert, Sales Europe (neubert@mips.com), Erhalt der Daten am 23. April 2009.

[120] *www.arm.com/markets/mobile_ solutions/app.html.*

[121] T. Baumann, Doris Schmitt-Landsiedel, Christian Pacha: *Impact of Technology and Microarchitecture on the Robustness of Embedded Low-Power Microprocessors*; In: Vortrag auf Kleinheubacher Tagung, Miltenberg, September 2008.

[122] T. Baumann, D. Schmitt-Landsiedel, C. Pacha: *Architectural Assessment of Design Techniques to Improve Speed and Robustness in Embedded Microprocessors*; In: IEEE/ACM Design Automation Conference, Juli 2009.

[123] M. S. Floyd, S. Ghiasi, T. W. Keller: *System Power Management Support in the IBM POWER6 Microprocessor*; In: IBM Journal of Research and Developement, Vol. 51, No. 6, November 2007, pp. 733-746.

[124] B. Curran, E. Fluhr, J. Paredes: *Power-Constrained High-Frequency Circuits for the IBM POWER6 Microprocessor*; In: IBM Journal of Research and Developement, Vol. 51, No. 6, November 2007, pp. 715-732.

[125] C. Chiang, J. Kawa: *Design for Manufacturability and Yield for Nano-Scale CMOS*; In: Springer Verlag, 2007, ISBN: 978-1-4020-5187-6.

[126] I. Nitta, T. Shibuya, K. Homma: *Statistical Static Timing Analysis Technology*; In: Fujitsu Science Technology Journal 43, Vol. 4, Oktober 2007, pp.516-523.

[127] D. Goswami, K. Tsai, M. Kassab, J. Rajski: *Test Generation in the Presence of Timing Exceptions and Constraints*; In: IEEE/ACM Design Automation Conference, Juni 2007, pp. 688-693.

[128] R. C. Dorf: *The VLSI Handbook*; In: CRC Press, 2007, ISBN: 0-8493-4199-X, pp. 63.18-63.19.

[129] D. Blaauw, K. Chopra, A. Srivastava, L. Scheffer: *Statistical Timing Analysis: From Basic Principles to State of the Art*; In: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 27, No. 4, April 2008, pp. 589-604.

[130] A. Ripp, M. Bühler, J. Koehl, J. Bickord, J. Hibbeler, U. Schlichtmann, R. Sommer, M. Pronath: *DFM/DFY Design for Manufacturability and Yield - Influence of Process Variations in Digital, Analog and Mixed-Signal Circuit Design*; In: Design, Automation and Test in Europe, 2006, pp. 387-392.

[131] U. Schlichtmann, M. Schmidt, M. Pronath, H. Kinzelbach, V. Glöckel, M. Dietrich, J. Haase, and U. Eichler: *Digital Design at a Crossroads*; In: Design, Automation and Test in Europe, April 2009.

[132] S. Shi, A. Ramalingam, Daifeng Wang, D. Pan: *Latch Modeling for Statistical Timing Analysis*; In: Design, Automation and Test in Europe, März 2008, pp. 1136-1141.

[133] H. Kinzelbach: *Statistical Variation Analysis for Digital Design*; In: Workshop on Design of Future Reliable Systems from Unreliable Components am Lehrstuhl für Entwurfsautomatisierung, TU München, 9. Juli 2009.

[134] V. G. Oklobdzija: *Clocking and Clocked Storage Elements in a Multi-Gigahertz Environment*; In: IBM Journal for Research and Developement Vol.47, No.5/6, September/November 2003, pp. 567-583.

[135] K. Bowman, J. Tschanz, M. Khellah, M. Ghoneima, Y. Ismail, V. De: *Time-Borrowing Multi-Cycle On-Chip Interconnects for Delay Variation Tolerance*; In: International Symposium on Low-Power Electronics and Design, Oktober 2006, pp. 79-84.

[136] M. Garg: *High Performance Pipelining Method for Static Circuits Using Heterogenous Pipelining Elements*; In: European Solid-State Circuits Conference, September 2003, pp. 185-188.

[137] J. G. Xi, W. W.-M. Dai: *Useful-Skew Clock Routing With Gate Sizing for Low Power Design*; In: IEEE/ACM Design Automation Conference, Juni 1996, pp. 383-388.

[138] J.-K. Wu, T.-Y. Wu, L.-Y. Lu, K.-Y. Chen: *IR Drop Reduction via a Flip-Flop Resynthesis Technique*; In: International Symposium on Quality Electronic Design, März 2008, pp. 78-83.

[139] E. Takahashi, Y. Kasai, M. Murakawa, T. Higuchi: *Post-Fabrication Clock-Timing Adjustment Using Genetic Algorithms*; In: IEEE Journal of Solid-State Circuits, Vol. 39, No. 4, April 2004, pp. 643-650.

[140] A. Jain, D. Blaauw: *Slack Borrowing in Flip-Flop Based Sequential Circuits*; In: ACM Great Lakes Symposium on VLSI, 2005, pp. 96-101.

[141] S. Henzler, S. Koeppe, D. Lorenz, W. Kamp, R. Kuenemund, D. Schmitt-Landsiedel: *Variation Tolerant High Resolution and Low Latency Time-to-Digital Converter*; In: European Solid-State Circuit Conference, September 2007, pp. 194-197.

[142] E. Borch, E. Tune, S. Manne, J. Emer: *Loose Loops Sink Chips*; In: International Symposium on High-Performance Computer Architecture, Februar 2002, pp. 299-310.

[143] Z. Chishti, T. N. Vijaykumar: *Wire Delay is Not a Problem for SMT (in the near future)*; In: ACM International Symposium on Computer Architecture, Juni 2004, pp. 40-51.

[144] A. Tiwari, S. R. Sarangi, J. Torrellas: *ReCycle: Pipeline Adaptation to Tolerate Process Variation*; In: ACM International Symposium on Computer Architecture, Juni 2007, pp. 323-334.

[145] K. A. Bowman, X. Tang, J. C. Eble, J. D. Meindl: *Impact of Extrinsic and Intrinsic Parameter Fluctuations on CMOS Circuit Performance*; In: IEEE Journal of Solid-State Circuits, Vol. 35, No. 8, August 2000, pp. 1186-1193.

[146] K. A. Bowman, S. G. Duvall, J. D. Meindl: *Impact of Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution of Gigascale Integration*; In: IEEE Journal of Solid-State Circuits, Vol. 37, No. 2, Februar 2002, pp. 183-190.

[147] S. Nadarajah, S. Kotz: *Exact Distribution of the Max/Min of Two Gaussian Random Variables*; In: IEEE Transactions on Very Large Scale Integration Systems, Vol. 16, No. 2, Februar 2008, pp. 210-212.

[148] L. S. Nielsen, C. Niessen, J. Sparso, K. van Berkel: *Low-power Operation Using Self-timed Circuits and Adaptive Scaling of the Supply Voltage*; In: IEEE Transaction on Very Large Scale Integration Systems, Vol. 2, No. 4, Dezember 1994, pp. 391-397.

[149] V. Ventktachalam, M. Franz: *Power Reduction Techniques For Microprocessor Systems*; In: ACM Computing Surveys, Vol. 37, No. 3, September 2005, pp. 195-237.

[150] M. Elgebaly, M. Sachdev: *Variation-Aware Adaptive Voltage Scaling System*; In: IEEE Transaction on Very Large Scale Integration Systems, Vol. 15, No. 5, Mai 2007, pp. 560-571.

[151] A. Wang, S. Naffziger: *Adaptive Techniques for Dynamic Processor Optimization*; Springer Verlag 2008, ISBN: 978-0-387-76471-9.

[152] M. Togo, T. Fukai, Y. Nakahara, S. Koyama, M. Makabe, E. Hasegawa, M. Nagase, T. Matsuda, K. Sakamoto, S. Fujiwara, Y. Goto, T. Yamamoto, T. Mogami, M. Ikeda, Y. Yamagata, and K. Imai: *Power-aware 65 nm Node CMOS Technology Using Variable V_{DD} and Back-bias Control with Reliability Consideration for Back-bias Mode*; In: International Symposium on VLSI Technology, Juni 2004, pp. 88-89.

[153] A. Montree, A. van Brandenburg, D. Klaasen, R. Peset Llopis, Y. Ponomarev, R. Roes, A. Scholten, R. van Veen: *Limitations to Adaptive Back Bias Approach for Standby Power Reduction in Deep Sub-micron CMOS*; In: European Solid-State Device Research Conference, September 1999, pp. 580-583.

[154] M. Anis, M. Aburahma: *Leakage Current Variability in Nanometer Technologies*; In: International Database Engineering & Application Symposium, Juli 2005, pp. 60-63.

[155] S. Narendra, A. Keshavarzi, B. A. Bloechel, S. Borkar, V. De: *Forward Body Bias for Microprocessors in 130nm Technology Generation and Beyond*; In: IEEE Journal of Solid-State Circuits, Vol. 38, No. 5 , Mai 2003, pp. 696-701.

[156] A. Hokazono, S. Balasubramanian, K. Ishimaru, H. Ishhiuchi, C. Hu, T. Liu: *MOSFET Hot-Carrier Reliability Improvement by Forward-Body Bias*; In: IEEE Electron Device Letters, Vol. 27, No. 7, Juli 2006, pp. 605-608.

[157] T. Chen, S.Naffziger: *Comparison of Adaptive Body Bias (ABB) and Adaptive Supply Voltage (ASV) for Improving Delay and Leakage Under the Presence of Process Variation*; In: IEEE Transactions on Very Large Scale Integration Systems, Vol. 11, No. 5, Oktober 2003, pp. 888-899.

[158] P. Huang, S. Ghiasi: *Leakage-Aware Intraprogram Voltage Scaling for Embedded Processors*; In: ACM/IEEE Design Automation Conference, Juli 2006, pp. 364-369.

[159] J. T. Tschanz, J. T. Kao, S. G. Narendra, R. Nair, D. A. Antoniadis, A. P. Chandrakasan, V. De: *Adaptive Body Bias for Reducing Impacts of Die-to-Die and Within-Die Parameter Variations on Microprocessor Frequency and Leakage*; In: IEEE Journal of Solid-State Circuits, Vol. 37, No. 11, November 2002, pp. 1396-1402.

[160] S. Mutoh, T. Douseki, Y. Matsuya, T. Aoki, S. Shigematsu, J. Yamada: *1-V Power Supply High-speed Digital Circuit Technology with Multithreshold-Voltage CMOS*; In: IEEE Journal of Solid-State Circuits, Vol. 30, No. 8, August 1995, pp. 847-854.

[161] L. Su, R. Schulz, J. Adkisson, K. Beyer, G. Biery, W. Cote, E. Crabbe, D. Edelstein, J. Ellis-Monaghan, E. Eld, D. Foster, R. Gehres, R. Goldblatt, N. Greco, C. Guenther, J. Heidenreich, J. Herman, D. Kiesling, L. Lin, S.-H. Lo, J. McKenna, C. Megivern, H. Ng, J. Oberschmidt, A. Ray, N. Rohrer, K. Tallman, T. Wagner, B. Davari: *A High-Performance Sub-0.25 pm CMOS Technology with Multiple Thresholds and Copper Interconnects*; In: Symposium on VLSI Technology, Juni 1998, pp. 18-19.

[162] L. Wei, Z. Chen, K. Roy, M. C. Johnson, Y. Ye, V. K. De: *Design and Optimization of Dual-Threshold Circuits for Low-Voltage Low-Power Applications*; In: IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol. 7, No. 1, März 1999, pp. 16-24.

[163] N. Sirisantana, L. Wei, K. Roy: *High-Performance Low-Power CMOS Circuits Using Multiple Channel Length and Multiple Oxide Thickness*; In: IEEE International Conference on Computer Design, 2000, pp. 227-232.

[164] P. Gupta, A. B. Khang, P. Sharma, D. Sylvester: *Selective Gate-Length Biasing for Cost-Effective Runtime Leakage Control*; In: Design Automation Conference, 2004, pp. 327-330.

[165] S. Rusu, S. Tam, H. Muljono, D. Ayers, J. Chang: *A Dual-Core Multi-Threaded Xeon Processor with 16MB L3 Cache*; In: IEEE International Solid-State Circuits Conference, 2006, p. 118.

[166] T. Baumann, J. Berthold, T. Niedermeier, T. Schoenauer, J. Dienstuhl, D. Schmitt-Landsiedel, C. Pacha: *Performance Improvement of Embedded Low-Power Microprocessor Cores by Selective Flip Flop Replacement*; In: European Solid State Circuits Conference, September 2007, pp. 308-311.

[167] L.T. Clark, E.J. Hoffman, J. Miller, M. Biyani, L. Luyun, S. Strazdus, M. Morrow, K.E. Velarde, M.A. Yarch: *An Embedded 32-b Microprocessor Core for Low-Power and High-Performance Applications*; In: IEEE Journal of Solid-State Circuits, Vol. 36, No. 11, November 2001, pp. 1599-1608.

[168] J. Warnock, D. Wendel, T. Aipperspach, E. Behnen, R. A. Cordes, S. H. Dhong, K. Hirairi, H. Murakami, S. Onishi, D. C. Pham, J. Pille, S. D. Posluszny, O. Takahashi, H. Wen: *Circuit Design Techniques for a First-Generation Cell Broadband Engine Processor*; In: IEEE Journal of Solid-State Circuits, Vol. 41, No. 8, August 2006, pp. 1692-1706.

[169] B. Curran, E. Fluhr, J. Paredes: *Power-constrained High-Frequency Circuits for the IBM POWER6 Microprocessor*; In: IBM Journal of Research and Developement, Vol. 51, No. 6, November 2007, pp. 715-732.

[170] D. Sylvester, D. Blaauw, E. Karl: *ElastIC: An Adaptive Self-Healing Architecture for Unpredictable Silicon*; In: IEEE Design and Test of Computers, Vol. 23, No. 6, Juni 2006, pp. 484-490.

[171] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, T. Mudge: *Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation*; In: IEEE/ACM Annual Symposium on Microarchitecture, Dezember 2003, pp. 7-18.

[172] D. Blaauw, S. Kalaiselvan, K. Lai, W. Ma, S. Pant, C. Tokunaga, S. Das, D. Bull: *RazorII: In Situ Error Detection and Correction for PVT and SER Tolerance*; In: IEEE International Solid-State Circuit Conference, Februar 2008, pp. 400-401.

[173] S. Mitra: *Globally Optimized Robust Systems to Overcome Scaled CMOS Reliability Challenges*; In: Design, Automation and Test in Europe, März 2008, pp. 941-946.

[174] G. Wei, M. Horowitz: *A Fully Digital, Energy-Efficient, Adaptive Power-Supply Regulator*; In: IEEE Journal of Solid-State Circuits, Vol. 34, No. 4, April 1999, pp. 520-528.

[175] T. Brusev, P. Goranov, M. Hristov: *Buck Converter for Low Power Applications*; In: International Conference on Microelectronics, Mai 2008, pp. 447-450.

[176] J. Lee, G. Hatcher, L. Vandenberghe, C. K. Yang: *Evaluation of Fully-Integrated Switching Regulators for CMOS Process Technologies*; In: IEEE Transaction on Very Large Scale Integration Systems, Vol. 15, No. 9, September 2007, pp. 1017-1027.

[177] M. Manninger: *Power Management for Portable Devices*; In: European Solid State Circuit Conference, September 2007, pp. 167-173.

[178] Y. Taur, T. H. Ning: *Fundamentals of Modern VLSI Devices*; In: Cambridge University Press, 2002, ISBN: 0-521-55056-4.

[179] J. Tschanz, K. Bowman, S. Walstra, M. Agostinelli, T. Karnik, V. De: *Tunable Replica Circuits and Adaptive Voltage-Frequency Techniques for Dynamic Voltage, Temperature, and Aging Variation Tolerance*; In: IEEE International Symposium on VLSI Circuits, Juni 2009.

[180] M. B. Ketchen, M. Bhushan: *Product-representative 'at Speed' Test Structures*; In: IBM Journal of Research and Developement, Vol. 50, No. 4/5, Juli/September 2006, pp. 452-468.

[181] A. J. Drake, R. M. Senger, H. Deogun, G. Carpenter, S. Ghiasi, T. Nguyen, N. James, M. Floyd, V. Pokala: *A Distributed Critical-Path Monitor for a 65nm High-Performance Microprocessor*; In: IEEE International Solid-State Conference, Februar 2007, pp. 398-399.

[182] A. J. Drake, R. M. Senger, H. Singh, G. D. Carpenter, N. K. James: *Dynamic Measurement of Critical-Path Timing*; In: IEEE International Conference on Integrated Circuit Design and Technology, Juni 2008, pp. 249-252.

[183] J. Keane, D. Persaud, Chris H. Kim: *An All-In-One Silicon Odometer for Separately Monitoring HCI, BTI, and TTDB*; In: IEEE International Symposium on VLSI Circuits, Juni 2009.

[184] S. Das, D. Roberts, L. Seokwoo, S. Pant, D. Blaauw, T. Austin, K. Flautner, T. Mudge: *A Self-tuning DVS Processor Using Delay-error Detection and Correction*; In: IEEE Journal of Solid-State Circuits Conference, Vol. 41, No. 4, April 2006, pp. 792-804.

[185] J. Gregg, T. W. Chen: *Post Silicon Power/Performance Optimization in the Presence of Process Variations Using Individual Well-Adaptive Body Biasing*; In: IEEE Transactions on Very Large Scale Integration Systems, Vol. 15, No. 3, März 2007, pp. 366-376.

[186] B. Stefano, D. Bertozzi, L. Benini, E. Macii: *Process Variation Tolerant Pipeline Design Through a Placement-Aware Multiple Voltage Island Design Style*; In: Design, Automation and Test in Europe, März 2008, pp. 967-972.

[187] F. Sill, F. Grassert, D. Timmermann: *Reducing Leakage with Mixed-VT (MVT)*; In: IEEE International Conference on VLSI Design, 2005, pp. 874-877.

[188] J. Tschanz, S. Narendra, Z. Chen, S. Borkar, M. Sachdev, V. De: *Comparative Delay and Energy of Single Edge-Triggered and Dual Edge-Triggered Pulsed Flip-Flops for High-Performance Microprocessors*; In: International Symposium on Low Power Electronics and Design, 2001, pp. 147-152.

[189] A. Venkatraman, R. Garg, S. P. Khatri: *A Robust, Fast Pulsed Flip-Flop Design*; In: ACM Great Lakes Symposium on VLSI, Mai 2008, pp. 119-122.

[190] F. Rade, B. Westergren: *Mathematische Formeln*; In: 3. Auflage, Springer Verlag, 2000.

Publikationsliste

T. Baumann, J. Berthold, T. Niedermeier, T. Schoenauer, J. Dienstuhl, D. Schmitt-Landsiedel, C. Pacha: *Performance Improvement of Embedded Low-Power Microprocessor Cores by Selective Flip Flop Replacement*; In: European Solid State Circuits Conference, September 2007, pp. 308-311.

C. Pacha, K. von Arnim, F. Bauer, T. Schulz, W. Xiong, K.T. San, A. Marshall, T. Baumann, C.-R. Cleavelin, K. Schruefer, J. Berthold: *Efficiency of Low-Power Design Techniques in Multi-Gate FET CMOS Circuits*; In: European Solid State Circuits Conference, September 2007, pp. 111-114.

T. Baumann, Doris Schmitt-Landsiedel, Christian Pacha: *Impact of Technology and Microarchitecture on the Robustness of Embedded Low-Power Microprocessors*; In: Vortrag auf Kleinheubacher Tagung, Miltenberg, September 2008.

T. Baumann, D. Schmitt-Landsiedel, C. Pacha: *Architectural Assessment of Design Techniques to Improve Speed and Robustness in Embedded Microprocessors*; In: IEEE/ACM Design Automation Conference, Juli 2009.

K. von Arnim, K. Schruefer, T. Baumann, K. Hofmann, T. Schulz, C. Pacha, J. Berthold: *A Voltage Scaling Model for Performance Evaluation in Digital CMOS Circuits*; In: Accepted Paper, IEEE International Electron Devices Meeting, Dezember 2009.

Abbildungsverzeichnis

1.1 Gliederung der Arbeit nach verschiedenen Abstraktionsebenen. 14

2.1 Überblick über die einzelnen Komponenten der konzipierten Vorgehensweise zur Bewertung des Einflusses von Variationen auf die Geschwindigkeit und Robustheit von integrierten Schaltungen. 18

3.1 Kriterien zur Klassifizierung von Variationen. 22
3.2 Schalttrajektorien von Inverter und 2-fach NAND in 65nm. Im Hintergrund ist die globale, prozessbedingte 1σ Schwankung des Drainstroms I_D eines NMOS Transistors gezeigt. 23
3.3 Räumliche Klassifizierung von Prozessvariationen. 24
3.4 Modellierung zur Bestimmung des Beitrags des Leitungswiderstands zur Laufzeit eines Pfades. 28
3.5 Häufigkeitsverteilung von Leitungskapazität, -widerstand und RC Laufzeit der Verdrahtung (nach Glg. 3.6, 3.7) in 65nm CMOS unter der Annahme globaler, normalverteilter Geometrieschwankungen. 30
3.6 Vergleich von Monte Carlo Simulationen mit standardmäßigem und reduziertem Parametersatz ($V_{DD} = V_{DD}^{nom}$, T=27°C). 31
3.7 Schematische Darstellung der einzelnen IR-Drop Komponenten. 32
3.8 Schematische Darstellung eines Crosstalk relevanten Netzes in einem Fanout-4 Inverter Pfad. 34
3.9 Simulierte crosstalk-bedingte Laufzeiterhöhung der in Bild 3.8 gezeigten Testschaltung in 65nm CMOS ($V_{DD} = V_{DD,nom} + 10\%$, T=27°C). 35
3.10 Gemessene Frequenz einer Crosstalk-Struktur in 45nm low-power CMOS Technologie bei T=27°C. 36
3.11 Zeitkonstanten von Prozess-, Umgebungsvariationen und Alterungseffekten. 37
3.12 Laufzeitsensitivitäten einer NAND2-NOR2 Kette gegenüber L, V_T, μ und V_{DD} Schwankungen. 40
3.13 Änderung der Laufzeitsensitivitäten einer NAND2-NOR2 Kette bei reduzierter Versorgungsspannung. 41
3.14 Anteile an der von L, V_T und μ Variationen induzierten Laufzeitschwankung. 42
3.15 Skalierungsverhalten der statistischen Einsatzspannungsschwankung. 44
3.16 Simulierte Crosstalk-induzierte Laufzeitänderung für 90nm, 65nm und 40nm CMOS auf Basis extrahierter Netzlisten ($V_{DD} = V_{DD,nom} + 10\%$, T=27°C). 47
3.17 Simulierte relative Laufzeitänderung aufgrund von Temperaturschwankungen in 40nm CMOS auf Basis extrahierter Netzlisten. 48
3.18 Simulierte technologie- und schaltungstechnikabhängige Unterschiede der Laufzeitschwankung auf Basis extrahierter Netzlisten (Nomineller Betrieb bei: $V_{DD} = V_{DD}^{nom}$, T=27°C). 49

3.19 Vergleich der Laufzeitschwankung von Gattermix und ND2-NR2 Äquivalent für verschiedene Technologieknoten auf Gatterebene (Einzelgatter). Simulationsergebnisse auf Basis extrahierter Netzlisten (Nomineller Betrieb bei: $V_{DD} = V_{DD}^{nom}$, T=27°C). 50

4.1 Pfadverteilung des untersuchten ARM926 Designs in 90nm CMOS. 54
4.2 Pfadspektrum der geschwindigkeitskritischen Pfade des ARM926 Designs. . 55
4.3 Akkumulierte Verteilung der Laufzeitbeiträge aller kritischen Logikpfade. . 56
4.4 Verteilung der Logiktiefe n_{Log} bzw. Gatteranzahl n_{Gatter} in den kritischen Pfaden. 57
4.5 Logiktiefe und Treiberstärken im untersuchten ARM926. 58
4.6 Treiber zu Last Verhältnis für alle Gatter im kritischen Timing-Bereich. . . 59
4.7 Verteilung von Verdrahtungs- und Koppelkapazitätsanteilen an der Gesamtkapazität der Netze innerhalb des kritischen Timing Bereichs. 60
4.8 Wahrscheinlichkeitsverteilung für das Eintreten einer um den Faktor x reduzierten crosstalk-bedingten worst-case Laufzeiterhöhung. 61
4.9 Verteilung der maximalen Anzahl an Clock Buffern nach dem Aufspalten im Taktbaum. 63
4.10 Pfadspektrum der Hold-Zeit kritischen Pfade des ARM926 Designs. 64
4.11 Verteilung der Logiktiefe und der mittleren Treiberstärke in den Hold-Zeit kritischen Pfaden. 65
4.12 Verteilung der Anzahl an Clock Buffern nach dem Aufspalten im Taktbaum für Hold-Zeit kritische Pfade. 66
4.13 Skalierungsverhalten der Fanout-4 Laufzeit im Vergleich zu einem repräsentativen Gattermix. 67
4.14 Generische Pipelinestruktur: Registerelemente, Logik und Taktbaum. . . . 68
4.15 NAND2-NOR2 Kette: Generische Struktur zur Nachbildung des Laufzeitverhaltens kritischer Pfade. 69
4.16 Spannungsabhängigkeit von Inverter-Kette, NAND2-NOR2 Kette und repräsentativem kritischen Pfad des ARM926 Mikroprozessors in 65nm CMOS (T=27°C). 70
4.17 Gatter- und pfadtopologieabhängige Dämpfung von statistischen Laufzeitvariationen. 73
4.18 Spannungsabhängiges Laufzeit-Modell für kombinatorische Logik und Inverter Kette in 65nm CMOS. Die Simulation basiert auf extrahierten Netzlisten (T=27°C). 74
4.19 Gemessenes Spannungsverhalten logik-, leitungs- und crosstalkdominierter Pfade in 45nm CMOS (T=27°C). 75
4.20 Modellierung des Taktverteilungsnetzes. 77
4.21 Berücksichtigung von WID Laufzeitschwankungen im Taktverteilungsnetz. 79
4.22 Schematische beispielhafte Darstellung der Versorgungsspannungsschwankungen in den verschiedenen Schaltungsteilen, die zur Clock Jitter Abschätzung berücksichtigt werden müssen. 81
4.23 Marktdaten eingebetteter Mikroprozessoren: Hersteller und Architekturen [119]. 84
4.24 Struktur der Integer-Pipeline aller untersuchten ARM Mikroprozessoren. . 86
4.25 Laufzeitbeiträge von Pipelineregistern, Logik und Timing Unsicherheit für ARM9, ARM11 und ARM Cortex A8. 87

4.26 Beiträge zur Timing Unsicherheit. 87
4.27 Ergebnisse des Mikroprozessormodells für ARM9, ARM11 und ARM Cortex A8 in 65nm low-power CMOS. 89
4.28 Semi-Custom low-power und Full-Custom high-speed Prozessoren im Vergleich. 90

5.1 Schematische Darstellung des Schaltverhaltens von Flip Flop Zellen. 97
5.2 Übertragung von Laufzeitschwankungen auf nachfolgende Pipelinestufen. . 98
5.3 Klassifizierung von Pfadtopologien nach Lage und Umgebung von kritischen Pfaden. 99
5.4 Einfluss der Pfadtopologie auf die Verteilung von Setup-Zeit Verletzungen bei Überlagerung globaler systematischer und lokaler statistischer Laufzeitvariationen. 101
5.5 Gatter- und Pfadspektrum des ARM926 Produktdesigns in 90nm CMOS Technologie. 104
5.6 Schematische Darstellung der topologischen Korrelation von Pfaden durch die gemeinsame Nutzung von Gattern. 105
5.7 Schematische Darstellung der in [145, 146] getroffenen Annahmen und die mögliche Erweiterung dieses Ansatzes. 106
5.8 Untersuchung des Einflusses topologischer Korrelationen auf die statistische Laufzeitschwankung unter der Annahme zeitlich gleich langer kritischer Pfade und der Vernachlässigung sub-kritischer Pfade. 107
5.9 Ergebnisse eines erweiterten Matlab Modells zur Analyse des Einflusses der absoluten Anzahl unkorrelierter und topologisch korrelierter Pfade auf die akkumulierte Verteilung der maximalen Pfadlaufzeit. 108
5.10 Wahrscheinlichkeitsverteilung der maximalen Pfadlaufzeit für die Nachbildung verschiedener Pfadspektren und Vergleich mit der Methodik unkorrelierter Pfade mit maximaler nomineller Laufzeit. 110
5.11 Einfluss lokaler Laufzeitschwankungen bei topologisch korrelierten und unkorrelierten Pfaden. 111
5.12 Veranschaulichte Definition der Schaltungssensitivität. 113
5.13 Darstellung des Gewichtungsfaktors für einen kritischen Timingbereich von 10%. 114

6.1 Klassifizierung verschiedener Techniken zur Kompensation von Laufzeitvariationen. 118
6.2 Aufwand zur Implementierung verschiedener Kompensationstechniken. . . 119
6.3 Abschätzung der dynamischen Energieersparnis durch PVS bei T=27°C. . 121
6.4 Gemessene normierte Taktfrequenz und normierte Energieaufnahme eines ARM926 kritischen Pfades in 65nm und 45nm low-power CMOS bei T=27°C ($\alpha_{Schalt} = 0.1$). 123
6.5 Technologietrend: Body-Effekt-bedingtes ΔV_T bei nominellem V_{DD} normiert auf 130nm low-power CMOS. 124
6.6 Technologietrend: Einfluss der Body-Source Spannung auf die Laufzeit von CMOS Digitalschaltungen ($V_{DD} = 1.2V$, T=27°C). 125
6.7 Messergebnisse eines ARM926 kritischen Pfades in 65nm low-power CMOS bei symmetrischem Reverse und Forward Body Biasing ($V_{DD} = V_{DD,nom}$, T=27°C) 126

6.8 Normierte Energieaufnahme und Maximalfrequenz bei ABB eines ARM926 kritischen Pfads auf nominellem Die ($\alpha_{Schalt} = 0.1$). 127
6.9 Vergleich der Messergebnisse von PVS/AVS und ABB hinsichtlich Energieerhöhung zur Adaption der Pfadlaufzeiten in 45nm CMOS (nomineller Die, 27°C, $V_{DD} = V_{DD}^{nom}$). 128
6.10 Grundprinzip verschiedener Monitorkonzepte zur Überwachung des Schaltungszustands. 130
6.11 Normierte Laufzeitsensitivitäten von Reg-V_T, High-V_T und Long-L_{Poly} NAND2-NOR2 Pfaden in 40nm CMOS Technologie ($V_{DD} = V_{DD,nom}, T = 27°C$). . . 135
6.12 Simulierte Erhöhung bzw. Abnahme der aufgenommen Energie in Abhängigkeit des Leckstromanteils an der Gesamtenergieaufnahme in 40nm CMOS bei gleicher Geschwindigkeit($V_{DD} = V_{DD}^{nom} + 60mV$, T=85°C). . . . 136
6.13 Laufzeitsensitivitäten einer low-V_T FO4 Inverter-Kette normiert auf eine reg-V_T Implementierung ($V_{DD} = V_{DD}^{nom}$, T=27°C). 138
6.14 Mikroprozessormodell - Vergleich der Beiträge des Taktverteilungsnetzes zur Laufzeitschwankung für reg-V_T und low-V_T Taktverteilungsnetze. . . . 139
6.15 Schematisches Ablaufdiagramm des robustheitsorientierten Ersetzungsalgorithmus. 143
6.16 Gatter- und Pfadspektrum des ARM926 vor und nach selektivem Einsatz von low-V_T Gattern. 144
6.17 P-FF mit aufgespaltetem Propagationspfad zur Beschleunigung der Clock-Q Laufzeit. 147
6.18 Simulierte Clock-Q Laufzeit von MS-FF und P-FF in Abhängigkeit der Data-Clock Laufzeit auf Basis extrahierter Netzlisten in 65nm CMOS ($V_{DD} = V_{DD,nom}$, T=27°C). 148
6.19 Schematische Darstellung der Ersetzungsstrategie. 150
6.20 Pfad- und akkumulierte Gatterverteilung des ARM926 in 90nm CMOS vor und nach dem Ersetzen von MS-FFs in den kritischen Pfaden. 151
6.21 Gemessene maximale Taktfrequenz der obigen Anordnung für den Einsatz von MS-FFs und P-FFs in loop-internen kritischen Pfaden am Beispiel eines 65nm CMOS Testchips bei T=27°C. 152
6.22 Simulation der Energieaufnahme von P-FF und MS-FF in 65nm CMOS auf Basis extrahierter Netzlisten ($V_{DD} = V_{DD}^{nom}$, T=27°C). 153
6.23 Schematische Darstellung der Ersetzungsstrategie. 155
6.24 Gepulstes Latch mit externem Pulsgenerator. 156
6.25 Simulierte Energieaufnahme von MS-FF und P-L in 65nm CMOS in Abhängigkeit der Schaltaktivität ($V_{DD} = V_{DD}^{nom}$, T=27°C). 158
6.26 Pfad- und Gatterspektrum des ARM926 für das ursprüngliche MS-FF Design und nach globalem Einsatz von P-Ls. 159
6.27 Veranschaulichung von Time-Borrowing in Schaltungen mit gepulsten Latches. 160
6.28 Ergebnisse des Mikroprozessormodells für den globalen Einsatz von gepulsten Latches in ARM926, ARM1176 und ARM Cortex A8. 161
6.29 Zeitliche Sicherheitsmarge und normierter Schaltungssensitivitätsfaktor der einzelnen präventiven Kompensationstechniken für den ARM926 in 90nm CMOS. 163
6.30 Modell-Genauigkeit und Nachbildung des Gatterspektrums durch die Verwendung von Templates als Mehrfach-Instanzen. 166

6.31 Validierung des Schaltungssensitivitätsfaktors. 167
6.32 Methodik zur Validierung des Schaltungssensitivitätsfaktors. 168

Tabellenverzeichnis

2.1 Übersicht über die technologischen Kernparameter von 180nm bis 45nm low-power CMOS Technologien (reg-V_T Transistoren). 19

3.1 Relevanter Temperaturbereich von CMOS Digitalschaltungen. 33
3.2 In der Literatur zu findende Aussagen zu relativen Schwankungsbreiten der wichtigsten Einflussparameter. 38
3.3 Relativer Fehler der linearen Approximation für eine 3σ Auslenkung der wichtigsten Transistorparameter. 40

4.1 Abschätzung der Wahrscheinlichkeit des worst-case Crosstalk Szenarios in Abhängigkeit der Aggressoranzahl bei p=0.5. 62
4.2 Prozessschwankungsbedingte, relative 1σ Laufzeitschwankung für NAND2-NOR2 Kette und Mittelwert eines für kritische Pfade repräsentativen Gattermixes. 69
4.3 Laufzeitbeiträge im Logikpfad für verschiedene Pfadtypen. 71
4.4 Pipelining und Parallelität eingebetteter Mikroprozessoren. 84

5.1 Die mittels Multi-Stage STA identifizierten Pfadtypen des untersuchten ARM926 in den oberen 5% des Pfad Timings. 101
5.2 Topologischer Korrelationsfaktor in den geschwindigkeitskritischen Pfaden des untersuchten ARM926 Produktdesigns. 106

6.1 Relative Laufzeitänderung eines low-V_T Taktpfades im Vergleich zur reg-V_T Implementierung. 138
6.2 Relative Leckstromerhöhung für low-V_T Taktverteilungsnetze in ARM926, ARM1176 und ARM Cortex A8. 139
6.3 Normierte Sensitivitätsfaktoren für den Einsatz von low-V_T Gattern. 142
6.4 Zusammenfassung zum low-V_T Gatter Einsatz in einem 90nm ARM926. . . 145
6.5 Zusammenfassung für den Einsatz von P-FFs in geschwindigkeitskritischen Pfaden eines 90nm ARM926. 154
6.6 Übersicht über die Ergebnisse des globalen Einsatzes von P-Latches in einem 90nm ARM926. 160
6.7 Mikroprozessormodellbasierte Kosten-Nutzen Analyse für den globalen Einsatz gepulster Latches in eingebetteten Mikroprozessoren. 161