Alterungsanalyse komplexer analoger integrierter Schaltungen aus Systemsicht

M.Sc. Theodor Hillebrand

Universität Bremen 2018

Alterungsanalyse komplexer analoger integrierter Schaltungen aus Systemsicht

Dem Fachbereich Physik / Elektrotechnik
der Universität Bremen

zur Erlangung des akademischen Grades eines
DOKTOR-INGENIEUR (Dr.-Ing.)
vorgelegte Dissertation

von

M.Sc. Theodor Hillebrand
aus Bremen

Referent:	Professor Dr.-Ing. St. Paul
Korreferent:	Apl. Prof. Dr.-Ing. habil. H. Gräb
Tag der Einreichung:	13.11.2018
Tag des Promotionskolloquiums:	22.05.2019

Abstract

The design of analog circuits ranges from the specifications on system level, the selection of a suitable circuit topology up to the choice of the concrete physical dimensions of components like transistors. The individual steps are performed within computer-aided design environments. These environments are based on a database made available by the semiconductor manufacturers containing process parameters and influences on the components. In particular, the influences to be considered in the design have increased in recent years due to the continuous reduction of the producible structural sizes. Thus, it must be possible to analyze the deviations due to process, temperature, time degradation and, for special applications, radiation influences during the design phase. Conventional approaches regard these additional effects as standing next to the actual design process. As a result, the latter is no longer consistent and it is much more complex to consider different circuits and effects on different abstraction levels within the design flow.

The focus of this work lies on the development of a consistent consideration of process, voltage, temperature, aging and radiation influences (PVTAR) during the entire design process of analog circuits to the initial measurement of manufactured circuits. To achieve this goal, a transistor model was extended by the influences to be considered. Thereby, the analysis of the additional effects is seamlessly integrated into conventional design processes and methods. In addition, the possibility of a structured analog design is evaluated. This approach allows the estimation of PVTAR influences on dedicated analog function blocks and their propagation on circuit level. Thus, the enormous simulation effort associated with aging analyses can be reduced.
The design and manufacture of circuits is always followed by the measurement of the core properties of these circuits. In the context of this work a method was developed which makes it possible to use all insights from the design of a circuit for the improvement of the measuring results. In addition, the internal parameter sets of individual components can be inferred from the terminal behavior of circuits and systems. Finally, the results of the measurement method can be used for the automated calculation of circuit reliability parameters.

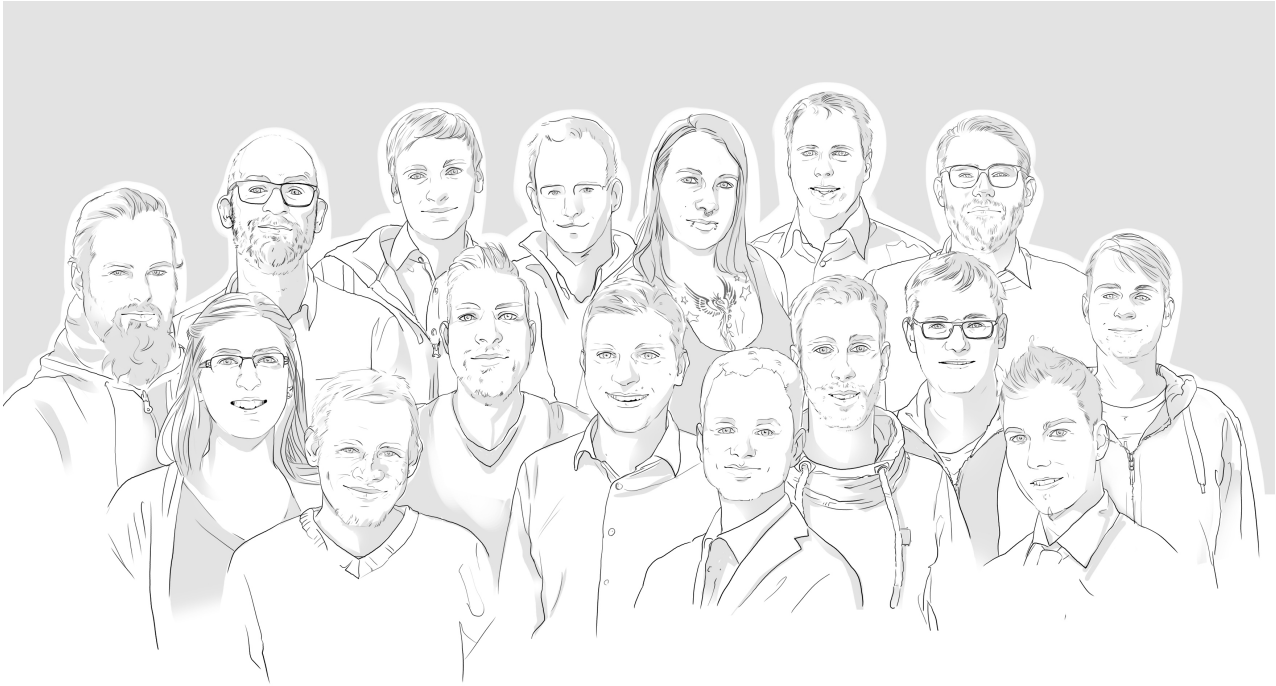
Kurzfassung

Der Entwurf von analogen Schaltungen reicht von den Spezifikationen auf Systemebene, über die Auswahl einer geeigneten Schaltungstopologie bis hin zur Wahl der konkreten physikalischen Abmessungen von Bauteilen wie Transistoren. Die einzelnen Schritte werden innerhalb von computergestützten Entwurfsumgebungen durchgeführt. Diese Entwurfswerkzeuge haben als Basis für den Entwurf eine Datenbank des Halbleiterherstellers mit relevanten Prozessparametern und Einflüssen auf die Bauteile. Insbesondere die im Entwurf zu berücksichtigenden Einflüsse haben in den letzten Jahren durch die kontinuierliche Verkleinerung der herstellbaren Strukturgrößen zugenommen. So müssen die Abweichungen durch die reine Herstellung ebenso wie Temperatur, zeitliche Degradierung und für spezielle Anwendungen Strahlungseinflüsse während des Entwurfs analysiert werden können. Herkömmliche Ansätze betrachten diese zusätzlichen Effekte als neben dem eigentlichen Entwurfsablauf stehend. Dadurch ist dieser nicht mehr konsistent und es ist deutlich aufwrndiger, verschiedene Schaltungen und Effekte auf unterschiedlichen Abstraktionsebenen zu betrachten.

Der Fokus dieser Arbeit liegt auf der Entwicklung einer konsistenten Betrachtung von Prozess-, Spannungs-, Temperatur-, Alterungs- und Strahlungseinflüssen (*engl. Process, Voltage, Temperature, Ageing, and Radiation*, PVTAR) während des gesamten Entwurfsprozess von analogen Schaltungen bis zur initialen Vermessung von gefertigten Schaltungen. Zur Erreichung dieses Ziels wurde ein Transistormodell um die zu betrachtenden Einflüsse erweitert. Somit bettet sich die Analyse der zusätzlichen Effekte nahtlos in herkömmliche Entwurfsprozesse ein. Zudem wird die Möglichkeit eines strukturierten analogen Entwurfs evaluiert. Dieser Ansatz ermöglicht die Abschätzung von PVTAR Einflüssen auf dedizierte analoge Funktionsblöcke und deren Propagierung auf Schaltungsebene. Somit kann der enorme Simulationsaufwand, der mit Alterungsanalysen einhergeht, reduziert werden.

Im Anschluss an das Design und die Fertigung von Schaltungen schließt immer eine messtechnische Erfassung der Kerneigenschaften ebendieser an. Im Rahmen dieser Arbeit wurde eine Methode entwickelt, die es ermöglicht, sämtliche Erkenntnisse aus dem Entwurfsprozess einer Schaltung zur Verbesserung der Messergebnisse zu verwenden. Zudem kann vom Klemmenverhalten von Schaltungen und Systemen auf die internen Parametersätze einzelner Bauelemente rückgeschlossen werden. Schlussendlich können die Ergebnisse der Messmethode zur automatisierten Berechnung von Zuverlässigkeitsparametern von Schaltungen verwendet werden.

Danksagung



» On my own, but not alone «

Inhaltsverzeichnis

1 Einleitung . . . 1
1.1 Eigener Beitrag . . . 4
1.2 Gliederung . . . 5

2 Einführende Grundlagen . . . 7
2.1 Integrierte Transistoren . . . 7
2.2 Transistormodelle . . . 15
2.2.1 Strom-Spannungs-Modellierung . . . 15
2.2.2 Oberflächenpotential-basierte Modellierung . . . 16
2.2.3 Ladungsbasierte Modellierung . . . 20
2.3 Zuverlässigkeit . . . 29
2.3.1 Einflüsse auf integrierte Schaltungen . . . 30
2.3.1.1 Prozessvariationen . . . 30
2.3.1.2 Temperatur und Spannung . . . 32
2.3.1.3 Alterungsmechanismen . . . 35
2.3.2 Skalierung und Zuverlässigkeit . . . 39
2.3.3 Strahlungseinflüsse . . . 40
2.3.4 Stochastische Auswertung . . . 42
2.4 Entwurfsprozess analoger integrierter Schaltungen . . . 47
2.4.1 Festlegung des nominalen Arbeitspunktes . . . 50
2.4.2 Berücksichtigung der Zuverlässigkeit . . . 51
2.5 Zustandsschätzung von Systemen . . . 54
2.5.1 Systembeschreibung . . . 54
2.5.2 Wahrscheinlichkeiten und Stochastik . . . 56
2.5.2.1 Wahrscheinlichkeiten . . . 56
2.5.2.2 Zufallsvariablen . . . 57
2.5.2.3 Statistische Momente . . . 58
2.5.2.4 Statistische Unabhängigkeit . . . 59
2.5.2.5 Stochastische Prozesse . . . 60
2.5.3 Kalman Filter . . . 61
2.5.4 $\boldsymbol{H}_\infty$ Filter . . . 64

2.5.5 Partikelfilter . . . 65
2.5.6 Vergleich der Filteralgorithmen . . . 70

3 Entwurfsprozess unter Berücksichtigung der Zuverlässigkeit . . . 73
3.1 REL-MOS-Transistormodell . . . 75
3.1.1 Alterungseinflüsse . . . 76
3.1.2 Strahlungseinflüsse . . . 84
3.1.3 Gesamtmodell . . . 85
3.2 Stochastischer Schaltungsentwurf . . . 88
3.2.1 Analyse von integrierten Schaltungen . . . 88
3.2.2 Erweiterte g_m/I_D-Methode . . . 93
3.3 Zusammenfassung . . . 100

4 Schaltungsentwurf unter Berücksichtigung von PVTAR-Einflüssen 103
4.1 Topologieabhängige Zuverlässigkeit . . . 104
4.1.1 Funktionsblöcke . . . 105
4.1.2 Schaltungsdesign mit Funktionsblöcken . . . 112
4.1.2.1 Redesign von Schaltungen . . . 112
4.1.2.2 Fallstudie zu Biasschaltungen . . . 115
4.1.2.3 Fallstudie zu Verstärkern . . . 124
4.2 Auswertung der Systemzuverlässigkeit . . . 130
4.3 Gefertigter Verstärker und Referenzquelle . . . 135
4.4 Zusammenfassung . . . 136

5 Messplatz und Extraktionsablauf . . . 139
5.1 Modellparametrierung mittels Partikelfilter . . . 142
5.1.1 Bestimmung der Schwellspannung . . . 142
5.1.1.1 FDSOI FET . . . 146
5.1.1.2 Bulk-MOSFET . . . 148
5.1.1.3 FinFET . . . 148
5.1.2 Messungen am Transistor . . . 149
5.1.3 Messungen am Beta-Multiplier . . . 161
5.2 Beschleunigte Alterung und Vorhersagemöglichkeiten . . . 170
5.3 Zusammenfassung . . . 176

6 Zusammenfassung und Ausblick . . . 177

A Rechenvorschriften der Transistormodelle . . . 181
A.1 Berechnung des Oberflächenpotentials . . . 181
A.2 Oberflächenpotential am Source-Anschluss . . . 182
A.3 Oberflächenpotential am Drain-Anschluss . . . 185
A.4 Hilfsfunktion für die Berechnung des Oberflächenpotentials . . . 185
A.5 BSIM6-Ladungsberechnung . . . 187

A.6 Szenarien zum Vergleich der Filteralgorithmen 189
A.6.1 Beispiel 1 189
A.6.2 Beispiel 2 191
A.7 Parameter des BTI-Modells 192

B Verstärkertopologien 193
B.1 Miller-OTA 195
B.2 Operationsverstärker 198
B.3 Differentieller Kaskode-Operationsverstärker 201
B.4 Folded-Cascode-Operationsverstärker 204
B.5 Folded-Cascode-OTA 207
B.6 Zweistufiger Verstärker mit Kompensation 210
B.7 Dreistufiger Verstärker 213
B.8 Zweistufiger Verstärker 216

Abkürzungsverzeichnis 219

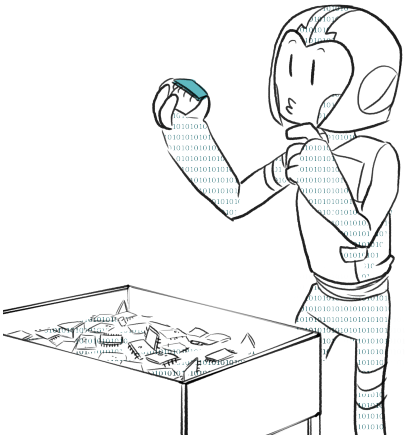
Symbolverzeichnis 221

Wissenschaftliche Veröffentlichungen 227

Literaturverzeichnis 231

1 Einleitung

» Wenn man nur einen Hammer hat, sehen alle Probleme aus wie Nägel. «



Zukünftige Anwendungsgebiete für integrierte elektronische Schaltungen erfordern ein Umdenken im Designablauf ebendieser Komponenten. Eine wichtige Neuerung in der Entwicklung elektronischer Schaltungen ist die Verwendung von commercial off-the-shelf (COTS) Bauelementen in sicherheitskritischen Anwendungen bzw. in Bereichen wie Automotive [Aal16] und Luft- und Raumfahrt [BS13]. Dies ist ebenfalls daran zu erkennen, dass der Anteil der weltweit verkauften analogen integrierten Schaltungen an allen Verkäufen von integrierten Schaltungen von 2017 auf 2018 um fast 5% auf 24,4% angestiegen ist [II18b]. Bereits in 2017 betrug der Umsatz mit analogen Schaltungen in diesem Bereich über 13 Mrd. US-Dollar [II17b], was mehr als einem Drittel aller im Automotive-Bereich verkauften integrierten Schaltungen entspricht [II18a]. Zudem wird ein jährliches Wachstum zwischen 6%-10% für die nächsten Jahre in diesen Marktsegmenten vorausgesagt [II17a]. Somit ist der Bereich der analogen Schaltungen der am stärksten wachsende Markt im Bereich der integrierten Schaltungen [II18c]. Um eben diese Wachstumsraten zu ermöglichen, müssen vermehrt COTS Bauteile zur Anwendung kommen, da diese die Time-to-Market bzw. die Entwicklungsdauer von Komponenten in sicherheitskritischen Anwendungen reduzieren können. Diese

Entwicklung macht jedoch eine stetige und aufwendige Analyse der Zuverlässigkeit und der Ausbeute notwendig [CL12].

Das Design von COTS Baugruppen grenzt sich deutlich von anwendungsspezifischen Designs für sicherheitskritische Anwendungen ab. Die Spezifikationen enthalten in der Regel eine deutlich geringe Lebensdauer, keine Ausfallsicherheit und äußere Einflüsse wie zum Beispiel Strahlung werden vollständig ausgelassen. Für die jeweiligen kommerziellen Anwendungen sind diese Maßnahmen sinnvoll und führen zu einer deutlichen Kostenreduktion sowohl in der Entwicklung als auch in der Fertigung. Durch die einhergehende Massenherstellung können die Kosten abermals gesenkt werden. Dem gegenüber steht die längere Entwicklungsdauer von hochzuverlässigen Schaltungen, die meist in deutlich geringerer Stückzahl hergestellt und vertrieben werden [PMP14a, Sam17]. Um ebendiese Lücke zu schließen, sollen in den kommenden Jahren verstärkt COTS Bauteile und Baugruppen in sicherheitskritischen Anwendungen und rauen Umgebungen zur Anwendung kommen. Diese Arbeit zielt auf den Kern der oben genannten Problematik ab. Der Unterschied der beiden Entwicklungsabläufe beginnt bereits beim jeweils verwendeten Transistormodell. Modelle für sicherheitskritische Anwendungen werden für eine höhere Temperatur-Bandbreite vermessen und beinhalten zudem je nach Anwendungsgebiet zusätzliche Effekte wie Strahlung und Alterung. Die beiden letztgenannten Einflüsse betten sich jedoch nicht nahtlos in bestehende Design-Werkzeuge ein. Strahlung muss in der Regel außerhalb der Entwicklungsumgebung in Technology Computer Aided Design (TCAD) Software separat analysiert werden und wird mittels Kompaktmodellierung an dedizierten Stellen zurück in den Designprozess integriert. Alterungseinflüsse werden im Moment von führenden Software-Herstellern (Cadence, Synopys und Mentor) durch eigene Simulationsabläufe berücksichtigt. Diese können jedoch nicht frei mit allen, für den Designprozess notwendigen, Analysen kombiniert werden. Dabei ist die durchgehende Kombination aller relevanten Effekte über den gesamten Designprozess und sämtliche Entwurfsebenen von entscheidender Bedeutung. Nur dadurch können die inneren Zusammenhänge zwischen allen beteiligten Effekten betrachtet werden. Zu den wichtigen Einflüssen zählen Prozessvariationen, Spannungsschwankungen, Temperatureinflüsse, Alterung und Strahlung; im Weiteren kurz als PVTAR benannt. Die einzelnen Effekte beeinflussen das Verhalten des einzelnen Bauelements und können somit ebenfalls Auswirkungen auf das Systemverhalten der Schaltung haben. Üblicherweise müssen somit umfangreiche Analysen auf Transistorebene durchgeführt werden, um daraus ein geändertes Systemverhalten abzuleiten. Dabei ist immer das nach außen sichtbare Systemverhalten und nicht einzelne Transistoreigenschaften die sicherheitskritische Eigenschaft, die es gilt, möglichst unempfindlich gegenüber den PVTAR Einflüssen auszulegen. Die jeweilige Stärke und Art hängt dabei maßgeblich vom Belastungsprofil der Schaltung über ihre Lebens- und Verwendungszeit ab. Dieses Belastungsprofil muss während des Designprozesses bereits berücksichtigt werden können, um eine Überdimensionierung der Komponenten zu verhindern. Für den Fall, dass die Kombination der Effekte und das Belastungsprofil nicht auf allen Ebenen des Entwurfs betrachtet werden

können, kommt es zum sogenannten Margin-Stacking. Dabei werden die Sicherheitsspannen, die aus den isolierten Effekten für ein konstantes Belastungsprofil gewonnen werden, auf die jeweilige Systemeigenschaft gerechnet. Diese einfache Akkumulation der Effekte bildet eine Überabschätzung der resultierenden Gesamtauswirkung, die weit über dem geforderten 6σ-Bereich liegt. Der Brückenschlag von COTS zu sicherheitsrelevanten Anwendungen wird durch diese diffuse Anordnung von Tools und Effekten verhindert. Designer sind durch die unterschiedlichen Modelle und Modellierungen dazu gezwungen, für sicherheitsrelevante integrierte Schaltungen im Designprozess von vorne zu beginnen.

» Aus einem nahtlosen Transistormodell folgt ein durchgängiger Entwurfsprozess «

Ein in dieser Arbeit vorgestellter Lösungsentwurf ist die Einbettung aller relevanten Effekte direkt in ein Transistormodell, das durch seine physikalische Begründung und mathematische Herleitung die Möglichkeit bietet, PVTAR Effekte intrinsisch zu berücksichtigen. Durch ein solches Modell kann der Designablauf für sämtliche Anwendungsgebiete vereinheitlicht werden. Ursprüngliche COTS Designs können nahtlos für sicherheitsrelevante Anwendungen in der gleichen Entwicklungsumgebung evaluiert werden. Zudem können sämtliche Korrelationen und innere Zusammenhänge zwischen den PVTAR Einflüssen, dem nominellen Entwurfspunkt der Schaltung und dem Systemverhalten analysiert werden.
An das Design schließt sich für integrierte Schaltungen in sicherheitskritischen Anwendungen immer eine messtechnische Evaluierung der Zuverlässigkeit an. Diese Messungen sind für spätere Zertifizierungen durch diverse Standards und Normen vorgeschrieben (JEDEC JESD28 und JEP122H, IEC 61508, DIN 61508) und daher obligatorisch. Je nach Definition des Belastungsprofils wird die Schaltung maximalen Stressbedingungen oder sogar, für den Fall der beschleunigten Messungen, Einflüssen über den Maximalwerten (Temperatur, Spannungen, o.ä.) ausgesetzt. Bei diesen Messungen bestehen zwei, sich gegenseitig bedingende, Probleme. Zum einen dauern die nicht beschleunigten Messungen sehr lange (Tage, Wochen, Monate). Eine Beschleunigung der Messung wird üblicherweise durch eine Erhöhung der Spannungen und Temperaturen erreicht. Falls diese Messungen durch überspezifizierte Belastungen beschleunigt werden, können physikalische Effekte aktiviert werden, die im spezifizierten Betrieb nicht auftreten würden und somit das Messergebnis deutlich verfälschen. Zum anderen können nicht alle Effekte sinnvoll beschleunigt werden und zudem können die Wechselwirkungen zwischen den Effekten durch die Beschleunigung verzerrt werden. Für einzelne Bauelemente, bei denen sämtliche Anschlussspannungen und die Temperatur individuell geregelt werden können, besteht die Möglichkeit, die Stressbedingungen teilweise sinnvoll zu beschleunigen. Jedoch sind in integrierten Schaltungen nur sehr wenige oder gar keine Spannungen innerhalb der Schaltung individuell steuerbar. Dadurch können bei beschleunigter Alterung lediglich die Eingangsspannungen und Versorgungsspannungen erhöht werden. Dies kann wiederum zu Schaltungszuständen führen, die im Betrieb nicht eingenommen werden und somit wird das Belastungsprofil und die De-

gradation stark verzerrt. Diese Art der Beschleunigung ist besonders problematisch für Referenz-Schaltungen. Diese Schaltungen sollen Effekte durch Temperatur und Spannungsschwankungen im Nominalbetrieb kompensieren, somit kann ihre eigentliche Funktion bei erhöhten Stressbedingungen nicht gewährleistet werden. In einem integrierten System sind Referenz-Schaltungen für Arbeitspunkte der nachgeschalteten Baugruppen verantwortlich, weshalb Abweichungen ihres Ausgangsverhaltens starke Auswirkungen auf das gesamte Systemverhalten haben. Daher ist eine Beschleunigung der Alterung für integrierte Schaltungen nicht ohne weiteres möglich.

» Die Übertragung von Kenntnissen aus dem Entwurfsprozess in die Messmethode erhöht die Möglichkeiten und die Qualität von Messungen «

Diese Kontradiktion aus Zeitaufwand und Messungenauigkeit wird in dieser Arbeit mit einem neuen Ansatz angegangen. Ein großer Nachteil bisheriger Verfahren ist, dass sämtliche Erkenntnisse über die Schaltung und deren Verhalten aus dem Entwurfsprozess keinen Einzug in die Messungen halten. Der Ansatz ist, dass während der Messung des Systemverhaltens bei den maximalen spezifizierten Betriebsbedingungen eine Vorhersage der Zuverlässigkeit des Systems berechnet wird. Um dies zu tun, wird am Ende des Designprozesses bzw. zum Tapeout der Schaltung ein Verhaltensmodell des Systems erstellt. Dieses Modell wird während der Messung mittels Markov-Chain-Monte-Carlo-Verfahren (MCMC) [AMGC02] an die Messdaten angepasst. Der Vorteil dieses Vorgehens ist, dass sämtliche Erkenntnisse, die bereits über das System vorliegen, in das Modell mit einfließen können und zudem keine Annahmen über das Systemverhalten in Bezug auf PVTAR Einflüsse zu Beginn der Messung getroffen werden müssen. Der MCMC-Algorithmus stellt alle notwendigen Parameter sowie deren Verteilungsfunktionen bereit, um eine Vorhersage der Zuverlässigkeit berechnen zu können. Mithilfe dieses Verfahrens können neue Abwägungen zwischen Messaufwand, Messgenauigkeit und Probenumfang getroffen werden. In der Arbeit wird der Designprozess mit den vereinheitlichten Transistormodellen sowie die Vorhersagemethodik anhand von einzelnen Transistoren und selbst entwickelter Schaltungen evaluiert.

1.1 Eigener Beitrag

In dieser Arbeit wurde ein bestehendes Transistormodell mit allen für die Zuverlässigkeit von Schaltungen relevanten Einflüssen erweitert. Dieses bettet sich direkt in die kommerzielle Entwicklungsumgebung für integrierte analoge Schaltungen ein. Aus den Möglichkeiten des Modells heraus wurde eine topologiebasierte Entwurfsmethode entwickelt, die auf zuvor simulierten Ergebnissen auf Transistorebene beruht. Somit kann vor der Dimensionierung der Bauteile bereits die Schaltung resistent gegenüber einzelnen Einflüssen ausgelegt werden. Darüber hinaus wurde eine Messmethode entwickelt, die die stochastischen Eigenschaften von Alterungsprozessen und Prozessvariationen

berücksichtigt und somit Entwurfsparameter wie die Schwellspannung deutlich genauer bestimmen kann oder dazu verwendet werden, um die Verteilungsfunktionen und Zeitabhängigkeiten von für die Zuverlässigkeit relevanten Parametern zu bestimmen.

1.2 Gliederung

Die vorliegende Arbeit gliedert sich entlang des eigenen Beitrags. Zuerst werden die theoretischen Grundlagen der verwendeten Modellierung, Entwurfsmethoden und Messmethoden dargestellt. Anschließend wird in Kapitel 3 der Entwurfsprozess von analogen integrierten Schaltungen unter Berücksichtigung der Zuverlässigkeit erläutert. Danach wird in Kapitel 4 der Entwurf von Systemen basierend auf zuvor analysierten Funktionsblöcken beschrieben. In Kapitel 5 wird die entwickelte Messmethode anhand von transienten Simulationen und Messungen von Schaltungen veranschaulicht. Die Arbeit schließt in Kapitel 6 mit einer Zusammenfassung und dem Ausblick ab.

2 Einführende Grundlagen

» Auf den Schultern von Riesen... «

In diesem Kapitel werden die theoretischen Grundlagen der vorliegenden Arbeit beschrieben. Dazu zählen der Entwurfsprozess integrierter, insbesondere analoger Schaltungen und die darin verwendeten Bauteilmodelle. Des Weiteren wird die Zuverlässigkeit von integrierten Schaltungen definiert sowie zentrale Einflüsse auf diese. Anschließend wird mittels mathematischer Methoden die Auswertung der Zuverlässigkeit dargestellt. Danach wird der Entwurfsprozess analoger Schaltungen näher beleuchtet und auf die Abbildung der relevanten Einflüsse auf die Zuverlässigkeit der Schaltungen in diesem Prozess eingegangen. Abschließend werden Markov-Ketten und darauf aufbauende Monte-Carlo-Verfahren erläutert, da diese zentraler Bestandteil der Modellparametrierung in dieser Arbeit sind.

2.1 Integrierte Transistoren

Bei integrierten Schaltungen wird zwischen analogen, digitalen und Mixed-Signal Schaltungen unterschieden. Dabei werden auf digitale Schaltungen zumeist mathematische,

logische oder arithmetische Funktionen abgebildet, wohingegen analoge Schaltungen vorrangig für die Ansteuerung von Sensoren, Erzeugung von Referenzströmen und -spannungen sowie der Verarbeitung von analogen Signalen aus der elektrischen und anderen physikalischen Domänen verwendet werden. Mixed-Signal Schaltungen bilden das Bindeglied zwischen diesen beiden Schaltungsarten. Sie dienen häufig zur Konvertierung von Signalen zwischen der analogen und digitalen Domäne. Zudem können größere Schaltungen häufig, sobald analoge und digitale Komponenten vorhanden sind, als Mixed-Signal Schaltungen beschrieben werden.

Die grundlegenden Bauelemente dieser Schaltungsarten unterscheiden sich, entgegen ihrer Einsatzzwecke, nicht. Die Grundelemente sind Widerstände, Kapazitäten, Transistoren und Spulen. Letztere sind in Herstellungsprozesse für integrierte Schaltungen schwierig zu implementieren und resultieren in einem enormen Platzbedarf. Daher wird bei dem Entwurf integrierter Schaltungen stets versucht, Spulen zu vermeiden. Falls nicht anders möglich, werden diese als diskrete Bauelemente oder in Form planarer MEMS- oder Oberflächen-Spulen realisiert. Die beiden verbleibenden passiven Bauelemente Widerstände und Kapazitäten können je nach Herstellungsverfahren und Technologie auf unterschiedliche Weisen hergestellt werden. Wie auch bei der Spule wird stets versucht, den Platzbedarf dieser Bauelemente möglichst gering zu halten. Bei einigen Herstellungsverfahren besteht zum Beispiel die Möglichkeit, unterhalb von Kapazitäten weitere Bauelemente zu platzieren. Die Art der Widerstände wird dabei durch den geforderten Widerstandswert festgelegt, so können diese Bauelemente als Metall, dotiertes Silizium oder Polysilizium hergestellt werden. Eine weitere Möglichkeit der Implementierung von Widerständen und Kapazitäten ist durch die Verwendung von Transistoren gegeben. Je nach Art des Betriebs und Anschlussspannungen zeigen Transistoren das Klemmenverhalten der beiden vorbezeichneten passiven Bauelemente. Daher wird der Transistor im Weiteren näher beschrieben.

In dieser Arbeit werden vorrangig planare Feldeffekttransistoren untersucht. Eine kurze Zusammenfassung über andere Arten von Transistoren und ihre jeweiligen Einsatzgebiete befindet sich am Ende dieses Abschnitts. Abbildung 2.1 zeigt den Querschnitt eines planaren Metalloxid-Halbleiter Feldeffekttransistors (MOSFET) mit Bulkanschluss (B). Daneben besitzt der Transistor noch die Anschlüsse Gate (G), Drain (D) und Source (S). Die Gebiete unter dem Source- und Drainkontakt sind n-leitend dotiert und bilden mit dem p-dotierten Substrat eine sperrende Diode, solange die Spannungen V_d und V_s größer als V_b sind. Wird nun eine, in diesem Fall positive Spannung V_{gs} an das Gate gegenüber Source angelegt, beginnt das Halbleitermaterial unterhalb des Gateoxids zu verarmen. Wird die Spannung weiter erhöht stellt sich ein Inversionskanal ein und die Bandkanten innerhalb des Gateaufbaus folgen dem Verlauf wie er in Abbildung 2.2 gezeigt ist. Die Fläche eines Transistors wird durch seine Kanallänge L, hier in x-Richtung, und seine Weite W in y-Richtung bestimmt. Des Weiteren sind Φ_m und Φ_s die Austrittsarbeiten der jeweiligen Materialien und Φ_{ms} deren Differenz. V_{ox} ist die Spannung, die über dem Gateoxid abfällt und zu einem elektrischen Feld E_{ox} innerhalb

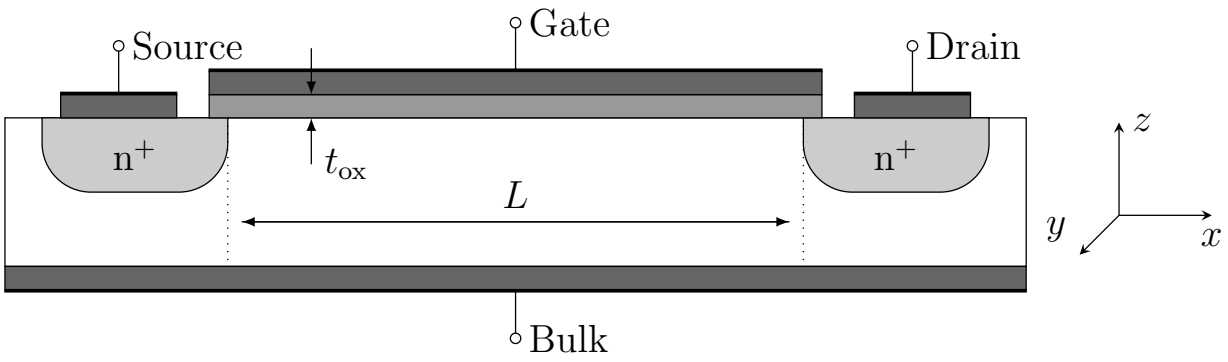


Abbildung 2.1: Querschnitt eines planaren N-MOSFET mit Bulkanschluss.

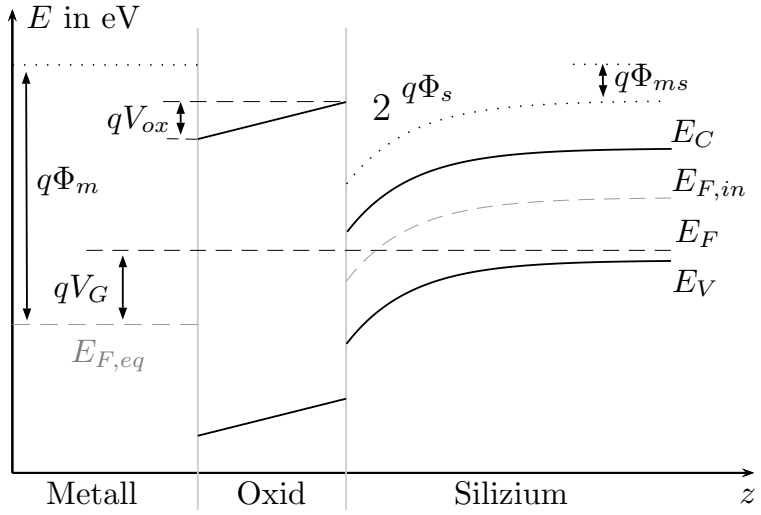


Abbildung 2.2: Verlauf der Bandkanten innerhalb des Gateaufbaus eines N-Kanal MOSFET bei eintretender Inversion.

des Oxids führt, das für spätere Betrachtungen in dieser Arbeit von zentraler Bedeutung ist. Die Feldstärke hängt dabei maßgeblich von der Dicke des Oxids t_{ox} und dessen Permittivität ab. $E_{F,eq}$ ist das Ferminiveau im Gleichgewichtszustand, welches nur für illustrative Zwecke mit in der Grafik gezeigt ist, da es, sobald eine Spannung angelegt, wird nicht mehr existiert. E_F ist das Ferminiveau und $E_{F,in}$ das intrinsische Ferminiveau. E_C und E_V sind die Energien des Leitungs- und Valenzbands. Wie dargestellt, wird das Leitungsband im p-dotieren Substrat durch die angelegte Gatespannung am Gateoxid in die Nähe des Ferminiveaus gebracht. Dadurch werden die Elektronen in diesen Bereich zu den Majoritätsträgern und das p-dotierte Material wurde n-leitend invertiert. Auf der z-Achse ist der Ort aufgetragen, ein Gateoxid ist typischerweise weniger als 10 nm stark. Die räumliche Ausdehnung des Inversionskanals

$$W_{dep} = \sqrt{\frac{2\varepsilon_0\varepsilon_{si}}{qN_A}\Psi_s} \tag{2.1}$$

ist auf das Gebiet direkt unter dem Gateoxid beschränkt. Neben der Beschreibung anhand von Spannungen kann die Funktion in Form von Ladungen geschehen. Die Ladungsverteilung in einem N-Kanal-MOSFET mit ausgeprägtem Inversionskanal ist in Abbildung 2.3 gezeigt. Dabei ist Q_g die Gateladung des gesamten Gatekontaktes. Q_{fc}

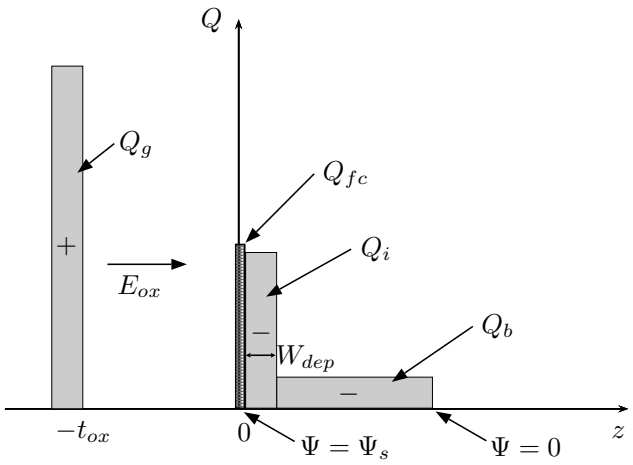


Abbildung 2.3: Ladungsverteilung in einem N-Kanal-MOSFET mit ausgeprägtem Inversionskanal.

beschreibt sämtliche Ladungsträger innerhalb des Gateoxids als konzentrierte Schicht an der Grenze zwischen Oxid und Silizium. Diese Ladungsträger sammeln sich während des Oxidationsprozesses bei der Halbleiterfertigung an[Fra10]. Die Darstellung der einzelnen Fehlstellen als eine Ladungsschicht beschreibt die Auswirkungen der Ladungsträger auf das nominelle Verhalten des Transistors. Im Zusammenhang mit verschiedenen Defekt- und Tunnelmechanismen kann die Lage einzelner Fehlstellen im Oxid von Bedeutung sein. Q_i ist die Inversionsladung und Q_b die Bulkladung. Das Oberflächenpotential Ψ_s stellt sich direkt an der Grenzschicht zwischen Oxid und Silizium ein. Aus der Ladungserhaltung folgt:

$$Q_g = Q_{fc} + Q_i + Q_b \ . \tag{2.2}$$

Der Ladungsanteil Q_{fc} wird im Laufe der vorliegenden Arbeit noch für weitere Modellierungszwecke verwendet.
Wird nun an den Source- oder üblicher an den Drainkontakt eine Spannung V_{ds} angelegt, beginnt ein Strom I_{ds} durch den Kanal zu fließen. Wird die Spannung V_{ds} weiter erhöht, verschiebt sich der Abschnürpunkt des Kanals in Richtung Source, da die bewegliche negative Kanalladung durch das vom Draingebiet ausgehende elektrische Feld verdrängt wird. Dies ist in Abbildung 2.4 in Form eines Querschnitts dargestellt. Hinter dem Abschnürpunkt tragen nur noch sehr wenige ortsfeste Dotierzentren mit einer geringen Anzahl von möglichen Zuständen für Ladungsträger zum Stromtransport bei. Dadurch erhöht sich die Geschwindigkeit und somit die kinetische Energie der Ladungsträger in dem Bereich enorm. Die hohen kinetischen Energien sind für im

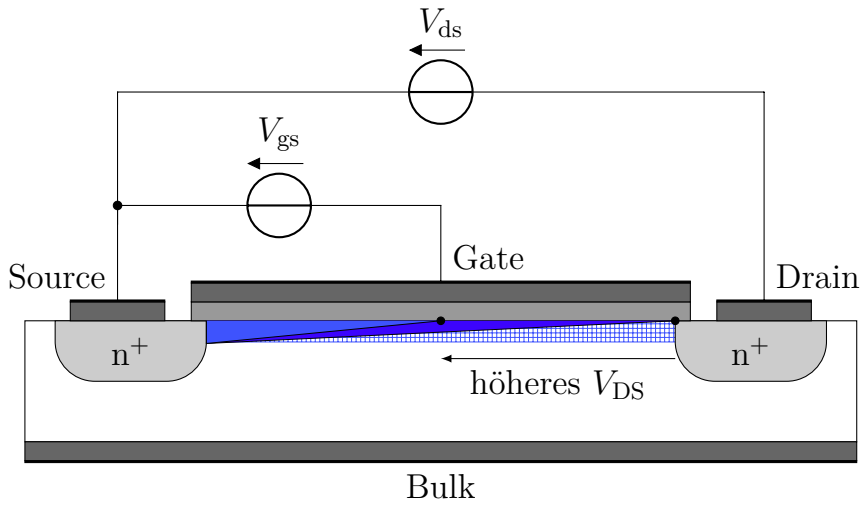


Abbildung 2.4: Verschiebung des Abschnürpunkts durch Erhöhung der Drain-Source-Spannung in einem N-Kanal MOSFET.

Weiteren betrachtete Einflüsse auf den Transistor sehr wichtig. Zudem verringert sich der äquivalente Kanalwiderstand und ceteris paribus erhöht sich der Drain-Source-Strom I_{ds}. Dieser Effekt wird als Kanallängenmodulation bezeichnet. Die Gesamtfunktion des Transistors lässt sich durch zwei Charakteristiken ausdrücken. Zum einen mit Hilfe der Übertragungscharakteristik $I_{ds}(V_{gs})$ und zum anderen durch die Ausgangscharakteristik $I_{ds}(V_{ds})$, beide sind in Abbildung 2.5 für einen N-Kanal MOSFET mit Kanallängenmodulation gezeigt. Die Übertragungscharakteristik zeigt einen exponentiellen Anstieg

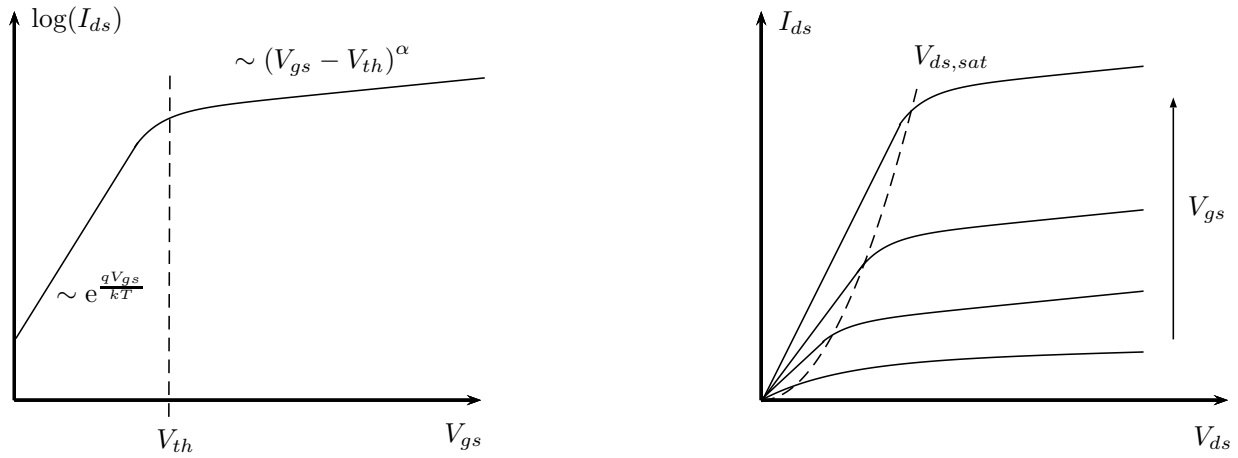


Abbildung 2.5: Übertragungs- (links) und Ausgangscharakteristik (rechts) eines N-Kanal MOSFET mit Kanallängenmodulation.

bis zur Schwellspannung V_{th}. Ab diesem Punkt ist der Inversionskanal ausgeprägt und beginnt zu leiten. Danach ist der Strom mit dem Exponenten $1 \leq \alpha \leq 2$ von der Spannungsdifferenz $V_{gs} - V_{th}$ abhängig. Bei dieser Charakteristik werden die Zustände des Transistors in Abhängigkeit der Stärke der Inversion beschrieben. Darauf wird im Weiteren noch detaillierter eingegangen. Zentraler Parameter der Beschreibung ist die

Schwellspannung. Diese wird für die Auslegung von Transistoren und Schaltungen sowie für die Beschreibung einzelner Effekte in Transistoren verwendet. Dadurch, dass V_{th} von wesentlicher Bedeutung ist, ist dem Parameter in dieser Arbeit ein eigener Abschnitt gewidmet. Die Ausgangscharakteristik hängt zwar maßgeblich von V_{ds} ab, jedoch muss zur Beschreibung des Zustandes ebenfalls die Gate-Source-Spannung verwendet werden. Es werden insgesamt drei Betriebsbereiche unterschieden.

1. **Subschwellbereich:** In diesem Bereich ist $V_{gs} < V_{th}$, V_{ds} kann im Rahmen der Technologie beliebige Werte annehmen. Durch den exponentiellen Zusammenhang zwischen Gate-Spannung und Drain-Strom weist der Transistor in diesem Bereich sehr hohe Verstärkungen auf.

2. **Linearer oder Triodenbereich:** In diesem Bereich ist $V_{gs} \geq V_{th}$ und $V_{ds} < V_{ds,sat}$, wobei $V_{ds,sat}$ die Drain-Source-Spannung beschreibt, bei der der Abschnürpunkt gerade am Drainkontakt ist. Der Transistor kann in diesem Zustand als einstellbarer Widerstand betrachtet werden, da der Strom-Spannungs-Zusammenhang nahezu linear ist.

3. **Sättigung:** Es gilt $V_{gs} > V_{th}$ und $V_{ds} \geq V_{ds,sat}$. Wie in der Abbildung dargestellt, weist I_{ds} eine gute Steuerbarkeit durch V_{gs} auf. Dadurch kann der Transistor als spannungsgesteuerte Stromquelle aufgefasst werden.

Durch diese Grundfunktionen kann der Transistor als spannungsgesteuerte Stromquelle oder einstellbarer Widerstand verwendet werden. Zudem bildet der Gateaufbau einen Plattenkondensator, der in Schaltungen als Kapazität verwendet werden kann. Art und Weise des Klemmenverhaltens des Transistors hängt daher sehr stark von den Klemmenspannungen sowie als zentrale Parameter von der Schwellspannung V_{th} und der Sättigungsspannung $V_{ds,sat}$ ab.
Somit ergeben sich für die spätere Verwendung die beiden Symbole für P- und N-Kanal MOSFET wie in Abbildung 2.6. Mit dieser Konvention für die Spannungspfeile gilt für den P-Kanal MOSFET dieselbe Herleitung wie für die N-Kanal Variante. Zentraler Unterschied ist hingegen die Ladungsträgerbeweglichkeit μ, die Beweglichkeit der Elektronen ist deutlich höher als die der Löcher. Daher erreichen Ladungsträger in N-Kanal MOSFET höhere Geschwindigkeiten und höhere kinetische Energien.

Sekundäreffekte Neben der eigentlichen Funktion des Transistors treten noch weitere, häufig unerwünschte Effekte in Erscheinung [SSS18, Sah14, KSA+09]. Diese beeinträchtigen entweder das Schaltverhalten oder reduzieren die Leistungsfähigkeit in einzelnen Betriebsbereichen. Hier werden die für diese Arbeit relevanten Effekte kurz eingeführt.

1. Geschwindigkeitssättigung: Durch die Ladungsträgerbeweglichkeit und das laterale elektrische Feld im Kanal stellt sich eine maximale Geschwindigkeit v_{sat} ein, wie Abbildung 2.7 zeigt. Diese wird im Sättigungsbereich erreicht und führt zu einem maximalen Strom, der durch den Kanal fließen kann.

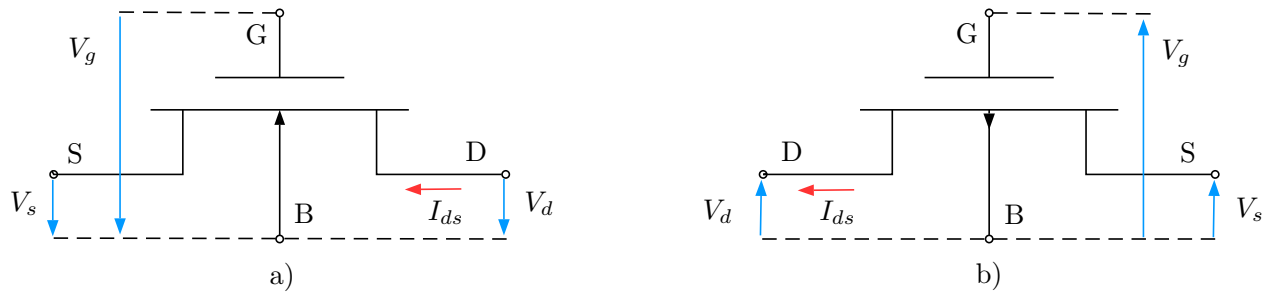


Abbildung 2.6: Schaltzeichen eines a) N-Kanal und b) P-Kanal MOSFET mit Klemmenspannungen und Strompfeil.

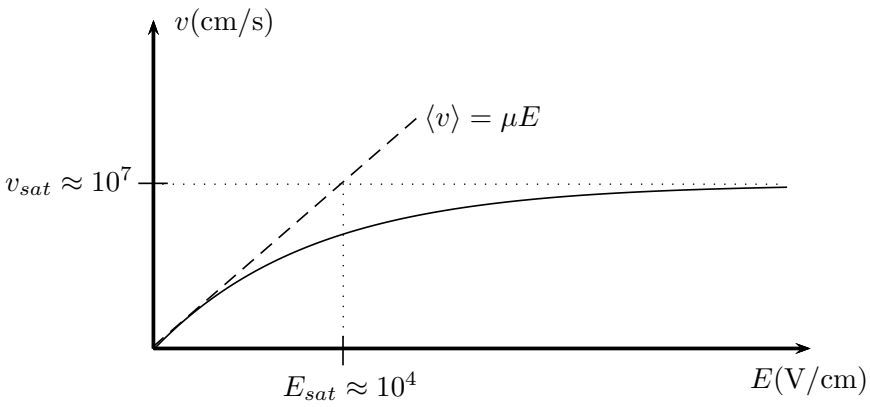


Abbildung 2.7: Geschwindigkeitssättigung von Ladungsträgern für Silizium in Abhängigkeit des lateralen elektrischen Feldes im Kanal.

2. Kurzkanaleffekte: Mit abnehmender Kanallänge L steigen die Leckströme durch die Struktur. Zum einen steigt durch die ebenfalls reduzierte Oxiddicke der Leckstrom des Gates. Zum anderen führen Gate Induced Drain Leakage (GIDL) und Drain Induced Barrier Lowering (DIBL) zu ungewollten oder zu höheren Strömen in der MOS-Struktur [SSS18]. Nähert sich die Kanallänge dem jeweiligen technologischen Minimum, treten weitere fertigungsbedingte Nichtidealitäten auf.

   Neben diesen Effekten führt die Skalierung der Transistoren zu weit mehr und zum Teil komplexen Effekten die ihrerseits wechselwirken. Einige von diesen Effekten wie in [SSS18] aufgeführt, treffen jedoch erst für Transistoren mit Kanallängen von weniger als 65nm zu oder werden durch Arbeitspunkte ausgelöst, die im Rahmen der Arbeit nicht betrachtet werden. In späteren Kapiteln werden zur Evaluierung einiger Messmethoden TCAD-Simulationen von 14nm-FinFET und 28nm-FDSOI-Transistoren betrachtet. Dort sind weitere Effekte in den Simulationsdaten zu erkennen. Diese sind jedoch für die weitere Arbeit irrelevant.

Weitere Transistorarten Da die Herstellungskosten einer integrierten Schaltung proportional zu dem Flächenbedarf sind, ist ein stetiges Drängen auf immer kleinere Technologieknoten zu beobachten. Ein Technologieknoten bezeichnet dabei die kleinste

herstellbare geometrische Anordnung, meist die Kanallänge eines Transistors, und folgte bis vor einigen Jahren Moore's-Law [Sch97]. So werden neben den Bipolar-Transistoren und planaren Feldeffekttransistoren auch andere Arten von Transistoren für verschiedene Einsatzzwecke interessant. In den für diese Arbeit relevanten Einsatzgebieten, Automotive und Luft-/Raumfahrt, finden sich diese Technologien noch nicht wieder. Daher werden diese nur knapp beschrieben, um die betrachteten Transistoren in den allgemeinen Entwicklungskontext zu setzen.

1. Bipolar: Bipolare Transistoren finden sich heute noch in vielen Anwendungen. Für integrierte Schaltungen spielen sie, aufgrund der Notwendigkeit eines Basisstroms und der daraus folgenden Verlustleistung eine untergeordnete Rolle. In einigen CMOS-Herstellungsprozessen sind Bipolar-Transistoren verfügbar (BiCMOS), um zum Beispiel eine Temperaturkompensation zu erreichen [TS02].

2. FinFET: Die Idee der Fin-Feldeffekttransistoren, wobei engl. *fin* Flosse meint, ist das Gate in 3-Dimensionen auszuführen, um so eine höhere steuerbare Gate-Fläche auf weniger Grundfläche zu realisieren. Somit können kleine Transistoren genau so gut und schnell angesteuert werden, wie solche in planarer Ausführung [PMP14b].

3. FDSOI: Ein Nachteil der planaren Transistoren ist es, dass nicht nur die Inversionsladung Q_i, sondern bei jedem Schaltvorgang ebenfalls die Bulk-Ladung Q_b umgeladen werden muss. Dies verlangsamt den Schaltvorgang und führt zu weiteren Verlusten. Fully-Depleted-Silicon-On-Insulator umgeht diese Probleme, indem sich unter dem Gateoxid eine sehr dünne Schicht Substrat befindet, an das sich sofort ein Isolator anschließt. So wird Q_b verringert und der Transistor benötigt weniger Leistung und schaltet schneller [MFC13].

4. Ballistische FET: Bei diesen Transistoren ist die Länge L so gering, dass es unterhalb der mittleren freien Weglänge der Ladungsträger ist. Dadurch ist der Kanal quasi ideal leitend [LG06].

5. Tunnel FET: Das Substrat ist in diesem Fall undotiert und L ist deutlich geringer als bei normalen MOSFET. Wird eine ausreichend hohe Spannung an das Gate angelegt, gleichen sich Valenzband und Leitungsband in Source und Drain an und die Ladungsträger tunneln durch den Kanal. Vorteil dieses Ansatzes ist es, dass das Schalten dieser Transistoren schneller möglich ist als mit jeder anderen Art von Transistoren [BI07].

6. Double-Gate / Gate-Allaround: Wie der Name bereits nahelegt, befindet sich rund um den Kanal das Gateoxid. Dies ist eine Weiterentwicklung der FinFET und zielt auf dasselbe Ergebnis wie ebendiese [PMP14b].

2.2 Transistormodelle

Damit mit diesen Bauelementen ein Entwurf von Schaltungen und Systemen durchgeführt werden kann, benötigen Transistoren Modelle, die rechnergestützt simuliert werden können. Dadurch kann vor der Herstellung der Schaltung das gewünschte Verhalten validiert und verifiziert werden. Im Laufe der Zeit haben sich drei verschiedene Arten der Transistormodellierung durchgesetzt. Die Entwicklung der einzelnen Modelle ist in Tabelle 2.1 anhand der jeweiligen Technologieknoten gezeigt.

Tabelle 2.1: Entwicklung der Transistormodelle mit kleinerwerdenden Technologieknoten. Übernommen und erweitert aus [DH11].

Technologie-knoten	MOSFET Modell-generation	MOSFET Modell	CMC Standard	Modell basierend auf
$>> 1\,\mu$m	1. Generation	Level 1		V_{th}
		Level 2		V_{th}
		Level 3		V_{th}
$\geq 1\,\mu$m	2. Generation	BSIM 1		V_{th}
		BSIM 2		V_{th}
350 nm		HSPICE Level 28		V_{th}
250 nm	3. Generation	BSIM 3.x	X(1996)	V_{th}
180 nm		BSIM 4.x	X(2000)	V_{th}
		MOS 9, MOS 11		V_{th}
130 nm		EKV		Q_i (iterativ)
90 nm	4. Generation	BSIM 5		Q_i (iterativ)
65 nm		PSP	X(2006)	Ψ_s (explizit)
45 nm		HiSIM	X(2007)	Ψ_s (iterativ)
≤ 45 nm	5. Generation	BSIM-BULK	X(2013)	Q_i (explizit)

So waren bis einschließlich 180nm alle Transistormodelle auf der Schwellspannung basierend. Für kleinere Technologieknoten sind die Unzulänglichkeiten der V_{th}-Modelle [Liu01] so stark, dass ausschließlich ladungsbasierte [EV06, CVC+14a, CH99] als auch Oberflächenpotential-basierte [GLW+05] Modelle für moderne Transistormodelle in Frage kommen [TRB+03]. Die drei unterschiedlichen Arten von Transistormodellen werden im Weiteren beschrieben.

2.2.1 Strom-Spannungs-Modellierung

Diese phänomenologische Strom-Spannungs-Modellierung der Transistorfunktion ist die älteste [PMP14b]. Die ersten und einfachsten Modelle heißen Level1 (MOS1, SPICE Level1) oder quadratisches Modell [SSKJ87]. Aus ihnen gingen die BSIM3 [Yuh02] und

BSIM4 [XDH+03] Modelle hervor, die die Grundgleichungen um zusätzliche Effekte erweitern. Die Grundgleichungen der Modelle unterscheiden sich dabei nicht. Wie bereits im vorherigen Abschnitt beschrieben, unterscheiden diese Modelle zwischen den drei Betriebsbereichen des Transistors in Abhängigkeit der Ausgangscharakteristik und modellieren die einzelnen Bereiche wie folgt:

1. Subschwellbereich:

$$I_{ds} = I_{ds,0} \cdot \mathrm{e}^{\frac{V_{gs}-V_{th}}{U_t \cdot n}} \tag{2.3}$$

   Dabei ist $I_{ds,0}$ der Strom der fließt wenn $V_{gs} = V_{th}$ ist, n die Spannungssteigung und U_T die Temperaturspannung.

2. Linearer Bereich:

$$I_{ds} = \mu C_{ox} \frac{W}{L} \left((V_{gs} - V_{th}) \cdot V_{ds} - \frac{1}{2} V_{ds}^2 \right) \tag{2.4}$$

   Dabei ist C_{ox} die flächenbezogene Kapazität des Gateoxids.

3. Sättigung: Für diesen Bereich gilt als weitere Einschränkung $V_{gs} \geq V_{th}$ und für den Drainstrom

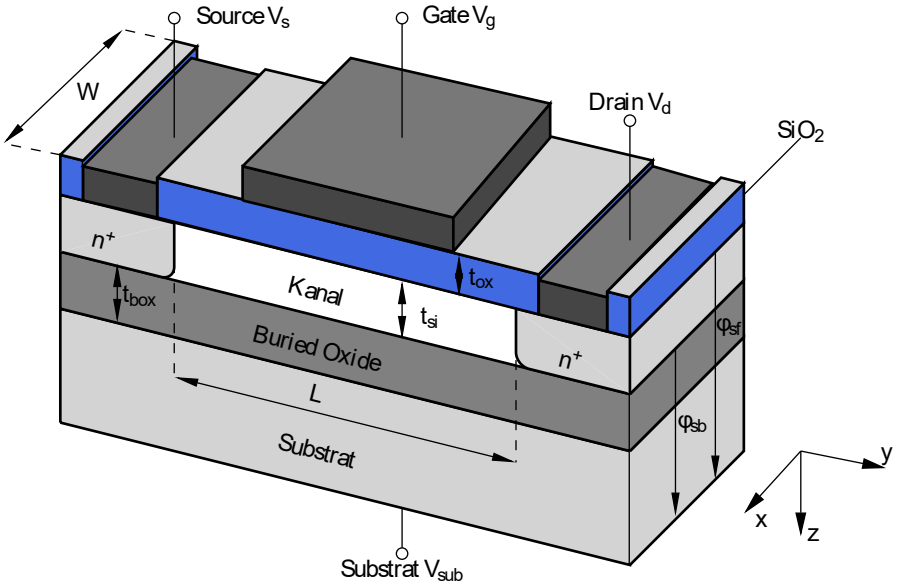
$$I_{ds} = \frac{1}{2} \mu C_{ox} \frac{W}{L} \cdot (V_{gs} - V_{th})^2 \ . \tag{2.5}$$

Die Exponenten der einzelnen Arbeitsbereiche, insbesondere in Sättigung, können sich in Abhängigkeit von Transistortyp und Herstellungsprozess unterscheiden [SN90]. Der Vorteil dieser Modellierung ist, dass die Gleichungen sogar "per Hand"berechenbar sind. Dies erleichtert die ersten Entwurfsschritte sehr. Die dargestellten Gleichungen sind ein Beispiel für die Beschreibung in den tatsächlich simulierbaren Modellen und in späteren Modellversionen wurden diese Gleichungen leicht modifiziert. Neben der groben Approximation des Drainstroms in den einzelnen Bereichen, welche durch die Weiterentwicklung der Modelle verbessert wurde, ist der entscheidende Nachteil dieser Art der Modellierung der Übergang zwischen den einzelnen Betriebsbereichen. In dem Bereich $V_{gs} \approx V_{th}$ weichen beide modellierten Bereiche, subschwell und linear, sehr weit vom eigentlichen Wert des Drainstroms ab. Dahingegen ist gerade dieser Bereich ein guter Kompromiss zwischen Leistungsfähigkeit und Effizienz [EV06]. Durch die Schwäche des Modellierungsansatzes wurden weitere Transistormodelle entwickelt, die den Transistor in allen Arbeitsbereichen konsistent abbilden können.

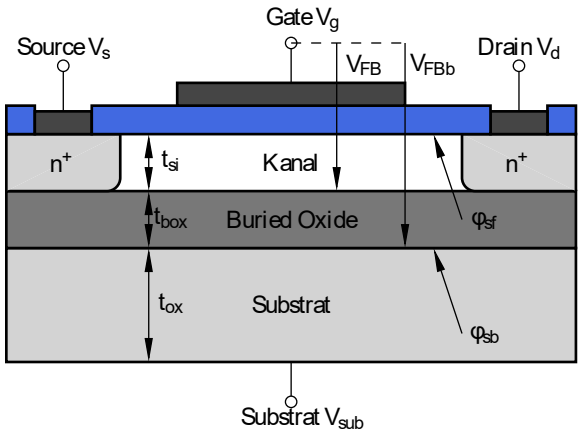
2.2.2 Oberflächenpotential-basierte Modellierung

Zur Beschreibung des Transistorverhaltens wird in diesem Transistormodell das Oberflächenpotential genutzt, welches bestimmt werden soll, um Abhängigkeiten von Pro-

zessparametern aufzuzeigen [GWC+04, GLW+06]. Zudem soll gezeigt werden, dass unterschiedliche Transistormodelle für verschiedene Strukturen auf ähnlichen Modellannahmen beruhen. Dazu wird in diesem Abschnitt die Oberflächenpotential-basierte Modellierung anhand eines FDSOI-Transistors gezeigt. Dafür wird von der dreidimensionalen Poisson-Gleichung im Gebiet der Inversion ausgegangen. Alle Angaben beziehen sich auf die gezeigte FDSOI-Struktur aus Abbildung 2.8. Für SOI Transistoren existieren sowohl PSP- als auch BSIM-Modelle [WYG10, GNS06]. Zudem werden von einzelnen FDSOI-Herstellern eigene Modelle entwickelt [PRS+15a, PRS+15b].



(a)



(b)

Abbildung 2.8: N-Kanal FDSOI Transistor. a) Dreidimensionale Ansicht mit eingezeichneten Anschlüssen und Transistorgrößen b) Querschnitt mit eingezeichneten Potentialen und Spannungen.

Ausgehend von der Poisson-Gleichung im Gebiet des Inversionskanals

$$\frac{\partial^2 \Psi_s}{\partial x^2} + \frac{\partial^2 \Psi_s}{\partial y^2} + \frac{\partial^2 \Psi_s}{\partial z^2} = -\frac{\rho}{\varepsilon_{si}} \ , \tag{2.6}$$

können die x- und $y-$Abhängigkeit vernachlässigt werden, wenn der Kanal ausreichend weit und lang ist, so dass Kurz- und Schmalkanaleffekte ausgeschlossen werden können. Somit folgt die eindimensionale Poisson Gleichung

$$\frac{\partial^2 \Psi_s}{\partial z^2} = \frac{\mathrm{q} \cdot n_\mathrm{i}}{\varepsilon_\mathrm{si}} \cdot \exp\left(\frac{\Psi_s - V_\mathrm{c}}{U_\mathrm{T}}\right). \tag{2.7}$$

Dabei ist $\varepsilon_\mathrm{si} = \varepsilon_\mathrm{r} \cdot \varepsilon_0$ die Permittivität des Siliziums, Ψ_s das Potential im Kanal unter dem Gate, q die Elektronenladung, n_i die intrinsische Ladungsträgerdichte, V_c die Spannung in y-Richtung im Kanal und U_T die thermische Spannung, für die gilt:

$$U_\mathrm{T} = \frac{k_\mathrm{B} \cdot T}{\mathrm{q}} \tag{2.8}$$

wobei k_B die Boltzmann Konstante und T die Temperatur in Kelvin ist.

Um das elektrische Feld innerhalb des Kanals zu bestimmen, wird das Front- und Backsurface Potential jeweils eingesetzt und voneinander subtrahiert.

$$E_\mathrm{sf}^2 - E_\mathrm{sb}^2 = \frac{2 \cdot q \cdot n_\mathrm{i} \cdot V_\mathrm{T}}{\varepsilon_\mathrm{si}} \cdot \underbrace{\left[\exp\left(\frac{\Psi_\mathrm{sf} - V_\mathrm{c}}{U_\mathrm{T}}\right) - \exp\left(\frac{\Psi_\mathrm{sb} - V_\mathrm{c}}{U_\mathrm{T}}\right)\right]}_{x} \tag{2.9}$$

Aus dem Maschenumlauf angewendet auf Abbildung 2.8 ergeben sich die Randbedingungen für das elektrische Feld.

$$E_\mathrm{sf} = -\frac{C'_\mathrm{ox}}{\varepsilon_\mathrm{si}} \cdot (V_\mathrm{g} - V_\mathrm{FB} - \Psi_\mathrm{sf}) \tag{2.10}$$

$$E_\mathrm{sb} = -\frac{C'_\mathrm{box}}{\varepsilon_\mathrm{si}} \cdot (V_\mathrm{sub} - V_\mathrm{FBb} - \Psi_\mathrm{sb}) \tag{2.11}$$

Dabei sind V_FB bzw. V_FBb die Flachbandspannung am Gate bzw. am Si-Substrat. Die Kapazitäten $C'_\mathrm{ox} = {}^{\varepsilon_\mathrm{ox}}/_{t_\mathrm{ox}}$ und $C'_\mathrm{box} = {}^{\varepsilon_\mathrm{box}}/_{t_\mathrm{box}}$ sind flächenbezogene Kapazitäten, wodurch der erste Term auf der rechten Seite in Gleichung (2.10) und Gleichung (2.11)

einheitenlos ist und beide Gleichungen die Einheit Volt/Meter besitzen. Das Einsetzen der Randbedingungen in Gleichung (2.9) liefert

$$\frac{C_{\mathrm{ox}}'^2}{\varepsilon_{\mathrm{si}}^2} \cdot (V_{\mathrm{g}} - V_{\mathrm{FB}} - \Psi_{\mathrm{sf}})^2 - \frac{C_{\mathrm{ox}}'^2}{\varepsilon_{\mathrm{si}}^2} \cdot (V_{\mathrm{sub}} - V_{\mathrm{FBb}} - \Psi_{\mathrm{sb}})^2 = \frac{2 \cdot q \cdot n_{\mathrm{i}} \cdot V_{\mathrm{T}}}{\varepsilon_{\mathrm{si}}} \cdot x \tag{2.12}$$

$$(V_{\mathrm{g}} - V_{\mathrm{FB}} - \Psi_{\mathrm{sf}})^2 - (V_{\mathrm{sub}} - V_{\mathrm{FBb}} - \Psi_{\mathrm{sb}})^2 = \frac{2 \cdot q \cdot n_{\mathrm{i}} \cdot U_{\mathrm{T}} \cdot \varepsilon_{\mathrm{si}}}{C_{\mathrm{ox}}'^2} \cdot x \tag{2.13}$$

$$(V_{\mathrm{g}} - V_{\mathrm{FB}} - \Psi_{\mathrm{sf}})^2 - (V_{\mathrm{sub}} - V_{\mathrm{FBb}} - \Psi_{\mathrm{sb}})^2 = \gamma^2 \cdot U_{\mathrm{T}} \cdot x, \tag{2.14}$$

wobei γ den Substratfaktor darstellt:

$$\gamma = \frac{\sqrt{2 \cdot q \cdot n_{\mathrm{i}} \cdot \varepsilon_{\mathrm{si}}}}{C_{\mathrm{ox}}'}. \tag{2.15}$$

Bei dem verwendeten Modell wird angenommen, dass sich das Backinterface, also die Grenzfläche zwischen Buried Oxide und Substrat, immer in Verarmung befindet. Somit gilt $V_{\mathrm{sb}} > V_{\mathrm{FB}}$, wobei V_{sb} die Spannung zwischen Substrat und Buried Oxide und V_{FB} die Flachbandspannung ist. Dadurch ist es nicht möglich, den FDSOI-Transistor als Double-Gate-Transistor zu betreiben. Die beiden unbekannten Potentiale werden über iterative numerische Lösungswege bestimmt. Ein Auszug des Ablaufs zur Berechnung des Potentials befindet sich im Anhang A.1. Ausgehend von Drift- und Diffusionsanteil des Drain-Source-Stroms

$$I_{ds} = \mu W \left(\underbrace{-Q_i \frac{\mathrm{d}\Psi_s}{\mathrm{d}x}}_{\text{Drift}} + \underbrace{U_T \frac{\mathrm{d}Q_i}{\mathrm{d}x}}_{\text{Diffusion}} \right) \tag{2.16}$$

kann dieser mit

$$Q_i = Q_{im} - \alpha(\Psi_s - \Psi_m) \tag{2.17}$$

berechnet werden. Ψ_m ist dabei Durchschnittspotential zwischen Source und Drain. Q_{im} ist die Inversionsladung beim Potential $\Delta\Psi_m$. Anschließend kann die Temperaturabhängigkeit mit der effektiven Inversionsladung q_{im}^* modelliert werden. Insgesamt ergibt sich der Drain-Source-Strom zu

$$I_{ds} = \mu \cdot \frac{W}{L} \cdot q_{im}^* \Psi_m \ . \tag{2.18}$$

als Gleichung für die Berechnung des Stroms. Der Drain-Source Strom I_{ds} des Transistormodells hängt im Wesentlichen von der Länge und den Oberflächenpotentialen ab. Um zusätzliche Effekte zu berücksichtigenn kann die Gleichung erweitert werden:

$$I_{\mathrm{ds}} = \frac{\beta}{\Delta C'_{\mathrm{ox,qm}}(q)} \cdot F_{\Delta \mathrm{L}}(L) \cdot A(\Psi_m) \cdot \Psi_m(x_{\mathrm{ds}}) \ . \tag{2.19}$$

Die Funktionen F und A sind weitere zu modellierende funktionale Abhängigkeiten. Somit können Kurzkanaleffekte, Veränderung der Gatekapazität, Skalierung des Stroms bzw. verschiedene Abhängigkeiten und nichtlineare Abhängigkeiten der Potentiale modelliert werden. Da der Strom auch von der Differenz x der Oberflächenpotentiale x_{d} und x_{s} abhängt (vgl. Gleichung (2.9)), gelten für diesen Teil der Gleichung die selben Abhängigkeiten wie für die jeweiligen Einzelpotentiale (siehe dazu Anhang Abschnitt A.2). Der Berechnungsablauf zur Bestimmung des Potentials ist ebenfalls im Anhang zu finden. Besonders aufwendig wird die Berechnung durch die Schleifen in den Quelltext-Auszügen, die in jedem Simulationsschritt durchlaufen werden muss. Für jede der Abhängigkeiten in Gleichung (2.19) existieren in den jeweiligen Modellen verschiedene Ansätze zur Berechnung. Die grundlegende Berechnung der Potentiale an Drain und Source ist jedoch bei den Modellen gleich, wobei sich der numerische Lösungsweg unterscheiden kann. Wie an der Beschreibung dieses Modellierungsansatzes zu erkennen ist, sind die verwendeten Gleichungen deutlich komplizierter als die bei einer reinen Strom-Spannungs-Modellierung. Diese Gleichung des Oberflächenpotential-basierten Modells eignen sich nicht mehr für eine Berechnung "per Hand", können jedoch den Transistor über alle Arbeitsbereiche konsistent beschreiben.

2.2.3 Ladungsbasierte Modellierung

Eine weitere Möglichkeit der Modellierung ist die Verwendung der Inversionsladung als zentralen Parameter. Diese Herangehensweise findet sich im EKV [EV06] und im BSIM-Bulk Modell [CVC+14a] wieder. Ausgangspunkt ist hierbei ebenfalls die Poisson-Gleichung im Raumgebiet der Inversion welche in Abbildung 2.3 gezeigt ist. Diese ist für drei Raumrichtungen:

$$\frac{\partial^2 \Psi_s}{\partial x^2} + \frac{\partial^2 \Psi_s}{\partial y^2} + \frac{\partial^2 \Psi_s}{\partial z^2} = -\frac{\rho}{\varepsilon_{si}} \ , \tag{2.20}$$

dabei ist $\rho = q(p_p - n_p - N_b)$ wobei p_p die Konzentration von Löchern ist, n_p die von Elektronen ist und N_b die Dotierkonzentration. Durch die Tatsache, dass sich das Potential in x- und y-Richtung deutlich schwächer ändert als in z-Richtung, können diese

beiden Komponenten vernachlässigt werden. Das Umschreiben der eindimensionalen Poisson-Gleichung in Abhängigkeit von Φ_F führt zu:

$$\frac{\mathrm{d}^2\Psi_s}{\mathrm{d}z^2} = \frac{qN_b}{\varepsilon_{si}}\left(\underbrace{\exp\left(\frac{\Psi_s - 2\Phi_F - V_c}{U_T}\right)}_{\text{Elektronen}} - \underbrace{\exp\left(-\frac{\Psi_s}{U_T}\right)}_{\text{Löcher}} + \underbrace{1}_{\text{Ortsfeste Ladung}}\right) \tag{2.21}$$

$$= \frac{qN_b}{\varepsilon_{si}} G(\Psi_s, 2\Phi_F + V_c)\ . \tag{2.22}$$

Dabei ist $V_c = \Phi_{fn} - \Phi_F$ die Spannung im Kanal in y-Richtung. Φ_{fn} ist das quasi-Ferminiveau, das durch eine Spannung V_s oder V_d entsteht. Mit der Beziehung

$$E_z = -\frac{\mathrm{d}\Psi_s}{\mathrm{d}z} \tag{2.23}$$

können beide Seiten der Poisson-Gleichung integriert werden. Es entsteht folgender Ausdruck für das elektrische Feld in z-Richtung E_z:

$$E_z = \mathrm{sign}(\Psi_s)\frac{U_T}{L_D} F(\Psi_s, 2\Phi_F + V_c) \tag{2.24}$$

Dabei ist L_D die extrinsische Debye-Länge

$$L_D = \sqrt{\frac{\varepsilon_{si} U_T}{2qN_b}} \tag{2.25}$$

und

$$F(\Psi_s, 2\Phi_F + V_c) = \int_0^{\Psi} G(\Psi_s, 2\Phi_F + V_c)\,\mathrm{d}\Psi \tag{2.26}$$

$$= \sqrt{\left(\exp\frac{\Psi_s}{U_T} - 1\right)\exp\frac{-(2\Phi_f + V_c)}{U_T} + \left(\exp\frac{-\Psi_s}{U_T} - 1\right) + \frac{\Psi_s}{U_T}}\ . \tag{2.27}$$

Die flächenbezogene Ladung unter dem Oxid wird nun nach dem Satz vom Hüllenfluss wie folgt berechnet:

$$Q_{si} = \varepsilon_{si} \oint \vec{E} \cdot \mathrm{d}\vec{A}\ . \tag{2.28}$$

Dadurch, dass die elektrischen Felder in der x-y-Ebene vernachlässigt werden können, folgt:

$$Q_{si} = -\varepsilon_{si} E_z = -\frac{\varepsilon_{si} U_T}{L_D} F(\Psi_s, 2\Phi_F + V_c)\ . \tag{2.29}$$

Nachdem ein Ausdruck für die Ladung Q_{si} entwickelt wurde, muss dieser in Abhängigkeit von der Gatespannung ermittelt und anschließend der Drain-Source-Strom berechnet werden. Durch erneutes Anwenden des Satzes vom Hüllenfluss ergibt sich folgende Relation:

$$Q_{si} + Q_{fc} = -\varepsilon_{ox} E_{ox} = -C_{ox}(V_g - \Phi_{ms} - \Psi_s) \tag{2.30}$$

dabei ist $C_{ox} = {}^{\varepsilon_{ox}}/_{t_{ox}}$ die Gate-Oxid-Kapazität pro Einheitsfläche. Dieser Parameter, wird im Vergleich zum vorherigen Abschnitt, anders benannt um so den Formalismus der Quelle einzuhalten. Wird in Gleichung (2.30) Q_{si} aus (2.29) eingesetzt und nach V_g umgestellt ergibt sich

$$\frac{V_g - V_{FB}}{U_T} = \frac{\Psi_s}{U_T} + \frac{\varepsilon_{si}}{C_{ox} L_D} F(\Psi_s, 2\Phi_F + V_c) \ . \tag{2.31}$$

Für die Flachbandspannung V_{FB} gilt

$$V_{FB} = \Phi_{ms} - \frac{Q_{fc}}{C_{ox}} \ . \tag{2.32}$$

Zur Berechnung der Inversionsladung unter dem Oxid muss ein Integral gelöst werden, das nicht explizit berechnet werden kann (Gleichung (2.33)). Daher wird die Näherung eingebracht, dass sich alle Ladungsträger direkt am Ort des Oberflächenpotentials Ψ_s befinden. Dadurch gilt:

$$Q_i = -q \int_0^{\Psi_s} \frac{n_p}{E_z} \, \mathrm{d}\Psi \approx Q_{si} - Q_b = -C_{ox}(V_g - V_{TB}) \tag{2.33}$$

V_{TB} ist die Schwellfunktion, für die

$$V_{TB} = V_{FB} + \Psi_s + \Gamma_b \sqrt{\Psi_s} \tag{2.34}$$

gilt. Der Substratfaktor ist $\Gamma_b = {}^{\sqrt{2qN_b\varepsilon_{si}}}/_{C_{ox}}$. Die Steigung von V_{TB} ist

$$n = \frac{\mathrm{d}V_{TB}}{\mathrm{d}\Psi_s} = 1 + \frac{\Gamma_b}{2\sqrt{\Psi_s}} \tag{2.35}$$

Für ein gegebenes V_g gilt

$$\frac{\mathrm{d}{}^{Q_i}/_{C_{ox}}}{\mathrm{d}\Psi_s} = n \ . \tag{2.36}$$

Eine weitere Beschreibung kann über das Abschnürpotential Ψ_P erfolgen. Wenn $\Psi_s = \Psi_P$ dann folgt $Q_i = 0$ und somit ist

$$V_g = V_{FB} + \Psi_P + \Gamma_b \sqrt{\Psi_P} \ . \tag{2.37}$$

Das Abschnürpotential ist

$$\Psi_p = V_g - V_{FB} - \Gamma_b^2 \left(\sqrt{\frac{V_g - V_{FB}}{\Gamma_b^2} + \frac{1}{4}} - \frac{1}{2} \right) \ . \tag{2.38}$$

Mit den eingeführten Vereinfachungen kann die Inversionsladung wie folgt aufgeschrieben werden

$$Q_i = -\Gamma_b C_{ox} \sqrt{U_T} \left(\sqrt{\frac{\Psi_s}{U_T} + \exp \frac{\Psi_s - 2\Phi_F - V}{U_T}} - \sqrt{\frac{\Psi_s}{U_T}} \right) \ . \tag{2.39}$$

Die Inversionsladung kann nicht in Abhängigkeit der Gatespannung ausgedrückt werden, da der Ausdruck für Ψ_s nicht nach V_g aufgelöst werden kann. Durch die geringe Abhängigkeit von n und Ψ_s wird es als konstant angenommen. Diese Vereinfachung hat Gültigkeit, solange $\Psi_s > 10U_T \approx 260\,\text{mV}$ gilt (vgl. Abb(2.9)) [EV06].

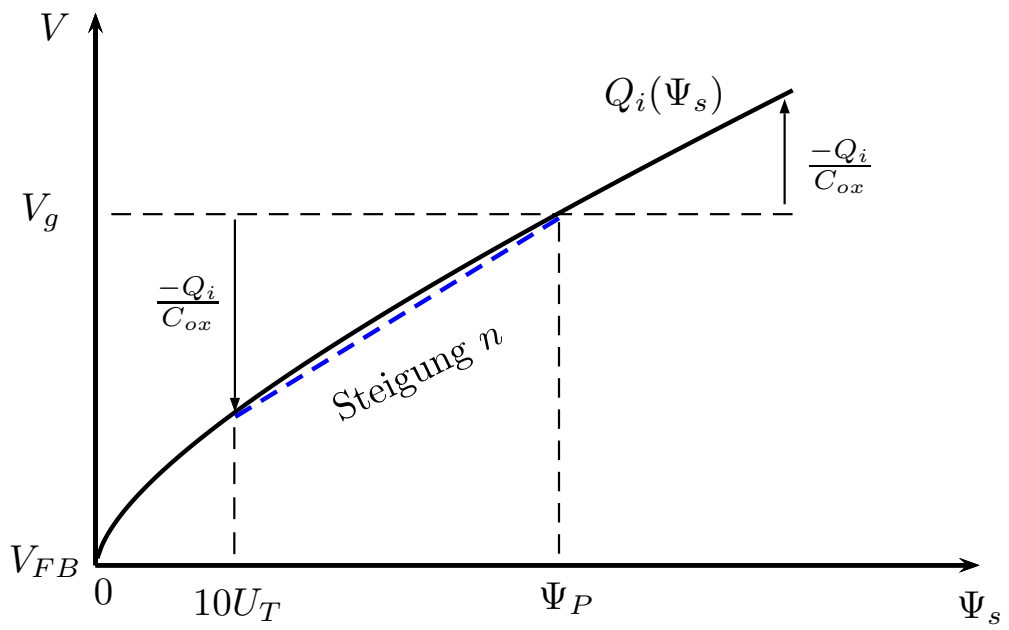


Abbildung 2.9: Linearisierung der Inversionsladung Q_i in einen Ψ_s-Intervall.

Daraus folgt

$$Q_i = nC_{ox}(\Psi_s - \Psi_P) \qquad \text{oder} \qquad \Psi_s = \Psi_P + \frac{Q_i}{nC_{ox}} \ . \tag{2.40}$$

Durch das Einsetzen in Gleichung (2.39) ergibt sich der folgende analytische Ausdruck:

$$\frac{\Psi_P - 2\Phi_F - V_c}{U_T} = \frac{-Q_i}{nC_{ox}U_T} + \ln\left[\frac{-Q_i}{\Gamma_b C_{ox}\sqrt{U_T}}\left(\frac{-Q_i}{\Gamma_b C_{ox}\sqrt{U_T}} + 2\sqrt{\frac{Q_i}{nC_{ox}U_T} + \frac{\Psi_P}{U_T}}\right)\right] . \tag{2.41}$$

Diese Gleichung bildet den allgemeinen Zusammenhang zwischen der Inversionsladung Q_i und den Spannungen V_c und V_g. Der durch die verschiedenen Vereinfachungen entstandene Fehler ist im Gültigkeitsbereich gering [EV06]. Für sehr kleine Werte des Oberflächenpotentials Ψ_s steigt der Fehler an und vor allem fehlt dem Modell in diesem Bereich die nichtlineare Abhängigkeit zwischen n und Ψ_s. Im nächsten Abschnitt werden die zur Lösung der Gleichung (2.41) benötigten Umformungsschritte beschrieben. Dies ist wichtig, da im verwendeten Transistormodell exakt diese Gleichungen implementiert wurden.

Berechnung der Inversionsladung Q_i

Um die Berechnung so allgemein wie möglich zu halten, werden sämtliche Größen normiert. Diese normierten Größen werden durch Kleinbuchstaben repräsentiert.

$$q_i = \frac{Q_i}{-2nU_T C_{ox}} = \frac{Q_i}{Q_{spec}} \tag{2.42}$$

$$\frac{V}{v} = \frac{\Psi}{\psi} = \frac{\Phi}{\phi} = \left(\frac{\Gamma_b}{\gamma_b}\right)^2 = U_T \tag{2.43}$$

Daraus folgt

$$2q_i + \ln q_i + \ln\left[\frac{2n}{\gamma_b}\left(q_i\frac{2n}{\gamma_b} + 2\sqrt{\psi_p - 2q_i}\right)\right] = \psi_p - 2\phi_f - v_c . \tag{2.44}$$

Die Terme mit q_i im Logarithmus können vernachlässigt werden, da ihr Einfluss auf den Wert des Logarithmus sehr gering ist. Daraus folgt:

$$2q_i + \ln q_i + \ln\left(\frac{4n}{\gamma_b}\sqrt{\psi_p}\right) = \psi_p - 2\phi_f - v_c . \tag{2.45}$$

Mit dieser Gleichung kann numerisch $Q_i(V_c, V_G)$ berechnet werden. Da Ψ_p bei einem Transistor nicht messbar ist, ist es sinnvoller, das Abschnürpotential in eine Abschnürspannung V_P der Kanalspannung V_c umzurechnen.

$$v_p = \frac{V_P}{U_T} = v_c(2q_i + \ln q_i = 0) = v_c(q_i = 0.4263) \tag{2.46}$$

Der Wert für q_i am Punkt $2q_i + \ln q_i = 0$ muss lediglich ein einziges Mal numerisch bestimmt werden. Mit der Definition der Abschnürspannung folgt aus Gleichung (2.45)

$$2q_i + \ln q_i = v_p - v_c \ . \tag{2.47}$$

Diese Gleichung ist analytisch nicht lösbar. Es ist jedoch möglich, eine Fallunterscheidung zu treffen. Wenn $q_i \ll 1$ dann ist $2q_i \ll \ln q_i$ und es folgt:

$$q_i = \exp(v_p - v_c) \tag{2.48}$$

und falls $q_i \gg 1$ dann $2q_i \gg \ln q_i$ und somit folgt:

$$q_i = \frac{v_p - v_c}{2} \ . \tag{2.49}$$

Berechnung des Drainstroms I_{ds}

Der für die Funktionalität wichtigste elektrische Strom ist der Drain-Source-Strom. Dieser fließt, sobald zwischen Drain und Source-Gebiet eine Spannung anliegt. Er setzt sich zusammen aus einem Drift- und einem Diffusionsstrom und fließt entlang der x-Richtung. Die gesamte Inversionsladung Q_i wird als konzentrierte Schicht unter dem Oxid angenommen und somit die z-Abhängigkeit vernachlässigt. Des Weiteren wird angenommen, dass der Kanal ausreichend weit ist und somit die Integration über y zu einer reinen Multiplikation mit der Kanalweite W reduziert werden kann. Diese Annahmen unterscheiden sich von denen für die Berechnung der Inversionsladung. Somit wird deutlich, dass im Modell keine konsistente Lösung der ursprünglichen Poisson-Gleichung berechnet wird, sondern vielmehr eine sequentielle Berücksichtigung der einzelnen Effekte.

$$I_{ds} = \mu W \left(\underbrace{-Q_i \frac{\mathrm{d}\Psi_s}{\mathrm{d}x}}_{\text{Drift}} + \underbrace{U_T \frac{\mathrm{d}Q_i}{\mathrm{d}x}}_{\text{Diffusion}} \right) \ . \tag{2.50}$$

Mit der Vereinfachung, dass die gesamte Ladung Q_i an der Oberfläche konzentriert ist, folgt:

$$Q_i \propto \exp \frac{\Psi_s - \Phi_F - V_c}{U_T} \tag{2.51}$$

und daher

$$\frac{\mathrm{d}Q_i}{\mathrm{d}x} = \frac{Q_i}{U_T}\left(\frac{\mathrm{d}\Psi_s}{\mathrm{d}x} - \frac{\mathrm{d}V_c}{\mathrm{d}x}\right) \ . \tag{2.52}$$

Daraus folgt für den Strom durch Einsetzen von Gleichung (2.52) in (2.50) und der Annahme, dass der Strom konstant ist:

$$I_{ds} = \beta \int_{V_s}^{V_d} \frac{-Q_i}{C_{ox}} \,\mathrm{d}V = I_f - I_r \tag{2.53}$$

wobei $\beta = \mu C_{ox} {}^{W}/_{L}$ der Transferparameter des Transistors ist. I_f ist der vorwärts und I_r der rückwärts fließende Stromanteil. Mit der Linearisierung der Ladung ergibt sich

$$I_{ds} = \underbrace{2n\beta U_T^2}_{I_{spec}} \int_{v_s}^{v_d} q_i \,\mathrm{d}v \ . \tag{2.54}$$

Eine wichtige Kenngröße im EKV-Modell zur Bestimmung des Betriebsbereiches eines Transistors ist der Inversionskoeffizient IC. Dieser ergibt sich aus

$$IC = \frac{I_{ds}}{I_{spec}} \tag{2.55}$$

$$\approx \frac{2n\mu C_{ox} U_T^2 \left(\frac{W}{L}\right) \ln\left(1 + \exp\left(\frac{V_g - V_P}{2nU_T}\right)\right)^2}{2n_0\mu_0 C_{ox} U_T^2 \left(\frac{W}{L}\right)} \tag{2.56}$$

$$= \frac{n\mu \ln\left(1 + \exp\left(\frac{V_g - V_P}{2nU_T}\right)\right)^2}{n_0\mu_0} \ . \tag{2.57}$$

Mit diesem Parameter kann der Betriebsbereich gemäß [EV06] wie folgt klassifiziert werden:

- schwache Inversion, $IC \ll 1$
- moderate Inversion, $IC \approxeq 1$
- starke Inversion, $IC \gg 1$

Die jeweiligen Grenzen sind nicht absolut fest und können je nach Drain-Source-Spannung, Herstellungsprozess und Transistorart leicht variieren. Abbildung 2.10 zeigt die Betriebsmodi in Abhängigkeit des Inversionskoeffizienten.
Es können je nach Betriebsbereich verschiedene Terme in der Berechnung einzelner Parameter vernachlässigt werden, was eine grobe Berechnung der Transistorparameter

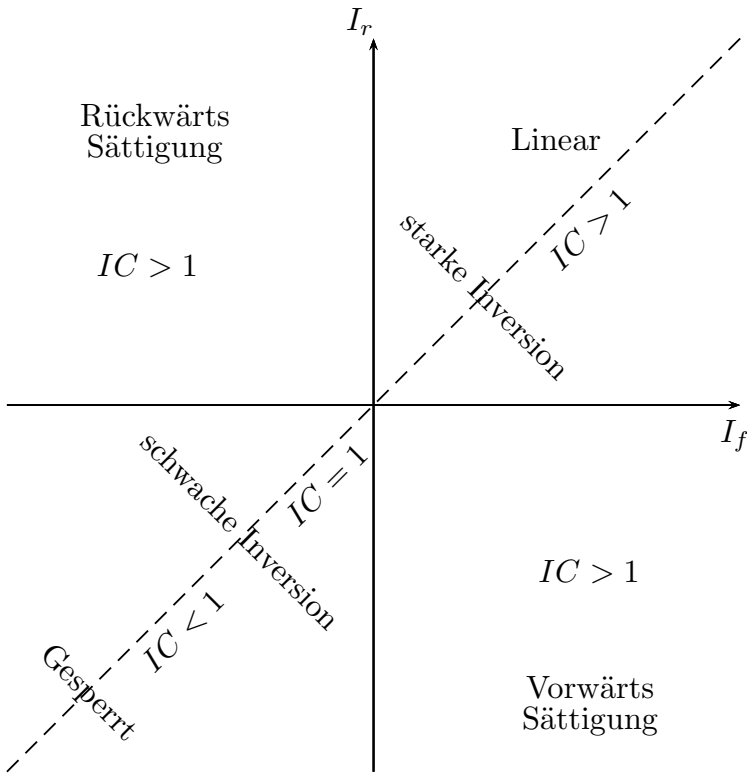


Abbildung 2.10: Einteilung der Betriebsbereiche eines MOSFET anhand des Inversionskoeffizienten IC

stark beschleunigt. Da die Kleinsignalgrößen in dieser Arbeit jedoch direkt in SPICE berechnet werden, nachdem der Arbeitspunkt mit dem vollständigen Großsignalmodell bestimmt wurde, ist diese Unterscheidung nicht notwendig.

Das BSIM6-Transistormodell Das BSIM6.1.0 MOSFET Kompaktmodell wurde 2014 an der University of California, Berkeley, entwickelt. Das Modell steht als Verilog-A Implementierung zur Verfügung [CVC+14b]. Dieses Transistormodell ist in der Lage, einen Bulk-MOSFET in allen Arbeitsbereichen konsistent zu beschreiben. Der Kern des Modells sind die im vorherigen Abschnitt vorgestellten Gleichungen zur Berechnung der Inversionsladung in Abhängigkeit von außen angelegter Spannungen. Aus den berechneten Ladungen werden im Anschluss Ströme, Spannungen und Kapazitäten des Transistors berechnet. Da diese Berechnungen eine oder mehrere Ableitungen beinhalten, ist es wichtig, dass die die Ladungen beschreibenden Funktionen mindestens zweifach stetig differenzierbar sind. Andernfalls würden einige Werte des Modells unphysikalische Wertesprünge und Inkonsistenzen aufweisen. Andere Einflüsse auf den Transistor wie parasitäre Widerstände oder Kurzkanaleffekte werden phänomenologisch behandelt, indem die globalen Größen I, V, R nach ihrer Berechnung durch diese zusätzlichen Einflüsse modifiziert werden. Die Modellierung dieser Einwirkungen wurde dabei aus dem BSIM4-Modell übernommen. Abbildung 2.11 zeigt den schematischen Ablauf der Berechnung.

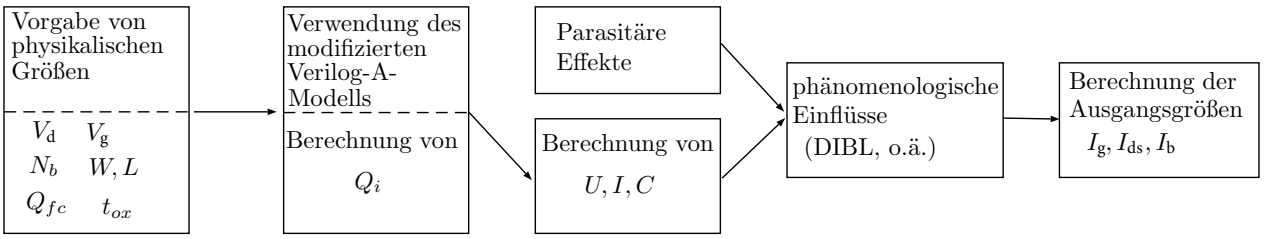


Abbildung 2.11: Berechnungsablauf im BSIM6-Transistormodell.

Die Notation in den Beschreibungen ist leicht unterschiedlich, so heißt die Kanalspannung v_c im BSIM6-Modell v_{ch} und die Steigung n heißt Ladungssteigung n_q. Die Bezeichnungen werden in dieser Arbeit so verwendet, wie in ihren jeweiligen Quellen, damit ein Vergleich leichter fällt.
Um in diesem Modell Gleichung (2.44) lösen zu können, wird im ersten Schritt ein initialer Wert für q_i berechnet und im Anschluss wird in Abhängigkeit des Arbeitspunktes dieses Ergebnis mit jeweils einem Schritt des Halley- und des Newton-Verfahrens verbessert. Der Startwert wird berechnet, indem in Gleichung (2.45) die lineare Abhängigkeit von q_i in eine Taylorreihe um den Abschnürpunkt entwickelt wird. Anschließend wird die Ladung q_p von dem Punkt $|\ln(q_i)| = |2q_i|$ berechnet. Eine detaillierte Beschreibung der Berechnung findet sich im Anhang A.5. Für den Drain-Source-Strom folgt:

$$I_{ds} = 2n_q\mu_{eff}\frac{W_{eff}}{L_{eff}}C_{ox}nU_T^2\left[(q_s - q_{deff})(q_s + q_{deff} + 1)\right] \ . \tag{2.58}$$

Alle effektiven Größen, bei denen $\square_{eff}$ im Index steht, sind solche, die durch weitere Effekte von ihrem Ursprungswert abweichen. Alle weiteren Ströme und Spannungen sowie die Kapazitäten können aus den bis hierhin berechneten Größen ermittelt werden. Für die Berechnung der phänomenologischen Einflüsse werden die für die Grundfunktion nicht benötigten Größen Schwellspannung V_{th} und Subschwellbereich-Steigung n eingeführt. Die sich aus Einflüssen wie *Drain Induced Barrier Lowering* (DIBL) ergebende ΔV_{th} wird direkt von der Flachbandspannung V_{FB} abgezogen. Der Parameter n wird in den Gleichungen benötigt, die aus dem BSIM4-Modell übernommen worden sind. Diese Art der Modellierung bereitet weniger Rechenaufwand als das Oberflächenpotential-basierte Modell. Für die Größen, die für einen Entwurf von integrierten Schaltungen notwendig sind, existieren, wie in späteren Abschnitten beschrieben, analytische Gleichungen. Somit ist diese Art der Modellierung ein Kompromiss aus den beiden zuvor beschriebenen Ansätzen.

2.3 Zuverlässigkeit

Die Zuverlässigkeit einer Schaltung oder eines Systems gibt die Fähigkeit an, Spezifikationen, die zum Zeitpunkt der Herstellung festgelegt wurden, in bestimmten Verhältnissen über eine spezifizierte Zeitdauer einzuhalten [Han88]. Somit beschäftigt sich die Zuverlässigkeit vornehmlich mit dem Ausfall oder der Fehlfunktion eines Systems nach einem bestimmten aber zufälligen Zeitpunkt $T > 0$. Dafür wird die zeitabhängige Zufallsvariable $F(t) = P(T \leq t)$ mit der Wahrscheinlichkeitsdichteverteilung f eingeführt. Die Fehlerrate ist durch

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{2.59}$$

definiert, wobei die Überlebensfunktion $S(t) = 1 - F(t)$ ist [AJ99]. Die Fehlerrate ist die einzige dieser Größen, die bei technischen Systemen überwacht oder gemessen werden kann. Alle Aussagen über die Zuverlässigkeit eines Systems sind stets auf $\lambda(t)$ zurückzuführen. Eine typische Darstellung der Fehlerrate von integrierten Systemen ist die Badewannenkurve, welche in Abbildung 2.12 gezeigt wird. Zu erkennen sind drei Bereiche mit unterschiedlichen Ausfallraten. Direkt nach der Herstellung ist die Fehlerrate hoch und fällt schnell ab. Die dort auftretenden Fehler sind zumeist durch Abweichungen während der Herstellung bedingt.

Diese Abweichungen werden Prozessvariationen genannt, üblicherweise führen sie nur zu einer statistischen Verteilung der Herstellungsparameter, können jedoch auch zu Fehlern führen. Diese Fehlerrate wird gesenkt, indem direkt nach der Herstellung sogenannte Burn-In Tests durchgeführt werden. Dabei werden Systeme und Schaltungen mit Temperaturen und Spannungen oberhalb ihrer spezifizierten Bereiche betrieben. Schaltungen, die bereits nach der Herstellung fehleranfällig waren, werden durch diesen Test zum Auslauf gebracht und werden erst gar nicht an den Kunden geliefert. Anschließend bildet die Fehlerrate über die Lebzeit der Schaltung ein Plateau. Alle Fehler die während dieser Zeit auftreten sind rein zufällig. Systematische Fehler wie der Divide-by-zero Fehler der Intel Pentium Prozessoren oder die Probleme mit den Akkumulatoren in Samsung Mobiltelefonen bleiben in der Badewannenkurve unberücksichtigt. Ab einem bestimmten Punkt setzt der Verschleiß ein und die Fehlerrate in der Badewannenkurve steigt an. Die Systeme verlassen nach und nach den spezifizierten Bereich bis schlussendlich keine Schaltung mehr funktioniert.

In diesem Abschnitt werden Mechanismen beschrieben, die die Transistoren einer integrierten Schaltung betreffen und zu einem Anstieg der Fehlerrate führen. Anschließend werden die resultierenden stochastischen Auswertungsmöglichkeiten und Implikationen dargestellt.

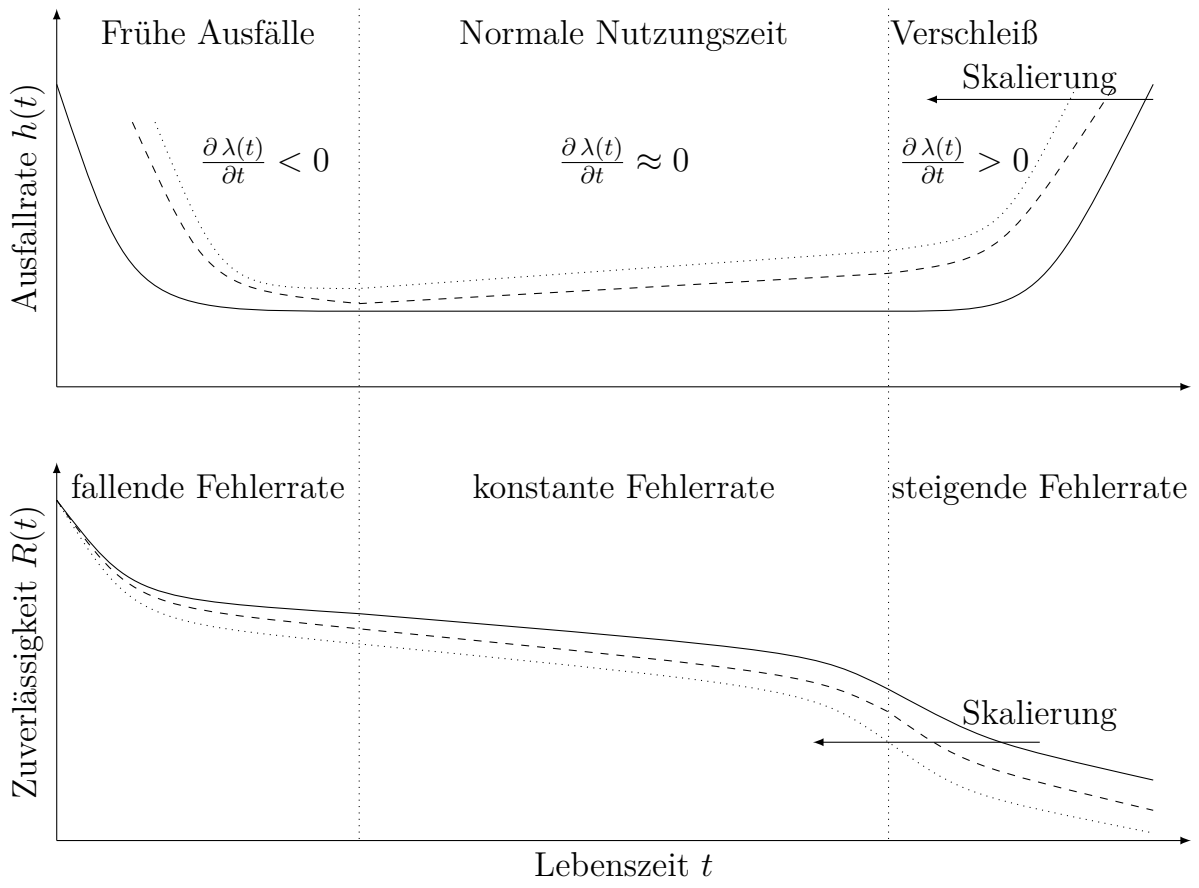


Abbildung 2.12: Typische Badewannen-förmige Fehlerraten-Kurven einer integrierten Schaltung.

2.3.1 Einflüsse auf integrierte Schaltungen

In diesem Abschnitt werden die Mechanismen beschrieben, die sich auf die Fehlerrate von Transistoren in integrierten Schaltungen auswirken. Dazu zeigt Abbildung 2.13 die Anordnung der Herstellungslagen unterteilt in das sogenannte Front-End-Of-Line (FEOL), Back-End-Of-Line (BEOL) sowie das Advanced Packaging [DH11, Sah10].
In dieser Arbeit werden nur solche Effekte betrachtet, die die FEOL und deren Komponenten und Herstellungsverfahren betreffen. In diesen Bereich fallen alle aktiven und die meisten passiven Bauelemente. In einigen Herstellungsverfahren werden MIM-Kapazitäten und Widerstände im BEOL eingebettet. Zu den wichtigsten Bauteilen für analoge Schaltungen zählen heutzutage MOSFET, daher sind Effekte, die diese Bauteile betreffen, maßgeblich für die Zuverlässigkeit der Schaltung verantwortlich.

2.3.1.1 Prozessvariationen

Prozessvariationen sind Abweichungen bei der Herstellung von integrierten Bauelementen [DH11, Zja16]. Aus Sicht der Transistormodellierung sind davon alle Herstellungsparameter betroffen. Dazu zählen die Länge L, die Weite W, die Oxiddicke t_{ox},

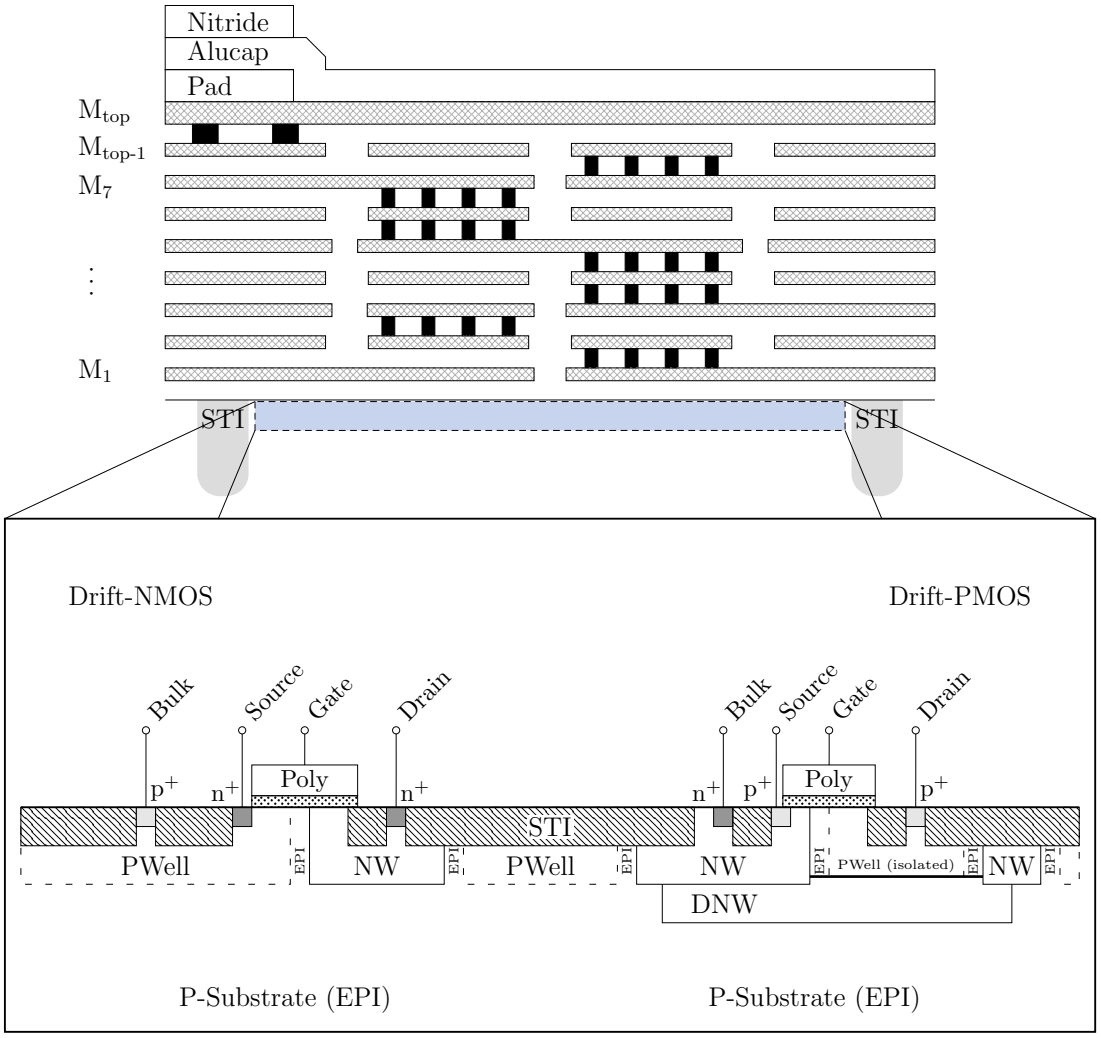


Abbildung 2.13: Anordnung der Herstellungslagen einer integrierten Schaltung und deren Einordnung.

sowie die Dotierkonzentrationen im Drain- und Source Bereich $N_{D/S}$ und im Bulk N_B [BN00, DM03, CC07]. Diese Abweichungen von den spezifizierten Werten haben unterschiedliche Gründe, zudem sind durch kleiner werdende Technologieknoten weitere Ursachen hinzugekommen. Die wichtigsten Einflüsse in der in dieser Arbeit betrachteten 65nm Technologie sind im Weiteren knapp beschrieben.

Random Dopant Fluctuation Aufgrund dessen, dass die Anzahl der Dotieratome im Kanal für Technologieknoten mit der Kanallänge abnimmt, ist die örtliche Position der Dotieratome maßgeblich für die Leitungseigenschaften des Transistors verantwortlich. Die Abweichungen werden dabei mit der Stolk-Gleichung nach [Sah10] modelliert:

$$\sigma_{V_{th}} = \frac{\sqrt[4]{4q^3\varepsilon_{si}\phi_b}}{2} \cdot \frac{t_{ox}}{\varepsilon_{ox}} \cdot \frac{\sqrt[4]{N_b}}{\sqrt{W \cdot L}} \tag{2.60}$$

Dabei ist $\sigma_{V_{th}}$ die durch diesen Effekt hervorgerufene Standardabweichung der Schwellspannung. An der Gleichung wird deutlich, dass die einzelnen Effekte miteinander in Verbindung stehen. Variiert zum Beispiel t_{ox} durch Prozessvariationen ändert sich $\sigma_{V_{th}}$.

Line-edge und Line-width Roughness Von dieser Variation sind Transistoren betroffen, deren Gate-Kontakt nicht aus Metall, sondern aus Polysilizium besteht. Dabei werden die Kanten und Oberflächen des Polysilizium so rau, dass sich W und L inhomogen im Transistor verändern. Dadurch wird die Steuerbarkeit der Kanalladung durch die Gate-Spannung beeinträchtigt und somit erhöht sich die Schwellspannung und der Leckstrom über die Drain-Source-Strecke. Zudem sind alle geometrischen Abhängigkeiten davon betroffen und leicht verschoben.

Oxiddicke Bedingt durch das Aufwachsen des Oxids ist dieses nicht gleich hoch. Die Dicke kann sich innerhalb eines Bauelements wie auch zwischen verschiedenen Bauelementen ändern. Für sehr geringe Oxiddicken werden zusätzlich Schichten mit höheren Permittivitäten als Siliziumdixoid auf das Gate aufgebracht. Dadurch wird vor allem der Leckstrom durch das Gate reduziert. Diese zusätzlichen Schichten können ortsfeste Ladungsträger enthalten. In Zusammenhang mit der Variation der Oxiddicke sind die Schwellspannung, die Ladungsträgerbeweglichkeit sowie der Sättigungsstrom betroffen.
Allgemein können Einflüsse durch Prozessvariationen auf einzelne Transistorparameter $\mathcal{P}$ durch

$$\sigma_{\mathcal{P}} = \frac{c_{\mathcal{P}}}{\sqrt{W \cdot L}} \tag{2.61}$$

modelliert werden [PDW89]. Dabei ist $c_{\mathcal{P}}$ ein Modellparameter, der für jeden Effekt und jeden Parameter empirisch ermittelt werden muss.

2.3.1.2 Temperatur und Spannung

Die Temperatur eines Transistors kann sich zum einen durch die Umgebungstemperatur ändern oder zum anderen durch die Erwärmung der Schaltung durch joulesche Verluste in der Schaltung selbst. Nahezu alle Parameter eines Transistors sind von der Temperatur abhängig. Begründet ist dies in der Tatsache, dass die Leitfähigkeit des Halbleiterkristalls selbst exponentiell von der Temperatur abhängt [AM07, LF12]. Der grundlegendste Wert, der durch die Temperatur verändert wird, ist die intrinsische Ladungsträgerkonzentration n_i. Für diese gilt

$$n_i^2 = N_C N_V \exp\left(\frac{-E_G}{k_b T/q}\right) \ . \tag{2.62}$$

Dabei ist E_G die Energie der Bandlücke und $N_{C/V}$ sind die jeweiligen Zustandsdichten im Leitungs- und Valenzband. Ändert sich n_i so ändert sich ebenfalls das Ferminiveau Φ_F

$$\Phi_F = U_T \ln\left(\frac{N_D}{n_i}\right) \ . \tag{2.63}$$

Dabei ist $U_T = k_b T$ und N_D die Dotierkonzentration. Wie im vorherigen Kapitel gezeigt wurde, basieren alle weiteren Berechnungen der Transistoreigenschaften auf dem Ferminiveau. Daher sind alle Eigenschaften, bis auf die Prozesseigenschaften (t_{ox}, W, L), von der Temperatur abhängig. Bei einem MOSFET werden vornehmlich die Variationen der Drain-Spannung V_d und der Gate-Spannung V_g betrachtet, da diese maßgebend für das Ausgangsverhalten des Transistors sind. Die Spannungen an den anderen beiden Anschlüssen Source V_s und Bulk V_b beeinflussen das Verhalten des Transistors zwar ebenfalls, dabei kann die Änderung von V_s in die Variation von V_d umgerechnet werden. Die Bulk-Spannung hat maßgeblichen Einfluss auf die Bulk-Ladung Q_b und Änderungen von V_b wirken sich nur gering auf das Ausgangsverhalten eines Transistors aus. Außerdem ist der Bulk-Anschluss in vielen Anwendungsfällen mit V_{ss} (NMOS) oder V_{dd} (PMOS) verbunden. Daher können Änderungen der Bulk-Spannung wiederum in eine Variation der Versorgungsspannung umgerechnet werden.

Es werden die Variationen der beiden Spannungen V_d und V_g betrachtet. Da es statistisch *viele* Einflüsse auf das Rauschen der Quellen gibt und dadurch der zentrale Grenzwertsatz [Pól20] Anwendung finden kann, können die Spannungen als normalverteilt angenommen werden. Der zusätzliche Rauschterm wird zum aktuellen Wert der jeweiligen Spannung addiert. Die Normalverteilung besitzt zwei Freiheitsgerade, zum einen den Mittelwert μ und zum anderen die Standardabweichung σ. Die Gleichung der Verteilung ist

$$\Delta V_{g/d} = \mathcal{N}(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \ . \tag{2.64}$$

Für beide verwendeten Verteilungsfunktionen der Spannung gilt $\mu_{V_d} = \mu_{V_g} = 0$, da es sich um additive Rauschterme handelt. Dies folgt daraus, dass beide additiv den jeweiligen Spannungen hinzugefügt werden. Die jeweilige Standardabweichung richtet sich nach der jeweiligen initialen Versorgungsspannung, Temperatur und etwaigen internen Rauschquellen innerhalb einer Schaltung.

Der schwerwiegendste Fall im Zusammenhang mit Temperatur ist der thermische Drift. Es beruht auf der Eigenerwärmung eines Bauteils durch den fließenden Strom und ohmsche Verluste. Dieser Effekt kann zur thermischen Zerstörung des Bauteils oder der Schaltung führen. In späteren Kapiteln wird bei der MOS-Diode ebenfalls gezeigt, dass die Wechselwirkung zwischen einzelnen Effekten wie Temperatur und Alterung den thermischen Drift beeinflussen und verstärken kann. Neben diesem Extremfall sind, bedingt durch die Änderung der Mobilität und der Leitfähigkeit, nahezu alle Transistorparameter von der Temperatur abhängig. Neben der Temperatur haben

unerwünschte Spannungsschwankungen an den Kontakten des Transistors Einfluss auf seine einwandfreie Funktion.
Um den Einfluss dieser Variationen zu verdeutlichen, werden die Abhängigkeiten in einem stark vereinfachten MOS1 Transistormodell ausgerechnet. Durch das Bilden des jeweiligen totalen Differenzials der Spannung kann die Auswirkung der Änderung in den verschiedene Arbeitsbereichen des Transistors bestimmt werden.

1. Subschwellbereich

$$I_{ds} = I_{ds,0} \cdot \mathrm{e}^{\frac{V_g - V_{th}}{k_b T/q}} \tag{2.65}$$

$$\Delta I_{ds} = nU_T I_{ds,0} \cdot \mathrm{e}^{\frac{V_g - V_{th}}{k_b T/q}} \cdot \Delta V_g \tag{2.66}$$

2. Linearer Bereich

$$I_{ds} = \mu C_{ox} \frac{W}{L} \left((V_g - V_{th}) V_d - \frac{1}{2} V_d^2 \right) \tag{2.67}$$

$$\Delta I_{ds} = -\mu C_{ox} \frac{W}{L} V_d \cdot \Delta V_g \tag{2.68}$$

$$+ \mu C_{ox} \frac{W}{L} (V_g - V_{th} - V_d) \cdot \Delta V_d \tag{2.69}$$

3. Sättigungsbereich

$$I_{ds} = \frac{1}{2} \mu C_{ox} \frac{W}{L} (V_g - V_{th})^2 (1 + \lambda V_d) \tag{2.70}$$

$$\Delta I_{ds} = \mu C_{ox} \frac{W}{L} (V_g - V_{th})(1 + \lambda V_d) \cdot \Delta V_g \tag{2.71}$$

$$+ \frac{1}{2} \mu C_{ox} \frac{W}{L} \lambda (V_g - V_{th}) \cdot \Delta V_d \tag{2.72}$$

Wie erkennbar ist, sind die Auswirkungen von Spannungsschwankungen sowohl am Gate- als auch am Drain-Anschluss deutlich unterschiedlich, in Abhängigkeit der jeweiligen Arbeitspunkte. Die sehr einfachen Modellierungsansätze liefern ungenaue Ergebnisse sobald $V_g \approx V_{th}$ oder $V_{ds} \approx V_{ds,sat}$ gilt, da dann Rauscheinflüsse die eindeutige Zuordnung des Betriebspunktes erschweren. Dies führt insbesondere bei der initialen Bestimmung der Schwellspannung eines Transistors zu Problemen. Abbildung 2.14 zeigt die resultierenden Veränderungen des Drainstroms anhand eines Kennlinienfeldes eines NMOS-Transistors und die sich ergebenden Verteilungsfunktionen für ΔI_{ds}. Die Verteilungsfunktionen sind dabei Simulationswerte von 200 NMOS-Transistoren in einer 65nm Technologie. Klar zu erkennen ist, dass durch die jeweilige Asymmetrie keine der Verteilungsfunktionen einer Normalverteilung folgt. Besonders im Schwellbereich und im linearen Bereich sind die Verteilungen für ΔI_{ds} stark abweichend von den Verteilungen für $\Delta V_{g/d}$.

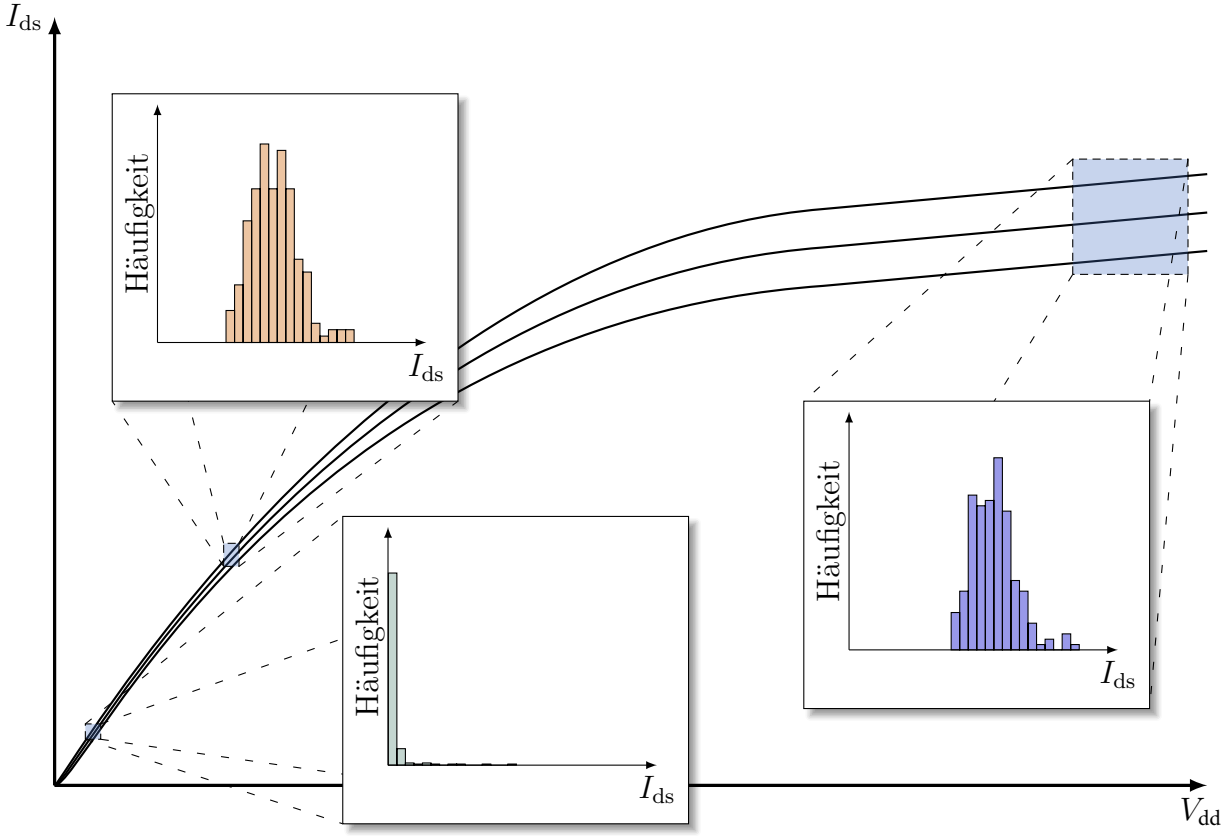


Abbildung 2.14: Schematische Darstellung der Abhängigkeiten von I_{ds} von V_d und V_g und Wahrscheinlichkeitsdichteverteilungen für ΔI_{ds} in verschiedenen Betriebsbereichen.

2.3.1.3 Alterungsmechanismen

In diesem Abschnitt werden die betrachteten Alterungsmechanismen sowie die zu ihrer jeweiligen Beschreibung verwendeten Modelle eingeführt und erläutert. Alterung von MOSFET meint eine Änderung der Transistorparameter über der Zeit, abhängig von den jeweiligen Betriebsbedingungen der Transistoren. Alterungsmechanismen werden in zwei Gruppen eingeteilt, solche, die abrupt zu einer Fehlfunktion oder einem Totalausfall des Transistors führen und solche, die im Laufe der Zeit die Parameter so verändern, dass der Transistor nicht mehr sein ursprüngliches Betriebsverhalten zeigt. Zur ersten Kategorie gehören zum Beispiel der *Time Dependent Dielectric Breakdown* (TDDB). Bei diesem Fehlermechanismus bildet sich ein elektrisch leitender Pfad durch das Gateoxid. Ein Bestehen dieses Pfades führt zum Totalausfall des Transistors. Mechanismen, die der ersten Kategorie zuzuordnen sind, werden in dieser Arbeit nicht betrachtet, da keine zeitabhängige Beschreibung, sondern lediglich Auftrittswahrscheinlichkeitsverteilungen als Modelle existieren. Zur zweiten Kategorie gehören die *Hot Carrier Injection* (HCI) [Gra14] und die *Bias Temperature Instability* (BTI) [Gra13, GKG+11]. Bei diesen Effekten handelt es sich um entscheidende Alterungsmechanismen. Beide Effekte führen

zu einer in Teilen permanenten Erhöhung von Q_{fc}. Dies geschieht durch Ladungsträger, die entweder vorhandene Fehlstellen besetzen oder so energiereich sind, dass diese weitere ungesättigte Bindungen (engl. *Dangling Bonds*) erzeugen können, indem vorhandene Bindungen zerstört werden.
In der Literatur finden sich verschiedene Beschreibungen für Ursache und Wirkung von HCI und BTI. Beide Effekte weisen klare Abhängigkeiten gegenüber der Gate-, Drain-, Source- und Bulk-Spannung sowie der Temperatur auf [Gra13, Gra14, MMM+14, MG13, MP18]. Abbildung 2.15 zeigt Stressbedingungen für HCI und BTI über V_{ds} und V_{gs}. Wie zu erkennen ist, finden die stärksten Beeinflussungen beider Effekte deutlich oberhalb der jeweils nominellen Spannungen statt. Zum einen wird HCI durch hohe Spannungen begünstigt, da somit viele (V_{gs}) und stark beschleunigte (V_{ds}) Ladungsträger im Kanal vorhanden sind. Des Weiteren sind die Dauer einzelner Stressbedingungen und die

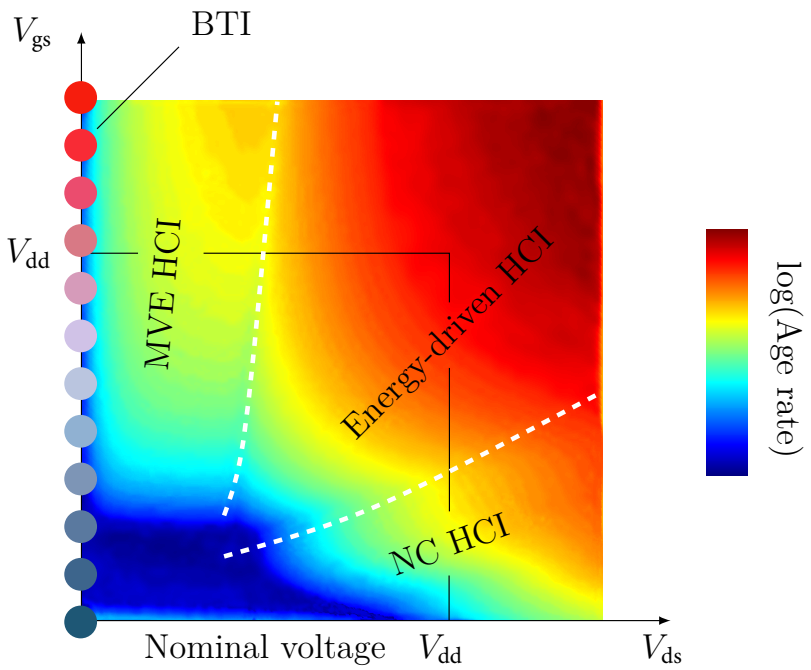


Abbildung 2.15: Stressbedingungen für HCI (MVE - Multi Vibrional Exaltation, NC - Non Conducting)und BTI über V_{ds} und V_{gs}. Nach [HCF+15]

Wiederherstellungsphasen wichtig [RGE+11]. Zudem hängen beide Effekte von der Prozessvariation [SBP+17], W, L, dem W/L-Verhältnis [SPB+15], t_{ox} und E_{ox} ab. Es besteht Konsens, dass beide Mechanismen zu einer Erhöhung der Ladungsträgerdichte in und unterhalb des Gateoxids führen. Jedoch werden die Ursachen dafür ambivalent diskutiert [SMG18]. Einige Arbeiten legen nahe, dass die Effekte zumindest teilweise zufällig sind [GKG+11, KFW+18] und dass BTI mit dem Random Telegraph Noise (RTN) zusammenhängt [ZDJZ16]. Die Auswirkungen einzelner Fehlstellen können, besonders bei kleinen Technologieknoten, stark schwankend sein. Dieses wird durch das sogenannte *percolation path model* beschrieben [GHA+15, FKM+16]. Abhängigkeiten

zwischen HCI, BTI und TDDB können durch verschiedene Tunnelmechanismen verursacht werden. Zudem kann die effektive elektrische Feldstärke im Oxid durch zusätzliche Ladungsträger erhöht werden [SAG+12, MJC+14].

Hot Carrier Injection Dieser Alterungsmechanismus basiert auf Ladungsträgern mit hohen kinetischen Energien [TMJ+18]. Die kinetische Energie entsteht dabei durch ein starkes laterales elektrisches Feld im Kanal des MOSFET, welches die Ladungsträger beschleunigt. Da die Ladungsträger vom Source-Kontakt zum Drain-Kontakt hin beschleunigt werden, treten die höchsten kinetischen Energien nahe der Drainregion auf [UJT+17]. In diesem Gebiet kann es passieren, dass diese hochenergetischen Ladungsträger mit Gitteratomen kollidieren und ihre Energie an Bindungselektronen abgeben, was wiederum dazu führen kann, dass eine Bindung zerstört wird. Dazu müssen die Ladungsträger höhere Energien besitzen als solche, die am normalen Ladungstransport beteiligt sind. Dieser Mechanismus begünstigt, sobald niedrige Temperaturen die mittlere freie Weglänge erhöhen und zudem die Ladungsträger durch hohe laterale elektrische Felder stark beschleunigt werden [TJS+15, TMJ+18]. Falls die Umgebung nach der Zerstörung der Bindung relaxiert, ist die Bindung permanent zerstört und ein zusätzlicher Dangling Bond ist entstanden. Durch diesen Effekt erhöht sich die Ladung Q_{fc} und/oder Q_{it} und beeinflusst dadurch das gesamte Verhalten des Transistors. Diese Einflüsse führen zu einer Asymmetrie der elektrischen Charakteristik des Transistors [Gra14]. Dies wurde bereits in einem Kompaktmodell in [LK12] abgebildet. Durch HCI zerstörte Bindungen sind irreversibel, da bei diesem Effekt stets eine Relaxierung stattfindet. Wie anfällig eine Bindung gegenüber diesem Effekt ist, hängt von der Bindungsenergie ab. Je geringer diese ist, desto wahrscheinlicher ist es, dass diese durch einen energiereichen Ladungsträger zerstört wird.
Es existieren drei Möglichkeiten, wie durch Hot Carrier Injection ein Bindungselektron in das Leitungsband angeregt werden kann und somit die Bindung zerstört wird [Gra14].

1. **Einzel-Ladungsträger-Prozess**
   Bei diesem Prozess wird ein einzelner Ladungsträger durch ein starkes elektrisches Feld so stark beschleunigt, dass dieser durch eine Kollision die vollständige Energie an das Bindungselektron abgeben kann, die benötigt wird, um das Bindungselektron in das Leitungsband anzuregen. Daher hängt dieser Vorgang von der Stärke des lateralen elektrischen Feldes im Kanal, der Ladungsträgerbeweglichkeit und der Sättigungsgeschwindigkeit der Ladungsträger ab. Daraus folgt, dass dieser Prozess besonders stark in Kurzkanal-NMOSFET auftritt.

2. **Mehrfach-Ladungsträger-Prozess** Beim Mehrfach-Ladungsträger-Prozess werden die Bindungselektronen durch mehrfache Kollision mit am Ladungstransport durch den Kanal beteiligten Ladungsträgern in Zwischenenergieniveaus angeregt. Dadurch ist es nicht notwendig, dass ein einzelner Ladungsträger die zur Zerstörung der Bindung benötigte Energie bereitstellt. Dieser Prozess hängt, zusätzlich zu den

gleichen Größen des Einzel-Ladungsträger-Prozesses, von der Anzahl der möglichen Zwischenzustände ab. Je mehr Zwischenzustände existieren, desto geringer ist der jeweilige energetische Abstand und desto wahrscheinlicher ist es, dass das Bindungselektron angeregt wird.

3. **Ladungsträger-Streuung** Bei diesem Streuungsvorgang kollidieren zwei Ladungsträger elastisch im Kanal und es entsteht ein sehr energiereiches Elektron. Dieser Vorgang hängt maßgeblich von der Anzahl der Ladungsträger ab und nicht, wie die anderen Prozesse, von der Transistorlänge und dem lateralen elektrischen Feld im Kanal. Je mehr Ladungsträger vorhanden sind, desto wahrscheinlicher ist eine solche Kollision. Das auf diese Weise entstandene energiereiche Elektron ist bei einer Kollision mit einem Gitteratom in der Lage, das Bindungselektron in das Leitungsband anzuregen.

Abbildung 2.16 zeigt alle drei Mechanismen, die zur Zerstörung einer Bindung führen können. Wie bereits genannt, ist diejenige Bindung anfällig, die eine geringe Bindungs-

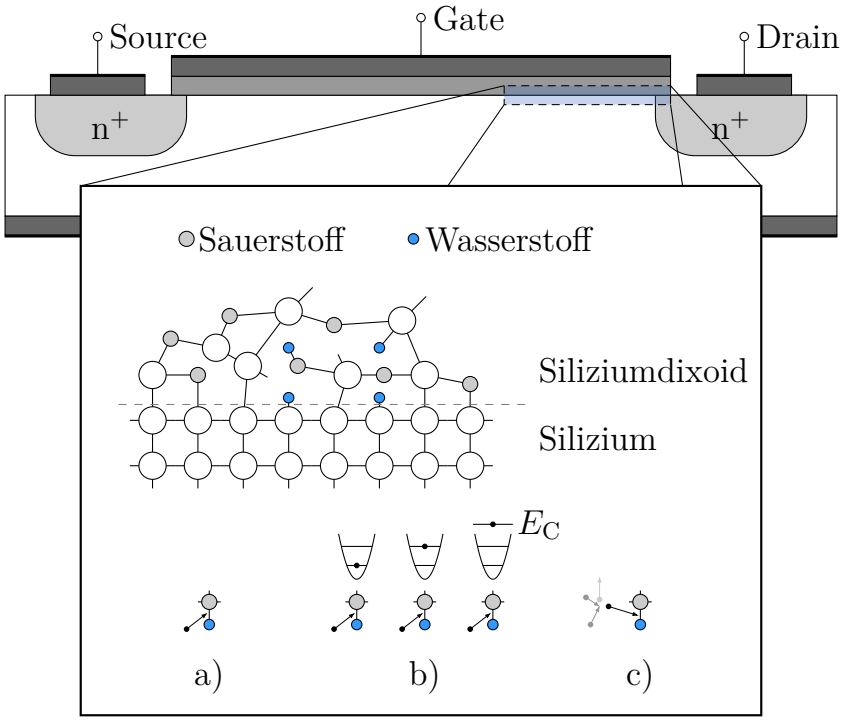


Abbildung 2.16: Querschnitt durch die Grenzschicht zwischen Substrat und Oxid, sowie die drei HCI-Mechanismen. a) Einzel-Ladungsträger b) Multi-Ladungsträger und c) Streuungs-Prozess

energie besitzt. In der Umgebung des Inversionskanals existieren folgende Bindungen:

- Silizium-Silizium, $\mathrm{Si} - \mathrm{Si}$
- Silizium-Phosphor (N-Dotierung), $\mathrm{Si} - \mathrm{P}$

- Silizium-Bor (P-Dotierung), $Si - B$
- Silizium-Sauerstoff, $Si - O$
- Silizium-Wasserstoff, $Si - H$
- Sauerstoff-Sauerstoff, $O = O$, $Si - O - O - Si$
- Sauerstoff-Wasserstoff, $O - H$

Phosphor und Bor sind beispielhaft als Dotieratome genannt. Die geringsten Bindungsenergien besitzen Bindungen mit Wasserstoff. Diese treten durch das Passivieren bei der MOSFET-Fertigung an der Grenzschicht zwischen Substrat und Oxid auf. Somit tritt HCI in integrierten Schaltungen vor allem an der Oxid-Grenzschicht bei Kurzkanal-NMOSMET auf. Der physikalische Effekt tritt jedoch in jedem MOSFET auf, nur die Intensität und die Auswirkungen unterscheiden sich.

Bias Temperature Instability Der physikalische Mechanismus ist bei der Bias Temperature Instability derselbe wie bei der Hot Carrier Injection. Durch energiereiche Ladungsträger werden Bindungen mit geringer Bindungsenergie zerstört. Der Unterschied besteht in der Ursache und damit im Ort des Auftretens und den daraus resultierenden Abhängigkeiten von Betriebsparametern. Bei BTI werden Ladungsträger durch eine negative Bias-Spannung am Gate-Kontakt in Richtung der Oxid-Schicht beschleunigt. An der Grenzschicht können diese Ladungsträger Bindungen zerstören und somit Dangling Bonds erzeugen (Q_{it}) oder tieferliegende Fehlstellen (Q_{trap}) besetzen, die zuvor thermisch ionisiert wurden. Beides führt zu einer Erhöhung der Ladung unter dem Gate-Oxid. Diese Komponenten sind nach Beendigung der Stressbedingung (starkes vertikales elektrisches Feld) wiederherstellbar. Einige Zusammenhänge werden bei BTI in der Literatur immer noch kontrovers diskutiert [SMG18].
Dadurch, dass die Ladungsträger in Richtung des Oxids beschleunigt werden müssen, tritt dieser Effekt stärker bei PMOSFET auf. Die starke Temperaturabhängigkeit resultiert aus der thermischen Ionisation der tieferliegenden Fehlstellen. Je höher die Temperatur, desto mehr besetzbare Zustände existieren für die Ladungsträger. Abbildung 2.17 zeigt beide modellierten Effekte. Für NMOS-Transistoren mit einem high-k Gate tritt ebenfalls BTI auf [MMM+14, CAD+10].

2.3.2 Skalierung und Zuverlässigkeit

Durch die immer kleiner werdenden Technologieknoten ist die Zuverlässigkeit von integrierten Schaltungen immer weiter in den Fokus von Forschung und Entwicklung gerückt [HCF+15, KSM+07]. Ursache dafür sind die unterschiedlichen Skalierungen von entscheidenden Technologieparametern (z.B. Versorgungsspannung, Schwellspannung) wie in Abbildung 2.18 dargestellt ist.

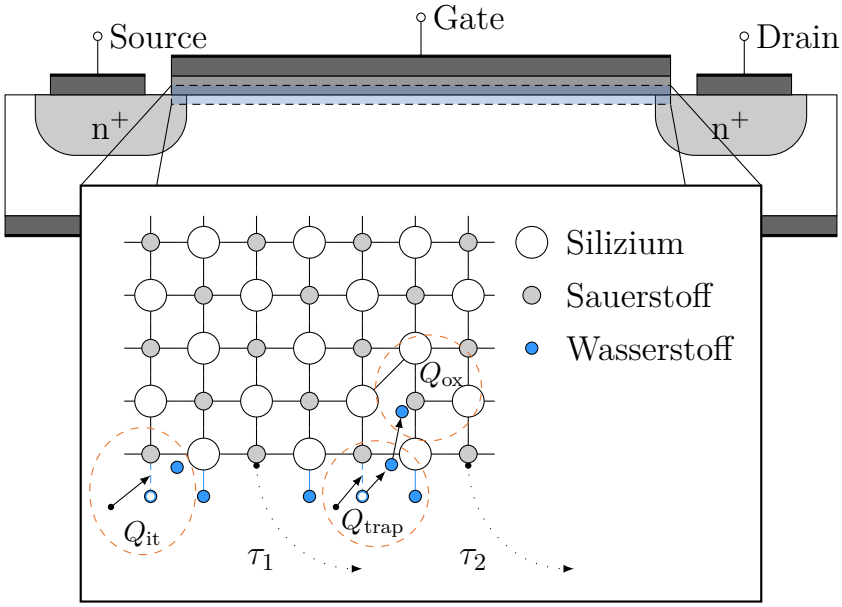


Abbildung 2.17: BTI-Mechanismen, Besetzung vorhandener Fehlstellen Q_{trap}, Q_{it} und Q_{ox}

Wie zu erkennen ist, nehmen V_{ds} und V_{th} unterhalb von 100nm nur noch geringfügig ab. Dagegen fällt die Oxiddicke weiterhin mit der gleichen Steigung wie für größere Technologien. Diese unterschiedliche Skalierung führt dazu, dass die elektrischen Felder sowohl im Oxid als auch lateral zwischen Drain- und Sourceanschluss zunehmen. Des Weiteren nimmt die Fläche des Transistors durch kürzere Längen L ab, bei gleichem V_{ds} erhöht sich somit die Stromdichte. Dies führt zu erhöhten ohmschen Verlusten und zu einer Eigenerwärmung. Zudem werden die Einflüsse durch Prozessvariationen verstärkt. Gleichzeitig bleibt die Schwellspannung ebenfalls nahezu unverändert unter 100nm Gatelänge. Dies bedeutet bei kleinerem t_{ox} höhere Leckströme im normalen operativem Einsatz des Transistors. Zusammengefasst werden alle zuvor beschriebenen Mechanismen, die die Zuverlässigkeit von Transistoren beeinflussen, durch den anhaltenden Trend der Verkleinerung von Technologieknoten verstärkt.

2.3.3 Strahlungseinflüsse

In einigen Einsatzgebieten wie der Luft- und Raumfahrt aber auch in Teilchenbeschleunigern sind Schaltungen verschiedenen Strahlungseinflüssen ausgesetzt. Die beiden Haupteinflüsse sind die Total Ionizing Dose (TID) [OM03, Bar06] und Single Event Upsets (SEU) [ABKF10]. Dabei beschreibt TID eine permanente Änderung des Transistorverhaltens durch anhaltenden Beschuss durch Partikel und daraus resultierende zusätzliche positive Ladungen im Gateoxid. Je nachdem ob es sich um einen N- oder PMOS handelt, steigt bzw. sinkt die Schwellspannung dieses Bauteils graduell, bis es sich nicht mehr ein bzw. ausschalten lässt. Ein SEU führt zu einer sehr kurzfristigen Änderung des Drainstroms eines Transistors durch einen Teilcheneinschlag in unmit-

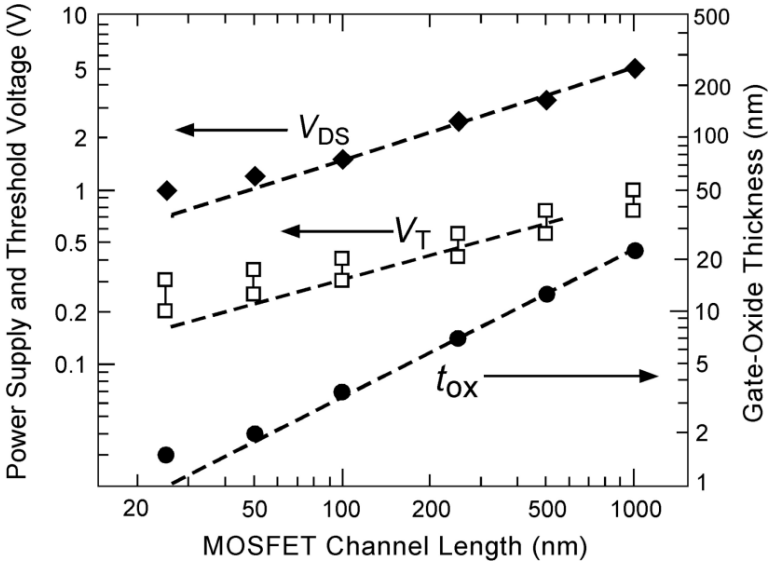


Abbildung 2.18: Skalierung der Versorgungsspannung, der Schwellspannung und der Oxiddicke über verschiedene Technologieknoten. Datenquelle:[TBC+97] Bildquelle:[Whi12]

telbarer Nähe zum Transistor. Abbildung 2.19 zeigt die Auftrittspunkte für TID und SEU anhand des Querschnitts eines N-MOS Transistors. Im Bild ist X der Abstand zwischen dem Ort des Einschlags und der betrachteten MOS-Struktur und r der Radius des erzeugten Ionenkanals. Beide Parameter werden später im Simulationsmodell des SEU in Abschnitt 3.1.2 wieder aufgegriffen. Die Schäden durch SEU sind jedoch nicht-

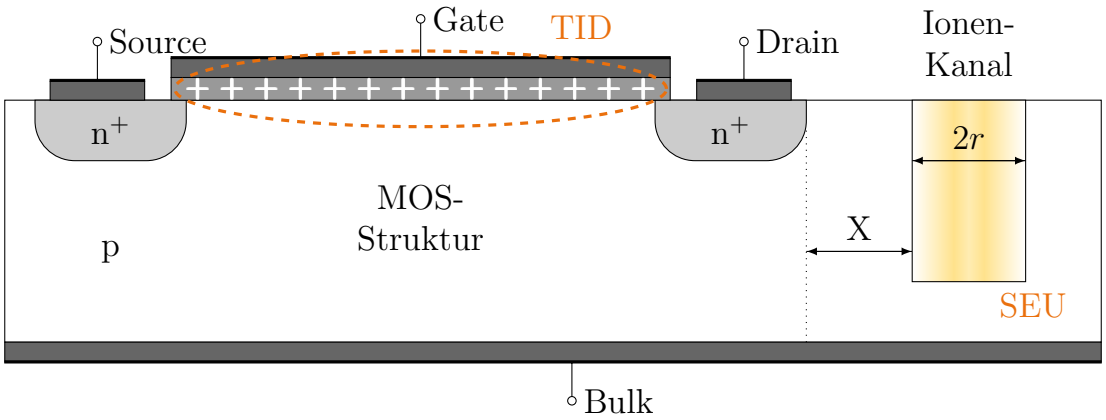


Abbildung 2.19: Querschnitt einen NMOS Transistor mit den Auftrittspunkten für TID und SEU.

permanent. Die Stärke der Auswirkungen hängt dabei von der Energie des Teilchens, ebenso jedoch vom Arbeitspunkt des Transistors, den Prozessvariationen und dem Alter des Bauelements ab. Dies ist in Abbildung 2.20 für einen NMOS Transistor gezeigt. Wie gezeigt, wechselwirken alle Einflüsse sehr unterschiedlich miteinander. So ist der Einfluss in schwacher Inversion gegenüber den anderen Betriebspunkten am stärksten.

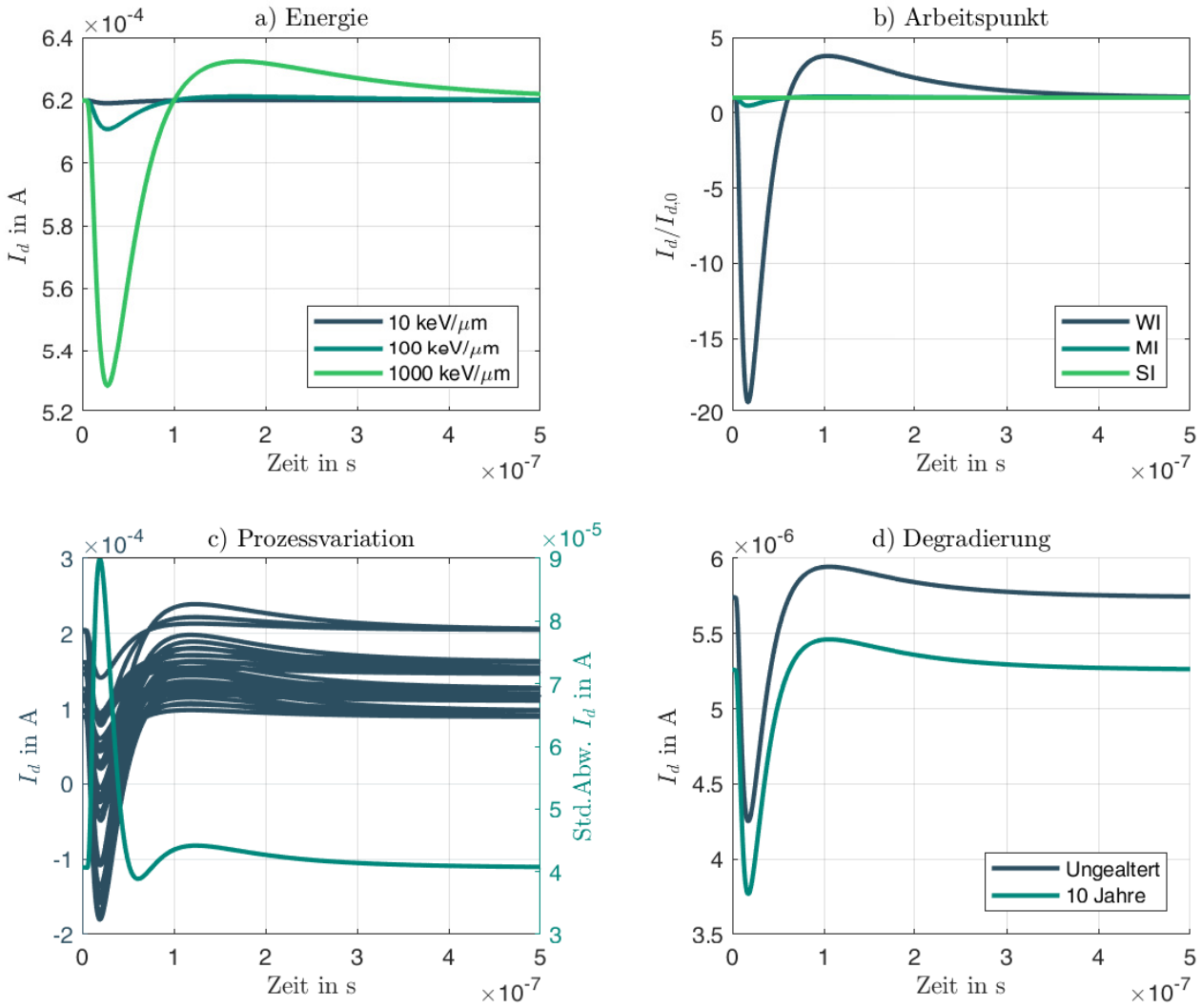


Abbildung 2.20: Einfluss auf den Drainstrom durch a) Energie des Teilchens, b) Arbeitspunkt des Transistors, c) Prozessvariation und d) Degradierung

Prozessvariationen führen nicht nur zu unterschiedlichen Ruheströmen der Transistoren sondern auch zu unterschiedlich starken Auswirkungen von SEU.

2.3.4 Stochastische Auswertung

Die Zuverlässigkeit einer integrierten Schaltung ergibt sich aus den spezifizierten Eigenschaften wie zum Beispiel der Verstärkung, der Verzögerung oder der Leistungsaufnahme. Die Eigenschaften streuen durch die Prozessvariation zum Zeitpunkt der Herstellung und weisen eine bestimmte Wahrscheinlichkeitsverteilungsdichte auf. Durch die zuvor beschriebenen Degradationsmechanismen verändern sich die Schaltungseigenschaften im Laufe der Zeit und in Abhängigkeit von äußeren Einflüssen und der Benutzung der Schaltung. Dabei ändern sich die Eigenschaften graduell und somit ihre Verteilungen. Durch diese graduelle Änderung entsteht die steigende Ausfallrate in der Verschleißphase. Dies ist in Abbildung 2.21 anhand der Verteilung einer Eigenschaft über der

Zeit gezeigt. Dabei zeigt die Gerade in der Verschleißphase die angenommene konstante Fehlerrate über diesen Bereich. Um eine Aussage über die Zuverlässigkeit einer Schaltung

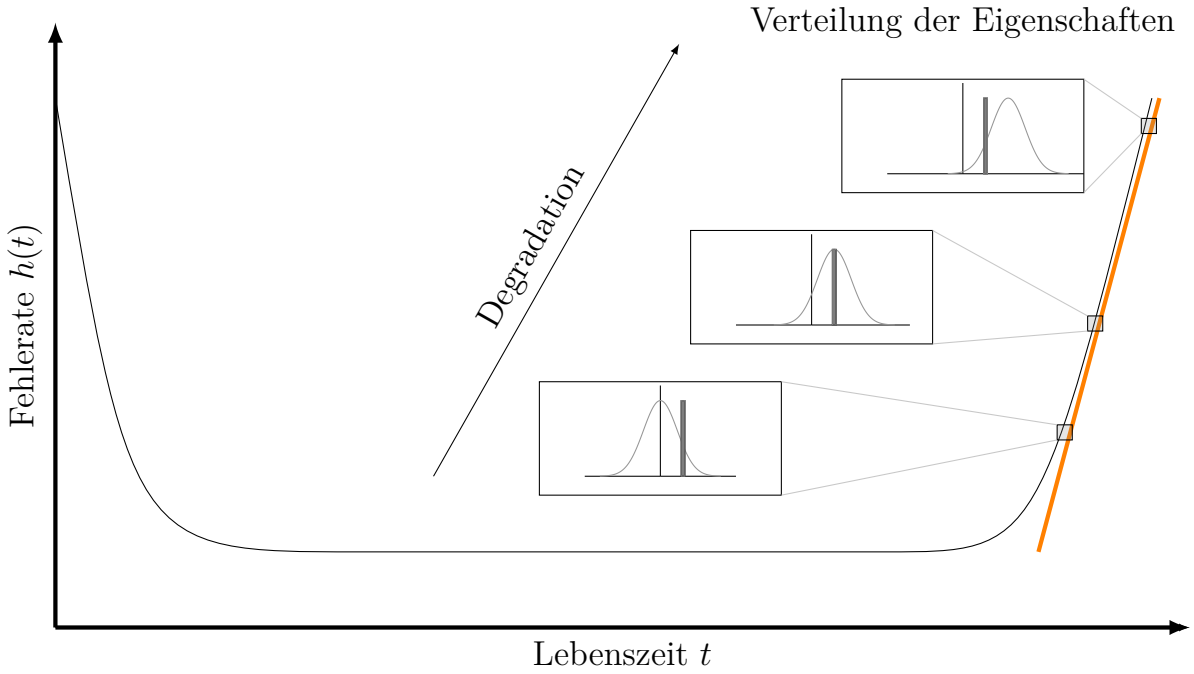


Abbildung 2.21: Verschiebung der Verteilung einer Schaltungseigenschaft über der Zeit und resultierender Anstieg der Ausfallrate durch Überschreiten der Spezifikationsgrenzen.

zu treffen, muss die Ausfallrate gemessen oder geschätzt werden. Eine Messung der Ausfallrate würde so lange dauern wie die spezifizierte Lebensdauer der Schaltung. Da dies einige Jahre sind, kann die Ausfallrate in praktischen Anwendungen nur geschätzt werden. Zur Schätzung der Ausfallrate müssen die zuvor beschriebenen Einflüsse in der Modellierung der Zuverlässigkeit berücksichtigt werden. Da die Ausfallrate zudem von äußeren Einflüssen wie der Temperatur und der Art des Betriebs abhängt, müssen diese Faktoren bei der Schätzung mit berücksichtigt werden. Die Umsetzung geschieht zumeist durch die Verwendung von sogenannten Highly-Accelerated-Lifetime-Tests (HALT) [WCJ11, McP10, Sch06]. Dabei wird ausgehend von der Arrhenius-Gleichung

$$FR = A\mathrm{e}^{\frac{E_A}{kT}} \tag{2.73}$$

ein Beschleunigungsfaktor für Messungen bei erhöhten Temperaturen bestimmt [KCK13]. A ist eine Konstante, k die Boltzmann-Konstante, T die Temperatur und E_A die Aktivierungsenergie der zu beschleunigenden Reaktion. Die Modellannahmen beinhalten, dass alle Reaktionen rein thermisch begründet sind, was für quantenmechanische Zusammenhänge nicht immer zutreffend ist. Insgesamt kann durch eine deutlich verkürzte Messung bei hohen Temperaturen die Degradation der Schaltung beschleunigt werden, um somit die Ausfallrate zu bestimmen. Typischerweise wird als Aktivierungs-

energie 0,5eV - 0,7eV angenommen, wobei die Bandbreite 0,1eV bis 1,2eV umfasst [HPR+07, GHB09, Gra13, Gra14]. Daraus folgen Beschleunigungsfaktoren von über 10.000, wodurch die Messung einer spezifizierten Lebenszeit von 10 Jahren einer Schaltung auf 21,6 Tage verkürzt werden kann. Somit kann die Fehlerrate einer Schaltung tatsächlich gemessen werden.
Nachteil dieser Methode ist, dass nicht nur ein Effekt zur Alterung beiträgt, sondern verschiedene und es sich somit um eine Vielzahl von unterschiedlichen Aktivierungsenergien handelt. Diese werden unterschiedlich stark beschleunigt und führen zu einer Art Vorhersage-Dispersion. Somit wird die Lebensdauer für Effekte mit hohen Aktivierungs-

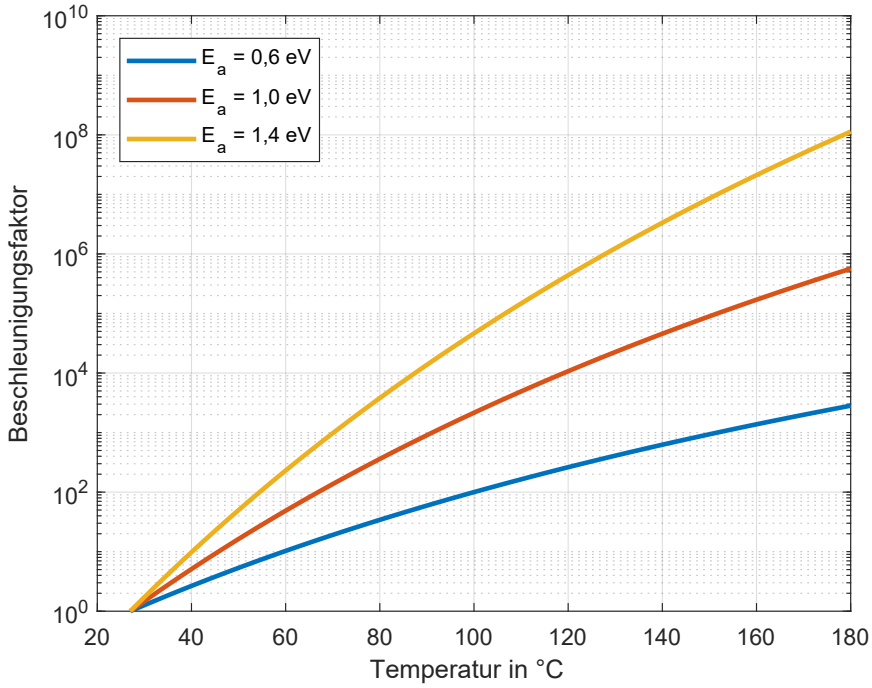


Abbildung 2.22: Beschleunigungsfaktoren für verschiedene Aktivierungsenergien E_A über der Temperatur.

energien zu stark beschleunigt, um innerhalb derselben Messung Effekte mit niedrigeren Aktivierungsenergien zu beschleunigen [GP16]. Zudem zeigen aktuelle Forschungsergebnisse, dass das Verhalten einiger Alterungsmechanismen in Halbleiterstrukturen nicht der Arrhenius-Gleichung entspricht [PFV+18]. Somit können diese Einflüsse mit herkömmlichen Messmethoden nicht korrekt berücksichtigt werden.
Fällt eine Schaltung während der Messung durch einen zu stark beschleunigten Effekt aus, verfälscht dies die Messung. Zudem werden durch sehr hohe Temperaturen von über 150°C Mechanismen und Wechselwirkungen ausgelöst, die in den spezifizierten Bedingungen niemals auftreten können. Dazu zählen zum Beispiel mechanische Spannung durch die Erwärmung und Erhöhung der Leitfähigkeit von Halbleitermaterialien und Senkung der Leitfähigkeit eines Metalls. Zudem können Ausheilungsprozesse durch diese Methode während der Betriebszyklen unzureichend in der Messung berücksichtigt werden. Eine Erweiterung dieses Modells ist das Eyring-Modell [KCK13, CTF+06],

das unterschiedliche Arten von Stressbedingungen berücksichtigen kann. Dabei werden ausgehend vom Arrhenius-Modell zusätzliche Terme hinzugefügt

$$FR = AT^{\alpha} \exp\left[\frac{E_A}{kT} + \left(B + \frac{C}{T}\right) \cdot S_1 + \left(D + \frac{E}{T}\right) \cdot S_2\right] \ . \tag{2.74}$$

Dabei sind A, B, C, D, α empirisch zu ermittelnde Modellparameter und S_1 und S_2 sind die Stressterme durch andere Einflüsse, zum Beispiel elektrische oder mechanische Spannungen. Der Nachteil dieses Modells ist die schwierige Bestimmbarkeit der Modellparameter. So müssen die Temperaturabhängigkeiten der zusätzlichen Stressbedingungen ermittelt werden. Diese beeinflussen jedoch ihrerseits die Empfindlichkeit der Schaltung gegenüber der Temperatur. Diese Wechselwirkungen sind nahezu kaum korrekt messtechnisch zu erfassen. Zudem befindet sich während der Messung somit nicht nur die Temperatur, sondern auch zum Beispiel die Versorgungsspannung außerhalb des spezifizierten Betriebsbereiches, dies hat ähnliche Auswirkungen wie eine alleinige Erhöhung der Temperatur. Trotz ihrer Nachteile wird diese Methode wegen der hohen Praktikabilität häufig verwendet und findet sich ebenfalls in einigen Messvorschriften für Degradationseffekte wieder.

Neben der Möglichkeit der beschleunigten Messung besteht die Möglichkeit, anhand physikalischer Modelle der Schaltung und der Degradationsmechanismen die Alterung zu beobachten und den weiteren Verlauf zu extrapolieren, um somit eine Aussage über die verbleibende Lebensdauer der Schaltung treffen zu können [RF15]. Diese Verfahren werden aktuell verstärkt für die Zuverlässigkeitsanalyse von Akkumulatoren verwendet [MXC+13, SG09]. Das zugrunde liegende Prinzip kann jedoch auf jedes physikalische System übertragen werden. Bei dieser Herangehensweise werden keine Annahmen oder Einschränkungen bezüglich der Degradationsmechanismen getroffen. Aufwendig bei diesem Verfahren ist hingegen die Entwicklung geeigneter Modelle. Für die Übertragung dieses Verfahrens auf integrierte Schaltungen kann jedoch auf die bestehenden Transistormodelle sowie Degradationsmodelle zurückgegriffen werden. Besonderer Vorteil ist, dass während der Betriebszeit der Schaltung die Schätzung der verbleibenden Lebensdauer (*engl. Remaining Useful Lifetime* RUL) regelmäßig aktualisiert wird [Suh11, SWHZ11]. Somit können in dieser Schätzung die tatsächlichen Belastungsprofile und Umwelteinflüsse berücksichtigt werden [BK15]. Zudem wird die Schaltung zu keinem Zeitpunkt außerhalb des erlaubten Betriebsbereiches betrieben. Ausgangspunkt der Schätzung ist die Verteilung der Schaltungseigenschaften nach der Herstellung. Anschließend werden die Veränderungen dieser Eigenschaften in regelmäßigen Abständen überprüft. Auf Grundlage der beobachteten Verschiebung des physikalischen Degradationsmodells wird der Zeitpunkt berechnet, an dem die Schaltung die Spezifikationen verlassen würde und es zu einem Ausfall käme. Ein zeitlicher Ablauf einer solchen Schätzung ist in Abbildung 2.23 gezeigt.

Dabei unterliegt die Güte der Schätzung der Anzahl der Messungen sowie der Variabilität der Degradationsmechnismen. In jedem Falle kann, gegensätzlich zu der beschleunigten

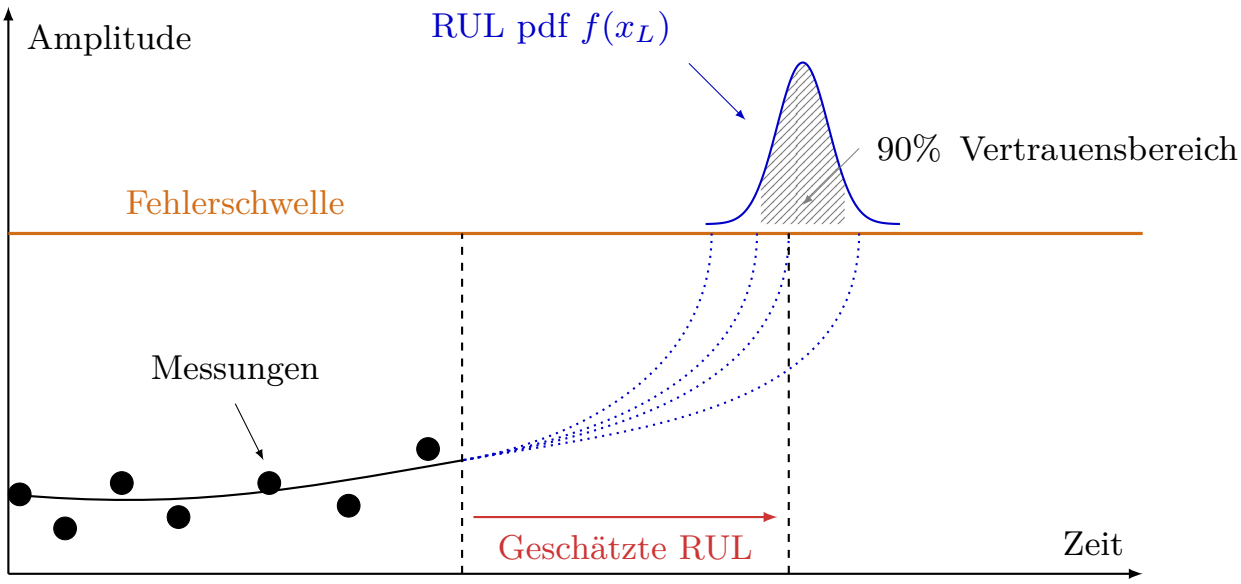


Abbildung 2.23: Schätzung der restlichen Betriebszeit einer Schaltung auf Grundlage zurückliegender Messungen.

Messung, eine Aussage über die Güte getroffen werden. Beide Arten der Bestimmung der Ausfallraten können kombiniert werden. So kann die Schaltung an den jeweiligen oberen Grenzen des Betriebsbereiches betrieben werden. Dadurch können Beschleunigungsfaktoren von einigen 100 erreicht werden. Diese Messungen werden anschließend verwendet, um die restliche Lebenszeit mittels Modellansatzes zu schätzen. Diese Schätzung wird mit jedem neuen vorhandenen Messwert aktualisiert. So können die Vorteile beider Methoden miteinander kombiniert werden: Deutlich verkürzte Messzeiträume und gleichzeitig eine bewertbare Güte der Schätzung. Die restliche Lebensdauer kann als Zufallsvariable X_L mit der Wahrscheinlichkeitsdichteverteilung x_L aufgefasst werden. Werden die vergangenen Betriebszustände der Schaltung als Y_s definiert, ergibt sich die Schätzung

$$f(x_L|Y_s) = f(x_L) = \frac{f(\tau + x_L)}{S(\tau)} \ , \tag{2.75}$$

wobei τ eine bestimmte Zeitdauer darstellt [AJ99]. Diese Wahrscheinlichkeiten werden mit Hilfe des Partikelfilteralgorithmus, der in Abschnitt 2.5.5 vorgestellt wird, ermittelt. Somit kann anhand mindestens zweier Messungen die Überlebensfunktion $S(t)$ berechnet werden, welche anschließend für die Extrapolation verwendet werden kann. Dabei ist die Varianz der Schätzung proportional zur Weite der Extrapolation sowie antiproportional zur Anzahl der Messungen. Einen schematischen Überblick über die modellbasierte Methode zeigt Abbildung 2.24.

Dabei wird das initiale Modell anhand von Schätzung oder Messdaten aus vorangegangenen Messungen parametriert. Zu jedem Messzeitpunkt wird das Modell reparametriert und somit verfeinert. Auf dessen Basis wird anhand der Spezifikationen die restliche

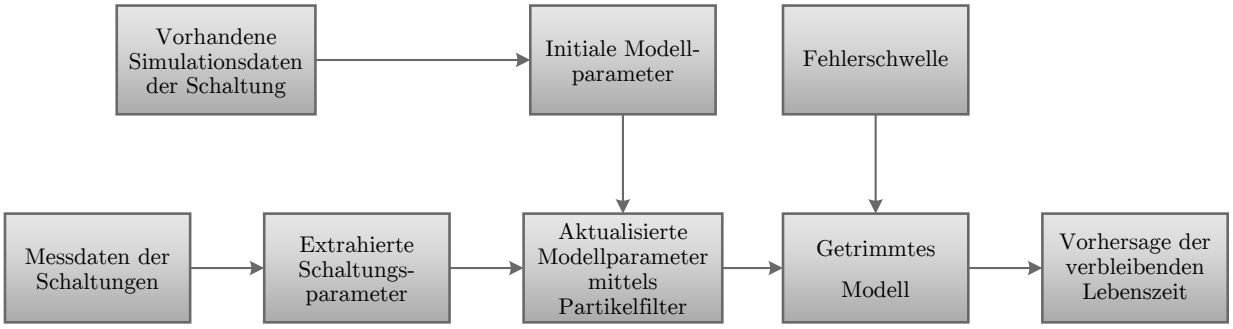


Abbildung 2.24: Blockdiagramm der Schätzung der restlichen Lebensdauer einer Schaltung auf Grundlage eines modellbasierten Ansatzes.

Lebensdauer geschätzt. Dabei können mehrere Modellparameter sowie Spezifikationen überwacht werden. Notwendige Bedingung ist, dass die relevanten Systemgrößen im Betrieb beobachtbar sind. Die einzelnen in diesem Abschnitt vorgestellten Konzepte und Methoden werden in folgenden Abschnitten näher beschrieben und im weiteren Verlauf das Vorgehen an verschiedenen Messobjekten erläutert.

2.4 Entwurfsprozess analoger integrierter Schaltungen

Um von einzelnen Bauelementen wie Widerständen und Transistoren zu Schaltungen und Systemen zu gelangen, müssen diese bereits vor ihrer Fertigung untersuchbar sein. Zentrale Aufgabe des Entwurfsprozess ist es, von den Spezifikationen auf die notwendigen Bauelemente und deren Anordnung zu gelangen. Zudem müssen Zwischenschritte validierbar und verifizierbar sein [RG12]. Wie bereits eingangs erwähnt, wird zwischen digitalen und analogen Schaltungen unterschieden, selbiges setzt sich auch bei den Entwurfsprozessen fort. Daher werden im Weiteren die beiden Prozesse getrennt beschrieben.

Bei dem Entwurfsprozess für digitale Schaltungen steht zu Beginn die mathematische Beschreibung der Funktionalität in Form der Spezifikationen. Anschließend werden die Algorithmen in eine Hardware-Beschreibungssprache übersetzt [Kae08]. Zu diesem Zeitpunkt findet der erste Abgleich mit dem Algorithmus statt, indem die Simulationsergebnisse der beiden Implementierungen verglichen werden. Falls Fehler auftauchen, können diese sofort behoben werden. Anschließend findet die Synthese der Hardware-Beschreibungssprache statt. Dabei werden die Operationen automatisiert auf Speicherelemente und Logikgatter abgebildet. Die verwendeten Elemente werden von Halbleiterfabriken als vorgefertigte Schaltungen bereitgestellt. Die verwendete Softwareumgebung berücksichtigt zudem das Timing der Schaltung und führt automatisiert mehrere Iterationen durch, um die Spezifikationen zu erfüllen. Nach diesem Schritt

können abermals Vergleiche mit den vorherigen zwei Implementierungen stattfinden. Zum Schluss findet das automatisierte Layout der Schaltung statt. Dabei werden die Speicherelemente und Logikgatter automatisiert in ein Raster gesetzt und entsprechend ihrer Funktion verdrahtet. Viele der Schritte sind somit mindestens teilautomatisiert und auf die Wiederverwendbarkeit der Ergebnisse ausgelegt. Für den Entwurf von entscheidender Bedeutung sind die Güte und der Umfang der vom Hersteller bereitgestellten Bibliotheken. Nach der Herstellung kann die gefertigte Schaltung durch automatisierte Tests auf Fehler und Abweichungen hin untersucht werden. Somit haben die Designer die Möglichkeit, mehr Zeit für die Algorithmen und die Übersetzung in eine Hardwarebeschreibungssprache zu investieren.

Im Gegensatz zum digitalen Entwurfsablauf gibt es im Analogen kaum automatisierte Schritte. So muss der Designer von den Spezifikationen auf das gewünschte Eingangs- und Ausgangsverhalten der Schaltung schließen. Dies kann zum Beispiel ein bestimmter Verstärkungsfaktor oder Referenzspannung sein. Für die Realisierung desselben gewünschten Ausgangsverhaltens existieren verschiedene innere Aufbauten von einem Verstärker oder einer anderen Schaltung. Abbildung 2.25 zeigt die Abstraktionsebenen und Entwurfswerkzeuge in einem Entwurfsprozess für analoge Schaltungen. Zudem sind die Fehlermodelle auf den verschiedenen Ebenen gezeigt. Ein wichtiger Punkt für den Entwurf analoger Schaltungen ist das Propagieren der Spezifikationen durch die unterschiedlichen Ebenen.

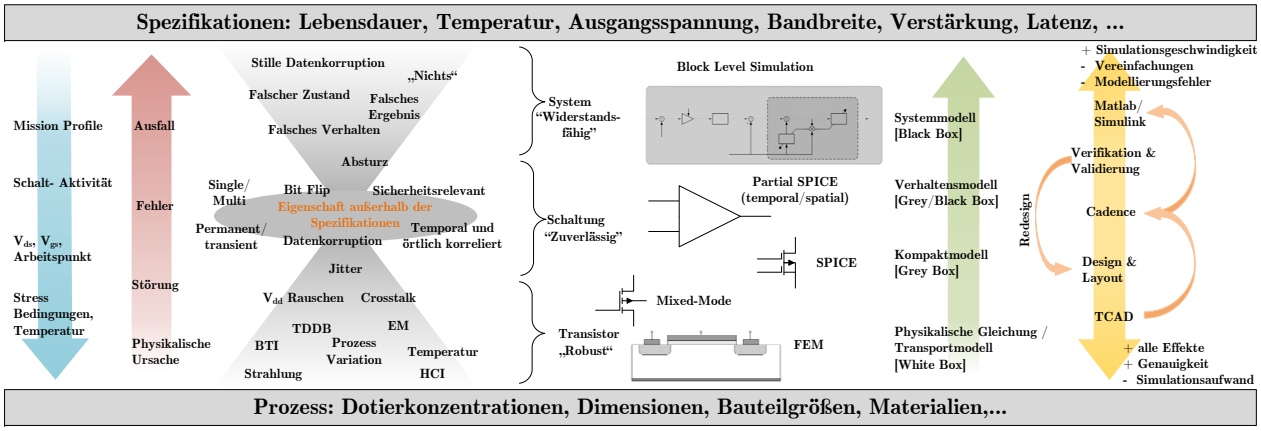


Abbildung 2.25: Abstraktionsebenen und Entwurfswerkzeuge in einem Entwurfsprozess für analoge Schaltungen.

Die einzelnen Entwurfsschritte für analoge Schaltungen lassen sich grob in vier Stufen gliedern [Hab13]:

1. *Auswahl der Schaltungstopologie:* Dabei ist die Topologie eine strukturelle Anordnung von Schaltungelementen mit mindestens zwei Anschlüssen. Für die einzelnen Elemente müssen zudem Modelle für analytische oder numerische Simulationen vorhanden sein.

2. *Größenbestimmung der Schaltung:* Festlegung der einzelnen Bauteilgrößen in Abhängigkeit geforderter Eigenschaften auf Basis der zuvor gewählten Topologie.

3. *Layoutsynthese der Schaltung:* Geometrische Anordnung der tatsächlich zu fertigen Formen zusammen mit der notwendigen Verdrahtung.

4. *Post-Layout Extraktion und elektrische Verifikation:* Um als parasitäre Bauelemente modellierte physikalische Effekte erweiterte simulierbare Schaltung zur Verifikation der Funktionalität nach der Herstellung.

Jeder dieser Schritte engt die zum Design verfügbaren Freiheitsgrade weiter ein, dabei sind diese zum Zeitpunkt der Auswahl der Schaltungstopologie am größten. Diese Arbeit fokussiert sich daher auf eben diesen ersten Schritt eines jeden analogen Designablaufs. Alle weiteren Aufgaben sind bereits ohne die Betrachtung von Zuverlässigkeit für sich herausfordernde Aufgaben, für die eine Vielzahl von Arbeiten existieren [Gra07, Wil06, Bak08, SGM10, DW12].
Zur Schließung dieser analogen Lücke wurden verschiedene Herangehensweisen entwickelt. Eine Möglichkeit ist es, komplexere Schaltungen in modulare Subblöcke [Nic15, JM17b, SK08, HRC89, MH15] mit einer dedizierten Funktion innerhalb einer größeren Schaltung zu unterteilen. Diese Subblöcke müssen einmal entworfen werden und können anschließend in diversen Schaltungen mehrfach verwendet werden. So muss der Entwurfsprozess nicht von vorn beginnen. Eine andere Methode ist die sogenannte g_m/I_d-Methode [Bin07]. Diese basiert auf der Transkonduktanzeffizienz eines Transistors, dessen Arbeitspunkt durch das Tupel aus Länge, Inversionskoeffizient und Drainstrom bestimmt ist. Anhand von, im Vorhinein simulierten und in Look-Up-Tabellen gespeicherten, Werten werden optimale Weiten der Transistoren bestimmt [MMD15, TCSB16, SFJ96, HHPDP13]. Anstelle von Look-Up-Tabellen können ebenfalls Response Surface Modelle (RSM) [THT+16, THPPD17, HHPPD15, MMB+06] verwendet werden. Zudem kann der gesamte Ablauf mit Tools [HHE+15, SHH+17] automatisiert werden. Die Kanallänge L muss, falls diese nicht selbst durch Spezifikationen beschränkt ist, zuvor als übergeordneter Parameter vom Designer durch Flächen-, Geschwindigkeits- und Energieabwägungen gewählt werden. Diese Methode funktioniert vor allem für solche Schaltungen, bei denen Verstärkungseigenschaften im Vordergrund stehen. Die dritte Möglichkeit ist die Verwendung von Kleinsignalgleichungen für die Schaltungseigenschaften, um aus diesen die Parameter einzelner Transistoren zu gewinnen. Wird diese Methode simulationsbasiert durchgeführt, wird auch vom SPICE-Monkey-Flow gesprochen da dies meist zu einer Vielzahl von notwendigen Simulationen und Iterationen im Entwurfsprozess führt [Mao13, GRLVAI16, MG10]. Eine deutlich strukturiertere Herangehensweise verwendet anstelle von initialen Simulationen analytische Transistorgleichungen, um die Transistorgrößen anzunähern. Das dort auftretende Problem ist die unzureichende Modellierung moderner integrierter Transistoren durch simple analytische Gleichungen. Mit Hilfe des EKV Modells ist es jedoch möglich, für häufige Betriebsbereiche (Sättigung, moderate und starke Inversion) das Bauteilverhalten durch einfache Gleichungen

gut abzubilden [ECP17a, ECP17b, Bin07]. Die g_m/I_d-Methode kann ebenfalls mit den EKV-Modellgleichungen kombiniert werden [HHPDP13, HHT+14, TFMLCL+17]. Ein generelles Problem dieser Erfassung sind Spezifikationen $\partial g_m/\partial V_{dd}$ mit $\partial V_{dd}/\partial t \neq 0$, die auf transienten Schaltungseigenschaften wie dem Startup-Verhalten beruhen. Die Funktion der Startup-Schaltungen beruht darauf, dass in der Regel zum Einschaltzeitpunkt ein leitender Pfad zwischen internen Knoten der funktionalen Schaltung und einem äußeren Potential (zum Beispiel V_{dd}) besteht. Dieser Pfad soll, sobald die Schaltung ihren Arbeitspunkt erreicht hat, durch eine Rückkopplung in der Startup-Schaltung, unterbrochen werden. Dieses setzt voraus, dass einzelne Transistoren während des Anstiegs der Versorgungsspannung selbst einen ausreichend hohen Strom leiten, da ansonsten die Startup-Schaltung versagt. Diese Anforderung beinhaltet ein notwendiges Verhalten in Bezug auf $\partial g_m/\partial V_{dd}$. Wird der spezifizierte Strom zum Einschaltzeitpunkt durch eine geringe Gate-Spannung-Drain-Strom-Steuerung bei steigender Versorgungsspannung nicht erreicht, so startet die Schaltung mit Verzögerung oder überhaupt nicht. Diese Auswirkungen können in den Kleinsignalgleichungen kaum dargestellt werden, da diese nur für einen Arbeitspunkt mit geringer Auslenkung gültig sind. Um dieses Verhalten zu erfassen müssten die resultierenden Arbeitspunkte während steigender Versorgungsspannung neu berechnet werden. Dies würde einer transienten SPICE oder DC sweep Analyse entsprechen. Somit wäre der Simulationsaufwand enorm und die Vorteile der Methoden deutlich reduziert.

2.4.1 Festlegung des nominalen Arbeitspunktes

Für die Festlegung des nominalen Arbeitspunktes sind die Prozessvariationen von entscheidender Bedeutung. Durch ihren Einfluss variieren alle im Entwurfsprozess verwendeten Größen und bedingen sich zum Teil gegenseitig [MAD+17]. Dazu zeigt Abbildung 2.26 die Standardabweichung des Drainstroms aus TCAD Simulationen, durch Prozessvariation für N- und PMOS FDSOI Transistoren über der Gate-Spannung für verschiedene Betriebszustände und Gate-Längen. Zu erkennen ist, dass die Variation des Drainstroms für $V_g \approx V_{th}$ am höchsten ist. In Sättigung bildet dieser Bereich das globale Maximum. Ebendieser Betriebspunkt der moderaten Inversion in Sättigung ist nach der g_m/I_d-Methode zusammen mit den EKV-Modellgleichungen der optimale Arbeitspunkt. Somit entsteht bereits nur durch Prozessvariationen die Möglichkeit der Änderung des Arbeitspunkts einzelner Transistoren. Neben der Erfüllung der Spezifikationen dürfen die Prozessvariationen, und insbesondere deren Wechselwirkungen, in den jeweiligen Arbeitspunkten der Transistoren nicht dazu führen, dass die Ausbeute der gesamten Schaltung darunter leidet. Einfache Maßnahmen, um den Prozessvariationen entgegenzuwirken, leiten sich direkt aus ihren Beschreibungen ab (vgl. Abschnitt 2.3.1.1). So kann der Einfluss durch Vergrößerung der Transistoren sowie durch Erhöhung des Drainstroms verringert werden. Beide Maßnahmen haben starken Einfluss auf die Effizienz des Designs, daher ist es stets besser, die Auswirkungen der Prozessvariationen durch

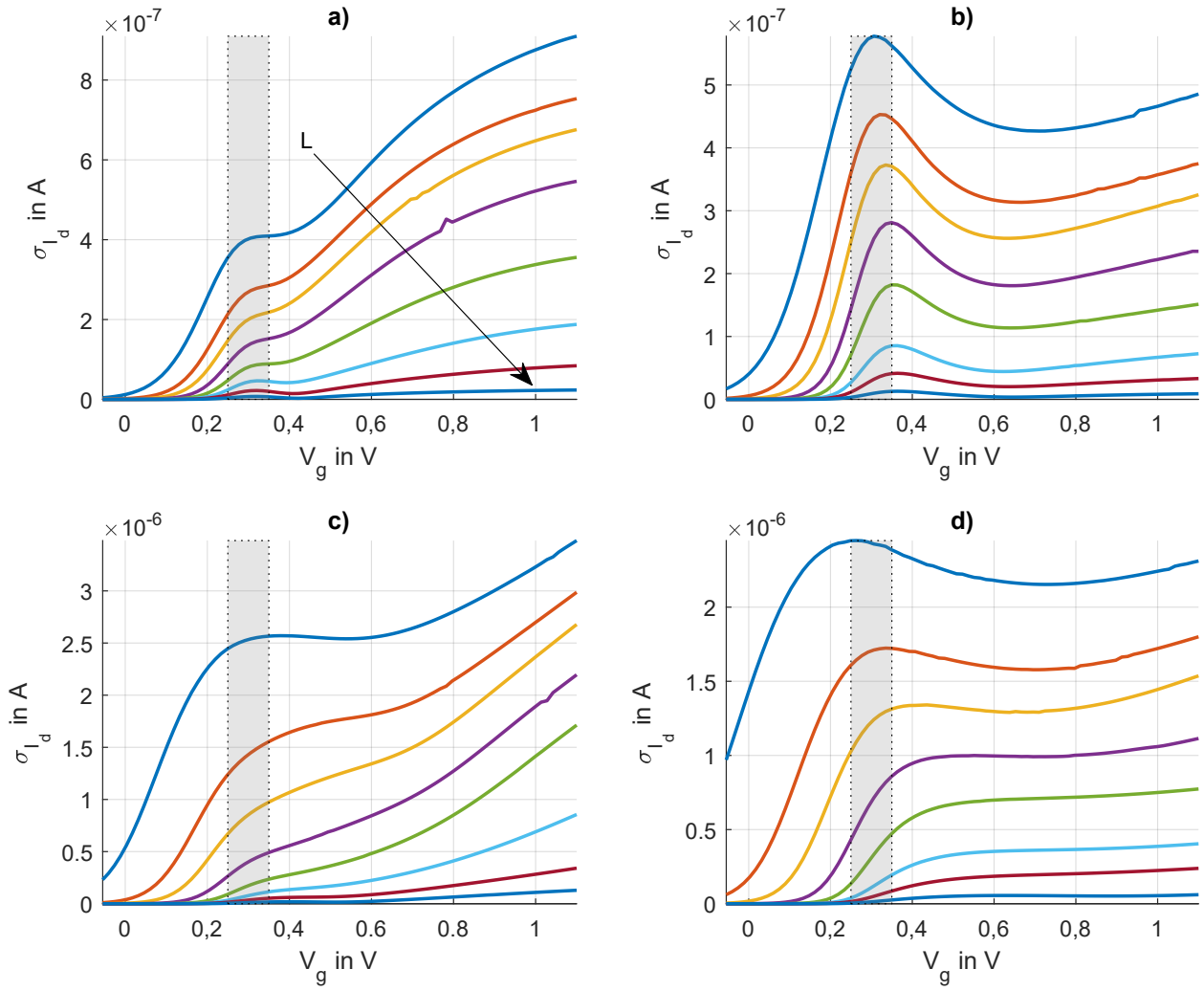


Abbildung 2.26: Variation des Drainstroms durch Prozessvariation über der Gate-Spannung für 28nm FDSOI Transistoren für verschiedene Gate-Längen. NMOS im a) linearen Bereich c) Sättigung sowie PMOS im b) linearen Bereich d) Sättigung. Die grau hervorgehobenen Bereiche zeigen die Lage der jeweiligen Schwellspannung.

die Wahl des korrekten Arbeitspunktes zu minimieren. Gleichzeitig ist aus Abbildung 2.26 zu erkennen, dass für kleinere Gate-Längen die Variationen stark ansteigen. Dies hat vor allem Auswirkungen auf sehr schnelle Schaltungen und solche mit geringer Versorgungsspannung, die ihrerseits geringe Schwellspannungen benötigen. Insgesamt ist festzustellen, dass die effizienten Arbeitspunkte eben solche sind, die besonders anfällig für Prozessvariationen sind.

2.4.2 Berücksichtigung der Zuverlässigkeit

Wie bereits beschrieben, müssen neben den Prozessvariationen zusätzlich die Alterungseffekte bei dem Schaltungsentwurf berücksichtigt werden [PMP14a, MHJ+16, MG13,

McP10]. Aus den Stressbedingungen der Alterung wird klar, dass hohe Gate- und Drainspannungen zu erhöhten Einflüssen durch Alterung führen [AKGH16]. Dies sind die Arbeitspunkte Sättigung und starke Inversion. Somit führt die Wahl des Arbeitspunktes zu einer möglichen Verstärkung einzelner Effekte, die die Zuverlässigkeit einer Schaltung beeinflussen können. Für ein optimales Design müssen daher im gesamten Entwurfsprozess stets alle Einflüsse betrachtet werden [JLK14]. Ansonsten kann es passieren, dass gewählte Arbeitspunkte sehr anfällig für einzelne Einflüsse sind. Als Beispiel ist das Verstärkungs-Bandbreite-Produkt (engl. Gain-Bandwidth-Product, GBW) über der Flankensteilheit (*engl. Slew Rate*) eines Verstärkers für drei Alterungszustände in Abbildung 2.27 gezeigt. Zu erkennen ist, dass sich der Mittelwert beider Größen stark verschiebt. Des Weiteren nimmt die Standardabweichung der Slew Rate zuerst ab und danach wieder zu und die des GBW nimmt über das Alter ab. Diese Zusammenhänge

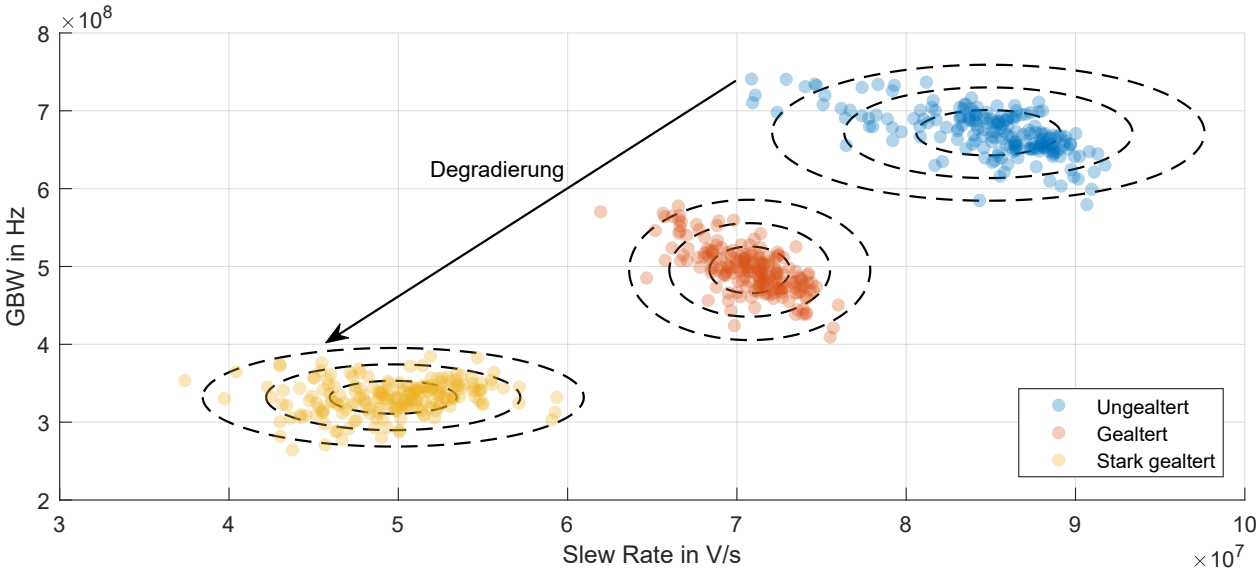


Abbildung 2.27: Gain-Bandwidth-Product über der Slew Rate eines Verstärkers für drei Alterungszustände.

erschweren den Entwurf von Schaltungen enorm, da sich diese mit jeder Änderung an der Schaltung ändern können und zudem von den tatsächlichen Betriebszuständen der Schaltung im betrachteten Zeitraum abhängen. So ist eine Worst-Case Abschätzung der Alterung ebenso wie bei Prozessvariationen wenig sinnvoll und führt zu nicht optimalen Entwurfsergebnissen. Analog zu den Cornern eines Prozesses liegen die Worst-Case Belastungsprofile einer Schaltung über 10 Jahre deutlich außerhalb den 6σ-Bereiches. Somit führen diese einzelnen Aspekte zu der Notwendigkeit, jeden einzelnen Entwurfsschritt mit unterschiedlichen und realistischen Belastungsprofilen zu evaluieren. Nur somit kann eine alterungs- und prozessvariationsabhängige Entwurfszentrierung durchgeführt werden. Die Grundidee dieses Ansatzes ist in Abbildung 2.28 gezeigt. Dabei wird der jeweilige minimale gemäß der statistischen Parameterverteilung gewichtete Abstand $S(t)$ zwischen dem Mittelwert der Verteilung und der Spezifikationsgrenze

betrachtet. Zum Zeitpunkt t_0 direkt nach der Fertigung der Schaltung wird durch $S(t_0)$

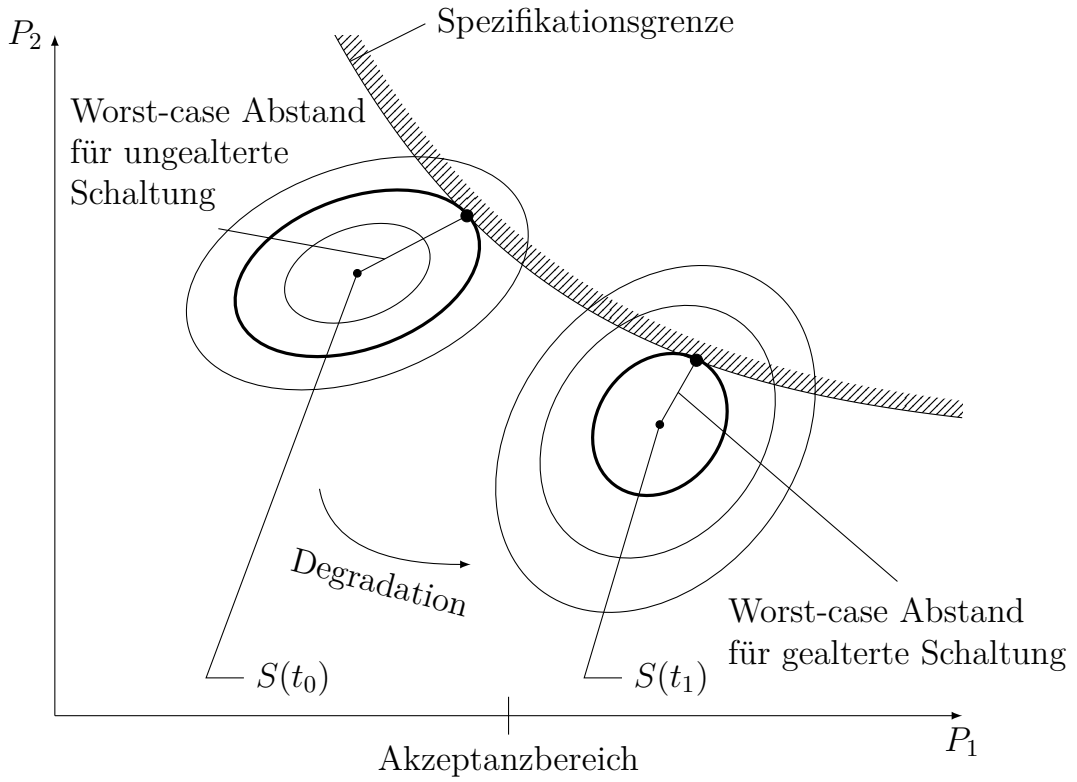


Abbildung 2.28: Entwurfszentrierung unter Berücksichtigung von Prozessvariationen und Alterung (nach [PG11]).

die Ausbeute bestimmt. Wie bereits in Abbildung 2.27 gezeigt, wandert die Verteilung durch Alterungseinflüsse. Zum Zeitpunkt t_1 kann mit $S(t_1)$ die Überlebensfunktion $S(t)$ des Designs bestimmt werden. $S(t_1)$ hängt dabei von der Alterung, den Prozessvariationen und dem Belastungsprofil ab. Daher kann diese Größe verwendet werden, um die Optimalität eines Entwurfs bezüglich der Zuverlässigkeit zu bewerten. Daraus folgt gleichzeitig, dass die Optimalitätskriterien in anderen Methoden nicht dafür geeignet sind, zuverlässige Schaltungen zu entwerfen. Jedoch können die Kriterien wie der Inversionskoeffizient und die Transkonduktanzeffizienz als von Prozessvariationen und Alterung abhängig betrachtet werden, dies wird in Abschnitt 3.2.1 näher erörtert. Die eigentlich analytischen Gleichungen müssen dadurch numerisch gelöst werden, jedoch findet so eine Verbindung aller Ansätze statt.

Eine Erweiterung des modularen Subsystemansatzes um diese Analyse- und Bewertungsmethode ist ebenso möglich. Dies ermöglicht die Konstruktion komplexer zuverlässiger Schaltungen aus zuverlässigen Subsystemen. Wie bereits beschrieben steigt damit die Wiederverwendbarkeit und der Entwurfsaufwand sinkt.

2.5 Zustandsschätzung von Systemen

Für die Zustandsschätzung von System existieren verschiedene Ansätze in der Literatur. Die am häufigsten verwendeten Methoden werden in diesem Abschnitt mit ihren jeweiligen wichtigsten Grundlagen vorgestellt. Zudem werden die Vor- und Nachteile der einzelnen Algorithmen im Kontext dieser Arbeit diskutiert. Anschließend werden die in dieser Arbeit verwendeten Partikelfilter vorgestellt und anhand von Beispielen evaluiert.
Im Rahmen dieser Arbeit soll die Zustandsschätzung von System dazu verwendet werden, die Messergebnisse von gefertigten Bauelementen und Schaltungen zu verbessern und automatisiert ein Modell von diesen zu erstellen. Diese Modelle können ohne Weiteres für Schaltungssimulationen verwendet werden. Des Weiteren wird die gewählte Methode so angepasst, dass eine Beobachtung des Schaltungs- oder Systemzustands im Rahmen von Langzeitmessungen deutlich vereinfacht ist. Diese Langzeitmessungen treten im Zusammenhang mit der Analyse von Alterungsprozessen auf und sind daher von besonderem Interesse.

2.5.1 Systembeschreibung

Viele technische, naturwissenschaftliche oder wirtschaftliche Systeme lassen sich durch eine Zustandsraumdarstellung beschreiben. Ist bei einer solchen Beschreibung der Zustand des Systems sowie alle bisherigen und zukünftigen Eingangswerte bekannt, können sämtliche Ausgangsgrößen berechnet werden. In einer Zustandsraumdarstellung kann allgemein zwischen einem linearen und einen nichtlinearen Teil unterschieden werden. In der Praxis auftretende Systeme sind zumeist nichtlinear, können jedoch, zumindest abschnittsweise, linear angenähert werden. Die Zustandsraumbeschreibung für ein kontinuierliches, deterministisches lineares System wird durch zwei Gleichungen bestimmt:

$$\dot{x} = Ax + Bu \tag{2.76}$$

$$y = Cx \tag{2.77}$$

dabei ist x der Zustandsvektor, u der Kontrollvektor und y der Ausgangsvektor. Die Matrizen A, B und C sind entsprechend der Vektoren dimensioniert und beschreiben das Übertragungs- und Ausgangsverhalten. So wird A Systemmatrix, B Eingangsmatrix und C Ausgangsmatrix genannt. Die Einträge der Matrizen können zeitlich variieren, ihre Dimensionen und somit die Art der Systembeschreibung wird als zeitlich invariant angenommen. Wie zuvor beschrieben, sind die meisten Systeme tatsächlich nichtlinear. Eine häufige Art der Nichtlinearität sind Sättigungseffekte. Diese treten in technischen Systemen für alle physikalischen Größen auf. Ist das Verhalten des Systems für kleine Eingangsgrößen linear, können für hohe Eingangswerte Sättigungseffekte auftreten, die

die Linearität des Systems, zumindest abschnittsweise, stören. Ein nichtlineares System kann durch

$$\dot{x} = f(x, u, w) \tag{2.78}$$
$$y = h(x, v) \tag{2.79}$$

beschrieben werden. $f(\cdot)$ und $h(\cdot)$ sind beliebige vektorwertige Funktionen. Als Annahme für eine, nicht rein mathematische, Systembetrachtung ist w das Prozessrauschen und v das Messrauschen. Um im Anschluss mit nichtlinearen Systemen umgehen zu können, werden die Systeme um einen bestimmten Punkt, häufig der jeweilige Arbeitspunkt des Systems, durch eine Taylorreihe [BHL+12] linearisiert. Sei $\bar{x}$ der Zustandsvektor im Arbeitspunkt, so kann die Funktion f um diesen Punkt mit

$$f(x) = f(\bar{x}) + \left.\frac{\partial f}{\partial x}\right|_{\bar{x}} (x - \bar{x}) + \frac{1}{2\,!} \left.\frac{\partial^2 f}{\partial x^2}\right|_{\bar{x}} (x - \bar{x})^2 + \frac{1}{3\,!} \left.\frac{\partial^3 f}{\partial x^3}\right|_{\bar{x}} (x - \bar{x})^3 + \dots \tag{2.80}$$

angenähert werden, wobei sich für eine endliche Anzahl von Summengliedern $\tilde{f}(x) \approx f(x)$ ergibt. Somit kann das nichtlineare System um die Vektoren $\bar{x}$, $\bar{u}$ und $\bar{w}$ als

$$\dot{\tilde{x}} = A\tilde{x} + B\tilde{x} + L\tilde{w} \tag{2.81}$$
$$\tilde{y} = X\tilde{x} + Dv \tag{2.82}$$

dargestellt werden. L beschreibt dabei die Rauschübertragung des Prozesses. Damit werden nichtlineare Systeme als linearisierte Systeme in einem Arbeitspunkt angenähert. Vorausgesetzt diese Funktion ist in diesem Punkt differenzierbar.

Für die weitere Betrachtung sind weitere grundlegende Eigenschaften eines Systems von Bedeutung. Dazu zählen die Stabilität, die Steuerbarkeit und die Bobachtbarkeit. Diese drei zentralen Begriffe werden im Folgenden eingeführt und definiert.

Stabilität Unter der Annahme eines linearen, kontinuierlichen, zeitinvarianten Systems ohne Eingangsgrößen ist dieses System grenzstabil, falls alle Zustände $x(t)$ für alle Zeiten t und begrenzte Anfangszustände $x(0)$ begrenzt sind. Dieses System heißt asymptotisch stabil, falls für alle begrenzten Anfangszustände

$$\lim_{t \to \infty} x(t) = 0 \tag{2.83}$$

gilt. Dabei schließt asymptotische Stabilität Grenzstabilität ein. Des Weiteren können die Eigenwerte der Matrix A für eine Stabilitätsaussage verwendet werden. Das beschriebene System ist asymptotisch stabil, falls alle Eigenwerte einen negativen Realteil besitzen und grenzstabil, falls dieser Realteil negativ oder null ist. Zudem müssen für Grenzstabilität die Eigenwerte einfach auf der imaginären Achse vorhanden sein.

Für technische Systeme kann zudem die BIBO-Stabilität (Bounded-Input,Bounded-Output) verwendet werden. Dies bedeutet, dass ein System, diskret oder kontinuierlich, auf ein beschränktes Eingangssignal ein beschränktes Ausgangssignal liefert. Dies trifft für alle technischen Systeme ohne interne Energiequellen zu.

Steuerbarkeit Damit ein kontinuierliches System steuerbar ist, muss für jeden Anfangszustand $x(0)$ und jeden Endzeitpunkt $t > 0$ eine Steuergröße existieren, die den Zustand x zu einem beliebigen Wert ändern kann.

Beobachtbarkeit Ein kontinuierliches System heißt beobachtbar, falls für jeden Anfangszustand $x(0)$ und jeden Endzeitpunkt $t > 0$ der Anfangszustand $x(0)$ eindeutig durch die Kenntnis vom Eingang $u(\tau)$ und Ausgang $y(\tau)$ für alle $\tau \in [0, t]$ bestimmt werden kann.

2.5.2 Wahrscheinlichkeiten und Stochastik

In diesem Abschnitt werden die später verwendeten Konzepte und Definitionen für Wahrscheinlichkeit und Stochastik beschrieben.

2.5.2.1 Wahrscheinlichkeiten

Die Wahrscheinlichkeit für ein Ereignis A sei definiert als:

$$P(A) = \frac{\text{Anzahl des Auftretens von } A}{\text{Gesamtzahl der Ergebnisse}} \tag{2.84}$$

Diese Definition wird relative Häufigkeit genannt. Neben dieser existieren noch mathematisch genauere Definitionen [Dev11, LXF+17], die jedoch für diese Arbeit keinen Mehrwert haben. Die bedingte Wahrscheinlichkeit, dass das Ereignis A unter der Bedingung des Auftretens B eintritt ist durch

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.85}$$

gegeben, mit $P(B) \neq 0$. Dabei ist $P(A|B)$ die bedingte Wahrscheinlichkeit und $P(A \cap B)$ die vereinigte Wahrscheinlichkeit beider Ereignisse. Durch das Einsetzen dieser Beziehungen in Gleichung (2.85) folgt der Satz von Bayes

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \ . \tag{2.86}$$

Die Wahrscheinlichkeiten für Einzelereignisse ($P(A)$, $P(B)$) heißen *a priori* Wahrscheinlichkeiten, da diese die Wahrscheinlichkeit eines Ereignisses ohne vorher bekannte Informationen angeben. Bedingte Wahrscheinlichkeiten ($P(A|B)$, $P(B|A)$) werden *a*

posteriori genannt, da Informationen über das zusammenhänge Ereignis bekannt sein müssen. Eine weitere wichtige Eigenschaft für die Berechnung von Wahrscheinlichkeiten ist die statistische Unabhängigkeit zweier Wahrscheinlichkeiten. Diese kann durch die folgenden drei Gleichungen auf verschiedene Weisen ausgedrückt werden:

$$P(A \cap B) = P(A) \cdot P(B) \tag{2.87}$$

$$P(A|B) = P(A) \tag{2.88}$$

$$P(B|A) = P(B) \tag{2.89}$$

2.5.2.2 Zufallsvariablen

Eine Zufallsvariable ist die funktionale Abbildung von möglichen Ereignissen auf reelle Zahlen [PP02]. Dabei ist das Ereignis selbst nicht die Zufallsvariable, sondern diese repräsentiert die Eigenschaften des Experiments. Eine Zufallsvariable kann diskret oder kontinuierlich sein, ihre grundlegende Eigenschaft der Zufallsvariable X ist ihre Wahrscheinlichkeitsverteilungsfunktion (PDF) $F_X(x)$, welche als

$$F_X(x) = P(X \leq x) \tag{2.90}$$

definiert ist, dabei ist x eine unabhängige nicht-Zufallsvariable oder Konstante. Zentrale Eigenschaften der PDF sind

$$F_X(x) \in [0, 1] \tag{2.91}$$

$$F_X(-\infty) = 0 \tag{2.92}$$

$$F_X(\infty) = 1 \tag{2.93}$$

$$F_X(a) \leq F_X(b) \quad \text{falls } a \leq b \tag{2.94}$$

$$P(a \leq x \leq b) = F_X(b) - F_X(a) \tag{2.95}$$

Des Weiteren sei die Wahrscheinlichkeitsdichtefunktion (pdf) definiert als

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x} \ . \tag{2.96}$$

Aus den Eigenschaften der PDF und der Definition der pdf ergeben sich folgende Eigenschaften:

$$F_X(x) = \int_{-\infty}^{x} f_X(z)\,\mathrm{d}z \tag{2.97}$$

$$f_X(x) \geq 0 \tag{2.98}$$

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1 \tag{2.99}$$

$$P(a < x \leq b) = \int_a^b f_X(x)\,\mathrm{d}x \tag{2.100}$$

Mit dem Satz von Bayes aus Gleichung (2.85) folgt somit für die bedingten Wahrscheinlichkeiten:

$$F_X(x|A) = \frac{P(X \leq x \cap A)}{P(A)} \ . \tag{2.101}$$

Dies kann für zwei Zufallsvariablen X_1 und X_2 geschrieben werden als

$$f_{X_1|X_2}(x_1|x_2) = P\left[(X_1 \leq x_1)|(X_2 = x_2)\right] \tag{2.102}$$

$$= \frac{f_{X_1|X_2}(x_1 \cap x_2)}{f_{X_2}(x_2)} \ . \tag{2.103}$$

Der Satz von Bayes kann für weitere Zufallsvariablen erweitert werden. Die Chapman-Kolmogorov Gleichung, welche als Grundlage der Bayes'schen Zustandsschätzung dient [PP02], ist folgend beispielhaft gezeigt:

$$f[x_1|(x_2 \cap x_3 \cap x_4)] \cdot f[(x_2 \cap x_3)|x_4] = \frac{f(x_1 \cap x_2 \cap x_3 \cap x_4)}{f(x_2 \cap x_3 \cap x_4)} \cdot \frac{f(x_2 \cap x_3 \cap x_4)}{f(x_4)} \tag{2.104}$$

$$= \frac{f(x_1 \cap x_2 \cap x_3 \cap x_4)}{f(x_4)} \tag{2.105}$$

$$= f[(x_1 \cap x_2 \cap x_3)|x_4] \ . \tag{2.106}$$

Diese Gleichung kann für eine beliebige Anzahl von Zufallsvariablen erweitert werden.

2.5.2.3 Statistische Momente

Der Erwartungswert oder Mittelwert (μ) für die Anzahl der Wiederholung $N \to \infty$, einer Zufallsvariable X ist definiert als

$$E(X) = \frac{1}{N} \sum_{i=1}^{N} A_i n_i \ . \tag{2.107}$$

Dabei ist m die Anzahl unterschiedlicher Ereignisse A_i und n_i deren jeweilige Auftrittsanzahl. Mit der Defintion der pdf ergibt sich somit

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x \ . \tag{2.108}$$

Die Standardabweichung σ bzw. die Varianz σ^2 einer Zufallsvariable gibt an, wie stark diese um ihren Mittelwert streut. Dabei ist

$$\sigma_X^2 = E[(X - E(X))^2] \tag{2.109}$$

$$= \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x)\,\mathrm{d}x \ . \tag{2.110}$$

Diese Gleichung kann zu $\sigma^2 = E(X^2) - E(X)^2$ umgestellt werden. Wobei $E(X^2) = E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x$ ist. Somit kann eine Zufallsvariable anhand ihres Mittelwerts und Varianz geschrieben werden als

$$X \sim (\mu, \sigma^2) \ . \tag{2.111}$$

2.5.2.4 Statistische Unabhängigkeit

Als zentrale Eigenschaften zweier Zufallsvariablen und ihrer Verteilungsfunktionen wird die statistische Unabhängigkeit in diesem Abschnitt näher beschrieben. Zwei Zufallsvariablen X und Y werden unabhängig genannt, falls gilt:

$$P(X \leq x \cap Y \leq y) = P(X \leq x) \cdot P(Y \leq y) \quad \forall\, x, y \ . \tag{2.112}$$

Für die gemeinsame Wahrscheinlichkeitsverteilungsfunktion und Wahrscheinlichkeitsdichtefunktion folgt daraus, dass

$$F_{xy}(x, y) = F_X(x) F_Y(y) \tag{2.113}$$

$$f_{xy}(x, y) = f_X(x) f_Y(y) \ . \tag{2.114}$$

Als zentrales Maß der linearen Abhängigkeit zweier Zufallsvariablen wird der Korrelationskoeffizient verwendet, welcher als

$$\rho = \frac{C_{XY}}{\sigma_x \sigma_y} \tag{2.115}$$

definiert ist. Darin ist $C_{XY} = E(XY) - E(X) \cdot E(Y)$ die Kovarianz. Sind die Variablen unabhängig bzw. unkorreliert, ist $\rho = 0$, bei rein linearer Abhängigkeit gilt $\rho = \pm 1$. Die Korrelation zweier Variablen darf dabei nicht als Maß für funktionale Abhängigkeit der Variablen verstanden werden. Besteht ein nichtlinearer Zusammenhang zwischen den Variablen ist der Korrelationskoeffizient ohne Aussage. Dieser Umstand ist in Abbildung 2.29 durch die beispielhafte Berechnung des Korrelationskoeffizienten für verschiedene lineare und nichtlineare Zusammenhänge dargestellt. Wie in der untersten Zeile gezeigt, gilt hier $\rho = 0$, obwohl der jeweilige funktionale Zusammenhang klar ersichtlich ist.

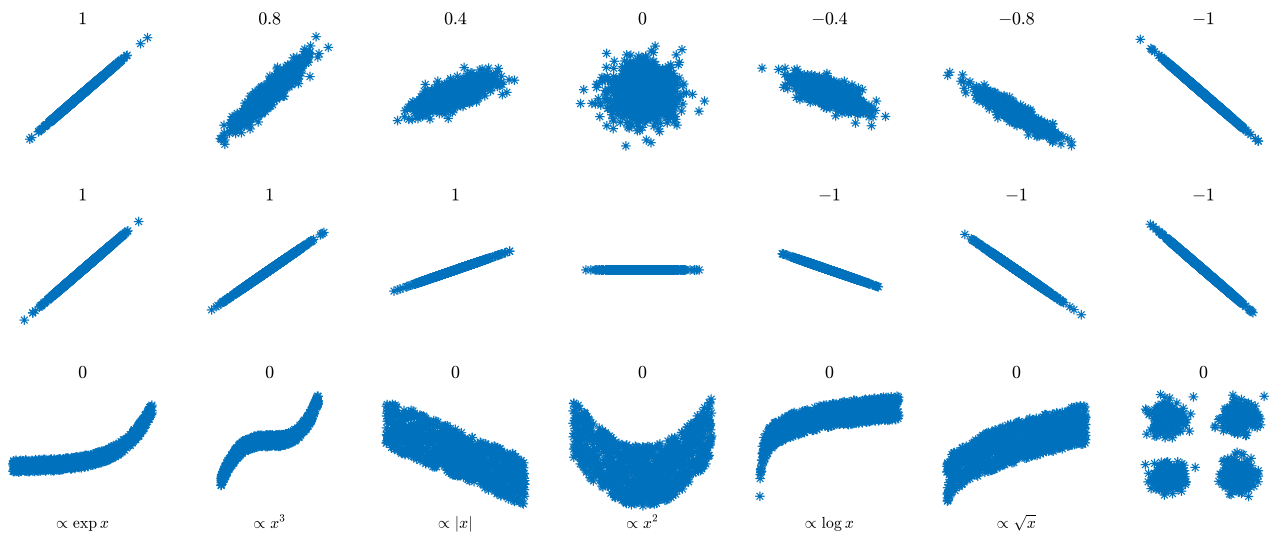


Abbildung 2.29: Beispielhafte Darstellung des Korrelationskoeffizienten für lineare und nichtlineare Zusammenhänge

Gerade in Bezug auf Systeme können dadurch Eingangs- und Ausgangsgrößen unkorreliert sein, obwohl sie einen starken nichtlinearen funktionalen Zusammenhang aufweisen.

2.5.2.5 Stochastische Prozesse

Ein stochastischer Prozess kann als eine Zufallsvariable beschrieben werden, die sich über einer unabhängigen Variable ändert. Die unabhängige Variable kann z.B. die Zeit t sein oder auch jede andere Größe, die den Prozess beeinflusst. Durch diese Erweiterung werden die PDF und pdf ebenfalls abhängig von dieser zusätzlichen Variablen und es folgt:

$$F_X(x,t) = P(X(t) \leq x) \tag{2.116}$$

und des Weiteren

$$f_X(x,t) = \frac{\mathrm{dF_X(x,t)}}{\mathrm{d}x} \ . \tag{2.117}$$

Ebenfalls werden die Momente der einzelnen Verteilungen wie auch die Beziehungen zwischen Verteilung von der Variablen abhängig. Daraus folgt

$$\rho(t) = \frac{C_{X(t),Y(t)}}{\sigma_x(t)\sigma_y(t)} \ , \tag{2.118}$$

womit die Korrelation zweier stochastischer Prozesse auf einzelne Abschnitte $\tau \in (0,t)$ beschränkt werden kann.

2.5.3 Kalman Filter

Das Kalman Filter ist das beste bzw. Bayes-optimal lineare Filter, um die Zustände eines Systems zu schätzen [MZ09]. Ein zeit-diskretes linearen System, wobei k die diskreten Zeitpunkte bezeichnet, ist gegeben durch

$$x_k = F_{k-1}x_{k-1} + G_{k-1}u_{k-1} + w_{k-1} \tag{2.119}$$

$$y_k = H_k x_k + v_k \ , \tag{2.120}$$

dabei sind F und H die Matrizen, die die Systemfunktion beschreiben. Für zeitinvariante Systeme ändern sich diese Matrizen nicht. Für die Rauscheinflüsse gilt

$$w_k \sim (0, Q_k) \tag{2.121}$$

$$v_k \sim (0, R_k) \tag{2.122}$$

$$E[w_k w_j^T] = Q_k \delta_{k-j} \tag{2.123}$$

$$E[v_k v_j^T] = R_k \delta_{k-j} \tag{2.124}$$

$$E[v_k w_j^T] = 0 \ . \tag{2.125}$$

Dies bedeutet, dass die Rauschprozesse weiß, mittelwertfrei und unkorreliert sind. Zudem sind ihre Kovarianzmatrizen bekannt. Das Kronecker-Delta ist definiert als

$$\delta_{k-j} = \begin{cases} 1 & , \text{ falls } k = j \\ 0 & , \text{ falls } k \neq j \end{cases} \tag{2.126}$$

Ziel ist es, die Zustände x_k des Systems auf Grundlage der Kenntnis der Systemdynamik und verrauschter Messungen y_k zu schätzen. Um zwischen den einzelnen Arten der Schätzung zu unterscheiden, sind die Zeitpunkte der Schätzungen in Abbildung 2.30 in Relation zum aktuellsten Messzeitpunkt $k = 5$ gezeigt.

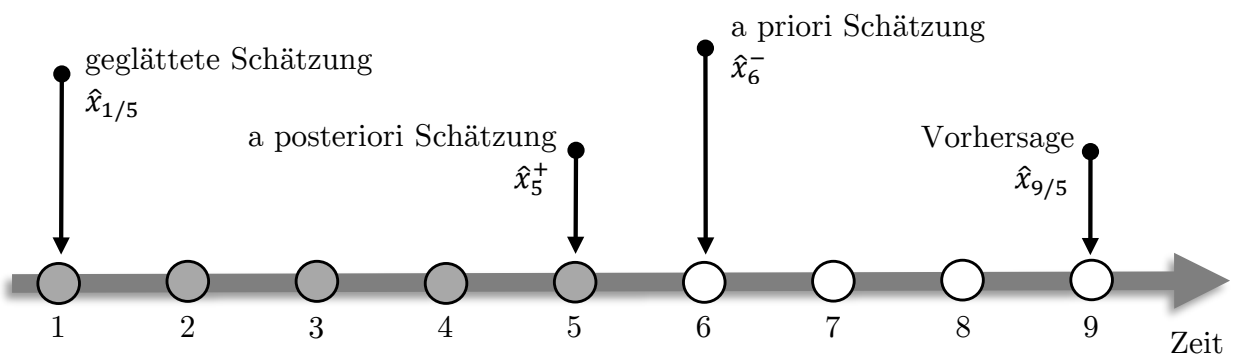


Abbildung 2.30: Zeitpunkte einzelner Schätzungen relativ zum aktuellsten Messzeitpunkt $k = 5$.

Die einzelnen mathematischen Definitionen der Schätzungen sind wie folgt:

1. *a posteriori* Schätzung

$$\hat{x}_k^+ = E[x_k|y_1, y_2, \ldots, y_k] \tag{2.127}$$

   Schätzung der Zustände x_k nach der Verarbeitung der Messung zum Zeitpunkt k.

2. *a priori* Schätzung

$$\hat{x}_k^- = E[x_k|y_1, y_2, \ldots, y_{k-1}] \tag{2.128}$$

   Schätzung der Zustände x_k vor der Verarbeitung der Messung zum Zeitpunkt k.

3. Geglättete Schätzung

$$\hat{x}_{k|k+N} = E[x_k|y_1, y_2, \ldots, y_k, \ldots, y_{k+N}] \tag{2.129}$$

   Dabei ist N eine positive Integerzahl. Diese Schätzung verwendet N Messungen nach dem Zeitpunkt k, welche bekannt sein müssen, und alle vorherigen Messungen um die Schätzung zu glätten.

4. Vorhergesagte Schätzung

$$\hat{x}_{k|k-M} = E[x_k|y_1, y_2, \ldots, y_{k-M}] \tag{2.130}$$

   Mit dieser Gleichung lässt sich eine Vorhersage des Zustandes in Abhängigkeit sämtlicher vorangegangener Messungen berechnen.

Da zum Zeitpunkt $k = 0$ keine Messwerte vorhanden sind, wird $X_0^+ = E(x_0)$ gesetzt. Die Kovarianzmatrizen, die als Maß der Genauigkeit der jeweiligen Schätzung dienen, sind durch

$$P_k^- = E[(x_k - \hat{x}_k^-)(x_k - \hat{x}_k^+)^T] \tag{2.131}$$
$$P_k^+ = E[(x_k - \hat{x}_k^+)(x_k - \hat{x}_k^+)^T] \tag{2.132}$$

gegeben. Die Filter-Gleichungen, die in jedem Zeitschritt k berechnet werden müssen, nach Gleichung (2.119), sind:

$$P_k^- = F_{k-1}P_{k-1}^+F_{k-1}^T + Q_{k-1} \tag{2.133}$$
$$K_k = P_k^- H_k^T (H_k P_k^- H_k^T + R_k)^{-1} = P_k^+ H_k^T R_k^{-1} \tag{2.134}$$
$$\hat{x}_k^- = F_{k-1}\hat{x}_{k-1}^+ + G_{k-1}u_{k-1} \quad \text{a priori Zustandsstschätzung} \tag{2.135}$$
$$\hat{x}_k^+ = x_k^- + K_k(y_k - H_k\hat{x}_k^-) \quad \text{a posteriori Zustandsstschätzung} \tag{2.136}$$
$$P_k^+ = (I - K_k H_k)P_k^- \tag{2.137}$$

Dabei ist I die Einheitsmatrix und K_k wird Kalman-Verstärkungsfaktor genannt. Zur Laufzeit müssen nur die beiden $\hat{x}_k$ Gleichungen ausgewertet werden, alle anderen

Gleichungen sind unabhängig von den Messwerten eines bestimmten Systems. Falls die Zustände des Systems x_k konstant sind, reduziert sich das Kalman-Filter zu einem rekursiven Kleinste-Quadrate Filter.

Anmerkungen zum Kalman Filter Das Kalman-Filter ist allgemein für lineare Systeme mit weißem unkorrelierten Rauschen entwickelt worden. Es existieren verschiedene Ansätze für das Kalman-Filter, um diese Einschränkungen zu umgehen. Für nichtlineare Systeme können die Gleichungen mittels Taylorreihen linearisiert werden. Dadurch entsteht ein zusätzlicher Approximationsfehler, der sich besonders für stark nichtlineare System negativ auf die Güte der Schätzung auswirkt. Zudem müssen sämtliche Update-Gleichungen in jedem Schritt berechnet werden. Diese Filter werden Extended Kalman-Filter (EKF) genannt. Eine Weiterentwicklung der EKF ist das Unscented Kalman-Filter (UKF), dabei wird anhand eines Sigma-Bereiches der wahrscheinlichste Zustandsvektor durch das Propagieren von Mittelwert und Kovarianz durch das nichtlineare System berechnet. Dadurch kann der Linearisierungsfehler des EKF deutlich reduziert werden. Jedoch werden die Ergebnisse ebenfalls um einen bestimmten Punkt angenähert. Für die Schätzung von konstanten Zuständen ist das Kalman-Filter ebenfalls nicht geeignet, es muss künstliches Rauschen hinzugefügt werden, um die Qualität der Schätzung zu verbessern. Eine weitere Möglichkeit, um mit Nichtlinearitäten oder korreliertem Rauschen umzugehen, ist die künstliche Erhöhung der Systemordnung. Um die rein numerische Qualität das Kalman-Filterung zu verbessern, existieren verschiedene Ansätze, die die notwendigen Matrix-Operationen wie die Invertierung durch vorherige Zerlegung oder Pivotisierung erleichtern. Für den Fall nicht perfekter Systemkenntnis oder unbekannten Einflussfaktoren ist die Kalman-Schätzung stark beeinträchtigt, für diesen Fall können mehrfache Kalman-Filter verwendet werden. Mit Hilfe des Satz von Bayes kann die Schätzung

$$P(X|Y) = \frac{\mathrm{pdf}(y|x)P(x)}{\sum\limits_{i=1}^{N} \mathrm{pdf}(y|x_i)P(x_i)} \tag{2.138}$$

in N einzelne Schätzungen zerlegt werden.

Limitierungen des Kalman-Filters Der Mittelwert und die Korrelation beider Rauschprozesse muss zu jedem Zeitpunkt k bekannt sein. Zudem müssen die Kovarianzen der Rauschprozesse bekannt sein. Da Q_k und R_k Designvariablen des Kalman-Filters sind, kann bei Unkenntnis die generelle Verwendung nahezu unmöglich werden. Das Kalman-Filter stellt den Schätzer mit minimaler Varianz für Gauss-verteiltes Rauschen und den Schätzer mit linearer minimaler Varianz dar, falls das Rauschen nicht Gauß-verteilt ist. Soll eine andere Kostenfunktion minimiert werden, kann das Kalman-Filter gegebenenfalls gar nicht verwendet werden. Schlussendlich müssen die Systemmatrizen F_k und H_k bekannt sein.

Bei den in der Arbeit betrachteten Systemen kommen alle genannten Einschränkungen des Kalman-Filters zusammen. In einigen Betriebsbereichen sind die Zustände konstant. Die Korrelation von Rauschen sowie Ein- und Ausgangsgrößen hängt ebenfalls vom Arbeitspunkt ab und die Systeme weisen extrem starke Nichtlinearitäten wie exponentiellen Anstieg und kurz darauf Sättigung auf. Wie später in der Arbeit gezeigt wird, ist der wichtigste zu schätzende Zustand genau an der Grenze dieser beiden Bereiche. Daher kann das System nicht mit einer Taylorreihe ausreichend genau approximiert werden. Somit werden im weiteren keine Kalman-Filter für die Zustandsschätzung verwendet, auch wenn diese die am weitesten erforschten und am besten dokumentierten Zustandsschätzer sind [Sim06].

2.5.4 H_∞ Filter

Das H_∞, oder auch Minimax-Filter, ist eine Weiterentwicklung des Kalman-Filters, der für robuste Schätzungen in Fällen von Modellierungsungenauigkeiten und Rauschprozessunsicherheiten gedacht ist [Sim06]. Ausgehend von einem linearen System soll eine Schätzung einer Linearkombination z_k der Zustände x_k zum Zeitpunkt k gegeben durch

$$z_k = L_k x_k \tag{2.139}$$

berechnet werden. Die Rauschprozesse des Systems unterliegen keinen Einschränkungen. Die zu minimierende Kostenfunktion ergibt sich zu

$$J_1 = \frac{\sum\limits_{k=0}^{N-1} ||z_k - \hat{z}_k||^2_{S_k}}{||x_0 - \hat{x}_0||^2_{P_0^{-1}} + \sum\limits_{k=0}^{N-1} \left(||w_k||^2_{Q_k^{-1}} + ||v_k||^2_{R_k^{-1}} \right)} \tag{2.140}$$

mit

$$||z_k - \hat{z}_k||^2_{S_k} = \sqrt{(z_k - \hat{z}_k)^T S_k (z_k - \hat{z}_k)} \ . \tag{2.141}$$

x_0 ist dabei der frei wählbare initiale Zustand. J_1 soll unter der Bedingung von maximalen Rauscheinflüssen und x_0 minimiert werden. S_k ist eine Matrix zur Gewichtung der einzelnen Zustände während der Schätzung, alle anderen Größen (P_0, Q_k, R_k) entsprechen denen aus der Herleitung des Kalman-Filters. Jedoch müssen diese nicht

zwingend bekannt sein. Unter Angabe der benutzerdefinierten Schwelle $\frac{1}{\theta}$ gelten für das H_∞-Filter folgende Update-Gleichungen:

$$\bar{S}_k = L_k^T S_k L_k \tag{2.142}$$

$$K_k = P_k \left[I - \theta \bar{S}_k P_k + H_k^T R_k^{-1} H_k P_k \right]^{-1} H_k^T R - k^{-1} \tag{2.143}$$

$$\hat{x}_{k+1} = F_k \hat{x}_k + F_k K_k (y_k - H_k \hat{x}_k) \tag{2.144}$$

$$P_{k+1} = F_k P_k \left[I - \theta \bar{S}_k P_k + H_k^T R_k^{-1} H_k P_k \right]^{-1} F_k^T + Q_k \tag{2.145}$$

Zudem muss in jedem Schritt k $P_k^{-1} + \theta \bar{S}_k + H_k^T R_k^{-1} H_k > 0$ erfüllt sein. Wird $\theta = 0$ gewählt, geht das H_∞-Filter in das Kalman-Filter über.

Anmerkungen zum H_∞-Filter Im Gegensatz zum Kalman-Filter ist das H_∞-Filter robuster gegenüber Fehlern im Systemmodell und Systemen, die sich unvorhergesehen ändern. Die Schätzung ist dabei sehr pessimistisch. Zudem ist die Herleitung sehr aufwendig, dies erschwert die Anwendung auf nichtlineare Systeme mit Hilfe von Taylorreihen. Zudem besteht die Möglichkeit der Kombination von Kalman- und H_∞-Filtern, dies dient maßgeblich dazu, die sehr pessimistische Schätzung des H_∞-Filters abzumildern.

2.5.5 Partikelfilter

Im Gegensatz zu den beiden bisher vorgestellten Filtern ist das Partikelfilter ein nichtlinearer Filter der auf der Bayes'schen Zustandsschätzung beruht [AMGC02]. Neben der Bezeichnung Partikelfilter werden sequential importance sampling, bootstrap filtering, condensation algorithm, interacting particle approximations, Monte-Carlo Filterung, sequentielle Monte-Carlo Filterung und survival of the fittest verwendet [C$^+$03]. Der Ablauf des Algorithmus ist wie folgt: Der Zustandsraum wird in eine Vielzahl von Abschnitten unterteilt, diese werden mit Partikeln entsprechend einer Wahrscheinlichkeitskennzahl aufgefüllt. Je höher die Wahrscheinlichkeit, desto dichter sind die Partikel konzentriert. Das System entwickelt sich zeitlich den Gleichungen des Zustandsraummodells entsprechend. Die jeweiligen Verteilungsdichten entwickeln sich währenddessen entsprechend weiter. Die Verteilungsdichten können dabei mit Punktmassen-Histogrammen, durch zufälliges Abtasten des Zustandsraumes, approximiert werden. Dadurch repräsentieren diese abgetasteten Partikel die sich verändernde Verteilungsdichte. Des Weiteren ist die posteriori Verteilung unbekannt oder sehr schwierig zu schätzen, daher ist die Wahl einer anderen repräsentativen Verteilung für ein effizientes Abtasten wünschenswert. Für die einzelnen beschriebenen Aufgaben, von der Wahl der Wahrscheinlichkeitskennzahl bis zur Auswahl der Ersatzverteilung, gibt es sehr viele verschiedene Implementierungen von Partikelfiltern. In dieser Arbeit wird eine Implementierung gewählt, die relativ simple Algorithmik verwendet, da das Zustandsraummodell bereits sehr aufwendig ist. Im Folgenden werden die verwendeten Gleichungen des implementierten Algorithmus

kurz vorgestellt. Für eine tiefere Analyse der gesamten Thematik sei auf [C+03] und [Sim06] verwiesen. In beiden Quellen werden alle in dieser Arbeit betrachteten Filteralgorithmen von Beginn an beschrieben. Besonders [C+03] zeigt die verschiedenen Arten der Partikelfilter-Implementierung und eine umfassende Betrachtung der jeweiligen theoretischen und praktischen Vor- und Nachteile der einzelnen Aspekte.

Die Herangehensweise benötigt keinerlei Annahme für das System und funktioniert für Systeme, die stark nichtlinear sind und für die Kalman-Filter versagen. Ausgehend von einem nichtlinearen System gegeben durch die folgenden Gleichungen

$$x_{k+1} = f_k(x_k, w_k) \tag{2.146}$$

$$y_k = h_k(x_k, v_k) \tag{2.147}$$

sollen die Zustände x_k geschätzt werden. $f_k(\cdot)$ und $h_k(\cdot)$ sind zeitvariante nichtlineare Funktionen. Das Rauschen wird als weiß und unabhängig mit bekannten Verteilungen angenommen. Das Ziel der Bayes'schen Schätzung ist die Annäherung der bedingten Wahrscheinlichkeitsdichtefunktionen $p(x_k|Y_k)$ der Zustände x_k in Abhängigkeit der Messungen $y_1, y_2, \ldots, y_k$. Zum Zeitpunkt der Schätzung des initialen Wertes

$$p(x_0) = p(x_0|Y_0) \tag{2.148}$$

liegen, wie beim Kalman-Filter, keine Messwerte vor. Im Gegensatz zu den vorherigen Filtern kann bei dem Partikelfilter von einer beliebigen, auch multimodalen, Wahrscheinlichkeitsdichtefunktion ausgegangen werden. Eine multimodale Wahrscheinlichkeitsdichtefunktion kann dabei mehr als ein Maximum aufweisen. Gerade diese Eigenschaft ist von entscheidender Bedeutung für die Anwendung der Partikelfilter für die messtechnischen Zwecke in dieser Arbeit. Ziel des Algorithmus, der dem Partikelfilter zu Grunde liegt, ist die Berechnung der bedingten Wahrscheinlichkeitsdichtefunktion $p(x_k|Y_k)$. Diese pdf beschreibt die Wahrscheinlichkeit, mit der ein bestimmter Ausgangswert zu einem Systemzustand zuzuordnen ist. Somit kann von beobachteten Ausgangswerten auf interne Systemzustände geschlossen werden. Für die Berechnung eben dieser bedingten pdf werden alle Messungen bis zum Zeitpunkt k verwendet, um die a priori pdf zu bestimmen:

$$p(x_k|Y_{k-1}) = \int p[(x_k, x_{k-1})|Y_{k-1}]\,\mathrm{d}x_{k-1} \tag{2.149}$$

$$= \int p[x_k|(x_{k-1}, Y_{k-1})]p(x_{k-1}|Y_{k-1})\,\mathrm{d}x_{k-1} \tag{2.150}$$

Für ein technisches System ist es sehr schwierig, die bedingte pdf $p[(x_k, x_{k-1})|Y_{k-1}]$ zu berechnen, da diese alle Zustandsübergänge in Zusammenhang mit einem bestimmten Ausgangswert beschreibt. Um diese zu bestimmen, sind sehr viele Messungen nötig. Im Gegensatz dazu enthalten die pdfs in Gleichung (2.150) nur solche Wahrscheinlichkeiten,

die iterativ bestimmt werden können. Falls das System rückkopplungsfrei ist, kann Gleichung (2.150) weiter zu

$$p(x_k|Y_{k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|Y_{k-1})\,\mathrm{d}x_{k-1} \tag{2.151}$$

vereinfacht werden. Die im Rahmen dieser Arbeit betrachteten Systeme sind rückkopplungsfrei da es sich um einzelne Bauelemente oder ungeregelte Schaltungen handelt. Für die a posteriori pdf folgt aus Gleichungen (2.103) und (2.106)

$$p(x_k|Y_k) = \frac{p(Y_k|x_k)}{p(Y_k)}p(x_k) = \frac{p(y_k|x_k)p(x_k|Y_{k-1})}{p(y_k|Y_{k-1})} \ . \tag{2.152}$$

In dieser Gleichung wird $p(y_k|x_k)$ durch die Messfunktion $h_k(\cdot)$ bestimmt, und $p(x_k|Y_{k-1})$ ist die zuvor bestimmte a priori pdf. Des Weiteren gilt

$$p(y_k|Y_{k-1}) = \int p(y_k|x-k)p(x_k|Y_{k-1})\,\mathrm{d}x_k \ , \tag{2.153}$$

womit alle notwendigen pdf aus den Messwerten und den Systemgleichungen bestimmbar sind. Analytische Lösungen für die Berechnungen der a priori und a posteriori pdf existieren jedoch nur unter den Annahmen eines linearen Systems und gaußverteiltem unkorreliertem Rauschen. Dies führt zurück auf die Lösung mit Hilfe des Kalman-Filters. Daher basiert das Partikelfilter auf der numerischen Annäherung beider Funktionen. Zu Beginn werden N zufällige Zustandsvektoren $x^+_{(0,i)}$ $(i = 1, \ldots, N)$ erzeugt. Mit diesen Zustandsvektoren wird die Prozessgleichung ausgewertet, um mögliche Folgezustände zu berechnen:

$$x^-_{(k,i)} = f_{k-1}(x^+_{k-1,i}, w_{k-1,i}) \tag{2.154}$$

Mit Hilfe der Messgleichung kann nun die relative Wahrscheinlichkeit $q_i = P[(y_k = y^*)|(x_k = x^-_{k,i})]$ bestimmt werden, welche angibt, dass ein bestimmter Zustand zu einem bestimmten Ausgang führt. Dies ist je nach Komplexität des Systems und der Partikelanzahl N mit einem hohen Rechenaufwand verbunden. Die relative Wahrscheinlichkeit über alle Partikel in einem Zeitschritt ergibt sich zu

$$q_i = \frac{q_i}{\sum\limits_{j=1}^{N} q_j} \ . \tag{2.155}$$

Anschließend werden neue Zustandsvektoren $x^+_{k,i}$ auf Basis von q_i generiert. Um von der beliebigen pdf q_i eine Stichprobe erhalten zu können, werden verschiedene Ansätze wie order statisics, stratified sampling, residual sampling oder systematic resampling verwendet. Entscheidend sind die verschiedenen Verfahren nur, falls der Rechenaufwand

für das Erstellen einer Stichprobe auf Grundlage von q_i zu hoch wird. Abschließend wird der Erwartungswert

$$E(x_k|y_k) \approx \frac{1}{N}\sum_{i=0}^{N} x_{k,i}^{+} \tag{2.156}$$

durch den Mittelwert der Partikel angenähert. Dies entspricht dem wahrscheinlichsten Systemzustand zu einem beobachteten Ausgangsvektor.

Anmerkungen zum Partikelfilter Neben dem hohen Rechenaufwand durch die N-fache Auswertung der Systemgleichungen besteht das Problem der Partikelverarmung. Dabei kann es passieren, dass nach einigen Zeitschritten nur sehr wenige Partikel ein $q_j \neq 0$ besitzen. Dies geschieht, falls nur wenige a priori Partikel in Regionen liegen, in denen $p(y_k|x_k)$ signifikante Werte annimmt. Eine einfache Lösung ist die Erhöhung von N, dies führt zu einem erhöhten Rechenaufwand und verschiebt das Problem häufig nur zu späteren Zeitpunkten k. Weitere Möglichkeiten sind das Hinzufügen von künstlichem Rauschen zu $x_{k,i}^{+}$, oder die Implementierung eines regularisierten Partikelfilters, bei dem von einer kontinuierlichen pdf

$$\hat{p}(x_k|y_k) = \sum_{i=1}^{N} w_{k,i} K_h(x_k - x_{k,i}) \tag{2.157}$$

a posteriori Partikel gezogen werden. Für die Gewichte gilt $w_{k,i} = q_i$ und K_h sind Kernel-Funktionen von Verteilungsdichten, häufig Epanechnikov [Epa69].

$$K_h = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right) & , |x| \leq \sqrt{5} \\ 0 & , |x| > \sqrt{5} \end{cases} \tag{2.158}$$

Der Vorteil dieser Kernel-Funktion ist es, dass sie den mittleren quadratischen Fehler des Schätzers minimiert. Durch die kontinuierliche pdf besitzen weniger Partikel verschwindende Gewichte und das Verarmungsproblem wird abgeschwächt. Eine weitere Möglichkeit ist das Markov-Ketten-Monte-Carlo-Resampling, dabei werden anhand der Übergangswahrscheinlichkeit $p(x_{k,i}^{-}|x_{k-1,i}^{+})$ zufällig einige $x_{k,i}^{-}$ zu neuen Zustandsvektoren $\tilde{x}_{k,i}^{-}$. Die Akzeptanzwahrscheinlichkeit ist dabei durch

$$\alpha = \min\left[1, \frac{p(y_k|\tilde{x}_{k,i}^{-})p(\tilde{x}_{k,i}^{-}|x_{k-1,i}^{+})}{p(y_k|x_{k,i}^{-})p(x_{k,i}^{-}|x_{k-1,i}^{+})}\right] \tag{2.159}$$

gegeben. Somit werden Partikel mit geringem Gewicht erneut generiert. Zudem kann die Berechnung von der relativen Wahrscheinlichkeit um einen statischen Faktor abhängig von x_k und $x_{k,i}^-$ erweitert werden. Die Gleichung (2.155) wird somit zu

$$q_i = \frac{P[(y_k = y^*)|(x_k = x_{x_k,i}^-)]}{P[(y_k = y^*)|\mu_{k,i}]} \ . \tag{2.160}$$

Dabei kann $\mu_{k,i} = E(x_k|x_{k,i}^-)$ oder $\mu_{k,i} = \mathrm{pdf}(x_k|x_{k,i}^-)$ gelten. Eine einfachere Möglichkeit ist ebenso

$$\tilde{q} = \frac{(\alpha - 1)q_i + \bar{q}}{\alpha} \tag{2.161}$$

mit $\alpha \in (0, \infty)$ und $\bar{q} = E(q_i)$. Beide Maßnahmen führen dazu, dass Partikel mit hohen Wahrscheinlichkeiten reduzierte Gewichte erhalten, um der Verarmung entgegen zu wirken. Diese Art der Erweiterung wird Auxiliary Partikel-Filterung genannt. Die einzige limitierende Bedingung bei der Implementierung eines Partikelfilters ist der Rechenaufwand, alle anderen Designparameter können an beliebige Systeme und Verteilungsdichten angepasst werden.

2.5.6 Vergleich der Filteralgorithmen

Um die Filteralgorithmen miteinander zu vergleichen, werden zwei häufig verwendete Szenarien aus der Literatur verwendet. Beide Anwendungsbeispiele sind im Anhang A.6 näher erläutert. Das erste Beispiel ist der freie Fall eines Objektes aus großer Höhe. Das System ist schwach nichtlinear und das Rauschen jeweils Gauß-verteilt und weiß. In Abbildung 2.31 sind die Fehler der Schätzung der Position (P), Geschwindigkeit

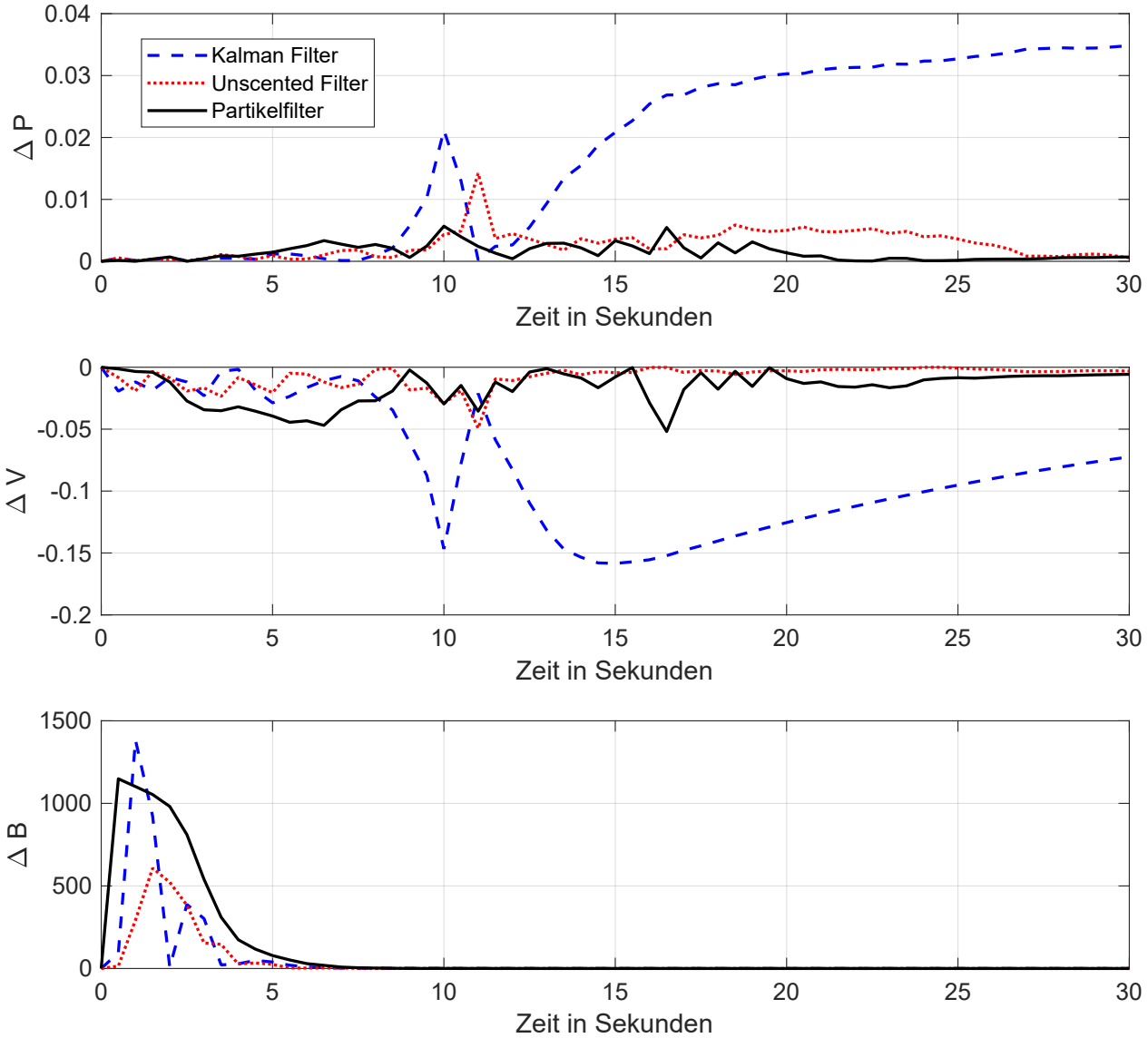


Abbildung 2.31: Vergleich der dimensionslosen, relativen Fehlern von Kalman-Filter, unscented Kalman-Filter und Partikelfilter der Schätzung der Position (P), Geschwindigkeit (V) und ballistischen Koeffizienten (B) eines Objekts im freiem Fall.

Kalman-Filter die größten Schätzfehler verursacht. Dies ist damit zu erklären, dass bereits die schwache Nichtlinearität des Systems das Schätzergebnis des Kalman-Filters

beeinflusst. Das unscented Kalman-Filter (UKF) approximiert das nichtlineare System abschnittsweise linear und kann somit den Schätzfehler deutlich reduzieren. Dies tritt besonders ab dem Zeitpunkt $t = 10$s stark auf, da sich zu diesem Zeitpunkt das Fallverhalten ändert und der Übergang zwischen diesen beiden Bereichen nichtlinear ist. Dadurch erhöht sich kurzzeitig der Fehler der beiden Kalmanfilter. Für die Schätzung der Position und der Geschwindigkeit bleibt beim Kalmanfilter eine Abweichung von bis zu 10% bestehen. Das Ergebnis des Partikelfilters ist ungefähr dem des UKF gleichzusetzen. Dies ist zu erwarten, da beide Algorithmen für schwach nichtlineare Systeme gleichwertig sind. In Abbildung 2.32 ist das Konvergenzverhalten der drei verschiedenen Algorithmen gezeigt. Die Position wird von allen drei Algorithmen von Beginn an am genauesten angenähert. Für die anfängliche Schätzung der Geschwindigkeit bestehen bei allen

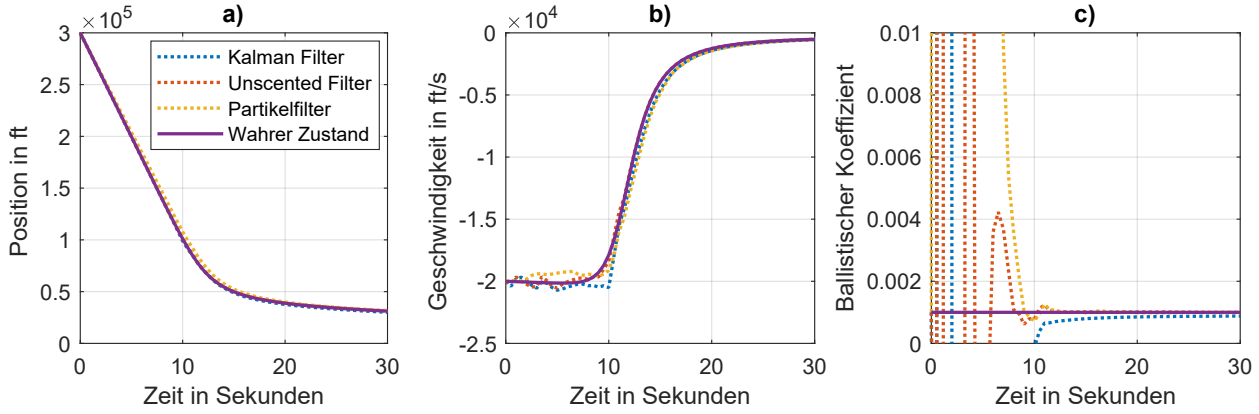


Abbildung 2.32: Konvergenzverhalten der betrachteten Filteralgorithmen für die Schätzung der Position (a), Geschwindigkeit (b) und ballistischen Koeffizienten (c) eines Objekts im freien Fall.

Filtern leichte Abweichungen für den Zeitraum bis 10 Sekunden. Beim ballistischen Koeffizienten sind die Abweichungen bis zu diesem Zeitpunkt für alle Filter sehr groß. Dies liegt ebenfalls daran, dass dieser Wert mit 10^{-3} betragsmäßig im Vergleich zu den anderen Größen sehr klein ist und sich deshalb Abweichungen deutlich stärker auswirken.

Der zweite Testfall ist ein stark nichtlineares System. Der Vergleich von Partikel- und Kalman-Filter ist in Abbildung 2.33 für ein stark nichtlineares System gezeigt. Wie klar zu erkennen ist, gehen die Zustandsschätzungen beider Filter ab dem initialen Messwert sofort deutlich auseinander. Das Partikelfilter ist in der Lage, den Zustand mit sehr geringen Abweichungen über den gesamten Verlauf anzunähern. Das Kalman-Filter hingegen liegt bis zum Zeitpunkt $k = 20$ weit ab vom eigentlichen Systemzustand. Danach werden teilweise einige Zustände richtig durch das Kalman-Filter geschätzt. Diese Abweichungen treten trotz dessen auf, dass bereits ein Extended Kalman-Filter (EKF) verwendet wird, welcher Nichtlinearitäten des Systems berücksichtigen kann. Der Grund für die starken Abweichungen wird deutlich bei der Betrachtung der geschätzten Zustandsverteilungsdichten zum Zeitpunkt $k = 20$. Das Partikelfilter approximiert eine

multimodale Verteilung, wohingegen das Kalman-Filter nur eine einzige Gaußverteilung approximieren kann. Das System besitzt zwei dedizierte Zustände zwischen denen unvorhersehbar, bedingt durch die Nichtlinearitäten, gewechselt werden kann.

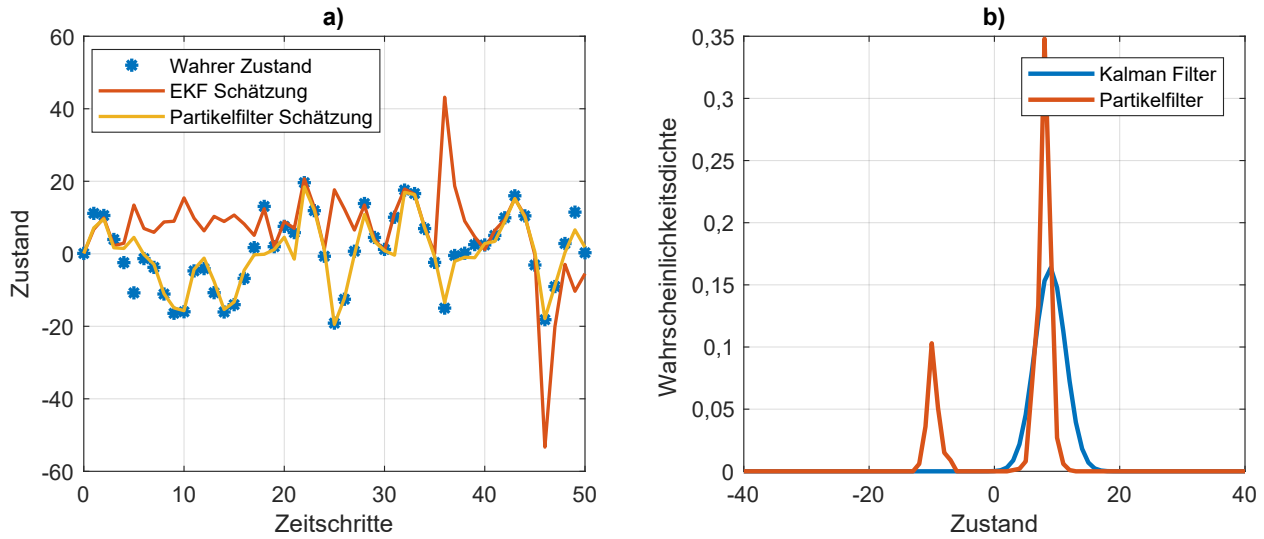
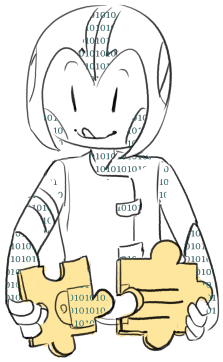


Abbildung 2.33: Vergleich von erweiterter Kalman-Filter und Partikelfilter für ein stark nichtlineares System. a) Zustandsschätzung b) Wahrscheinlichkeitsdichtefunktion der Zustände beim Zeitpunkt $k = 20$

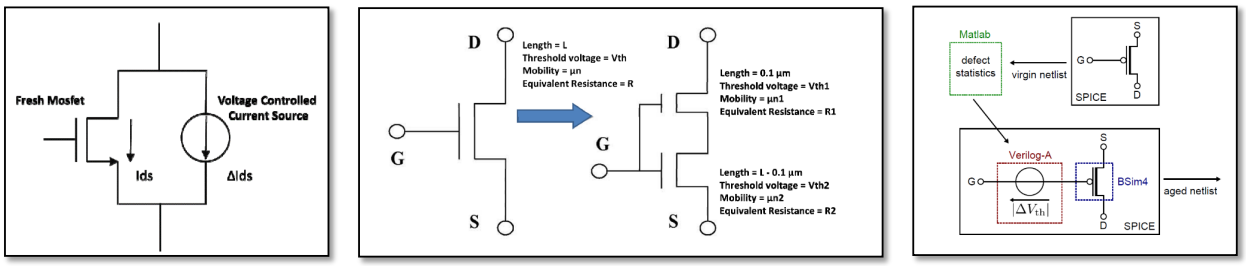
Nur das Partikelfilter ist dadurch, dass keine Annahmen über die initialen Verteilungsdichtefunktionen angenommen werden müssen, in der Lage die Zustandsvektoren nichtlinearer Systeme, bei denen verschiedene voneinander getrennte Betriebszustände existieren, korrekt zu schätzen. Zudem können allgemein bimodale oder multimodale Verteilungsdichtefunktionen so korrekt geschätzt werden. Bei allen anderen Algorithmen, ob nun Kalman, EKF, UKF oder H_∞ müssen stets Annahmen über einzelne Eigenschaften des zugrunde liegenden Systems getroffen werden. Da dies, wie in späteren Kapiteln gezeigt, für die in dieser Arbeit betrachteten Bauelemente und Schaltungen nicht möglich ist, werden im Weiteren Partikelfilter für die Zustandsschätzung eingesetzt.

3 Entwurfsprozess unter Berücksichtigung der Zuverlässigkeit

» A problem well stated is a problem half solved «



Wie in den vorangegangenen Kapiteln beschrieben, besteht die Notwendigkeit, zur Auslegung von integrierten Schaltungen die Einflüsse von Alterung, Prozessvariationen und Strahlung gleichzeitig mit realistischen Belastungsprofilen analysieren zu können. Vom Compact Model Council (CMC) wurde eine Schnittstelle für die Modellierung von Alterung vorgestellt. Diese bindet externe Alterungsmodelle mit Hilfe von zusätzlichen Simulatoren (UDRM, MOSRA, RelXpert) in den eigentlichen Entwurfsprozess ein [HCF+15, Ram13]. Bei diesem Vorgehen können nur periodische Belastungsprofile analysiert werden, da die Alterung auf Basis einer kurzen transienten Simulation extrapoliert wird. Anschließende Simulationen verwenden eine neu erstellte degradierte Netzliste für die Analyse anderer Einflüsse. Zudem werden die Alterungseffekte als vom Transistor getrennte Effekte simuliert, wie in Abbildung 3.1 gezeigt ist. Wie zu sehen ist, wird jede einzelne Auswirkung, Senkung von I_d sowie Erhöhung von V_{th}, der Alterung getrennt modelliert. Diese Ersatzschaltbilder bilden die Wechselwirkungen und Abhängigkeiten nicht korrekt ab, so zum Beispiel ist die Senkung des I_d teilweise durch die



Alterung von I_{ds} **Alterung durch HCI** **Simulationsablauf**

Abbildung 3.1: Modellierung verschiedener Alterungseffekte nach dem CMC Standard [Ram13].

Erhöhung von V_{th} bedingt, beides tritt demnach gleichzeitig auf. Die Modellierung von HCI mit einem zusätzlichen sehr kurzen Transistor führt dazu, dass Prozessvariationen und Temperatur/Spannung-Einflüsse falsch vorhergesagt werden, da in das Design ein Kurzkanaltransistor eingeführt wurde. Für die Analyse von Prozessvariationen müssen zudem die Abhängigkeiten von ΔI_d und ΔV_{th} bekannt sein, gleiches gilt für die Temperatur. Für einen analogen Entwurfsprozess ist die Veränderung der Netzliste und der Topologie durch das Einbringen zusätzlicher Schaltungselemente problematisch, da diese innerhalb der SPICE-Simulation zu parasitären Strömen oder Spannungsabfällen führen können, die das Gesamtergebnis verfälschen. Wie in Abbildung 3.2 gezeigt, wird mit jedem zu betrachtenden Effekt der übliche Modellierungsansatz problematischer. Zudem haben Parameter wie V_{th} in modernen Transistormodellen keine tatsächliche Funktion [CVC+14a] mehr und eignen sich somit nicht für die Modellierung der Einflüsse. Im BSIM6-Modell wurde die Schwellspannung erst 2 Jahre nach Veröffentlichung in der Version 6.1.1. wieder eingeführt, um Designern ihren gewohnten Parameter wiederzugeben. Für die Modellierung und die Berechnungen im Modell besitzt die Schwellspannung keine Funktion.

Um diesen gewachsenen Nachteilen entgegenzuwirken, wird in dieser Arbeit ein vollständig neuer Ansatz verfolgt. Die Idee ist das Implementieren sämtlicher relevanter Effekte an dem Ort ihres Entstehens, im Transistormodell, wie es für Temperatur längst getan wird. Somit können während des Schaltungsentwurfs alle relevanten Einflüsse gleichzeitig, ohne Änderung der Schaltung, betrachtet und analysiert werden. Wie in Abbildung 3.2 durch die idealen Spannungsquellen veranschaulicht, wird mit jedem zu betrachtenden Effekt der übliche Modellierungsansatz problematischer.

Durch den neuen Ansatz werden Alterung und Strahlungseinflüsse in das Transistormodell verschoben und somit innerhalb einer SPICE-Simulation im Entwurfsprozess analysierbar. Zudem können alle Wechselwirkungen im Transistor betrachtet werden. Im Folgenden wird dazu das Transistormodell vorgestellt und die Implikationen für den Entwurfsprozess evaluiert.

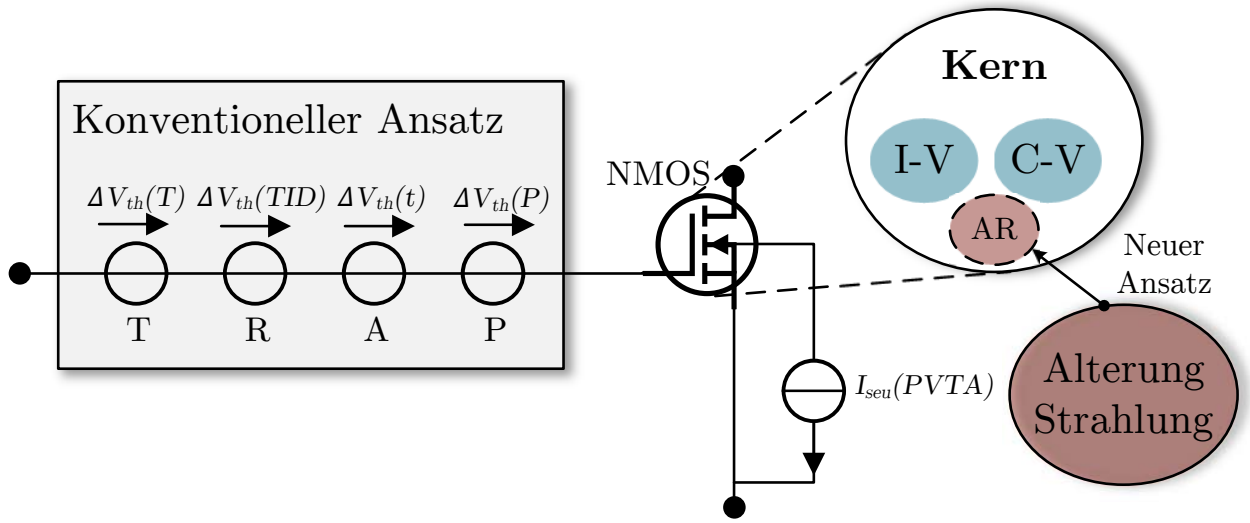


Abbildung 3.2: Neuartiger Modellierungsansatz durch Erfassung der Einflüsse direkt im Transistormodell.

3.1 REL-MOS-Transistormodell

In diesem Abschnitt wird das Transistormodell zusammen mit den jeweiligen Effekten und der Verilog-A-Implementierung beschrieben. Als Grundlage für die Entwicklung dient das BSIM-BULK-Transistormodell. Das BSIM6-Modell wurde nach seiner Einführung als industriell nutzbares Modell in BSIM-BULK umbenannt, um den Einsatzbereich klarer herauszustellen. Der Vorteil dieses Modells ist, dass es die Ladungsberechnungen aus dem EKV beinhaltet und um sämtliche relevanten Sekundäreffekte moderner Transistoren bei Bedarf erweiterbar ist. Dadurch, dass das Modell ladungsbasiert ist, können Alterungs- und Strahlungseffekte direkt modelliert werden, und es müssen keine Hilfsvariablen in Form von Strömen oder Spannungen eingeführt werden.
Durch das Compact Model Council wurden Standards und Richtlinien veröffentlicht, die beinhalten, welche Simulationsarten und Modellierungen für Zuverlässigkeitsanalysen auf Transistorebene notwendig sind [Ram13]. Als absolute Notwendigkeit wird die Portierbarkeit zwischen verschiedenen SPICE-Simulatoren, Betriebssystemen und Entwicklungsumgebungen definiert. Zudem müssen sich die Zuverlässigkeitsanalysen in ein bestehendes Process Design Kit (PDK) einbetten lassen. Ein PDK umfasst dabei Parametersätze für eine spezifische Herstellungstechnologie und Bauelementmodelle. Des Weiteren müssen sich die Alterungseinflüsse ausschalten lassen, um Simulationen ohne Zuverlässigkeitsaspekt nicht zu beeinflussen und unnötig zu verlängern. Die Simulationen selbst müssen Monte-Carlo-Analysen unterstützen und in der Lage sein, alle Alterungseinflüsse und Ausheilungsprozesse abzubilden. Zudem muss eine Wiederaufnahme von Simulationen, die Speicherung eines gealterten Zustands und die Speicherung von Alterungskenngrößen ermöglicht werden. Als für diese Arbeit konträr anzusehende Maßgaben sind, dass das Kompaktmodell nicht geändert werden soll und

dass davon ausgegangen wird, dass eine transiente Simulation über mehrere Jahre nicht möglich ist. Diese beiden Punkte können ambivalent diskutiert werden, da es in der Literatur Vorstöße in Richtung der Ermöglichung von transienten Simulationen gibt [HG16, GSR+14, GLKJ16]. Das REL-MOS-Kompaktmodell ist eine Änderung des BSIM-BULK-Modells, zur Laufzeit oder Nutzungsdauer in einem PDK muss dieses Kompaktmodell nicht verändert werden. Des Weiteren besteht durch das REL-MOS-Modell die Möglichkeit, eine initiale transiente Simulation über z.B. 10% der Nutzungsdauer der Schaltung in der Anwendung durchzuführen und anschließend die weitere Dauer zu extrapolieren. Somit können jedoch in der initialen Simulation bereits umfangreiche und beliebige Nutzungsprofile analysiert werden. Zudem zeigte sich im Rahmen der Arbeit, dass analoge Schaltungen mit bis zu 300 Transistoren sehr wohl, innerhalb einiger Stunden, direkt über ihre Lebenszeit von 10 Jahren transient simuliert werden können. Insgesamt bietet das REL-MOS-Modell alle durch das CMC notwendigen Eigenschaften für die Analyse von Zuverlässigkeit von integrierten Schaltungen und, soweit notwendig und gewünscht, deutlich erweiterte und genauere Analysemöglichkeiten.

3.1.1 Alterungseinflüsse

Für die Modellierung der Alterung wird der Parameter Q_{fc} verwendet (vgl. Abschnitt 2.2.3). Dieser beschreibt innerhalb des Transistormodells Defekte im Gateoxid und betrachtet diese als an der Oberfläche zum Substrat befindlich. Genau das beschreibt die Auswirkungen der Alterung. Die einzelnen Komponenten der Alterung werden zu einer Ladung

$$Q_{\text{age}}(t, \mathcal{P}, \mathbb{T}) = Q_{\text{trap}} + Q_{\text{it}} + Q_{\text{ox}} \tag{3.1}$$

zusammengefasst, wobei $Q_{\text{age}} \subset Q_{fc}$ gilt. t ist die Zeit, $\mathcal{P}$ beinhaltet alle relevanten Parameter, von denen die Alterung abhängt, und $\mathbb{T}$ beschreibt den Arbeitspunkt des Transistors. In Abbildung 3.3 ist schematisch die Änderung von $Q_{\text{age}}(t, \mathcal{P}, \mathbb{T})$ für mehrere Stress- und Erholungsphasen gezeigt. Die Auswirkungen der zusätzlichen Ladung lassen sich über die Oberflächenladung

$$Q'_s = Q_s + Q_{\text{age}}(t, \mathcal{P}, \mathbb{T}) \tag{3.2}$$

erfassen. Durch weitere Transistoreffekte wie Poly-Depletion kann die Alterungsladung nicht durch eine reine Erhöhung der Flachbandspannung dargestellt werden [EV06]. Sie muss in der Ladungsberechnung vollständig berücksichtigt werden. Durch die Verknüpfung von Ψ_s und Q_i in Gleichung (2.41) zeigt sich, dass die Auswirkungen der Alterung Q_{age} arbeitspunktabhängig sind. Die Modellierung im Transistormodell selbst erfolgt durch die Erhöhung von Q_{fc}. Dadurch verändert sich die Flachbandspannung gemäß Gleichung (2.32). Zudem verändert sich die Ladungsverteilung im Transistor in Gleichung (2.30). Somit ändern sich das Oberflächenpotential und die Inversionsla-

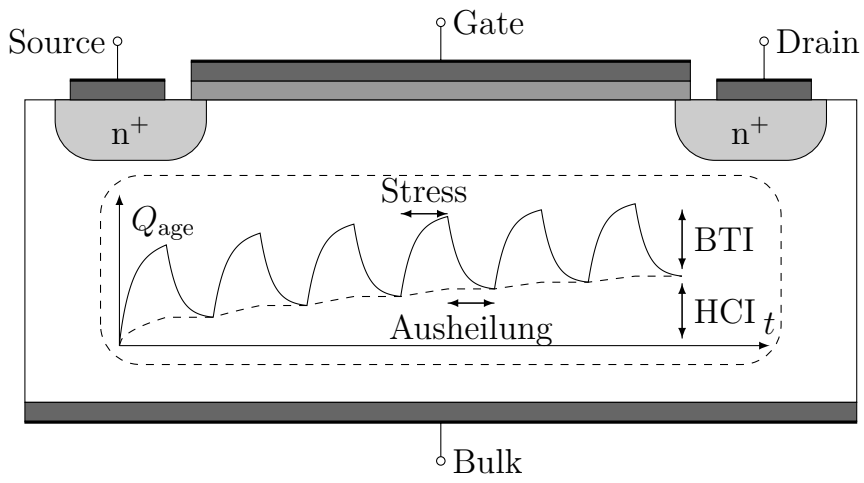


Abbildung 3.3: Änderung der alterungsabhängigen Ladung durch Stress- und Erholungsphasen innerhalb des Transistors.

dung. Daraus folgt eine Änderung des q_{ini} als Ausgangspunkt für die Berechnung des Verilog-A-Transistormodells wie in Abschnitt A.5 gezeigt ist. Durch die Funktionalität des Transistormodells ändern sich als Folge der Alterung sämtliche betroffene Transistorparameter wie z.B. die Gate-Ladung, Gate-Kapazität, Transkonduktanz. Die erneute Berechnung findet zu sämtlichen Simulationszeitpunkten und zu iterativen Zwischenschritten des Simulators statt. Somit können sämtliche transiente Eigenschaften der Alterung direkt beobachtet werden. Zudem kann die Auswirkung der Alterung in allen SPICE-Simulationsarten untersucht werden.

Um mit herkömmlichen Alterungsmodellen vergleichbar zu sein, kann die Ladung $Q_{\text{age}}(t, \mathcal{P}, \mathbb{T})$ in eine Verschiebung der Schwellspannung mit

$$\Delta V_{th} \approx \frac{Q_{\text{age}}}{C_{ox}} \ , \tag{3.3}$$

wie in Abbildung 3.4 gezeigt ist, umgerechnet werden. Der Punkt an dem $Q_{si} \approx Q_i$ gilt, verschiebt sich durch die Alterung und somit V_{th}. Die Gatekapazität C_{ox} ist ihrerseits von Prozessvariationen abhängig, daher kann diese Approximation zu einer Überschätzung der Variabilität der Alterung führen. Zusammen mit Strahlungseffekten ergibt sich die neu eingeführte Ladungskomponente zu

$$Q_{age}(t, \mathcal{P}, \mathbb{T}) = Q_{HCI}(t, \mathcal{P}, \mathbb{T}) + Q_{BTI,str}(t, \mathcal{P}, \mathbb{T}) - Q_{BTI,rec}(t, \mathcal{P}, \mathbb{T}) + Q_R(t, \mathcal{P}, \mathbb{T}, TID) \ .$$

Dabei sind $Q_{HCI}(t, \mathcal{P}, \mathbb{T})$ und $Q_{BTI,str}(t, \mathcal{P}, \mathbb{T})$ die durch beide Effekte produzierten zusätzlichen Ladungen. $Q_{BTI,rec}(t, \mathcal{P}, \mathbb{T})$ modelliert den anteiligen Ausheilungsprozess und $Q_R(t, \mathcal{P}, \mathbb{T}, TID)$ beinhaltet die kurz- und langfristigen Ladungsträger durch SEU und TID. Die Berechnungen dieser vier Komponenten wird im Folgenden erläutert.

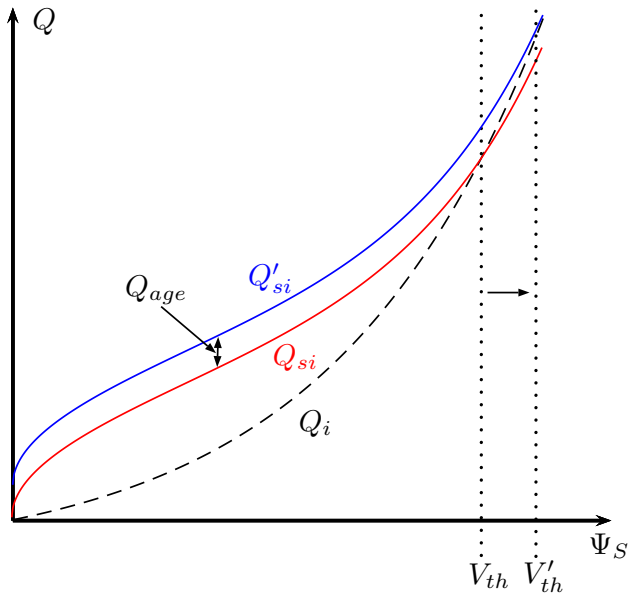


Abbildung 3.4: Verschobenes V_{th} durch die zusätzlichen Ladungen verursacht durch HCI und BTI.

Bias Temperature Instability Aufgrund dessen, dass BTI drei verschiedene Komponenten beinhaltet, die unterschiedliche Stressbedingungen besitzen und nur teilweise reversibel sind, werden diese getrennt modelliert. Der Algorithmus 1 zeigt die implementierte Berechnungen des verwendeten BTI Modells auf Grundlage von [MMM+14]. Für die Berechnung werden die relevanten Transistor- und Umweltparameter durch $\mathcal{P} \in \mathbb{R}^M$ übergeben. Der Arbeitspunkt des Transistors wird charakterisiert durch $\mathbb{T} \in \mathbb{R}^3$. Die initiale Ladung, welche bereits vorhanden ist, ist gegeben durch $\mathbf{Q}_{BTI,0}$, und t_{sim} gibt den betrachteten Zeitraum an und wird vom Simulator übergeben. t_{sim} gliedert sich in zwei Komponenten t_s und t_r, die die Stress- und Wiederherstellungszeiträume beinhalten. Für die Beschreibung des degradierten Arbeitspunktes einzelner Transistoren kann die initiale Ladung als Parameter verwendet werden. Diese bereits vorhandenen Ladungen werden zuerst initialisiert. Die Stressbedingung selbst hängt von dem elektrischen Feld im Gateoxid ab, da dieses proportional zur Bindungsenergie ist. Anschließend werden die Zuwachsraten der drei Komponenten berechnet und danach über den Stresszeitraum integriert. Die Modellierungsgleichungen entsprechen unterschiedlichen Besetzungwahrscheinlichkeiten der Fehlstellen und somit ihrer Aktivierungsenergien. Dadurch ist $\Delta N_{int,s}$ durch ein logarithmisches Wachstum, $\Delta N_{trap,s}$ durch ein exponentielles Wachstum und $\Delta N_{ox,s}$ durch eine Potenzfunktion dargestellt. Die unterschiedlichen mathematischen Modellierungen begründen sich in den Zeitkonstanten des gewünschten Verhaltens. Durch die logarithmische Abhängigkeit steigt die Ladung zu Beginn schnell an. Dies wird verwendet, um die Fehlstellen mit geringen Zeitkonstanten darzustellen, diese müssen in der Lage sein, im Bereich von Piko- bis Mikrosekunden Fehlstellen zu modellieren. Dies ist insbesondere für digitale Schaltungen mit steilen Flanken interessant. Die Fehlstellen tief im Oxid haben höhere Zeitkonstanten und

Algorithmus 1 BTI-Modell

Eingabe: $\mathcal{P} \in \mathbb{R}^M$, $\mathbb{T} \in \mathbb{R}^3$, $\mathbf{Q}_{BTI,0}$, $t_{sim} \in \mathbb{R}^{\nu \times 2}$
Ausgabe: $\hat{\mathbf{Q}}_{BTI}$

1: $\hat{\mathbf{Q}}_{BTI}(\mathcal{P}) \leftarrow \mathbf{Q}_{BTI,0}$
2: **if** $\mathrm{E_{ox}}(\mathbb{T}, \mathcal{P}) \geq 0.5\,{}^{MV}/_{\mathrm{cm}}$ **then**
3: $\quad \Delta N_{int,s} = R_{str}(\mathcal{P}) \cdot (\tau_s(\mathbb{T}) + t_s))^{-1}$
4: $\quad \Delta N_{trap,s} = N_{trap,0}(\mathcal{P}) \cdot \tau_c(\mathbb{T}))^{-1} \cdot \exp(-t_s/\tau_c(\mathbb{T})))$
5: $\quad \Delta N_{ox,s} = N_0(\mathcal{P}) \cdot n_{exp}(\mathbb{T}) \cdot t_s^{1-n_{exp}(\mathbb{T})}$
6: $\quad \mathbf{Q}_{BTI}(1) = \int_0^{t_s} \Delta N_{int,s}\,\mathrm{d}t_s$
7: $\quad \mathbf{Q}_{BTI}(2) = \int_0^{t_s} \Delta N_{trap,s}\,\mathrm{d}t_s$
8: $\quad \mathbf{Q}_{BTI}(3) = \int_0^{t_s} \Delta N_{ox,s}\,\mathrm{d}t_s$
9: **else**
10: $\quad \Delta N_{int,r} = \mathbf{Q}_{BTI} \cdot R_r(\mathcal{P}) \cdot (t_r + \tau_r(\mathbb{T}))^{-1}$
11: $\quad \Delta N_{trap,r} = \mathbf{Q}_{BTI} \cdot \tau_e(\mathbb{T})^{-1} \cdot \exp(-t_r/\tau_e(\mathbb{T}))$
12: $\quad \mathbf{Q}_{BTI}(1) = \mathbf{Q}_{BTI}(1) - \int_0^{t_r} \Delta N_{int,r}\,\mathrm{d}t_r$
13: $\quad \mathbf{Q}_{BTI}(2) = \mathbf{Q}_{BTI}(2) - \int_0^{t_r} \Delta N_{trap,r}\,\mathrm{d}t_r$
14: **end if**
15: $\hat{\mathbf{Q}}_{BTI} \leftarrow \sum_i \mathbf{Q}_{BTI}[i]$

sind durch die exponentielle Abhängigkeit für kurze Zeiträume als träge modelliert. $\Delta N_{int,s}$ modelliert dadurch Fehlstellen mit sehr geringen Bindungsenergien, die bereits im Bereich von Piko- und Nanosekunden besetzt werden. $\Delta N_{trap,s}$ wächst dagegen erst für längere Zeiträume im Mikrosekundenbereich signifikant an. Schließlich modelliert $\Delta N_{ox,s}$ das permanente oder Langzeitverhalten. Im Gegensatz zu den beiden anderen Komponenten besitzt $\Delta N_{ox,s}$ keine Ausheilungsprozesse. Durch diese unterschiedlichen Zeitfenster kann das Modell als eine Capture-Emission-Time-Map aufgefasst werden [GWR+11], die in drei größere Bereiche unterteilt ist und nicht einzelne Fehlstellen, sondern alle Fehlstellen mit ähnlichen τ_e und τ_c als eine Ladung zusammenfasst. Die weiteren Faktoren und Exponenten sind vom Arbeitspunkt und Transistorparameter abhängig, ihre Abhängigkeiten sind im Anhang A.7 beschrieben, und modellieren die Wechselwirkungen zwischen den einzelnen Einflüssen. Für den Fall einer Ausheilungsphase regenerieren sich die Fehlstellen mit den gleichen mathematischen Abhängigkeiten wie im Fall von Stress. Somit heilen einige Fehlstellen schneller aus als andere, was dazu führt, dass der Transistor nach einer langen Stressphase deutlich längere Zeit benötigt um auszuheilen. In jedem Simulationschritt wird $\hat{\mathbf{Q}}_{BTI}$ für die Berechnung sämtlicher Transistorparameter innerhalb des BSIM-BULK-Modells berücksichtigt. Die Simulationsschritte können dabei Zeitschritte oder Iterationsschritte durch den Simulator sein,

dieser gibt ebenfalls die Schrittweite vor. So können ebenfalls Rückkopplungsschleifen in den Wechselwirkungen berücksichtigt werden und Effekte, die über der Zeit zu einer Verschlechterung oder Verbesserung der Degradation führen, korrekt dargestellt werden.

Hot Carrier Injection Im Gegensatz zu BTI besitzt HCI nur einen physikalischen Mechanismus, der zur Zerstörung von Bindungen führt. Die dem Ladungsträger bzw. der Bindung zugeführte Energie kann dabei unterschiedliche Ursachen haben. Es wird zwischen der Zerstörung durch einen einzelnen Ladungsträger (SE), mehrere (MVE) und durch Streuung (SC) unterschieden. Jeder dieser Mechanismen besitzt eigene Stressbedingungen. Die Zerstörung durch Streuungsereignisse findet statt, wenn der Transistor nicht leitend ist und die Ladungsträger der Leckströme kollidieren. MVE tritt auf, wenn einige Ladungsträger mit mittlerer Energie vorhanden sind und schließlich SE, wenn der Transistor leitend betrieben wird. Der Algorithmus 2 zeigt die implementierte Berechnung des verwendet HCI-Modells auf Grundlage von [Gra14]. Das ursprüngliche Modell basiert auf dem Drain-Source Sättigungsstrom. Dies wurde für das BSIM-BULK-Modell in eine Abhängigkeit der Sättigungsladung im Kanal übersetzt. Bei diesem Modell handelt es sich um eine phänomenologische Beschreibung der Einflüsse durch HCI. Die einzelnen Stressbedingungen für die Gate- und Drain-Spannung sind Modellparameter, die sich für verschiedene Transistorarten und Herstellungsverfahren unterscheiden können. Die Parameter C_1, C_2, C_3 sind reine Fitting-Parameter, um die funktionale Abhängigkeit an die gemessene Parameterabweichung anzupassen.

Wie beim BTI Modell wurden die Zuwachsraten der jeweiligen Ladungskomponenten modelliert und diese anschließend integriert, um die numerische Stabilität der Simulation zu erhöhen.

Algorithmus 2 HCI-Modell

Eingabe: $\mathcal{P} \in \mathbb{R}^M$, $\mathbb{T} \in \mathbb{R}^3$, $\mathbf{Q}_{HCI,0}$, $t_{sim} \in \mathbb{R}^{\nu \times 2}$
Ausgabe: $\hat{\mathbf{Q}}_{HCI}$

1: $\hat{\mathbf{Q}}_{HCI}(\mathcal{P}) \leftarrow \mathbf{Q}_{HCI,0}$
2: **if** $\mathrm{V_{ds}} \leq 0.01\,\mathrm{mV}$ **then**
3: $\quad \Delta Q_{SC} = C_1 \cdot \left(\frac{q_{d,sat}(\mathbb{T},\mathcal{P})}{W_{eff}}\right) \cdot \left(\frac{I_{ds}(\mathbb{T},\mathcal{P})}{q_{d,sat}(\mathbb{T},\mathcal{P})}\right)^m \cdot n \cdot t_s^{n-1}$
4: **end if**
5: **if** $\mathrm{V_{ds}} \geq 0.01\,\mathrm{V}$ **then**
6: $\quad \Delta Q_{MVE} = C_2 \cdot \sqrt{V_{ds,eff}(\mathbb{T},\mathcal{P})} \cdot \left(\frac{I_{ds}(\mathbb{T},\mathcal{P})}{W_{eff}}\right)^{oszi_{lvl}} \cdot n \cdot t_s^{n-1}$
7: **end if**
8: **if** $(V_{gs} > 0\,\mathrm{V})$ && $(\mathrm{V_{ds}} \geq 0.01\,\mathrm{V})$ **then**
9: $\quad \Delta Q_{SE} = C_3 \cdot \left(\frac{q_{d,sat}}{W_{eff}}\right) \left(\frac{I_{ds}}{q_{d,sat}}\right)^m \cdot n \cdot t_s^{n-1}$
10: **end if**
11: $\hat{\mathbf{Q}}_{HCI} = \int_0^{t_s} \Delta Q_{SC} + \Delta Q_{MVE} + \Delta Q_{SE}\,\mathrm{d}t$

Die Zuwachsraten der einzelnen Mechanismen sind jeweils mit einem Potenzgesetz von der Zeit abhängig, skalieren jedoch mit einzelnen Transistorparametern. Dabei kann die Sättigungsladung nach [EV06, Aga14] ausgedrückt werden als

$$q_{d,sat} = \frac{\lambda_c}{2} \cdot \frac{q_i + q_i^2}{1 + {}^{\lambda_c}/_{2}(1 + q_i)} \ . \tag{3.4}$$

Des Weiteren ist $V_{ds,eff} = V_{gs} - V_{th}$ die Overdrive-Spannung. Diese kann für verschiedene Inversionszustände durch

$$V_{ds,eff} = \begin{cases} 2nU_t \ln\left(\exp(\sqrt{IC}) - 1\right) & \text{allgemein} \\ 2nU_t \ln\left(IC\right) & \text{schwache Inversion} \\ 2nU_t \sqrt{(IC)} & \text{starke Inversion} \end{cases} \tag{3.5}$$

ausgedrückt werden. Diese analytischen Gleichungen werden im Kompaktmodell selbst nicht zur Berechnung verwendet [Aga14]. Sie verdeutlichen jedoch, dass das gewählte Modell die Möglichkeit bietet, einzelne Parameter angenähert analytisch berechnen zu können. Die einzelnen Zuwächse werden integriert und als $\hat{\mathbf{Q}}_{HCI}$ zusammengefasst. Wichtiger Teilaspekt der beiden implementierten Alterungsmechanismen ist, dass sich BTI und HCI in der Nähe des Drainkontaktes die Fehlstellen teilen. Das bedeutet, dass vorhandene Fehlstellen nur durch einen der beiden Prozesse besetzt werden können und dieser für den anderen Prozess nicht mehr zur Verfügung steht. Abbildung 3.5 zeigt die Schätzung der Alterung mit und ohne Interaktion beider Mechanismen. Im Modell teilen sich die BTI-Ladung am Interface und HCI dieselbe maximal mögliche Ladungskomponente. Ist diese erschöpft, werden im Modell durch diese beiden Mechanismen keine weiteren Fehlstellen erzeugt. Die BTI-Komponenten besitzen selbstverständlich eine wiederherstellbare Komponente. So unterscheiden sich die Auswirkungen der Alterung für unterschiedliche Reihenfolgen der Alterung.

Wie zu erkennen ist, kommt es bei der Vernachlässigung der Interaktion zu einer Überschätzung um das Doppelte. Somit ist es zwingend notwendig, dass Alterungsmodelle diese Interaktion zumindest phänomenologisch berücksichtigen. Zudem zeigt dieser Zusammenhang der beiden Alterungsmechanismen, dass eine getrennte Modellierung oder Betrachtung von HCI und BTI spätestens auf Schaltungsebene nicht zielführend ist und zu starken Abweichungen zum tatsächlichen Alterungsverhalten führt. Um die Fähigkeiten des REL-MOS-Modells zu demonstrieren, zeigt Abbildung 3.6 beispielhaft die Verläufe für die Ladung durch HCI bei verschiedenen NMOS-Transistoren mit Prozessvariationen für ein konstantes Stressprofil. Des Weiteren sind PMOS-Transistoren mit Prozessvariationen und BTI bei einer oszillierenden Gatespannung gezeigt. Die Periodendauer der rechteckförmigen Gatespannung ist dabei 100.000 Sekunden. Für BTI ist zudem die Standardabweichung der Ladung gezeigt.

Die Prozessvariationen führen bei der initialen Verteilung zu leicht unterschiedlichen Stressbedingungen bezüglich der Alterung. Daraus resultiert für jeden Transistor ein an-

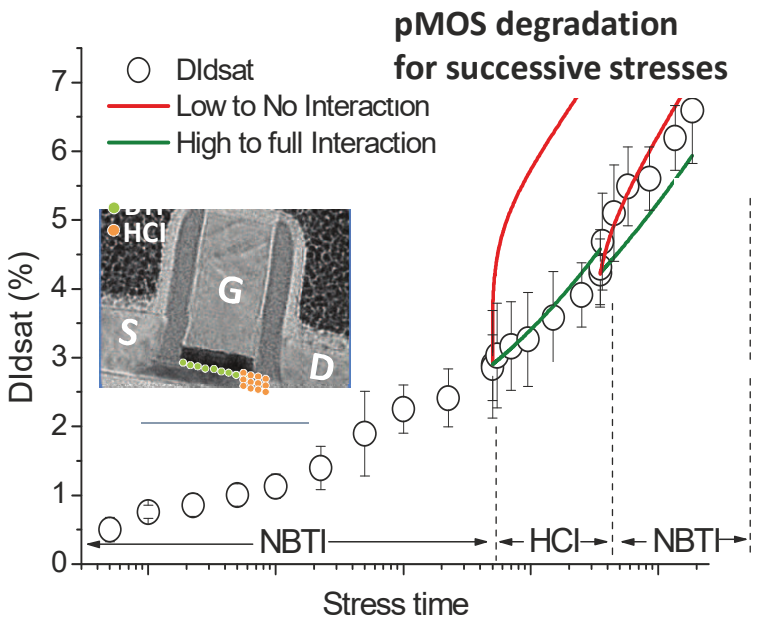


Abbildung 3.5: Interaktion von HCI und BTI bei der Fehlstellenbesetzung in der Nähe des Drainkontaktes. Dies wird durch die Änderung des Sättigungs-Drain-Strom *DIdsat* veranschaulicht. Bildquelle: [HCF+15]

derer Verlauf der Degradierung. Besonders bei BTI sind die Hübe des Ladungszuwachses stark durch diese Einflüsse bestimmt. Die Standardabweichung zeigt, dass diese im Laufe der Zeit immer weiter ansteigt, sowohl für Stress- als auch Ausheilungsphasen. Diese Graphen zeigen die Gesamtfunktionalität des REL-MOS-Modells in Bezug auf Alterung. Um die dargestellte Modellierung der Alterung mit physikalisch begründeten Werten zu vergleichen, zeigt Abbildung 3.7 TCAD-Simulationen und REL-MOS-Ergebnisse für PMOS-Transistoren, die von BTI betroffen sind, für zwei verschiedene Transistorkanallängen über einen Zeitraum von 10 Jahren. In TCAD sind 2D-Simulationen durchgeführt worden, dadurch werden Schmalkanaleffekte vernachlässigt. Die REL-MOS-Transistoren wurden mit W = 1μm so weit gewählt, dass hier ebenfalls Schmalkanaleffekte keine Rolle spielen. Die TCAD-Simulationen benötigen eine enorme Rechenleistung und viel Zeit. Daher wurde ein konstantes Stressprofil ohne Phasen der Ausheilung simuliert.
Wie zu erkennen ist, wird der starke Anstieg sowie die abfallende Änderungsrate durch das REL-MOS-Modell korrekt abgebildet. Zudem wird der absolute Anstieg der Schwellspannung für beide Kanallängen bis auf kleine Abweichungen gut angenähert. Dadurch können die im REL-MOS-Modell dargestellten Abhängigkeiten der Alterung als qualitativ richtig betrachtet werden. Da das REL-MOS-Modell selbst physikalische Zusammenhänge der einzelnen Parameter verknüpft, können die alterungsbedingten Änderungen der anderen Parameter des Modells ebenso als qualitativ richtig angenommen werden.

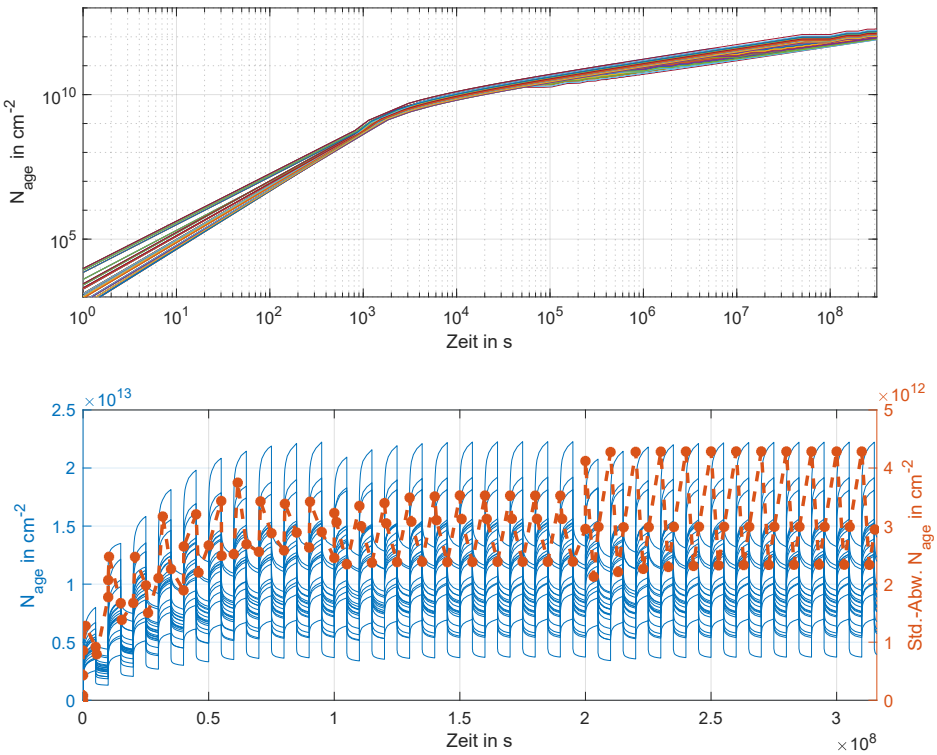


Abbildung 3.6: Simulationsmöglichkeiten der Alterungsmodelle für HCI (oben) und BTI (unten) von NMOS- respektive PMOS-Transistoren mit Prozessvariationen über einen Zeitraum von 10 Jahren. Dargestellt ist die zusätzliche Ladungsträgerkonzentration N_{age} über der Zeit.

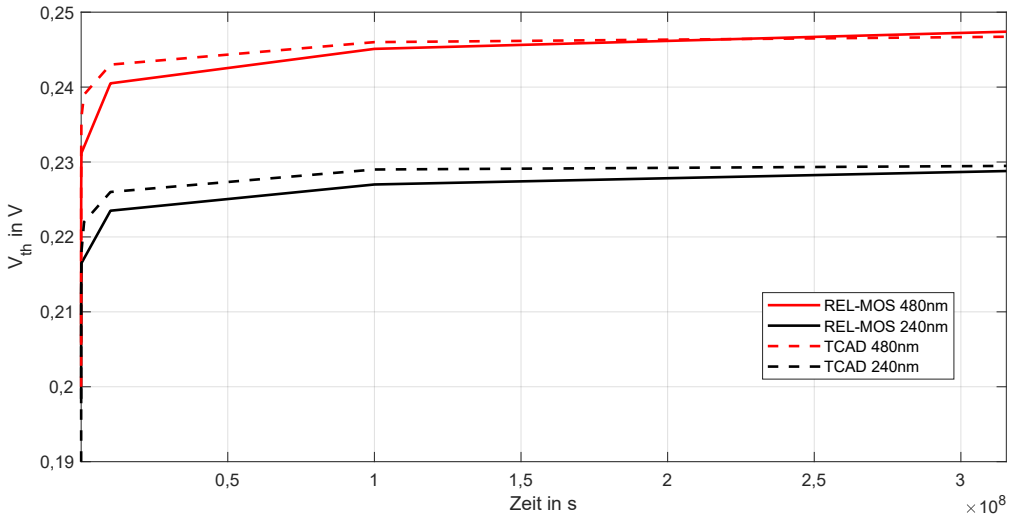


Abbildung 3.7: Vergleich von TCAD-Simulationen mit REL-MOS-Ergebnissen für PMOS-Transistoren, die von BTI betroffen sind, für zwei verschiedene Transistorkanallängen über einen Zeitraum von 10 Jahren.

3.1.2 Strahlungseinflüsse

Bei den Einflüssen durch Strahlung wird Totel Ionizing Dose (TID) ähnlich zu den Alterungseffekten modelliert, da dieser Effekt zu einer Erhöhung von, wenn auch ausschließlich positiven, Ladungsträgern im Gateoxid führt. Da die Ursache der Strahlung nichts mit dem Transistor, seinem Zustand oder der Schaltung zu tun hat, ist diese Modellierung deutlich einfacher. Während der transienten Simulation kann der Landungsträgerzuwachs durch TID festgelegt, während der Simulation kumuliert, und wie Q_{age} im Transistormodell berücksichtigt werden. Die Auswirkung und Stärke von TID hängt dabei von der Gatefläche $A_{it} = W \cdot L$ des Transistors ab. Je kleiner diese Fläche desto unwahrscheinlicher ist der Einschlag eines Teilchens, jedoch steigt die Stärke der Auswirkung eines Einschlags. Die Gleichung, die zur Modellierung von TID nach [ABKF10] verwendet wird, ist

$$Q_{TID} = qN_{ot} + qN_{it} \tag{3.6}$$

$$= q[N_{ot} - D_{it}(\Psi_s - \Phi_b - \Phi_n)A_{it}] \ . \tag{3.7}$$

qN_{ot} sind bereits vorhandene Ladungen und D_{it} ist die flächenbezogene Defektdichte, deren Ableitung $\partial D_{it}/\partial t$ vorgegeben werden muss. Φ_b ist dabei das Potential im Bulk und Φ_n das Potential der Ladungsträger. Die Differenz $\Psi_s - \Phi_b - \Phi_n$ gibt die Besetzungswahrscheinlichkeit der Defekte an und ist durch das Oberflächenpotential bestimmt, welches wiederum vom Arbeitspunkt, Alter des Transistors und Prozessvariationen abhängt. Beim SEU hängt die Stärke der Auswirkung vom Einschlagsort des Teilchens sowie den Eigenschaften des naheliegenden Transistors ab. Die Auswirkung wird durch einen zusätzlichen Drain-Strom, nicht Drain-Source-Strom, modelliert. Der zusätzliche Strom I_{SEU} tritt innerhalb des Transistors auf und kann nicht als externe Quelle modelliert werden, da der Strom je nach Arbeitspunkt des Transistors zum Source, Drain oder Bulk-Anschluss abfließen kann und somit die Auswirkungen auf das Verhalten des Transistors sehr unterschiedlich sind. Die Modellierungsgleichungen für einen NMOS-Transistor ergeben sich nach [ABKF10] zu:

$$I_{SEU} = \pi r^2 (J_n + J_p) \tag{3.8}$$

$$J_n = qD_n \nabla \rho_n + q\rho_n \mu_n E \tag{3.9}$$

$$J_p = -qD_p \nabla \rho_p + q(N_A + \rho_p)\mu_n E \ . \tag{3.10}$$

Darin sind $\rho_\square$ die Elektronen-Loch-Paare, die durch den Ionenkanal erzeugt werden. Diese führen zum folgenden Gradienten:

$$\Delta\rho_{n,p} = \frac{2\rho}{\pi R^3} \sum_k \lambda^{-2} \exp\left(\frac{-\lambda^2(2k-1)^2\pi}{4} - \frac{r^2}{t_{si}^2 \lambda^2}\right) \ , \tag{3.11}$$

dabei ist

$$\lambda = \sqrt{4D_{n,p} \cdot \frac{t}{R^2}} \tag{3.12}$$

und

$$R = \frac{W}{L}\sqrt{k_{x1} \cdot X^{k_{x2}}} \ . \tag{3.13}$$

k_{x1} und k_{x2} sind Parameter, die aus Messungen oder TCAD-Simulationen extrahiert werden können. Sie modellieren den Einfluss des Abstands zwischen Ionenkanal und Transistorstruktur.

Anmerkungen zur Verilog-A-Implementierung Damit die Best-Practice-Regeln für Verilog-A-Implementierungen eingehalten werden, wurden von allen Ausgangsgrößen jeweils die Änderungsraten modelliert und diese integriert:

$$Q_\square = \int_0^t \underbrace{\frac{\partial}{\partial t} Q_\square}_{\Delta Q_\square} \, \mathrm{d}t \ . \tag{3.14}$$

Dies hat den Vorteil, dass die Ausgangsgrößen immer mindestens einmal stetig differenzierbar sind, was wichtig für die Berechnungen von Strömen, Spannungen und Kapazitäten ist. Zudem wurden alle *If*-Abfragen durch logistische Funktionen der Form

$$f(\mathcal{P}) = \frac{1}{1 + \mathrm{e}^{-k(\mathcal{P} - \mathcal{P}_0)}} \tag{3.15}$$

ersetzt. k ist der Wachstumsfaktor und gibt die Steigung vor, $\mathcal{P}$ stellt die jeweilige Bedingung dar, z.B. V_d, V_g, und $\mathcal{P}_0$ ist der Schwellwert, ab dem die Entscheidung getroffen werden soll. Diese agieren als ein geglätteter Schalter und verhindern Simulationsartefakte und Unstetigkeiten. Dadurch können im Modell leichte Abweichungen entstehen [MCG+15].

3.1.3 Gesamtmodell

Das Modell wurde mit einem prädiktiven PDK für einen Bulk-CMOS-Prozess in 65nm-Technologie gefittet. Die einzelnen Parameter der Alterungs- und Strahlungseffekte sind ebenfalls prädiktiv und wurden aus den entsprechenden Veröffentlichungen übernommen. Das REL-MOS-Modell wurde innerhalb von Cadence Virtuoso getestet und entwickelt. Die besonderen Eigenschaften sind:

1. Alle Einflüsse in einer transienten Simulation gleichzeitig betrachten
2. Alterungsimulationen anhalten, abspeichern und fortsetzen

CDF Parameter	Value	Display
t_irradation_start	1n	off
X_LET	X_LET	off
t_si	1u	off
Rad_r	Rad_r	off
LET	LET	off
Rad_sw	1	off
L	L M	off
W	W M	off
HCI_sw_n	nHCI	off
BTI_sw_n	nBTI	off
HCI_Q_int_ini_n	0	off
BTI_Q_ox_ini_n	0	off
BTI_Q_trap_ini_n	0	off
BTI_Q_int_ini_n	0	off

Abbildung 3.8: Erweitere Modellparametereingabe während der Schaltplanerstellung in Cadence Virtuoso.

3. Kombination von Alterung und SEU in einer Simulation, da dies zwischen Zeiträumen von Piko- und Megasekunden variiert

4. Verwendung in Standard-Entwurfsprozess

5. Kombination von Monte-Carlo, DC, AC und transienten Simulationen zur Analyse der Zuverlässigkeit

Durch die erweiterte Funktionalität ändern sich zudem die Modellparameter, die über die Schaltfläche im Simulator einstellbar sind. Dazu zeigt Abbildung 3.8 die erweiterte Modellparametereingabe während der Schaltplanerstellung in Cadence Virtouso. Somit bettet sich die deutlich erweiterte Funktionalität des Modells direkt in den üblichen Designablauf ein. Jeder einzelne Transistor einer Schaltung verfügt über diesen erweiterten Parametersatz, somit können beliebige Kombinationen von Analysen realisiert werden. Die ersten 6 Parameter steuern das Strahlungsmodell, wobei *Rad_sw* angibt, ob dieser Einfluss betrachtet werden soll. *t_irradiation_start* definiert den Zeitpunkt, zu dem ein Partikeleinschlag simuliert werden soll. *X_LET* gibt den Abstand zwischen dem Einschlag und der MOS-Struktur an. *t_si* ist die Dicke des Substrats, *Rad_r* der Radius des Einschlags und *LET* die durch den Partikeleinschlag übertragende Energie. *HCI_sw* und *BTI_sw* sind ebenfalls Schaltervariablen, wenn diese boolschen Werte auf 0 gesetzt werden, wird die entsprechende Alterung nicht berücksichtigt und der Transistor verhält sich nominell wie im PDK definiert. Die _ini_-Werte stellen die jeweiligen zum Start der Simulation bereits vorhandenen alterungsabhängigen Ladungskomponenten dar. Eine interessante Eigenschaft des BSIM-BULK-Modells ist, dass die Temperatur eines jeden Transistors durch z.B. Selbsterhitzung getrennt beeinflusst werden kann. Somit besteht die Möglichkeit, in einer einzigen transienten Simulation

einer Schaltung das Temperaturverhalten der Alterung von einzelnen Transistoren zu beobachten. Aus praktischen Gründen müssen die extrahierten Werte mit dem Faktor $6 \cdot 10^{10}$ normiert werden, da diese Werte sonst die SPICE interne Blowup-Grenze für elektrische Modellparameter überschreiten. Der nominelle Simulationsaufwand steigt durch die zusätzlichen Modellgleichungen. Jedoch kann dies durch das weitere methodische Vorgehen, das in dieser Arbeit beschrieben wird, im Gesamtentwurfs-Prozess ausgeglichen werden. Die einzelnen Aussagemöglichkeiten der Aufzählung sind in Abbildung 3.9 illustriert.

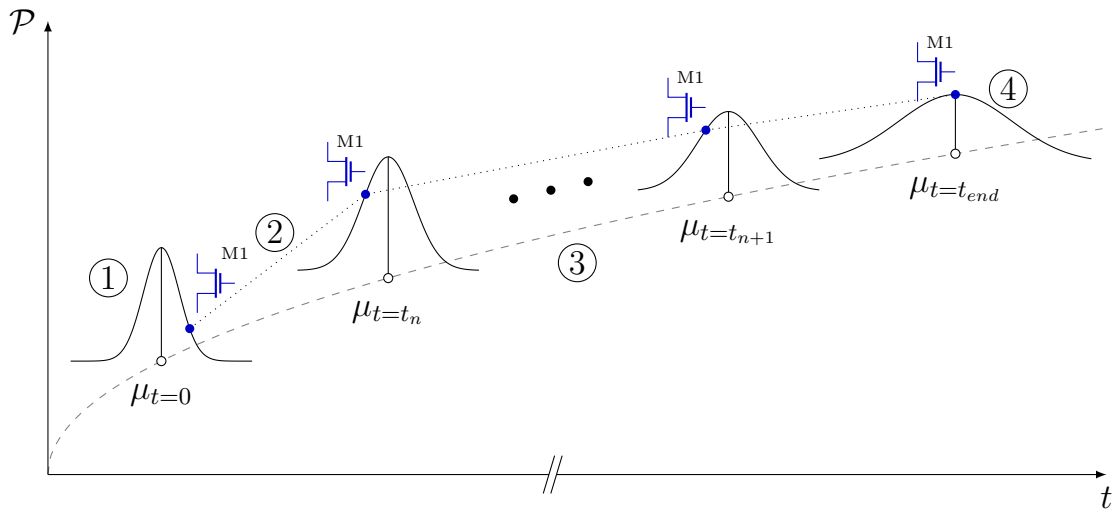


Abbildung 3.9: Verschiedene statistische Aussagen aus den Verteilungsfunktionen, die mit dem REL-MOS-Modell ermittelt werden können.

1. Initiale Verteilung durch Prozessvariationen von Schaltungs- und Bauteilparametern. Gleichzeitig können ebenfalls einzelne Transistoren getrennt von der Grundgesamtheit betrachtet werden, um z.B. Ausläufer der Verteilungen zu untersuchen.

2. Analyse des individuellen Alterungsverhalten einzelner Transistoren in Abhängigkeit von Stressbedingungen, Belastungsprofilen, Umgebungseinflüssen und der Variabilität der Alterung selbst.

3. Veränderung der statistischen Momente der Verteilungsfunktion der Stichprobe über längere Zeiträume. Diese Informationen können z.B. in Entwurfsmethoden verwendet werden.

4. Verteilungsfunktion am Ende der Lebensdauer der Schaltung als auch jeweils den vollständig individuellen Verlauf der Alterung einzelner Transistoren.

Die aufgezeigten Vorteile und Möglichkeiten des REL-MOS-Modells sind nicht kostenlos. Es erfüllt zwar alle geforderter Standards, jedoch steigt der Messaufwand durch die

betrachteten Effekte im Vergleich zu einem herkömmlichen Modell deutlich an. Wobei moderne Transistormodelle bereits mehrere hundert Transistorparameter besitzen können. Somit könnte ein modularer Ansatz, bei dem nur relevante Effekte berücksichtigt werden, eine Lösung sein. Des Weiteren steigt der Simulationsaufwand bei der Analyse von Alterung und Strahlung enorm an. Die Periodendauer kann zwar skaliert werden, um den Aufwand etwas zu verringern, dennoch steigt die Simulationsdauer insgesamt an. Für jede Anwendung muss abgewogen werden, ob der Simulationsaufwand den Erkenntnismehrwert rechtfertigt.

3.2 Stochastischer Schaltungsentwurf

Das entwickelte Transistormodell kann innerhalb von Analyse- und Entwurfsabläufen von integrierten Schaltungen auf verschiedene Weise zur Anwendung kommen. Zum einen können Schaltungen auf ihre Anfälligkeit gegenüber den beschriebenen Einflüssen untersucht werden. Um diese Analyse durchführen zu können, muss eine entworfene und dimensionierte Schaltung vorliegen. Zum anderen können die REL-MOS-Modelle für den initialen Schaltungsentwurf verwendet werden. Dabei werden aufbauend auf der g_m/I_D-Methode im Vorfeld simulierte Werte einzelner Transistoren mittels der Kleinsignalgleichungen der Schaltungen für den Schaltungsentwurf genutzt. Beide Anwendungsszenarien des REL-MOS-Modells werden in den nachfolgenden Abschnitten beschrieben.

3.2.1 Analyse von integrierten Schaltungen

Wie bereits beschrieben, ist die Analyse der PVTAR-Einflüsse auf eine integrierte Schaltung von entscheidender Bedeutung, um eine Überdimensionierung auf der einen oder eine Verletzung von Spezifikationen auf der anderen Seiten zu vermeiden [GRB+09, FKT13]. Der Designkorridor, der aus den Spezifikationen entsteht, wird durch Prozessvariationen geschmälert und verschiebt sich über der Zeit durch Alterungseinflüsse. Somit muss ein Transistormodell in der Lage sein, eben diesen Designkorridor durch geeignete Simulationen und Analysen zu erfassen. Dies gilt für einzelne Transistoren, Schaltungen und ganze Systeme.
Abbildung 3.10 zeigt die Mittelwerte und Standardabweichungen des Inversionskoeffizienten über V_d, V_g, W und L. Da der Inversionskoeffizient für die Definition des Arbeitspunktes des Transistors verwendet wird, ist sein Verlauf über diese Designparameter und die Abweichungen durch Prozessvariationen für die weiteren Betrachtungen entscheidend.
Der Mittelwert des Inversionskoeffizienten bildet über V_g den Verlauf des Stroms der Gatediode nach und zeigt über V_d die Ausgangscharakteristik des Transistors. Dabei ist der IC in starker Inversion am höchsten. Gleiches gilt für die Standardabweichung. Für den Bereich 0,6V$\geq V_g \geq$ 0,3V ist ein erneuter Anstieg der Standardabweichung

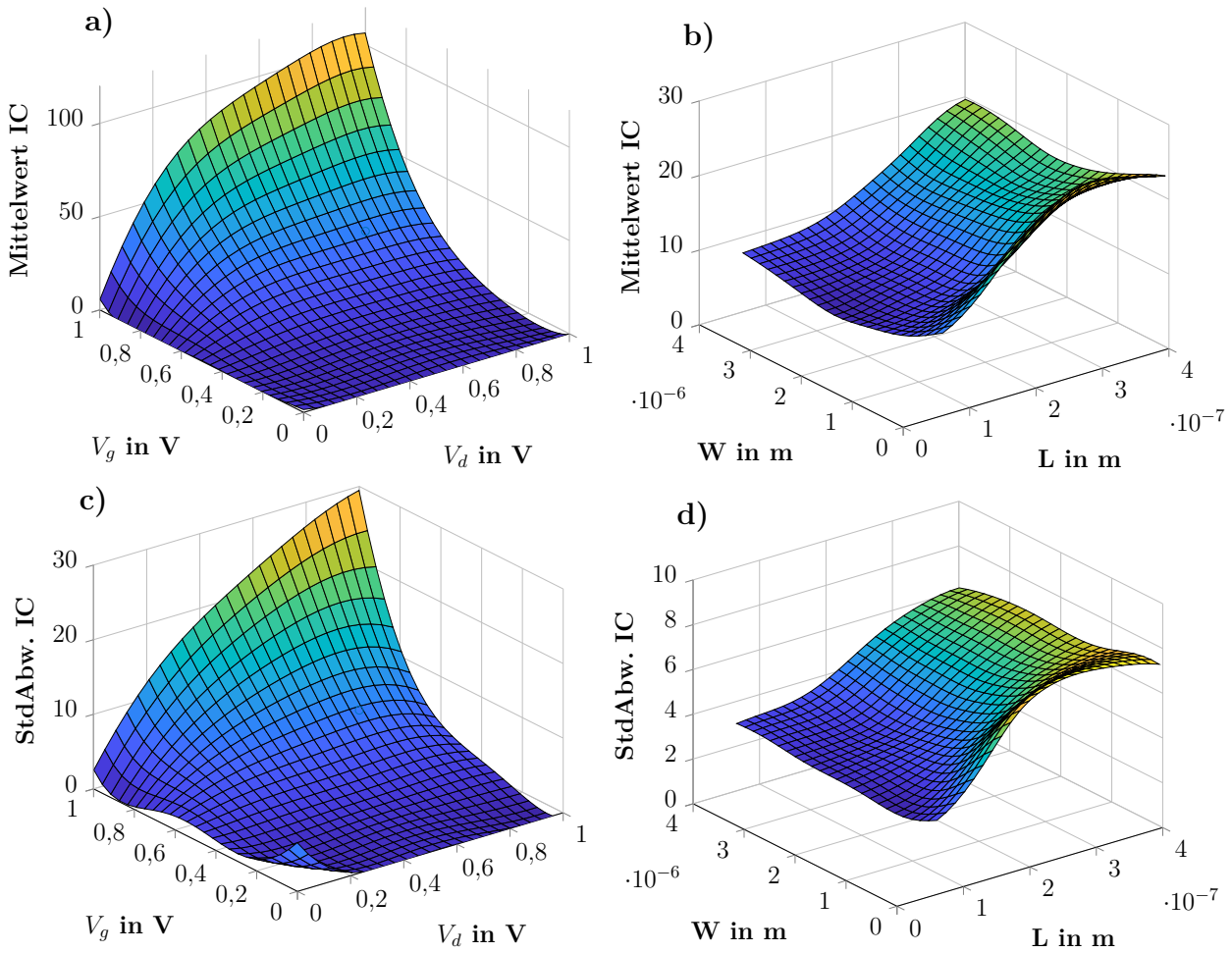


Abbildung 3.10: Inversionskoeffizient eines NMOSFET. a) Mittelwert von IC(V_d,V_g), b) Mittelwert von IC(W,L), c) und d) die entsprechenden Standardabweichungen bei Berücksichtigung von Prozessvariationen. Die jeweiligen Mittelwerte der Wertepaare sind für die jeweils andere Darstellung gewählt.

zu erkennen. Dies ist durch die erhöhte Auswirkung der Prozessvariation in moderater Inversion zu erklären. Für sehr kleine Werte beider Spannungen steigt die Standardabweichung abermals an. Da der IC verwendet wird, um den Inversionszustand eines Transistors eindeutig in drei Kategorien zu unterteilen, sind solche Bereiche besonders relevant, in denen die Prozessvariationen zu einer Verschiebung des Inversionszustands führen können. Dies würde bedeuten, dass sich der Arbeitspunkt bereits direkt nach der Herstellung verschiebt und somit seine Funktionalität deutlich vom spezifizierten Verhalten abweicht. Die Bereiche, in denen dies auftreten kann, liegen zwischen $20 \geq IC \geq 0{,}01$. Je nach Stärke der Prozessvariationen kann dabei der Arbeitspunkt zwischen moderater und starker Inversion sowie moderater und schwacher Inversion wandern. In sehr starker Inversion sowie in sehr schwacher Inversion sind die Standardabweichungen zwar relativ hoch, jedoch nicht hoch genug, um einen Wechsel des Arbeitspunktes zu verursachen.

Der Inversionskoeffizient sollte nach Gleichung (2.57) von der Weite und der Länge des Transistors unabhängig sein. Wie dargestellt, gilt diese Annahme jedoch ausschließlich für Bereiche, in denen keine Sekundäreffekte wie Kurz- oder Schmalkanal-Effekte auftreten. Dies beschränkt die Anwendbarkeit des Inversionskoeffizienten während des Designs auf Transistoren für die gilt $W \geq 2 \cdot W_{min}$ und $5 \cdot L_{min} \geq L \geq 2 \cdot L_{min}$. Dadurch können sehr schnelle Schaltungen, bei denen meist $L \approx L_{min}$ angestrebt wird und sehr energiesparende Schaltungen, für die große Kanallängen benötigt werden, nicht korrekt über IC dimensioniert werden. Besonders zu beachten sind dabei die Verschiebungen der Grenzen zwischen den Inversionsbereichen. So arbeiten sehr schmale Transistoren bereits vor der Grenze von $IC = 0{,}1$ in moderater und vor $IC = 10$ in starker Inversion. Dies kann während des Entwurfsprozesses zu einem systematischen Fehler bei der Auslegung der Bauelemente führen.

Neben den Prozessvariationen ist die Alterung und die damit einhergehende Verschiebung entwurfsrelevanter Parameter entscheidend. Hinzukommt die zusätzliche Variabilität, der Quotient aus Standardabweichung und Mittelwert, welche über den Nutzungszeitraum der Schaltung von hoher Bedeutung ist. Abbildung 3.11 zeigt die Verläufe von Schwellspannung V_{th}, Inversionskoeffizient, Kanallängenmodulationsfaktor λ_c und Transkonduktanz g_m eines NMOS (L=150nm, W=1μm) über 10 Jahre für verschiedene Gate-Spannungen in Sättigung. Ein Transistor wurde jeweils über die gesamte Zeit transient simuliert. Die elektrischen Spannungen sind, in Abhängigkeit des jeweiligen Arbeitspunktes, über die gesamte Simulationsdauer konstant. Somit altert jeder initiale Arbeitspunkt individuell den Stressbedingungen entsprechend. Wie an den Graphen des Inversionskoeffizienten zu erkennen ist, korrespondieren die verschiedenen Gate-Spannungen mit Inversionszuständen des Transistors. Der untersuchte Transistor wird in allen drei Bereichen betrieben, wobei dieser sich bei vier Verläufen in starker Inversion befindet. bei Schwellspannung und Inversionskoeffizient ist zu erkennen, dass die Stärke der jeweiligen Auswirkung der Alterung proportional zur Gate-Spannung bei beiden Parametern steigt. In starker Inversion fällt der IC um bis zu 10, wobei V_{th} um 7mV steigt. Beides beschreibt eine Abnahme des Drainstroms, welche durch HCI bei NMOS zu erwarten ist. Für λ_c, welches die Geschwindigkeitssättigung charakterisiert, ist ein anderes Verhalten zu beobachten. Für alle Betriebspunkte außerhalb der starken Inversion zeigt sich über das Alter keine Veränderung. Die Ursache dafür ist, dass der Parameter in diesen Arbeitspunkten keine Auswirkung auf das Verhalten des Transistors hat, da die Sättigungsgrenze der Ladungsträgergeschwindigkeit erst gar nicht erreicht wird. Wie erwartet fällt λ_c mit steigender Gate-Spannung, was einem Anstieg der Sättigungsgrenze gleichkommt. Über die Zeit ist in starker Inversion ein Anstieg von λ_c zu beobachten. Durch die Alterung fällt die maximale Geschwindigkeit im Kanal, daher ist dieses Verhalten zu erwarten. Bei der Transkonduktanz zeigt sich, dass dieser Parameter für Arbeitspunkte in moderater Inversion am stärksten durch Alterung beeinflusst wird. In schwacher und starker Inversion zeigt sich hingegen für das g_m nahezu keine Abweichung. Besonders die moderate Inversion ist ein häufig

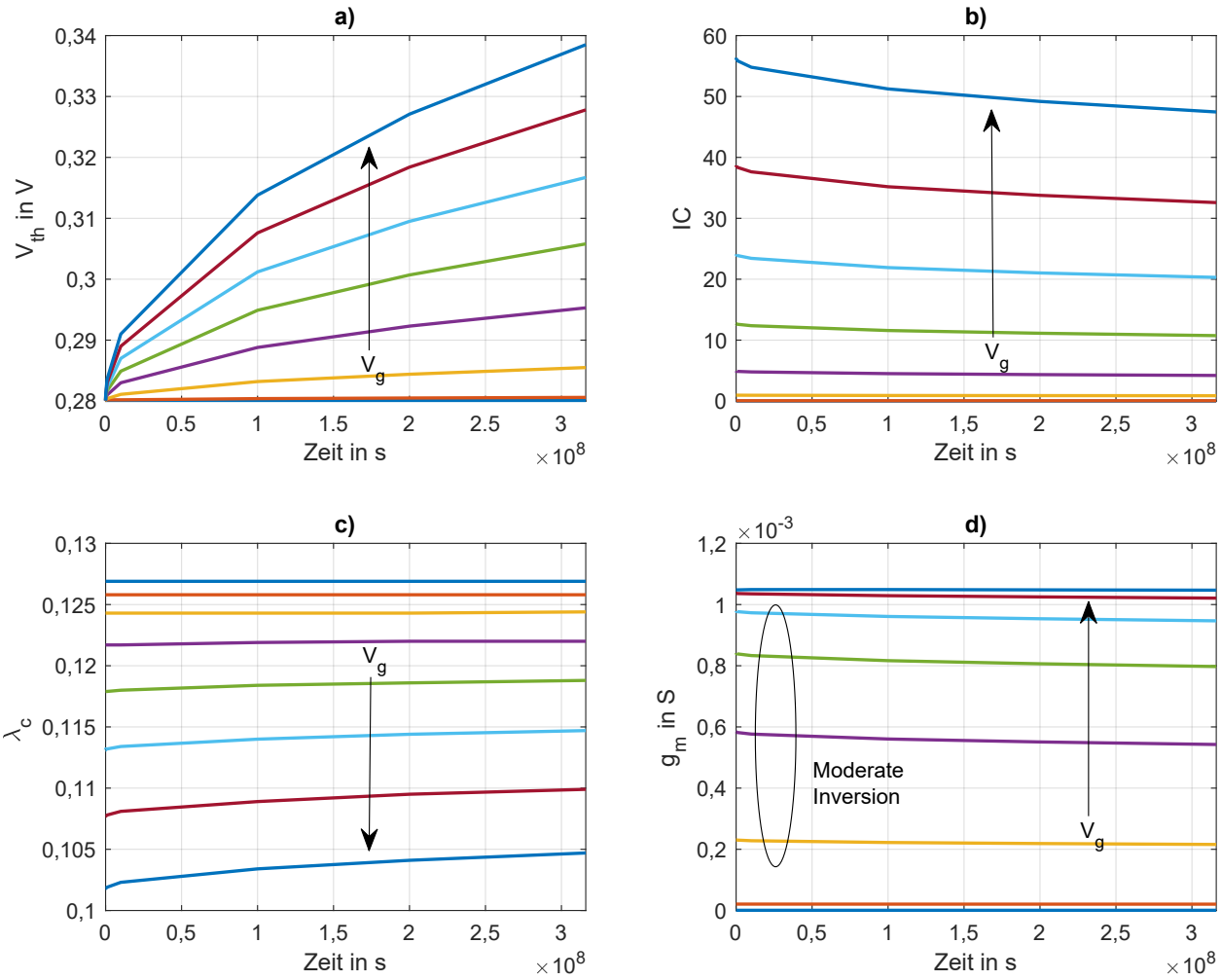


Abbildung 3.11: Degradierung der Schwellspannung V_{th} a), Inversionskoeffizient IC b), λ_c c) und Transkonduktanz g_m d) eines NMOS über 10 Jahre für verschiedene Gate-Spannungen.

verwendeter Arbeitspunkt, da dort die Transkonduktanzeffizienz g_m/I_D am höchsten ist. Gleichzeitig würde ein solcher Arbeitspunkt stärkste Auswirkungen durch Alterung und Prozessvariation bedeuten.

Bereits bei dieser sehr beschränkten Auswahl an Parametern zeigt sich, dass die Alterungseinflüsse keinem klaren Trend folgen. Zudem beeinflusst der Arbeitspunkt des Transistors, respektive der Inversionszustand, ob und wie stark die Alterung den jeweiligen Parameter ändert. Gleichzeitig kann eine starke Verschiebung durch Alterung bedingt durch den Arbeitspunkt kaum Auswirkung auf die Funktion des Transistors haben, wie am Beispiel des IC in starker Inversion gezeigt. Dahingegen können bereits leichte Abweichungen in moderater Inversion zu einer vollständigen Verschiebung des Arbeitspunktes und des Ausgangsverhalten des Transistors führen.

Durch die Vielzahl von Wechselwirkungen ist eine eindeutige Beschreibung der Auswirkungen der Alterung und Prozessvariation durch einen einzigen Parameter nicht

möglich. Bisherige Versuche, die Alterung allein durch die Schwellspannung zu charakterisieren, sind an bestimmte Transistormodelle gebunden und können dadurch nicht alle aufgezeigten Einflüsse der Alterung beschreiben. Sämtliche Auswirkungen der Alterung werden im REL-MOS-Modell physikalisch motiviert durch Q_{age} dargestellt, dessen Verhalten sich an physikalisch korrekte Capture-Emission-Time-Maps anlehnt. Somit kann die Modellierung der Ursache auf einen einzigen Parameter beschränkt werden und die korrekte Abbildung sämtlicher Auswirkungen findet durch das Transistormodell statt. Hierdurch reduziert sich der Modellierungsaufwand auf eine minimale Anzahl von Parametern bei gleichzeitiger allgemeiner Verwendbarkeit. Abbildung 3.12 zeigt die alterungsabhängige Ladung Q_{age} für schwache, moderate und starke Inversion für drei verschiedene Alterungszustände und 200 Transistoren, die von Prozessvariationen betroffen sind. Wie gezeigt, ist die Verbreiterung der Verteilung der zusätzlichen La-

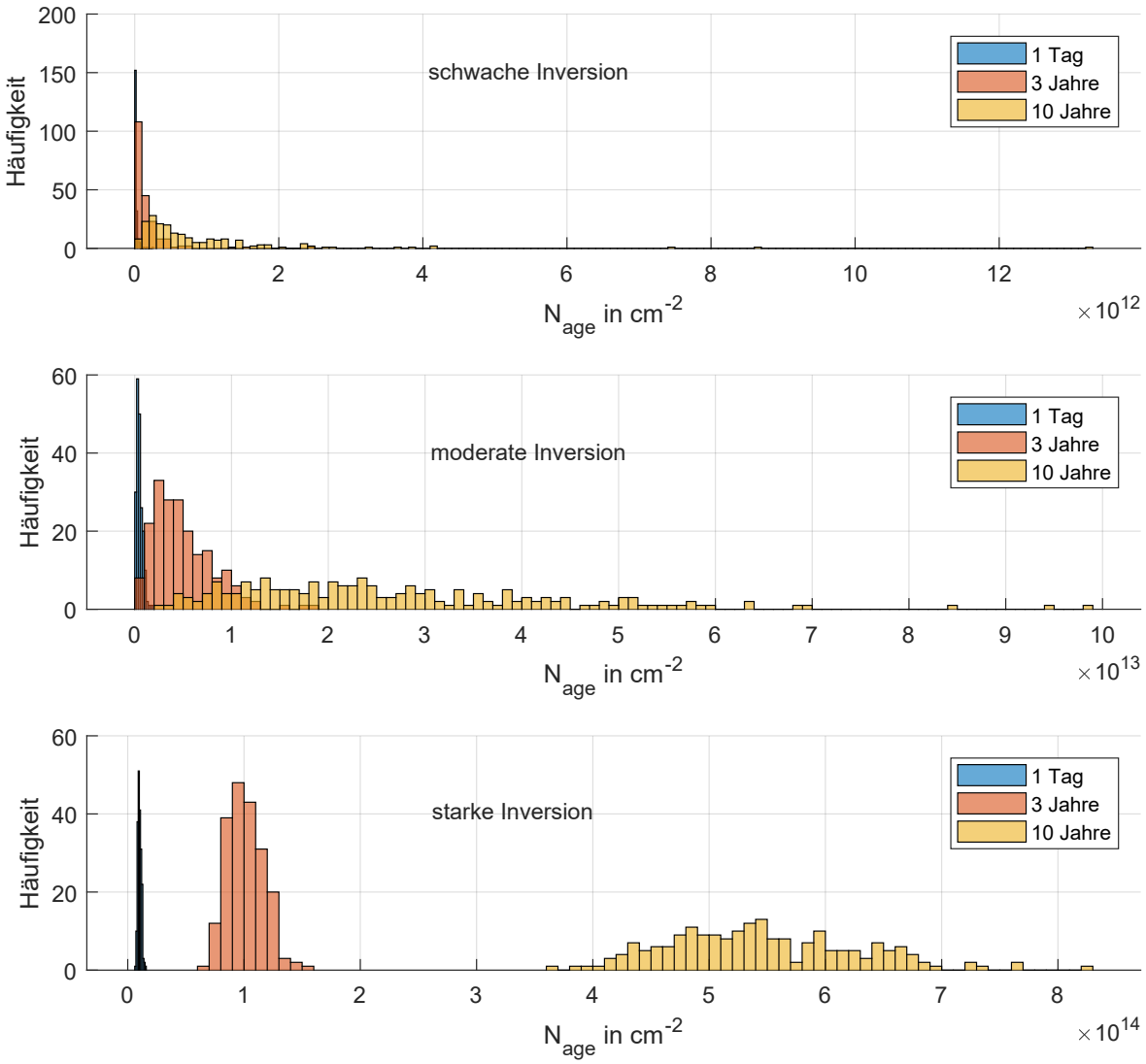


Abbildung 3.12: Alterungsabhängige Ladung N_{age} für schwache, moderate und starke Inversion (von oben nach unten) für drei verschiedene Alterungszustände.

dungsträgerdichte N_{age} nach einem Tag für die betrachteten Betriebsbereiche (schwache, moderate und starke Inversion) äußerst gering. Dies lässt darauf schließen, dass in solch kurzen Alterungsperioden kaum Unterschiede im Alterungsverhalten der Transistoren festzustellen ist. In schwacher Inversion ist nach 3 Jahren und nach 10 Jahren eine sehr deutliche Aufweitung der Verteilung von N_{age} zu erkennen. Wobei sich die Mitte der Verteilung von N_{age} der Alterung der Transistoren nur sehr schwach verschoben hat, gibt es jedoch auch einzelne Transistoren, die mehr als das 12-fache der durchschnittlichen Ladung angehäuft haben. Diese, deutlich stärker gealterten Transistoren, sind den Ausläufern der Verteilungsfunktionen zu finden. Zudem existieren nach 10 Jahren einzelne Transistoren, die nur rund 5% der mittleren Alterung erfahren haben. Maßgeblich für diese Unterschiede sind Abweichungen der Oxiddicke, welche zu anderen Stressbedingungen für die Alterung führt. Vernachlässigbar ist dabei keine der Prozessvariationen. In moderater Inversion ist nach 10 Jahren eine deutliche Verschiebung der Verteilung und eine enorme Aufweitung ebendieser festzustellen. Zudem sind die Auswirkungen in absoluten Zahlen etwa 10 mal so stark wie in schwacher Inversion. Dies resultiert aus dem erhöhten BTI-Stress durch eine höhere Gate-Spannung sowie den durch den höheren Drainstrom angestiegenen Auswirkungen durch HCI. Im Gegensatz zur schwachen Inversion gibt es keine Exemplare, bei denen nach 10 Jahren N_{age} weniger als $2 \cdot 10^{12}\ {}^{1}/_{\mathrm{cm}^2}$ beträgt. Durch diese Spannweite ist die alterungsabhängige Variabilität, bedingt durch die starken Auswirkungen durch die Prozessvariationen, am größten. In starker Inversion zeigt sich, dass die absoluten Auswirkungen am stärksten sind und die Verteilung sich stark verschiebt. Die Aufweitung ist in Relation zu denen der anderen Inversionsbereiche gering. Es zeigen sich drei getrennte Verteilungen über das Alter des Transistors. Nach 3 Jahren ist die Verteilung relativ kompakt um $N_{age} = 1 \cdot 10^{14}\ {}^{1}/_{\mathrm{cm}^2}$ verteilt, was dem Maximalwert in moderater Inversion entspricht. Nach 10 Jahren ist die Verteilung deutlich breiter und der Mittelwert hat sich versechsfacht. Insgesamt zeigt sich ebenfalls beim Verhalten von N_{age}, dass Arbeitspunkt, Prozessvariation und Alterung nicht getrennt betrachtet werden können.

Neben analogen Schaltungen, die im Weiteren fokussiert betrachtet werden, können die REL-MOS Modelle ebenfalls für die Analyse von digitalen Schaltungen und Systemen verwendet werden. Im Rahmen der vorliegenden Arbeit wurden Standardgatter, SRAM, FlipFlops und Oszillatoren auf PVTA Einflüsse hin untersucht. Die Ergebnisse können anschließend verwendet werden, um die Einflüsse durch weitere Modellierungsschritte auf höheren Abstraktionsebenen zu betrachten und somit für die Auswahl geeigneter Schaltungstypen angewandt werden.

3.2.2 Erweiterte $g_{\mathrm{m}}/I_{\mathrm{D}}$-Methode

In diesem Abschnitt die häufig für das Design von analogen Schaltungen verwendete $g_{\mathrm{m}}/I_{\mathrm{D}}$-Methode durch Verwendung von REL-MOS-Modellen um die Berücksichtigung von PVTA-Einflüssen erweitert. Bei dieser Methode werden einzelne Transistoren für

verschiedene Längen, Weiten sowie Gate- und Drainspannungen in einer Testbench simuliert. Diese im Vorhinein simulierten Werte werden verwendet, um die in Abbildung 3.13 gezeigten Graphen zu erstellen. Wie zu erkennen ist, sind alle Kenngrößen über

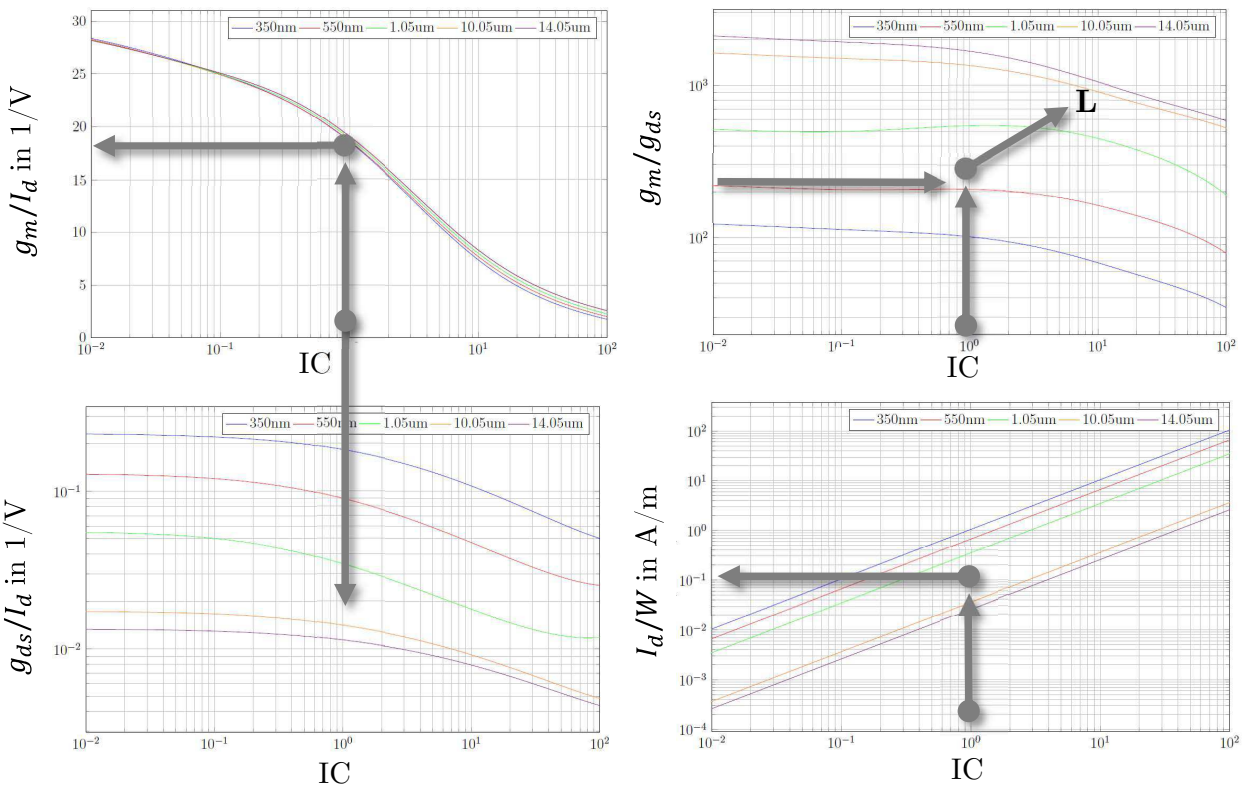


Abbildung 3.13: Bespielhafter Dimensionierungsablauf mit der tabellengestützten g_m/I_D-Methode. Die Graphen zeigen die einzelnen Designparameter über dem Inversionskoeffizienten für verschiedene Kanallängen.

dem Inversionskoeffizienten aufgetragen. Wie aus dem vorherigen Abschnitt folgt, verliert dieser Parameter für einige Transistordimensionierungen seine Allgemeingültigkeit, daher beschränkt sich die g_m/I_D-Methode auf eben diesen Bereich. Die Länge ist in dieser Methode ein globaler Parameter, der zuvor vom Designer anhand übergeordneter Abwägungen getroffen werden muss. Als erstes wird ausgehend vom IC die entsprechende Transkonduktanzeffizienz ermittelt. Aus IC und g_m/I_D ergibt sich bereits der Arbeitspunkt des Transistors. Anschließend wird der Ausgangsleitwert sowie die intrinsische Verstärkung ermittelt. Schließlich kann aus der Stromdichte die nötige Transistorweite abgelesen werden. Die Reihenfolge der Schritte und ob alle durchgeführt werden müssen, hängt dabei stark von der zu entwerfenden Schaltung und den vorgegebenen Spezifikationen ab. Neben dem grafischen Vorgehen per Hand, können die notwendigen Berechnungen selbstverständlich auch computergestützt durchgeführt werden. Werden diese Schritte für jeden Transistor einer Schaltung durchgeführt, beschreibt eines der Tupel $\mathbb{T} = \{I_d, L, g_m/I_D\}$ oder $\mathbb{T} = \{I_d, L, IC\}$ eindeutig den Arbeitspunkt des jeweiligen Transistors. In der Literatur werden zum Teil beide Beschreibungen synonymhaft verwen-

det, wobei manche Autoren die Beschreibung ohne den IC als für einen weiteren Bereich anwendbar beschreiben, da dieser ohne die Definition des I_{spec} auskommt, dem einige Annahmen vorausgehen. Generell können, abgesehen von bestimmten Sonderfällen, beide Beschreibungen verwendet werden. Die Methode eignet sich durch das Vorgehen besonders für die Auslegung von energieeffizienten Verstärkerschaltungen [Bin07]. In der Literatur gibt es zudem weitere Anwendungsgebiete [JM17a]. Des Weiteren wird versucht, die Methode weiter zu automatisieren und so weitere Aspekte in die Auslegung der Schaltung einfließen lassen zu können [HHPDP13]. Wie im vorherigen Abschnitt beschrieben, können Alterung, Prozessvariation und der Arbeitspunkt eines Transistors nicht getrennt betrachtet werden. Genau dieses Problem tritt bereits bei der initialen Dimensionierung einer Schaltung auf. Ein gewählter Arbeitspunkt in moderater Inversion kann sehr anfällig gegenüber Alterung sein und somit das gesamte Design. Dies kann zu einer Verletzung der Spezifikationen während des Nutzungszeitraums führen. Daher wurde in dieser Arbeit die g_m/I_D-Methode dahingehend erweitert, dass diese Einflüsse direkt während des Designprozesses berücksichtigt werden können. Da Prozessvariationen mit einbezogen werden, wird jeder Tabelleneintrag zu einer Verteilungsfunktion, deren Momente durch das Alter von der Zeit abhängig sind. Die Testbench, welche in Abbildung 3.14 gezeigt ist, bleibt dabei genau so simpel und auf zwei Elemente beschränkt wie in der zugrundeliegenden g_m/I_D-Methode. Zusätzlich zu den Variationen

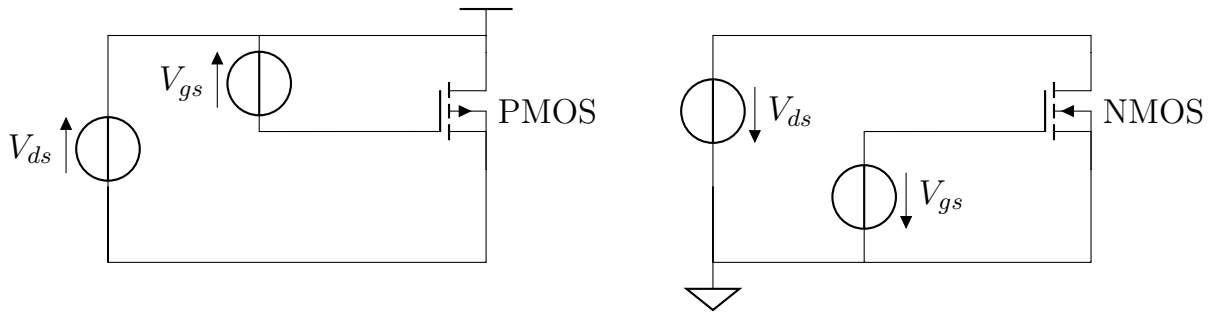


Abbildung 3.14: PMOS und NMOS Transistor Testbench zur Erstellung der g_m/I_D-Tabelleneinträge.

von W, L, V_g, V_d und der Temperatur T werden transiente Simulationen über 10 Jahre sowie Monte-Carlo Simulationen durchgeführt. Dadurch, dass keine Extrapolationen notwendig sind, können verschiedene Anwendungsprofile für die Transistoren analysiert werden, was eine zusätzliche Dimension in die Tabellen einführt. Dabei gilt zu beachten, dass die Temperatur, besonders auf BTI, bereits einen starken Einfluss besitzt und daher im Rahmen von Alterungsanalysen nicht außer Acht gelassen werden kann. Zur

Erweiterung der g_m/I_D-Methode werden die Verteilungsfunktionen durch Normal- und Lognormal-Verteilungen beschrieben, welche durch

$$f_n(\mathbf{x}, t) = \frac{1}{\sqrt{2\pi}\sigma(t)} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x} - \mu(t)}{\sigma(t)}\right)^2\right) \tag{3.16}$$

$$f_{ln}(\mathbf{x}, t) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma(t)\mathbf{x}} \exp\left(-\frac{1}{2}\left(\frac{\ln(\mathbf{x}) - \mu(t)}{\sigma(t)}\right)^2\right) & \mathbf{x} > 0 \\ 0 & \mathbf{x} \leq 0 \end{cases} \tag{3.17}$$

gegeben sind. $\mathbf{x}$ ist der Vektor des jeweiligen Parametersatzes $\{W, L, V_g, V_d, T\}$ und t die Zeit. Die Lognormal-Verteilung ist zusätzlich zur Normalverteilung notwendig, da einige Parameter keine symmetrische Verteilungsdichte besitzen. Für diese Parameter kann die Lognormal-Verteilung verwendet werden, um Mittelwert und Standardabweichung realistischer zu schätzen. Vorteil dieser beiden Verteilungen gegenüber anderen ist, dass ihre Momente eine physikalische Bedeutung besitzen und nicht reine Fitting-Parameter darstellen. Somit können diese Momente direkt für einen anschließenden Schaltungsentwurf verwendet werden. Wären die Momente nur reine Fitting-Parameter, müsste ergänzend ein funktionaler Zusammenhang zwischen z.B. dem Mittelwert und den Fitting-Parametern gefunden werden, was einen zusätzlichen Aufwand bedeuten würde. Wie bereits in den Gleichungen (3.16) und (3.17) gezeigt, sind die Momente der Verteilungsfunktion durch die Alterung von der Zeit abhängig. Diese Abhängigkeiten werden entweder durch Polynome

$$\mu(t) = p_{1,\mu}t^4 + p_{2,\mu}t^3 + p_{3,\mu}t^2 + p_{4,\mu}t + p_{5,\mu} \tag{3.18}$$

$$\sigma(t) = p_{1,\sigma}t^4 + p_{2,\sigma}t^3 + p_{3,\sigma}t^2 + p_{4,\sigma}t + p_{5,\sigma} \tag{3.19}$$

oder Potenzfunktionen

$$\mu(t) = \alpha_\mu + \beta_\mu \cdot t^{\gamma_\mu} \tag{3.20}$$

$$\sigma(t) = \alpha_\sigma + \beta_\sigma \cdot t^{\gamma_\sigma} \tag{3.21}$$

abgebildet. Welchen Grad das Polynom hat und ob Polynom oder Potenzfunktion verwendet werden, richtet sich nach der Größe des Fehlers der Approximation. Durch dieses Vorgehen entsteht eine N-dimensionale Tabelle, deren Einträge die Koeffizienten der Funktion der Momente der Verteilungsfunktionen sind. In Abbildung 3.15 ist eine solche Tabelle dargestellt. Um eine statistische Aussage über die Prozessvariation treffen zu können, müssen für jede Parameterkombination $\mathcal{P}_x$,$\mathcal{P}_y$,$\mathcal{P}_z$ aus $\{W, L, V_g, V_d, T\}$ 200 Transistoren je Typ (NMOS, PMOS), also insgesamt 400 Bauteile simuliert werden, und dies innerhalb einer transienten Simulation über 10 Jahre. Daher sind für die einmalige Erstellung der Tabelle mindestens 500.000 Simulationen notwendig. Diese müssen für eine Fertigungstechnologie jedoch nur einmal durchgeführt werden. Vorteil

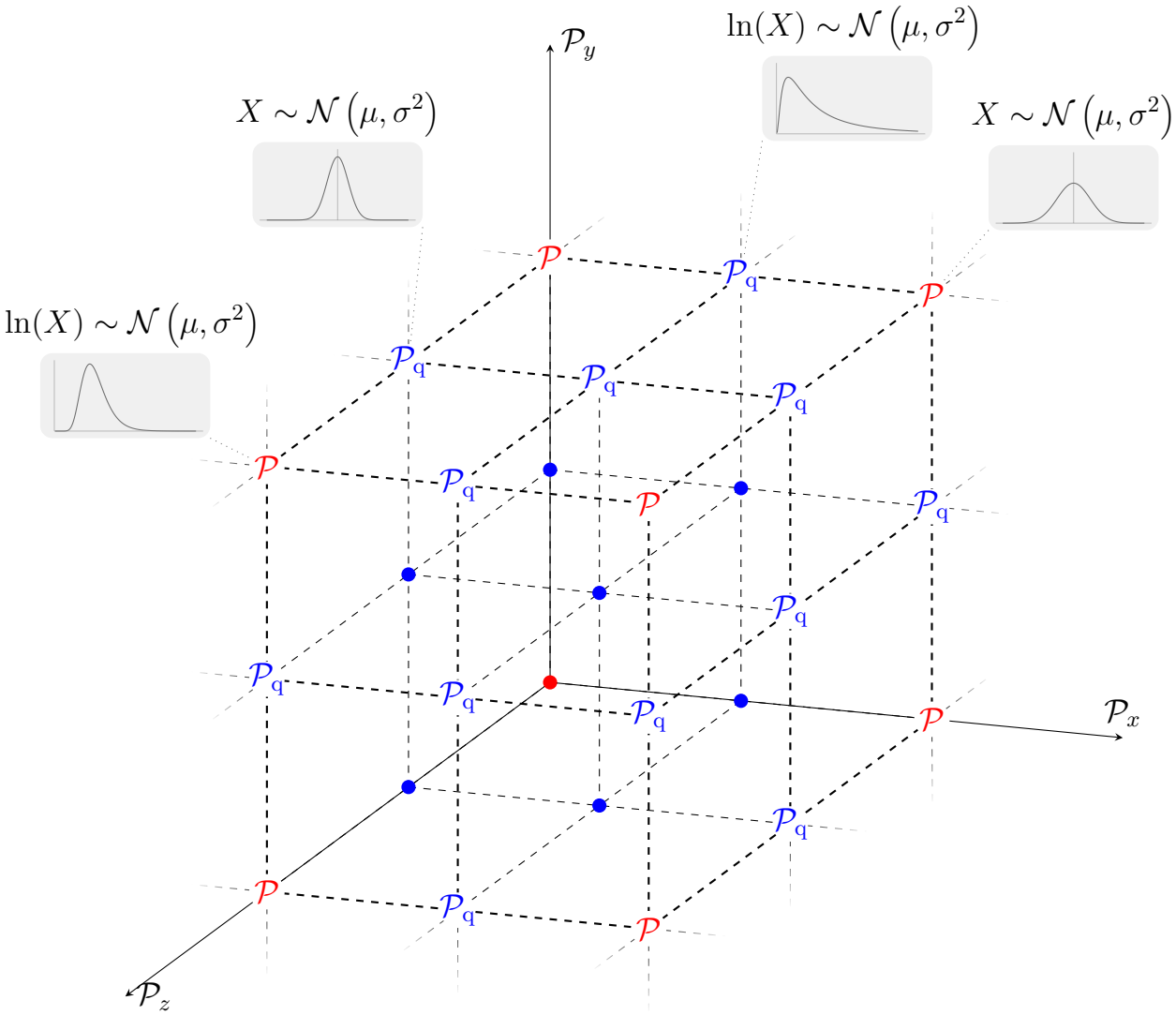


Abbildung 3.15: N-dimensionale Tabelle zur Speicherung der zeitabhängigen Momente der Verteilungsfunktionen über verschiedene Parametersätze $\mathcal{P}_x$,$\mathcal{P}_y$,$\mathcal{P}_z$ aus $\{W, L, V_g, V_d, T\}$.

der N-dimensionalen Tabelle ist die Reduktion des Speicherbedarfs und der Zugriffszeiten im Vergleich zur Speicherung der reinen Simulationswerte.

Mit den so gewonnenen Simulationswerten kann die $g_\mathrm{m}/I_\mathrm{D}$-Methode für die Mittelwerte und einen Zeitpunkt grafisch durchgeführt werden. So können zum Beispiel schnell die Parameter am Ende des Nutzungszeitraums einer Schaltung analysiert werden oder ein Design auf Grundlage der End Of Life (EOL) Parameter durchgeführt werden. Sollen die tatsächlichen Verteilungen und Zeitabhängigkeiten betrachtet werden, muss die Methode computergestützt durchgeführt werden. Dabei kann zur Festlegung des Arbeitspunktes eines Transistors sofort die Veränderung durch Prozessvariation und Alterung mit betrachtet werden. Für die Auslegung von Schaltungen werden die zuvor

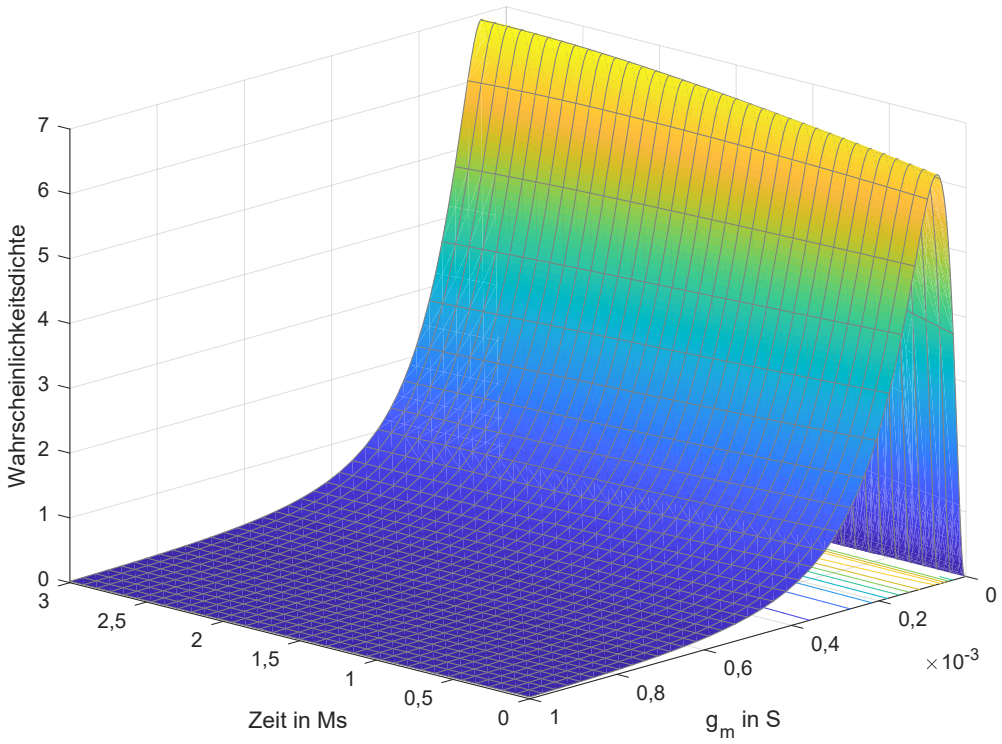


Abbildung 3.16: Zeitabhängige Verteilungsfunktion der Transkonduktanz g_m eines NMOS Transistors.

simulierten Verteilungen der Parameter in die Kleinsignalgleichungen der Schaltung eingesetzt. Für die Verstärkung einer Common-Source Stufe wäre dies

$$A_{DC} = \frac{\left(\frac{\mathbf{g_{m,1}}}{\mathbf{I_D}}\right)_{sim}}{\left(\frac{\mathbf{g_{ds,1}}}{\mathbf{I_D}}\right)_{sim} + \left(\frac{\mathbf{g_{ds,2}}}{\mathbf{I_D}}\right)_{sim}} = \frac{\mathbf{g_{m,1}}}{\mathbf{g_{ds,1}} + \mathbf{g_{ds_2}}} \ . \tag{3.22}$$

Die zeitabhängigen Verteilungsfunktionen werden, wie in Abbildung 3.16, für die Leitwerte aus der Tabelle entnommen. Die Gleichung für die Verstärkung A wird anschließend computergestützt mit Monte-Carlo-Sampling angenähert und somit die Verteilung und Zeitabhängigkeit der Verstärkung. Für diese Auswertung sind keine weiteren SPICE-Simulationen notwendig. Dieses Vorgehen kann auf alle spezifizierten Schaltungseigenschaften angewendet werden, die durch eine Kleinsignalgleichung beschrieben werden können. Der Vergleich der Mittelwerte für die Berechnung und Simulation der Verstärkung ist in Tabelle 3.1 gezeigt. Wie der Tabelle entnommen werden kann, wird

Tabelle 3.1: Vergleich der Mittelwerte der Berechnung und Simulation für $t = 0$ und $t = 10$ Jahre von A_{DC}.

Ungealtert	Wert	Gealtert	Wert
$A_{DC,pre}$	93.1 dB	$A_{DC,pre}$	13.7 dB
$A_{DC,sim}$	82.3 dB	$A_{DC,sim}$	19.12 dB

die Verstärkung durch die Berechnung überschätzt und nach 10 Jahren unterschätzt. Dies ist bedingt durch die Tatsache, dass die schaltungsbedingten Wechselwirkungen zwischen den beiden Transistoren über die jeweiligen Ausgangsleitwerte g_{ds} nicht in den Kleinsignalgleichungen berücksichtigt werden können und es somit zu den Abweichungen kommt. Da sich durch die Alterung die Drainspannung des einen Transistors verändert, diese aber über die Maschengleichung mit der des anderen Transistors verknüpft ist, ändern sich die Stressbedingungen durch die Alterung selbst. Zudem können sich die Sättigungszustände der Transistoren durch die alterungsbedingte Rückkopplung ändern. Nichtsdestotrotz bietet diese Erweiterung der g_m/I_D-Methode dem Designer die Möglichkeit, die komplexen Wechselwirkungen von Alterung und Prozessvariation vor der ersten SPICE-Simulation der Schaltung für verschiedene Dimensionierungen der Transistoren zu untersuchen.

Für komplexere Schaltungen wie zum Beispiel einem zweistufigen Miller-OTA können, wenn diese relevant für die Spezifikationen sind, einzelne Aspekte sehr einfach analysiert und dargestellt werden. So zeigt sich im Verlauf des Mittelwertes und der Standardabweichung der Verstärkung in Abbildung 3.17, dass das Maximum zusammen mit der minimalen Standardabweichung um eine Temperatur von 50°C liegt. Die Dimensionierung des Verstärkers wurde dabei mit der zuvor beschriebenen g_m/I_D-Methode durchgeführt. Das Maximum der Streuung wird dabei für sehr kleine Temperaturen erreicht, wobei die Gesamtverstärkung sinkt. Die Alterungsanalyse mit den REL-MOS-Modellen zeigt, dass für verschiedene Temperaturen die Verstärkung im Laufe der Zeit steigen oder fallen kann (vgl. Abb. 3.18). Die Kleinsignalgleichung für die Verstärkung lautet:

$$A_{ges} = \frac{g_{m,1}}{g_{ds,2} + g_{ds,4}} \cdot \frac{g_{m,7}}{g_{ds,6} + g_{ds,7}} \tag{3.23}$$

Die Arbeitspunkte der Transistoren M1, M2 und M7 liegen wegen der höheren Effizienz in moderater Inversion. Falls diese altern, verschiebt sich der Arbeitspunkt in Richtung schwacher Inversion, wodurch die Verstärkung zuerst steigen würde. Der Transistor M6 wird als Stromspiegel in starker Inversion betrieben, und der Einfluss durch BTI verstärkt sich mit steigender Temperatur. Dadurch steigt $g_{ds,6}$ im Laufe der Zeit. Je nachdem, welcher der Einflüsse überwiegt, steigt oder fällt die Verstärkung. An diesem Beispiel ist sehr deutlich zu erkennen, dass die Umwelteinflüsse, denen die Schaltung während ihrer Nutzung ausgesetzt ist, enormen Einfluss auf die Alterungscharakteristik haben. Des Weiteren kann durch die Analyse mit den REL-MOS-Modellen der Worst-Case für den Verstärker gefunden werden, der bei $V_{dd} = 0{,}9V$ und $T = -40$°C liegt. In diesem Arbeitspunkt haben Alterung und Prozessvariation zusammen die stärksten Auswirkungen und erniedrigen die Verstärkung nach 10 Jahren auf unter 20dB. Entgegen den Erwartungen für die Auswirkungen, für die die nominellen Stressbedingungen für hohe Temperaturen und Spannungen stärker sind, ist dieses Alterungsverhalten nicht von einzelnen Transistoren der Schaltung auf die gesamte Schaltung übertragbar.

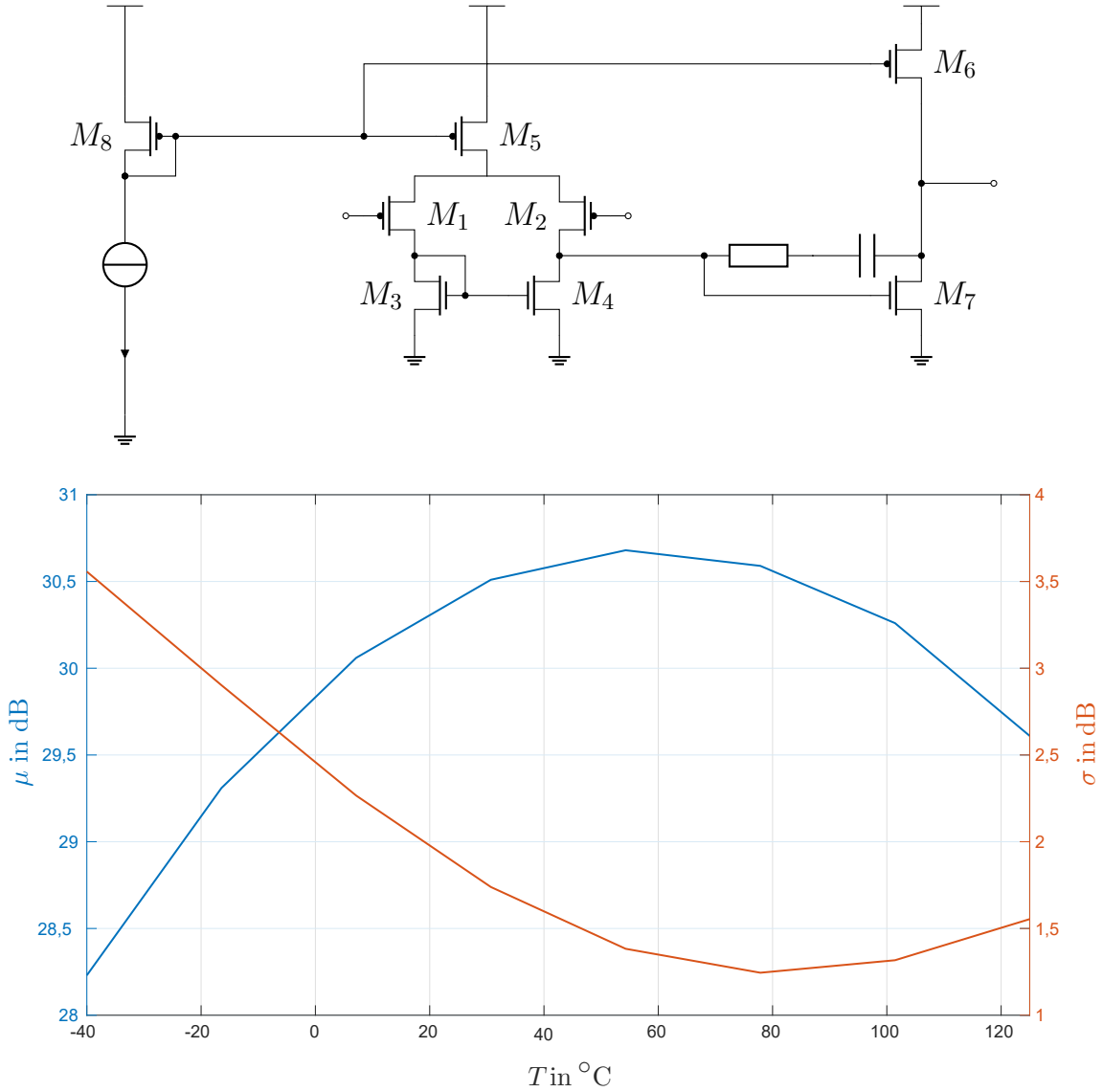


Abbildung 3.17: Schaltplan des Miller-OTA und Verlauf von Mittelwert und Standardabweichung der Verstärkung über der Temperatur.

3.3 Zusammenfassung

In diesem Kapitel wurde das entwickelte REL-MOS-Modell, welches erweitert und Alterung und Strahlung ist, beschrieben. Es wurde gezeigt, dass alle in der Literatur beschriebenen Effekte der Alterung qualitativ korrekt durch das Modell wiedergegeben werden können und dass der Modellparameter Q_{age} und seine Modellierung physikalisch motiviert sind. Des Weiteren wurden die deutlich erweiterten Analysemöglichkeiten gegenüber der herkömmlichen Modellierung von Alterung gezeigt. Beispielhaft wurden an zentralen Entwurfsparametern die Auswirkungen und Wechselwirkungen von Prozessvariation und Alterung gezeigt. Anschließend wurde eine erweiterte Entwurfsmethodik gezeigt, die es ermöglicht, sämtliche Effekte bereits bei der initialen Dimensionierung zu

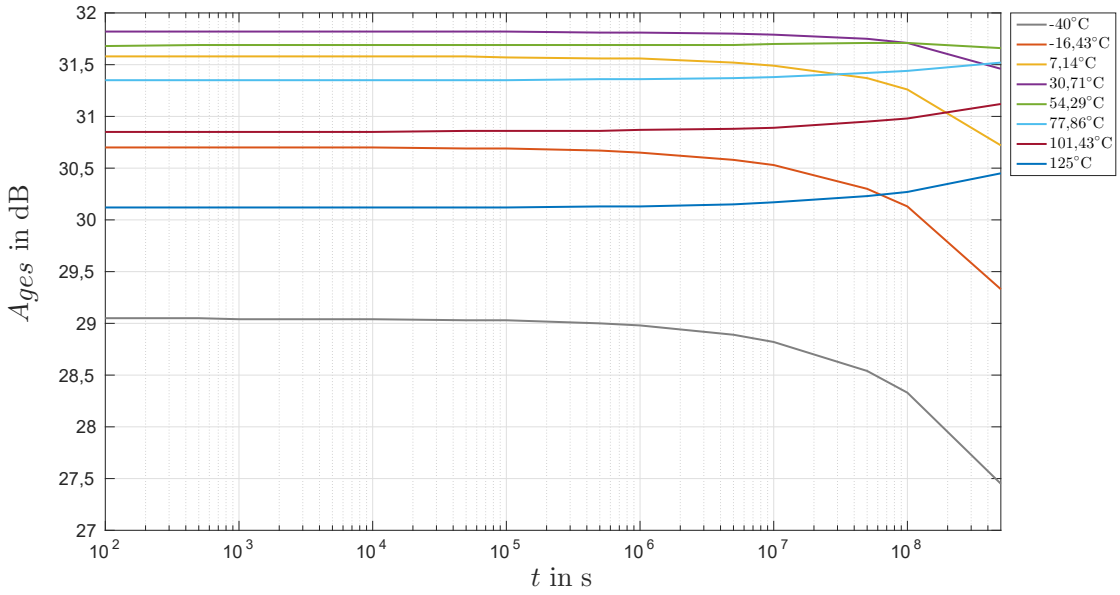
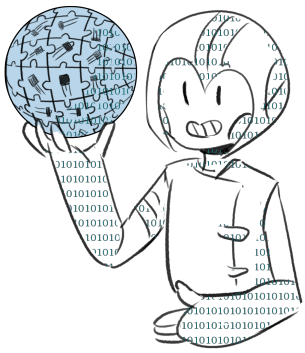


Abbildung 3.18: Verstärkung des Miller-OTA über 10 Jahre für verschiedene Temperaturen.

berücksichtigen. Abschließend wurden anhand eines zweistufigen Verstärkers die starken Verschränkungen zwischen den Einflüssen, der Schaltungsfunktion und den Umweltbedingungen gezeigt. Durch die Verschränkungen ist die Übertragung der Erkenntnisse von einzelnen Transistoren auf Schaltungs- oder Systemebene stark limitiert.

4 Schaltungsentwurf unter Berücksichtigung von PVTAR-Einflüssen

» The better is the enemy of the good «



Im vorherigen Kapitel wurden die Anwendungsmöglichkeiten des REL-MOS-Modells für die Analyse und den Entwurf von analogen integrierten Schaltungen gezeigt. Der Ansatz, bei der Auslegung der Schaltung von einzelnen Transistoren mittels Kleinsignalgleichungen auf die Spezifikationen zu schließen, ist durch die reine Komplexität größerer Schaltungen beschränkt. Zudem zeigte sich, dass die Abhängigkeiten von PVTAR-Einflüssen einzelner Transistoren innerhalb einer Funktionseinheit wie der Common-Source-Stufe nicht durch Kleinsignalgleichungen abgebildet werden können. Daher wurde im Rahmen dieser Arbeit eine Block-basierte Entwurfsmethodik für große analoge Schaltungen entwickelt. Die Methode verwendet für die Auslegung eine vordefinierte Gruppe von funktionalen Einheiten z.B. zum Verstärken. Diese Einheiten werden zuvor mit der $g_\mathrm{m}/I_\mathrm{D}$-Methode auslegt. Die systemrelevanten Parameter dieser Funktionsblöcke beschränken sich auf einzelne Aspekte und nicht auf alle Transistorparameter. Dadurch können Analysen durch den Designer zielgerichteter durchgeführt werden. Des Weiteren bietet dieser Ansatz die Möglichkeit, die Erkenntnisse auf Blockebene auf Schaltungsebene zu übertragen, was sich bei Transistoren verbietet. Dadurch können To-

pologien schnell nach einzelnen geforderten Aspekten klassifiziert werden. Dies wurde im Rahmen dieser Arbeit für 8 Verstärkertopologien und 7 Bias-Schaltungen durchgeführt. Die jeweils robustesten Schaltungen wurden anschließend in einem 65nm-CMOS-Prozess gefertigt und vermessen.

4.1 Topologieabhängige Zuverlässigkeit

In diesem Abschnitt wird die entwickelte Methode zum Design von analogen Schaltungen, mit einem Fokus auf Verstärkerschaltungen, basierend auf zuvor dimensionierten und simulierten Funktionsblöcken beschrieben. In analogen Schaltungen ist es meist das Ziel, die spezifischen Eigenschaften einzelner Transistoren hervorzuheben und diese störungsfrei zu den Ausgängen zu propagieren. Dazu müssen die Transistoren durch Biasschaltungen in den gewünschten Arbeitspunkt gebracht werden und das Nutzsignal korrekt verarbeitet werden. Dabei können einzelne Spezifikationen wie eine niedrige Leistungsaufnahme, geringer Platzbedarf oder sehr hoher Ausgangswiderstand große Auswirkungen auf die generelle Topologie haben. So wird etwa direkt eine differentielle Topologie gewählt, falls die Einflüsse durch Rauschen reduziert werden sollen. Um diese generellen Anforderungen mit den Blöcken abdecken zu können, wurden verschiedene Biasschaltungen und Verstärkerstufen verwendet. Die Idee, den Entwurf von analogen Schaltungen in kleine Subschaltungen zu unterteilen, besteht bereits für das nominelle Verhalten als auch für einige ausgewählte Schaltungen im Hinblick auf Zuverlässigkeit. In [SK08] wird eine Möglichkeit beschrieben, in Abhängigkeit von Funktionsblöcken verschiedene Topologien von Schaltungen zu entwerfen. Diese Methode basiert jedoch auf analytischen Kleinsignalgleichungen. Die Berücksichtigung von PVTA-Variationen für die einzelnen Blöcke und zudem weitere Effekte wie Kurzkanaleinflüsse und Wechselwirkungen zu berücksichtigen, ist selbst eine sehr aufwendige Aufgabe [Aro12]. So können mit diesem Ansatz nur Topologien entwickelt werden, die einzelne spezifizierte Kennzahl (*engl. Figure of Merit*, FoM) erfüllen. Eine Möglichkeit der generellen Analyse ist durch dieses Vorgehen nicht möglich. Wie bereits in vorherigen Abschnitten beschrieben, können insbesondere die Wechselwirkungen zwischen den einzelnen Effekten ausschließlich auf Transistorebene korrekt betrachtet werden. Ein anderes Vorgehen wird in [JM17b] beschrieben. In diesem wird die Dimensionierung der Transistoren ebenfalls berücksichtigt. Dabei wird die g_m/I_D-Methode verwendet und zudem werden Rauscheinflüsse und Fehlanpassung während des Designs berücksichtigt. Dieser Ansatz basiert nicht ausschließlich auf analytischen Gleichungen, um solche Effekte, die nur korrekt durch Simulationen von ganzen Transistormodellen abgebildet werden können, im Designprozess berücksichtigen zu können. Der Fokus liegt dabei jedoch auf der Dimensionierung der Transistoren und nicht auf der Auswahl von vornherein geeigneter Topologien. In [JM17b] werden weiterhin als Beispielanwendungen eine Biasquelle und ein Verstärker dimensioniert. Diese Beispiele verdeutlichen, dass die Verwendung von vorsimulierten Werten für den Designprozess von großem Nutzen sein kann. Jedoch

führt die Fokussierung auf die Dimensionierung bei der Erfüllung der Schaltungsspezifikationen (z.B. Verstärkung) zu, sich an der Grenze zum Praktikablen befindlichen, Dimensionen von $W > 2\text{mm}$ und $W/L > 1000$. Ein solcher Transistor würde bereits alleine die Fläche einer gesamten, in der Arbeit entworfenen Schaltung, einnehmen. Daher muss bei einem solchem Vorgehen zwischen Fläche und Effizienz der Schaltung abgewogen werden.

Ein ähnliches Vorgehen wird in [MG13] und [Nic15] beschrieben. Dabei werden vorsimulierte Werte auf Transistorebene in der $g_\mathrm{m}/I_\mathrm{D}$-Methode verwendet, um eine Berücksichtigung von Zuverlässigkeitseinflüssen im Entwurfsprozess zu ermöglichen. In [Nic15] wird zudem die Möglichkeit beschrieben, die vorsimulierten Ergebnisse für eine Sensitivitätsanalyse zu verwenden. Beide Arbeiten legen die Schwerpunkte dabei jedoch ebenfalls auf die Analyse und die Dimensionierung einzelner Transistoren mit den bereits beschriebenen einhergehenden Nachteilen und Einschränkungen.

Die Idee in dieser Arbeit ist es nicht, einzelne Parameterwerte durch das Design zu propagieren, um so eventuell einige Simulationsschritte auszulassen, sondern dass der Designer eine umfangreiche Analyse dieser Funktionsblöcke durchführt um somit Topologien verschiedener Schaltungen hinsichtlich der qualitativen Anfälligkeit gegenüber Alterung analysieren zu können. Dazu ist es notwendig, dass die alterungsrelevanten Informationen sich von der Block- auf die Schaltungsebene übertragen lassen. Somit werden die Ergebnisse der zuvor beschrieben Arbeiten kombiniert. Die Betrachtung der Einflüsse auf Topologieebene ermöglicht, bereits vor der Dimensionierung der Transistoren entscheidende FoM zu erfüllen bzw. das gesamte Design alterungsresistent auszulegen. Dabei kommen vorsimulierte Werte von Funktionsblöcken zum Einsatz. Dadurch können analytische Gleichungen auf System- und Schaltungsebene verwendet werden, welche mit den vorsimulierten Werten evaluiert werden, um eine geeignete Topologie zu bestimmen oder eine Sensitivitätsanalyse durchzuführen.

4.1.1 Funktionsblöcke

Die Schaltpläne der NMOS-Varianten der genutzten Schaltungen sind in Abbildung 4.1 gezeigt. Die PMOS Varianten sind nicht gezeigt, werden in der Methode jedoch verwendet. Für die Stromspiegel ist der Biasstrom als Eingangsgröße angenommen. Da jede Schaltung in den Blöcken jeweils eine einzige Aufgabe besitzt, kann diese in der $g_\mathrm{m}/I_\mathrm{D}$-Methode als wichtigste Zielvorgabe verwendet werden. Generell gibt es für diese Methode zwei mögliche Ansätze. Zum einen können die Blöcke bereits mit der erweiterten $g_\mathrm{m}/I_\mathrm{D}$-Methode robust gegenüber PVTA-Einflüssen ausgelegt werden. Dadurch werden die Designs jedoch größer und weniger energieeffizient. Zum anderen kann bei der Dimensionierung der Blöcke ein Fokus die jeweilige Hauptfunktionalität gelegt werden, und die PVTA-Einflüsse werden durch die Topologie und die Wahl der Blöcke auf Systemebene kompensiert. Im Rahmen dieser Arbeit wurden beide Varianten

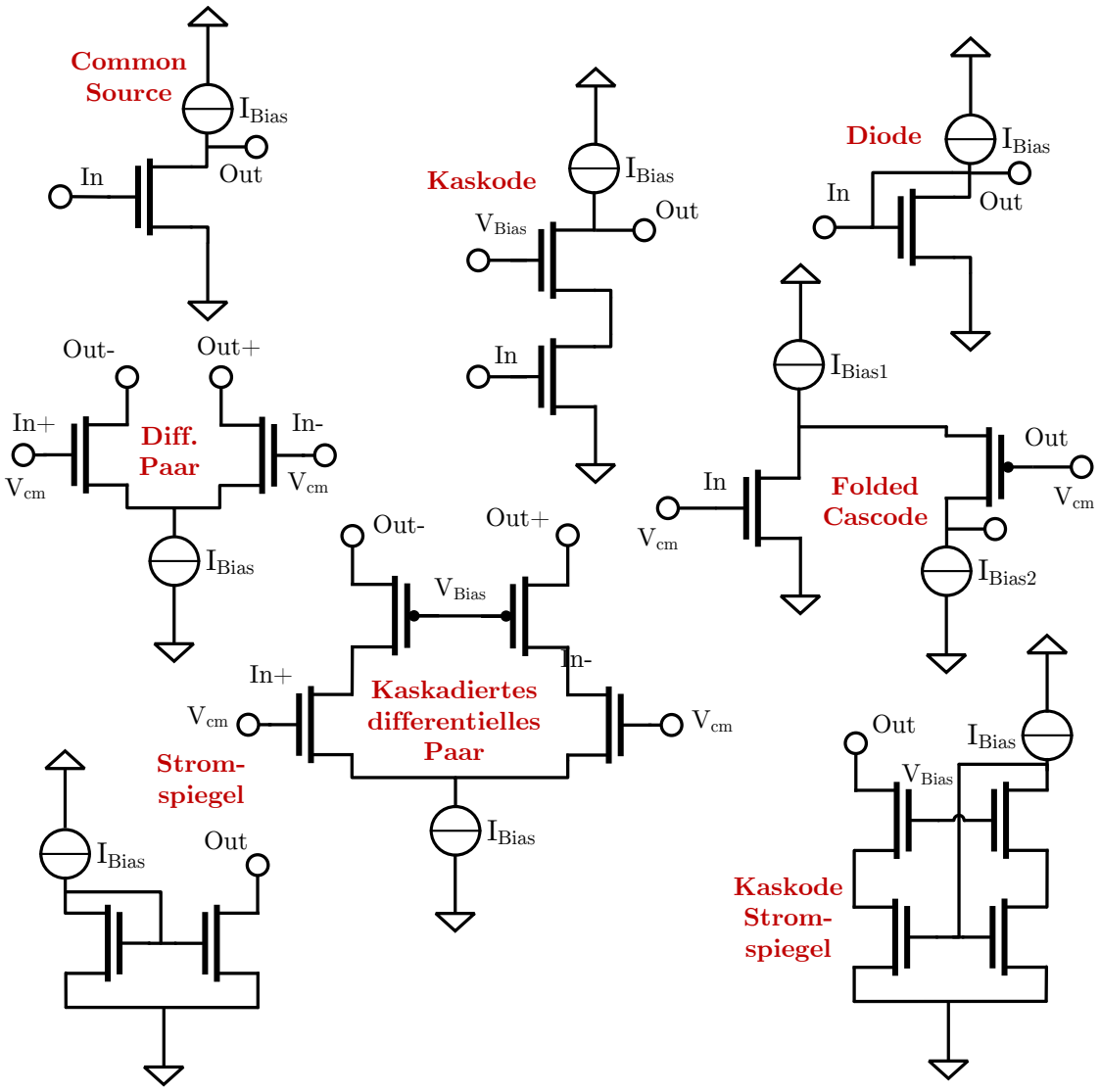


Abbildung 4.1: Schaltungen der Funktionsblöcke für die Block-basierte Entwurfsmethodik. Gezeigt sind jeweils nur die NMOS-Varianten.

von Funktionsblöcken entworfen. Im Weiteren werden zunächst die Ergebnisse der einzelnen Blöcke beschrieben.

Common-Source-Stufe Die Grundfunktion einer Common-Source-Stufe ist die Einspannungs-Ausgangsspannungs-Verstärkung. Dabei wird mit Hilfe der Transkonduktanz des Transistors die Eingangsspannung in einen antiproportionalen Ausgangsstrom gewandelt, welcher über den Ausgangswiderstand der Stufe in die Ausgangsspannung umgesetzt wird. Damit gilt für die Kleinsignalverstärkung

$$A = -g_m \cdot r_{out} \ . \tag{4.1}$$

Für die Transkonduktanz des Transistors in Sättigung wird vereinfachend $g_m = \sqrt{2\mu C_{ox} {}^{W}/_{L} I_d}$ angenommen. Der Ausgangswiderstand kann mit $r_{out} = {}^{1}/_{\lambda I_d}$ berechnet werden. Diese Gleichungen stammen aus dem MOS1-Modell und geben somit nicht den quantitativen, sondern nur den qualitativen Zusammenhang wieder. Dabei zeigt sich, dass eine Erhöhung des Drainstromes nicht zu einer Erhöhung der Verstärkung führen würde. Daher verbleiben die Länge und die Weite als Stellgrößen. Für die beiden Auslegungsarten wird für das in Bezug auf PVTA-Einflüsse zuverlässigkeitsoptimierte Design der Arbeitspunkt des Transistors in starke Inversion gelegt und für das funktionsoptimierte Design in moderate Inversion. Das funktionsoptimierte Design erfüllt die Spezifikationen unter nominellen Bedingungen bei geringster Fläche und Leistungsaufnahme. Abbildung 4.2 zeigt die Simulationsergebnisse der intrinsischen Verstärkung der NMOS-Common-Source Stufe für verschiedene Biasströme über 10 Jahre. Wie zu erwarten war, zeigt sich, dass für höhere Ströme die Alterung durch

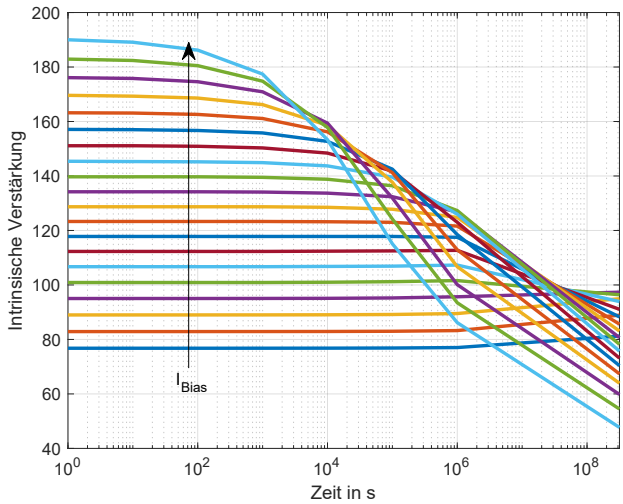


Abbildung 4.2: Simulationsergebnis der intrinsischen Verstärkung der NMOS-Common-Source-Stufe für verschiedene Biasströme (500 nA . . . 50 μA) über 10 Jahre.

HCI zunimmt und somit die Verstärkung durch den resultierenden starken Anstieg des Ausgangswiderstands zunimmt. Für geringe Biasströme befindet sich der Transistor in niedriger moderater Inversion. Durch den gewählten Arbeitspunkt überwiegt die alterungsbedingte Abnahme der Transkonduktanz, und die Verstärkung fällt. Nominal zeigt die Schaltung das Verhalten wie aus den Gleichungen vorhergesagt. Einige Arbeitspunkte sind dabei sehr sensitiv gegenüber Alterung, was bereits durch das Verhalten des NMOS-Transistors, dass in Abbildung 3.11 dargestellt ist, klar ist. Von besonderem Interesse für die Auslegung sind die Biaskonditionen, bei denen sich die Verstärkung nicht ändert. In diesen Arbeitspunkten heben sich die gegenläufigen Effekte gerade auf. Je nach Anforderung an die Schaltung kann ein kompensierter Arbeitspunkt gewählt werden oder ein hoher Biasstrom, bei dem die Verstärkung ansteigt. Beides würde die Schaltungs- und Systemfunktion für den gesamten Nutzungszeitraum gewährleisten.

MOS-Diode Die MOS Diode wird häufig als Last- oder Biaselement verwendet, da diese einen Strom in eine Spannung wandelt. Dabei ist die Gatespannung maßgeblich von der Drain-Sättigungsspannung $V_{ds,sat}$ abhängig. Die Biasspannung ist gegeben durch

$$V_{bias} = V_{th} + 2nU_t \ln\left(\exp\left(\sqrt{I_{bias}/I_s}\right) - 1\right) \ . \tag{4.2}$$

Die Alterung bewirkt, dass der Drainstrom durch die Erhöhung der Schwellspannung sinkt. In diesem Fall ist der Biasstrom jedoch fest eingeprägt, dadurch steigt die Biasspannung über das Alter stärker, als es ohne eine ideale Stromquelle passieren würde. Der generelle Trend bleibt davon jedoch unberührt. Da für den NMOS- und PMOS-Fall die Auswirkungen deutlich verschieden sind, sind in Abbildung 4.3 die beiden Fälle getrennt gezeigt. Wie zu erkennen ist, altert die PMOS-Diode fast gar nicht.

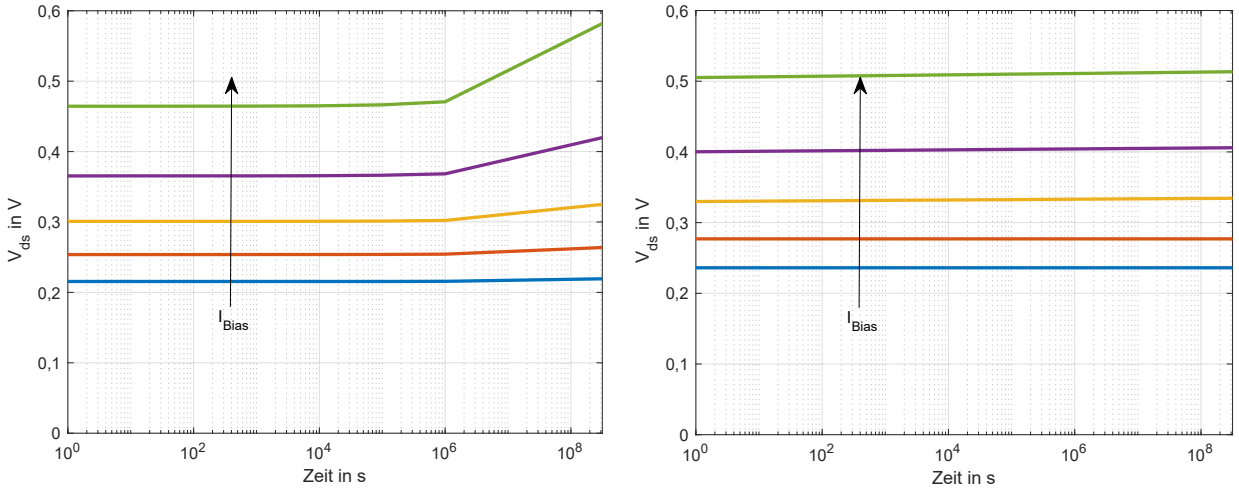


Abbildung 4.3: NMOS (links) und PMOS (rechts) Diode für verschiedene Biasströme ($1\,\mu$A ... $50\,\mu$A) über 10 Jahre.

Dies liegt daran, dass $V_{gs} \approx 0{,}5\,$V beträgt und es somit kaum zu BTI-Stress kommt (vgl. Abb. 2.15). Die Effekte durch HCI bei der NMOS-Diode skalieren hingegen mit dem Biasstrom. Somit kann der Einfluss durch Alterung bereits durch die Wahl der Bias- und Lastelemente stark beeinflusst werden. Für Entwürfe, bei denen die Alterungseinflüsse reduziert werden sollen, sollten für das Bias-Schema PMOS-Transistoren verwendet werden. Dadurch vergrößert sich bei gleichem Strom jedoch die Fläche der Schaltung.

Kaskode Bei der Kaskode werden zwei Transistoren im selben Zweig in Serie angeordnet, um den Ausgangswiderstand einer Schaltung zu erhöhen. Da die Transistoren in Sättigung betrieben werden sollten, um die V_{ds}-Abhängigkeit so gering wie möglich zu halten, ist die Anzahl der Kaskode-Transistoren durch das technologisch gegebene Verhältnis $V_{ds,sat}/V_{dd}$, begrenzt. Der Inversionszustand des Kaskode-Transistors kann dabei über die Biasspannung eingestellt werden. Der simulierte Ausgangsleitwert über einen Zeitraum von 10 Jahren ist in Abbildung 4.4 dargestellt. Ähnlich zur Common-

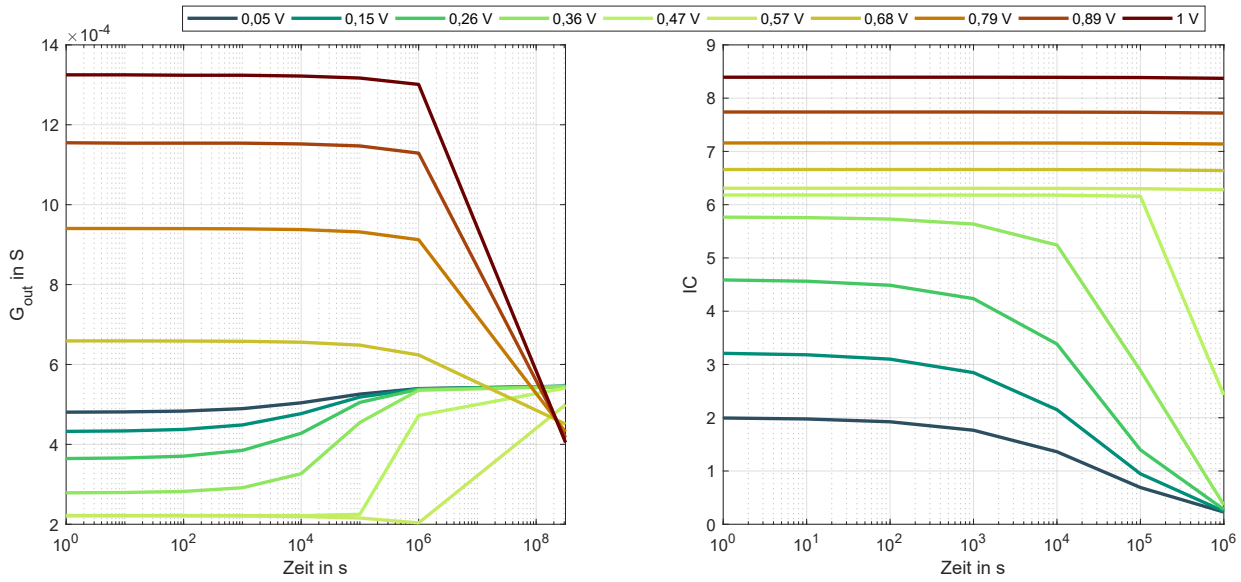


Abbildung 4.4: Ausgangsleitwert und Inversionskoeffizient einer NMOS-Kaskode für verschiedene Biasspannungen über 10 Jahre.

Source-Stufe können abhängig vom Inversionszustand des Kaskodetransistors die Alterungseinflüsse getrimmt werden. So existiert eine Biasspannung, die die Veränderung des Ausgangsleitwerts über der Zeit minimieren kann. Zudem ist an den unterschiedlichen Verläufen für den Ausgangswiderstand und den Inversionskoeffizienten zu erkennen, dass die Alterung sich deutlich unterschiedlich auf diese beiden Parameter auswirkt. Je nachdem, welche Parameter für das Schaltungsverhalten maßgeblich ist, muss ein anderer Arbeitspunkt gewählt werden, um der Alterung entgegen zu wirken.

Folded-Cascode Folded-Cascode-Schaltungen werden dazu verwendet, eine differenzielle Eingangstufe mit einem Kaskode-Ausgangszweig zu verbinden. Dadurch können die Vorteile beider Schaltungskomponenten verbunden werden. Im Gegensatz zur normalen Kaskode ist die Ausgangsspannung nicht durch die Drain-Source-Spannung von zwei Transistoren erniedrigt.

Differentielles Paar Ein differentielles Eingangspaar wandelt die Differenz der jeweiligen Gatespannungen und teilt den Biasstrom proportional zu den Gatespannungen in Zweigströme. Der Vorteil ist die höhere Verstärkung und die Unterdrückung von Gleichtaktsignalen. Die Transkonduktanz ist dieselbe wie für die Common-Source-Stufe. Beide Transistoren müssen in einem bestimmten Arbeitspunkt betrieben werden, um diese Eigenschaften zu gewährleisten. Für eine energieeffiziente Schaltung ist die moderate Inversion am besten geeignet. Der Inversionszustand an den Gatekontakten der Transistoren wird mittels der angelegten Gleichspannung V_{cm} und dem Biasstrom eingestellt. Befindet sich, wie in der Testbench, eine ideale Quelle im Pfad, ist allein der Biasstrom für die Regelung des Inversionszustands verantwortlich. Die intrinsische

Verstärkung ist in Abbildung 4.5 dargestellt. Wie bei der Common-Source-Stufe ist

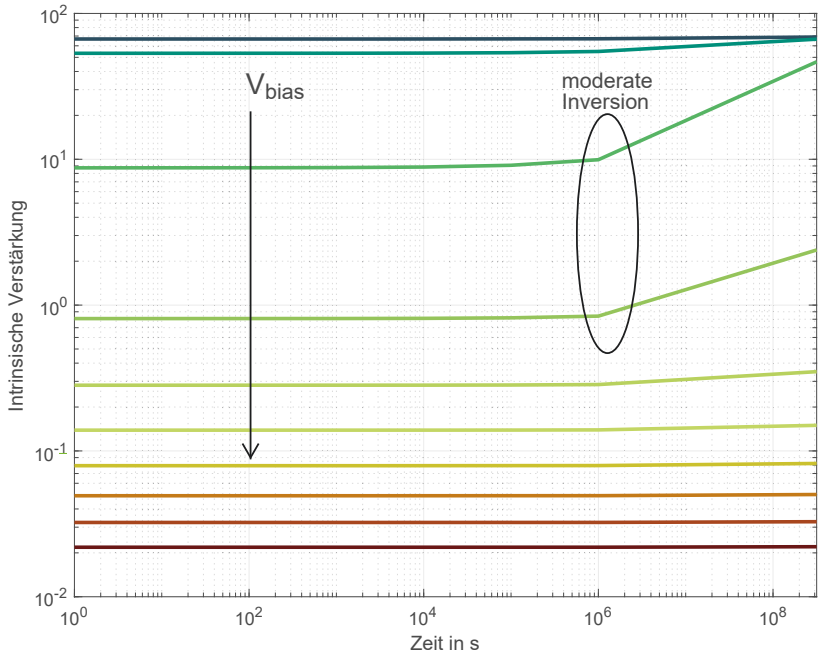


Abbildung 4.5: Instrinsische Verstärkung des differentiellen Eingangspaar für verschiedene Biasspannungen ($I_{bias} = 100\,\mu$A, $V_{bias} = 0,05 \ldots 1$ V) über 10 Jahre.

die intrinsische Verstärkung im Bereich der moderaten Inversion am stärksten von der Alterung betroffen. Daher ist für eine alterungsresistente Blockauslegung für das differentielle Eingangspaar die starke Inversion zu wählen.

Kaskadiertes differentielles Paar Für diese Schaltung gelten die gleichen Aussagen wie für die Kaskode und das differentielle Eingangspaar.

Stromspiegel Stromspiegel werden in vielen analogen Schaltungen verwendet, um die Arbeitspunkte der Transistoren über verschiedene Zweige hinweg einzustellen. Dabei wird das Gatepotential des als Diode verschalteten Transistors zum anderen Transistor gespiegelt. Bei gleichen Transistordimensionen stellt sich im Ausgangszweig somit der gleiche Strom ein. Ausgehend vom Drainstrom in Sättigung

$$I_{ds} = \frac{1}{2}\mu C_{ox} {}^{W}/_{L}(V_{gs} - V_{th}) \cdot (1 + \lambda(V_{ds} - V_{ds,sat})) \tag{4.3}$$

nach [Bak08], folgt für das Spiegelverhältnis der Ströme

$$\frac{I_{out}}{I_{ref}} = \frac{\mu_1 {}^{W_1}/_{L_1}(V_{gs} - V_{th,1}) \cdot (1 + \lambda(V_{ds,1} - V_{ds,sat,1}))}{\mu_2 {}^{W_2}/_{L_2}(V_{gs} - V_{th,2}) \cdot (1 + \lambda(V_{ds,2} - V_{ds,sat,2}))} \,. \tag{4.4}$$

Für exakt die gleichen Transistoren ohne Kurzkanaleffekt ist das Verhältnis 1. In einer realen Schaltung können jedoch die Gate-Source-Spannung, durch parasitäre Widerstände, sowie Drain-Source Spannungen unterschiedlich sein. Zudem weichen die Mobilität und die Schwellspannung durch Prozessvariation in der Regel leicht voneinander ab. Da die Stressbedingungen für beide Transistoren in Bezug auf Alterung unterschiedlich sind, führt dies zu einer ungleichen Veränderung von μ, V_{th} und $V_{ds,sat}$. Dabei sind die Stressbedingungen für den spiegelnden Transistor erhöht, dadurch ist zu erwarten, dass das Spiegelverhältnis mit dem Alter sinkt. Die Stressbedingungen können dabei ebenfalls von der Last abhängig sein. Das simulierte Verhalten ist in Abbildung 4.6 gezeigt. Wie erwartet, sinkt das Verhältnis fast unabhängig vom eingestellten Strom über das Alter. Dabei ist ein Abfall um bis zu 30% festzustellen. Für nachgeschaltete Transistoren,

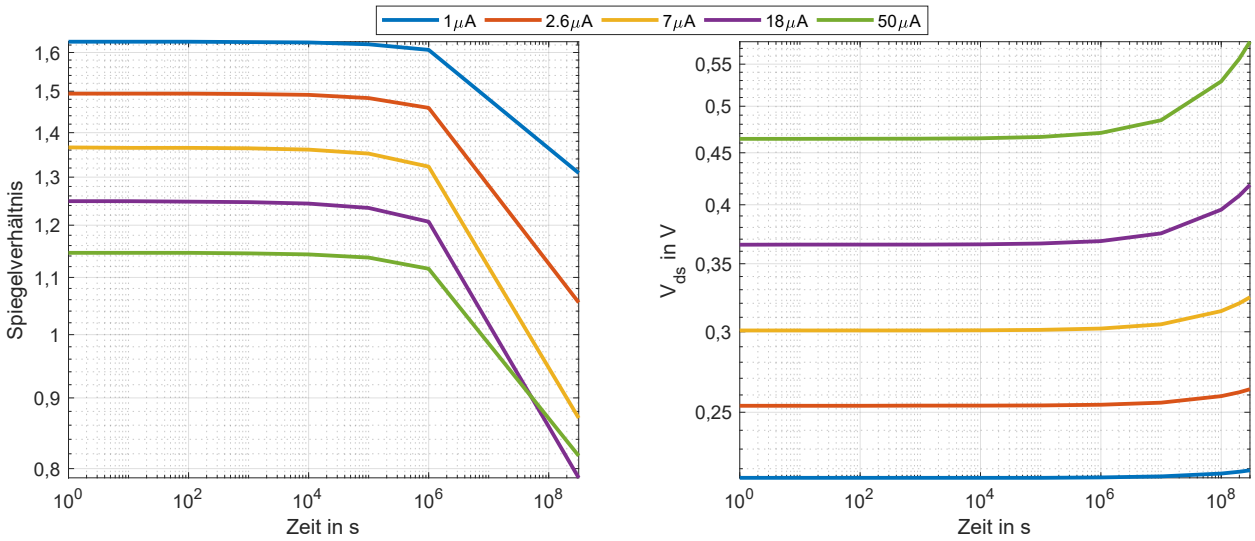


Abbildung 4.6: Spiegelverhältnis und Drain-Source-Spannungen eines NMOS Stromspiegel für verschiedene Biasströme für einen Zeitraum von 10 Jahren.

deren Arbeitspunkt über den Stromspiegel eingestellt werden sollte, folgt daraus eine deutliche Abweichung vom gewünschten Verhalten. Da der Stromspiegel die üblichste Art einer Bias-Schaltung darstellt, sich jedoch kein alterungsunempfindlicher Arbeitspunkt einstellen lassen kann, sind diese Funktionsblöcke nicht für hochzuverlässige Schaltungen zu empfehlen.

Kaskadierter Stromspiegel Der kaskadierte Stromspiegel folgt der Idee der Kaskode zur Erhöhung des Ausgangswiderstands. Dadurch wird die Abhängigkeit der Spiegelverhältnisse vom Biasstrom bzw. von den Drain-Source-Spannungen reduziert. Zudem bietet die Biasspannung einen Paramter zum Einstellen der Abhängigkeit von der Alterung. Der Verlauf des Spiegelverhältnisses über der Zeit ist in Abbildung 4.7 gezeigt. Die Biasspannungen wurden zwischen 0,4V und 1V variiert. Dabei zeigt sich, dass, die Einflüsse der Alterung reduziert werden sobald der Kaskodetransistor in starker Inversion betrieben wird. Für Biasspannungen oberhalb von 600mV wird

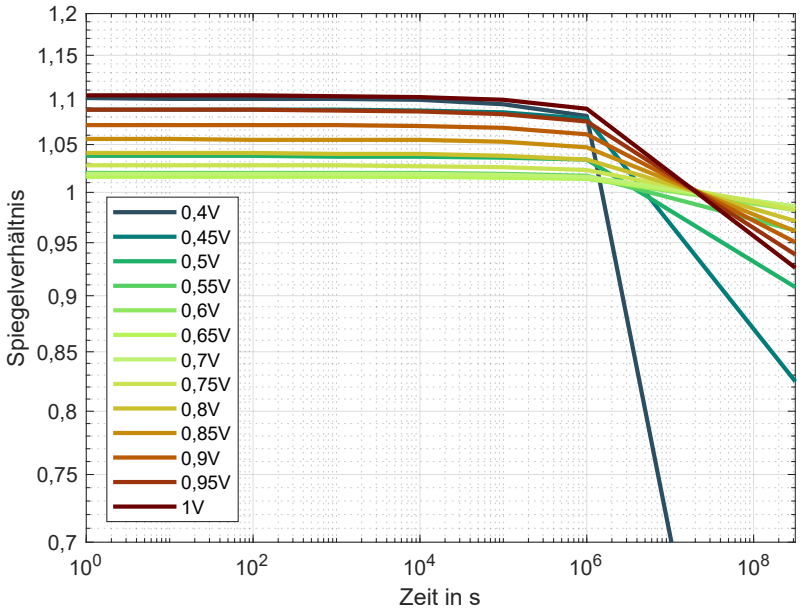


Abbildung 4.7: Spiegelverhältnis des kaskadierten NMOS Stromspiegels über 10 Jahre für verschiedene Biasspannungen.

die Änderung des Spiegelverhältnisses auf unter 3% abgesenkt. Werden die Kaskodetransistoren in starker Inversion und Sättigung betrieben, hat die Alterung auf diese nur leichte Auswirkungen, da nur die Reduktion von v_{sat} Einfluss auf den Drainstrom hat. Zudem ist die Drain-Source-Spannung der Spiegeltransistoren geringer, was zu geringeren Stressbedingungen führt.

4.1.2 Schaltungsdesign mit Funktionsblöcken

In diesem Abschnitt werden Möglichkeiten des Schaltungdesigns und Redesigns mit den zuvor vorgestellten Funktionsblöcken beschrieben. Zunächst wird beschrieben, wie eine bestehende Schaltung schnell verbessert werden kann. Anschließend werden in einer umfangreichen Fallstudie mit 7 Biasschaltungen und 8 Verstärkern die Designmöglichkeiten des Block-basierten Vorgehens beschrieben. Diese werden auf Grundlage der Funktionsblöcke bewertet und anschließend mit Simulationsergebnissen verglichen.

4.1.2.1 Redesign von Schaltungen

Die Funktionsblöcke können ebenfalls dazu verwendet werden, einen bestehenden Schaltungsentwurf zu verbessern oder diesen resistenter gegenüber einzelnen Einflüssen zu gestalten. Als Beispiel dient der Miller-OTA aus Abschnitt 3.2.2. Bei diesen sinkt die Verstärkung über den betrachteten Zeitraum ab und der niedrigste Verstärkungswert wird ebenfalls durch Alterung hervorgerufen. Wie bereits beschrieben, sind die Stromspiegel die Komponenten, die durch ihr Vorhandensein in jedem Zweig der Schaltung insgesamt die stärksten Auswirkungen auf die Alterung haben.

Um die Einflüsse durch Alterung abzuschwächen, werden nicht, wie in anderen Methoden übliche, einzelne Parameter der Transistoren untersucht oder die Dimensionierungen verändert, sondern es wird zuerst die Topologie verändert (vgl. Abb. 4.8). Nur so kann der Entwurfsraum vollständig untersucht werden. Eine andere Dimensionierung der

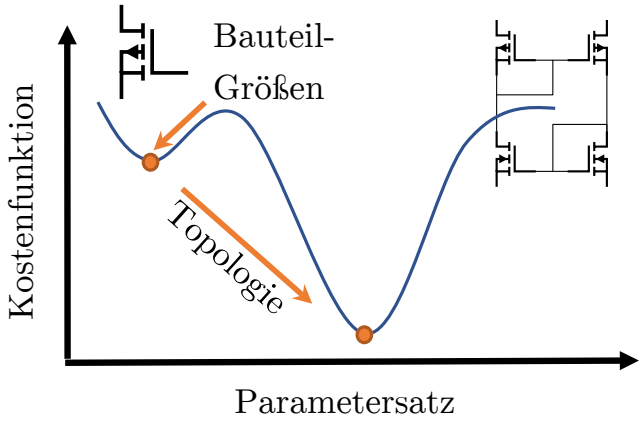


Abbildung 4.8: Beispielhafte Dimensionierung und Topologiewechsel, falls das Optimum nicht durch Dimensionierung allein erreichbar ist.

Stromspiegel würde in keinem Fall dazu führen, dass die Alterung abgeschwächt wird. Dies liegt daran, dass die relative Änderung der Stressbedingungen beider Transistoren zur Abweichung des gespiegelten Stroms führt. Eine nominelle Reduktion der Alterung würde das Ungleichgewicht nicht aufheben. Die vorangegangenen Analysen der einzelnen Blöcke führen dazu, dass die Stromspiegel durch kaskadierte Stromspiegel ersetzt werden sollten, um so eine deutliche Reduktion der Alterung der Verstärkung zu erzielen. Der in seine Funktionsblöcke unterteilte Miller-OTA ist in Abbildung 4.9 gezeigt. Neben

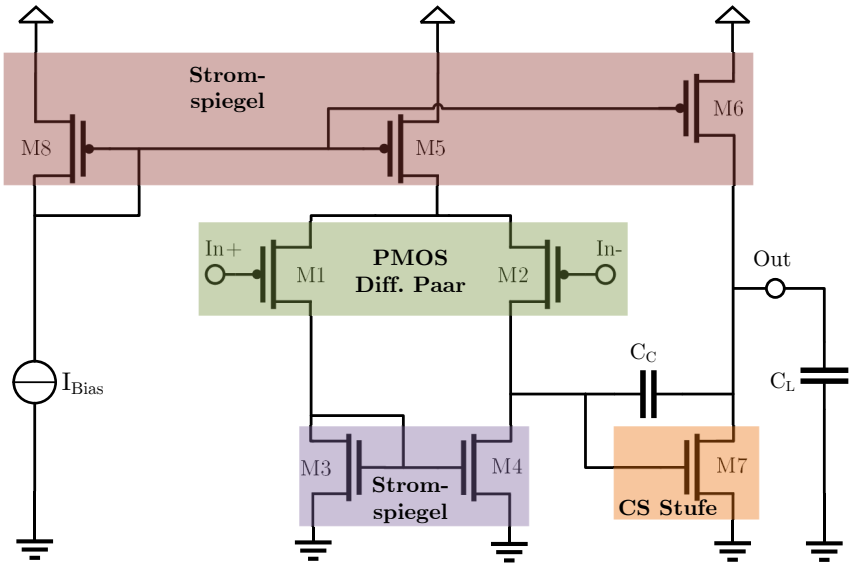


Abbildung 4.9: Miller-OTA unterteilt in funktionale Schaltungsblöcke.

dem Austausch der PMOS-Stromspiegel könnten weitere Maßnahmen ergriffen werden. Diese würde jedoch umfangreiche Neuauslegungen der Schaltungen nach sich ziehen. Solange alle Transistoren nach dem Einbringen der Kaskodetransistoren weiterhin in Sättigung arbeiten, besteht kein weiterer Änderungsbedarf. Eventuell muss die Miller-Kapazität durch den erhöhten Ausgangswiderstand angepasst werden, alle weiteren Spezifikationen bleiben jedoch unberührt. Nach Austausch der Stromspiegel muss noch die beste Biasspannung für die Kaskodetransistoren bestimmt werden. Die für die Bewertung verwendete Figure of Merit ist

$$FoM = \frac{\Delta\mu \cdot \max(\sigma(t))}{\sigma \cdot FoM_{ref}} \ , \tag{4.5}$$

dabei ist σ die Standardabweichung der Streuung der FoM im jeweiligen Arbeitspunkt. Als Referenz wird der nicht geänderte Miller-OTA (FoM_{ref}) verwendet. Mit dieser FoM werden Prozessvariation und Alterung gleichzeitig erfasst. Zur Bestimmung der Biasspannung wird diese im Bereich von 0V bis 1V variiert. Wie aus den Ergebnissen

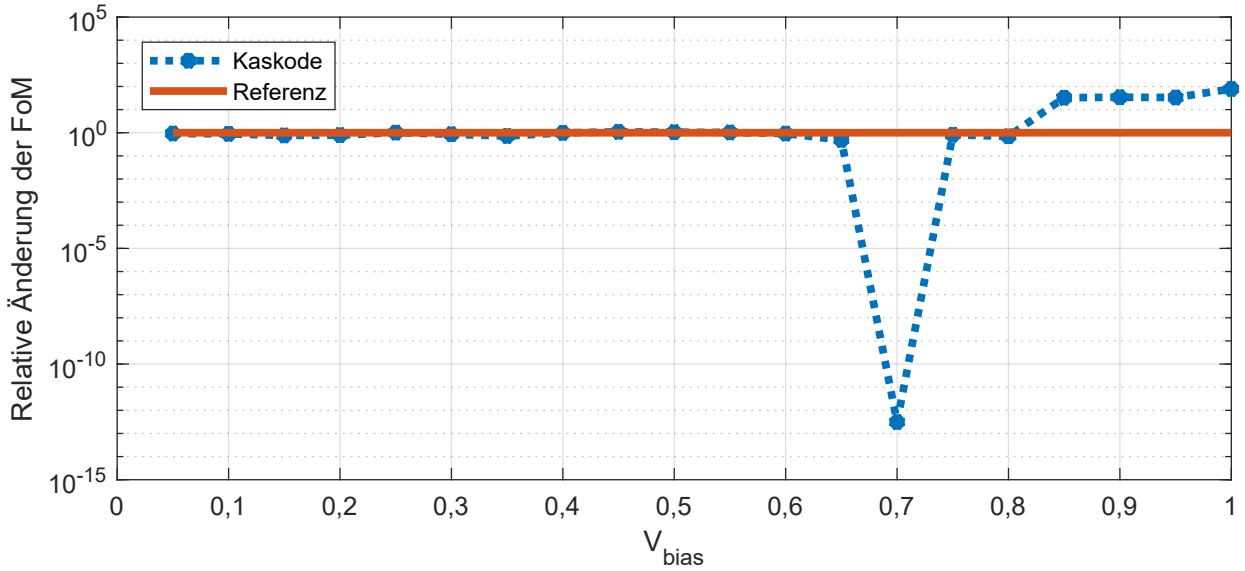


Abbildung 4.10: FoM für verschiedene Biasspannungen des Miller-OTA mit kaskadierten PMOS-Stromspiegeln.

der einzelnen Blöcke zu erwarten, zeigt sich das beste Ergebnis für etwa 0,7V. Danach steigt die FoM wieder an, da in sehr starker Inversion die Variation der Sättigungsgeschwindigkeit Einfluss auf den Drainstrom nimmt. Das Redesign basierend auf den Schaltungsblöcken versucht nicht, einzelne Werte zu verbessern, sondern verschiebt die gesamte Schaltung in einen Bereich, der resistenter gegenüber den Einflüssen ist. Zudem kann von Ergebnissen auf Blockebene auf solche auf Schaltungsebene geschlossen werden. Aussagen über das Verhalten einzelner Transistoren können hingegen nicht direkt auf Schaltungen verallgemeinert werden. Somit kann der Designer deutlich schneller und gezielter die Schaltung anpassen oder verbessern. Im Anschluss an den Blockaustausch

können weitere Verbesserungen durch Dimensionierung der Transistoren, falls notwendig, vorgenommen werden.

4.1.2.2 Fallstudie zu Biasschaltungen

In diesem Abschnitt werden 7 Topologien (8 Schaltungen) anhand ihrer Blockstruktur klassifiziert und die Ergebnisse mit Simulationen verglichen. Somit können direkt für einzelne Anwendungsbereiche die besten Topologien ohne die Notwendigkeit von Simulationen bestimmt werden. Ausgehend von den präsentierten Evaluationsergebnissen wurde die alterungsresistenteste Schaltungskombination ausgesucht und gefertigt.
Die Arbeitspunkte von allen analogen Schaltungen benötigen dedizierte Referenzspannungen. Diese können mit unterschiedlich hohem Schaltungsaufwand erzeugt werden. Dieser reicht von einer einfachen MOS-Diode bis hin zu temperaturkompensierten Schaltungen mit mehreren Dutzend Transistoren. Abbildung 4.11 zeigt die verwendeten Biasquellen. Die beiden einfachsten Quellen sind der NMOS- und PMOS-Stromspiegel,

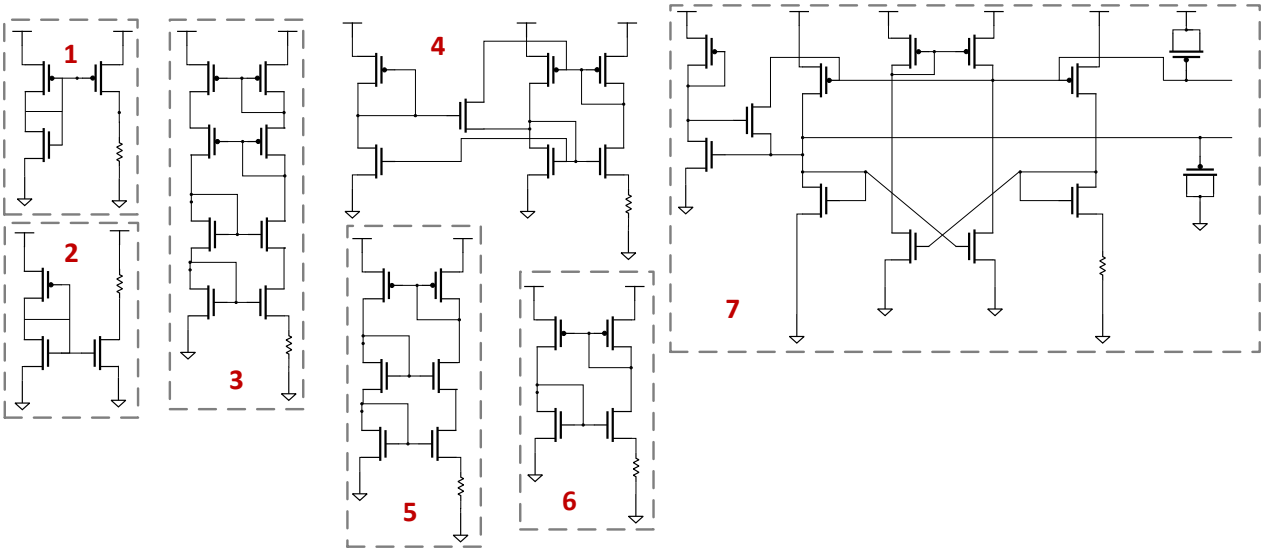


Abbildung 4.11: Verwendete Biasschaltungen. 1-2: Stromspiegel, 3-6: kaskadierte Stromspiegel, 7: Beta-Muliplier.

die als Lastelement jeweils den anderen Transistortyp als Diode verschaltet besitzen. Als weitere Quellen dienen vier geregelte Stromspiegel, bei denen jeweils ein PMOS-Stromspiegel in Serie mit einem NMOS-Stromspiegel angeordnet ist. Die Spiegelrichtungen sind dabei in den jeweils anderen Zweig gerichtet. Bei diesen Schaltungen ist zu beachten, dass sie eine Startup-Schaltung benötigen, da sämtliche Gates an Knoten angeschlossen sind die im ausgeschalteten Zustand ein nicht definiertes Potential besitzen. Dadurch können die Schaltungen ohne Startup-Schaltung nicht zuverlässig eingeschaltet werden. Als Variante der geregelten Stromspiegel werden zwei Schaltungen mit je einfacher und zweifacher Kaskode eingesetzt. Als aufwendigste Stromreferenz wird ein Beta-Multiplier mit zwei unterschiedlichen Dimensionierungen eingesetzt (Lang-

und Kurzkanal). Diese werden folgend als Topologie 7a (Kurzkanal) und Topologie 7b (Langkanal) benannt.

Der Beta-Multiplier ist in Abbildung 4.12 größer mit benannten Transistoren gezeigt [Bak08]. M1, M2 und M3 bilden die Startup-Schaltung. M1 ist als Diode verschaltet und es gilt $W/L << 1$. Dadurch fließt in dem Zweig des Transistors M1 nur ein sehr geringer Strom. Im ausgeschalteten Zustand liegt die Gatespannung von M3nahe 0V und somit die von M2 ebenfalls. Wird die Versorgungsspannung eingeschaltet, beginnt die Gate-Spannung von M2 zu steigen. Dadurch sind die beiden internen Knoten des Beta-Mulipliers über die Drain-Source-Strecke von M2 leitend verbunden. Unabhängig von den vorherigen Potentialen der inneren Knoten kann die Schaltung somit starten. Sobald sich durch M4 und M10 ein Zweigstrom einstellt, erhöht sich der Spannungsabfall über dem Transistor M5. Dadurch steigt die Gatespannung von M3, dieser Transistor steuert auf, und gleichzeitig verringert sich dadurch die Gatespannung von Transistor M2 und dieser steuert zu. Wird $V_{ds,5} = V_{gs,5} > V_{th,3}$ erreicht, wird die Startup-Schaltung

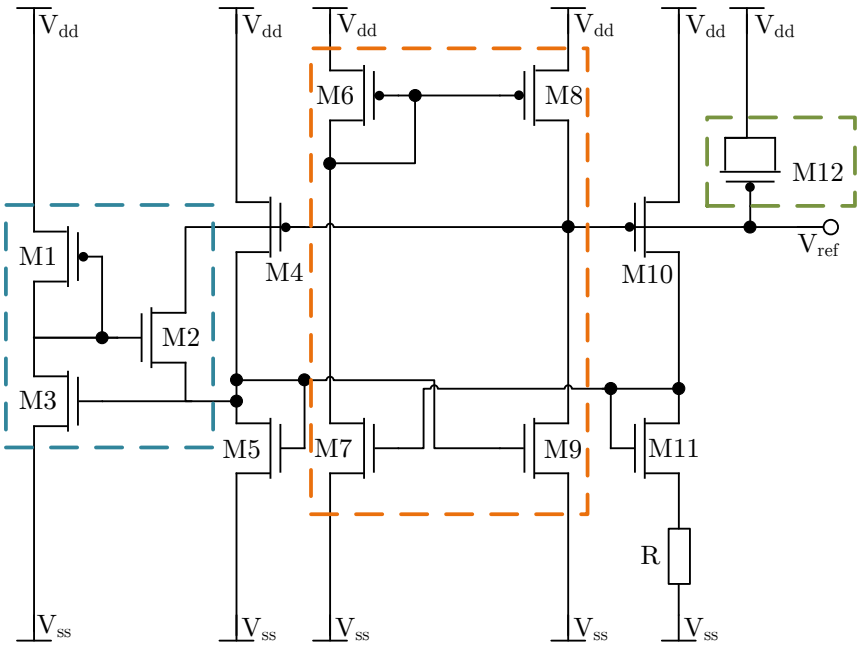


Abbildung 4.12: Beta-Multiplier mit MOS-Kapazität und Startup.

ausgeschaltet, und zwischen den internen Knoten existiert keine leitende Verbindung mehr. M12 ist eine MOS-Kapazität, die das Aufschwingen der Schaltung verhindern soll, da der obere interne Knoten einen sehr hohen Widerstand besitzt. Die Transistoren M6-M9 bilden einen Operationsverstärker, der Schwankungen der Versorgungsspannung ausregelt. Der eigentliche Beta-Multiplier besteht aus den Transistoren M4, M5, M10 und M11. Die Funktion der Schaltung folgt aus der Maschengleichung

$$V_{g,5} = V_{g,11} + I_{ref} \cdot R \tag{4.6}$$

Daraus folgt $V_{g,5} > V_{g,11}$ welches mit der Beziehung

$$V_{g,i} = \sqrt{\frac{2I_{d,i}}{\beta_i}} + V_{th,i} \tag{4.7}$$

nur erfüllt werden kann, wenn $\beta_5 > \beta_{11}$ gilt. Um das zu erreichen, reicht mit $\beta_i = \mu_i C_{ox,i} {}^{W_i}/_{L_i}$ aus, wenn $W_{11} = k \cdot W_5$ gilt. Somit ist $\beta_{11} = k \cdot \beta_5$. In dieser Beispielschaltung ist $k = 4$. Dieser Beziehung verdankt die Schaltung ihren Namen. Mit Umstellen von Gleichung (4.7) folgt:

$$I_{ref} = \frac{2}{R^2 \mu C_{ox} {}^{W_5}/_{L_5}} \left(1 - \frac{1}{\sqrt{k}}\right)^2 \tag{4.8}$$

Damit folgt für den Übergangsleitwert:

$$G_m = \sqrt{2\mu C_{ox} {}^{W}/_{L} I_{ref}} = \frac{1}{R} \tag{4.9}$$

und somit ist dieser unabhängig von MOSFET-Prozessstreuungen. Die Transistoren $M10$ und $M4$ spiegeln den Strom zurück in den Eingangszweig. Daraus folgt, dass die Transistoren $M4$, $M5$, $M10$ und $M11$ vom gleichen Strom I_{ref} durchflossen werden, der prozessstreuungs-unabhängig mittels R eingestellt werden kann. Gleichzeitig dient der Widerstand zur Begrenzung des positiven Feedbacks, da sonst ein erhöhter Strom in einem Zweig zu einer Erhöhung des Stromes im anderen Zweig führen würde. Ohne den Widerstand würde der Arbeitspunkt der Schaltung somit weglaufen. Als Ausgangsspannung wird

$$V_{ref} = V_{g,4} = \frac{2}{R\mu C_{ox} {}^{W_4}/_{L_4}} \left(1 - \frac{1}{\sqrt{k}}\right) + V_{th,4} \tag{4.10}$$

verwendet. Das Verhalten der Schaltungen über der Versorgungsspannung für verschiedene Temperaturen $T = \{-40, -20, 0, 27, 50, 80, 100, 120\}$°C ist in Abbildung 4.13 dargestellt. Einige Referenzgrößen zeigen eine starke Abhängigkeit von der Versorgungsspannung und/oder der Temperatur. Für eine Referenzquelle sind beide Abhängigkeiten so gering wie möglich zu halten, damit die Arbeitspunkte der nachgeschalteten Bauelemente möglichst konstant bleiben. Die stärksten Abhängigkeiten zeigen die Schaltungen, bei denen eine MOS-Diode die Referenzgröße einstellt. Dies folgt direkt aus den Ergebnissen der Block-Analyse. Des Weiteren liefern die Schaltungen mit ein- und zweifachen Kaskode-Stromspiegeln stabilere Referenzspannungen. Nichtsdestotrotz sind die Variationen zu stark für eine zuverlässige Referenzgröße. Die verschiedenen Beta-Multiplier halten, um den Punkt der nominellen Versorgungsspannung von 1V, den Arbeitspunkt über der Temperatur relativ konstant. Bei den beiden Beta-Mulipliern handelt es sich um eine Lang- und eine Kurzkanal-Variante der ansonsten gleichen Schaltungen. Eine weitere wichtige Einflussgröße ist die Alterung der Referenzquellen über verschiede-

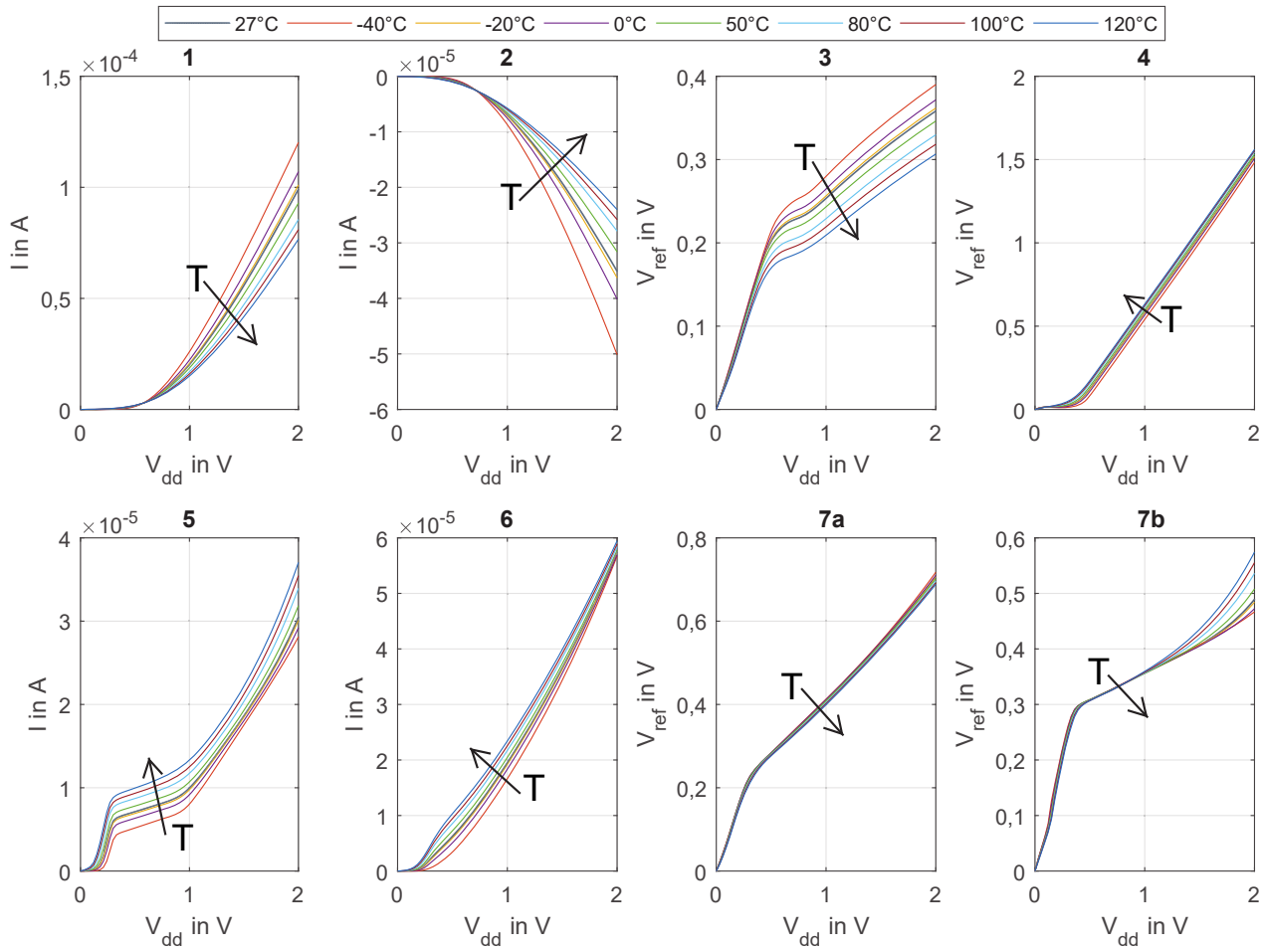


Abbildung 4.13: Referenzströme und Referenzspannung über der Versorgungsspannung für verschiedene Temperaturen $T = \{-40, -20, 0, 27, 50, 80, 100, 120\}$°C. Die Nummerierung entspricht den Schaltungen in Abbildung 4.11.

ne Temperaturen. Über einen Zeitraum von 10 Jahren sind die Ausgangsgrößen der betrachteten Quellen in Abbildung 4.14 gezeigt. Die Transistoren sind von HCI und BTI betroffen und die Alterung wurde durch eine einzige transiente Simulation bestimmt. Die Versorgungsspannung ist für die Alterungsanalyse auf 1V festgesetzt. Wie zu erkennen ist, ist der ungeregelte NMOS-Stromspiegel stärker von HCI betroffen. Die Ursache ist, dass die Stressbedingungen für BTI in dieser Schaltung, ein hohes V_{gs}, nicht vorhanden ist. Zudem ist der Strom durch den NMOS in diesem Design durch die sehr große Länge von $16,2\,\mu$m sehr gering und somit kommt es nur zu einer sehr geringen Alterung durch HCI. Bei den geregelten Stromspiegeln ist die Schaltung ohne Kaskode am stärksten von der Alterung betroffen. Bei den Beta-Multipliern ist zu erkennen, dass die Temperaturabhängigkeit der Alterung selbst dazu führt, dass sich die

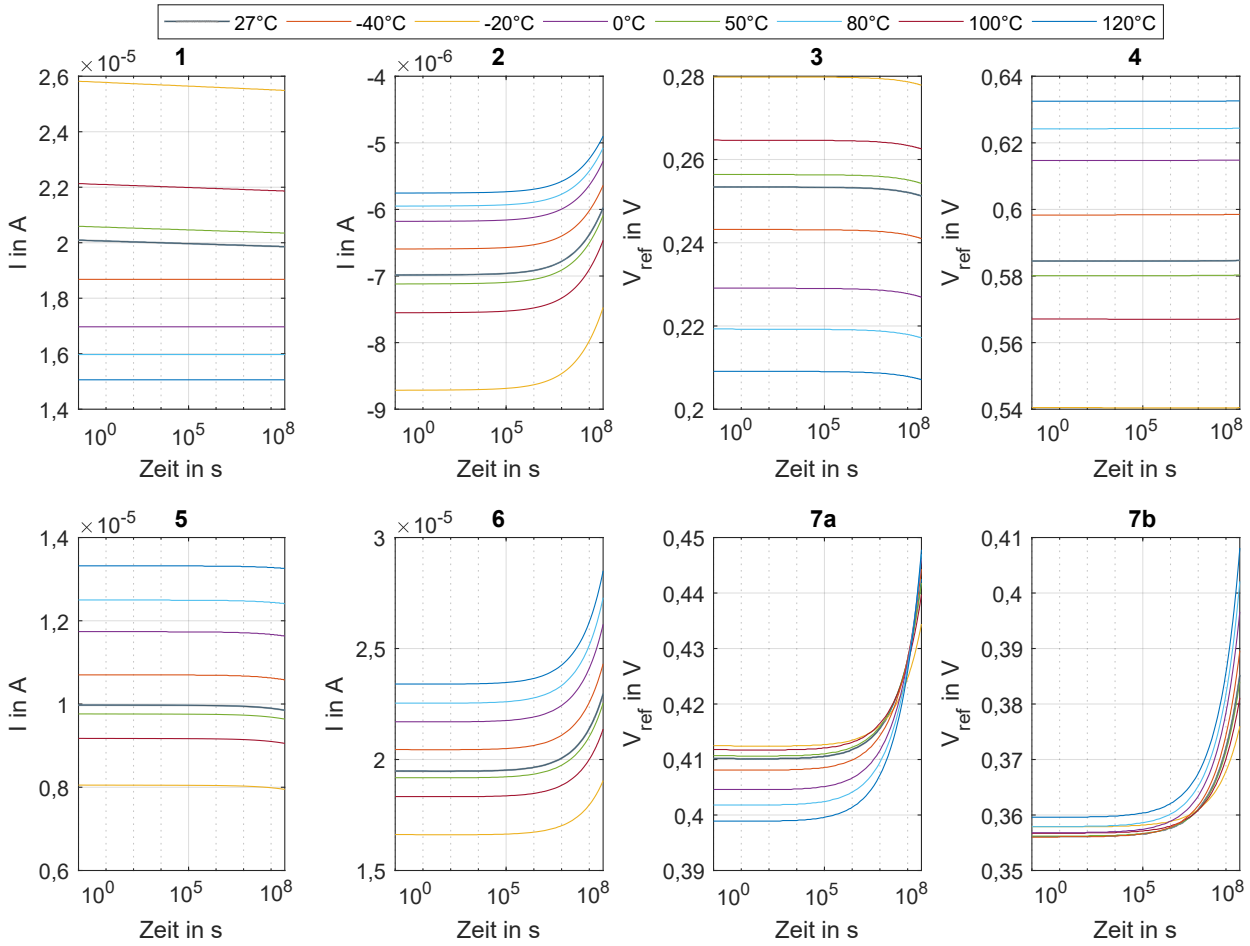


Abbildung 4.14: Referenzströme und Spannungen über 10 Jahre für verschiedene Temperaturen. Die Nummerierung entspricht den Schaltungen in Abbildung 4.11.

Temperaturabhängigkeit der Referenzspannung über der Zeit umkehrt. Die nominelle Temperaturabhängigkeit ist durch

$$\frac{\partial V_{ref}}{\partial T} = -\frac{2}{R\mu^2 C_{ox} {}^{W_4}/_{L_4}} \left(1 - \frac{1}{\sqrt{k}}\right) \cdot \left(\frac{1}{R}\frac{\partial R}{\partial T} + \frac{\partial \mu}{\partial T}\right) + \frac{\partial V_{th,4}}{\partial T} \tag{4.11}$$

gegeben. Durch die Alterung werden vor allem die Mobilität μ und die Schwellspannung V_{th} sowie die jeweiligen Temperaturabhängigkeiten verändert. Bei dieser Schaltung ist HCI der stärkere Einfluss, da die $V_{gs,i}$ der Transistoren ungefähr bei der halben Versorgungsspannung liegen und kaum BTI-Stress auftritt. Die Einflüsse durch HCI sind für niedrige Temperaturen verstärkt, da die Ladungsträger höhere Energien erreichen, da es seltener zu Gitterkollisionen kommt. Durch diesen Effekt ist die Referenzspannung nach 10 Jahren für geringe Temperaturen am niedrigsten, ohne Alterung wäre dies genau andersherum. Dies hat des Weiteren starke Auswirkungen auf eventuelle zusätzliche Schaltungen zur Temperaturkompensation. Diese würde im Laufe der Zeit ihre Wirkung verlieren bzw. einer Kompensation sogar entgegenwirken.

Die geringere Abhängigkeit der Beta-Multiplier von der Temperatur und der Versorgungsspannung geht jedoch auf Kosten des Flächen- und Leistungsbedarfs der Schaltungen. Ein Vergleich der Referenzquellen sowie die qualitativen und quantitativen Abhängigkeiten von Versorgungsspannung und Temperatur ist in Tabelle 4.1 gezeigt. Die qualitativen Abhängigkeiten wurden dabei aus den Blöcken der Schaltungen abgeleitet und die quantitativen Variationen aus der Schaltungssimulation bestimmt. Die Pfeile der qualitativen Bewertungen zeigen die jeweilige Abhängigkeit der Schaltung von Zeit oder Temperatur. Die Skala reicht von $\uparrow$ - sehr starker, $\nearrow$ - starker, $\rightarrow$ - mittlerer , $\searrow$ - schwacher bis $\downarrow$ - nahezu keiner Abhängigkeit. Wie aus den Werten der Tabelle

Tabelle 4.1: Vergleich der Referenzquellen sowie der qualitativen (Vorhersage mit Funktionsblöcken) und quantitativen (SPICE-Simulation) Abhängigkeiten von Alterung und Temperatur.

Nr.	Fläche in μm^2	Leistung in μW	Quali. Zeit	Quali. Temp.	Quant. Zeit in %	Quant. Temp. in %
1	10,152	23,9	$\downarrow$	$\uparrow$	1,68	53,95
2	3,384	12,09	$\uparrow$	$\uparrow$	14,48	41,5
3	5,796	29,25	$\rightarrow$	$\rightarrow$	0,87	27,87
4	4,5	61,9	$\nearrow$	$\rightarrow$	0,03	15,96
5	5,896	35,8	$\rightarrow$	$\rightarrow$	1,29	53,8
6	4,356	111,58	$\nearrow$	$\nearrow$	17,9	35,38
7a	42,876	68,67	$\uparrow$	$\searrow$	7,9	3,31
7b	40,74	108,9	$\uparrow$	$\searrow$	8,17	0,96

hervorgeht, können die quantitativen Ergebnisse der Simulationen recht genau qualitativ durch die Verwendung der Block-Analysen bestimmt werden. Die Abweichungen für Schaltung 4 resultieren aus der Rückkopplung der enthaltene Stromspiegel, welche nicht vollständig auf Block-Level abgebildet werden können. Für Schaltung 6 kommt es zu einer leichten Unterschätzung, die aus den gewählten sehr hohen Strömen bzw. geringen Kanallängen folgt, die zu einem verstärkten Einwirken durch HCI führen. Aus den Block-Level-Analysen folgt, dass die Kaskodetransistoren zu einer Reduktion der Alterung führen. Für die Block-Analyse der Beta-Multiplier ist maßgeblich der NMOS-Transistor im Ausgangszweig bzw. die Veränderung des g_m relevant sowie der PMOS-Stromspiegel. Letzterer ist jedoch durch die geringe Gatespannung V_{gs} kaum von Alterung beeinflusst. Das Temperaturverhalten wird insgesamt korrekt abgeschätzt, wobei die geringe Temperaturabhängigkeit des Kurzkanal-Beta-Multipliers eine Eigenart eben dieser Schaltung ist. Insgesamt steigt die Robustheit der Referenzgröße gegenüber Alterungs- und Spannungseinflüssen mit dem Schaltungsaufwand.

Wie aus den vorherigen Analysen zu erkennen ist, sind die Beta-Multiplier-Schaltungen relativ stark von Alterung betroffen, wenngleich sie am geringsten durch Temperatur

und Versorgungsspannung beeinflusst werden. Somit kann keine Schaltung bestimmt werden, die gegenüber PVTA ausreichend resistent ist. Um diese Lücke zu schließen, wurde in dieser Arbeit eine Online-Überwachung für Referenzquellen entworfen, die Startup-Schaltungen verwendet, um die Alterung der Quelle zu überwachen und ggf. mit einem Body-Bias den Auswirkungen entgegenwirkt. Beim Body-Bias werden die Bulk-Kontakte der Transistoren mit einer Spannung oberhalb von V_{dd} für PMOS und unterhalb von V_{ss} für NMOS versorgt, um die zusätzliche Ladung durch Alterung im Gateoxid auszugleichen. Dies folgt aus der Gleichung für die Bulkladung nach [Aga14]

$$Q_b = -\text{sign}(\Psi_s) \cdot \gamma \cdot C_{ox} \sqrt{U_t \cdot (\exp(-\Psi_s/U_t) - 1) + \Psi_s} \tag{4.12}$$

und damit gilt weiter für die Gate-Spannung

$$V_g = V_{FB} + \Psi_s - \frac{Q_i + Q_b}{C_{ox}} + \left(\frac{Q_i + Q_b}{\gamma \cdot C_{ox}}\right)^2 . \tag{4.13}$$

Um diesen Zusammenhang auszunutzen, wird ein Operationsverstärker an die Drain- und Source-Kontakte vom Transistors M2 wie in Abbildung 4.15 gezeigt angeschlossen.

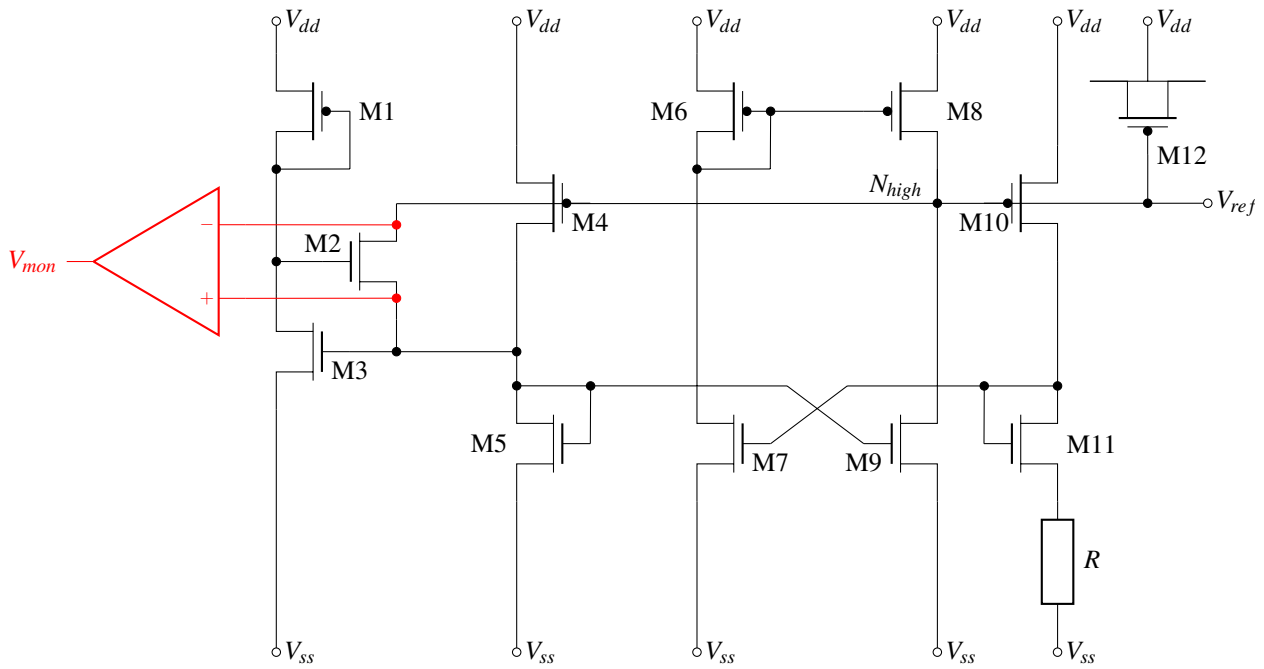


Abbildung 4.15: Beta-Multiplier-Schaltung mit zusätzlichem Verstärker für online Alterungsüberwachung.

Der dortige Spannungsabfall ist für einen weiten Temperaturbereich proportional zur Änderung der Referenzspannung, wie in Abbildung 4.16 dargestellt ist. Diese Spannung ist proportional zum alterungsabhängigen Leckstrom durch die Drain-Source-Strecke von M2. Gleichzeitig wird die Veränderung des Offsets des Verstärkers (M6-M9) mit erfasst. Der Spannungsabfall über der Drain-Source-Strecke von M2 beträgt 450mV

bei 20°C und $V_{dd} = 1\,\mathrm{V}$. Für Temperaturen oberhalb von 120°C und für sehr starke Alterung geht der lineare Zusammenhang verloren.

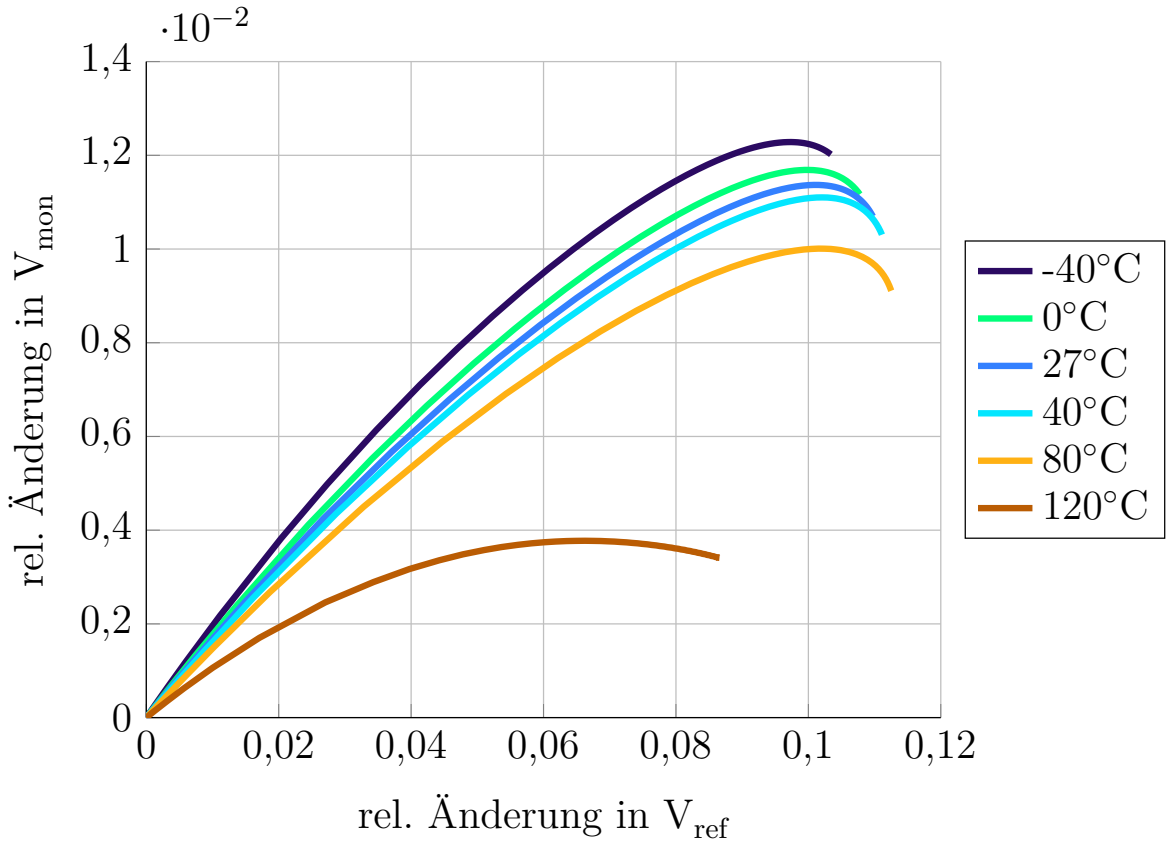


Abbildung 4.16: Verhältnis zwischen V_{ref} und der Monitorspannung V_{mon} für 10^6 Sekunden.

Mit diesem Referenzsignal wird anschließend die Spannung des Body-Anschlusses der PMOS- und NMOS-Transistoren gesteuert. Diese Maßnahme ist durch die notwendigen Ladungspumpen sehr flächen- und leistungsintensiv, jedoch die einzige Möglichkeit, eine Referenzquelle PVTA resistent zu entwerfen. Aus den Verläufen der kompensierten Schaltung in Abbildung 4.17 ist zu erkennen, dass es durch die Regelung des Body-Bias zu einer nahezu vollständigen Unterdrückung der Alterungsabhängigkeit kommt.

Die erzeugte Referenzspannung kann anschließend zur Arbeitspunkteinstellung anderer Schaltungsteile verwendet werden. Dadurch, dass bereits die Referenzquelle alterungsresistent entworfen wurde, werden die Arbeitspunkte der weiteren Schaltungsanteile nur durch die eigene Alterung verändert.

In diesem Abschnitt wurde gezeigt, dass mit den REL-MOS-Modellen alle notwendigen Analysen und vor allem Kombinationen dieser Analysen für eine umfassende Betrachtung der PVTA-Einflüsse während des Schaltungsentwurfs durchgeführt werden können. Zudem können die qualitativen Erkenntnisse auf Block-Ebene direkt auf Schaltungsebene verwendet werden, um die relevanten Eigenschaften (z.B. Referenzspannung) einer Schaltungstopologie mit reduzierten Simulationsaufwand erfassen zu können. Aufbauend

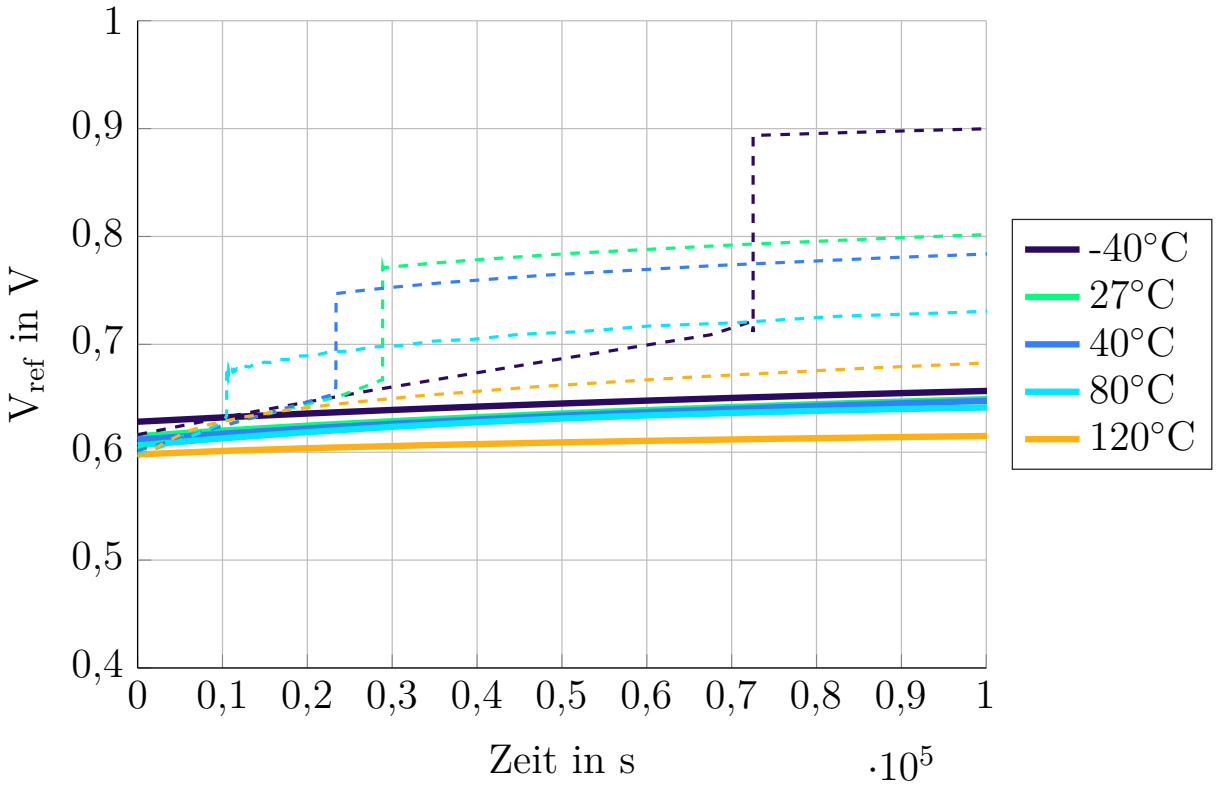


Abbildung 4.17: Vergleich zwischen kompensierter (durchgezogene Linie) und unkompensierter (gestrichelte Linie) Ausgangsspannung V_{ref}.

auf diesen Ergebnissen wurde eine Kompensationstechnik für Alterungseinflüsse für Beta-Multiplier-Referenzquellen entwickelt, die ohne oder mit geringen Eingriffen in die Schaltungstopologie auskommt. Somit bleibt die Resistenz des Beta-Multipliers gegenüber PVT-Einflüssen unberührt.

4.1.2.3 Fallstudie zu Verstärkern

Für die Auswahl des zu fertigenden Verstärkers wurden insgesamt neun verschiedene Topologien betrachtet. Diese sind im Anhang B detailliert mit ihren jeweiligen Funktionsblöcken gezeigt. Abbildung 4.18 zeigt das Verstärkungs-Bandbreite-Produkt der Verstärkertopologien für verschiedene Temperaturen $T = \{-40, -20, 0, 27, 50, 80, 100, 120\}$°C über 10 Jahre. Die Nummerierung entspricht der, die im Anhang verwendet wird. Wie

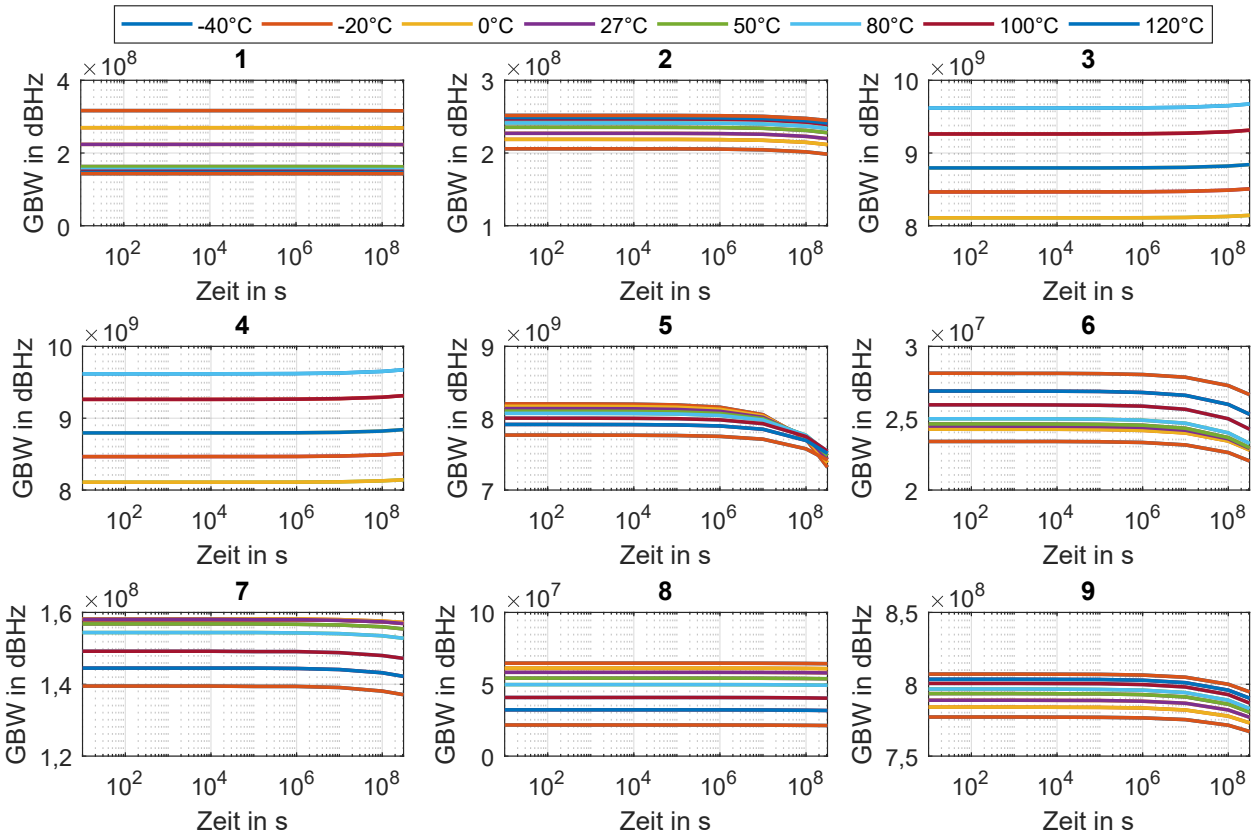


Abbildung 4.18: Verstärkungs-Bandbreite-Produkt der Verstärkertopologien für verschiedene Temperaturen $T = \{-40, -20, 0, 27, 50, 80, 100, 120\}$°C über 10 Jahre. 1: Miller-OTA, 2: Operationsverstärker, 3/4: Differentieller Kaskode-Operationsverstärker , 5. Folded-Cascode-Operationsverstärker, 6: Folded-Cascode-OTA, 7: Zweistufiger Verstärker mit Kompensation, 8: Dreistufiger Verstärker, 9: Zweistufiger Verstärker.

zu erkennen ist, zeigen die betrachteten Topologien sehr unterschiedliche Veränderungen über dem Alter und der Temperatur. So nimmt das Verstärkungs-Bandbreite-Produkt (*engl. Gain-Bandwidth-Product*, GBW) bei einigen Verstärkern über der Zeit ab und bei einigen anderen zu. Die Veränderung des GBW für verschiedene Temperaturen variiert stark zwischen den Topologien. So existieren einige Schaltungen, bei denen sich das GBW unter 10% über der Temperatur verändert. Entgegengesetzt dazu streuen andere Toplogien um bis zu 50%. Ähnliches gilt für die Degradation. Bei einigen Schaltungen

ist das Alterungsverhalten selbst stark von der Temperatur abhängig (Topologie 5) und bei anderen sind die absoluten Änderungen fast temperaturunabhängig. Zudem kann es zu sehr komplexem Verhalten kommen, wie der dargestellte Versorgungsspannungsdurchgriff (*engl. Power Supply Rejection Ratio*, PSRR) der Topologien über der Zeit und Temperatur in Abbildung 4.19 zeigt. Das Alterungsverhalten des PSRR hängt für

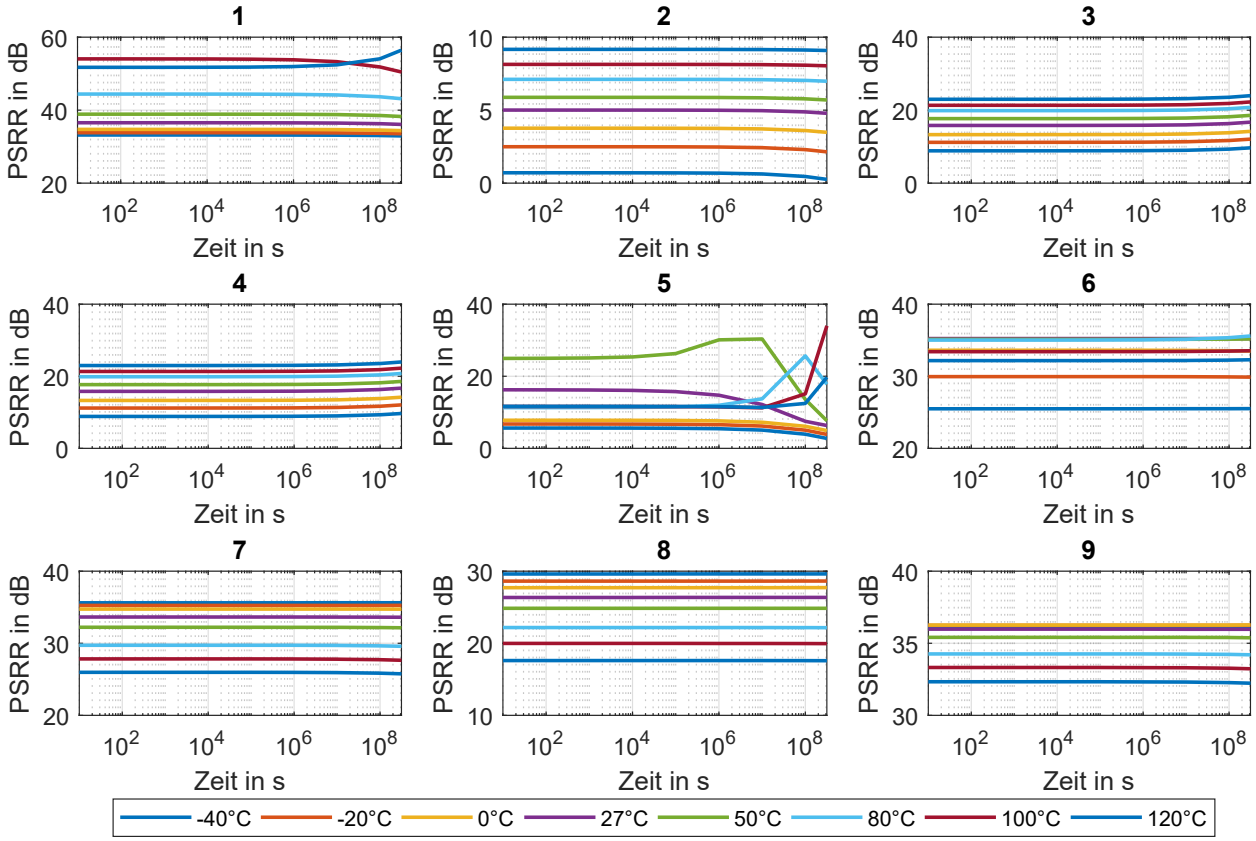


Abbildung 4.19: PSRR der Verstärkertopologien für verschiedene Temperaturen $T = \{-40, -20, 0, 27, 50, 80, 100, 120\}$°C über 10 Jahre.

einige Toplogien stark von der Temperatur ab. Dies zeigt abermals, dass eine reine Vorhersage der Abhängigkeiten durch Extrapolation zu falschen Ergebnissen führen würde. Des Weiteren zeigt sich im Vergleich von GBW und PSRR, dass eine Ableitung zwischen den Eigenschaften der Topologien nicht möglich ist. Somit muss jede Kennzahl getrennt analysiert werden. Das PSRR ist besonders interessant, weil es selbst angibt, wie sensitiv eine Schaltung gegenüber Störungen der Versorgungsspannungen reagiert. Erwartungsgemäß weisen die verschiedenen Schaltungen deutliche Unterschiede im PSRR auf. Jedoch zeigt sich, dass einige Schaltungen im Laufe der Zeit sensitiver bzw. weniger sensitiv auf Schwankungen von V_{dd} reagieren. Diese Kreuzabhängigkeit ist vor allem für Schaltungen und Anwendungsgebiete interessant, bei denen beide Einwirkungen zu erwarten sind.

Neben dem Verhalten, dass abhängig vom Belastungsprofil der Schaltung ist, sind die Prozessvariationen ein weiterer wichtiger Punkt für die Auswahl einer geeigneten Topo-

logie. Dazu zeigt Abbildung 4.20 die Verstärkung der Verstärkertopologien für jeweils 200 simulierte Schaltungen mit Prozessvariationen über einen Zeitraum von 10 Jahren. Wie bei den Topologien 3, 4 und 5 zu erkennen ist, beeinflussen die Prozessvariationen

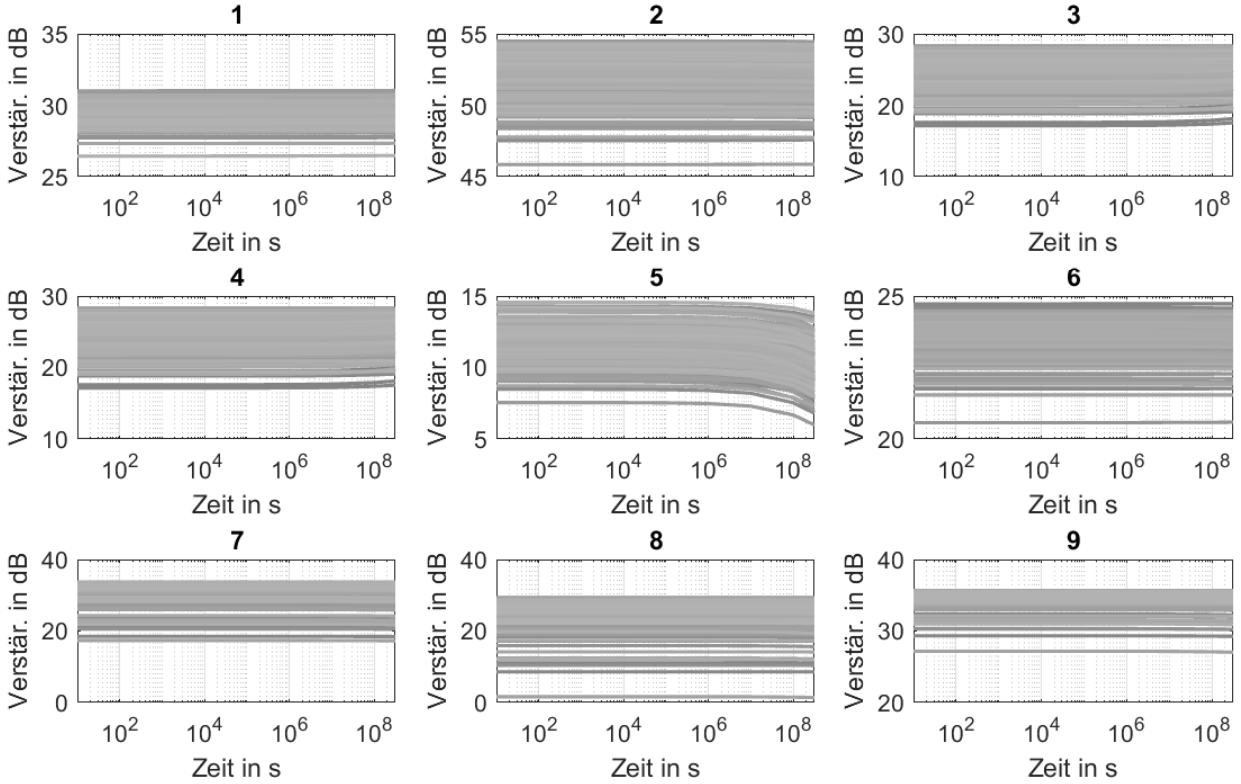


Abbildung 4.20: Verstärkung der Verstärkertopologien mit Prozessvariationen jeweils 200 Schaltungen über 10 Jahre.

die Alterung so stark, dass sich ihr Verlauf über die 200 betrachteten Schaltungen verändern kann. Besonders stark betroffen sind dabei die Ausläufer der jeweiligen Verteilung. Als Beispiel sind die Histogramme der Verteilungsfunktionen der Verstärkung der Topologien 3 und 5 für den ungealterten und gealterten Zustand in Abbildung 4.21 gezeigt. Wie zu erkennen ist, sind beide Verteilungen nicht normalverteilt. Dies ist für die entwickelte Messmethode im folgenden Kapitel wichtig, da diese speziell dazu gedacht ist, nicht-normalverteilte Funktionen korrekt schätzen zu können. Die obere Verteilungsfunktion weist eine Häufung um 12dB Verstärkung auf und eine weitere um 25dB. Des Weiteren verschieben sich die unteren Werte zu niedrigen Verstärkungen. Demgegenüber zeigt sich, dass bei den Schaltungen um 25dB Verstärkung über das Alter die Verstärkung erhöht wird. Somit ist das korrekte Abbilden der Verteilungsfunktion sehr wichtig für die Bestimmung der Zuverlässigkeit von Schaltungen, da sich die Ausläufer der Verteilung entgegengesetzt oder zumindest anders verhalten können als die überwiegenden Teile der Schaltungen. Bei der zweiten Verteilungsfunktion ist festzustellen, dass diese asymmetrisch ist und einen starken einseitigen Ausläufer besitzt. Wird eine solche Funktion mit einer Normalverteilung angenähert, wird fälschlicherweise

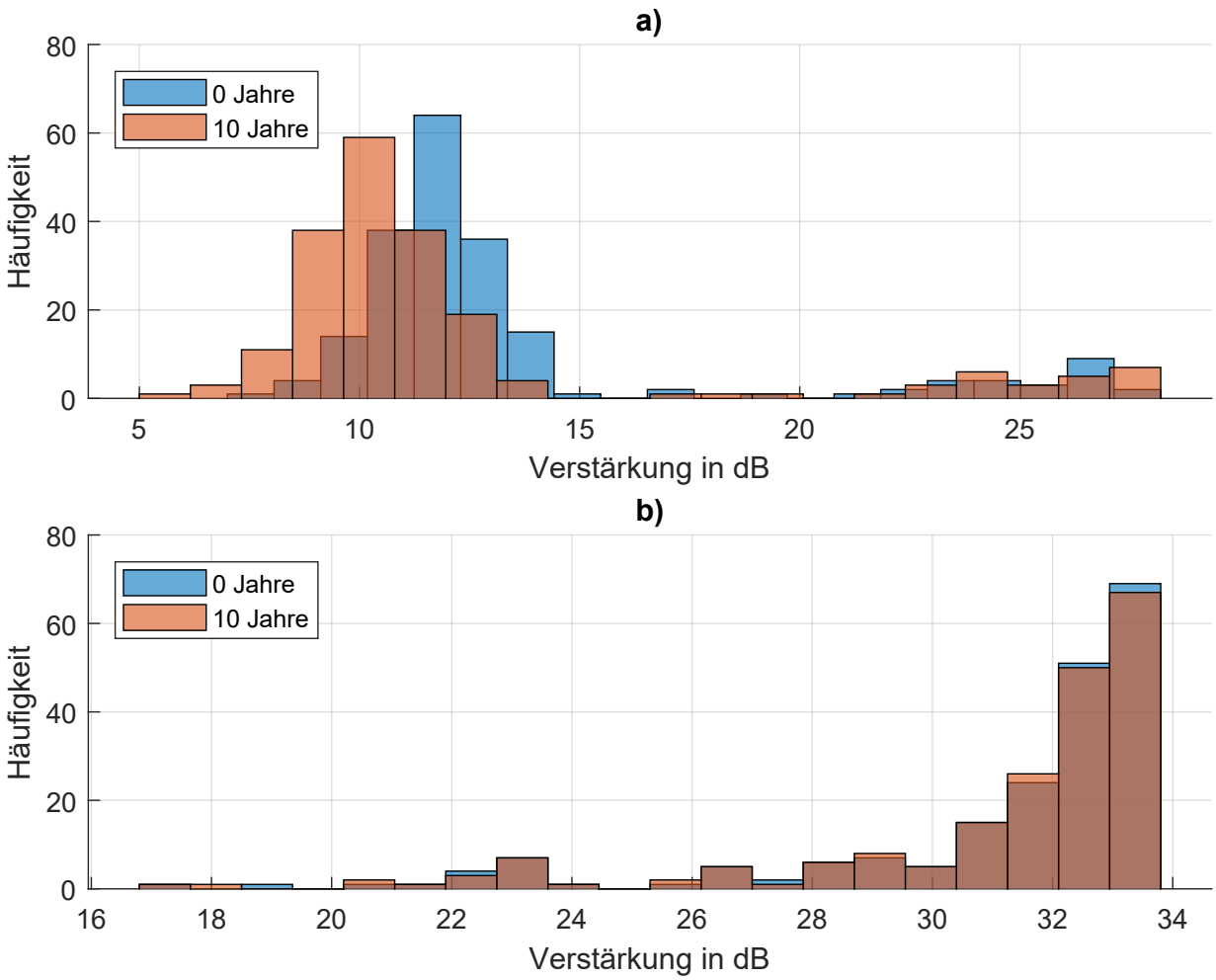


Abbildung 4.21: Histogramme der Verteilungsfunktionen de Verstärkung der Topologien 3 (b) und 5 (a) für den ungealterten und gealterten Zustand.

angenommen, dass auch Schaltungen oberhalb von 34dB Verstärkung auftreten können. Dies führt ebenfalls zu falschen Annahmen über die Ausbeute und die Zuverlässigkeit von Schaltungen. Im Gegensatz zur ersten Verteilung ändert sich die der Topologie 3 kaum über der Zeit. Eine nähere Betrachtung jeder einzelnen Topologie kann dem Anhang B entnommen werden.

Die einzelnen Topologien weisen entweder eine zu hohe Temperaturabhängigkeit einzelner Parameter auf oder altern zu stark. Daher wird eine Topologie anhand der Block-Strukturen entwickelt, die besonders resistent ist. Dazu werden für jeden Teil der Schaltung immer solche Blöcke verwendet, deren Analyse ergeben hat, dass diese kaum von der Alterung betroffen sind. Dazu muss der Ausgangszweig sowohl PMOS- als auch NMOS-Kaskoden enthalten, damit die Transistoren mit der jeweiligen Biasspannung in einen alterungsunabhänigigen Arbeitspunkt gebracht werden können. Um gleichzeitig einen annehmbaren Bereich für die Ein- und Ausgangspannungen zu erhalten, wird eine gefaltete Kaskode als Eingangspaar verwendet, deren Arbeitspunkte wiederum über

eine Kaskode eingestellt werden. Die entworfene Topologie ist in Abbildung 4.22 gezeigt und folgt allen Erkenntnissen aus der Fallstudie zu den Verstärkertopologien.

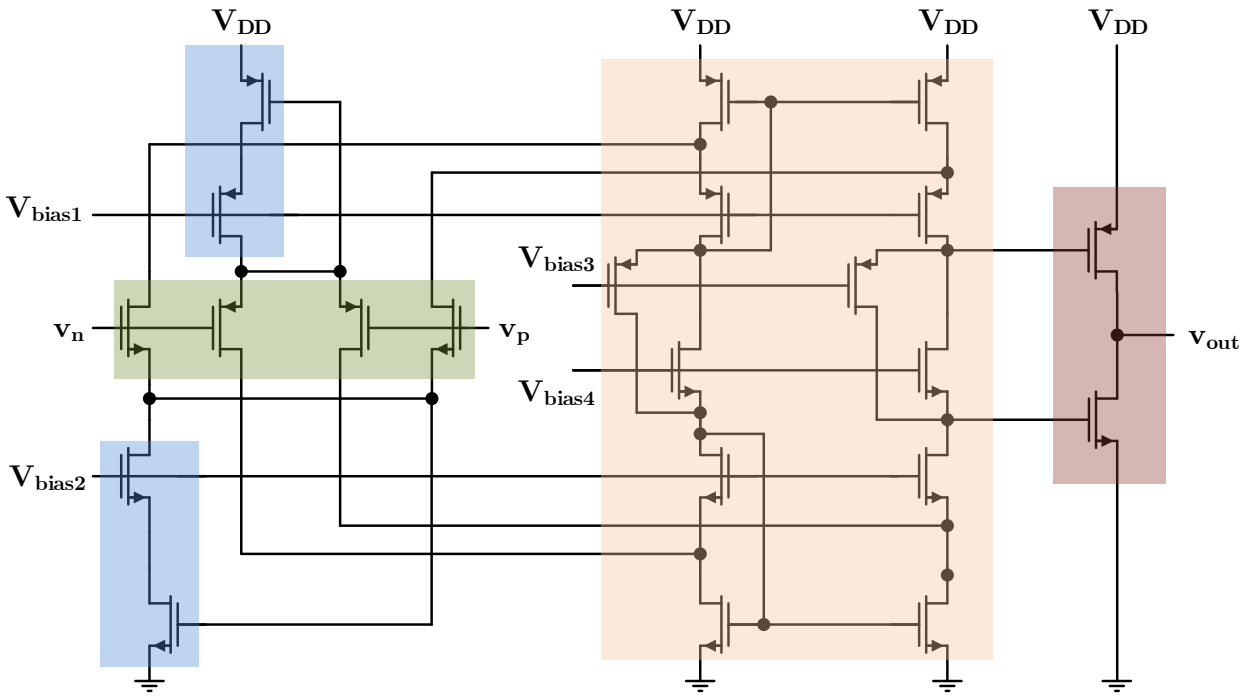


Abbildung 4.22: Alterungsresistente Verstärkertopologie mit den jeweiligen Funktionsblöcken.

Die verschiedenen Schaltungseigenschaften sind über das Alter in Abbildung 4.23 gezeigt. Dabei sind die Bandbreite, die Verstärkung, das Verstärkung-Bandbreite-Produkt, die Phasenreserve und die Transitfrequenz gezeigt. Für die gewählte Topologie sind die Werte vom Gleichspannungsverhalten sowie vom Frequenzverhalten der Schaltung abhängig. Somit geben diese Parameter eine Übersicht des Alterungsverhaltens aller Schaltungskomponenten. Wie erwünscht, sind alle Änderungen der Parameter unterhalb von 1%, somit hat die Alterung praktisch keinen Einfluss mehr auf die Schaltungseigenschaften dieser Topologie.

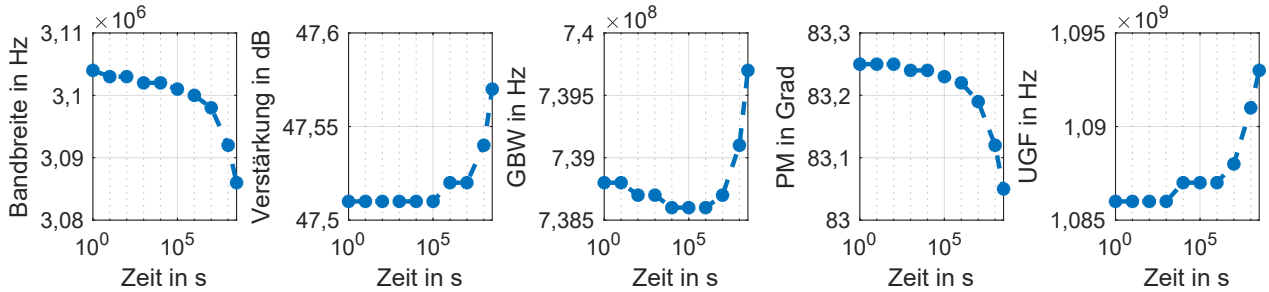


Abbildung 4.23: Bandbreite, Verstärkung, GBW, Phasenreserve (PM) und UGF der alterungskompensierten zweistufigen Verstärkertopologie über einen Zeitraum von 10 Jahren.

In diesem Abschnitt wurde gezeigt, dass das herkömmliche Vorgehen, bei dem verschiedene Topologien simulationsbasiert analysiert werden, nicht zu optimalen Ergebnissen führen, wenn einzelne Einflüsse gezielt unterdrückt werden sollen. Mit dem blockbasierten Entwurf wurde eine Verstärkertopologie ohne vorherige Simulationen gewählt, bei der jeder Block bereits kaum vom Alter abhängig ist, diese Eigenschaft überträgt sich auf die Gesamtschaltung.

4.2 Auswertung der Systemzuverlässigkeit

Neben dem Entwurf von Schaltungen können die zuvor vorgestellten Schaltungsblöcke ebenfalls für die Vorhersage der Ausbeute von Schaltungen verwendet werden. Somit kann der Designer direkt nach der Wahl einer Topologie überprüfen, ob diese die gewünschten Eigenschaften über den gesamten Nutzungszeitraum der Schaltung einhalten kann. Der Berechnungsablauf ist in Abbildung 4.24 gezeigt. Durch die blockweise Betrachtung ist es möglich direkt wie im vorherigen Abschnitt gezeigt einzelne Blöcke auszutauschen, falls die Ausbeute zu gering ausfällt. Während des Schaltungsentwurfes stehen somit transparente Stellschrauben für das Design bereit. Die Analyse

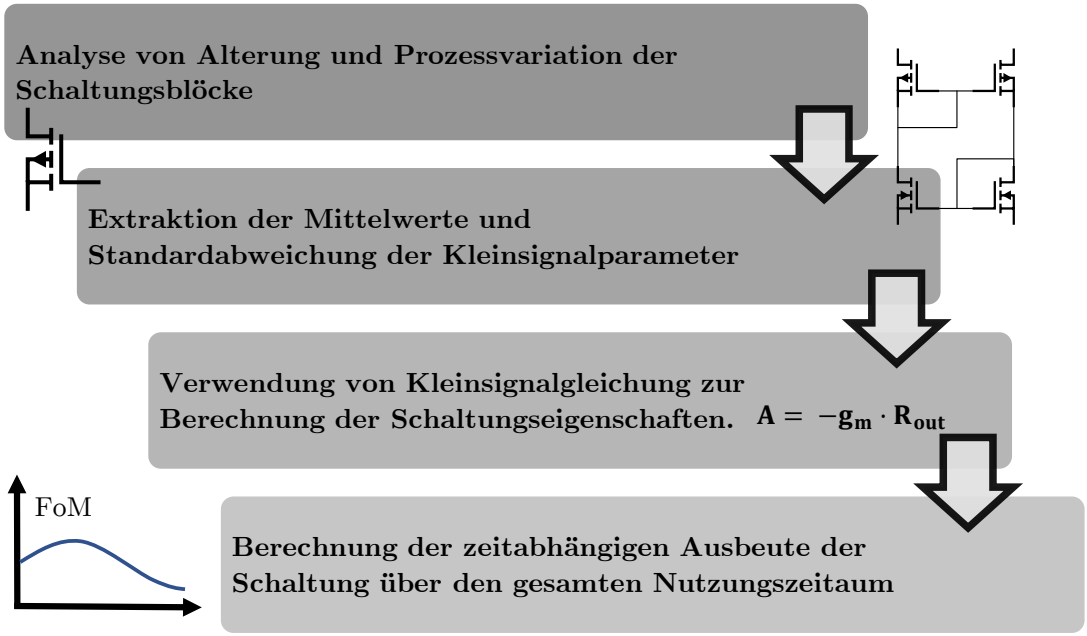


Abbildung 4.24: Ablauf der Berechnung der zeitabhängigen Ausbeute einer Schaltung.

und Speicherung der zeitabhängigen Momente geschieht in ähnlicher Form wie für die Transistorparameter in der erweiterten g_m/I_D-Methode. Durch die Speicherung dieser Teilergebnisse müssen zur Analyse verschiedener Topologien keine weiteren Simulationen durchgeführt werden. Falls in der jeweiligen Fertigungstechnologie bereits Schaltungen mit dieser block-basierten Entwurfsmethode erstellt worden sind, liegen diese Daten bereits vor und können vollständig wiederverwendet werden. Die Kleinsignalgleichungen der zu analysierenden Topologie werden anschließend verwendet, um einen analytischen Ausdruck für Schaltungseigenschaften zu erstellen. Anschließend wird mit Monte-Carlo-Sampling die zeitabhängige Verteilungsfunktion der Schaltungseigenschaft angenähert. Dies kann zum Beispiel in Computeralgebra-Programmen wie MATLAB durchgeführt werden.

Als Beispiel wird zum Vergleich von Verstärkertopologien

$$FoM = \frac{\text{Verstärkung} \cdot \text{Bandbreite} \cdot \text{Phasenreserve}}{\text{Leistung} \cdot \text{Fläche}} \tag{4.14}$$

als Figure-of-Merit verwendet werden. Diese FoM erfasst sowohl die nominelle Funktion des Verstärkers als auch die Effizienz der Schaltung. Diese enthält alle Schaltungseigenschaften, die die Funktion eines Verstärkers in erster Linie charakterisieren. Die Kleinsignalgleichungen der einzelnen Eigenschaften sind durch

$$\text{Verstärkung} = \frac{g_{m1}}{g_{ds,1} + g_{ds,3}} \cdot \frac{g_{m,7}}{g_{ds,7} + g_{ds_6}} \tag{4.15}$$

$$\text{Bandbreite} = \frac{1}{2\pi \text{Verstärkung}/\mathrm{g_{m,1}} \cdot \mathrm{C_c}} \tag{4.16}$$

$$\text{Phasenreserve} = 180^\circ - \sum_i \arctan\left(\frac{f_{pole,i}}{\text{Bandbreite}}\right) \tag{4.17}$$

$$\text{Leistung} = \sum_i I_i \cdot V_{dd} \tag{4.18}$$

$$\text{Fläche} = \sum_i W_i \cdot L_i \tag{4.19}$$

gegeben. Für unterschiedliche Verstärkerarten unterscheiden sich die tatsächlich verwendeten Gleichungen für die Verstärkung, Bandbreite und Phasenreserve. In diesem Fall werden die Gleichungen für den Miller-OTA aus den vorherigen Abschnitten verwendet. I_i sind die jeweiligen Zweigströme und W_i und L_i die Weiten und Längen der jeweiligen Transistoren. Leistungsaufnahme und Fläche werden durch das Layout und darin enthaltene notwendige Schritte für analoge Schaltungen, wie das Einfügen von Dummy-Strukturen, von den tatsächlichen Werten abweichen. Da alle Topologien im ähnlichen Maße durch diese Maßnahmen verändert werden, können diese Werte für einen qualitativen Vergleich der Topologien verwendet werden. Abbildung 4.25 zeigt die berechneten und simulierten Werte für die FoM und die Verstärkung des Miller-OTA bei gleichzeitiger Betrachtung der Prozessvariationen und Alterung über 8 Jahre. In den einzelnen Graphen sind zudem der Mittelwert und die 1σ-Grenze eingezeichnet. Bei der berechneten FoM werden die Prozessvariationen leicht überschätzt. Dies ist bedingt durch die Vernachlässigung von Korrelationen auf Block- bzw. Schaltungsebene. Der geschätzte Mittelwert liegt innerhalb der 1σ-Grenzen des tatsächlichen Wertes, daher sind die Abweichungen innerhalb einer statistischen Auswertung gering. Wichtiger ist, dass der komplexe Verlauf der FoM mit dem frühen Anstieg und dem leicht konvexen Verlauf über der Zeit durch die Berechnung nachgebildet werden kann. Eine reine Extrapolation der Werte für den wirklichen Schlimmstfall der Schaltung zu ihrer Mitte des Nutzungszeitraums kann nicht korrekt erfasst werden. Selbst eine Berücksichtigung des Verhaltens bis 100.000 Sekunden würde zu einer vollständigen Fehleinschätzung führen. Zudem kann nicht allein die EOL-Verteilung dazu verwendet werden, um das

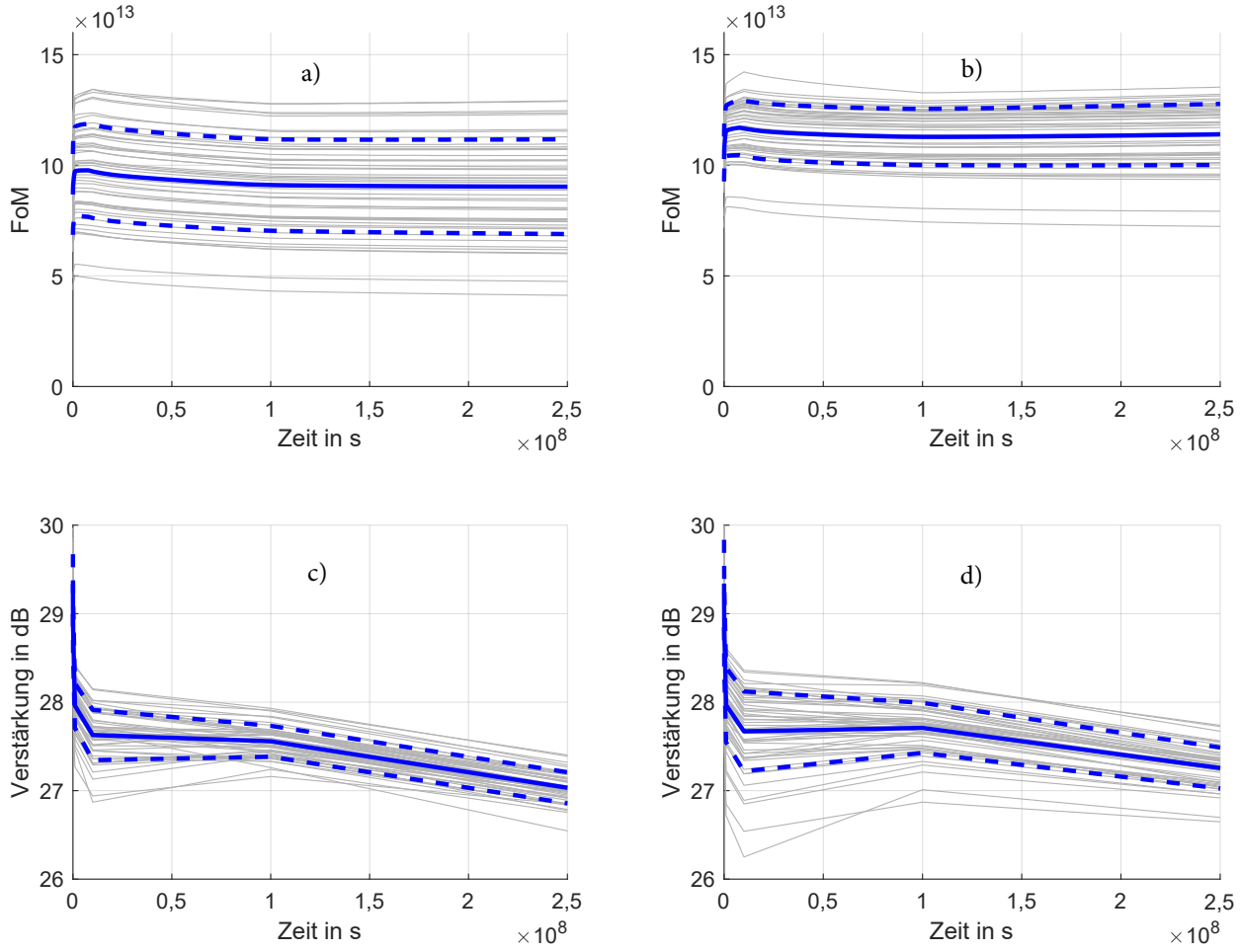


Abbildung 4.25: Vergleich der simulierten (b&d) und berechneten (a&c) FoM und der Verstärkung bei gleichzeitiger Betrachtung von Prozessvariationen und Alterung.

Maximum der Alterungseinflüsse auf Schaltungsebene zu charakterisieren. Obwohl alle Transistoren die gesamte Zeit über den Stressbedingungen ausgesetzt sind, schwächen sich die Auswirkungen ab. Der Anstieg der FoM direkt zu Beginn ist durch den Abfall der Leistungsaufnahme der Schaltung bedingt und durch die Alterung der Stromspiegel und damit fallende Zweigströme zu erklären. Zu diesem Zeitpunkt sind alle Fehlstellen der Transistoren mit geringeren Capture-Zeiten bereits besetzt und gleichzeitig nehmen die Stressbedingungen für HCI durch die geringeren Ströme ab. Somit verlangsamt sich die Alterung ab diesem Zeitpunkt selbst.

Die Verstärkung wird deutlich präziser als die FoM angenähert, dies liegt an der Tatsache, dass weniger Einflussfaktoren auf diese einwirken. Die Alterung bewirkt zu Beginn einen starken Abfall der Verstärkung. Anschließend steigt die Verstärkung, bis nach 3 Jahren erneut ein Abfall eintritt. So wie im Fall der FoM kann dieses Verhalten sehr genau durch die Berechnungen nachgebildet werden. Die zeitabhängige Variabilität und die sich dadurch verändernde Standardabweichung werden gut approximiert. Ein

weiterer wichtiger Aspekt ist die Betrachtung der Temperatur, da diese sowohl auf die Schaltungsfunktion als auch auf die Alterungseinflüsse starken Einfluss hat. Dafür zeigt Abbildung 4.26 die berechnete und simulierte Verstärkung für verschiedene Temperaturen $T = \{-40, 0, 27, 120\}$°C. 200 Samples wurden simuliert bzw. berechnet, welche von Prozessvariation und Alterung betroffen sind. In der oberen Reihe sind die berechneten

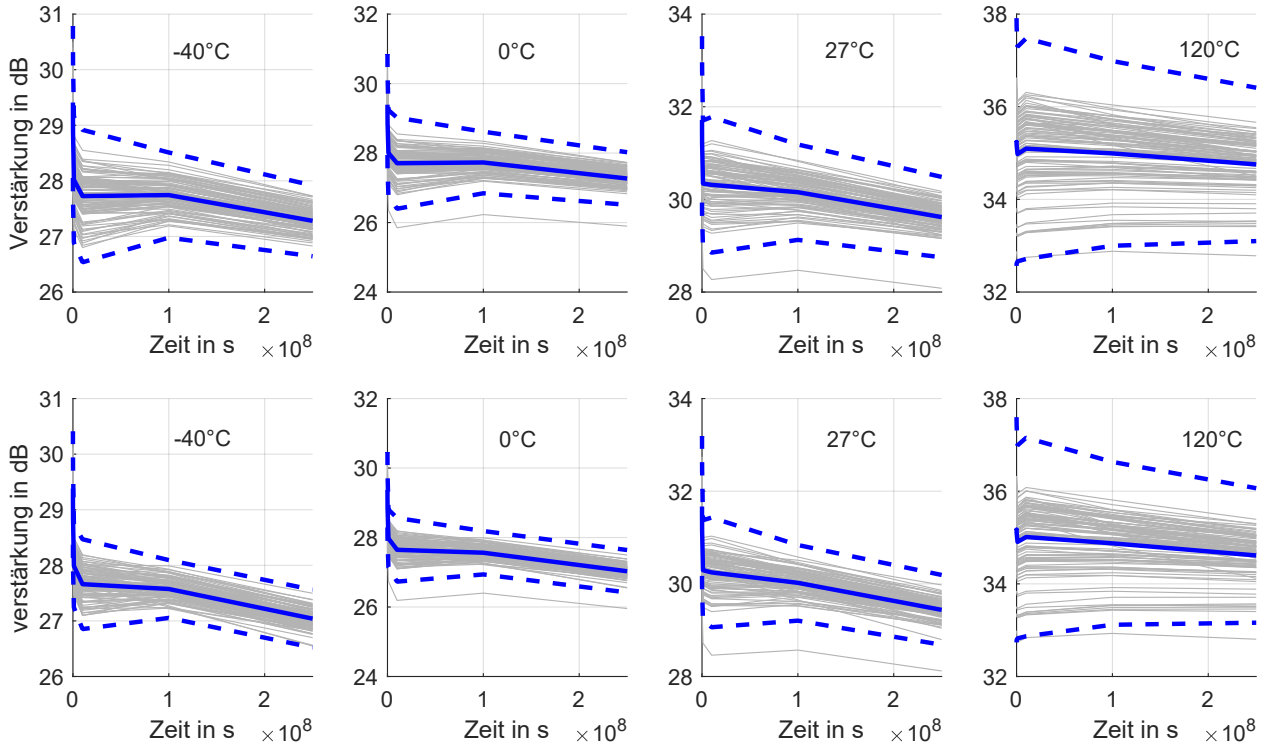


Abbildung 4.26: Simulierte (untere Reihe) und berechnete (obere Reihe) Verstärkung für $T = \{-40, 0, 27, 120\}$°C mit Prozessvariationen und Alterung über 8 Jahre.

Werte (Funktionsblöcke) und in der unteren die simulierten Werte (SPICE Simulation). Die Standardabweichung ist, wie gezeigt, stark von der Temperatur abhängig und für hohe Temperaturen am höchsten. Wie durch die vorherigen Untersuchungen des Miller-OTA zu erwarten ist, zeigt sich, dass die Alterung sich am stärksten für sehr geringe Temperaturen auswirkt. Sowohl die Änderung des Mittelwerts als auch die der Standardabweichung über der Zeit und den Temperaturen werden durch das vorgestellte, blockbasierte Vorgehen sehr gut nachgebildet.

Für die Auswertung der zeitabhängigen Ausbeute der Schaltung werden eine obere und untere Schranke für die FoM verwendet. Das Ergebnis ist in Abbildung 4.27 für verschiedene Temperaturen dargestellt. Aus den Verläufen der Graphen ist zu erkennen, dass die EOL-Verteilung nicht die Ausbeute der Schaltung definiert. Da für alle Anwendungsszenarien der Worst-Case, stärkste Absenkung der betrachteten Größe, nach ungefähr 3 Jahren eintritt. Somit würden Extrapolationen zu einer Überschätzung

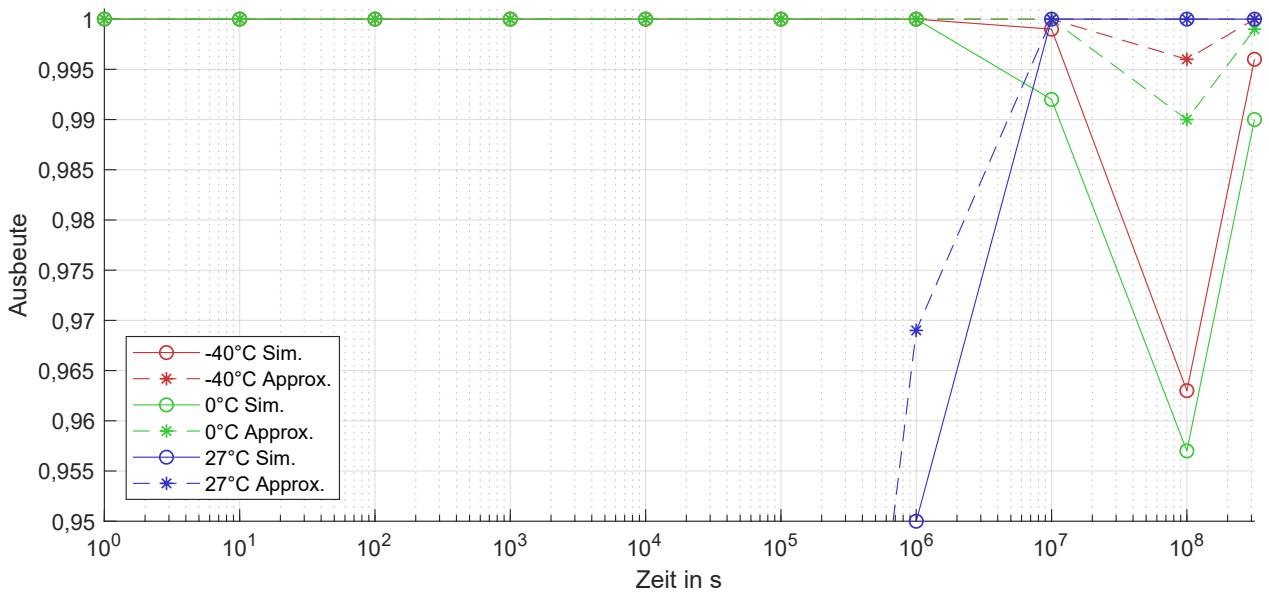


Abbildung 4.27: Zeitabhängige Ausbeute des Miller-OTAs für verschiedene Temperaturen über 8 Jahre.

der Ausbeute führen. Die Abweichungen zwischen Vorhersage und Simulation liegen mit qualitativ korrekten Zeitabhängigkeiten bei unter 4%.

4.3 Gefertigter Verstärker und Referenzquelle

Auf Grundlage der in Abschnitt 4.1.2 durchgeführten Analyse von verschiedenen Biasschaltungen und Verstärkertopologien wurden die beiden gewählt, die die besten Abwägungen zwischen Fläche, Leistungsaufnahme, nominellen Eigenschaften sowie der Resistenz gegenüber PVTA Einflüssen darstellen. Für diese beiden Schaltungen wurde ein Layout entwickelt und in einer 65nm-CMOS-Technologie gefertigt.

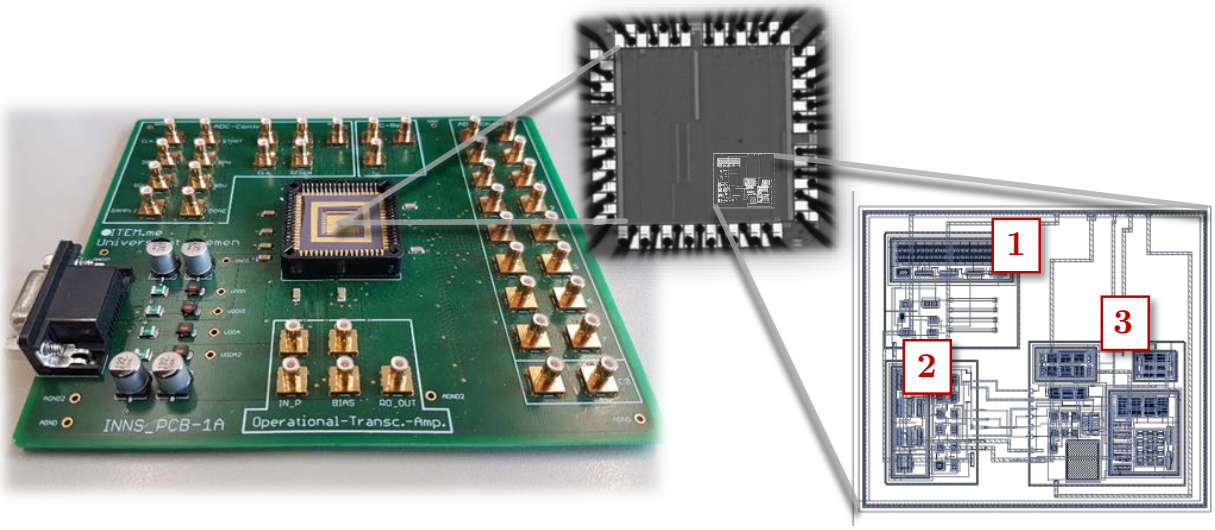


Abbildung 4.28: Testplatine, Chipfoto und Layout der gefertigten Schaltung bestehend aus Beta-Multiplier, Bias-Netzwerk und Verstärker.

Die entworfene Schaltung selbst besteht aus drei Teilen, welche in Abbildung 4.28 nummeriert sind:

1. Beta-Multiplier als Referenzspannungsquelle im Langkanal-Design wie in Abschnitt 4.1.2.2 beschrieben (vgl. Abbildung 4.29). Die Schaltung selbst belegt eine Fläche von 288 μm^2 und nimmt eine Leistung von 38,63 μW auf.

2. Das Bias-Netzwerk stellt die verschiedenen Spannungen für die Stromspiegel und Kaskodetransistoren des Verstärkers bereit (Schaltplan wird nicht gezeigt). Die Schaltung selbst belegt eine Fläche von 210 μm^2 und nimmt eine Leistung von 39,98 μW auf.

3. Verstärkerschaltung deren nominelle Resistenz gegenüber Alterung auf Grundlage der Blockanalyse am höchsten ist. Der Schaltplan ist in Abbildung 4.22 auf Seite 128 gezeigt. Die Schaltung selbst belegt eine Fläche von 409 μm^2 und nimmt eine Leistung von 59,7 μW auf.

Für die weitere Arbeit wurde die Referenzquelle als Messobjekt für die entwickelte Messmethode verwendet. Daher wird diese nochmals genauer mit den jeweiligen Bau-

teilgrößen beschrieben. Der Schaltplan ist in Abbildung 4.29 gezeigt und Tabelle 4.2 enthält alle Bauteildimensionen der Schaltung. Der Widerstandswert von R beträgt insgesamt 5,504 kΩ.

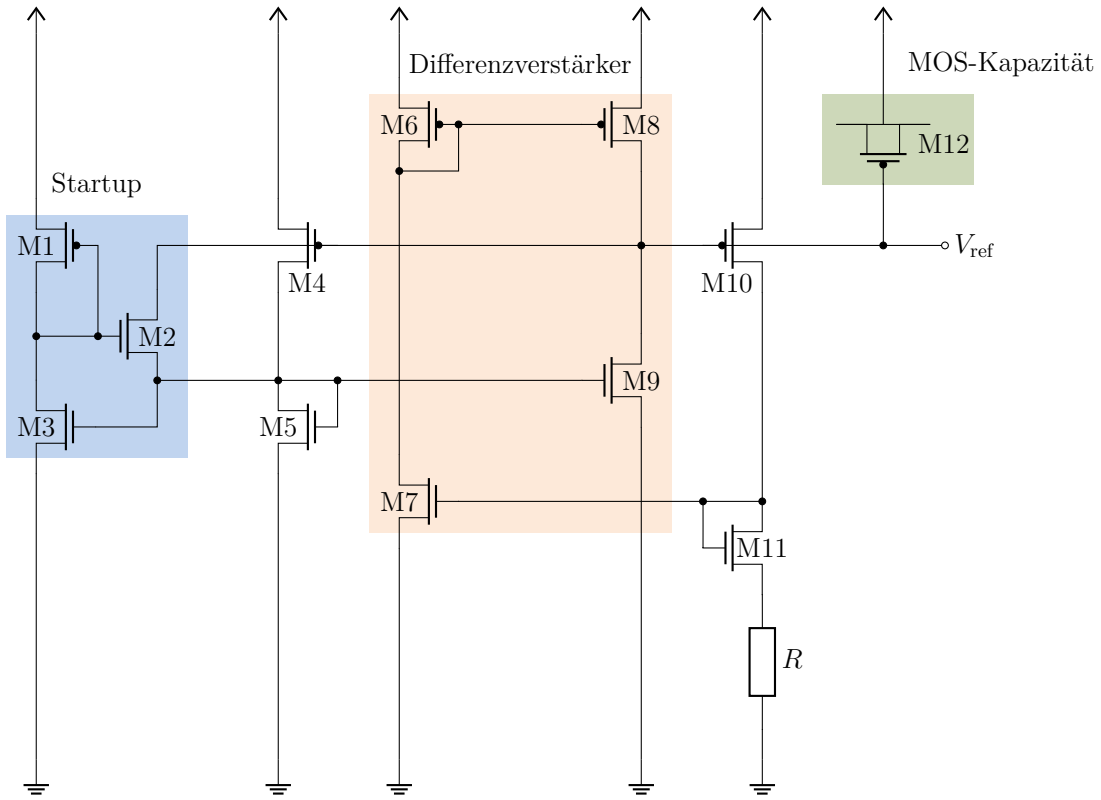


Abbildung 4.29: Schaltplan der gefertigten Spannungsreferenz mit MOS-Kapazität, Startup-Schaltung und Differenzverstärker.

Tabelle 4.2: Bauteilgrößen der gefertigten Referenzquelle

	M1	M2	M3	M4	M5	M6
W in μm	0,6	0,6	3	6	3	6
L in μm	1,2	0,06	0,12	0,12	0,12	0,12

	M7	M8	M9	M10	M11	M12	R
W in μm	3	6	3	6	12	6	0,06
L in μm	0,12	0,12	0,12	0,12	0,12	6	23,58

4.4 Zusammenfassung

In diesem Kapitel wurden die Möglichkeiten evaluiert, die die Verwendung des REL-MOS-Transistormodells beim Entwurf von integrierten analogen Schaltungen bietet.

Aufbauend auf diesen Erkenntnissen wurde eine funktionsblock-basierte Entwurfsmethode vorgestellt, bei der die Abhängigkeiten der Funktionsblöcke direkt auf Schaltungsebene übertragbar sind. Dies hat zum Vorteil, dass Topologien schnell anhand der Ergebnisse auf Blockebene analysiert oder entworfen werden können. Es wurden zwei Fallstudien für Biasschaltungen und Verstärker mit 7 und 8 Schaltungstopologien durchgeführt, um die Anwendbarkeit der Methode zu demonstrieren. Die jeweils widerstandsfähigste Schaltung gegenüber PVTA-Einflüsse wurde anschließend in einer 65nm-CMOS-Technologie gefertigt.

5 Messplatz und Extraktionsablauf

» One accurate measurement is worth a thousand expert opinions «

In diesem Abschnitt wird der in der Arbeit entwickelte Messaufbau sowie der Extraktionsablauf für Bauteile und Schaltungsparameter vorgestellt. Ein generelles Problem bei der Messung von integrierten Schaltungen ist, dass Informationen, die während des Entwurfs der Schaltung erhoben werden, nicht innerhalb der Messungen verwendet werden können. Für den Entwurf der meisten integrierten analogen Schaltungen ist die Betrachtung des Temperaturverhaltens sowie der Prozessvariationen und ggf. Alterung innerhalb der Simulationen notwendig. Für die Messungen bzw. messtechnische Verifikation stehen diese Informationen im Anschluss nicht mehr zur Verfügung. Treten während der Messung Abweichungen auf, kann kaum Rückschluss auf die Ursache gezogen werden. Wichtig wäre, dass eine Methode Informationen liefern könnte, ob das gemessene Verhalten durch Prozessvariationen oder das Temperaturverhalten erklärbar ist oder ob während der Fertigung ein Anpassungsfehler entstanden ist. Des Weiteren muss eine solche Methode in der Lage sein, das Rauschen der Bauteile und des Messaufbaus als solches berücksichtigen zu können. Andernfalls könnten große Messabweichungen auftreten.

Zudem sollte die Methode in der Lage, mit sehr wenigen Messungen zu arbeiten. Für die Analyse von sehr kleinen Stichproben werden zum Beispiel durch Bootstrapping [Efr92] oder Jackkniffing [ES81] die vorhandenen Messpunkte mehrfach verwendet, um die stochastischen Momente genauer schätzen zu können [ST12]. Durch die Partikelfilter müssen diese zusätzlichen Messungen jedoch nicht rein mathematisch erzeugt werden. Durch die Kenntnis der Prozessvariationen und dem vorhandenen Zustandsraumodell können kleine Stichproben durch weitere Messungen erweitert werden.

Die in dieser Arbeit entwickelte Methode ist in der Lage, all diese Randbedingungen zu erfüllen. Die Idee ist, dass ein Zustandsraummodell der Schaltung oder des Bauelements mit den Daten aus den Simulationen parametriert wird. Dieses Modell ist somit in der Lage, das Verhalten des Bauteils, welches aus den Simulationsdaten folgt, nachzubilden (vgl. Abb. 5.1).

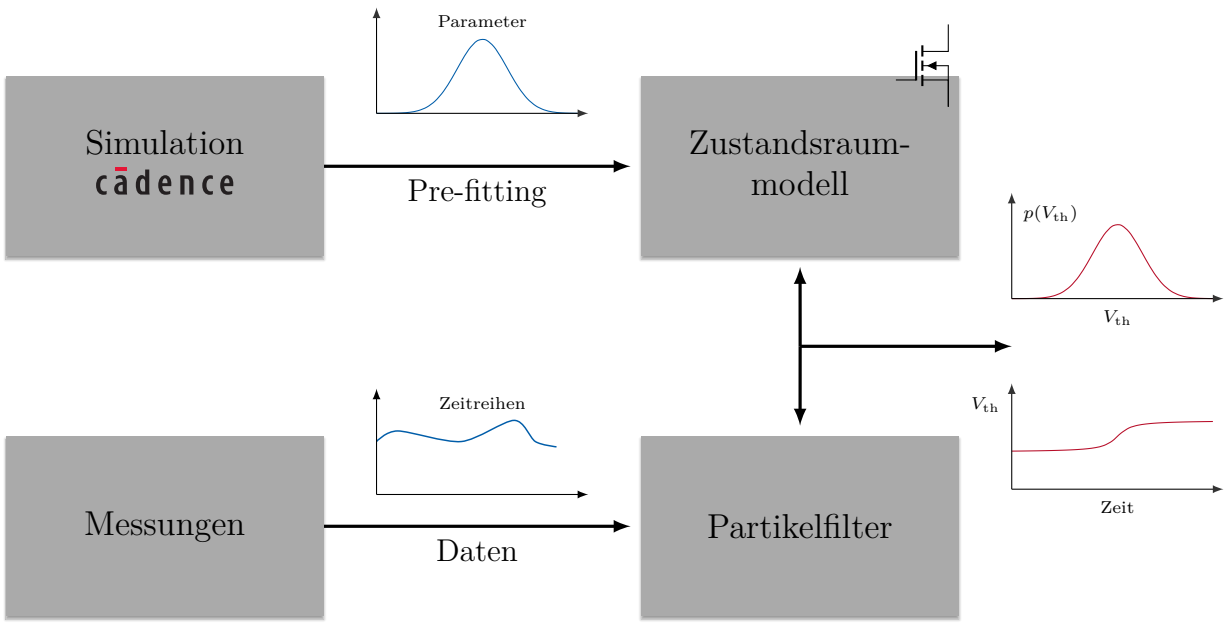


Abbildung 5.1: Ablauf der Gesamtmessmethode.

Dabei können auch einzelne Effekte modelliert oder ausgelassen werden, falls diese für die spätere Messung irrelevant sind. Dieses Zustandsraummodell enthält zudem die Informationen der Prozessvariation aus dem jeweiligen PDK. Ein Partikelfilter-Algorithmus wird im Anschluss verwendet, um zusammen mit den Informationen der Umgebungseinflüsse (Versorgungsspannung, Temperatur) das Zustandsraummodell anhand der aufgenommen Messwerte zu parametrieren. Durch dieses Vorgehen kann die Parametrierung zeitgleich zur Messung durchgeführt werden. Des Weiteren ist sofort erkennbar, ob das gemessene Verhalten zu den aus der Simulation stammenden Werten passt.

Unbekanntes Verhalten kann ebenfalls während der Messung direkt für die Modellierung verwendet werden. Dazu können im Algorithmus einzelne Parameter des Zustandsraummodells als Variablen definiert werden. Dadurch wird das Verhalten des Modells anhand

dieser Variablen parametrisiert. Eine Übersicht über den Ablauf der Methode ist in Abbildung 5.2 gezeigt.

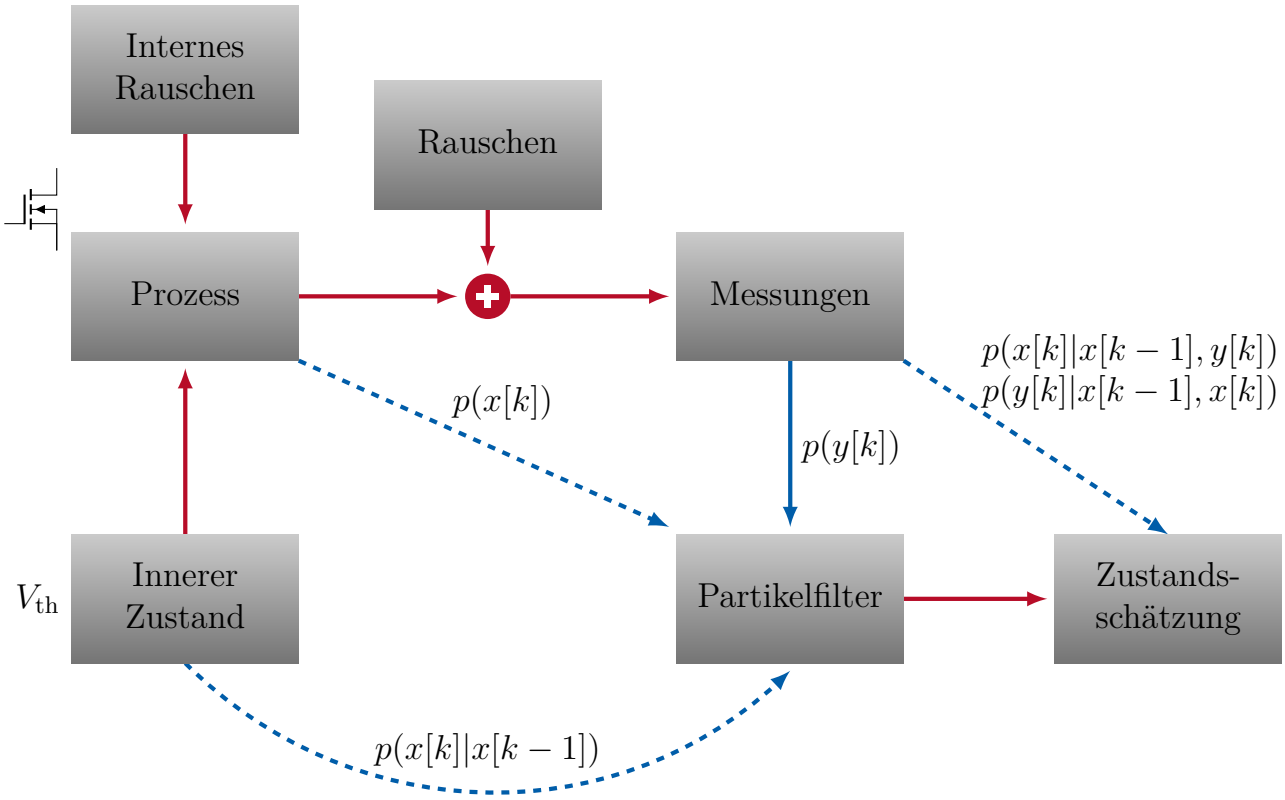


Abbildung 5.2: Ablauf der Zustandsschätzung mit dem Partikelfilter.

Darin ist der Prozess das Zustandsraummodell des Bauteils oder der Schaltung. Die internen Zustände sind all diejenigen, die nicht messbar oder beobachtbar sind. Für einen Transistor ist der Drainstrom eine messbare Zustandsgröße, die Schwellspannung hingegen nicht. Die interne Rauschquelle kann dabei einer beliebigen Verteilung folgen. Gleiches gilt für das Messrauschen. Der Berechnungsablauf des Partikelfilters ist in Algorithmus 3 gezeigt. Die Berechnung der Gewichte geschieht dabei durch die numerische Invertierung des Modells, um auf die internen Zustände schließen zu können. Die gesamte Berechnung und Auswertung findet in MATLAB statt. Die Anzahl der verwendeten Partikel N_s wurde, wenn nicht anders angegeben, zu 200 gewählt. Durch das Resampling wird das zuvor beschriebene Problem der Verarmung der Gewichte umgangen. Der bereits beschriebene Vorteil des Partikelfilters ist, dass die inneren Zustände eine beliebige Verteilungsdichtefunktion besitzen können. Dies ist besonders für die Vermessung von Schaltungen wichtig. Ein weiterer Vorteil dieses Ansatzes ist, dass bereits während der Langzeitmessungen, die für die Bestimmung der Alterung notwendig sind, das parametrierte Modell zur weiteren Vorhersage der Alterung bzw. der EOL-Verteilung und somit zur Prognose der Zuverlässigkeit der Schaltung verwendet werden kann.

Algorithmus 3 Implementierter Partikelfilter Algorithmus

Eingabe: $\mathbf{x}_0, p(\mathbf{x}_0), \mathbf{z}_k$
Ausgabe: $\{\mathbf{x}_k^i, \mathbf{w}_k^i\}_{i=1}^{N_s}$
1: **for** $i = 1 : N_s$ **do**
2: Ziehe $x_k^i \sim p(\mathbf{x}_k|\mathbf{x}_{k-1}^i)$
3: Berechne $w_k^i = p(\mathbf{y}_k|\mathbf{x}_k^i)$
4: **end for**
5: Berechne Gesamtgewichte: SUM $= \sum_{i=1}^{N_s} w_k^i$
6: **for** $i = 1 : N_s$ **do**
7: Gewichte normalisieren: $w_k^i = w_k^i/\text{SUM}$
8: **end for**
9: Resample

Im weiteren Verlauf dieses Kapitels werden die verwendeten Messgeräte und der konzipierte Messplatz vorgestellt. Anschließend wird die Modellparametrierung mittels Partikelfilter an Transistoren und Schaltungen gezeigt. Abschließend werden die Vorhersagemöglichkeiten in Kombination mit beschleunigter Alterung beschrieben.

5.1 Modellparametrierung mittels Partikelfilter

Der Inhalt dieses Abschnitts gliedert sich in drei Teile. Zuerst werden die herkömmlichen Messverfahren zur Bestimmung der Schwellspannung beschrieben. Anschließend wird die Messung der Schwellspannung mit der entwickelten Messmethode durchgeführt und mit den anderen Methoden verglichen. Danach werden die gefertigten Schaltungen mit Hilfe der Methode überwacht.

5.1.1 Bestimmung der Schwellspannung

Die Schwellspannung stellt einen der wichtigsten Parameter für Transistoren dar. Sie wird zur Beschreibung der Performanz von Transistoren verwendet, zudem zur Klassifizierung unterschiedlicher Transistortypen innerhalb einer Herstellungstechnologie. Des Weiteren werden mit der Schwellspannung Effekte wie Temperatur und Alterung beschrieben. Neben der Funktion als Modellparameter in veralteten Transistormodellen wird sie sehr oft für den Entwurf von analogen Schaltungen verwendet. Da diese Größe so wichtig ist, wurden die gängigsten Messverfahren der Schwellspannung untersucht. Die betrachteten Methoden verwenden alle die Übertragungskennlinie, Drainstrom I_d über Gatespannung V_g, des Transistors. Die gängigste Definition für die Schwellspannung ist der Punkt der Gatespannung, bei dem der Drainstrom signifikant ansteigt für ein $V_{ds} \neq 0$ [OCSL+02, OCGSM+13]. Ähnliche Definitionen sind die deutliche Abweichung vom exponentiellen Verhalten unterhalb der Schwellspannung bzw. vom linearen oberhalb von V_{th} für eine steigende Gate-Spannung. Abbildung 5.3 zeigt die

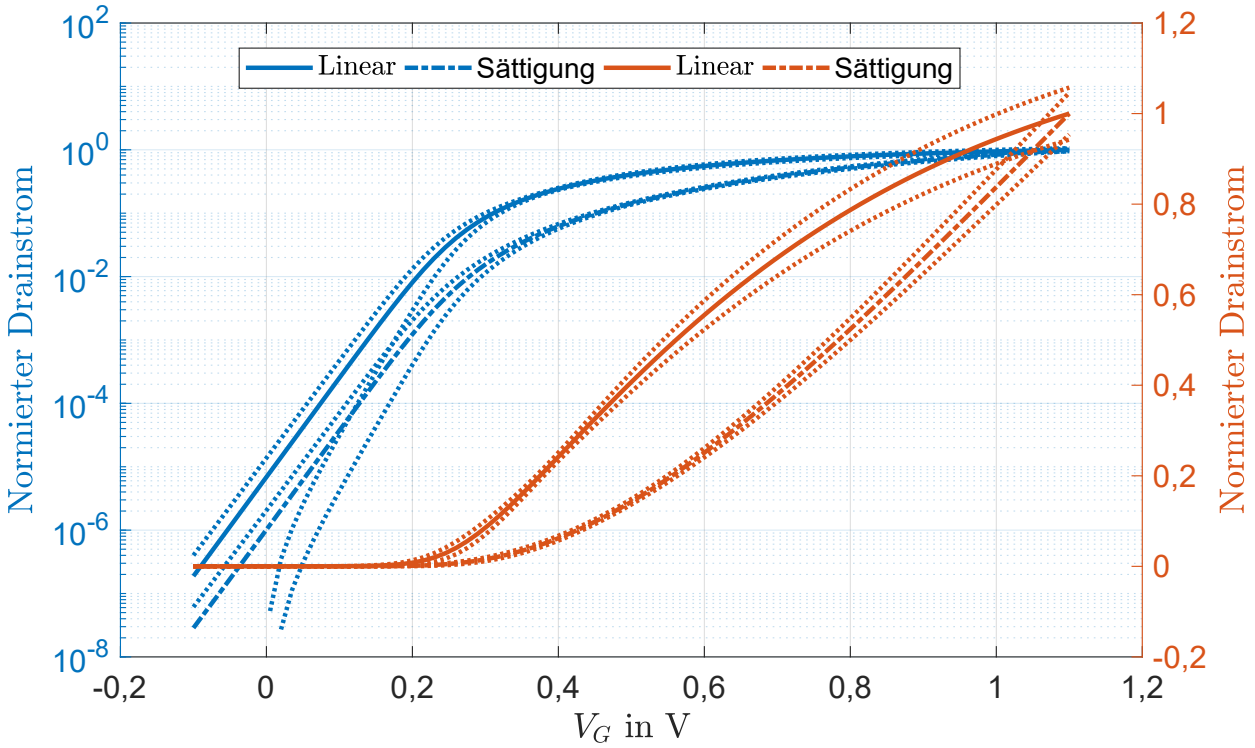


Abbildung 5.3: Übertragungskennlinie von NMOS-Transistoren im linearen Bereich und in Sättigung.

Übertragungskennlinie eines 28nm-FDSOI-NMOS-Transistors im linearen Bereich und in Sättigung. Des weiteren sind die 3σ-Grenzen der Prozessvariationen durch gestrichelte Linien dargestellt. Sowohl auf der linearen als auch logarithmischen Skala ist der Punkt $V_G = V_{th}$ nicht eindeutig abzulesen. Neben den messtechnisch erfassbaren Definitionen ist die Schwellspannung definiert als der Punkt, bei dem $Q_i = 0$ gilt. Dies entspricht eintretender Inversion. Dieser Punkt kann jedoch nur innerhalb von Simulationen (z.B. TCAD) bestimmt werden, da die Inversionsladung Q_i bei einem Transistor nicht gemessen werden kann. Wie in Abbildung 5.3 zu erkennen ist, unterscheiden sich die Schwellspannung in Sättigung und im linearen Bereich deutlich, egal welche Definition gewählt wird. Durch die verschiedenen Definitionen in Kombination mit den Betriebsbereichen des Transistors sind Literaturwerte für die Schwellspannung immer unter Berücksichtigung der jeweiligen Messbedingungen zu betrachten. So kann sich die Schwellspannung desselben Transistors nur durch die Verwendung einer anderen Messmethode oder der Messung in einem anderem Betriebspunkt unterscheiden [Sch06, OCSL+02, OCGSM+13, BBA+11]. Diese Tatsache erschwert insbesondere den Vergleich der Auswirkungen von Alterungseinflüssen durch und auf die Schwellspannung, da eine absolute Abweichung von $\Delta V_{th} = 10\text{mV}$ ihre Aussage erst erhält, wenn Betriebspunkt und Messmethode der Schwellspannung mitangegeben werden.

Die zwölf betrachteten Messmethoden sind in Tabelle 5.1 aufgelistet. Die Current-to-square-root-of-the-Transconductance Methode wird in dieser Arbeit in zwei Variationen

Tabelle 5.1: Übersicht der betrachteten Messmethoden

Nummer	Abkürzung	Name
1	MaxGM	Maximum g_m [OCSL+02]
2	LE	Linear Extraction [Sch06]
3	ACC	Adjusted Constant Current [BBA+11]
4	SD	Second Derivative [WWKB87]
5	TD	Third Derivative [WMYD01]
6	CsrTR	Current-to-square-root-of-the-Transconductance [Ghi88]
7	CsrTRF	Current-to-square-root-of-the-Transconductance (modifiziertes Fitting)
8	TransM	Transition Method [SOCDM+00]
9	NMID	Normalized Mutual Integral Difference [HXC+02]
10	NRH	Normalized Reciprocal H function [OCCE+01]
11	TCR	Transconductance-to-Current-Ratio [SSGM12]
12	RH	Reciprocal H function [OCGSLH10]

analysiert. Dabei besitzt die Methode mit selbsterstellten modifiziertem Fitting eine MATLAB-basierte Spline-Interpolation, da die originale Funktion sehr anfällig gegenüber Rauscheinflüssen ist. Die verschiedenen Messtechniken lassen sich in drei Kategorien einteilen. Solche, die Ableitungen der Übertragungscharateristik verwenden, solche, die Integrationen benutzen, und erweiterte Methoden die Modellgleichungen des Transistors verwenden, um die Messungen zu verbessern.

Ableitungsbasierte Messtechniken Diese Methoden verwenden die erste, zweite oder dritte Ableitung des Drainstroms und definieren die Schwellspannung an den Nullstellen oder den Maxima dieser Ableitungen. Anders ausgedrückt definieren die Methoden die Schwellspannung am Punkt der höchsten oder geringsten Sensitivität des Drainstroms gegenüber der Gatespannung. Die Werte der Schwellspannung unterscheiden sich dadurch sehr stark. Für die erste Ableitung des Drainstroms gilt

$$g_m = \frac{\partial I_d}{\partial V_g} \ , \tag{5.1}$$

sowie

$$\frac{g_m}{I_d} = \frac{\partial \log(I_d)}{\partial V_g} = \frac{1}{I_d}\frac{\partial I_d}{\partial V_g} \ . \tag{5.2}$$

Daher nehmen einige Methoden auch Bezug auf die Transkonduktanz. Die Berechnung dieser Methoden ist unter idealen Bedingungen relativ simpel. Ein großer Nachteil dieser

Methoden ist jedoch, dass diese die Auswirkungen von Rauschen verstärken. So sind die Auswertungen der zweiten und dritten Ableitung unter reellen Bedingungen kaum durchzuführen [OCGSM+13]. Zu diesen Methoden zählen MaxGM (1), LE (2), SD (4), TD (5), CsrTR (6), CsrTRF (7), TCR (11) aus Tabelle 5.1.

Integrationsbasierte Messtechniken Diese Techniken summieren den Drainstrom und normieren diesen auf spezifische Spannungen oder Ströme. An der Stelle des Maximums der entstehenden Kurve wird die Schwellspannung definiert.

$$V_{th} = V_g \left(\max \left\{ \frac{1}{k_{norm}(V_g, I_d)} \int_0^{V_g} I_d(V_g)\, \mathrm{d}V_g \right\} \right) \tag{5.3}$$

An diesem Punkt ist der Drainstrom über die einzelne signifikant größer gegenüber vorherigen Werten. Die Nachteile dieser Methoden sind parasitäre Spannungsabfälle über den Drain- und Source-Kontaktwiderständen. Des Weiteren liefern sie unzuverlässige Ergebnisse für sehr geringe Verhältnisse zwischen Sperrstrom und Sättigungsstrom der Transistoren in sehr modernen Technologien [OCGSM+13]. Zu diesen Methoden zählen TransM (8), NMID (9), NRH (10) and RH (12) aus Tabelle 5.1.

Erweiterte Methoden Diese Methode (ACC (3), Tab. 5.1) verwendet einzelne Gleichungen von Transistormodellen wie dem EKV-Modell, um einzelne Effekte zu unterdrücken oder das Messergebnis zu verbessern. Die Schwellspannung wird dabei über einen temperaturabhängigen Strom

$$I_{th} = \frac{\alpha_{th}\left(\frac{V_{DS}}{U_T}\right)}{\alpha_{th,sat}} \cdot I_{th,sat} \tag{5.4}$$

definiert, dabei ist α_{th} ein Koeffizient

$$\alpha_{th} = \frac{I_{th}}{I_{spec}} \tag{5.5}$$

wobei $I_{spec} = 2nU_T^2 \mu C_{ox} {}^{W}/_{L}$ der spezifische Technologiestrom aus dem EKV-Modell ist und

$$I_{th,sat} = 0,608 \cdot I_{spec} \tag{5.6}$$

gilt. Der Faktor 0,608 folgt aus der Definition $Q_i = 0$ wobei im EKV gilt $q_i^2 + \ln(q_i) = 0$ und für den Drainstrom folgt $[(q_i^2 + \ln(q_i))^2 + q_i^2 + \ln(q_i)] \cdot I_{spec} = I_d$ gilt. Somit kann die Temperaturabhängigkeit der Schwellspannung aus einzelnen Messungen unterdrückt werden.

Zur Auswertung der Methoden werden TCAD-Simulationen von jeweils 10.000, von Prozessvariationen betroffenen NMOS- und PMOS-FDSOI Transistoren, Bulk-Transistoren sowie FinFET mit verschiedenen Längen $L = 10\text{nm} \ldots 600\text{nm}$ verwendet. Als Maß für die Güte der Messtechniken wird zunächst die Variabilität der einzelnen Methoden

$$V_{th,i,var} = \frac{\sigma_{V_{th,i}}}{\mu_{V_{th,i}}} \tag{5.7}$$

herangezogen. Dabei ist σ die Standardabweichung und μ der Mittelwert der Schwellspannung über die jeweils 10.000 Transistoren. Als weiteres Bewertungskriterium wird der Wiedererkennungsfehler

$$V_{th,i,recog} = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{j=1}^{M} \left| \frac{V_{th,i,0} - V_{th,i,j}}{V_{th,i,0}} \right| \tag{5.8}$$

benutzt. $V_{th,i,0}$ ist dabei der initiale Wert der Schwellspannung des $i$ten Transistors. Diese gibt an, inwieweit eine Methode in der Lage ist, N verschiedene Transistoren in M wiederholten Messungen während Rauschen die Messgrößen verändert, zu bestimmen. Diese Fähigkeit ist besonders wichtig für Langzeitmessungen, um Alterungseinflüsse charakterisieren zu können.

5.1.1.1 FDSOI FET

In Abbildung 5.4 ist der Vergleich von 28nm-FDSOI-PMOS- und NMOS-Transistoren simulierter Wiedererkennungsfehler gezeigt. Es werden 500 Bauteile mit 15 wiederholten Messungen je Methode untersucht. Jede Messung unterliegt einem Messrauschen von $\sigma_{noise} = 200\text{pA}$. Wie zu erkennen ist, ist der Fehler für die MaxGM und TD Methoden am höchsten. Die nächst niedrigen Fehler weisen die SD und LE Methode auf. All diese vier Methoden sind ableitungsbasiert und daher sehr anfällig gegenüber Messrauschen. Dies ist insbesondere für Messungen bei höheren Temperaturen problematisch, da dort das thermische Rauschen ansteigt. Die weiteren ableitungsbasierten Methoden sind durch ihre weiteren Rechenschritte, zum Beispiel durch die Berechnung der Wurzel, in der Lage, trotz des Rauschens annehmbare geringe Fehler bei Wiedererkennung zu gewährleisten. Die Abweichungen für PMOS-Transistoren liegen insgesamt oberhalb derer der NMOS Transistoren. Ursächlich dafür ist die geringere Ladungsträgerbeweglichkeit und somit ist der Drainstrom bei sonst gleichen Bedingungen niedriger als bei NMOS-Transistoren. Dies führt zu einer stärkeren Auswirkung des Rauschens. Die geringsten Fehler werden durch die erweiterte Adjusted Constant Current Methode erreicht. In Abbildung 5.5 sind die Wiedererkennungsfehler der einzelnen Methoden für NMOS-FDSOI-Transistoren mit $L = 600\text{nm}$ für unterschiedliche Rauschleistungen gezeigt. Hier zeigt sich, dass die ACC und CsrTR Methoden gut mit der Rauschleistung skalieren. Die ableitungsbasierten Messmethoden weisen dabei die größten Abweichungen auf. Dies schließt eine verlässliche Messung der Schwellspannung voll-

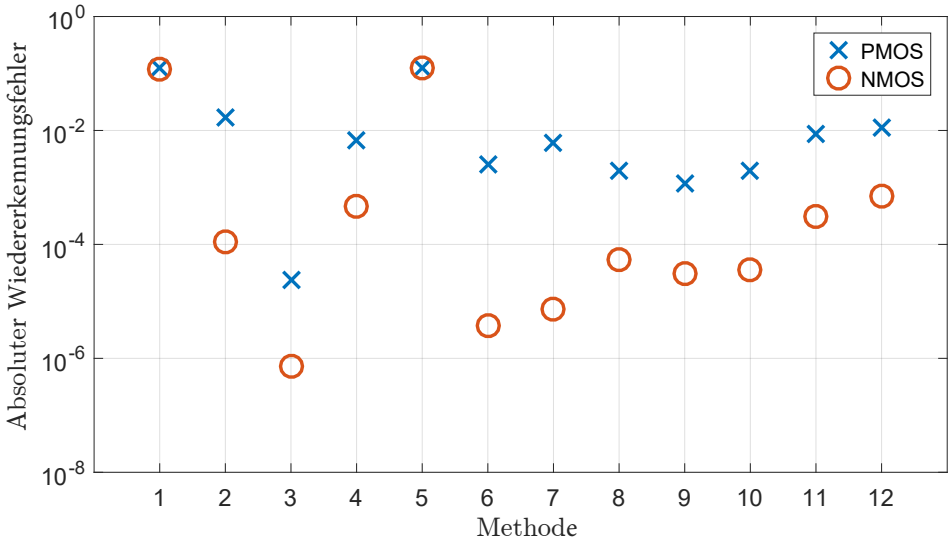


Abbildung 5.4: Simulierter Wiedererkennungsfehler von 28nm-FDSOI-PMOS und NMOS-Transistoren bei einem Rauschen von σ_{noise} = 200pA. 500 Bauteile mit 15 wiederholten Messungen je Methode.

ständig aus. Die integralbasierten Methoden skalieren ebenfalls, zeigen jedoch für kleine Rauschleistungen insgesamt einen größeren Fehler auf. Dies liegt an der Anfälligkeit gegenüber parasitären Spannungsabfällen, die bei solch geringen Transistorgrößen starke Auswirkungen haben. Ein Überblick über die Messleistungen der einzelnen Methoden zeigt Abbildung 5.6. Dargestellt sind die Mittelwerte und die Variabilität μ/σ der zwölf verschiedenen Methoden für verschiedene Kanallängen der FDSOI-Transistoren (vgl. Tabelle 5.1). Die Bestimmung von V_{th} wurde dabei sowohl in Sättigung als auch im linearen Bereich durchgeführt. An den Mittelwerten lässt sich erkennen, wie weit sich die einzelnen Methoden in dem bestimmten Wert der Schwellspannung unterscheiden. Dabei sind die Unterschiede in Sättigung deutlicher ausgeprägt. Zudem zeigen sich bei den Methoden unterschiedliche Abhängigkeiten von der Kanallänge. Die integralbasieren Methoden zeigen eine fallende Tendenz über höhere Längen und im Gegensatz dazu steigt die Schätzung der ableitungsbasierten Methoden an. In Sättigung ist die größte Abweichung der Schwellspannung über 300mV für ansonsten gleiche Transistoren. Im linearen Bereich sind die Schätzungen der Schwellspannungen deutlich einheitlicher, dafür insgesamt höher. Es zeigt sich nahezu keine Abhängigkeit der Schwellspannung von der Kanallänge. Die Variabilität zeigt, dass die Messungen in Sättigung deutlich stärker streuen als im linearen Bereich. Dies ist besonders für die Bestimmung der Schwellspannug im Rahmen von Langzeitmessungen zur Alterungsbestimmung interessant. Diese Messungen sollten, wenn möglich, im linearen Bereich durchgeführt werden. Somit entfallen weiterer Stress und Sättigungseffekte während der Messung. Wie bei den vorherigen Auswertungen zeigt die ACCMethode über alle Betriebsbereiche die besten Ergebnisse. Neben den Abweichungen unterscheiden sich die Methoden im notwendigen Messaufwand. Die integralbasierten Methoden benötigen meist die voll-

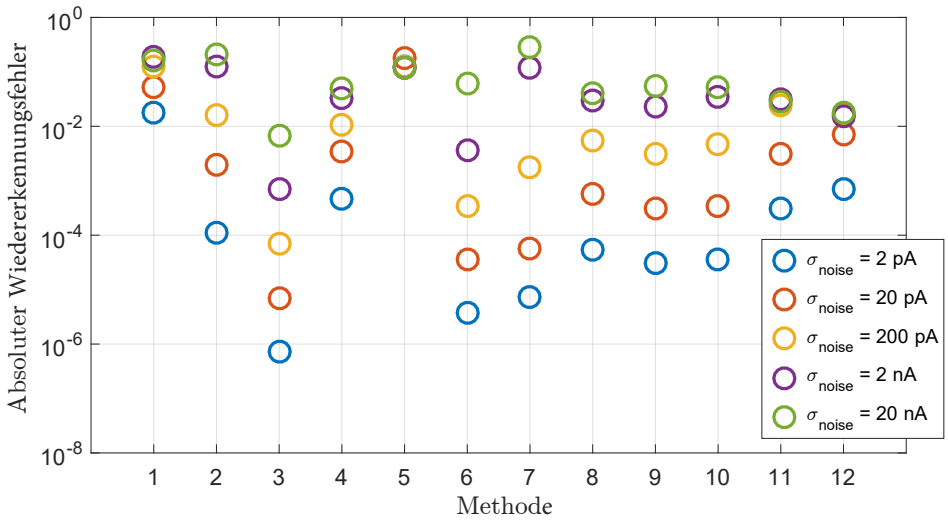


Abbildung 5.5: Wiedererkennung von NMOS-Transistoren mit $L = 600$nm für verschiedenes Rauschen σ_{noise}. 500 Transistoren mit 15 wiederholten Messungen je Methode.

ständige Messung der Übertragungscharakteristik. Dabei bestimmt die Schrittweite der Messungen meist die Auflösung der gemessenen Schwellspannung. Durch Interpolation können diese Ergebnisse verbessert werden, sie können ihrerseits jedoch ebenfalls zu weiteren Abweichungen führen. Die ableitungsbasierten Methoden benötigen weniger Messpunkte. Die lineare Extrapolation benötigt zum Beispiel nur zwei Messungen. Bei einer so geringen Anzahl sinkt gleichzeitig die Genauigkeit.

5.1.1.2 Bulk-MOSFET

Bei den Messwerten für Bulk-MOSFET-Bauelemente zeigt sich, dass die integralbasierten Methoden für Fertigungstechnologien mit größeren Technologieknoten, bei denen die parasitären Spannungsabfälle über den Drain- und Source-Gebieten weniger ins Gewicht fallen, deutlich bessere Ergebnisse liefern (vgl. Abb. 5.7). Insgesamt reduzieren sich die Messabweichungen durch die größeren Ströme in dieser Technologie. Bei der Wiedererkennung zeigt sich, dass die integralbasierten Methoden in diesen großen Fertigungstechnologien wirkliche Vorteile gegenüber den anderen Methoden besitzen. Daher sollten vor jeder Messung in einer bestimmten Technologie die unterschiedlichen Methoden evaluiert werden, da nicht nur Kanallänge und Betriebsbereich, sondern auch die Technologie als solche starken Einfluss auf die Leistungsfähigkeit der Methoden nehmen kann.

5.1.1.3 FinFET

Die untersuchten FinFET sind die kleinsten untersuchten Transistoren mit einer kleinstmöglichen Kanallänge von 20nm. Wie in Abbildung 5.8 zu erkennen ist, steigen die Abweichungen für alle Methoden an. Die integralbasierten Messungen zeigen im Ver-

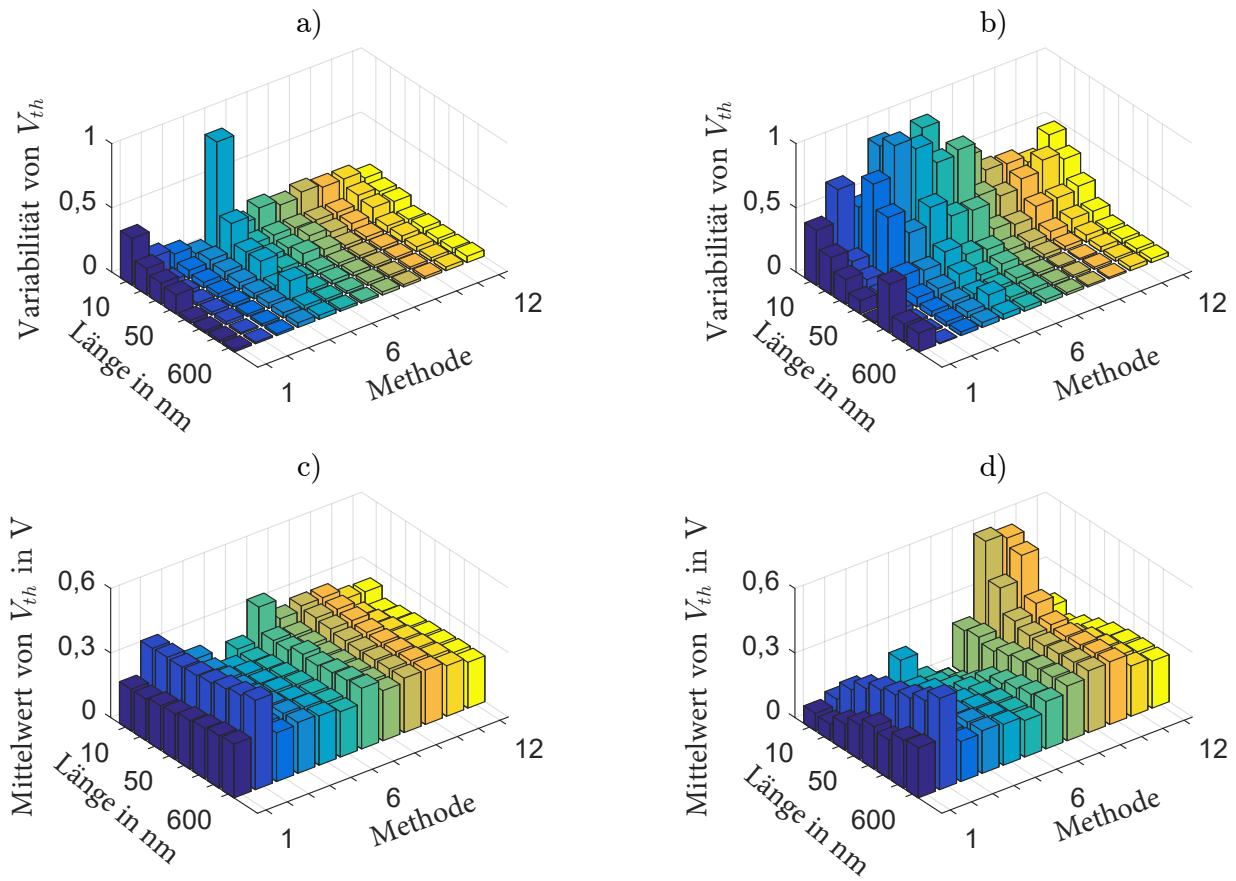


Abbildung 5.6: Variabiliät und Mittelwerte von NMOS-Transistoren im linearen Bereich (a, c) und in Sättigung (b, d) für verschiedene Kanallängen. (10000 Transistoren pro Säule)

gleich zu den Bulk-MOSFET-Messungen des vorherigen Abschnitts erneut höhere Fehler. Wie über alle Messungen hinweg zeigt die ACC-Methode gute Ergebnisse und eine gute Skalierung bzgl. des Rauschens. Somit kann geschlussfolgert werden, dass Methoden, die zusätzliche Effekte durch Transistormodellgleichungen während der Messung unterdrücken, zu deutlich besseren Ergebnissen führen, als solche Methoden, die rein auf den Messwerten operieren. Des Weiteren ist festzustellen, dass neben der Wahl der Methode der Arbeitspunkt eine wesentliche Rolle bei der Bestimmung der Schwellspannung spielt. Zudem kann keine generelle Aussage über die beste Messmethode getroffen werden, da diese sich zwischen den Technologien unterscheidet. Insgesamt zeigen ableitungsbasierte Methoden die geringste Zuverlässigkeit der Messungen, da diese sehr stark durch Rauschen beeinflusst werden.

5.1.2 Messungen am Transistor

Aus dem vorherigen Abschnitt ist klar geworden, dass die Messmethoden, die auf den Modellgleichungen der Transistoren basieren, für alle Messungen die besten Ergebnisse erzielen. Daher wird in diesem Abschnitt die entwickelte Messmethode verwendet, um

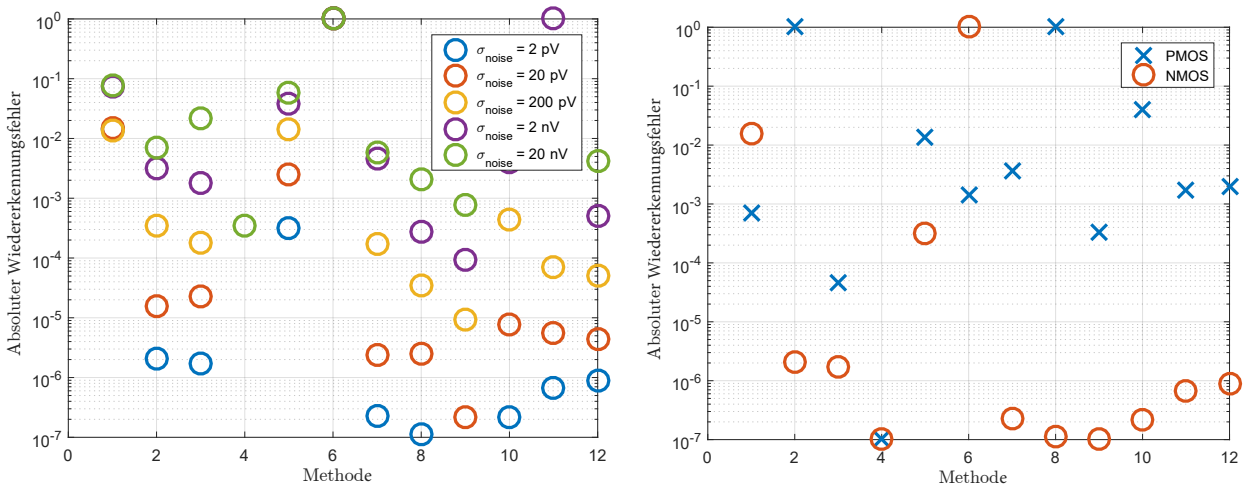


Abbildung 5.7: Vergleich von 90nm-Bulk-PMOS- und NMOS-Transistoren für verschiedene Rauschleistung sowie die Wiedererkennungsfehler. 500 Transistoren mit 15 wiederholten Messungen je Methode.

die Schwellspannung von Transistoren zu bestimmen. Als Modell für den Transistor innerhalb des Partikelfilers wird das EKV-2.6-Transistormodell [BLE+98] verwendet. Es ist in der Lage, mit nur wenigen Parametern das wesentliche Verhalten eines Transistors nachzubilden. Der wesentliche Teil des Verilog-Codes des Modells ist im Code-Ausschnitt 5.1 gezeigt.

Listing 5.1: Vereinfachte EKV-2.6-Modellgleichungen in Verilog-A.

```
analog begin
VG = V(g, b); VS = V(s, b); VD = V(d, b);
VGprime = VG - VTO + PHI + GAMMA * sqrt(PHI);
VP = VGprime - PHI - GAMMA
* (sqrt(VGprime+(GAMMA/2.0)*(GAMMA/2.0))-(GAMMA/2.0));
n = 1.0 + GAMMA / (2.0*sqrt(PHI + VP + 4.0*$vt));
beta = KP * (W/L) * (1.0/(1.0 + THETA * VP));

x=(VP-VS)/$vt;
iff = (ln(1.0+exp( x /2.0)))*(ln(1.0+exp( x /2.0)));
x=(VP-VD)/$vt;
ir  = (ln(1.0+exp( x /2.0)))*(ln(1.0+exp( x /2.0)));

Ispec = 2 * n * beta * $vt * $vt;
Id = Ispec * (iff - ir);
I(d,s) <+ Id;
end
```

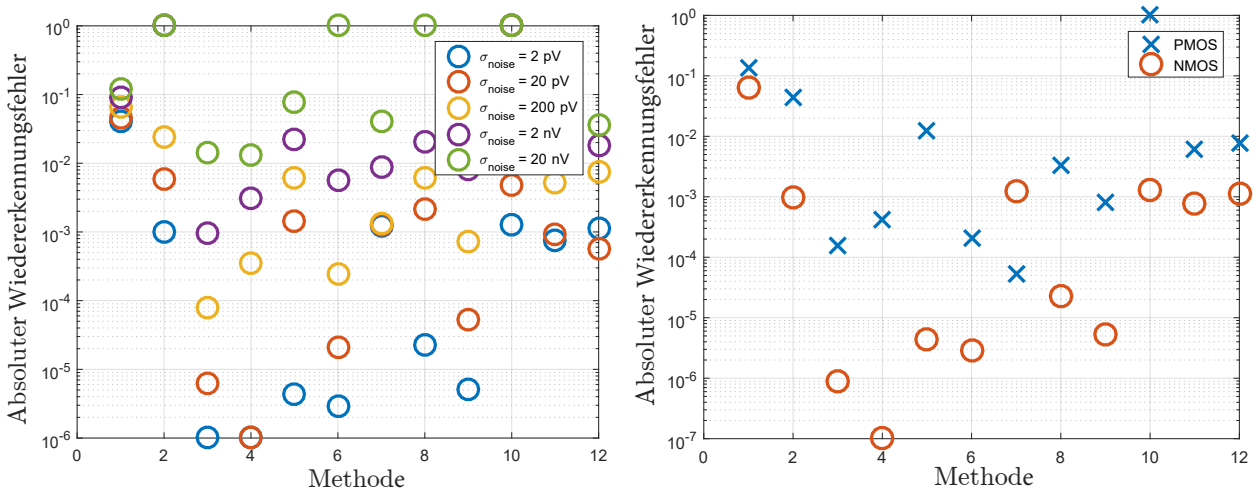


Abbildung 5.8: Vergleich von 20nm-FinFET-PMOS- und NMOS-Transistoren für verschiedene Rauschleistung sowie die Wiedererkennungsfehler. 500 Transistoren mit 15 wiederholten Messungen je Methode.

Dabei ist $VT0$ die Schwellspannung, $GAMMA$ und PHI sind Potentiale, KP der Mobilitätsparameter und $THETA$ gibt den Sättigungsbereich an. Neben diesem vereinfachten Modell gibt es noch weitere, die in der Lage sind, sekundäre Effekte zu modellieren. Zu Beginn muss dieses Modell parametriert werden, dies kann wie bereits beschrieben durch Simulations- oder Messwerte geschehen. Dazu wird die Gleichung

$$V_g(I_d) = \frac{nU_T \ln(I_d) + \frac{nU_T}{I_0} \cdot I_d - nU_T \ln(K \cdot I_0)}{1 - \frac{nU_T \Theta}{I_0} I_d} \tag{5.9}$$

verwendet [OCGSM+13]. Dabei ist n der Steigungsfaktor, U_T die Temperaturspannung, I_0 der Technologiestrom. Des Weiteren gilt $KP = K = \mu C_{ox}$ und $\Theta = THETA$. Falls vorhanden, können einige der Größen ebenfalls direkt ohne Simulationen aus dem PDK entnommen werden. Die Gleichung (5.9) kann anhand einer Übertragungscharakteristik parametriert werden. Wichtig ist, dass die Schwellspannung zu diesem Zeitpunkt nicht bestimmt worden ist. Alle Extraktionsabläufe, bei denen zu Beginn die Schwellspannung gemessen wird und darauf aufbauend die weiteren Parameter, würden die Ergebnisse verfälschen. Bei dem hier vorgestellten Vorgehen bleibt die Schwellspannung bis zur Bestimmung durch den Partikelfilteralgorithmus außen vor. Daher wird nicht der für das EKV vorgeschlagene Parametrisierungsablauf verwendet [GTJ+15]. Die beiden Parametersätze können ineinander umgerechnet werden. Somit wird das EKV-2.6-Modell in diesem Beispiel für einen FDSOI-NMOS-Transistor mit einer Weite $W = 1\mu m$ und einer Länge von $L = 100nm$ angepasst. Die Daten stammen aus TCAD-Simulationen. Die Werte, die für beide Parametersätze aus diesem Beispiel resultieren, sind in Tabelle 5.2 gezeigt. Der Wert für die Schwellspannung ist nur eine initiale Annahme und wurde, wie bereits beschrieben, nicht innerhalb der Methode für die Extraktion der anderen Parameter verwendet. Zum Vergleich zeigt Abbildung 5.9 die

Tabelle 5.2: Typische Parameterwerte für einen NMOS-FDSOI-Transistor in Sättigung

n	I_0	K	Θ	
$1,5$	$1,4\,\mu$A	$366\,{}^{\mu A}/\mathrm{V}^2$	44 mV^{-1}	
$GAMMA$	PHI	KP	$THETA$	VTO
$0,4\,\mathrm{V}^{0,5}$	$0,3$ V	$366\,{}^{\mu A}/\mathrm{V}^2$	44 mV^{-1}	250 mV

in TCAD simulierte Übertragungskennlinie und die durch das parametrierte Modell nachgebildete Übertragungskennlinie des NMOS-FDSOI-Transistor aus Tabelle 5.2. Für den linear-skalierten Ausschnitt um den Bereich der Schwellspannung zeigen sich äußerst geringe Abweichungen. Für den Bereich in starker Inversion besteht zwischen

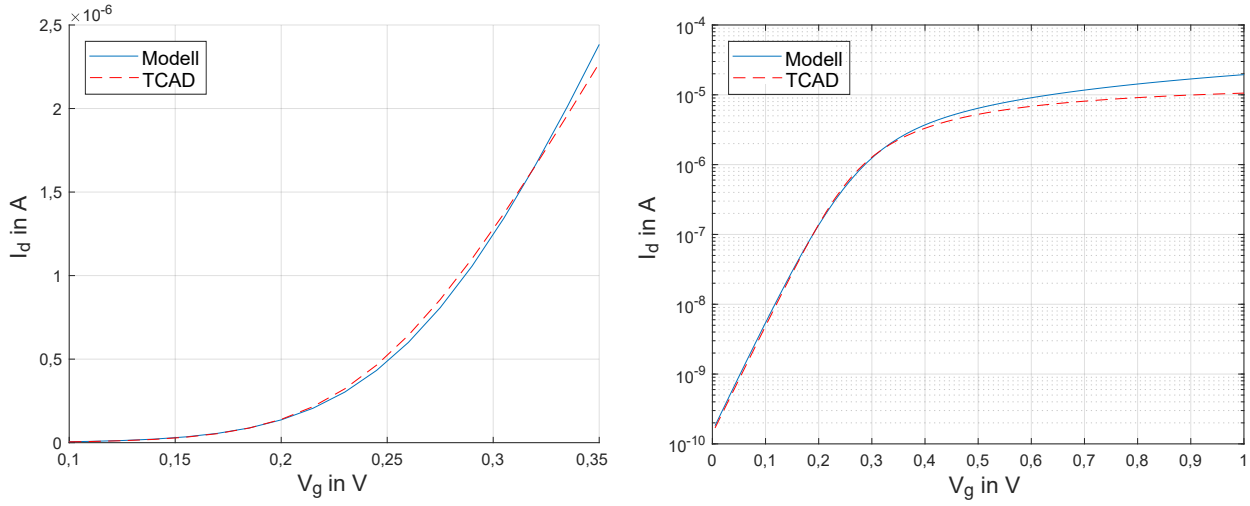


Abbildung 5.9: Übertragungskennlinie eines FDSOI NMOS Transistor in linearer und logarithmischer Skalierung.

dem Modell und den tatsächlichen TCAD-Daten eine Abweichung. Diese ist dadurch bedingt, dass das Modell die Geschwindigkeitssättigung nicht berücksichtigt. Dies kann durch die Erweiterung des Modells um den Parameter $UCRIT$ behoben werden, welches modellintern zu einer reduzierten Drain-Source-Spannung V'_{dss} führt und somit den Drainstrom in starker Inversion reduziert. Für die Bestimmung der Schwellspannung ist der Bereich, in dem die Abweichungen auftreten, irrelevant und kann daher vernachlässigt werden.

Im Weiteren wird ein 65nm-NMOS-Transistor mit $L = 250$nm und $W = 3\mu$m als Device-unter-Test (DUT) verwendet. Die bekannten Schwellspannungen dieses Transistors sind $V_{th,lin} = 249$ mV im linearen Bereich und $V_{th,sat} = 251{,}2$ mV in Sättigung. Zuerst wird das Konvergenzverhalten des Algorithmus betrachtet. Dazu werden mit dem Algorithmus die Drainstromwerte während der Aufnahme der Übertragungskennlinie übergeben. Die Zeitabschnitte korrespondieren dabei zu Erhöhungen der Gate-Spannung

von 0V auf 1V in 2,5mV-Schritten. In Abbildung 5.10 ist das Konvergenzverhalten des [illegible] r

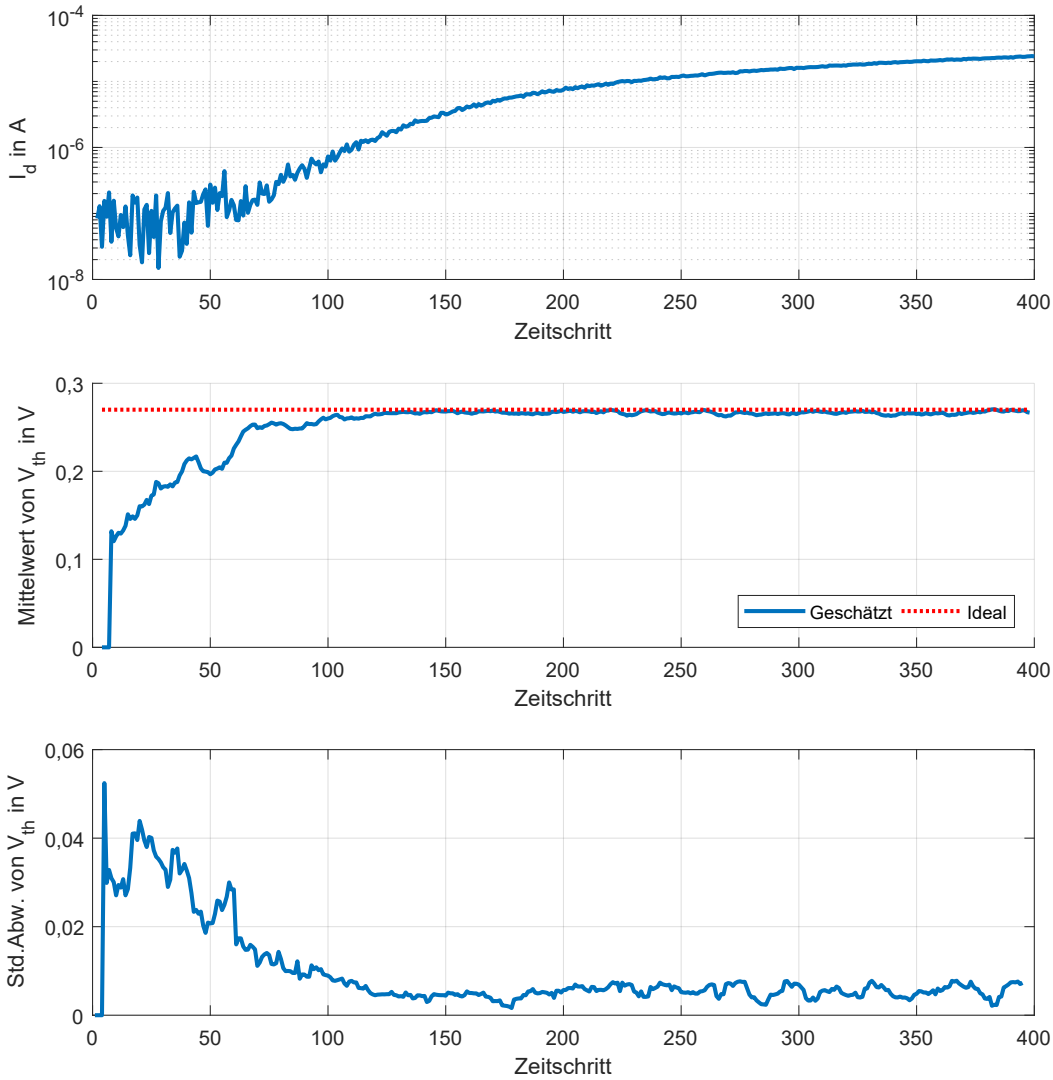


Abbildung 5.10: Konvergenz des Partikelfilters für linearen Bereich mit geringem Rauschen. Mittelwert und Standardabweichung der Schätzung von V_{th}.

Algorithmus den korrekten Wert der Schwellspannung nach ungefähr 100 Messpunkten. Zu diesem Zeitpunkt ist die Standardabweichung bei ungefähr 1mV. Die Gatespannung ist zu diesem Zeitpunkt ungefähr 250mV. Damit ist der Transistor nur in moderater Inversion. Für spätere Zeitpunkte fällt die Standardabweichung auf 0,6mV im Mittel, was durch das Rauschen bedingt ist. Die Geschwindigkeit der Konvergenz kann durch die Erhöhung der Partikelzahl erhöht werden. Da dies das übliche Szenario ist, um die Schwellspannung zu bestimmen, kann diese Extraktion mit der entwickelten Methode direkt mit den bereits vorgestellten Methoden verglichen werden. Für die CsrTR-Methode wurde die Variante mit abgeändertem Fitting verwendet. Den anderen Methoden standen dieselben Messwerte zur Verfügung. Für einen breiteren Vergleich

wurden insgesamt vier verschiedene Fälle unterschieden. Dazu zählen der lineare Bereich und die Sättigung mit jeweils einem Fall mit hoher $\sigma_{I_d} = 10^{-6}A$ und einem mit niedriger $\sigma_{I_d} = 10^{-7}A$ Rauschleistung. Die Ergebnisse sind in Tabelle 5.3 zusammengefasst. Messwerte die mit einem Asterisk ($*$) gekennzeichnet sind, werden als fehlgeschlagene Messung gewertet, da die jeweilige Methode in diesen Fällen kein verwertbares Ergebnis geliefert hat. Die entwickelte Methode ist in der Lage, die Schwellspannung für alle

Tabelle 5.3: V_{th} Extraktionergebnisse für verschiedene Messmethoden.

Nummer	**Abkürzung**	Linear niedriges σ_{I_D}	Linear hohes σ_{I_D}	Sättigung niedriges σ_{I_D}	Sättigung hohes σ_{I_D}
1	MaxGM	0,4306	0,2102	0,2145	0,1957
2	LE	0,2261	0,2261	0,1763	0,1777
3	ACC	0,0756*	0,0681*	0,1918	0,1619
4	SD	0,2581	0,2982	-0,0025*	-0,0028*
5	TD	0,0585*	0,0861*	0,0316*	0,0501*
6	CsrTR	0,2079	0,2385	0,1658	0,4762
7	TM	0,2534	0,2006	0,3264	0,2156
8	NMID	0,3024	0,2007	0,3912	0,4048
9	NRH	0,3283	0,3795	0,4385	0,4707
10	TCR	0,215	0,2035	0,2153	0,1994
11	RH	0,215	0,2035	0,2152	0,1995
12	**PF**	**0,25**	**0,245**	**0,25**	**0,248**

Testfälle mit einem Fehler von weniger als 2mV zu bestimmen. Die Ergebnisse der meisten anderen Methoden streuen sehr stark. Wie zu erwarten war, liefern die integrationsbasierten Methoden wegen ihrer Rauschtoleranz reproduzierbare Ergebnisse. Die Ergebnisse der ableitungsbasierten Methoden weichen sehr stark vom eigentlichen Wert der Schwellspannung ab. Keine der Vergleichsmethoden ist in der Lage, eine ausreichend genaue Schätzung der Schwellspannung zu liefern. Zur Veranschaulichung der vorgestellten Methode zeigen Abbildungen 5.11 und 5.12 die zeitlichen Verläufe der Rauscheinflüsse und der Systemausgänge, der Messung sowie des durch das Partikelfilter extrahierten Wertes der Schwellspannung. Die Werte sind für den linearen Bereich und die Sättigung gezeigt, wobei die Drainspannung 50mV bzw. 1V beträgt. Zudem sind die jeweiligen Histogramme gezeigt. Durch die verschiedenen Arbeitspunkte unterscheiden sich die Verläufe der beiden Drainströme. Das Prozessrauschen ist als normalverteilt angenommen und das Messrauschen als lognormalverteilt.
Trotz der unterschiedlichen Voraussetzungen ist der Algorithmus bereits nach jeweils 100 Messwerten in der Lage, die Schwellspannung präzise anzunähern. Gegenüber den anderen Methoden zur Messung der Schwellspannung ist der Partikelfilteransatz in der Lage, die Schwellspannung aus beliebigen Verläufen des Drainstromes zu extra-

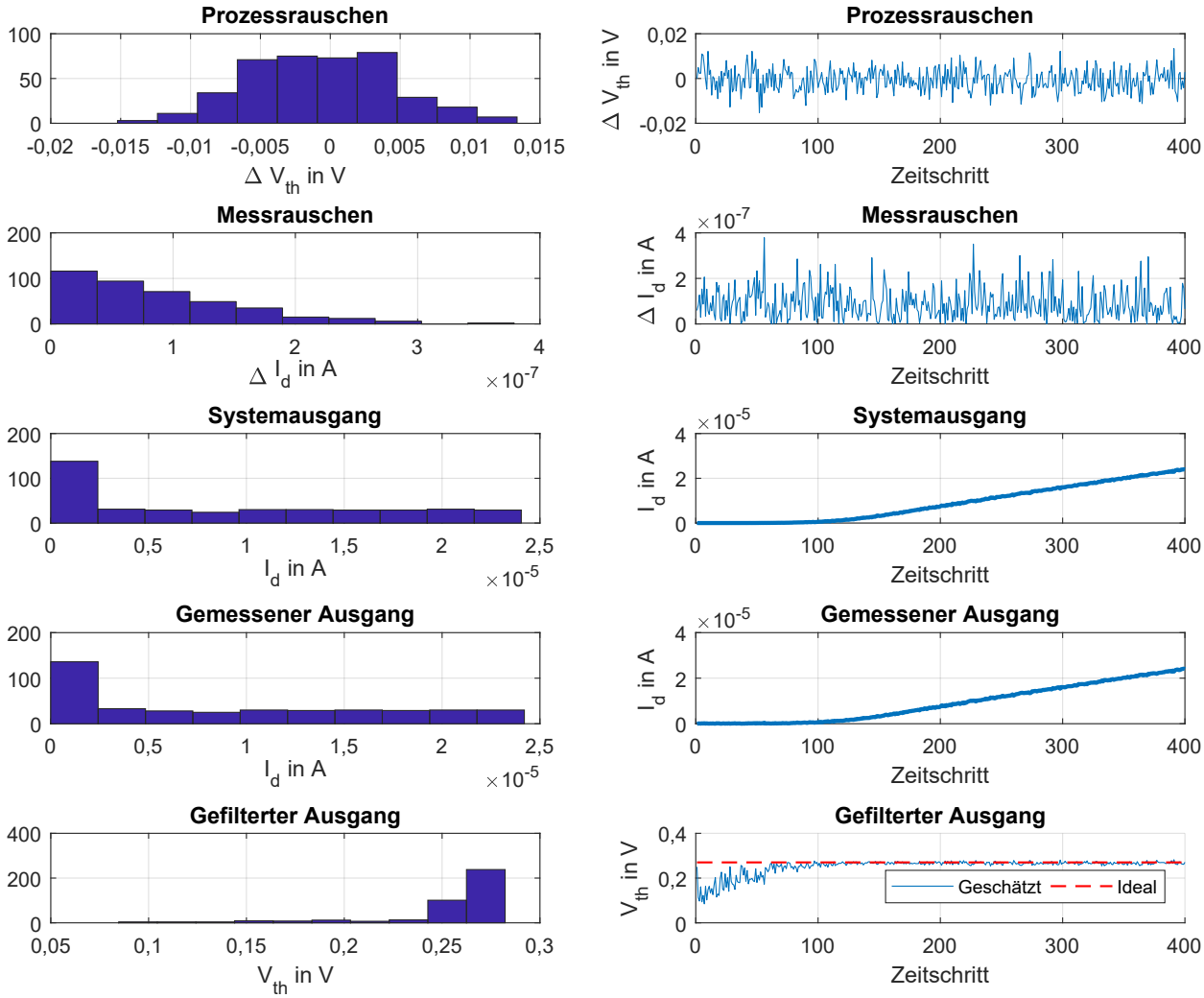


Abbildung 5.11: Zeitliche Verläufe und Histogramme der Ein- und Ausgänge für den linearen Betriebsbereich $V_{ds} = 0,05\,\mathrm{V}$.

hieren. Als Beispiel wird bei einem NMOS-Transistor mit einer Schwellspannung von $V_{th} = 270\mathrm{mV}$ nur die Drainspannung variiert bzw. die Ausgangscharakteristik verwendet. Dieses Szenario eignet sich besonders für die Messung von Einflüssen durch BTI, da die Stressbedingungen am Gate für die Bestimmung der Schwellspannung nicht verändert werden müssen. Somit können keine Ausheilungsprozesse eintreten, die das Messergebnis in kürzester Zeit verfälschen würden. In diesem Beispiel ist $V_g = 1\mathrm{V}$ und $V_d = 0 \ldots 1\mathrm{V}$, siehe Abbildung 5.13. Durch den starken Anstieg des Stroms bei dieser Messung verlässt der tatsächliche Drainstrom nach wenigen Messungen die Größenordnung des Rauschens. Dadurch zeigt sich eine sehr schnelle Konvergenz zum korrekten Wert der Schwellspannung in unter 20 Messungen. Eine derartige Bestimmung der Schwellspannung in so kurzer Zeit bzw. mit einer so geringen Anzahl von Messpunkten kann sehr gut für die Charakterisierung von Alterungseinflüssen verwendet werden. Die Standardabweichung der Schätzung bleibt in diesem Fall relativ hoch, dies ist jedoch bedingt durch das sehr hohe Rauschen. Wie aus dem Histogramm der Schwellspannung abgelesen werden kann,

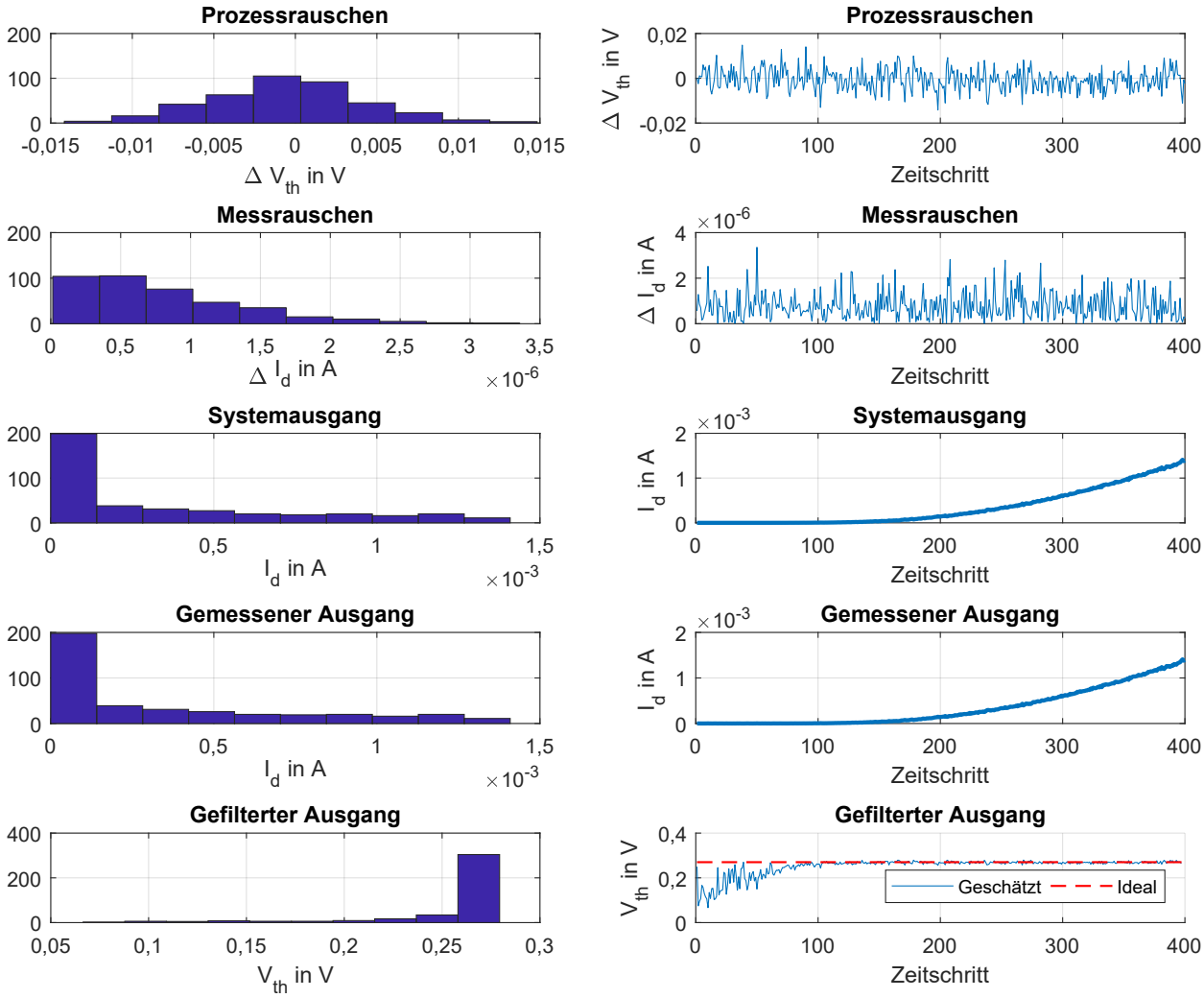


Abbildung 5.12: Zeitliche Verläufe und Histogramme der Ein- und Ausgänge für den gesättigten Betriebsbereich $V_{ds} = 1\,\mathrm{V}$.

ergibt eine Mittelung der Schätzwerte exakt den Schwellspannungswert von 270mV. Für die tatsächliche Messung von Alterungseinflüssen muss der Algorithmus in der Lage sein, eine Veränderung der Schwellspannung über der Zeit korrekt als solche zu filtern. Dazu wird die Schwellspannung des Transistors künstlich von 270mV auf 300mV ab dem 50. Messwert durch eine logistische Funktion angehoben. Weiterhin stehen dem Algorithmus lediglich das parametrisierte EKV-2.6-Modell und die Messungen des Drainstroms zur Verfügung. Der Transistor wird mit $V_g = 0{,}65\mathrm{V}$ an der Grenze zur starken Inversion und mit $V_d = 50\mathrm{mV}$ im linearen Bereich betrieben. Wie in Abbildung 5.14 gezeigt ist, ist das Messrauschen sehr stark ausgeprägt. Dies soll die Umwelteinflüsse während einer beschleunigten Alterungsmessung nachbilden. Der in der Abbildung dargestellte Drainstrom zeigt somit ein stark verrauschtes Verhalten. Nichtsdestotrotz ist das Filter in der Lage, den Verlauf der Schwellspannung korrekt aus den Messungen des Drainstroms zu extrahieren. Um ein realistisches Szenario nachzubilden, wird für die kommende Analyse die Schwellspannung wie zuvor erhöht, zudem wird die Drain-

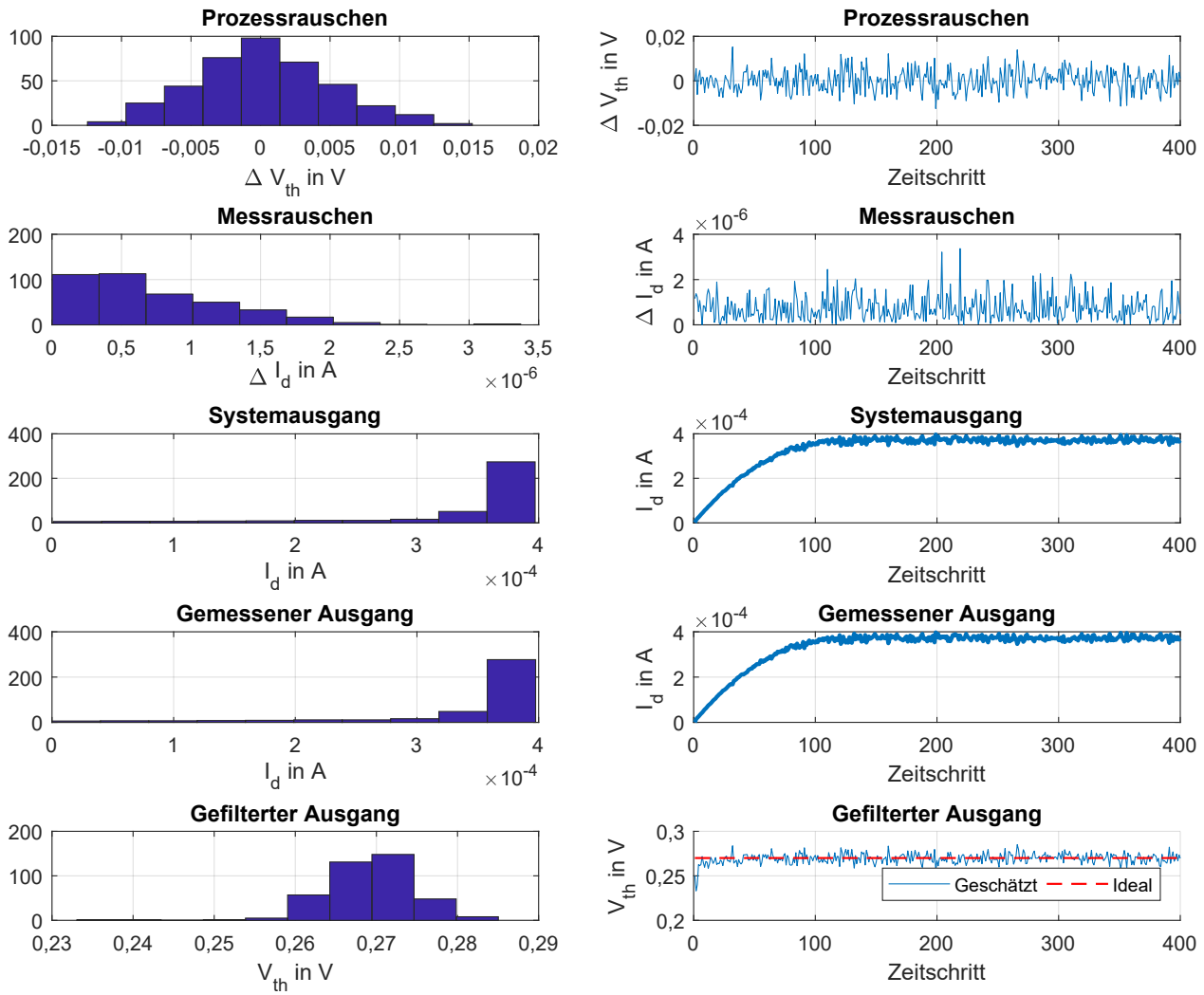


Abbildung 5.13: Extraktion von V_{th} mit ID-VD Charakteristik ($V_g = 1$V).

spannung von 0V auf 1V erhöht und die Temperatur steigt von 300K auf 350K. Des Weiteren wird die Gatespannung mit 0,28V genau in die moderate Inversion an die Grenze der Schwellspannung gelegt. Wie in den vorherigen Kapiteln beschrieben, folgt somit ein sehr sensitiver Arbeitspunkt in Bezug auf äußere Einflüsse für den Transistor. Wie die Verläufe in Abbildung 5.15 zeigen, ist der Algorithmus in der Lage, die einzelnen Einflüsse zu separieren und aus dem sehr komplexen Verlauf des Drainstroms die korrekte Veränderung der Schwellspannung zu extrahieren. Dabei steigt der Drainstrom durch die Erhöhung von V_d stark an. Danach stellt sich ein Gleichgewicht mit der einsetzenden Erhöhung der Schwellspannung ein. Da durch die weitere Erhöhung von V_d der Sättigungsstrom erreicht ist, beginnt der Drainstrom durch den weiteren V_{th}-Anstieg zu fallen. Nachdem $V_{th} = 300$mV erreicht wurde, steigt der Strom wieder leicht durch die weitere Temperaturerhöhung an. Diese Fähigkeit des Algorithmus macht es möglich, aus beliebigen Stressprofilen während Langzeitmessungen oder aus Messungen im Rahmen beschleunigter Alterung die Schwellspannung jederzeit extrahieren zu können. Zudem müssen die Stressbedingungen für die Messung nicht verändert werden, dies verbessert

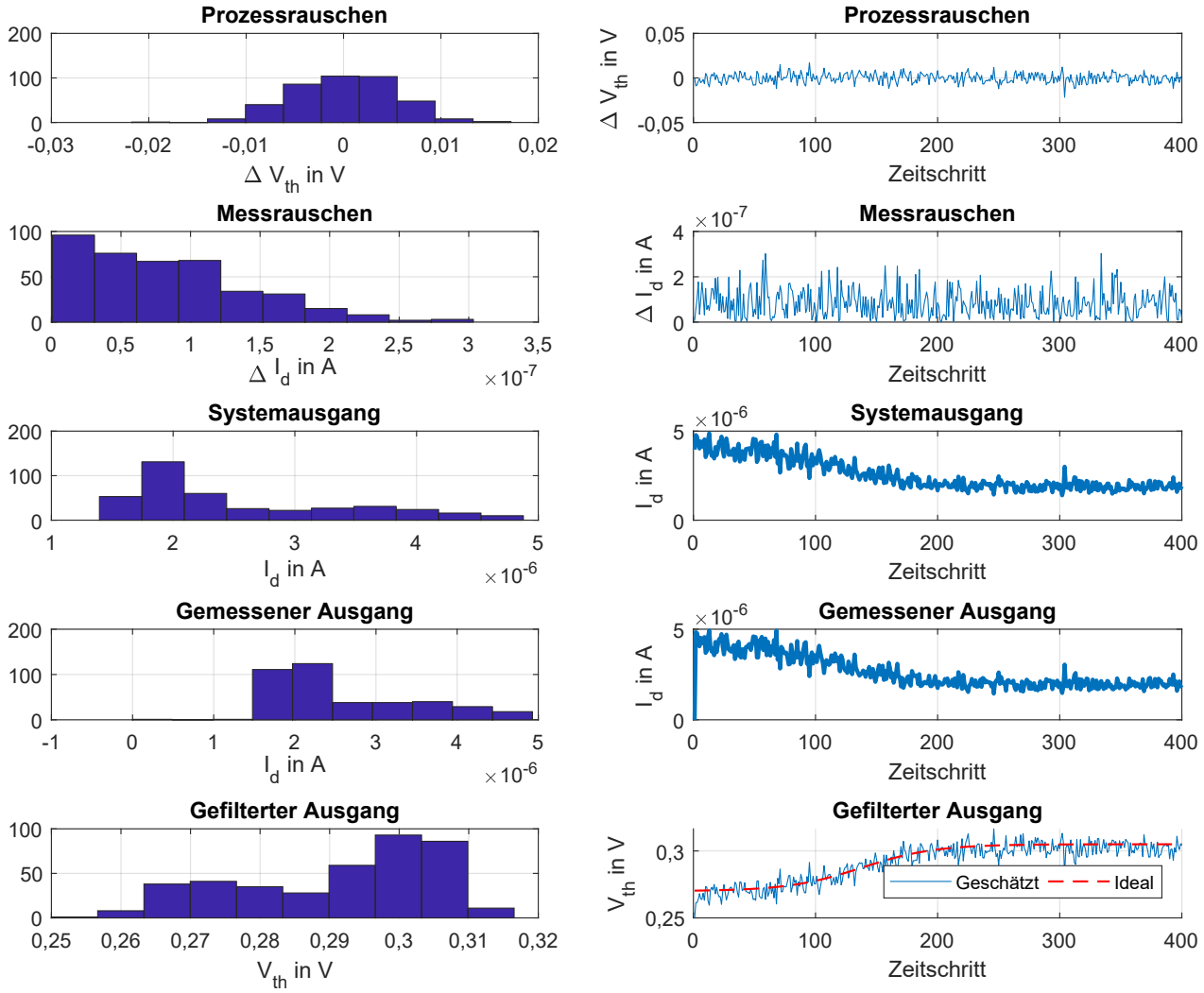


Abbildung 5.14: Extraktion der steigenden Schwellspannung V_{th} von 270mV auf 300mV, während sämtliche Umweltparameter konstant gehalten werden und starkes Rauschen angenommen wird.(T=300K, V_g = 0,65V)

das Messergebnis, da keine Ausheilprozesse stattfinden können. Des Weiteren kann die Schwellspannung in beliebigen Arbeitspunkten extrahiert werden und somit die Alterungseinflüsse dediziert für einzelne Arbeitspunkte untersucht werden.

Eine genauere Betrachtung des Konvergenzverhaltens und des Verlaufes der Schwellspannung für das letzte Szenario ist in Abbildung 5.16 gezeigt. Dieser Verlauf ist das Entscheidende bei der Messung von Alterungseinflüssen, da dieser für die weitere Modellierung verwendet wird. Nach einem kurzen Anstieg der Standardabweichung durch die Konvergenz des Algorithmus bleibt diese deutlich unter 0,5mV. Somit weist die Schätzung für praktische Anwendungen nahezu keine Abweichungen auf. Der Mittelwert folgt dem idealen Verlauf nach der Konvergenz über den gesamten Messzeitraum.

Im diesem Abschnitt wurde gezeigt, dass der entwickelte Partikelfilter-Algorithmus besser für die Messung der Schwellspannung geeignet ist als elf andere herkömmliche

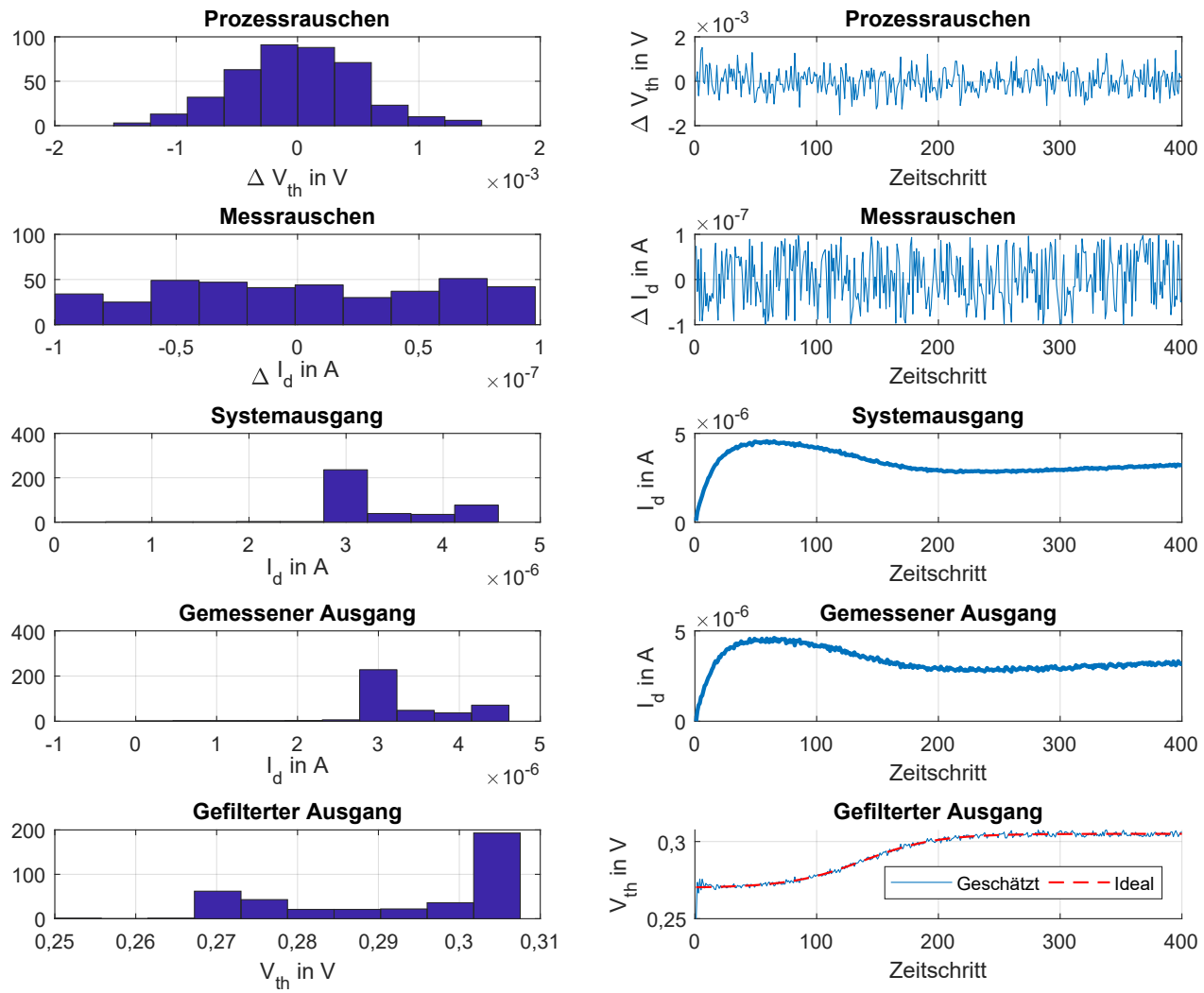


Abbildung 5.15: Extraktion der steigenden V_{th} von 270mV auf 300mV, während V_d von 0 auf 1V verändert wird und die Temperatur von 300k auf 350K für geringes Rauschen.($V_g = 0,28$V)

Messmethoden. Zudem kann das Filter dazu verwendet werden, die Schwellspannung an beliebigen Arbeitspunkten zu bestimmen. Darüber hinaus können mit dem Algorithmus einzelne Einflüsse durch Spannungs- und Temperaturänderung von Alterungseinflüssen getrennt werden. Zusammen mit der Parametrierung des EKV Models für den jeweiligen Transistor ist es erstmalig innerhalb von Messungen möglich, Prozessvariation, Temperatureinflüsse, Spannungsschwankungen und Alterung getrennt zu betrachten und unterscheidbar zu machen.

Neben der reinen Analyse können die zeitgleich zur Messung stattfindenden Schätzungen für die Vorhersage verwendet werden. Dies führt in Kombination mit der beschleunigten Alterung zu verschiedenen Möglichkeiten zur frühzeitigen Bestimmung möglicher System- oder Schaltungsausfälle, die im letzten Kapitel der Arbeit beschrieben werden. Mit dem Algorithmus können zudem nicht nur einzelne Transistoren, sondern ebenfalls

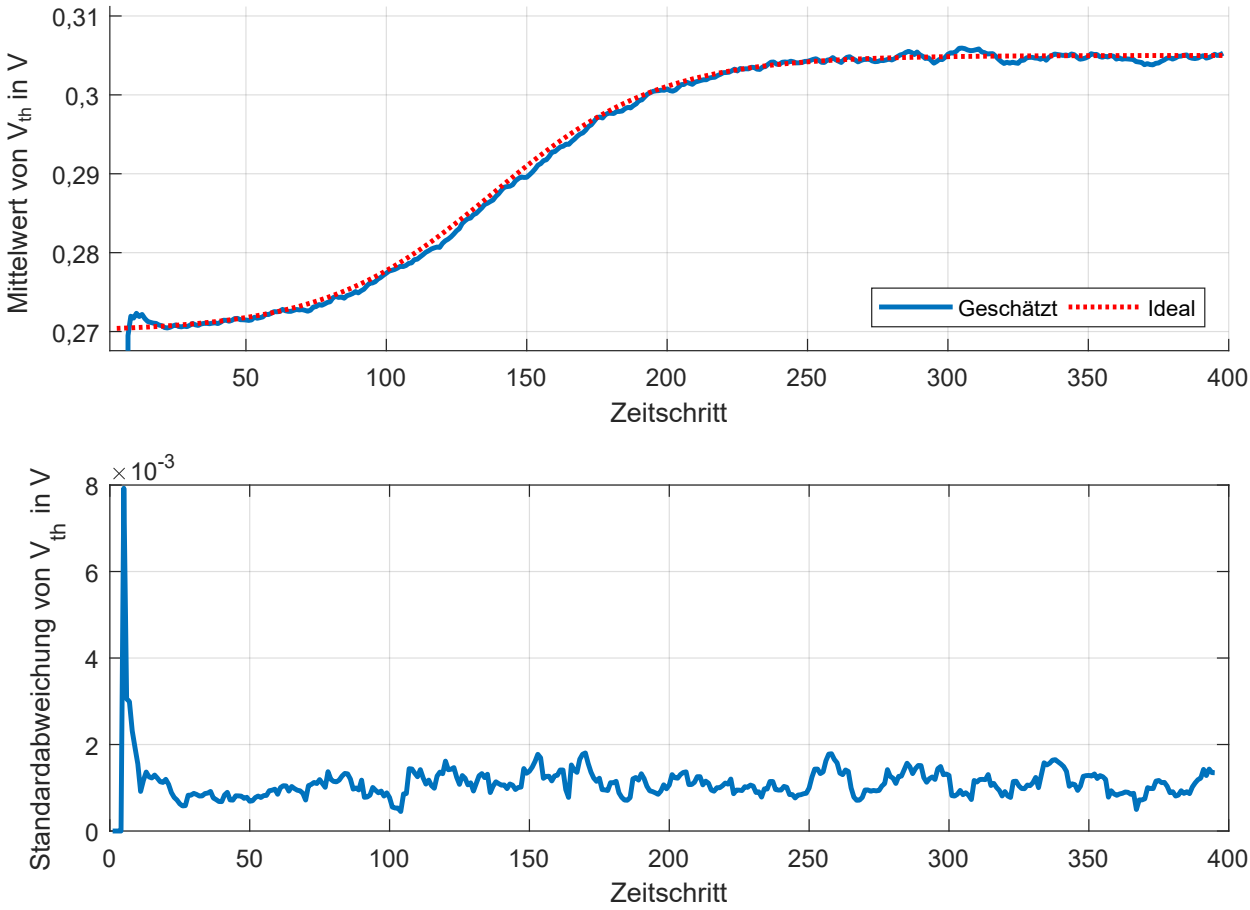


Abbildung 5.16: Mittelwert und Standardabweichung von V_{th} über den gesamten Messzeitraum mit einer Mittlung über jeweils acht Datenpunkte.

Schaltungen und ihr Verhalten untersucht werden. Dies ist im folgenden Abschnitt anhand des gefertigten Beta-Mulitpiers gezeigt.

5.1.3 Messungen am Beta-Multiplier

In diesem Abschnitt wird das Temperaturverhalten der gefertigten Referenzquelle mit Hilfe des Partikelfilter-Algorithmus untersucht und extrahiert. Wie bereits beschrieben, ist die Referenzspannung durch

$$V_{ref} = V_{dd} - \frac{2}{R \cdot \beta_5}\left(1 - \frac{1}{\sqrt{k}}\right) - V_{th,5} \tag{5.10}$$

gegeben. Dabei ist die Referenzspannung die einzige messbare Größe. Alle anderen Parameter können bei einer gefertigten Schaltung ohne weiteren schaltungstechnischen Aufwand (test pads o.ä.) nicht bestimmt werden. Die Parameter R, β, k und V_{th} sind von Interesse zur Untersuchung des nominellen Verhaltens der Schaltung und für die Modellierung des Temperaturverhaltens. Dadurch, dass zur Herleitung von Gleichung (5.10) das MOS1-Modell verwendet wurde, muss zuerst die Validität der Gleichung überprüft werden. Dazu wird für verschiedene Temperaturen die Referenzspannung simulativ bestimmt und anhand der Gleichung berechnet. Die Tabelle 5.4 zeigt den Vergleich der simulierten und berechneten Referenzspannungen. Die Abweichung bleibt zur Berechnung bleibt unter 2%, daher kann im Weiteren die Gleichung für die Berechnung der Referenzspannung verwendet werden. Als nächstes muss das Modell (Gleichung

Tabelle 5.4: Berechnete und simulierte Werte für die Referenzspannung V_{ref} in mV über den Temperaturbereich $1 - 120°C$.

T in $°C$	1	27	80	120
$V_{ref,calc}$ in mV	551	556,5	566.3	576
$V_{ref,sim}$ in mV	556	563,77	576,48	585,7
ΔV_{ref} in mV	5	7	10	9.7
ΔV_{ref} in %	1	1,25	1,76	1,68

(5.10)) der Schaltung durch Simulationen parametriert werden. Dazu werden zur Berücksichtigung der Prozessvariationen 200 Monte-Carlo-Simulationen durchgeführt. Die Ergebnisse für die relevanten Parameter sind in Tabelle 5.5 gezeigt. Dabei sind die spezifizierten Parameter die, die beim Entwurf angestrebt worden waren. Besonders beim Widerstand fällt bereits auf, dass 100Ω zwischen dem Mittelwert und dem spezifizierten Wert liegen. Diese Werte bilden die initialen Werte für das Partikelfilter. Durch die vorherige Monte-Carlo-Simulation ist sichergestellt, dass die Startwerte für das System die am wahrscheinlichsten Werte sind. Neben der reinen Modellbestimmung ist die Korrelation zwischen der Referenzspannung und einzelnen Parametern wichtig. Falls einzelne Parameter korreliert sein sollten, müsste dies in der Zustandsraumdarstellung mit Hilfe

Tabelle 5.5: Spezifizierte Werte und Ergebnisse der Monte-Carlo-Analyse für die Parameter R, β, k, und V_{th}, und für die Spannung V_{ref}

Name	Spezifiziert	Mittelwert	Std.-Abw.
R in kΩ	5,5	5,604	0,1933
β in $\mu A/V^2$	3,6	3,719	0,1236
k	2	2,081	0,052
V_{th} in mV	414	417,3	14,56
V_{ref} in mV	500	572	53,39

eines Korrekturterms berücksichtigt werden. Dazu sind die Korrelationskoeffizienten in Tabelle 5.6 gezeigt.

Tabelle 5.6: Korrelationskoeffizienten des Temperaturverhaltens von V_{ref} und Entwurfsparameter über den Temperaturbereich $1 - 120°C$.

$\rho_{V_{ref}}$	$1°C$	$27°C$	$80°C$	$120°C$
V_{th}	0,33	0,33	0,23	0,31
k	0,13	0,1	0,02	0,05
β	0,18	0,22	0,3	0,35
R	0,16	0,18	0,23	0,27

Dabei muss zwischen dem Zusammenhang der Größen aus (5.10), dem Temperaturverhalten der Größen und dessen Zusammenhang mit dem Temperaturverhalten der Ausgangsgröße V_{ref} unterschieden werden. Klar ist, dass jede der Größen eine Temperaturabhängigkeit aufweist und der funktionale Zusammenhang ebenfalls gegeben ist. Bei der Korrelation wird, wie in Abschnitt 2.5.2.4 beschrieben, der lineare Zusammenhang zwischen Temperaturverhalten der einzelnen Größen untersucht. Da keiner der Werte oberhalb von 0,5 liegt, werden die Parameter über alle Temperaturen als unkorreliert angenommen. Für dieses Szenario wird der zuvor extrahierte Mittelwert und die Wahrscheinlichkeitsverteilungsfunktion aus der Simulation verwendet. Für den Zustandsvektor zum Messzeitpunkt $k = 0$

$$\mathbf{x} = [V_{th}, R, \beta, k]_{k=0} \tag{5.11}$$

folgt für den Mittelwert

$$\bar{\mathbf{x}} = [414\,\mathrm{mV}, 5.5\,\mathrm{k}\Omega, 3.6\,{}^{\mu\mathrm{A}}/\mathrm{V}^2, 2]_{k=0} \tag{5.12}$$

und die Standardabweichung

$$\mathrm{std}(\mathbf{x}) = [15\,\mathrm{mV}, 190\,\Omega, 0, 16\,{}^{\mu\mathrm{A}}/\mathrm{V}^2, 100\mathrm{m}]_{k=0}\ . \tag{5.13}$$

Für die Verteilung werden Gaussverteilungen angenommen. Zusätzlich zu den Prozessvariationen wird von einem Messrauschen von $\sigma_{meas} = 15\,\mathrm{mV}$ ausgegangen. Um für diesen Fall die Konvergenz des Filters zu untersuchen, werden zwei verschiedene Partikelzahlen $N_s = 5$ und $N_s = 40$ für die initiale Bestimmung verwendet. Die Verläufe für die Parameter sowie deren Standardabweichung sind in den Abbildungen 5.17 und 5.18 gezeigt.
Wie zu erkennen ist, führt eine geringe Anzahl an Partikeln zu einer deutlich erhöhten Standardabweichung und sogar, wie im Fall für den Widerstandswert, zu einer Divergenz des Filters. Für eine ausreichend große Zahl von Partikeln kann allein durch die Messung der einen Ausgangsgröße V_{ref} korrekt auf die ansonsten nicht messbaren Größen R, β, k und V_{th} geschlossen werden. Dabei ist zu beachten, dass das Filter zu der durch die Prozessvariation wahrscheinlichsten Kombination der Abweichungen der Parameter konvergiert. Für eine Partikelzahl von 40 stellen sich für alle vier Größen Standardabweichungen von unter 1% ein. Dabei ist die Schwellspannung am stärksten von den Prozessvariationen betroffen.

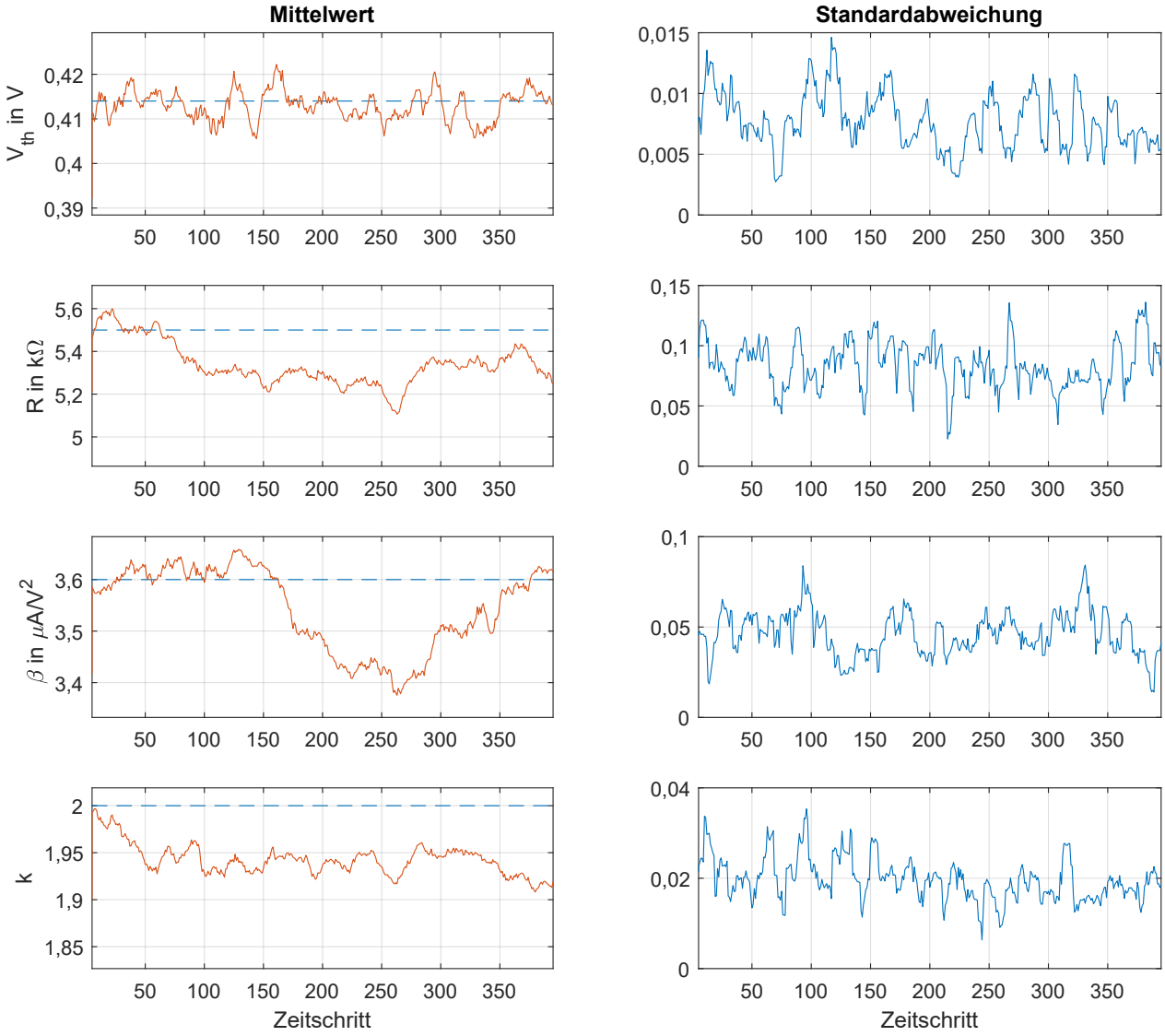


Abbildung 5.17: Simulationsergebnisse für $\mathbf{x} = [V_{th}, R, \beta, k]$ für eine Partikelanzahl von $N_s = 5$ und konstantes $V_{dd} = 1\,\mathrm{V}$ und $T = 300\,\mathrm{K}$. Die jeweiligen Idealwerte sind als gestrichelte Linien dargestellt.

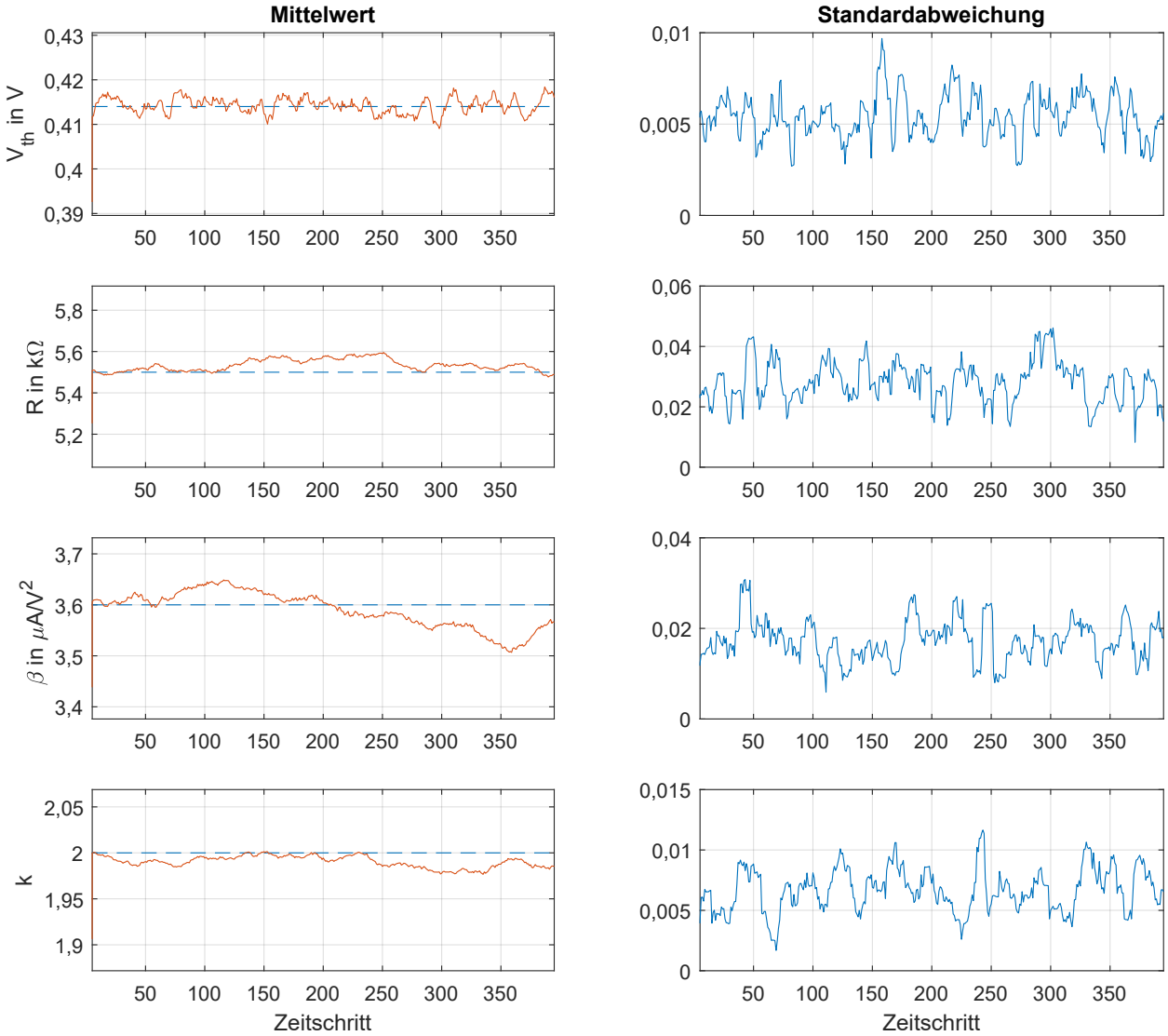


Abbildung 5.18: Simulationsergebnisse für $\mathbf{x} = [V_{th}, R, \beta, k]$ für eine Partikelanzahl von $N_s = 40$ und konstantes $V_{dd} = 1\,\mathrm{V}$ und $T = 300\,\mathrm{K}$. Die jeweiligen Idealwerte sind als gestrichelte Linien dargestellt.

Ähnlich zum Vorgehen bei einzelnen Transistoren soll auch die Schwellspannung eines Transistors der Referenzquelle bestimmt werden. Dabei soll die Versorgungsspannung V_{dd} starken Schwankungen ausgesetzt sein. Modelliert wird dies durch einen sinusförmigen Offset und beim Zeitschritt $k = 200$ durch einen steilen Anstieg um 100mV. Zudem wird sehr starkes Messrauschen angenommen. Da bei der Messung der Alterung stets die Langzeitveränderung im Mittelpunkt steht, müssen kurzzeitige Störungen herausgefiltert werden können. Die resultierenden Ein- und Ausgänge des Systems und des Filters sind

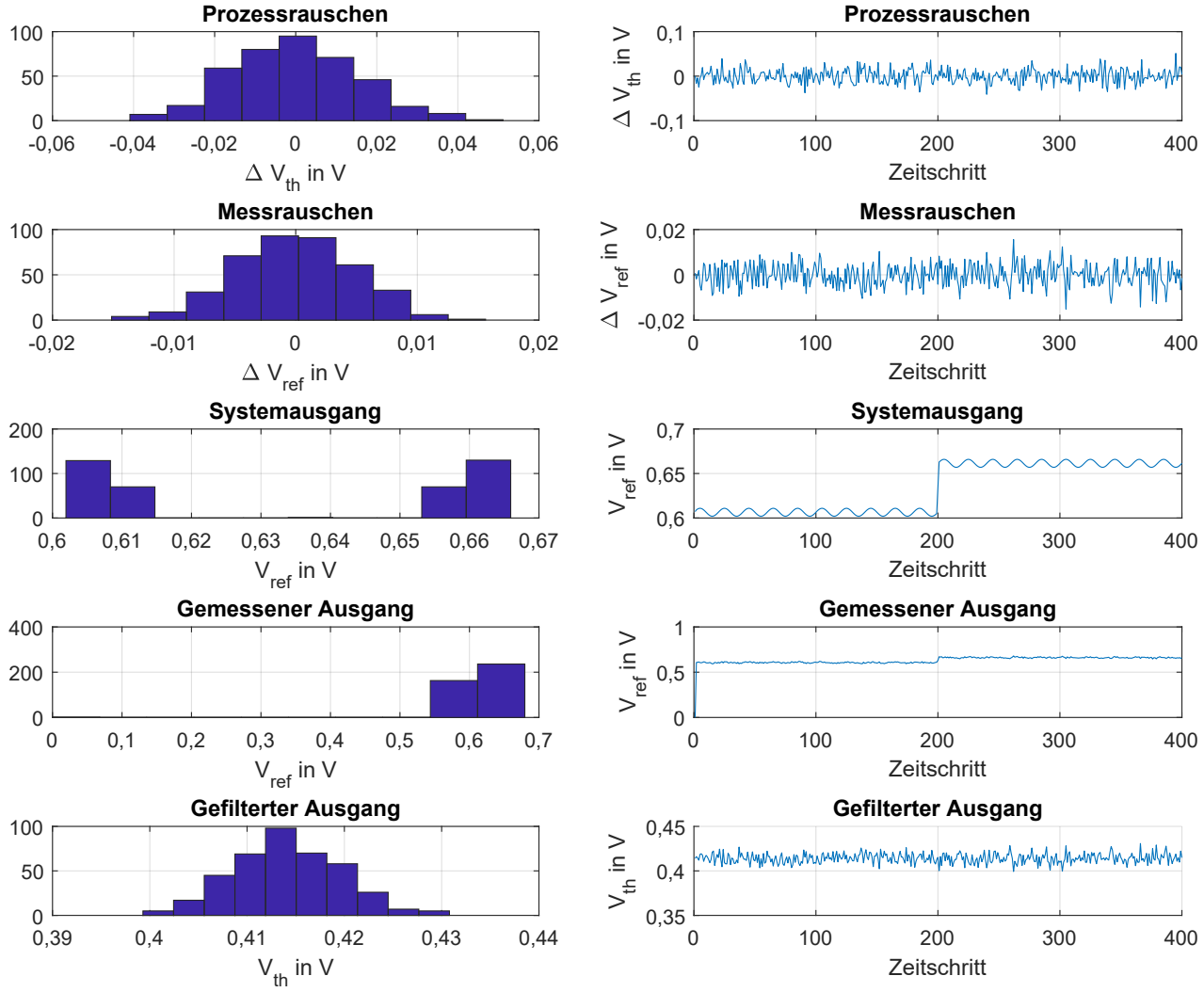


Abbildung 5.19: Simulationsergebnisse von V_{th} für eine alternierende Versorgungsspannung V_{dd} und $T = 300\,\mathrm{K}$.

Durch das sehr hohe schaltungsinterne Rauschen sowie das Messrauschen ist die gefilterte Schwellspannung ebenfalls verrauscht. Am Histogramm kann jedoch gesehen werden, dass der Mittelwert von 414mV sehr genau geschätzt wird. Insgesamt zeigt sich, dass die Möglichkeiten des Algorithmus von einzelnen Bauelementen auf Schaltungen übertragbar sind. Abschließend wird das Modell anhand des bereits für den nominalen Betrieb parametrisierten Teilmodells um eine zusätzliche Abhängigkeit erweitert. Dazu werden

zuerst die Temperaturabhängigkeiten aus der Simulation gewonnen. Für β und die Schwellspannung V_{th} sind die Werte in der Abbildung 5.20 gezeigt. Wie zu erkennen ist, fällt β um 34% und V_{th} um 8%. Durch den funktionalen Zusammenhang auf die Referenzspannung haben die Abweichungen von V_{th} stärkere Auswirkungen auf diese.

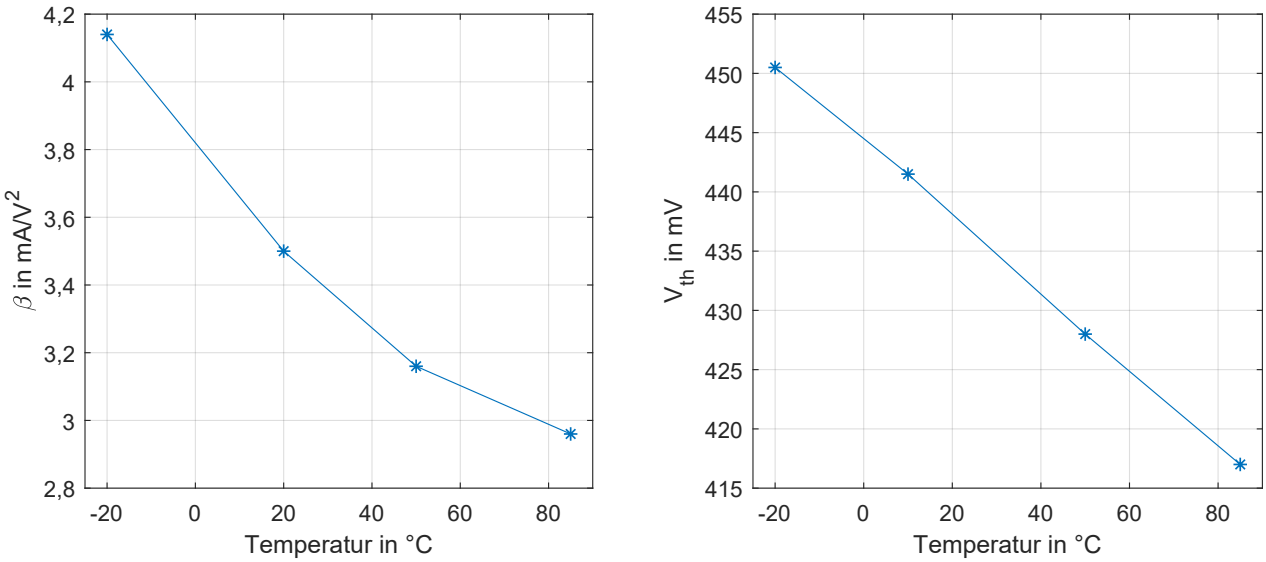


Abbildung 5.20: Simuliertes Temperaturverhalten von β und V_{th} des Transistors M5.

Zur Modellierung der Temperaturabhängigkeit der Referenzspannung kann somit

$$V_{ref}(T) = V_{dd} - \frac{2}{R \cdot a \cdot (1 + {}^{T}/_{T_{ref}})^2 \cdot \beta_5 \cdot b \cdot (1 + {}^{T}/_{T_{ref}})} \cdot \ldots \\ \left(1 - \frac{1}{\sqrt{k}}\right) - V_{th,5} \cdot c \cdot (1 + {}^{T}/_{T_{ref}}) \tag{5.14}$$

verwendet werden. a, b, c sind dabei die Temperaturkoeffizienten der einzelnen Bauelemente. Nun kann mit tatsächlichen Messwerten fortgefahren werden. Die Referenzspannung der gefertigten Schaltung wurde gemessen und die Parameter auf Grundlage des parametrierten Models geschätzt. Die Ergebnisse zeigen die Verläufe in Abbildung 5.21.

Durch die tatsächlichen Abweichungen durch Messgeräte und Messaufbauten sowie den Klimaschrank sind die Abweichungen gegenüber Simulationen erhöht. Zudem konvergieren die Parameter k und β durch ihren relativen schwachen funktionalen Zusammenhang zur Referenzspannung deutlich später. Die 250 Messungen, die notwendig sind, um dem Filter die Konvergenz zu ermöglichen, können mit dem Messaufbau in weniger als einer Sekunde durchgeführt werden. Somit relativiert sich diese Erhöhung für die Durchführung echter Messungen. Mit diesem an die jeweilige Schaltung angepassten Modell kann im Anschluss das Temperaturverhalten untersucht werden. Dazu wurden insgesamt 5 Chips über einen weiten Temperaturbereich von -20°C bis 85°C vermessen. Die Messergebnisse sind in Abbildung 5.22 gezeigt.

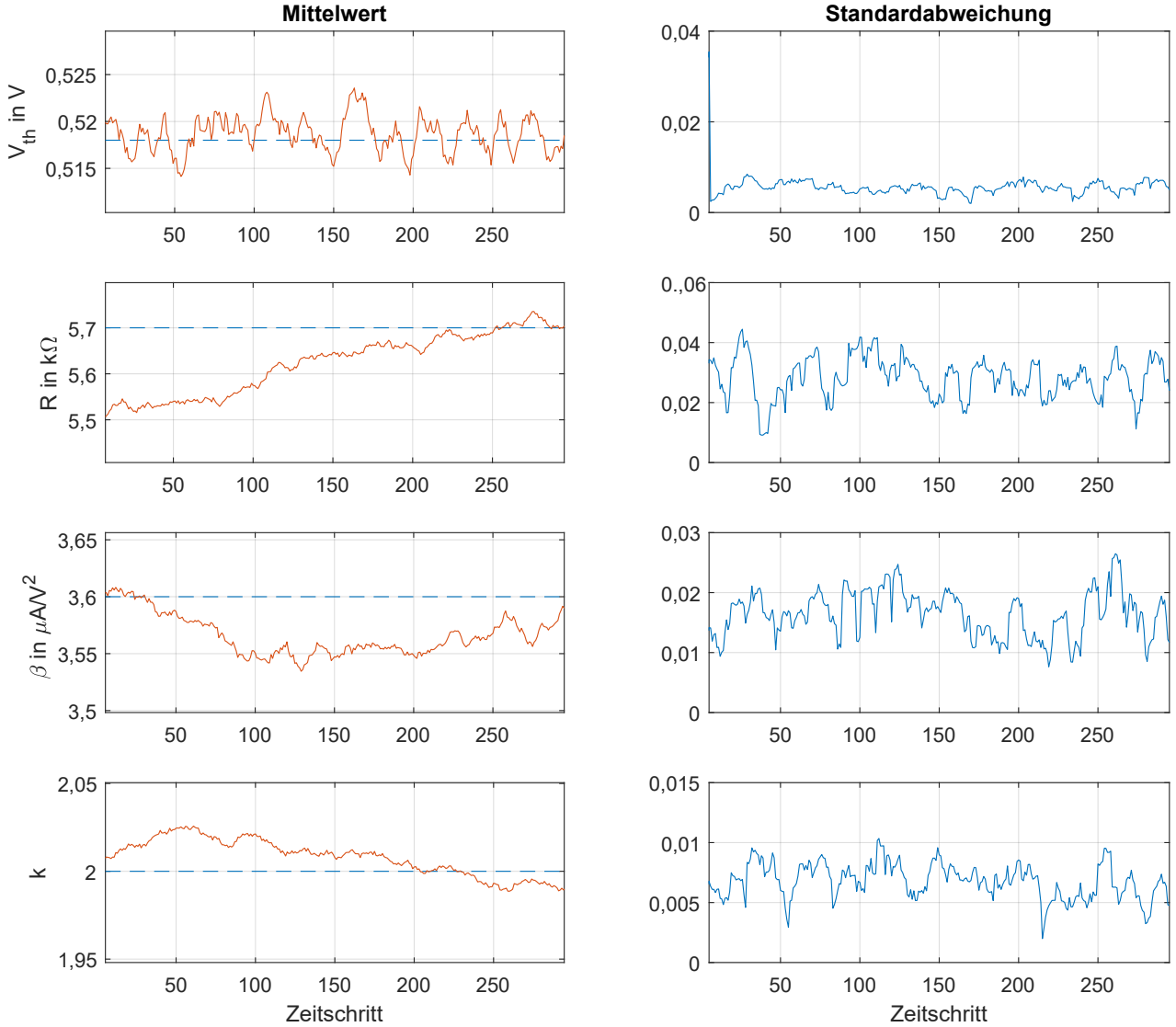


Abbildung 5.21: Mittelwert und Standardabweichung der Schätzung von $\mathbf{x} = [V_{th}, R, \beta, k]$ für den vermessenen Chip 5. Gestrichelte Linien markieren den wahrscheinlichsten Wert.

Die Prozessvariationen führen zu den großen Streuungen zwischen den einzelnen Chips. Zudem ist zu beobachten, dass das Temperaturverhalten über die verschiedenen Schaltungen andere Verläufe zeigt. Zusammen mit den Messungen über der Temperatur kann das Modell um das bisher unbekannte Temperaturverhalten ergänzt werden. Das vollständig parametrierte Modell von Chip 5 zeigt Abbildung 5.23 zusammen mit den Messwerten.

Insgesamt entsteht somit eine schrittweise Methode, um nacheinander einzelne Effekte in ein bestehendes Modell zu integrieren, ohne vorherige Annahmen über die Abhängigkeiten treffen zu müssen. Gleichzeitig können alle durch den Entwurfsprozess vorhandenen Simulationsergebnisse für die Verbesserung der Messergebnisse sowohl für einzelne Transistoren als auch für ganze Schaltungen verwendet werden.

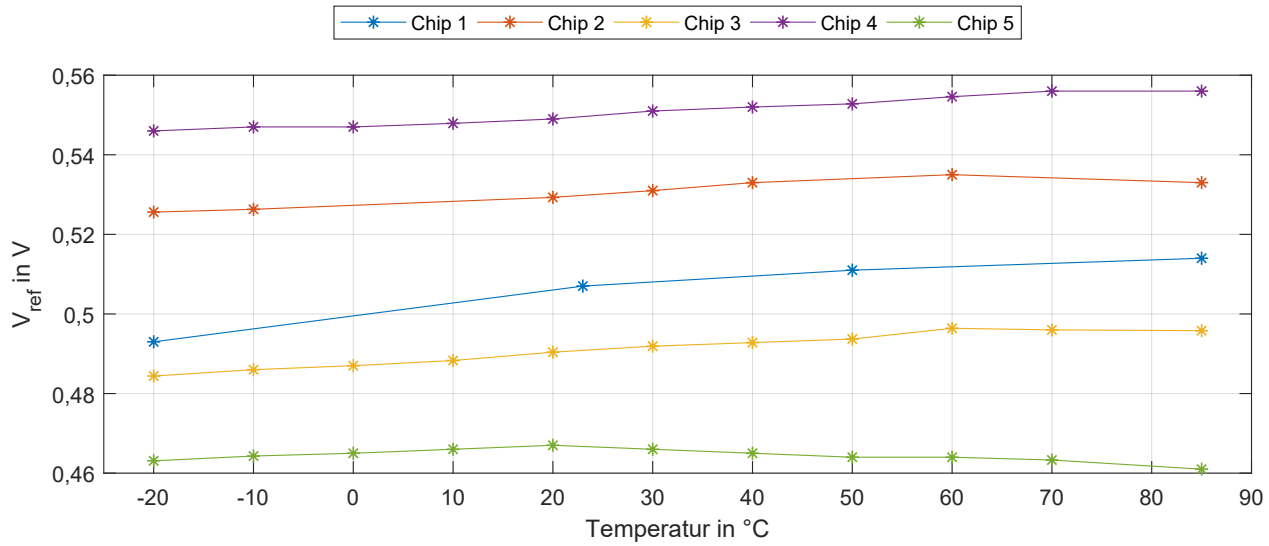


Abbildung 5.22: Gemessenes Temperaturverhalten der Referenzspannung V_{ref} für fünf verschiedene Chips.

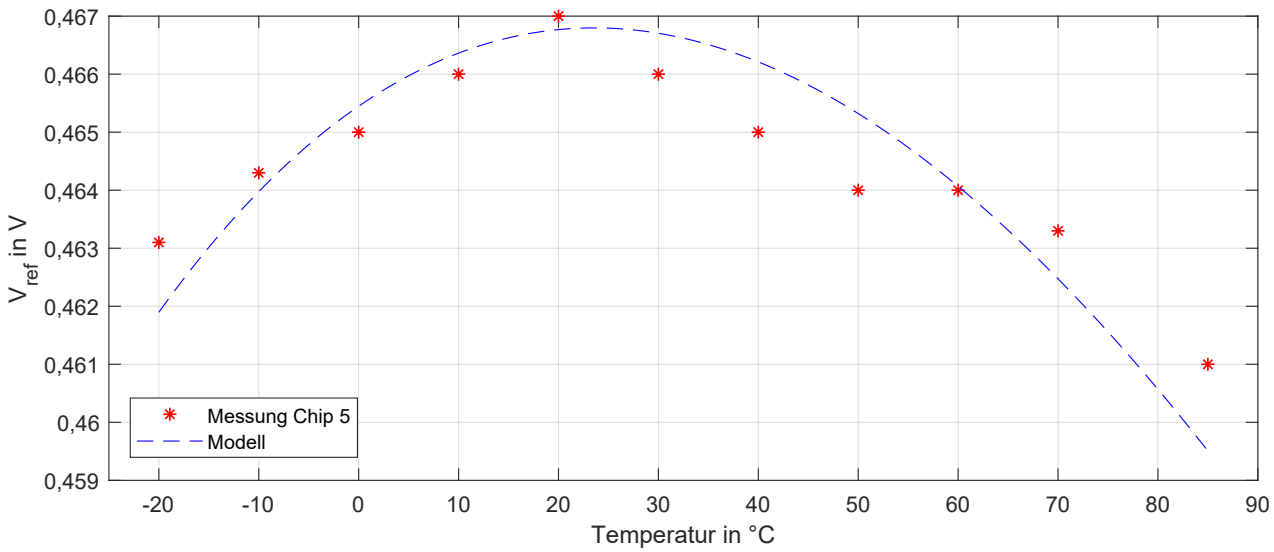


Abbildung 5.23: Parametriertes Temperaturverhalten innerhalb des Verhaltensmodells für Chip 5.

5.2 Beschleunigte Alterung und Vorhersagemöglichkeiten

In diesem Abschnitt werden die Möglichkeiten vorgestellt, die die Kombination aus einer beschleunigten Alterung, jedoch nicht hochgradig beschleunigten Alterungen, in Kombination mit der Parameterextraktion durch den Partikelfilteralgorithmus darstellen. Wie in Abschnitt 2.3.4 beschrieben, werden sämtliche Simulationen und Messungen nur im Bereich der jeweils spezifizierten Betriebsbedingungen der Schaltungen und Bauteile durchgeführt. Somit werden keine Effekte angeregt, die im Betrieb der Schaltung nicht auftreten können. Des Weiteren entfällt die Notwendigkeit der Rückrechnung der beobachteten Veränderung auf den nominellen Betriebsbereich.

Ausgangspunkt für die Schätzung der verbleibenden Lebenszeit sind die Ergebnisse für die Parameterschätzung von Transistoren aus Abschnitt 5.1.2. Für die Evaluierung der Vorhersagemöglichkeiten wird in diesem Beispiel die Schwellspannung durch ein Potenzgesetz

$$\Delta \mathrm{V}_{th} = a \cdot k^b \tag{5.15}$$

für 400 Zeitschritte k modelliert. Der Parameter a modelliert die mögliche Verschiebung der Schwellspannung und b verändert die zeitliche Abhängigkeit. Diese Gleichung eignet sich für die Modellierung der Alterung sowohl für PMOS- als auch für NMOS-Transistoren. Abbildung 5.24 zeigt die Veränderung der Schwellspannung von NMOS- und PMOS-Transistoren über einen Zeitraum von 10.000 Sekunden für verschiedene Kanallängen. Zu erkennen ist die funktionale Abhängigkeit durch ein Potenzgesetz. Die Schwellspannung ändert sich in den ersten Sekunden sehr schnell und danach mit abnehmender Rate. Zu unterscheiden ist, dass die Modellierung mit Gleichung (5.15) nur rein phänomenologisch ist und sich durch die bereits beschriebenen unterschiedlichen Mechanismen nicht zur Modellierung der Ursache eignet. Für die Bestimmung der Remaining Useful Lifetime (RUL) muss jedoch nur eine Funktion gefunden werden, die die beobachtete Veränderung nachbilden kann.

In Abbildung 5.25 ist der zeitliche Verlauf des Mittelwertes und der Standardabweichung der Schätzung der Schwellspannung mit Hilfe des Partikelfilters für $N_s = 150$ gezeigt. Es wurde für die Umgebung $T = 300K$ angenommen. Für die Rauscheinflüsse wurde $\sigma_{I_d} = 10^{-6}$ A als Messrauschen und $\sigma_{V_{th}} = 10^{-4}$ V als Prozessrauschen angenommen. Dies entspricht einem sehr starken Messrauschen. Der Arbeitspunkt des Transistors ist mit $V_{gs} = 0,28$V und $V_{ds} = 0,5$V eingestellt.

In Abbildung 5.26 ist der zeitliche Verlauf des beobachteten Drainstroms I_d während der Vorhersage der verbleibenden Lebensdauer gezeigt. Wie zu erkennen ist, fällt der Strom, entsprechend der steigenden Schwellspannung, zu Beginn stark ab. Danach verschwindet die weitere Änderung nahezu vollständig im Rauschen.

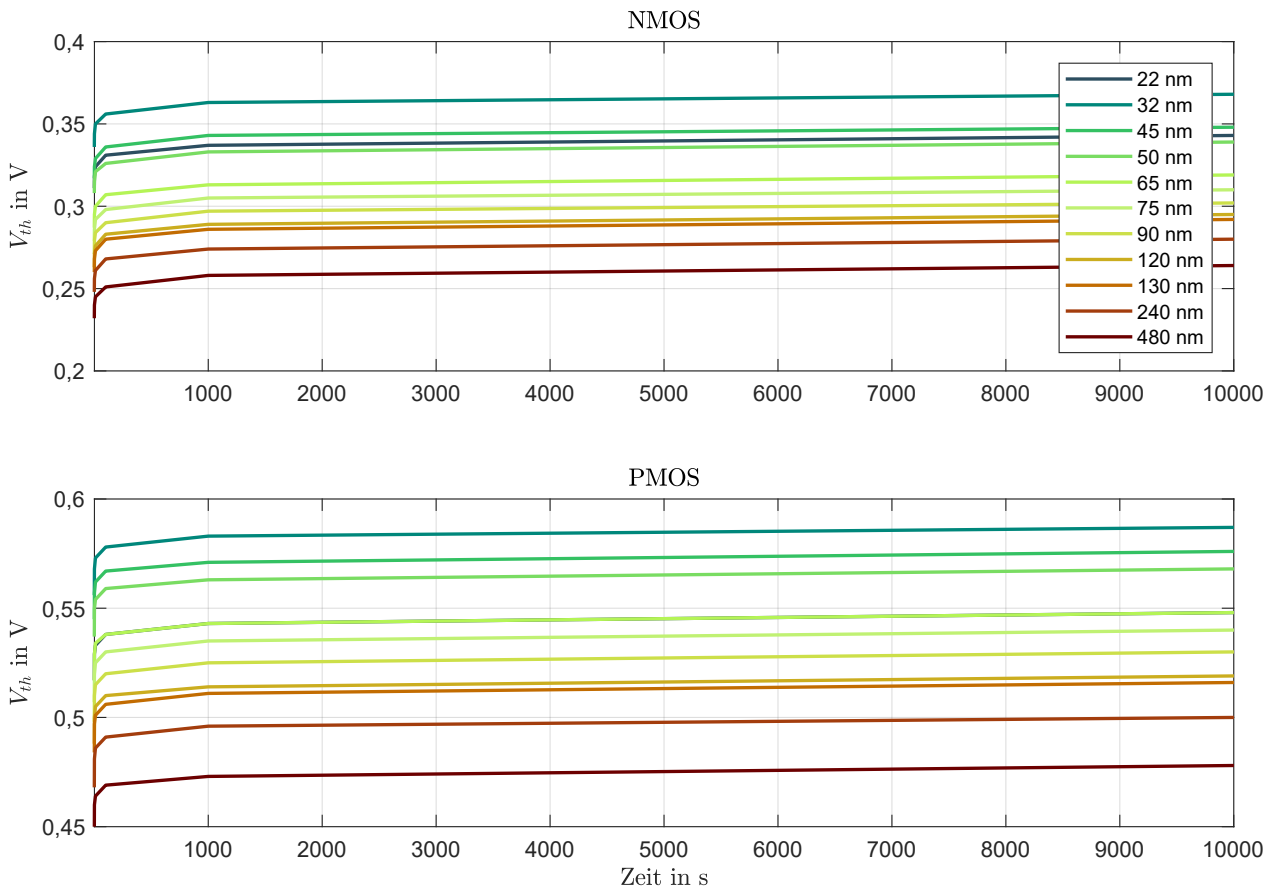


Abbildung 5.24: In TCAD simulierte Veränderung der Schwellspannung durch Alterung von planaren NMOS und PMOS Transistoren über einen Zeitraum von 10.000 Sekunden für verschiedene Kanallängen (W = 1μm).

Während der Messung und der Parameterschätzung durch das Partikelfilter kann in jedem Zeitschritt k eine Schätzung des weiteren Verlaufs durchgeführt werden. Für die Vorhersage wurde die Partikelanzahl des Filter auf $N_s = 150$ erhöht, um zum einen die Rauscheinflüsse besser herausfiltern zu können und zum anderen eine bessere Vorhersage berechnen zu können. Durch die Verwendung des Partikelfilters stehen für die Schätzung der Modellparameter a und b $i = 1, \ldots, k$ und in jedem Zeitschritt $j = 1, \ldots, N_s$ Messpunkte zur Verfügung. Somit besteht die Möglichkeit, die Mittelwerte zuerst über die Messwerte der Partikel oder über die Zeitschritte zu berechnen, da die Schätzungen, die auf einzelnen Partikeln beruhen sehr starken Schwankungen unterworfen sind, wie in Abbildung 5.27 gezeigt ist. Die jeweilige Schätzung ist mit den Messwerten bis zum jeweiligen Zeitschritt k durchgeführt. Wie zu erkennen ist, variiert die Schätzung zwischen den Partikeln für kurze Beobachtungszeiträume besonders stark.

Daher werden für die Schätzung in jedem Zeitschritt der Mittelwert der Veränderung der Schwellspannung über die einzelnen Partikel berechnet und anschließend über die Zeitschritte k die Parameter a und b bestimmt. Für die tatsächliche Bestimmung

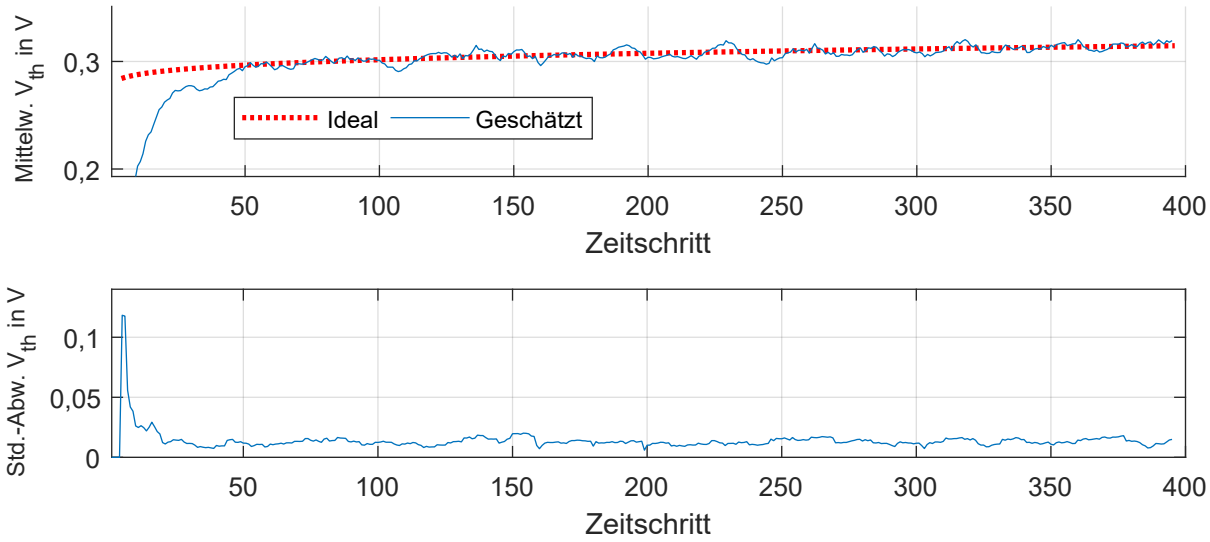


Abbildung 5.25: Zeitlicher Verlauf des Mittelwertes und der Standardabweichung der Schätzung der Schwellspannung mit Hilfe des Partikelfilters für $N_s = 150$ über 400 Zeitschritte k.

wird das, in MATLAB implementierte, Broyden-Fletcher-Goldfarb-Shanno (BFGS) Verfahren [Bro70b, Bro70a, Fle70, Gol70, Sha70] verwendet. Dieses Verfahren ist ein Quasi-Newton-Verfahren um iterativ nicht-lineare Optimierungsprobleme zu lösen. Anstelle der ersten Ableitung wird die zweite Ableitung bzw. die Hesse-Matrix nummerisch approximiert. Abbildung 5.28 zeigt den zeitlichen Verlauf der Schätzung der Schwellspannung ausgehend von verschiedenen Zeitpunkten k.

Für die Vorhersage zum Zeitpunkt $k = 20$ ist zu erkennen, dass der vorhergesagte Verlauf sehr stark vom tatsächlichen Verhalten der Schwellspannung abweicht. Dies ist dadurch begründet, dass wie in Abbildung 5.25 gezeigt, das Partikelfilter erst zum Zeitschritt $k = 50$ eingeschwungen ist. Die Ursache dafür ist der hohe Rauscheinfluss. Somit ist die Vorhersage durch das Filterverhalten verfälscht. Für die Schätzung zum Zeitpunkt $k = 100$ ist der Verlauf schon sehr nahe am tatsächlichen Verlauf der Schwellspannung, es besteht jedoch eine Überschätzung der Alterung. Für die Berechnung der Vorhersage zum Zeitpunkt $k = 300$, was einer Vorhersage von 100 Messzeitpunkten entspricht, beträgt die maximale Vorhersageabweichung nur 4mV. Die Messzeitpunkte müssen für die Berechnung nicht äquidistant sein. Somit können zu Beginn der Messung die ersten Messungen zeitlich kompakt durchgeführt werden, damit das Filter zeitlich schnell einschwingt. Anschließend können die Abstände der Messpunkte logarithmisch gewählt werden. Somit entspräche die Schätzung zum Zeitpunkt $k = 300$ einer deutlich geringeren Zeitspanne als die der folgenden 100 Schritte.

Werden die Ergebnisse bzw. Vorhersagen von verschiedenen Transistoren miteinander kombiniert, ergibt sich für jedes Bauteil ein individueller Verlauf der Schwellspannung über der Zeit, dies ist in Abbildung 5.29 gezeigt. Die initiale Variabilität ist allein bedingt durch die Prozessvariationen. Über die Alterung steigt die Variabilität der Stichprobe

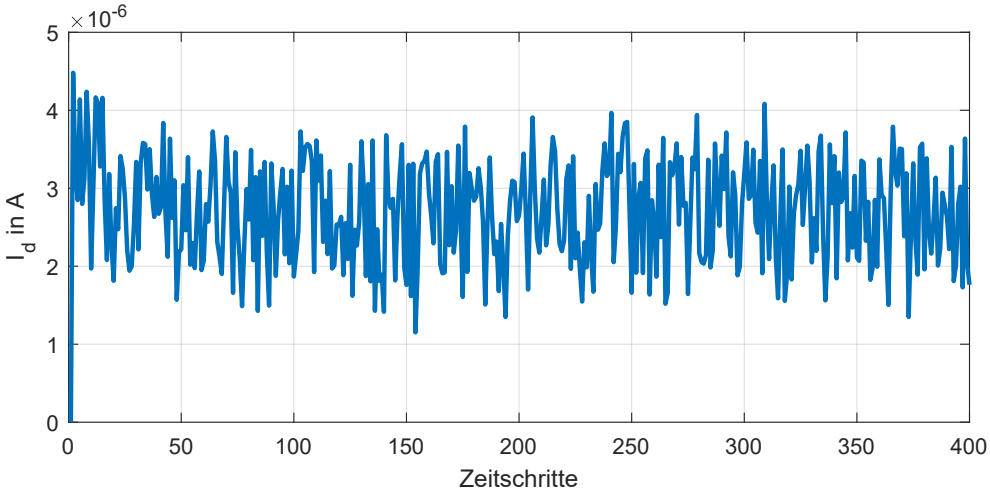


Abbildung 5.26: Zeitlicher Verlauf des beobachteten Drainstroms während der Vorhersage der verbleibenden Lebensdauer.

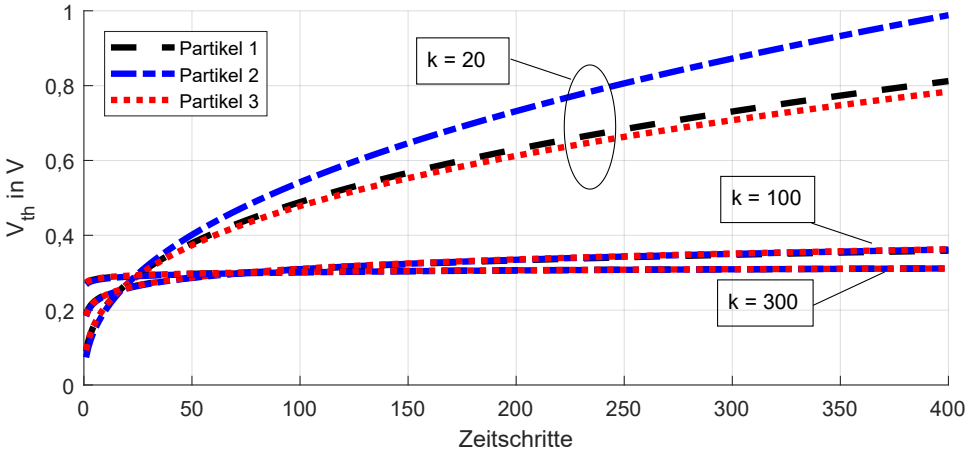


Abbildung 5.27: Schätzung der weiteren Änderung der Schwellspannung für drei verschiedene Partikel zu unterschiedlichen Zeitschritten k.

an. Die Ursache dafür ist, dass die Stressbedingungen selbst von den Prozessvariationen abhängen. Somit zeigen die Transistoren unterschiedliche zeitliche Abhängigkeiten der Alterung.

Für den gesamten Algorithmus und die Berechnung der Vorhersage können die Partikel durch wirkliche Messungen von unterschiedlichen Transistoren verwendet werden. Das Resultat wäre, falls Messungen miteinander kombiniert werden, ein im Mittel zutreffendes Alterungsmodell. Ebenfalls können bestehende Modelle einfach im Nachhinein durch weitere Messungen verbessert werden. Für Langzeitmessungen bedeutet dies, dass Messungen von Bauteilen, die während der Messung ausfallen, bis zu diesem Zeitpunkt unproblematisch zur Modellerstellung verwendet werden können. Zudem besitzt die Vorhersage durch das Partikelfilter die Möglichkeit, auf sich ändernde Umweltbedingun-

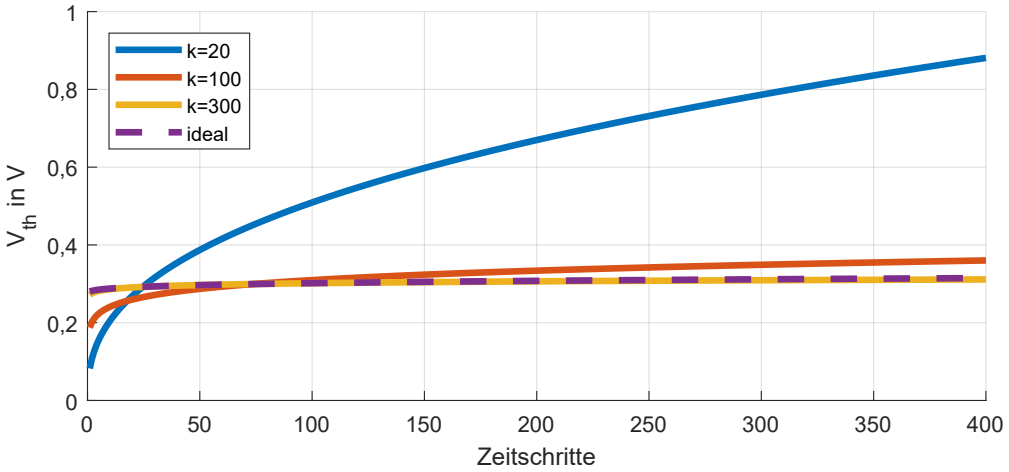


Abbildung 5.28: Zeitlicher Verlauf der Schätzung der Schwellspannung ausgehend von verschiedenen Zeitpunkten k.

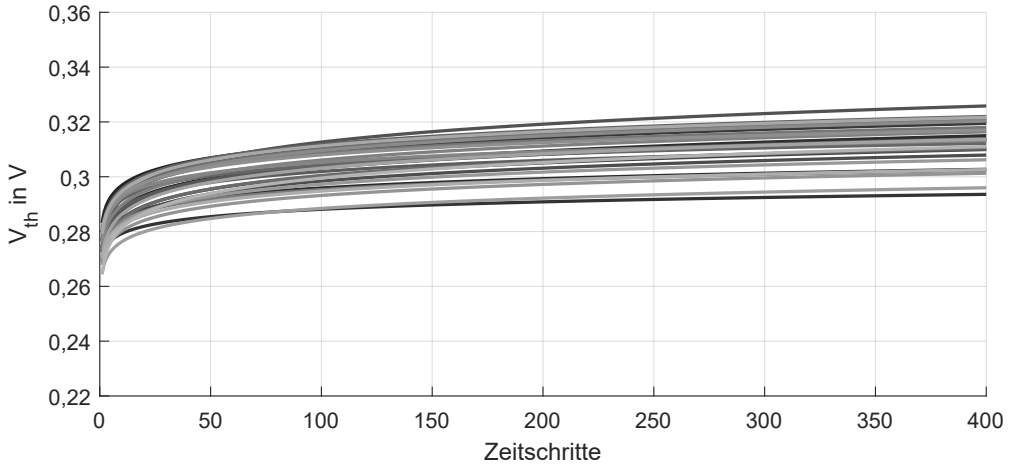


Abbildung 5.29: Vorhersage des Verlaufs der Schwellspannung zum Zeitpunkt $k = 300$ für 30 gleiche Transistoren die von Prozessvariationen betroffen sind.

gen wie Temperatur und Spannungsschwankungen zu reagieren, ohne das die Messung erneut begonnen werden muss.
Sowohl für die Erstellung von Alterungsmodellen als auch für die Berechnung der verbleibenden Lebensdauer ist es notwendig, dass die Verteilungsfunktionen der sich ändernden Parameter in jedem Zeitschritt bekannt sind. Dadurch kann für Modellierungszwecke zwischen der initialen Variabilität und der durch Alterung induzierten unterschieden werden. In Abbildung 5.30 sind Verteilungen der Schwellspannung für verschiedene Zeitpunkte von 150 von Prozessvariationen betroffene Transistoren gezeigt.
Wie zu erwarten ist, steigt der Mittelwert der Schwellspannung bedingt durch die Alterung über der Zeit. In diesem Beispiel steigt die Standardabweichung der Schwellspannung von 4mV auf 5mV an. Die Differenz kann nur durch die Alterung induziert worden sein und kann anschließend getrennt von den Prozessvariationen modelliert werden. Um die zeitabhängige Fehlerrate $\lambda(t)$, gemäß Gleichung (2.59), berechnen zu

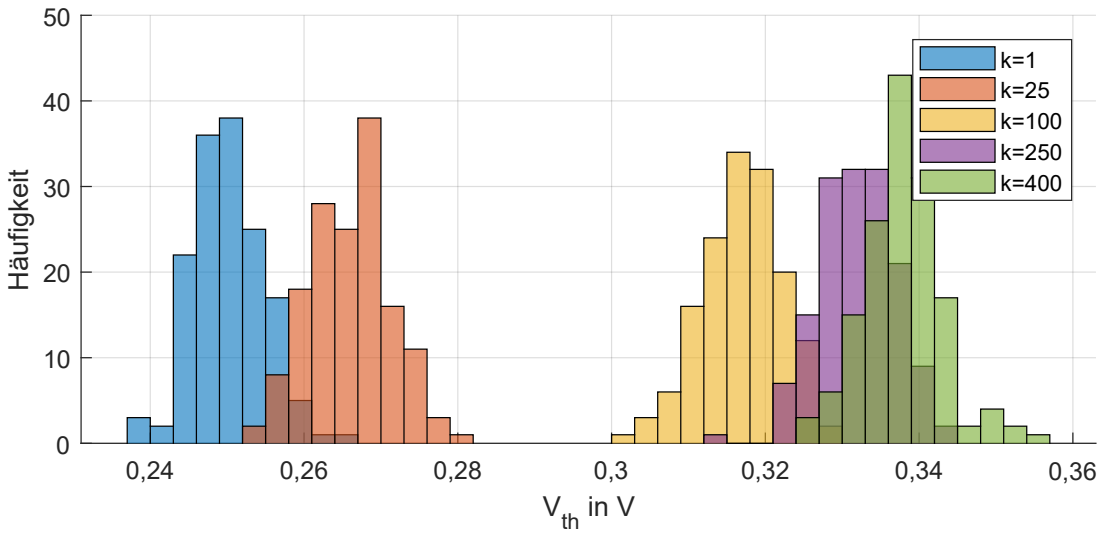


Abbildung 5.30: Verteilungen der Schwellspannung für verschiedene Zeitpunkte von 150 von Prozessvariationen betroffene Transistoren.

können, müssen ebenfalls die Verteilungen der spezifizierten Parameter bekannt sein. Als Fehlerschwelle wird in diesem Beispiel eine Schwellspannung von $V_{th} = 350\,\mathrm{mV}$ gewählt. Die Wahl der Fehlerschwelle richtet sich selbstverständlich nach der jeweiligen Anwendung und Spezifikation. Wie aus den Quantil-Quantil-Diagrammen in Abbildung 5.31 zu erkennen ist, entspricht die Verteilung besonders in Bezug auf die Ausläufer der Verteilungen weder einer Normalverteilung noch einer Weibullverteilung.

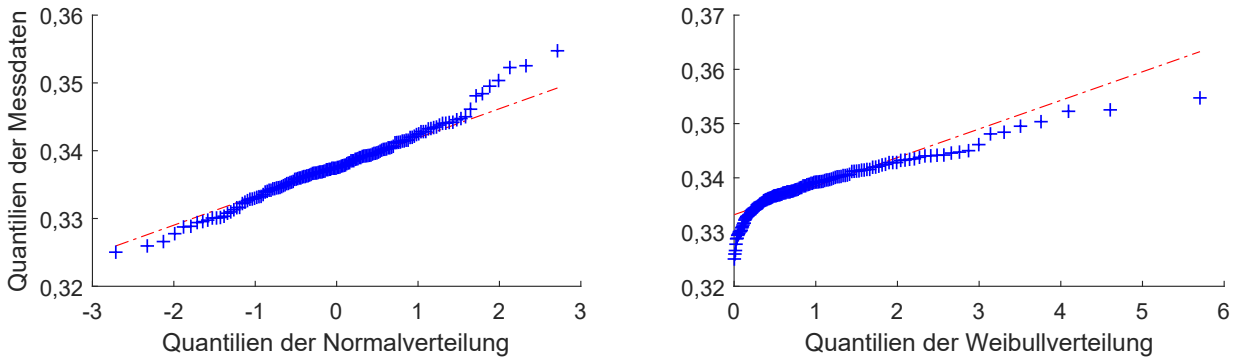


Abbildung 5.31: Quantil-Quantil-Diagramme für Normal- und Weibullverteilung der Messungen der Schwellspannung zum Zeitpunkt $k = 400$.

Die Ausläufer der Verteilungsfunktionen sind für die Bestimmung von $\lambda(t)$ von hoher Wichtigkeit. Bei diesen Transistoren handelt es sich um solche, die besonders stark gealtert sind, und somit ist die Ausfallrate höher als mit einer Normalverteilungsannahme. Diese würde etwa durch einen Kalman-Filter berechnet werden [CBTD01]. Für die tatsächliche Verteilungsfunktion ergibt sich eine Fehlerrate

$$\lambda_{\mathrm{mess}} = 0,0267 \tag{5.16}$$

und für die Fehlerrate unter einer Normalverteilungsannahme

$$\lambda_{\text{norm}} = 0,0084 \ . \tag{5.17}$$

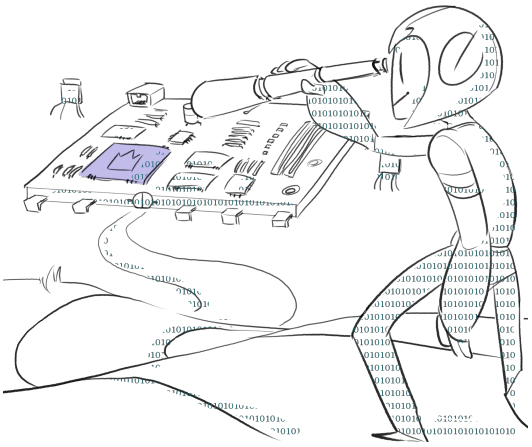
Somit beträgt die echte Fehlerrate das Dreifache der Fehlerrate die durch die Vereinfachung der Verteilungsfunktion berechnet worden wäre. Dies zeigt gleichzeitig die Möglichkeiten der entwickelten Methode als auch deren Notwendigkeit.

5.3 Zusammenfassung

In diesem Kapitel wurde gezeigt, dass Messmethoden, die einzelne Modellgleichungen verwenden, um Sekundäreffekte zu unterdrücken, die besten Messergebnisse für einzelne Transistorparameter liefern. Ausgehend von diesem Ergebnis wurde mit Hilfe eines Partikelfilter-Algorithmus gezeigt, dass dieses Vorgehen auf Verhaltensmodelle von Bauelementen und Schaltungen anwendbar ist. Dadurch entsteht eine Vielzahl neuer Mess- und Beobachtungsmöglichkeiten für die Bestimmung der Auswirkungen durch PVTA-Einflüsse auf integrierte Schaltungen. Des Weiteren wurde gezeigt, dass die so gewonnenen Erkenntnisse verwendet werden können, um eine Vorhersage der zeitabhängigen Ausbeute von Bauelementen und Schaltungen zu treffen. Ein Alleinstellungsmerkmal ist, dass innerhalb der Methode beliebige Anwendungs- und Belastungsszenarien analysiert werden können. Zudem müssen keine Annahmen über die Abhängigkeiten bisher nicht charakterisierter Phänomene getroffen werden.

6 Zusammenfassung und Ausblick

» The humble man makes room for progress. The proud man believes he is already there «



Im Rahmen dieser Arbeit wurde ein modifiziertes BSIM-BULK-Transistormodell erstellt, das in der Lage ist, Alterungseinflüsse und Effekte durch Strahlung in einem industriellen Entwurfsablauf auf jeder Stufe des analogen Entwurfs zu berücksichtigen. Sowohl Alterung als auch Strahlung wurden auf Basis der Änderung der Ladungsverteilung innerhalb der MOS-Struktur modelliert. Somit dient die Schwellspannung nicht als relevanter Modellparameter und kann als reine Indikatorgröße aufgefasst werden. Entgegen den üblichen Ansätzen können mit diesem Transistormodell transiente Simulationen über die vollständige Lebensdauer eines Bauteils oder einer Schaltung durchgeführt werden. Dadurch müssen keine Extrapolationen der Parameteränderungen berechnet werden. Somit können insbesondere die Wechselwirkungen zwischen diesen Effekten betrachtet werden und Worst-Case-Abschätzungen sind nicht mehr notwendig. Das Transistormodell kann konsistent in den gängigsten Entwurfsmethoden wie der $g_\mathrm{m}/I_\mathrm{D}$-Methode verwendet werden. Mit den Ergebnissen aus den Entwurfsprozessen ist ein Schaltungsentwurf möglich, das resistent gegenüber Alterungseinflüssen ist. Des Weiteren können bestehende Entwürfe einer Sensitivitätsanalyse im Hinblick auf Alterungseinflüsse unterzogen werden. Als Resultat können sämtliche, für die Schaltung

relevanten, Eigenschaften, wie Verstärkung der Bandbreite, über die gesamte spezifizierte Lebensdauer bereitgestellt.

Da die transiente Analyse von großen analogen Schaltungen mit einem hohen Simulationsaufwand einhergeht, wurde eine Funktionsblock-basierte Topologieanalyse für analoge Schaltungen entwickelt. Diese Methode basiert auf grundlegenden analogen Blöcken mit einer jeweilig eindeutigen Funktion innerhalb einer Schaltung, zum Beispiel zur Bereitstellung von Biaspunkten oder zur Signalverstärkung. Diese Blöcke werden mit dem entwickelten Modell sehr genau in Bezug auf Prozessvariationen, Temperaturverhalten und Alterung analysiert. Mit diesen Daten können qualitative Aussagen über die jeweilige, aus den Blöcken konstruierte Schaltungstopologie getroffen werden. Es wurden zwei Fallstudien für Biasquellen und Verstärker durchgeführt, um die Methode zu evaluieren. Eine alterungsresistenteste Verstärkertopologie wurde anschließend mit einer passenden Biasquelle in einer 65nm-CMOS-Technologie gefertigt. Die Schaltung diente anschließend als Messobjekt.

Neben der simulationsbasierten Analyse und des Entwurfs von Bauteilen und Schaltungen wurden in der vorliegenden Arbeit ein Messplatz und eine Methode für die Alterungsanalyse von Schaltungen und Bauteilen entwickelt. Die Methode basiert auf einem Partikelfilteralgorithmus, der mit Hilfe eines Zustandsraummodells des jeweiligen Testobjekts in der Lage ist, Umwelteinflüsse aus den eigentlichen Messungen herauszurechnen. Somit können einzelne Parameter von besonderem Interesse, wie die Schwellspannung, deutlich genauer bestimmt werden als mit anderen Methoden. Des Weiteren besteht die Möglichkeit, mehrere Parameter mit nur einer Messung zu schätzen. Dies geht mit einer leicht erhöhten Ungenauigkeit einher, verbindet jedoch alle Informationen, die aus dem Entwurf der Schaltung stammen, mit den jeweiligen Messdaten. Somit können sämtliche Simulationsergebnisse zur Verbesserung von Messungen herangezogen werden. Zudem existiert am Ende jeder Messung ein parametriertes Modell der Schaltung oder des Bauteils. Darüber hinaus kann zu jedem Messzeitpunkt eine Schätzung der verbleibenden Lebensdauer berechnet werden.

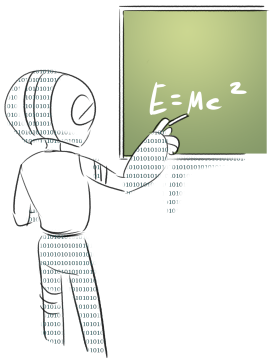
Welche Anknüpfungspunkte existieren? Aus der vorliegenden Arbeit können insgesamt drei verschiedene Linien weiter verfolgt werden. Das REL-MOS-Transistormodell kann von seinem prädiktiven Status aus weiter entwickelt werden. Dazu könnte in TCAD eine prädiktive Technologie zusammen mit Alterung untersucht werden, so wären im Modell die Alterung und das nominelle Verhalten vom gleichen Ursprung. Eine weitaus aufwendigere Möglichkeit wäre es, dies mit Messungen zu tun. Die Herangehensweise der Modellierung von Alterung und Strahlung könnte zudem auf andere Transistormodelle wie das in der Industrie weit verbreite BSIM4-Modell übertragen werden oder auf Modelle ganz anderer Transistorarten wie FinFET oder FDSOI.

Neben der Verbesserung oder Erweiterung könnten für den Entwurfsprozess von Digital- oder Hochfrequenzschaltungen Analysen durchgeführt werden. Besonders für den digitalen Bereich können diese Analysen, wie bereits einige Vorarbeiten zeigen, viele Vorteile

durch die konsistente Betrachtung aller Einflüsse während des Entwurfs haben. Für die Bewertung der Alterung ist besonders der kritische Pfad einer digitalen Schaltung von Interesse. Um diesen zum Beispiel durch Monitorstrukturen während der Lebensdauer der Schaltung zu überwachen, müssen neben dem nominellen kritischen Pfad die alterungskritischen Pfade zum Entwurfszeitpunkt auffindbar sein. Genau an dieser Stelle können die im Rahmen dieser Arbeit entwickelten Modelle eingesetzt werden. Denn diese sind in der Lage, alle Wechselwirkungen von Prozessvariationen, Alterung und Belastungsprofilen zu analysieren. Für Hochfrequenzschaltungen sind zum Teil andere Entwurfsparameter von Interesse als für analoge Schaltungen. Diese könnten, ähnlich wie die in dieser Arbeit vorgestellten Methoden, untersucht werden und somit zum Beispiel eine alterungstolerante Impedanzanpassung realisiert werden.
Des Weiteren können die Modelle dazu verwendet werden, um die Effekte von Strahlung auf digitale Gatter und Speicherelemente zu untersuchen. Die Ergebnisse können durch weitere Modellierungsschritte für ganze digitale Systeme auf höheren Abstraktionsebenen verwendet werden. Erste Arbeiten dazu wurden bereits durchgeführt. Diese können auf weitere Elemente und umfassendere Analysen ausgedehnt werden. Besonders interessant sind in diesem Zusammenhang die Wechselwirkung von Alterung und Strahlung auf Speicherelemente wie SRAM und Flash-Speicher.
Als Weiteres kann die gefertigte Schaltung umfassend auf Alterungseinflüsse untersucht werden und somit die in dieser Arbeit vorgestellte Entwurfsmethodik nicht nur simulativ durch Fallstudien, sondern auch messtechnisch verifiziert werden. Der dazu nötige Messaufbau und Methode wurden im Rahmen dieser Arbeit entwickelt. Jedoch standen für die Analyse der Verstärkerschaltung nicht alle notwendigen Messinstrumente für einen teilautomatisierten Messablauf zur Verfügung. Der Messplatz sowie der Extraktionsablauf basierend auf dem Partikelfilter können auf weitere Schaltungen erweitert werden.
Schlussendlich kann die entwickelte Messmethode erweitert werden. Dadurch, dass die Extraktionsmethode mit dem Partikelfilter, entgegen anderer Filteralgorithmen, keinen Einfluss auf den Frequenzbereich nimmt, kommt es insbesondere zu keiner Tiefpassfilterung und Einschwingvorgänge und Startup-Verhalten verschiedener Schaltungen können untersucht werden. Darüber hinaus kann der Algorithmus als solcher evaluiert und verbessert werden. Die verschiedenen Arten von Implementierungen von Partikel-Filter-Algorithmen können für einzelne Anwendungen Vorteile bieten. Zudem kann die Messung mit geringem Stichprobenumfang weiter untersucht werden.

A Rechenvorschriften der Transistormodelle

» If it seems easy,
you're doing it wrong.«



In diesem Kapitel werden die zusätzlichen, zur Berechnung der Transistormodelle nötigen, Gleichungen aus dem Abschnitt 2.2.1 näher beschrieben. Dabei werden die tatsächlich zur Berechnung des Ladungsträger-basierten (BSIM6 [Aga14]) und des Oberflächenpotential-basierten (UTSOI [SSKvdT13]) Transistormodells beschrieben. Da diese Gleichungen in jedem Simulationsschritt berechnet werden müssen, geben sie einen Eindruck vom Rechenaufwand des jeweiligen Modells.

A.1 Berechnung des Oberflächenpotentials

Bei dem verwendeten Modell wird angenommen, dass sich das Backinterface, also die Grenzfläche zwischen Buried Oxide und Substrat, immer in Verarmung befindet. Somit gilt $V_{\mathrm{sb}} > V_{\mathrm{FB}}$, wobei V_{sb} die Spannung zwischen Substrat und Buried Oxide und V_{FB} die Flachbandspannung ist. Dadurch ist es nicht möglich, den FDSOI Transistor als

Double-Gate Transistor zu betreiben. Für das Backsurface Potential ϕ_{sb} ergibt sich somit [RJPB12]:

$$\phi_{\mathrm{sb}} = \alpha_{\mathrm{c}} \cdot \phi_{\mathrm{sf}} + \varepsilon \tag{A.1}$$

$$\alpha_{\mathrm{c}} = \frac{C_{\mathrm{si}}}{C_{\mathrm{si}} + C_{\mathrm{box}}} \tag{A.2}$$

$$\varepsilon = \frac{C_{\mathrm{box}}}{C_{\mathrm{si}} + C_{\mathrm{box}}} \cdot V_{\mathrm{sub}}. \tag{A.3}$$

Um Gleichung (2.14) zu lösen, wird der folgende Formalismus eingeführt [RJPB12, CG01]:

$$x = \frac{\phi_{\mathrm{sf}}}{V_{\mathrm{T}}} \tag{A.4}$$

$$x_{\mathrm{gf}} = \frac{V_{\mathrm{g}} - V_{\mathrm{FB}}}{V_{\mathrm{T}}} \tag{A.5}$$

$$x_{\mathrm{gb}} = \frac{V_{\mathrm{sub}} - V_{\mathrm{FBb}}}{V_{\mathrm{T}}} \tag{A.6}$$

$$x_{\mathrm{n}} = \frac{V_{\mathrm{c}}}{V_{\mathrm{T}}} \tag{A.7}$$

$$\varepsilon' = \frac{C_{\mathrm{box}}}{C_{\mathrm{si}} + C_{\mathrm{box}}} \cdot x_{\mathrm{gb}} \tag{A.8}$$

$$G = G_{\mathrm{f}} = \frac{\gamma}{\sqrt{V_{\mathrm{T}}}}. \tag{A.9}$$

Das Einsetzen in Gleichung (2.14) ergibt:

$$(x_{\mathrm{gf}} - x)^2 - \frac{C_{\mathrm{box}}^2}{C_{\mathrm{ox}}'^2} \cdot (x_{\mathrm{gb}} - \alpha_{\mathrm{c}} \cdot x - \varepsilon')^2 = G^2 \cdot \left[\exp(x - x_{\mathrm{n}}) - \exp(\alpha_{\mathrm{c}} \cdot x + \varepsilon' - x_{\mathrm{n}})\right]. \tag{A.10}$$

Die implizite Gleichung (A.10) muss nach der Unbekannten x aufgelöst werden, was nach Gleichung (A.4) das Oberflächenpotential ϕ_{sf} im Kanal, normiert auf die Temperaturspannung V_{T}, darstellt. Daher wird im Folgendem x als Synonym für das Oberflächenpotential genutzt.

A.2 Oberflächenpotential am Source-Anschluss

Die iterative Lösung von Gleichung (A.10) ist in Algorithmus 4 dargestellt. Die genutzten Hilfsfunktionen sind in Anhang A.4 definiert.
Der Ausgangspunkt für Algorithmus 4 ist Gleichung (A.10) und die Annahme, dass $x_{\mathrm{s}} = x_{i0} + \eta$ ist. Dabei liefert $\eta(\cdot)$ eine erste Näherung für das Oberflächenpotential. Der zusätzliche Term x_{i0} ist ein Korrekturterm, der die Genauigkeit der Berechnung von x_{s}

erhöht. Ausgang des Algorithmus ist die gesuchte Größe x_s, das Oberflächenpotential auf der Source-Seite des Kanals.

Im nächsten Schritt werden wichtige Hilfsgrößen bestimmt, die im weiteren Verlauf des Algorithmus gebraucht werden. x_1 stellt dabei einen Startwert für das Oberflächenpotential dar, welcher sich ergibt, wenn man den Maschenumlauf auf den Querschnitt aus Abbildung 2.8 anwendet. Die Größen aus Zeile 3 folgen den Gleichungen (A.4)-(A.9).

Die Berechnung des Oberflächenpotentials wird für zwei Fälle betrachtet. Der erste Fall liegt vor, wenn $x_\mathrm{gf} > 0$ ist. Aus Gleichung A.5 ergibt sich daher, dass die Spannung am Gate größer als die Flachbandspannung zwischen Gate und Kanal ($V_\mathrm{g} > V_\mathrm{FB}$) sein muss.

In Zeile fünf wird die Funktion $\eta(\cdot)$ berechnet, welche eine erste Näherung für das Oberflächenpotential darstellt. In den Zeilen 6-8 werden die Hilfsvariablen a, c und τ berechnet. Dabei gilt für $x_{\ell\mathrm{j}}$

$$x_{\ell 0} = \frac{1}{G_\mathrm{f}^2 \cdot k_{0,0}} \cdot \left[x_\mathrm{gf}^2 - \frac{C_\mathrm{box}^2}{C_\mathrm{ox}'^2} \cdot (x_\mathrm{gb} - \varepsilon)^2 \right] \tag{A.11}$$

$$x_{\ell 1} = -\frac{2}{G_\mathrm{f}^2 \cdot k_{0,0}} \cdot \left[x_\mathrm{gf} - \frac{C_\mathrm{box}^2}{C_\mathrm{ox}'^2} \cdot \alpha_\mathrm{c} \left(x_\mathrm{gb} - \varepsilon \right) \right] \tag{A.12}$$

$$x_{\ell 2} = \frac{1}{G_\mathrm{f}^2 \cdot k_{0,0}} \cdot \left[1 - \frac{C_\mathrm{box}^2}{C_\mathrm{ox}'^2} \cdot \alpha_\mathrm{c}^2 \right] \tag{A.13}$$

$$k_{0,0} = 1 - \exp\left[(\alpha_\mathrm{c} - 1) \cdot x_\mathrm{vtf0} + \varepsilon \right]. \tag{A.14}$$

Aus diesen Hilfsvariablen und der Funktion $\sigma(\cdot)$ (siehe Anhang A.4) lässt sich nun eine erste Schätzung für den Korrekturterm angeben. Um eine noch größere Genauigkeit für das Oberflächenpotential zu erreichen, wird in mehreren Schleifendurchläufen der Korrekturterm und die Hilfsvariablen Δ_i bzw. k_i neu berechnet. Als Abbruchkriterium der Schleife findet man in der Literatur sowohl eine Maximalanzahl als auch eine zu erreichende Genauigkeit. Zusätzlich werden die Funktionen p_i und q_i bestimmt, welche als Hilfsfunktionen Bestandteil des Oberflächenpotentials sind.

In Zeile 18 wird die Schätzung des Korrekturterms $x_{(i+1)0}$ und die Berechnung der Funktion $\eta(\cdot)$ der Eingangsfunktion $x_\mathrm{s} = x_{j0} + \eta$ übergeben.

Für den Fall, dass $x_\mathrm{gf} \leq 0$ ist, ergibt sich das Oberflächenpotential zu dem Initialisierungswert x_1. Somit ist das Oberflächenpotential auf der Source-Seite des Kanals eindeutig bestimmt.

Algorithmus 4 Berechnung des Oberflächenpotentials x_s auf der Source-Seite des Kanals

Eingabe: $x_\mathrm{s} = x_{j0} + \eta$; Eq. (A.10)
Ausgabe: x_s

1: **Initialisierung:**
2: $x_1 = \dfrac{C_\mathrm{box} \cdot C_\mathrm{si}}{C'^2_\mathrm{ox} \cdot C_\mathrm{box} + C_\mathrm{si} \cdot C_{box} \cdot C_{ox} \cdot C_{si}} \cdot (x_\mathrm{gb} - x_\mathrm{gf}) + x_\mathrm{gf}$ ▷ Berechne Startwert für x_s
3: G_f^2, x_gb, x_gf, x_vtf0 ▷ Berechne Gleichung (A.4)-(A.9)
4: **if** $x_\mathrm{gf} > 0$ **then**
5: $\eta = MINFCT(x_1, x_\mathrm{vtf0}, 15) - \dfrac{1}{2} \cdot \left(x_\mathrm{vtf0} - \sqrt{x_\mathrm{vtf0}^2 + 5}\right)$ ▷ Schätzung für x_s (Anhang A.4)
6: $a = x_{\ell 0} + x_{\ell 1} \cdot \eta + x_{\ell 2} \cdot \eta^2$ ▷ Berechne Hilfsvariablen a,c,τ
7: $c = -x_{\ell 1} - 2 \cdot x_{\ell 2} \cdot \eta$
8: $\tau = -\eta + \ln(a)$
9: $x_{00} \approx \sigma(a, x_{\ell 2}, c, \tau, \eta)$ ▷ Schätzung für x_{j0}
10: **for** $i = 1, \ldots, i_\mathrm{max}$ **do** ▷ Verbesserung der Schätzung des Korrekturterms x_{j0}
11: $i = i - 1$
12: $\Delta_{i0} = \exp(x_{i0})$ ▷ Berechne Hilfsvariable Δ_{i0}, k_{i0}
13: $k_{i0} = \exp((\alpha_\mathrm{c} - 1) \cdot x_{i0} + \varepsilon)$
14: $p_i = 2 \cdot (x_\mathrm{gf} - x_{i0}) - 2 \cdot \alpha_\mathrm{c} \cdot \dfrac{C_\mathrm{box}^2}{C'^2_\mathrm{ox}} \cdot (x_\mathrm{gb} - \alpha_\mathrm{c} \cdot x_{i0} - \varepsilon) + G_\mathrm{f}^2 \cdot \Delta_{i0} \cdot (1 - k_{i0} \cdot \alpha_\mathrm{c})$
15: $q_i = (x_\mathrm{gf} - x_{i0})^2 - \dfrac{C_\mathrm{box}^2}{C'^2_\mathrm{ox}} \cdot (x_\mathrm{gb} - \alpha_\mathrm{c} \cdot x_{i0} - \varepsilon)^2 + G_\mathrm{f}^2 \cdot \Delta_{i0} \cdot (1 - k_{i0})$
16: $x_{(i+1)0} = x_{i0} + 2q_i \left[p_i + \sqrt{p_i^2 - 2q_i \cdot \left(2\left(1 - \dfrac{C_\mathrm{box}^2}{C'^2_\mathrm{ox}}\alpha_\mathrm{c}^2\right) - G_\mathrm{f}^2 \Delta_{i0}\left(1 - k_{i0}\alpha_\mathrm{c}^2\right)\right)}\right]^{-1}$
17: **end for**
18: $x_\mathrm{s} = x_{i0} + 2 \cdot q_0 \cdot \left[p_i + \sqrt{p_i^2 - 2 \cdot q_i \cdot \left(2 \cdot \left(1 - \dfrac{C_\mathrm{box}^2}{C'^2_\mathrm{ox}} \cdot \alpha_\mathrm{c}^2\right) - G_\mathrm{f}^2 \cdot \Delta_{i0} \cdot (1 - k_{i0} \cdot \alpha_\mathrm{c}^2)\right)}\right]^{-1}$
19: **else** $x_\mathrm{gf} \leq 0$
20: $\mathrm{x_s} = \mathrm{x_1}$
21: **end if**

A.3 Oberflächenpotential am Drain-Anschluss

Für die Drain-Seite des Kanals ergibt sich das gleiche Vorgehen wie für die Source-Seite. Aus physikalischer Sicht sind Änderungen an der Bedingung des if-Statements und neue Hilfsgleichungen nötig. Es folgt:

$$f_{\mathrm{x}} = (x_{\mathrm{gf}} - x_{\mathrm{s}})^2 - \frac{C_{\mathrm{box}}^2}{C_{\mathrm{ox}}'^2} \cdot (x_{\mathrm{gb}} - x_{\mathrm{gf}})^2 \tag{A.15}$$

$$V_{\mathrm{dse}} = \frac{V_{\mathrm{ds}}}{\left[1 + \left(\frac{V_{\mathrm{ds}}}{V_{\mathrm{dsat}}}\right)^{AX}\right]^{1/AX}} \tag{A.16}$$

$$b_{\mathrm{x}} = \frac{V_{\mathrm{dse}}}{\phi_{\mathrm{T}}^*} + x_{\mathrm{vtf0}}. \tag{A.17}$$

Die Berechnung für das Oberflächenpotential auf der Drain Seite des Kanals ist in Algorithmus 5 dargestellt, wobei für k_0 gilt:

$$k_0 = 1 - exp\left[(\alpha_{\mathrm{c}} - 1) \cdot \left(\frac{V_{\mathrm{dse}}}{\psi_{\mathrm{T}}^*} + x_{\mathrm{vtf0}}\right) + \varepsilon\right]. \tag{A.18}$$

Insgesamt zeigt sich, dass die tatsächliche Berechnung des Oberflächenpotential-basierten Transistormodells zu einem sehr hohen iterativen Rechenaufwand führt.

A.4 Hilfsfunktion für die Berechnung des Oberflächenpotentials

$$MINFCT(x, y, a) = \frac{1}{2}\left(x + y - \sqrt{(x-y)^2 + a}\right) \tag{A.19}$$

$$MAXFCT(x, y, a) = \frac{1}{2}\left(x + y + \sqrt{(x-y)^2 + a}\right) \tag{A.20}$$

$$\nu(a, c) = a + c \tag{A.21}$$

$$\mu(a, b, c, \tau) = \frac{\nu(a, c)^2}{\tau} + \frac{c^2}{2} - a \cdot b \tag{A.22}$$

$$\sigma(a, b, c, \tau, \eta) = \frac{a \cdot \nu(a, c)}{\mu(a, b, c, \tau) + \frac{\nu(a, c)}{\mu(a, b, c, \tau)} \cdot c \cdot \left(\frac{c^2}{3} - a \cdot b\right)} \tag{A.23}$$

Algorithmus 5 Berechnung des Oberflächenpotentials x_d auf der Drain Seite des Kanals

Eingabe: $x_\mathrm{d} = x_{j0} + \eta$; Eq. (A.10)
Ausgabe: x_d

1: **Initialisierung:**
2: $x_1 = \dfrac{C_\mathrm{box} \cdot C_\mathrm{si}}{C'_\mathrm{ox} \cdot C_\mathrm{box} + C_\mathrm{si} \cdot C_\mathrm{box} \cdot C'_\mathrm{ox} \cdot C_\mathrm{si}} \cdot (x_\mathrm{gb} - x_\mathrm{gf}) + x_\mathrm{gf}$ ▷ Berechne Startwert für x_s
3: $b_\mathrm{x} = \dfrac{V_\mathrm{dse}}{\phi^*_\mathrm{T}} \cdot x_\mathrm{vtf0}$ ▷ Berechne Gleichung (A.17)
4: G^2_f, x_gb, x_gf, x_vtf0 ▷ Berechne Gleichung (A.4)-(A.9)
5: **if** $f_\mathrm{x} > 10^{-6}$ **then**
6: $\eta = MINFCT(x_1, b_\mathrm{x}, 15) - \dfrac{1}{2} \cdot \left(b_\mathrm{x} - \sqrt{b^2_\mathrm{x} + 5}\right)$ ▷ Schätzung für x_s
7: $a = x_{\ell 0} + x_{\ell 1} \cdot \eta + x_{\ell 2} \cdot \eta^2$ ▷ Berechne Hilfsvariablen a,c,τ
8: $c = -x_{\ell 1} - 2 \cdot x_{\ell 2} \cdot \eta$
9: $\tau = \dfrac{V_\mathrm{dse}}{\phi^*_\mathrm{T}} - \eta + \ln(a)$
10: $x_{00} \approx \sigma(a, x_{\ell 2}, c, \tau, \eta)$ ▷ Schätzung für x_{j0}
11: **for** $i = 1, \ldots, i_{max}$ **do** ▷ Verbesserung der Schätzung des Korrekturterms x_{j0}
12: $i = i - 1$
13: $\Delta_{i0} = \exp\left(x_{i0} - \dfrac{V_\mathrm{dse}}{\phi^*_\mathrm{T}}\right)$ ▷ Berechne Hilfsvariable Δ_{i0}, k_{i0}
14: $k_{i0} = exp((\alpha_\mathrm{c} - 1) \cdot x_{i0} + \varepsilon)$
15: $p_i = 2 \cdot (x_\mathrm{gf} - x_{i0}) - 2 \cdot \alpha_\mathrm{c} \cdot \dfrac{C^2_\mathrm{box}}{C'^2_\mathrm{ox}} \cdot (x_\mathrm{gb} - \alpha_\mathrm{c} \cdot x_{i0} - \varepsilon) + G^2_\mathrm{f} \cdot \Delta_{i0} \cdot (1 - k_{i0} \cdot \alpha_\mathrm{c})$
16: $q_i = (x_\mathrm{gf} - x_{i0})^2 - \dfrac{C^2_\mathrm{box}}{C'^2_\mathrm{ox}} \cdot (x_\mathrm{gb} - \alpha_\mathrm{c} \cdot x_{i0} - \varepsilon)^2 + G^2_\mathrm{f} \cdot \Delta_{i0} \cdot (1 - k_{i0})$
17: $x_{(i+1)0} = x_{i0} + 2q_i \left[p_i + \sqrt{p_i^2 - 2q_i \left(2\left(1 - \dfrac{C^2_\mathrm{box}}{C'^2_\mathrm{ox}} \cdot \alpha^2_\mathrm{c}\right) - G^2_\mathrm{f} \Delta_{i0} \left(1 - k_{i0} \alpha^2_\mathrm{c}\right)\right)}\right]^{-1}$
18: **end for**
19: $x_\mathrm{s} = x_{i0} + 2 \cdot q_i \cdot \left[p_i + \sqrt{p_i^2 - 2 \cdot q_i \cdot \left(2 \cdot \left(1 - \dfrac{C^2_\mathrm{box}}{C'^2_\mathrm{ox}} \cdot \alpha^2_\mathrm{c}\right) - G^2_\mathrm{f} \cdot \Delta_{i0} \cdot (1 - k_{i0} \cdot \alpha^2_\mathrm{c})\right)}\right]^{-1}$
20: **else** $f_x \leq 10^{-6}$
21: $b_\mathrm{fds} = 2 \cdot \left[\alpha_\mathrm{c} \cdot \dfrac{C^2_\mathrm{box}}{C'^2_\mathrm{ox}} \cdot (x_\mathrm{gb} - x_\mathrm{s,b}) - (x_\mathrm{gf} - x_\mathrm{s})\right]$
22: $x_\mathrm{ds} = -G^2_\mathrm{f} \cdot \dfrac{exp(x_\mathrm{s}) - exp(x_\mathrm{s,b})}{b_\mathrm{fds}} \cdot \left[1 - exp\left(-\dfrac{V_\mathrm{dse}}{\psi^*_\mathrm{T}}\right)\right]$
23: $x_\mathrm{d} = x_\mathrm{ds} + x_\mathrm{s}$
24: **end if**

A.5 BSIM6-Ladungsberechnung

Die Notation in den Beschreibungen ist leicht unterschiedlich, so heißt die Kanalspannung v im BSIM6-Modell v_{ch} und die Steigung n heißt Ladungssteigung n_q. Die Bezeichnungen werden in dieser Arbeit so verwendet, wie in ihren jeweiligen Quellen, damit ein Vergleich leichter fällt.

Um in diesem Modell Gleichung (2.44) lösen zu können, wird im ersten Schritt ein initialer Wert für q_i berechnet und im Anschluss wird in Abhängigkeit des Arbeitspunktes dieses Ergebnis mit jeweils einem Schritt des Halley- und des Newton-Verfahrens verbessert. Die Idee des Vorgehens, das im BSIM-Modell verwendet wird, stammt aus [PKD+06]. Der Startwert wird berechnet, indem in Gleichung (2.45) die lineare Abhängigkeit von q_i in eine Taylorreihe um den Abschnürpunkt entwickelt wird. Anschließend wird die Ladung q_p von dem Punkt $|\ln(q_i)| = |2q_i|$ berechnet. Diese Ladung q_p hat somit den Zahlenwert $q_p = 0,301$ [Aga14]. Daraus folgt für die initiale Ladung

$$q_{ini} = \exp\left(\frac{v - 0,201 - \sqrt{(v + 0.402982)v + 2,446562}}{2}\right) . \tag{A.24}$$

Dabei ist

$$v = \phi_p - 2\phi_f - v_{ch} - \ln\left(\frac{4n_q\sqrt{\psi_p}}{\gamma_b}\right) . \tag{A.25}$$

Diese Spannung kann für einen bestimmten Punkt berechnet und in Gleichung (A.24) eingesetzt werden. Somit ist

$$q_{ini} = \exp\left(\frac{v_{ini} - 0,201 - \sqrt{(v_{ini} + 0.402982)v_{ini} + 2,446562}}{2}\right) . \tag{A.26}$$

Anhand des Wertes von $\ln q_{ini}$ werden verschiedene Fallunterscheidungen getroffen.

- falls $\ln q_{ini} \leq -80$, dann

$$q_i = q_{ini} \cdot \left[1 + \psi_p - 2\phi_F - v_{ch} - \ln q_{ini} - \ln\left(\frac{2n_q}{\gamma_b}\left(\frac{2q_{ini}n_q}{\gamma_b} + 2\sqrt{\psi_p}\right)\right)\right]$$

  - falls $\ln q_{ini} < -110$, dann $q_{ini} = \exp(-100)$, zur Anpassung der Rechengenauigkeit des Simulators
  - falls $\ln q_{ini} > -90$, dann $q_{ini} = q_{ini}$
  - sonst, $q_{ini} = \exp(-100 + 20\left(\frac{5}{64} + \frac{z}{2} + z^2\left(\frac{15}{16} - z^2\left(1,25 - z^2\right)\right)\right))$
    Dabei ist $z = {}^{(\ln q_{ini}+100)}/_{20}$

- falls $\ln q_{ini} > -80$, wird der initiale Wert q_{ini} für q_i mit Hilfe des Newton- und des Halley-Verfahrens verbessert. Zu Beginn wird ein Schritt mit dem Newton-Verfahren berechnet; somit sind die Anforderungen an den Startwert relaxierter. Im Anschluss daran wird ein weiterer Schritt mit dem Halley-Verfahren berechnet. Dieses Verfahren ist ein numerisches Verfahren zur Bestimmung von Nullstellen der Art $f(x) = 0$. Es ähnelt dem bekannten Newton-Verfahren, besitzt jedoch die Konvergenzordnung $\mathcal{O}\left(f(x)^3\right)$, besser bekannt als kubische Konvergenz, benötigt dafür jedoch die zweite Ableitung der Zielfunktion. Die Zielfunktion entsteht durch Umstellen der Gleichung (2.44) unter Vernachlässigung von q_i im Wurzelterm. So ergeben sich die Zielfunktion und die Ableitungen zu:

$$f(q_i) = 2q_i + \ln q_i + \ln\left(\frac{2q_i n_q}{\gamma_b}\left(\frac{2q_i n_q}{\gamma_b} + 2\sqrt{\psi_p}\right)\right) - \psi_p + 2\phi_f + v_{ch} \tag{A.27}$$

$$f'(q_i) = 2 + \frac{1}{q_i} + \frac{\frac{n_q}{\gamma_b} - \frac{1}{\sqrt{\psi_p}}}{\frac{n_q}{\gamma_b} q_i + \sqrt{\psi_p}} \tag{A.28}$$

$$f''(q_i) = -\frac{1}{q_i^2} - \frac{1}{\psi_p^{3/2}\left(\frac{n_q}{\gamma_b} q_i + \sqrt{\psi_p}\right)} - \left(\frac{\frac{n_q}{\gamma_b} - \frac{1}{\sqrt{\psi_p}}}{\frac{n_q}{\gamma_b} q_i + \sqrt{\psi_p}}\right)^2 . \tag{A.29}$$

Für die Inversionsladung q_i folgt nun

$$q_i = \underbrace{q_{ini} - \frac{f(q_{ini})}{f'(q_{ini})}}_{q_{new}} \underbrace{- \frac{f(q_{new})}{f'(q_{new})} \cdot \left(1 - \frac{f(q_{new}) \cdot f''(q_{new})}{2 \cdot f'(q_{new})^2}\right)^{-1}}_{\text{Halley-Verfahren}} \tag{A.30}$$

q_{new} ist der Wert, der durch den ersten Iterationsschritt des Newton-Verfahrens berechnet wird.

Mit diesen Fallunterscheidungen wird für den gesamten Funktionsbereich des Transistors die Inversionsladung konsistent und mindestens zweifach stetig differenzierbar. Sehr kleine Werte von q_i würden unter die Rechengenauigkeit des Simulators fallen, daher wird dieser Fall mit einer eigenen Fallunterscheidung abgefangen. Für den Bereich $-110 \leq \ln q_{ini} \leq -90$ wird eine Glättungsfunktion konstruiert, die die beiden angrenzenden Bereiche ebenfalls zweimal stetig differenzierbar miteinander verbindet. Des Weiteren kann durch die unterschiedlichen Spannungen an Drain und Source an diesen Kontakten jeweils ein q_s beziehungsweise ein q_d berechnet werden. Die Gleichung zur Berechnung des Drain-Source-Stromes I_{ds} weicht durch die phänomenologische Behandlung von weiteren Transistoreffekten von der Beschreibung im EKV-Modell ab.

A.6 Szenarien zum Vergleich der Filteralgorithmen

Beide Beispiele sind aus [Sim06] übernommen. Bei dem ersten Beispiel handelt sich um einen durch die Atmosphäre fallenden Körper, dessen Position, Geschwindigkeit und ballistischer Koeffizient geschätzt werden soll. Das zweite Beispiel ist eine sehr stark nichtlineare analytische Funktion, deren Parameter geschätzt werden sollen.

A.6.1 Beispiel 1

Das Beispiel stammt aus [JUDW00] und stammt ursprünglich aus [AWB68]. Die initialen Größen sind:

Listing A.1: Initiale Siumlationswerte für den fallenden Körper

```
1        rho0 = 2; % lb-sec^2/ft^4
2        g = 32.2; % ft/sec^2
3        k = 2e4; % ft
4        R = 10^4; % measurement noise variance (ft^2)
5        Q = diag([0 0 0]); % process noise covariance
6        M = 10^5; % horizontal range of position sensor
7        a = 10^5; % altitude of position sensor
8        P = diag([1e6 4e6 10]); % initial estimation error
            covariance
9
10       x = [3e5; -2e4; 1e-3]; % initial state
11       xhat = [3e5; -2e4; 1e-3]; % initial state estimate
12
13       N = 1000; % number of particles
```

Zu beachten ist, dass in diesem Beispiel das imperiale Maßsystem verwendet wird. Die eigentliche Berechnung des Modells findet in folgender Schleife statt:

Listing A.2: Systembeschreibung in Matlab

```
1 for t = T : T : tf
2         for tau = dt : dt : T
3                 % Fourth order Runge Kutta integration
4                 dx1(1,1) = x(2);
5                 dx1(2,1) = rho0 * exp(-x(1)/k) * x(2)^2 / 2 *
                     x(3) - g;
6                 dx1(3,1) = 0;
7                 dx1 = dx1 * dt;
8                 xtemp = x + dx1 / 2;
9                 dx2(1,1) = xtemp(2);
```

```
10                    dx2(2,1) = rho0 * exp(-xtemp(1)/k) * xtemp(2)
                        ^2 / 2 * xtemp(3) - g;
11                    dx2(3,1) = 0;
12                    dx2 = dx2 * dt;
13                    xtemp = x + dx2 / 2;
14                    dx3(1,1) = xtemp(2);
15                    dx3(2,1) = rho0 * exp(-xtemp(1)/k) * xtemp(2)
                        ^2 / 2 * xtemp(3) - g;
16                    dx3(3,1) = 0;
17                    dx3 = dx3 * dt;
18                    xtemp = x + dx3;
19                    dx4(1,1) = xtemp(2);
20                    dx4(2,1) = rho0 * exp(-xtemp(1)/k) * xtemp(2)
                        ^2 / 2 * xtemp(3) - g;
21                    dx4(3,1) = 0;
22                    dx4 = dx4 * dt;
23                    x = x + (dx1 + 2 * dx2 + 2 * dx3 + dx4) / 6;
24                    x = x + sqrt(dt * Q) * [randn; randn; randn] *
                          dt;
25            end
26 end
```

Dabei ist t der jeweilige Zeitschritt, zu dem Messwerte zur Verfügung stehen und τ die Variable für das numerische Integrationsverfahren zur Lösung der Systemgleichung. Anschließend wird die verrauschte Messung für die Filteralgorithmen berechnet:

Listing A.3: Berechnung der verrauschten Messwerte der Höhe

```
1  z = sqrt(M^2 + (x(1)-a)^2) + sqrt(R) * randn;
```

Anschließend werden die [Sim06] als MATLAB-Implementierung gegebenen Filteralgorithmen zum Vergleich verwendet. Der ballistische Koeffizient ist dabei definiert als

$$B = \frac{m}{c_w \cdot A} \ , \tag{A.31}$$

wobei m die Masse, c_w der Luftwiderstandsbeiwert und A die Querschnittsfläche ist. Für den Fall aus großer Höhe wird diese Definition erweitert als

$$B^* = \frac{1}{2} \cdot B \cdot \rho_0 \ . \tag{A.32}$$

Dabei ist ρ_0 die Dichte der Atmosphäre in der Höhe h_0.

A.6.2 Beispiel 2

Diese Beispiel verwendet ein stark nichtlineares, rein mathematisches System nach [GSS93]. Das System ist gegeben durch

$$x_k = 0,5x_{k-1} + \frac{25x_{k-1}}{1 + x_{k-1}^2} + 8\cos(k-1) + w_k \tag{A.33}$$

$$y_k = \frac{x_k^2}{20} + v_k \ . \tag{A.34}$$

k ist der zeitliche Laufindex, w_k das Prozessrauschen und v_k das Messrauschen mit jeweils beliebigen Verteilungsfunktionen. Dieses System wird sehr häufig zur Evaluierung von nichtlinearen und nicht-gaußverteilen Filteralgorithmen verwendet.

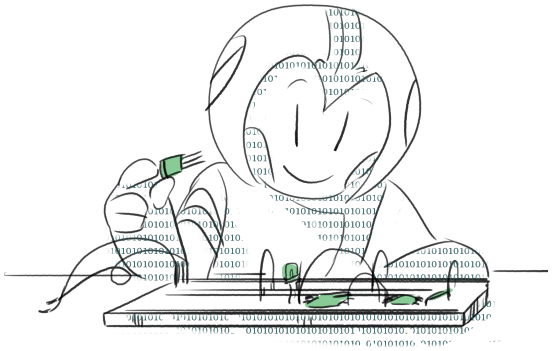
A.7 Parameter des BTI-Modells

In diesem Abschnitt werden die Abhängigkeiten und zusätzlichen Modellparameter des verwendeten BTI-Modells aus [MMM+14] beschrieben.

Parameter	**Definition und Gleichung**
τ_s	Generierungskonstante für Oberflächenzustände: 3e-4 s
τ_r	Rekombinierungskonstante für Oberflächenzustände: 1e-6 s
τ_e	Zeitkonstante der Loch-Emission: 1e-2 s
N_0 in cm^3	Dichte der ins Oxid eingetragenden Löcher $\frac{qN_0}{C_{ox}} \propto I_g = GL_1 \exp\left(-{}^{GL_2}/_{E_{ox}}\right)$ $\Rightarrow \frac{1}{E_{ox}} = \frac{\ln\left(\frac{C_{ox}}{qN_0}\right)+\ln(GL_1)}{GL_2} = \frac{\ln\left(\frac{C_{ox}}{qN_0}\right)+\ln(M_{N0})}{N_{N0}}$ Es gilt $M_{N0} = 17,6$ und $N_{N0} = 4,18$. Beide Parameter sind von den Materialien und dem Prozess abhängig. Die Gleichung kann nach N_0 umgestellt werden oder implizit in Verilog implementiert werden.
$N_{trap,0}$ in cm^2	Existierende Fehlstellen der Oberfläche $N_{trap,0} = \frac{C_{ox}q}{\cdot} M_{Ntrap} \cdot \exp(N_{Ntrap} \cdot E_{ox})$ Dabei sind $M_{Ntrap} = 0,012$ und $N_{Ntrap} = 1,31$.
τ_c in s	Einfang-Zeitkonstante von Löchern $\tau_c = M_{tc} \cdot \exp(N_{tc} \cdot E_{ox})$ Dabei gilt für $M_{tc} = 0,0012$ und $N_{tc} = -5,55$.
R_{str} in V	Generierungsrate der Fehlstellen an der Oberfläche $R_{str} = \frac{C_{ox}q}{\cdot} M_{Rstr} \cdot \exp(N_{Rstr} \cdot E_{ox})$ Mit $M_{Rstr} = 0,0016$ und $N_{Rstr} = 2,54$.
R_{rec}	Rekombinierungsrate der Fehlstellen an der Oberfläche $R_{rec} = M_{Rrec} \cdot t_r^{N_{Rrec}}$ Mit $M_{Rrec} = 3,09e-3$ und $N_{Rrec} = -0,117$

B Verstärkertopologien

» Errors are not in the art but in the artificers. «



In diesem Abschnitt werden die verschiedenen Verstärkertopologien, die in der Fallstudie in Kapitel 4.1.2.3 verwendet werden, vorgestellt. In den Schaltplänen sind die relevanten Funktionsblöcke farblich hervorgehoben. Zudem sind für jeden Verstärker die relevanten Figure of Merits ausführlich über Temperatur, Prozessvariationen und Alterung dargestellt.
Die betrachtenden Größen sind:

- BW: Bandbreite des Verstärkers, definiert als die Frequenz bei der die Verstärkung um 3dB abgefallen ist. Dies ist eine äquivalente Definition der Eckfrequenz, jedoch weisen die Verstärker im betrachteten Frequenzbereich kein Hochpassverhalten auf, daher entsprechen sich diese Größen.

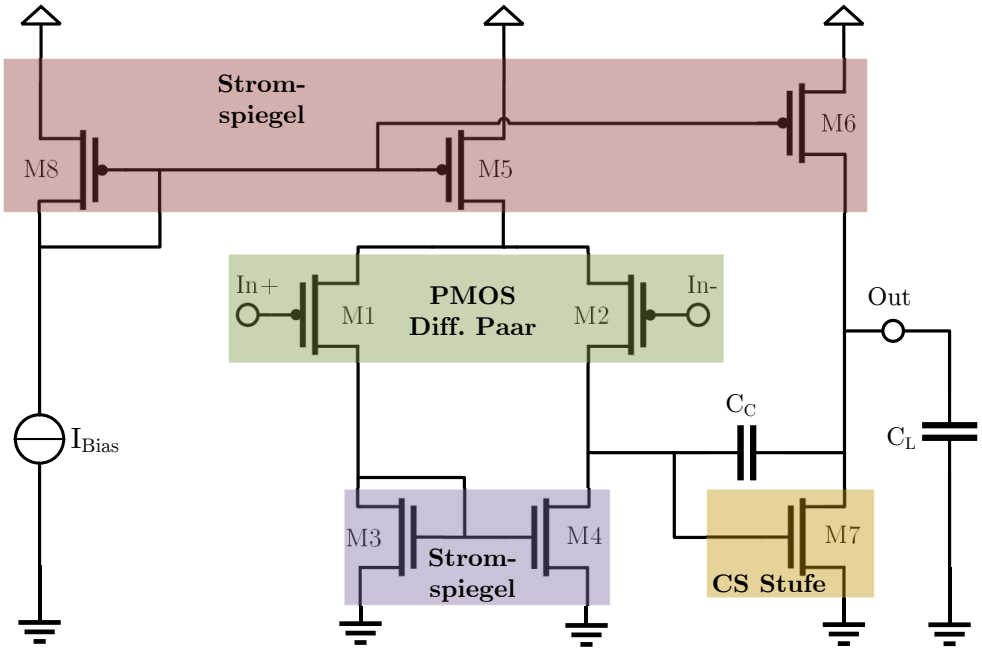
- GBW: Gain-Bandwidth-Product, Produkt aus der linearen Verstärkung und der Bandbreite.

- σ_{-}: Äquivalente Rauschspannung des invertierenden Eingangs auf den Ausgang.

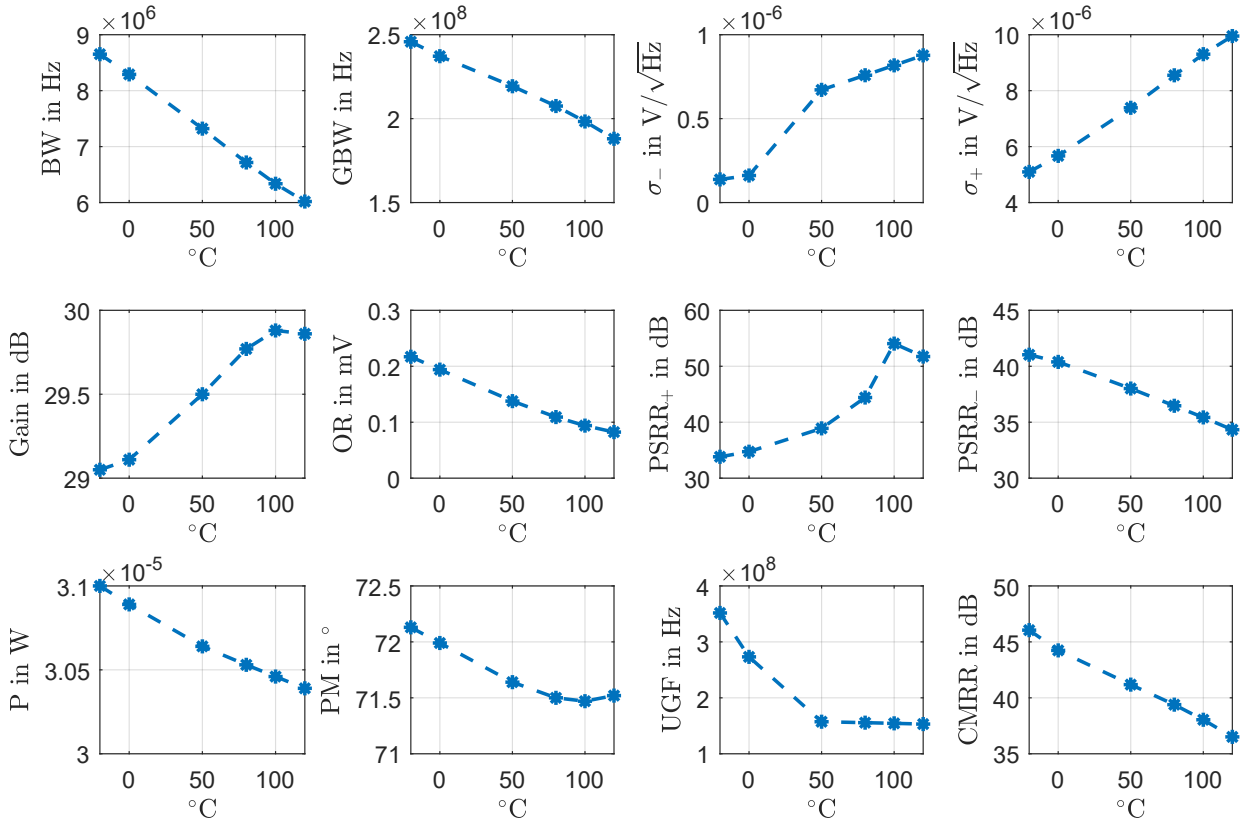
- σ_+: Äquivalente Rauschspannung des nicht invertierenden Eingangs auf den Ausgang.
- Gain: Verstärkung für eine Konfiguration in offener Schleife.
- OR: Output Range, mögliche Auslenkung der Spannung am Ausgang des Verstärkers
- PSRR$_+$,PSRR$_-$: Power Supply Rejection Ration, die Sensitivität des Verstärkers gegenüber Schwankungen von V_{dd} und V_{ss}
- P: Leistungsaufnahme des Verstärkers mit der jeweiligen Biasquelle, falls notwendig.
- PM: Phase Margin, Phasenreserve des Verstärkers
- UGF: Unity Gain Frequenz, die Frequenz bei der die Verstärkung 0dB beträgt.
- CMRR: Common Mode Rejection Ration, Sensitivität des Verstärkers gegenüber einer Veränderung des Gleichtaktsignals am Eingang.

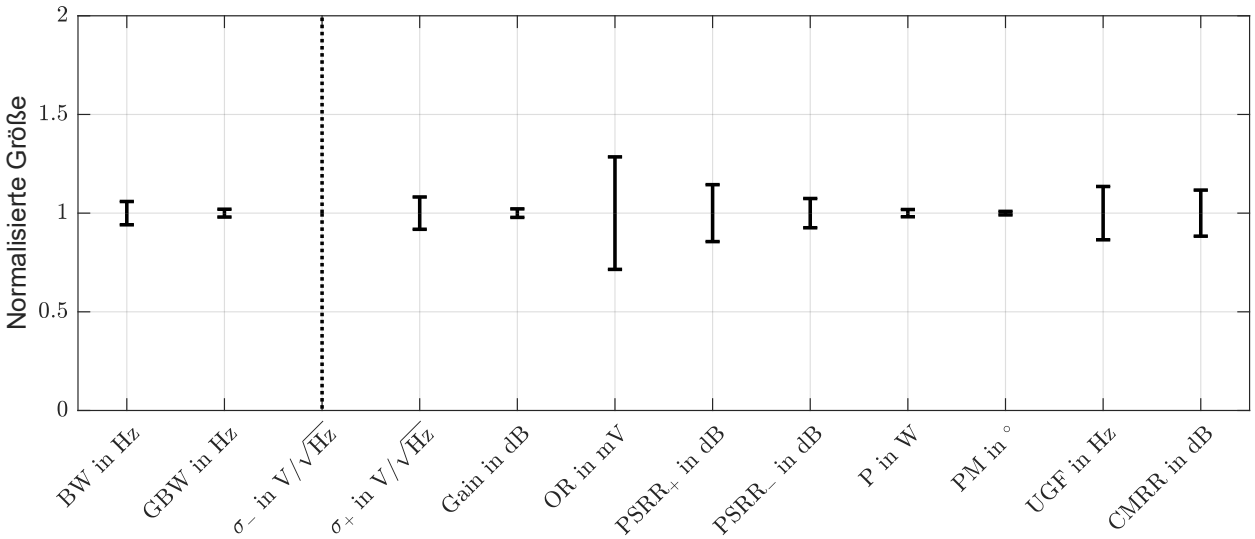
Für die weitere Analyse werden alle FoM über der Temperatur des jeweiligen Verstärkers gezeigt. Anschließend zeigt ein vereinfachter Boxplot die Variabilität der einzelnen Größen, verursacht durch Prozessvariationen. Danach wird Beispielhaft das GBW über 10 Jahre bei 3 verschiedenen Temperaturen und 5 Monte-Carlo Samples mit Prozessvariationen gezeigt, um die Auswirkungen von Wechselwirkungen auf die jeweilige Topologie zu zeigen. Anschließend werden 9 der Größen für zwei verschiedene Temperaturen über 10 Jahre mit Alterung dargestellt. Diese Analyse bzw. Darstellungen werden für jede der Topologien gezeigt.
Alle Schaltungen wurden aus [Bak08] entnommen, um zu gewährleisten, dass die Wahl der Transistorgrößen definitiv zu stabilen und gut evaluierten Arbeitspunkten führt. Dies ist entscheidend, da in diesem Abschnitt nur die Wahl der Topologie die Änderungen im Verhalten gegen Temperatur, Prozessvariationen und Alterung verursachen soll. Die einzelnen Verstärker sind selbstverständlich für verschiedene Anwendungen ausgelegt und somit unterscheiden sich die nominellen Werte der FoM. Es soll jedoch nur die jeweilige Sensitivität der FoM auf Umwelteinflüsse untersucht werden. Daher sind die absoluten Werte zweitrangig.
Sämtliche Schaltungen wurden in Cadence Virtouso aufgebaut und mit den REL-MOS Modellen analysiert. Die Alterung entspricht einer transienten Simulation über die gesamte Dauer von 10 Jahren. Die jeweilige Temperatur der Alterungssimulationen gilt für den gesamten Alterungszeitraum. V_{dd} ist für alle Schaltungen zu 1V gewählt. Wenn keine Temperatur angeben ist, gilt $T = 27°$C. Als Biasquellen wurden die Beta-Multiplier aus dem vorherigen Abschnitt verwendet.

B.1 Miller-OTA



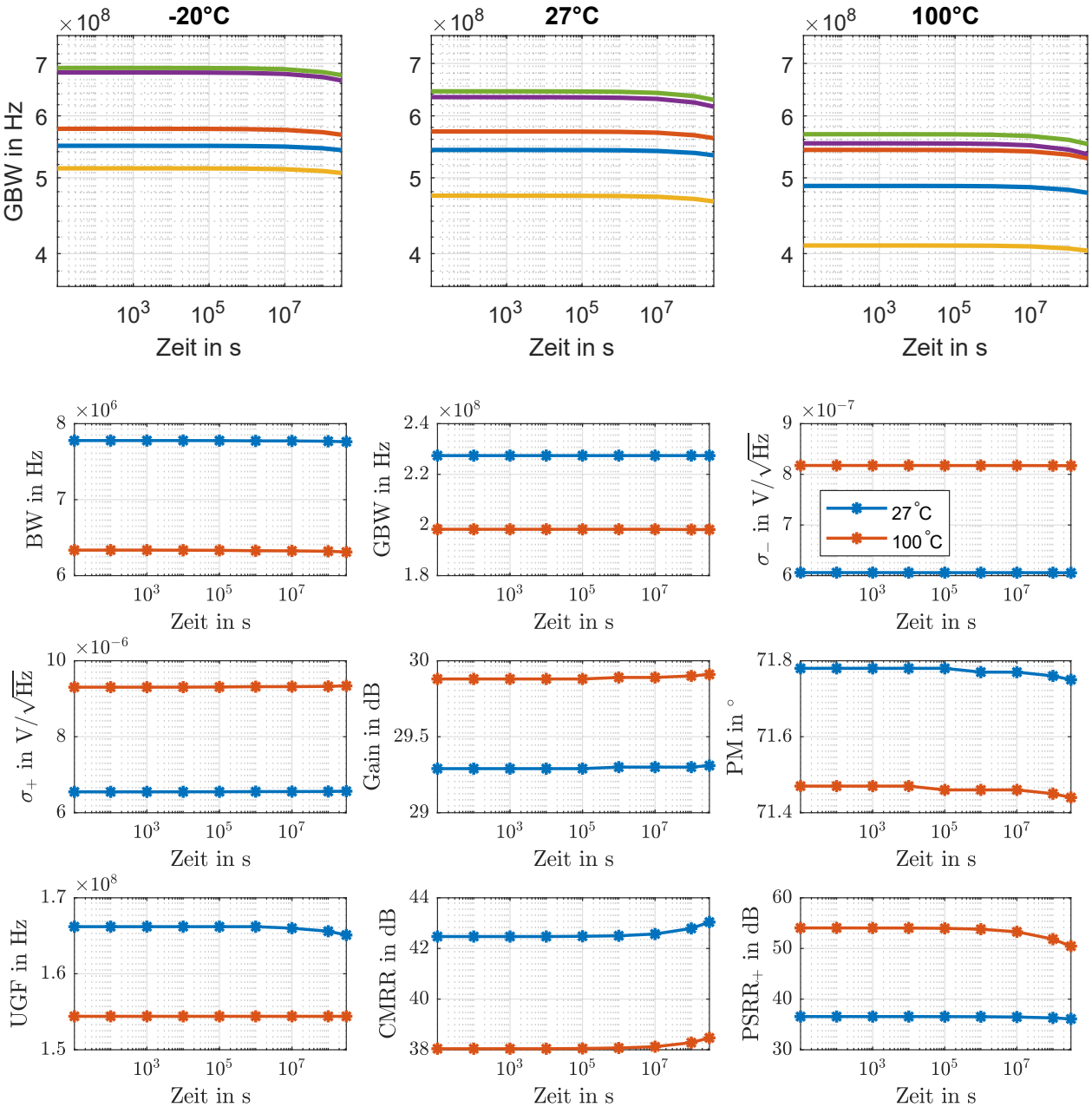
Bei diesem simplen Verstärker gibt es keine Kaskode und somit keine trimmbare Biasspannung. Dies bedeutet zum einen, dass der Arbeitspunkt einzelner Transistoren und somit der gesamten Schaltung über das Alter und der Temperatur stark verschieden sein kann. Besonders an der Output Range, die durch die Ausgangswiderstände der NMOS- und PMOS-Transistoren bestimmt wird, ist zu erkennen, dass sich deren Arbeitspunkte über der Temperatur sehr stark verschieben. Zum anderen führt das Fehlen von Biasspannungen dazu, dass keine Transistoren aufeinander abgestimmt sein müssen. Somit zeigt sich die Schaltung insgesamt robust gegenüber Prozessvariationen. Dadurch, dass die Alterung sich nur sehr gering mit der Temperatur verschiebt, ist HCI die Hauptursache und diese wie die Block-Analyse gezeigt hat vor allem durch die Common Source Stufe und den nicht kaskadierten NMOS Stromspiegel.



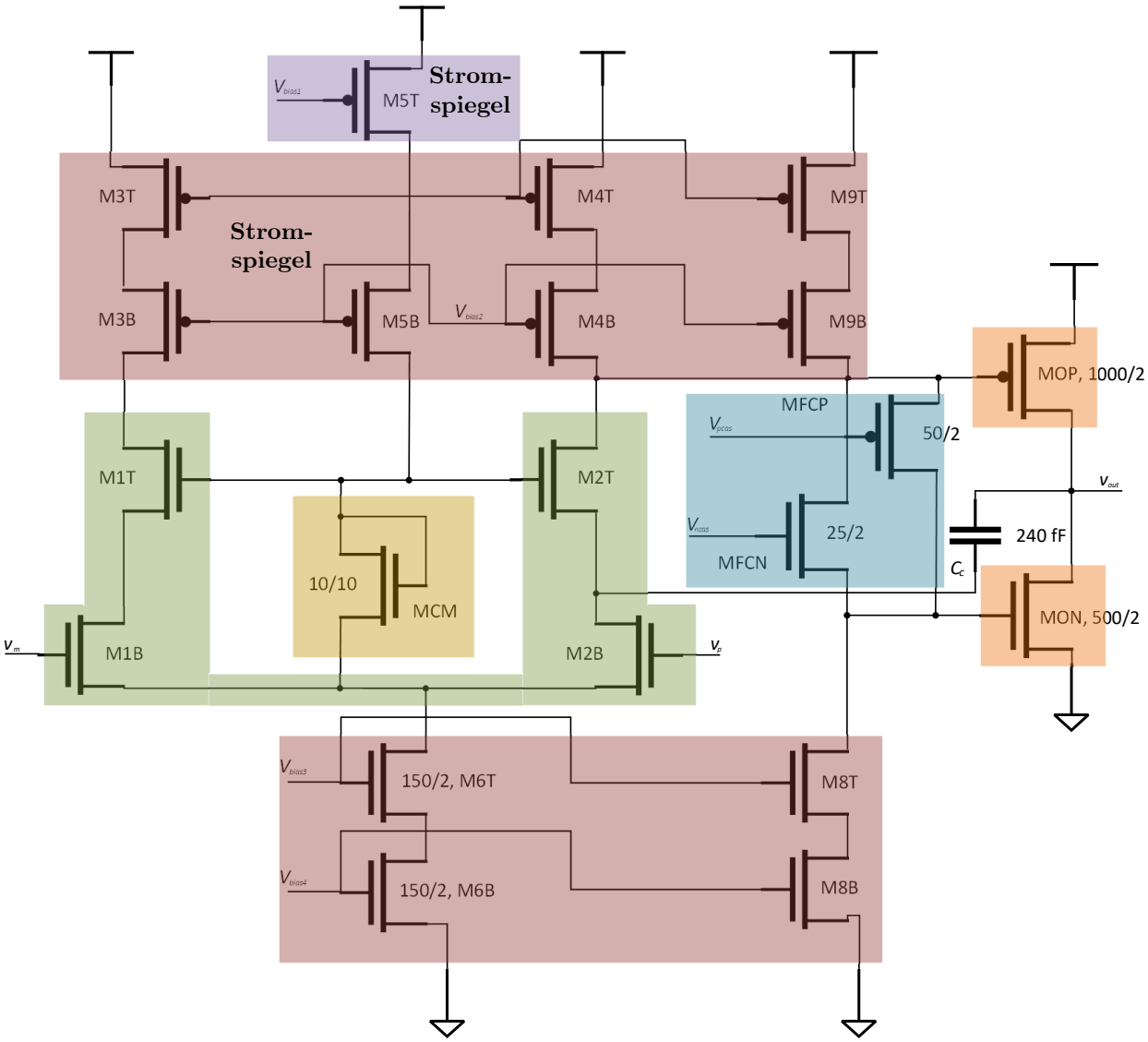


	BW in MHz	GBW in MHz	σ_- in $\mu V/\sqrt{\text{Hz}}$	σ_+ in $\mu V/\sqrt{\text{Hz}}$	Gain in dB	OR in mV
Mittelwert	7,792	226,7	634,1m	6,568	29,26	162,5
Std.Abw.	0,459	4,485	0,956	0,537	0,632	46,32

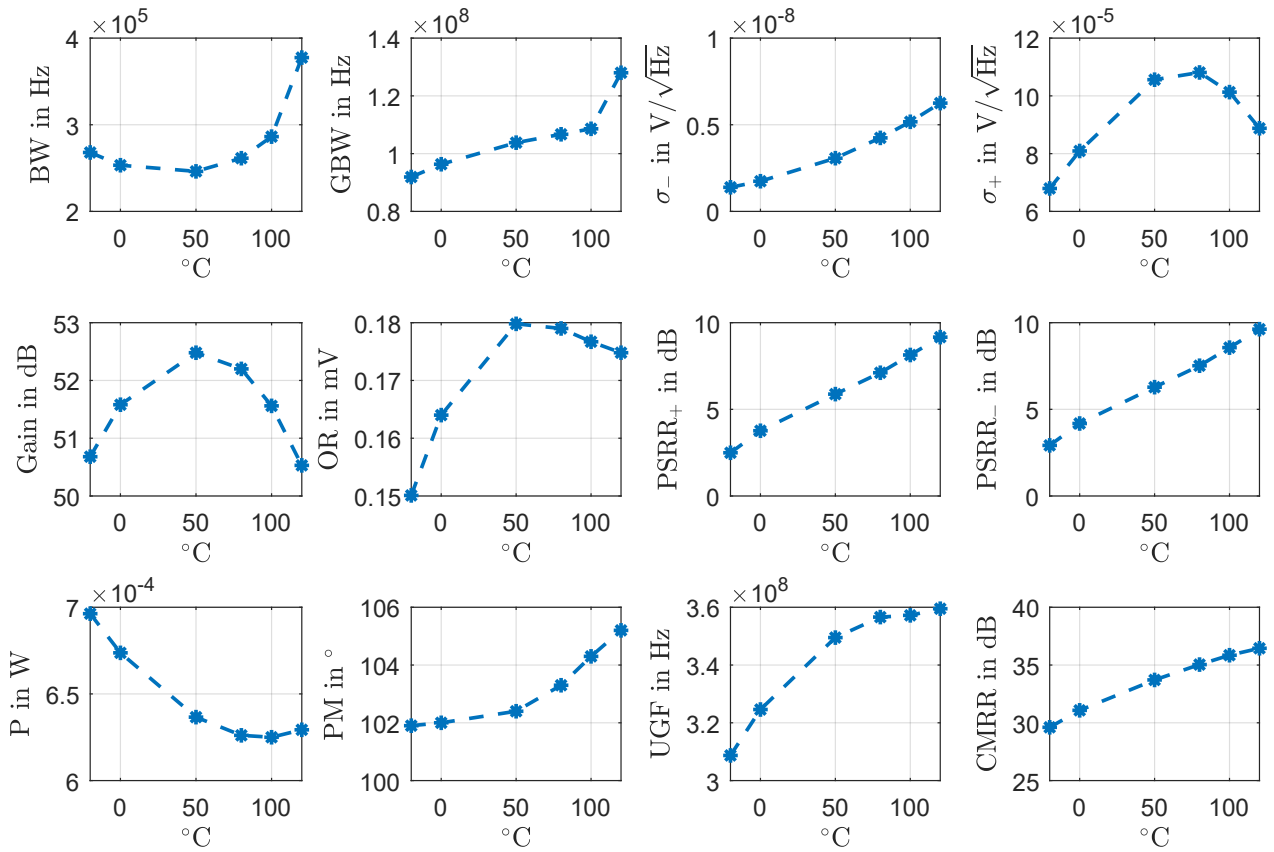
	PSRR$_-$ in dB	PSRR$_+$ in dB	P in μW	PM in °	UGF in MHz	CMRR in dB
Mittelwert	38,32	39,14	30,78	71,81	176,9	42,84
Std.Abw.	5,527	2,901	0,564	0,628	23,89	5

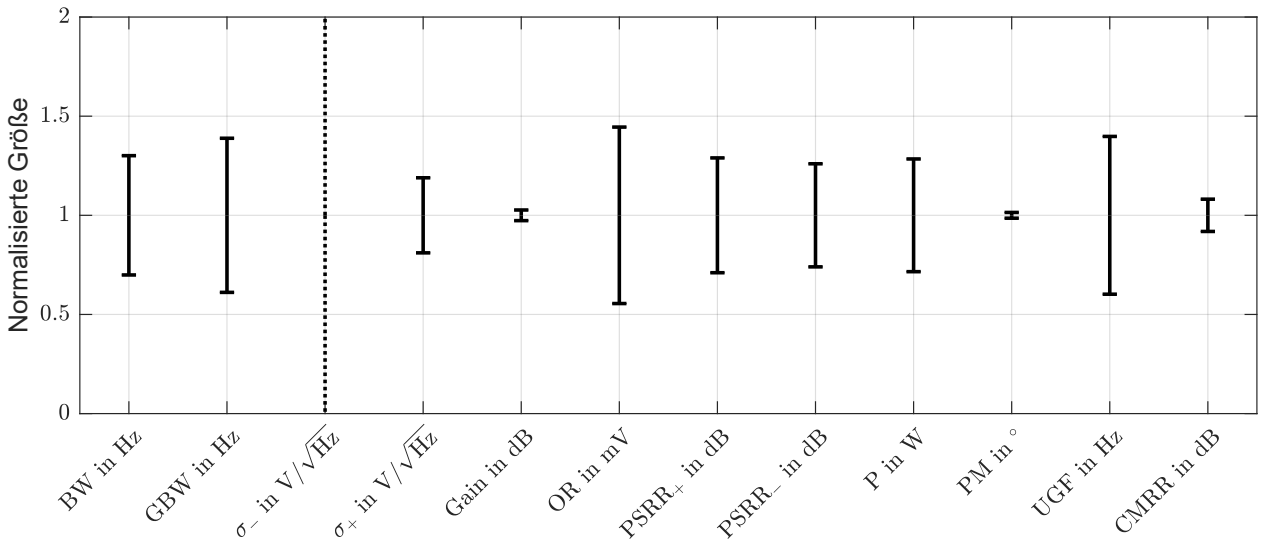


B.2 Operationsverstärker



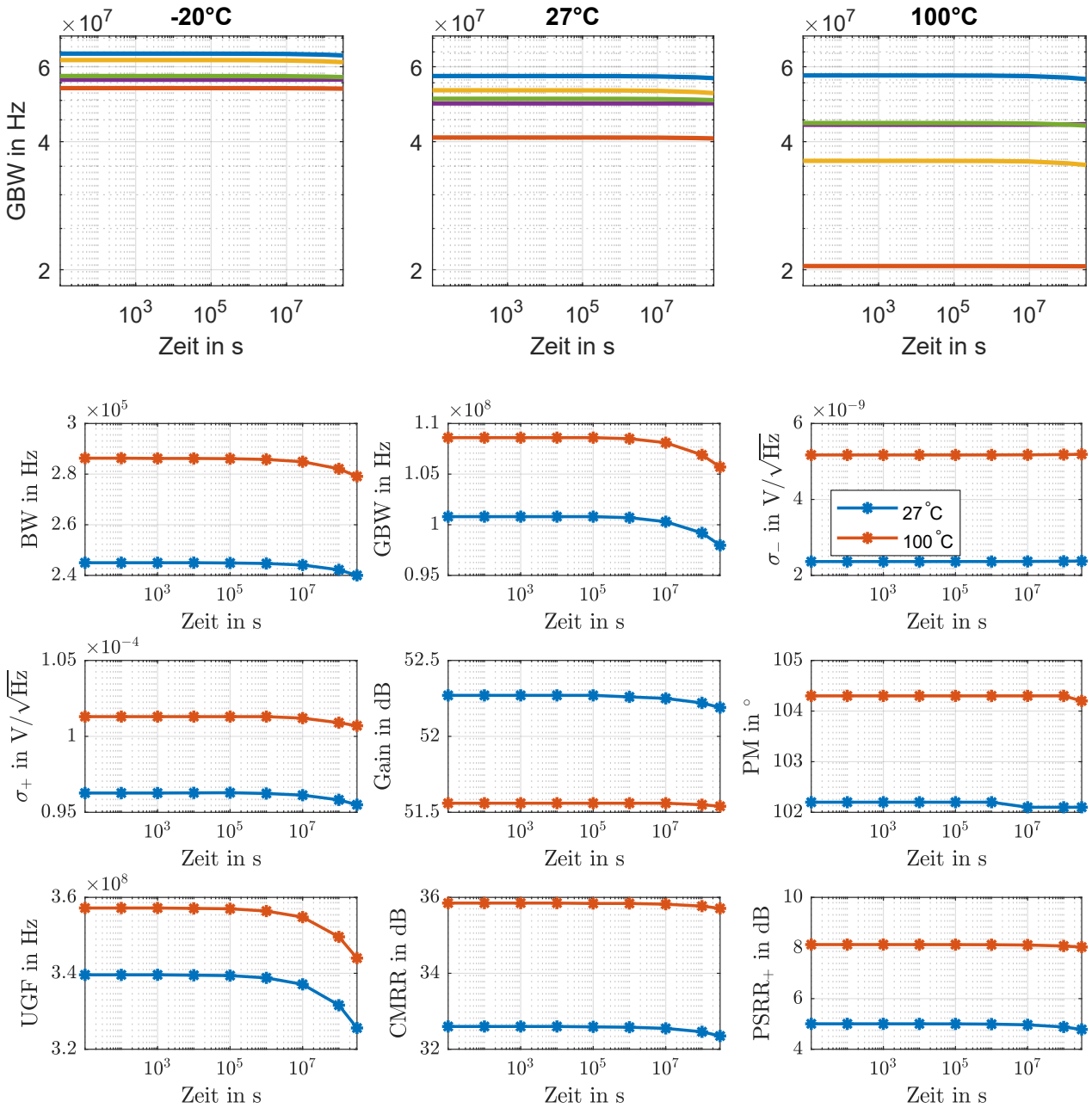
Obwohl bei diesem Verstärker alle relevanten Funktionseinheiten als Kaskode ausgeführt sind, sind die Kennzahlen über Alterung, Temperatur und Prozessvariation sehr veränderlich. Die Ursache dafür ist die NMOS-Diode, die den Arbeitspunkt der Eingangstransistoren maßgeblich mitbestimmt. Durch die starke Abhängigkeit dieses Blocks von den genannten Einflüssen wird ebenfalls der gesamte Verstärker anfällig.



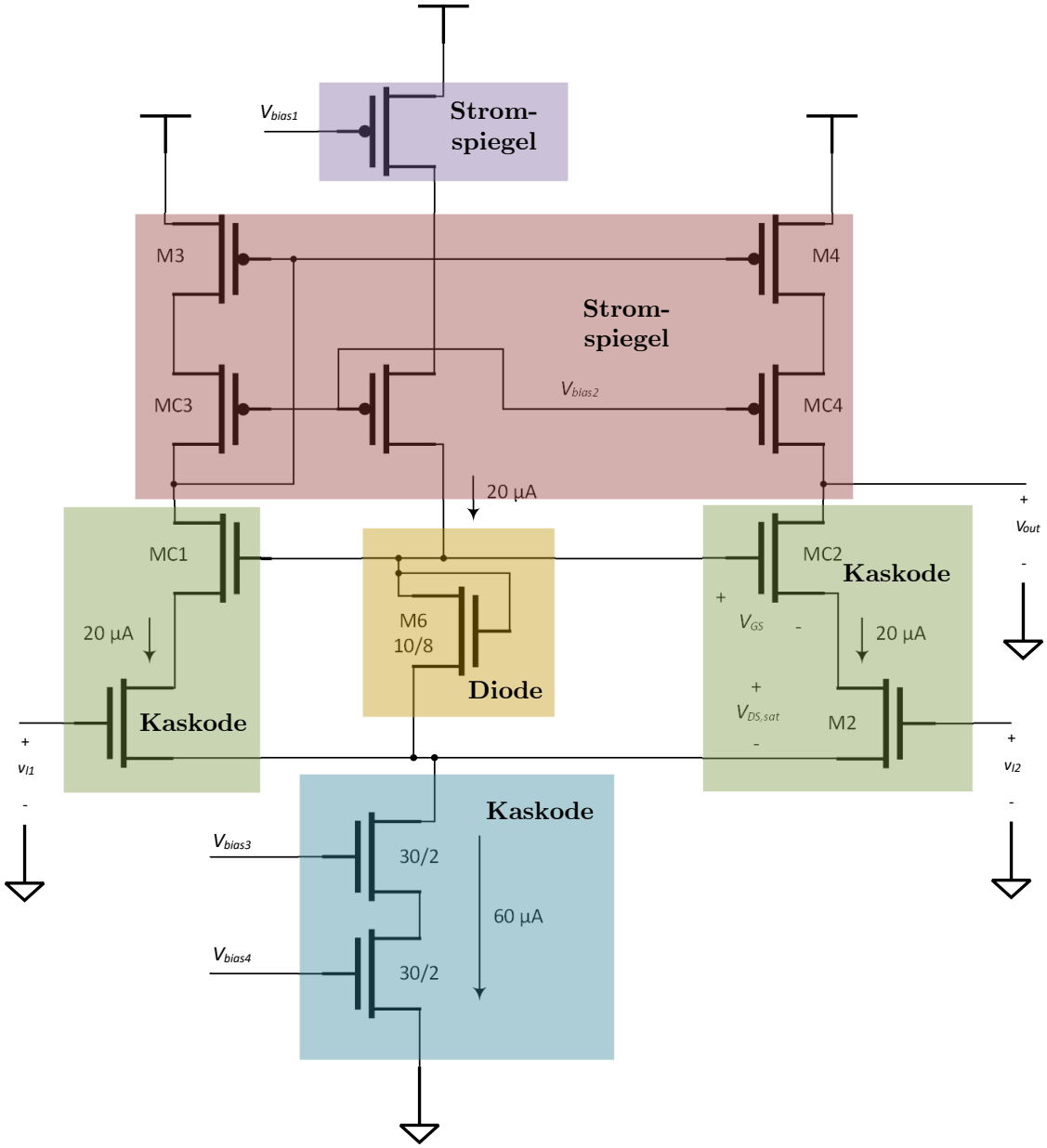


	BW in MHz	GBW in MHz	σ_- in $\mu V/\sqrt{\text{Hz}}$	σ_+ in $\mu V/\sqrt{\text{Hz}}$	Gain in dB	OR in mV
Mittelwert	265,3	109,3	0,738	95,83	52,01	143,6
Std.Abw.	0,798	42,48	10,4	18,14	1,395	63,9

	PSRR$_-$ in dB	PSRR$_+$ in dB	P in μW	PM in °	UGF in MHz	CMRR in dB
Mittelwert	5,524	5,108	682,2	102,4	349,6	32,72
Std.Abw.	1,436	1,479	193,9	1,485	0,798	2,663



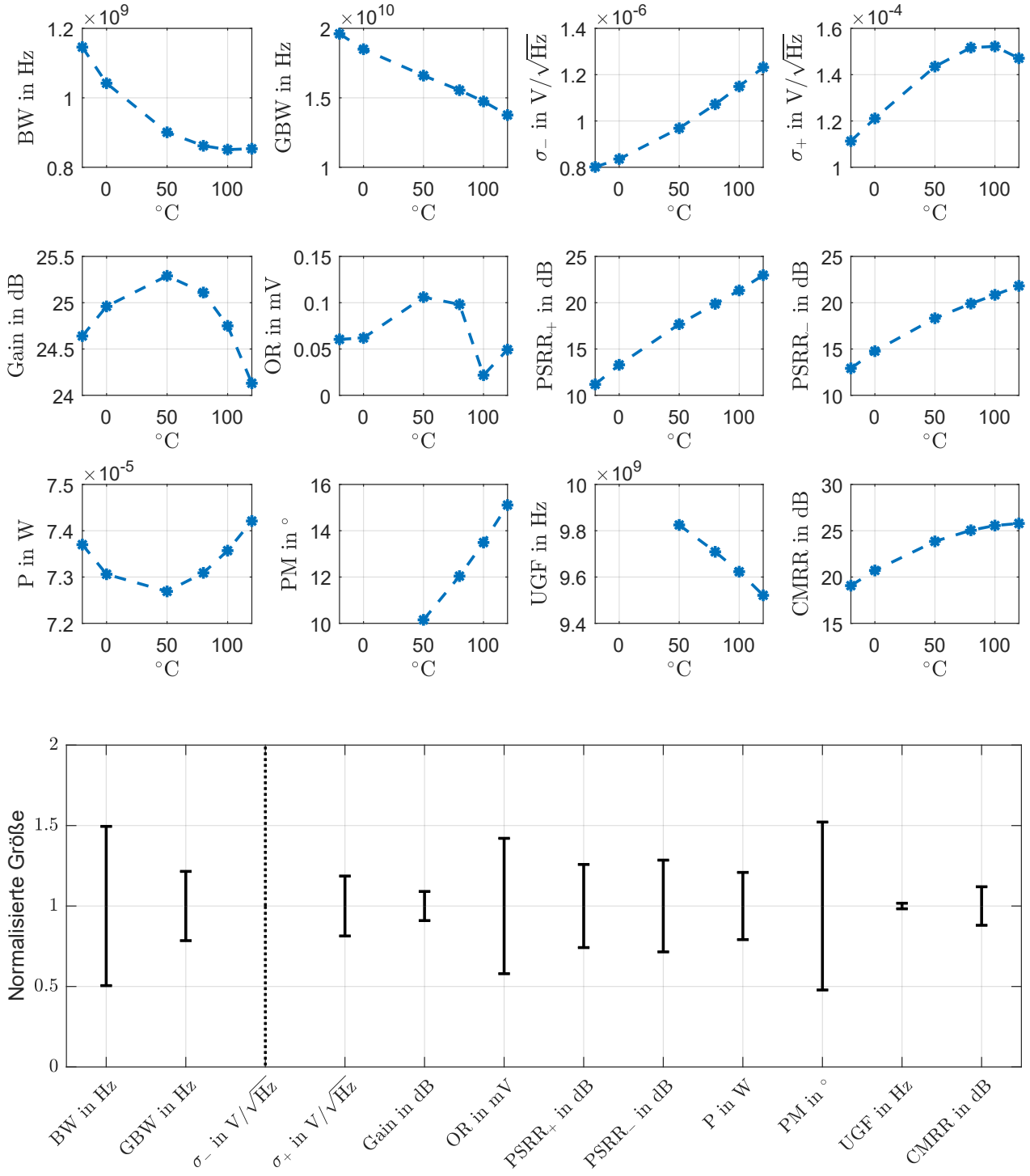
B.3 Differentieller Kaskode-Operationsverstärker



Für diesen Verstärker gilt dasselbe wie für den vorherigen. Durch die Arbeitspunkteinstellung der Eingangstransistoren mittels einer NMOS-Diode wird der gesamte Verstärker anfällig gegenüber äußeren Einflüssen und Prozessvariationen. Besonders die Größen, die maßgeblich durch den Eingang bestimmt sind, Gain und CMRR, sind somit am stärksten von der Alterung betroffen. Auch hier legt der geringe Unterschied der Alterung für verschiedene Temperaturen nahe, dass HCI der Haupteinflussfaktor ist. Zudem zeigt die starke Variabilität der Alterung des GBW, dass die Wechselwirkungen die einzelnen Effekte verstärken.

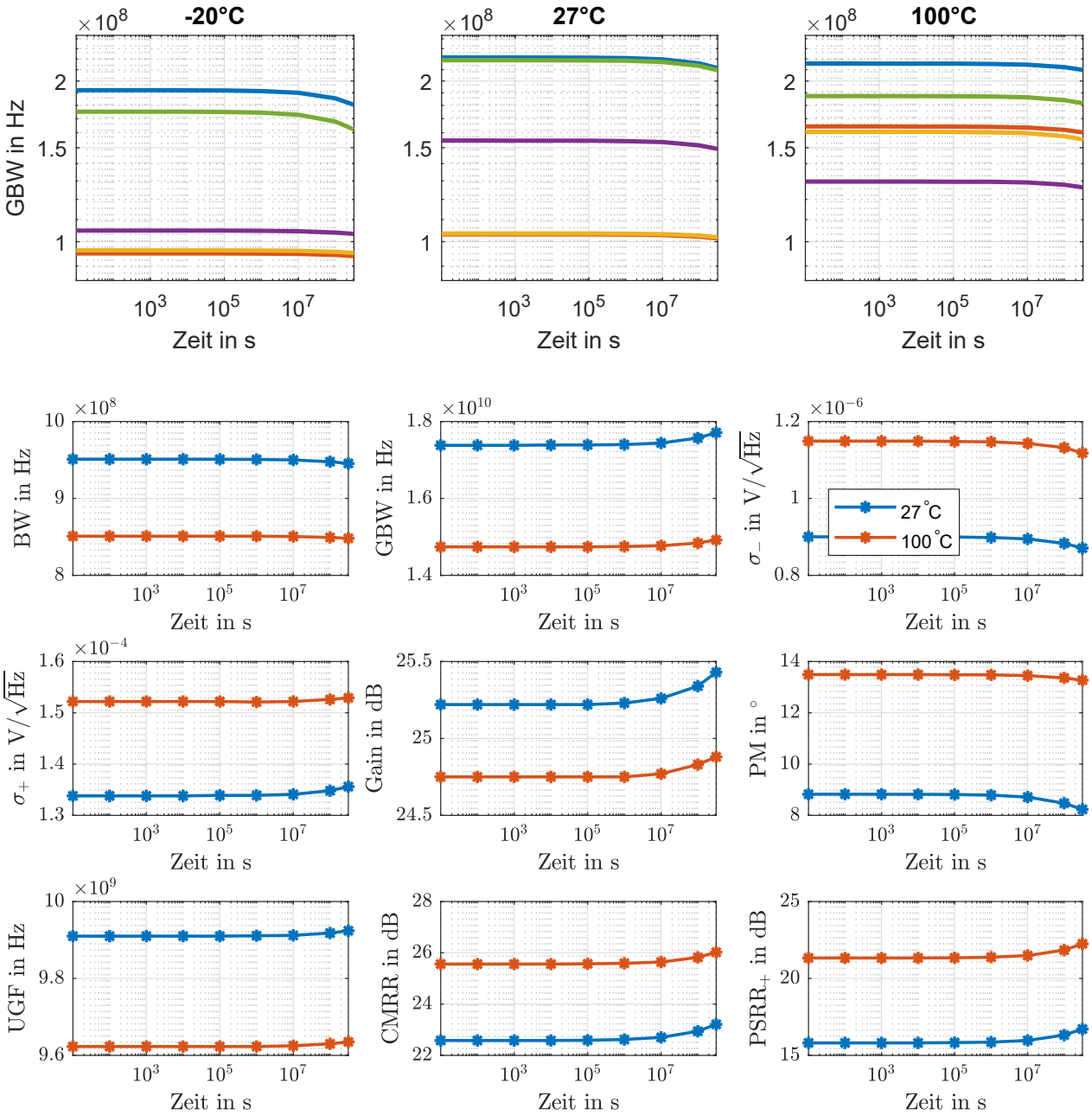
Dieser Verstärker wurde mit einer Lang- und einer Kurzkanalreferenzquelle analysiert. Die Wahl der Quelle hat jedoch keine Auswirkungen auf die Funktion des Verstärkers,

solange die Quelle selbst temperaturkompensiert ist. Daher sind die Ergebnisse in Kapitel 4.1.2.3 für die Topologien 3 und 4 nahezu identisch.

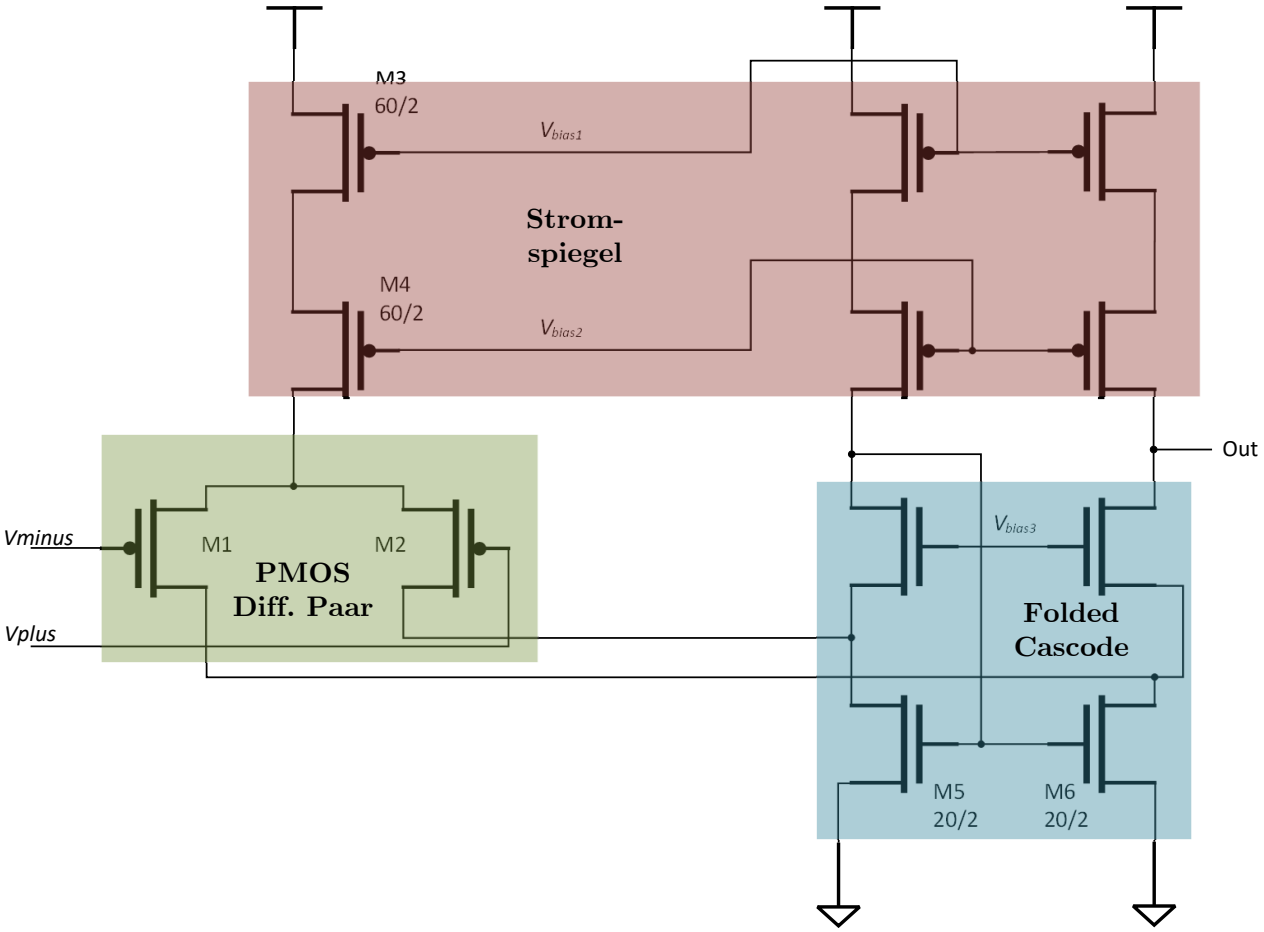


	BW in MHz	GBW in MHz	σ_- in $\mu V/\sqrt{\text{Hz}}$	σ_+ in $\mu V/\sqrt{\text{Hz}}$	Gain in dB	OR in mV
Mittelwert	1150	18530	1,641	129,1	24,69	90,36
Std.Abw.	569	3987	6,981	24,04	2,234	38,01

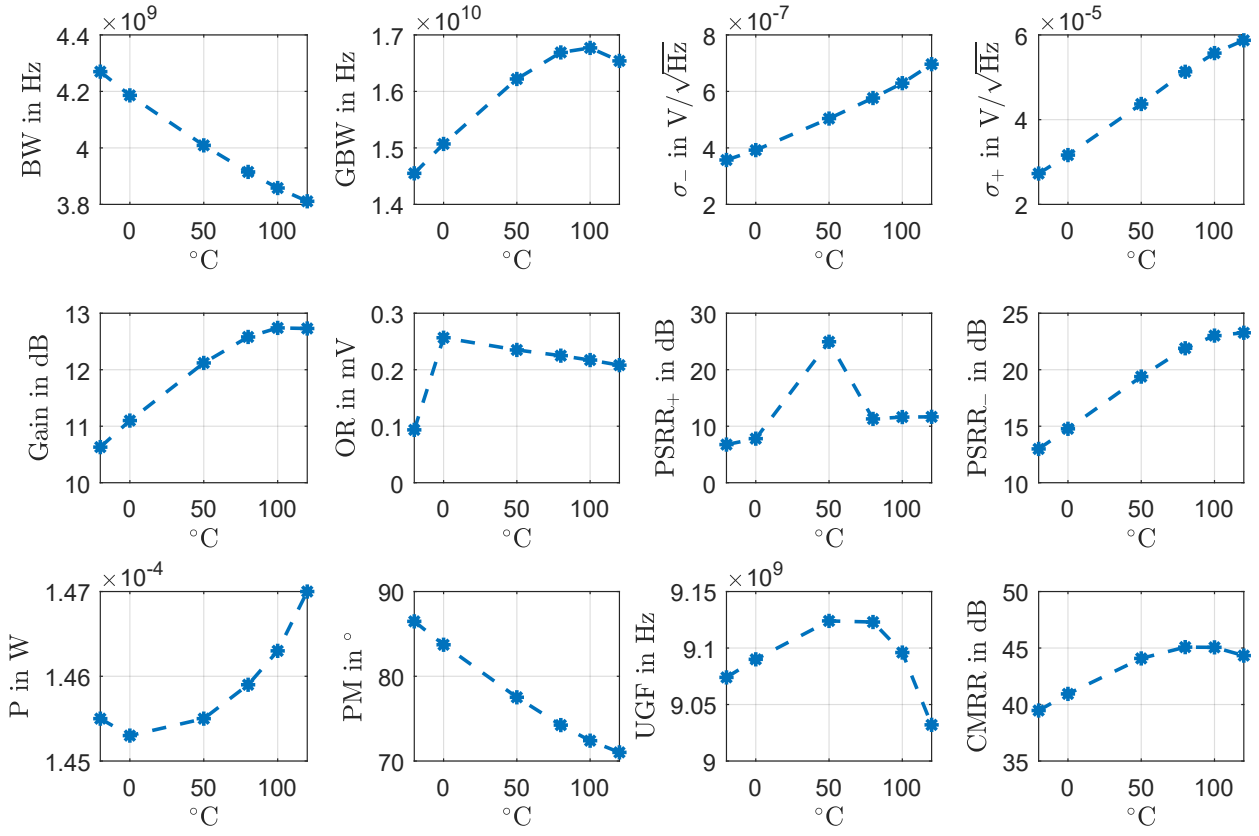
	PSRR$_-$ in dB	PSRR$_+$ in dB	P in μW	PM in °	UGF in MHz	CMRR in dB
Mittelwert	17,94	17,34	74,63	11,63	9745	23,31
Std.Abw.	5,114	4,478	15,58	6,068	168,5	2,79

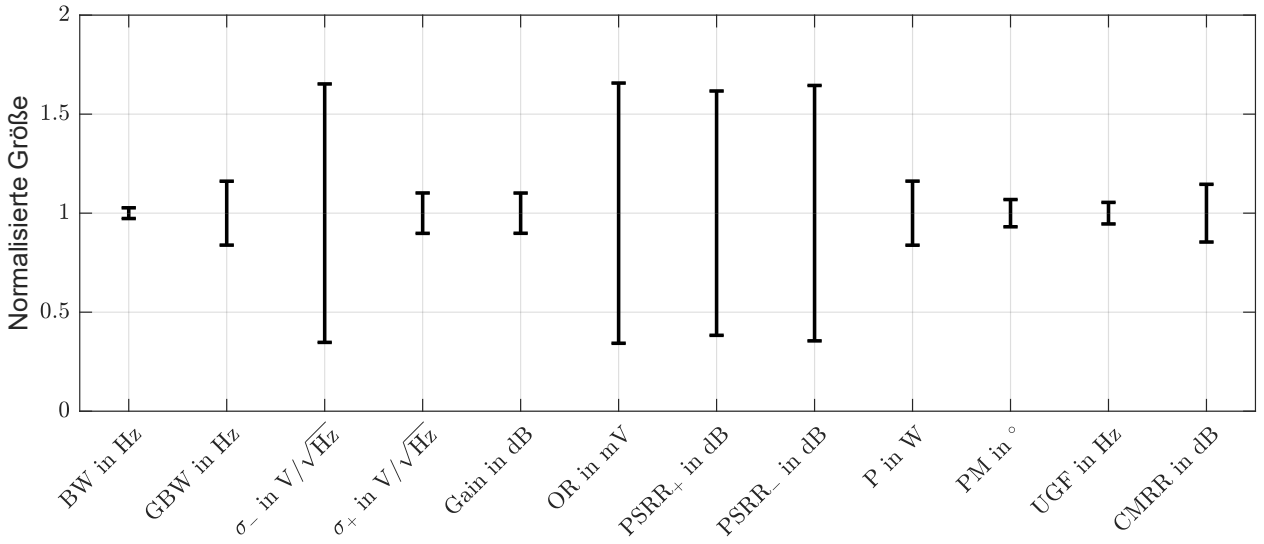


B.4 Folded-Cascode-Operationsverstärker



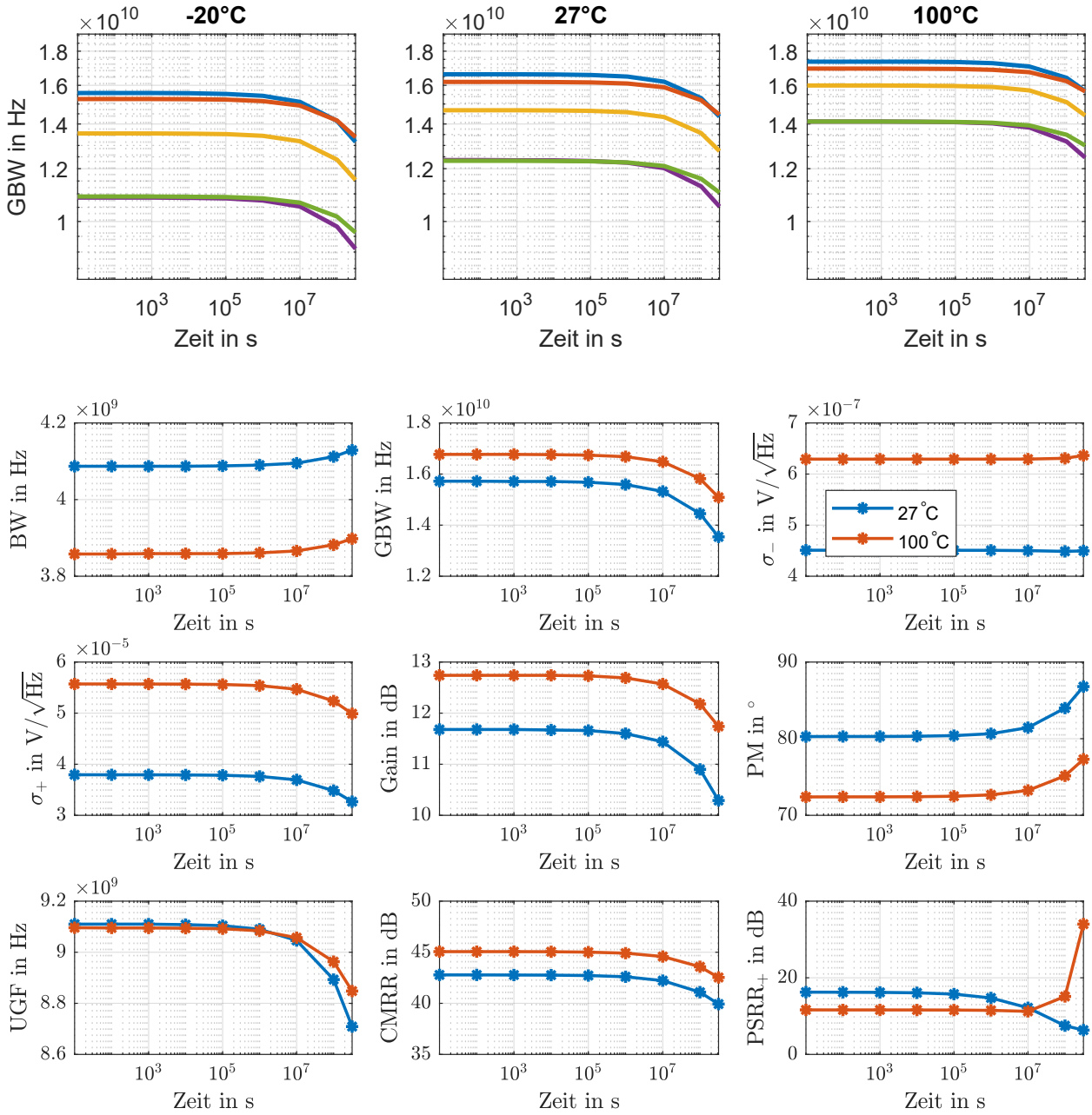
Durch die Kaskoden im Ausgangszweig sind die Arbeitspunkte der vier Transistoren stark von der Versorgungsspannung abhängig. Dies zeigt sich insbesondere beim PSRR gegenüber V_{dd}. Gleichzeitig ist die Alterung dieses Parameters stark von der Temperatur abhängig, dies lässt für diese FoM auf einen Einfluss durch BTI schließen. Gleiches gilt für die Output Range die sich sehr stark mit der Temperatur ändert. Die Wechselwirkung zwischen Temperatur, Alterung und Prozessvariationen sind durch die Kaskodeschaltungen eher gering und verstärken sich gegenseitig kaum. Es herrschen leichte Stressbedingungen für BTI in der Folded Cascode mit PMOS Transistoren, wodurch sich die Verstärkung und das CMRR im Laufe der Zeit ändert.



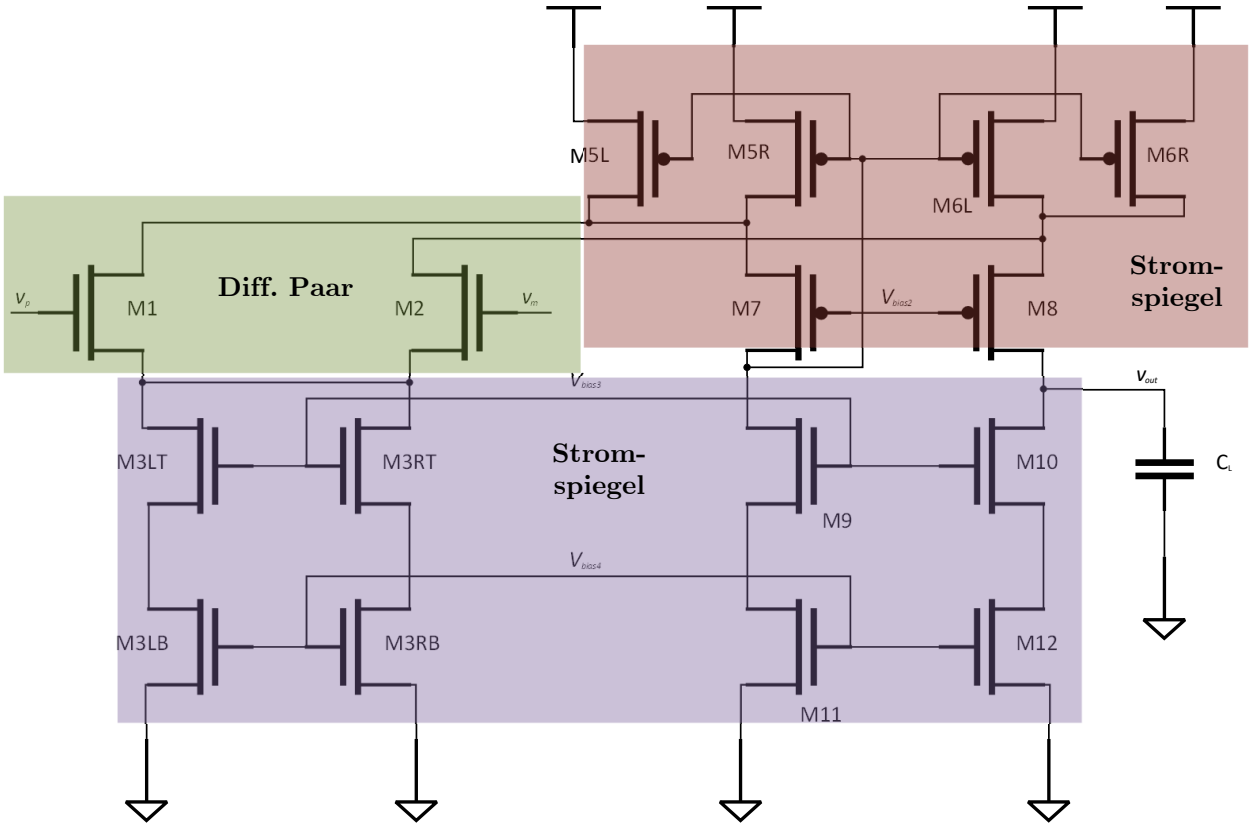


	BW in MHz	GBW in MHz	σ_- in $\mu V/\sqrt{\text{Hz}}$	σ_+ in $\mu V/\sqrt{\text{Hz}}$	Gain in dB	OR in mV
Mittelwert	4083	15910	0,536	38,55	11,69	147,1
Std.Abw.	110,5	2567	0,349	3,931	1,188	96,64

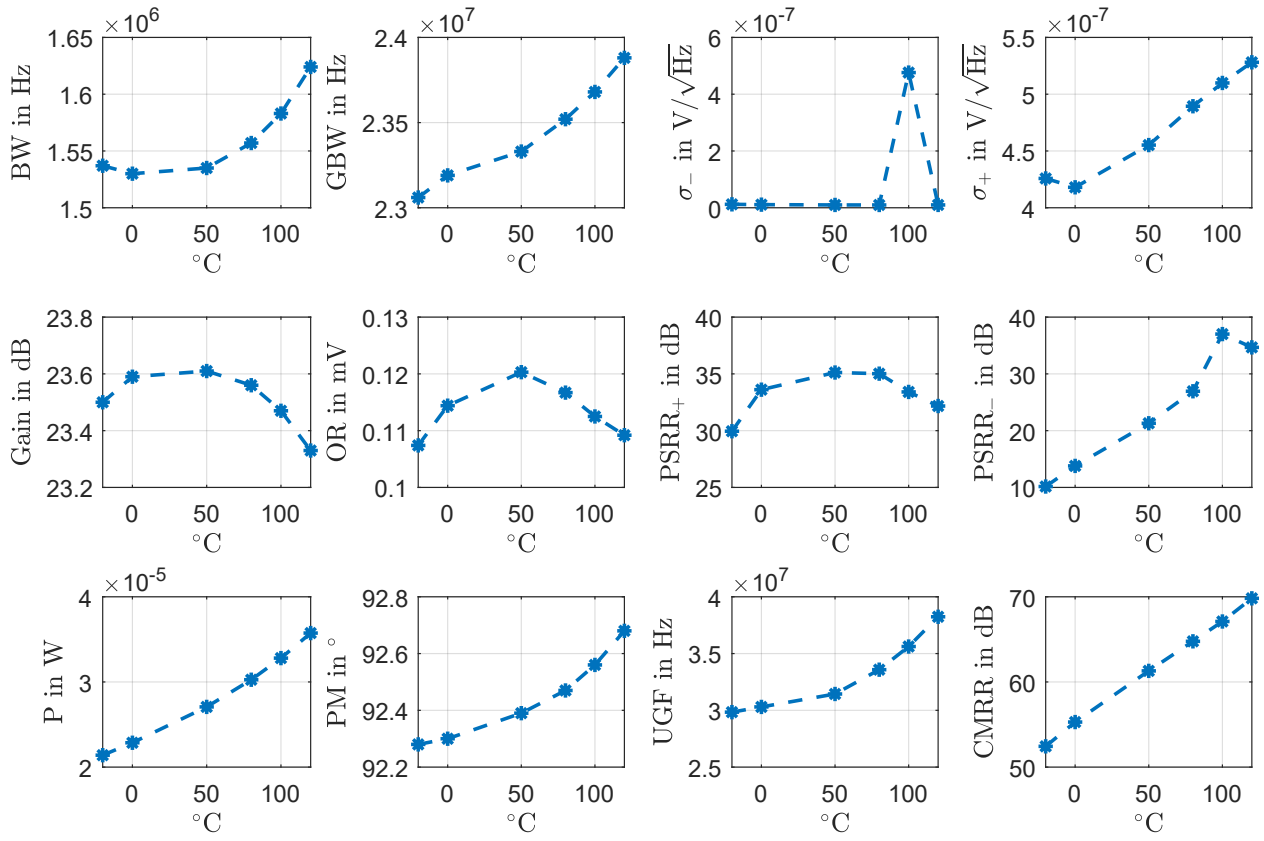
	PSRR$_-$ in dB	PSRR$_+$ in dB	P in μW	PM in °	UGF in MHz	CMRR in dB
Mittelwert	20,13	12,88	147,9	80,58	9038	44,43
Std.Abw.	12,98	7,945	23,92	5,542	491,9	6,476

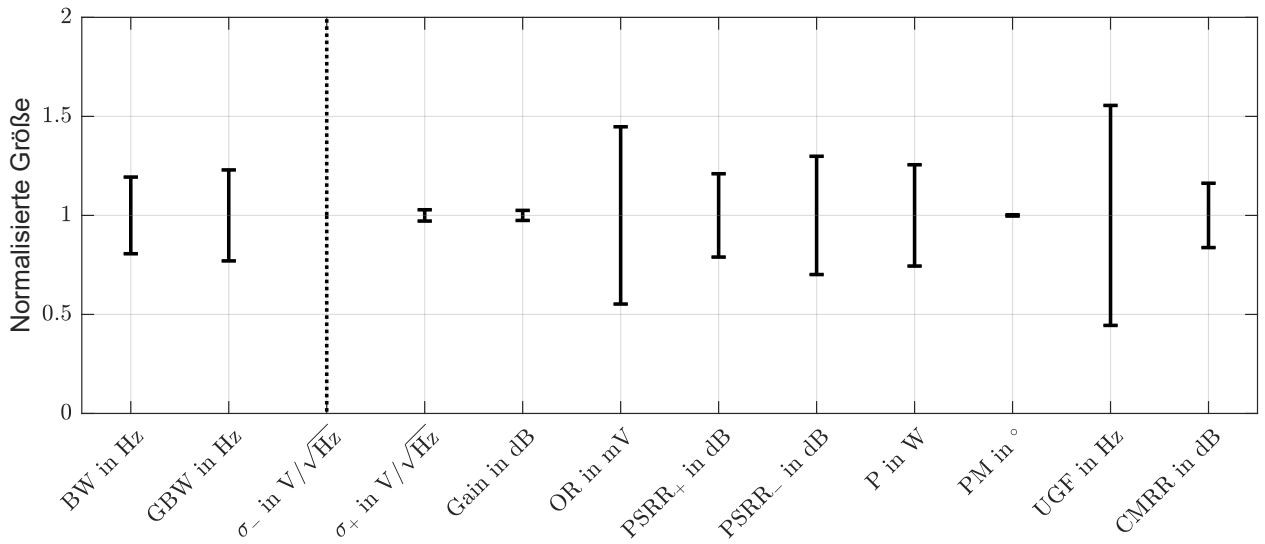


B.5 Folded-Cascode-OTA



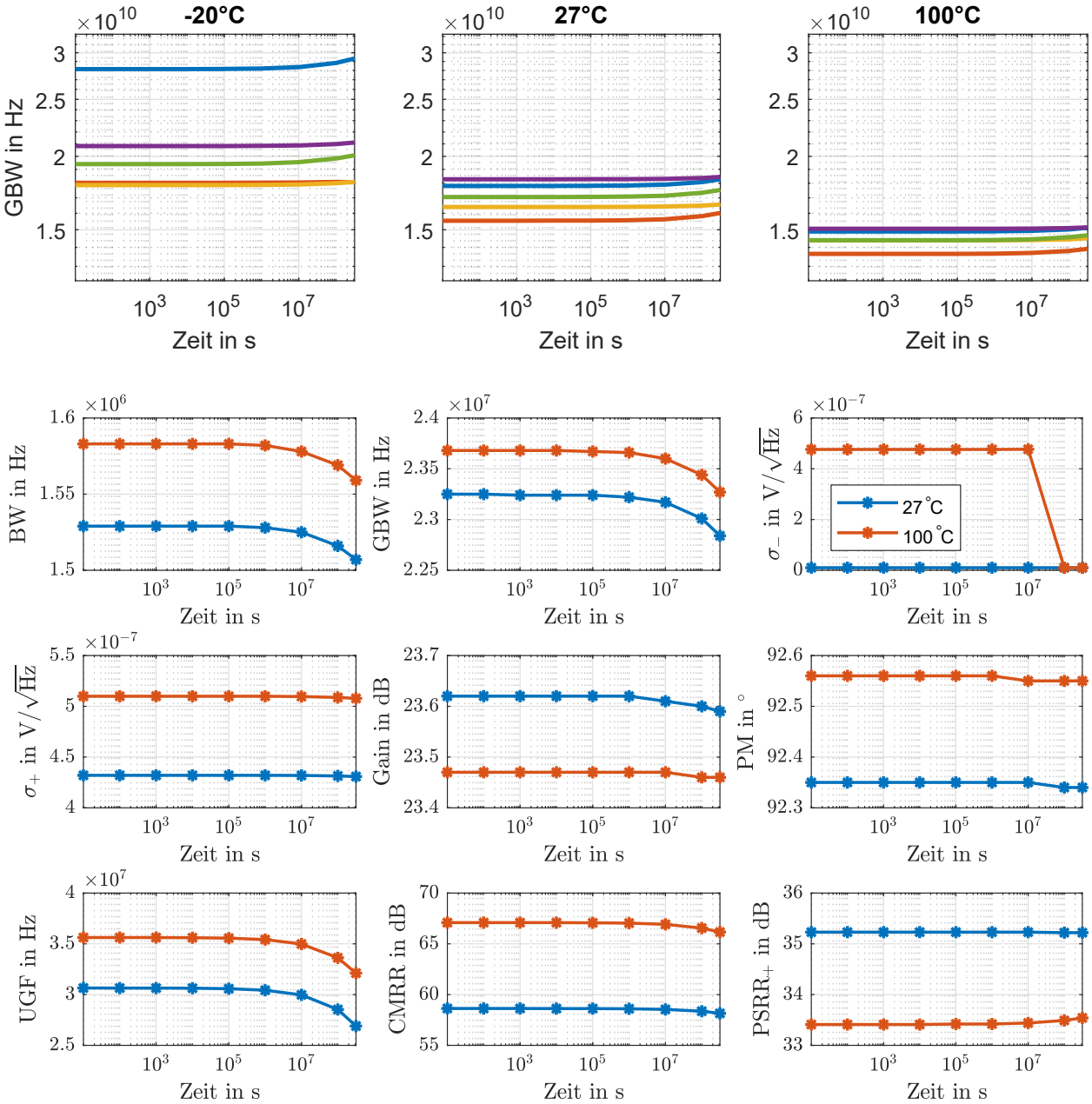
Dieser Verstärker besitzt eine Folded-Cascode am Ausgang und jeweils NMOS- und PMOS-Kaskoden im Ausgangszweig. Dadurch, dass eine NMOS-Kaskode verwendet wird, findet keine Alterung des Eingangspaars durch BTI statt. Dies zeigt sich in dem kaum von Temperatur und Alterung abhängigen Gain und in dem von der Alterung nicht betroffenen CMRR. Die Ouput Range ist durch die vier Transistoren im Ausgangszweig stark reduziert. Prozessvariationen haben durch die symmetrische Kaskodeschaltungen eine geringe Auswirkung.



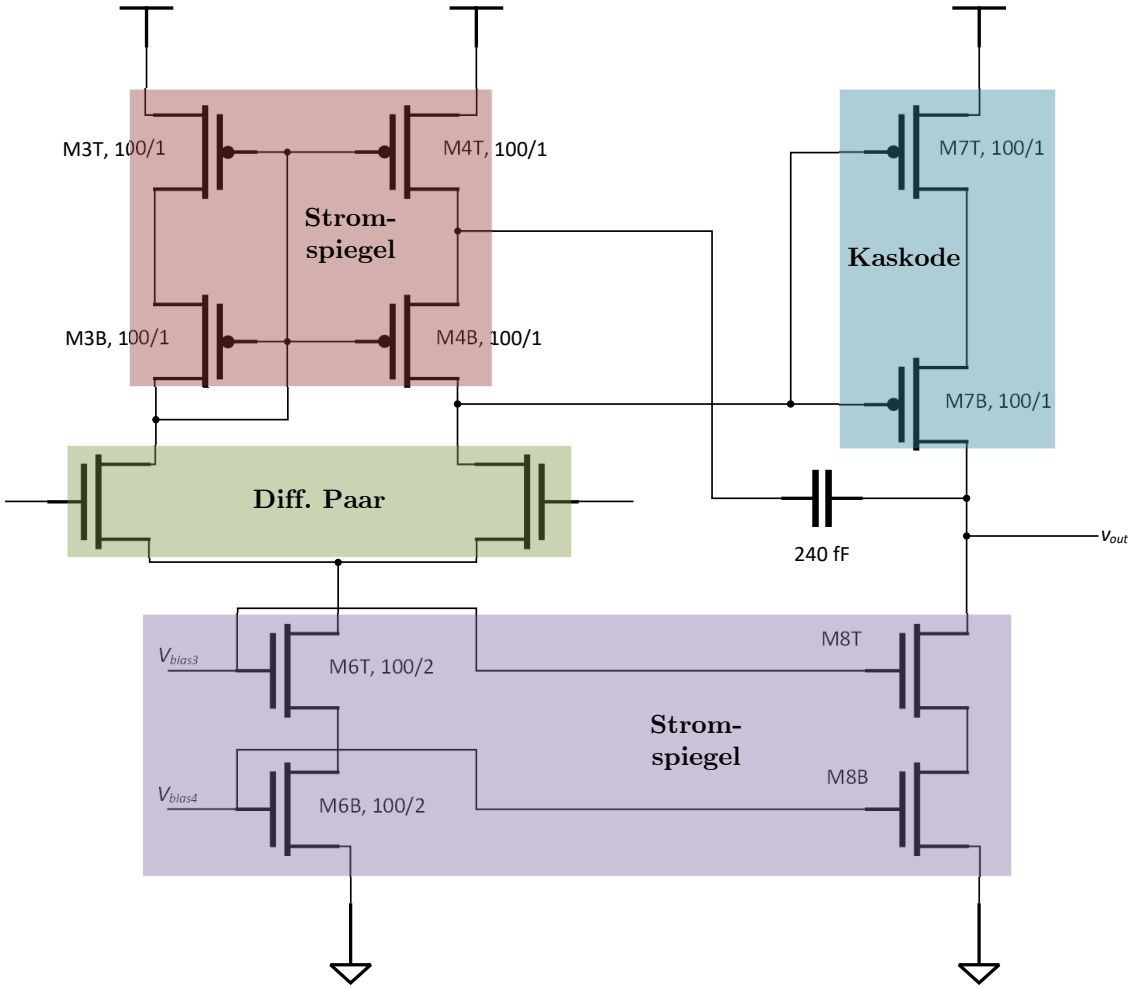


	BW in MHz	GBW in MHz	σ_- in $\mu V/\sqrt{\text{Hz}}$	σ_+ in $\mu V/\sqrt{\text{Hz}}$	Gain in dB	OR in mV
Mittelwert	1,601	24,15	0,123	0,431	23,48	98,97
Std.Abw.	0,309	5,548	0,183	0,122	0,596	44,25

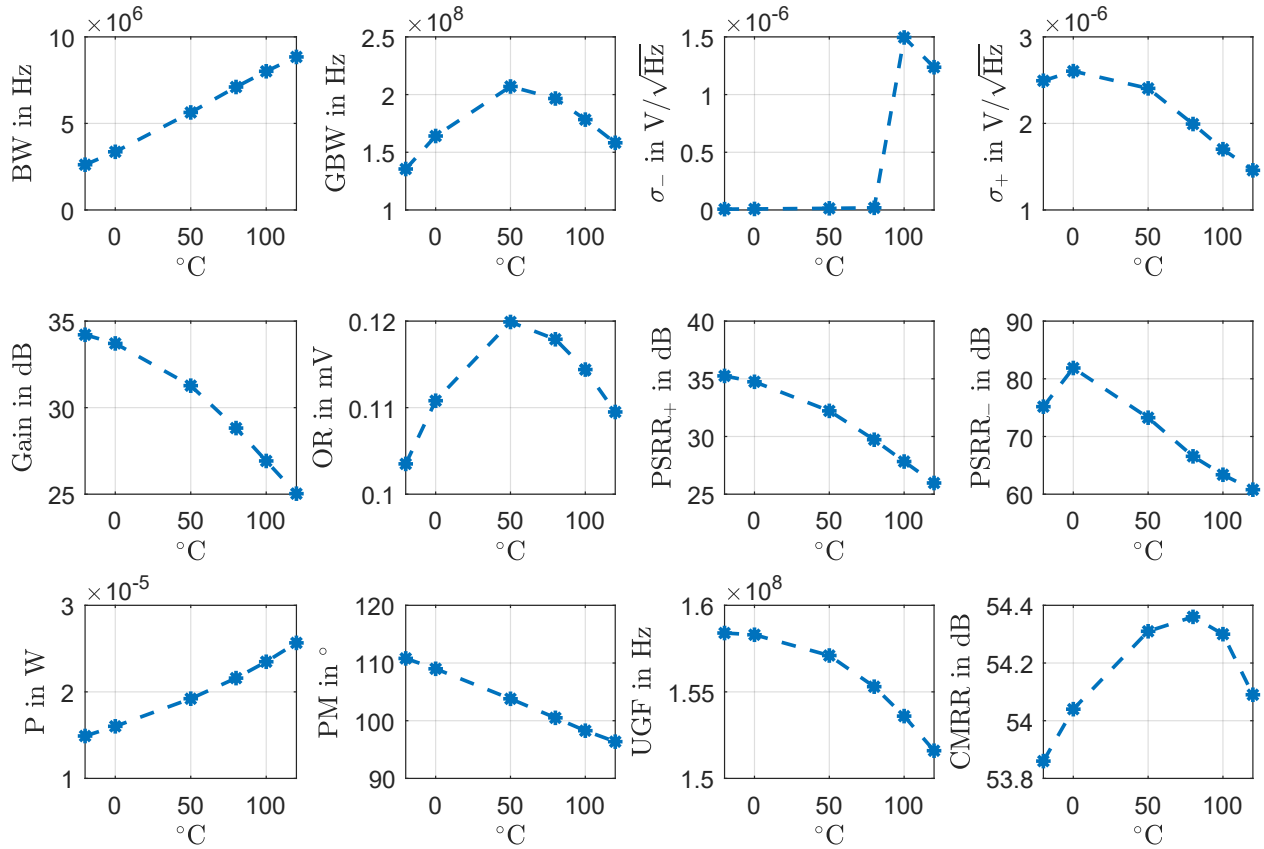
	PSRR$_-$ in dB	PSRR$_+$ in dB	P in μW	PM in °	UGF in MHz	CMRR in dB
Mittelwert	19,28	36,64	26,31	92,46	39,56	60,84
Std.Abw.	5,757	7,708	6,73	0,304	21,97	9,884

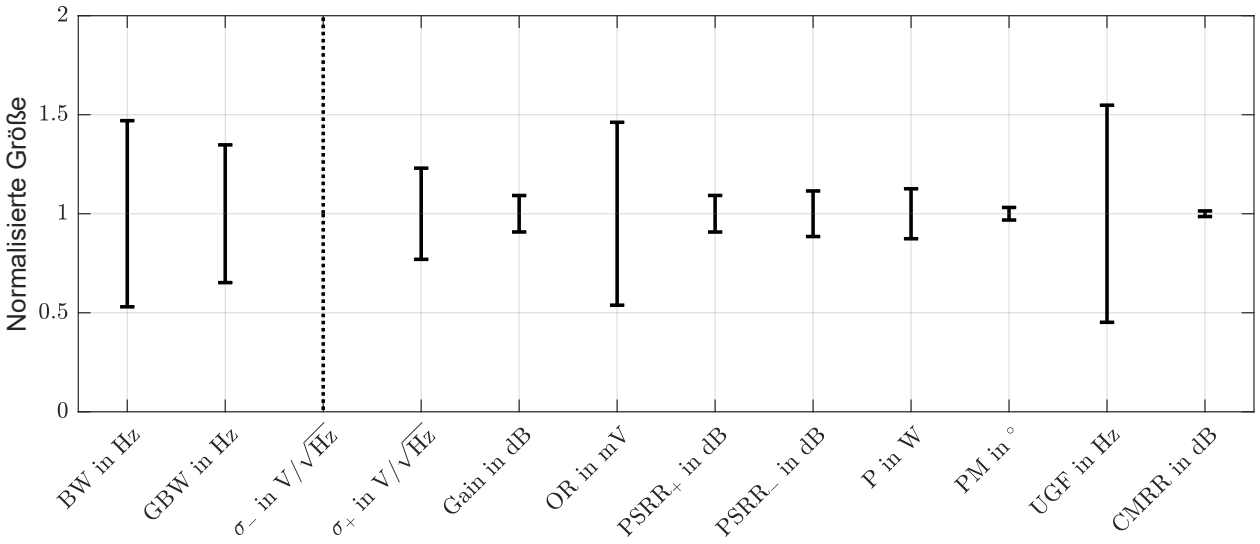


B.6 Zweistufiger Verstärker mit Kompensation



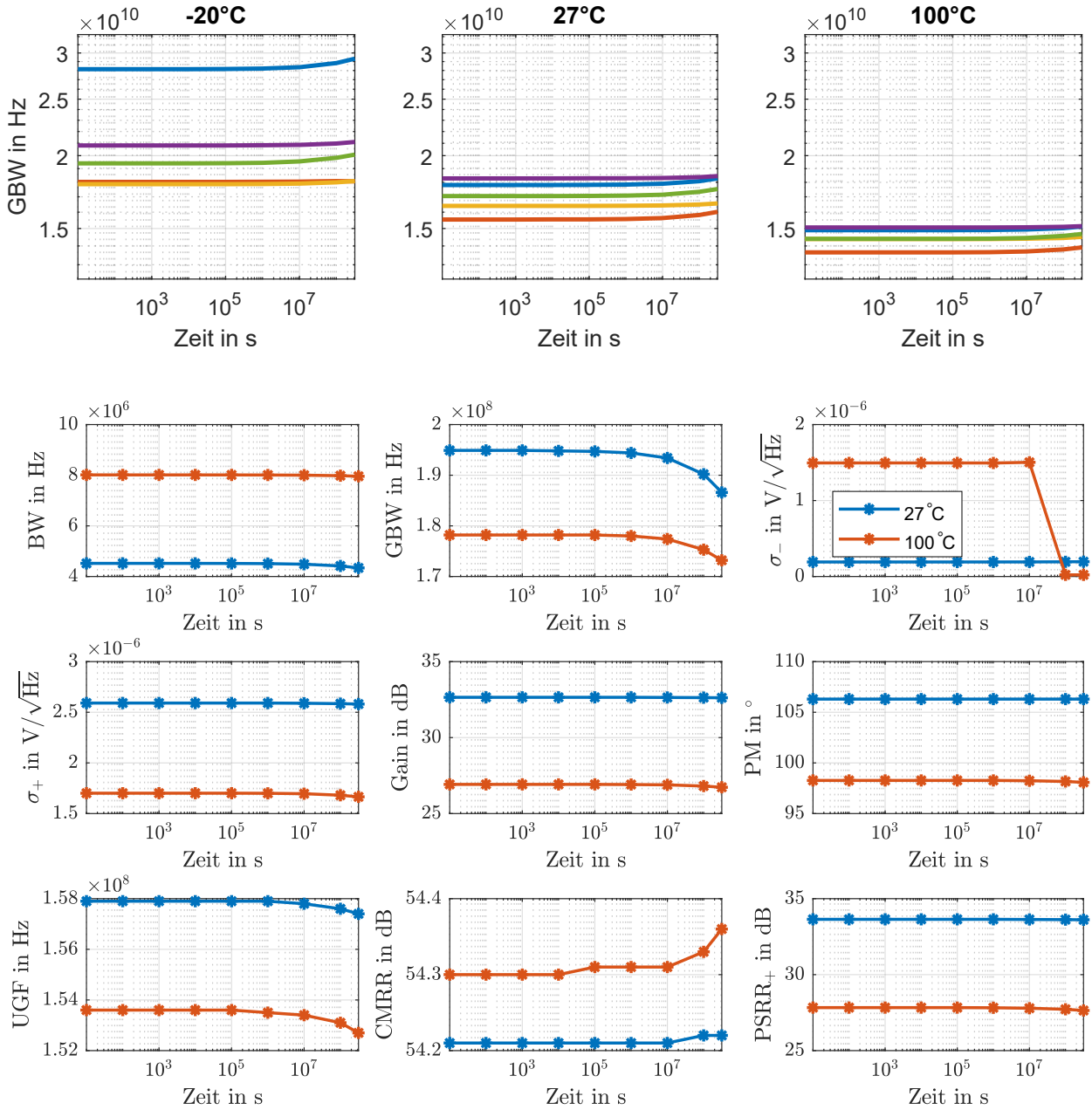
Die Topologie dieses Verstärkers ist sehr ähnlich zu der des Miller-OTA. Der entscheidende Unterschied ist jedoch, dass alle Stromspiegel als Kaskode ausgeführt sind. Somit zeigen sich etwas höhere Einflüsse durch Prozessvariationen, jedoch deutlich geringere Auswirkungen der Alterung. Dies verdeutlicht, dass alleine die Wahl der Topologie direkte und sehr starke Auswirkungen auf das gesamte Verhalten einer Schaltung gegenüber einzelnen Einflüssen hat. Die sehr gute Widerstandsfähigkeit gegenüber Alterung über alle FoMs hinweg resultiert in einer höheren Anfälligkeit durch Temperatureinwirkungen. Dies liegt vor allem an der starken gegenseitigen Abhängigkeit der Sättigung der Transistoren. Bei vier Transistoren ist bereits bei $V_{dd} = 1$ V $V_{ds} \approx V_{ds,sat}$ erreicht und somit kann sich der Arbeitspunkt des Transistors bereits bei kleinen Änderungen der Biasbedingungen verschieben. Dadurch ist das PSRR relativ gering, jedoch zeigt es ebenfalls keine Änderung über das Alter. Somit zeigt sich, dass die Arbeitspunkte der Kaskoden zwar von anderen Einflüssen, jedoch nicht durch Alterung verändert werden. Dies ist der Fall, solange die Biasspannungen der Kaskodetransistoren entsprechend alterungsresistent gewählt werden.



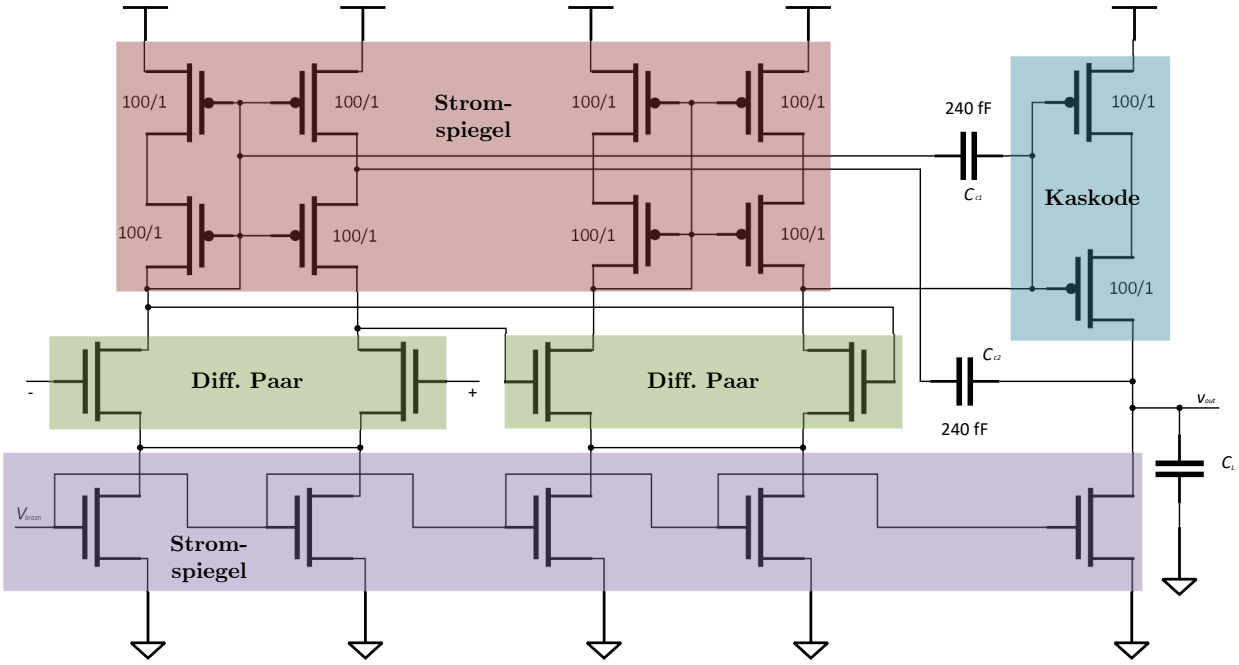


	BW in MHz	GBW in MHz	σ_- in $\mu V/\sqrt{\text{Hz}}$	σ_+ in $\mu V/\sqrt{\text{Hz}}$	Gain in dB	OR in mV
Mittelwert	5,135	183,8	0,164	2,359	31,43	98,4
Std.Abw.	2,413	63,92	0,530	0,543	2,894	45,44

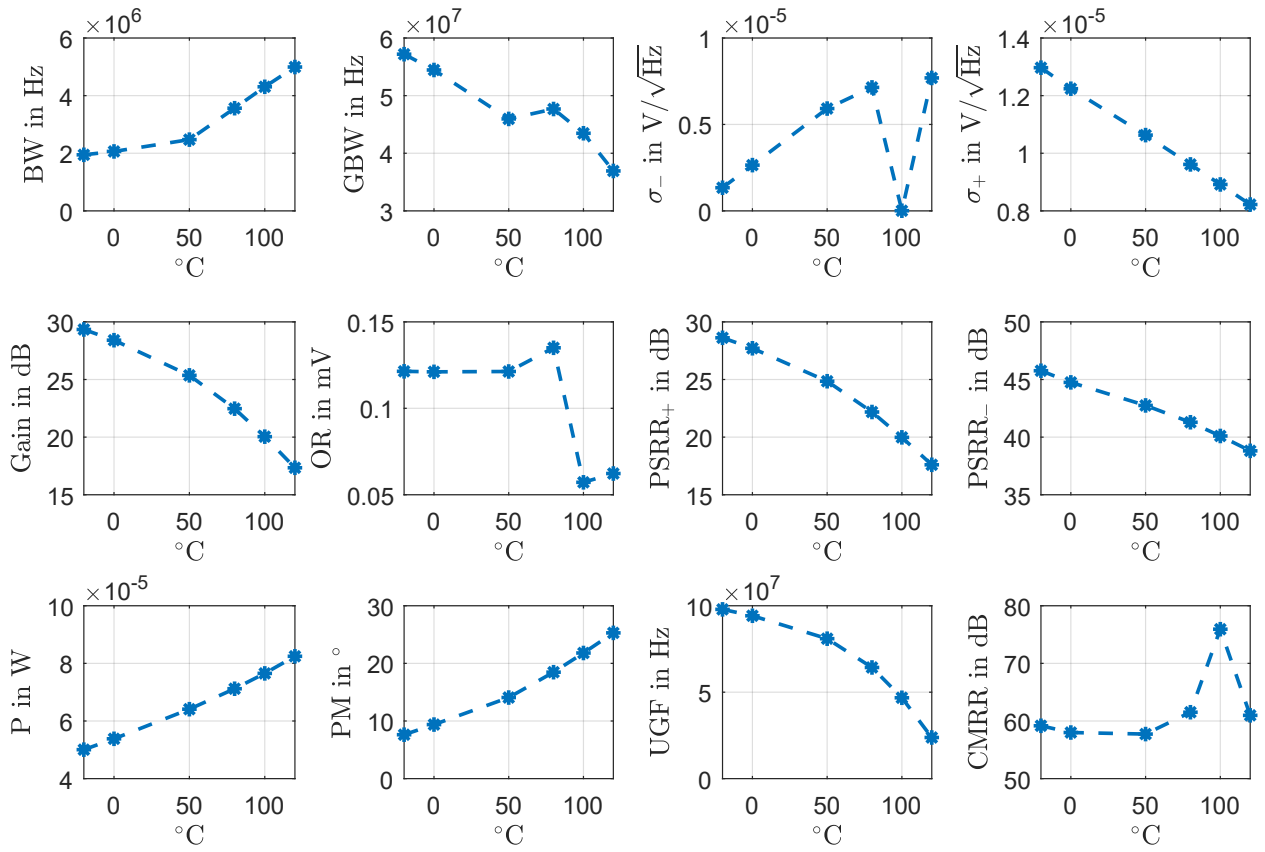
	PSRR$_-$ in dB	PSRR$_+$ in dB	P in μW	PM in °	UGF in MHz	CMRR in dB
Mittelwert	80,54	32,4	15,89	105,3	224,4	54
Std.Abw.	9,268	2,991	2,004	3,341	123	0,765

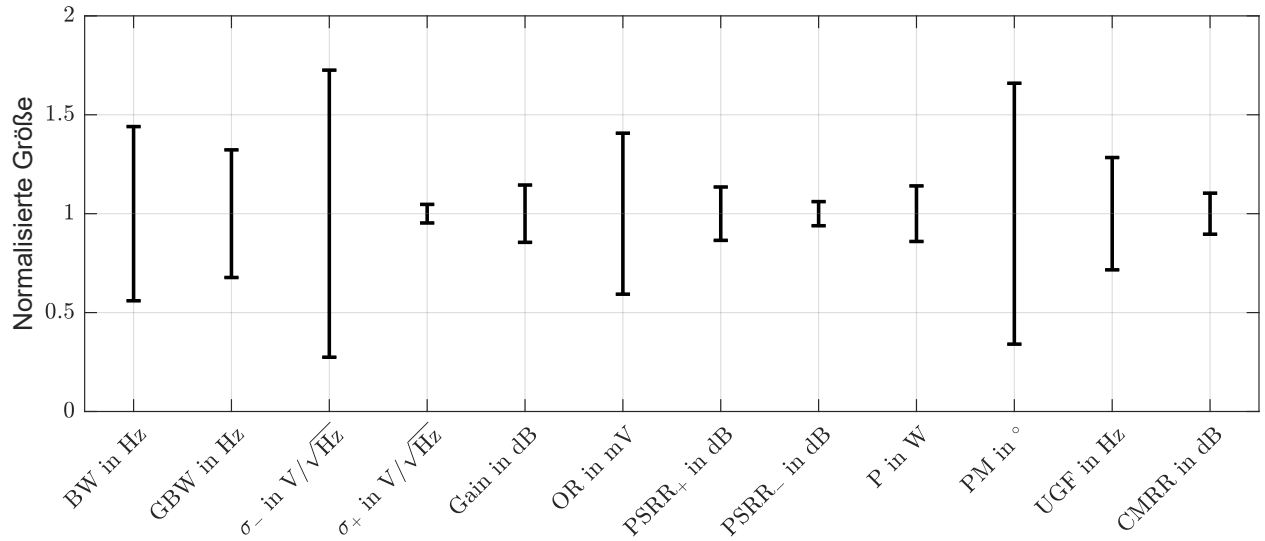


B.7 Dreistufiger Verstärker



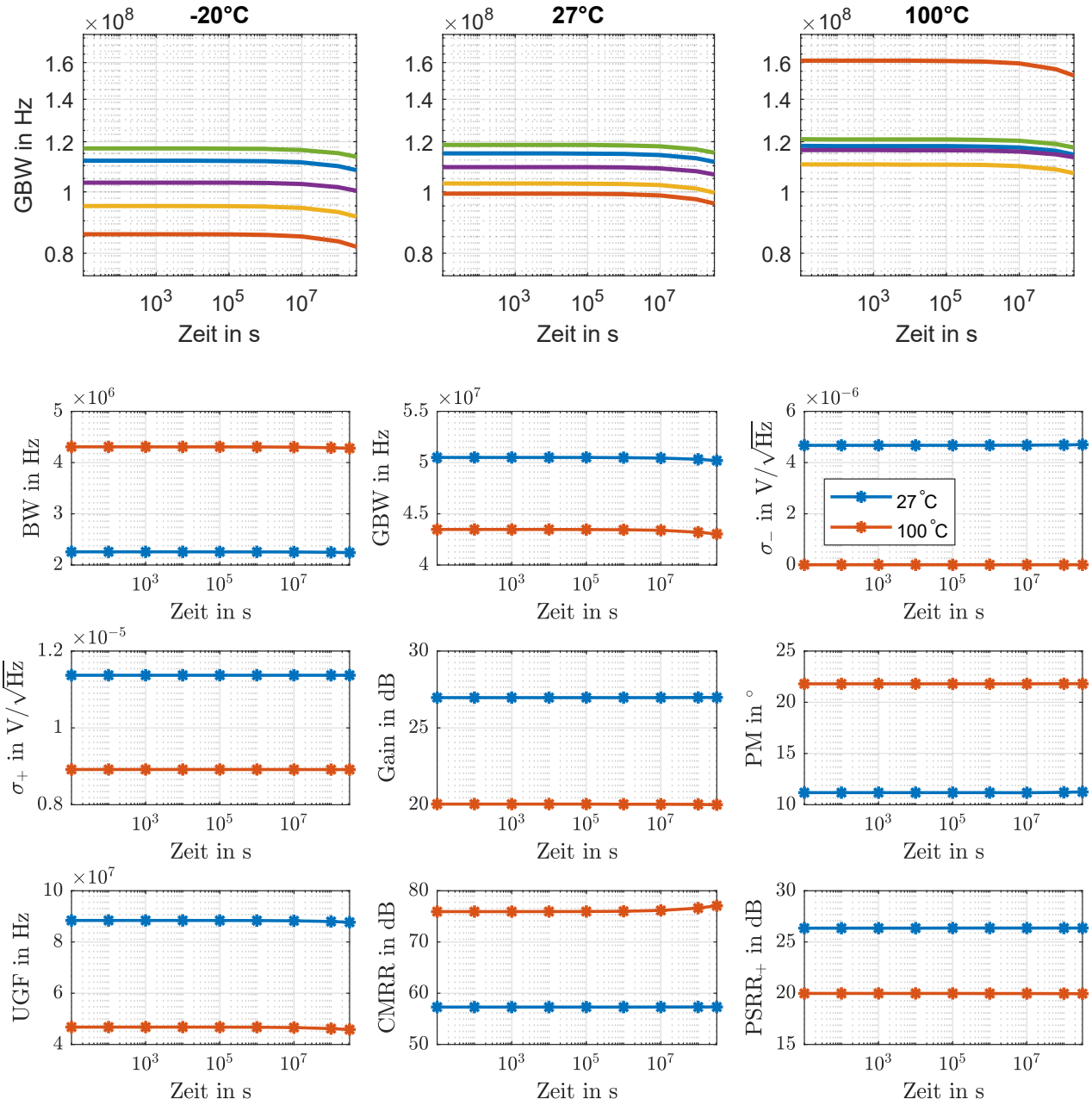
Als Referenz wurden bei dieser Topologie nur die NMOS-Transistorkomponenten jeweils als Kaskode ausgeführt. Wie zu erwarten ist, sinkt der Einfluss der Prozessvariation durch die geringe V_{ds}-Abhängigkeit der Transistoren. Gleichzeitig ist die Temperaturabhängigkeit im Vergleich zu den beiden vorherigen Topologien reduziert, jedoch immer noch stärker als bei der ersten Topologie (Miller-OTA) ohne Kaskoden. Kennzahlen wie CMRR und Gain zeigen keine Alterung, da diese maßgeblich durch die NMOS Transistoren, welche als Kaskode ausgeführt werden, abhängig sind. Für die FoM, die sich über dem Alter verändern (z.B. PM, UGF, PSRR,) zeigt sich, dass sich die Alterung über der Temperatur ändert. Somit ist BTI, also die PMOS Transistoren, maßgeblich für die Veränderung verantwortlich. Somit können mit der Blockanalyse ebenfalls einzelne Teile der Schaltung bzw. spezifische Eigenschaften, so lange die funktionalen Abhängigkeiten bekannt sind, alterungsresistent durch die Wahl der Topologie beeinflusst werden.



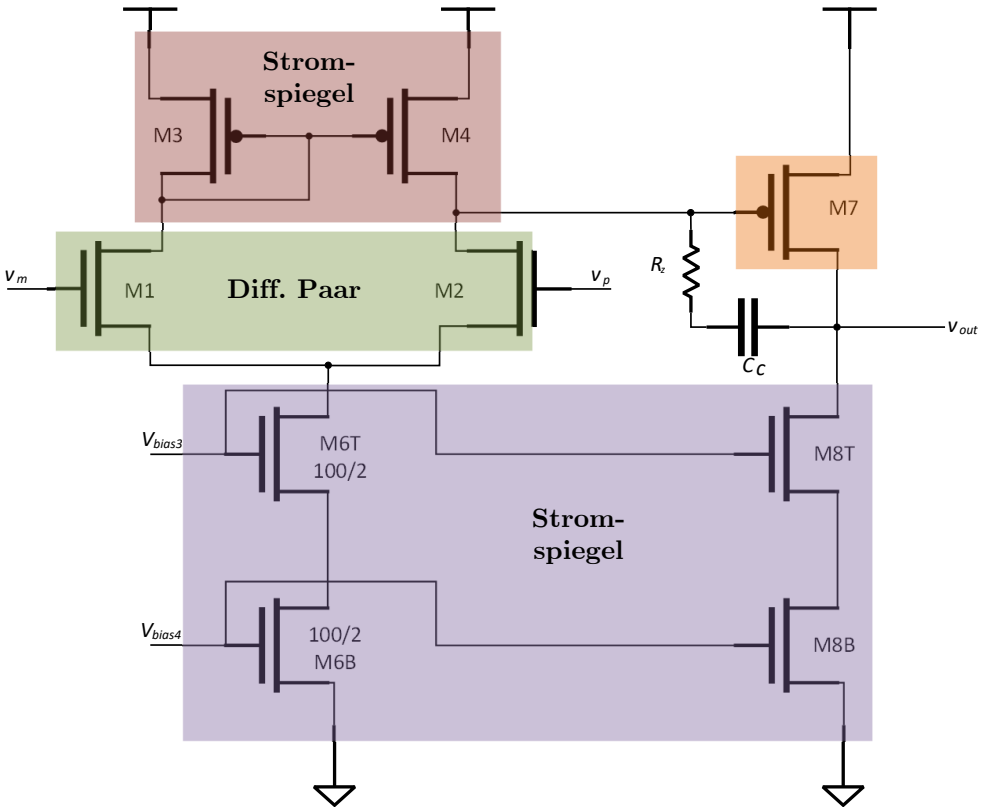


	BW in MHz	GBW in MHz	σ_- in $\mu V/\sqrt{\text{Hz}}$	σ_+ in $\mu V/\sqrt{\text{Hz}}$	Gain in dB	OR in mV
Mittelwert	2,855	53,37	4,414	11,46	25,59	103,9
Std.Abw.	1,257	17,22	3,203	0,538	3,707	42,28

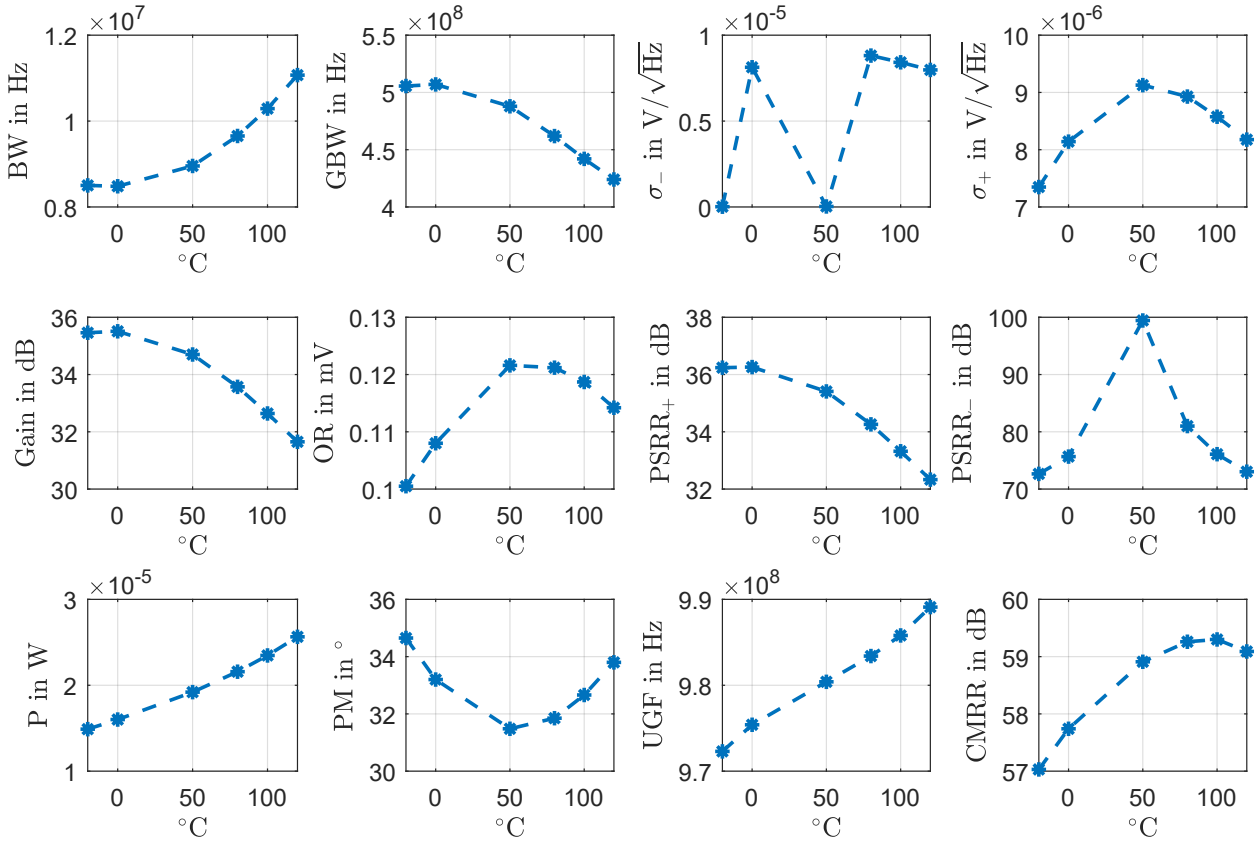
	PSRR$_-$ in dB	PSRR$_+$ in dB	P in μW	PM in °	UGF in MHz	CMRR in dB
Mittelwert	43,3	25,09	59,2	14,08	82,04	58,66
Std.Abw.	2,64	3,384	8,316	9,286	23,29	6,087

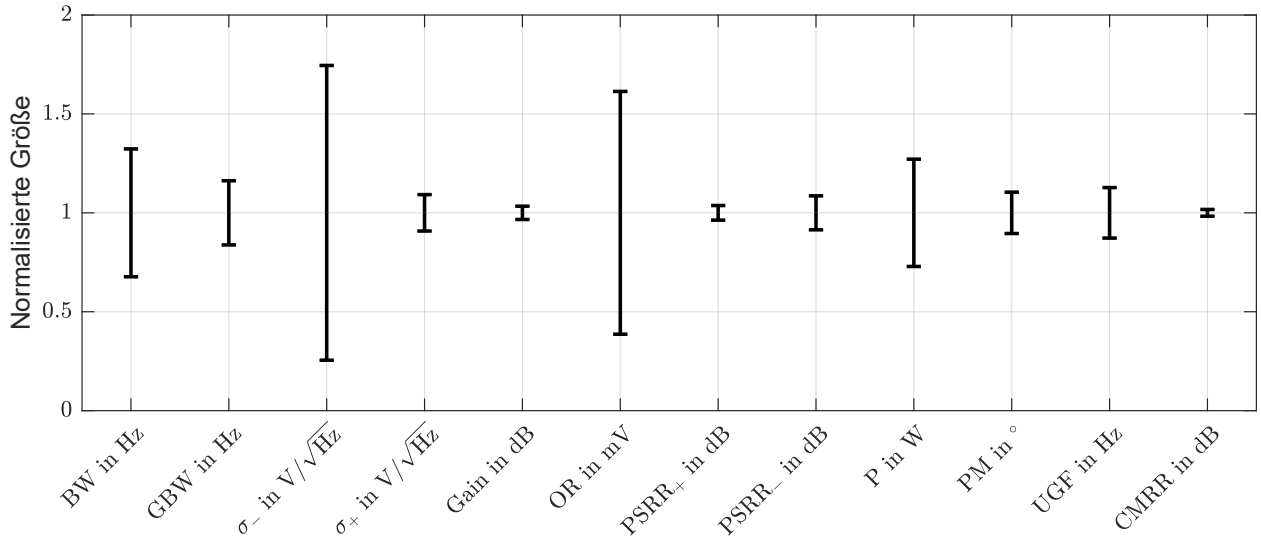


B.8 Zweistufiger Verstärker



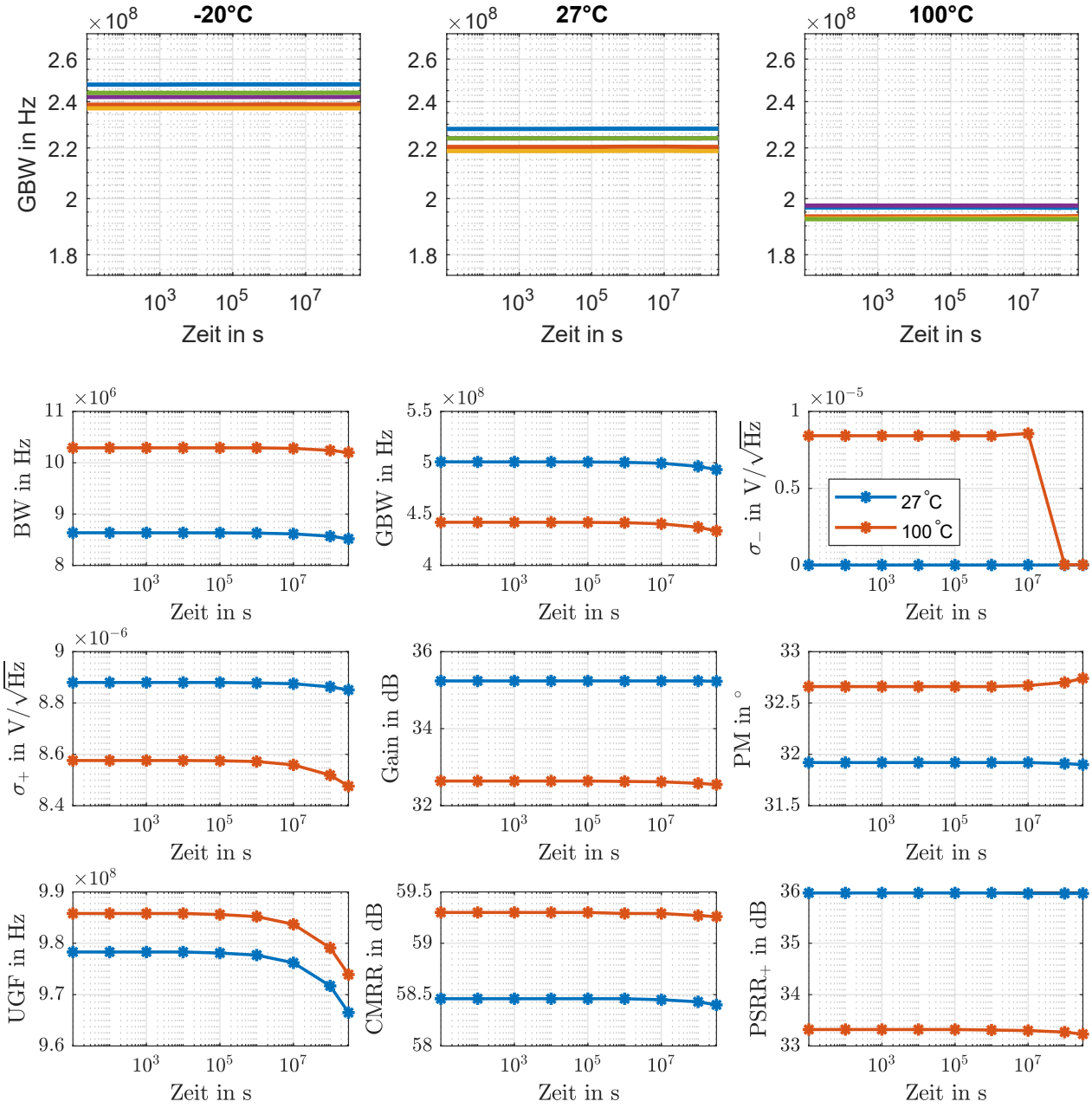
Dieser zweistufige Verstärker folgt derselben Idee wie der vorherige dreistufige Verstärker. Daher zeigt sich die gleiche Alterungsresistenz. Somit lässt sich sagen, dass sich die Ergebnisse der Blockanalyse zum einen auf eine Topologie übertragen lassen. Es lassen sich die für Alterung sensitiven Elemente ermitteln und konsequent durch die Anwendung der Erkenntnisse aus der Blockanalyse beheben. Zum anderen können mit diesem Vorgehen bestehende Topologien direkt um weitere Stufen erweitert werden, die den gleichen Paradigmen folgen. Gleichzeitig ist im Vergleich zwischen zweistufigen und dreistufigen Verstärker zu erkennen, dass die Widerstandsfähigkeit mit jedem weiteren resistenten Schaltungsblock erhöht wird. Dieses Verhalten konnte im Rahmen der Fallstudien mit der begrenzten Anzahl von Topologien nicht verallgemeinert werden.





	BW in MHz	GBW in MHz	σ_- in $\mu V/\sqrt{\text{Hz}}$	σ_+ in $\mu V/\sqrt{\text{Hz}}$	Gain in dB	OR in mV
Mittelwert	9,188	496	5,435	8,488	34,73	89,87
Std.Abw.	2,969	80,45	4,049	0,781	1,168	55,15

	PSRR$_-$ in dB	PSRR$_+$ in dB	P in μW	PM in °	UGF in MHz	CMRR in dB
Mittelwert	87,29	35,43	21,06	32,83	962,4e6	58,24
Std.Abw.	7,494	1,309	5,711	3,433	122,9e6	0,998



B Abkürzungsverzeichnis

Abkürzung	Bezeichnung
BEOL	Backend of line
BiCMOS	Bipolar Complementary metal-oxide-semiconductor
BSIM	Berkeley Short-channel IGFET Model
BTI	Bias Temperature Instability
BW	Bandwidth
CAD	Computer Aided Design
cdf	cumulative distribution function
CMC	Compact Modeling Council
CMOS	Complementary metal-oxide-semiconductor
CMRR	Commonmode Rejection Ratio
COTS	commercial off-the-shelf
CS	Common Source
DIBL	Drain Induced Barrier Lowering
DIN	Deutsches Institut für Normung
DUT	Device Under Test
EKF	Extended Kalman Filter
EKV	Enz Krummenacher Vittoz
EM	Elektromigration
EOL	End of Life
FDSOI	Fully Depleted Silicon in Insulator
FEM	Finite Elemente Methode
FEOL	Frontend of line
FET	Fin Field Effect Transistor
FinFET	Fin Field Effect Transistor
FoM	Figure of Merit
GBW	Gain-Bandwidth-Product
HALT	Highly Accelerated Lifetime Test

HCI	Hot Carrier Injection
HiSIM	Hiroshima-University STARC IGFET Model
IEC	International Electrotechnical Commission
IGFET	Insulated-Gate Field-Effect Transistor
JEDEC	Joint Electron Device Engineering Council
LET	Linear Energy Transfer
MCMC	Markov Chain Monte Carlo
MI	Medium Inversion
MIL-STD	military standard
MOSFET	Metal-Oxid-Semiconductor Fieldeffect Transistor
MOSRA	MOS Reliability Analysis
MVE	Multi Vibration Excitation
PDF	probability distribution function
pdf	probability density function
PDK	Process Design Kit
PM	Phase Margin
PSP	Penn State Philips transistor model
PSRR	Power Supply Rejecion Ratio
PVTA(R)	Process Voltage Temperature Aging (Radiation)
REL	Reliability
RUL	Remaining Useful Lifetime
SC	Scatter Event Excitation
SE	Single Event Excitation
SEU	Single Event Upset
SI	Strong Inversion
SPICE	Simulation Program with Integrated Circuit Emphasis
SR	Slew Rate
SRAM	Static Random Access Memory
TCAD	Technology Computer Aided Design
TDDB	Time Dependent Dielectric Breakdown
TID	Total Ionizing Dose
UDRM	user-defined reliability model
UGF	Unity Gain Frequency
UKF	Unscented Kalman Filter
WI	Weak Inversion

B Symbolverzeichnis

Physikalische Größen

Symbol	Bezeichnung
q	Einheitsladung
t	Zeit
T	Temperatur
k_b	Boltzmann-Konstante
N_d	Dotierkonzentration am Drain
N_s	Dotierkonzentration am Source
N_b	Dotierkonzentration im Bulk
n_i	intrinsische Ladungsträger
N_C	Ladungsträger im Leitungsband
N_V	Ladungsträger im Valenzband
ε_0	Permittivität des Vakuums
ε_{ox}	relative Permittivität des Oxids
ε_{box}	relative Permittivität des *buried* Oxids
ε_{si}	relative Permittivität des Siliziums
μ	Ladungsträgerbeweglichkeit
μ_0	Referenz-Ladungsträgerbeweglichkeit
v_{sat}	Sättigungsgeschwindigkeit der Ladungsträger
ρ	Ladungsträgerdichte

Spannungen

Symbol	Bezeichnung
V_{dd}	Versorgungsspannung
V_{ss}	Bezugspotential
V_g	Gate-Spannung
V_{gs}	Gate-Source-Spannung

V_s	Source-Spannung
V_d	Drain-Spannung
V_{ds}	Drain-Source-Spannung
V_b	Bulk-Spannung
V_{FB}	Flachbandspannung
V_{TB}	Schwellfunktion
V_{th}	Schwellspannung
ΔV_{th}	Änderung der Schwellspannung
V_{ref}	Referenzspannung
$V_{ds,eff}$	effektive Drain-Source-Spannung
$V_{ds,sat}$	Drain-Source Sättigungsspannung
V_{bias}	Biasspannung
V_{mon}	Monitorspannung
V_{ref}	Referenzspannung
V_{ox}	Oxidspannung
V_c	Kanalspannung
V_{sub}	Substratspannung
U_T	thermische Spannung
$v_\square$	auf U_T normierte Spannung
v_{ini}	initiale Spannung zur q_i Berechnung

Potentiale und Felder

Symbol	**Bezeichnung**
Ψ_p	Abschnürpotential
Ψ	allgemeines Potential
ψ	normiertes Potential
Ψ_s	Oberflächen-Potential
Φ_b	Potential im Bulk
Φ_n	Potential der Ladungsträger
Ψ_m	mittleres Oberflächen-Potential
Φ_F	Ferminiveau
Φ_{fn}	Quasi-Ferminiveau
ϕ	normiertes Potential
Φ_s	Source-Potential
Φ_d	Drain-Potential
E_{ox}	elektrisches Feld im Oxid
E_z	elektrisches Feld im Kanal in z-Richtung

Energien

Symbol	Bezeichnung
Φ_m	Austrittsarbeit im Metall
Φ_s	Austrittsarbeit im Silizium
Φ_{ms}	Austrittsarbeitsunterschied
$E_{F,eq}$	Fermienergie im Gleichgewicht
$E_{F,in}$	intrinsische Fermienergie
E_F	Fermienergie
E_C	Energie der Leitungsbandkante
E_V	Energie der Valenzbandkante
E_G	Energie der Bandlücke
E_A	Aktivierungsenergie

Ströme

Symbol	Bezeichnung
I_{ds}	Drain-Source-Strom
$I_{ds,0}$	Bezugs-Drain-Source-Strom
I_d	Drain-Strom
ΔI_{ds}	Änderung des Drain-Source-Stroms
I_{ref}	Referenzstrom
I_{Bias}	Biasstrom
I_{spec}	Bezugsstrom
I_s	Sperrstrom
I_{th}	Schwellstrom
I_{th}	Schwellstrom in Sättigung
I_{SEU}	Durch SEU hervorgerufener Strom
J_n, J_p	Stromdichten der Elektronen und Löcher durch SEU

Ladungen

Symbol	Bezeichnung
Q_g	Gate-Ladung
Q_{fc}	Fehlstellen-Ladung
Q_i	Inversionsladung
Q_b	Bulk-Ladung
Q_s	Source-Ladung
Q_{si}	Ladung im Silizium

Q_{spec}	Bezugsladung
Q_{it}	Grenzschicht-Ladung
Q_{trap}	Ladungsträgerfallen-Ladung
Q_{ox}	Ladungsträgerfallen-Ladung im Oxid
Q_{age}	Ladung durch Alterung
Q_{HCI}	Ladung durch HCI
Q_{BTI}	Ladung durch BTI
q_i	normierte Inversionsladung
q_{ini}	normierte initiale Inversionsladung
q_{new}	Ladung durch Newton-Verfahren berechnet
q_s	normierte Source-Ladung
q_d	normierte Drain-Ladung
$q_{ds,sat}$	Drain-Ladung beim Erreichen der Sättigung
N_{age}	Ladungsträgerkonzentration durch Alterung
N_{it}	Ladungsträgerkonzentration an der Grenzschicht
N_{trap}	Ladungsträgerkonzentration der Fallen
D_{it}	Ladungsträgerdichte durch TID

Transistorparameter

Symbol	**Bezeichnung**
Γ_b, γ	Substratfaktor
γ_b	normierter Substratfaktor
n	Steigungsparameter
n_0	Referenz-Steigungsparameter
β	Transferparameter
IC	Inversionskoeffizient
g_m	Transkonduktanz
g_{ds}	Drain-Source-Leitwert
r_{out}	Ausgangswiderstand
g_m/I_d	Transkonduktanzeffizenz
λ_c	Kanallängenmodulationsfaktor
k	Stromübersetzungverhältnis
C_{ox}	Oxidkapazität
C'_{ox}	Flächenbezogene Oxidkapazität
μ_{eff}	Effektive Ladungsträgerbeweglichkeit

Geometrische Größen

Symbol	Bezeichnung
W	Kanalweite
W_{min}	kleinstmögliche Kanalweite
W_{eff}	effektive Weite
W_{dep}	Tiefe des Inversionskanals
L	Kanallänge
L_{min}	kleinstmögliche Kanallänge
L_{eff}	effektive Länge
L_D	Debye-Länge
t_{ox}	Dicke des Oxids
t_{box}	Dicke des *buried* Oxids
t_{si}	Dicke des Substrats
A_{it}	Aktive Transistorfläche

Zeitkonstanten

Symbol	Bezeichnung
τ_s	Stresszeitkonstante
τ_r	Wiederherstellungszeitkonstante
τ_{sim}	Simulationszeitraum
τ_e	BTI Fehlstellenkonstante
τ_c	BTI Fallenkonstante
t_r	Wiederherstellungszeitraum
t_s	Stresszeitraum
k, T	Diskrete Zeitpunkte
τ	Zeitraum

Stochastik

Symbol	Bezeichnung
μ	Mittelwert
σ	Standardabweichung
C_{xy}	Kreuzkorrelation
ρ	Korrelationskoeffizient
X_L, F	Zufallsvariable
x_L, f	pdf der Zufallsvariable
P	Wahrscheinlichkeit

λ	Fehlerrate
S	Überlebensfunktion
h	Ausfallrate
R	Zuverlässigkeit

Sonstige Symbole

Symbol	**Bezeichnung**
x, y, z	kartesische Raumrichtungen
C_1, C_2, C_3	Normierungsgrößen HCI - Modell
α	Fittingparameter
$p_0, ..., p_5$	Polynomkoeffizienten
$\mathbb{T}$	Parameter-Tupel
$\mathcal{P}$	Parametersatz
FR	Beschleunigungsparameter
r	Radius des SEU
X	Abstand zum SEU
$c_{\mathcal{P}}$	Parameter des Pelgrom-Modells
A, B, C, D, E	empirische Parameter des Eyring-Modells
S_1, S_2	Modellparameter des Eyring-Modells

B Wissenschaftliche Veröffentlichungen

Die folgenden Listen enthalten die wissenschaftlichen Veröffentlichungen und Konferenzpräsentationen des Autors. Die Listen sind antichronologisch nach Erstautor und Mitautor sortiert.

Konferenz- & Journalbeträge

1. Theodor Hillebrand, Nico Hellwege, Nils Heidmann, Steffen Paul, Dagmar Peters-Drolshagen, *Charge-based stochastic aging analysis of CMOS circuits*, IEEE International Integrated Reliability Workshop (IIRW), 2015

2. Theodor Hillebrand, Nico Hellwege, Nils Heidmann, Steffen Paul, Dagmar Peters-Drolshagen, *Stochastic analysis of degradation and variations in CMOS-Transistors*, Zuverlässigkeit und Entwurf (ZuE), 8. GMM/ITG/GI-Symposium Reliability by Design, 2015

3. Theodor Hillebrand, Nico Hellwege, Maike Taddiken, Konstantin Tscherkaschin, Steffen Paul, Dagmar Peters-Drolshagen, *Stochastic LUT-based reliability-aware design method for operation point dependent CMOS circuits*, 23rd International Conference Mixed Design of Integrated Circuits and Systems (MIXDES), 2016

4. Theodor Hillebrand, Maike Taddiken, Konstantin Tscherkaschin, Steffen Paul, Dagmar Peters-Drolshagen, *Unified Stochastic Reliability Aware Model for SONOS Memories in Harsh Environments*, Conference on Radiation Effects on Components and Systems (RADECS), 2016

5. Theodor Hillebrand, Timur Schäfer, Nico Hellwege, Marco Erstling, Dagmar Peters-Drolshagen, Steffen Paul, *Design and Verification of Analog CMOS Circuits Using the gm/ID-Method with Age-Dependent Degradation Effects*, 7th European Workshop on CMOS Variability (VARI), 2016

6. Theodor Hillebrand, Maike Taddiken, Konstantin Tscherkaschin, Steffen Paul, Dagmar Peters-Drolshagen, *Vth Is Dead - Long Live The Threshold Voltage*, IEEE International Integrated Reliability Workshop (IIRW), 2016

7. Theodor Hillebrand, Maike Taddiken, Konstantin Tscherkaschin, Steffen Paul, Dagmar Peters-Drolshagen, *Online Monitoring of NBTI and HCD in Beta-Multiplier Circuits*, IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS), 2016

8. Theodor Hillebrand, Ludwig Karsthof, Steffen Paul, Dagmar Peters-Drolshagen, *Reliability-aware Multi-Vth Domain Digital Design Assessment*, 21st IEEE International Symposium on Design and Diagnostics of Electronic Circuits and Systems (DDECS), 2018

9. Theodor Hillebrand, Steffen Paul, Dagmar Peters-Drolshagen, *A New Approach to Threshold Voltage Measurements of Transistors*, IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS), 2018

10. Theodor Hillebrand, Steffen Paul, Dagmar Peters-Drolshagen, *Topology-Driven Reliability Assessment of Integrated Circuits*, 24th International Conference Mixed Design of Integrated Circuits and Systems (MIXDES), 2018

11. Theodor Hillebrand, Maike Taddiken, Steffen Paul, Dagmar Peters-Drolshagen, *Yield Approximation of Analog Integrated Circuits under Time-Dependent Variability*, 15th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2018

12. Theodor Hillebrand, Konstantin Tscherkaschin, Steffen Paul, Dagmar Peters-Drolshagen, *On-line Parameter Extraction Technique for Integrated Circuits*, 16. GMM/ITG-Fachtagung ANALOG, 2018

13. Theodor Hillebrand, Konstantin Tscherkaschin, Sebastian Schmale, Steffen Paul, Dagmar Peters-Drolshagen, *On-line Surveillance Techniques for Reliable Active Pixel Sensor Systems*, 16. GMM/ITG-Fachtagung ANALOG, 2018

14. Theodor Hillebrand, Steffen Paul, Dagmar Peters-Drolshagen, *REL-MOS - A Reliability-Aware MOS Transistor Model*, IEEE Transactions on Electron Devices, 2018

15. Maike Taddiken, Theodor Hillebrand, Konstantin Tscherkaschin, Steffen Paul, Dagmar Peters-Drolshagen, *Parameter identification for behavioral modeling of analog components including degradation*, 23rd International Conference Mixed Design of Integrated Circuits and Systems (MIXDES), 2016

16. Konstantin Tscherkaschin, Theodor Hillebrand, Maike Taddiken, Steffen Paul, Dagmar Peters-Drolshagen, *Degradation and temperature analysis of voltage-controlled ring oscillators for robust and reliable oscillator designs in a 65nm bulk*

CMOS process, 23rd International Conference Mixed Design of Integrated Circuits and Systems (MIXDES), 2016

17. Konstantin Tscherkaschin, Theodor Hillebrand, Maike Taddiken, Steffen Paul and Dagmar Peters-Drolshagen, *Temperature- and Aging-Resistant Inverter for Robust and Reliable Time to Digital Circuit Designs in a 65nm Bulk CMOS Process*, IEEE International Symposium on On-Line Testing and Robust System Design (IOLTS), 2016

18. Timur Schäfer, Theodor Hillebrand, Nico Hellwege, Marco Erstling, Dagmar Peters-Drolshagen, Steffen Paul, *Design and Verification of Analog CMOS Circuits Using the gm/I D-Method with Age-Dependent Degradation Effects*, Journal of Low Power Electronics, 2017

19. Sascha Heinßen, Theodor Hillebrand, Maike Taddiken, Steffen Paul, Dagmar Peters-Drolshagen, *On-line monitoring and error correction in sensor interface circuits using digital calibration techniques*, IEEE 36th VLSI Test Symposium (VTS), 2018.

20. Sascha Heinßen, Theodor Hillebrand, Maike Taddiken, Steffen Paul, Dagmar Peters-Drolshagen, *Design for reliability of generic sensor interface circuits*, Microelectronics Reliability, 2018.

21. Sascha Heinssen, Maike Taddiken, Theodor Hillebrand, Steffen Paul, Dagmar Peters-Drolshagen, *Behavioral modeling of a sensor interface circuit including various non-idealities*, 14th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2017.

22. Sascha Heinssen, Maike Taddiken, Theodor Hillebrand, Steffen Paul, Dagmar Peters-Drolshagen, *On-Line Error Correction in Sensor Interface Circuits by Using Adaptive Filtering and Digital Calibration*, Eurosensors, 2018.

23. Kris Niederkleine, Theodor Hillebrand, Steffen Paul, *Realistic Radiation and PVTA fault simulation for OFDM synchronization* , Conference on Radiation Effects on Components and Systems (RADECS), 2018

Präsentationen

1. Theodor Hillebrand, Nico Hellwege, Steffen Paul, Dagmar Peters-Drolshagen, *Reliability-Aware Design Method for CMOS Circuits*, Workshop on Early Reliability Modeling for Aging and Variability in Silicon Systems (ERMAVSS), 2016

2. Theodor Hillebrand, Nico Hellwege, Steffen Paul, Dagmar Peters-Drolshagen, *Online Monitoring in Beta-Multiplier Circuits*, Test und Zuverlässigkeit (TUZ), 2016

3. Theodor Hillebrand, Steffen Paul, Dagmar Peters-Drolshagen, *Unified charge-based Transistor Model including Degradation Mechanisms*, Arbeitskreis Modellierung von Systemen und Parameterextraktion (MOS-AK), 2017

4. Steffen Paul, Dagmar Peters-Drolshagen, Theodor Hillebrand, Konstantin Tscherkaschin, Maike Taddiken, Heiner Lange, *Charge based design methods for CMOS circuits*, The Munich Workshop on Design Technology Coupling (DTC), 2016

5. Konstantin Tscherkaschin, Sascha Heinssen, Theodor Hillebrand, Maike Taddiken, Dagmar Peters-Drolshagen, Steffen Paul, *Reliability in Advanced CMOS Designs: Tackle Systems down to Electrons in a Single Workflow*, Campeon InnoWeek Universitätsabend, 2016

6. Konstantin Tscherkaschin, Theodor Hillebrand, Steffen Paul, Dagmar Peters-Drolshagen, *Aging Monitoring Concept for Mixed Signal Circuits in Advanced Technologies*, Analog Workshop, 2017

7. Kris Niederkleine, Nils Hülsmeiner, Theodor Hillebrand, Jochen Rust, Steffen Paul, *Radiation Tolerant Stochastic Fourier-Transformation*, 7th International Workshop on Analogue and Mixed-Signal Integrated Circuits for Space Applications (AMICSA), 2018

B Literaturverzeichnis

[Aal16] Andreas Aal. Reliability management - The central enabler for advanced Technologies in Automotive. *Microelectronics Reliability*, 64:13–18, 2016.

[ABKF10] Joaquín Alvarado, E Boufouss, Valeria Kilchytska, and Denis Flandre. Compact model for single event transients and total dose effects at high temperatures for partially depleted SOI MOSFETs. *Microelectronics Reliability*, 50(9):1852–1856, 2010.

[Aga14] et al. Agarwal. *BSIM6.1.0 MOSFET Compact Model*. University of Calirfornia, 2014.

[AJ99] Terje Aven and Uwe Jensen. *Stochastic models in reliability*. Springer, 1999.

[AKGH16] Hussam Amrouch, Behnam Khaleghi, Andreas Gerstlauerz, and Jörg Henkel. Reliability-aware design to suppress aging. In Unknown, editor, *Proceedings of the 53rd Annual Design Automation Conference on - DAC '16*, pages 1–6, New York, New York, USA, 2016. ACM Press.

[AM07] G Alastair Armstrong and Chinmay K Maiti. *Technology computer aided design for Si, SiGe and GaAs integrated circuits*, volume 21. IET, 2007.

[AMGC02] M Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on signal processing*, 50(2), 2002.

[Aro12] Narain D Arora. *MOSFET models for VLSI circuit simulation: theory and practice*. Springer Science & Business Media, 2012.

[AWB68] Michael Athans, Richard Wishner, and Anthony Bertolini. Suboptimal state estimation for continuous-time nonlinear systems from discrete noisy measurements. *IEEE Transactions on Automatic Control*, 13(5):504–514, 1968.

[Bak08] R Jacob Baker. *CMOS: circuit design, layout, and simulation*, volume 1. John Wiley & Sons, 2008.

[Bar06] HJ Barnaby. Total-ionizing-dose effects in modern CMOS technologies. *IEEE Transactions on Nuclear Science*, 53(6):3103–3121, 2006.

[BBA+11] Antonios Bazigos, Matthias Bucher, Joachim Assenmacher, Stefan Decker, Wladyslaw Grabinski, and Yannis Papananos. An adjusted constant-current method to determine saturated and linear mode threshold voltage of MOSFETs. *IEEE Transactions on Electron Devices*, 58(11):3751–3758, 2011.

[BHL+12] Ilja N Bronstein, Juraj Hromkovic, Bernd Luderer, Hans-Rudolf Schwarz, Jochen Blath, Alexander Schied, Stephan Dempe, Gert Wanka, and Siegfried Gottwald. *Taschenbuch der mathematik*, volume 1. Springer-Verlag, 2012.

[BI07] Kathy Boucart and Adrian Mihai Ionescu. Double-gate tunnel fet with high-k gate dielectric. *IEEE Transactions on Electron Devices*, 54(7):1725–1733, 2007.

[Bin07] David M Binkley. Tradeoffs and optimization in analog CMOS design. In *Mixed Design of Integrated Circuits and Systems, 2007. MIXDES'07. 14th International Conference on*, pages 47–60. IEEE, 2007.

[BK15] Manjul Bhushan and Mark B Ketchen. *CMOS Test and Evaluation*. Springer, 2015.

[BLE+98] Matthias Bucher, Christophe Lallement, Christian Enz, Fabien Théodoloz, and François Krummenacher. The epfl-ekv mosfet model equations for simulation. *Version 2.6*, 1998.

[BN00] Duane Boning and Sani Nassif. Models of process variations in device and interconnect. *Design of high performance microprocessor circuits*, page 6, 2000.

[Bro70a] Charles G Broyden. The convergence of a class of double-rank minimization algorithms: 2. the new algorithm. *IMA journal of applied mathematics*, 6(3):222–231, 1970.

[Bro70b] Charles George Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[BS13] Alain Bensoussan and Ephraim Suhir. Design-for-reliability (DfR) of aerospace electronics: Attributes and challenges. In *Aerospace Conference, 2013 IEEE*, pages 1–17. IEEE, 2013.

[C+03] Zhe Chen et al. Bayesian filtering: From kalman filters to particle filters, and beyond. *Statistics*, 182(1):1–69, 2003.

[CAD+10] Moonju Cho, Marc Aoulaiche, Robin Degraeve, Ben Kaczer, Jacopo Franco, Thomas Kauerauf, Philippe Roussel, Lars Å Ragnarsson, Joshua Tseng, Thomas Y Hoffmann, et al. Positive and negative bias temperature instability on sub-nanometer EOT high-K MOSFETs. In *Reliability Physics Symposium (IRPS), 2010 IEEE International*, pages 1095–1098. IEEE, 2010.

[CBTD01] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[CC07] Yu Cao and Lawrence T Clark. Mapping statistical process variations toward circuit performance variability: an analytical modeling approach. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 26(10):1866–1873, 2007.

[CG01] TL Chen and G Gildenblat. Analytical approximation for the mosfet surface potential. *Solid-State Electronics*, 45(2):335–339, 2001.

[CH99] Yuhua Cheng and Chenming Hu. *MOSFET modeling & BSIM3 user?s guide*. Springer Science & Business Media, 1999.

[CL12] Wei-Ting Kary Chien and Ming Li. Reduce time-to-market by considering reliability tradeoffs. *IEEE Transactions on Reliability*, 61(4):837–845, 2012.

[CTF+06] Carroll Croarkin, Paul Tobias, JJ Filliben, Barry Hembree, Will Guthrie, et al. Nist/sematech e-handbook of statistical methods. *NIST/SEMATECH, 13.03.2018*, 2006.

[CVC+14a] Yogesh Singh Chauhan, Sriramkumar Venugopalan, Maria-Anna Chalkiadaki, Muhammed Ahosan Ul Karim, Harshit Agarwal, Sourabh Khandelwal, Navid Paydavosi, Juan Pablo Duarte, Christian C Enz, Ali M Niknejad, et al. BSIM6: Analog and RF compact model for bulk MOSFET. *IEEE Transactions on Electron Devices*, 61(2):234–244, 2014.

[CVC+14b] Yogesh Singh Chauhan, Sriramkumar Venugopalan, Maria-Anna Chalkiadaki, Muhammed Ahosan Ul Karim, Harshit Agarwal, Sourabh Khandelwal, Navid Paydavosi, Juan Pablo Duarte, Christian C. Enz, Ali M. Niknejad, and Chenming Hu. BSIM6: Analog and RF Compact Model for Bulk MOSFET. *IEEE Transactions on Electron Devices*, 61(2):234–244, 2014.

[Dev11] Jay L Devore. *Probability and Statistics for Engineering and the Sciences.* Cengage learning, 2011.

[DH11] Manfred Dietrich and Joachim Haase. *Process Variations and Probabilistic Integrated Circuit Design.* Springer Science & Business Media, 2011.

[DM03] Patrick G Drennan and Colin C McAndrew. Understanding MOSFET mismatch for analog design. *IEEE Journal of Solid-State Circuits*, 38(3):450–456, 2003.

[DW12] Bob Dobkin and Jim Williams. *Analog circuit design volume 2: immersion in the black art of analog design*, volume 2. Newnes, 2012.

[ECP17a] Christian Enz, Francesco Chicco, and Alessandro Pezzotta. Nanoscale MOSFET Modeling: Part 1: The Simplified EKV Model for the Design of Low-Power Analog Circuits. *IEEE Solid-State Circuits Magazine*, 9(3):26–35, 2017.

[ECP17b] Christian Enz, Francesco Chicco, and Alessandro Pezzotta. Nanoscale MOSFET Modeling: Part 2: Using the Inversion Coefficient as the Primary Design Parameter. *IEEE Solid-State Circuits Magazine*, 9(4):73–81, 2017.

[Efr92] Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[Epa69] Vassiliy A Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969.

[ES81] Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

[EV06] Christian C Enz and Eric A Vittoz. *Charge-based MOS transistor modeling: the EKV model for low-power and RF IC design.* John Wiley & Sons, 2006.

[FKM+16] J Franco, B Kaczer, S Mukhopadhyay, P Duhan, P Weckx, Ph J Roussel, T Chiarella, L-Å Ragnarsson, L Trojman, N Horiguchi, et al. Statistical model of the NBTI-induced threshold voltage, subthreshold swing, and transconductance degradations in advanced p-FinFETs. In *Electron Devices Meeting (IEDM), 2016 IEEE International*, pages 15–3. IEEE, 2016.

[FKT13] Farshad Firouzi, Saman Kiamehr, and Mehdi B Tahoori. Statistical analysis of BTI in the presence of process-induced voltage and temperature variations. In *Design Automation Conference (ASP-DAC), 2013 18th Asia and South Pacific*, pages 594–600. IEEE, 2013.

[Fle70] Roger Fletcher. A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322, 1970.

[Fra10] Sami Franssila. *Introduction to microfabrication*. John Wiley & Sons, Chichester West Sussex England, 2nd ed. edition, 2010.

[GHA+15] Louis Gerrer, Razaidi Hussin, Salvatore M Amoroso, Jacopo Franco, Pieter Weckx, Marco Simicic, Naoto Horiguchi, Ben Kaczer, Tibor Grasser, and Asen Asenov. Experimental evidences and simulations of trap generation along a percolation path. In *Solid State Device Research Conference (ESSDERC), 2015 45th European*, pages 226–229. IEEE, 2015.

[GHB09] C Guerin, V Huard, and A Bravaix. General framework about defect creation at the si/ si o 2 interface. *Journal of Applied Physics*, 105(11):114513, 2009.

[Ghi88] Gérard Ghibaudo. New method for the extraction of MOSFET parameters. *Electronics Letters*, 24(9):543–545, 1988.

[GKG+11] Tibor Grasser, Ben Kaczer, Wolfgang Goes, Hans Reisinger, Thomas Aichinger, Philipp Hehenberger, Paul-Jürgen Wagner, Franz Schanovsky, Jacopo Franco, María Toledano Luque, et al. The paradigm shift in understanding the bias temperature instability: From reaction–diffusion to switching oxide traps. *IEEE Transactions on Electron Devices*, 58(11):3652–3666, 2011.

[GLKJ16] Kay-Uwe Giering, André Lange, Ben Kaczer, and Roland Jancke. Bti variability of sram cells under periodically changing stress profiles. In *Integrated Reliability Workshop (IIRW), 2016 IEEE International*, pages 1–5. IEEE, 2016.

[GLW+05] G Gildenblat, X Li, H Wang, W Wu, R Van Langevelde, AJ Scholten, GDJ Smit, and DBM Klaassen. Introduction to PSP MOSFET model. In *Proc. the MSM 2005 Int. Conf., Nanotech 2005*, 2005.

[GLW+06] G. Gildenblat, X. Li, W. Wu, H. Wang, A. Jha, R. van Langevelde, G.D.J. Smit, A. J. Scholten, and D.B.M. Klaassen. PSP: An Advanced Surface-Potential-Based MOSFET Model for Circuit Simulation. *IEEE Transactions on Electron Devices*, 53(9):1979–1993, 2006.

[GNS06] Wladyslaw Grabinski, Bart Nauwelaers, and Dominique Schreurs. *Transistor level modeling for analog/RF IC design.* Springer, 2006.

[Gol70] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26, 1970.

[GP16] Kirk A Gray and John J Paschkewitz. *Next Generation HALT and HASS: Robust Design of Electronics and Systems.* John Wiley & Sons, 2016.

[Gra07] Helmut E Graeb. *Analog design centering and sizing*, volume 64. Springer, 2007.

[Gra13] Tibor Grasser. *Bias Temperature Instability for Devices and Circuits.* Springer Science & Business Media, 2013.

[Gra14] Tibor Grasser. *Hot Carrier Degradation in Semiconductor Devices.* Springer, 2014.

[GRB+09] Meeta S Gupta, Jude A Rivers, Pradip Bose, Gu-Yeon Wei, and David Brooks. Tribeca: design for PVT variations with local recovery and fine-grained adaptation. In *Microarchitecture, 2009. MICRO-42. 42nd Annual IEEE/ACM International Symposium on*, pages 435–446. IEEE, 2009.

[GRLVAI16] Fernando García-Redondo, Marisa López-Vallejo, Hernan Aparicio, and Pablo Ituero. Reliable design methodology: The combined effect of radiation, variability and temperature. In *Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2016 13th International Conference on*, pages 1–4. IEEE, 2016.

[GSR+14] Kay-Uwe Giering, Christoph Sohrmann, Gerhard Rzepa, Leonhard Heiß, Tibor Grasser, and Roland Jancke. Nbti modeling in analog circuits and its application to long-term aging simulations. In *Integrated Reliability Workshop Final Report (IIRW), 2014 IEEE International*, pages 29–34. IEEE, 2014.

[GSS93] Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F (Radar and Signal Processing)*, volume 140, pages 107–113. IET, 1993.

[GTJ+15] Wladek Grabinski, Daniel Tomaszewski, Farzan Jazaeri, Anurag Mangla, Jean-Michel Sallese, Maria-Anna Chalkiadaki, Antonios Bazigos, and Matthias Bucher. FOSS EKV 2.6 parameter extractor. In *2015 22nd International Conference Mixed Design of Integrated Circuits & Systems (MIXDES)*, pages 181–186. IEEE, 25.06.2015 - 27.06.2015.

[GWC+04] G. Gildenblat, Hailing Wang, Ten-Lon Chen, Xin Gu, and Xiaowen Cai. SP: An advanced surface-potential-based compact MOSFET model. *IEEE Journal of Solid-State Circuits*, 39(9):1394–1406, 2004.

[GWR+11] T Grasser, P-J Wagner, H Reisinger, Th Aichinger, G Pobegen, M Nelhiebel, and B Kaczer. Analytic modeling of the bias temperature instability using capture/emission time maps. In *Electron Devices Meeting (IEDM), 2011 IEEE International*, pages 27–4. IEEE, 2011.

[Hab13] Husni Habal. *Constraint-Based Layout-Driven Sizing of Analog Circuits*. PhD thesis, Technische UniversitÃ¤t MÃ¼nchen, Dissertation, 2013, MÃ¼nchen, 2013.

[Han88] Military Handbook. Electronic reliability design handbook. Technical report, MIL-HDBK-338, DoD, 1988.

[HCF+15] V Huard, F Cacho, X Federspiel, W Arfaoui, M Saliva, and D Angot. Technology scaling and reliability: Challenges and opportunities. In *Electron Devices Meeting (IEDM), 2015 IEEE International*, pages 20–5. IEEE, 2015.

[HG16] Husni Habal and Helmut Graeb. A step-accurate model for the trapping and release of charge carriers suitable for the transient simulation of analog circuits. *Microelectronics Reliability*, 61:17–23, 2016.

[HHE+15] Nico Hellwege, Nils Heidmann, Marco Erstling, Dagmar Peters-Drolshagen, and Steffen Paul. An aging-aware transistor sizing tool regarding BTI and HCD degradation modes. In *Mixed Design of Integrated Circuits & Systems (MIXDES), 2015 22nd International Conference*, pages 272–277. IEEE, 2015.

[HHPDP13] Nico Hellwege, Nils Heidmann, Dagmar Peters-Drolshagen, and Steffen Paul. Using operating point-dependent degradation and g m/I D method for aging-aware design. In *Integrated Reliability Workshop Final Report (IRW), 2013 IEEE International*, pages 113–116. IEEE, 2013.

[HHPPD15] Nils Heidmann, Nico Hellwege, Steffen Paul, and Dagmar Peters-Drolshagen. NBTI and HCD aware behavioral models for reliability analysis of analog CMOS circuits. In *Reliability Physics Symposium (IRPS), 2015 IEEE International*, pages CA–1. IEEE, 2015.

[HHT+14] Nils Heidmann, Nico Hellwege, Maike Taddiken, Dagmar Peters-Drolshagen, and Steffen Paul. Analog behavioral modeling for age-dependent degradation of complex analog circuits. In *Mixed Design of Integrated Circuits & Systems (MIXDES), 2014 Proceedings of the 21st International Conference*, pages 317–322. IEEE, 2014.

[HPR+07] V Huard, C Parthasarathy, N Rallet, C Guerin, M Mammase, D Barge, and C Ouvrard. New characterization and modeling approach for nbti degradation from transistor to product level. In *Electron Devices Meeting, 2007. IEDM 2007. IEEE International*, pages 797–800. IEEE, 2007.

[HRC89] Ramesh Harjani, Rob A Rutenbar, and L Richard Carley. OASYS: A framework for analog circuit synthesis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 8(12):1247–1266, 1989.

[HXC+02] Jin He, Xuemei Xi, Mansun Chan, Kanyu Cao, Chenming Hu, Yingxue Li, Xing Zhang, Ru Huang, and Yangyuan Wang. Normalized mutual integral difference method to extract threshold voltage of MOSFETs. *IEEE Electron Device Letters*, 23(7):428–430, 2002.

[II17a] Statista The Statistics Portal IC Insights. Five-year constant annual growth rate (cagr) of the automotive integrated circuit (ic) market worldwide, from 2016 to 2021, by region., 2017.

[II17b] Statista The Statistics Portal IC Insights. Size of the automotive integrated circuit (ic) market worldwide in 2016 and 2017, by device type (in million u.s. dollars)., 2017.

[II18a] Statista The Statistics Portal IC Insights. Leading analog integrated circuit (ic) suppliers worldwide from 2013 to 2017, by revenue (in million u.s. dollars)., 2018.

[II18b] Statista The Statistics Portal IC Insights. Market share of analog integrated circuit (ic) sales worldwide from 2014 to 2018, by application., 2018.

[II18c] Statista The Statistics Portal IC Insights. Worldwide integrated circuit (ic) market compound annual growth rates (cagrs) from 2017 to 2022, by product categories., 2018.

[JLK14] Seobin Jung, Jiho Lee, and Jaeha Kim. Variability-Aware, Discrete Optimization for Analog Circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 33(8):1117–1130, 2014.

[JM17a] Paul G. A. Jespers and Boris Murmann, editors. *Systematic Design of Analog CMOS Circuits*. Cambridge University Press, Cambridge, 2017.

[JM17b] Paul GA Jespers and Boris Murmann. *Systematic Design of Analog CMOS Circuits*. Cambridge University Press, 2017.

[JUDW00] Simon Julier, Jeffrey Uhlmann, and Hugh F Durrant-Whyte. A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on automatic control*, 45(3):477–482, 2000.

[Kae08] Hubert Kaeslin. *Digital integrated circuit design: from VLSI architectures to CMOS fabrication.* Cambridge University Press, 2008.

[KCK13] Way Kuo, Wei-Ting Kary Chien, and Taeho Kim. *Reliability, yield, and stress burn-in: a unified approach for microelectronics systems manufacturing & software development.* Springer Science & Business Media, 2013.

[KFW+18] B Kaczer, J Franco, P Weckx, Ph J Roussel, V Putcha, E Bury, M Simicic, A Chasin, D Linten, B Parvais, et al. A brief overview of gate oxide defect properties and their relation to MOSFET instabilities and device and circuit time-dependent variability. *Microelectronics Reliability*, 81:186–194, 2018.

[KSA+09] Jeffrey S Kauppila, Andrew L Sternberg, Michael L Alles, A Matt Francis, Jim Holmes, Oluwole A Amusan, and Lloyd W Massengill. A bias-dependent single-event compact model implemented into BSIM4 and a 90 nm CMOS process design kit. *IEEE Transactions on nuclear Science*, 56(6):3152–3157, 2009.

[KSM+07] Michihiro Kanno, Akira Shibuya, Masao Matsumura, Kazuhiro Tamura, Hitoshi Tsuno, Shigetaka Mori, Yuzo Fukuzaki, Tetsuo Gocho, Hisahiro Ansai, and Naoki Nagashima. Empirical characteristics and extraction of overall variations for 65-nm mosfets and beyond. In *VLSI Technology, 2007 IEEE Symposium on*, pages 88–89. IEEE, 2007.

[LF12] Simon Li and Yue Fu. *3D TCAD Simulation for Semiconductor Processes, Devices and Optoelectronics.* Springer New York, New York, NY, 2012.

[LG06] Mark S Lundstrom and Jing Guo. The ballistic nanotransistor. *Nanoscale Transistors: Device Physics, Modeling and Simulation*, pages 83–114, 2006.

[Liu01] William Liu. *MOSFET Models for SPICE Simulation including BSIM3v3 and BSIM4.* Wiley-Interscience Publication, 2001.

[LK12] Yusuf Leblebici and Sung-Mo Steve Kang. *Hot-carrier reliability of MOS VLSI circuits*, volume 227. Springer Science & Business Media, 2012.

[LXF+17] Bangjun Lei, Guangzhu Xu, Ming Feng, Ferdinand Van der Heijden, Yaobin Zou, Dick de Ridder, and David MJ Tax. *Classification, parameter estimation and state estimation: an engineering approach using MATLAB*. John Wiley & Sons, 2017.

[MAD+17] Chhandak Mukherjee, Bertrand Ardouin, Jean-Yves Dupuy, Virginie Nodjiadjim, Muriel Riet, Thomas Zimmer, François Marc, and Cristell Maneux. Reliability-aware circuit design methodology for beyond-5g communication systems. *IEEE Transactions on Device and Materials Reliability*, 17(3):490–506, 2017.

[Mao13] James Mao. *CircuitBook: A framework for analog design reuse*. PhD thesis, Stanford University, 2013.

[MCG+15] Colin C McAndrew, Geoffrey J Coram, Kiran K Gullapalli, J Robert Jones, Laurence W Nagel, Ananda S Roy, Jaijeet Roychowdhury, Andries J Scholten, Geert DJ Smit, Xufeng Wang, et al. Best practices for compact modeling in Verilog-A. *IEEE Journal of the Electron Devices Society*, 3(5):383–396, 2015.

[McP10] Joe W McPherson. *Reliability physics and engineering*. Springer, 2010.

[MFC13] Philippe Magarshack, Philippe Flatresse, and Giorgio Cesana. Utbb fd-soi: A process/design symbiosis for breakthrough energy-efficiency. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 952–957. EDA Consortium, 2013.

[MG10] Elie Maricau and Georges Gielen. Efficient variability-aware nbti and hot carrier circuit reliability analysis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 29(12):1884–1893, 2010.

[MG13] Elie Maricau and Georges Gielen. *Analog IC Reliability in Nanometer CMOS*. Springer Science & Business Media, 2013.

[MH15] Markus Meissner and Lars Hedrich. Feats: Framework for explorative analog topology synthesis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 34(2):213–226, 2015.

[MHJ+16] S Mhira, V Huard, A Jain, F Cacho, D Meyer, S Naudet, A Bravaix, and C Parthasarathy. Mission profile recorder: An aging monitor for hard events. In *Reliability Physics Symposium (IRPS), 2016 IEEE International*, pages 4C–3. IEEE, 2016.

[MJC+14] S Mukhopadhyay, K Joshi, V Chaudhary, N Goel, S De, RK Pandey, KVRM Murali, and S Mahapatra. Trap generation in IL and HK layers during BTI/TDDB stress in scaled HKMG N and P MOSFETs. In *Reliability Physics Symposium, 2014 IEEE International*, pages GD–3. IEEE, 2014.

[MMB+06] François Marc, Benoit Mongellaz, Corinne Bestory, Herve Levi, and Yves Danto. Improvement of aging simulation of electronic circuits using behavioral modeling. *IEEE Transactions on Device and Materials Reliability*, 6(2):228–234, 2006.

[MMD15] Agnish Mal, Ashis Kumar Mal, and Sanjit Kumar Datta. CMOS Amplifier Design Using Simplified g m/I D Technique. In *Intelligent Computing and Applications*, pages 537–544. Springer, 2015.

[MMM+14] Chenyue Ma, Hans Jürgen Mattausch, Kazuya Matsuzawa, Seiichiro Yamaguchi, Teruhiko Hoshida, Masahiro Imade, Risho Koh, Takahiko Arakawa, and Mitiko Miura-Mattausch. Universal NBTI compact model for circuit aging simulation under any stress conditions. *IEEE Transactions on Device and Materials Reliability*, 14(3):818–825, 2014.

[MP18] Souvik Mahapatra and Narendra Parihar. A review of NBTI mechanisms and models. *Microelectronics Reliability*, 81:127–135, 2018.

[MXC+13] Qiang Miao, Lei Xie, Hengjuan Cui, Wei Liang, and Michael Pecht. Remaining useful life prediction of lithium-ion battery with unscented particle filter technique. *Microelectronics Reliability*, 53(6):805–810, 2013.

[MZ09] Howard Musoff and Paul Zarchan. *Fundamentals of Kalman filtering: a practical approach.* American Institute of Aeronautics and Astronautics, 2009.

[Nic15] NicoÂ Hellwege. *Aging-Aware Design Methods for Reliable Analog Integrated Circuits using Operating Point-Dependent Degradation.* PhD thesis, UniversitÃ¤t Bremen, Dissertation, 2015, Bremen, 2015. 1 Online-Ressource (ix, 131 Seiten) : Illustrationen.

[OCCE+01] Adelmo Ortiz-Conde, Antonio Cerdeira, Magali Estrada, Francisco J Garcıa Sánchez, and Rodolfo Quintero. A simple procedure to extract the threshold voltage of amorphous thin film mosfets in the saturation region. *Solid-State Electronics*, 45(5):663–667, 2001.

[OCGSLH10] Adelmo Ortiz-Conde, Francisco J García-Sánchez, Juin J Liou, and Ching-Sung Ho. Integration-based approach to evaluate the sub-

threshold slope of MOSFETs. *Microelectronics Reliability*, 50(2):312–315, 2010.

[OCGSM+13] Adelmo Ortiz-Conde, Francisco J García-Sánchez, Juan Muci, Alberto Terán Barrios, Juin J Liou, and Ching-Sung Ho. Revisiting MOSFET threshold voltage extraction methods. *Microelectronics Reliability*, 53(1):90–104, 2013.

[OCSL+02] Adelmo Ortiz-Conde, FJ Garcıa Sánchez, Juin J Liou, Antonio Cerdeira, Magali Estrada, and Y Yue. A review of recent MOSFET threshold voltage extraction methods. *Microelectronics Reliability*, 42(4):583–596, 2002.

[OM03] Timothy R Oldham and FB McLean. Total ionizing dose effects in MOS oxides and devices. *IEEE Transactions on Nuclear Science*, 50(3):483–499, 2003.

[PDW89] Marcel JM Pelgrom, Aad CJ Duinmaijer, and Anton PG Welbers. Matching properties of mos transistors. *IEEE Journal of solid-state circuits*, 24(5):1433–1439, 1989.

[PFV+18] Vamsi Putcha, Jacopo Franco, Abhitosh Vais, Sonja Sioncke, Ben Kaczer, Dimitri Linten, and Guido Groeseneken. On the apparent non-arrhenius temperature dependence of charge trapping in iiiv/high-k mos stack. *IEEE Transactions on Electron Devices*, 9(99), 2018.

[PG11] Xin Pan and Helmut Graeb. Lifetime yield optimization of analog circuits considering process variations and parameter degradations. In *Advances in analog circuits*. IntechOpen, 2011.

[PKD+06] Fabien Prégaldiny, François Krummenacher, Birahim Diagne, François Pêcheux, Jean-Michel Sallese, and Christophe Lallement. Explicit modelling of the double-gate mosfet with vhdl-ams. *INTERNATIONAL JOURNAL OF NUMERICAL MODELLING: electronic networks, devices and fields*, 19(3):239–256, 2006.

[PMP14a] Soumya Pandit, Chittaranjan Mandal, and Amit Patra. *Nano-scale CMOS Analog Circuits: Models and CAD Techniques for High-level Design*. CRC Press, 2014.

[PMP14b] Soumya Pandit, Chittaranjan Mandal, and Amit Patra. *Nano-scale CMOS Analog Circuits: Models and CAD Techniques for High-level Design*. CRC Press, 2014.

[Pól20] Georg Pólya. Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung und das momentenproblem. *Mathematische Zeitschrift*, 8(3-4):171–181, 1920.

[PP02] Athanasios Papoulis and S Unnikrishna Pillai. *Probability, random variables, and stochastic processes.* Tata McGraw-Hill Education, 2002.

[PRS+15a] T Poiroux, O Rozeau, P Scheer, S Martinie, MA Jaud, M Minondo, A Juge, JC Barbé, and M Vinet. Leti-utsoi2. 1: A compact model for utbb-fdsoi technologies - part i: Interface potentials analytical model. *IEEE Transactions on Electron Devices*, 62(9):2751–2759, 2015.

[PRS+15b] Thierry Poiroux, O Rozeau, Patrick Scheer, Sébastien Martinie, Marie-Anne Jaud, M Minondo, André Juge, JC Barbé, and Maud Vinet. Leti-utsoi2. 1: A compact model for utbb-fdsoi technologies - part ii: Dc and ac model description. *IEEE Transactions on Electron Devices*, 62(9):2760–2768, 2015.

[Ram13] Ahmed Ramadan. Compact model council's standard circuit simulator interface for reliability modeling. In *Reliability Physics Symposium (IRPS), 2013 IEEE International*, pages 2A–5. IEEE, 2013.

[RF15] Nagarajan Raghavan and Daniel D Frey. Particle filter approach to lifetime prediction for microelectronic devices and systems with multiple failure mechanisms. *Microelectronics Reliability*, 55(9-10):1297–1301, 2015.

[RG12] Dev G Raheja and Louis J Gullo. *Design for reliability.* John Wiley & Sons, 2012.

[RGE+11] Hans Reisinger, Tibor Grasser, Karsten Ermisch, Heiko Nielen, Wolfgang Gustin, and Christian Schlünder. Understanding and modeling AC BTI. In *Reliability Physics Symposium (IRPS), 2011 IEEE International*, pages 6A–1. IEEE, 2011.

[RJPB12] Oliver Rozeau, Marie-Anne Jaud, Thierry Poiriux, and Mohamed Benosman. *UTSOI Model 1.1.4. Surface Potential Model for Ultra Thin Fully Depleted SOI MOSFET*, 2012.

[SAG+12] Christian Schlunder, Stefano Aresu, Georg Georgakos, Werner Kanert, Hans Reisinger, Karl Hofmann, and Wolfgang Gustin. HCI vs. BTI?-Neither one's out. In *Reliability Physics Symposium (IRPS), 2012 IEEE International*, pages 2F–4. IEEE, 2012.

[Sah10] Samar K. Saha. Modeling Process Variability in Scaled CMOS Technology. *IEEE Design & Test of Computers*, 27(2):8–16, 2010.

[Sah14] Samar K Saha. Compact MOSFET modeling for process variability-aware VLSI circuit design. *IEEE Access*, 2:104–115, 2014.

[Sam17] Michael J Sampson. Echoes of the Past, Glimpses of the Future-Ongoing Trends in Assurance of EEE Parts for Spaceflight. *21st Annual Components for Military and Space Electronics (CMSE) Training and Exhibition*, 2017.

[SBP+17] Christian Schlünder, Jörg Berthold, Fabian Proebster, Andreas Martin, Wolfgang Gustin, and Hans Reisinger. On the influence of BTI and HCI on parameter variability. In *Reliability Physics Symposium (IRPS), 2017 IEEE International*, pages 2E–4. IEEE, 2017.

[Sch97] Robert R Schaller. Moore's law: past, present and future. *IEEE spectrum*, 34(6):52–59, 1997.

[Sch06] Dieter K Schroder. *Semiconductor material and device characterization.* John Wiley & Sons, 2006.

[SFJ96] Flandre Silveira, Denis Flandre, and PGA Jespers. A g/sub m//I/sub D/based methodology for the design of CMOS analog circuits and its application to the synthesis of a silicon-on-insulator micropower OTA. *IEEE journal of solid-state circuits*, 31(9):1314–1319, 1996.

[SG09] Bhaskar Saha and Kai Goebel. Modeling li-ion battery capacity depletion in a particle filtering framework. In *Proceedings of the annual conference of the prognostics and health management society*, pages 2909–2924, 2009.

[SGM10] Márcio Cherem Schneider and Carlos Galup-Montoro. *CMOS analog design using all-region MOSFET modeling.* Cambridge University Press, 2010.

[Sha70] David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656, 1970.

[SHH+17] Timur Schäfer, Theodor Hillebrand, Nico Hellwege, Marco Erstling, Dagmar Peters-Drolshagen, and Steffen Paul. Design and Verification of Analog CMOS Circuits Using the g m/I D-Method with Age-Dependent Degradation Effects. *Journal of Low Power Electronics*, 13(1):135–147, 2017.

[Sim06] Dan Simon. *Optimal state estimation: Kalman, H infinity, and nonlinear approaches.* John Wiley & Sons, 2006.

[SK08] Danica Stefanovic and Maher Kayal. *Structured analog CMOS design*. Springer Science & Business Media, 2008.

[SMG18] James H Stathis, Souvik Mahapatra, and Tibor Grasser. Controversial issues in negative bias temperature instability. *Microelectronics Reliability*, 81:244–251, 2018.

[SN90] Takayasu Sakurai and A Richard Newton. Alpha-power law mosfet model and its applications to cmos inverter delay and other formulas. *IEEE Journal of solid-state circuits*, 25(2):584–594, 1990.

[SOCDM+00] FJ Garcıa Sánchez, A Ortiz-Conde, G De Mercato, JA Salcedo, JJ Liou, and Y Yue. New simple procedure to determine the threshold voltage of MOSFETs. *Solid-State Electronics*, 44(4):673–675, 2000.

[SPB+15] Christian Schlünder, Fabian Proebster, Jörg Berthold, Wolfgang Gustin, and Hans Reisinger. Influence of MOSFET geometry on the statistical distribution of NBTI induced parameter degradation. In *Integrated Reliability Workshop (IIRW), 2015 IEEE International*, pages 81–86. IEEE, 2015.

[SSGM12] Osmar Franca Siebel, Marcio Cherem Schneider, and Carlos Galup-Montoro. MOSFET threshold voltage: Definition, extraction, and some applications. *Microelectronics Journal*, 43(5):329–336, 2012.

[SSKJ87] B. J. Sheu, D. L. Scharfetter, P.-K. Ko, and M.-C. Jeng. BSIM: Berkeley short-channel IGFET model for MOS transistors. *IEEE Journal of Solid-State Circuits*, 22(4):558–566, 1987.

[SSKvdT13] GDJ Smit, AJ Scholten, DBM Klaassen, and R van der Toorn. Psp 103.3. Technical report, Technical Note NXP-TN-2012-0080, NXP Semiconductors, 2013.

[SSS18] Himangi Sood, Viranjay M Srivastava, and Ghanshyam Singh. Advanced mosfet technologies for next generation communication systems-perspective and challenges: A review. *Journal of Engineering Science and Technology Review*, 11(3):180–195, 2018.

[ST12] Jun Shao and Dongsheng Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

[Suh11] Ephraim Suhir. Remaining useful lifetime (rul): Probabilistic predictive model. *International Journal of Prognostics and Health Management Volume 2 (color)*, page 140, 2011.

[SWHZ11] Xiao-Sheng Si, Wenbin Wang, Chang-Hua Hu, and Dong-Hua Zhou. Remaining useful life estimation–a review on the statistical data driven approaches. *European journal of operational research*, 213(1):1–14, 2011.

[TBC+97] Yuan Taur, Douglas A Buchanan, Wei Chen, David J Frank, Khalid E Ismail, Shih-Hsien Lo, George A Sai-Halasz, Raman G Viswanathan, H-JC Wann, Shalom J Wind, et al. Cmos scaling into the nanometer regime. *Proceedings of the IEEE*, 85(4):486–504, 1997.

[TCSB16] E Tlelo-Cuautle and AC Sanabria-Borbon. Optimising operational amplifiers by evolutionary algorithms and gm/Id method. *International Journal of Electronics*, 103(10):1665–1684, 2016.

[TFMLCL+17] A Toro-Frías, P Martin-Lloret, R Castro-López, E Roca, FV Fernández, J Martin-Martinez, R Rodriguez, and M Nafria. Including a stochastic model of aging in a reliability simulation flow. In *Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2017 14th International Conference on*, pages 1–4. IEEE, 2017.

[THPPD17] Maike Taddiken, Theodor Hillebrand, Steffen Paul, and Dagmar Peters-Drolshagen. Variation-and degradation-aware stochastic behavioral modeling of analog circuit components. In *Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD), 2017 14th International Conference on*, pages 1–4. IEEE, 2017.

[THT+16] Maike Taddiken, Theodor Hillebrand, Konstantin Tscherkaschin, Steffen Paul, and Dagmar Peters-Drolshagen. Parameter identification for behavioral modeling of analog components including degradation. In *Mixed Design of Integrated Circuits and Systems, 2016 MIXDES-23rd International Conference*, pages 336–340. IEEE, 2016.

[TJS+15] S Tyaginov, Markus Jech, Prateek Sharma, J Franco, Ben Kaczer, and T Grasser. On the temperature behavior of hot-carrier degradation. In *Integrated Reliability Workshop (IIRW), 2015 IEEE International*, pages 143–146. IEEE, 2015.

[TMJ+18] SE Tyaginov, AA Makarov, M Jech, MI Vexler, J Franco, B Kaczer, and T Grasser. Physical Principles of Self-Consistent Simulation of the Generation of Interface States and the Transport of Hot Charge Carriers in Field-Effect Transistors Based on Metal–Oxide–Semiconductor Structures. *Semiconductors*, 52(2):242–247, 2018.

[TRB+03] S. C. Terry, J. M. Rochelle, D. M. Binkley, B. J. Blalock, D. P. Foty, and M. Bucher. Comparison of a BSIM3V and EKV MOSFET model

for a 0.5um CMOS process and implications for analog circuit design. *IEEE Transactions on Nuclear Science*, 50(4):915–920, 2003.

[TS02] Ulrich Tietze and Christoph Schenk. *Halbleiter-Schaltungstechnik:[neuer Teil: Nachrichtentechnische Schaltungen]*. Springer, 2002.

[UJT+17] Bianka Ullmann, Markus Jech, S Tyaginov, Michael Waltl, Yury Illarionov, Alexander Grill, Katja Puschkarsky, H Reisinger, and T Grasser. The impact of mixed negative bias temperature instability and hot carrier stress on single oxide defects. In *Reliability Physics Symposium (IRPS), 2017 IEEE International*, pages XT–10. IEEE, 2017.

[WCJ11] Mark White, Mark Cooper, and Allan Johnston. Users guide on scaled cmos reliability: Nasa electronic parts and packaging (nepp) program office of safety and mission assurance. Technical report, Pasadena, CA: Jet Propulsion Laboratory, National Aeronautics and Space Administration, 2011., 2011.

[Whi12] Mark White. Scaled cmos reliability and considerations for spacecraft systems: bottom-up and top-down perspectives. In *2012 IEEE International Reliability Physics Symposium (IRPS)*, pages 4B–4. IEEE, 2012.

[Wil06] MC Sansen Willy. Analog design essentials. *Springer*, 2006.

[WMYD01] Jen Shuang Wong, Jian Guo Ma, Kiat Seng Yeo, and Manh Anh Do. A New Approach for the Extraction of Threshold Voltage for MOSFETs. In *Technical Proceedings of the 2001 International Conference on Modeling and Simulation of Microsystems*, volume 1, pages 534–537, 2001.

[WWKB87] Hon-Sum Wong, Marvin H White, Thomas J Krutsick, and Richard V Booth. Modeling of transconductance degradation and extraction of threshold voltage in thin oxide MOSFET's. *Solid-State Electronics*, 30(9):953–968, 1987.

[WYG10] Weimin Wu, Wei Yao, and Gennady Gildenblat. Psp-soi: A surface-potential-based compact model of soi mosfets. In *Compact Modeling*, pages 41–74. Springer, 2010.

[XDH+03] X Xi, Mohan Dunga, Jin He, Weidong Liu, Kanyu M Cao, Xiaodong Jin, Jeff J Ou, Mansun Chan, Ali M Niknejad, Chenming Hu, et al. BSIM4. 3.0 MOSFET Model. *UC Berkley*, 2003.

[Yuh02] Yuhua Cheng & Chenming Hu. *Mosfet Modeling & BSIM3 User's Guide*. Kluwer Academic Publishers, Boston, 2002.

[ZDJZ16] JF Zhang, M Duan, Z Ji, and W Zhang. Defects for random telegraph noise and negative bias temperature instability. In *Semiconductor Technology International Conference (CSTIC), 2016 China*, pages 1–3. IEEE, 2016.

[Zja16] Amir Zjajo. *Stochastic Process Variation in Deep-Submicron CMOS*. Springer, 2016.