Untersuchung der kapazitiven Eigenschaften von 3D-Clusterstrukturen in einer 0,35 µm CMOS-Technologie und deren mögliche kryptografische Anwendungen

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von
Dipl. Inf. Matthias Harter
aus Offenburg

Mannheim, 2007

Dekan: Professor Dr. Matthias Krause, Universität Mannheim
Referent: Professor Dr. Peter Fischer, Universität Mannheim
Korreferent: Professor Dr. Reinhard Männer, Universität Mannheim

Tag der mündlichen Prüfung: 11. September 2007

Abstract

In this work, a novel method for generating reproducible random on-chip keys for cryptographic applications is presented. The technique proposed herein is based on the finding, that complex and irregular structures of randomly intertwined interconnect lines, the so-called 3D-clusters, can be regarded as the teeth of a key and can be used as such, if the secret key is derived from the capacitance of these structures. For this purpose, analog circuitry has been developed which is able to measure capacitances in the Femto-Farad region and has a resolution of 0.1 fF and below.

Zusammenfassung

In dieser Arbeit wird ein neuartiges Verfahren zur Generierung von reproduzierbaren, zufälligen „on-chip"-Schlüsseln für kryptografische Anwendungen präsentiert. Die hierbei vorgeschlagene Technik basiert auf der Erkenntnis, dass komplexe und irreguläre Strukturen von zufällig ineinander verwobenen Verbindungsleitungen, die sogenannten 3D-Cluster, als Bart eines Schlüssels aufgefasst und in dieser Weise verwendet werden können, wenn der geheime Schlüssel aus der Kapazität dieser Strukturen abgeleitet wird. Zu diesem Zweck wurde eine analoge Schaltung entwickelt, die in der Lage ist Kapazitäten im Femtofarad-Bereich zu messen und eine Auflösung von 0,1 fF und darunter aufweist.

Untersuchung der kapazitiven Eigenschaften von 3D-Clusterstrukturen in einer 0,35 µm CMOS-Technologie und deren mögliche kryptografische Anwendungen

Matthias Harter

Vorwort

Hirnforscher und Psychologen sind bekanntermaßen der Ansicht, dass Erfahrungen in unserer Kindheit und Jugend einen besonderen Einfluss auf unsere Entscheidungen und Interessen im Erwachsenenalter haben. Pädagogen versuchen möglichst früh das Interesse der Kleinen für bestimmte Dinge zu wecken und sie auch für zunehmend kompliziertere Sachverhalte zu begeistern.

In den achtziger Jahren des vergangenen Jahrhunderts erlebte ich wie viele Heranwachsende das aufkommende Computer- und Internet-Zeitalter als eine neue und faszinierende Welt, die es zu entdecken galt. Die sogenannten „Homecomputer“ ermöglichten den preiswerten Einstieg und bildeten den Ausgangspunkt für eine eigene Subkultur. Namen wie „C64“ und „Atari“ sind längst zur Legende geworden.

Bild 1.1. Der Schneider CPC 6128 Homecomputer aus dem Jahre 1985. (Quelle: „HCM - The HomeComputer Museum“).

In dieser Zeit erhielt ich meinen ersten Computer der Marke „Schneider CPC 6128“. Er war wie der C64 ein 8-Bit Homecomputer und besaß als Besonderheit einen Hauptspeicher von 128 Kilobyte. Er konnte nicht an einen Fernseher angeschlossen werden, sondern nur an einen Monochrom-Monitor von Schneider, der sein Bild in grüner Farbe darstellte. Aus heutiger Sicht würde man ihn als wenig leistungsfähig bezeichnen und kein Jugendlicher

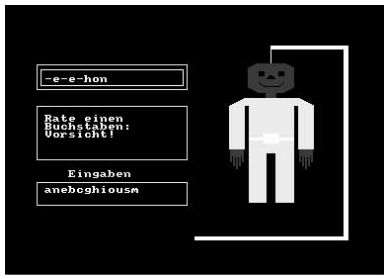


Bild 1.2. Das Ratespiel „Wordhang“ auf dem CPC. Damals faszinierten selbst einfachste Programme, heute regt es mehr zum Schmunzeln an.

Bild 1.3. Der „magische“ Befehl, der den Textfeldrahmen des CPC zum Blinken brachte.

würde sich für die magere Grafik und die schlechte Tonwiedergabe begeistern. Selbst heutige Mobiltelefone bieten dagegen ein wahres Feuerwerk an audiovisuellen Eindrücken.

Diesen Computer verstand ich bis an die Grenzen seiner Speicherkapazität mit endlosen Zeilen an Programmtext zu füllen. Das Ende war mit einer lapidaren Fehlermeldung erreicht: „*Out of memory*“. Nun galt es den Computer nicht mehr nur durch die Größe der Programme an seine Grenzen zu bringen, sondern mit wenig Anweisungen möglichst bunte und abwechslungsreiche Effekte zu erzielen. Als Vorbild hatte ich ein Demonstrationsprogramm des Herstellers, das gerne auf Werbemessen gezeigt wurde und als Verkaufsargument dienen sollte. Es bot mir eine Fülle an Programmieranregungen, schließlich hatten seine Programmierer versucht, das Beste aus dem Computer herauszuholen. Es zeigte einige Ausschnitte aus einem Textverarbeitungs- und Tabellenkalkulationsprogramm, um den möglichen Einsatz im Büro zu unterstreichen. Als Heimanwender war ich jedoch besonders von dem blinkenden Textfeldrahmen und den animierten Grafikelementen fasziniert.

Wie hatte der Programmierer diese Effekte nur erreicht ohne eine „*Out of memory*“ Fehlermeldung zu bekommen? Wie erzeugte man einen blinkenden Rahmen, ohne eine umständliche Endlosschleife einzusetzen, die zwischen hell und dunkel hin- und herschaltete? Die einzig verfügbare Programmiersprache war BASIC und stellte keine Anweisung zur Verfügung, mit der dieses Blinken sinnvoll hätte erzeugt werden können.

Dieser faszinierende Effekt des Blinkens weckte also meine Neugierde und es galt die magischen Befehle ausfindig zu machen, die ein solches Wunderwerk bewirken konnten. Ich hatte mir fest vorgenommen, am Beispiel des Demonstrationsprogrammes möglichst viel für meine eigenen Programmierkünste zu lernen.

Was sollte einfacher sein, als sich den Programmtext des Demonstrationsprogrammes anzuschauen, um sich aus den Befehlsfolgen die Kniffe des Programmierers abzuschauen und jenen geheimnisvollen Befehl zu finden, der einen Teil des Bildschirms wie von Geisterhand bedient zum Blinken bringen konnte, während das eigentliche Programm unabhängig davon weiterlief? Es musste ein ganz besonderer Befehl sein, eine Art Zauberspruch oder digitales Mantra. Nur eine moderne Beschwörungsformel vermochte in der Lage sein, das Blinken zu bewirken während gleichzeitig andere Anweisungen ausgeführt wurden. Die Idee der gleichzeitigen Abarbeitung von Befehlen, das Multitasking, zählte schließlich zu dieser Zeit bei Heimcomputern noch zum Science Fiction.

Programme im Speicher des Schneider CPC ließen sich gewöhnlich ganz einfach mit dem Kommando „LIST“ anzeigen. Man erhielt ein sogenanntes „Listing“ des Programmtextes, also jede Zeile der Anweisungsfolge nach Zeilennummern aufgelistet.

Nicht jedoch bei dem Demonstrationsprogramm! Hier verweigerte der Computer die Preisgabe des Quelltextes. Es stellte sich heraus, dass es einen speziellen Schutz gegen das Anzeigen des Programmtextes gab, mit dem jenes Programm versehen war. Es verhinderte jede Form der Anzeige des Quelltextes, sei es über das LIST-Kommando oder über einen indirekten Weg. Der Schutz lies sich über ein sogenanntes „Protect“-Attribut aktivieren, das mit dem Kommando „SAVE "dateiname", P“ der Datei beim Speichern auf dem Datenträger gegeben wurde.

Natürlich versuchte ich den Schutz auf irgendeine Weise zu umgehen, um doch noch an die versteckten Geheimnisse zu kommen. Doch der Schutz erwies sich als unknackbar für einen Elfjährigen und entsprechend machte sich bei mir Enttäuschung breit. Sollte es mir tatsächlich verwehrt bleiben, von den Programmierkünsten anderer zu lernen? War es nicht gerade das Lernen aus Beispielen, das Vormachen und Nachmachen, was bei Kindern und Jugendlichen besonders schnell zum Lernerfolg führen sollte?

Die im Entstehen begriffene Subkultur der „Hacker" und „Computerfreaks" hatte ihre eigenen Antworten parat. Der freie Zugang zu Informationen wurde gefordert und im sogenannten Hackerethos verbanden sich die Vorstellungen einer anarchischen Spielplatz-Ideologie mit den romantisierenden Allmachtsfantasien der Jugendlichen. Es galt als Sport, weniger noch als ein Kavaliersdelikt, an versteckte, geheime oder auf andere Art geschützte Informationen zu gelangen. Man hatte das Gefühl, der Protagonist eines Spionagethrillers zu sein und der Kalte Krieg bot genug Stoff für die Einbildungskraft der Helden des digitalen Sandkastens.

„I recall that same "kid in a candy store" feeling the first time I hacked a real BASIC app to do something really exotic like remove the copyright statement or include my name or something like that! In those days, we were all a lot easier to amuse." (Anonym, Auszug aus einer Internet-Diskussionsgruppe.)

So war es nur eine Frage der Zeit, bis ich die Wunderwaffe fand, die in der Lage war, die Tür in jene verschlossen geglaubte Welt zu öffnen. Ein kleines, gewitztes Programm, das ich eher zufällig fand, erlaubte es, den BASIC Quellcode-Schutz zu entfernen. Es war ein Schlüsselerlebnis, per Knopfdruck eine imaginäre Mauer einreißen zu können und eine neue Welt jenseits der Mauer entdecken zu dürfen. Triumphierend rannte ich durch das Haus, war begeistert, beglückt, berauscht. Mit einem Mal verschob sich die Grenze des Möglichen weit hinaus in die Ferne, der Himmel war die neue Grenze.

Und so war jenes Erlebnis ein sehr prägendes gewesen. Es lenkte mein Interesse auf alles „Gewitzte", Geheimnisvolle und Unbekannte. Ich wollte von nun an verstehen, wie all die Wunder der Technik dieser Zeit funktionierten, welche raffinierten Ideen manchen von ihnen einen ganz besonderen Zauber verliehen. Heutzutage würde man nach dem „Intellectual Property" fragen, nach der Innovation. Und in diesem Kontext taucht die Frage nach dem Eigentum an der Innovation und seinem Schutz auf. Genauso wie die Programmierer jenes Demonstationsprogrammes ihr Wissen, Können und ihre Ideen mit dem Quellcode-Schutz für sich beanspruchten, stellt sich heute die Frage nach dem Eigentumsschutz in der Wissens- und Informationsgesellschaft, in der nicht mehr materielle Produktionsfaktoren und die Förderung von Rohstoffen entscheidend sind, sondern immaterielle, geistige Güter. Diese Frage ist Ausgangspunkt der vorliegenden Arbeit.

Matthias Harter im Januar 2005

Inhaltsverzeichnis

Anmerkungen

Einige Punkte bedürfen einer gesonderten Behandlung. Am wichtigsten scheint es mir, zu betonen, dass diese Arbeit nicht nur ein wissenschaftliches Werk sein soll, sondern auch eine Art Anleitung für Studenten, Diplomanden und fachlich wenig versierte Leser. Aus diesem Grund habe ich an einigen Stellen Sachverhalte behandelt, die für den Experten selbstverständlich sind. In den grau unterlegten Boxen habe ich zudem Beispiele und Hilfen für den Einsteiger gegeben, der Fachmann kann sie überspringen, ohne wesentliche Inhalte zu verpassen.

Desweiteren möchte ich auf das ungewöhnliche Layout dieser Arbeit eingehen. Es entstand mit dem Ziel, eine hohe Zahl an Abbildungen in ansprechender Weise auf den Seiten unterzubringen und gleichzeit den Bezug zum dazugehörenden Text herzustellen. Über die durchgehenden Randspalten auf jeder Seite schien mir dies am ehesten möglich. Dadurch konnte ich auch im Nachhinein Abbildungen für textuell beschriebene Sachverhalte einfügen, ohne den Mengentext aufbrechen zu müssen, um den erforderlichen Platz zu schaffen. Störende Zeilen- und Seitenumbrüche konnte ich so vermeiden. Mit der hohen Zahl an Abbildungen wollte ich dem populären „PowerPoint-Stil“ Tribut zollen, frei nach dem Motto: „Ein Bild sagt mehr als tausend Worte“. Trotzdem habe ich versucht, die Zahl der oft wenig aussagekräftigen Mindmap-Charts und Blockdiagramme zugunsten inhaltlichen Gehalts auf ein Mindestmaß zu beschränken. Als Idealvorstellung schwebte mir eine Sammlung von Abbildungen vor, über die sich der Inhalt der Arbeit schnell, umfassend und mit wenigen Rückgriffen auf den Mengentext erfassen lassen sollte. Ich hoffe, diesem Ziel möglichst nahe gekommen zu sein.

Für den interessierten Leser möchte ich kurz einige technische Details nennen: Das spezielle Layout erforderte den Einsatz von Adobe FrameMaker (Version 7.0) für den Textsatz, obwohl bei sehr tiefgehenden Kenntnissen auch LaTeX möglich gewesen wäre. Der Grad der verwendeten Schriftart Linux Libertine im Mengentext betrug 10 Punkt bei einfachem Zeilenabstand. *Alle* Abbildungen, Graphen, Fotos etc. habe ich selbst erstellt. Als Werkzeug dienten mir dazu die Mal- und Zeichenfunktionen von FrameMaker selbst, sowie Origin von OriginLab und Mathematica von Wolfram Research (Graphen).

Zu guter Letzt ein stilistischer Hinweis: Im Zusammenhang mit dem Thema „wissenschaftliches Schreiben“ werden häufig zwei konträre Positionen vertreten. Die eine Partei ist für das Schreiben aus der Ich-Perspektive, die andere für die neutrale Formulierung über Passiv-Konstruktionen. Ich habe mich für letztere Philosophie entschieden, obwohl meines Erachtens nach keine objektiven Gründe eindeutig dafür sprechen. Letzten Endes ist es eine Frage des persönlichen Geschmacks. Der häufige Wechsel des Tempus zwischen Gegenwarts- und Vergangenheitsform sollte den Unterschied zwischen zwei Sachverhaltsarten verdeutlichen: Generelle Aussagen mit unbeschränkter Gültigkeit (Präsens) und Aussagen über von mir getroffene „design choices“, also Entscheidungen während des Entwicklungsprozesses (Präteritum).

Kapitel 1
Einführung

Ausgangspunkt dieser Arbeit ist in Abschnitt 1.1 die Formulierung einer Aufgabenstellung, die zum Ziel hat, Verfahren für kryptografische Schlüssel mit bestimmten, neuartigen Eigenschaften zu entwickeln. Es wird gezeigt, dass die wichtigste physikalische Größe solcher Schlüssel im Informationsgehalt besteht und ihr fundamentaler Zusammenhang mit den Begriffen „Zufall und Wahrscheinlichkeit“ erklärt. Als Motivation dafür, Lösungsansätze für die genannten Probleme zu finden, werden mögliche Anwendungen im Bereich des Erfindungs- und Innovationsschutzes, des Schutzes von Multimedia-Inhalten und privaten Daten und der Identifikation und Autorisation individueller Chips genannt.

In Abschnitt 1.2 wird der Stand der Technik skizziert, angefangen bei den klassischen Zufallsgeneratoren, die sich in die arithmetischen und die physikalischen Generatoren einteilen lassen. Letzteres sind echte Zufallsgeneratoren, die beispielsweise Rausch- oder radioaktive Zerfallsprozesse ausnutzen. Hierzu werden Beispiele in Form von Schaltungen gegeben und bewertet. Bei den arithmetischen Zufallsgeneratoren werden ebenso Standardbeispiele genannt und die typischen Anwendungen erörtert. Es wird gezeigt, dass diese Form der Zufallszahlen-Generierung berechenbar und daher für kryptografische Zwecke ungeeignet ist. Das Teilgebiet der Techniken zur Chip-Identifikation wird mit dem Prinzip des „eingravierten Zufalls“ eingeleitet und zwei Realisierungsformen vorgestellt. Schließlich wird jeweils eine der zahlreichen Techniken zum Einfügen von Seriennummern und digitalen Wasserzeichen in Hardware erklärt, wobei letztere hauptsächlich zum Erfindungs- und Innovationsschutz dient.

Den Abschluss des Kapitels bildet Abschnitt 1.3 mit der Vorstellung des neuartigen Lösungsansatzes in dieser Arbeit. Die Idee der 3D-Cluster wird erläutert und die kapazitätsbasierte Ableitung des kryptografischen Schlüssels skizziert. Es wird gezeigt, dass eine strukturelle und damit kapazitive Unbestimmtheit als Quelle der Entropie dient und schaltungstechnisch nutzbar gemacht werden kann. Schließlich wird anhand der Organisation der vorliegenden Arbeit ein Wegweiser zur Orientierung geboten, um die zu diesem Thema geleisteten Beiträge schnell und einfach ausfindig machen zu können.

* * *

1.1 Ausgangslage

1.1.1 Problemstellung

Der Schutz von Informationen vor unerlaubtem Zugriff ist eine Aufgabe mit ständig zunehmender Bedeutung. Durch Verschlüsselungstechniken wurden schon lange vor den ersten Rechnersystemen Daten geschützt, die grundsätzliche Idee ist also schon alt: Die Informationen werden mit geheimen Zusatzdaten – dem Schlüssel – nach wohldefinierten Regeln verknüpft, so dass der Inhalt nur mit Kenntnis des Schlüssels ermittelt werden kann. Der Schutz hängt also nur von der Sicherheit des Schlüssels ab. Dies entspricht dem berühmten in Kerckhoffs 1883 formulierten Prinzip.

Immer dann, wenn nicht der Besitzer eines informationsverarbeitenden Systems, sondern eine dritte Instanz, z.B. der Inhaber der Rechte an den Informationen, die letzte Kontrolle über den Schutz haben soll, wird eine ganz spezielle Lösung benötigt: Der Schlüssel muss dem System vorliegen, vor dem Besitzer aber geheim bleiben. In besonderen Anwendungsfällen muss der Schlüssel sogar zusätzlich dem Rechteinhaber bekannt sein, obwohl dieser beispielsweise aufgrund der räumlichen Distanz keinen direkten Zugriff auf das System hat.

Damit besteht die technische Aufgabe in der „*Bereitstellung eines Verfahrens zum Einprägen eines Schlüssels in ein informationsverarbeitendes Objekt („Hardware"), derart, dass der Schlüssel auch vor dem Besitzer geheim bleibt und idealerweise einen hohen Informationsgehalt (Schlüssellänge) aufweist, der von einer dritten Person (Rechteinhaber) oder der Hardware selbst stammt.*"

Bild 1.1. Die Zacken und Vertiefungen stellen im übertragenen Sinn die Schlüsseldaten in der Kryptografie dar. Neu ist die direkte Koppelung an die Hardware wie beim Bart eines echten Schlüssels. Die individuelle Form kann vorgegeben werden oder den zufallsverteilten Herstellungsschwankungen entspringen, so dass jeder Schlüssel ein Unikat darstellt. Mikroskopisch kleine, komplexe 3D-Strukturen aus Drähten auf einem Chip ähneln in dieser Weise dem Schlüsselbart.

Die Entropie

Der Informationsgehalt einer Quelle ist also von zentraler Bedeutung. In der berühmten Arbeit von Shannon aus dem Jahr 1948 wird der Zusammenhang mit der Wahrscheinlichkeit p_i für das Auftreten des i-ten Zeichens des von der Quelle verwendeten n-elementigen Alphabets hergestellt:

$$H = -\sum_{i=1}^{n} p_i log(p_i) \tag{1.1}$$

Aufgrund der Verwandtschaft mit dem Entropie-Begriff in der Thermodynamik wurde H von Shannon als (informationstheoretische) Entropie bezeichnet. Bei Verwendung des 2er-Logarithmus hat H die Einheit Bit pro Zeichen. Der Informationsgehalt eines geheimen Schlüssels hängt also direkt mit der Wahrscheinlichkeitsverteilung der Zeichen – allgemein formuliert der Statistik der Daten – zusammen. Je höher die Entropie, desto umfangreicher ist die im Schlüssel verborgene Information und desto stärker geschützt ist alles, was damit verschlüsselt wird.

Anforderungen

Die Entropie des Schlüssels kann im Wesentlichen zwei Quellen entstammen: Dem Eigentümer der zu schützenden Informationen (Person, Institution, etc.) oder dem informationsverarbeitenden System in der Hand eines potentiell

vertrauensunwürdigen Besitzers. Im ersten Fall findet eine Entscheidung über die Wahl des Schlüssels statt, während im zweiten Fall eine technisch vollzogene Generierung erfolgt, beispielsweise durch einen Zufallsgenerator.

Die harten statistischen Anforderungen, die an Zufallsgeneratoren typischerweise gestellt werden, sind nur bei den „klassischen" Zufallsgeneratoren anzusetzen, da sie einem Angreifer erlauben, eine große Zahl an Zufallszahlen zu generieren und durch statistische Analysen Informationen über die Wahrscheinlichkeitsverteilung zu bekommen. Für die Entropie des kryptografischen Schlüssels bedeutet dies den Verlust einer gewissen Anzahl an Bits. Beim Unterschreiten einer gewissen Grenze kann der Schlüssel dann durch Ausprobieren („brute force") ermittelt werden.

Bei den im Sinne der Aufgabenstellung geforderten Schlüssel-Generatoren hat ein Angreifer nur auf eine beschränkte Zahl an Schlüsselwerten Zugriff, da diese durch Einprägung unmittelbar an jedes einzelne Hardwareobjekt (z.B. Chip) gekoppelt sind. Geht man von einer gewissen Mengenbeschränkung (z.B. wegen der Kosten) bei diesen Objekten aus, steht somit die für statistische Analysen nötige breite Datenbasis nicht zur Verfügung und etwaige Schwächen wie z.B. nicht perfekte Gleichverteilung (systematischer „bias") bleiben dem Angreifer verborgen.

Den Extremfall stellt ein einmaliger Schlüssel (z.B. „one-time pad") dar, also ein in allen Objekten vorhandener, identischer Schlüssel. Solange dieser einem Angreifer vollständig unzugänglich bleibt, sind die statistischen Anforderungen minimal: Es kann sich um einen ausgewählten Datenwert in der Art eines Passworts handeln. Dieses Konzept soll in der vorliegenden Arbeit ebenfalls verfolgt werden.

1.1.2 Motivation

Denkt man bei der so formulierten Aufgabenstellung an konkrete Anwendungsfälle, so lässt sich eine Liste der verschiedensten Motivationsgründe erstellen. Ein spezieller Hauptaugenmerk stellt dabei (aus persönlichem Interesse) der Schutz solcher Informationen dar, die eine Erfindung oder zumindest Innovation darstellen. Alle weiteren Anwendungsfelder werden im Folgenden ohne Anspruch auf Vollständigkeit beispielhaft erwähnt.

Erfindungs- und Innovationsschutz

Der Schutz technischer Innovationen und Erfindungen wird im Allgemeinverständnis zunächst mit dem Patentschutz in Verbindung gebracht. Die Geschichtsforschung (siehe Kurz 2000) datiert die Entstehung des ersten Patentgesetzes in der Republik Venedig auf das Jahr 1474 und führt dieses auf eine bereits länger vorherrschende Praxis der Privilegienvergabe an Erfinder zurück. Diese Patent- und Privilegienvergabe entsprang bis zur Entwicklung einer Naturrechtstheorie einer rein wirtschaftspolitischen Motivation und schloss auch die Vergabe von Einführungsprivilegien ein, die an Personen vergeben wurden, die eine fremde Technik in den Staat einführten.

Neben dem rechtlichen Schutz von Erfindungen durch Patente war jedoch auch der Schutz durch Geheimhaltung von Bedeutung. Man versuchte zu vermeiden, dass Unberechtigte oder gar fremde Staaten am Erfolg einer Erfindung teilhaben konnten. Man hinderte sie entweder am Zugang zu allen

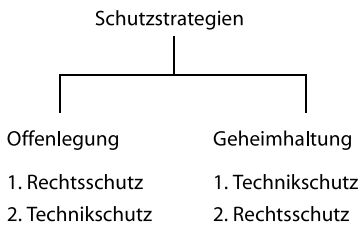


Bild 1.2. Es gibt grundsätzlich zwei Schutzstrategien. Bei der Offenlegung steht der Schutz durch Patente, Gebrauchsmuster o.ä. (Rechtsschutz) im Vordergrund, bei der Geheimhaltung der Schutz durch technische Maßnahmen wie z.B. kryptographische Verfahren oder Obfuskation.

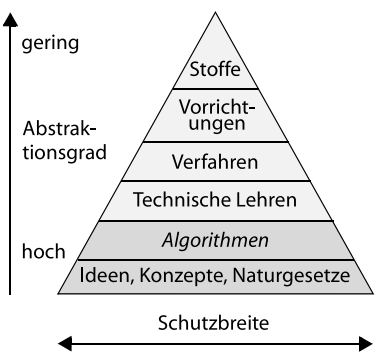


Bild 1.3. Der pyramidenförmige Aufbau der Gegenstände des geistigen Eigentums. Patente schützen meist nur konkrete Formen von Erfindungen (hellgrau), während die Kryptografie bereits Algorithmen zugänglich ist. Entsprechend breiter ist die Schutzwirkung.

erfinderischen Bestandteilen oder versuchte sie ganz geheim zu halten. Der Erfindungs- und Innovationsschutz wurde also bereits im 15. Jahrhundert unter dem Aspekt der Geheimhaltung gesehen.

In Bild 1.2 sind die beiden Herangehensweisen einander gegenübergestellt. Bei der Offenlegungsstrategie ist der Schutz hauptsächlich rechtlicher Art, meistens in Form von Patenten, Gebrauchsmustern, usw. Der Technikschutz ist dem nachgeordnet, ohne den rechtlichen Rahmen ist er wirkungslos. Beispiele hierfür sind die sog. Watermarking-Techniken, durch die robuste, schwer entfernbare Muster in Programme, logische Funktionen, Zustandsmaschinen und die Chip-Verdrahtungstopologie eingefügt werden, die dann zum Nachweis der Urheberschaft dienen und eine Rechtsverletzung (z.B. bei Produktpiraterie, Plagiaten) beweisbar machen (siehe Qu & Potkonjak 2003).

Dem gegenüber steht die Geheimhaltungsstrategie. Sie bedient sich in erster Linie der technischen Möglichkeiten, eine Erfindung geheimzuhalten, in der Regel durch Verschlüsselung, sofern es sich um Programme (siehe Kuhn 1996) oder die Konfiguration von programmierbaren Logikbausteinen handelt (z.B. Yip & Ng 2000). Aber auch festverdrahtete Schaltungen in Mikrochips können durch technische Tricks geschützt werden, etwa durch Verschleierungstaktiken („Obfuskation“, z.B. in Baukus et al. 1999), Zugriffsschutz („shielding“) oder -detektion (z.B. in Anderson 2001 und Taddiken & Laackmann 2000). Erst an zweiter Stelle kommen dann rechtliche Aspekte mit ins Spiel: Betriebs- bzw. Geschäftsgeheimnisse erfordern die Kennzeichnung des Schutzgegenstandes durch eine technische Barriere, um den Schutzwillen zu dokumentieren. Nur wenn etwas geschützt ist, kann der Unberechtigte dies als Unterlassungsaufforderung verstehen. Ebenso sind Gesetze gegen Datenspionage durch das sog. „Hacken“, gegen den Straftatbestand des elektronischen „Hausfriedensbruchs“ und gegen Urheberrechtsverletzung eng an die Existenz wirksamer (nicht „unumgehbarer“) technischer Schutzmaßnahmen geknüpft (ausführlich in Ernst 2004 und Wodtke & Richters 2004).

Um zum eigentlichen Thema der Arbeit zurückzukehren, sei ein Blick auf Bild 1.3 geworfen. Algorithmen haben darin den höchsten Abstraktionsgrad unten den Formen geistigen Eigentums, die in der Praxis noch als schutzfähig angesehen werden dürfen. Während ihnen in ihrer allgemeinen Formulierung aufgrund mangelnder Technizität[1] der Patentschutz meist versagt bleibt, können kryptografische Verfahren eingesetzt werden, um jedwede Ausgestaltung des Algorithmus in Form von Programmen oder rekonfigurierbarer Logik zu schützen. Durch den hohen Abstraktionsgrad wird so eine große Zahl an denkbaren Varianten vom Schutz erfasst.

Schutz von Multimedia- und Privat-Inhalten

Schützenswerte Informationen sind nicht nur Innovationen in Form von Programmen und Schaltungen, sondern allgemein Dateninhalte, die bei der Speicherung, Verarbeitung oder Übertragung potentiellen Gefahren ausgesetzt sind. In erster Linie sind dies Informationen, die der Privatsphäre entspringen und in fremde Hand geraten können, oder Multimedia-Inhalte, die ein Rechteinhaber in Form von Nutzungslizenzen an Endkunden weitergibt.

1. Technizität im Sinne des dt. Rechtsbegriffs. Der Algorithmus muss einen konkreten Bezug zur physikalisch-technischen Welt haben.

Als konkretes Beispiel ist der Kopierschutz bei der DVD und ihren Nachfolgern HD-DVD und Blue-Ray zu nennen. Hier wäre der Einsatz von geheimen Schlüsseln im Sinne der Aufgabenstellung denkbar, je nach Art und Ausgestaltung der gewünschten Sicherheits-Infrastruktur. Selbstverständlich gilt dies auch, wenn entsprechende Lösungsansätze durch Standards und fertige Produkte bereits vorliegen. Der erfolgreiche „Hack" des DVD-Kopierschutzes zeigt, dass hier Bedarf an konzeptionell neuen Ansätzen bestehen könnte.

Als weiteres Beispiel ist die Verteilung von Multimedia-Daten zu nennen, zum einen im kleinen Maßstab, etwa von einem Speichermedium zum Prozessor und von diesem zum Display. Zum anderen über größere räumliche Distanzen, z.B. von einem virtuellen Filmverleih über das Internet auf ein mobiles Endgerät. Ebenso besteht möglicherweise Bedarf bei der Verteilung von zukünftigen digitalen Filmrollen an Kinos[2]. Bei der Speicherung privater Inhalte auf Speichersticks und Festplatten werden Sicherheitsfragen inzwischen ebenfalls verschlüsselungstechnisch angegangen (z.B. „BitLocker"-Konzept der Fa. Microsoft).

Letztendlich sorgen die in dieser Arbeit anvisierten geheimen Schlüssel immer dafür, dass Informationen nur innerhalb des informationsverarbeitenden Chips unverschlüsselt vorliegen. In der Gefahrenzone außerhalb, also auf den Übertragungswegen, etwa zum Speicher, über das Internet, per Funk oder über Datenträger, sind die Daten geschützt. Das neuartige Schlüsselkonzept sorgt also dafür, dass der Informationsschutz unmittelbar an die Hardware, in der die Schlüssel eingeprägt sind, gekoppelt wird.

Im Rahmen der sog. „Trusted Computing" Initiative wird gegenwärtig die Entwicklung und Umsetzung einer umfassenden Sicherheitsarchitektur vorangetrieben, die eine digitale Rechteverwaltung („digital rights management", DRM) ermöglichen soll und dabei – wie im Namen enthalten – die Vertrauensfrage in die Informationsverarbeitung einbezieht. Der Endanwender verliert im Endeffekt die Kontrolle über die Schlüssel des Cryptosystems, indem diese in einem speziellen Chip, dem TPM-Baustein („Trusted Platform Module"), über herkömmliche Zufallsgeneratoren (siehe Abschnitt „Physikalische Zufallsgeneratoren") erzeugt und gespeichert werden. Fast alle Soft- und Hardware-Hersteller von Rang und Namen haben hierzu Produkte entwickelt[3]. Die in dieser Arbeit angestrebten in Hardware gegossenen Schlüssel könnten hierfür einen Beitrag leisten.

Autorisations- und Authentizitätsprüfung

Ein weiteres Anwendungsfeld ist bei einfacheren Sicherheitsarchitekturen zu finden. Sollen nicht wie im vorangehenden Fall große Datenmengen oder gar -ströme in einer komplexen, heterogenen Umgebung geschützt werden, sondern simple „Challenge-Response"-Protokolle implementiert werden, z.B. zur Autorisations- und Authentizitätsprüfung, so verschieben sich die Anforderungsprofile an die Verschlüsselungstechnik. Das in der Aufgabenstellung

2. „Digital Cinema Initiatives" unter http://www.dcimovies.com/
3. Codename von Microsoft: „Next-Generation Secure Computing Base" (NGSCB). Vormals „Palladium".

formulierte Konzept kann dadurch besonders attraktiv werden, insbesondere im Zusammenhang mit Kleinstrechnern in Smart-Cards und Zugangs-Kontrollsystemen, kurz bei den „embedded systems“.

Unter den in jüngster Zeit häufig anzutreffenden Begriff „Ambient Intelligence“ fallen Systeme, bei denen eine große Zahl an verteilten, kostengünstigen Funktionseinheiten miteinander intelligent interagieren und so in ihrem Zusammenspiel komplexes Verhalten aufweisen. Beispiele sind Sensor-Netzwerke aus „silicon dust“, also einfache, winzige Chips mit Mess- und Kommunikationsfunktion. Werden auf diese Weise zwischen den Chips Daten ausgetauscht, stellt sich auch hier wieder die Sicherheitsfrage. Ähnliche Ansätze werden bei den RFID-Chips und den intelligenten Etiketten („smart labels“) verfolgt.

* * *

1.2 Stand der Technik

Ausgehend von der im vorangehenden Abschnitt formulierten Aufgabenstellung soll nun kurz beleuchtet werden, welche Techniken und Verfahren heute existieren, um Zufallszahlen zu generieren, die als Grundlage für kryptografische Schlüssel dienen. Darüber hinaus soll ein Überblick gegeben werden über bereits bekannte Wege zum Einprägen solcher Zufallszahlen in Halbleiterchips.

1.2.1 Zufallsgeneratoren

In der Kryptografie ist der Zufallsgenerator ein wichtiges Kernelement. Insbesondere bei der asymmetrischen Verschlüsselung („public-key encryption") wie z.B. beim RSA-Algorithmus wird er zur Erzeugung des Paares aus öffentlichem und geheimem Schlüssel verwendet.

Je nach Anwendung haben diese Zufallsgeneratoren verschiedene statistische Eigenschaften. In der Regel wird eine Gleichverteilung der Zahlen erwünscht, eine schnelle Erzeugung und die Unvorhersagbarkeit der Ergebnisse, also indeterministisches Verhalten. In der Praxis können diese Vorgaben nicht alle eingehalten werden, insbesondere bei rein rechnergestützter Kryptografie. Hierbei bereitet der konstruktionsbedingte Determinismus Schwierigkeiten, Computer sind schließlich so ausgelegt, dass sie gerade nicht unvorhersehbare Ergebnisse liefern.

Arithmetische Zufallszahlengeneratoren

Hierbei handelt es sich um Verfahren zur *Berechnung* von Zufallszahlen, also um deterministische Generatoren. Das Ergebnis ist bei Kenntnis der Eingangswerte vollständig reproduzierbar, so dass sie auch als „Pseudo-Zufallszahlengeneratoren" bezeichnet werden. Das Berechnungsverfahren sorgt dabei für eine annähernde Gleichverteilung.

Theoretisch denkbar sind Zufallszahlen, die aus irrationalen Zahlen wie π hervorgehen. In der praktischen Realisierung auf dem Rechner scheidet dieser Weg jedoch aus, da sich irrationale Zahlen nur als Näherungswerte mit endlicher Genauigkeit darstellen lassen. Stattdessen kommen rekursive Funktionen zum Einsatz, die aus einer Menge von Startwerten iterativ Zufallszahlen errechnen. Diese dienen zur Initialisierung des Verfahrens („randomseed") und werden beispielsweise durch Benutzerinteraktion (Mauszeiger, Tastendruck) oder den Millisekunden-Teil der Rechneruhr gebildet. Auf diese Weise wird der Determinismus der Berechnung mit dem (bedingten) Indeterminismus der Eingaben von Außen kombiniert und das Ergebnis so weitestgehend unvorhersehbar gemacht.

Rechnergestützte Generatoren. In die Klasse der Kongruenzgeneratoren fallen die meisten der heutzutage in Programmiersprachen verwendeten Zufallsgeneratoren. Beispiele sind die linearen, multiplikativen und inversen Kongruenzgeneratoren, sowie die Fibonacci-Generatoren. Im Standardwerk von Knuth 1969 wird der lineare Kongruenzgenerator ausführlich beschrieben (siehe auch Bild 1.4). Der Mersenne-Twister ist ein relativ neuer Generator und wurde in Matsumoto & Nishimura 1998 vorgestellt. Er weist eine extrem lange Periode von $2^{19937} - 1$ auf, ist sehr schnell und hat hervorragende statistische Eigenschaften.

Parameter:

Modulus	m	$0 < m$
Faktor	a	$0 \le a < m$
Inkrement	c	$0 \le c < m$

Startwert:

Seed	X_0	$0 \le X_0 < m$

Linear kongruente Sequenz:

$$X_{n+1} = (aX_n + c) \bmod m$$

Bild 1.4. Linearer Kongruenzgenerator (nach Knuth 1969). Für $X_0 = a = c = 7$ und $m = 10$ liefert der Generator die Sequenz 7, 6, 9, 0, 7, 6, 9, 0, ... Nur wenn der Satz von Knuth erfüllt wird, entspricht die Periodenlänge dem Maximum m.

Keiner dieser Algorithmen wird gegenwärtig in der Kryptografie verwendet, da sie den harten Anforderungen nicht genügen. So lassen sich im Fall des linearen Kongruenzgenerators die Parameter a und c bereits nach wenigen Werten bestimmen und damit die ganze Sequenz. Stattdessen werden spezielle Einweg-Funktionen (z.B. SHA-512 und MD5) oder Blockchiffren (z.B. DES) verwendet, um eine Folge von Seed-Werten in eine Zufallszahlensequenz mit bestimmten kryptografischen Eigenschaften zu überführen. In der NIST Publikation 800-90 aus dem Jahr 2006 wird dieses Thema ausführlich behandelt.

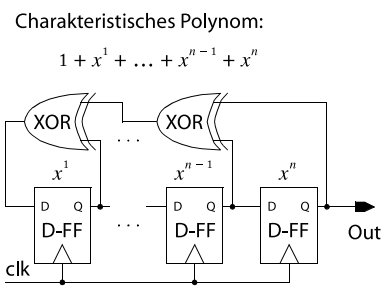


Bild 1.5. LFSR als Pseudo-Zufallszahlengenerator. Die Koeffizienten des charakteristischen Polynoms bestimmen, ob an der durch den Exponenten gegebenen Position ein Abgriff erfolgt (1) oder nicht (0). Die Zählung beginnt links bei x^1, der Term $x^0 = 1$ bleibt unberücksichtigt.

LINEAR FEEDBACK SHIFT-REGISTER. Bei vielen Anwendungen steht keine Recheneinheit zur Verfügung, um komplizierte Algorithmen zur Zufallszahlen-Generierung auszuführen. Stattdessen sollen mit minimalem Hardware-Aufwand zufällige Bitsequenzen erzeugt werden. Die einfachste schaltungstechnische Realisierung stellt ein lineares Schieberegister mit Rückkoppelungspfad („linear feedback shift-register", LFSR) dar, eine Lösung, die auch zu anderen Zwecken sehr häufig eingesetzt wird, beispielsweise als schnelle, platzsparende Zähler. Als Zufallsgenerator in der Kryptografie eignen sich LFSR in ihrer Reinform aufgrund des Determinismus nicht.

In Bild 1.5 ist der prinzipielle Aufbau eines LFSR zu sehen. Kernelement ist die Kette aus n Registern (D-FF), die mit einem beliebigen Bitmuster (außer dem Nullvektor) initialisiert wird. Das Flipflop an vorderster Stelle wird mit dem Ergebnis einer logischen XOR-Verknüpfung der Bits an bestimmten Positionen gefüttert. Ob an einer bestimmten Bit-Position ein Abgriff vorgenommen wird, hängt von den Koeffizienten des charakteristischen Polynoms ab. Dieses wiederum kann so gewählt werden, dass die Periodenlänge der Bitsequenz des LFSR dem theoretischen Maximum von $2^n - 1$ entspricht. Bei einem 32-Bit LFSR kann das charakteristische Polynom beispielsweise die Form $1 + x^2 + x^6 + x^7 + x^{32}$ haben, die Abgriffe befinden sich also an den Stellen 2, 6, 7 und 32 und führen in ein 4-fach XOR-Gatter. Erst nach ca. 4,3 Mrd. Zyklen wiederholt sich das Bitmuster bei einem derartigen 32-Bit LFSR. Da es sich hierbei um ein sogenanntes „primitives Polynom" handelt, ist die Sequenzlänge maximal.

Die Berechnung des primitiven Polynoms eines n-Bit LFSR kann mit Hilfe der Theorie endlicher Körper (Galois-Felder) erfolgen. Im Standardwerk von Lin & Costello 1983 über Fehler-korrigierende Codes wird dies ausführlich beschrieben. Zu den Details der kryptografischen Eigenschaften LFSR-basierender Pseudo-Zufallsgeneratoren ist die Dissertation von Zenner aus dem Jahr 2004 zu empfehlen.

Physikalische Zufallsgeneratoren

Durch die Kombination der arithmetischen Pseudo-Zufallsgeneratoren mit indeterministischen Startwerten werden in der Praxis kryptografisch sichere Zufallszahlen erzeugt. Nur physikalische Vorgänge können eine absolute Zufälligkeit und damit Unvorhersagbarkeit garantieren, beispielsweise bei radioaktiven Zerfallsprozessen[4], in der Quantenphysik oder durch Ausnutzung des thermischen Rauschens von Widerständen.

4. http://www.fourmilab.ch/hotbits/

Während die erstgenannten Möglichkeiten außer bei Studienobjekten und teueren Speziallösungen kaum in Standardprodukten realisiert wurden, finden sich Zufallsgeneratoren auf der Basis von Rauschprozessen oder dem chaotischen Verhalten spezieller Schaltungen in einigen Massenprodukten, darunter der Zufallsgenerator der Firma Intel.

Oszillator-basierend. Hierbei handelt es sich um ein Paar aus Oszillatoren mit verschiedenen Frequenzen. Der langsame Oszillator dient als Taktsignal für ein D-Flipflop, das den Zustand des schnellen Oszillators abtastet (siehe Bild 1.6). Aufgrund der Frequenzinstabilität und des Flanken-Jitters ist der Abtastzeitpunkt einer statistischen Verteilung unterworfen, so dass entweder eine Null oder eine Eins in das Flipflop übernommen wird.

Ganz so einfach wie in der Abbildung wird der Zufallsgenerator in der Regel jedoch nicht implementiert, da systematische Effekte beispielsweise den Phasenversatz der beiden Signale beeinflussen können. Generell ist nach Bucci & Luzzi 2005 darauf zu achten, dass jede als Zufallsgenerator eingesetzte Schaltung zurückgesetzt wird, um alle Zustandvariablen des Systems neu zu initialisieren und dadurch das interne „Gedächtnis“ zu löschen. Die Korrelation der Bits untereinander wird so minimiert. Der pragmatische Ansatz hierfür ist, die beiden Oszillatoren mit einem Start- und Stopp-Signal zu versehen und sie zusammen mit dem Flipflop nach jedem Bit in den Ausgangszustand zurückzuversetzen.

In Bock et al. 2004 wird der wohl neueste Stand der Technik zu diesem Verfahren dargestellt. Die Autoren sind Mitarbeiter der Firma Infineon, die im Bereich der Sicherheit eingebetteter Systeme und Chipkarten stark vertreten ist. Der in der Publikation beschriebene Zufallsgenerator wurde zum Patent angemeldet.

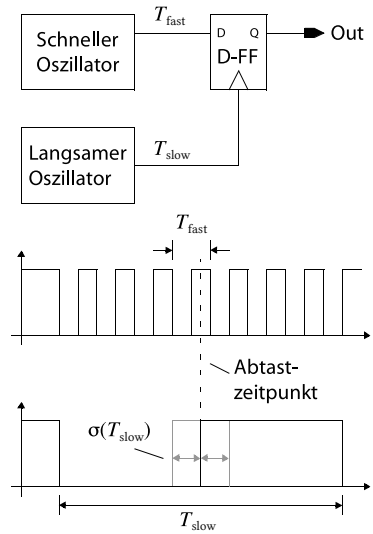


Bild 1.6. Oszillator-basierender Zufallsgenerator nach Bucci & Luzzi 2005.

Chaotische Schaltungen. In der Digitaltechnik sind metastabile Zustände eine bekannte Fehlerquelle für das indeterministische Verhalten von Schaltungen und werden deshalb entwurfstechnisch vermieden. Einen chaotischen Zufallsgenerator erhält man, indem diese Zustände stattdessen explizit erzwungen werden.

In der Literatur finden sich eine ganze Reihe von Schaltungsvorschlägen, darunter einige, die für Anwendungen in der Kommunikationstechnik entwickelt wurden. In Mandal & Banerjee 2003 wird der Chaos-Generator in Bild 1.7 vorgestellt. Auch hier empfiehlt es sich, ein Reset-Signal hinzuzufügen, mit dem der Kondensator nach Masse entladen werden kann, um jede „Erinnerung“ an vorangehende Bits zu löschen.

Das Verhalten der Schaltung wird durch die Wahl zweier Zeitkonstanten $\tau_1 = R_1 C$ und $\tau_2 = R_2 C$ und die Periode des Taktsignals bestimmt. An der steigenden Taktflanke wird der Kondensator mit τ_1 auf die Spannung V_1 aufgeladen. Erreicht die Spannung über C den Referenzwert V_R, so wird der Kondensator mit τ_2 bis zum Eintreffen der nächsten Flanke (teilweise) entladen. Da die Schaltungsanordnung sensibel auf Schwankungen im zeitlichen Verlauf der Auf- und Entladevorgänge reagiert, wechseln sich diese in chaotischer Weise ab: T_{clk} gibt ein festes Zeitraster für den Abbruch der Entladevorgänge vor, so dass eine Zwischenspannung erreicht wird, die als Startpunkt für den nächsten Aufladevorgang dient. Schwankungen der Zeitkonstanten wirken sich so im Gegensatz zum vollständigen Entladen nicht erst nach vielen Perioden durch eine langsam voranschreitende Phasenverschie-

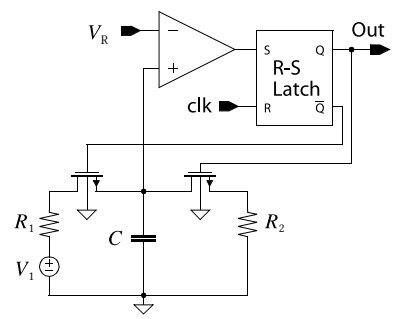


Bild 1.7. Chaos-Generator nach Mandal & Banerjee 2003.

bung zum Taktsignal aus, sondern werden sofort in einen indeterministischen Spannungswert umgewandelt, der wiederum über den Komparator den Beginn des nächsten Entladevorgangs bestimmt.

Der Latch-Ausgang dieser Schaltung kann wie in Bucci & Luzzi 2005 vorgeschlagen nochmals mit einem heruntergetakteten Flipflop abgetastet werden, um das Ergebnis einer gewissen Anzahl an Auf- und Entladevorgängen abzuwarten. Nach der Evaluation sollte die Schaltung durch den bereits erwähnten Reset-Eingang wieder vollständig zurückgesetzt werden, um eine (wenn auch geringfügige) Abhängigkeit bzw. Korrelation der Bits untereinander auszuschließen.

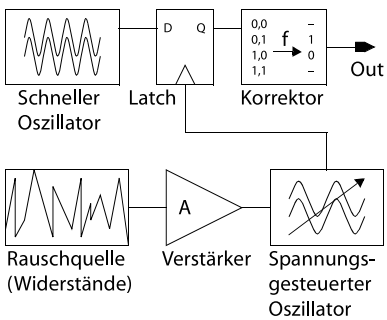


Bild 1.8. Schematische Darstellung des Zufallsgenerators der Fa. Intel (nach Jun & Kocher 1999).

INTELS ZUFALLSZAHLENGENERATOR. Beim Zufallsgenerator der Firma Intel wird das Prinzip der zwei Oszillatoren mit einer weiteren, häufig verwendeten Entropiequelle kombiniert: Das thermische Rauschen von Widerständen (siehe Bild 1.8). Um den gemeinsamen Einfluss von Schwankungen der Umgebungsbedingungen (Temperatur, Versorgungsspannung, etc.) auf die Bits des Zufallsgenerators zu reduzieren und damit die Korrelation zu minimieren, wird das Rauschen von zwei benachbarten Widerständen gemessen und differentiell weiterverarbeitet. Ein breitbandiger und hochgradig linearer Verstärker steuert damit einen spannungsgesteuerten Oszillator (VCO) an, der die Rolle des langsamen Schwingkreises in Bild 1.6 übernimmt. Entsprechend legt er über den Takteingang am Latch den Zeitpunkt der Abtastung des schnellen Oszillators fest. In der Publikation von Jun & Kocher 1999 wird die Standardabweichung der Frequenz des VCO in Relation zur Frequenz des schnellen Oszillators mit 10–20 angegeben, so dass die Kombination aus Rauschquelle und Verstärker offensichtlich wesentlich zur Unbestimmtheit des Gesamtsystems beiträgt.

Ein nachgeschalteter digitaler Korrektor (sog. „von Neumann corrector“) sorgt für eine ausgeglichene Verteilung der Nullen und Einsen am Ausgang des Generators. Dazu werden jeweils zwei Bits zu einem Ergebnis kombiniert, falls sie sich unterscheiden. Sind beide gleich, wird kein Bit ausgegeben, so dass die Datenrate am Ausgang variiert. Im Schnitt werden 75 Kilobit pro Sekunde erreicht.

1.2.2 Chip-Identifikation und Autorisation

Zu einem steigenden Bedarf an speziellen kryptografischen Hardware-Lösungen in den letzten Jahren führten auch Anwendungsfälle, in denen die Identifikation einzelner Chips nötig wurde. Damit sollte ermöglicht werden, die Herkunft, Authentizität und Autorisation von Endgeräten in unsicherer Hand zu überprüfen.

Eingravierter Zufall

Das wohl wichtigste Konzept in dieser Arbeit stellt der in mikroskopischen Effekten enthaltene und durch Einprägung reproduzierbar gemachte Einfluss von physikalischen Zufallsprozessen dar. Diese Idee wurde unabhängig von Vorarbeiten anderer Autoren zu Beginn entwickelt. Insbesondere das Ausnutzen der Verteilung von Komparatoroffsets bzw. der Transistor-Schwellenwertunterschiede wurde zunächst als eine Konzeptvariante verfolgt und zu

einen Testchip weiterentwickelt, ehe die Publikation von Lofstrom et al. aus dem Jahre 2000 ausfindig gemacht wurde, was die Fortentwicklung dieses konkreten Lösungsweges überflüssig machte.

Transistor-schwellenwertdispersion. In Bild 1.9 ist der Schaltungsvorschlag von Lofstrom, Daasch & Taylor zu sehen. Kernelement ist eine Reihe oder Matrix aus NMOS-Transistoren, die bei einer gemeinsam vorgegebenen Gate-Spannung aufgrund der Unterschiede ihrer Schwellenwerte zu verschiedenen Drain-Strömen führen. Jeder dieser Transistoren ist über statische Schalter einzeln ansprechbar bzw. zu einem Lastwiderstand hinzuschaltbar. Die Stromunterschiede führen über diesen zu Spannungsunterschiedenen, die über einen speziellen Komparator in ein Bitmuster überführt werden, falls die Transistoren auf diese Weise sukzessive durchmessen werden. Er dient im Wesentlichen zur Erkennung der Spannungsdifferenz zwischen jeweils zwei nacheinander selektierten Transistoren. Das resultierende Bit gibt also an, ob beim Wechsel von einem Transistoren zum nächsten eine positive Flanke an der Drain-Spannung auftrat oder eine negative.

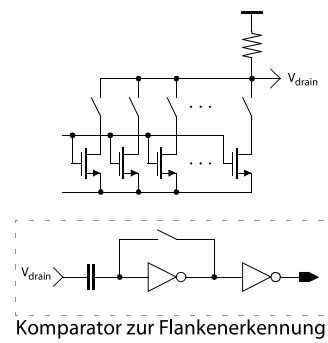


Bild 1.9. Schaltung zur IC-Identifikation durch Nutzung der Transistor-Schwellenwertdispersion („mismatch“). Nach Lofstrom, Daasch & Taylor 2000.

Die auf diese Weise gewonnene Bitsequenz repräsentiert letzten Endes den Mismatch der Prozessparameter, vor allem die Fluktuation der Dotierungsstärke und aller Faktoren, die Einfluss auf die Schwellenwertspannung haben. Diese Parameter können sich – beispielsweise durch das Eindringen von geladenen Fremdatomen – mit der Zeit ändern, so dass einige der Bits umkippen können. Darüber hinaus führt der Vergleich nicht immer zu eindeutigen Ergebnissen: Mit einer gewissen Wahrscheinlichkeit treten Transistorpaare auf, die sehr dicht beeinanderliegende Schwellenwerte aufweisen, so dass Bits durch Störsignale und Rauschen instabil sein können. Die Autoren schätzen, dass beide Effekte zusammengenommen zu weniger als 5 Prozent Bitfehlern führen. Bei einer Serienmessung an Testchips mit einer Sequenzlänge von 112 Bit traten nach 100-stündigem Erhitzen (250° C) maximal 6 Bitfehler auf (5,4 Prozent). Die Sequenzen zweier verschiedener Chips unterschieden sich immer um mindestens 27 Bit (Hamming-Distanz).

Mit der vorgeschlagenen Technik ist es also möglich, eine Datenbank mit den individuellen Bitsequenzen aller gefertigten Chips einer Produktionsreihe anzulegen, vergleichbar mit digitalen Fingerabdrücken. Die registrierten Chips können dann später durch Vergleichen der Hamming-Distanzen identifiziert werden. Die Anwendbarkeit dieser Technik als Grundlage für kryptografische Schlüssel wurde in der Publikation nicht untersucht. Als problematisch ist das Auftreten von Bitfehlern durch Umkippen anzusehen. Die Frage nach der Statistik der Sequenzen (Gleichverteilung, Korrelati0n, usw.) und damit der Entropie ist ebenfalls offen. Zu vermuten ist, dass im Falle eines Chip-weiten Gefälles (Gradienten) bei der Dotierungsstärke (detailliert in Abschnitt 2.1 ff.) starke Bit-zu-Bit Korrelationen auftreten, so dass die Gesamtentropie wesentlich sinkt.

Signallaufzeit-basierend. Einen etwas anderen Ansatz verfolgt die Gruppe um S. Devadas am Massachusetts Institute of Technology im Rahmen eines von zahlreichen hochrangigen Firmen finanzierten Projektes mit Namen „Oxygen“. Die grundlegende Idee besteht darin, die von einer Menge von Eingangsvektoren („challenge“) abhängige Signallaufzeit durch eine spezielle Anordnung von Logikgattern hindurch zu ermitteln und als individuelle, unvorherbestimmbare Antwort („response“) zu interpretieren. Hauptanwen-

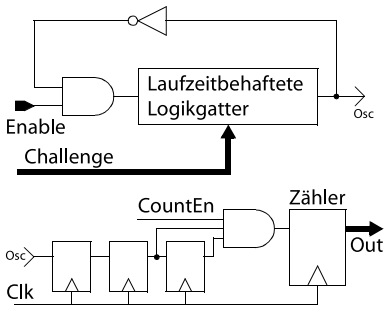


Bild 1.10. Vereinfachte Darstellung eines Ringoszillator (oben) mit Taktsynchronisation und Flanken-Zähler (unten). Wird die Schaltung für eine gewisse Anzahl Takte aktiviert, so misst sie indirekt über die Frequenz des Oszillators die Signallaufzeit durch die Logikgatter (nach Gassend et al. 2002).

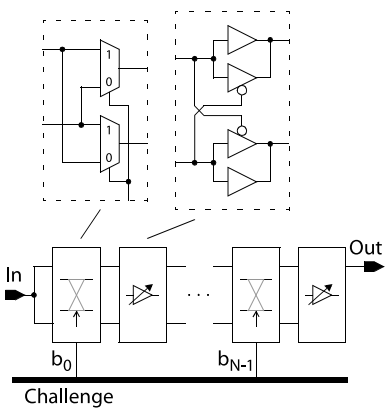


Bild 1.11. Kette aus Logikgattern mit einer Signallaufzeit, die von dem angelegten Eingangsvektor („challenge“) abhängt. Durch die einstellbaren Verzögerungselemente ist die Laufzeit eine nicht-monotone Funktion des Challenge (nach Gassend et al. 2002).

dungsgebiet ist dementsprechend die Implementierung eines „challenge-response“-Protokolls (siehe hierzu Kapitel 2 „Protocols“ im Standardwerk von Anderson 2001).

In Bild 1.10 ist zunächst die Schaltungsanordnung zu sehen, die mithilfe eines taktsynchronen Binärzählers (unten) die Frequenz eines Ringoszillators misst, die umgekehrt proportional ist zur Signallaufzeit durch die Kette von Logikgattern, aus der sich der Ring zusammensetzt (oben), und daher als Maß für die physikalischen Effekte dient, die Einfluss auf die Verzögerung der Gatter haben. Es sind dies die mikroskopischen, unvorhersagbaren Zufallsprozesse, die in jeden Chip bei der Fertigung ein individuelles Muster einprägen und die Identifikation erst ermöglichen. Darüber hinaus sind die Verzögerungsglieder durch den Eingangsvektor parametrisierbar, d.h. eine Funktion des Zufalls *und* des Challenge.

Eine Kette an Logikgattern, die eine solche Funktion erfüllt, ist in Bild 1.11 zu sehen. Der Eingangsvektor bestimmt im Wesentlichen den Signalpfad durch die Kette: Jedes Bit $b_0 \ldots b_{N-1}$ des N-bit Challenge gibt an, ob die beiden Signale (die zwei „Kopien“ der zirkulierenden Flanke) in der entsprechenden Stufe ihren Weg kreuzen oder parallel zueinander verlaufen. In Stufe $N-1$ wird schließlich das an der letzten Abzweigung selektierte Signal in den Rückkoppelungspfad des Oszillators eingespeist.

Neben den Kreuzpunkt-Elementen besteht jedes Glied der Kette aus Buffern, die mit einer a-priori unbekannten und zufallsverteilten Verzögerung (z.B. wegen Schwankungen der Leitungskapazitäten) behaftet sind. Um zu verhindern, dass ein potentieller Angreifer ein additives Modell (lineares Gleichungssystem) der Signallaufzeit der gesamten Kette erstellt und die Unbekannten durch wenige Messungen ermittelt, wurden die Verzögerungsglieder etwas komplizierter gestaltet: Jedes Glied besteht aus der Kombination von einem langsamen Buffer geringer Treiberstärke mit einem schnellen Tri-State Buffer hoher Stärke. Der Steuereingang („active low“) wird mit dem jeweils anderen Eingang der beiden Signalpfade kreuzweise verbunden. Ist ein Tri-State Buffer aktiv, bestimmt er den Signalpegel am Ausgang[5]. Dadurch wird die Modellierung wesentlich erschwert, hängt die Verzögerung doch vom Zustand ab, in dem sich jede Stufe befindet. Die Signallaufzeit der Kette ist damit von links nach rechts nicht-monoton steigend, eine schnelle Stufe kann zu einer größeren Gesamtverzögerung führen.

Durch die Auswahl des Signalwegs über den Challenge wird die Verzögerung der Kette also in komplizierter, schwer modellierbarer Weise variiert. Die Geschwindigkeit der im Ringoszillator zirkulierenden Signalflanke und damit seine Frequenz hängen also sowohl vom Challenge, als auch von den zufallsbedingten Schwankungen des Herstellungsprozesses ab. Weiterhin wird die Frequenz von den Betriebsparametern (Temperatur, Versorgungsspannung, etc.) stark beeinflusst, so dass eine zusätzliche Kompensation nötig ist: Statt die absolute Frequenz (der Stand des Zählers) als „Response“ im Sinne der Protokolls zu interpretieren, wird das Zähler*verhältnis* von zwei verschiedenen Oszillatoren gebildet.

5. Die Buffer treiben gegeneinander. Durch die unterschiedliche Treiberstärken-Dimensionierung wird dennoch ein fester Pegel erzwungen.

Um durch den Zählerstand eine hinreichend genaue Information über die Frequenz der Oszillatoren zu erhalten, ist es nötig, sehr lange zu messen, d.h. eine hohe Zahl an Taktperioden abzuwarten: Aus den Experimenten geht hervor, dass ca. 20 Millisekunden, also Hunderttausende Taktzyklen nötig sind (2^{20} Zyklen bei 50 MHz, siehe Gassend 2003).

Die auf diese Weise gewonnenen Messwerte sind jedoch einer Reihe von Störungen unterworfen, die zu Fehlern bei Wiederholung der Messungen führen, so dass die Antwort auf einen bestimmen Challenge falsch sein kann. Auch durch den Einsatz fehlerkorrigierender Codes in einer nachgeschalteten Stufe liegt die Wahrscheinlichkeit, die richtige Antwort zu bekommen, unter 50 Prozent. Daher müssen eine ganze Reihe von Eingangsvektoren durchgetestet werden, um die Identifikation einer großen Zahl Chips zu ermöglichen, die Geschwindigkeit des Systems insgesamt wird weiter reduziert.

Gründe für die Störung der Messungen liegen zum einen trotz der genannten Kompensation in der Abhängigkeit von der Versorgungsspannung und der Betriebstemperatur, zum anderen im Genauigkeitslimit der Messmethode (z.B. durch Rauschen). Darüber hinaus beeinflussen sich benachbarte Oszillatoren gegenseitig, beispielsweise durch elektromagnetische Abstrahlung (die zudem von Angreifern leicht von außen zur indirekten Frequenzmessung genutzt werden kann). Schließlich ist das unter dem Stichwort „Aging“ bekannte Problem des Parameterdrifts zu nennen, ein Phänomen, das auch im Schaltungsvorschlag von Lofstrom et al. 2000 (siehe Bild 1.9) Probleme bereitet.

Rechnet man diese Störeinflüsse verallgemeinernd dem Rauschen zu, so führt das Verfahren bei den experimentell ermittelten Signalen bezüglich zweier Chips zu einem Signal-Rausch-Verhältnis von 100. Umgerechnet bedeutet dies, dass pro Messung 3,3 Bit Entropie gewonnen werden. Die Identifikation einer großen Zahl an Chips kann deshalb nur über eine entsprechend langsame Mehrfachmessung ermöglicht werden. Ähnliches gilt für eine mögliche Adaption des Verfahrens zur Erzeugung kryptografischer Schlüssel.

Seriennummern und digitale Wasserzeichen

Die Identifikation von Chips anhand von individuellen Merkmalen kann auch über die Einprägung einer beim Entwurf oder der Herstellung festgelegten, eindeutigen Signatur erfolgen. Den einfachsten Fall stellen die Seriennummern (oder beliebig andere Datenfolgen) dar, die über eine Vielzahl von Vorgehensweisen in einen Chip eingeschrieben, eingespeichert oder sogar eingebrannt werden können. In den meisten Fällen wird die Bitfolge in die Topografie eines Chips eingebracht, indem der Knoten, der den Wert eines bestimmten Bits repräsentiert, entweder mit der Versorgungsspannung oder der Masse elektrisch verbunden wird.

Eine recht naheliegende Erweiterung dieses Prinzips besteht in der Idee, die Vorrichtung zur Festlegung der Bits über alle Verdrahtungsebenen (oder sogar bis in die Dotierungsschichten) zu verteilen. In Bild 1.12 ist eine solche Schaltungsvariante zu sehen. Als Wechselschalter kommen neben den abgebildeten Durchkontaktierungen im anderen Fällen auch spezielle Sicherungen („fuses“) in Frage, die z.B. aus dünnen Polysiliziumstücken bestehen und durch selektives Beschicken mit hohen Strömen durchbrennen, so dass die

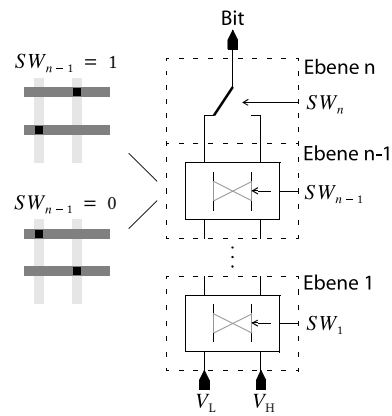


Bild 1.12. Simpler Seriennummer-Generator, der mit nur *einer* Maskenänderung eine Anpassung der Bitsequenz ermöglicht. Die als Kreuzschienen-Verteiler ausgelegten Wechselschalter sind in allen Ebenen des Chips vorhanden und können z.B. über die links gezeigte Anordnung von Durchkontaktierungen realisiert werden (nach Wagner 2003).

elektrische Verbindung aufgetrennt wird. Die Zahl der technischen Möglichkeiten hierfür, wie allgemein zur Seriennummern-Fixierung ist so groß, dass Bild 1.12 als Beispiel genügen soll.

Unter dem Begriff „Steganografie“ wurden in den letzten Jahren zahlreiche Techniken entwickelt, um Seriennummern, Kennzeichnungen und Urhebernachweise in die verschiedensten Formen digitaler Inhalte einzufügen, und zwar möglichst unbemerkt vom Nutzer der Daten. Diese als digitale „Wasserzeichen“ oder „Fingerabrücke“ bezeichneten Verfahren zielen in der Regel darauf ab, vorhandene oder absichtlich hinzugefügte Redundanzen in den Daten für die Signaturinformationen zu benutzen. Vergleichbar mit einer Unterschrift sollen diese die Urheberschaft der Daten beweisen, sollte es zu Rechtsverletzungen wie unerlaubtes Kopieren kommen. Je nach Anwendungsgebiet sollen die Wasserzeichen auch bei Manipulation der Daten erhalten bleiben, so dass eine gewisse Robustheit benötigt wird.

Der vielleicht bekannteste Fall, bei dem solche unsichtbaren Markierungen in Daten eingefügt werden, ist der Urhebernachweis bei digitalen Bildern. Auf diesem Gebiet ist die technische Entwicklung bereits soweit fortgeschritten, dass sich entsprechende Verfahren bereits in handelsüblichen Fotoapparaten wiederfinden. Weniger bekannt dagegen sind die Maßnahmen, mit denen digitale Wasserzeichen in Hardware-Produkte eingeprägt werden. Einen guten Überblick über den Stand der Technik bietet hierfür das Buch von Qu & Potkonjak 2003. Wegen der Fülle der Möglichkeiten sei hier nur ein beliebiges Beispiel herausgegriffen:

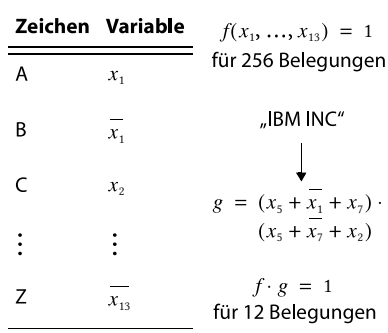


Bild 1.13. In vielen EDA-Algorithmen muss das SAT-Problem gelöst werden, also eine Variablenbelegung einer boole'schen Funktion f gefunden werden, für die $f = 1$ ist. Werden zur Funktion „constraints“ g hinzugefügt, die eine Signatur codieren, so reduziert sich der Lösungsraum beträchtlich. Der Nachweis der Urheberschaft geschieht dann über das Verhältnis der Wahrscheinlichkeiten, in den beiden Fällen die im Produkt realisierte Belegung per Zufall gefunden zu haben (nach Qu & Potkonjak 2003).

Bei vielen Entwicklungswerkzeugen der „Electronic Design Automation“ (EDA) kommen NP-harte oder NP-vollständige Algorithmen zum Einsatz, z.B. in der Logiksynthese bzw. -optimierung (siehe Entrena et al. 1993). Die optimale Lösung des zugrundeliegenden Berechnungsproblems kann deshalb nur in sehr einfachen Fällen gefunden werden, so dass die Berechnungsergebnisse häufig Redundanzen aufweisen. Das SAT-Problem beispielsweise fragt nach der Erfüllbarkeit einer boole'schen Funktion. Hängt diese von einer großen Zahl an Variablen ab, so kann es sehr viele Variablenbelegungen geben, die sie lösen, und das Auffinden einer einzigen trotzdem sehr rechenintensiv sein (siehe Beispiel in Bild 1.13).

Diese Tatsache kann dazu benutzt werden, den Lösungsraum durch zusätzliche Bedingungen („constraints“) einzuschränken, so dass der Algorithmus eine Belegung findet, die auch unter diesen Constraints zum Ergebnis „wahr“ führt. Sie beinhalten dabei eine codierte Signatur, so dass z.B. bei der Logiksynthese Schaltkreise entstehen, die in ihrer Struktur ein Abbild (Wasserzeichen) dieser Signatur aufweisen.

Der Nachweis einer Urheberrechts-Verletzung geschieht nun über die folgende Argumentation: Der Rechteinhaber kann für sich reklamieren, die Constraints g beim Entwurf der Schaltung(en) verwendet zu haben, da die strittige Hardware sie erfüllt, für das Funktionieren der Schaltung dieses aber nicht nötig gewesen wäre, sondern nur die Erfüllbarkeit von f. Die Wahrscheinlichkeit, dass der Rechteinhalber eine Belegung findet, die f *und* g löst, ist mit Kenntnis der Constraints sehr viel größer, als die Wahrscheinlichkeit, dass ein unberechtigter Nutzer der Hardware genau diese Belegung unter der viel größeren Zahl der Lösungen für f findet.

* * *

1.3 Neuartiger Lösungsansatz

In vorangehenden Abschnitt 1.2.2 wurde gezeigt, welche Techniken es nach heutigem Stand gibt, um wiedergewinnbare Hardware-Zufallszahlen zu erzeugen oder digitale Wasserzeichen in diese einzuprägen, z.B. durch Einflechten in die Verdrahtungsressourcen, das Einbringen zusätzlicher FSM-Zustände oder das Ausnutzen der Redundanz logischer Gleichungen. Für die reproduzierbaren Zufallszahlen werden die herstelltechnisch bedingten Schwankungen einiger physikalischer Parameter ausgenutzt, bei Lofstrom et al. 2000 ist dies hauptsächlich die Dotierungsstärke bzw. der Transistor-Schwellenwert. Bei Gassend et al. ist es die Signal-Durchlaufzeit von speziellen Verzögerungsglieder, die gleich von mehreren Parametern abhängt. In erster Linie sind dies die parasitären Leitungskapazitäten und -widerstände und zu einem gewissen Grad wiederum die Schwellenwerte der Transistoren.

Novum – kapazitätsbasierte Ableitung

Gerade dieser letzte Einflussfaktor bringt jedoch einen entscheidenden Nachteil mit sich: Die Schwellenwerte verschieben sich im Laufe der Zeit durch das als „Aging“ bezeichnete Phänomen, so dass die daraus abgeleiteten Zufallssequenzen nicht vollständig erhalten bleiben. Einige Bits werden instabil oder kippen um. Beim MIT-Ansatz von Gassend et al. ist die Abhängigkeit von der Versorgungsspannung und der Temperatur besonders stark, so dass nur eine differentielle Messmethode in Frage kommt. Dies geschieht über das Frequenzverhältnis der aus den Verzögerungsgliedern zusammengesetzten Ringoszillatoren, was eine sehr lange Messperiode erforderlich macht.

Der neue Ansatz, der in dieser Arbeit vorgeschlagen wird, besteht darin, allein die elektrische Kapazität zu messen – und zwar von sehr speziellen, dreidimensionalen Strukturen von Verbindungsleitungen, den 3D-Clustern (siehe Bild 1.14). Die Beschränkung auf die Kapazität hat den Vorteil der völligen Unabhängigkeit von den Betriebsbedingungen und allen Alterungsprozessen[6]. Nur die Messelektronik ist diesen Effekten ausgesetzt, nicht jedoch das Messobjekt selbst. Die Sicherheit der aus den Kapazitätswerten der Cluster gebildeten Bitsequenzen setzt sich dabei im Wesentlichen aus folgenden Formen der Unbestimmtheit zusammen:

- Strukturelle Unbestimmtheit.
  Bei der lithografischen Herstellung der Verbindungsleitungen eines Chips verursachen physikalische Zufallsprozesse auf mikroskopischer Ebene Unregelmäßigkeiten, die zu individuellen, statistischen Abweichungen von der beim Entwurf vorgegebenen exakten Struktur führen.
- Kapazitive Unbestimmtheit.
  Bei hinreichender Komplexität ist das elektrische Feld bzw. die Verteilung der Ladung mathematisch nur durch eine Differentialgleichung (Laplace-Gleichung) beschreibar. Für sie gibt es nur bei einfachen geometrischen Gebilden analytische Lösungen, bei komplizierten Strukturen sind dagegen rechenintensive numerische Verfahren nötig.

Bild 1.14. Vorbild für die Gestalt der 3D-Kapazitätscluster. Statt eines Kabelknäuels bestehen die Cluster aus irregulären, ineinander greifenden Metallbahnen. Das elektrische Feld von aufgebrachten Ladungsträgern ist auf komplizierte Weise verknotet und daher schwer zu berechnen.

6. Allein das als Elektronenmigration bezeichnete Phänomen könnte eine Rolle spielen.

Hinzu kommt für einen potentiellen Angreifer die Schwierigkeit, die Kapazität mit ausreichender Genauigkeit zu messen, um auf die daraus abgeleitete Bitsequenz zu schließen. Alle bekannten Messmethoden setzen hierfür voraus, dass eine elektrische Verbindung zum Messobjekt hergestellt wird, beispielsweise durch Kontaktnadeln auf einem Spitzenmessplatz. Aufgrund der Eigenkapazität solcher Nadeln kann das Ergebnis bereits signifikant verfälscht werden. Auch ist das Messverfahren selbst aufwendig.

Wollte ein Angreifer die individuellen Cluster-Strukturen eines vorliegenden Chips durch numerische Berechnung auf einem Rechner (sog. „Extraktion“, siehe Abschnitt 2.2.3 auf Seite 43) ermitteln, so müsste er die dreidimensionale Form jedes Objekts *exakt* modellieren und sehr teuere, spezialisierte Software einsetzen. Keines der gängigen Extraktionswerkzeuge unterstützt zur Zeit die Modellierung der durch Beugungseffekte bei der Lithografie entstehenden Kantenabrundungen der Metallisierung. Stattdessen werden die sonst üblichen Winkel von 90 Grad (manchmal 45 Grad) vorausgesetzt. Der Angreifer hätte darüber hinaus die Schwierigkeit, die Geometrie der Cluster überhaupt genau genug dem Chip entnehmen zu können. Er müsste alle Metallisierungsebenen sukzessive abtragen, mikroskopisch analysieren und im Rechner wieder detailgetreu zu einem dreidimensionalen Gebilde zusammenfügen.

Auf der anderen Seite sehen die schaltungstechnischen Möglichkeiten, die elektrische Kapazität auf einem Chip integriert zu messen, vielversprechend aus. Gängige Verfahren erfordern nur wenige aktive Bauelemente und sind in der Lage, Kapazitäten bis hinab in den Bereich von wenigen Attofarad zu messen (siehe ausführlich in Abschnitt 2.3 ab Seite 45). Dies ist freilich nur möglich, wenn schon zur Entwurfszeit entsprechende Vorkehrungen getroffen werden.

Obwohl die Cluster bezüglich ihrer prinzipiellen Gestalt das Kabelknäuel in Bild 1.14 zum Vorbild haben, ist der konkrete Aufbau damit noch nicht festgelegt. Es existiert eine ganze Reihe von Freiheitsgraden, neben den Leiterbahndicken und -abständen die Zahl der Ebenen und Kreuzungen, die Größe (Volumen), die strukturelle Dichte, der Grad der Zufälligkeit, usw. Alle folgenden Untersuchungen in dieser Arbeit beschränken sich auf eine bestimmte Grundform der Cluster (typisches Beispiel in Bild 3.8 auf Seite 67), da zeitliche Gründe und das Fehlen zusätzlicher Arbeitskräfte gegen eine vergleichende Analyse aller denkbaren Varianten sprachen.

Es sei schließlich noch darauf hingewiesen, dass bis zum Tag der Fertigstellung dieser Arbeit die Cluster und das zugrundeliegende Konzept völlig neuartig sind. Dies geht aus den Recherchen des Deutschen Patent- und Markenamts hervor, bei dem die vorliegende Erfindung zum Patent[7] angemeldet wurde.

Organisation dieser Arbeit

Im folgenden Kapitel „Theoretische Grundlagen“ wird zunächst auf die allgemeinen Ursachen, Arten und Auswirkungen der Prozessschwankungen bei der Chip-Herstellung eingegangen, sowie deren mathematische Modellierung (Abschnitt 2.1). Die Behandlung dieses Themas bildet die Basis für das

7. Deutsche Anmeldung unter der Nummer DE 10 2005 024 379, internationale Anmeldung unter dem Aktenzeichen PCT/DE2006/000909.

Verständnis des Herstellprozesses als Entropiequelle und die Ursachen der strukturellen Unbestimmtheit der Cluster. Die Mathematik der Kapazitätsberechnung wird in Abschnitt 2.2 erläutert, um einen Eindruck vom Rechenaufwand und damit der kapazitiven Unbestimmtheit zu vermitteln. Die theoretischen Aspekte der Kapazitätsmessung werden schließlich in Abschnitt 2.3 behandelt, um die prinzipiellen Techniken mit ihrer jeweiligen Messgüte vorzustellen.

Im anschließenden Kapitel „Implementierung“ wird als erstes gezeigt, wie die dreidimensionalen Kapazitätscluster in der Praxis erzeugt werden können (Abschnitt 3.1). Ein eigens entwickelter Random-Walk Algorithmus wird dabei präsentiert, der in Form eines Programmes in einer speziellen Sprache auf die Layout-Funktionen der Chip-Entwicklungsumgebung zugreift. Im Anschluss daran wird erläutert, wie die Messung von Clustern auf eigens entwickelten und gefertigten Testchips vonstatten ging (Abschnitt 3.2). Schließlich wird in Abschnitt 3.2 eine mögliche Auswerteelektronik vorgeschlagen, die aus dem Kapazitätsverhältnis von Clusterpaaren die Bitsequenz eines Schlüssels erzeugt.

Das Kapitel „Ergebnisse“ beginnt mit einem Vergleich der Kapazitätswerte aus den Berechnungen verschiedener rechnergestützter Werkzeuge der „Electronic Design Automation“ und bietet damit ein quantitatives Maß für die Unbestimmtheit bei der Kapazitätsberechnung (Abschnitt 4.1). Einen Vergleich dieser Werte mit den gemessenen Kapazitäten der Testchips ermöglicht Abschnitt 4.2. Dort werden die Messergebnisse von Clustern und von regulären Strukturen (z.B. Plattenkondensatoren) gegenübergestellt und die Streuungseigenschaften (besonders das „Matching“) analysiert. In Abschnitt 4.3 werden schließlich die Ergebnisse der vorgeschlagenen Auswerte- bzw. Schlüsselelektronik vorgestellt.

Den Abschluss bildet das Kapitel „Zusammenfassung und Ausblick“, in dem alle Teile dieser Arbeit zu einem Gesamtbild zusammengefügt werden und im Kontext der Aufgabenstellung bewertet wird. Bild 1.15 bietet nochmals einen Überblick über die Grobstruktur der Arbeit.

* * *

Bild 1.15. Aufbau und Organisation dieser Arbeit.

Kapitel 2
Theoretische Grundlagen

Grundlage der Entropiegewinnung zur Erzeugung geheimer Schlüssel ist die Streuung der Prozessparameter bei der Herstellung von Halbleiterchips. Daher werden in Abschnitt 2.1 theoretische Aspekte der Prozessschwankungen behandelt, beginnend mit einer Klassifikation der Ursachen und Effekte. Der Unterschied zu deterministischen Fehlern wird erläutert und ihre Entstehung erklärt. Schließlich wiederholt der Unterabschnitt „Modellierung“ die Herleitung der in der Vergangenheit von verschiedenen Autoren entwickelten theoretischen Modelle zur mathematischen Beschreibung von Randeffekten und Oxydschichtschwankungen. Hier ist vor allem das Pelgrom-Modell zu nennen, das in der Praxis breite Anwendung findet. Es wird jedoch gezeigt, dass dieses Modell nur bei einfachen Plattenkondensatoren eingesetzt werden kann und wie es hierfür bei Schaltkreissimulationen genutzt wird.

Abschnitt 2.2 widmet sich der Berechnung der elektrischen Kapazität im allgemeinen Fall, ausgehend von der physikalischen Definition des elektrischen Feldes von (ruhenden) Ladungsträgern. Zu diesem Zweck wird die Laplace-Formel hergeleitet und es wird argumentiert, dass die analytische Lösung dieser Gleichung nur bei geometrisch einfach aufgebauten Objekten möglich ist. Aus diesem Grunde werden die numerischen Verfahren zur Approximation der Kapazität von beliebig geformten Leitern vorgestellt und ein Überblick über den Stand der Forschung anhand von Literaturverweisen geboten. Schließlich werden die heute verfügbaren
Software-Werkzeuge der EDA-Sparte zur Berechnung (Extraktion) der Kapazität aufgelistet und ihre Besonderheiten erklärt.

Am Schluss des Kapitels steht in Abschnitt 2.3 die Behandlung der Kapazitätsmessung durch experimentelle Methoden. Die vielfach erprobte Technik der Ladungspumpen wird anhand ihres Prinzips erläutert und Fragen wie Auflösung und schaltungstechnische Verbesserungen werden angegangen. In diesem Zusammenhang wird durch mathematische Herleitung gezeigt, dass die nicht-idealen Schalter der Ladungspumpe für die Messfehler verantwortlich sind. Abschließend wird das auf einer Modifikation klassischer Ladungspumpen basierende und zur Zeit genaueste Verfahren zur Kapazitätsmessung vorgestellt.

* * *

2.1 Prozessstreuung

2.1.1 Prozessstreuung und Mismatch

Allgemeines

Der Herstellungsprozess von Chips in der Mikroelektronik setzt sich aus einer Reihe hochkomplizierter und aufwendiger Einzelschritte zusammen, die eine große Zahl an Einflussfaktoren und Parameter aufweisen. Die heute herstellbaren Strukturen liegen größenordnungsmäßig im Mikrometerbereich und reichen bei den modernsten Chipfabriken bis unter einige Nanometer (z.B. Gate-Oxydschichtdicke).

Aus diesen Gründen sind die Anlagen und Maschinen einer solchen Fabrik bzw. der von diesen ausgeführte Herstellungsprozess mit unvermeidbaren Ungenauigkeiten und Schwankungen behaftet, die sich in einer Streuung der elektrischen und geometrischen Parameter der hergestellten Produkte äußern. Die Auswirkungen dieser Schwankungen auf die elektronischen Schaltungen eines Chips sind insbesondere im analogen Schaltungsentwurf von Bedeutung, da sie einen direkten Einfluss auf die Leitungsdaten und das Funktionieren der Gesamtelektronik haben. Die Wirkungsweise und das Ausmaß der Prozessstreuung sind daher Gegenstand einer Vielzahl von wissenschaftlichen Untersuchung, sowohl theoretischer Art, als auch empirisch-messtechnischer Art.

Im Rahmen dieser Arbeit soll im Folgenden die Prozessstreuung unter dem Aspekt der Schlüsselerzeugung und der dadurch entstehenden Fragestellungen untersucht werden. Speziell die Auswirkung der Schwankungen auf die Kapazität der vorgestellten 3D-Cluster und die Stabilität der aus ihnen abgeleitete Schlüssel sind von Interesse.

Begriffsbestimmung

MISMATCH. Dieser Begriff bezeichnet die zeitunabhängige Differenz ΔP der physikalischen Parameter identisch angelegter Bauteile bzw. funktionsgebender Strukturen bei der Realisierung durch den Herstellungsprozess. In Pelgrom 1989 wurde definiert:

„«Mismatch» ist der Vorgang, der zeitunabhängige, zufällige Variationen in physikalischen Größen bei identisch entworfenen Bauteilen bewirkt."

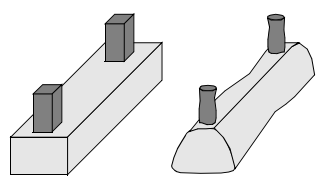


Bild 2.1. Die unvermeidbaren Ungenauigkeiten und Schwankungen des Prozesses führen zu Abweichungen der hergestellten Strukturen von der Idealform beim Entwurf. Identisch angelegte Bauteile weisen dadurch physikalische Unterschiede auf, die sich durch abweichende elektrische Parameter bemerkbar machen können (Mismatch).

In Bild 2.1 ist auf der linken Seite ein Leitungsstück in der Entwurfsansicht (Layout) zu sehen, das einem Bauteil entstammen könnte. Eine Vielzahl von physikalischen Einflussfaktoren führt bei der Herstellung zu einerseits systematischen, das heißt deterministischen Regeln folgenden Abweichung, andererseits auch zu zufallsbedingten Variationen. Die rechte Seite zeigt das Ergebnis beider Wirkungsarten. Identisch entworfene Bauteile bzw. identische Strukturen weisen indessen die gleiche deterministische Abweichung auf, so dass nur der zufallsbedingte Anteil dem „Mismatch" zugerechnet werden kann. Neben den in Bild 2.1 gezeigten geometrischen Abweichungen und Variationen, die sich durch Deformationen der Struktur bemerkbar machen, gibt

es physikalisch-elektrische Parameter (z.B. Dotierungsstärke und Materialbeschaffenheit), die Abweichungen vom angenommen Idealwert aufweisen. Letztere wirken sich ebenfalls auf den Mismatch aus.

Da der Absolutwert des Parameters P_i *eines* Bauteils eine sehr viel größere Variation erfährt, als der durch ΔP gegebene *relative Fehler* bzw. Mismatch von *Paaren* von Bauteilen, werden bei analoger Elektronik in der Regel relative elektrische Parameter schaltungstechnisch genutzt, beispielsweise das Kapazitätsverhältnis von Kondensatoren oder das Stromstärkeverhältnis eines Stromspiegels. Der Mismatch ist geringer als der Absolutwertfehler, da globale Variationen durch sogenannte Matching-Techniken wie zum Beispiel common-centroid Anordnungen reduziert werden können (siehe hierzu die Abschnitte „Lokale und globale Variationen“).

PROZESSSTREUUNG. Darunter soll im Folgenden die Gesamtheit der *zufallsbedingten* (stochastischen) Streuungen jener physikalischen Parameter $q_1, q_2, \ldots, q_N$ verstanden werden, die einen Einfluss auf die elektrischen Eigenschaften $P_1, P_2, \ldots, P_M$ funktionsgebender Strukturen oder Bauteile haben. Hierunter fallen sowohl mittelbare elektrische Größen wie die Dotierungsstärke, als auch geometrische Parameter, die sich über einen bestimmten Funktionszusammenhang $P_i = f(q_1, q_2, \ldots, q_N)$ in elektrischen Parametern wie zum Beispiel Kapazität und Widerstand bemerkbar machen. Die Bezeichnung „Parameter“ wird im Folgenden also sowohl für die physikalischen Prozessparameter $q_1, q_2, \ldots, q_N$ verwendet, als auch für die elektrischen Eigenschaften $P_1, P_2, \ldots, P_M$ eines Bauteils.

Ebene	Paar entstammt	Wirkung
Bauteil	Punkten	lokal
Chip	Bauteilen	global
Wafer	Chips	global
Los	Wafern	global
Prozess	Losen	global

Tabelle 2.1. Der Mismatch wirkt auf die elektrischen Parameter von Bauteilen desselben Chips, Wafers, Loses oder Prozesses (Ebene). Seine Wirkung kann lokal begrenzt sein oder sich als Parametergefälle über weite Bereiche bemerkbar machen.

Klassifikation

Die Streuung der Prozessparameter und der damit einhergehende Mismatch geht, wie bereits erwähnt, auf eine Vielzahl von Einflussfaktoren zurück, die gewissen Fehlerquellen zugeordneten werden können. Es existieren deterministische Einflüsse und rein zufallsbasierte, global auftretende Parametergefälle und lokal wirkende Fluktuationen, miteinander korrelierte Parameter und unabhängige, und so weiter. Aus diesem Grund soll eine Klasseneinteilung und -sortierung vorgenommen werden:

HIERARCHIE. Prozesstreuungen existieren auf verschiedenen Hierarchieebenen. So kann der Mismatch Bauteile betreffen, die sich auf demselben Chip befinden oder aber auf einer übergeordneten Ebene wie dem Wafer, einem Stapel an Wafern (Los) oder dem gesamten Prozess. Die Streuung bezieht sich auf Paare (P_i,P_j) $(i \neq j)$ von Bauteilparametern, die aus zwei verschiedenen Bauteilen, Chips, Wafern oder Waferstapeln stammen (Tabelle 2.1, zweite Spalte). Aus ihnen wird die Differenz $\Delta P = P_i - P_j$ berechnet, die den Mismatch $\Delta P_{i,j}$ angibt.

Auf der untersten Ebene der einzelnen Bauteile wird die Paarbildung durch Punkte innerhalb eines Bauteils oder einer Struktur vorgenommen. Da der Parameter P meist nur für ganze Bauteile Sinn macht, wird die Differenz der Prozessparameterpaare (q_i,q_j) betrachtet. Dies stellt einen wichtigen Startpunkt für die theoretische Herleitung des Einflusses lokal eng begrenzter Variationen auf die Bauteilparameter dar (siehe die Abschnitte „Randeffekte und Oxydschicht-Schwankungen“ und „Das Kondensatormodell für den Mismatch“ auf Seite 31 ff.).

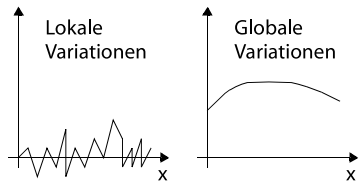


Bild 2.2. Die Prozessstreuungen können lokal, d.h. über kurze Distanz (x-Achse) wirken oder einen globalen Trend aufweisen, der sich nur über große Entfernungen bemerkbar macht.

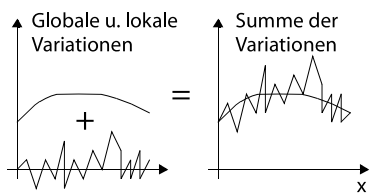


Bild 2.3. Der Mismatch geht auf die Summe lokaler und globaler Variationen zurück. Der globale Teil kann durch sog. Matching-Techniken reduziert werden, der lokale Teil stellt die unterste erreichbare Genauigkeitsgrenze dar.

LOKALE UND GLOBALE VARIATIONEN. In Shyu 1984 wird zum ersten Mal zwischen lokalen und globalen Variationen unterschieden, das heißt solche Parameterstreuungen, die innerhalb eines Bauteils vorkommen und solchen, die zwischen weit entfernten Bauteilen (zwischen Bauteilen auf verschiedenen Chips, Wafern oder Waferstapeln) auftreten.

In Bild 2.2 ist dargestellt, wie sich ein beliebiger Prozessparameter q_i in Abhängigkeit vom Ort in den beiden Fällen verhält. Links variiert der Parameter in zufälliger Weise und über kurze Distanzen. Würde die x-Achse als Zeitachse aufgefasst, so entspräche die Funktion weißem Rauschen. Im Frequenzbereich wären alle Frequenzen von Null bis (theoretisch) Unendlich vertreten. Die Autokorrelationsfunktion fällt bereits für sehr kleine Distanzen stark ab („short distance correlation“), typischerweise ist sie sehr viel kleiner als die Ausmaße der Bauteile. Diese Aussage bedeutet, dass – bezogen auf lokale Variationen – keine Beziehung zwischen dem Abstand der Bauteile und dem Mismatch besteht (Pelgrom 1989). Durch das Nahebringen der Bauteile kann der von lokalen Variationen herrührende Mismatch also nicht verbessert werden. Die Ursache für diese Variationen liegen in feingranularen Fluktuationen der Verteilung von implantierten oder diffundierten Ionen, der Ladungsträgermobilität, der Permittivität und so weiter. Insgesamt wirken sich die einzelnen Parameter auf den Mismatch ΔP mittelwertneutral aus, das heißt sie bewirken keine systematische Verschiebung in positiver oder negativer Richtung.

Globale Variationen ergeben sich dagegen über große Entfernungen auf einem Chip oder von Chip zu Chip (Wafer zu Wafer, Lot zu Lot). Es handelt sich nicht mehr um Rauschen mit hochfrequentem Anteil und kurzer Korrelationsdistanz, sondern um trendmäßige Verschiebungen der Prozessparameter, die sich nur über weitere Entfernungen bemerkbar machen (Bild 2.2, rechte Seite). Entsprechend weisen globale Variationen eine „long distance correlation“ auf, so dass nahe beisammen liegende Bauteile einen geringeren diesbezüglichen Mismatch aufweisen, als weit entfernte. Die Ursache für globale Variationen liegt beispielsweise in der zeitlichen Änderungen der Konzentration von Gasen und Ätzflüssigkeiten (Drift) oder in räumlichen Temperatur- und Konzentrationsunterschieden (Gradienten) bei der Herstellung.

UNABHÄNGIGE UND ABHÄNGIGE VARIATIONEN. Voneinander unabhängige Variationen werden durch grundsätzlich verschiedene physikalische Vorgänge und Verfahrenschritte hervorgerufen. Die Prozessparameter sind bezüglich des Anteils der unabhängigen Variationen nicht miteinander korreliert. Dies bedeutet, dass man aus der Kenntnis eines Parameters nicht auf den Wert eines anderen Parameters schließen kann, falls beide nur unabhängigen Variationen unterworfen sind. Dies ist beispielsweise bei den Prozessparametern Dotierungsstärke und Oxydschichtdicke der Fall. Eine dünne Oxydschicht bedeutet nicht, dass die Dotierung stark oder schwach sein muss, beide Größen stehen in keinem funktionalen Zusammenhang.

Im Gegensatz dazu gibt es Prozessparameter, die mehr oder weniger stark miteinander korreliert sind. Dies ist hauptsächlich dann der Fall, wenn Parameter durch dasselbe herstellungstechnische Verfahren beeinflusst werden, beispielsweise das Aufbringen von Metallschichten beim sogenannten „Sputtering“. Zu einem gewissen Anteil kann in diesem Fall die Dicke aller Lagen vom typischen Wert in die eine oder andere Richtung abweichen, wenn etwa die Anlagen für das Sputtering schlecht eingestellt sind oder Ver-

schleißerscheinungen zeigen. In der Praxis ist der funktionale Zusammenhang der Parameter unbekannt, nur mit Hilfe stochastischer Modelle und von Erfahrungswerten lässt sich eine Beziehung herstellen.

Stochastische und deterministische Variationen. Stochastische Variationen ergeben sich aufgrund *zufälliger* Schwankungen und Fluktuationen. Sie entstehen durch eine Vielzahl an Fehlerquellen während des Herstellungsprozesses, einige wohlbekannte von ihnen sind in Tabelle 2.2 aufgelistet. Neben den rein zufallsbedingten Variationen, die auf physikalische Rauschprozesse oder chaotische Wechselwirkungen zurückzuführen sind, gibt es Variationen, die zwar gewissen festen Regeln unterworfen sind, diese jedoch unbekannt sind. Durch die Unkenntnis der Gesetzmäßigkeit der Schwankungen scheinen diese zufälliger Natur zu sein oder werden bewusst als zufällig angenommen, um sie einer statistischen Analyse zu unterziehen. Ergebnis einer solchen Analyse sind die sogenannten statistisch *kontrollierbaren* Schwankungen. Dies sind jene Variationen, die vorherbestimmbaren statistischen Wahrscheinlichkeitsverteilungen folgen und so in den Schaltungsentwurf einkalkuliert werden können.

Deterministische Variationen auf der anderen Seite folgen einer festen, in der Regel bekannten Gesetzmäßigkeit. Dadurch können sie bereits zur Entwurfszeit einer Schaltung berücksichtigt bzw. eliminiert werden. Sie sind Gegenstand zahlreicher wissenschaftlicher Untersuchungen und werden verstärkt in diverse Entwurfswerkzeuge der Electronic Design Automation (EDA) einbezogen. Dazu gehört insbesondere die Korrektur von Abbildungsfehlern durch Beugungseffekte bei der lithografischen Übertragung sehr kleiner Maskenstrukturen auf den Wafer („Optical Proximity Correction“, OPC und „Phase Shift Masks“, PSM). Bisher unberücksichtig von den EDA-Werkzeugen bleiben hingegen struktur- bzw. musterabhängige Unterschiede in den Oxydschichtdicken trotz eingehender wissenschaftlicher Behandlung. Im Abschnitt „Deterministische Fehler“ auf Seite 27 werden diese genauer behandelt.

Ebene	Fehlerquellen
Bauteil 	Punktdefekte Feingranulare Fluktuationen Linsenfehler Maskenfehler
Chip 	Gaskonzentrationsgrad. Temperaturgradienten Umgebungsabhängige F. Orientierungsabhängige F. Maskenfehler
Wafer 	Gaskonzentrationsgrad. Temperaturgradienten Ätzfehler Chemikaliensaturierung Abscheidungsfehler Maskenausrichtungsfehler
Los 	Temperaturdrift Gaskonzentrationsdrift
Prozess 	Materialdrift Gerätedrift

Tabelle 2.2. Auf allen Hierarchieebenen existieren Fehlerquellen, die den Mismatch bewirken. Liegen zwei Bauteile auf einer niedrigen Ebene sehr nahe beieinander, wird der Mismatch reduziert, da alle übergeordneten Fehlerquellen beide gleich stark beeinflussen.

Der Zentrale Grenzwertsatz

Eine grundlegende Eigenschaft einiger Prozessparameter besteht in der *normalverteilten* Streuung. Beispielsweise häufen sich die Fläche und der Umfang eines Plattenkondensators um einen gewissen Wert, einige wenige dagegen liegen mit abnehmender Wahrscheinlichkeit weiter davon entfernt. Ein Histogramm dieser Verteilung entspricht also annähernd einer Gauß'schen Glockenkurve.

Diese Eigenschaft liegt in der Tatsache begründet, dass sich solche Parameter aus vielen sekundären Prozessparametern zusammensetzen, die wiederum von einer noch viel größeren Zahl an physikalischen Effekten bestimmt werden. Insgesamt ist die Menge der Einflussfaktoren auf die Parameter so groß, dass der Zentrale Grenzwertsatz gilt (Box 2.1). Dieser besagt, dass eine Größe, die aus dem kumulativen Effekt vieler unabhängiger Variablen hervorgeht, eine Normalverteilung annähert, egal wie die Verteilung der Variablen aussieht (Barlow 1989). Diese Unabhängigkeit ist immer dann gewährleistet, wenn die den Variablen zugrundeliegenden physikalischen Prozesse grundverschieden sind.

Box 2.1 Zentraler Grenzwertsatz

Sei X die Summe von N unabhängigen Variablen x_i mit $i = 1,2,\ldots,N$ jeweils aus einer beliebigen Verteilung mit Mittelwert μ_i und Varianz V_i. Dann gilt für die Verteilung von X

1. Der Erwartungswert ist

$$E(X) = \sum \mu_i$$

2. Die Varianz ist

$$V(X) = \sum V_i$$

3. Die Verteilung nähert sich einer Normalverteilung an für $N \to \infty$.

(Beweis in Barlow 1989, Anhang 2).

Häufig wird in der Literatur das Kriterium der identischen Verteilung der Variablen als Voraussetzung für die Anwendbarkeit des Zentralen Grenzwertsatzes genannt. Es existieren jedoch Verallgemeinerungen des Theorems, die dies nicht fordern, sondern nur, dass keine der Variablen zu großen Einfluss auf das Resultat hat (z.B. Lyapunov-Bedingung).

Diese Bedingung wird durch diverse Matching-Regeln hergestellt, indem dafür gesorgt wird, dass die dominierenden Einflussfaktoren auf Paare von Bauteilen in möglichst gleichem Maße wirken und sich so relativ gesehen gegenseitig eliminieren. Beispielsweise führt das Nebeneinanderplatzieren von zwei Bauteilen dazu, dass globale Drifts und Gradienten auf diese gleich stark wirken und sich bei subtraktiver Verwendung der Bauteile aufheben.

Diese globalen Prozessschwankungen sind im Gegensatz dazu häufig nicht normalverteilt, da sie sich nicht aus einer Vielzahl anderer physikalischer Parameter zusammensetzen oder diese nicht unabhängig voneinander sind. So wird der globale Verlauf der Isolationsschichtdicke zwischen den Metallisierungsbahnen von wenigen Parametern dominiert, hauptsächlich von der Planarität von Wafer und Polierscheibe beim chemisch-mechanischen Polieren des Chips („chemical-mechanical polishing“, CMP).

Die Verteilung der globalen Prozessparameter unter- und überschreitet (von Ausnahmen abgesehen) gewisse Grenzen nicht (Fail-/Pass-Parameter). Diese Grenzen werden als die ungünstigsten („worst-case“), bzw. günstigsten („best-case“) Parameter bezeichnet, obwohl diese qualitative Zuordnung nicht immer eindeutig ist. Erreicht wird dies durch die ständige Kontrolle des Prozesses während der Herstellung („process control monitoring“, PCM) und das Aussortieren der Chips oder Wafer, die außerhalb der durch das best-case/worst-case Parameterpaar definierten Spezifikation liegen[8]. Dadurch ergeben sich für die elektrischen Parameter ebenfalls gewisse Ober- und Untergrenzen P^{wc} und P^{bc}.

8. Eine andere Methode besteht darin, alle Chips bzw. Wafer zu akzeptieren. Die best-/worst-case Parameter stellen dann keine Fail-/Pass-Parameter mehr dar, sondern haben rein informatorischen Charakter.

2.1.2 Deterministische Fehler

Streuen die Prozessparameter in deterministischer, also in reproduzierbarer und gewissen Mustern folgender Weise, so führt das zu Fehlern, die sich zwar als Prozessschwankungen schaltungstechnisch auswirken, jedoch während des Entwurfs, der Herstellung oder sogar nach der Fertigung herausrechnen lassen. Voraussetzung hierfür ist, die Gesetzmäßigkeiten zu kennen, denen diese Fehler folgen. Im Folgenden werden einige dieser Fehlerquellen vorgestellt ohne jedoch Anspruch auf Vollständigkeit zu erheben.

Beugungsfehler

Hierbei handelt es sich um eine Fehlerquelle, die bei Prozessen der neuesten Generation auf Bauteilebene wirkt, genauer gesagt auf die Geometrie von Leiterbahnen und Durchkontaktierungen. Der erste Effekt, der von den Prozessingenieuren erkannt und softwaretechnisch korrigiert werden konnte, ist die Variation der Leiterbahndicke in Abhängigkeit vom Abstand zu den benachbarten Leitungen. In Wolf 2004 ist ein Beispiel genannt: 0,35 Mikrometer breite Leiterbahnen, die weit voneinander entfernt liegen, sind in diesem Fall 90 Nanometer breiter, als nahe beieinander liegende Leitungen. Da es sich um einen Effekt handelt, der abhängig ist vom Abstand der Strukturen, wird die Korrekturmethode „Optical Proximity Correction“ genannt: Die Breite der Leitungen wird bereits auf der Maske vor der Belichtung des Wafers erhöht, und zwar umso mehr, je näher diese bei benachbarten Strukturen liegen.

Ursache für diesen Fehler ist die Verzerrung des Maskenbildes bei der Belichtung aufgrund von Beugungsfehlern an den Strukturen der Maske. Diese sind in modernen Prozessen so klein, dass sich an Kanten und in Ecken quantenmechanische Effekte bemerkbar machen und die Auflösungsgrenze der Linsen überschritten wird. Heutzutage wirken sich diese Effekte nicht nur auf die Leiterbahnbreite aus, sondern generell auf die Kantenschärfe von Leitungen und Vias, so dass Ecken abgerundet werden wie in Bild 2.1 zu sehen.

Sonstige Fehler

Eine Reihe weiterer Faktoren können die geometrischen Strukturen und elektrischen Eigenschaften der Bauteile eines Chips in deterministischer Weise beeinflussen. Einige von ihnen sind den Prozessingenieuren bereits bekannt, darunter z.B. strukturabhängige Über-, Unterätzungen und Metall- und Oxydauswaschungen („erosion“ und „dishing“ während des CMP). Andere Effekte (z.B. Unebenheiten des Fotolacks, Linsen- und Maskenfehler) mögen noch unbekannt oder schwer beherrschbar sein, entfalten aber dennoch ihre Wirkung in reproduzierbarer Weise.

Zu diesem Themenkomplex gibt es eine Reihe von Forschungsarbeiten z.B. Mehrotra 2001, Park 2002 und Gbondo-Tugbawa 2002. Die Publikation von Boning et al 1998 bietet eine gute Einführung, hauptsächlich in Hinblick auf die Modellierung und Korrektur der Effekte. Auf Beugungseffekte und andere derartige deterministische Einflüsse zurückgehende Fehler sind also keine „echten“ Prozessschwankungen, sie sind nicht zufällig, sondern (zumindest theoretisch) a-priori bekannt und berechenbar.

2.1.3 Ausbeute (Yield)

Definition

ALLGEMEINE AUSBEUTE. Generell ist die Ausbeute („Yield“) definiert als das Verhältnis aus der Anzahl N akzeptierter Chips, im Folgenden „Nutzen“ genannt („yield body“), und der Gesamtzahl S an Chips:

$$Y = \frac{N}{S} \tag{2.1}$$

Welche Chips nun akzeptiert werden und dem Nutzen zugerechnet werden, kann von verschiedenen Kriterien abhängen, je nachdem, ob nur defekte Chips aussortiert werden (funktionelle Ausbeute), oder auch solche, die außerhalb von vorgegebenen Fail-/Pass-Parameterbereichen liegen (parametrische Ausbeute).

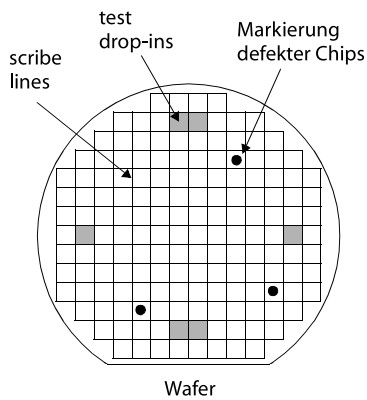


Bild 2.4. Der Wafer wird durch Messungen an Teststrukturen in den Zwischenräumen angrenzender Chips („scribe lines“) getestet und dort, wo die Strukturen reguläre Chips ersetzen. Fehlerhafte Wafer werden nicht weiterverarbeitet. Bei Einzeltests werden nur die defekten Chips durch Farbpunkte markiert und aussortiert.

FUNKTIONELLE AUSBEUTE (FUNCTIONAL YIELD). Hierunter versteht man gemeinhin die Anzahl Chips, die spezielle Funktionstests bestanden haben, pro Gesamtzahl an Chips. Der Nutzen ist also gegeben aus der Differenz der Zahl aller Chips und der Zahl nicht funktionierender Chips F:

$$N = S - F \quad \Rightarrow \quad Y_F = \frac{S - F}{S} \tag{2.2}$$

Die Funktionstests werden typischerweise in mehreren Schritten durchgeführt: Vor dem Zersägen des Wafers und dem Einsetzen in die Gehäuse, sowie in mehreren Phasen danach, wenn die Chips über alle Anschlussleitungen verfügen und einfacher elektrisch kontaktiert werden können. Im ersten Schritt, Wafersortierung genannt, werden die einzelnen Chips auf dem Wafer über eine spezielle Nadelkarte („probe card“) kontaktiert und der Reihe nach einfachen Gleichstrommessungen unterzogen. Dies geschieht im Fall der Massenfertigung automatisiert und aufgrund der hohen mechanischen Anforderungen an die Positionierungsgenauigkeit und wegen messtechnischen Erfordernissen auf einer Probe-Station oder Wafer-Prober (siehe Abschnitt “Messaufbau” auf Seite 68 ff.). Für jeden getesteten Chip wird das Resultat in einer Karte des Wafers („wafer map“) eingetragen oder durch Tintenkleckse farblich markiert, um die fehlerhaften Chips aussortieren zu können (siehe Bild 2.4).

Nach dem Einsetzen der Chips in die Gehäuse und den anschließenden Assemblierungsphasen können eine Reihe weiterer Funktionstests erfolgen, beispielsweise dynamische Hochgeschwindigkeitstests oder intensive Belastungstests („burn-in tests“). Der Nutzen reduziert sich dann um die Zahl der Chips, die diese Tests nicht bestehen, je nachdem ob sie in der Definition des Yield berücksichtigt werden sollen oder nicht.

PARAMETRISCHE AUSBEUTE (PARAMETRIC YIELD). Der Nutzen kann sich weiter reduzieren, wenn noch mehr Chips aus der Menge der akzeptierten Chips aussortiert werden. Dies ist beispielsweise dann der Fall sein, wenn bestimmte Prozessparameter außerhalb des durch ein best-case/worst-case Parameterpaar gegebenen Bereichs liegen. Dadurch überschreiten auch einige der elektrischen Bauteilparameter die best-case/worst-case Werte. Im Gegensatz zum Functional Yield, bei dem nur zwischen funktionstüchtigen und defekten Chips unterschieden wird, geht es hier nur um die Frage, ob gewisse Parameter eine vorgegebene Spezifikation erfüllen oder nicht. Ist dies bei einem Chip

nicht der Fall, kann er trotzdem funktionieren und die funktionale Ausbeute verbessern, wenn auch seine Leistung in der Regel gering ist. Sei also P die Zahl der Chips, die außerhalb der Parameterspezifikation liegen. Dann kann die parametrische Ausbeute Y_P definiert werden als:

$$N = S - F - P \quad \Rightarrow \quad Y_P = \frac{S - F - P}{S} \tag{2.3}$$

Gleichung 2.3 erweist sich in der Praxis allerdings häufig als wenig nützlich, da nicht nur jene Chips aussortiert werden, die den spezifizierten Prozessparameterbereich überschreiten, sondern *alle* Chips des betroffenen Wafers. Der Grund liegt darin, dass P nicht bekannt ist, bzw. welche Chips tatsächlich außerhalb des Bereichs liegen, da der parametrische Test immer für einen ganzen Wafer durchgeführt wird.

Realisiert wird dies durch besondere Teststrukturen, die in den Zwischenräumen angrenzender Chips („scribe lines“) angeordnet werden oder an Stellen, die speziell für diesen Zweck ausgespart wurden (siehe Bild 2.5). An diesen Strukturen werden über geeignete Schaltungen Messungen vorgenommen, aus denen die Werte der zu überprüfenden Prozessparameter zurückgerechnet werden können. Ähnlich wie beim Funktionstest geschieht dies durch Einsatz einer Probe-Station und hochpräziser Messinstrumente (i.d.R. „source monitoring units“, SMU).

Als Grund für die Entsorgung der kompletten Wafer sind die Kosten für jeden weiteren Prozessschritt zu nennen, die eingespart werden können, wenn die Messungen bereits vor Vollendung des gesamten Herstellungsprozesses durchgeführt werden. Dies kann zum Beispiel vor dem Aufbringen weiterer Metallisierungslagen geschehen, wenn die für die Messschaltungen benötigten Lagen bereits vorhanden sind.

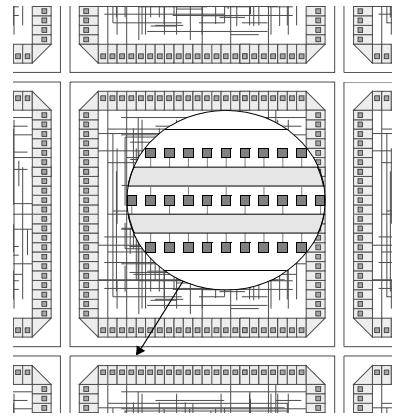


Bild 2.5. In den scribe lines werden Teststrukturen angeordnet, durch die über geeignete Schaltungstechniken die wichtigsten Prozessparameter errechnet werden können.

Parameter-, Leistungs- und Funktionsbereich

Ziel eines jeden Entwicklers im Schaltungsentwurf ist es, leistungsfähige Chips zu produzieren. Die Leistungsspezifikation erfolgt häufig durch Marktvorgaben oder dem Einsatzzweck entsprechend und stellt damit ein Entwurfsziel dar, das es durch die Kunst des Schaltungsentwurfs zu erreichen gilt. Diese Leistungsvorgaben werden auch beim besten Schaltungsdesign nur dann erreicht, wenn die Prozessparameter bei der Fertigung des Chips in der Fabrik die beim Entwurf angenommen Parameterbereiche nicht überschreitet. Mit anderen Worten: Die Leistungsspezifikation, festgelegt durch Marktvorgaben, führt zu einem Bereich akzeptierbarer Prozessparameter im Parameterbereich („parameter domain“), den Nutzen („yield body“). Dieser Zusammenhang ist in Bild 2.6 links dargestellt. Der Nutzen ist dabei nicht immer deckungsgleich mit dem Bereich der Prozessstreuungen („process spread“). Es ist die Aufgabe des Entwicklers, beide Flächen zur Deckung zu bringen, um dadurch die Ausbeute zu erhöhen und die Leistungsvorgaben einzuhalten.

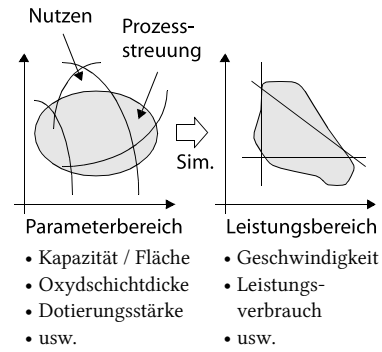


Bild 2.6. Durch hochkomplexe Simulationen (siehe Bild 2.7) wird der Übergang vom Parameterbereich zum Leistungsbereich vollzogen. Der Nutzen ist jedoch meist im Leistungsbereich spezifiziert und der Weg zurück zum Parameterbereich sehr schwierig.

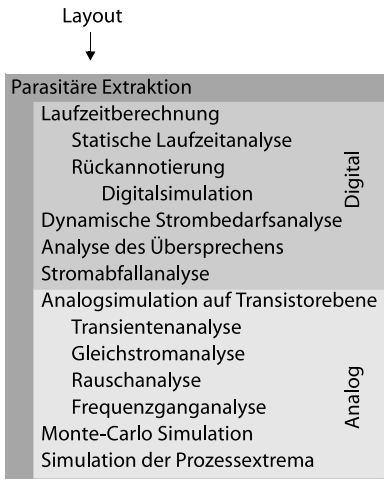


Bild 2.7. Es existiert eine große Zahl an komplexen Verfahren zur Simulation und Analyse der Leistungsfähigkeit und Funktion der Schaltungen. Ausgangspunkt ist immer die Schaltungsrückerkennung (Extraktion) aus den geometrischen Entwurfsdaten (Layout), bei der eine Netzliste mit parasitären Bauelementen erzeugt wird. Diese wirken sich leistungsmindernd aus und können die Funktion beeinträchtigen.

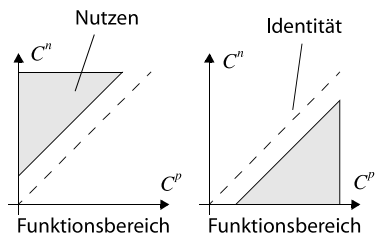


Bild 2.8. Beim Einsatz der Kapazitätscluster im Rahmen dieser Arbeit ist der Nutzen im Funktionsbereich gleich den grauen Flächen. Der Abstand zur Winkelhalbierenden ergibt sich aus der Messgenauigkeit der Elektronik.

Der Entwickler wird dabei durch rechnerbasierte Simulationen und automatische Verifikationswerkzeuge unterstützt, durch sie werden aus den Prozessparametern Leistungsdaten errechnet und der Übergang zum Leistungsbereich („performance domain“, Bild 2.6 rechts) vollzogen. Dieser Vorgang ist heute sehr kompliziert und nur noch mithilfe ausgefeilter, aufwendiger Algorithmen und komplexer Softwarepakete zu bewerkstelligen. Er ist in Bild 2.7 exemplarisch dargestellt.

Dem Leistungsbereich in Bild 2.6 sehr ähnlich ist der Funktionsbereich. In ihm ist der Teilbereich der funktionstüchtigen Chips eingetragen, er definiert den Nutzen. Dieser sollte mit dem Bereich der Prozessstreuung möglichst deckungsgleich sein, da das Verhältnis aus beiden gerade der funktionalen Ausbeute Y_F entspricht. Auch hier ist es Aufgabe des Entwicklers, die Elektronik eines Chips so auszulegen, dass die beiden Flächen zur Deckung kommen, und um so den Yield zu erhöhen. Während indessen der Graph im Funktionsbereich bei rein digitalen Schaltungen in der Regel zwei- oder dreidimensional ist[9], kann die Anzahl der Dimensionen bei analogen Chips je nach Anwendungsbereich und Einsatzzweck durch weitere Kriterien höher sein, so dass sich der Funktionsbereich nicht mehr grafisch darstellen lässt.

Die im Rahmen dieser Arbeit behandelten Kapazitätscluster als „Schlüssel“-Element führen zu dem in Bild 2.8 dargestellten Funktionsbereich, falls das kapazitive Größen*verhältnis* von Clusterpaaren schaltungstechnisch genutzt werden soll. Prinzipiell sind auch andere Formen der Auswertung möglich und sinnvoll, konnten jedoch in den Untersuchungen der vorliegenden Arbeit mangels Zeit nicht berücksichtigt werden. Der *relative* Ansatz wurde verfolgt, da er entsprechend den etablierten Matching-Techniken gute Chancen versprach, systematische Effekte und den Einfluss von Schwankungen der Betriebsparameter zu unterdrücken.

Da in dieser Variante das Vorzeichen der Kapazitätsdifferenz jeweils zweier Cluster C_b^p und C_b^n den binären Wert der Schlüsselbits b bestimmt, entsprechen die Achsen des Graphen gerade C_b^p und C_b^n. Die Punkte in der Ebene repräsentieren dann jeweils die Clusterpaare aller Bits und aller Chips. Der Übersichtlichkeit halber kann der Graph jedoch auf die einzelnen Bits b des Schlüssels aufgeteilt werden, so dass bei einer Schlüssellänge von M Bits M Graphen entstehen (oder entsprechend ein $M+1$-dimensionaler Einzelgraph). Der Abstand des Nutzens zur Winkelhalbierenden ist durch die Trennschärfe des Auswerteelektronik gegeben. Punkte auf oder nahe bei der Identitätslinie führen zu instabilen Bits, da die Elektronik die Cluster größenmäßig nicht unterscheiden kann und zwischen zwei Entscheidungen hin- und herschwankt.

9. Entscheidendes Kriterium für das Funktionieren ist der sog. „hold-slack“, der von parasitären Leitungskapazitäten und -widerständen abhängt und positiv sein muss. Leistungsdaten wie die Geschwindigkeit stehen nicht unmittelbar mit ihm in Zusammenhang.

Statt der Darstellungsweise über die Kapazitäts*differenz* in Bild 2.8 ist es auch möglich, das Kapazitäts*verhältnis* zu betrachten und im resultierenden Graphen den Nutzen zu bestimmen. Im Abschnitt „Statistische Analyse“ auf Seite 94 wird dies für die in dieser Arbeit entwickelte Auswertelektronik durchgeführt, im Abschnitt 4.3 basierend auf Messergebnissen von Testchips.

2.1.4 Modellierung

Die mathematische Modellierung der Prozessschwankungen und deren Auswirkungen auf die Bauteilparameter wurde erst in den achtziger Jahren des vergangenen Jahrhunderts in der Literatur behandelt, beginnend mit Untersuchungen an integrierten Plattenkondensatoren. In McCreary 1981 wurden die Messergebnisse von 32.000 Kondensatorbänken aus 16 Wafergruppen und fünf verschiedenen Technologien analysiert, die theoretische Behandlung folgte wenig später in Shyu et al. 1982. Erst in Pelgrom et al. 1989 wurde ein allgemeingültiges Konzept entwickelt, das auch heute noch zur Modellierung des Matchings aller Arten von Bauteilen dient.

Randeffekte und Oxydschicht-Schwankungen

In Shyu et al. 1982 wurde anhand wahrscheinlichkeitstheoretischer Überlegungen untersucht, welchem grundsätzlichen Trend die Kapazitätsschwankungen von Plattenkondensatoren folgen. Der Ausgangspunkt ähnelt dem später entwickelten Modell von Pelgrom (siehe Abschnitt „Das Kondensatormodell für den Mismatch“), die Durchführung erfolgt jedoch nicht im Frequenzbereich, sondern über die (Korrelations-)Distanz (entspricht der Zeit in anderen Fällen). Die mathematische Herleitung der Matching-Eigenschaften erfolgte in der Publikation in sehr knapper Weise. Aus diesem Grund wird der erste Teil des Rechenwegs in Box 2.2 auf Seite 32 nochmals ausführlicher durchgeführt.

Fehler durch Randeffekte. Das Ergebnis sei an dieser Stelle nochmals zusammengefasst: Der absolute Kapazitätsfehler aufgrund von zufälligen Schwankungen des Randes eines Plattenkondensators ist proportional zu $C^{1/4}$, der relative Fehler zu $C^{-3/4}$ (siehe Bild 2.9).

$$\Delta C \sim C^{1/4} \qquad \frac{\Delta C}{C} \sim \frac{1}{C^{3/4}} \tag{2.9}$$

Fehler durch Schwankungen der Oxydschicht. Im zweiten Teil der Publikation wurde untersucht, wie Variationen der Oxydschichtdicke aufgrund der unregelmäßigen Kornstruktur und Schwankungen der Dielektrizitätskonstante Einfluss auf die Kapazität haben. Die Vorgehensweise ist analog zum vorherigen Fall, es wurde wieder angenommen, dass die Dicke Δt und Permittivität $\Delta\varepsilon$ stochastische Prozesse mit Mittelwert Null darstellen. Die Analyse besagt, dass der absolute Kapazitätsfehler proportional zu $\sqrt{C}$ und der relative Fehler proportional zu $1/\sqrt{C}$ ist.

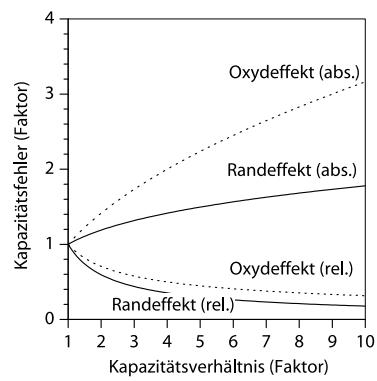


Bild 2.9. Verlauf des relativen und absoluten Fehlers aufgrund der Oxydschichtvariationen (gestrichelt) und der Randeffekte in Abhängigkeit vom Verhältnis zweier Kapazitäten.

Box 2.2 Zufallsbedingte Fehler aufgrund von Randeffekten.

In Shyu et al. 1982 wurde eine Formel zur Abschätzung von zufallsbedingten Fehlern bei quadratischen Plattenkondensatoren hergeleitet und als Grundlage zur Abschätzung einer oberen Schranke verwendet. Im Folgenden wird diese Herleitung für Plattenkondensatoren beliebiger Form durchgeführt.
Annahme: Ein Plattenkondensator der Fläche A habe den Umfang U. Die Kanten des Kondensators sind aufgrund der prozesstechnischen Ungenauigkeiten unscharf begrenzt und seien durch den stochastischen Prozess $l(x)$ modelliert.
Frage: Wie hängt der absolute (relative) Fehler ΔC bzw. $\Delta C/C$ vom Umfang U des Kondensators ab?
Betrachtet man einen beliebigen Randpunkt x der Kondensatoren, kann die Unschärfe der Kanten über unendlich viele Kondensatoren gemittelt als Null angenommen werden, d.h. $l(x)$ ist ein stationärer Prozess mit Erwartungswert Null: $\mathrm{E}\{l(x)\} = 0$. Die Fläche eines beliebigen Kondensators wird durch die unscharf begrenzen Kanten um die Fläche ΔA vergrößert (verkleinert): $A' = A + \Delta A$ mit

$$\Delta A = \int_0^U l(x)dx \tag{2.4}$$

Für die Varianz einer Zufallsvariablen X gilt weiterhin:

$$\begin{aligned}\sigma^2 &= \mathrm{E}\{(X-\mu)^2\} = \mathrm{E}\{X^2 - 2X\mu + \mu^2\} \\ &= \mathrm{E}\{X^2\} - 2\mu\mathrm{E}\{X\} + \mu^2 \\ &= \mathrm{E}\{X^2\} - \mathrm{E}^2\{X\}\end{aligned} \tag{2.5}$$

Da $l(x)$ für einen beliebigen Randpunkt x eine mittelwertfreie Zufallsvariable darstellt, entspricht also die Varianz dem Erwartungswert des Quadrats der Zufallsvariablen und man erhält durch Einsetzen von Gleichung 2.4 in 2.5:

$$\begin{aligned}\sigma^2_{\Delta A} &= \mathrm{E}\{\Delta A^2\} = \mathrm{E}\left\{\left(\int_0^U l(x)dx\right)^2\right\} \\ &= \mathrm{E}\left\{\int_0^U\int_0^U l(x_1)l(x_2)dx_1dx_2\right\} \\ &= \int_0^U\int_0^U \mathrm{E}\{l(x_1)l(x_2)\}dx_1dx_2\end{aligned} \tag{2.6}$$

In Papoulis 1991 wird gezeigt, dass die Autokorrelationsfunktion $R(t_1, t_2)$ eines stochastischen Prozesses $x(t)$ dem Erwartungswert des Produkts $x(t_1)x(t_2)$ entspricht:

$$\mathrm{R}(t_1, t_2) = \mathrm{E}\{x(t_1)x(t_2)\} \tag{2.7}$$

Falls der Mittelwert von $x(t)$ über die Zeit t konstant bleibt, dann ist $x(t)$ im weiteren Sinn stationär und die Autokorrelationsfunktion hängt nur von der Zeitdifferenz $z = t_1 - t_2$ ab. Angewandt auf Gleichung 2.6 ergibt sich (Zeitindex t entspricht x_1 bzw. x_2):

$$\sigma^2_{\Delta A} = \int_0^U\int_0^U \mathrm{R}(x_1 - x_2)dx_1dx_2 \tag{2.8}$$

Um Gleichung 2.8 zu vereinfachen, bietet sich ein Wechsel der Koordinaten bzw. eine Variablensubstitution an. Ausgangspunkt ist dabei die durch Gleichung 2.8 beschriebene Integration des Volumens $R(x_1 - x_2)$ über der quadratischen Grundfläche mit der Seitenlänge U in Bild 2.10. Durch die Substitution

$$z = z(x_1, x_2) = x_1 - x_2$$

wird die Koordinatenachse x_1 auf z geändert, Achse x_2 wird beibehalten. Die Integrationsgrenzen werden angepasst, indem der Integrationsbereich zunächst in die beiden Flächen A_1 und A_2 aufgespalten wird. Die Integration von A_1 beginnt, wie in Bild 2.10 dargestellt, in der linken oberen Ecke des Quadrats und geht bis in die Mitte (graue Fläche), d.h. von $z(-U, 0) = -U$ bis $z(0, 0) = 0$. Die Fläche A_2 (Bild 2.11) wird von der Mitte bis zur rechten unteren Ecke des Quadrats integriert, also von $z(0, 0) = 0$ bis $z(U, 0) = U$. Diese Aufteilung ist nötig, um die Grenzen der Integration in x_2-Richtung ausdrücken zu können. Im ersten Fall reicht sie von $-z$ bis U, im zweiten Fall von 0 bis $U - z$ (schraffierte Bereiche).
Damit wird aus Gleichung 2.8

$$\begin{aligned}
\sigma^2_{\Delta A} &= \int_0^U \int_0^U \mathrm{R}(x_1 - x_2)\,dx_1 dx_2 = \int_{A_1} \mathrm{R} + \int_{A_2} \mathrm{R} \\
&= \int_{-U}^{0} dz \int_{-z}^{U} dx_2 \mathrm{R}(z) + \int_0^U dz \int_0^{U-z} dx_2 \mathrm{R}(z) \\
&= \int_{-U}^{0} \mathrm{R}(z)\,dz\,(U + z) + \int_0^U \mathrm{R}(z)\,dz\,(U - z) \\
&= \int_{-U}^{0} \mathrm{R}(z)\,dz\,U + \int_0^U \mathrm{R}(z)\,dz\,U + \int_{-U}^{0} \mathrm{R}(z)\,dz \cdot z + \int_0^U \mathrm{R}(z)\,dz\,(-z)
\end{aligned}$$

Im letzten Ausdruck können die beiden rechten Integrale unter der Verwendung der Betragsfunktion zusammengefasst werden:

$$\begin{aligned}
\sigma^2_{\Delta A} &= \int_{-U}^{U} \mathrm{R}(z)\,dz\,U - \int_{-U}^{U} \mathrm{R}(z)\,dz\,|z| \\
&= \int_{-U}^{U} (U - |z|)\,\mathrm{R}(z)\,dz
\end{aligned} \tag{2.10}$$

Das Maximum der Autokorrelationsfunktion $\mathrm{R}(z)$ ist bei $z = 0$ und entspricht der Varianz σ_l^2 von $l(x)$ (siehe Gl. 10–43 bei Papoulis 1991). Ab einer gewissen Distanz d von Null entfernt fällt die Funktion stark ab, d.h. $|\mathrm{R}(z)| \ll \mathrm{R}(0)$ für $|z| > d$. Die Distanz d ist gegeben durch die Granulation der Kantenbeschaffenheit (typischerweise ist $d < 1\mu\mathrm{m}$). Da der Umfang U der Kondensatoren in der Regel sehr viel größer ist als d, vereinfacht sich Gleichung 2.10 also weiter:

$$\sigma^2_{\Delta A} \approx U \int_{-d}^{d} \mathrm{R}(z)\,dz \tag{2.11}$$

Diese Näherung dient nun zur Abschätzung einer oberen Schranke für die Varianz von ΔA, direktes Einsetzen ist ohne Kenntnis von $\mathrm{R}(z)$ nicht möglich. Da $\mathrm{R}(z)$ jedoch sein Maximum bei Null mit $\mathrm{R}(0) = \sigma_l^2$ hat, gilt folgende Abschätzung:

$$\sigma^2_{\Delta A} < U \int_{-d}^{d} \mathrm{R}(0)\,dz = U \cdot \mathrm{R}(0) \cdot [z]_{-d}^{d} = 2dU\sigma_l^2 \tag{2.12}$$

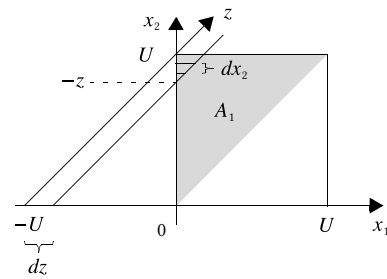


Bild 2.10. Integration des ersten Teilvolumens über einer dreieckigen Grundfläche. Durch Substitution wird ein Koordinatenwechsel durchgeführt, der die Vereinfachung von Gleichung 2.8 erleichtert.

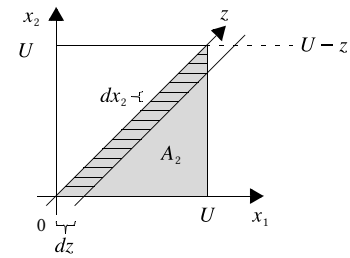


Bild 2.11. Integration des zweiten Teilvolumens über der Grundfläche A_2.

Aus der Formel für die Kapazität eines Plattenkondensators wird ersichtlich, dass die Abweichung ΔC bzw. der RMS-Fehler σ_C proportional zur Streuung der Flächenabweichung ist:

$$C = \frac{\varepsilon}{t} \cdot A' = \frac{\varepsilon}{t} \cdot (A + \Delta A) \quad \Rightarrow \quad \Delta C = \sigma_C \sim \sigma_{A'} \sim \sigma_{\Delta A}$$

Damit gilt schließlich folgende Abschätzung:

$$\boxed{\sigma_C < \frac{\varepsilon}{t} \cdot \sigma_l \cdot \sqrt{2dU}} \tag{2.13}$$

Da die Kapazität eines Plattenkondensators proportional zum Quadrat seines Umfangs ist, gilt:

$$\boxed{\Delta C \sim C^{1/4} \qquad \frac{\Delta C}{C} \sim \frac{1}{C^{3/4}}} \tag{2.14}$$

$$\Delta C \sim \sqrt{C} \qquad \frac{\Delta C}{C} \sim \frac{1}{\sqrt{C}} \tag{2.15}$$

Aus Bild 2.9 wird deutlich, dass der durch Unregelmäßigkeiten der Oxydschicht bedingte Fehler stärker zu Buche schlägt, als die von Randeffekten verursachten Variationen. Zu einem ähnlichen Ergebnis kommt das auf Pelgrom zurückgehende Kondensatormodell für den Mismatch.

Das Pelgrom-Modell

Die Herleitung des Pelgrom-Modells wurde in Pelgrom et al. 1989 im Frequenzbereich durchgeführt. Für das Verständnis hilfreiche Zwischenschritte wurden ausgelassen und eine Reihe von Annahmen getroffen und Vereinfachungen gemacht. Deshalb soll das Modell hier nochmals hergeleitet werden, diesmal im Distanzbereich (nach Terrovitis et al. 1996).

Ausgangspunkt ist die Frage nach der Varianz der Bauteilparameterdifferenz ΔP zweier Bauteile, wobei $P = f(q_1, q_2, \ldots, q_N)$ eine Funktion von N Prozessparametern q_i ist. Mit dem Fehlerfortpflanzungsgesetz von Gauß (siehe Box 4.1 auf Seite 102) kann die Varianz berechnet werden:

$$\sigma^2_{\Delta P} = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial q_i} \right)^2 \sigma^2_{\Delta q_i} \tag{2.16}$$

ANNAHMEN. Die eigentliche Modellbildung beginnt also mit der Frage nach der Varianz der Prozessparameterdifferenz Δq_i. Folgende Annahmen sollen hierbei getroffen werden:

- Der Prozessparameter ist eine Funktion des Ortes (x, y) innerhalb des Bauteils, abgekürzt $q(x, y)$.
- $q(x, y)$ ist ein stochastischer Prozess.
- Die statistischen Eigenschaften von $q(x, y)$ sind konstant, der Mittelwert (an der Stelle (x, y)) ist Null (Prozess ist im weiteren Sinn stationär). Der Prozess q repräsentiert also zunächst nur den Anteil der *lokalen* Schwankungen.
- Der Prozess stellt spatiales, weißes Rauschen dar, d.h. die Werte von q an

zwei verschiedenen Stellen sind völlig unkorreliert.
- Der Mittelwert von q über die Bauteilfläche entspricht näherungsweise dem effektiven Wert q_{eff}, der zum gleichen Wert des Parameters P führt, wie $q(x, y)$.

Der Anteil der globalen Variationen wird getrennt durch einen additiven Term der Prozessparameterdifferenz Δq_i hinzugeschlagen. Annahmen hierbei:
- Die Distanz zwischen den beiden Bauteilen ist klein genug, dass die Variationen als linear angenommen werden können.
- Der Gradient λ ist eine Zufallsvariable mit Mittelwert Null.
- Die beiden Bauteile liegen auf einer der beiden Koordinatenachsen in der Entfernung D.

Lokale Variationen. Wir gehen also davon aus, dass sich der Prozessparameter hauptsächlich in Form des über die Bauteilfläche gemittelten Wertes $\bar{q}$ auf den Bauteilparameter P auswirkt, alle Orte innerhalb der Fläche also gleich starken Einfluss haben:

$$q_{\text{eff}} = \bar{q} = \frac{1}{WL}\int_0^{WL}\int q(x,y)\,dx\,dy \tag{2.17}$$

Hierbei sind W und L die Weite bzw. Länge des Bauteils mit $A = WL$. Der Mismatch (Differenz) des Parameters zweier nicht überlappender, benachbarter Bauteile beträgt dann:

$$\Delta q = \frac{1}{A_1}\int_{A_1} q(x,y)\,dA - \frac{1}{A_2}\int_{A_2} q(x,y)\,dA \tag{2.18}$$

Aus Gleichung 2.6 in Box 2.2 ist bekannt, dass die Varianz von Δq dem Erwartungswert des Quadrats entspricht:

$$\sigma^2_{\Delta q} = \mathrm{E}\{\Delta q^2\} \tag{2.19}$$

Quadrieren von Gleichung 2.18 liefert also:

$$\Delta q^2 = \frac{1}{A_1^2}\left(\int_{A_1} q(x,y)\,dA\right)^2 + \frac{1}{A_2^2}\left(\int_{A_2} q(x,y)\,dA\right)^2 - \frac{2}{A_1^2 A_2^2}\int_{A_1} q(x,y)\,dA \int_{A_2} q(x,y)\,dA \tag{2.20}$$

Der erste Teil von Gleichung 2.20 lässt sich folgendermaßen schreiben:

$$\frac{1}{A_1^2}\left(\int_{A_1} q(x,y)\,dA\right)^2 = \frac{1}{A_1^2}\int_{A_1}\int_{A_1} q(x_a,y_a)\,q(x_b,y_b)\,dx_a\,dy_a\,dx_b\,dy_b \tag{2.21}$$

Da der Prozess $q(x, y)$ im weiteren Sinn stationär ist, hängt die Autokorrelationsfunktion nur von der Differenz $x_a - x_b$ bzw. $y_a - y_b$ ab. Entsprechend Gleichung 2.7 gilt:

$$\mathrm{E}\left\{\frac{1}{A_1^2}\left(\int_{A_1} q(x,y)\,dA\right)^2\right\} = \frac{1}{A_1^2}\int_{A_1}\int_{A_1} \mathrm{R}(x_a - x_b, y_a - y_b)\,dx_a\,dy_a\,dx_b\,dy_b \tag{2.22}$$

Generell gilt, dass weißes Rauschen zu verschiedenen Zeitpunkten t_1 und t_2 völlig unabhängig voneinander ist. Die Autokorrelation ist also nur an der Stelle $\tau = t_1 - t_2 = 0$ ungleich Null. Zusammen mit dem Wiener-Khintchine-Theorem ergibt sich also (S_0 Leistungsdichte):

$$R(\tau) = S_0 \delta(\tau) \tag{2.23}$$

Weil es sich bei $q(x, y)$ um weißes Rauschen handelt, gilt die auf zwei Dimensionen erweiterte Form von Gleichung 2.7:

$$E\left\{\frac{1}{A_1^2}\left(\int_{A_1} q(x,y)dA\right)^2\right\} = \frac{1}{A_1^2}\int_{A_1}\int_{A_1} S_0\delta(x_a - x_b, y_a - y_b)dx_a dy_a dx_b dy_b \tag{2.24}$$

Das innere Integral auf der rechten Seite dieser Gleichung lässt sich lösen, indem überprüft wird, ob der Punkt $(0,0)$ Teil der Integrationsfläche ist und der Dirac-Impuls durch eine Eins ersetzt wird. Da die Punkte x_a, x_b, sowie y_a und y_b derselben Fläche entnommen werden, ist dies der Fall und man erhält:

$$E\left\{\frac{1}{A_1^2}\left(\int_{A_1} q(x,y)dA\right)^2\right\} = \frac{S_0}{A_1^2}\int_{A_1} dx_b dy_b = \frac{S_0}{A_1} \tag{2.25}$$

Der zweite Term in Gleichung 2.20 liefert auf dieselbe Art und Weise:

$$E\left\{\frac{1}{A_2^2}\left(\int_{A_2} q(x,y)dA\right)^2\right\} = \frac{S_0}{A_2} \tag{2.26}$$

Der dritte Term ist ein Mischterm aus den Integralen jeweils über die Fläche A_1 und A_2 der beiden Bauteile. Da diese sich nicht überlappen, ist der Punkt $(0,0)$ nicht Teil der Integrationsfläche und es folgt:

$$E\left\{\frac{2}{A_1^2 A_2^2}\int\ldots\right\} = \frac{2}{A_1^2 A_2^2}\int_{A_1}\int_{A_2} S_0\delta(x_1 - x_2, y_1 - y_2)dA_2 dA_1 = 0 \tag{2.27}$$

Die Varianz der Prozessparameterdifferenz ist also die Summe der Erwartungswerte in Gleichung 2.25 und Gleichung 2.26. Die Flächen der beiden Bauteile können als fast identisch angenommen werden ($\alpha \approx 1$). Durch Einführung einer für den Prozess charakteristischen, experimentell zu ermittelnden Konstante S_i^2 des Parameters q_i folgt schließlich:

$$\sigma^2_{\Delta q_i} = \frac{S_0}{A_1} + \frac{S_0}{A_2} = \frac{S_0(\alpha + 1)}{A} = \frac{S_i^2}{A} \tag{2.28}$$

Führt man diese Herleitung für Prozessparameter aus, die nur von einer Dimension abhängen, beispielsweise für W und L, so erhält man:

$$\sigma^2_{\Delta W} = \frac{S_W^2}{L} \qquad \sigma^2_{\Delta L} = \frac{S_L^2}{W} \tag{2.29}$$

Hierbei sind S_L^2 und S_W^2 wieder experimentell zu ermittelnde Konstanten.

GLOBALE VARIATIONEN. Nun sei angenommen, dass der Prozessparameter q_i auch die globalen Variationen umfasst, so dass noch ein additiver Term zu den lokalen Schwankungen hinzukommt. Dieser Teil besteht in einer systematischen Abhängigkeit von der Distanz D aufgrund des zufälligen, mittelwertfreien Gradienten λ:

$$\Delta q = \Delta q_{\text{local}} + \lambda D \tag{2.30}$$

Die Varianz beträgt schließlich mit $G^2 = \sigma_\lambda^2$:

$$\sigma_{\Delta q_i}^2 = \frac{S_i^2}{A} + G_i^2 D^2 \tag{2.31}$$

Gleichung 2.31 stellt das Gesamtmodell nach Pelgrom dar.

Das Kondensatormodell für den Mismatch

In der Praxis wird das Pelgrom-Modell z.B. zur Beschreibung der Matching-Eigenschaften von Plattenkondensatoren verwendet. Dabei handelt es sich in der Regel um Poly1-Poly2 Plattenkondensatoren oder spezielle MIM-Plattenkondensatoren („metal-insulator-metal"), sofern die Prozesstechnik über solche verfügt. Auf kompliziertere Strukturen, insbesondere die in dieser Arbeit vorgestellten 3D-Cluster kann das Modell kaum angewandt werden, da die Gleichung(en) für die Kapazität (siehe Abschnitt 2.2) im allgemeinen Fall analytisch nicht lösbar sind.

Modellierung des Plattenkondensators. Dies stellt den einfachsten Fall eines integrierten Kondensators dar, insbesondere wenn Streufelder an den Rändern vernachlässigt werden. Angenommen die Kapazität pro Flächeneinheit sei C_A, dann führt die Anwendung des Fehlerfortpflanzungsgesetzes auf die Kapazität $C = C_A WL$ zu:

$$\left(\frac{\sigma_{\Delta C}}{C}\right)^2 = \left(\frac{\sigma_{\Delta C_A}}{C_A}\right)^2 + \left(\frac{\sigma_{\Delta W}}{W}\right)^2 + \left(\frac{\sigma_{\Delta L}}{L}\right)^2 + \frac{2}{WL}\varrho_{WL}\sigma_{\Delta W}\sigma_{\Delta L} \tag{2.32}$$

Der Korrelationskoeffizient ϱ_{WL} soll berücksichtigen, dass Weite und Länge z.B. einer Leiterbahn im selben Prozessschritt festgelegt werden und insofern korreliert sind. Dies gilt jedoch nur für die globalen Variationen. Die lokal begrenzte Rauheit der Kanten dagegen kann als unkorreliert angenommen werden, so dass sich der letzte Term in Gleichung 2.32 durch einen Ausdruck ersetzen lässt, der die Korrelation des globalen Gradienten von W und L, sowie die Entfernung der Bauteile berücksichtigt. Da die Varianz von C_A nur von der Permittivität und der Oxydschichtdicke, jedoch nicht von W und L abhängt, hat sie die Form von Gleichung 2.31 und man erhält:

$$\left(\frac{\sigma_{\Delta C}}{C}\right)^2 = \left(\frac{S_{C_A}^2}{WL} + G_{C_A}^2 D^2\right) + \frac{1}{W^2}\left(\frac{S_W^2}{L} + G_W^2 D^2\right) + \frac{1}{L^2}\left(\frac{S_L^2}{L} + G_L^2 D^2\right) + \frac{2}{WL}\varrho_{\lambda_W \lambda_L} D^2 \tag{2.33}$$

Lokaler Mismatch. In der Praxis wird üblicherweise versucht, durch Einhalten spezieller Layoutregeln (z.B. „common-centroid" Anordnung) den Einfluss von globalen Variationen zu minimieren. Betrachtet man oben in Gleichung 2.33 nur den Anteil des lokalen Mismatch, so vereinfacht sich der Ausdruck weiter:

$$\left(\frac{\sigma_{\Delta C}}{C}\right)^2 = \frac{S_{C_A}^2}{WL} + \frac{1}{W^2} \cdot \frac{S_W^2}{L} + \frac{1}{L^2} \cdot \frac{S_L^2}{L} \tag{2.34}$$

Falls die Kondensatoren quadratisch und sehr groß sind, so fallen die letzten beiden Terme in dieser Gleichung wenig ins Gewichtung und man erhält als Schätzung:

$$\boxed{\left(\frac{\sigma_{\Delta C}}{C}\right)^2 \approx \frac{S_{C_A}^2}{WL}} \tag{2.35}$$

Simulation

MONTE-CARLO SIMULATION. In der Praxis wird der Mismatch zwischen zwei Kondensatoren üblicherweise über Gleichung 2.35 modelliert und der Parameter S_{C_A} experimentell ermittelt. Damit verbunden sind – wie aus der Herleitung hervorgeht – folgende Einschränkungen:

- Nur der Anteil der lokalen Variationen für die Berechnung des Mismatch wird modelliert.
- Globale Parameterdrifts wirken auf alle Bauteile gleichermaßen.
- Es handelt sich um einfache Plattenkondensatoren.
- Streufelder an den Rändern werden vernachlässigt.
- Die Kondensatoren sind groß.

In Box 2.3 ist ein typisches Beispiel für die Simulation der Prozessschwankungen eines Plattenkondensators gemäß Gleichung 2.35 gegeben. In jedem Zyklus n der Monte-Carlo Iterationsschleife wird ein neuer Wert C_n berechnet und die Gesamtschaltung erneut simuliert:

$$C_n = (C_A A + C_P U) + \Delta C = C + C \cdot \frac{S}{\sqrt{A}} \cdot \text{NormalDist}(0, 1) \tag{2.36}$$

Hierbei entsprechen A und U der Fläche bzw. dem Umfang des Kondensators und S der experimentell ermittelten, für den Prozess charakteristischen Konstante. Eine normalverteilte Zufallszahl mit Mittelwert Null und Standardabweichung Eins sorgt multiplikativ für die eigentliche Streuung.

SIMULATION DER PROCESS-CORNERS. Eine andere Simulationsform besteht in der Simulation einer Schaltung (oder einzelner Bauteile) unter den Extrema der Prozessparameter („process-corners“). Je nachdem, ob das Parametermaximum oder -minimum für ein günstiges, typisches oder ungünstiges Schaltungsverhalten verantwortlich ist, werden diese Extrema in die Kategorien „best-case“, „typical-case“ und „wort-case“ eingeordnet (manchmal auch mehr). Der Benutzer kann dann zwischen diesen zu Beginn eines Simulationsdurchlaufs auswählen.

Da die Auswirkungen von Prozessschwankungen nur bei Plattenkondensatoren mittels des Pelgrom-Modells simuliert werden können, besteht die einzige Alternative zur Untersuchung des Einflusses der Parametervariationen auf die Kapazität von Verbindungsleitungen und komplexen 3D-Strukturen in der Simulation an den „Ecken“ des Prozesses.

Da die Kapazität solcher Strukturen von einer ganzen Reihe von Prozessparametern abhängt, die in den process-corners nicht berücksichtigt werden (siehe Bild 2.12), ist diese Simulationsform jedoch ebenfalls wenig geeignet, um die kapazitiven Schwankungen der in dieser Arbeit vorgestellten 3D-Cluster korrekt und im vollen Umfang wiederzugeben.

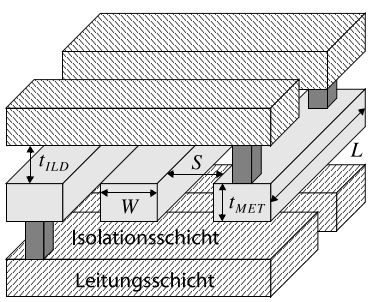


Bild 2.12. Die geometrischen Einflussgrößen, die für Kapazitätsschwankungen verantwortlich sind. Bei Simulation der worst-, typical- und best-case Prozessextrema („corners“) werden nur die Dicken variiert (t_{ILD} und t_{MET}).

* * *

Box 2.3 Modellierung beim AMS 0,35 µm Analogprozess.

Als Beispiel für die Modellierung der Prozessschwankungen eines Poly1-Poly2 Plattenkondensators beim 0,35 µm Analogprozess der Firma Austria Microsystems (AMS) dienen die Monte-Carlo Simulationsmodelle für das Spice-Derivat „Spectre“ von Cadence.

Aus dem Schaltplan werden dabei die Fläche `A` und der Umfang `U` des Bauteils übernommen, sowie vom Benutzer die Art der Simulation: `process` (globale, gemeinsame Parameterdrifts), `mismatch` (lokaler Mismatch) oder beide. Die Parameter werden dann in jeder Monte-Carlo Iteration entsprechend der Definition des jeweiligen Abschnitts zufällig variiert und `C` neu berechnet.

```
...
C = Ca * A + Cp * U + mc_cpoly * 6.9e-9 *
   (Ca * A + Cp * U) / sqrt( A )
...
parameters Ca = 8.89e-4
parameters Cp = 8.7e-11
parameters mc_cpoly = 0
...
statistics {
  process {
    vary Ca dist=unif N=8.9e-5  percent=no
    vary Cp dist=unif N=3.1e-12 percent=no
  }
  mismatch {
    vary mc_cpoly dist=gauss std=1
  }
}
```

Hinweis: Die Werte im Beispiel wurden aufgrund einer Verschwiegenheitserklärung („non-disclosure agreement“, NDA) geändert. Sie entsprechen nicht den Daten von AMS, Ähnlichkeiten wären rein zufällig.

2.2 Kapazitätsberechnung

Formal ist die Kapazität als Konstante definiert, die Ladung und Spannung miteinander verknüpft und von der Form des Körpers abhängt, auf dem die Ladung sitzt. Die Herleitung der Kapazität geschieht meist über das elektrische Potential einer gegebenen Ladungsmenge und führt auf eine fundamentale Beziehung, die Poisson-Gleichung.

2.2.1 Die Poisson-Gleichung

Für die Berechnung der elektrischen Kapazität eines Körpers genügt es, vom elektrostatischen Spezialfall auszugehen, d.h. der Annahme, dass keine elektrischen Ströme fließen, sondern nur ruhende Ladungsträger vorhanden sind. Aus der Definition des elektrischen Feldes folgt, dass die Ladungsträger Quellen der elektrischen Feldlinien sind, genauer gesagt: Die Feldlinien beginnen in positiven Ladungen und enden in negativen.

ELEKTRISCHE FLUSSREGEL. Legt man eine geschlossene Fläche A um eine Ladung Q, so erzeugen die aus der Fläche austretenden Feldlinien des elektrischen Feldes $\vec{E}$ einen elektrischen Fluss ϕ, der proportional zur eingeschlossenen Ladung ist:

$$\phi = \frac{1}{\varepsilon_0} Q = \oint_A \vec{E} d\vec{A} \tag{2.37}$$

Der Gauß'sche Integralsatz verknüpft nun diese Regel mit der Divergenz eines beliebigen Vektorfeldes $\vec{E}$, indem er die Äquivalenz des Durchflusses durch eine geschlossene Oberfläche mit dem Integral über die Divergenz des Vektorfeldes im Inneren des Volumens ausnutzt. Dadurch wird aus dem Oberflächeintegral ein Volumenintegral, man stellt also eine Bilanz über die Quellen (positive Ladungsträger) und Senken (negative Ladungsträger) infinitesimal kleiner Volumenstücke im Inneren auf:

$$\oint_A \vec{E} d\vec{A} = \int_V \operatorname{div} \vec{E} dV \tag{2.38}$$

Definiert man nun eine Ladungsdichte ϱ als Ladung Q pro Volumen V, so folgt – weil die Gleicheit für beliebige Oberflächen gilt – mit

$$Q = \int_V \varrho dV \tag{2.39}$$

aus Gleichung 2.37 und Gleichung 2.38 die Poisson-Gleichung. Sie repräsentiert im Endeffekt nur eine mathematisch andere Form der Flussregel:

$$\operatorname{div} \vec{E} = -\operatorname{div} \operatorname{grad} U = -\Delta U = \frac{1}{\varepsilon_0} \varrho \tag{2.40}$$

Hierbei wurde ausgenutzt, dass ein elektrisches Feld durch den Gradienten des skalaren Potentials U ausgedrückt werden kann.

LAPLACE-GLEICHUNG. Für den Spezialfall $\Delta U = 0$ erhält man die Laplace-Gleichung, sie gibt die Beziehung für raumladungsfreie Kapazitäten. Alle im Rahmen dieser Arbeit vorgestellen rechnergestützten Extraktionsprogramme beschränken sich auf diesen Fall.

2.2.2 Berechnungsverfahren

Eine Vielzahl von Ansätzen zur Lösung der Poisson-Gleichung oder eines Spezialfalls wurden in der Vergangenheit vorgeschlagen. Darunter finden sich analytische Lösungen, die teilweise vereinfachende Annahmen voraussetzen, und numerische Verfahren, die in der Regel zur Berechnung auf dem Computer (Extraktion) entwickelt wurden.

Analytische Lösung und Näherungsformeln

Der analytische Ansatz liefert zwar einen geschlossenen Ausdruck, stellt jedoch keine allgemeingültige Lösung der Poisson- oder Laplace-Gleichung dar. Es gibt zwei Strategien, die in der Literatur verfolgt werden: Der in Chang et al. 1976 gewählte Ansatz stellt die genaueste Lösung dar, jedoch nur im geometrisch eingeschränkten Fall einer oder zweier übereinander verlaufender Metallbahnen, die über ein Dielektrikum vom Substrat getrennt sind, das dicker ist als die Breite der Leiterbahnen.

Bei anderen Geometrieformen finden sich in der Literatur diverse Ansätze, die mehr Näherungsformeln darstellen, statt Lösungen der Laplace-Gleichung. Meistens wird der Leiter als Zusammensetzung aus einem Plattenkondensator und zweier Zylinderhälften modelliert, die den kapazitiven Anteil der Streufelder („fringe“) an den Leitungsrändern berücksichtigen sollen.

Letztendlich gibt es eine exakte analytische Lösung der Poisson- bzw. Laplace-Gleichung also nur für sehr einfache, regelmäßige Objekte, z.B. Kugeln oder Zylinder unendlicher Länge. Andere Strukturen müssen auf diese abgebildet werden, so dass die Formel für die Kapazität nur eine Näherung darstellt und bei komplizierten Objekten vollends versagt.

Numerische Verfahren

Bei den numerischen Verfahren gibt es zwei Hauptkategorien, in die sich die Algorithmen einsortieren lassen: Die als „Finite Difference“ und „Finite Element“ bezeichneten Methoden lösen die differentielle Form der Laplace-Gleichung (Gleichung 2.40), die anderen Verfahren die Integralform (Gleichung 2.37) oder kombinieren beide (siehe Bild 2.13).

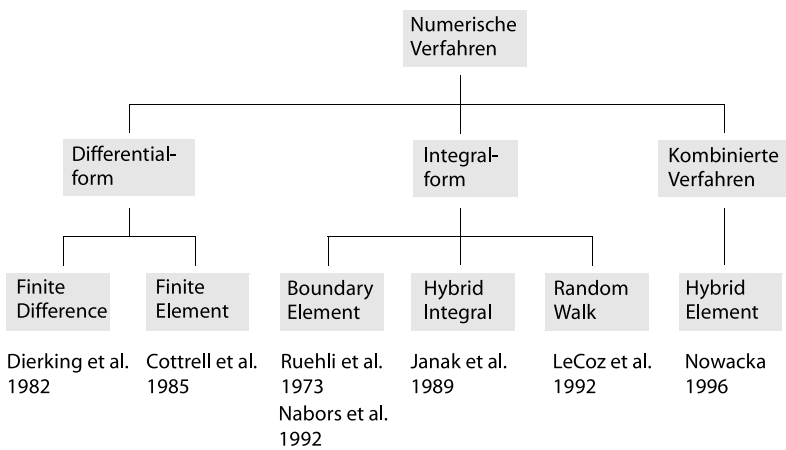


Bild 2.13. Die Verfahren zur numerischen Kapazitätsberechnung lösen entweder die differentielle Form der Poisson- bzw. Laplace-Gleichung (Gleichung 2.40), die Integralform (Gleichung 2.37) oder basieren auf einer Kombination aus beiden.

DIFFERENTIALFORM. Die Finite-Elemente-Methode ist generell ein Verfahren, um partielle Differentialgleichungen unter Randbedingungen zu lösen. Es findet sich daher eine Vielzahl an Anwendungen in den Ingenieurswissenschaften, Standardwerke wie Norrie et al. 1973 bieten allgemeine Einführungen hierzu an. In Cottrell et al. 1985 wird das Verfahren auf die Kapazitätsberechnung der Chipverdrahtung angewandt.

Kennzeichnend ist die Diskretisierung des Berechnungsgebietes über das gesamte Volumen, wodurch sich ein großes, dünnbesetztes Gleichungssystem ergibt. Die Berechnung ist damit für große Chips sehr aufwendig und problematisch bei falscher Wahl der Gitterpunkte. Schwierig ist auch die Behandlung des Rands des zu berechnenden Gebietes, elektrische Felder, die ins Unendliche reichen, sind schlecht zu modellieren. Vorteilhaft ist, dass keine Einschränkung bei den Isolationsschichten gemacht werden muss. Die Dielektrika können verschiedene Werte aufweisen und von völlig unregelmäßiger Struktur sein.

INTEGRALFORM. Hierbei wird die *Oberfläche* der Leiter und Dielektrika diskretisiert, entsprechend also eine Lösung für das Oberflächenintegral in Gleichung 2.37 berechnet. Es wird von perfekten Leitern ausgegangen, d.h. die Ladung befindet sich komplett am Rand. Dadurch ergibt sich ein sehr viel kleineres und komplett besetztes Gleichungssystem, das effizienter gelöst werden kann. Elektrische Felder, die ins Unendliche reichen, können mit dieser Methode korrekt behandelt werden, solange die Randbedingungen einfach sind. Bei variierenden und ungleichmäßigen Isolationsschichten versagt die Methode jedoch.

Aus diesem Grund wird beispielsweise bei Ruehli et al. 1973 auf ein „hybrid boundary element method“ genanntes Verfahren zurückgegriffen, das auch unregelmäßige Dielektrika behandeln kann, jedoch höhere Kosten verursacht. Ändert sich zusätzlich noch die Dielektrizitätskonstante, so wird meist eine Kombination aus BEM- und FEM-Methoden angewandt. Einige der in kommerziellen Extraktionstools verwendeten Algorithmen koppeln diese beiden Ansätze, in der Dissertation von Nowacka 1996 wird ausführlich darauf eingegangen.

STATISTISCHE METHODEN. In Haji-Sheikh et al. 1966 wird zum ersten Mal ein sogenannter „floating random-walk“ Algorithmus zur Lösung der Laplace-Gleichung im Fall einer Hitzeverteilung vorgestellt. Zwar waren Monte Carlo Verfahren zur Lösung von Differentialgleichungen schon lange davor bekannt, jedoch wurde die praktische Anwendbarkeit erst im Zuge der fortschreitenden Computertechnik Gegenstand der Forschungstätigkeit.

Der Algorithmus wurde dann in LeCoz et al. 1991 für die Kapazitätsberechnung von Leiterbahnen komplexer ICs weiterentwickelt. Charakteristisch für das Verfahren ist die statistische Schätzung der Kapazität, und zwar nur dort, wo es nötig ist: die Oberfläche jedes elektrischen Leiters, entsprechend handelt es sich um ein Verfahren zur Lösung der Integralform (Gleichung 2.37). Je länger die Berechnung läuft, desto genauer werden die Schätzungen. Unregelmäßige Dielektrika und abrupte Permittivitätswechsel werden durch stückweise lineare Approximation modelliert.

2.2.3 Extraktion

Einige der vorgestellten Algorithmen wurden von führenden EDA-Herstellern (Cadence, Synopsys, Magma) in ihre Softwareprodukte für den Chipentwurf integriert. Diese als Extraktionstools bezeichneten Programme lösen die Integralform der Laplace-Gleichung, meist um die Kapazität kritischer Leitungsnetze sehr genau zu berechnen (TCAD-Klasse, sog. „Field-Solver“).

Für den Anwender dieser Tools ist es in der Regel sehr schwer, technische Detailinformationen über die verwendeten Algorithmen zu bekommen, die Dokumentation beschränkt sich fast ausschließlich (mit Ausnahme von Quickcap) auf die Benutzungsanleitung. Weder der Name des Algorithmus, noch die Originalpublikationen zu den mathematischen Hintergründen werden angegeben.

Durch intensive Literaturrecherchen konnte in einigen Fällen rekonstruiert werden, welchen Weg der Algorithmus von der Theorie (Originalpublikation aus Bild 2.13) über erste mit inoffiziellen Namen versehene Software-Implementierungen bis hin zu den kommerziell vertriebenen Extraktionstools nahm.

Gängige Extraktionstools

Assura-FS. Unter der Produktbezeichnung „Assura“ vertreibt die Firma Cadence eine ganze Reihe von Verifikationsprogrammen. Über einen (etwas verstecken) Knopf in der Benutzeroberfläche der IC-/Virtuoso-Plattform lässt sich der sogenannte „Field-Solver“ aktivieren. Dahinter verbirgt sich ein Programm, das in der Literatur als „Nebula“ oder anfänglich „IES3“ bezeichnet wurde, der Algorithmus basiert auf der Boundary-Elements Methode in Kombination mit einer Fast-Multipole Optimierung. Er wurde in Kapur et al. 1998 zwar nicht das erste Mal vorgestellt, doch finden sich die meisten Referenzen auf diese Publikation. In Kapur et al. 2000 wird die erste Implementierung (Nebula) zusammen mit Zahlen zum Rechenaufwand vorgestellt.

Quickcap. Unter diesem Namen vertreibt die Firma Magma den einzigen nicht-deterministischen Extraktor auf dem Markt, er geht auf den „random-walk“ Algorithmus zurück, der in LeCoz et al. 1992 und in Iverson et al. 2001 beschrieben wird. Grundlage des Algorithmus ist die Integralform der Laplace-Gleichung, die durch einen Monte-Carlo Ansatz gelöst wird.

FastCap. Hierbei handelt es sich um frei verfügbare Software, die für die gängigsten Betriebssysteme und im Quelltext aus dem Internet bezogen werden kann. Da es sich um das Ergebnis der Forschungstätigkeit einer (oder mehrerer) öffentlicher Einrichtungen handelt, ist die Integration in die üblichen Layoutwerkzeuge der führenden EDA-Hersteller mangelhaft bzw. nicht vorhanden. Auch existieren keine der üblichen Schnittstellen (z.B. GDSII), das Design muss über spezielle Geometrie-Primitiven modelliert werden.

Sind diese Hürden jedoch überwunden, so ist FastCap das einzige kostenlose Extraktionsprogramm, das sich etabliert hat. Im Gegensatz dazu werden für die Produkte in Tabelle 2.3 teilweise 100.000 Euro und mehr verlangt. Die mathematischen Grundlagen des Boundary-Elements Algorithmus sind gut dokumentiert, z.B. in Nabors et al. 1991 und Nabors et al. 1992, die praktische Benutzbarkeit des Programms jedoch eher schlecht.

Produkt	Hersteller	Algorithmus
Quickcap	Magma	Floating Random Walk
Assura-FS (Nebula, IES3)	Cadence	Boundary Elements u. Fast Multipole
Exact	Silvaco	Unbekannt
Raphael	Synopsys	Boundary Elements u. Finite Difference
Q3D Extractor	Ansoft	Boundary Elements u. Method of Moments
CELL-AN, NET-AN	OEA	Unbekannt (Cheetah II)

Tabelle 2.3. Die gängigsten Extraktionswerkzeuge aus der TCAD-Klasse der Field-Solver im Überblick (kommerziell).

Projekt	Quelle	Algorithmus
SAP	TU Wien	Finite Elements
CapCal	Siemens	Finite Difference
Space	Uni Delft	Hybrid Elements
FastCap	MIT	Boundary Elements und Fast Multipole

Tabelle 2.4. Einige Algorithmen wurden im Rahmen von Forschungsprojekten zu eigenständigen Programmen zur Kapazitätsextraktion weiterentwickelt.

DIVA, CALIBRE-XRC. Das Produkt Diva der Firma Cadence fällt aus dem Rahmen der hier vorgestellen Algorithmen, da das Programm weniger zur (hochgenauen) Extraktion parasitärer Leitungskapazitäten entwickelt wurde, sondern mehr zur Schaltungsrückerkennung aktiver Bauteile für den Layout-versus-Schematic Vergleich (LVS). Er fällt also nicht in die TCAD-Klasse der Extraktoren, die numerisch die Laplace-Gleichung lösen, sondern arbeitet mit einem gänzlich unbekannten Algorithmus. Die Laufzeit am Rechner selbst für große Designs ist sehr gering, die Extraktionsgenauigkeit ebenfalls.

Ähnliches gilt für Calibre-xRC von Mentor Graphics, wenngleich der zugrundeliegende Algorithmus etwas besser geeignet zu sein scheint, parasitäre Leitungskapazitäten zu extrahieren. Auch hier gibt es keine Informationen über die mathematischen Hintergründe. Zur Frage der Laufzeit und zur Genauigkeit bei Diva und Calibre sei auf Abschnitt 4.1 („Extraktion“) auf Seite 100 ff. verwiesen.

* * *

2.3 Kapazitätsmessung

2.3.1 Klassische Ladungspumpen

In Chen et al. 1996 wurde zum ersten Mal ein Schaltungsprinzip vorgestellt, das die Messung von kleinsten Kapazitäten mit einer Auflösung von bis zu 10 Attofarad ermöglicht. Damit können winzige Überlappkapazitäten ermittelt werden, wie sie beispielsweise am Kreuzungspunkt zweier Leiterbahnen auftreten oder zwischen benachbarten Leitungen derselben Ebene. Die Messungen finden dabei weitgehend integriert statt, einzig ein Amperemeter ist vonnöten, sowie neben der Spannungsversorgung ein Taktgenerator, der zwei Rechtecksignale („non-overlapping“) erzeugt.

Schaltungsprinzip

Prinzipiell nutzt das Verfahren die Tatsache aus, dass die Menge an Ladung, die bei einer gegebenen Spannung auf einem zu ermittelnden Kondensator gespeichert wird, proportional zu seiner elektrischen Kapazität ist:

$$C = \frac{Q}{U} \quad \Rightarrow \quad Q = U \cdot C$$

Diese Ladung macht sich als kurzzeitiger Stromstoß bemerkbar, sobald die Spannung U an eine entladene Kapazität C angeschlossen wird und die Ladung auf den Kondensator übergeht (in diesen „gepumpt“ wird). Wird dieser Vorgang fortwährend wiederholt, indem der Kondensator nach jeder Pumpphase entladen wird, so ergibt sich daraus ein Stromfluss, der über einen längeren Zeitraum gemittelt annähernd konstant ist.

Bei einer Ladungsmenge Q im Zeitraum t (gegeben durch die Frequenz f) und bei Spannung U ergibt sich also folgende Beziehung:

$$\frac{Q}{t} = U \cdot C \cdot \frac{1}{t} \quad \Leftrightarrow \quad I = U \cdot C \cdot f \quad \Leftrightarrow \quad C = \frac{I}{U} \cdot \frac{1}{f} \tag{2.41}$$

Die Pump- und Entladezyklen werden dabei über PMOS- bzw. NMOS-Schalter realisiert, die über zwei Taktsignale gesteuert werden (siehe Bild 2.14). Um zu gewährleisten, dass nicht beide Schalter gleichzeitig leiten, werden die Transistoren durch sogenannte „non-overlapping“ Pulse angesteuert (Bild 2.14, rechte Seite), d.h. es gibt keinen Zeitpunkt, an dem beide Signale aktiv sind ($V_{1n} = V_{DD}$ und $V_{2p} = 0$). Die Zeitspanne, in der die Transistoren jeweils leiten, muss dabei lang genug sein, um den Kondensator vollständig aufzuladen bzw. zu entladen.

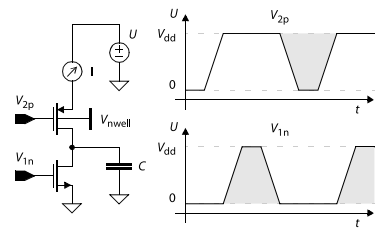


Bild 2.14. Ladungspumpe zur integrierten Kapazitätsmessung.

Für f und U in Gleichung 2.41 können theoretisch beliebige Werte eingesetzt werden, so dass sich für jede zu messende Kapazität eine unbegrenzte Zahl an Messpunkten ergibt. In der Praxis sind Spannung und Frequenz durch eine Reihe von Parametern (z.B. die „Overdrive“-Spannung der Transistoren, sowie die bereits erwähnte minimale Zeit zum Auf-/Entladen) auf bestimmte, technologieabhängige Bereiche beschränkt. Ein konkretes Beispiel ist im Abschnitt „Auswertesystematik“ auf Seite 82 oder in Bild 2.15 zu finden.

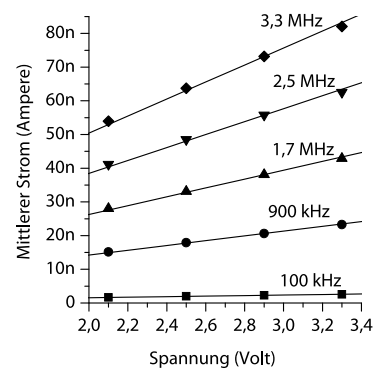


Bild 2.15. Mittlerer Strom als Funktion der Spannung für 5 Frequenzen. Die Steigung der Geraden durch die Frequenz ergibt jeweils einen Kapazitätswert (hier Mittelwert 7,762 fF).

Jeder Messpunkt kann schließlich bei gegebener, fester Frequenz in ein Diagramm als Funktion der Spannung eingetragen werden (alternativ kann der Strom gegen die Frequenz bei gegebener Spannung aufgetragen werden). Die zu einer festen Frequenz gehörenden Messwerte müssen dabei auf einer Geraden durch den Ursprung liegen, deren Steigung I/U beträgt. Durch li-

neare Interpolation der Messpunkte einer Frequenz lässt sich die Steigung dieser Geraden leicht bestimmen, wodurch sich Ungenauigkeiten bei der Bestimmung von I und beim Einspeisen der Versorgungsspannung und der Taktsignale reduzieren lassen. Für mehrere Frequenzen ergibt sich so eine Geradenschar, aus der pro Frequenz jeweils ein Kapazitätswert resultiert. Durch Mittelung dieser Werte minimiert sich die Fehleranfälligkeit weiter. In Bild 2.15 ist das Ergebnis einer solchen Messreihe für jeweils vier Spannungswerte bei fünf Frequenzen zu sehen. In diesem Beispiel beträgt die mittlere Kapazität 7,762 Femtofarad.

Kapazitätsauflösung

Fehlerquelle	**Art**
Mismatch der Gate-Drain/Gate-Source Überlappkapazitäten.	zufällig
Mismatch der Gate-Drain/Gate-Source Diodenkapazitäten	zufällig
Mismatch der Schwellenspannung	zufällig
Ladungsumverteilung	systematisch

Tabelle 2.5. Die Auflösung der Ladungspumpe limitierende Fehlerquellen (oben). Die Ladungsinjektion/-umverteilung ist systematisch, hebt sich jedoch aufgrund der Abhängigkeit von C im Nettostrom nicht vollständig auf.

MISMATCH. Ein wichtiges Prinzip ist, dass die Messungen differentiell vorgenommen werden, d.h. dass vom mittleren Strom I, der bei der zu ermittelnden Kapazität C gemessen wurde, der mittleren Strom I' einer Ladungspumpe *ohne* Messkapazität abgezogen wird. Der Nettostrom $I_{\text{net}} = I - I'$ enthält damit keine Ladungen mehr, die auf die parasitären Kapazitäten der Transistoren gepumpt werden, sowie keine Fehler, die durch Störungen des Ladungsbudgets (z.B. Ladungsverlust, „Leakage“) entstehen und dabei beide Ladungspumpen *in gleichem Umfang* betreffen.

Die prozessbedingten Schwankungen der parasitären Source- bzw. Drain-Überlapp- und Diodenkapazitäten, sowie der Schwellenspannung, führen jedoch dazu, dass I und I' unterschiedlich stark verfälscht werden, selbst bei identischer Messkapazität. Im Nettostrom findet sich also ein gewisser Ladungsanteil, der sich durch das differentielle Prinzip nicht heraushebt. Auf diese Weise stellt im wesentlichen der Mismatch der Transistoren die Grenze der Auflösung dar.

LADUNGSUMVERTEILUNG. Neben dem unvermeidbaren, prozessbedingten Mismatch führt die Umverteilung der Ladung („charge redistribution“) aus dem Kanal zu einem *systematischen* Fehler, der die Genauigkeit beeinträchtigt, sofern keine Gegenmaßnahmen ergriffen werden (durch späteres Herausrechnen oder schaltungstechnische Mittel). Im Moment des Abschaltens des Transistors verteilt sich die Kanalladung zu gewissen Teilen in die an Source und Drain des Transistors angeschlossenen Netze. In Bild 2.16 sind diese Effekte für zwei Schaltzyklen in der 0,35 µm AMS-Technologie gezeigt, die Transistoren haben jeweils minimale Strukturgröße.

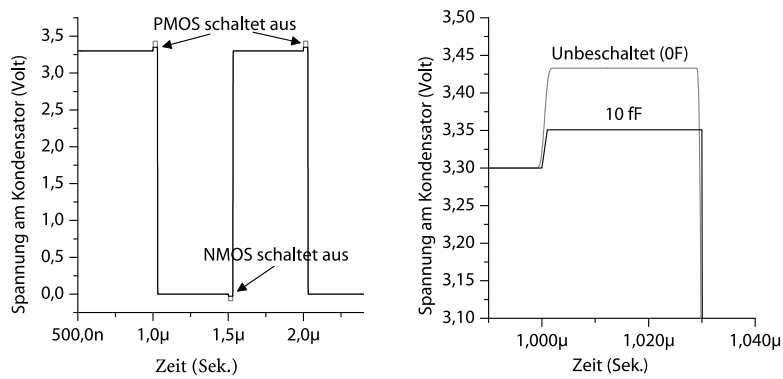


Bild 2.16. Ladungsinjektion und - umverteilung im Moment des Ausschaltens der Transistoren (Simulation). Ladungsträger im Kanal führen zu Spannungsspitzen an der Messkapazität C_x. Der Spannungssprung ist bei der unbeschalteten Ladungspumpe deutlich größer, als bei einer Beschaltung mit 10 fF.

In der Vergrößerung erkennt man, dass der durch die eingebrachte Ladung verursachte Spannungssprung am zu messenden Kondensator C_x bei einer unbeschalteten Ladungspumpe ($C_x = 0$) größer ist als bei einer Beschaltung mit $C_x = 10$ fF. In Bild 2.17 ist der durch die zusätzliche Ladung verursachte relative Fehler für drei verschiedene Transistorgrößen (NMOS und PMOS) zu sehen, d.h. die Kapazitätsdifferenz bezogen auf die zu messende Kapazität C_x (Anstiegs-/Abfallszeit der Taktsignale V_{1n} und V_{2p} beträgt 1 ns). Der Fehler ist bei kleinen Transistoren weniger gravierend, da der leitende Kanal kleiner ist und damit weniger Ladungsträger beinhaltet, die in die angeschlossenen Knoten abgegeben werden. Mit steigender Kapazität C_x nimmt der Fehler wie zu erkennen ist ab, da die Ladungsmenge bezogen auf den zu messenden Kondensator anteilig immer unbedeutender wird, der absolute Fehler nimmt jedoch zu.

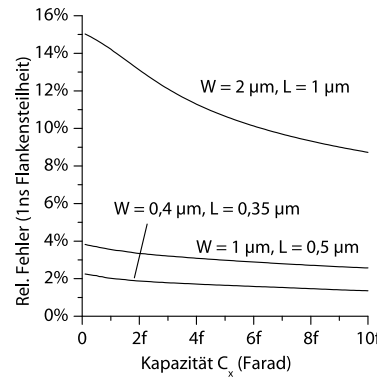


Bild 2.17. Relativer Fehler (bezogen auf C_x) in Abhängigkeit von C_x.

Die Ursache für diesen Fehler liegt jedoch nicht in dem in Bild 2.16 gezeigten Spannungssprung, da die in Gleichung 2.41 vorkommende Spannung nicht der Spannung entspricht, die am Kondensator erreicht wird, *nachdem* der PMOS-Transistor abgeschaltet wird, sondern der Spannung, mit der er aufgeladen wird. Es ist somit die Spannung ausschlaggebend, die *vor* dem Abschalten erreicht wird.

Die aus dem Kanal des PMOS-Transistors in die angeschlossenen Knoten eingebrachte Ladung ist indes sehr wohl für den Fehler der gemessenen Kapazität verantwortlich, nämlich durch Beeinflussung der Strombilanz I_{net} auf der Seite der Spannungsquelle (Source-Seite). Sie führt zu einem kleinen in die Quelle zurückfließenden mittleren Strom. Dieser zurückfließende Strom ist dabei abhängig von der Größe des Kondensators C_x, der an der gegenüberliegenden Seite des Transistors (Drain) angeschlossen ist. Die Beeinflussung findet offenbar durch den Kanal des sich im Abschalten befindlichen Transistors statt. Bild 2.18 zeigt diesen Strom bei einer Flankensteilheit von 1 Nanosekunde.

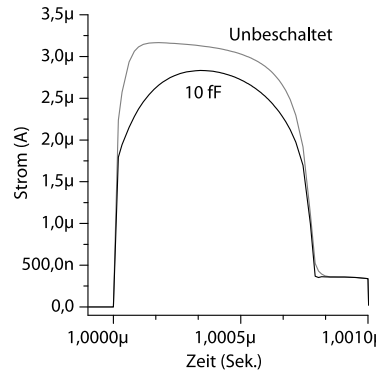


Bild 2.18. In die Spannungsquelle zurückfließender Strom, der auf Ladungsträger aus dem Kanal des PMOS-Transistors zurückgeht. Die Ladungsmenge ist von der Kondensatorgröße abhängig.

Eine Reihe weiterer Simulationen der Ladungspumpe wurde durchgeführt, aus der die Ladungsdifferenz zwischen der mit C_x beschalteten Ladungspumpe und der unbeschalteten Ladungspumpe bestimmt wurde. Trägt man diese in Abhängigkeit von C_x und der Flankensteilheit (genauer Anstiegs- und Abfallszeit) auf, so ergibt sich folgendes Bild (Bild 2.19):

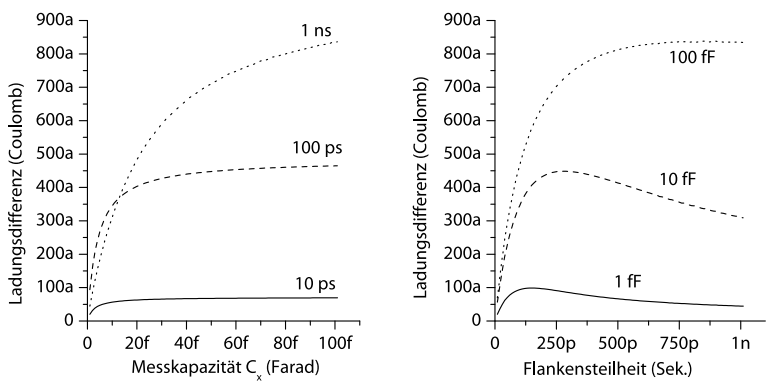


Bild 2.19. Ladungsdifferenz zwischen der unbeschalteten Ladungspumpe (Referenz) und der mit der zu messenden Kapazität beschalteten Ladungspumpe als Funktion der Messkapazität (links) und der Anstiegs-/Abfallszeit (rechts).

Die Abhängigkeit der Ladung (bzw. des gemessenen mittleren Stroms) von C_x ist im linken Graphen deutlich zu erkennen, sie variiert mit der Flankensteilheit der Schaltimpulse (V_{1n} und V_{2p}).

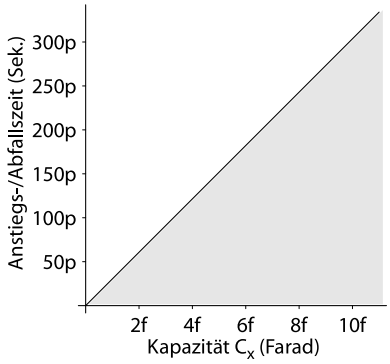


Bild 2.20. Kriterium für die Charakterisierung der Steuersignale der Ladungspumpe als „schnell“. Die Gerade entstammt Gl. 9 bzw. Gl. 10 in Sheu et al. 1984, das Signal gilt als „schnell“, falls es weit darunter liegt.

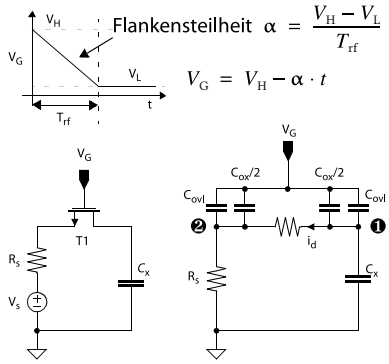


Bild 2.21. Sample & Hold-Glied (links). Der Transistor wird im Kleinsignalmodell durch ein „lumped model“ ersetzt (rechts).

Eine wohlbekannte Regel[10] für das Abschalten von MOS-Schaltern postuliert jedoch die gleichen Ladungsmengen für Source und Drain, *falls* der Transistor schnell genug abgeschaltet wird. Nur im Falle des langsamen Abschaltens beeinflussen sich Drain und Source und die jeweils abgegebene Ladungsmenge ist vom Verhältnis der angeschlossenen Kapazitäten abhängig. Da in das Kriterium für „schnell“ auch die zu messende Kapazität mit eingeht, können im Femtofarad-Bereich Anstiegs- bzw. Abfallszeiten von wenigen Picosekunden erforderlich sein, um das Kriterium zu erfüllen, bei typischen Parametern (aus Tabelle 2.7) ergibt sich die Situation in Bild 2.20. Zu beachten ist, dass die Anstiegs- bzw. Abfallszeit der Steuersignale *sehr viel* kleiner sein muss, als die gezeichnete Gerade.

Aufgrund dieses Zusammenhangs soll nun im Folgenden untersucht werden, wie Flankensteilheit, Messkapazität und Fehler zusammenhängen. Hierzu ist es erforderlich, zunächst den einfachsten denkbaren Fall eines sog. Abtast-Halteglieds („sample & hold“) zu betrachten und daraus Rückschlüsse auf die vorliegende Schaltung mit den Ladungspumpen zu ziehen.

LADUNGSUMVERTEILUNG BEI S&H-GLIEDERN. In analogen „switched-capacitor“ Schaltungen, insbesondere bei Analog-Digital Wandlern, kommt das Abtast-Halteglied häufig vor. Es dient dazu, die abzutastende Spannung über einen als Schalter fungierenden Transistor auf einen Speicherkondensator zu übertragen, um sie der weiteren Signalverarbeitung zugängig zu machen. Es gibt dabei also zwei Phasen:

1. Abtastphase. Der Transistor ist im leitenden Zustand, die Eingangsspannung wird auf den Speicherkondensator übertragen.
2. Haltephase. Der Transistor ist abgeschaltet, die Spannung wird vom Speicherkondensator gehalten.

Beim Übergang von Phase 1 zu Phase 2 wird durch das Abschalten des Transistors T1 Ladung aus dem Kanal in die Source- und Drain-Gebiete abgegeben, so dass die Spannung auf dem Speicherkondensator C_x nicht exakt dem gewünschten Wert der Spannungsquelle V_s entspricht (Bild 2.21, links). Da dieser Fehler in vielen Anwendungen die Genauigkeit der Gesamtschaltung limitiert, wurden in der Vergangenheit einige Versuche unternommen, das Phänomen mathematisch zu beschreiben, u.a. in Sheu & Hu 1984, Shieh, Patil & Sheu 1987 und Wegmann et al. 1987.

Die ersten beiden Publikationen haben für die Berechnungen denselben Ausgangspunkt, der auch für die Untersuchung des Fehlers bei der Ladungspumpe geeignet ist. Statt der Herleitung des Spannungsfehlers am Kondensator C_x soll im Folgenden ein Ausdruck für die Ladung hergeleitet, nämlich jene Ladung, die vom Kanal des Transistors in die Spannungsquelle abgegeben wird und dadurch einen Fehlerstrom verursacht, der die Auflösung der Ladungspumpe begrenzt.

10. Die Regel geht auf Arbeiten von Sheu et al. 1984, Shieh et al. 1987 und Wegmann et al. 1987 zurück, siehe folgenden Abschnitt.

Die Nomenklatur[11] (Tabelle 2.6) orientiert sich weitgehend an den Originalpublikationen. Die Herleitung beginnt am Anfang, da sich in beide Publikationen (obwohl Jahre dazwischen liegen) derselbe Fehler[12] eingeschlichen hat. Zunächst wird der Transistor T1 im Schaltbild (Bild 2.21, links) durch ein einfaches Modell ersetzt, das im wesentlichen aus den Überlappkapazitäten C_{ovl} der Source- und Drain-Anschlüsse und der Gate-Kapazität C_{ox} besteht. Letztere wird aufgeteilt und zu gleichen Teilen jeweils Source und Drain zugeordnet. Der Kanal wird durch einen Widerstand ersetzt. Es ergibt sich der Ersatzschaltplan in Bild 2.21 rechts. Dieses als „lumped model" bezeichnete Konzept wird in Sheu et al. 1984, Anhang I, analytisch hergeleitet.

Im Ersatzschaltbild kann man nun nach Kirchhoffs Knotenregel die Stromgleichungen für die Punkte 1 und 2 aufstellen:

$$C_{\mathrm{x}}\frac{dv_1}{dt} = -i_{\mathrm{d}} + \left(C_{\mathrm{ovl}} + \frac{C_{\mathrm{ox}}}{2}\right)\frac{d}{dt}(V_{\mathrm{G}} - v_1) \tag{2.42}$$

für Punkt 1 und

$$\frac{v_2}{R_{\mathrm{s}}} = i_{\mathrm{d}} + \left(C_{\mathrm{ovl}} + \frac{C_{\mathrm{ox}}}{2}\right)\frac{d}{dt}(V_{\mathrm{G}} - v_2) \tag{2.43}$$

für Punkt 2.

In diesen Gleichungen stellen v_2 und v_1 Kleinsignalgrößen dar, sie repräsentieren die Fehlerspannungen. Unmittelbar vor dem Abschalten sind v_2 und v_1 Null, da die Spannung an den Punkten 1 und 2 V_{s} entspricht, die Drain-Source Spannung beträgt also Null. Der Transistor befindet sich somit im linearen Bereich und der Strom i_{d} ist gegeben durch:

$$i_{\mathrm{d}} = \beta(V_{\mathrm{HT}} - \alpha t)(v_1 - v_2) \qquad t = 0\ldots(V_{\mathrm{HT}}/\alpha) \tag{2.44}$$

wobei β der Technologiefaktor (Transkonduktanz) des Transistors ist, und V_{HT} als Abkürzung dient:

$$\beta = \mu C_{\mathrm{ox}}\frac{W}{L} \quad \text{und} \quad V_{\mathrm{HT}} = V_{\mathrm{H}} - V_{\mathrm{s}} - V_{\mathrm{T}} \tag{2.45}$$

Der Einfachheit wegen soll angenommen werden, dass die Steuerspannung V_{G} am Gate des Transistors einer Rampenfunktion entspricht, die vom Zeitpunkt Null an linear abnimmt. Der Transistor ist also zu Beginn leitend und wird in der Zeit T_{rf} abgeschaltet:

$$V_{\mathrm{G}}(t) = V_{\mathrm{H}} - \frac{V_{\mathrm{H}} - V_{\mathrm{L}}}{T_{\mathrm{rf}}} \cdot t = V_{\mathrm{H}} - \alpha t \tag{2.46}$$

Durch Differentiation und Einsetzen in Gleichung 2.42 und 2.43 erhält man unter der Annahme $|dV_{\mathrm{G}}/dt| \gg |dv_1/dt|$ und $|dV_{\mathrm{G}}/dt| \gg |dv_2/dt|$:

$$C_{\mathrm{x}}\frac{dv_1}{dt} = -\beta(V_{\mathrm{HT}} - \alpha t)(v_1 - v_2) - \left(C_{\mathrm{ovl}} + \frac{C_{\mathrm{ox}}}{2}\right)\alpha \tag{2.47}$$

Variable	Bedeutung
β	Konduktanzkoeffizient
C_{ox}	Gatekapazität
C_{ovl}	Source- bzw. Drain-Überlappkapazität
C_{x}	Speicherkondensator
L	Effektive Kanallänge
μ	Ladungsträgermobilität
R_{s}	Innenwiderstand der Spannungsquelle
α	Flankensteilheit von V_{G}
V_{G}	Steuerspannung am Gate
V_{H}	Maximum der Steuersp.
V_{L}	Minimum der Steuersp.
V_{s}	Eingangsspannung der Quelle
V_{T}	Effektive Schwellenspannung
v_1	Fehlerspannung an Punkt 1
v_2	Fehlerspannung an Punkt 2
W	Transistorweite

Tabelle 2.6. Nomenklatur für die folgenden Rechnungen.

11. Die Zuordnung der Transistoranschlüsse zu Source und Drain hängt vom Vorzeichen von V_{DS} ab. Bei $V_{\mathrm{DS}} = 0$ sind Drain und Source nicht eindeutig definiert, so dass hier o. B. d. A. Punkt 1 als Drain und Punkt 2 als Source angenommen werden können.
12. In Sheu et al. 1984, Gl. 6 und Shieh et al. 1987, Gl. 5 vermutlich Druckfehler; in beiden Arbeiten ein Vorzeichenfehler beim Aufstellen der Differenzialgleichung.

Variable	Größe	Einheit	Art
β	198µ	A/V²	fest
C_{ox}	0.69f	Farad	fest
C_{ovl}	44a	Farad	fest
µ	29µ	m²/Vs	fest
C_x	1f, 10f, 100f	Farad	frei
W	0.55µ	Meter	frei
L	0.29µ	Meter	frei
R_s	0	Ohm	frei
U	3.3G	Volt/Sek.	frei
V_H	3.3	Volt	frei
V_L	0	Volt	frei
V_s	1.5	Volt	frei
V_T	855m	Volt	frei

Tabelle 2.7. Technologieparameter des AMS 0,35 µm Prozesses (aus „design rule manual“, oben) und Designparameter (unten), wie sie für die Rechnungen gewählt wurden. Die Werte der ersten vier Zeilen wurden aufgrund einer Verschwiegenheitsvereinbarung mit AMS geändert.

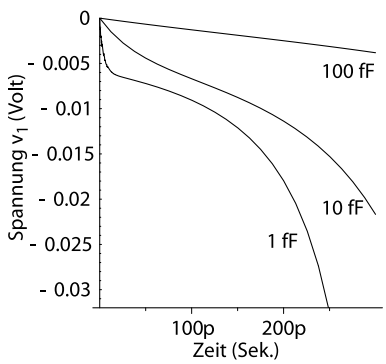


Bild 2.22. Von Mathematica auf Basis der analytischen Lösung (Gleichung 2.52) berechneter Spannungsverlauf von v_1 für drei verschiedene Größen des Speicherkondensators C_x.

$$\frac{v_2}{R_s} = \beta(V_{HT} - \alpha t)(v_1 - v_2) - \left(C_{ovl} + \frac{C_{ox}}{2}\right)\alpha \tag{2.48}$$

Gleichung 2.48 kann man nun nach v_2 auflösen:

$$v_2 = \frac{R_s(2C_{ovl}\alpha + C_{ox}\alpha + 2\beta v_1(\alpha t - V_{HT}))}{-2 + 2\beta R_s(\alpha t - V_{HT})} \tag{2.49}$$

Eingesetzt in Gleichung 2.48 erhält man eine Differentialgleichung erster Ordnung:

$$C_x\frac{dv_1}{dt} = -\frac{\beta(V_{HT} - \alpha t)}{1 + \beta R_s(V_{HT} - \alpha t)}v_1 - \left(C_{ovl} + \frac{C_{ox}}{2}\right) \cdot \left(2 - \frac{1}{1 + \beta R_s(V_{HT} - \alpha t)}\right)\alpha \tag{2.50}$$

Diese Differentialgleichung vereinfacht sich unter der Annahme, dass der Innenwiderstand R_s der Stromquelle (Source-Meter!) Null ist:

$$\frac{dv_1}{dt} = -\beta(V_{HT} - \alpha t)v_1 - \left(C_{ovl} + \frac{C_{ox}}{2}\right)\frac{\alpha}{C_x} \tag{2.51}$$

Die Lösung dieser Differentialgleichung wurde durch Einsatz der Mathematiksoftware „Mathematica“ der Firma „Wolfram Research“ gefunden:

$$v_1(t) = -\frac{(2C_{ovl} + C_{ox})}{2\sqrt{\beta C_x}}e^{\frac{(V_{HT} - \alpha t)^2\beta}{2\alpha C_x}}\sqrt{\alpha\pi/2} \cdot \left[\mathrm{Erf}\left(\sqrt{\frac{\beta}{2\alpha C_x}}(\alpha t - V_{HT})\right) + \mathrm{Erf}\left(\sqrt{\frac{\beta}{2\alpha C_x}}V_{HT}\right)\right] \tag{2.52}$$

In Box 2.4 wird exemplarisch gezeigt, wie Mathematica eingesetzt wurde, um dem Leser den Einstieg bei Bedarf zu erleichtern. In Tabelle 2.7 sind darüber hinaus die Werte aller Parameter (Transistor- u. Betriebsparameter) aufgelistet, die für die Berechnungen der Schaubilder verwendet wurden.In Bild 2.22 ist v_1 über die Zeit aufgetragen, vom Beginn des Abschaltvorgangs bis zum Erreichen der Schwellenspannung des Transistors. Die Anstiegs- bzw. Abfallszeit der Steuerspannung V_G beträgt 1 Nanosekunde. Man erkennt, dass sich offensichtlich eine *kapazitätsabhängige* Potentialbarriere für die Kanalladung aufbaut, so dass diese stärker zum Punkt 2 drängt.

Die nun folgenden Ausführungen sind nicht mehr Teil der genannten Originalpublikation oder (nach bestem Wissen des Autors) irgend einer anderen Publikation, vor allen Dingen wird in keiner Abhandlung über die Verbesserung von Ladungspumpen hinsichtlich des durch Ladungsumverteilung verursachten Fehlers bei Kapazitätsmessungen die Ursache des Fehlers genauer untersucht bzw. analytisch hergeleitet. Die Autoren begnügen sich mit dem Hinweis auf Ladungsinjektion (gemeint ist die Ladungsumverteilung) als Fehlergrund.

Der Ausdruck v_s/R_s in Gleichung 2.48 repräsentiert den gesuchten Strom, der aus der Spannungsquelle bzw. in sie hinein fließt (je nach Vorzeichen), so dass v_d aus Gleichung nur noch in Gleichung 2.48 eingesetzt werden muss:

Box 2.4 Mathematica

Mit den folgenden Anweisungen kann die Differenzialgleichung 2.50 analytisch mit Mathematica gelöst werden. Die vom Benutzer eingegegebenen Zeilen werden vom Editor automatisch mit „In[1]:=“, „In[2]:=“ usw. markiert, entsprechend der Reihenfolge der Eingabe. Falls eine Anweisung eine Ausgabe bewirkt, stellt Mathematica jeweils ein „Out[x]“ voran, wobei x der Zeilennummer der Eingabe entspricht.

```
In[1]:=myEq := V1'[t] == -\[Beta]/Cx * (Vht - \[Alpha]*t) *
       V1[t] - (Cov + Cox/2) * \[Alpha]/Cx
```

Durch diese Zeile wird die Differentialgleichung definiert, das Hochkomma bewirkt die Ableitung des vorangehenden Ausdrucks. Durch die Zeilenfolge „\[Beta]“ wird der griechische Buchstabe Beta erzeugt (alternativ per Mausklick).

```
In[2]:=myEqSol := DSolve[{myEq, V1[0] == 0}, V1[t], t]
```

„DSolve“ löst die Differentialgleichung analytisch per Integration unter der Randbedingung (Anfangswert) „V1[0]==0“. Durch den Doppelpunkt wird die Berechnung noch nicht ausgeführt, vielmehr wird „myEqSol“ der Ausdruck nach dem Gleichheitszeichen zugewiesen.

```
In[3]:=myV1[t_] = V1[t] /. First[Simplify[myEqSol]]
```

```
Out[3]:= -1/(2 Sqrt[Cx] Sqrt[β]) ( (2 Cov + Cox) e^((Vht - t α)^2 β)/(2 Cx α) Sqrt[π/2] Sqrt[α] ( Erf[ Vht Sqrt[β] / (Sqrt[2] Sqrt[Cx] Sqrt[α]) ] + Erf[ (-Vht + t α) Sqrt[β] / (Sqrt[2] Sqrt[Cx] Sqrt[α]) ] ) )
```

Die rechte Seite des Ausdrucks wird evaluiert und der linken Seite zugewiesen. Der Rückgabewert von „DSolve“ ist eine Liste, in der die Lösung an erster Stelle steht, auf sie wird mit „First“ zugegriffen. Der Lösungsausdruck x wird dabei in der Form „V1[t]->x“ zurückgegeben, nämlich einer Transformationsvorschrift („V1[t]“ wird zu „x“). Der „/.“-Operator wendet diese Vorschrift dann auf alle passenden Teile der linken Seite an, ersetzt also „V1[t]“ durch den Lösungsterm.

```
In[4]:=myIs[t_, Cx_] = \[Beta]*(Vht - \[Alpha]*t)*myV1[t] -
       (Cov + Cox/2)*\[Alpha]
```

```
Out[4]:= -(Cov + Cox/2) α - 1/(2 Sqrt[Cx]) ( (2 Cov + Cox) e^((Vht - t α)^2 β)/(2 Cx α) Sqrt[π/2] Sqrt[α]
          (Vht - t α) Sqrt[β] ( Erf[ Vht Sqrt[β] / (Sqrt[2] Sqrt[Cx] Sqrt[α]) ] + Erf[ (-Vht + t α) Sqrt[β] / (Sqrt[2] Sqrt[Cx] Sqrt[α]) ] ) )
```

Definiert die Gleichung für den Strom und evaluiert diese. Der Ausdruck „myV1[t]“ wird dabei durch die Lösung der Differentialgleichung ersetzt.

```
In[5]:=Cov = 4.4 10^-17; Cox = 6.9 10^-16;
```

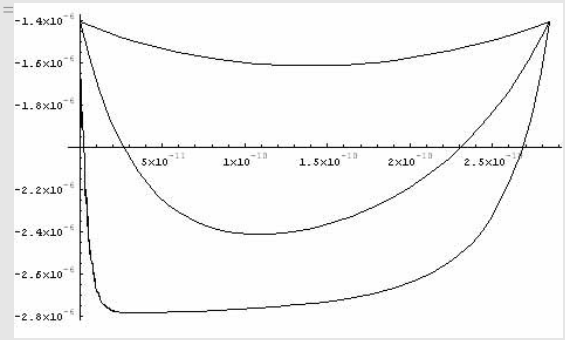
```
In[6]:=\[Beta] = 1.98 10^-4; Vht = 0.94; \[Alpha] := 3.3/Trf;
```

```
In[7]:=Tvt = Vht/U; Trf = 1 10^-9;
```

Definiert einige Konstanten zur numerischen Berechnung (siehe auch Tabelle 2.7).

```
In[8]:= plot1 = Plot[{myIs[t, 1 10^-15], myIs[t, 10 10^-15],
        myIs[t, 100 10^-15]}, {t, 0, Tvt}];

Out[8]=
```



$$i_s = \frac{1}{4}(2C_{ov} + C_{ox})\sqrt{\alpha}\left\{-2\sqrt{\alpha} + \frac{1}{\sqrt{C_x}}e^{\frac{\beta(V_{HT}-\alpha t)^2}{2\alpha C_x}}\sqrt{2\pi}\right. \tag{2.53}$$

$$\left.(\alpha t - V_{HT})\sqrt{\beta}\left[\mathrm{Erf}\left(\frac{(\alpha t - V_{HT})\sqrt{\beta}}{\sqrt{2\alpha C_x}}\right) + \mathrm{Erf}\left(\frac{V_{HT}\sqrt{\beta}}{\sqrt{2\alpha C_x}}\right)\right]\right\}$$

In Bild 2.23 ist i_s für drei Werte von C_x vom Beginn des Abschaltens bis zum Erreichen der Schwellenspannung bei $t' = V_{HT}/\alpha$ gezeichnet (Anstiegs-/Abfallzeit 1 Nanosekunde). Es ergibt sich eine charakteristische Kurve, die vor allem bei $C_x = 1$ fF eigentümlich erscheint.

Zum Zeitpunkt Null erkennt man einen Sprung von $i_s = 0$ zu $i_s = -1.5\ldots-2$ µA , der aus dem Abschneiden der Kurve bei negativen Werten herrührt. Man beachte, dass die hier diskutierten Gleichungen *nur* für den Zeitbereich von Null bis t' ihre Gültigkeit haben, so dass der Sprung eine direkte Folge aus dem eingeschränkten Definitionsbereich ist. Für Werte von $t > V_{HT}/\alpha$ existieren keine freie Ladungsträger mehr im Kanal, so dass zwischen den beiden Anschlüssen des Transistors keine gegenseitige Beeinflussung mehr stattfindet[13]. Damit besteht auch keine Abhängigkeit des Stroms von der zu messenden Kapazität C_x ab diesem Zeitpunkt. Der Strom rechts von diesem Punkt (nicht gezeigt) wird nur noch von der Ladungsinjektion durch die Überlappkapazitäten C_{ov} hervorgerufen (unabhängig von C_x).

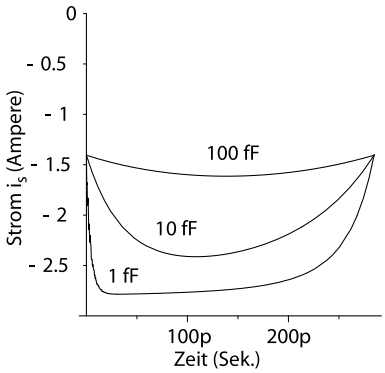


Bild 2.23. Fehlerstrom i_s der Stromquelle aus Mathematica (Punkt 2, Bild 2.21) für drei Werte von C_x.

13. Diese Aussage gilt nicht uneingeschränkt. In der Realität gibt es keine abrupten Übergänge an den Grenzen des Definitionsbereiches, zum anderen existiert ein i.d.R. kleiner sog. „sub-threshold current“, also Strom, der unterhalb der Schaltschwelle fließt. Beide Effekte haben nur geringen Einfluss und sollen deshalb vernachlässigt werden.

Die jeweilige Fläche unter den Kurven in Bild 2.23 repräsentiert schließlich jene Ladung, die den Fehler im Nettostrombudget bei der Kapazitätsmessung verursacht. Genauer gesagt ist es die Ladungs*differenz* zwischen der unbeschalteten Ladungspumpe und der mit C_x beschalteten, die für den Fehler verantwortlich ist.

Um einen Ausdruck für diesen Fehler zu erhalten, muss zunächst Gleichung 2.53 integriert werden, und zwar von Null bis zum Erreichen der Schwelle. Da der Strom negativ ist, zusätzlich unter Anwendung der Betragsfunktion:

$$Q_s(C_x) = \int_0^{t'} |i_s| = \frac{2C_{ov} + C_{ox}}{4\sqrt{\beta}} \left[4V_{HT}\sqrt{\beta} - \sqrt{2\pi\alpha C_x}\,\mathrm{Erf}\left(\frac{V_{HT}\sqrt{\beta}}{\sqrt{2\alpha C_x}} \right) \right] \quad (2.54)$$

In Bild 2.24 ist diese Gleichung als Funktion der Abfallszeit[14] der Steuerspannung V_G gezeichnet, die Berechnung wurde wieder mit Mathematica durchgeführt. In Bild 2.25 ist ebenfalls Gleichung 2.54 abgebildet, diesmal als Funktion der zu messenden Kapazität C_x. Man erkennt in beiden Abbildungen die starke Abhängigkeit von C_x für kleine Werte. Geht C_x gegen Null, so strebt der Kehrwert gegen Unendlich und es folgt:

$$\lim_{C_x \to 0} \mathrm{Erf}\left(\frac{V_{HT}\sqrt{\beta}}{\sqrt{2\alpha C_x}} \right) = 1 \quad \Rightarrow \quad \lim_{C_x \to 0} Q_s(C_x) = (2C_{ov} + C_{ox})V_{HT} \quad (2.55)$$

Der rechte Ausdruck in Gleichung 2.55 stellt damit die Ladung dar, die bei einem unbeschalteten Sample&Hold-Glied (S&H-Glied) in die Stromquelle zurückfließt (gesamte Kanalladung plus Ladungsinjektion bis zum Erreichen der Schwellenspannung). Der Einfachheit halber sei diese Ladung mit Q_{ref} abgekürzt. Bei Verwendung der in Tabelle 2.7 gegebenen Werte ergibt sich für Q_{ref} so eine Ladung von ca. 800 Attocoulomb.

Die Ladungsdifferenz zwischen dem „offenen" S&H-Glied und dem Abtast-Halteglied mit C_x gibt schließlich den absoluten Fehler an. Setzt man diese Differenz in Beziehung zur Gesamtladung, die innerhalb eines S&H-Zyklusses aus der Stromquelle geflossen ist, so ergibt sich zu guter Letzt ein Ausdruck für den relativen Fehler:

$$Er = \frac{Q_{ref} - Q_s(C_x)}{V_s C_x} = \frac{(2C_{ov} + C_{ox})\sqrt{\alpha\pi/2}}{2V_s\sqrt{\beta C_x}}\,\mathrm{Erf}\left(\frac{V_{HT}\sqrt{\beta}}{\sqrt{2\beta C_x}} \right) \quad (2.56)$$

In Bild 2.26 ist zu sehen, wie der relative Fehler mit steigender Kapazität C_x abnimmt. Der Grund liegt darin, dass die in die Stromquelle zurückfließende Ladung (siehe Bild 2.25) mit zunehmendem C_x gegen einen Grenzwert strebt, die durch den Abtastvorgang geflossene Gesamtladung jedoch linear mit C_x steigt. Im Verhältnis zu dieser Gesamtladung wird der absolute Ladungsfehler also immer unbedeutender. Den Grenzwert für $C_x \to \infty$ erhält man durch Reihenentwicklung der Fehlerfunktion:

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)n!} \Rightarrow \lim_{C_x \to \infty} Q_s(C_x) = \frac{1}{2}(2C_{ov} + C_{ox})V_{HT} \quad (2.57)$$

14. Die Abfallszeit steht mit der Flankensteilheit über den Ausdruck $\alpha = (V_H - V_L)/T_{rf}$ in Beziehung, siehe auch Bild 2.21

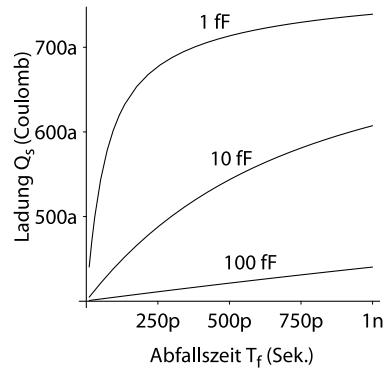


Bild 2.24. Die zurückfließende Ladung Q_s aus Gleichung 2.54 als Funktion der Abfallzeit der Steuerspannung (für drei Werte von C_x). Für die Berechnung in Mathematica wurden die Werte aus Tabelle 2.7 benutzt.

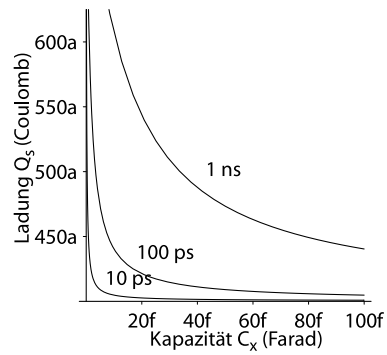


Bild 2.25. Die zurückfließende Ladung Q_s aus Gleichung 2.54 in Abhängigkeit von der zu messenden Kapazität C_x.

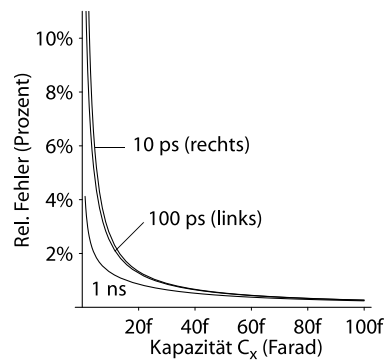


Bild 2.26. Ladungsdifferenz zwischen unbeschaltetem und dem mit C_x beschalteten Abtast-Halteglied, normiert auf die Gesamtladung eines S&H-Zyklusses (Ergebnis aus Mathematica).

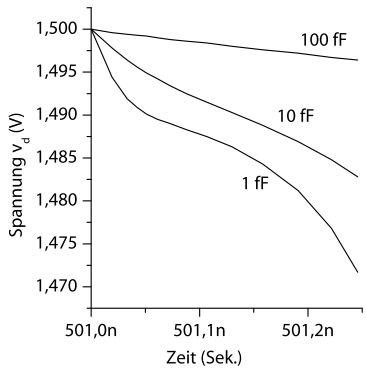


Bild 2.27. Simulierter Spannungsverlauf von v_d für drei Werte von C_x. Bild 2.22 auf Seite 50 stellt die entsprechende analytische Lösung dar.

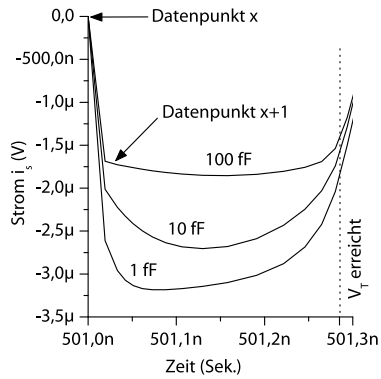


Bild 2.28. Simulierter Verlauf des Fehlerstroms i_s für drei Werte von C_x zum Vergleich mit Bild 2.23 auf Seite 52.

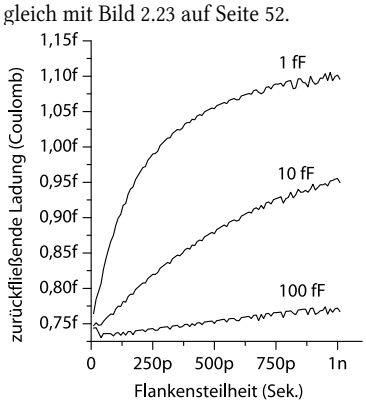


Bild 2.29. Aus der Simulation stammender Verlauf der Ladungsmenge, die in die Quelle V_s zurückfließt (zweiter Parameter C_x). Vgl. hierzu Bild 2.24 auf Seite 53.

ANALYTISCHE LÖSUNG VERSUS SIMULATION. Die bisherigen analytischen Berechnungen zum Einfluss der Kanalladungsumverteilung bei Abtast-Haltegliedern wurden auch mit einem Schaltkreissimulator numerisch verifiziert. Zum Einsatz kam wieder Spectre in Kombination mit den Transistormodellen der 0,35 µm Technologie von AMS.

In Bild 2.27 ist der simulierte Spannungsverlauf v_d zu sehen, also die Spannung an jenem Transistoranschluss, an dem der Speicherkondensator C_x angeschlossen ist. Gezeichnet wurde der Verlauf für drei Werte von C_x vom Beginn des Abschaltvorgangs bei 501 Nanosekunden bis zum Erreichen der Schwellenspannung (Wert wurde analytisch bestimmt). Strenggenommen müsste die Spannung statt v_d einen anderen Namen bekommen, da sie nicht wie in Bild 2.22 der *Fehler*spannung entspricht, sondern der tatsächlich an diesem Knoten anliegenden Spannung (entspricht zu Beginn V_s, hier 1,5 Volt). Man erkennt eine gute Übereinstimmung.

Der für die Ladungspumpe relevante Fehlerstrom i_s zurück in die Quelle V_s ist in Bild 2.28 zu erkennen. Wieder wurden drei Werte für C_x eingesetzt, der gezeichnete Bereich beginnt wieder mit dem Start des Abschaltvorgangs. Die analytisch berechnete Schwellenspannung von 855 Millivolt wurde zum Zeitpunkt erreicht, der mit der gestrichelten Linie gekennzeichnet ist. Dieser Zeitpunkt scheint nicht ganz zur Simulation zu passen, die Kurven konvergieren erst ein Stück weit rechts der Linie. Grund für die Differenz der Schaubilder ist vermutlich der Einfluss der Sperrschichtkapazitäten, d.h. die in der analytischen Rechnung nicht modellierte, spannungsabhängige Kapazität der Source- bzw. Draindioden.

Am linken Rand des Graphen ist ein scharfer Übergang zwischen den beiden mit Pfeilen markierten Punkten zu erkennen. Ein solches Verhalten ist im analytischen Graphen nicht vorhanden. Der Grund liegt einerseits in der großen Schrittweite der Simulation bzw. dem Fehlen eines Zwischenwertes und zum anderen im unstetigen Übergang vom Gleichgewichtszustand ($i_s = 0$) zum Bereich, der durch das analytische Modell abgedeckt ist.

In Bild 2.29 ist nun die in die Quelle V_s zurückfließende Ladung abgebildet wie sie aus der Simulation stammt. Der qualitative Verlauf entspricht zwar dem durch Gleichung 2.54 analytischen beschrieben bzw. in Bild 2.24 gezeigten Verlauf, weist jedoch größenmäßig andere Werte auf. Der Grund hierfür liegt im größen Integrationsbereich, der in der Simulation verwendet wurde. Statt nur die Ladung zu ermitteln, die vom Beginn des Abschaltens bis zum Erreichen der Schwellenspannung geflossen ist, wurde auch die Ladung berücksichtigt, die unterhalb der Schwellenspannung bis zum Erreichen der Spannung $V_G = V_L = 0$ abgegeben wurde. Der Integrationsbereich wurde bis zu diesem Punkt ausgedehnt, da dieser Zeitpunkt einfach zu ermitteln war, während die Bestimmung der Schwellenspannung in jeweils getrennten Simulationen hätte erfolgen müssen und anschließend umständlich manuell in die eigentliche Simulation hätte eingetragen werden müssen. Die im erweiterten Integrationszeitraum geflossene Ladung stellt jedoch nur eine Verschiebung (Offset) der Kurven in Bild 2.24 nach oben dar. Die fehlende quantitative Übereinstimmung wird ohnehin auch durch andere Faktoren beeinflusst, wie zu sehen sein wird.

Trägt man den Verlauf der zurückfließenden Ladung wie in Bild 2.25 über die Kapazität C_x auf, so ergibt sich aus der Simulation das in Bild 2.30 gezeigte Verhalten. Der Offset ist freilich wieder vorhanden, er kann ignoriert werden. Für Werte unter 10 Femtofarad nimmt die Simulation einen wesent-

lich flacheren Verlauf als die analytische Rechnung. Auch für dieses Phänomen lassen sich mögliche Gründe nennen. Erstens reicht bei sehr kleinen Kapazitäten das ursprünglich angesetzte „lumped model“ nicht mehr aus. Zahlreiche „second-order effects“ der Transistoren, also Effekte von normalerweise untergeordneter Bedeutung, scheinen durch eine Verstärkungswirkung sichtbar zu werden, bleiben jedoch unberücksichtigt. Und zweitens wird die Diodenkapazität des mit C_x verbundenen Transistoranschlusses (Punkt 1 in Bild 2.21) in der Rechnung nicht gesondert berücksichtigt, sondern C_x beaufschlagt. In der Simulation dagegen sind beide getrennt, so dass bei $C_x = 0$ eine Restkapazität von einigen Femtofarad übrigbleibt, die von der Sperrschichtkapazität des Transistoranschlusses herrührt. Damit ist C_x in der Simulation um diesen Kapazitätsbetrag höher als in der Rechnung.

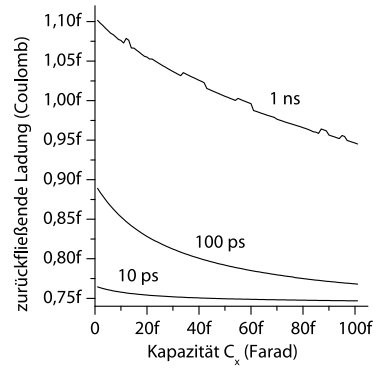


Bild 2.30. Ladungsmenge, die in die Quelle zurückfließt, als Funktion von C_x für drei Werte der Abfallszeit. Bild 2.25 auf Seite 53 ist die analytische Entsprechung.

Konsequenzen für die Ladungspumpe. Ausgangspunkt der vorangehenden Betrachtungen war die Frage nach dem durch Kanalladungsumverteilung bedingten Fehler bei dem in Bild 2.14 gezeigten Schaltungsprinzip der Ladungspumpe zur Kapazitätsmessung. Die Reduktion auf den einfacheren Fall des Abtast-Halteglieds ermöglichte eine analytische Herangehensweise und zeigte eine qualitative Übereinstimmung mit der Simulation. Die Anwendbarkeit der Ergebnisse auf den ursprünglichen Fall der Ladungspumpe ergibt sich durch folgende Argumentation:

Die Ansteuerung der Ladungspumpe in Bild 2.14 über die Signale V_{1n} und V_{2p} geschieht in zwei Phasen, die sich fortlaufend abwechseln. In der ersten Phase wird der untere NMOS-Transistor über die Eingangsspannung V_{1n} in den leitenden Zustand versetzt, um den Messkondensator C_x zu entladen. In der zweiten Phase wird er über die Steuerspannung V_{2p} und den PMOS-Transistor auf das Potential der Spannungsquelle gebracht. Diese beiden Phasen sind in Bild 2.31 nochmal deutlicher abgebildet. Für jede der beiden Phasen kann die Ladungspumpe durch das Ersatzschaltbild im jeweils unteren Teil der Abbildung substituiert werden, da der grau unterlegte Teil inaktiv bzw. ohne schaltungstechnische Relevanz ist.

Zusammenfassend kann gesagt werden, dass der Gesamtfehler, der durch die Ladungsumverteilung der beiden Transistoren in der Ladungspumpe verursacht wird, Ladungsträger aus dem Kanal des PMOS-Transistors und des NMOS-Transistors umfasst, die sich über die gemeinsame Masse als Fehlerstrom bemerkbar machen. Der durch die Umverteilung der Kanalladung bedingte *relative* Fehler bei der Kapazitätsmessung macht sich in erster Linie bei kleinen Kapazitäten (unter 20 fF) bemerkbar und geht bei steigender Kapazität gegen Null. Der Anteil der Ladung, der als Fehlerstrom im Messgerät sichtbar wird, ist abhängig von der zu messenden Kapazität, da sich durch die Spannung, die sich über der Kapazität aufgrund der Ladungsumverteilung aufbaut, ein Rückkoppelungseffekt durch den Kanal hindurch ergibt, der die Ladungsabgabe aus dem Kanal beeinflusst. Diese Beeinflussung geschieht bei kleinen Kapazitäten hinreichend schnell, da geringe Ladungsmengen ausreichen, um das erforderliche Potentialgefälle aufzubauen, und zwar während weitere Ladungen aus dem „schrumpfenden“ Kanal an den energetisch günstigeren Anschluss wandern.

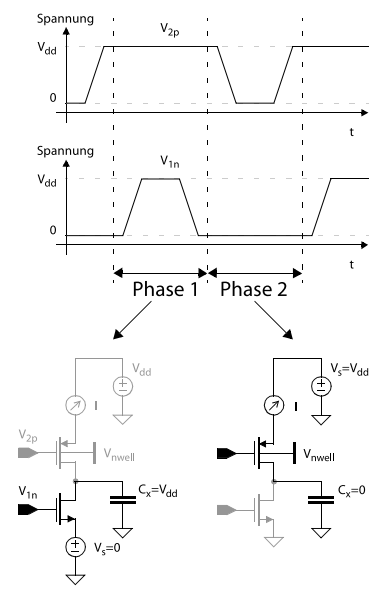


Bild 2.31. Ladungspumpe als Kombination zweier spezieller S&H-Glieder.

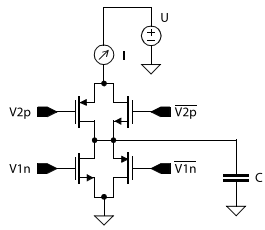


Bild 2.32. Ladungspumpe mit „Pass-gate“ Schaltern zur Kompensation der Kanalladung.

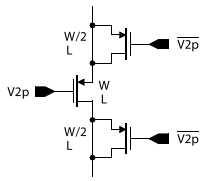


Bild 2.33. MOS-Schalter mit Dummy Switches zur Minimierung der Ladungsinjektion und -um verteilung.

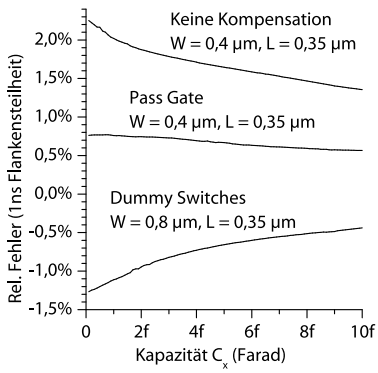


Bild 2.34. Vergleich der Lösungsansätze zur Kompensation der Kanalladungsumverteilung bei Ladungspumpen.

Schaltungstechnische Verbesserungen

Eine Reihe von Verbesserungen der Ladungspumpe ist in der Literatur vorgeschlagen worden. Dazu gehören die Arbeiten von Froment et al. 1999, Brambilla et al. 2003 und Chang et al. 2004. In der Publikation von Brambilla et al. 2003 wird eine modifizierte Version der Ladungspumpe vorgestellt, die den durch Ladungsumverteilung bedingten Fehler verringert. Die Lösung des Problems wird durch Einsatz von sog. „Pass-Gate“ Schaltern erreicht. Dabei handelt es sich um Paare von NMOS- und PMOS-Transistoren, die parallelgeschaltet werden und möglichst die gleichen geometrischen Eigenschaften aufweisen. Die aus dem Kanal des einen Transistortyps injizierte Ladung wird so durch die Kanalladung des jeweils anderen Typs kompensiert, im Idealfall sogar aufgehoben. Bild 2.32 zeigt den schematischen Aufbau der Schaltung.

Diese Vorgehensweise ist keineswegs neu, die Verwendung von Pass-Gate Schaltern ist ein bereits seit langem bekanntes Mittel zur Kompensation der Kanalladungsumverteilung von MOS-Schaltern. Es existieren sogar noch bessere Verfahren, den Fehler zu minimieren, da die Pass-Gate Schalter nur unter der Annahme gut funktionieren, dass die geometrischen Eigenschaften der beiden Transistortypen gleich sind. In der Realität werden sich die effektive Länge und Weite der Transistoren aufgrund der Schwankungen des Herstellungsprozesses (Stichwort „matching“, siehe “Prozessstreuung und Mismatch” auf Seite 22) gerade zwischen NMOS- und PMOS-Transistoren unterscheiden. Dadurch sind auch die abgegebenen Ladungsmengen der beiden Typen verschieden, so dass keine vollständige Kompensation der Ladungsabgabe erfolgt.

Statt der Verwendung von Pass-Gate Schaltern wird häufig Gebrauch von sog. „dummy switches“ gemacht, d.h. zusätzlichen, mehr oder weniger inaktiven Transistoren desselben Typs. An die beiden Anschlüsse des aktiven Schalters wird jeweils ein Transistor desselben Typs mit kurzgeschlossener Source und Drain angeschlossen. Die Weite dieser beiden Transistoren beträgt jeweils die Hälfte des eigentlichen Schalters. Angesteuert werden sie über das komplementäre Steuersignal. In Bild 2.33 ist ein solcher Schalter gezeigt.

Die Verwendung von Dummy-Switches der halben Weite des eigentlichen Schalttransistors begründet sich in der Annahme, dass die vom Schalter abgegebene Ladung zu gleichen Teilen in die beiden Anschlüsse abgegeben wird. Da diese Annahme im Falle der Ladungspumpen bei kleinen zu messenden Kapazitäten nicht gültig ist, können auch Dummy-Switches das Problem der Ladungsumverteilung nicht vollständig lösen. Darüber hinaus ist es erforderlich, den Schalttransistor doppelt so groß zu machen, wie die Entwurfsregeln des Herstellungsprozesses für die kleinsten Transistoren („minimum size“) erlauben. Dadurch ist die Kanalfläche größer und damit die Menge der abgegebenen Ladung, als bei Pass-Gate Schaltern.

Vergleicht man die beiden Lösungsansätze hinsichtlich des Messfehlers, der sich bei der Messung einer bekannten Kapazität durch die Simulation einer Ladungspumpe mit Spectre/Spice ergibt, so erhält man das Schaubild in Bild 2.34. Man erkennt, dass die Lösung mit Pass-Gate Schaltern gegenüber der Variante mit Dummy-Switches bei kleinen Kapazitäten geringfügig besser abschneidet. Aus diesem Grund und angesichts der Verwendung dieser

Lösung in der Publikation von Brambilla et al. aus dem Jahr 2003 wurde diese Variante für alle in den folgenden Abschnitten vorgestellt Messungen verwendet.

Conclusio

Die erreichbare Auflösung bei der Kapazitätsmessung mittels Ladungspumpen (siehe Bild 2.14 auf Seite 45) wird hauptsächlich durch den Mismatch der Transistoren beschränkt. Dieser ist aufgrund seiner zufälligen Natur a priori unbekannt und kann daher nicht einfach aus dem Messergebnis herausgerechnet werden.

Im Gegensatz dazu gibt es den systematischen Effekt der Kanalladungsumverteilung, der sich mathematisch analysieren lässt und so bei hinreichender Modellierungsgenauigkeit im Endergebnis eliminieren liese. Die mathematische Herleitung zeigte, dass dieser Effekt tatsächlich für den in der Simulation sichtbaren Fehler verantwortlich ist. Statt den Fehler herauszurechnen wurde eine schaltungstechnische Lösung über Pass-Gate Transistoren vorgeschlagen, die in einem relativen Fehler von ca. 0,75 Prozent resultiert (im relevanten Messbereich), statt einem relativen Fehler von ca. 1,5 Prozent ohne eine solche Kompensation.

2.3.2 Alternative Messverfahren

Varianten der Ladungspumpe

In Chang et al. 2004 wurde eine Variante der Ladungspumpentechnik vorgestellt, die prinzipiell keine Fehler durch den Mismatch und die Ladungsinjektion und -umverteilung der Transistoren aufweist. Die Technik ist ebenso einfach und platzsparend wie die herkömmliche Ladungspumpe in Bild 2.14 auf Seite 45, d.h. sie kommt ohne Pass-Gate Schalter aus. Darüber hinaus benötigt das Verfahren keine unbeschaltete Referenzladungspumpe zur Nettostrombildung.

Aufgrund dieser Vorteile ist davon auszugehen, dass der Schaltungsvorschlag in Zukunft die Standardtechnik zur Prozesscharakterisierung und -überwachung sein wird. Die einzig problematische Einschränkung ist, dass nur Querkopplungskapazitäten („cross-coupling“) gemessen werden können. Die Kapazität eines Knotens zur Masse hingegen ist auf diese Weise nicht ermittelbar. Dem könnte entgegengehalten werden, dass die herkömmliche (und auch die durch Pass-Gates verbesserte) Ladungspumpe dafür keine Koppelkapazitäten (exakt genug) ermitteln kann. In Froment et al. 1999 wurde jedoch gezeigt, dass es hierfür eine Lösung gibt, die als „Single Pattern Driver“ (SPD) bezeichnet wird und die ursprüngliche Ladungspumpe im Wesentlichen unverändert lässt.

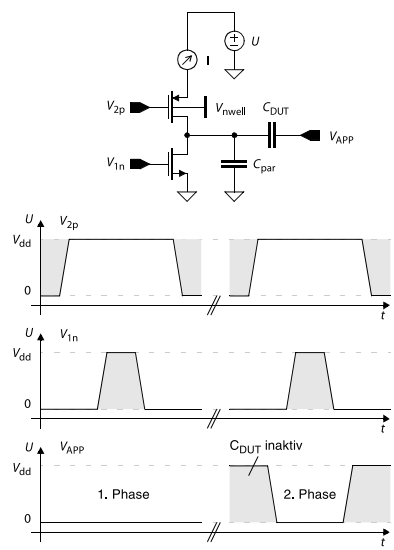


Bild 2.35. Variante der Ladungspumpe zur Messung von Querkopplungskapazitäten. In Phase 1 wird der mittlere Strom I_1 wie bei der herkömmlichen Ladungspumpe gemessen, in Phase 2 wird die zu bestimmende Kapazität C_{DUT} über die Spannung V_{APP} gleichstrommäßig deaktiviert und der Strom I_2 gemessen.

Funktionsweise. Auch beim Vorschlag von Chang et al. 2004 kommen Ladungspumpen zum Einsatz, jedoch erweitert um einen zusätzlichen Anschluss („probe pad“), der die Elektroden aller zu ermittelnden Kapazitäten auf einer Seite verbindet. In Bild 2.35 ist die Situation dargestellt: C_{DUT} soll gemessen werden, V_{APP} wird über die zusätzliche Kontaktfläche eingespeist und C_{PAR} repräsentiert die parasitäre Kapazität der Drain-Gebiete der Transistoren. Für $V_{APP} = 0$ erhält man die herkömmliche Ladungspumpe (Bild 2.14).

Statt nun wie bisher *einen* mittleren Strom pro Ladungspumpe zu messen, werden *zweimal* mittlere Ströme bei *derselben* Ladungspumpe in zwei Phasen gemessen. In der ersten Phase wird $V_{\mathrm{APP}} = 0$ gesetzt und der Strom $I_1 = f \cdot V_{\mathrm{dd}} \cdot (C_{\mathrm{DUT}} + C_{\mathrm{PAR}})$ gemessen. In der zweiten Phase wird die Messung wiederholt, diesmal mit $V_{\mathrm{APP}} = V_{\mathrm{dd}}$ während der aktiven Phase des PMOS-Transistors (graue Flächen oben). Der gemessene mittlere Strom $I_2 = C_{\mathrm{PAR}} \cdot f \cdot V_{\mathrm{dd}}$ umfasst dabei nur noch die parasitäre Kapazität C_{PAR}, da C_{DUT} keinen Spannungsabfall aufweist und somit „deaktiviert“ wurde.

Die Nettostrombildung zur Subtraktion der parasitären Kapazitäten findet nun nicht durch eine „leere“ Referenzladungspumpe statt, die dem Mismatch unterliegt, sondern über *dieselbe* Ladungspumpe wie bei der Messung der zu ermittelnden Kapazitäten. Das Strombudget liefert:

$$I_1 - I_2 = C_{\mathrm{DUT}} \cdot f \cdot V_{\mathrm{dd}} \tag{2.58}$$

Darüber hinaus wirkt sich die Ladungsinjektion und -umverteilung der Transistoren in beiden Phasen im gleichen Umfang aus, so dass sich der Effekt herausrechnet und nicht mehr als Fehler im Ergebnis bemerkbar macht. Der Grund für die gleich starke Wirkung ist, dass C_{DUT} in beiden Phasen unverändert bleibt und die Ladungsverteilung nur wechselstrommäßig beeinflusst. Die verschiedenen Elektrodenpotentiale am Pad-Anschluss von C_{DUT} wirken sich als reiner Gleichstromanteil („DC bias“) aus.

Kapazitiv arbeitende Sensoren

Eine ganze Fülle an weiteren Schaltungstechniken zur Kapazitätsmessung ergibt sich, wenn das Einsatzgebiet keine solch extrem genauen Messungen bei zudem kleinsten Kapazitäten im Attofarad-Bereich erfordert oder die Schaltung wesentlich größer sein darf. In der Sensorik ist dies häufig der Fall, da in der Regel nur ein Kondensator gemessen werden soll, dessen variable Kapazität als sekundärer Messwert für eine primäre sensorische Größe, z.B. Druck oder Beschleunigung, steht. Im letzten Fall werden großflächige, beweglich aufgehängte Kammstrukturen aus Metall auf dem Chip erzeugt, die wegen der Trägheit ihrer Masse auf Beschleunigung mit Auslenkungsbewegungen reagieren, die in einer veränderten Kapazität resultieren.

Solche Kapazitätsänderungen können beispielsweise über einen Kapazitäts-Frequenzwandler ermittelt werden, wie in Krummenacher 1985 vorgestellt. Ein anderer Ansatz besteht darin, an die zu messende Kapazität auf einer Seite ein Sinus- oder Rechtecksignal anzulegen, an die andere Seite einen Transimpedanz- oder ladungsempfindlichen Verstärker. Das Ausgangssignal weist (nach eventueller Demodulation) einen funktionalen jedoch in der Regel nicht-linearen Zusammenhang mit der fraglichen Kapazität auf.

* * *

Kapitel 3
Implementierung

Die Implementierung des Cluster-Konzepts und der messtechnischen Analyseverfahren beginnt mit der Erzeugung der 3D-Clusterstrukturen in Abschnitt 3.1. Zunächst wird der Unterschied zwischen den Clustern und herkömmlichen Kondensatoren erläutert und aufgelistet, welche Anforderungen ein Verfahren zum Layoutentwurf erfüllen muss, darunter die Automatisierbarkeit, Komplexität und Zufälligkeit. Danach wird ein Algorithmus vorgestellt, der diese Forderungen erfüllt. Anhand von Flussdiagrammen wird die Funktionsweise erklärt und erläutert, warum es sich um eine Variante der vielfach verwendeten Random-Walk Strategie handelt. Schließlich wird gezeigt, wie sich dieser Algorithmus mit einer speziellen Skriptsprache in der Form eines Programms für die Chip-Entwicklungsumgebung von Cadence umsetzen lässt. Die Erzeugung einer ganzen Bibliothek von Clustern, inklusive der Extraktion der (parasitären) Kapazität wird demonstriert.

In Abschnitt 3.2 wird die Vorgehensweise bei der Implementierung des experimentellen Teils der Arbeit präsentiert. Nachdem der Messaufbau in Form eines zu diesem Zweck entworfenen Testchips, einer speziellen Leiterplatte und des Mess-Equipments (Spitzenmessplatz und Source-Meter) erklärt wird, geht es im anschließenden Unterabschnitt um die praktischen Aspekte der Durchführung. So wird gezeigt, wie die Steuerung bzw. Automatisierung des Messvorgangs vonstatten ging, sowie welche Probleme bei der speziellen Form des Messablaufs auftraten und wie diese gelöst wurden. Insbesondere wird gezeigt, wie systematische Fehler herausgerechnet wurden und welche Wiederholbarkeit des Messungen erreicht wurde. Schließlich werden erste Ergebnisse präsentiert.

Im dritten Teil des Kapitels (Abschnitt 3.3) geht es um einen Schaltungsvorschlag zur integrierten Kapazitätsmessung, um damit die Cluster auszuwerten und eine digitale Schlüsselsequenz zu erzeugen. Das Schaltungsprinzip wird erklärt und die Verwandtschaft mit den Ladungspumpen aufgezeigt. Grundlegende Eigenschaften der Schaltung wie die Auflösung werden anschließend hergeleitet. Den Abschluss bildet die Erläuterung des Aufbaus des zweiten Testchips, mit dem das Funktionieren des Schaltungsvorschlags in der Praxis bewiesen wurde.

* * *

3.1 Erzeugung der 3D-Cluster

3.1.1 Einführung

Zur Erzeugung der großen Zahl an Layouts für die 3D-Kapazitätscluster wurde ein spezieller Algorithmus eingesetzt, der auf einem iterativen Zufallsverfahren basiert. Seine Aufgabe ist es, innerhalb einer vorgegebenen Fläche komplexe, dreidimensionale Verbindungsstrukturen zu erzeugen. Diese Strukturen sollen über eine möglichst ungenau bekannte oder schwer berechenbare elektrische Kapazität verfügen.

Der integrierte Kondensator

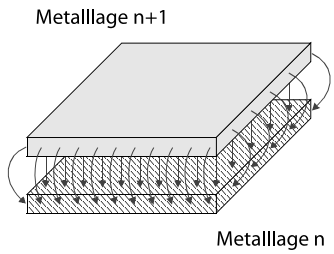


Bild 3.1. Ein einfacher Plattenkondensator auf einem Chip. Das elektrische Feld bildet sich zwischen den zwei Metalllagen aus (Pfeile).

Metallleitungen werden im analogen Schaltungsentwurf nicht nur zum Signalaustausch eingesetzt, sondern auch, um passive Bauteile wie Spulen oder Kondensatoren zu realisieren. Wohldefinierte Regeln beim geometrischen Entwurf der Metallstrukturen werden eingehalten, so dass die elektrische Kapazität bzw. Induktivität möglichst genau kontrolliert werden kann.

Ein solches Bauelement stellt der Kondensator in Bild 3.1 exemplarisch dar. Das elektrische Feld des Kondensators bildet sich zwischen den beiden übereinander liegenden Metallplatten aus. Die Entwurfsregel war in diesem Fall, beide Metalllagen gleich groß zu wählen und für eine vorgegebene Kapazität die möglichst größte Fläche bei geringstem Umfang zu wählen. Auf diese Weise wird der Kondensator für seinen typischen Einsatzzweck optimiert: Die elektrische Kapazität soll möglichst genau den Annahmen beim Entwurf und der Simulation entsprechen, damit sich das Verhalten der Schaltungen auf dem Chip innerhalb der Spezifikationsgrenzen bewegt.

Die Kapazitätscluster

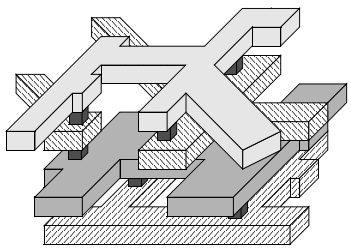


Bild 3.2. 3D-Ansicht eines typischen Kapazitätsclusters. Sein Aufbau weist eine komplexe Irregularität auf, Breite, Länge und Richtung der Metallstücke sind zufällig.

Einem völlig anderen Einsatzzweck sollen jedoch die parasitären Kapazitätscluster dienen. Zwar wird auch hier die elektrische Kapazität schaltungstechnisch ausgenutzt, jedoch soll ihr genauer Wert für einen Außenstehenden möglichst schwer ermittelbar sein. Dadurch soll das Verhalten der Schaltung in starker Weise unvorhersehbar und letztlich uneinsehbar gemacht werden.

Für diesen Zweck wird absichtlich gegen die Entwurfsregeln für Kondensatoren verstoßen, so dass es sich bei den Kapazitätsclustern nicht mehr um „herkömmliche“ Kondensatoren handelt. Es kommen keine Metallplatten mehr zum Einsatz, sondern eine Vielzahl mehr oder weniger dünner Metallleitungen, die zu einer komplexen, irregulären und zufälligen Struktur zusammengesetzt sind. Bild 3.2 zeigt exemplarisch den Aufbau eines solchen Clusters. Die elektrische Kapazität ist im Normalfall schaltungstechnisch ungewollt oder sogar nachteilig, man spricht daher von der parasitären Kapazität.

Eben diese parasitäre Kapazität von Metallleitungen soll innerhalb der Kapazitätscluster dem Einsatzzweck entsprechend optimiert werden und schaltungstechnisch nutzbar gemacht werden. Da eine große Zahl an Clustern benötigt werden, soll der Entwurf automatisiert werden. Hierzu wurde der Random-Walk Algorithmus entwickelt und in der Skriptsprache SKILL umgesetzt (siehe Abschnitt 3.1.4).

3.1.2 Anforderungen

Vollständige Automatisierbarkeit

Als wichtiges Kriterium für den Einsatz der parasitären Kapazitätscluster ist zunächst die vollständige Automatisierbarkeit des Entwurfsvorganges zu nennen. Die Erstellung des geometrischen Aufbaus (Maskenlayout) jedes einzelnen Clusters soll ohne wiederholten Eingriff oder manuelle Steuerung möglich sein. Die einzige Form der Benutzerinteraktion besteht in der Festlegung von gewissen Startparametern oder Einstellungen, die nur zu Beginn des automatischen Entwurfvorgangs vorgenommen werden.

Der Grund für diese Anforderung liegt in der hohen Zahl an Clustern, die für einen geheimen Schlüssel mit einer realistischen Anzahl Bits vonnöten sind: In der Regel werden 200 Bit Entropie und mehr benötigt. Da für jedes zusätzliche Bit je nach Schaltungsvariante bis zu zwei weitere Cluster benötigt werden, ist die Gesamtzahl an Clustern bereits so hoch, dass die manuelle Erstellung des Layouts jedes einzelnen Clusters aus Zeitaufwandsgründen ausscheidet.

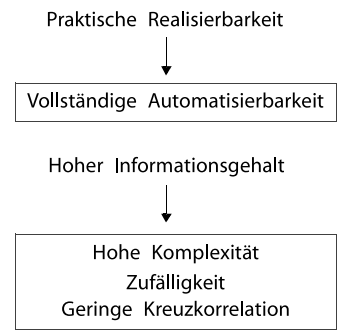


Bild 3.3. Anforderungsprofil an die parasitären Kapazitätscluster.

Hohe Komplexität

Das Erfordernis der Komplexität ergibt sich direkt aus der wichtigsten Eigenschaft der Kapazitätscluster: Der hohe Grad an Informationsgehalt[15] jedes einzelnen Clusters. Diese informationstheoretische Aussage bedeutet, in einfache Worte übersetzt, dass jeder Cluster ein hohes Maß an unbekannten Informationen in sich tragen soll, in diesem Fall ist dies die elektrische Kapazität. Je weniger bekannt ist über den genauen Wert, desto mehr Informationen sind in ihm enthalten.

Eben diese Information stellt die Grundlage für die Erzeugung des digitalen Fingerabdrucks mit der in dieser Arbeit beschriebenen schaltungstechnischen Realisierung dar. Die in den Clustern enthaltene Information stellt den geheimen Schlüssel in „Rohform“ dar. Wäre die elektrische Kapazität aller Cluster eines Chips für Außenstehende mit hoher Genauigkeit bekannt, so könnte der geheime Schlüssel daraus abgeleitet werden. Diesen Fall stellt der einfache Plattenkondensator in Bild 3.1 dar. Seine Kapazität ist aufgrund des einfachen Aufbaus sehr exakt berechenbar. Ein Schlüssel-Chip, der nur mit solchen einfachen Plattenkondensatoren aufgebaut wäre, hätte nur einen sehr geringen Informationsgehalt: Jeder Kondensator gerade soviel Information wie der tatsächliche Kapazitätswert vom berechneten Wert abweicht.

Zufälligkeit

Ein möglichst zufälliger struktureller Aufbau der Kapazitätscluster ist ebenfalls aus Gründen des Informationsgehaltes nötig. Die Zufälligkeit stellt sicher, dass kein systematischer „Bias“ vorhanden ist, d.h. keine Bevorzugung bestimmter Strukturen oder regelmäßiger Muster. Im Idealfall ist die Wahrscheinlichkeit für das Auftreten einer bestimmten Struktur gleich wahrscheinlich wie für das Auftreten einer völlig anderen Struktur.

15. An dieser Stelle sei auf den Abschnitt „Die Entropie“ verwiesen, in dem die Rolle des Informationsgehaltes und der Zusammenhang mit Wahrscheinlichkeiten in der Kryptographie behandelt wird.

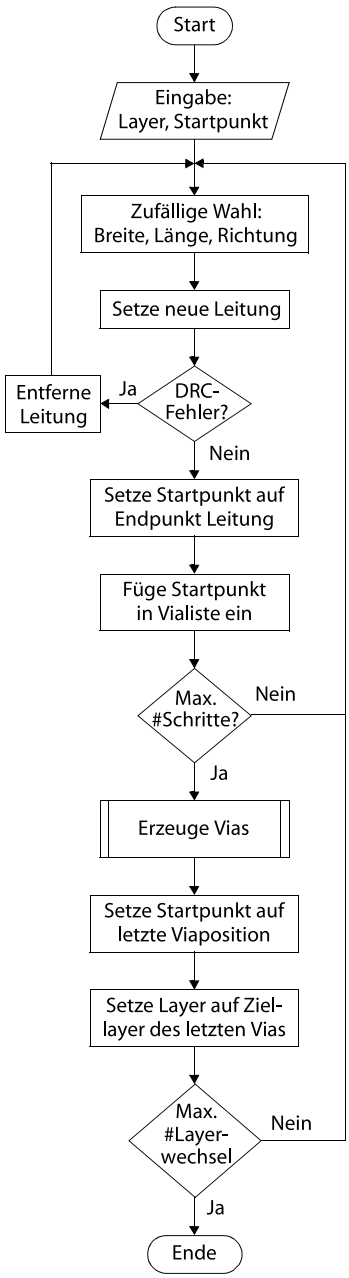


Bild 3.4. Der Random-Walk Algorithmus. Zentraler Bestandteil ist die zufällige Auswahl von Parametern und der DRC-Check, der Test auf Gültigkeit.

Diese Gleichverteilung kann sich freilich nur innerhalb der Grenzen bewegen, die durch Flächenvorgabe, schaltungstechnischen Einsatz oder Prozesstechnischen Einschränkungen vorgegeben sind - kurz alle Randbedingungen, die von vornherein bekannt sind.

Geringe Kreuzkorrelation

Als dritte Einflussgröße für den Informationsgehalt der Kapazitätscluster ist der Verwandsschaftsgrad von jeweils zwei beliebigen Clustern auf einem Chip zu nennen, d.h. die Kreuzkorrelation. Eine geringe Korrelation bedeutet, dass man nicht von einem Cluster auf die anderen schließen kann, d.h. dass kein Angreifer aus der Kenntnis der Kapazität eines Clusters Informationen über die Kapazität der anderen Cluster ableiten kann.

3.1.3 Der Random-Walk Algorithmus

Die Layouterzeugung der Kapazitätscluster wurde rechnergestützt automatisiert. Dazu wurde ein Algorithmus entwickelt, der alle genannten Forderungen hinreichend erfüllt. Der Algorithmus basiert auf einem iterativen „trial-and-error“ Verfahren, bei dem zufällig Leiterbahnen und Durchkontaktierungen gesetzt werden, die anschließend auf Verletzungen der Entwurfsregeln (DRC-Fehler) überprüft werden. Ergibt sich ein Fehler, so wird die letzte Änderung rückgängig gemacht und eine andere Variante ausprobiert.

Erzeugung der Leiterbahnen

Der Algorithmus beginnt mit der Erzeugung der Leiterbahnen auf einer bestimmten Metallisierungsebene und an einem vorgegebenen Startpunkt. Das Flussdiagramm in Bild 3.4 zeigt den funktionellen Ablauf. Der Startpunkt und die Startebene sind für den Anschluss des Kapazitätsclusters an die Auswerteelektronik nötig, die aus dem Kapazitätswert der Cluster die Bits des Schlüssels generiert.

Der nächste Schritt steht am Anfang jeder Iteration des Algorithmus: Die zufällige Wahl geeigneter Parameter für die Erzeugung eines Metallstückes. Hierunter fällt die Breite, Länge und Richtung der Leiterbahn. Gewisse Mindestbreiten und -längen, sowie die Beschränkung auf Winkel von 45 Grad sind dabei prozesstechnisch gedingt vorgegeben. Bei der zufälligen Auswahl der Parameter werden dabei idealerweise diese prozesstechnischen Vorgaben („design-rules“) berücksichtigt, um die Fehlerwahrscheinlichkeit bei der späteren Abstandsregelprüfung zu minimieren.

Diese Prüfung auf Regelverletzung, „design-rule check“ (DRC) genannt, wird nach jedem Neusetzen eines Leitungsstückes durchgeführt. Im Falle eines Regelverstoßes wird die Leitung entfernt und eine andere Parameterkombination getestet, der Algorithmus kehrt zurück zur Parameterwahl. Wurde ein Metallstück fehlerfrei gesetzt, so repräsentiert das Ende der Leitung den Anfangspunkt für das nächste Metallstück, entsprechend wird der Startpunkt für die nächste Leitung neu gesetzt.

Diese Start- und Endpunkte stellen gleichzeitig Positionen dar, an denen ein Wechsel der Metallisierungsebene über eine Durchkontaktierung (Via) nach oben oder unten möglich ist. Sie werden deshalb in eine spezielle Vialiste eingetragen, die bei der Erzeugung der Vias in einer Unterroutine verwendet wird. Diese Unterroutine wird immer dann aufgerufen, wenn eine

Metallisierungsebene komplett - bezogen auf die Zielfläche - abgearbeitet wurde, d.h. mit Leitungsstücken zufälliger Größe und Orientierung gefüllt wurde. Die Fläche gilt dann als gefüllt, wenn die maximale Anzahl an Schritten ausgeführt wurde bzw. die maximale Anzahl Metallstücke erzeugt wurde.

Nach der Erzeugung der Vias verfügt der aktuelle Layer über ein oder mehrere Durchkontaktierungen nach oben oder unten. Eines dieser Vias stellt somit einen geeigneten Ausgangspunkt für die nächste Metallisierungsebene dar, auf der erneut Metallstücke auf die selbe Art und Weise erzeugt werden sollen. Aus diesem Grund wird die neue Startposition auf die Koordinaten eines der Vias gesetzt, z.B. das letzte erzeugte Via. Ebenso wird die Metalllage, zur der die Durchkontaktierung wechselt, als neue Startebene gesetzt.

Die Anzahl der Wechsel der Metallisierungsebene kann über die Angabe eines Maximalwertes gesteuert werden. Er muss nicht gleich der Anzahl der zur Verfügung stehenden Metalllagen eines Prozesses sein, sondern sollte darüber liegen. Dadurch wird sichergestellt, dass der Algorithmus mit der Generierung von Metallstücken zu einem Layer zurückkehrt, auf dem bereits Metallstücke erzeugt wurden. Auf diese Weise wird verhindert, dass der Algorithmus nur Strukturen erzeugt, die in der Vertikalen keine Richtungsänderung aufweisen, d.h. einem Turm oder Stapel gleich aufgebaut sind.

Wurde die maximale Anzahl an Lagenwechsel also noch nicht erreicht, so kehrt der Algorithmus an den Anfang des Programmflusses zurück, um im akutellen Layer erneut Metallstücke zu generieren. Der Algorithmus wird beendet, wenn die Maximalzahl erreicht wurde.

Setzen der Vias

Die Erzeugung der Durchkontaktierungen ist in Bild 3.5 schematisch dargestellt. Der Routine wird zu Beginn der aktuelle Layer mitgeteilt, die maximale Anzahl zu erzeugender Vias festgelegt und die Liste mit gültigen Viapositionen übergeben.

Die darauf folgenden Schritte befinden sich im Inneren einer Programmschleife, die solange ausgeführt wird, bis die maximale Anzahl an erzeugten Durchkontaktierungen erreicht wurde. Zunächst wird der erste Eintrag in der Vialiste aus der Liste entfernt, er stellt die Position des nächsten zu erzeugenden Vias dar. Das Via kann sodann einen Kontakt zur nächsten, darüberliegenden Metalllage herstellen, oder zur darunterliegenden Ebene. Dies ist freilich nur möglich, wenn es sich nicht um die unterste Metallisierungsebene handelt, in diesem Fall ist nur ein Wechsel nach oben möglich. Das gleiche gilt in umgekehrter Weise für die oberste Lage.

Aus diesem Grund überprüft der Algorithmus die aktuelle Metalllage und entscheidet, ob das zu erzeugende Via eine Durchkontaktierung nach oben oder nach unten darstellen soll. Sind beide Richtungen möglich, so wird mit gleicher Wahrscheinlichkeit zufällig eine der beiden Möglichkeiten gewählt.

Daraufhin wird das Via erzeugt und der Abstandsregelprüfung unterzogen. Besteht die Änderung den Test, so wurde das Via regelkonform erzeugt. Wurde die maximale Anzahl an Durchkontaktierungen noch nicht erreicht, so springt der Algorithmus an den Beginn der Schleife zurück. Die Bearbeitung des nächsten Element der Vialiste beginnt, d.h. ein weiteres Via wird erzeugt.

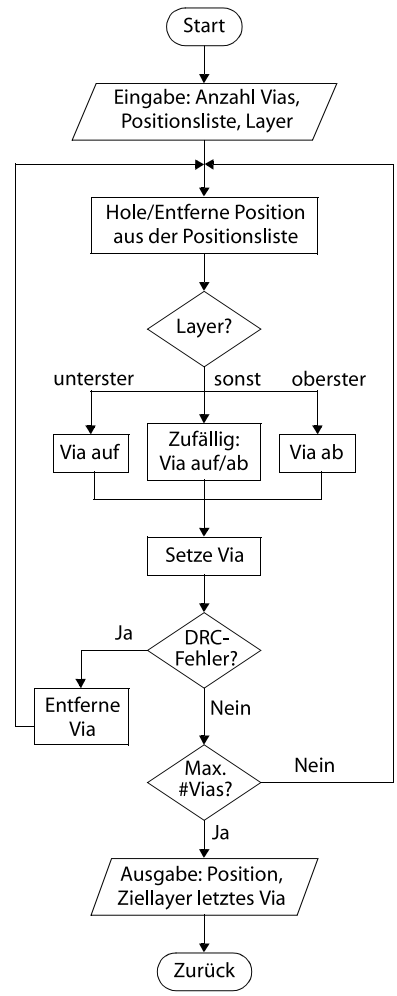


Bild 3.5. Die Generierung von Durchkontaktierungen beim Random-Walk Algorithmus. Auch hier spielt der DRC-Check die zentrale Rolle der Gültigkeitsprüfung.

War die Generierung des Vias nicht regelkonform, beispielsweise weil der DRC-Check eine Verletzung des Mindestabstands zwischen dem Via und einem benachbarten Metallstück festgestellt hat, so wird das soeben erzeugte Via entfernt. Der Programmfluss kehrt nun ebenfalls an den Schleifenanfang zurück und bearbeitet die nächste Position in der Positionsliste.

3.1.4 EDA-Umsetzung

WERKZEUGAUTOMATION DURCH SKRIPTSPRACHEN. Der im letzten Abschnitt vorgestellte Algorithmus zur Erzeugung der parasitären Kapazitätscluster stellt wie jeder Algorithmus lediglich ein Handlungsrezept dar, das prinzipiell auch von Hand durch einen Anwender ausgeführt werden könnte. Doch die Entwicklung eines formellen Algorithmus findet meist mit dem Ziel der Automatisierung und insbesondere Beschleunigung statt, indem der Algorithmus für die Abarbeitung auf einem Rechner entworfen wird.

Gewöhnlicherweise werden Algorithmen also in konkrete Befehlsfolgen überführt, die wiederum durch einen Compiler in einen lauffähigen Programmcode übersetzt werden. Das kompilierte Programm, der sog. Binärcode, läuft dann typischerweise direkt auf dem Prozessor des Rechners, von gelegentlichen Sprüngen ins Betriebssystem zur Nutzung von Standardroutinen und von Systemfunktionen abgesehen.

Ein völlig anderes Verfahren wird bei den Skriptsprachen verwendet. Die Befehlsfolgen werden nicht mehr komplett und einmalig übersetzt, sondern während der Laufzeit und sukzessive durch den sogenannten Interpreter[16]. Häufig ist er Teil eines größeren Programmgerüsts („framework“) und bietet direkten Zugriff auf seine speziellen Funktionen. Damit ist der Anwender in der Lage, einzelne Bedienungsschritte zu automatisieren und den Bedarf an Interaktionen zu minimieren.

Diese Eigenschaft der Programmierbarkeit mittels Skriptsprachen weisen die meisten modernen rechnergestützen Entwicklungswerkzeuge auf, insbesondere jene der Electronic Design Automation, den Softwarewerkzeugen für die Entwicklung elektronischer Produkte. Dabei orientieren sich die Programmkonstrukte und sprachlichen Merkmale häufig an den Vorgaben der TCL-Spezifikation („Tool Command Language“), einer Definition zur Vereinheitlichung der Syntax von anwendungsbezogenen Skriptsprachen. Kennzeichnend für solche auf TCL basierenden Entwicklungswerkzeuge ist die syntaktische Einfachheit, die Bereitstellung spezifischer Spezialkommandos und die Mächtigkeit der Befehle. Zum letzten Punkt gehört die Verfügbarkeit von Listen und Operationen zur Manipulation und zum Durchlauf von Mengen und Listen.

Die Cadence Virtuoso Custom Design Plattform und SKILL

Die Firma Cadence ist als eine der drei Marktführer[17] für EDA-Software in Fachkreisen bestens bekannt. Die Virtuoso Custom Design Plattform von Cadence stellt eine Reihe von Werkzeugen für den analogen Schaltungsentwurf

16. Bei manchen Skriptsprachen können die Programme zur Geschwindigkeitsoptimierung optional ebenfalls kompiliert werden.
17. Cadence Design Systems Inc., Mentor Graphics Inc., Synopsys Inc.

von Chips zur Verfügung, mit Schwerpunktsetzung auf der manuellen Eingabe von Layoutdaten und hohem Maß an Freiheitsgraden bei Entwurfsdetails („full-custom design“).

Zur Automatisierung von Standardaufgaben und zur Lösung anwendungsspezifischer Aufgaben verfügt die Virtuoso Plattform über eine in Eigenregie entwickelte Skriptsprache mit dem Namen SKILL. Ihre Syntax ist einfach zu erlernen (Orientierung an den Programmiersprachen C und Lisp), der Befehlsumfang ist reichhaltig - die Sprache stellt eine Vielzahl von Listen- und Stringoperationen zur Verfügung - und bietet einen direkten Zugriff auf die speziellen Funktionen der Plattform zum analogen Schaltungsentwurf.

Layoutspezifische Kommandos. Zur Umsetzung des Generierungsverfahrens für die parasitären Kapazitätscluster kamen fast ausschließlich Kommandos für den Layout-Editor der Entwicklungsplattform zum Einsatz. Diese Kommandos bieten einen direkten Zugriff auf alle geometrischen Objekte des Layouts wie Metallbahnen, Vias, dotierte und polykristalline Bereiche und deren Eigenschaften. Die Befehle erlauben es, gezielt einzelne geometrische Formen oder Gruppen davon in ihrer Länge, Breite, Richtung usw. zu ändern oder den Vorgaben entsprechend zu erzeugen.

Tabelle 3.1 gibt einen Überblick über die zentralen Befehle in Virtuoso, wie sie zur Umsetzung der Flussdiagramme in Bild 3.4 und Bild 3.5 verwendet wurden. So wurde der Punkt „Setze neue Leitung“ im Diagramm über das erste Kommando realisiert, der Test auf Abstandsregelverletzungen über den zweiten Befehl. Im Falle eines DRC-Fehlers wurden die letzten beiden Anweisungen eingesetzt, um die letzte Änderung, d.h. das fehlerhafte Leitungsstück, zu selektieren und zu entfernen. Die Erzeugung der Vias in Bild 3.5 lies sich schließlich über das Kommando in der dritten Zeile verwirklichen.

Kommando	Funktion
dbCreatePath	Leitung erzeugen
ivDRC	DRC durchführen
leCreateContact	Via erzeugen
deGetEditCellView	Aktuelles Fenster
geSelectObject	Objekt auswählen
dbDeleteObject	Objekt löschen

Tabelle 3.1. Grundlegende SKILL-Kommandos zur Layouterzeugung in Virtuoso.

Die Fülle der Einsatzmöglichkeiten bzw. Mächtigkeit der in Tabelle 3.1 aufgelisteten Befehle bedeutet auch, dass eine Vielzahl von Parametern beim Aufruf der Kommandos übergeben werden müssen. So werden typischerweise Positionsangaben in Form von Koordinaten und Angaben zum verwendeten Layer gemacht. Hinzu kommt in der Regel die Auswahl von speziellen, kommandospezifischen Optionen.

Ausführung, Ausgabe und Weiterverarbeitung

Die Generierung eines Kapazitätsclusters typischer Größe (5 auf 10 µm) mit dem hier vorgestellten Verfahren und bei Umsetzung durch SKILL auf der Virtuoso-Plattform ist in der Regel nach einer Laufzeit von 2-3 Minuten abgeschlossen, je nach Rechenleistung und Wahl der Parameter. Werden beim Programmstart beispielsweise sehr viele Lagenwechsel und zu erzeugende Vias vorgegeben, so werden häufiger Abstandsregeln verletzt, d.h. der Algorithmus muss häufiger alternative Leitungsparameter auf Gültigkeit testen. Dementsprechend länger ist die Laufzeit.

Nach der Erzeugung eines Clusters liegt seine Geometrie in Form einer zweidimensionalen Entwurfsansicht im Layout-Editor vor. Bild 3.7 zeigt eine solche Ansicht (hier schwarz-weiß, die einzelnen Ebenen werden im Editor farblich unterschieden). Der Cluster wird als eigenständige Entwurfseinheit (Zelle) in einer Bibliothek abgespeichert und einer Kapazitätsanalyse (Extraktion) unterzogen (ausführlich in Abschnitt 4.1). Zu diesem Zweck und zur

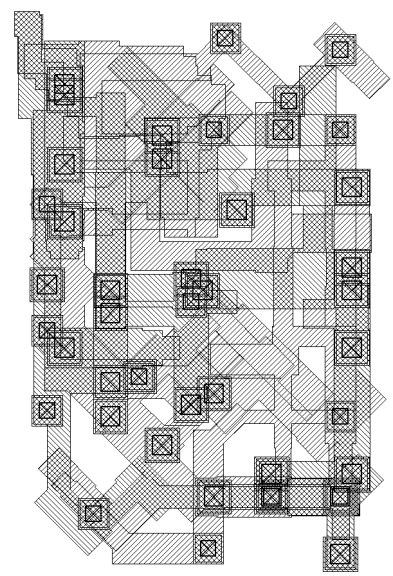


Bild 3.7. Entwurfsansicht eines parasitären Kapazitätsclusters. Zu sehen sind Metallleitungen auf den Ebenen 1-4, sowie Leitungen aus Polysilizium und Durchkontaktierungen (Vias).

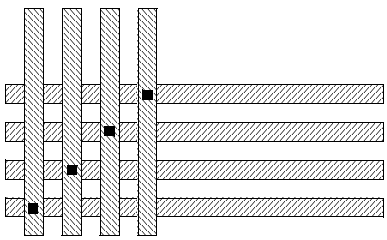


Bild 3.6. Automatisch erzeugter Leitungsbus mit Stichleitungen zu nebenstehendem Programm (Ausschnitt).

Box 3.1 Exemplarisches Programmfragment in SKILL

Im Folgenden wird ein kurzes Beispiel eines bereits lauffähigen Programmfragmentes in SKILL vorgestellt. Es kann direkt in die Konsole („Command Interpreter Window", CIW) der Entwicklungsplattform eingegeben werden und erfordert nur ein bereits geöffnetes Fenster des Layout-Editors:

```
SpacingH = 3 LengthH = 260 WidthH = 1
SpacingV = 2 LengthV = 200 WidthV = 1
Origin = 0:0 BusWidth = 64

for( n 0 BusWidth-1

  SrcV = xCoord(Origin)+n*SpacingV : yCoord(Origin)
  SrcH = xCoord(Origin) : yCoord(Origin)+n*SpacingH
  DestV = xCoord(SrcV) : yCoord(SrcV)+LengthV
  DestH = xCoord(SrcH)+LengthH : yCoord(SrcH)

  dbCreatePath( deGetEditCellView() list("MET1" "drawing")
                list(SrcH DestH) WidthH "extendExtend")
  dbCreatePath( deGetEditCellView() list("MET2" "drawing")
                list(SrcV DestV) WidthV "extendExtend")
  leCreateContact( deGetEditCellView() "VIA1_C"
                   list(xCoord(SrcV) yCoord(SrcH)) "R0"
                   0.5 0.5 1 1 0.9 0.9 "left" "bottom" )
)
```

Die Befehlsfolge erzeugt das Layout eines auf der untersten Verdrahtungsebene (MET1) horizontal verlaufenden 64-Bit Leitungsbusses (siehe Bild 3.6), von dem auf der zweiten Ebene (MET2) in vertikaler Richtung Stichleitung abgehen. Die beiden Leitungen sind über Vias (VIA1_C) verbunden. Außer der Angabe von Längen, Breiten und Abständen werden Parameter (z.B. 'extendExtend') angegeben, die hier von sekundärer Bedeutung sind.

Bild 3.8. 3D-Ansicht des Clusters aus Bild 3.7, die Kapazität beträgt ca. 17 fF (prozessabhängig). Weitere Beispiele (farbig) finden sich auf Seite 147 ff.

externen Weiterverarbeitung mit Werkzeugen anderer EDA-Plattformen wird das Layout in ein Standardformat konvertiert (sog. GDSII- oder Stream-Format).

3D-VISUALISIERUNG. Für einige der erzeugten Kapazitätscluster wurde eine dreidimensionale Schrägansicht erstellt, um den strukturellen Aufbau besser erkennen zu können. Die zweidimensionale Entwurfsansicht, die typisch ist für alle Layout-Editoren, bietet bei den Kapazitätsclustern dagegen wenig Anschaulichkeit, da sie für die manuelle Eingabe geometrischer Formen gedacht ist und eher bei regelmäßigen Strukturen Vorteile aufweist.

Die 3D-Ansicht in Bild 3.8 wurde mit dem Programm POVRay erstellt, nachdem das Layout vom GDSII-Format mit einem Konvertierungsprogramm (GDS2POV) in das interne Datenformat von POVRay umgewandelt wurde. POVRay gehört in die Klasse der Raytracing- und Rendering-Programme, mit dem dreidimensionale Szenen unter Berücksichtigung von Lichtausbreitung, Schattenwurf und Reflexion berechnet werden können. Dadurch erzeugen Licht- und Schatteneffekte einen räumlichen Tiefeneindruck, der dem Betrachter die Dreidimensionalität der betrachteten Objekte vermittelt.

Der Kapazitätscluster in Bild 3.8 (schwarz-weiß, im Original farbig) entspricht dem Cluster in Bild 3.7. In der linken, hinteren Ecke des Clusters ist die Anschlussstelle für die Auswerteelektronik zu sehen, ein kleiner, rechteckiger Bereich auf der obersten Metalllage. Neben den typischerweise horizontal und vertikal verlaufenden Strukturen erstecken sich einige Leitungsstücke in schräger Richtung. Der grundsätzliche Aufbau ist völlig willkürlich und die einzelnen Metallfragmente in ihrer Größe und Lage zufällig. Als besonderes Merkmal sind schließlich noch die kleinen Ecken und Ausbuchtungen zu nennen. Es sind in erster Linie diese letzten drei Eigenschaften der Kapazitätscluster, die sie von herkömmlichen, regelmäßigen Strukturen wie sie aus der Hand eines Ingenieurs stammen oder von Verdrahtungswerkzeugen („Router“) erzeugt werden, unterscheidet. Auf den Farbtafeln der Seiten 147 bis 151 sind weitere Cluster in ihrer 3D-Ansicht zu sehen. Weitere Details zu den Kapazitätswerten verschiedener Extraktionswerkzeuge und den von Testchips stammenden Messwerten finden sich in Tabelle 4.3 auf Seite 110, sowie in Tabelle 4.9 auf Seite 118 und Tabelle 4.10 auf Seite 121.

Erzeugung der Clusterbibliothek und parasitäre Extraktion

Neben dem Random-Walk Algorithmus zur Clustererzeugung wurde die automatische Extraktion jedes erzeugten Clusters mit den Tools Quickcap, Assura, Assura-FS, Calibre und Diva komplett in SKILL für das Cadence IC-Framework umgesetzt. Über eine eigene Benutzeroberfläche (siehe Bild 3.9) konnten die wichtigsten Parameter festgelegt werden, so z.B. die maximale Anzahl an Iterationen, die der Algorithmus pro Cluster ausführen sollte. Im Dialogfeld „Configurations“ lies sich eine von mehreren vordefinierten Strategien bezüglich der Schrittweite, Anzahl Durchkontaktierungen und Lagenwechsel auswählen.

Ein mehr der untergeordneten Information dienendes Merkmal war die Option, ein akustisches Signal bei jedem aufgetretenen DRC-Fehler auszulösen. Dadurch war es möglich, die Fehlerhäufigkeit in Abhängigkeit von den aktuellen Parametern zu überprüfen.

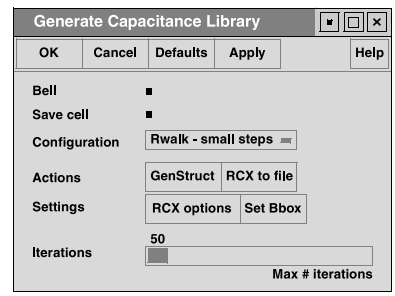


Bild 3.9. Die Benutzeroberfläche zur automatischen Erzeugung der Clusterbibliothek. Sie ist Teil des Cadence Custom IC/ Virtuoso-Gespanns und wurde ebenfalls in SKILL programmiert.

* * *

3.2 Messungen

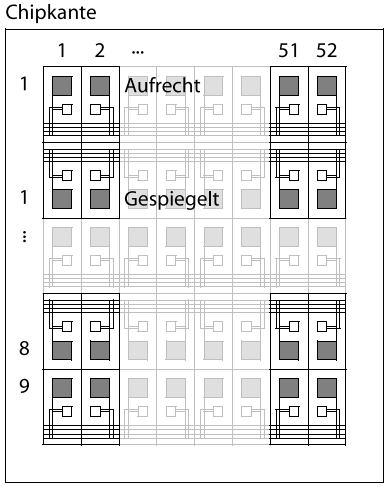


Bild 3.10. Organisation der Teststrukturen auf dem Testchip in Form einer Matrix zur Messung der Kapazität mittels Ladungspumpen.

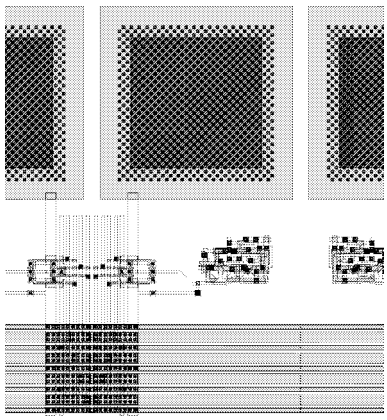


Bild 3.11. Layout einer Ladungspumpe mit Teststruktur. Die angrenzenden Ladungspumpen sind jeweils gespiegelt. Die großen Quadrate oben stellen die Kontaktflächen für die Messspitze dar.

Auf Basis der im Abschnitt 2.3.1 diskutierten Ladungspumpen mit Pass-Gate Schaltern wurde ein Testchip in einer 0,35 µm Technologie von Austria Microsystems (AMS) erstellt und gefertigt. Es handelt sich dabei um den gleichen Prozess, der auch an anderer Stelle in dieser Arbeit als Grundlage dient, insbesondere wurden die in Abschnitt 3.1 vorgestellen Kapazitätscluster für diese Technologie ausgelegt.

Der Testchip verfügt über eine Größe ca. 5 mm^2 und besteht aus einer Matrix von Ladungspumpeneinheiten, die mit Teststrukturen beschaltet sind. Die Organisation der Matrix ist in Abbildung Bild 3.10 zu erkennen: Die Zeilen der Matrix sind vom Aufbau her immer identisch, jede zweite Zeile ist jedoch an der Horizontalen gespiegelt. Es gibt neun Zeilen in aufrechter Orientierung und 8 nach unten gespiegelte Zeilen. Innerhalb einer Zeile sind die linken beiden Ladungspumpen unbeschaltet, sowie zwei in der Mitte und zwei am rechten Rand. Sie dienen der Nettostrombildung wie in Abschnitt „Schaltungsprinzip“ diskutiert. Jede Zeile enthält 52 Ladungspumpen, so dass sich summa summarum $(9 + 8) \cdot 52 = 884$ Einheiten pro Chip ergeben, die jeweils kontaktiert und durchgemessen wurden (siehe hierzu Abschnitt „Messaufbau“).

In Bild 3.11 ist die Geometrie (Layout) einer einzelnen Ladungspumpe aus Bild 3.10 gezeigt. Im oberen Bereich sind die großen rechteckigen Kontaktflächen („probe pads“) für die Messspitze zu sehen, über die der Kontakt zwischen dem Messgerät (Source-Meter) und dem Testchip hergestellt wird, in der Mitte liegen die Kapazitätsstrukturen. Die breiten horizontal verlaufenden Leitungsbahnen im unteren Bildbereich führen die Steuersignale V_{1n}, V_{2p} , ihre Komplementärsignale, den Wannenanschluss[18] (V_{nwell}) sowie die Masseleitung. Um die Steuersignale gegen die Einkopplung von Störungen durch Abschirmung („shielding“) voreinander zu schützen, liegt zwischen jeweils zwei Steuerleitungen eine Masseleitung.

Die linke Abbildung in Bild 3.12 zeigt den ganzen, um 90 Grad nach rechts gedrehten Chip, die rechte Abbildung eine Ausschnittsvergrößerung.

3.2.1 Messaufbau

Aufgrund der großen Zahl (884) von Ladungspumpen bzw. Teststrukturen pro Chip konnten diese nicht komplett über eine Festverdrahtung (Bonddrähte) mit dem Gehäuse verbunden werden. Nur die gemeinsamen Anschlüsse wie V_{1n}, V_{2p} , V_{nwell}, sowie die Versorgungsspannung und Masse wurden, wie in Bild 3.12 ganz links zu sehen, mit der Außenwelt elektrisch fest verbunden. Die Messung des Stroms, der während des Pumpvorgangs fließt, erforderte aufgrund des Schaltungsprinzip eine getrennte Kontaktfläche für jede einzelne Ladungspumpe, die dann über eine Messspitze auf einem Spitzenmessplatz

18. Nwell-Spannung. Die PMOS-Transistoren sitzen in einer negativ dotierten „Wanne“ (engl. „well“), die üblicherweise mit V_{dd} verbunden ist, so auch hier.

(„wafer-prober“) kontaktiert werden musste. Als Messgerät zu Erzeugung der Spannung V_{dd} bei gleichzeitiger Messung des Stromes diente ein Source-Meter der Firma Keithley mit der Bezeichnung 4200-SCS.

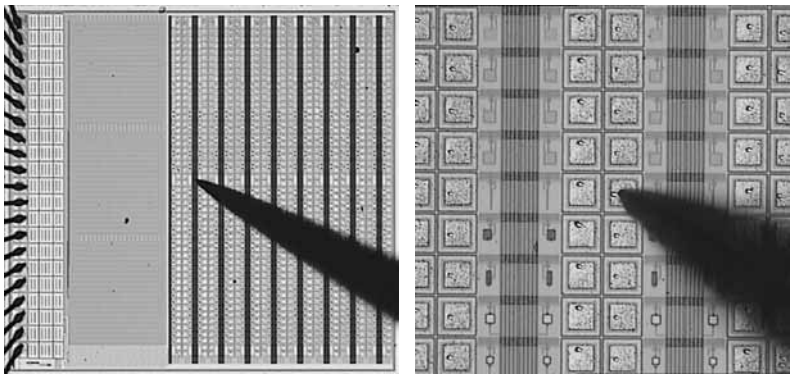


Bild 3.12. Mikroskopbild des Testchips bei 20-facher Vergrößerung. Die Anschlüsse der Stromversorgung und Eingangssignale sind im linken Bild links zu erkennen, die Kontaktflächen der Messspitze in Spalten rechts (der gleichförmig wirkende Block im linken Bild beinhaltet andere Testschaltungen). Das rechte Bild stellt eine Ausschnittvergrößerung des linken Bildes dar.

Bestandteile

Die Leiterplatte. Der Einsatz eines Wafer-Probers führt üblicherweise dazu, dass *alle* Eingangssignale über Kontaktnadeln („probes“) eingespeist werden müssen, einschließlich der Versorgungsspannung. Zudem sind die einzelnen Testchips eines Wafers häufig noch nicht in einzelne Stücke („dies“) zersägt, geschweige denn in ein Gehäuse eingesetzt und elektrisch verbunden. In diesen Fällen müssen also viele Nadeln bzw. Messspitzen eingesetzt werden, um die erforderlichen Anschlüsse zu realisieren. Bei dieser Vorgehensweise ist es erforderlich, dass eine Reihe von Nadeln bei der Messung eines Chips an festen Positionen stehen bleiben, nämlich alle jene, die für die Einspeisung der *gemeinsamen* Signale zuständig sind, während andere Nadeln (hier eine einzelne) von Testschaltung zu Testschaltung bewegt werden müssen. Im Falle der vorliegenden Ladungspumpen müssten (fast) alle in Bild 3.12 ganz links abgebildeten Drahtverbindungen über feststehende Kontaktnadeln realisiert werden, während die Verbindung zum Source-Meter über eine motorisierte Nadel („flying probe“) hergestellt werden müsste.

Soll ein kompletter Wafer auf diese Weise durchgemessen werden, so ist dies der adäquate Weg, um mit der beweglichen Nadel die einzelnen Probe Pads anzufahren, während der ebenfalls motorisierte Probenteller („chuck“), auf dem der Wafer liegt, dazu dient, von Testchip zu Testchip zu springen.

Diese Vorgehensweise besitzt jedoch den Nachteil, dass die große Zahl der erforderlichen Nadeln in der Praxis zu Kontaktschwierigkeiten führt, nämlich aufgrund der ungenauen Positionierung, insbesondere der motorisierten Nadel. Selbst wenn die Position der Nadeln exakt mit den Kontaktflächen übereinstimmt, ist der elektrische Kontakt, vermutlich durch Ablagerungen auf der Oberfläche, häufig schlecht. Dieses Problem wird im Abschnitt “Durchführungsprobleme” auf Seite 75 genauer behandelt.

Aus diesem Grund wurde versucht, die Anzahl der Nadeln so gering wie möglich zu halten und alle anderen Leitungen als Festverdrahtung zu realisieren. Dies war möglich, da ohnehin kein kompletter Wafer produziert werden konnte. Die 20 produzierten Chips wurden (vom Hersteller) in einzelne Dies zersägt, jeweils in ein Gehäuse (PLCC-68) eingesetzt und über Bonddrähte mit dem Gehäuse verbunden. Jeder einzelne der Testchips wurde schließlich auf eine einseitige Leiterplatte („printed circuit board“, PCB) ge-

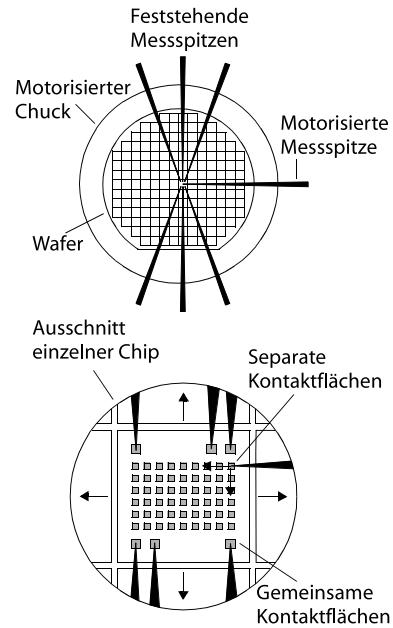


Bild 3.13. In Anwendungsfällen, in denen eine Vielzahl an Teststrukturen innerhalb eines Chips durchgemessen (elektr. charakterisiert) werden soll, ist es notwendig, die Pads, die alle Strukturen gemein haben, über feststehende Nadeln zu kontaktieren und die individuellen Pads über eine motorisierte Nadel anzufahren. Der Übergang von Chip zu Chip geschieht dann über den motorisierten Chuck.

lötet, die auf dem Chuck platziert wurde. Der Anschluss der gemeinsamen Signale und Versorgungsleitungen wurde also über das PCB realisiert, das dann über dünne, flexible Kabel an den Taktgenerator und die Stromversorgung angeschlossen wurde.

Auf diese Weise konnte die gesamte Leiterplatte mit dem Chuck bewegt werden, um über die einzig benötigte Nadel den Kontakt zu den 884 Probe Pads herzustellen. Die wesentlich höhere Positionierungsgenauigkeit des Chucks im Vergleich zu der motorisierten Probe und der Einsatz einer einzigen Nadel statt einer Vielzahl erhöhte die Erfolgswahrscheinlichkeit beim späteren Messdurchlauf beträchtlich. Angesichts der hohen Zahl an anzufahrenden Positionen (884 Positionen pro Chip, 20 Testchips) musste das Messverfahren weitgehend automatisiert werden. Hierzu sei auf "Durchführungsprobleme" auf Seite 75 verwiesen.

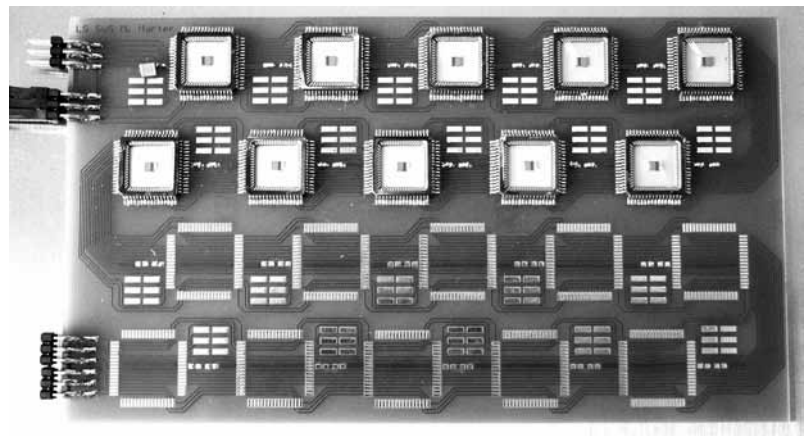


Bild 3.14. Einlagige Leiterplatte (PCB) mit zehn Testchips (die unteren beiden Reihen sind noch unbestückt). Jeder einzelne verfügt über 884 Ladungspumpen bzw. Teststrukturen.

In Bild 3.14 ist die Leiterplatte zu sehen, wie sie für den halbautomatischen Messdurchlauf verwendet wurde. Sie wurde aus Vorsichtsmaßnahmen zunächst nur mit zehn Testchips bestückt, um im Fall eines gravierenden Designfehlers nicht alle Chips zu gefährden. Außer jeweils zwei Abblockkondensatoren pro Chip befinden sich keine weiteren Bauteile auf der Platine, einzig die Stiftleiste für die Kabelanschlüsse zur Signaleinspeisung sind links zu erkennen.

Jede Form von Unregelmäßigkeiten durch Leitungsstrukturen auf der Unterseite würde dazu führen, dass die Leiterplatte nicht plan auf dem Chuck aufliegt. Da der zu messende Gegenstand – also normalerweise der Wafer, hier die Platine – durch ein extern erzeugtes Vakuum an den Chuck gesaugt wird, um ihn am Verrutschen zu hindern, würde jede Unebenheit zu Lufteinlässen führen, die eine Fixierung auf dem Teller erschweren würde. Deshalb konnte die Unterseite der Platine nicht benutzt werden. Eine einseitig genutzte Leiterplatte reichte im hier vorliegenden Fall jedoch völlig aus.

Bild 3.15. Messaufbau zur Kapazitätsbestimmung mittels Ladungspumpen. In der Mitte ist der Spitzenmessplatz („wafer prober“) PA200 der Firma Suess zu sehen, links auf dem Container der Taktgenerator zu Erzeugung der Steuersignale und ein Oszilloskop. Der Source-Meter zur Messung des Stroms steht rechts auf einem Tisch (nicht abgebildet).

Der Wafer-Prober. Der eingesetzte Spitzenmessplatz ist in Bild 3.15 zu sehen, zum Einsatz kam das Modell PA200 der Firma Suess Microtech. Der Messplatz verfügt unter anderem über einen programmierbaren, beweglichen Chuck (nicht sichtbar), eine Abschirmung („probe shield“) zur Verdunkelung und zum Schutz vor Störsignalen und ein Mikroskop (Mitte oben). Der Prober steht auf einem durch eine pressluftgelagerte Granitplatte schwingungsisolierten Tisch (unten).

Die Steuerung des Probers (Position des Chucks, Ein-/Ausschalten des Lichts, usw.) geschieht durch einen externen Rechner (rechts zu erkennen), an den ein Kontrollpult mit Joystick angeschlossen ist. Damit kann die PA200 manuell bedient werden, was vor allem die Grobjustage am Anfang erleichtert.

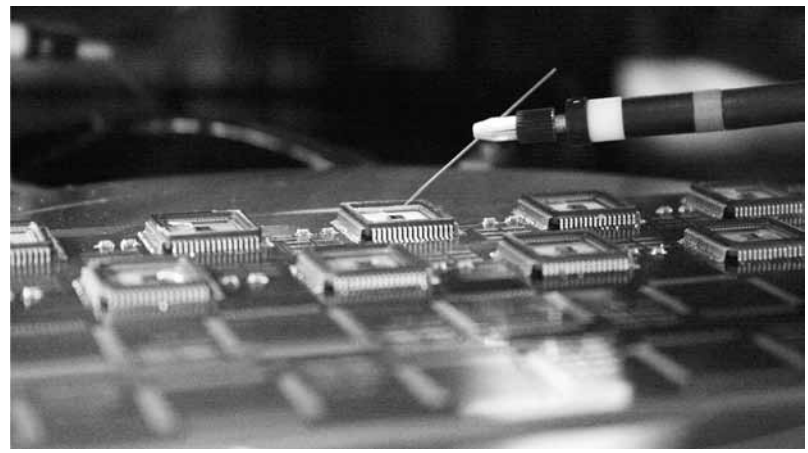


Bild 3.16. Kontaktnadel aus Wolfram mit einem Durchmesser von 2 µm an der Spitze. Der Deckel der Chipgehäuse wurde geöffnet, um die auf die Leiterplatte gelöteten Chips mit der Messspitze zu kontaktieren. Die Platine selbst liegt auf dem beweglichen Chuck und wird dort durch Unterdruck festgesaugt.

In Bild 3.16 ist die Messspitze zu sehen, wie sie für die Durchführung der Messungen verwendet wurde. Sie besteht aus Wolfram und hat einen Spitzendurchmesser von 2 µm. Die Nadel ist in die Führung des Arms der Sondenhalterung („probe head“) geklemmt, wobei der Führungswinkel eine nicht zu unterschätzende Rolle spielt. Ist der Winkel flach, so bewegt sich die Nadel beim Aufsetzen auf die Kontaktfläche weit vom Zielort weg, befreit dadurch jedoch in vorteilhafter Weise die Oberfläche von störenden Ablagerungen. Ist der Winkel auf der anderen Seite steil, so wir der Zielpunkt leichter getroffen, die Nadel gräbt sich aber weniger unter die Verunreinigungen.

Statt einer Nadel mit Spitzendurchmesser von 2 µm hätte auch eine mit 7 µm zur Auswahl gestanden. Versuche mit diesem Nadeltyp lieferten jedoch kein befriedigendes Ergebnis, da der größere Durchmesser das Eindringen der Nadel in die elektrisch leitfähige Schicht der Kontaktflächen durch die Verunreinigungen der Oberfläche hindurch erschwert. Darüber hinaus ist die Wahrscheinlichkeit größer, mit dem Rand der Nadelspitze auf der überhöhten Kante des Probe Pads aufzusetzen und so gar keinen Kontakt zu erhalten.

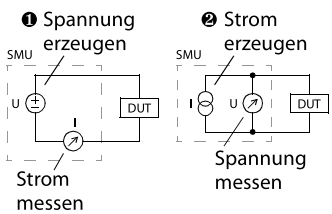


Bild 3.17. Ein Source-Meter (oder eine SMU) hat im wesentlichen zwei Betriebsmodi: Spannung erzeugen und Strom messen oder umgekehrt.

DER SOURCE-METER. Als zentrales Messgerät diente das Modell 4200-SCS von Keithley Instruments. Es handelt sich dabei um ein komplettes System zur Charakterisierung von Halbleiterbauelementen. Zu diesem Zweck verfügt das Geräte über acht Einschübe, die mit sogenannten „source monitoring units“ (SMU) bestückt werden können. Wahlweise können diese dann zusätzlich über einen Aufsatz mit Vorverstärkern ausgerüstet werden.

Hauptaufgabe der SMU ist die Erzeugung und das Messen einer Spannung oder eines Stromes. Es gibt im wesentlichen zwei Betriebsparameter (siehe Bild 3.17):

1. Eine Spannung wird erzeugt („source“) und der dabei fließende Strom gemessen („monitor“ oder „meter“).
2. Ein Strom wird erzwungen („source“) und die am Messobjekt anliegende Spannung gemessen („monitor“ oder „meter“).

Eine zweite wichtige Eigenschaft eines Source-Meters ist die Fähigkeit, eine Vierpunkt-Messung durchzuführen. Diese Merkmal wird im zweiten Fall häufig benötigt, d.h. bei der Messung der Spannung, die bei einem vorgegebenen Strom am Testobjekt („device under test“, DUT) anliegt. Mit einem Leitungspaar, „Force“ genannt, wird der gewünschte Strom erzeugt, mit einem zweiten Leitungspaar, „Sense“, wird die anliegende Spannung *strom-*

los gemessen. Durch die stromlose Messung kommt es auf diesen Kabeln zu keinem Spannungsabfall („IR-drop“), der das Ergebnis signifikant beeinflussen würde (siehe Bild 3.18).

Im vorliegenden Fall wurde der 4200-SCS gemäß Fall eins benutzt. Die SMU erzeugt dabei die Spannung U in Bild 2.32 auf Seite 56, sie entspricht der Versorgungsspannung V_{dd}. Der durch den fortwährenden Pumpvorgang fließende Strom wird von der SMU gemessen und am Ende ausgewertet. Auf die Vierpunkt-Methode konnte verzichtet werden, da die für die Berechnung der Kapazität maßgebliche Spannung (in Gleichung 2.41 die Variable U) in jenen Phasen an der Ladungspumpe anliegen muss, in denen keine Ladung in den zu messenden Kondensator gepumpt wird, also auch kein Strom fließt.

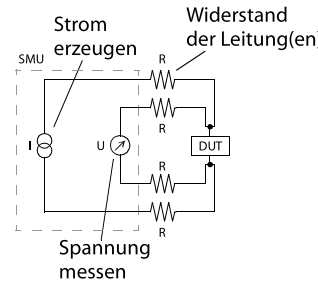


Bild 3.18. Die Vierpunkt-Messung dient dazu, eine Spannung stromlos zu messen, um den Spannungsabfall aufgrund des ohmschen Widerstands der Leitungen zu umgehen.

3.2.2 Durchführung der Messungen

Steuerung des Messvorgangs

Die zentrale Kontrolle aller Elemente des Messaufbaus, also der Versorgungsspannung (Netzgerät), der Steuersignale (Taktgenerator), des Probers und des Source-Meters übernahm eine in C++ geschriebene Steuerungs- und Testapplikation, die auf einem externen Rechner lief. Die Verbindung mit dem Source-Meter, dem Netzgerät und dem Taktgenerator wurde über den IEEE-488 Standard (GPIB) hergestellt, der Anschluss an den Hostrechner des Probers erfolgte über einen herkömmlichen Ethernet Netzwerkanschluss. Bild 3.19 zeigt das Zusammenspiel der Komponenten in der Übersicht.

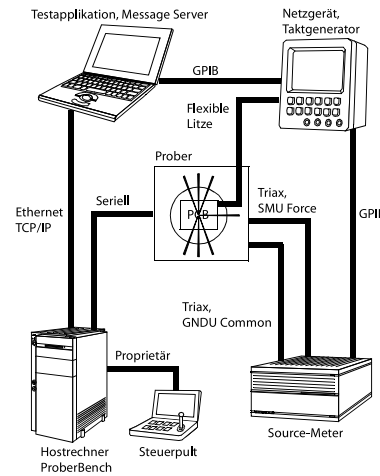


Bild 3.19. Aufbau der Messumgebung. Netzgerät und Taktgenerator sind mit der Leiterplatte verbunden, der Hostrechner mit dem Prober und der Source-Meter mit der Sondenhalterung.

Der Anschluss des Source-Meters an die Sondenhalterung der Messspitze im Prober geschieht typischerweise über Triax-Kabel, einer mit Coaxial-Kabeln verwandten, mit einer zusätzlichen Abschirmung versehenen Kabelsorte hoher Güte. Im Rahmen des Messaufbaus wurden nur zwei dieser Kabel benötigt, nämlich die Force-Leitung (sog. „Force-Hi“-Leitung) einer SMU-Einheit und die gemeinsame Masse des Gerätes, die sich an der Masse-Einheit („ground unit“, GNDU) befindet und mit „common“ gekennzeichnet ist. Zu beachten ist, dass die Ground Unit über vier Anschlüsse verfügt, „common“, „chassis“, „force“ und „sense“. Wird keine Vierpunkt-Messung durchgeführt, kann auf die letzten beiden verzichtet werden, die gemeinsame Masse genügt („chassis“ ist mit dem Gehäuse des Geräts verbunden und über eine abnehmbare Metallplatte mit „common“ verbunden).

Aufgabe des Hostrechners ist im allgemeinen Anwendungsfall die Kontrolle des Probers mittels der von Suess bereitgestellten Software „ProberBench“. Kernstück dieses Softwarepakets ist der „Suss Message Server“ über den die Kommunikation mit der Hardware des Probers erfolgt. Als Interface dient entweder die Benutzeroberfläche der Software oder das externe Bedienpult, das über einen Joystick zur präzisen Steuerung verfügt und die aktuelle Position des Chucks oder der motorisierten Messspitze anzeigt. Darüber hinaus zeigt der Hostrechner das Bild der Kamera an, die an das Mikroskop des Probers angeschlossen ist.

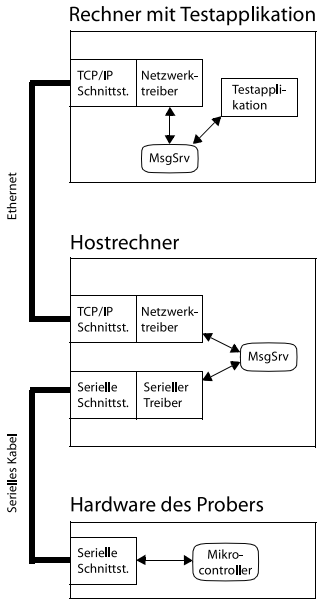


Bild 3.20. Die Kommunikation der Testapplikation mit dem Prober geschieht über mehrere Stufen. Kernelement ist der „Suss Message Server", auf den die Testapplikation über Bibliotheksfunktionen zugreifen kann.

Um eine möglichst weitgehende Automatisierung des Messverfahrens zu verwirklichen, ist es erforderlich, über entsprechende Schnittstellen eine Verbindung zum Message Server herzustellen, der dann die Steuerungsbefehle umsetzt und an die Hardware des Probers weiterleitet. Eine Möglichkeit hierzu ist, die Test- und Steuerungsapplikation direkt auf dem Hostrechner laufen zu lassen und über Bibliothekszugriffe die Kommunikation mit dem Message Server zu realisieren.

Eine andere Möglichkeit besteht darin, das ProberBench Paket auf einem weiteren Rechner zu installieren und alle Befehle der Testapplikation an den lokalen Message Server zu übergeben, der diese dann via Ethernet bzw. TCP/IP an den Message Server des Hostrechners weiterleitet. Vorteil dieser Variante ist, dass der Hostrechner „unangetastet" bleibt und die Programmierung der Testapplikation in der dem Anwender vertrauten Umgebung auf einem persönlich eingerichteten Rechner erfolgen kann. Schematisch ist diese Möglichkeit in Bild 3.20 dargestellt.

Nach der Installation der ProberBench Software kann die Testapplikation über gewöhnliche Bibliotheksfunktionen, die Suess als Teil der „Programmer Tools" zur Verfügung stellt, auf den lokalen Message Server zugreifen, der die Befehle dann weiterschickt. Zur Illustration der typischen Vorgehensweise bei der Programmierung des Probers ist in Box 3.2 ein einfaches Beispiel gegeben.

Alle weiteren anzusteuernden Messgeräte und Signalgeneratoren können üblicherweise über die GPIB-Schnittstelle angesprochen werden. Im konkreten Fall wurde auch der Source-Meter von Keithley über GPIB gesteuert, obwohl das Gerät einen eigenständigen PC darstellt, so dass die komplette Testsoftware auch auf dem Source-Meter hätte laufen können.[19] Der Zugriff auf die Hardware des Geräts wäre dann über Bibliotheksroutinen erfolgt. Da diese jedoch in keiner für die verwendete Entwicklungsumgebung angepassten Form von Keithley bereitgestellt wurde und um den Rechner des Source-Meters nicht als Plattform zur Programmierung zu „missbrauchen", wurde die Testapplikation also auf dem externen Rechner entwickelt und benutzt.

Die Testapplikation hatte als zentrales Instrument zur Steuerung, Messung und Auswertung des Messvorgangs eine Reihe von Aufgaben zu erfüllen:

1. Initialisierung durchführen und Höhenprofil des Chips erstellen.
2. Versorgungsspannung und Frequenz der Eingangssignale einstellen.
3. Chuck des Probers an die aktuelle Kontaktflächenposition fahren.
4. Nadel aufsetzen und Kontaktgüte überprüfen. Bei schlechtem Kontakt den Chuck innerhalb gewisser Grenzen zufällig hoch- und runterbewegen und Kontaktgüte erneut überprüfen.
5. Einschalten der Spannung am Source-Meter, mittleren Strom im vorgegebenen Intervall messen.

19. Keithley bietet für den 4200-SCS eine spezielle Programmierumgebung mit der Bezeichnung „KULT" an, die eine vorgefertigte Benutzerschnittstelle zur Verfügung stellt und über die der Source-Meter in C programmiert werden kann. Der Prober wird über die zentrale Charakterisierungssoftware „KITE" gesteuert. In vielen Fällen ist dieser Weg der bequemste. Soll der Prober jedoch in spezieller Weise gesteuert werden, bietet dieser Weg zu wenig Flexibilität.

Box 3.2 Kommunikation mit dem Prober unter C++ (Windows)

Die Befehle zur Kontrolle des Probers ähneln der Form nach den GPIB-Anweisungen vieler Messgeräte, d.h. es handelt sich um einfache Zeichenfolgen (strings) die den Kommandonamen und die Argumente enthalten. Die Befehle wurden von Suess in drei Kategorien eingeteilt, die relevanten Steuerungsbefehle finden sich in der Dokumentation unter „Kernel Commands".

Zum Auslösen eines Befehls genügt der Aufruf einer einzigen Bibliotheksfunktion mit dem Namen „DoProberCommand", die das eigentliche Kommando als Parameter übergeben bekommt. Im Folgenden dazu ein einfaches Beispiel, das bereits eigenständig lauffähig ist:

```
// Reserviert Puffer für das Kommando und die Antwort
char CmdBuf[255];
char RespBuf[255];

// Meldet diese Anwendung bei der ProberBench an
if ( !RegisterProberApp("SCI_BCB5_SIMPLE",
     "SCI_BCB5_SIMPLE",0) ) {
  Application->MessageBox("No access!","Error",MB_OK);
  return;
}

// Senkt den Chuck um 10 Micron relativ zur aktuellen Position
// mit der Geschwindigkeit Vel.
sprintf( CmdBuf, "MoveChuckZ -%d R Y 10", Vel);
DoProberCommand( CmdBuf, RespBuf );

// Bewegt den Chuck an die Position (Posx, Posy)
// mit der Geschwindigkeit Vel.
sprintf( CmdBuf, "MoveChuck %g %g H Y %.2f", Posx, Posy, Vel);
DoProberCommand( CmdBuf, RespBuf );

// Meldet die Anwendung bei der ProberBench ab:
CloseProberApp();
```

6. Ergebnisse in Tabellenform zwischenspeichern, einen Graphen der Messpunkte einer Ladungspumpe erstellen und beide auf Festplatte speichern.

Einige dieser Punkte erwiesen sich erst im Praxisdurchlauf als notwendig und mussten mehrmals den Erfordernissen angepasst werden. Als Hauptproblem erwies sich der teilweise nur schwer herzustellende elektrische Kontakt. Auf die Details der Testsoftware wird deshalb im folgenden Abschnitt („Durchführungsprobleme") eingegangen.

Durchführungsprobleme

HÖHENKOMPENSATION. Ein spezielles Problem zeichnete sich schon vor der ersten Messung ab. Bedingt durch die Verwendung einer Leiterplatte mussten alle zu messenden Chips (von Hand) in ein Gehäuse eingesetzt und auf die Platine gelötet werden. Dadurch ergaben sich kleine, für das Auge kaum sichtbare Unterschiede in der Ausrichtung der Chips zueinander und zwar in mehreren Punkten:

1. Die Drehung um die Vertikale (Winkel Theta).
2. Abstand und
3. Neigung der Chipebene zur Leiterplatte bzw. zum Chuck.

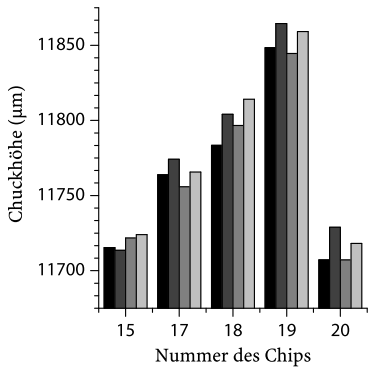


Bild 3.21. Chuck-Höhe am Kontaktpunkt der Nadel in der linken unteren, linken oberen, rechten oberen und rechten unteren Ecke (schwarz bis hellgrau) von fünf gemessen Chips.

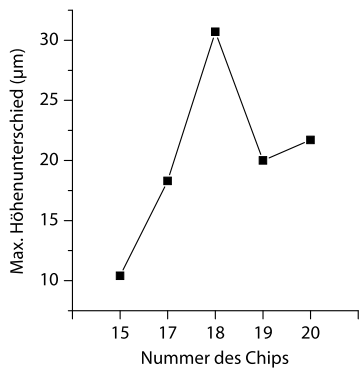


Bild 3.22. Maximaler Höhenunterschied des Chucks am Kontaktpunkt der Nadel innerhalb eines Chips. Der Unterschied ist bei allen so groß, dass eine fortwährende Höhenanpassung während des Durchmessens der jeweils 884 Kontaktflächen nötig ist.

Der erste Punkt führte dazu, dass bei jedem Übergang von Chip zu Chip der Chuck um die Z-Achse (Vertikale) gedreht werden musste.

Die Punkte zwei und drei zeigten sich durch den Höhenunterschied des Probentellers am Kontaktpunkt der Nadel, sowohl von Chip zu Chip, als auch innerhalb eines Chips. Der Unterschied war als Folge der Neigung der Chipebene und der prozesstechnisch bedingten nahezu perfekten Ebenheit der Dies an den Ecken am größten. In Bild 3.21 sind die vier Z-Werte von fünf Chips zu sehen, beginnend in der linken unteren Ecke (schwarzer Balken) über die linke obere, rechte oberen bis zur rechten unteren Ecke (hellgrauer Balken). Der maximale Höhenunterschied innerhalb eines Chips ist in Bild 3.22 zu sehen, d.h. die Schiefe der Chipebene relativ zum Chuck. Die Werte zwischen 10 und 30 Mikrometern führten dazu, dass sich die Kontaktpunkthöhe beim Durchschreiten der einzelnen Positionen auf dem Chip von einer Ecke zur anderen maximal um den angegebenen Werte verschob.

Diese Werte sind für die hochpräzise Feinmechanik des Probers und der Kontaktnadel bereits so groß, dass sie ohne Höhenkorrektur beim Durchwandern der Pad-Positionen dazu führen, dass die Nadel entweder den Kontakt verliert (Höhenabnahme), oder immer stärker an die Chipoberfläche gedrückt wird und dadurch seitlich verrutscht (Höhenzunahme). Aus diesem Grund musste bei jedem Kontaktflächenwechsel eine Anpassung der Chuck-Höhe vorgenommen werden, die Schieflage der Chips also herausgerechnet werden.

Diese Höhenkompensation wurde auf Grundlage eines Höhenprofils vorgenommen, das zu Beginn der Messreihe eines neuen Chips erstellt wurde. Dazu ermittelte die Steuerungssoftware die Chuck-Höhe an den vier Ecken des Chips, indem der Benutzer jeweils aufgefordert wurde, den Chuck manuell so weit hochzufahren, bis die Nadel durch leichte Bewegungen den physischen Kontakt zum Pad anzeigte. Die Software prüfte den elektrischen Kontakt und übernahm im positiven Fall den Höhenwert des Probentellers.

Im Idealfall einer absolut planen Chipoberfläche liegen dabei alle vier Z-Werte in einer Ebene, in der Praxis jedoch lag ein Punkt häufig um ein oder zwei Mikrometer höher oder tiefer als die durch die anderen drei Punkte aufgespannte Ebene. Ursache war meist die Tatsache, dass die Nadel an den vier Ecken unterschiedlich stark angedrückt werden musste, um einen zufriedenstellenden elektrischen Kontakt zu erreichen, weniger die Unebenheit der Oberfläche. Als Lösung des Problems berechnete die Software aus den Z-Werten der vier Ecken eine Ausgleichsebene (bilineare Regression), die zur Approximation der Höhenwerte jedes einzelnen Pads diente. Als Ebene kam die folgende Funktion $f(x, y)$ zum Einsatz, die Koeffizienten a_0 bis a_3 wurden über die Methode der kleinsten Quadrate („least squares“) aus den vier Eckpunkten berechnet:

$$f(x, y) = a_0 + a_1 x + a_2 y + a_3 xy \tag{3.1}$$

DER ELEKTRISCHE KONTAKT. Der physische Kontaktpunkt ist immer dann erreicht, wenn die Messspitze bei jeder weiteren Annäherung eine deutlich sichtbare Bewegung zu Seite erkennen lässt. An diesem Punkt berühren sich Kontaktfläche und Nadel zwar rein mechanisch, der *elektrische* Kontakt ist dabei jedoch noch nicht notwendigerweise hergestellt. Zum einen wird ein gewisser Anpressdruck durch Überschreiten des Kontaktpunkts („overtravel“) benötigt, zum anderen muss die Nadel einige Atomlagen Schmutz und Ablagerungen durchdringen, um zur eigentlichen Kontaktflä-

che zu gelangen. In der Praxis hat sich herausgestellt, dass die besten Ergebnisse zu erzielen sind, wenn die Nadel durch manuelles Eingreifen die Pad-Oberfläche regelrecht „freikratzt“. In Bild 3.23 ist die mikroskopische Aufnahme einiger Kontaktfläche zu sehen, an denen auf diese Weise vorgegangen wurde, um den elektrischen Kontakt zu verbessern bzw. herzustellen.

Als pragmatische Lösung wurde die Güte des elektrischen Kontakts anhand der Zielmessung beurteilt, d.h. die Schaltung wurde wie bei der eigentlichen Messung in Betrieb genommen und die Wiederholbarkeit bewertet. Bei einer gegebenen Versorgungsspannung und Taktfrequenz wurde der Strom wiederholt gemessen und Standardabweichung und Mittelwert bestimmt. Je nach Größe der Kapazität, die an die jeweils gemesse Ladungspumpe angeschlossen war, galt es, bestimmte Bereiche für den Mittelwert und die Standardabweichung einzuhalten. In Tabelle 3.2 sind diese Werte aufgelistet. Bei den unbeschalteten Ladungspumpen ist der mittlere Strom niedriger als in den anderen Fällen. An Ladungspumpe Spalte 3 hängt z.B. ein relativ großer Plattenkondensator (ca. 74 fF), Mittelwert und Standardabweichung müssen deshalb größer sein. Die Werte aus der Tabelle entstammt Messungen, bei denen Messwerte über den gesamten Parameterbereich zu konsistenten Ergebnissen (Kapazitäten) geführt hatten.

Anhand der Kriterien in Tabelle 3.2 konnte nun ein iteratives Verfahren implementiert werden, das bei schlechtem elektrischem Kontakt dafür sorgte, dass die Messspitze wiederholt in die Oberfläche hineingetrieben wurde. Dazu wurde der Chuck zufällig um einige wenige Mikrometer hoch- und runterbewegt, solange, bis der gemessene Strom die gestellten Bedingungen erfüllte. Die Anzahl der Bewegungen konnte durch die Benutzeroberfläche der Steuerungssoftware vorgegeben werden, ebenso wie der Bewegungsbereich H_{overTr} des Chucks:

$$H^{n+1} = H^n + H_{\text{overTr}} \cdot \text{RandomRange}(-¼, 1) \tag{3.2}$$

Die neue Höhe des Chucks ergab sich also aus der alten, zuzüglich des Offsets, der aus dem Parameter H_{overTr} errechnet wurde, sowie einem Skalierungs- bzw. Gewichtungswert, der zufällig zwischen -¼ und +1 gewählt wurde. Die stärke Gewichtung hin zu positiven Werten rührt daher, dass ein stärkerer Anpressdruck erfolgversprechender war, als ein größerer Abstand zwischen Pad-Oberfläche und Nadel. In Bild 3.24 ist der Verlauf der Standardabweichung des Stroms und die Höhe des Probentellers während dieser Prozedur zu sehen. Man erkennt, dass bloßes Hineindrücken der Nadel in das Pad (bzw. des Chucks in die Messspitze) nicht ausreichte. Stattdessen bewegte sich der Chuck wiederholt von der Nadel weg, um sich danach wieder zu nähern. Durch das Auf *und* Ab „kratzt“ die Nadel also die Oberfläche frei.

Laufzeit. Durch eine Reihe von Ursachen war die Laufzeit des Messdurchlaufs eines jeden Chips beträchtlich, im Schnitt betrug sie ca. 10 Stunden! Einige dieser Umstände sind unvermeidbar, andere mögen mit zusätzlichem zeitlichem Aufwand ein lösbares Problem darstellen, konnten im Rahmen dieser Arbeit jedoch nicht weiter verfolgt werden, um den Focus der Arbeit nicht zu weit auf die Messtechnik zu verschieben. Als Ursachen sind zu nennen:

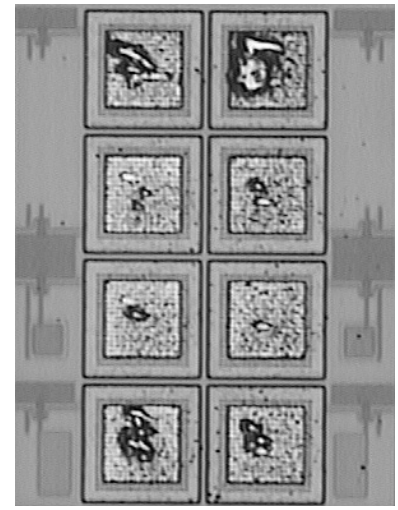


Bild 3.23. Mikroskopbild der Kontaktflächen nach einer Reihe von Kontaktversuchen.

Kapazität	$\mu_{\text{min}}(I)$	$\mu_{\text{max}}(I)$	$\sigma_{\text{max}}(I)$
keine	10 nA	14 nA	9.9 pA
~74 fF	200 nA	350 nA	30 pA
sonst	20 nA	110 nA	9.9 pA

Tabelle 3.2. Kriterium zur Beurteilung der Güte des elektrischen Kontakts durch wiederholtes Messen des Stroms bei realen Testbedingungen.

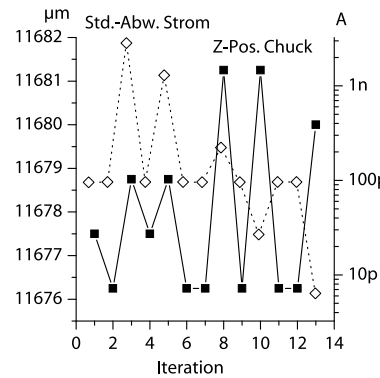


Bild 3.24. Z-Verlauf (Höhe) des Probentellers gemäß der Optimierungsroutine bei schlechtem elektrischem Kontakt (linke Achse, durchgezogene Linie). Erst bei Iteration 13 fällt die Standardabweichung des gemessenen Stroms (rechte Achse) unter 10 pA, d.h. der elektrische Kontakt ist hergestellt.

1. Physischer Kontakt ungleich elektrischer Kontakt. Das „Kratzen“ und wiederholte Nachmessen kostet Zeit.
2. Große Anzahl anzufahrender Positionen (884 Pads) bei geringer Geschwindigkeit des Chucks.
3. Verrutschen bzw. Drift der Nadel nach einer gewissen Zeit (einigen Stunden).
4. Ausreißer bei einigen Messpunkten trotz offenbar gutem elektrischem Kontakt. Die Messung des entsprechenden Pads musste komplett wiederholt werden.
5. Geringe Geschwindigkeit bei der Programmierung des Taktgenerators und des Netzgerätes.
6. Langsame Strommittelung im Source-Meter.

Punkt 1 wurde bereits im vorhergehenden Abschnitt diskutiert und kann als unvermeidbar angesehen werden.

Punkt 2 könnte durch eine höhere Chuck-Geschwindigkeit gelöst werden. Der Wert wurde sehr konservativ gewählt, da die auf dem Probenteller liegende Leiterplatte eine recht große Masse darstellt (und damit Trägheit) und deshalb durch ruckartige Bewegungen verrutschen könnte. Auch wird die Messspitze durch die Bewegungen des Chucks in Schwingungen versetzt, trotz eines Abstands von 100 Mikrometern (sog. „separation height“). Dadurch ist die Wahrscheinlichkeit für ein Wegdriften der Nadel nach einigen hundert Schritten größer.

Damit zu Punkt 3: Er kann das Ergebnis einer zu hohen Bewegungsgeschwindigkeit sein, jedoch auch andere Ursachen haben. So bewegen sich die Nadel und der Chuck auch bei Nichtbenutzung über mehrere Stunden auseinander. Eine Nadel, die am Abend über einem Pad positioniert wird, befindet sich am nächsten Morgen nicht mehr exakt an derselben Stelle. Dies könnte die Folge von Erschütterungen des vorbeifahrenden Verkehrs sein, trotz schwingungsisoliertem Tisch (der über Nacht allerdings nicht mehr mit Pressluft betrieben wird). Darüber hinaus führt die oben erwähnte „Kratzprozedur“, vermutlich jedoch schon das bloße Überfahren des physischen Kontaktpunkts durch den Overtravel, zu einem Wegdrücken der Nadel vom Ursprungsort weg. Wird die Belastung schließlich weggenommen, so kehrt sie an diesen nicht mehr exakt zurück. Um ein solches Verrutschen der Messspitze während des Messablaufs zu entdecken, wurde das Licht beim Pad-Wechsel immer eingeschaltet, so dass der Zielort der Nadel überprüft werden konnte. Stimmt er mit der Pad-Mitte nicht mehr überein, so wurde der Messvorgang angehalten und die Messspitze nachjustiert.

Punkt 4 kann nicht abschließend erklärt werden. Trotz geringer Schwankungen bei der Wiederholung der Messung eines Spannung-Frequenz Paares und plausiblem Mittelwert traten Ausreißer bei anderen Parameterpaaren auf. Möglich ist, dass der elektrische Kontakt während des Messvorgangs durch verkehrsbedingte oder andere Erschütterungen verschlechtert wurde. Um die Ausreißer zu identifizieren wurde eine pragmatische Lösung gewählt: Die von der Steuerungssoftware erstellen Graphen der Gesamtmessung jeder Ladungspumpe wurden automatisch abgespeichert und am Ende aller Messungen inspiziert. Visuell konnten Ausreißer so schnell gefunden werden und die entsprechenden Messungen wiederholt werden. In Tabelle 3.3 ist exemplarisch zu sehen, an welchen Stellen und wieviele Ausreißer auftraten. Mit ca. 9 Prozent ist die Zahl so hoch, dass es sich lohnen würde, mit etwas mehr Zeit dem Problem auf den Grund zu gehen.

Reihe	Orientierung	Ausreißer Spalte
1	Aufrecht	13, 23, 31, 33, 35-37, 39, 45, 48-52
3	Aufrecht	15
4	Aufrecht	8, 10, 24
5	Aufrecht	42
6	Aufrecht	12, 19, 42, 48
7	Aufrecht	3, 6, 19, 49
8	Aufrecht	7,20, 24, 28, 36, 48
9	Aufrecht	3, 41, 49
1	Gespiegelt	3, 4, 9, 18, 30, 34, 46, 49
2	Gespiegelt	7, 12, 24
4	Gespiegelt	4
5	Gespiegelt	24, 30, 31, 34, 40, 41, 45
6	Gespiegelt	4, 14, 15, 17, 19, 21, 22, 33, 37, 38, 47, 49
7	Gespiegelt	3, 8, 10, 22, 34, 36-38, 48
8	Gespiegelt	4, 9, 17, 24, 30

Tabelle 3.3. Beispiel für die Anzahl und Verteilung der Ausreißer bei einigen Messpunkten. Ca. 9% der Pads mussten wiederholt gemessen werden.

Punkt 5 ist sehr wahrscheinlich unvermeidbar. Die Programmierung der Versorgungsspannung und des Taktsignals dauert eine gewisse Weile, da zu der langsamen Befehlsübertragung als Zeichenfolge die Latenz der Geräte beim Umschalten des Betriebsbereichs hinzukommt. Beim Übergang von einem Spannungsbereich zum nächsten und zwischen den Frequenzbereichen ist das Klicken der Relais im Ausgangsteil der Geräte, über die der Bereich gewechselt wird, deutlich zu hören. Dieses Umschalten dauert aufgrund der mechanischen Bauweise beträchtlich. Eine mögliche Lösung besteht in der Einschränkung des Spannungs- und Frequenzbereichs auf einige Millivolt bzw. hundert Kilohertz, was jedoch sicherlich einen Verlust bei der Messgenauigkeit zur Folge hat.

Doch nicht nur auf der Seite der Stimuli ergeben sich Verzögerungen. Auch bei der Messung des mittleren Stroms durch den Source-Meter (Punkt 6) vergeht eine gewisse Zeit, die über den Parameter „integration time“ eingestellt werden kann. Je länger das Integrationsintervall, desto genauer ist die Messung, da sich sporadische Störung mit der Zeit wegmitteln. Aus diesem Grund wurde der 4200-SCS auf die Einstellung „quiet“ gesetzt, was eine lange Integrationszeit zur Folge hatte.

Diese letzten beiden Punkte liefern die Hauptbeiträge zur Gesamtlaufzeit der Messungen, sind jedoch kaum beeinflussbar. Generell lässt sich sagen: Will man sehr kleine Ströme über einen weiten Parameterbereich hochgenau messen, so dauert dies notgedrungenermaßen lange.

Anstiegs-/Abfallszeiten. Die Flankensteilheit der Eingangsimpulse hat, wie in Abschnitt „Kapazitätsauflösung“ diskutiert, Einfluss auf die Genauigkeit des Messergebnisses. Die tatsächliche Flankensteilheit an den Gate-Anschlüssen der Transistoren konnte jedoch nicht direkt beeinflusst werden, da die Eingangssignale durch eine Kaskade von Signalverstärkern („buffer“) geschickt wurden, bevor sie an den Ladungspumpen antreffen.

Trotzdem hatte die Flankensteilheit einen durchschlagenden Einfluss auf das Messergebnis, nämlich in der Frage, ob die Messung überhaupt zu sinnvollen Werten führte oder nicht. Bei zu hohen Geschwindigkeiten (z.B. 1,6 ns) ließen sich keine brauchbaren Messergebnisse erzielen. Die gemessenen Ströme einer Spannungs- bzw. Frequenzreihe bildeten keine Gerade (siehe Bild 3.25). Erst bei 40 Nanosekunden und darüber blieben die willkürlichen Störungen aus. Der Grund für das geschilderte Verhalten liegt in den Reflexionen auf den unterminierten Anschlussleitungen vom Pulsgenerator.

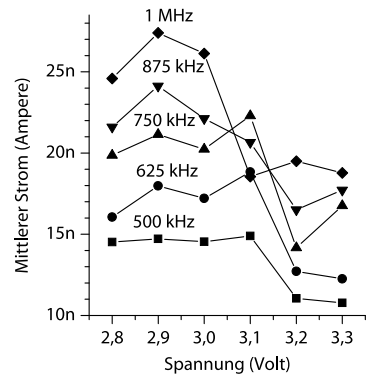


Bild 3.25. Ergebnis eines Messdurchlaufs bei einer Anstiegs-/Abfallszeit von 1,6 ns.

Die Benutzeroberfläche. Der gesamte Messvorgang zergliederte sich in mehrere Schritte, die teilweise aufgrund der diskutierten Probleme nötig waren, teilweise aber einfach nur die normale Messprozedur wiederspiegeln. Dank der Automatisierung des Messvorgangs durch die Steuerungssoftware konnten die Messungen weitgehend eigenständig vonstatten gehen. Völlig unbeaufsichtigt konnte der Messaufbau jedoch nicht gelassen werden, da eine Benutzerinteraktion in den Fällen nötig war, in denen die Messspitze verrutschte, sowie am Ende bei der Wiederholung der fehlerhaften Messungen.

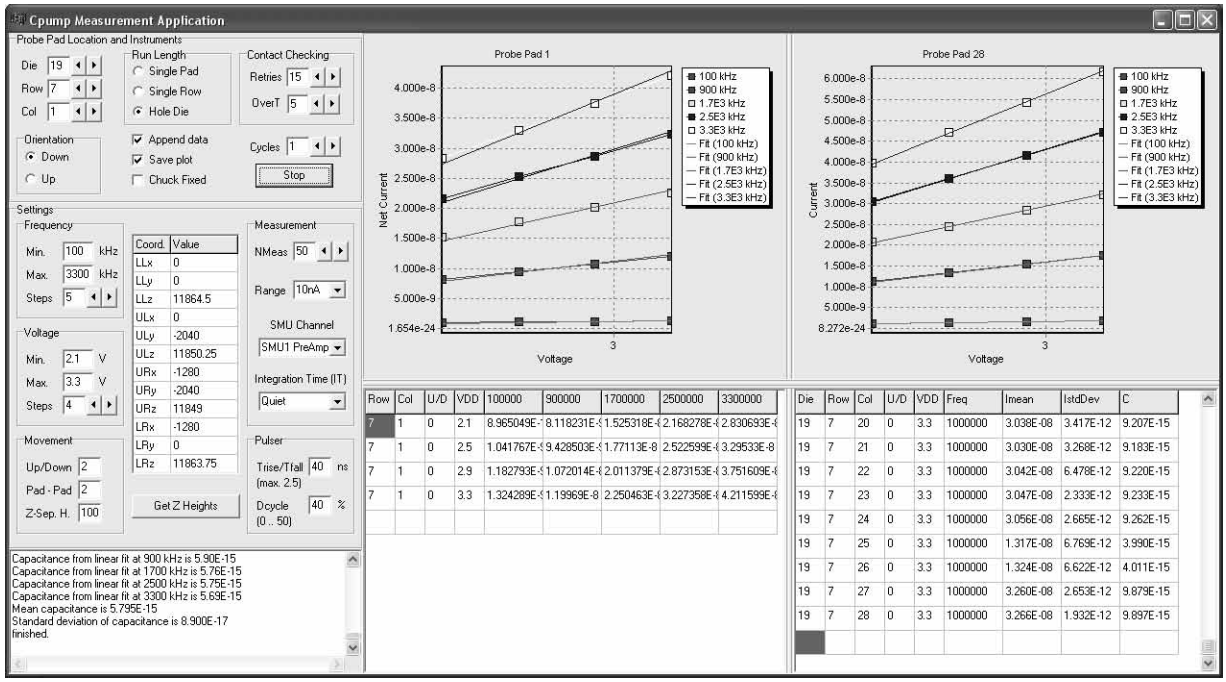


Bild 3.26. Testapplikation zur Programmierung der Versorgungsspannung und Taktsignale, sowie zum Auslesen des gemessenen Stromes. Die Kontaktflächen der einzelnen Teststrukturen werden automatisch angefahren, indem die Applikation die entsprechenden Koordinaten berechnet und an den Wafer Prober sendet. Neben der Speicherung der Messwerte wird eine erste Auswertung der Daten vorgenommen und daraus ein Schaubild erstellt.

Schließlich sei noch die Benutzeroberfläche erwähnt, über die das komplette Messverfahren gesteuert, überwacht und ausgewertet wurde. Die meisten der in Bild 3.26 gezeigten Eingabefelder wurden bereits diskutiert. Nicht sichtbar sind die Ausgabedateien, in denen der gemesse Absolutstrom, der Nettostrom, der Mittelwert und die Standardabweichung des Stroms im Zuge der elektrischen Kontaktprüfung und schließlich die ermittelte Kapazität gespeichert wurden.

3.2.3 Auswertung

Deterministische Fehler

Abgesehen von den Fehlern, die durch Ladungsinjektion und -umverteilung in den Transistoren entstehen, ergab sich bei den Messungen unter dem Stichwort „Signalintegrität“ ein weiterer systematischer Messfehler. Ursache war der strombedingte Spannungsabfall („IR-drop“) auf den widerstandsbehafteten Masseleitungen der einzelnen Zeilen. Zwar hat der Widerstand der Masse normalerweise keinen signifikanten Einfluss auf das Verhalten einer Schaltung, sofern er bei der Layouterstellung klein gehalten wird. Auch im vorliegenden Fall wurde auf eine gewisse Mindestbreite geachtet. Jedoch zeigte sich, dass die Empfindlichkeit der Ladungspumpentechnik so hoch ist, dass selbst kleinste Stromspitzen auf den Masseleitungen zu Spannungsabfäl-

len führte, der sich in einem auflösungslimitierenden Messfehler bemerkbar machte. Die Stromspitzen wurden dabei von den digitalen Signalverstärkern („Buffer“) in jeder Zeile an den Flanken der Steuersignale verursacht.

Zurückblickend ergab sich dieser Schluss folgendermaßen: In Bild 3.27 ist der Verlauf der Kapazität eines Plattenkondensators über alle Zeilen eines der gemessenen Chips zu sehen. Der starke Anstieg wurde zunächst als tatsächliche Kapazitätszunahme aufgrund eines vermuteten starken Gradienten bei der Isolationsschichtdicke interpretiert. Diese Annahme erwies sich als falsch, da es einen systematischen, spannungsabhängigen Effekt gab, der über die Zeilen hinweg abnahm. In Bild 3.28 ist dieser zu erkennen, die vier Kapazitätswerte pro Zeile ergaben sich aus den vier Ausgleichsgeraden („linear fit“), jeweils gebildet aus dem Ladungspumpenstrom über die Frequenz bei vier Werten der Versorgungsspannung (siehe Gleichung 2.41 auf Seite 45). Die Kapazitätswerte in Bild 3.27 entstanden dem Prinzip der Ladungspumpen gemäß aus dem Mittelwert der jeweils vier Punkte in Bild 3.28, so dass der Anstieg über die Zeilen hinweg allein durch das Abflachen der Verbindungslinien zustande kommen musste. Ohne diesen systematischen Effekt hätten die jeweils vier Punkte immer eine horizontale Linie bilden müssen, allein das Rauschen hätte zu geringfügigen Abweichungen führen können, so dass die Punkte zufällig über und unter der Horizontalen verteilt gewesen wären.

Die einzige plausible Erklärung für diesen Effekt ist der strombedingte Spannungsabfall über widerstandsbehaftete Leitungen, hier der Masseleitung. Der Widerstand von Zeile zu Zeile war mit ca. 4 Ohm bereits hoch genug, um maximale Spannungsabfälle von mehreren hundert Millivolt durch den kurzzeitigen Impuls beim Stromverbrauch der Signalbuffer[20] an den Flanken der Steuersignale zu bewirken. Die Spannungsabhängigkeit der Kapazitätswerte in Bild 3.28 war also eine Folge der zunehmenden Verschiebung des Massepotentials, so dass in Gleichung 2.41 ein immer kleiner werdender Spannungswert hätte eingesetzt werden müssen. Der Anstieg des Massepotentials ergab sich dabei aus dem Spannungsabfall über die Leitung, der bei steigender Versorgungsspannung über den größer werdenden Strom zustande kam.

In Bild 3.29 ist ein einfaches Modell der Situation zu sehen. Der Widerstandswert R zwischen jeweils zwei benachbarten Zeilen wird als konstant angenommen, ebenso der Strombedarf I der einzelnen Signalbuffer. Innerhalb einer Zeile gibt es keinen Leitungswiderstand und der zusätzliche Messstrom wird vernachlässigt, da er im Bereich weniger Nanoampere liegt. Da sich die Strombeiträge der Buffer zeilenweise addieren, beträgt der Spannungsabfall U_x in Zeile x $R \cdot I \cdot x$. Bezogen auf die nach außen führende, (widerstandslose) gemeinsame Masseleitung am Summationspunkt I_{sum} beträgt der gesuchte Spannungsabfall in Zeile x schließlich (bei insgesamt N Zeilen):

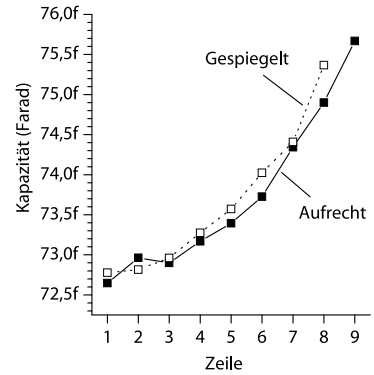


Bild 3.27. Kapazität eines Poly1-Poly2 Plattenkondensators in Abhängigkeit von der Zeile. Der exponentielle Anstieg war viel stärker als durch einen chipweiten Gradienten bei der Oxydschichtdicke zu erklären wäre.

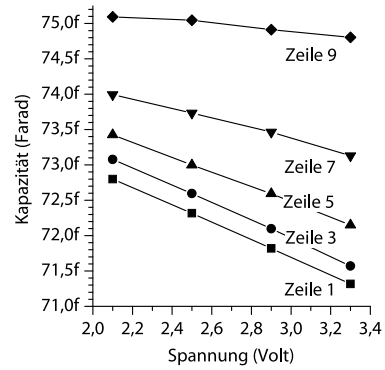


Bild 3.28. Kapazität des Plattenkondensators aus Bild 3.27 über die Versorgungsspannung (ohne die gespiegelten Zeilen).

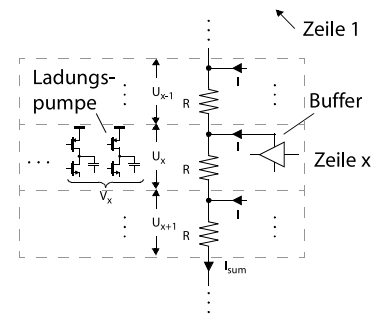


Bild 3.29. Modellierung des zeilenabhängigen Spannungsabfalls über die widerstandsbehaftete Masseleitung.

20. Es kamen insgesamt 68 CMOS-Buffer mit 15-facher Treiberstärke und mittlerem Leistungsverbrauch von 4,5 Mikrowatt pro Megahertz zum Einsatz.

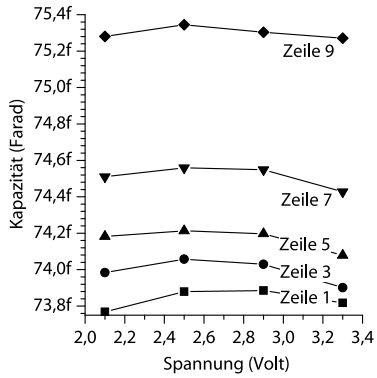


Bild 3.30. Kapazität des Poly1-Poly2 Plattenkondensators nach der Korrektur des Fehlers durch Spannungsabfall. Zum Vergleich siehe Bild 3.28.

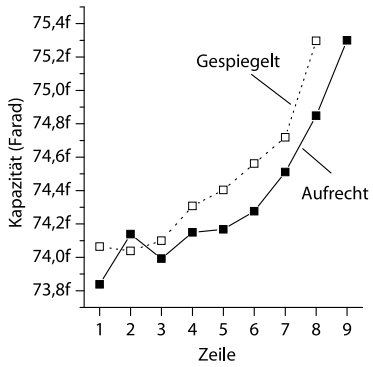


Bild 3.31. Kapazität des Poly1-Poly2 Plattenkondensators aus Bild 3.27 nach der Korrektur des Fehlers durch Spannungsabfall.

$$V_x = \sum_{k=x}^{N} U_k = RI \sum_{k=0}^{N-x} k + x = RI\left[x(N-x+1) + \sum_{k=0}^{N-x} k\right] = \frac{RI}{2}(N^2 - x^2 + N + x) \tag{3.3}$$

Obwohl alle Zeilen durch die Spiegelung an der Horizontalen (bis auf Zeile 9) doppelt vorkommen, der Widerstand innerhalb eines solchen Paares jedoch vernachlässigbar klein ist, beträgt die Gesamtzahl der Zeilen $N = 9$. In der ersten Zeile ist der Spannungsabfall nach Gleichung 3.3 also mit $V_x = 45 \cdot RI$ am größten, in Zeile 9 beträgt er nur $V_x = 9 \cdot RI$.

Da der Stromverbrauch I der Buffer vom zeitlichen Verlauf der Eingangssignale abhängt, wurde die Korrektur des Messfehlers jedoch anstatt mit dem analytischen Modell mithilfe einer Simulation durchgeführt. Die Simulationsergebnisse wurden in Tabellenform gespeichert und zur Korrektur der Messwerte in eine eigens erstelle Analysesoftware (ähnlich der in Bild 3.26 auf Seite 80 dargestellen) eingelesen.

Für die Korrektur selbst blieben die Ströme unangetastet, stattdessen erfolgte sie über die Anpassung der Frequenz, bei der die einzelnen Ströme gemessen wurden:

$$f^{\text{neu}} = f^{\text{alt}} \frac{V_{DD} - \alpha V_x}{V_{DD}} \tag{3.4}$$

Die Konstante α diente dazu, die in der Simulation verwendenten Werte für R und I per Skalierung an die Realität anzupassen, sowie den Einfluss des Spannungsabfalls insgesamt auf die Kapazitätswerte[21]. Bei $\alpha \approx 0{,}4$ war die Spannungsabhängigkeit der Kapazität im wesentlichen beseitigt, wie in Bild 3.30 zu sehen ist (im Abschnitt „Auflösung und Genauigkeit“ auf Seite 84 ff. wird α nochmals genauer beleuchtet).

Das Abflachen der Geraden durch die vier Messpunkte bestätigt also die Annahme, dass der Spannungsabfall über die widerstandsbehaftete Masseleitung für den systematischen Fehler verantwortlich war. Ein Blick auf die *korrigierte* Kapazität des Plattenkondensators in Abhängigkeit von der Zeile in Bild 3.31 zeigt, dass die Kapazitätszunahme wesentlich geringer wurde, aber immer noch vorhanden war. Dies bedeutet, dass es tatsächlich einen Gradienten bei der Dicke der Isolationsschicht über den Chip hinweg gibt.

Auswertesystematik

Im Abschnitt „Klassische Ladungspumpen“ auf Seite 45 ff. wurde bereits erwähnt, dass es mehrere Möglichkeiten gibt, aus den gemessenen Strömen gemäß Gleichung 2.41 die endgültige Kapazität zu berechnen. Entweder wird der Strom über die Frequenz aufgetragen oder die Versorgungsspannung. Aus den Steigungswerten der Geradenschar durch die Messpunkte (ohne Ursprung) ergibt sich im ersten Fall jeweils ein Kapazitätswert pro Spannungswert, im zweiten Fall jeweils ein Wert pro Frequenzwert.

21. V_x kann nicht direkt in Gleichung 2.41 eingesetzt werden, da der Wert nur das Maximum am Umschaltpunkt der Buffer darstellt. Der genaue Zusammenhang ist unbekannt und hängt vom zeitlichen Verlauf von V_x in der aktiven Phase der Ladungspumpe ab.

Wie in den Abbildungen 3.27 bis 3.31 wurde bei allen folgenden Auswertungen immer der Strom über die Frequenz (bei 100 kHz, 900 kHz, 1,7 MHz, 2,5 MHz und 3,3 MHz) als Ausgangspunkt für den Geradenfit verwendet, da die Kapazitätswerte dann eine geringere Streuung aufwiesen, als im anderen Fall. Ursache hierfür ist, dass Frequenzen generell messtechnisch präziser erzeugt und stabil gehalten werden können, als Spannungen und der vorangehend diskutierte Spannungsabfall auf den Masseleitungen einen geringeren Einfluss auf das Endergebnis hatte[22]. Aus den Kapazitätswerten der vier Versorgungsspannungspunkte (bei 2,0 V, 2,4 V, 2,9 V und 3,3 V) wurde dann der Mittelwert gebildet und als Ergebnis gewertet.

Wichtiges Kriterium für die Störanfälligkeit der Messungen ist die Wiederholbarkeit einer Gesamtmessung. Die Breite der Verteilung der ermittelten Kapazität gibt Aufschluss auf den Einfluss von Störungen bzw. Rauschen. Die wichtigste Messung innerhalb einer Zeile der Matrix ist die der sechs unbeschalteten Ladungspumpen, die als Referenz zur Nettostrombildung diente. Ist diese Messung störbehaftet, so wirkt sich dies auf alle anderen Messungen durch den Nettostrom ebenfalls aus.

In Bild 3.32 ist die Verteilung der Kapazität für die erste Spalte der Matrix zu sehen. Fast alle gemessenen Werte liegen beim Mittelwert von 3,997 Femtofarad, nur einige wenige daneben. Die maximale Kapazitätsdifferenz beträgt nur 10,6 Attofarad. Bezogen auf den Mittelwert sind das nur 0,3 Prozent, bei der Standardabweichung von 2,6 Attofarad sogar nur 0,07 Prozent. Zu beachten ist, dass die ermittelte Kapazität nicht als Kondensator an der Ladungspumpe existiert, sondern im wesentlichen durch parasitäre Kapazitäten und eventuelle Leckströme bedingt ist, die es von den übrigen Messungen abzuziehen gilt (Nettostrombildung).

Ähnliche Werte lassen sich auch für einige andere Ladungspumpen bestimmen, so z.B. für die mit einem Plattenkondensator hoher Kapazität beschaltete Ladungspumpe in Spalte 3. In Tabelle 3.4 sind diese Werte gegeben, zusammen mit den auf den Mittelwert bezogenen prozentualen Abweichungen (Standardabweichung und Spanne Min. zu Max.). Bei der Berechnung der mit einem Stern gekennzeichneten Zahlen wurde ein Ausreißer aus den Ausgangsdaten entfernt. Die prozentualen Abweichungen betragen in diesen Fällen weit weniger als ein Prozent, was als sehr gutes Ergebnis angesehen werden kann.

Um Ausreißer bei einer einmaligen Messung zu identifizieren, d.h. ohne die statistische Analyse einer größeren Zahl von Messdurchläufen, wurden zum einen die einzelnen Datenpunkte herangezogen, die den gemessenen Strom eines Spannungs-Frequenz Paares repräsentieren. War der Abstand eines Punktes weit von der zugehörenden Ausgleichsgerade entfernt, so wurde die Messung wiederholt. Entscheidend hierfür war das Ergebnis der visuellen Kontrolle des jeweiligen Schaubildes, wie es von der Testapplikation erstellt wurde (Bild 3.26 auf Seite 80, alternativ hätte ein quantitatives Abstandsmaß festgelegt und automatisch auf Konformität überprüft werden können).

Als zweites Kriterium für die Identifikation von Ausreißern diente die *Spanne* der Kapazitätswerte bei den vier Spannungen, die jeweils aus der Steigung der Ausgleichsgeraden durch die Stromwerte ermittelt wurden, d.h.

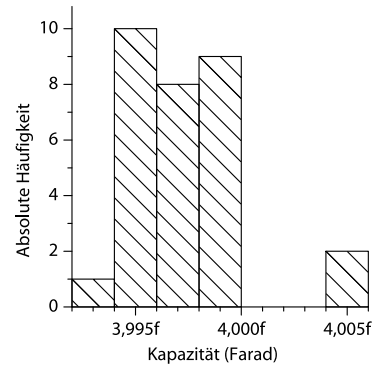


Bild 3.32. Verteilung der Kapazität, wie sie mithilfe der unbeschalteten Ladungspumpe (Referenz) in Spalte 1 bei 30 Wiederholungen ermittelt wird.

Spalte	μ(C)	σ(C)	Spanne(C)
1	3,997 fF	2,6 aF	10,6 aF
		0,07%	0,3%
3	73,60 fF	422 aF	2,29 fF
		0,5%	3,1%
3 (*)	73,53 fF	168 aF	403 aF
		0,2%	0,5%
4	7,84 fF	3,3 aF	11,54 aF
		0,04%	0,1%

Tabelle 3.4. Statistik der über die Ladungspumpen in Spalte 1,3 und 4 ermittelten Kapazität bei 30 Wiederholungen der Messung. In dem mit (*) gekennzeichneten Fall wurde ein Ausreißer entfernt.

22. Bei der Kapazitätsberechnung aus dem Strom bildet die fehlerbehaftete Versorgungsspannung den X-Achsenabschnitt beim linearen Geradenfit und hat damit einen viel größeren Einfluss auf Güte der Regression.

Splt.	$\mu(\Delta C)$ $\mu(\Delta C)/\mu(C)$	P_{75}+3IQR P_{95}+3IQR	max(ΔC)
1	104 aF 2,6%	112 aF 113 aF	107 aF
3	307 aF 0,42%	693 aF *854 aF*	*3,557 fF*
3 (*)	195 aF 0,27%	690 aF 791 aF	429 aF
4	106 aF 1,35%	123 aF 155 aF	141 aF

Tabelle 3.5. Statistik der maximalen Kapazitätsdifferenz ΔC (Spanne, „range“) bei den vier Spannungen. Alle Werte größer als das Perzentil P95 plus dem dreifachen Quartilabstand (IQR) wurden als Ausreißer angesehen (bei den Werten mit (*) wurden diese entfernt).

die maximale Kapazitätsdifferenz ΔC der vier zusammengehörenden Werte, wie sie in beispielsweise Bild 3.30 zu sehen sind. Lag die Differenz deutlich über einem bestimmten Wert, so wurde die Messung ebenfalls verworfen und wiederholt. In Tabelle 3.5 sind die statistischen Kenngrößen der maximalen Kapazitätsdifferenz in den bisher betrachteten drei Fällen angegeben. Es zeigte sich, dass die häufig verwendete Definition für extreme Ausreißer, nämlich als jene Werte, die über der Summe aus dem oberen Quartil und dem dreifachen Quartilabstand liegen, zu konservativ wäre, da z.B. im Falle von Spalte 4 bereits drei Messungen als Ausreißer anzusehen wären (mit 141 aF, 138 aF und 126 aF), obwohl der auf den Mittelwert bezogene Fehler des dazugehörenden Endergebnisses selbst beim größten Ausreißer nur 0,08 Prozent beträgt. Aus diesem Grund wurde stattdessen für alle weiteren Analysen die Perzentile P95 plus dem dreifachen Quartilabstand als Grenze zu Identifikation von Ausreißern festgelegt.

Auflösung und Genauigkeit

Die Tatsache, dass die Verteilung der Kapazitätsdifferenz bei Spalte 4 in der Tabelle 3.5 nach herkömmlicher Definition bereits drei extreme Ausreißer aufweist, deutet auf einen systematischen Effekt hin, der bisher nicht berücksichtigt wurde. Da die Werte in der Tabelle auf der Grundlage *derselben* Ladungspumpe bei 30 Wiederholungen, also *über die Zeit*, erstellt wurde, scheidet das Layout als Ursache aus. Stattdessen ist die zeitliche Änderung der Güte des elektrischen Kontakts eine mögliche Ursache, bedingt etwa durch ein Zusammenziehen der Leiterplatte oder des Klebers im Chipgehäuse.

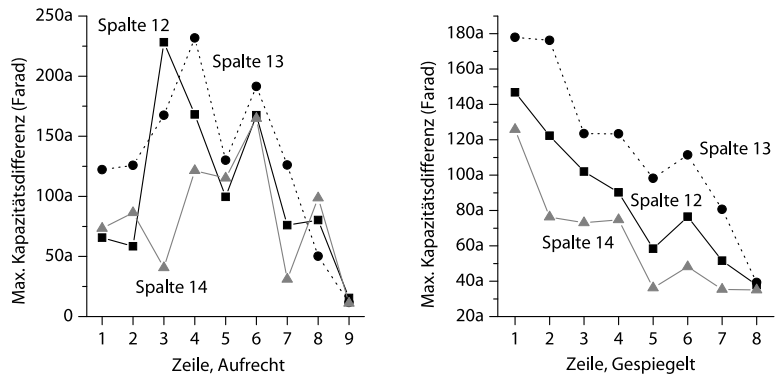


Bild 3.33. Maximale Differenz (Spanne) der Kapazitätswerte bei den vier Spannungen, aufgetragen über die Zeile. Der Skalierungsfaktor α betrug ca. 0,4. Die Kapazität der drei Spalten beträgt im Mittel (über alle Zeilen von Chip Nr. 20) 8,34 fF, 11,68 fF und 7,31 fF (Spalte 12 – 14).

Um zu untersuchen, in welchem Maße die Wahl des Faktors α in Gleichung 3.4 die maximale Kapazitätsdifferenz ΔC und damit die Unbestimmtheit (Genauigkeit) des Endwertes beeinflusst, hilft zunächst ein Blick auf Bild 3.33. Darin ist die maximale Kapazitätsdifferenz dreier Ladungspumpen *über die Zeilen* der Matrix aufgetragen, getrennt nach der Orientierung. Im Falle der aufrechten Zeilen ist kaum eine Systematik zu erkennen (lediglich in Zeile 9, in der die Werte zusammenfallen). Ganz im Gegenteil dazu verhalten sich die gespiegelten Zeilen. Es ist eine deutliche Abnahme der Werte von Zeile 1 bis 8 zu erkennen, eine Tendenz, die auch bei allen anderen Ladungspumpen vorhanden ist (nicht gezeigt).

Die Zeilenabhängigkeit im rechten Graphen deutet darauf hin, dass die Wahl von α nicht genau genug sein könnte, um den Spannungsabfall über die Masseleitung komplett zu kompensieren oder das einfache Modell (siehe Bild 3.29) die Realität nur unzureichend wiederspiegelt. Die erste Möglichkeit kann überprüft werden, indem α nicht als Konstante angesehen wird, sondern pro Ladungspumpe und Zeile der optimale Wert ermittelt wird, also der Faktor, bei dem die Kapazitätsdifferenz minimal wird. In Bild 3.34 ist das Ergebnis einer solchen Optimierung zu sehen.

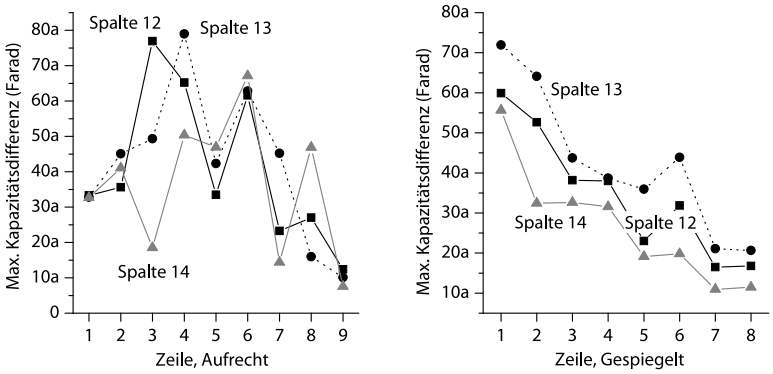


Bild 3.34. Maximale Differenz der Kapazitätswerte über die Zeile bei optimalem Skalierungsfaktor α. Statt einer Konstante ist der Faktor nun eine *Funktion* der Zeile und Spalte (Ladungspumpe). Die Kapazität der drei Spalten beträgt im Mittel (über alle Zeilen von Chip Nr. 20) 8,44 fF, 11,83 fF und 7,38 fF (Spalte 12, 13, 14).

Die Systematik ist weiterhin vorhanden, qualitativ hat sich an den Graphen nichts geändert. Damit scheidet α als alleinige Ursache aus, so dass die ungenügende Modellierung des Spannungsabfalls (oder anderer unbekannter Layouteffekte) wahrscheinlich ist: Die gespiegelten Ladungspumpen wurden im Modell wie die aufrechten Zeilen behandelt und der Widerstand dazwischen (< 1 Ω) vernachlässigt. Auf diese Weise erklärt sich die noch vorhandene Zeilenabhängigkeit im rechten Graphen.

Quantitativ hat sich in Bild 3.34 indes sehr wohl etwas geändert: Die Kapazitätsdifferenzen sind allesamt sehr viel geringer, als in Bild 3.33, eine Beobachtung, die auch für alle anderen Ladungspumpen (Spalten) gilt: In Tabelle 3.6 sind die Grenzen für die P95-Ausreißer auf der Grundlage eines kompletten Chips (Die 20) aufgelistet, sowie der 30-fachen Wiederholung der Messung in Spalte 4, Zeile 1. Tabelle 3.7 zeigt die veränderte Statistik unter Anwendung des Ausreißerkriteriums.

Insgesamt hat sich die Spanne verringert, so dass die Unbestimmtheit des Endwertes im Mittel nur rund 40 Attofarad beträgt. Die maximale Kapazitätsdifferenz beträgt nun 147 Attofarad. Unterstellt man für die anderen Chips ähnliche Werte, so folgt der Schluss, dass der Absolutwert des Endergebnisses des gesamten Messverfahrens in der vorliegenden Variante und Durchführung nicht genauer sein kann als ca. 150 Attofarad, d.h. der ermittelte Wert weist einen unbekannten, konstanten Offset auf. Da die sich die Werte insgesamt durch die Wahl des optimalen Skalierungsfaktors deutlich verbessert haben, wurden für alle folgenden Analysen in dieser Arbeit analog vorgegangen.

Spalte	P_{75}+3IQR # Ausreißer	P_{95}+3IQR # Ausreißer
1 – 52 Aufrecht	201 aF 5 von 468	254 aF 5 von 468
1 – 52 Gespiegelt	143 aF 5 von 416	189 aF 3 von 416
4, Zeile 1 (30 Wdh.)	60 aF 1 von 30	70 aF 1 von 30

Tabelle 3.6. Grenzen zur Identifikation von Ausreißern bei optimalem Skalierungsfaktor α (Chip 20). Grundlage ist die Spanne der intermediären Kapazitäten über alle Zeilen bzw. bei 30 Wiederholungen (unten).

Spalte	Mittel	Stdabw.	Max.
1 – 52 Aufrecht	40 aF	28 aF	128 aF
1 – 52 Gespiegelt	37 aF	26 aF	147 aF
4 (30 Wdh.)	48 aF	4 aF	59 aF

Tabelle 3.7. Statistik der Kapazitätsspanne/-differenz *ohne* die P95-Ausreißer (vgl. Tabelle 3.6).

Strom-bereich	Auflösung	Genauigkeit ±(% Anz. + Amp.)
1 µA	1 pA	0,05% + 100 pA
100 nA	100 fA	0,05% + 30 pA
10 nA	10 fA	0,05% + 1 pA
1 nA	3 fA	0,05% + 100 fA

Tabelle 3.8. Auflösung und Genauigkeit des Source-Meters je nach Messbereich (aus dem „4200-SCS QuickStart Manual", Keithley Instruments). Zur Genauigkeitsangabe kommt noch ein Faktor 1 – 5 hinzu, je nach Umgebungstemperatur und rel. Feuchtigkeit.

Chip	Min.	Typ./µ	Max
Die 15	74,7 fF	75,4 fF	76,3 fF
Die 17	76,0 fF	76,9 fF	78,2 fF
Die 18	75,0 fF	75,6 fF	76,4 fF
Die 19	75,7 fF	76,3 fF	77,0 fF
Die 20	73,4 fF	74,1 fF	74,9 fF
Alle 5	73,4 fF	75,6 fF	78,2 fF
Rechnung	78,7 fF	86,3 fF	95,7 fF

Tabelle 3.9. Spanne und Mittelwert der gemessenen Kapazität des Poly1-Poly2 Plattenkondensators im Vergleich zur Rechnung.

Anders als die Genauigkeit des Absolutwertes ist die Kapazitäts*auflösung* mit einer Standardabweichung von wenigen Attofarad sehr hoch, so dass relative Aussagen über Größenverhältnisse viel genauer getroffen werden können: Aus Tabelle 3.4 wird deutlich, dass bei Wiederholung der Messungen keine nennenswerten Schwankungen auftreten, einzig beim Poly1-Poly2 Plattenkondensator (Spalte 3) steigt die Standardabweichung auf 0,2 Prozent. Der Grund für den Anstieg liegt im Wechsel des Messbereichs beim Source-Meter. Der resultierende größere Strom (17 nA – 0,8 µA) wird vom Source-Meter mit geringerer Auflösung gemessen (siehe Tabelle 3.8) als die schwächeren Ströme der kleinen Kapazitäten in den anderen Spalten (0,8 nA – 0,3 µA).

Erste Ergebnisse

Der Plattenkondensator in Spalte 3 weist eine im Vergleich mit allen übrigen Ladungspumpen recht hohe Kapazität auf. Die meisten der getesteten Strukturen bewegen sich im Bereich von 3 bis 26 Femtofarad, der Plattenkondensator liegt mit durchschnittlich 75,6 Femtofarad (ermittelt über fünf Testchips) deutlich darüber, ohne dabei von der Geometrie entsprechend größer zu sein. Der Grund liegt im Aufbau, genauer in der Dicke der Isolationsschicht zwischen den beiden Platten. Diese ist prozesstechnisch für den Einsatz in Kondensatoren hoher Kapazität auf eine geringe Dicke hin optimiert. Als Platten fungieren 100 μm^2 große, quadratische Flächen aus polykristallinem Silizium („poly1" und „poly2"), die durch eine dünne Schicht aus Siliziumdioxid (Quarz, sog. „inter metal dielectric", IMD) als Isolator voneinander getrennt sind.

In die elektrische Kapazität eines solchen Plattenkondensators geht die Dicke t_{ox} der Isolationsschicht ein. Diese wird bei der Herstellung fortwährend bestimmt und auf Einhaltung der Spezifikation überprüft (siehe hierzu "Ausbeute (Yield)" auf Seite 28). Liegen die Werte außerhalb der Grenzen, wird der Wafer verworfen. Bei dem in dieser Arbeit verwendeten 0,35 µm Prozess garantiert der Hersteller eine Dicke von 37 bis 45 Nanometern, der typische Wert liegt mit 41 Nanometern in der Mitte. Damit berechnet sich die Kapazität im typischen Fall zu 86,3 Femtofarad. Die Werte für die beiden anderen Fälle (minimale und maximale Isolationsschichtdicke) sind in der letzten Zeile in Tabelle 3.9 gegeben.

Zu diesen Werten kommt noch die Kapazität der Verbindung zur oberen Poly2-Platte über eine Metallzuleitung („metal1"), die auf einer Fläche von 0,5 Quadratmikrometern über der unteren Poly1-Platte verläuft. Die Kapazität dieses Stücks ist mit 3,76 Attofarad jedoch vernachlässigbar klein. Auch die der Leitung zur anderen Kondensatorplatte ist im Bereich von wenigen Attofarad.

Bezieht man noch die Kapazitätswerte aufgrund von Streufeldern am Rand und seitlich zur unteren Anschlussleitung hin in die Rechnung mit ein, so *erhöht* sich der Wert nur noch, kleiner kann er nicht werden. Der höchste tatsächlich gemessene Wert von 78,2 Femtofarad kann somit selbst bei der maximalen Isolationsschichtdicke von 45 Nanometern – also minimaler Kapazität – nicht erreicht werden, auch nicht, wenn man die tatsächliche Fläche der Platten aufgrund von Rundungsverlusten an den Ecken etwas kleiner ta-

xiert. Die einzige Schlussfolgerung ist, dass die tatsächliche Dicke des Dielektrikums größer ist, als die Spezifikation erlaubt. Der Wafer hätte eigentlich aussortiert werden müssen!

In Bild 3.35 ist der Verlauf der Kapazität des Plattenkondensators über die Zeilen der Matrix bei drei verschiedenen Testchips zu sehen. Die Unterscheidung in die Kategorien „aufrechte Orientierung“ (durchgezogene Linien) und „nach unten gespiegelt“ (gestrichelt) ergibt erst später bei den speziellen Kapazitätsstrukturen (siehe Abschnitt „Erzeugung der 3D-Cluster“ auf Seite 60) Sinn, bei spiegelsymmetrischen Strukturen wird sie nur insofern benötigt, als die gespiegelten Zeilen zwischen je zwei aufrechten Zeilen liegen und damit die ortsbezogene Abhängigkeit der Kapazität verdeutlichen.

Diese Ortsabhängigkeit lässt sich nur durch einen Gradienten bei der prozesstechnischen Erzeugung der Isolationsschicht erklären. Von der ersten Zeile bis zu Zeile 9 scheint es eine kontinuierliche Abnahme der Dicke des Dielektrikums zu geben, so dass die Kapazität zunimmt.

* * *

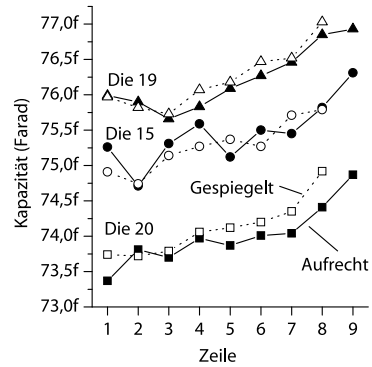


Bild 3.35. Kapazität des Poly1-Poly2 Plattenkondensators in den Zeilen 1 bis 9 bei drei verschiedenen Testchips. Die durchgezogenen Linien stellen die Werte für die Zeilen mit aufrechter Orientierung dar, die gestrichelten Linien die der nach unten gespiegelten Zeilen.

3.3 Die Schlüsselelektronik

Im Folgenden wird das schaltungstechnische Prinzip der Auswerteelektronik vorgestellt, mit der aus dem Größenverhältnis von jeweils zwei Clustern ein binärer „größer"- bzw. „kleiner"-Wert erzeugt werden kann. Aus einer entsprechenden Zahl an Clusterpaaren kann damit eine Bitfolge generiert werden, die als Grundlage für kryptografische Schlüssel verwendet werden kann.

3.3.1 Anforderungsprofil

Zunächst kann man sich klarmachen, dass die Auswerteelektronik eine Reihe von Bedingungen erfüllen muss, um für den Einsatzzweck geeignet zu sein. Diese Anforderungen ergeben sich aus dem Anwendungsgebiet des Cluster-Konzepts als digitaler, geheimer Schlüssel auf einem Mikrochip:

1. Integrierbarkeit.
   Die Auswerteelektronik muss komplett auf einem Chip integrierbar sein, so dass keine externen Geräte wie Strommesser und Signalgeneratoren benötigt wird.
2. Sicherheit.
   Mit der Sicherheit steht und fällt das ganze Konzept der Cluster. Bietet die Auswerteelektronik keine ausreichende Sicherheit gegen Attacken wie Reverse-Engineering oder Stromverbrauchsanalyse, so bedeutet dies die Preisgabe des Schlüssels.
3. Robustheit.
   Die Abhängigkeit der Schaltung von Betriebsparametern wie Spannung und Temperatur sollte so gering wie möglich sein, um ein breites Einsatzfeld zu ermöglichen, vom Bereich „Eingebette Systeme" über „Anwendungsspezifische ICs" (ASICs), bis hin zum Prozessor in Standardsystemen.
4. Schlüssellänge.
   Die Anzahl der Bits des Schlüssels ist eine wichtige Eigenschaft, da sie seinen Informationsgehalt bestimmt. Je höher dieser ist, desto größer ist der Aufwand, den ein Angreifer treiben muss, um an die geschützte Information, also die Bitfolge, durch Ausprobieren („Brute-Force") zu gelangen.
5. Sonstiges.
   Eine Reihe weiterer, untergeordneter Anforderungen sind denkbar, die Liste kann je nach dem konkreten Einsatzgebiet beliebig fortgesetzt werden. Als Beispiele sind die einfache Implementierbarkeit, der geringe Strom- und Flächenverbrauch, sowie geringe Umsetzungs- und Produktionskosten zu nennen.

3.3.2 Schaltungsprinzip

Zur Messung sehr kleiner Kapazitäten sind eine Reihe von Schaltungstechniken aus der Literatur bekannt. Die wichtigste von ihnen wurde bereits im Abschnitt "Kapazitätsmessung" auf Seite 45ff. vorgestellt. Trotz des einfachen Prinzips und der hohen Messgenauigkeit sind die Ladungspumpen alleine nicht in der Lage, als Auswerteelektronik für die Cluster zu dienen. Hauptmangel ist der Bedarf an externem Laborgerät wie Source-Meter und Signal-

generatoren. Zudem liefern Ladungspumpen eine Information über den absoluten Wert einer elektrischen Kapazität und treffen keine „größer“- bzw. „kleiner“-Entscheidung. Um aus einer absoluten Kapazitätsmessung eine relative Aussage über das Größenverhältnis zweier Kondensatoren zu machen, müssten die Kapazitätswerte digitalisiert, verglichen und in ein 0/1-Ergebnis umgewandelt werden. Dies stellt ein relativ aufwändiges Verfahren für ein einzelnes Bit dar.

Stattdessen wurde die grundlegende Technik des Pumpens von Ladung in die zu messenden (zu vergleichenden) Kapazitäten angepasst und weiterentwickelt (siehe Bild 3.36). Statt des externen Strommessgeräts (Source-Meter) fungiert nun ein (relativ) großer Messkondensator C_{L}, der über das Signal „load“ mit Ladungsträgern gefüllt wird. Diese Ladung wird anschließend sukzessive auf den ersten unbekannten Cluster C_1 (siehe Bild 3.37) übertragen, so dass die Spannung V_{Qin} über C_{L} analog zum Ladungsverlust nach und nach absinkt. Dieses Pumpen geschieht über den oberen PMOS-Transistor in Bild 3.37 mittels „Qinj“, das Auswählen des Clusters C_1 über das Signal „swC1“. Nach einer gewissen Zahl n an Schritten erreicht die Spannung einen Wert, der von der Größe des Clusters abhängt. Verfährt man mit dem zweiten Cluster C_2 in derselben Weise, so lassen sich die beiden Spannung über einen Komparator vergleichen und so eine „größer“- bzw. „kleiner“-Entscheidung treffen.

Da die beiden Spannungen zunächst nicht *gleichzeitig* an zwei verschiedenen Knoten anliegen, sondern *nacheinander* an ein und demselben Punkt (nämlich V_{Qin}), müssen sie noch zwischengespeichert werden. Diese Aufgabe übernehmen (über einen Source-Folger zu Entkoppelung) zwei Abtast-Halteglieder (Sample & Hold). Wird in der ersten Phase der rechte PMOS-Transistor über das Signal „swCshP“ eingeschaltet, so überträgt sich die Spannung V_{out} des ersten Clusters auf den Speicherkondensator C_{shP}. In der zweiten Phasen wird die Spannung des zweiten Clusters auf C_{shN} gespeichert, indem nun statt des rechten Transistors der linke PMOS (über „swCshN“) eingeschaltet wird.

Die Funktion des Source-Folgers in Bild 3.36 ist die Entkoppelung des Knotens V_{Qin} von den beiden Sample & Hold-Gliedern, um ihn kapazitiv so wenig zu belasten wie nötig. Andernfalls würde die Spannung an diesem Punkt sofort absacken, sobald eines der beiden Abtast-Halteglieder aktiviert werden würde, da sich die auf C_{L} gespeicherten Ladungsträger auf die großen Kondensatoren C_{shP} bzw. C_{shN} verteilen würden. Darüber hinaus isoliert der Source-Folger den Knoten von den zusätzlichen Ladungen, die von den Transistoren der Sample & Hold-Gliedern während des Umschaltens in die angeschlossenen Knoten injiziert wird. Neben dem eingezeichneten Source-Folger existieren in diesem Schaltungsentwurf noch jeweils ein zusätzlicher Source-Folger zwischen den Ausgängen der beiden S&H-Glieder und den Eingängen des Komparators. Diese wurden der Übersichtlichkeit in Bild 3.36 weggelassen, sie sind mit dem gezeigten identisch.

Die Funktion der drei NMOS-Transistoren (Bild 3.37), die über das Signal „clear“ gesteuert werden, besteht im Entfernen der Ladungsträger, die bei jedem Pumpvorgang auf den internen Knoten und die Cluster einer Zelle aufgebracht werden. In der Kombination aus dem PMOS-Pumptransistor, dem zu messenden Kondensator (Cluster) *und* dem dazugehörenden NMOS-Transistor erkennt man die ursprüngliche Ladungspumpe (CBCM), wie sie in Bild 2.14 auf Seite 45 zu sehen ist. Jede Zelle entsprechend Bild 3.37 stellt ge-

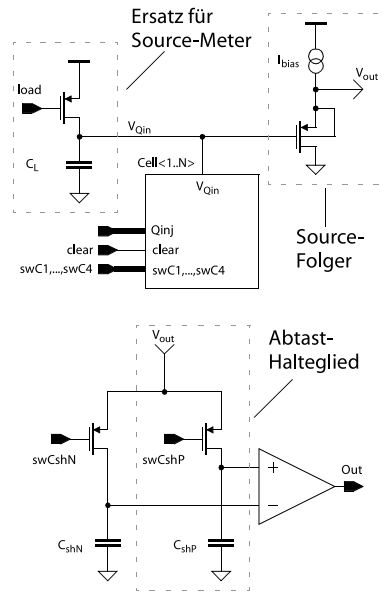


Bild 3.36. Auswerteelektronik zur Erzeugung eines binären Wertes aus dem Größenverhältnis der Kapazitäten in den Zellen. Das ausgegebene Bit steht für „größer“ bzw. „kleiner“ .

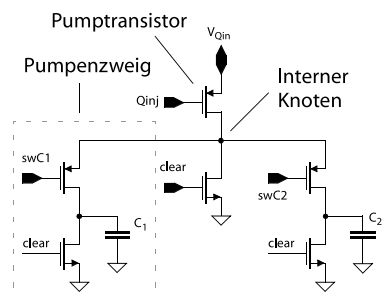


Bild 3.37. Eine Zelle für zwei Cluster (C_1 und C_2). Weitere Kapazitäten können angeschlossen werden, indem jeweils ein zusätzlicher Pumpenzweig an den internen Knoten angehängt wird.

wissermaßen eine Kombination aus zwei Ladungspumpen dar, die über die beiden PMOS-Schalter (via „swC1“ und „swC2“) ausgewählt bzw. aktiviert werden können. Jeder dieser beiden Ladungspumpen verfügen über den gemeinsamen Pumptransistor.

Anzahl Bits

Weitere Pumpenzweige können auf diese Weise an den Pumptransistor bzw. den internen Knoten angeschlossen werden. Praktikabel sind beispielsweise vier Zweige, mit denen sich insgesamt sechs Clusterpaare bilden und damit sechs (korrelierte) Bits gewinnen lassen. Allgemein gilt für die Anzahl B der Bits, die bei M Pumpzweigen ermittelt werden können:

$$B = \binom{M}{2} = \frac{1}{2} \cdot \frac{M!}{(M-2)!} = \frac{1}{2} \cdot \frac{\prod_{k=1}^{M} k}{\prod_{k=1}^{M-2} k} = \frac{1}{2}(M-1)M \tag{3.5}$$

Diese Gleichung ergibt sich direkt aus der Kombinatorik, deshalb wird auf den Beweis verzichtet. Die Anzahl der Cluster ist zwar prinzipiell unbeschränkt, sollte in der Praxis jedoch nicht höher liegen als vier bis sechs. Der Grund liegt im Verlust der Messgenauigkeit bzw. Auflösung, da die parasitäre Kapazität des internen Knoten von der Drain-Kapazität der angeschlossenen Transistoren und der Länge der Verbindungsleitung abhängt. Jeder weitere Transistor führt damit zu einer Zunahme dieser parasitären Kapazität, die unvermeidbarerweise mit den Clustern mitgemessen wird.

3.3.3 Eigenschaften

Im Folgenden sollen nun einige wesentliche Eigenschaften der Schaltung in Bild 3.36 bzw. Bild 3.37 untersucht werden, der Einfachheit halber immer unter der Annahme, dass nur zwei Pumpenzweige existieren. Ferner soll angenommen werden, dass nur eine der beiden Kapazitäten unbekannt sei, die andere dagegen vollständig bekannt. Ohne Beschränkung der Allgemeinheit sei $C_x = C_1$ der unbekannte Cluster und $C = C_2$ die vollständig bekannte Kapazität. Damit lässt sich folgender Satz formulieren:

Sei x das Verhältnis C_x/C zwischen zwei Kapazitäten zu vergleichenden Kapazitäten und l der Quotient aus dem Ladekondensator C_L und C. Dann beträgt die Spannungsdifferenz am Komparator nach n Pumpzyklen:

$$D(n, l, x) = V_{DD} \cdot l^n \cdot ((l+x)^{-n} - (l+1)^{-n}) \tag{3.6}$$

BEWEIS. In der Ladephase wird C_L auf V_{DD} aufgeladen, wodurch die Ladungsmenge $Q_L = C_L V_{DD}$ gespeichert wird. In der nächsten Phase wird ein Teil dieser Ladung auf C übertragen, indem C und C_L parallelgeschaltet werden. Die Spannung über der resultierenden Gesamtkapazität beträgt damit

$$V = \frac{Q_L}{C_L + C} = \frac{C_L V_{DD}}{C_L + C} = \frac{ClV_{DD}}{Cl + C} = V_{DD}\frac{l}{l+1}. \tag{3.7}$$

Nach n Pumpzyklen beträgt der Spannungsabfall

$$V(n) = V_{DD}\left(\frac{l}{l+1}\right)^n. \tag{3.8}$$

Verfährt man mit dem unbekannten Cluster C_x in der gleichen Weise, so beträgt die Spannung am Ende von n Pumpzyklen

$$V_x(n) = V_{DD}\left(\frac{l}{l+x}\right)^n. \tag{3.9}$$

Die führt schließlich zu

$$D(n,l,x) = V_x(n) - V(n) = V_{DD}\cdot l^n \cdot ((l+1)^{-n} - (l+x)^{-n}). \tag{3.10}$$

Womit der Satz bewiesen wäre.

Maximale Auflösung

Intuitiv scheint es nahezuliegen, dass D für $n \approx l$ das Maximum erreicht. Der Plot in Bild 3.38 unterstützt diese Annahme, er zeigt den Betrag der Spannungsdifferenz D für drei verschiedene Werte von l als Funktion der Pumpzyklen n. Das Verhältnis x der Kapazitäten beträgt 0,1 Prozent und die Versorgungsspannung V_{DD} wurde auf 1 Volt gesetzt.

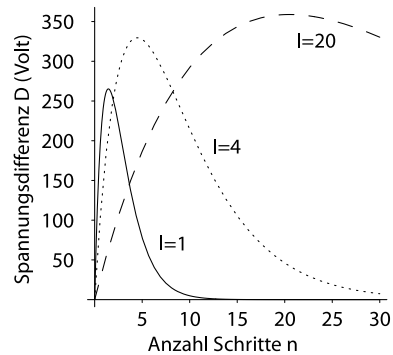


Bild 3.38. Betrag der Spannungsdifferenz am Kondensator bei einem Kapazitätsverhältnis x von 0,1% und $V_{DD} = 1$ Volt.

Zur analytischen Maximumsbestimmung, muss Gleichung 3.10 zunächst nach n abgeleitet werden:

$$\frac{\partial}{\partial n}D(n,l,x) = V_{DD}l^n\left[(l+x)^{-n} - (l+1)^{-n}\right]\ln(l) + V_{DD}l^n\left[\frac{\ln(l+1)}{(l+1)^n} - \frac{\ln(l+x)}{(l+x)^n}\right] \tag{3.11}$$

Gleichsetzen mit Null liefert:

$$\frac{\partial}{\partial n}D(n,l,x) = 0 \qquad \text{gdw.}$$
$$\frac{\ln(l)}{(l+x)^n} - \frac{\ln(l+x)}{(l+x)^n} = \frac{\ln(l)}{(l+1)^n} - \frac{\ln(l+1)}{(l+1)^n} \tag{3.12}$$

Der Faktor $V_{DD}l^n$ verschwindet, da angenommen werden kann, dass weder V_{DD} noch l Null sind. Durch Anwendung der Rechenregeln des Logarithmus erhält man:

$$(l+1)^n \ln\left(\frac{l}{l+x}\right) = (l+x)^n \ln\left(\frac{l}{l+1}\right) \tag{3.13}$$

Um Gleichung 3.13 nach n aufzulösen, bietet sich der Einsatz der Logarithmusfunktion an, so dass der Exponent auf beiden Seiten in einen Vorfaktor verwandelt wird. Diese Vorgegensweise ist legitim, wenn das Argument des Logarithmus nie negativ wird. Mit $l \gg 1$ und $x > 0$ folgt sowohl für den Ausdruck $l/(l+x)$ auf der linken Seite, als auch für $l/(l+1)$ auf der rechten Seite, dass die Logarithmusfunktionen in Gleichung 3.13 negative Werte liefern. Da $(l+1)^n$ und $(l+x)^n$ positiv sind, folgt in beiden Fällen, dass der Gesamtausdruck auf jeder Seite negativ ist. Multipliziert man Gleichung 3.13 auf beiden Seiten mit -1, darf der Logarithmus jeweils angewendet werden:

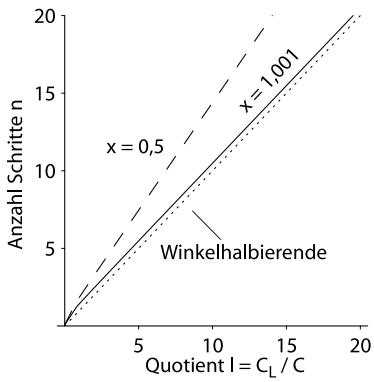


Bild 3.39. Maximum der Spannungsdifferenz $D(n, l, x)$ in Abhängigkeit von l. Für typische werte von $x \approx 1$ liegt das Maximum bei $n \approx l$ ($V_{DD} = 1$ Volt).

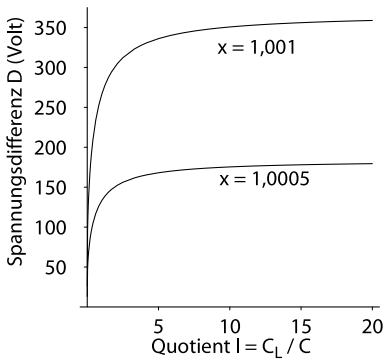


Bild 3.40. Verlauf des Maximums der Spannungsdifferenz $D(n, l, x)$ in Abhängigkeit von l ($V_{DD} = 1$ Volt). Bereits ab einem im Vergleich zu C zehnmal größeren Kondensator C_L nimmt die Differenz kaum noch nennenswert zu.

$$\ln(l+1)^n + \ln\left[\ln\left(\frac{l}{l+x}\right)\right] = \ln(l+x)^n + \ln\left[\ln\left(\frac{l}{l+1}\right)\right] \tag{3.14}$$

Schließlich liefert Auflösen nach n das gewünschte Endergebnis:

$$n = \ln\left[\frac{\ln\left(\frac{l}{l+1}\right)}{\ln\left(\frac{l}{l+x}\right)}\right] \ln\left(\frac{(l+1)}{(l+x)}\right)^{-1} \tag{3.15}$$

Dieser Ausdruck muss das gesuchte Extremum sein, da es sich um die einzige Nullstelle handelt. Dass es sich dabei um ein Maximum handeln muss, wird aus dem Plot in Bild 3.38 ersichtlich, auf den analytische Beweis über die zweite Ableitung wird hier verzichtet (da diese sehr unhandlich wird).

Trägt man Gleichung 3.15 über den Quotienten l auf, so erhält man den Graphen in Bild 3.39. Typischerweise sind die Kapazitätswerte der beiden Cluster C_x und C fast gleich, so dass x ungefähr Eins beträgt. In diesem Fall stimmt die eingangs geäußerte Vermutung, dass für $n \approx l$ das Maximum der Spannungsdifferenz D am Komparator erreicht wird. Zum Vergleich ist in Bild 3.39 die Winkelhalbierende eingezeichnet, sie entspräche dem Ergebnis $n = l$.

Durch Einsetzen des Maximums aus Gleichung 3.15 in den Ausdruck für die Spannungsdifferenz D (Gleichung 3.10) erhält man den Graphen in Bild 3.40. Zu sehen ist der Verlauf des Betrags der Spannungsdifferenz in Abhängigkeit von dem Quotienten l aus dem Messkondensator C_L und der als bekannt angenommenen Kapazität C. Die Kurve flacht sehr schnell ab, so dass bereits bei einem zehnmal größeren Messkondensator die maximal erreichbare Spannungsdifferenz fast vollständig erreicht wird. Trotzdem sind Werte von $l = 20 \ldots 500$ empfehlenswert, da das Rauschen der Spannung am Kondensator und dadurch an den Eingängen des Komparators durch eine große Kapazität vermindert wird. Noch höhere Werte sind, auf der anderen Seite, nicht sinnvoll, da sie in der Regel mit sehr hohem Platzbedarf einhergehen.

Im Rahmen diese Arbeit wurde ein Testchip für die bereits erwähnte 0,35 µm Technologie entworfen, der über die in Bild 3.36 gezeigte Schaltung verfügt. Zum Einsatz kam ein Messkondensator mit der Kapazität $C_L = 4$ pF und – bei Clustergrößen von $C = 4 \ldots 12$ fF – einem Quotienten von $l = 333 \ldots 1000$. Aus dem Plot in Bild 3.40 entnimmt man in diesem Fall eine Spannungsdifferenz von ca. $V_{DD} \cdot 360\ \mu \approx 1{,}2$ mV am Komparator, die es zu detektieren gilt, falls man von einem Clusterverhältnis von 0,1 Prozent bei einer Spannungsversorgung von 3.3 Volt ausgeht.

Der Komparator

Zunächst kann man sich überlegen, dass die Frage, welcher der beiden Kapazitäten in Bild 3.37 größer ist, gleichbedeutend ist mit der Frage, ob der Quotient aus C_1 und C_2 größer oder kleiner Eins ist. Die Entscheidung trifft der Komparator anhand des Spannungsunterschieds D aus Gleichung 3.10, wie sie im vorangehenden Abschnitt hergeleitet wurde. Ein solcher Komparator oder Diskriminator wird typischerweise als Differenzverstärker realisiert,

wie er in Bild 3.37 zu sehen ist. Es handelt sich dabei um eine Schaltungsvariante mit gefalteter Kaskode („folded cascode“), die auch auf dem Testchip zum Einsatz kam (es sei auf die einschlägige Literatur verwiesen).

Zwar gibt es eine Fülle von weiteren Möglichkeiten, einen solchen Verstärker zu implementieren, jedoch haben alle eine wesentliche gemeinsame Eigenschaft: Der Umschaltpunkt des Komparators liegt nicht, wie im idealen Fall, bei einer Spannungsdifferenz von $D = V_+ - V_- = 0$, sondern weist eine – quasi eingebaute – Spannungsdifferenz („offset“) auf, die von der Ungleichheit der Transistoren bei der Herstellung (siehe “Prozessstreuung und Mismatch” auf Seite 22) herrührt. Dieser Offset kann durch geeignete Kompensationstechniken minimiert werden. Falls, wie auf dem Testchip, eine solche Maßnahme fehlt, kann er durch Vertauschen der Eingänge detektiert werden. Ändert sich das Ergebnis durch die Vertauschung nicht, so ist die Spannungsdifferenz kleiner als der Offset.

In diesem Fall kann der Offset durch Anpassen der Schrittweite n in den beiden Phasen reduziert werden. Dies wird durch eine ganzzahlige Konstante α realisiert, die auf die Schrittweite in einer der Phasen aufgeschlagen wird. Sie wird zu Beginn auf einen mehr oder weniger beliebigen Wert gesetzt, z.B. $\alpha_0 = 5$. Die Vorgehensweise ist dabei die Folgende:

1. Offsetermittlung, erste Pumphase.
   Das rechte S&H-Glied in Bild 3.36 („+“-Eingang, Signal „swCshP“) wird aktiviert, sowie ein beliebiger Pumpenzweig in Bild 3.37, idealerweise ein Zweig einer Zelle, bei dem die angeschlossene Kapazität C bekannt ist (z.B. Plattenkondensator). Nun werden $n_1 = l = C_L/C$ Pumpzyklen durchgeführt.
2. Offsetermittlung, zweite Pumphase.
   Das linke S&H-Glied („-“-Eingang, Signal „swCshN“) wird aktiviert, sowie *derselbe* Pumpenzweig wie in der ersten Phase. Die Anzahl der Pumpzyklen beträgt $n_2 = n_1 + \alpha$.
3. Offsetermittlung, Diagnose.
   Ist das Ergebnis positiv (Komparator liefert Ergebnis „1“), so muss α erniedrigt werden ($\alpha = \alpha - 1$) und die Offsetermittlung wiederholt werden (zurück zu Punkt 1). Andernfalls ist die Offsetermittlung beendet und der eigentliche Messvorgang kann gestartet werden.
4. Messvorgang, erste Pumphase.
   Über das Signal „swCshP“ wird das rechte S&H-Glied aktiviert und der erste zu messende Cluster ausgewählt (Signal „swC1“). Es werden n_1 Pumpzyklen durchgeführt.
5. Messvorgang, zweite Pumphase.
   Über „swCshN“ wird das am negativen Komparatoreingang angeschlossene S&H-Glied aktiviert und der zweite zu messende Cluster über „swC2“ ausgewählt. Nun werden $n_2 = n_1 + \alpha$ Pumpzyklen durchgeführt, wobei α dem Wert aus der Offsetermittlung entspricht.

Durch diesen Algorithmus (siehe auch Bild 3.42) tastet man sich langsam an die Schaltschwelle heran, bis bei einem bestimmten Wert α der Komparator von Eins nach Null umschaltet. In allen folgenden Messungen kann dieser Wert nun verwendet werden, um den Offset „wegzueichen“. Welche Werte α in der Praxis annimmt und wie gut die Kompensation durch dieses Verfahren funktioniert, wird in Abschnitt „Schwellenwertdispersion“ ab Seite 125 diskutiert.

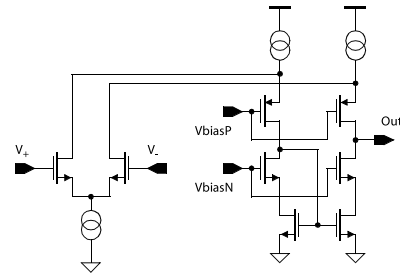


Bild 3.41. Schematische Darstellung eines sog. „folded cascode“ Differenzverstärkers.

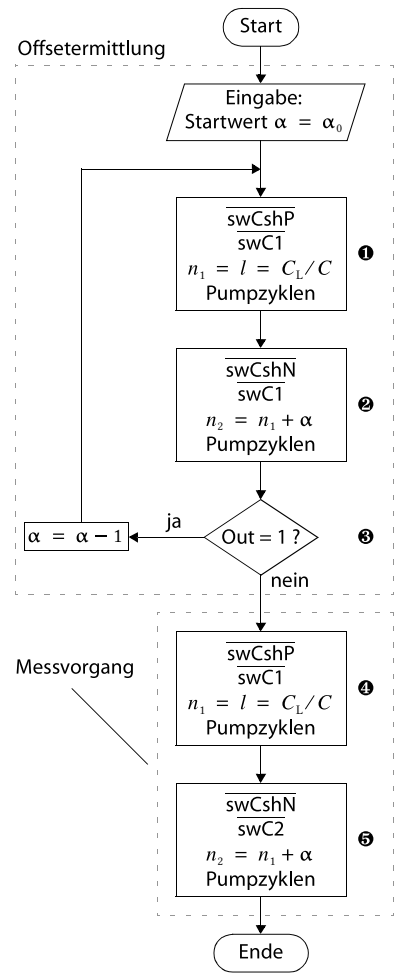


Bild 3.42. Offsetkompensation durch Anpassung der Schrittweite n. Die Schaltsignale „sw...“ sind „active low“.

Statistische Analyse

STATISTISCHE VS. DETERMINITISCHE KAPAZITÄT. Bei jeder Messung werden unweigerlich Fehler aufgrund verschiedener Störquellen[23] gemacht, so dass die gemessene Kapazität der beiden Cluster mit einer gewissen statistischen Varianz behaftet sind. Ebenso schwanken die tatsächlichen Kapazitätswerte von Chip zu Chip, weisen also ebenfalls eine statistische Verteilung mit einer bestimmten Standardabweichung auf. Ursache sind hier prozessbedingte Fluktuationen bei den Herstellungsbedingungen, wie in "Prozessstreuung und Mismatch" auf Seite 22 erläutert wird. Als dritte statistische Größe ist die Unbestimmtheit der beiden Kapazitätswerte aus Sicht eines Angreifers anzusehen, da dieser die genauen Werte nicht kennen kann (auch die Messelektronik kennt diese ja nicht exakt), zumindest aber über Schätzwerte verfügt. Man kann also den Messwert, den jeweiligen tatsächlichen Kapazitätswert und den Schätzwert des Angreifers als statistische Größen auffassen, die mit einem bestimmten Fehler behaftet sind.

Sind σ_{C1}, σ_{C2} die Standardabweichungen des (absoluten) Fehlers von C_1 und C_2, so gilt für den Quotienten $\tilde{x} = C_1/C_2$ über das Fehlerfortpflanzungsgesetz von Gauß (Beweis in Barlow, 1989):

$$\left(\frac{\sigma_{\tilde{x}}}{\tilde{x}}\right)^2 = \left(\frac{\sigma_{C1}}{C_1}\right)^2 + \left(\frac{\sigma_{C2}}{C_2}\right)^2 \tag{3.16}$$

Der relative Fehler bei der Bestimmung von $\tilde{x}$ ergibt sich also aus der Wurzel der Summe der Quadrate der relativen Fehler von C_1 und C_2.

Sei $C_x = C_1$ die Kapazität des ersten Clusters und $C = C_2$ die Kapazität des zweiten Clusters. Bis hierher handelt es sich somit nur um eine Umbenennung. Nun sei zusätzlich angenommen, dass C *bekannt* bzw. *reproduzierbar* ist, und zwar exakt ($\sigma_C = 0$). Die Unbestimmtheit der Kapazität der beiden Cluster sei komplett auf C_x aufgeschlagen, so dass $\sigma_{Cx}/C_x = \sigma_{\tilde{x}}/\tilde{x}$ gilt. Damit ergibt sich für den relativen Fehler des Quotienten x aus C_x und C:

$$\left(\frac{\sigma_x}{x}\right)^2 = \left(\frac{\sigma_C}{C}\right)^2 + \left(\frac{\sigma_{Cx}}{C_x}\right)^2 \Rightarrow \frac{\sigma_x}{x} = \frac{\sigma_{\tilde{x}}}{\tilde{x}} = \sqrt{\left(\frac{\sigma_{C1}}{C_1}\right)^2 + \left(\frac{\sigma_{C2}}{C_2}\right)^2} \tag{3.17}$$

Auf diese Weise genügt es, mit nur einer statistischen Kapazität zu rechnen, ihre Standardabweichung ergibt sich aus Gleichung 3.17. Die andere Kapazität kann dann je nach Situation als exakt gemessen, exakt reproduzierbar oder vollständig bekannt angenommen werden, also als vollkommen deterministisch.

TRENNSCHÄRFE UND STABILITÄT. Die Genauigkeit des Komparators im Zusammenhang mit der Bestimmtheit und Reproduzierbarkeit der Kapazitätswerte beeinflusst die Stabilität der einzelnen Bits. Liegen die Werte zu nahe beieinander, so führen Störeinflüsse zu Schwankungen in der Bitsequenz. Wenn diese Bitfolge direkt (ohne Nachbearbeitung, z.B. Filterung) als Schlüssel in einem System verwendet werden soll, das keine schwankenden Bits erlaubt, so reduziert sich die Ausbeute je nach der Wahrscheinlichkeit des Auftretens von instabilen Bits erheblich.

23. Thermisches Rauschen, 1/f-Rauschen, Übersprechen, usw. Zu den Ursachen sei auf die Literatur verwiesen.

Im Abschnitt „Parameter-, Leistungs- und Funktionsbereich“ auf Seite 29 ff. wurde dieser Themenkreis bereits diskutiert, insbesondere wurde die Frage der Stabilität der Bits in Abhängigkeit von der Kapazitäts*differenz* angeschnitten. Es wurde gezeigt, dass sich ein Graph mit den Kapazitätswerten der Clusterpaare aller möglichen Ausprägungen erstellen lässt, der – heruntergebrochen auf einzelne Bits – die funktionelle Ausbeute bzw. den Nutzen veranschaulicht (siehe Bild 2.8 auf Seite 30). In Bild 3.43 ist ein solcher Graph für das Kapazitäts*verhältnis* zu sehen. Liegt ein Clusterpaar außerhalb des Nutzens, so reduziert dies die Ausbeute an Bits oder – falls gar keine instabilen Bits auftreten dürfen – sogar die Ausbeute an funktionsfähigen Chips.

Um diese Zusammenhänge analytisch greifbar zu machen, sollen die folgenden Überlegungen dienen. Zunächst soll angenommen werden, dass der Komparator über eine perfekte Offsetkompensation verfügt. Er berechnet offensichtlich die Funktion

$$F = \begin{cases} 1 & \text{falls} \quad x > 1+\delta \\ 0 & \text{falls} \quad x < 1-\delta \end{cases} \qquad \text{mit} \qquad x = \frac{C_x}{C}\,. \tag{3.18}$$

Innerhalb des durch δ definierten Bereichs ist das Ergebnis nicht deterministisch, sondern statistischen Schwankungen unterworfen. Die Breite δ wird durch die Störanfälligkeit bzw. Trennschärfe des Komparators vorgegeben, ein Wert von 0,1 Prozent kann in der Praxis als plausibel angenommen werden[24].

Die Wahrscheinlichkeit P_{instabil}, dass ein beliebiges zu messendes Paar aus Clustern im instabilen Intervall $1-\delta \ldots 1+\delta$ liegt, kann leicht über die Verteilungsfunktion bzw. das Integral der Wahrscheinlichkeitsdichte berechnet werden, ebenso die Wahrscheinlichkeit $P_{1\to 0,\,1\to 0}$ für die Ausprägung eines stabilen Paares mit falschem Ergebnis (siehe Bild 3.44):

$$\begin{aligned} P_{\text{instabil}} &= \frac{1}{\sigma_x\sqrt{2\pi}} \int_{-\infty}^{1+\delta} e^{-\frac{1}{2}\left(\frac{t-\mu_x}{\sigma_x}\right)^2} dt - P_{1\to 0,\,1\to 0} \\ &= \frac{1}{2} + \frac{1}{2}\,\text{erf}\left[\frac{1+\delta-\mu_x}{\sqrt{2}\sigma_x}\right] - P_{1\to 0,\,1\to 0} \end{aligned} \tag{3.19}$$

$$P_{1\to 0,\,1\to 0} = \frac{1}{2} + \frac{1}{2}\,\text{erf}\left[\frac{1-\delta-\mu_x}{\sqrt{2}\sigma_x}\right]$$

Die Funktion $\text{Erf}[z]$ in Gleichung 3.19 entspricht dabei der in Mathematica definierten Fehlerfunktion. In der Literatur wird diese gelegentlich etwas anders definiert (z.B. in Papoulis 1991), durch Variablensubstitution kann man in diesen Fällen einfach zur hier verwendeten Version übergehen.

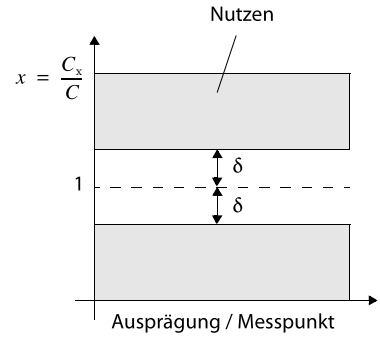


Bild 3.43. Liegt das Kapazitätsverhältnis x innerhalb der grauen Flächen (Nutzen), so ist das gewonnene Bit stabil. Andernfalls schwankt das Ergebnis und das entsprechende Bit muss verworfen werden.

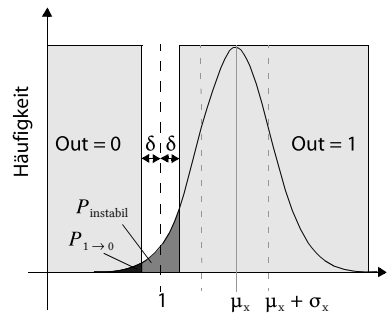


Bild 3.44. Berechnung der Wahrscheinlichkeit für das Auftreten falscher oder instabiler Ergebnisse aus der Wahrscheinlichkeitsdichte des Quotienten der beiden Kapazitätswerte.

24. Schätzwert ohne Beleg. Heutige CMOS-Komparatoren können Spannungen von wenigen Millivolt unterscheiden, falls man den Datenblättern Glauben schenkt.

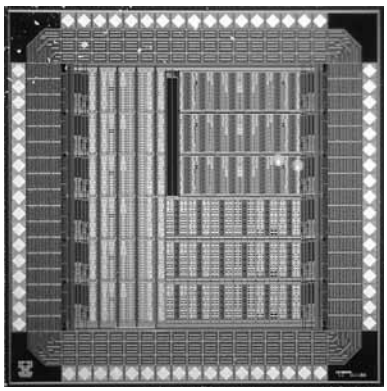


Bild 3.45. Auf dem Testchip wurden insgesamt 3264 Cluster integriert. Die Auswerteelektronik wurde in drei Varianten implementiert, erkennbar an den Strukturunterschieden der drei Bereiche in der Mitte des Chips (Farbversion des Bildes auf Seite 152).

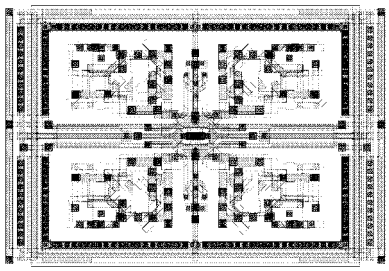


Bild 3.46. Die Layoutvariante mit 4 Clustern bzw. (Platten-)Kondensatoren pro Zelle. Die auflösungslimitierende parasitäre Kapazität des internen Knoten beträgt ca. 2,3 fF. Zellgröße: $45 \times 30\ \mu m^2$.

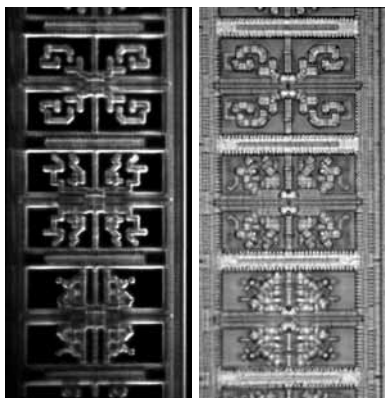


Bild 3.47. Mikroskopische Aufnahme einer Matrixspalte mit drei Zellen der Layoutvariante mit jeweils vier Clustern. In der Dunkelfeldaufnahme links ist nur die oberste Metalllage zu erkennen, rechts scheinen dagegen die unteren Lagen etwas durch.

3.3.4 Der Testchip

Bestandteile

Die in Abschnitt 3.3 vorgestellte Auswerteelektronik wurde auf einem Chip implementiert und getestet. Ziel des Tests war in erster Linie, das Funktionieren des Schaltungsvorschlags unter Beweis zu stellen und die Kapazitätsauflösung zu bestimmen. Hierfür wurden drei verschiedene mit Plattenkondensatoren beschaltete Layoutvarianten erstellt, erkennbar an den Strukturunterschieden der drei Bereiche im Chipkern (Bild 3.45). Zusätzlich wurden einige Cluster in den Test einbezogen, um die Wahrscheinlichkeit des Auftretens instabiler Bits zu untersuchen, falls die Cluster zur Generierung eines wiedergewinnbaren kryptografischen Schlüssels benutzt werden sollen.

Von jeder Layoutvariante wurden auf dem Chip jeweils mehrere Instanzen erstellt. Im Fall der Variante in Bild 3.46 wurden sechs identische Kopien auf der kompletten linken Chipseite (siehe Bild 3.45) platziert, die beiden anderen Varianten kamen jeweils dreimal zum Einsatz. Jede Instanz wurde dabei aus einer Vielzahl von Zellen entsprechend Bild 3.37 auf Seite 89 zu einer Matrix zusammengesetzt und mit jeweils einem Ensemble aus Source-Meter, Abtast-Halteglied und Komparator ergänzt (Bild 3.36). Zusätzlich wurde für die Spannung V_{out} ein extern einstellbarer Signalverstärker und ein analoger Buffer (Spannungsfolger) hinzugefügt, mit denen es möglich ist, kleine Spannungsintervalle auf einen sehr viel größeren Bereich aufzuspreizen, um geringste Spannungsunterschiede extern beobachten zu können (Beispiel in Bild 4.30). Zum Einsatz kam immer ein Messkondensator mit der Kapazität $C_{\text{L}} = 4\,\text{pF}$, die Clustergrößen bzw. Kapazitäten wurden im Bereich $C = 4 \ldots 12\,\text{fF}$ gewählt.

DIE 4-FACH VARIANTE. In der nebenstehenden Version einer Zelle mit vier zu vergleichenden Kapazitäten wurde auf eine möglichst kompakte Form geachtet, um das Minimum an Platzbedarf zu ermitteln. Die Ränder der Zelle sind durch auf Massepotential liegende „Wände“ aus Metallbahnen auf allen Ebenen und mit maximaler Zahl an Durchkontaktierungen voneinander getrennt (siehe Bild 3.46). Der Pumptransistor befindet sich in der Mitte, an den die vier Schalttransistoren der Cluster angeschlossen sind. Sie liegen geringfügig vom Zentrum entfernt jeweils zwischen den Anschlüssen der Cluster und dem Pumptransistor, also auf den beiden Diagonalen durch die Zelle.

Der Aufbau ist streng punktsymmetrisch zur Zellenmitte, so dass alle vier Messstrukturen eine vom Entwurf her exakt identische Umgebung besitzen. Keine Ecke ist „benachteiligt“. Die parasitäre Kapazität des internen Knotens, die bei jedem Pumpvorgang unvermeidbar mitgemessen wird, beträgt 2,3 Femtofarad und ist im Vergleich mit den 27,8 Femtofarad der Verarmungsgebiete der Transistor-Anschlussdioden sehr klein. In Bild 3.47 ist die mikroskopische Aufnahme dreier solcher Zellen innerhalb einer Matrixspalte des Testchips zu sehen. Außer dieser kompakten Version wurde das Layout als dritte Variante „auseinandergezogen“, so dass sich zwar die Abstände der Zellbestandteile vergrößerten, der prinzipielle Aufbau jedoch gleich blieb.

DIE 2-FACH VARIANTE. Im Gegensatz zur Punktsymmetrie der 4-fach Zelle gehen die beiden Teile der Layoutvariante mit zwei Pumpzweigen nicht durch Spiegelung an einem Punkt oder einer Geraden aus der jeweils anderen hervor, sondern stellen identische Kopien dar (siehe Bild 3.48). Dies verbessert

das Matching sowohl der Zellelektronik (Pumptransistor, Schalttransistoren etc.) und der Anschlussleitungen, als auch der Vergleichsstrukturen selbst. Damit sind *systematische* Fehler gemeint, nicht lokale, zufallsbedingte Ungenauigkeiten (z.B. Randunschärfen).

Durch die geringere Anzahl an Schalttransistoren sank die Diffusionskapazität der Transistoren am internen Knoten in dieser Variante auf 10,5 Femtofarad, während die parasitäre Leitungskapazität auf 5,5 Femtofarad aufgrund des höheren Verdrahtungsaufwands stieg. Als Besonderheit wurden die N-dotierten Wannen der Schalttransistoren auf das Potential des internen Knotens gelegt (statt V_{DD}), so dass die Sperrschichtkapazität der beiden Wanne-zu-Substrat Dioden von insgesamt 18 Femtofarad noch hinzukommt. Damit ist der Bulk-Anschluss der Schalttransistoren in dieser Variante immer auf Source-Potential, so dass kein Bulkeffekt (Verschiebung der Schwellenspannung) auftritt.

An der auflösungslimitierenden Kapazitätsbilanz des internen Knotens hat sich also kaum etwas geändert, während der vom (systematischen) Mismatch herrührende Messfehler geringer ist. Statt sechs Bits, die aus einer Zelle gewonnen werden können, liefert diese Variante nur noch ein Bit.

* * *

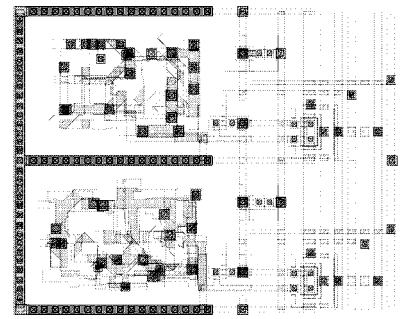


Bild 3.48. Variante mit zwei Vergleichsstrukturen. Die parasitäre Kapazität des internen Knoten beträgt hier ca. 5,5 fF, die Zellgröße liegt bei $36 \times 26\ \mu\mathrm{m}^2$.

Kapitel 4
Ergebnisse

Nach der Beschreibung der Vorgehensweise bei der Implementierung des Cluster-Konzepts im vorangehenden Kapitel werden im Folgenden die Ergebnisse vorgestellt. Den Anfang macht in Abschnitt 4.1 die Analyse der Extraktionswerte, die von einigen Untersuchungen grundlegender Eigenschaften des Referenz-Tools „Quickcap“ eingeleitet wird. Danach werden die unter typischen Prozessbedingungen extrahierten Werte von Standard-Extraktoren verglichen, gefolgt von den Werten aus der „worst- case“ Extraktion. Eine Tabelle mit den größen Abweichungen der 299 untersuchten Clustervarianten wird auf Seite 110 präsentiert, die wiederum auf die Farbbilder der jeweiligen Layouts verweist. Abgeschlossen wird die Analyse durch den Vergleich der Laufzeiten bei der Extraktion, die Hinweise auf die strukturelle Komplexität der Cluster liefert.

Die Ergebnisse der Prober-Messungen am ersten der beiden Testchip-Varianten werden in Abschnitt 4.2 dargestellt. Zunächst werden die Messwerte von einfachen Plattenkondensatoren und parallelen Metallbahnen analysiert und der Kapazitätsverlauf über die Chipfläche hinweg untersucht. Danach wird auf spezielle Strukturen eingegangen, die im wesentlichen Plattenkondensatoren mit horizontal verlaufenden Feldlinien entsprechen, jedoch besondere Eigenschaften aufweisen. Darauf folgen die Messergebnisse der Cluster, die hinsichtlich ihrer Streuung über fünf Testchips und des Extraktionsfehlers untersucht werden. Insbesondere das Matching der Cluster ist Gegenstand dieser Analysen.

In Abschnitt 4.3 folgt eine kurze Vorstellung des Testaufbaus der zweiten Testchip-Variante, auf der die Schlüssel-Schaltung zur Auswertung der Cluster-Kapazitätsverhältnisse untergebracht wurde. In der Analyse werden die Anzahl der Pumpzyklen des Ladungspumpen-Prinzips berücksichtigt, sowie Fragen nach den Komparatoreigenschaften und der Messauflösung beantwortet. Es wird gezeigt, dass Kapazitätsunterschiede von 111,5 Attofarad gemessen werden können. Abschließend wird anhand der Ergebnisse des Entropie-Tests plausibel gemacht, dass die relative Messmethode bei der Existenz stark ausgeprägter, globaler Parametergradienten versagt. Stattdessen muss die im Schlusskapitel geschilderte absolute Messtechnik eingesetzt werden.

* * *

4.1 Extraktion

Die Kapazitätscluster dienen - vereinfacht dargestellt - als Kondensatoren mit speziellen Eigenschaften. Als ein solcher stellt jeder von ihnen eine elektrische Kapazität bereit, deren genauer Wert eine entscheidende Rolle spielt. Aus diesem Grund sollen im Folgenden die Ergebnisse aus Simulationen und die Analyse von Messwerten vorgestellt werden.

4.1.1 Werkzeugspezifische Kapazitätswerte

Ist im Zusammenhang mit elektrischen Kapazitäten und mikroelektronischen Schaltungen von Simulationen die Rede, so ist hier die rechnerunterstützte Schaltungsrückerkennung (Extraktion) aus der Geometrie der Maskendaten (Layout) gemeint[25]. Bei dieser Rückerkennung wurde der Schwerpunkt auf die Extraktion der parasitären Schaltungselemente gelegt und die sonst übliche Erkennung der aktiven Teile (Transistoren, Dioden) vernachlässigt.

Diese Unterscheidung schlägt sich in der Auswahl der Werkzeuge nieder, mit denen die parasitäre Extraktion durchgeführt wurde. So gibt es Extraktionsprogramme, die in erster Linie zur Erkennung von Transistoren, Dioden und herkömmlichen Kondensatoren verwendet werden, beispielsweise das Produkt DIVA von Cadence. Es dient hauptsächlich zur Endkontrolle bzw. Vergleich der Schaltpläne mit den Layouts („layout versus schematic“, LVS) im analogen Schaltungsentwurf. Die Extraktion parasitärer Kapazitätswerte ist dabei zuschaltbar. Andere Produkte konzentrieren sich ausschließlich auf diese parasitären Bauteile, sind mit ihren Leistungsdaten hinsichtlich Genauigkeit und Geschwindigkeit jedoch mehr für digitale, zellbasierte Chips mittlerer Größe und darüber ausgelegt.

Eine kleinere Zahl an Extraktionsprogrammen ist für den Einsatz bei kleinen, analogen und handoptimierten Schaltungen ausgelegt, wodurch der Genauigkeit vor der Geschwindigkeit Vorrang eingeräumt wird. Sie verfügen über spezielle Algorithmen zur Berechnung von parasitären Kapazitätswerten (siehe Abschnitt 2.2.2), so dass je nach der zugrundeliegenden Mathematik und ihrer algorithmischen Umsetzung Unterschiede in der erzielbaren Genauigkeit liegen. Aus diesem Grund werden im Folgenden die Ergebnisse einiger ausgewählter Extraktionsprogramme vorgestellt und verglichen.

Der „Golden Standard“

Das Extraktionstool Quickcap der Firma Magma basiert auf einem nicht-deterministischen Näherungsverfahren, bei dem je nach Vorgabe und Laufzeitbeschränkung Kapazitäten beliebig genau extrahiert werden können, wobei die einzelnen Werte von Lauf zu Lauf innerhalb der vorgegeben Genauigkeitsschranke aufgrund des nicht-deterministischen „random-walk“ Algorithmus schwanken (siehe Abschnitt „Numerische Verfahren“ auf Seite 41).

25. Je nach Zusammenhang wird häufig von Extraktion, Simulation oder Berechnung der Kapazität gesprochen.

Die Genauigkeitsvorgabe entspricht dabei einer Schätzung der Standardabweichung σ und kann als absolutes Ziel (in Farad) oder relatives Ziel spezifiziert werden. Das Ergebnis jedes Extraktionsvorgangs einer Messreihe streut mit σ um den Mittelwert, der wiederum durch die jeweils extrahierte Kapazität geschätzt wird. Je öfter die Extraktion wiederholt wird, desto mehr nähert sich der Mittelwert der Messreihe dem wahren Wert an. Alternativ kann von vornherein ein viel strengeres Genauigkeitsziel vorgegeben werden, wodurch sich die Streuung des berechneten Wertes um den wahren Wert auf Kosten der Laufzeit reduziert.

Die Schätzung der (a priori unbekannten) Standardabweichung einer hypothetisch unendlichen Messreihe dient Quickcap zur Beurteilung, ob das vorgegebene Genauigkeitsziel erreicht wurde. Ist dies der Fall, bricht Quickcap die Berechnung ab. Die Schätzung selbst ist wiederum Schwankungen unterworfen. Wird bei einer Messreihe der Mittelwert der Schätzungen der Standardabweichung gebildet, so entspricht er sehr genau der Standardabweichung der einzelnen extrahierten Werte der Messreihe, wie im Folgenden zu sehen ist.

In Bild 4.1 ist die Streuung der Kapazität eines Clusters bei 1000 Extraktionsvorgängen zu sehen. Die vorgegebene Genauigkeit betrug dabei ±5 Prozent und wurde bei jedem Durchlauf schon beim ersten Iterationszyklus des Algorithmus mit 3,4 Prozent erreicht, so dass die Laufzeit sehr gering war. Die von Quickcap angegebene Abweichung von im Mittel ±250 Attofarad (±3,46%) entspricht recht genau der Standardabweichung des Histogramms mit ±254 Attofarad. Der Mittelwert der Verteilung mit 7,534 Femtofarad repräsentiert die genaueste Näherung, die sich aus einer solchen Extraktionsreihe ermitteln lässt.

Analog zu Bild 4.1 wurde in Bild 4.2 ein Histogramm für die Genauigkeitsvorgabe von ±0,5 Prozent erstellt. Wieder wurden 1000 Extraktionsdurchläufe vorgenommen, die Cluster waren dieselben. Durch die engere Genauigkeitsgrenze wurde das Ziel nicht sofort, sondern erst nach einigen Iterationsschritten des Algorithmus erreicht, so dass die Gesamtlaufzeit auf mehrere Stunden anstieg. Die angegebene mittlere Abweichung betrug nun ±37,46 Attofarad (±0,498%), sie repräsentiert das vorgegene Genauigkeitsziel von ±0,5 Prozent. Die Standardabweichung der Daten in Bild 4.2 berechnet sich zu ±37,81 Attofarad, ist also mit den aus Quickcap stammenden Werten konsistent. Der Mittelwert der Verteilung liegt mit 7,523 Femtofarad sehr nahe bei dem der Verteilung in Bild 4.1.

Die weitere Steigerung der Genauigkeit erhöht die Laufzeit jedes Extraktionsvorgangs exponentiell. In Bild 4.3 ist der Zusammenhang zwischen Genauigkeitsziel und Laufzeit für den betrachteten Cluster zu sehen. Die Gesamtlaufzeit steigt damit bei einer 1000-fachen Wiederholung der Extraktion sehr schnell an. Für die Vorgabe von ±0,1 Prozent beträgt die Rechenzeit eines Durchlaufs bereits über eine Stunde, so dass auf Wiederholungen der Berechnung verzichtet wurde. Das Ergebnis dieses Durchlaufs betrug 7,525 Femtofarad und liegt sehr genau beim Mittelwert der Histogramme in Bild 4.1 und Bild 4.2.

Bei sehr engen Genauigkeitsgrenzen soll Quickcap den wahren Wert sehr genau approximieren. Bei einer theoretisch unendlichen Laufzeit bleibt nur noch ein sogenannter „Bias“ übrig, der als inhärenter, deterministischer

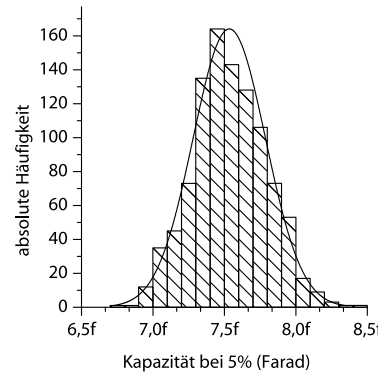


Bild 4.1. Verteilung der mit Quickcap bei 5% Genauigkeit extrahierten Werte für den Cluster in Farbtafel I (a-c). Der Mittelwert liegt bei 7,534 fF, die Standardabweichung beträgt 254 aF.

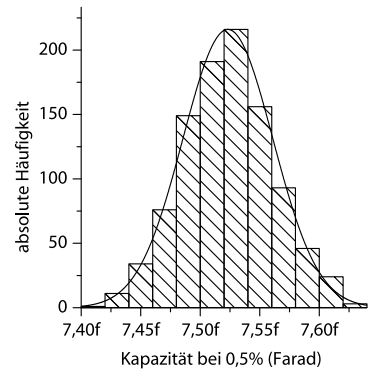


Bild 4.2. Verteilung der Werte bei 0,5% Genauigkeit. Der Mittelwert liegt nun bei 7,523 fF, die Standardabweichung beträgt 37,81 aF.

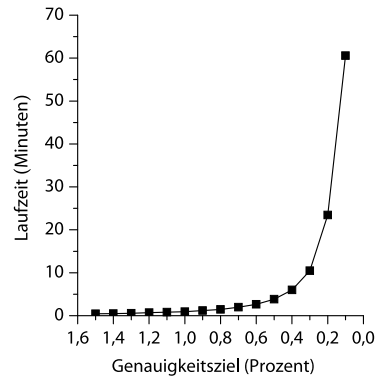


Bild 4.3. Laufzeit von Quickcap als Funktion des Genauigkeitsziels.

Problem	Bias (%)
1D: Parallele Platten	-0.002 ± 0.001
2D: Parallele runde Leitungen	-0.018 ± 0.004
3D: Kugel im freien Raum	+0.002 ± 0.004

Tabelle 4.1. Systematischer Fehler von Quickcap für einfache Probleme. Aus Iverson & LeCoz 2001.

Fehler des Algorithmus angesehen werden kann. Wie in Tabelle 4.1 ersichtlich, liegt dieser bei Vergleichen mit analytischen Lösungen weit unter 0,03 Prozent (siehe Iverson & LeCoz 2001).

Aufgrund dieser Eigenschaft ist Quickcap in der Vergangenheit zum „golden standard" avanciert und wird von anderen Softwareherstellern als Referenz für eine vergleichende Bewertung ihrer eigenen Produkte herangezogen. In den folgenden Abschnitten wird in ähnlicher Weise Quickcap als Referenz betrachtet, die Genauigkeitsvorgabe beträgt in der Regel ±0,2 Prozent.

Der Vergleich eines statistischen Extraktionsprogrammes mit Quickcap kann über die Berechnung der absoluten Fehlerwerte geschehen, wenn eine Vielzahl von Extraktionsvorgänge *derselben* Struktur durchgeführt werden, oder – falls es sich um einen deterministischen Extraktor handelt – über eine Anzahl *verschiedener* Strukturen. Aus diesen Absolutfehlern lässt sich dann der mittlere quadratische Fehler des Extraktors berechnen. Dieser RMS-Fehler stellt als einzelne, skalare Größe ein Maß für die Streuung der Fehlerverteilung dar und kann damit als charakteristische Genauigkeit des Extraktors angesehen werden.

Werden auf diese Weise Extraktoren mit Quickcap verglichen, so wirkt sich die bei Quickcap obligatorische Genauigkeitsvorgabe bzw. die Standardabweichung vom jeweils wahren Wert auf das Ergebnis aus. Je größer die Schwankungsbreite (Ungenauigkeit) der Werte aus Quickcap, desto größer ist auch der Gesamtfehler. Durch Anwendung des Fehlerfortpflanzungsgesetzes von Gauß kann eine Aussage über den Einfluss des Fehlers bzw. der Genauigkeitsvorgabe bei Quickcap auf den Vergleich getroffen werden. In Box 4.1 wird dies für die Differenzbildung bzw. den Absolutfehler durchgeführt. Zur Herleitung des Fehlerfortpflanzungsgesetzes selbst siehe Abschnitt 4.3 in Barlow 1989.

Box 4.1 Fehlerfortpflanzungsgesetz nach Gauß.

Frage: Es seien x_j und y_j die mit Quickcap und einem anderen Extraktor ermittelten Werte von Extraktionen eines Clusters mit den Standardabweichungen σ_x und σ_y. Welche Standardabweichung σ_f ergibt sich für die Differenz $f_j = y_j - x_j$ der von den Extraktoren berechneten Werte?
Antwort: Durch das Fehlerfortpflanzungsgesetz von Gauß lässt sich die Varianz der Funktion f ermitteln (ϱ = Korrelationskoeffizient):

$$\begin{aligned} V(f) &= \left(\frac{\partial f}{\partial x}\right)^2 V(x) + \left(\frac{\partial f}{\partial y}\right)^2 V(y) + 2\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right)\varrho\sigma_x\sigma_y \\ &= V(x) + V(y) - 2\varrho\sigma_x\sigma_y \end{aligned} \tag{4.1}$$

Gleichung 4.1 führt auf eine bekannte Regel, falls die Werte von x und y unkorreliert sind ($\varrho = 0$): Die Absolutfehler der Summe oder Differenz zweier Zufallsvariablen addieren (subtrahieren) sich quadratisch. Sind die Werte von x und y zu ±300 fF und zu ±400 fF bekannt, so sind $x + y$ und $x - y$ zu ±500 fF bekannt.

Gleichung 4.1 stellt also eine Formel für den Einfluss des Fehlers von Quickcap und des Vergleichsextraktors auf den Fehler der jeweils extrahierten Einzelwerte für eine Messreihe dar. Es wird vorausgesetzt, dass immer ein und dasselbe Layout (Cluster) extrahiert wird und der mit Quickcap zu ver-

gleichende Extraktor auf einem statistischen Verfahren basiert, also wie Quickcap selbst Schätzwerte für den wahren Wert berechnet, die einer normalverteilten Schwankung mit der Standardabweichung σ_y unterworfen sind und deren Mittelwert den wahren Wert für größer werdenden Messreihenumfang immer genauer approximiert.

Aus Gleichung 4.1 folgt also für die Varianz des RMS-Fehlers der Kapazitätsdifferenz (bei M Extraktionen):

$$\frac{1}{M}\sum_j (y_j - x_j)^2 = \sigma_x^2 + \sigma_y^2 \tag{4.2}$$

Hierbei wird angenommen, dass die Werte x_j und y_j unkorreliert sind, so dass $\rho = 0$ gesetzt werden darf. Diese Annahme kann dadurch gerechtfertigt werden, dass es sich bei x und y um Kapazitätswerte von verschiedenen Extraktionswerkzeugen handelt, die über jeweils eigene numerische Algorithmen ermittelt werden. Es besteht daher kein Grund davon auszugehen, dass wenn Extraktor X (Quickcap) ein Layout mit der Kapazität x_j angibt, die Kapazität y_j des Extraktors Y eher in der Nähe von x_j liegt, statt am wahren Wert w, und umgekehrt. Anders ausgedrückt machen grundverschiedene numerische Verfahren statistische Fehler, die nichts miteinander gemein haben. Sind also die Fehler $\mathrm{Er}(x) = w - x$ und $\mathrm{Er}(y) = w - y$ von X und Y unkorreliert, so ist die Differenz (analog Summe)

$$f = y - x = w - \mathrm{Er}(y) - (w - \mathrm{Er}(x)) = \mathrm{Er}(x) - \mathrm{Er}(y)$$

ebenfalls unkorreliert.

Gleichung 4.2 wirft jedoch noch eine weitere Frage auf: Welcher Wert ist für die Standardabweichung σ_x anzusetzen? Schließlich gibt Quickcap nur eine Schätzung von σ_x ab, die ihrerseits statistischen Schwankungen unterliegt. Wird ein relatives Genauigkeitsziel g_x gesetzt, so hängt σ_x darüber hinaus von den jeweils extrahierten Messwerten x_j ab: $\sigma_{x,j} = g_x \cdot x_j$. Da nun aus der vorangehenden Analyse der Histogramme in Bild 4.1 und Bild 4.2 deutlich wurde, dass der Mittelwert der Schätzungen der Standardabweichung sehr nahe an der tatsächlichen Standardabweichung der gesamten Messreihe liegt, kann für σ_x also der Mittelwert gesetzt werden:

$$\sigma_x = \frac{1}{M}\sum_j \sigma_{x,j} = \frac{1}{M}\sum_j g_x \cdot x_j = \frac{g_x}{M}\sum_j x_j = g_x \cdot \bar{x} \tag{4.3}$$

Damit erhält man durch Einsetzen in Gleichung 4.2 folgenden Ausdruck:

$$\frac{1}{M}\sum_j (y_j - x_j)^2 = (g_x \cdot \bar{x})^2 + \sigma_y^2 \tag{4.4}$$

Durch Division mit $\bar{x}^2$ bzw. Normierung auf den Mittelwert schließlich:

$$\frac{1}{M}\sum_j \left(\frac{y_j - x_j}{\bar{x}}\right)^2 = g_x^2 + \left(\frac{\sigma_y}{\bar{x}}\right)^2 \tag{4.5}$$

Gleichung 4.5 setzt also den „beobachteten“ RMS-Fehler des Extraktors Y in Beziehung zur Standardabweichung bzw. Genauigkeit des Extraktors X (Quickcap) und seiner selbst. Mit anderen Worten ist der beobachtete RMS-Fehler gleich der Genauigkeit des Extraktors Y, *korrigiert* um die Genauigkeit von Quickcap (jeweils quadratisch). Da in den folgenden Abschnitten Quickcap immer mit einer Genauigkeit von 0,2 Prozent zum Einsatz kam, beträgt der Korrekturterm für die Varianzen nur 0,04 (Quadratprozente). Die Genau-

igkeit ist also wie vermutet sehr hoch, so dass Quickcap zu Recht als „golden standard“ angesehen werden kann und bei Vergleichen die extrahierten Werte als die korrekten, wahren Kapazitäten angenommen werden können.

Typical-case

Unter normalen Prozessbedingung („typical-case“) werden die Strukturen eines Chips durch den lithografischen Herstellungsprozess mit bestimmten, *typischen* Eigenschaften erzeugt. Dazu gehören geometrische Größen wie die Breite und Dicke von Leiterbahnen, Isolationsschichten und von dotierten Bereichen, sowie *elektrische* Eigenschaften wie ohmscher Widerstand, Schwellenwert und Dotierungsstärke. Mit Ausnahme der Dielektrizitätskonstante gibt es keine elektrischen Größen, die einen Einfluss auf die Kapazitätswerte der Kapazitätscluster haben. Folglich machen sich Prozessschwankungen hier nicht bemerkbar. Die *geometrischen* Eigenschaften, insbesondere die Dicke der Isolationsschichten zwischen den Leiterbahnebenen, haben hingegen einen direkten Einfluss auf die Kapazität der Cluster. Chips, deren geometrische und elektrische Kenngrößen vom typischen Wert abweichen, weisen ein schaltungstechnisch günstiges oder nachteiliges Verhalten auf, entsprechend ist vom „best-case“ oder „worst-case“ bezüglich der Prozessbedingungen die Rede, wenn der günstigste bzw. ungünstigste Fall auftritt.

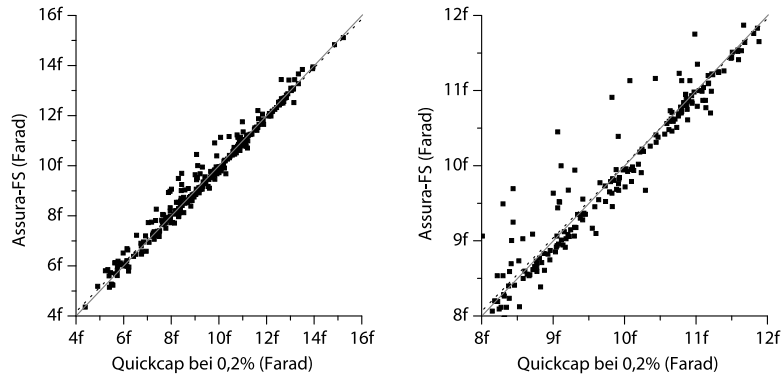


Bild 4.4. Assura-FS bei typischen Prozessbedingungen. Jeder Punkt repräsentiert einen Kapazitätscluster, die Achsen geben die jeweils extrahierten Werte an. (Rechte Seite ist Ausschnitt der linken.)

In Bild 4.4 ist das Ergebnis der Extraktion von 299 verschiedenen Kapazitätsclustern mit Assura-FS und Quickcap bei 0,2 Prozent Genauigkeit zu sehen (je ein Extraktionsvorgang pro Cluster). Je näher die einzelnen Punkte an der grauen Linie liegen, desto größer ist die Übereinstimmung der beiden Programme. Die Streuung um die Winkelhalbierende der Achsen verläuft ohne wesentliche Häufung in positiver oder negativer Richtung, was auf einen Mittelwert des Fehlers nahe bei Null hinweist.

Die gestrichelte schwarze Linie ist das Ergebnis einer linearen Regression, d.h. diejenige Gerade $y = mx + c$, zu welcher der quadratische Abstand der Punkte (x_i,y_i) minimal ist (Prinzip der kleinsten Quadrate, siehe Barlow 1989, Abschnitt 6.2). Diese Ausgleichsgerade ist fast deckungsgleich mit der grauen Winkelhalbierenden, was ebenfalls auf einen Mittelwert nahe Null hindeutet und einen nur geringen systematischen, kapazitätsabhängigen Fehler von Assura-FS. Wäre die Gerade steiler oder flacher als die Winkelhalbie-

rende, so würde das auf die systematische Tendenz von Assura-FS hinweisen, große Kapazitäten häufiger und stärker zu überschätzen bzw. zu unterschätzen, als kleine Kapazitäten.

Auffallend sind einige Punkte, die weiter von der grauen Winkelhalbierenden entfernt sind. So extrahierte Quickcap einen Cluster mit 9,07 Femtofarad, während Assura-FS ihn mit 10,45 Femtofarad angab. Ein anderer Cluster liegt im Schaubild bei 6,21 Femtofarad (Quickcap) und wurde mit 5,79 Femtofarad von Assura-FS extrahiert. Dies entspricht im ersten Fall einem Fehler von 15,3 Prozent und -6,9 Prozent im zweiten Fall.

In Bild 4.5 ist der Fehler für alle in Bild 4.4 gezeigten Daten als Histogramm zu sehen. Der Mittelwert der Fehlerverteilung liegt mit -0,4 Prozent sehr nahe am Idealfall von Null Prozent, wie bereits durch die Lage der Ausgleichsgeraden in Bild 4.4 vermutet wurde. Assura-FS weist also keinen systematischen Offset (Versatz) bzw. Fehler auf. Auffallend ist jedoch die linkssteile (rechtsschiefe, positiv schiefe) Verteilung[26], d.h. die Ausdehnung und das Abflachen der Verteilung hin zu größeren positiven Fehlern. Die Kapazität einiger weniger Cluster scheint von Assura-FS also ungewöhnlich stark überschätzt zu werden, nämlich um bis zu 1,38 Femtofarad (15,3%). Der diesem Extremwert entsprechende Cluster ist auf Farbtafel III auf Seite 149 zu sehen, die Werte selbst sind in Tabelle 4.3 auf Seite 110 aufgelistet.

Auf der anderen Seite gibt es Cluster, deren Kapazität unterschätzt wurde, wenn auch das Ausmaß und die Anzahl wesentlich geringer ist. So beträgt die Differenz im Extremfall -0,42 Femtofarad (-6,9%) für den Cluster in Farbtafel II auf Seite 148 (siehe ebenfalls Tabelle 4.3). Zwischen dem dreidimensionalen Aufbau der beiden Ausreißer besteht jedoch kein prinzipieller Unterschied, beide wurden mit dem in Abschnitt 3.1.3 vorgestellen Random-Walk Algorithmus erzeugt. Einzig und alleine die Verwendung von einigen Stücken Polysilizium beim positiven Ausreißer unterscheidet den Cluster vom negativen Fall, bei dem kein Polysilizium vorkommt. Ein Zusammenhang kann jedoch ausgeschlossen werden, da die Cluster in Farbtafel IV und Farbtafel V ebenfalls über Polysilizium verfügen, ihre Kapazität von Assura-FS jedoch diesmal unterschätzt wird.

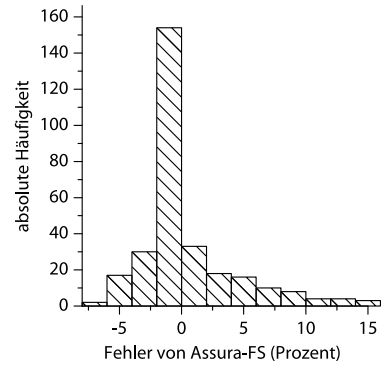


Bild 4.5. Verteilung des Fehlers von Assura-FS, normalisiert auf Quickcap bei 0,2%. Der Mittelwert liegt bei -0,4%.

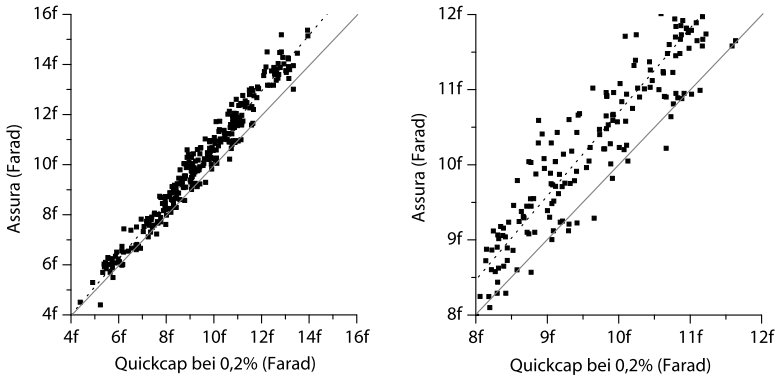


Bild 4.6. Assura bei typischen Prozessbedingungen. (Rechte Seite ist Ausschnitt der linken.)

26. Pearson'sches Schiefheitsmaß. Es verwendet die Differenz zwischen arithmetischem Mittel und Dichtemittel bezogen auf die Standardabweichung. Ist dieser Wert größer als Null, handelt es sich um eine rechtsschiefe Verteilung, andernfalls um eine linksschiefe Verteilung.

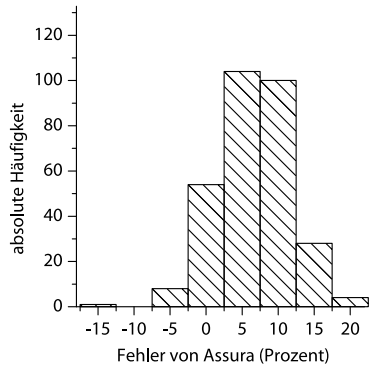


Bild 4.7. Verteilung des Fehlers von Assura, normalisiert auf Quickcap. Der Mittelwert beträgt 6,5%.

Produkt	Hersteller	Anwendung
Assura	Cadence	analog und mixed-signal
Dracula RCX	Cadence	digital
Diva	Cadence	worst-case Abschätzung
Fire&Ice QX	Cadence	digital
Calibre-xRC	Mentor	analog und mixed-signal
Metal	OEA	IC, PCB, etc.
HIPEX	Silvaco	mixed-signal

Tabelle 4.2. Gängige Extraktionswerkzeuge für komplette Chips bei Vernachlässigung der Genauigkeit.

Im Graphen von Bild 4.6 ist die Punktwolke für die extrahierten Werte von Assura zu sehen. Zwar handelt es sich hierbei um dasselbe Softwarepaket der Firma Cadence, der zugrundeliegende Algorithmus unterscheidet sich jedoch von Assura-FS. Es kommt kein numerischer Field-Solver- Algorithmus zur Lösung der Laplace'schen Gleichung („Fast Multipole Method", FFM) zum Einsatz, sondern ein schnelleres, für größere Datenmengen optimiertes Verfahren, das auf Nachschlagetabellen basiert („lookup-up table"). Zwischen beiden Versionen lässt sich durch Aktivieren der Field-Solver-Option in der Benutzeroberfläche umschalten. Für eine genauere Behandlung der algorithmischen Grundlagen sei an dieser Stelle auf Abschnitt 2.2.2 hingewiesen.

Analog zu den Graphen in Bild 4.4 wurde auch diesmal eine lineare Regression zur Bestimmung einer Ausgleichsgeraden (gestrichelte Linie) durchgeführt. Im Gegensatz zu Assura-FS liegt sie nicht auf der grauen Winkelhalbierenden, welche die Identität Assura-Quickcap darstellt, sondern erstreckt sich scherenartig öffnend oberhalb der Identitätslinie. Sie ist dabei kaum um einen konstanten Wert in positiver Richtung versetzt, sondern zeigt allein die Tendenz, bei größer werdenden Kapazitäten diese stärker zu überschätzen, als bei kleinen Kapazitäten. Dieses Auseinanderstreben bedeutet für Assura offensichtlich, dass sich die Teilfehler der bei der Extraktion durchgeführten Einzelschritte bzw. Iterationszyklen addieren, ohne sich dabei etwa in vorteilhafter Weise im Mittel gegenseitig aufzuheben. Bei kleinen Kapazitäten ist die Gesamtzahl der fehlbehafteten Operationen und damit die Summe der Fehlerbeiträge geringer, als bei großen Kapazitäten. Dort summieren sich diese zu einem immer größeren Gesamtfehler auf, so dass dieser bei 12,84 Femtofarad (Quickcap) absolut gesehen 2,34 Femtofarad (18,3%) beträgt (zweiter Punkt von oben, isoliert).

Betrachtet man nun die Verteilung des relativen Fehlers in Bild 4.7, so zeigt sich ein typischer, glockenförmiger Verlauf ohne eine ausgeprägte Links- oder Rechtssteilheit, der Mittelwert von 6,5% korrespondiert mit der Abweichung der Ausgleichsgeraden von der Winkelhalbierenden. Die beiden Extremwerte sind mit -16 Prozent (Absolutwert -0,84 fF) bei 5,24 Femtofarad (Quickcap) und mit 19,3 Prozent (absolut 1,21 fF) bei 6,22 Femtofarad in Tabelle 4.3 gegeben.

Bei Betrachtung der dreidimensionalen Struktur der Cluster scheint es keinen systematischen Zusammenhang zwischen der Struktur und dem Fehler in Assura zu geben. Alle in den Farbtafeln III bis V gezeigten Cluster weisen zum Beispiel eine hohe Komplexität auf und verfügen über Leitungsstücke auf allen vier Metallisierungsebenen sowie der Polysilizium-Ebene. Ihre Kapazität setzt sich aus elektrischen Feldern zusammen, deren Feldlinien sich sowohl in vertikaler, als auch in horizontaler Richtung erstrecken. Trotz dieser Gemeinsamkeiten bewegt sich der Fehler in Assura bei diesen Cluster in einem Bereich von -0,7 Prozent bis 14,5 Prozent.

Worst-case

Jene Prozessbedingungen, die sich am ungünstigsten auf die Elemente einer Schaltung auswirken, insbesondere Kondensatoren bzw. Kapazitäten, werden als „worst-case" Bedingungen bezeichnet. Im Folgenden wird der Vergleich zwischen den mit verschiedenen Werkzeugen extrahierten Werten auch für die ungünstigsten Prozessbedingungen durchgeführt und auf zwei weitere

Extraktionstools ausgedehnt (Calibre-xRC und Diva). Für diese wurde keine entsprechende typical-case Extraktion durchgeführt, da die dazugehörenden regelbasierten Prozessbeschreibungen und Verarbeitungsvorschriften (sog. „rule sets“ oder „technology files“) nicht vorhanden waren und nur in zeitaufwändiger Handarbeit erstellt werden können.

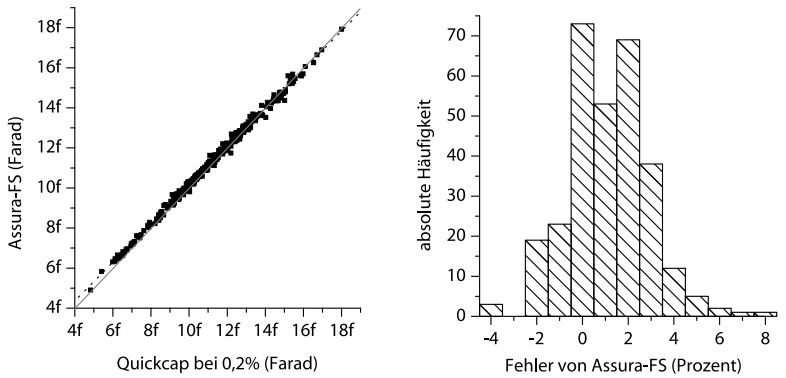


Bild 4.8. Verteilung der für die ungünstigsten Prozessbedingungen extrahierten Werte von Assura-FS (Punktwolke). Der relative Fehler (Normierung auf Quickcap) ist im rechten Bild zu sehen. Der Mittelwert beträgt 1,1%.

In Bild 4.8 und Bild 4.9 sind die mit Assura-FS und Assura extrahierten Werte als Punktwolken dargestellt, zusammen mit den resultierenden Fehlern. Jeder einzelne Punkt der beiden linken Abbildungen stellt wieder den einmalig extrahierten Wert eines Kapazitätsclusters dar (insgesamt 299). Wie bereits im vorherigen Abschnitt wurde Quickcap auf eine hohe Genauigkeit von 0,2 Prozent gesetzt, um die entsprechenden Werte als korrekt annehmen zu können. Für die Berechnung des relativen Fehlers in den beiden rechten Histogrammen wurden die Absolutfehler wieder auf die von Quickcap extrahierten Werte normiert.

Die geringere Breite der Verteilung im Vergleich zu jener unter typischen Prozessbedingungen ist schließlich darauf zurückzuführen, dass ein gewisser Anteil des Fehlers von Assura-FS von den Dicken (Höhen) der Metallbahnen und Isolationsschichten dazwischen abhängt. Durch Änderung der Prozessparameter vom typischen Fall zum ungünstigsten Fall ändert sich somit auch die Dicke aller Cluster, auch wenn die strukturelle Zusammensetzung in der Entwurfsansicht (2D, von oben) gleich bleibt. Dadurch sind die Cluster im worst-case nicht mehr identisch mit jenen des typical-case, aus einem Cluster werden zwei (wenn auch sehr ähnliche). Assura-FS macht für jeden der beiden Clustervarianten verschiedene Fehler, die in der Aufteilung der Geometrie in eine Gitterstruktur begründet liegen. Der Algorithmus nimmt eine Diskretisierung des dreidimensionalen Raumes vor, bei der die Dicke der einzelnen geometrischen Schichten so einen direkten Einfluss hat.

Bei Assura in Bild 4.9 ist im Vergleich zu den typischen Prozessbedingungen kein wesentlicher Unterschied zu erkennen. Die Ausgleichsgerade der Punktwolke zeigt die gleiche scherenartige Öffnung und die Fehlerverteilung im rechten Graphen erstreckt sich über einen Bereich von 4,5 Prozent (Standardabweichung des Fehlers) im Vergleich zu 4,9 Prozent unter typischen Bedingungen. Das Histogramm insgesamt ist um einige Prozentpunkte in die positive Richtung verschoben, der Mittelwert beträgt nun 9,2 statt 6,5 Prozent.

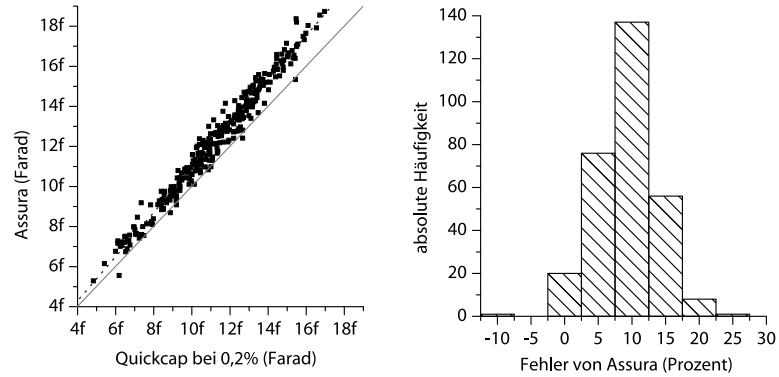


Bild 4.9. Die Verteilung der Punktwolke für Assura (ungünstigste Prozessbedingungen, worst-case). Der relative Fehler hat einen Mittelwert von 9,2%.

Ein bisher noch nicht analysiertes Extraktionswerkzeug wird von der Firma Mentor Graphics unter dem Namen Calibre-xRC angeboten. Dieser Extraktor gehört in die Klasse der auf Geschwindigkeit optimierten Algorithmen, die keine direkte Lösung der Laplace'schen Gleichung suchen, sondern die Lösung approximieren, indem alle Leitungsnetze in kleine Fragmente geteilt werden, deren Kapazität bereits vorberechnet wurde. Calibre-xRC baut zu diesem Zweck eine interne Datenbank auf, die als Nachschlagetabelle für die im Layout auftretenden Leitungsvarianten dient. Aus diesen Teilkapazitäten errechnet der Extraktor dann die Gesamtkapazität.

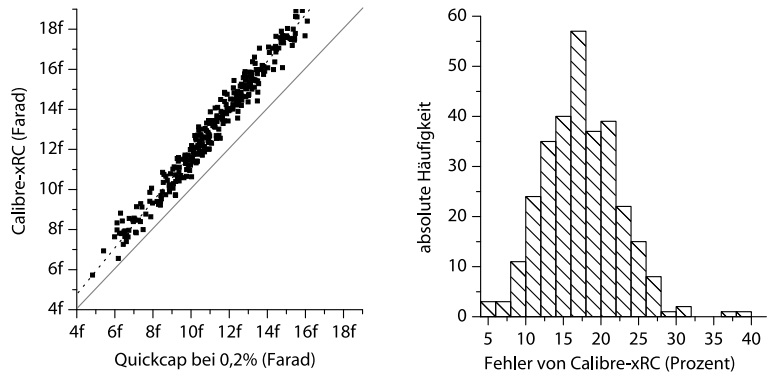


Bild 4.10. Calibre-xRC (worst-case) mit mittlerem Fehler von 17,4%

Beim Betrachten der Punktwolke in Bild 4.10 (links) fällt sofort der mutmaßliche Versatz der Ausgleichskurve (gestrichelte Linie) auf, der bisher (bei Assura/-FS) kaum vorhanden war. Selbst bei sehr keinen Kapazitäten von 4 Femtofarad extrahiert Calibre-xRC einige Cluster mit größeren Werten, als dies Quickcap tut. Überprüft man jedoch den Schnittpunkt der grauen Winkelhalbierenden mit der gestrichelten Ausgleichsgeraden, so stellt man fest, dass sich diese recht genau im Ursprung treffen, so dass es sich *nicht* um einen konstanten, systematischen Fehler bzw. Offset handelt.

Der Öffnungswinkel zwischen den beiden Geraden ist nun größer als bei Assura, die Geraden laufen erkennbar auseinander. Calibre-xRC hat also die zunehmende Tendenz, die Kapazität zu überschätzen. Die Fehlerverteilung rechts weist nur noch positive Werte auf, entsprechend sind keine Punkte links unterhalb der Winkelhalbierenden anzutreffen. Damit gibt es nun keine Cluster mehr, die von Calibre-xRC, mit der gleichen Kapazität extrahiert wurden, wie der Referenzextraktor Quickcap. Der Mittelwert der Verteilung steigt auf 17,4 Prozent, so dass im Mittel mit einem extraktionsbedingten

Messfehler gerechnet werden muss, der bereits in der Größenordnung der Schwankungen des Prozesses liegt: Bildet man die Differenz aus den worst-case Werten von Quickcap und den typical-case Werten, normiert auf die typical-case Werte, so ergibt sich im Mittel eine Abweichung von 17,7 Prozent (siehe hierzu Bild 4.11).

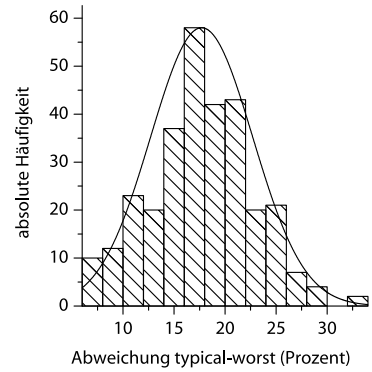


Bild 4.11. Abweichung der Kapazitätswerte der Cluster im ungünstigsten Fall von den typischen Prozessbedingungen (worst-case versus typical-case).

Das letzte Extraktionsprogramm in diesem Vergleich, Diva von Cadence, stellt ein relativ einfaches aber schnelles Werkzeug zur worst-case Abschätzung der Kapazität von primitiven Strukturen dar. Der Ausdruck Abschätzung bedeutet, dass es in erster Linie nicht um Präzision geht, sondern um eine zuverlässige Aussage, ob die Kapazität eines vorgegebenen Layouts eine bestimmte Schranke überschreitet oder nicht. Die Schätzwerte werden dabei von Diva sehr schnell ermittelt, d.h. im Bereich von wenigen Sekunden, selbst bei großen Chips. Möglich wird dies durch eine sehr einfache Vorgehensweise: Diva errechnet für ein gegebenes Leitungsstück das Produkt aus Größe und Kapazität pro Fläche der Metalllage. Die Werte für die Kapazität pro Fläche wiederum entnimmt Diva der Technologiedatei, in der gemessene oder vorausberechnete Werte für alle Ebenen vom Benutzer eingetragen werden müssen. Im Gegensatz zu anderen auf Nachschlagetabellen basierenden Verfahren wie Assura oder Calibre-xRC werden die Strukturen des Layouts nicht algorithmisch zerlegt, sondern immer als einfaches Rechteck angesehen.

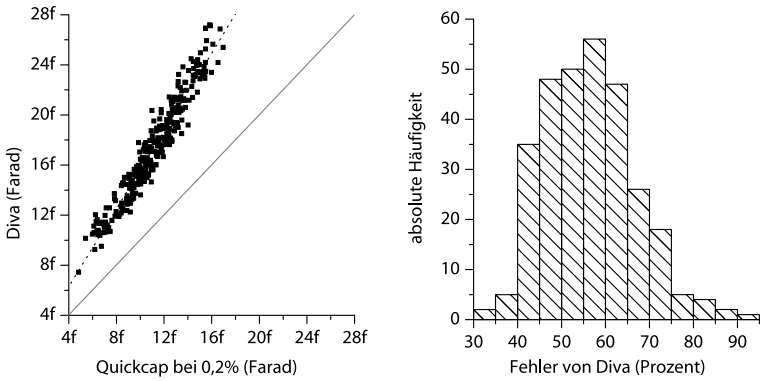


Bild 4.12. Diva (worst-case) mit einem mittleren Fehler von 56%

Die bei diesem Prinzip entstehenden Fehler sind im Falle der Kapazitätscluster in Bild 4.12 zu sehen. Es handelt sich bei den Clustern um besonders ungünstige Strukturen, für die Diva nicht ausgelegt wurde. Deutlich wird dies nicht nur am großen Öffnungswinkel der beiden Geraden in der linken Abbildung, sondern besonders am mittleren Fehler von über 50 Prozent (rechter Graph). Die Fehlerverteilung ist nicht normalverteilt und weist für einen Cluster einen Maximalfehler von 92,8 Prozent auf. Über ein Viertel der Cluster wurden von Diva mit einem Fehler von mehr als der Hälfte ihres wahren Wertes extrahiert. Auf der anderen Seite wurde keiner der Cluster von Diva *unterschätzt*. Als Richtwert für eine worst-case Abschätzung ist Diva also durchaus geeignet (obwohl die Extraktionsregeln auf den typischen Prozessparametern beruhen). Für Informationen über den wahren Kapazitätswert ist die Schätzung jedoch zu ungenau.

Quickcap (0,2%)		Assura-FS		Assura		Calibre	Diva	Abbildung
typ.	w.c.	typ.	w.c.	typ.	w.c.	w.c.	w.c.	
4,91 fF	5,41 fF	5,19 fF 5,5%	5,83 fF **7,8%**	5,29 fF 7,7%	6,15 fF 13,7%	6,93 fF 28,1%	10,15 fF 87,6%	
5,24 fF	6,19 fF	5,80 fF 10,8%	6,45 fF 4,3%	4,40 fF **-16%**	5,56 fF **-10,2%**	6,55 fF **5,9%**	9,25 fF 49,5%	
5,52 fF	6,30 fF	5,25 fF -4,8%	6,49 fF 3%	6,01 fF 9%	7,05 fF 11,9%	8,82 fF **39,9%**	11,48 fF 82,2%	Farbtafel I
5,55 fF	6,12 fF	5,23 fF -5,9%	6,46 fF 5,6%	6,29 fF 13,2%	7,28 fF 19%	8,33 fF 36,1%	10,60 fF 73,3%	
5,63 fF	6,24 fF	6,12 fF 8,6 %	6,66 fF 6,6%	6,01 fF 6,8%	7,01 fF 12,4%	7,82 fF 25,3%	12,03 fF **92,8%**	
6,21 fF	7,43 fF	5,79 fF **-6,9%**	7,51 fF 1,1%	6,50 fF 4,7%	7,99 fF 7,6%	9,12 fF 22,8%	12,57 fF 69,2%	Farbtafel II
6,22 fF	7,35 fF	5,95 fF -4,4%	7,57 fF 3%	7,43 fF **19,3%**	9,18 fF **25%**	8,39 fF 14,3%	10,94 fF 48,9%	
9,07 fF	11,08 fF	10,45 fF **15,3%**	11,34 fF 2,3%	9,00 fF -0,7%	11,29 fF 1,9%	12,30 fF 11%	15,61 fF 40,8%	Farbtafel III
10,25 fF	13,23 fF	10,22 fF -0,3%	13,03 fF -1,5%	11,73 fF 14,5%	15,05 fF 13,7%	16,37 fF 23,7%	23,23 fF 75,6%	Farbtafel IV
12,57 fF	15,02 fF	12,49 fF -0,6%	14,37 fF **-4,3%**	13,59 fF 8,1%	16,47 fF 9,6%	17,66 fF 17,6%	22,87 fF 52,3%	Farbtafel V

Tabelle 4.3. Überblick über die extrahierten Kapazitäten einiger ausgewählter Cluster. Die angegebenen Prozentzahlen repräsentieren den relativen Fehler der jeweils extrahierten Absolutwerte. Dabei wurden die Werte aus Quickcap als korrekt angenommen. Diese liegen mit einer Wahrscheinlichkeit von 68% im Bereich der Standardabweichung von 0,2%. Mit 68% Wahrscheinlichkeit beträgt der Fehler der Werte aus Quickcap also nur 0,2%, mit 95% Wahrscheinlichkeit 0,4% und mit mehr als 99% Wahrscheinlichkeit 0,6%.

4.1.2 Überblick und Vergleich.

Durch den Vergleich der mit den vorgestellten Werkzeugen extrahierten Kapazitäten mit den als korrekt angenommen Werten aus Quickcap bei 299 Clustern konnte ein mittlerer Fehler von -0,4 Prozent und 6,5 Prozent (Assura-FS und Assura) für typische Prozessbedingungen und ein mittlerer Fehler von 1,1 Prozent, 9,2 Prozent, 17,4 Prozent und 56 Prozent (Assura-FS, Assura, Calibre-xRC und Diva) für die ungünstigsten Prozessbedingungen ermittelt werden. Der Begriff Mittelwert suggeriert das Vorliegen einer Normalverteilung, er ist jedoch auch für beliebige Verteilungen definiert und auf solche anwendbar. Durch Anwendung eines Tests auf Normalverteilung nach Shapiro-Wilk wurde untersucht, ob bei den Fehlerverteilungen der analysierten Extraktoren eine Normalverteilung vorliegt. Das Resultat des Tests ist in Tabelle 4.4 zu sehen, wobei der W-Wert den berechneten Korrelationskoeffizienten der Verteilung darstellt. Den Ergebnissen liegt ein Signifikanzniveau von $\alpha = 5\%$ zugrunde, so dass die Hypothese der Normalverteilung bei

$P < 0{,}05$ verworfen wurde. Zu den Einzelheiten des Tests nach Shapiro-Wilk sei auf die weiterführende Literatur verwiesen, siehe Precht, Kraft, Bachmaier 1999.

In zwei Fällen wurde bei Anwendung des Tests die Hypothese schließlich verworfen: Bei Diva und Assura-FS für typische Prozessbedingungen. Im ersten Fall verwundert das Ergebnis weniger, da der Extraktor nur eine sehr ungenaue worst-case Schätzung abgibt. Bei Assura-FS überrascht jedoch das für die beiden Prozessbedingungen verschiedene Ergebnis. Beim Wechsel vom worst-case zum typical-case wird aus der Normalverteilung eine unbekannte Verteilung. Eine Erklärung für die unterschiedliche Fehlerverteilung können nur weiterführende Untersuchungen liefern. An dieser Stelle sei lediglich die Vermutung geäußert, dass die durch die verschiedenen Leitungs- und Isolationsdicken bedingte voneinander abweichende Geometrie zu unterschiedlichen Diskretisierungsfehlern führt, die sich wiederum in einer atypischen Fehlerverteilung niederschlagen kann. Ein Indiz dafür bietet die im Vergleich mit dem worst-case breitere Fehlerverteilung bei typischen Prozessparametern.

Extraktor		W	P	NV
Assura-FS	typ	0,828	0	–
	wc	0,986	0,742	✓
Assura	typ	0,985	0,657	✓
	wc	0,995	0,999	✓
Calibre	wc	0,979	0,137	✓
Diva	wc	0,973	0,011	–

Tabelle 4.4. Ergebnisse des Tests auf Normalverteilung (NV). Das Signifikanzniveau beträgt 5%.

In Tabelle 4.3 sind zusammenfassend einige Kapazitätscluster aufgelistet, bei denen die bisher betrachteten Extraktionstools besonders große Fehler machten. Es zeigt sich, dass besonders große Fehler bei einem Extraktor nicht zwangsläufig zu ebenfalls großen Fehlern bei einem anderen Programm führt. Sogar der Wechsel zwischen Prozessparametern unter Beibehaltung der Software kann zu völlig verschiedenen Fehlern führen. Daraus kann schließlich geschlossen werden, dass innerhalb der betrachteten 299 Cluster keine besonderen Cluster vorhanden sind, die bei *allen* Extraktoren zu besonders großen Fehlern führen, etwa durch eine besondere, von der Regel abweichende Struktur. Darüber hinaus kann ausgeschlossen werden, dass es einen algorithmischen oder numerischen Fehler gibt, den alle Programme gemein hätten.

Laufzeiten

Bei der bisherigen Analyse stand die Genauigkeit bzw. der Fehler der betrachteten Extraktionswerkzeuge im Vordergrund. Einen Hinweis auf die Laufzeit gab es nur im Zusammenhang mit Diva. Im Rahmen dieser Arbeit ist die Laufzeitfrage nicht von besonderer Bedeutung, soll hier aber dennoch kurz untersucht werden, um den Vergleich der Extraktionstools zu einem runden Abschluss zu bringen.

In Bild 4.13 ist die Laufzeit von Quickcap für die 299 bisher untersuchten Kapazitätscluster als Punkteverteilung gegen die mit 0,2% Genauigkeit extrahierten Werte aufgetragen (typische Prozessbedingungen). Zwischen der Größe des Kapazitätswertes und der Laufzeit besteht offensichtlich kein Zusammenhang, die Häufung der Punkte bei 15 Minuten erstreckt sich von ca. 8 Femtofarad bis hin zu 13 Femtofarad. Entgegen der Erwartung ist die Laufzeit bei einigen „kleinen" Clustern, also solchen, die Quickcap mit einer kleinen Kapazität angibt (4fF bis 8fF), sehr viel größer, als bei großen Clustern. Beispielsweise benötigte Quickcap bei einem Cluster mit 5,55 Femtofarad 91 Minuten, um das Genauigkeitsziel von 0,2 Prozent zu erreichen, während ein anderer Cluster mit 14,9 Femtofarad lediglich 2,95 Minuten benötigte. Betrachtet man die drei Cluster, die im Bild 4.13 bei ca. 90 Minuten zu sehen sind, im Detail, so stellt man fest, dass sich diese in den Zeilen 4 bis 6 der

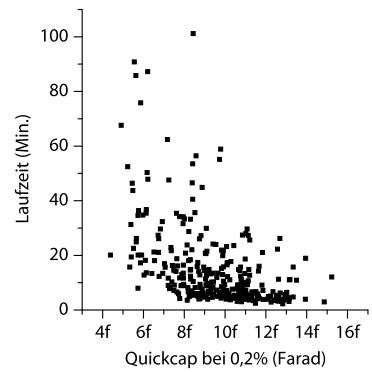


Bild 4.13. Laufzeit von Quickcap bei 0,2% Genauigkeit für die untersuchten Kapazitätscluster und bei typischen Prozessbedingungen.

Extraktor		Mittel	Std.-Ab.	Max
Quickcap bei 0,2%	typ	15,2m	15,4m	1,69h
	wc	11,5m	12,8m	1,42h
Assura-FS	typ	37,3s	15,5s	1,8m
	wc	35,9s	12,5s	1,5m
Assura	typ	11,3s	1,1s	22s
	wc	11,6s	1,2s	21s
Calibre	wc	4,0s	1,0s	18s
Diva	wc	0,4s	0,5s	2s

Tabelle 4.5. Laufzeit der Extraktoren bei den 299 Kapazitätsclustern. Gegeben sind Mittelwert, Standardabweichung und Maximalwert.

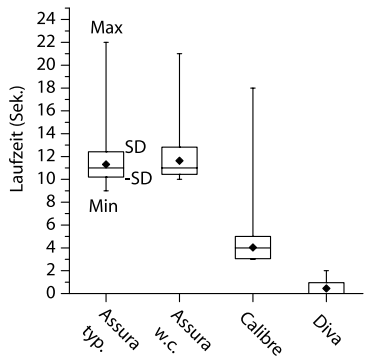


Bild 4.14. Laufzeit der auf Geschwindigkeit optimierten Extraktionswerkzeuge. Der schwarze Punkt in der Mitte der Boxen repräsentiert den Mittelwert.

Tabelle 4.3 wiederfinden (unter 5,55 fF, 5,63 fF und 6,21 fF), also einen sehr großen Fehler bei den anderen Extraktoren aufweisen (der Cluster mit 6,21 Femtofarad und 87 Minuten Laufzeit ist auch in Farbtafel II auf Seite 148 zu sehen). Dieser Zusammenhang lässt sich auch bei den anderen Clustern beobachten, für die Quickcap eine besonders lange Rechenzeit benötigte. Damit wird klar, warum die Laufzeit nicht mit der „Größe" der Cluster steigt: Die Rechenzeit in Quickcap orientiert sich in erster Linie an der strukturellen Komplexität, nicht an der Gesamtkapazität. Cluster mit großer Kapazität können einfacher aufgebaut sein, als solche mit kleiner Kapazität. Die Laufzeit in Quickcap stellt also ein einfaches Maß für die Komplexität der in dieser Arbeit vorgestellen Kapazitätscluster dar.

Die unregelmäßige Verteilung der Punkte in Bild 4.13 bedeutet jedoch auch, dass es schwierig ist, die Laufzeiten der einzelnen Extraktionswerkzeuge miteinander zu vergleichen. Allenfalls statistische Größen wie der Mittelwert der Laufzeitdaten jedes Extraktors, sowie die Standardabweichung und die Minimal-/Maximalwerte können als Grundlage für einen Vergleich dienen. In Tabelle 4.5 sind diese Kenngrößen für alle vorgestellen Programme zusammengefasst. Die Werte aus Quickcap sind dabei erwartungsgemäß sehr viel größer, als bei den anderen Extraktoren, schließlich wurde Quickcap auf eine sehr hohe Präzision auf Kosten der Laufzeit eingestellt. An zweiter Stelle rangiert Assura-FS, der Extraktor mit der größten Genauigkeit nach Quickcap. Alle anderen Programme fallen sowohl algorithmisch, als auch in Hinblick auf Exaktheit in eine andere Kategorie, ihre Laufzeit liegt immer bei wenigen Sekunden. Dies gilt insbesondere für Diva, dessen Aufgabe in der *schnellen* Abschätzung der Kapazitäten im worst-case liegt und dessen Eignung hierdurch bestätigt wird.

Da Quickcap und Assura-FS eine eigene Klasse an Extraktionswerkzeugen darstellen (TCAD-Klasse, siehe Abschnitt 2.2.2), die über aufwendige numerische Verfahren zur Lösung der Laplace'schen Gleichung verfügen, sollen im Folgenden beide Programme einem von den anderen Extraktoren getrennten Vergleich unterzogen werden.

Genauigkeitseichung

Ein Vergleich der Laufzeitunterschiede zwischen Quickcap und Assura-FS ist zwar prinzipiell möglich, wird jedoch durch die einstellbare Genauigkeit und die statistischen Schwankungen der Ergebnisse bei Quickcap erschwert: Beide Programme müssen mit der gleichen Genauigkeit extrahieren, damit die Vergleichbarkeit gewährleistet ist[27]. Wird Quickcap wie im vorherigen Abschnitt auf ein Genauigkeitsziel von ±0,2 Prozent gesetzt, so beträgt die durchschnittliche Laufzeit 15 Minuten, während Assura-FS nur 37 Sekunden im Schnitt benötigt, jedoch weniger genau ist. Ein Laufzeitvergleich wäre bei dieser Vorgabe also unfair.

Eine mögliche Lösung besteht darin, aus der Fehlerverteilung in Bild 4.5 einen Genauigkeitswert für Assura-FS zu errechnen. Wird dabei der mittlere Fehler („root-mean-square", RMS) als ein solcher Wert gewählt, so stellt er das Analogon zum einstellbaren Genauigkeitsziel in Quickcap dar. Wird Quickcap nun auf dieses Genauigkeitsziel gesetzt, so kann die Laufzeit beider

27. Darüber hinaus kann argumentiert werden, dass der Speicherplatzbedarf der Programme ebenso in den Vergleich eingehen sollte.

Programme verglichen werden. Wie schon im Zusammenhang mit dem Fehlerfortpflanzungsgesetz von Gauß in Box 4.1 auf Seite 102 diskutiert wurde, wirken sich die statistischen Schwankungen von Quickcap dabei auf den beobachteten bzw. berechneten RMS-Fehler aus. Die durch Gleichung 4.5 gegebene Beziehung zwischen dem RMS-Fehler auf der linken Seite und den Standardabweichungen der in den RMS-Fehler eingehenden Kapazitätswerte der Extraktoren X (Quickcap) und Y (hier Assura-FS) auf der rechten Seite setzt zunächst voraus, dass die einzelnen Punkte x_j und y_j einer Messreihe *eines* Layouts (Clusters) enstammen. Es handelt sich also nicht um eine Reihe *verschiedener* Cluster, für die jeweils eine Einzelmessung vorgenommen wird. Da jedoch Assura-FS aufgrund seines Determinismus für ein festes Layout immer den gleichen Kapazitätswert berechnet, ist die direkte Anwendung von Gleichung 4.5 nicht möglich. Stattdessen wurde in den bisherigen Vergleichen immer mit verschiedenen Kapazitätsclustern gearbeitet, für die pro Extraktor jeweils nur ein Wert bestimmt wurde. Da der Mittelwert der Einzelmessung (Extraktion) eines Clusters i gerade dem Messwert entspricht, kann $\bar{x} = x_i$ gesetzt werden. Schließlich wird der Übergang von der Messreihe eines Einzelclusters zu Einzelmessungen einer Reihe von Clustern durch Änderung des Laufindex $j = 1 \ldots M$ hin zu $i = 1 \ldots N$ vollzogen, wobei N der Gesamtzahl der Cluster entspricht, also wie bisher 299. Aus Gleichung 4.5 wird so:

$$\frac{1}{N}\sum_i \left(\frac{y_i - x_i}{x_i}\right)^2 = g_x^2 + \left(\frac{\sigma_y}{x_i}\right)^2 \tag{4.6}$$

Damit wird es nun möglich, den Ausdruck σ_y / x_i in Gleichung 4.6 als (auf Quickcap) normierte Standardabweichung des deterministischen Extraktors Y, hier also Assura-FS, zu interpretieren. Dem für einen festen Cluster immer gleichen Wert von Assura-FS wird also eine statistische Abweichung zugeordnet, die als charakteristischer Fehler des Extraktors aufgefasst wird. Durch Auflösen der Gleichung 4.6 erhält man schließlich:

$$\frac{\sigma_y}{x_i} = \sqrt{\frac{1}{N}\sum_i \left(\frac{y_i - x_i}{x_i}\right)^2 - g_x^2} \tag{4.7}$$

Gleichung 4.7 stellt damit eine Korrekturvorschrift für die Beeinflussung des beobachteten bzw. berechneten RMS-Fehlers von Programm Y durch die statistischen Schwankungen des Vergleichsextraktors X dar. Wird als Referenz wieder Quickcap mit 0,2 Prozent Genauigkeit eingesetzt, so ist der Einfluss der Schwankungen von Quickcap so gering, dass der RMS-Fehler von Assura-FS unter Vernachlässigung von Quickcap näherungsweise als charakteristischer Gesamtfehler angesehen werden kann.

Der RMS-Fehler von Assura-FS kann der Standardabweichung der Fehlerverteilung in Bild 4.5 und Bild 4.8 entnommen werden. Im Fall typischer Prozessbedingungen liegt dieser bei 3,8 Prozent, im worst-case bei 1,8 Prozent. Mit diesen beiden Werten wurden die 299 Kapazitätscluster erneut von Quickcap extrahiert und die resultierende Laufzeit aufgezeichnet. Die statistischen Kenngrößen der Laufzeitdaten sind in Bild 4.15 zu sehen. Der schwarze Punkt repräsentiert wieder den Mittelwert, die Standardabweichung ist durch SD markiert. Beim direkten Vergleich der Laufzeiten des typical-case schneidet Quickcap demnach sehr viel besser ab, als Assura-FS. Durch die geringe Genauigkeit von 3,8 Prozent verringerte sich die Rechenzeit von Quick-

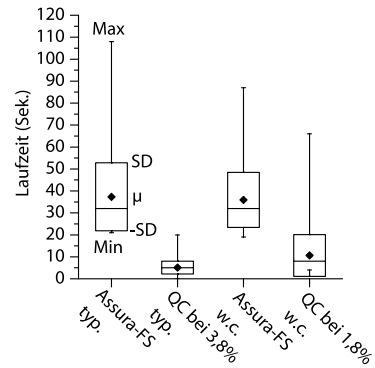


Bild 4.15. Laufzeit der TCAD-Klasse der Extraktionswerkzeuge.

cap beträchtlich. Anders verhält es sich im Fall der ungünstigsten Prozessbedingungen. Hier war Assura-FS sehr viel genauer bei der Berechnung, dementsprechend höher wurde das Genauigkeitsziel in Quickcap gesetzt. Dies schlägt sich in der etwas höheren Rechenzeit von Quickcap im Vergleich zum typical-case durch, so dass Quickcap nun größenordnungsmäßig im Bereich von Assura-FS liegt.

Der Unterschied im Ergebnis liegt darin, dass die Werte von Assura-FS durch die hohe Anlaufzeit des Programms nach oben versetzt sind. Dies liegt daran, dass Assura-FS vor dem eigentlichen Start der Berechnungen eine Reihe temporärer Dateien anlegt und einige Vorverarbeitungen vornimmt, die bei allen Extraktionsvorgängen anfallen. Für einen fairen Vergleich der Algorithmen beider Extraktoren muss diese Startzeit abgezogen werden.

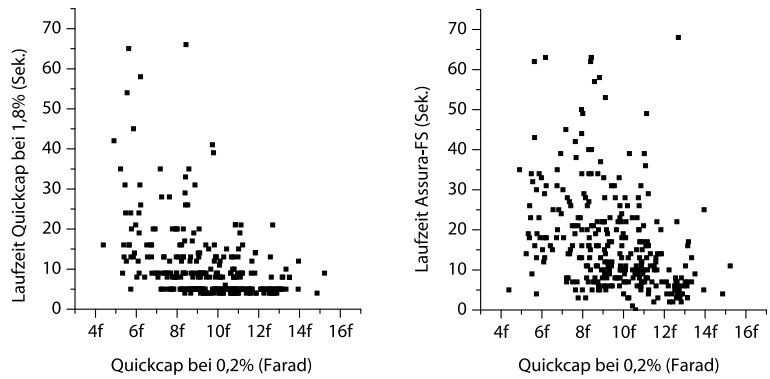


Bild 4.16. Vergleich der Laufzeiten von Quickcap und Assura-FS. Jeder Punkt repräsentiert einen Kapazitätscluster (ungünstigste Prozessbedingungen).

In Bild 4.16 ist das Ergebnis eines solchen Vergleichs zu sehen. Die Verteilung der Punktwolken ist in beiden Abbildungen sehr ähnlich und wie bereits in Bild 4.13 besteht bei beiden Extraktoren kein erkennbarer Zusammenhang zwischen Größe der Kapazität und Laufzeit. Auf einen ähnlichen Punktwolkenvergleich für die Laufzeiten des typical-case wurde verzichtet, da die Werte von Assura-FS weit über denen von Quickcap liegen. Unter typischen Prozessbedingungen scheint Assura-FS aufgrund der vermuteten Fehler bei der Diskretisierung der Geometrie generell sehr viel schlechter abzuscheiden, als Quickcap, was sich in einer viel geringeren Genauigkeit und höheren Laufzeit niederschlägt.

* * *

4.2 Der Prober-Testchip

Alle Zeilen des Testchips sind vom Layout her identisch. Innerhalb der Zeilen gibt es drei Gruppen von Ladungspumpen: Die unbeschalteten (Spalten 1, 2, 25, 26, 51, 52) und die mit Clustern versehenen Ladungspumpen (Spalten 15 bis 51), sowie einige, die mit einfachen Strukturen wie z.B. Plattenkondensatoren bestückt wurden (Spalten 3 bis 14). Die unbeschalteten Ladungspumpen wurden für das Messverfahren benötigt (Nettostrombildung), während die letzte Gruppe zum Vergleich der Cluster (Gruppe 2) mit ihrem komplexen und zufälligen Aufbau mit einfachen, regulären Strukturen dienen sollten.

4.2.1 Strukturvergleich

Einfache Strukturen

Unter diesen Strukturen befinden sich auch einige Plattenkondensatoren, die man idealerweise auch zur Berechnung der Oxydschichtdicke heranziehen kann. Zum Zwecke der Prozesskontrolle werden häufig sehr große Plattenkondensatoren an geeigneten Stellen auf den Wafern oder speziellen Testwafern untergebracht, die sehr gut als einfacher Plattenkondensator modelliert werden können, so dass die Isolationsschichtdicke aus der gemessenen Kapazität präzise berechnet werden kann. Hierfür sind die Strukturen jedoch *nicht* geeignet, da aufgrund ihres hohen Umfang-zu-Fläche Verhältnisses die Rückrechnung wesentlich erschwert wird (Randeffekte bzw. Streufelder haben einen zu hohen Anteil an der Gesamtkapazität).

Eine Möglichkeit besteht in der Theorie darin, in einem iterativen Extraktionsverfahren die Oxydschichtdicke solange zu variieren, bis die berechnete Kapazität der gemessen entspricht. Mit dem Extraktor Quickcap wäre dies hier möglich gewesen. Leider standen die Lizenzen für Quickcap zum Zeitpunkt der Analyse nicht mehr zur Verfügung[28], so dass nur noch ein anderes Werkzeug aus der Klasse der Field-Solver (aufgrund der hohen Genauigkeit) hierfür geeignet gewesen wäre. Assura-FS als einzig verfügbare Alternative schied jedoch auch aus, da eine Änderung der Prozessparameter in Assura-FS einen lang andauernden Vorverarbeitungsdurchlauf erfordert[29].

Somit war eine Eichung der Extraktoren auf die (mittleren) Prozessparameter der hergestellten Testchips nicht möglich. Im nun folgenden Vergleich der gemessenen Kapazitäten mit den extrahierten Werten ergibt sich deshalb ein teilweise großer Gangunterschied („offset“), den es zu berücksichtigen gilt.

Der in Tabelle 3.9 auf Seite 86 bereits behandelte Fall eines Poly1-Poly2 Plattenkondensators ist ein solches Beispiel. Hier war die im Prozessparameterhandbuch des Herstellers spezifizierte Oxydschichtdicke wesentlich geringer als die gemessene, die tatsächliche Kapazität also zu klein. In Tabelle 4.6 ist diese Messung oben nochmals aufgelistet. Alle Daten in der Tabelle wurden auf Grundlage der fünf gemessenen Testchips erhoben, spiegeln also die

	Messwerte über alle Dies		
Struktur	**Min.**	**Mittel**	**Max.**
Pol1–Poly2 (Spalte 3)	73,37 fF	75,57 fF	78,2 fF
Subs.–Met1 (Spalte 4)	7,83 fF	8,08 fF	8,33 fF
Subs.–Poly1 (Spalte 5)	24,51 fF	25,50 fF	26,41 fF
Poly1–Met1 (Spalte 6)	6,52 fF	6,77 fF	7,28 fF
Met1–Met2 (Spalte 7)	6,13 fF	6,32 fF	6,52 fF
Met1–Met3 (Spalte 8)	3,25 fF	3,40 fF	3,66 fF
Met2–Met3 (Spalte 9)	6,06 fF	6,28 fF	6,68 fF
Met1 ‖ Met1 (Spalte 10)	2,82 fF	3,05 fF	3,20 fF

Tabelle 4.6. Kapazitätswerte der Plattenkondensatoren (Spalte 3 bis 9) bzw. der parallelen Met1-Bahnen (Spalte 10) aus den Messwerten aller fünf getesteten Chips (Dies 15, 17 bis 20).

28. Die von Magma Inc. freundlicherweise unentgeldlich bereitgestellen Lizenzen waren auf ein Jahr beschränkt. Ein Kauf zum Zwecke der Lizenzverlängerung schied aufgrund des hohen Preises (ca. 100.000 Euro) aus.
29. Generierung der Datei „rcxfs.dat“ mit dem Programm „capgen“. Die Laufzeit beträgt jeweils mehrere Stunden.

	Kapazität		
Struktur	**best**	**typical**	**worst**
Pol1–Poly2 (Spalte 3)	90,38 fF	90,38 fF	90,51 fF
Subs.–Met1 (Spalte 4)	7,82 fF	8,99 fF	11,09 fF
Subs.–Poly1 (Spalte 5)	26,3 fF	28,58 fF	31,75 fF
Poly1–Met1 (Spalte 6)	6,80 fF	8,63 fF	13,66 fF
Met1–Met2 (Spalte 7)	5,26 fF	6,50 fF	12,35 fF
Met1–Met3 (Spalte 8)	3,4 fF	3,77 fF	4,99 fF
Met2–Met3 (Spalte 9)	5,29 fF	6,53 fF	12,37 fF
Met1 ‖ Met1 (Spalte 10)	3,05 fF	3,33 fF	3,69 fF

Tabelle 4.7. Extraktionswerte der Plattenkondensatoren in den Spalten 3 bis 9, sowie der parallel verlaufenden Met1-Bahnen (Spalte 10) aus Assura-FS (Field-Solver).

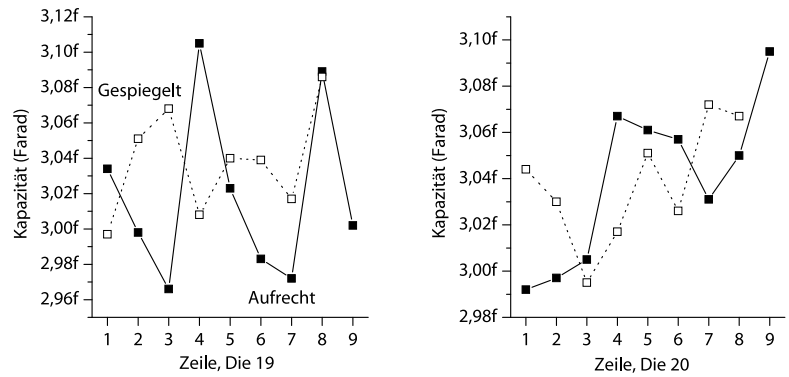
Bild 4.17. Kapazitätsverlauf zweier parallel verlaufenden Met1-Leiterbahnen in Spalte 10 (20 µm Länge, 0,45 µm Abstand) über die Zeilen zweier Testchips hinweg.

globale Chip-zu-Chip Schwankungsbreite wieder, ohne dabei den vollen Umfang der Wafer-zu-Wafer, Los-zu-Los oder gar Prozess-zu-Prozess Variationen bzw. Drifts zu erreichen (siehe auch Tabelle 2.2 auf Seite 25).

Zum Vergleich dieser Werte mit der Extraktion (Assura-FS) dienen die Daten in Tabelle 4.7. Hier ist der volle Umfang der Los-zu-Los Prozessschwankungen wiedergegeben, so wie der Hersteller diese spezifizierte. Aus den Werten wird deutlich, dass sämtliche Parameter, die auf die Kapazität von Strukturen unmittelbar über dem Substrat Einfluss haben, also die Feld-, Interpoly- und Poly-zu-Metall (Met1) Oxyddicken, außerhalb der Spezifikation liegen, wodurch die Extraktion in den ersten vier Fällen der Tabelle zu hohe Kapazitäten berechnete.

Die Oxyddicken auf den höheren Ebenen über Met1 stimmen dagegen mit den Messungen überein. Dort liegt die ermittelte Kapazität im Bereich zwischen der best-case Extraktion und den typical-case Werten. Im Fall der parallelen Metallbahnen in der letzten Tabellenzeile kommt zur horizontal verlaufenden Kapazität noch die Metall-zu-Substrat Kapazität hinzu, so dass der extrahierte Wert ebenfalls etwas zu hoch angesetzt ist.

Betrachtet man in diesem Fall den Kapazitätsverlauf über die Zeilen zweier Chips, so erhält man für Die 19 (links) und Die 20 (rechts) die Graphen in Bild 4.17. Die im Abschnitt „Erste Ergebnisse“ auf Seite 86 beobachtete Kapazitätszunahme zu hohen Zeilennummern hin ist nur noch bei Die 20 geringfügig vorhanden, bei allen anderen Testchips verläuft die Kapazität ähnlich unsystematisch wie in der linken Abbildung von Die 19. Erklären lässt sich dieses Verhalten mit der hauptsächlich horizontalen Ausrichtung der Feldlinien des Kondensators, so dass sich der Gradient in der Oxydschichtdicke kaum bemerkbar macht.



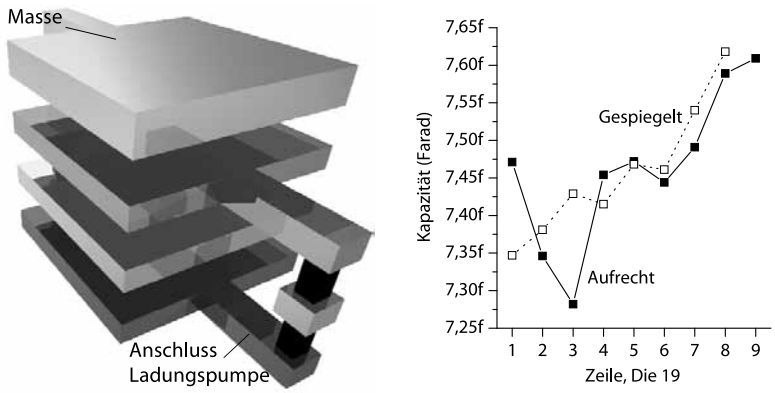
Der sehr geringe Unterschied zwischen den best-, typical- und worst-case Kapazitäten beim Poly1-Poly2 Plattenkondensator (erste Zeile von Tabelle 4.7) liegt daran, dass Assura-FS den Kondensator als Bauelement fester Größe erkennt, einzig und allein die parasitären Kapazitäten (z.B. Zuleitungen) werden den Parameterschwankungen des Prozesses entsprechend variiert. Schließlich verwundert noch der hohe worst-case Wert einiger Strukturen. Es handelt sich jedoch um keinen Berechnungsfehler, die Werte wurden mit Assura (ohne Field-Solver) auf Konsistenz überprüft. Ebenso kann ein Fehler bei den Technologiedaten ausgeschlossen werden.

Spezielle Strukturen

Neben den konventionellen Plattenkondensatoren (2 Ebenen) sind auch solche möglich, die aus mehreren, zusammengeschalteten Platten bestehen. Man erhält eine höhere Kapazitätsdichte, jedoch nicht unbedingt bessere Matching-Eigenschaften. Deshalb werden horizontal verlaufende Plattenkondensatoren („horizontal parallel plate", HPP) mit mehreren Ebenen in der Praxis kaum verwendet. In Bild 4.18 ist ein solcher Fall gezeigt, der Kapazitätsverlauf über die Zeilen eines Testchips exemplarisch für alle anderen daneben.

Bei Betrachtung der Schwankungsbreite dieser Struktur bestätigen sich die schlechten Matching-Eigenschaften: Ähnlich wie die parallelen Met1-Leiterbahnen und der sich über zwei Isolationsschichten erstreckende Met1-Met3 Plattenkondensator (siehe Tabelle 4.8) weist das HPP-Gebilde mit 11 Prozent eine recht hohe Spanne Minimum zu Maximum auf, der Matching-Fehler ist mit 1,26 Prozent (Er1) bzw. 1,11 Prozent (Er2) ebenfalls hoch.

„Matching-Fehler" ist in diesem Zusammenhang die Standardabweichung der relativen Kapazitätsdifferenz zwischen den aufrechten und den gespiegelten Strukturen jeweils *innerhalb einer* Zeile (Er1), bzw. zwischen direkt benachbarten Zeilen *derselben* Orientierung (Er2).



Im Gegensatz dazu ist die Schwankungsbreite des VPP-Plattenkondensators („vertical parallel plate", siehe Bild 4.20) mit einer Spanne von 8,8 Prozent und einem Matching-Fehler von 0,79 Prozent (Er1) wesentlich besser. Der Er2-Fehler ist mit 1,03 Prozent ebenfalls etwas kleiner. Die VPP-Struktur erreicht zwar nicht ganz die guten Matching-Eigenschaften eines simplen 2-fach Plattenkondensators auf den unteren Ebenen, weist jedoch eine recht hohe Kapazitätsdichte auf, so dass in Fällen, in denen der Herstellungsprozess keine zweite Polysilizium-Ebene mit dünnem Oxyd zur Verfügung stellt – beispielsweise in rein digitalen Prozessen – die VPP-Struktur eine echte Alternative darstellt. Der Substrat-Poly1 Kondensator scheidet trotz hoher Dichte je nach Anwendung und Prozess aus, da das Substrat (von getrennten Wannen abgesehen) zwingend auf Massepotential liegt und häufig hochohmig ist.

	Messwerte über alle Dies		
Struktur	**Er1**	**Er2**	**Spanne**
Pol1–Poly2	0,50%	0,39%	6,39%
Subs.–Met1	1,07%	1,02%	6,28%
Subs.–Poly1	0,55%	0,59%	7,45%
Poly1–Met1	1,27%	1,23%	11,13%
Met1–Met2	1,01%	0,90%	6,20%
Met1–Met3	2,81%	2,52%	12,15%
Met2–Met3	1,22%	1,09%	9,84%
Met1 ‖ Met1	2,65%	2,55%	12,42%

Tabelle 4.8. Matching-Fehler (Spalte 2 u. 3), d.h. Standardabweichung der relativen Kapazitätsdifferenz zwischen unmittelbar benachbarten Strukturen, sowie die Spanne Min.-Max. über alle Chips.

Bild 4.18. HPP-Struktur („horizontal parallel plate") in Spalte 14 von Die 19. Über alle Chips gemittelt beträgt die Kapazität 11,0 fF, mit einer Spanne von 1,21 fF, also 11% des Mittelwerts. Der Matching-Fehler beträgt 1,26% (Er1) bzw. 1,11% (Er2).

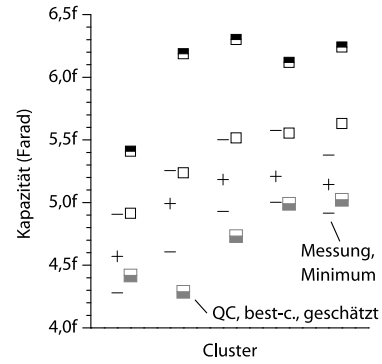


Bild 4.19. Mittelwert, Maximum und Minimum der Messwerte aller fünf Testchips im Vergleich mit den typical-, worst- und best-case Extraktionswerten aus Quickcap. Letztere wurden aus den worst-case Werten geschätzt. (Daten aus den ersten fünf Zeilen in Tabelle 4.9.)

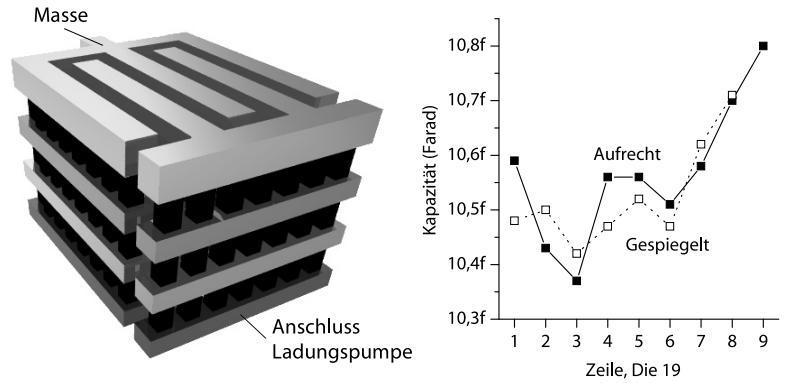


Bild 4.20. VPP-Struktur („vertical parallel plate“) in Spalte 11 von Die 19. Der Mittelwert beträgt 7,48 fF, die Spanne 685 aF (8,8%). Der mittlere Er1-Fehler liegt bei 0,79%, der mittlere Er2-Fehler bei 1,03%

Bei dieser Wertung sollte jedoch nicht vergessen werden: Die Plattenkondensatoren sind hier aufgrund des kleinmaschigen Matrixrasters der Ladungspumpen sehr klein, so dass die Anschlüsse und Randeffekte einen großen Anteil an der Gesamtkapazität haben. Sie spielen daher bei einfachen Plattenkondensatoren eventuell eine größere Rolle, als bei der VPP-Struktur. Für einen speziell auf die Vor- und Nachteile der verschiedenen Realisierungsformen gerichteten Vergleich sei daher auf Aparicio et al. 2002 verwiesen. Die Autoren kommen in dieser Publikation zu ähnlichen Ergebnissen hinsichtlich des Verhältnisses von Matching zu Kapazitätsdichte.

Quickcap	Messwerte über 5 Chips		
typ. ,wc.	µ(C)	Min, Max	Spanne
4,91 fF +0,5 fF	4,57 fF	4,28 fF 4,91 fF	13,72%
5,24 fF +0,95 fF	4,99 fF	4,61 fF 5,25 fF	13,00%
5,52 fF +0,78 fF	5,18 fF	4,93 fF 5,50 fF	11,05%
5,55 fF +0,57 fF	5,21 fF	5,00 fF 5,58 fF	11,02%
5,63 fF +0,61 fF	5,14 fF	4,92 fF 5,38 fF	9,00%
6,21 fF +1,22 fF	5,82 fF	5,56 fF 6,08 fF	8,92%
6,22 fF +1,13 fF	5,84 fF	5,57 fF 6,21 fF	10,95%
9,07 fF +2,01 fF	8,61 fF	8,32 fF 8,97 fF	7,56%
10,25 fF +2,98 fF	9,56 fF	9,20 fF 9,97 fF	8,01%
12,57 fF +2,45 fF	11,89 fF	11,43 fF 12,36 fF	7,82%

Tabelle 4.9. Vergleich der Werte aus der typical-case und worst-case Extraktion (best-case wurde nicht durchgeführt) mit Quickcap versus Messung. Es handelt sich um dieselben Cluster wie in Tabelle 4.3 auf Seite 110.

Die Cluster

Alle Cluster, für die im Rahmen des Vergleichs der Extraktionstools (Abschnitt 4.1) in Tabelle 4.3 auf Seite 110 Daten aufgelistet sind, wurden auf den Testchips in jeweils zwei Varianten (Spalten 15 bis 36) integriert und vermessen. Jede Struktur in Spalten mit geraden Nummern entspricht dem Spiegelbild des Clusters in den ungeraden Spalten (Spiegelung an der Vertikalen). Darüber hinaus wurden in den Spalten 37 bis 51 zusätzlich einzelne Cluster ohne Spiegelbild vorgesehen, sie finden sich nur in den Punktwolken-Diagrammen wieder.

Da die Cluster für die Extraktionstools bezüglich Spiegelungen isomorph sind, d.h. als identisch angesehen werden, werden die Spiegelbilder eines Clusters von den Programmen immer mit derselben Kapazität extrahiert[30]. Für einen Vergleich der Berechnungen mit den Messungen spielt die Unterscheidung also keine Rolle, beide Varianten wurden daher zunächst zusammengefasst.

In Tabelle 4.9 sind die Werte aus Quickcap den Messwerten gegenübergestellt. Da in den Werkzeugvergleich in Abschnitt 4.1 keine best-case Werte einbezogen wurden, kann nur die Differenz zwischen den typischen Werten und den worst-case Kapazitäten als Schätzungsgrundlage für die günstigsten Prozessparameter – also die kleinste Kapazität – herangezogen werden. Die Messwerte liegen dann allesamt im Bereich der Extraktion, mit einer starken Tendenz hin zu den best-case Kapazitäten.

30. Abgesehen von Quickcap, da dieses Tool statistisch arbeitet, d.h. bei jedem Extraktionsvorgang nur Kapazitätsschätzungen abgibt. Aber auch hier sind die Ergebnisse homogen.

Einen visuellen Eindruck vermittelt Bild 4.21. Darin sind die Cluster als Punktwolke abgebildet, wobei die x-Achse die gemessenen Maximalwerte angibt und die y-Achse den zugehörenden typical-case Wert aus der Extraktion. In dieser Konstellation ist die Übereinstimmung besser als beim Vergleich des Mittelwerts der Messungen mit Quickcap in Bild 4.22. Die Punkte wurden in diesem Graphen miteinander verbunden, so dass sich jeweils eine Linie ergab, die zunächst monoton steigend erscheint.

Bei genauerem Hinsehen lassen sich einzelne lokale Minima ausmachen, die weiteren Aufschluss über die Extraktionsgenauigkeit geben. Je stärker diese Minima ausgeprägt sind, desto größer ist der Berechnungsfehler des Tools, hier Quickcap. Wäre die Linie streng monoton steigend, so hätte die Software die Cluster in dieselbe Reihenfolge bezüglich der Kapazität einsortiert, wie die Messungen. Die *relative* Kapazität hätte Quickcap in diesem Fall richtig berechnet, unabhängig vom Absolutwert. Das Auftreten eines Minimums bedeutet jedoch, dass Quickcap einen (oder mehrere) Cluster kleiner extrahiert als den im Graphen links daneben liegenden Cluster, das Größenverhältnis in Wahrheit jedoch gerade anders herum ist.

Bei den 34 getesteten Clustern trat diese Situation im worst-case sechsmal auf, unter typischen Prozessbedingungen viermal. Das stärkste Minimum im typischen Fall ist in der Abbildung markiert und kennzeichnet ein Clusterpaar, bei dem Quickcap am weitesten von der Messungen entfernt lag. Mit Ausnahme einer einzigen Zeile eines Testchips lag der von Quickcap auf 9,66 Femtofarad taxierte Cluster (Spalte 40) in der gemessenen Kapazität immer um mindestens 100 Attofarad *über* dem Cluster (Spalte 41), den Quickcap mit 9,73 Femtofarad extrahierte. Bild 4.23 zeigt ihren Kapazitätsverlauf bei zwei der getesteten Dies, die Graphen ähneln dabei qualitativ denen der anderen.

Festzuhalten gilt, dass sich erstens mit der Änderung der Prozessparameter zwischen best-, typical- und worst-case die extrahierten Größen*verhältnisse* der Cluster ändern können, da die Zahl der Minima nicht konstant ist, und zweitens das Ausmaß des relativen Fehlers steigt, je weiter die in der Extraktion verwendeten Prozessparameter von den tatsächlichen abweichen. Dies wird aus den beiden Linien in Bild 4.22 deutlich.

Zusammenfassend kann man auf eine sehr hohe Extraktionsgenauigkeit für Quickcap schließen, falls die Prozessparameter stimmen. Unter den günstigsten Prozessbedingungen bzw. für Parameter zwischen best- und typical-case ist im vorliegenden Fall zu erwarten, dass die entsprechende Linie in Bild 4.22 weniger oder gar keine Minima aufweisen würde. Für einen quantitativen Vergleich müsste die Extraktion auf Basis der experimentell bestimmten Parameter nochmals durchgeführt werden, was aus den genannten Gründen nicht möglich war.

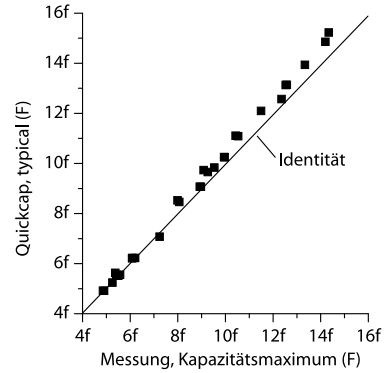


Bild 4.21. Vergleich der gemessenen Maximalkapazität der Cluster über alle fünf Testchips mit dem Wert aus der Extraktion bei typischen Prozessbedingungen. Jeder Punkt repräsentiert einen Cluster.

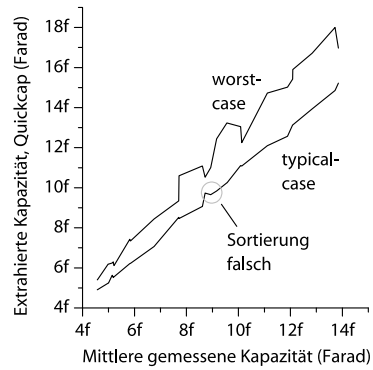


Bild 4.22. Trägt man die extrahierte Kapazität gegen die gemessene als Punktwolke auf (nicht gezeigt), so verläuft die Verbindungslinie streng monton, falls die größenmäßige Reihenfolge gleich ist. Andernfalls ergeben sich lokale Minima.

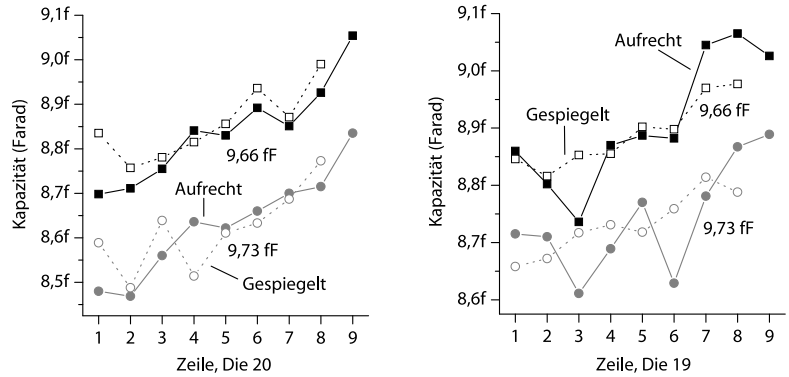


Bild 4.23. Clusterpaar, dessen Größenverhältnis in der Extraktion genau entgegengesetzt zur Messung war. Quickcap schätzte den größeren Cluster auf 9,66 fF, den kleineren auf 9,73 fF (typical-case).

4.2.2 Matching

Betrachtet man den Kapazitätsverlauf eines einzelnen Clusters in den beiden Varianten über die Zeilen eines Testchips hinweg, so findet man eine recht hohe Übereinstimmung bzw. gutes Matching zwischen der gespiegelten und der Ursprungsversi0n. In Bild 4.24 ist dies für den Cluster in den Spalten 35 und 36 zu sehen, seine typical-case Kapazität beträgt 12,57 Femtofarad (Quickcap). Auch hier gibt es wieder in jeder Zeile die aufrechten und die nach unten gespiegelten Ladungspumpen, so dass pro Cluster und Zeile alle vier möglichen Spiegelbilder vertreten sind. Denkt man sich die dazugehörenden vier Kurven übereinandergelegt, so wird die strukturelle Identität im gemeinsamen Kapazitätsverlauf deutlich.

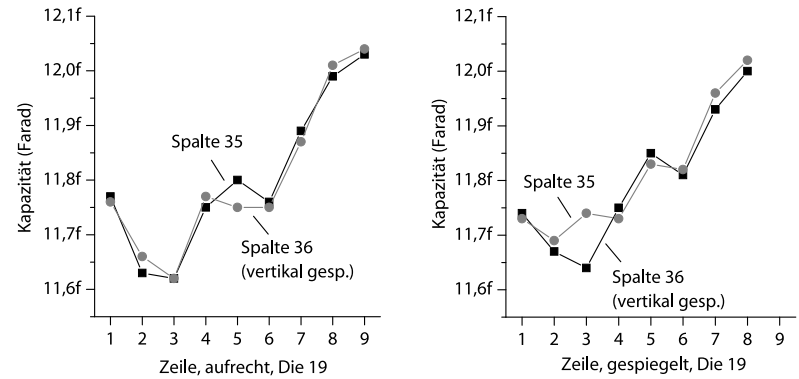


Bild 4.24. Kapazitätsverlauf des Clusters in den Spalten 35 und 36 (an der Vertikalen gespiegelt) bzw. aus Farbtafel V auf Seite 151. Die typical-case Kapazität aus Quickcap beträgt 12,57 fF, der worst-case Wert liegt bei 15,02 fF

Bei den anderen Testchips und anderen Clustern passen die Kurven dagegen weniger genau zusammen. Vielmehr handelt es sich bei dem in Bild 4.24 gezeigten Fall um ein seltenes Beispiel von außerordentlich gutem Matching. In Tabelle 4.10 sind die statistischen Größen der zehn Doppelcluster über alle fünf Testchips gegeben. Die Spiegelbilder der Cluster wurden bei der Berechnung wie getrennte Strukturen behandelt, so dass sich die Wertepaare geringfügig unterscheiden. Für die Berechnung des Matching-Fehlers wurden wie bisher Clusterpaare aus den jeweils an der Horizontalen gespiegelten Ladungspumpen gebildet (Er1) bzw. aus direkt übereinanderliegenden Zeilen derselben Orientierung (Er2). Das Matching der Paare aus geraden und ungeraden Zeilennummern wurde nicht berechnet.

Erwartungsgemäß ist die mittlere Kapazität dieser Clusterpaare nahezu identisch und ist damit eine Bestätigung der guten Reproduzierbarkeit und hohen Auflösung der Messmethode. Die größe Differenz besteht zwischen Spalte 17 und 18 mit nur 16 Attofarad, die der anderen ist noch weit geringer.

Bei der Analyse der beiden Kennzahlen des Matching-Fehlers fällt auf, dass die Werte mit der Spaltenzahl zu steigen scheinen. Es handelt sich dabei aber nicht um einen systematischen Messfehler, sondern um eine kapazitätsabhängige Tendenz. Große Cluster scheinen die zufälligen lokalen Prozessschwankungen prozentual gesehen besser ausgleichen zu können, als die kleinen Cluster. So liegen die Er1-Werte fast aller Cluster mit mehr als 8 Femtofarad bei 1,0 Prozent und darunter. Das Minimum liegt bei einem Er1-Wert von 0,5 Prozent (Spalte 49, mittlere Kapazität 13,7 fF), der niedrigste Er2-Fehler bei 0,68 Prozent in Spalte 32.

Da der Er1-Wert aus Clusterpaaren berechnet wurde, die jeweils spiegelsymmetrisch zur Horizontalen sind, der Er2-Wert jedoch aus Paaren gebildet wurde, die vom Entwurf her deckungsgleich bzw. identisch sind, wirken sich bei letzterem Prozessschwankungen weniger stark aus. Dadurch ist der Er1-Wert bei den kapazitiv kleineren Clustern meistens höher als der Er2-Wert.

Was den groben (tendenziellen) Zusammenhang des (lokalen) Matching-Fehlers mit der Clusterstruktur bzw. kapazitiven Größe angeht, so findet sich eine gewisse Übereinstimmung mit dem Drift der globalen Prozessparameter, die für die minimale und maximale gemessene Kapazität bzw. die Kapazitätsspanne verantwortlich sind (siehe Bild 4.25). Keine Übereinstimmung des Matching-Fehlers gibt es dagegen mit der Kapazitätsspanne aus den typical- und worst-case Extraktionen. Im Gegenteil, die Extraktion sagt bei den großen Clustern starke prozessbedingte Abweichungen voraus, wohingegen das lokale Matching laut Messergebnis besser wird.

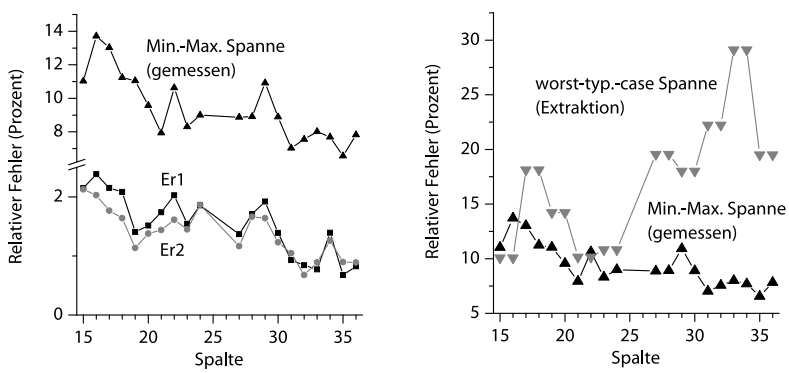


Bild 4.25. Spanne und die mittleren Er1- bzw. Er2- Fehlerwerte aus Tabelle 4.9 und Tabelle 4.10 in Abhängigkeit von den Clustern.

Damit gibt es keine Möglichkeit, aus den Extraktionswerten der Cluster a priori auf das *Matching* „in Realitas" zu schließen. Der Grund hierfür liegt in den unterschiedlichen Ursachen der Variationen: Für die best-, typical- und worst-case Kapazitäten ist der globale Gangunterschied der Leiterbahn- und Oxydschichtdicken über ganze Wafer oder Produktionsreihen verantwortlich, für das Matching die lokalen, zufallsbedingten Ungenauigkeiten, z.B. an Leitungsrändern. Hierzu bietet Abschnitt 2.1.1 weitere Grundlageninformationen.

Im Vergleich mit den herkömmlichen Metall-Metall Plattenkondensatoren kann zunächst festgehalten werden, dass die Cluster mit kleiner Kapazität ein deutlich schlechteres Matching aufweisen. Insbesondere der Poly1-Poly2

Messwerte über 5 Chips

Spalte	$\mu(C_u)$, $\mu(C_g)$	Er1	Er2
15	4,569 fF	2,15%	2,13%
16	4,572 fF	2,38%	2,03%
17	4,984 fF	2,15%	1,77%
18	5,000 fF	2,08%	1,64%
19	5,185 fF	1,40%	1,13%
20	5,182 fF	1,51%	1,38%
21	5,206 fF	1,74%	1,44%
22	5,212 fF	2,03%	1,61%
23	5,142 fF	1,54%	1,45%
24	5,148 fF	1,86%	1,86%
27	5,817 fF	1,37%	1,16%
28	5,825 fF	1,71%	1,66%
29	5,843 fF	1,92%	1,64%
30	5,846 fF	1,39%	1,23%
31	8,609 fF	0,93%	1,05%
32	8,616 fF	0,84%	0,68%
33	9,564 fF	0,77%	0,89%
34	9,556 fF	1,39%	1,26%
35	11,891 fF	0,68%	0,90%
36	11,887 fF	0,82%	0,89%

Tabelle 4.10. Pro Cluster gibt es zwei Varianten: Jene mit den ungeraden Spaltennummern (C_u), die wie beim Entwurf ausgerichtet sind, sowie die an der Vertikalen gespiegelten in den geraden Spalten (C_g). Er1 und Er2 sind wieder die Matching-Fehler.

Kondensator weist ein exzellentes Matching auf. Je größer die Cluster jedoch werden, desto geringer wird auch der Mismatch und erreicht Werte, die sogar unter denen einiger einfacher Plattenkondensatoren liegen (z.B. Substrat-Met1 und Poly1-Met1). Nur der über zwei Oxydschichten verlaufende Met1-Met3 Kondensator und die parallelen Met1-Leiterbahnen schwanken immer wesentlich stärker als die Cluster.

* * *

4.3 Der Schlüssel-Testchip

4.3.1 Testaufbau

Der Test des Schlüssel-Testchips wurde mithilfe einer kleinen Prototypen-Platine durchgeführt („Uxibo“[31], siehe Bild 4.26), die über einen programmierbaren Logikbaustein (FPGA „Spartan“ der Firma Xilinx) verfügt. Dieser diente zur Ansteuerung und zum Auslesen der Chips, die sich auf einer über die beiden Buchsenleiten mit dem Uxibo verbunde Erweitungsplatine befanden (nicht abgebildet). Diese zusätzliche Leiterplatte besaß im Wesentlichen einen PLCC-68 Sockel zum Einsetzen der Testchips, sowie einige Digital-Analog Konverter zur Erzeugung der Bias-Ströme und analogen Steuerungspannungen (z.B. für den Komparator) und Abblockkondensatoren zur Stabilisierung der Versorgungsspannung.

Die primäre Test- und Kontrolllogik wurde in Form einer Testapplikation auf einem Rechner realisiert, die über die USB-Schnittstelle mit dem FPGA des Prototypenboards kommunzierte. Die Funktionalität des Logikbausteins beschränkte sich in diesem Szenario letztlich auf die „low-level“-Steuerung (z.B. Takt- bzw. Pulsgenerierung), die Software übernahm die generelle Ablaufkontrolle (siehe Bild 4.27).

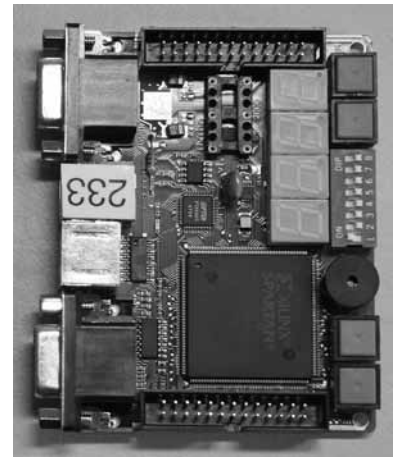


Bild 4.26. FPGA Prototypen-Testplatine „Uxibo“ mit USB-Schnittstelle und Buchsenleisten zum Anschluss von Erweitungshardware.

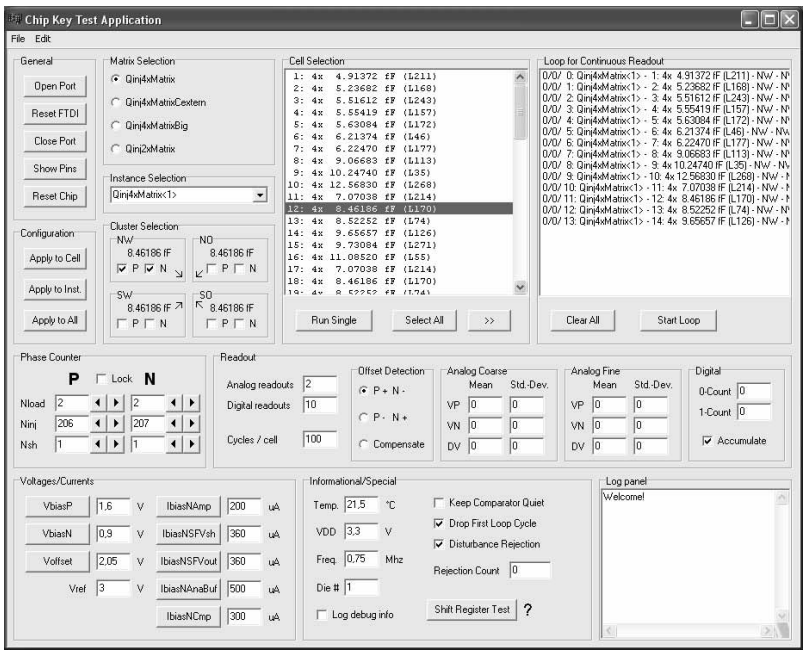


Bild 4.27. C++ Applikation zur Steuerung und zum Auslesen des Testchips über das „Uxibo“. Die auszulesenden Zellen können einzeln selektiert werden (Liste in der Mitte) und einer Jobliste (rechts) zur Auswertung hinzugefügt werden. Die Ergebnisse werden in eine separate Tabelle (nicht zu sehen) eingetragen und auf Wunsch in eine Datei gespeichert.

4.3.2 Auswertung

Nicht jede Matrix der drei Layoutvarianten konnte komplett ausgewertet werden, da die Ergebnisse durch (im nachhinein) unvermeidbare Störungen der Messungen verfälscht wurden. Erst im Laufe der Tests erwies sich näm-

31. Entwicklung am Lehrstuhl für Schaltungstechnik & Simulation, P. Fischer und M. Koniczek.

lich das Zusammenspiel aller Komparatoren, analogen Verstärker und Signalbuffer als Quelle solch starker Störungen[32], dass die Ergebnisse der Messungen an einer bestimmten Matrix durch das Schaltverhalten aller anderen beeinflusst wurden.

Der Grund für dieses Verhalten liegt darin, dass beim Entwurf des Testchips keine Vorkehrungen getroffen wurden, mit denen diese Störquellen hätten deaktiviert werden können. Alle Instanzen aller Layoutvarianten waren während der Tests gleichzeitig aktiv, unabhängig von der jeweils zu messenden, selektierten Zelle.

Anzahl Pumpzyklen

Gleichung 3.15 auf Seite 92 gibt das theoretische Optimum n_{opt} der Anzahl Pumpzyklen an, bei dem die Spannungsdifferenz $D(n,l,x)$ maximal ist, also die Messauflösung am höchsten ist. Näherungsweise kann $n_{\text{opt}} \approx l$ gesetzt werden.

In diese Rechnung geht allerdings nicht die Eigenschaft der Schalttransistoren ein, bei einer Drain-Source Spannung unterhalb der „Overdrive"-Spannung den Bereich der Sättigung zu verlassen. Die Simulation in Bild 4.28 zeigt einen Einbruch der Spannungsdifferenz D beim Erreichen des linearen Bereichs für mehr als 150 Pumpzyklen. Es handelt sich dabei um eine Zelle mit zwei Pumpzweigen (2-fach Layoutvariante), die mit Plattenkondensatoren der Größe 6,25 und 6,139 Femtofarad, also einer Kapazitätsdifferenz von 111 Attofarad, bestückt wurde. Die Simulation wurde auf Grundlage der extrahierten Leitungskapazitäten und der Transistormodelle des Prozesses durchgeführt.

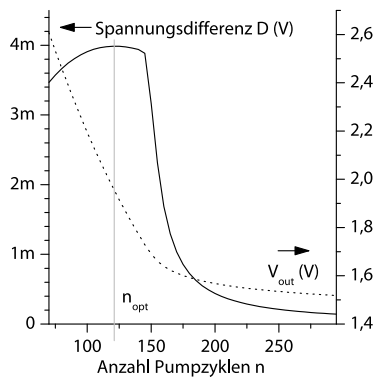


Bild 4.28. Simulation der Spannungsdifferenz $D(n, l, x)$ und von V_{out} in Abhängigkeit von der Anzahl der Pulse bei einer Kapazitätsdifferenz von 111,5 aF.

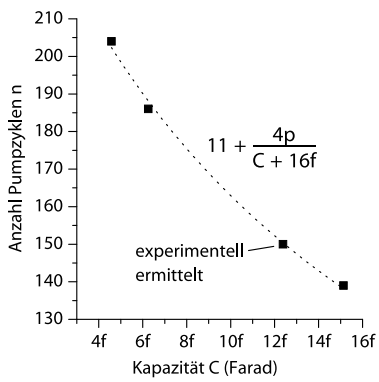


Bild 4.29. Ergebnis des nicht-linearen Kurvenfits (ermittelt mit der Software Origin). Der interne Knoten weist offenbar eine Kapazität von 16 fF auf.

Das starke Einbrechen der Spannungsdifferenz bedeutet, dass die Anzahl der Pumpzyklen einen gewissen Wert keinesfalls überschreiten sollte, um einen Verlust bei der Messauflösung zu vermeiden. Dieser Wert geht aus den bisher angestellten theoretischen Rechnung nicht hervor. Das Modell könnte zwar verfeinert werden und den Übergang der Transistoren zum linearen Bereich berücksichtigen, wäre jedoch nicht genau genug: Jede Ungenauigkeit bei der Bestimmung der Kapazität[33] C_{int} des internen Knotens hat starken Einfluss auf die optimale Zahl der Pumpzyklen:

$$n_{\text{opt}} \approx l = \frac{C_{\text{L}}}{C + C_{\text{int}}} \tag{4.8}$$

Setzt man für die Kapazität des internen Knotens die volle Diffusionskapazität der Transistor- und Wannendioden an, so erhält man in der Layoutvariante mit zwei Pumpzweigen zusammen mit der parasitären Leitungskapazität $C_{\text{int}} \approx 34\,\text{fF}$. Das Optimum n_{opt} liegt dann bei 100 Pumpzyklen und einer Spannungsdifferenz von $D \approx 3{,}4\,\text{mV}$. Wirken sich die Diffusionskapazitäten schwächer aus, so erhält man beispielsweise bei $C_{\text{int}} = 26\,\text{fF}$ ein Optimum von 124 Pumpzyklen und eine Spannungsdifferenz von $D \approx 4{,}2\,\text{mV}$, Werte die zur Simulation sehr gut passen.

32. Aufgrund hoher Ströme, digitaler Spannungspegel und oszillierendem Schaltverhalten wurden starke elektromagnetische Wechselfelder erzeugt, sowie Transienten auf den Stromversorgungsleitungen.
33. Den größten Anteil machen die Diffusionskapazitäten aus, ihr genauer Wert ist ohne Kenntnis der Spannungsverhältnisse unbestimmt.

Statt also die optimale Anzahl Pumpzyklen über Gleichung 4.8 zu bestimmen und dabei Gefahr zu laufen, das Optimum aufgrund der Unbestimmtheit von C_{int} zu überschreiten und einen Einbruch der Auflösung zu bewirken, wurde ein anderer Ansatz gewählt: Wie in Bild 4.28 zu sehen, sinkt die Spannung V_{out} (Ausgang des Source-Folgers in Bild 3.36 auf Seite 89) mit der Anzahl Pumpzyklen und erreicht bei n_{opt} einen Wert von ca. 1,9 Volt. Diese Spannung wird auf dem Testchip, wie bereits erwähnt, verstärkt und nach außen geführt. Auf diese Weise konnte n_{opt} experimentell für jede der drei Layoutvarianten und einige Werte der Messkapazität C ermittelt werden. Durch eine nicht-linearen Kurvenfit von Gleichung 4.8 an diese Werte ergaben sich schließlich drei Werte für C_{int} (sowie ein Offset-Wert[34]), die schließlich in der Steuerungssoftware benutzt wurden, um daraus n_{opt} für jeden Wert von C bestimmen zu können (siehe Bild 4.29).

In Bild 4.30 ist der Verlauf von V_{out} in den letzten drei Pumpzyklen zu sehen, wie ihn das Oszilloskop am Ausgang des Testchips aufzeichnete. Die Messung wurde wieder an einer mit Plattenkondensatoren der Kapazitätsdifferenz von 111,5 Attofarad bestückten Zelle der 2-fach Variante durchgeführt. Die Spannungsdifferenz D zwischen der grauen und der schwarzen Kurve ermittelt sich unter Berücksichtigung der Verstärkung zu ca. 6,8 Millivolt bei 186 Pumpzyklen.

Diese Werte sind höher als in der Simulation, da die Kapazität des internen Knotens in der Realität anscheinend recht klein ist: Die Rechnung ergibt bei $C_{\text{int}} = 15{,}4$ fF die experimentell verwendete, optimale Anzahl von Pumpzyklen von 186 bei einer Spannungsdifferenz von 6,3 Millivolt.

Alle anderen charakteristischen Eigenschaften der Messkurven in Bild 4.30, beispielsweise die Spannungssprünge aufgrund der Ladungsinjektion bzw. -umverteilung im Umschaltpunkt der Transistoren („Qinj“ und „clear“), entsprechen den Erwartungen. Der dynamische Bereich von V_{out} liegt im Bereich von ca. 1,2 Volt bis knapp unter die Versorgungsspannung, da der maximale Spannungshub des Verstärkerausgangs konstruktionsbedingt eingeschränkt ist (kein „rail to rail“-Verstärker). Dadurch stauchen sich die Messkurven am unteren Limit.

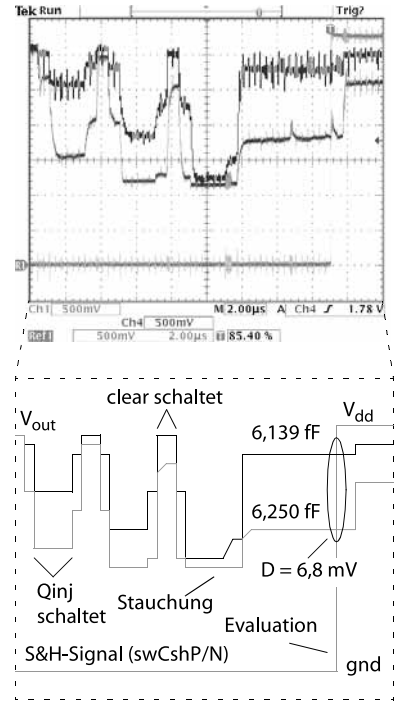


Bild 4.30. Oszilloskopbild des Spannungsverlaufs von V_{out} bei zwei Plattenkondensatoren mit unterschiedlicher Kapazität. Das Signal wurde intern verstärkt, so dass die Spannungsdifferenz am Komparator mit ca. 6,8 mV geringer ist als angezeigt (4,5 mV / Skalenteil).

Schwellenwertdispersion

Da die Komparatoren ohne schaltungstechnische Offsetkompensation realisiert wurden (siehe Bild 3.41 auf Seite 93), konnte die Schaltschwelle jedes einzelnen Komparators experimentell „ertastet“ werden, indem aufgezeichnet wurde, bei welcher Pumpzyklendifferenz das Ergebnis am Ausgang umkippte bzw. anfing, zu schwanken. Verglichen wurde dabei immer ein Plattenkondensator (6,25 fF) mit sich selbst. In Bild 4.31 sind die Werte eines Komparators zu sehen, dessen Offset sehr gering ist, da sich erst bei einer Zyklendifferenz von -1 am Ausgang sehr vereinzelte Ergebnisschwankungen ergaben. Bei einer Differenz von Null lieferte der Komparator in ca. 52 Prozent der Fälle das Ergebnis „1“, in 48 Prozent der Fälle „0“.

34. Die Verstärkung des Source-Folgers in Bild 3.36 ist nicht exakt $A_v = 1$ und die Messkapazität C aufgrund der Zellränder größer, als beim Kurvenfit angenommen. Dadurch ergibt sich eine Unsicherheit von mehreren Pumpzyklen, die sich als Offset bemerkbar macht.

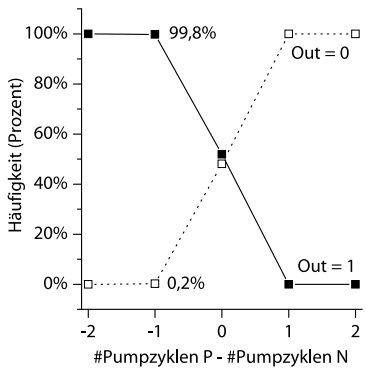


Bild 4.31. Häufigkeit des Auftretens einer „1“ bzw. „0“ in Abhängigkeit von der Differenz der Anzahl Pulse. Im vorliegenden Fall (Die 1, Matrix 1, 2er-Variante) befindet sich die Schaltschwelle des Komparators sehr nahe bei „0“, was auf einen geringen Offset hindeutet.

Wie im Abschnitt „Der Komparator“ auf Seite 92 vorgeschlagen wurde besteht die Möglichkeit, den Offset durch eine Anpassung der Zahl der Pumpzyklen zu kompensieren. Die eine der beiden zu vergleichenden Kapazitäten wird dabei ein oder zwei Zyklen häufiger mit Ladungsträgern des Messkondensators C_{L} gefüllt, als der andere. Geht man von identischen Kapazitäten aus, so wird durch die Anpassung trotz gleicher Kapazitätsverhältnisse eine Spannungsdifferenz am Komparatoreingang erzeugt, die (im Idealfall) dem Offset entspricht. Damit ist die Schaltschwelle erreicht und alle weiteren Vergleiche zwischen differierenden Kapazitäten führen zum Umkippen des Komparators nach „0“ oder „1“.

Durch die Messungen an fünf Schlüssel-Testchips zeigte sich, dass diese Art der Kompensation zu grobstufig bzw. ungenau funktioniert, um die sonst übliche schaltungstechnische Kompensation zu ersetzen. Bei der optimalen Zahl Pumpzyklen einer Kapazität mit 6,25 Femtofarad beträgt die Spannungsdifferenz durch einen zusätzlichen Zyklus im zweiten Durchlauf 3,85 Millivolt. Gangunterschiede unterhalb dieses Wertes können mit der Kompensationsmethode also nicht ausgeglichen werden.

Dieser Umstand stand freilich schon vor der Implementierung des Testchips fest. Trotzdem wurde auf diese Art der Kompensation zurückgegriffen, einerseits, um eine größtmögliche Flexibilität beim Experimentieren sicherzustellen und andererseits, da der Umfang der Offsetwerte höher eingeschätzt wurde, als dieser auf den Testchips ausfiel. Entweder lag er unterhalb des kompensierbaren Spannungsbereichs oder leicht darüber, also bei einer Pumpzyklendifferenz von 0, 1 oder 2.

Messauflösung

Um die Auflösung bzw. die minimale Kapazitätsdifferenz zu bestimmen, die von der Auswerteelektronik noch zuverlässig erkannt wird, wurden die Zellen mit Plattenkondensatoren verschiedener Größe bzw. Kapazität bestückt. Die Differenz wurde dabei in einem Bereich von 1,67 Femtofarad bis hinab zu 111,5 Attofarad variiert, kleinere Differenzen wurden nicht berücksichtigt.[35]

Auf allen fünf getesten Chips konnte diese Kapazitätsdifferenz korrekt erkannt werden (2-fach Layoutvariante). Hierfür wurden jeweils 396 Messwiederholungen pro Zelle durchgeführt, wobei jede Einzelmessung aus der zehnfachen Abtastung des digitalen Komparatorausgangs („Out“ in Bild 3.36) bestand. Somit lieferte jede Zelle 3960-mal das korrekte Ergebnis, nämlich „1“, falls die Kapazität des größeren Plattenkondensators mit der des kleineren vergleichen wurde, bzw. „0“ im umgekehrten Fall.

Durch das Vertauschen der Reihenfolge der Plattenkondensatoren bei der Differenzbildung wurde sichergestellt, dass es sich tatsächlich um die Ergebnisse des Kapazitätsvergleichs handelte, und nicht um einen sekundären, systematischen Effekt, z.B. den Offset der Komparatoren. Als zusätzliche Konsistenzprüfung wurde das Kondensatorpaar in jeder Matrix doppelt vorgesehen.

35. Es wurde angenommen, dass die Auflösung schlechter sein würde. Diese Schätzung erwies sich als zu konservativ.

Auf eine Kompensation des Komparatoroffsets wurde vollständig verzichtet, da dieser nur in vier Fällen größer war, als die kleinste zu messende Kapazitätsdifferenz (Matrixinstanzen 2 und 3 von Chip Nr. 3). In weiteren acht Fällen war diese Differenz klein genug, um die Schaltschwelle marginal zu überschreiten, so dass maximal ein Viertel der Ergebnisse eines Paarvergleichs falsch waren, d.h. zu instabilen Bits führte (990 Werte falsch, 2970 richtig, Chip Nr. 3 bei einer Kapazitätsdifferenz von 111,5 aF). Alle anderen Werte, insgesamt ca. 1,7 Millionen Messergebnisse aus 438 von 450 Paarvergleichen, waren immer korrekt, die minimale Kapazitätsdifferenz von 111,5 Attofarad wurde bei 19 von 30 Paaren immer korrekt gemessen.

Die Cluster

Hierfür wurden immer Paare von Clustern gebildet, die vom Entwurf her identisch sind und im Falle der beiden Matrixvarianten mit vier Kapazitäten (siehe Bild 3.46 bzw. Abschnitt „Der Testchip“ auf Seite 62) an der Horizontalen gespiegelt wurden. Im Falle der 2-fach Variante (Bild 3.48) fand keine Spiegelung statt.

Schwankende Bits. Zu erwarten ist, dass das Kapazitätsverhältnis der jeweils zu vergleichenden Cluster um den Wert Eins schwankt, idealerweise sogar gleichverteilt. Ein solches Verhalten wurde von der Theorie prognostiziert (siehe Abschnitt „Prozessstreuung und Mismatch“ auf Seite 22 ff.). Die Wahrscheinlichkeit für das Auftreten von Clusterpaaren mit exakt gleicher Kapazität ist demanch sehr gering, das Auftreten schwankender Bits entsprechend selten.

Tatsächlich trat dieser Fall nur bei insgesamt 32 von 2178 Paarvergleichen auf (1,47%), allein 28 davon in der 2-fach Layoutvariante. Der Grund hierfür liegt darin, dass die Cluster in dieser Variante mit höherer Wahrscheinlichkeit dicht beeinander liegende Kapazitätswerte d.h. ein besseres Matching aufweisen, als wenn sie gespiegelt werden. In Abschnitt 4.2.2 wurde dieser Sachverhalt bereits anhand der Kennwerte Er1 und Er2 diskutiert, die Analyseergebnisse stimmen insofern überein.

Genauer gesagt bedeutet das Auftreten schwankender Bits, dass der Kapazitätsvergleich zu einer Spannungsdifferenz geführt hat, die sehr nahe an der Schaltschwelle des Komparators liegt. Die Wahrscheinlichkeit für schwankende Bits entspricht somit der Wahrscheinlichkeit den Komparatoroffset „getroffen“ zu haben. Dieser ist im Allgemeinen nicht Null, somit die Kapazität der beiden Cluster nicht identisch, sondern nur dicht beisammenliegend.

Komparatoroffset. Da sich die Kompensation des Komparatoroffsets mithilfe der Pumpzyklenanpassung als ungenau und umständlich erwies, wurde nur eine Offset-Detektion durchgeführt. Damit konnten die Fälle identifiziert werden, in denen die Spannungs- bzw. Kapazitätsdifferenz der Cluster geringer war, als der Gangunterschied des Komparators. Realisiert wurde diese Detektion durch das Vertauschen der Eingänge. Änderte sich nichts am Ergebnis, so war der Offset größer, als die Spannungsdifferenz.

	Layoutvariante			
Wert	**4-fach (klein)**	**4-fach (groß)**	**2-fach**	**Alle**
X	483	203	212	898
	50,9%	35,6%	32,1%	41,2%
/	383	321	233	937
	40,4%	56,3%	35,3%	43,0%
n	0	1	20	21
	0%	0,18%	3,03%	0,96%
p	2	1	8	11
	0,21%	0,18%	1,21%	0,51%
0	5	10	177	192
	0,53%	1,75%	26,8%	8,82%
1	75	34	10	119
	7,9%	5,96%	1,52%	5,46%

Tabelle 4.11. Zahl der Fälle, in denen ein möglicherweise störbehaftetes Ergebnis aufgetreten ist (X), der Komparatoroffset unterschritten wurde (/) oder eine instabile 0 oder 1 (n oder p) gemessen wurde. Nur in der 2-fach Variante sind rund die Hälfte der Ergebnisse brauchbar (0 oder 1, Zeile 5 und 6).

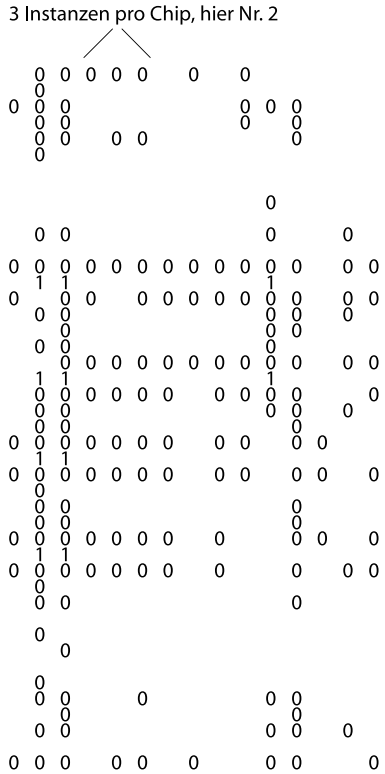


Bild 4.32. Messergebnisse der 2-fach Layoutvariante. In jeder Zeile sind die Werte eines Clusters für die jeweils 3 Instanzen der 5 Testchips aufgelistet, in den Spalten die Ergebnisse der insgesamt 64 Clusterpaare, die pro Instanz und Chip gemessen wurden.

In Tabelle 4.11 ist die Zahl dieser Fälle je Layoutvariante in der zweiten Zeile („/"-Symbol) aufgelistet. Die Statistik spiegelt die Wahrscheinlichkeit für Spannungsdifferenzen unterhalb des Offsets wieder *und* die Verteilung des Offsets über die insgesamt 11 Komparatoren je Chip (5 Matrix- bzw. Komparatorinstanzen in Spalte 2, je 3 Instanzen in den Spalten 3 und 4). Aufgrund der geringen Zahl an Komparatoren pro Layoutvariante lässt sich aus den Prozentzahlen in Tabelle 4.11 jedoch keine weitere Aussage treffen.

Zu beachten ist, dass die Zahlen in der Tabelle aus der jeweils 3960-fachen Messwiederholung bzw. Abtastung je Clusterpaar gebildet wurden. Die Zahl Eins in Zeile vier („p") der Spalte drei bedeutet also beispielsweise, dass von diesen 3960 Messergebnissen 20-mal eine Null gemessen wurde und 3940-mal eine Eins.

STÖRBEHAFTETE (UNSICHERE) BITS. Die Zahlen in der ersten Zeile der Tabelle („X"-Symbol) repräsentieren die Fälle, in denen während des Auslesens neben dem jeweils betrachteten (gemessenen) Komparator einige der restlichen Komparatoren auf dem Testchip ebenfalls umschalteten. Wie bereits erläutert, wurde während des Entwurfs keine Vorkehrung getroffen, die es erlaubt hätte, die anderen Komparatoren während des Messvorgangs zu deaktivieren. Dieser Umstand macht sich an dieser Stelle als sehr misslich bemerkbar, da in einer großen Zahl der Fälle Umschaltvorgänge zu verzeichnen waren.

Durch das Umschalten eines gewissen Teils der insgesamt elf Komparatoren und CMOS-Ausgangszellen wurden Störsignale erzeugt, die in einigen Fällen – wie aus dem Oszilloskopbild hervorging – Auswirkungen auf das Messergebnis hatten. Auf diese Weise konnten die Fälle mit Störungen beim automatischen Messdurchlauf jedoch nicht identifiziert werden, schon allein aufgrund des zusätzlichen Zeitaufwands, der bei Millionen von Abtastvorgängen stark zu Buche geschlagen hätte.

Stattdessen wurde eine pragmatische Lösung gewählt: *Alle* Fälle, in denen neben dem jeweils auszuwertenden Komparator auch andere Komparatoren umschalteten, wurden mit einem „X"-Symbol gekennzeichnet und als „unsicher" verworfen. Dieser rigorose Ansatz führte dazu, dass sehr viel mehr Ergebnisse als eigentlich nötig aussortiert wurden, wie aus den Zahlen in der ersten Zeile der Tabelle 4.11 ersichtlich wird.

STABILE BITS. Die Zahl der stabilen Bits in Tabelle 4.11 war auf diese Weise nur in der Layoutvariante mit zwei Pumpzweigen hoch genug, um eine sinnvolle Analyse der 0/1-Verteilung durchführen zu können. Die beiden anderen Layoutvarianten lieferten zu viele Ergebnisse, bei denen nicht sicher war, ob nun die gewollten Kapazitätsschwankungen der jeweils betrachteten Cluster gemessen wurden oder die Systematik der – freilich auch auf die Kapazitätsverteilung der Cluster zurückzuführenden – Umschaltvorgänge der Komparatoren.

In Bild 4.32 ist die Verteilung der stabilen und sicheren Bits in der 2-fach Layoutvariante der jeweils drei Matrixinstanzen aller fünf Testchips zu sehen, die anderen Ergebnisse wurden entfernt. In den Spalten sind die Bits von 44 verschiedenen Clusterpaaren (pro Paar identische Cluster) aufgelistet, die Reihenfolge entspricht der Anordnung der Zellen im Layout, zu sehen auf Seite 153, Abbildung (c): In jeder Matrixspalte befinden sich neuen Clusterpaare (jeweils übereinanderliegend), jede zweite Spalte ist an der Vertikalen gespiegelt. Die Reihenfolge der Bits verläuft im ständigen Wechsel 18-mal

von unten nach oben, dann umgekehrt. Immer wird jedoch die Kapazität des oberen Clusters einer Zelle mit der des unteren verglichen. Ist das Bit Null, so ist der obere Cluster größer als der untere, sonst kleiner.

Zunächst erkennt man in der 0/1-Verteilung in Bild 4.32, dass eine stark ausgeprägte Systematik vorherrscht, das Ziel einer gleichmäßigen, rein zufälligen Verteilung in dieser Schaltungsvariante also noch nicht erreicht wurde. Erst die modifizierte Schaltung in Bild 5.2 auf Seite 137 (Abschnitt 5.2.1) verspricht eine Lösung.

Stattdessen ist ein systematischer Effekt dafür verantwortlich, dass eine Ortsabhängigkeit der Bits existiert. Dieser Effekt geht jedoch nicht direkt aus der Matrix-Anordnung der Zellen im Layout hervor, da die Verteilung der Nullen und Einsen nicht mit der Ortsverteilung der Zellen übereinstimmt. Folglich sind ungewollte geometrische Einflüsse auf die Elektronik aufgrund verschiedener Umgebungen der Zellen an den Matrixrändern dafür *nicht* verantwortlich.

Die Bitverteilung gibt also tatsächlich die Kapazitätsverteilung über die Matrix wieder. Diese ist entgegen den Erwartungen nicht rein stochastisch aufgrund lokaler Prozessschwankungen, sondern weist eine starke systematische Komponente auf. Ursache hierfür ist sehr wahrscheinlich ein Gradient bei der Dicke der Isolationsschichten, der die Kapazitätsdifferenz direkt benachbarter Clusterpaare in systematischer Weise stärker beeinflusst, als die überlagerten, rein zufälligen und lokalen Prozessschwankungen, z.B. unregelmäßige Leiterbahnränder und feinkörnige Fluktuationen der Oxydschichtdicke. Ein solcher Isolationsschichtengang wurde bereits auf dem Prober-Testchip beobachtet (siehe Abschnitt 4.2 auf Seite 115 ff.). In Bild 4.33 ist der Kapazitätsverlauf zweier Cluster über die neun Zeilen des Prober-Testchips Nr. 20 zu sehen. Man erkennt, dass die systematische Zunahme der Kapazität von Zeile 1 bis 9 den Verlauf dominiert, die zufälligen Schwankungen zwischen benachbarten Zeilen fallen im Vergleich gering aus.

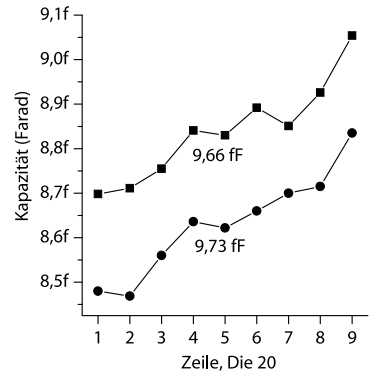


Bild 4.33. Systematische Zunahme der Kapazität zweier Cluster über die Zeilen des Prober-Testchips Nr. 20 hinweg. Ursache ist der Gradient der Isolationsschichtdicken, der den Kapazitätsverlauf dominiert (Werte neben den Kurven aus der Extraktion mit Quickcap).

4.3.3 Fazit

Zusammenfassend kann gesagt werden, dass die vorgestellte Schaltungstechnik in der Lage ist, sehr kleine Kapazitätsdifferenzen im Attofarad-Bereich zu erkennen. Dies zeigen zum einen die Messungen der Plattenkondensatoren mit einer Differenz von 111,5 Attofarad, zum anderen die Tatsache, dass die durch globale Gradienten der Oxydschichtdicken bedingte Kapazitätsdifferenz der Cluster gemessen werden kann. Diese ist, wie in Bild 4.33 beispielhaft gezeigt, im Allgemeinen noch sehr viel geringer, als die getesteten 111,5 Attofarad der Plattenkondensatoren.

Für den Einsatz als Entropiequelle in der Kryptographie eignen sich die hier vorgestellten Cluster nicht, falls das Vergleichspaar zur Erzeugung eines Bits wie in der Layoutvariante mit zwei Pumpzweigen aus identischen Clustern gebildet wird. In diesem Fall sorgen systematische Effekte für eine starke Korrelation der Bits. Falls das Paar aus spiegelsymmetrischen Clustern gebildet wird, so ist das Kapazitätsverhältnis sehr viel zufälliger, wie in Bild 4.23 auf Seite 120 anhand des Unterschieds zwischen aufrecht und gespiegelten Zeilen zu erkennen ist.

Die besten Resultate bezüglich Zufälligkeit und Kapazitätsdifferenz sind zu erwarten, wenn die Cluster strukturell so aufgebaut sind, dass die Kapazität aus weniger vertikal verlaufenden Feldlinien zusammengesetzt ist, son-

dern aus mehr lateralen, beispielsweise durch Beschränkung auf wenige Metallisierungsebenen. Auf diese Weise machen sich Gangunterschiede in der Dicke der Isolationsschicht zwischen den Ebenen weniger bemerkbar, zugunsten zufälliger, lithografischer Ungenauigkeiten, wie sie beispielsweise den Kapazitätsverlauf in Bild 4.17 auf Seite 116 bestimmen.

* * *

Kapitel 5
Zusammenfassung und Ausblick

Das Schlusskapitel dieser Arbeit wird durch die Zusammenfassung der gewonnenen Erkenntnisse und offenen Fragen, eine Diskussion der Anwendungsmöglichkeiten des vorgestellten Cluster-Konzepts und das Fazit gebildet. In Abschnitt 5.1 wird zunächst die strukturelle Unbestimmtheit der Cluster aufgrund systematischer Unschärfeeffekte und lokaler Kornstruktur-Fluktuationen zusammengefasst. Wie erläutert wird, leitet sich daraus eine kapazitive Unbestimmtheit ab, die durch Kompromisse bei der Berechnung aufgrund der numerischen Approximation der analytisch nicht lösbaren Laplace-Gleichung noch vergrößert wird. Danach wird auf die messtechnischen Aspekte des experimentellen Teils der Arbeit eingegangen, beispielsweise die weitverbreitete Ladungspumpen-Technik und die Implementierung und Durchführung der Messungen. Die Ergebnisse der Messwert-Analyse werden anschließend zusammengefasst. Den Abschluss des Abschnitts bildet eine Übersicht über die noch offenen Fragen, insbesondere sind dies die Sicherheits-Frage und die Zuverlässigkeit der vorgeschlagenen Lösungen.

In Abschnitt 5.2 wird die Eignung des Cluster-Konzepts in den beiden wichtigsten Anwendungsgebieten – die Generierung von „single chip keys“ und „all-chip keys“ – auf Grundlage der Ergebnisse dieser Arbeit bewertet. Es wird diskutiert, welche Probleme für den jeweiligen Einsatzzweck noch zu lösen sind und es werden Vorschläge hierfür gemacht. Es wird gezeigt, welche Palette an Möglichkeiten besteht, die Cluster auf die Erfüllung bestimmter kapazitiver Eigenschaften hin zu optimieren und welche Vorgehensweise dafür geeignet ist.

Abschnitt 5.3 bildet das Resümee und bietet abschließend einen Ausblick auf zukünftige Entwicklungen. Als Quintessenz könnte man sagen, dass die Konzepte, Methoden und Messergebnisse, die in dieser Arbeit vorgestellt wurden, das Rüstzeug bietet zur Weiterentwicklung einer Cluster-basierten Schlüsseltechnik.

* * *

5.1 Zusammenfassung

Ausgehend von der in Kapitel 1 beschriebenen Motivation und den skizzierten Anwendungsfällen wurde eine Aufgabenstellung formuliert, die einen neuartigen Lösungsansatz erforderte. Es wurde gezeigt, dass der Stand der Technik einige wenige interessante Techniken hierfür bereithält, die in dieselbe Richtung wie in der vorliegenden Arbeit gehen, z.B. bei der MIT-Gruppe um S. Devadas: Das Ausnutzen der zufallsbedingten Schwankungen einiger physikalischer Parameter des Herstellungsprozesses von Halbleiterchips zur Erzeugung einer individuellen Bitsequenz. Diese wird beim MIT-Ansatz in erster Linie zur Identifikation einzelner Chips verwendet, der Einsatz als Entropiequelle im Sinne eines Zufallsgenerators für kryptografische Schlüssel wird nur am Rande erwähnt.

Dieses Ziel wurde im Rahmen der vorliegenden Arbeit weiterverfolgt, um einigen der Hauptkritikpunkte bei den existierenden Verfahren ein Alternativkonzept entgegensetzen zu können: Die Ableitung der Schlüsselsequenz aus der Kapazität von komplexen, dreidimensionalen Strukturen aus Verbindungsleitungen, den 3D-Clustern. Dadurch wird das Problem des Parameterdrifts konzeptionell umgangen, um nur einen der Vorteile zu nennen. Ein weiteres Einsatzgebiet eröffnet sich durch die Möglichkeit, geheime Schlüssel in Form einer festen, vordefinierten Zahlenfolge in alle Chips einer Produktionsreihe einzuprägen, wie im Abschnitt „All-Chips Key“ in diesem Kapitel erläutert wird.

Inhalt und Ergebnisse dieser Arbeit:

- Entwicklung des Konzepts des „eingeprägten Zufalls“
- Recherche zur Theorie gängiger Extraktionswerkzeuge
- Theorie der Prozessstreuung und Herleitung der Matching-Modelle
- Erklärung der Auflösungsbeschränkung klassischer Kapazitäts-Messmethoden
- Entwicklung und Implementierung des Random-Walk Algorithmus
- Durchführung von automatisierten Prober-Messungen an Testchip-Clustern
- Vorschlag einer Schlüsselelektronik zur integrierten Kapazitätsauswertung
- Vergleich der Extraktionswerte verschiedener gängiger Tools
- Vergleich der Messwerte von Clustern und Kondensatoren, inkl. Matching
- Funktionsbeweis der Schlüsselelektronik

Bild 5.1. Stichwortliste der gelösten Probleme und abgeschlossenen Arbeiten.

5.1.1 Gewonnene Erkenntnisse

Es wurde anhand von Testchips gezeigt, dass die in der Einführung ins Auge gefassten, mit einem Kabelknäuel vergleichbaren 3D-Clusterstrukturen durch einen speziell entwickelten Random-Walk Algorithmus einfach und effizient erzeugt und prozesstechnisch hergestellt werden können. Es zeigte sich im Rahmen einer Patent-Anmeldung, dass diese Gebilde in ihrer Art und Anwendung völlig neuartig sind.

Die kapazitive Unbestimmtheit

Der Vergleich der Cluster-Layouts mit Mikroskopbildern zeigte, dass eine *systematische* Unschärfe (z.B. Abrundung der Kanten) durch den lithografischen Prozess zu verzeichnen ist, so dass die Cluster den Extremfall jener Chipstrukturen darstellen, die nach der Herstellung besonders stark vom Layout abweichen. Weiterhin wurde anhand der Theorie der Prozessschwankungen und durch Herleitung des Pelgrom-Modells gezeigt, dass eine Vielzahl vereinfachender Annahmen in die Modellierung eingeht und nur einfache Plattenkondensatoren in den Matching-Modellen der Schaltkreissimulatoren enthalten sind. Aus den Ergebnissen der Berechnung von Shyu et al. in Bild 2.9 auf Seite 31 geht hervor, dass die zufälligen Schwankungen der Isolationsschichtdicke *bei Plattenkondensatoren* stärker sind, als Randeffekte. Bei den 3D-Clustern kann diese Aussage aufgrund der Beschränkungen des Modells so einfach nicht getroffen werden.

Die Kombination dieser Faktoren, also systematische Effekte, globale Parametergefälle und lokale Fluktuationen sorgen für eine Abweichung der konkreten Form jedes einzelnen Clusters vom Layout, so dass zur Entwurfs-

zeit die Endgestalt der Cluster nach der Herstellung unbestimmt ist. Daraus ergibt sich sofort eine kapazitive Unbestimmtheit, da gezeigt wurde, dass die elektrische Kapazität unmittelbar von der genauen dreidimensionalen Gestalt eines Leiters abhängt.

Darüber hinaus wurden die verschiedenen Berechnungs- bzw. Extraktionsverfahren vorgestellt und verglichen. Es zeigte sich, dass eine numerische Lösung der Laplace-Gleichung rechenintensiv ist und die Kapazität selbst von den genauesten Programmen nur approximiert wird: Bei einem der 299 Cluster, die bei diesem Vergleich untersucht wurden, betrug die Kapazitätsdifferenz 15,3 Prozent. Der Unterschied zwischen dem wohl exaktesten Tool „Quickcap“ und den auf Schnelligkeit optimierten Extraktoren Assura, Calibre und Diva betrug in den extremsten Fällen gar 25 Prozent, 40 Prozent und 93 Prozent (siehe Tabelle 4.3 auf Seite 110). Damit ergibt sich zusätzlich eine kapazitive Unbestimmtheit, die ihre Ursache in der Schwierigkeit hat, die genaue Verteilung von elektrischen Ladungen auf komplizierten Leiterbahnstrukturen zu berechen.

Unter der Annahme, dass auch die hergestellten Cluster ein solches Verhalten aufweisen, ist es möglich, schon beim automatisierten Layoutentwurf solche Gebilde algorithmisch zu erzeugen oder auszuwählen, die eine besonders hohe kapazitive Abweichung zwischen den Extraktionswerkzeugen aufweisen, um damit die Unbestimmtheit weiter zu erhöhen. Für einen Außenstehenden, der nur die Form der Cluster nach der Herstellung (z.B. durch Mikroskopbilder) kennen kann, ist die Unbestimmtheit damit besonders groß.

Weiterhin zeigte sich im Zusammenhang mit den Ergebnissen in Bild 4.22 auf Seite 119, dass sich die Kapazitätsverhältnisse zweier Cluster umdrehen können, wenn die Dicke der Isolationsschichten z.B. durch Übergang von den typischen Prozessparametern zur „worst-case corner“ variiert wird. Der Einfluss dieser Änderung auf die Kapazität solcher spezieller Clusterpaare ist damit nicht-linear und sorgt dafür, dass in diesen Fällen die kapazitive Größensortierung zur Entwurfszeit unbestimmt ist. Auch hier kann das automatisierte Cluster-Generierungsverfahren dafür optimiert werden, solche speziellen Strukturen bevorzugt zu erzeugen.

Schließlich geht aus den gemessenen Matching-Werten der Cluster in Bild 4.25 bzw. Tabelle 4.10 auf Seite 121 und den Werten von Plattenkondensatoren in Tabelle 4.8 auf Seite 117 hervor, dass der auf die Kapazität bezogene relative Matching-Fehler mit der Größe abnimmt und ähnlich gute Werte erreicht, wie Plattenkondensatoren mit einer Isolationsschicht. Liegen die Kondensatorplatten weiter auseinander mit zusammengesetztem Oxyd oder handelt es sich um parallel verlaufende Metallbahnen, so ist der Matching-Fehler deutlich höher, als bei den Clustern. Dies legt die Vermutung nahe, dass Cluster hoher struktureller Dichte und mit ausgeglichenem Aufbau (Schwerpunkt in der Mitte) Prozessschwankungen kapazitiv besser ausgleichen können, als einfache Gebilde geringer Dichte. Somit wurde gezeigt, dass der geometrische Aufbau der Cluster im Generierungsverfahren noch optimiert werden kann, wenn beispielsweise starke kapazitive Streuungseigenschaften als Entropiequelle für die Bits eines Schlüsselgenerators gewünscht werden.

Auf der anderen Seite können solche Strukturen favorisiert werden, die kapazitiv sehr nahe beieinander liegen und dennoch wenig streuen. Unterscheided sich ihr Layout in nicht-trivialer Weise, so kann ein potentieller Angreifer aufgrund des Berechnungsproblems nur mit Schwierigkeiten das Grö-

ßenverhältnis der beiden aus dem Mikroskopbild des Chips ableiten. Damit können vordefinierte Bitsequenzen in ganze Produktionsreihen eingeprägt werden, die stabil bleiben und dennoch schwer angreifbar sind.

Messtechnische Verfahren und Ergebnisse

Im Kapitel „Theoretische Grundlagen" wurde erklärt, wie hochgenaue Kapazitätsmessungen schaltungstechnisch bewerkstelligt werden und welche Auflösungsgrenzen existieren. Es wurde gezeigt, dass der Einfluss von Ladungsträgern aus dem Kanal der Schalttransistoren und der Ladungsinjektion durch die Steuerelektrode limitierend wirken und die Ursache mathematisch hergeleitet. Weiter wurden schaltungstechnische Verbesserungen vorgestellt und auf den Testchips implementiert. Abgerundet wurde das Thema durch eine Recherche zu alternativen Messtechniken und der derzeit exaktesten Methode nach Chang et al. 2004 (siehe Bild 2.35, S. 57).

Die Durchführung der Messungen auf dem Spitzenmessplatz wurde automatisiert, wobei ein neuer Ansatz getestet wurde, der es ermöglichte, mit nur einer Kontaktnadel pro Messung auszukommen und dadurch die Anforderungen an die Positionierungsgenauigkeit wesentlich reduzierte: Alle Testchips wurden auf einer Leiterplatte aufgebracht (inkl. Gehäuse), die wiederum auf dem Probenteller Platz fand. Die durch diese Vorgehensweise sich nachteilig auswirkenden Probleme (Schiefe der Dies im Gehäuse, Spannungsabfall, etc.) wurden gelöst. Wege, diese bei zukünftigen Messreihen zu vermeiden, wurden dadurch aufgezeigt.

Ebenso wurden Lösungen für den in den Testchips auftretenden Spannungsabfall auf den Versorgungsleitungen erarbeitet. Es wurde die Ursache ergründet und eine Kompensationsmethode aufgezeigt, um den Störeffekt zu minimieren. Beim Entwurf zukünftiger Testchips können diese Hinweise von praktischem Wert sein und Probleme vermeiden helfen. Nach der Korrektur und Optimierung des Verfahrens wurden insgesamt 2720 Messwerte analysiert und mit den berechneten Werten des als am genauesten geltenden Extraktionswerkzeugs „Quickcap" verglichen.

Durch eine Untersuchungen des Kapazitäts*verlaufs* der getesteten Strukturen wurde nachgewiesen, dass die Dicke der isolierenden Oxydschichten aller fünf Testchips ein starkes Gefälle (globaler Gradient) aufweist, das zu dem Anstieg der Kapazität über die Zeilen hinweg führt, wie in mehreren Messkurven zu sehen (z.B. in Bild 4.24 auf Seite 120). Dieser Gradient existiert dabei Wafer-weit und ist so stark, dass nur ein prozesstechnisches Problem oder eine falsche Geräteeinstellung beim CMP-Verfahren („chemical-mechanical polishing") dafür verantwortlich sein kann. Gestützt wird diese These durch die Tatsache, dass die aus der Kapazität errechnete Oxydschichtdicke bei allen Testchips größer ist, als die Prozessparameter-Spezifikation des Herstellers zulässt.

Auf der anderen Seite konnte gezeigt werden, dass die Messwerte eine sehr gute Wiederholbarkeit bei Folgemessungen an denselben Strukturen aufweist. Der durch Rauschen bedingte Messfehler war im Falle des Substrat-Metal1 Plattenkondensators in Spalte 4 so gering, dass die maximale Kapazitätsdifferenz bei 30 Wiederholungen nur 11,5 Attofarad betrug, die Standardabweichung sogar nur 3,3 Attofarad (siehe Tabelle 3.4 auf Seite 83). Neben dem Auftreten eines einzelnen extremen Ausreißers beim Poly1-Poly2 Plattenkondensator wurden die restlichen „normalen" Ausreißer damit erklärt,

dass sich die Kontaktgüte im Laufe der Messungen z.B. durch Materialdehnungen verschlechterte. Selbst mit diesen Werten lag die maximale Kapazitätsabweichung der jeweiligen Messreihe bei unter 0,5 Prozent, die Standardabweichung unter 0,2 Prozent.

Schließlich wurde im Rahmen dieser Arbeit eine Messmethode zur integrierten Kapazitätsmessung bzw. zum Kapazitätsvergleich von Clustern entwickelt und in Abschnitt 3.3 vorgestellt. Es wurde eine Formel zur Berechnung der Spannungsdifferenz hergeleitet, die von der Anzahl Pumpzyklen und der Kapazitätsdifferenz des zu vergleichenden Clusterpaares abhängt. Die Rolle dieses Spannungsunterschieds im Zusammenhang mit dem Komparator zur Erzeugung eines Bits wurde erläutert.

Der Einsatz dieser Schaltung als Schlüsselelektronik wurde schließlich in Form eines weiteren Testchips erprobt und in Abschnitt 4.3 analysiert. Es konnte gezeigt werden, dass die Schaltung Kapazitätsdifferenzen bis mindestens 111,5 Attofarad zuverlässig erkennt, wobei der Test an kleineren Differenzen noch aussteht. Damit wurde bewiesen, dass der Schaltungsvorschlag prinzipiell dazu geeignet ist, kleinste Kapazitätsdifferenzen von Clusterpaaren in ein Bitmuster zu überführen.

Aus den Ergebnissen der Vergleiche von identischen Clustern ging hervor, dass die Generierung einer von Chip zu Chip verschiedenen, zufälligen Bitsequenz mit hoher Entropie durch einen reinen größer-/kleiner-Vergleich von Clusterpaaren nicht möglich war, da ein Gefälle der Isolationsschichtdicken einen wesentlich höheren, systematischen Einfluss auf die Kapazitätsverhältnisse der Paare hatte, als die zufälligen, lokalen Prozessschwankungen, die als Entropiequelle hätten dienen sollen. Da es sich bei diesem Oxydschichtgradienten um das gleiche Phänomen handelt, wie bei den Prober-Testchips, und beide vom selben Prozesslauf (und vermutlich Wafer) stammen, liegt die Vermutung nahe, dass auch hier ein Problem beim CMP-Verfahren Ursache ist.

Um dennoch das Cluster-Konzept als Entropiequelle nutzen zu können, bietet sich eine andere Form der Kapazitätsauswertung an: Statt der Umwandlung des Kapazitätsverhältnisses von Clusterpaaren in ein einzelnes Bit kann die vorgeschlagene Schlüsselelektronik so modifiziert werden, dass sie den *absoluten* Kapazitätswert *eines* Clusters ermittelt und über einen Analog-Digital-Wandel in mehrere Schlüsselbits umwandelt. Dieser Ansatz wird im Abschnitt 5.2.1, Bild 5.2 vorgestellt.

5.1.2 Offene Fragen

Da es sich bei den Clustern um gänzlich neuartige Strukturen handelt, die bisher in keiner Arbeit hinsichtlich ihrer kapazitiven Eigenschaften untersucht wurden und ihr Einsatz als Entropiequelle im Sinne des eingeprägten Zufalls ein zukunftsweisendes, aber dennoch wagnisreiches Neukonzept darstellt, sind viele Fragen noch offen oder konnten nur am Rande gestreift werden. Die Auswahl bestimmter Fragestellungen hängt teilweise damit zusammen, dass es sich beim Cluster-Konzept in weiten Teilen um gänzlich unbekanntes Terrain handelt, das sich nur in einer „trial and error“ Manier erkunden lies. Die wichtigsten der ausgelassenen Fragestellungen sollen im Folgenden erörtert werden.

Sicherheit

Die wichtigste Eigenschaft eines Schlüsselgenerators stellt die Sicherheit vor Angreifern dar. Der Schlüssel soll weder mit Methoden wie der differenziellen Stromverbrauchsanalyse, noch über die elektromagnetische Abstrahlung von außen ermittelbar oder mit invasiven Mitteln direkt messbar sein. Hierzu sind in der Praxis immer umfangreiche Tests nötig, die im Rahmen dieser Arbeit nicht durchgeführt werden konnten. Anders als bei den kryptografischen Algorithmen lässt sich die Sicherheit eines technischen Verfahrens grundsätzlich nicht mit theoretischen Mitteln beweisen, sondern kann nur indirekt getestet werden.

Neben direkten Messungen oder dem Abhorchen besteht eine Angriffsmethode darin, mit statistischen Analysen Aussagen über die Wahrscheinlichkeitsverteilung der Schlüsselbits zu treffen. Dadurch reduziert sich die Entropie des Schlüssels bzw. die Anzahl der unbekannten Bits und ein erfolgreicher „brute-force" Angriff wird wahrscheinlicher. Die resultierende Entropie unter Einbeziehung der statistischen Eigenschaften der Schlüsselbits wurde in dieser Arbeit nicht bestimmt, da die Bits durch die differentielle Kapazitäts-Messmethode ohnehin nur die globalen Gradienten wiedergaben. Erst die im folgenden Abschnitt vorgeschlagene Modifikation verspricht ein sinnvolles Maß an Entropie wiederzugeben.

Im günstigsten Fall erfüllt ein Zufallsgenerator die sogenannt „DIEHARD"-Spezifikation von George Marsaglia, die Kriterien für die Güte der statistischen Eigenschaften festlegt. Weitere Test sind auf Grundlage von Knuth's Test (siehe Knuth 1969) möglich, sowie im NIST-Standard FIPS 140-1 zu finden. Eine Überprüfung des Cluster-Konzepts nach diesen Maßstäben blieb in der vorliegenden Arbeit aus den genannten Gründen aus. Es sei jedoch darauf hingewiesen, dass diese harten Kriterien nur bei solchen Zufallsgeneratoren nötig sind, die es einem Angreifer erlauben, unbeschränkt viele Zufallssequenzen zu erzeugen, auf diese direkt zuzugreifen und auf statistische Auffälligkeiten hin zu untersuchen. Bei in Hardware gegossenen Schlüsseln ist dies nur sehr eingeschränkt möglich.

Schließlich lässt sich argumentieren, dass nicht in allen Anwendungsfällen die härtesten Sicherheitsanforderungen erfüllt werden müssen, um eine Technik einsetzbar zu machen. Für Bankensysteme gilt sicherlich ein höherer Standard, als für RFID-Chips in der Lagerhaltung.

Zuverlässigkeit

Die Abhängigkeit von Betriebsparametern (Spannung und Temperatur) konnte in dieser Arbeit nicht untersucht werden. Ebenso wurde nicht experimentell überprüft, welchen Einfluss das Aging-Phänomen auf die Schlüsselgenerierung hat. Im Gegensatz zum Vorschlag von Lofstrom et al. 2000 (Abschnitt 1.2.2) ist jedoch die Entropiequelle selbst nicht von diesen Einflüssen betroffen, da die Kapazität der Cluster nur von der Geometrie und den Dielektrizitätskonstanten abhängt, die wiederum fest sind.

Der Fall von instabilen Bits durch Clusterpaare mit zufälligerweise identischer Kapazität wird durch das im Folgenden vorgeschlagene absolute Messprinzip vollständig vermieden, solange das Messverfahren nicht die Rauschgrenze unterschreitet.

* * *

5.2 Anwendungsmöglichkeiten

Einige der ins Auge gefassten Anwendungsfelder der 3D-Cluster wurden bereits anfangs im Abschnitt „Motivation“ genannt. Neben dem Einsatz als Entropiequelle zur Schlüsselgenerierung in der Kryptografie sind jedoch auch alternative Anwendungen denkbar, wie im Abschnitt 5.2.2 gezeigt wird.

5.2.1 Schlüsselgenerierung

Single-Chip Keys

Unter dieser Bezeichnung soll das wohl wichtigste Konzept in dieser Arbeit verstanden werden: Der in Hardware „eingeprägte Zufall“ als Entropiequelle zur Erzeugung von kryptografischen Schlüsseln, die sich von Chip zu Chip unterscheiden. Um die Cluster zu diesem Zweck einsetzen zu können, ist eine Modifikation der Schlüsselelektronik nötig: Die Auswertung der Cluster sollte nicht über einen größer/kleiner-Vergleich von Paaren stattfinden, sondern über die Messung des Absolutswerts eines *einzelnen* Clusters.

In Bild 5.2 ist zu sehen, wie dies bewerkstelligt werden kann. Die ursprünglich aus zwei (oder mehreren) Clustern bestehende Zelle wird vereinfacht, indem nur noch ein auszuwertender Cluster vorgesehen wird. Dadurch können die vormals über „swCx“ gesteuerten PMOS-Schalter entfallen. Die Spannung V_{out} wird über einen analogen Verstärker (mit möglichst geringem Rauschen) in einem Spannungsbereich, der extern über V_{offset} vorgegeben wird, aufgespreizt und an einen Analog-Digital-Wandler (ADC) weitergegeben. Der Ausgang des ADC stellt bereits den digitalen Schlüssel dar, falls keine instabilen Bits durch Unterschreiten der Rauschgrenze auftreten. Andernfalls kann eine nachgeschaltete Dekodierungsstufe die Daten z.B. als Hamming-Code auffassen und so das Auftreten eines einzelnen Bitfehlers korrigieren.

Da die Bitsequenz am Ausgang keine Aussage über die nach physikalischer Definition korrekte Kapazität in Form einer exakten Zahl treffen muss, sind die Anforderungen an die Linearität und den Offset des Verstärkers und des ADC minimal. Auch muss der ADC das Ergebnis der Umwandlung nicht als binär kodierte Dezimalzahl (BCD) ausgeben. Er kann (und sollte sogar) intern über einen einfachen LFSR-Zähler (siehe Bild 1.5 auf Seite 10) implementiert sein. Die Dekodierung des Hamming-Codes (die Erzeugung entfällt) ist schließlich sehr effizient möglich und geschieht über einfache kombinatorische Logik.

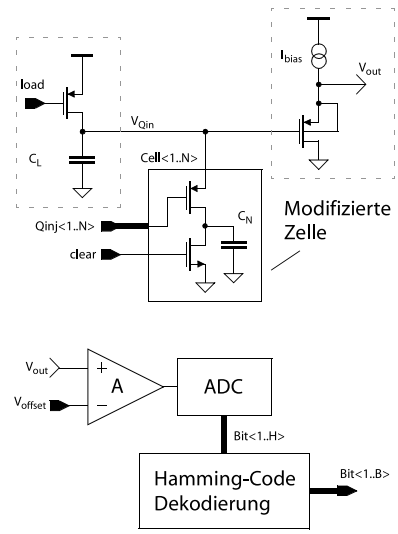


Bild 5.2. Modifizierte Auswerteelektronik zur Erzeugung einer mehrstelligen Binärzahl aus dem Absolutwert der Kapazität eines Clusters.

All-Chips Key

Im Gegensatz zu den individuellen Schlüsseln sollen im zweiten Anwendungsfall keine zufälligen Bitsequenzen erzeugt werden, sondern vordefinierte, auf allen Chips einer Produktionsreihe identische Schlüssel aus den Kapazitätsverhältnissen der Cluster ausgelesen werden. Dazu werden zur Entwurfszeit Paare von Clustern verschiedener Kapazität in einer Zelle zusammengefasst, für jedes Bit ein anderes Paar. Je nach dem kapazitiven Größenverhältnis ist das Ergebnis Null oder Eins. Die in Bild 1.5 auf Seite 10 vorgestellte Schlüsselelektronik kann zu diesem Zweck direkt verwendet werden. Um eine hohe Sicherheit vor Angreifern zu gewährleisten, sollte die Kapazitätsdifferenz der beiden Cluster so gering wie möglich ausfallen. Auf

der anderen Seite muss ein gewisser kapazitiver Mindestabstand eingehalten werden, um zu vermeiden, dass sich die Kapazitäten durch Prozessschwankungen bei einigen Chips der Produktionsreihe überschneiden und so zu instabilen oder fehlerhaften Bits führen.

Um dies zu vermeiden, können die Cluster im Idealfall in der Form von Testchips gefertigt und durch die Ladungspumpen-Technik vermessen werden. Sowohl der Entwurf der Cluster, als auch das Messverfahren kann durch Anwendung und Weiterentwicklung der in dieser Arbeit beschriebenen Vorgehensweisen automatisiert werden, so dass eine hohe Zahl realisiert werden kann. Durch eine Analyse der Schwankungseigenschaften können dann solche Clusterpaare ausgewählt werden, die kapazitiv dicht beieinander liegen aber dennoch durch die Schlüsselelektronik auf allen Chips eindeutig getrennt werden können.

Grundsätzlich ist diese Vorgehensweise auch durch eine reine Extraktionsanalyse möglich, erfordert aber den Einsatz von EDA-Software der TCAD-Klasse wie Quickcap oder Assura-FS (siehe Tabelle 2.3 auf Seite 43). Wird zusätzlich eines der neuen OPC-Modellierungsprogramme verwendet, so reduziert sich der nötige kapazitive Sicherheitsabstand, da die strukturelle Unbestimmtheit der Cluster durch Berechnung der Unschärfe und von Beugungseffekten verringert wird. Trotzdem ist eine größere Kapazitätsdifferenz einzuhalten, um eine Überschneidung aufgrund der Prozessstreuungen zu vermeiden.

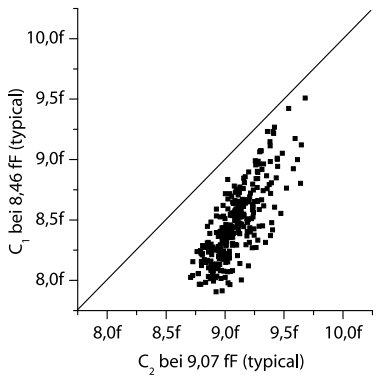


Bild 5.3. Kapazitätsverteilung zweier Cluster C_1 und C_2 bei Variation der Oxydschichtdicken. Jeder Punkt gibt die Extraktionswerte der beiden Cluster bei gleicher Dicke an. Der Korrelationskoeffizient beträgt 0,78.

Geht man davon aus, dass – ähnlich wie im vorangehenden Kapitel beschrieben – die globalen Gradienten der Isolationsschichtdicke stärkeren Einfluss auf die Kapazität von dicht platzierten Clustern haben, als die lokalen, nicht-deterministischen Randeffekte und Fluktuationen der Kornstruktur, so kann die Simulation der globalen Prozessschwankungen mit TCAD-Extraktoren wertvolle Hinweise auf die Streuungseigenschaften von Clusterpaaren liefern. Der Grund hierfür liegt darin, dass die Isolationsschichtdicken in die Berechnung eingehen und folglich variiert werden können, während die lokalen Effekte unberücksichtigt bleiben (siehe Bild 2.12 auf Seite 38). Wählt man also für die Dicken der Dielektrika einige Werte, so kann man jeweils den Einfluss auf die Kapazität der beiden Cluster berechnen. In Bild 5.3 ist das Ergebnis einer solchen Vorgehensweise zu sehen. Die Punkte geben jeweils die Kapazität der beiden (verschiedenen) Cluster bei einer bestimmten Kombination an Oxyddicken an (pro Metallisierungsebene). Die Wahl erfolgte zufällig und normalverteilt um den jeweils typischen Wert (Berechnung aus Quickcap bei 0,2%).

Wie zu sehen ist, besteht zwischen den Kapazitätswerten eine gewisse Korrelation, die sich durch die langestrecke Form der Punktwolke äußert. Der Korrelationskoeffizient berechnete sich zu 0,78. Keine der Punkte liegt über oder auf der Winkelhalbierenden, so dass die simulierten Schwankungen zu keinen Bitfehlern führen können (idealer Komparator vorausgesetzt). Damit eignet sich dieses spezielle Clusterpaar besonders gut, um für ein Bit der Schlüsselsequenz in einer der Zellen zum Einsatz zum kommen. Es zeigte sich, dass andere Clusterkombinationen einen wesentlich geringeren Korrelationskoeffizienten aufweisen, es sich dabei also um eine spezielle, strukturabhängige Eigenschaft der Cluster handelt und explizit im Generierungsverfahren bevorzugt erzeugt werden kann.

Generell besteht die Möglichkeit, die Eigenschaften der Cluster im Generierungsverfahren so zu optimieren, dass sie bestimmte Zielvorgaben erfüllen. Ein Mittel hierfür besteht in der Simulation der Einflüsse von Änderungen der Prozessparameter (soweit modelliert) durch TCAD-Extraktoren. Die kapazitive Bewertung der Änderungen kann dann als Korrekturwert bzw. Stellgröße interpretiert und durch Rückkoppelung in das Generierungsverfahren einbezogen werden. Nach der Neu-Parametrisierung des Random-Walk Algorithmus wird die geänderte Geometrie neu extrahiert und so weiter. Der schematische Aufbau dieses iterativen Verfahrens ist in Bild 5.4 zu sehen.

Die Zielvorgabe des Verfahrens kann schließlich eine bestimmte Kapazität sein, die erreicht werden soll, ein bestimmtes Oxydschichtdicken-bezogenes Streuungsverhalten oder eine hohe (oder niedrige) Korrelation mit anderen Clustern. Damit steht ein mächtiges Werkzeug zur Verfügung, das die Wahl und Generierung geeigneter Clusterpaare zur Bildung der Bits eines geheimen All-Chips Schlüssels ermöglicht.

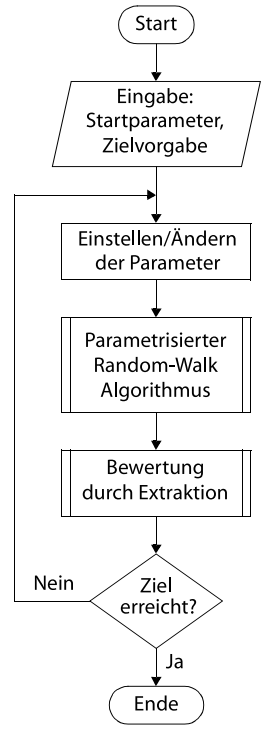


Bild 5.4. Iteratives Verfahren zur Generierung von Clustern mit bestimmten kapazitiven Eigenschaften.

5.2.2 Alternative Anwendungen

LAYOUT-OBFUSKATION. Einem völlig anderen Zweck als bei der Schlüsselerzeugung könnten die Cluster dienen, wenn nicht die elektrische Kapazität ausgenutzt wird, sondern die irreguläre, willkürlich wirkende Struktur selbst. Da es sich bei den Clustern um ineinander verwobene Verbindungsleitung handelt, können sie dazu benutzt werden, elektrische Signale zu transportieren, beispielsweise innerhalb einer Schaltung. Die Schaltungsrückerkennung durch „Reverse-Engineering“, also über das Abtragen, Fotografieren und (teilweise rechnergestützte) Wiederzusammensetzen, wird durch die wirre Leitungsführung wesentlich erschwert. Solche Verschleierungstechniken sind im Software-Bereich unter der Bezeichnung „Spaghetti-Code“ bekannt. Für diesen Einsatzweck kann der Random-Walk Algorithmus so eingestellt werden, dass er große Cluster mit weniger Ecken und schrägen Kanten erzeugt, die wahlweise in ein vorhandenes Leitungsnetz eingewebt werden.

SENSORNETZE (SHIELDING). Durch ein solches Einweben in die bereits vorhandenen Leitungen einer Schaltung können in einem weiteren Anwendungsfall kritische Schaltungsteile geschützt werden. Diese als „shielding“ bezeichnete Technik sieht vor, Änderungen der elektrischen Parameter eines Sensornetzes zu messen, um invasive Angriffsversuche, z.B. durch Microprobing auf einem Spitzenmessplatz, zu detektieren. Die Kapazität eines Clusters kann in dieser Weise genutzt werden, wenn seine Leitungen die kritischen Schaltungselemente umschließen oder überdecken.

Der Vorteil gegenüber traditionellen Sensornetzen (z.B. Gitterstrukturen auf verschiedenen Ebenen) besteht wiederum in der irregulären, verwirrenden Struktur, die es einem Angreifer erschwert, das Netz z.B. durch FIB-Eingriffe („Focused Ion Beam“) zu überbrücken oder zu deaktivieren. Weiterhin ist vorteilhaft, dass sich das Shield an die bereits vorhandenen Leitungen der zu schützenden Schaltung anpasst, während Sensornetze aus Gitterstrukturen (mindestens) eine komplett freie Metallisierungsebene voraussetzen.

* * *

5.3 Fazit und Ausblick

In dieser Arbeit wurde ein neues Verfahren zur Generierung kryptografischer Hardware-Schlüssel entwickelt, das als Entropiequelle spezielle und neuartige 3D-Mikrostrukturen nutzt, die Cluster. Diese irregulären, komplex strukturierten Ballen aus zufällig ineinander verwobenen Verbindungsleitungen weisen eine kapazitive Unbestimmtheit auf, die als Quelle der Entropie dient, entweder durch den „eingeprägten Zufall" der Schwankungen des Herstellungsprozesses, oder durch Wahl einer geheimen Schlüsselsequenz zur Entwurfszeit. Er wurde gezeigt, wie diese 3D-Cluster automatisiert und parametrisierbar entworfen und gefertigt werden können und welche kapazitiven Eigenschaften sie aufweisen.

Weiterhin wurde eine Schaltungstechnik zur Ableitung der Schlüsselsequenzen aus der Entropiequelle entwickelt, die sich am bewährten Ladungspumpen-Prinzip orientiert und ähnlich hochauflösend Kapazitäten misst, jedoch vollständig integrierbar ist. Es wurde gezeigt, dass die Kombination dieser Technik mit geeigneten Clustern kryptografische, geheime Schlüssel erzeugen kann. Für die Weiterentwicklung hin zu einer einsatzfähigen Lösung wurden aussichtsreiche Schaltungsvarianten vorgestellt.

ZUKÜNFTIGE ENTWICKLUNGEN. Eine solche Weiterentwicklung ist besonders interessant bei den zu erwartenden neuen Technologien der nächsten Jahrzehnte. Unter den Schlagworten „Plastik-Elektronik" und „druckbare Schaltkreise" werden Hardware-Lösungen verstanden, die sich durch Vorteile bei der Massenfertigung und den Kosten, sowie der Handhabung und Alltags-Integration kennzeichnen. Einschränkungen bei den Leistungsdaten und der Verfügbarkeit elektronischer Standardkomponenten, vor allem integriertem, nicht-flüchtigem Speicher (z.B. EEPROM), machen die Entwicklung spezieller Lösungen nötig. Hier können die Cluster als Quelle geheimer Schlüssel einen möglichen Beitrag liefern, da anders als bei herkömmlichen Verfahren die Speicherung der Schlüssel entfällt. Die 3D-Cluster bieten also ein Konzept, um auch mit zukünftigen lithografischen Prozessen mikroskopisch kleine Fingerabdrücke in informationsverarbeitende, elektronische Systeme einzugravieren.

* * *

Anhang

A1 Literaturverzeichnis

A1.1 Referenzwerke

Kryptografie, Erfindungsschutz

Anderson, R. J., 2001. *Security Engineering: A Guide to Building Dependable Distributed Systems*. Verlag John Wiley & Sons, 2001.

Ernst, S., 2004. *Hacker, Cracker & Computerviren*. Verlag Dr. Otto Schmidt, Köln, 2004.

Kerckhoffs, A., 1883. *La cryptographie militaire*. Journal des sciences militaires. Band 9, Januar 1883, S. 5–38 und Februar 1883, S. 161–191.

Kurz, P., 2000. *Weltgeschichte des Erfindungsschutzes*. Verlag Carl Heymanns, Köln, 2000.

Qu, G., Potkonjak, M., 2003. *Intellectual Property Protection in VLSI Designs*. Kluwer Academic Publishers, Boston, 2003.

Shannon, C. E., 1948. *A Mathematical Theory of Communication*. The Bell Systems Technical Journal. Band 27, Juli 1948, S. 397–423 und Oktober 1948, S. 623–656.

Wodtke, C., Richters, S., 2004. *Schutz von Betriebs- und Geschäftsgeheimnissen*. Verlag Erich Schmidt, Berlin, 2004.

Zufallsgeneratoren, reproduzierbarer Zufall

Bock, H., Bucci, M., Luzzi R., 2004. *An Offset-Compensated Oscillator-Based Random Bit Source for Security Applications*. Lecture Notes in Computer Science: Cryptographic Hardware and Embedded Systems – CHES 2004. Springer-Verlag. Band 3156, 2004, S. 268–281.

Gassend, B., Clarke, D., van Dijk, M., Devadas, S., 2002. *Silicon Physical Random Functions*. Proceedings of the 9th ACM Conference on Computer and Communications Security, November 2002.

Knuth, D. E., 1969. *The Art of Computer Programming: Seminumerical Algorithms*. 3. Auflage, Addison-Wesely, 2006.

Lin, S., Costello, D. J., 1983. *Error Control Coding*. Prentice Hall, 2. Ausgabe, April 2004.

Lofstrom, K., Daasch, W. R., Taylor, D., 2000. *IC identification circuit using device mismatch.* Technical Digest of IEEE International Solid-State Circuits Conference. Februar 2000, S. 372–373.

Matsumoto, M., Nishimura, T., 1998. *Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator.* ACM Transactions on Modeling and Computer Simulation. Band 8, Nr. 1. 1998, S. 3–30.

Prozesstechnik, Matching

McCreary J. L., 1981. *Matching Properties, and Voltage and Temperature Dependence of MOS Capacitors.* IEEE Journal of Solid-State Circuits. Band 16, Nr. 6, Dezember 1981, S. 608-616.

Papoulis A., 1991. *Probability, Random Variables and Stochastic Processes.* 3. Auflage, McGraw-Hill, New York, 1991.

Pelgrom M. J. M., Duinmaijer A. C. J., Welbers, A. P. G., 1989. *Matching Properties of MOS Transistors.* IEEE Journal of Solid-State Circuits. Band 24, Nr. 5, Oktober 1989, S. 1433-1440.

Shyu J., Temes G. C., Krummenacher F., 1984. *Random Error Effects in Matched MOS Capacitors and Current Sources.* IEEE Journal of Solid-State Circuits. Band 19, Nr. 6, Dezember 1984, S. 948-955.

Shyu J., Temes G. C., Yao K., 1982. *Random Errors in MOS Capacitors.* IEEE Journal of Solid-State Circuits. Band 17, Dezember 1982, S. 1070-1076.

Wolf, W., 2004. *Microchip Manufacturing*, Lattice Press, Sunset Beach, 2004.

Ladungsinjektion

Eichenberger C., Guggenbühl W., 1989. *Dummy Transistor Compensation of Analog MOS Switches.* IEEE Journal of Solid-State Circuits. Band 24, Nr. 4, August 1989, S. 1143-1145.

Sheu B. J., Hu C., 1984. *Switch-Induced Error Voltage on a Switched Capacitor.* IEEE Journal of Solid-State Circuits. Band 19, Nr. 4, August 1984, S. 519-525.

Shieh J.-H., Patil M., Sheu B. J., 1987. *Measurement and Analysis of Charge Injection in MOS Analog Switches.* IEEE Journal of Solid-State Circuits. Band 22, Nr. 2, April 1987, S. 277-281.

Kapazitätsberechnung

Chang, W. H., 1976. *Analytical IC Metal-Line Capacitance Formulas.* IEEE Transactions on Microwave Theory and Techniques. Band 24, Nr. 9, September 1976, S. 608-611 und Band 25, Nr. 8, August 1977, S. 712.

Cottrell, P. E., Buturla E. M., 1985. *VLSI wiring capacitance.* IBM Journal of Research and Development. Band 29, Nr. 3, Mai 1985, S. 277–288.

Haji-Sheikh, A., Sparrow, E. M., 1966. *The Floating Random Walk and Its Application to Monte Carlo Solutions of Heat Equations.* Band 14, Nr. 2, März 1966, S. 370–389.

Kapur, S., Long, D. E., 1998. *IES3: Efficient electrostatic and electromagnetic simulation.* IEEE Computational Science and Engineering. Band 5, Nr. 4, Oktober/Dezember 1998, S. 60–67.

LeCoz, Y. L., Iverson, R. B., 1992. *A stochastic algorithm for high speed capacitance extraction in integrated circuits.* Solid-State Electronics. Band

35, Nr. 7, Juli 1992, S. 1005–1012.
Nabors, K., White, J., 1991. *FastCap: A Multipole Accelerated 3-D Capacitance Extraction Program*. IEEE Transactions on Computer-Aided Design, Band 10, Nr. 11, November 1991, S. 1447–1459.
Norrie, D. H., de Vries, G. 1973. *The Finite Element Method*. Academic Press, New York, 1973.
Ruehli, A. E., Brennan, P. A., 1973. *Efficient Capacitance Calculations for Three-Dimensional Multiconductor Systems*. IEEE Transactions on Microwave Theory and Techniques. Band 21, Nr. 2, Februar 1973, S. 76–82.

Kapazitätsmessung

Brambilla A., Maffezzoni P., Bortesi L., Vendrame L., 2003. *Measurements and Extractions of Parasitic Capacitances in ULSI Layouts*. IEEE Transactions on Electron Devices. Band 50, Nr. 11, November 2003, S. 2236-2247.
Chang Y.-W., Chang H.-W., Hsieh C.-H., Lai H.-C., Lu T.-C., Ting W., Ku J., Lu C.-Y., 2004. *A Novel Simple CBCM Method Free From Charge Injection-Induced Errors*. IEEE Electron Device Letters. Mai 2004, Band 25, Ausgabe 5, S. 262-264.
Chen J.C., McGaughy B.W., Sylvester D., C. Hu, 1996. *An On-chip, Attofarad Interconnect Charge-Based Capacitance Measurement (CBCM) Technique*. Technical Digest of International Electron Devices Meeting. Dezember 1996, S. 69-72.
Chen, J.C. Sylvester, D. Chenming Hu, 1998. *An on-chip, interconnect capacitance characterization method with sub-femto-farad resolution*. IEEE Transactions on Semiconductor Manufacturing. Mai 1998, Band 11, Ausgabe 2, S. 204-210.
Froment B., Paillardet F., Bely M., Cluzel J., Granger E., Haond M., Dugoujon L., 1999. *Ultra Low capacitance measurement in multilevel metallisation CMOS by using built-in Electron-meter*. Technical Digest of International Electron Devices Meeting. S. 897-900.

A1.2 Weiterführende Literatur

Kryptografie, Erfindungsschutz

Baukus, J. P., Chow, L. W., Clark, W. M., 1999. *Digital Circuit with Transistor Geometry and Channel Stops providing Camouflage against Reverse Engineering*. United States Patent US6064110, Mai 2000.
Kuhn, M., 1996. *Sicherheitsanalyse eines Prozessors mit Busverschlüsselung*. Diplomarbeit, Universität Erlangen-Nürnberg, Juli 1996.
Taddiken, H., Laackmann, P., 2000. *Vorrichtung zum Schutz einer integrierten Schaltung*. Europäische Patentanmeldung EP1182702, Februar 2002.
Wagner, W., 2003. *Halbleiter-Chip mit einer Identifikationsnummer-Generierungseinheit*. Europäische Patentanmeldung EP1465254, Oktober 2004.
Yip, K. W., Ng, S., 2000. *Apparatus and Method for Protecting Configuration Data in a Programmable Device*. United States Patent Application US2001/0032318, Oktober 2001.

Zufallsgeneratoren, reproduzierbarer Zufall

Bagini, V., Bucci, M., 1999. *A Design of Reliable True Random Number Generator for Cryptographic Applications*. Lecture Notes in Computer Science: Cryptographic Hardware and Embedded Systems – CHES 1999. Springer-Verlag. Band 1717, 1999, S. 204–218.

Bucci, M., Luzzi, R., 2005. *Design of Testable Random Bit Generators*. Lecture Notes in Computer Science: Cryptographic Hardware and Embedded Systems – CHES 2005. Springer-Verlag. Band 3659, 2005, S. 147–1546.

Gassend, B., Clarke, D., van Dijk, M., Devadas, S., 2002. *Controlled Physical Random Functions*. Proceedings of the 18th Annual Computer Security Conference, Dezember 2002.

Gassend, B., 2003. *Physical Random Functions*. Master's Thesis, Massachusetts Institute of Technology, Cambridge. Januar 2003.

Jun, B., Kocher, P., 1999. *The Intel Random Number Generator*. White Paper. Cryptography Research, Inc. April 1999.

Mandel, S., Banerjee, S., 2003. *An Integrated CMOS Chaos Generator*. Proceedings of the 1. Indian National Conference on Nonlinear Systems & Dynamics. Dezember 2004, S. 313–316.

NIST FIBS 140-1, 1994. *Security Requirements for Cryptographic Modules*. National Institute of Standards and Technology, Januar 1994.

NIST Special Publication 800-22, 2001. *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*. National Institute of Standards and Technology, Mai 2001.

NIST Special Publication 800-90, 2006. *Recommendation for Random Number Generation Using Deterministic Random Bit Generators*. National Institute of Standards and Technology, Juni 2006.

Zenner, E., 2004. *On Cryptographic Properties of LFSR-based Pseudorandom Generators*. Dissertation, Universität Mannheim, Mannheim, 2004.

Logik-Optimierung

Entrena, L., Abascal, J. G., Cheng, K.-T., 1993. Sequential logic optimization by redundancy addition and removal. Proceedings of the 1993 IEEE/ACM International Conference on Computer-Aided Design. November 1993, S. 310–315.

Prozesstechnik, Matching

Aparicio R., Hajimiri A., 2002. *Capacity Limits and Matching Properties of Integrated Capacitors*. IEEE Journal of Solid-State Circuits. Band 37, Nr. 3, März 2002, S. 384-393.

Barlow, R. J., 1989. *Statistics*. Verlag John Wiley & Sons, Chichester, England, 1989.

Boning, D. S., 1991. *Semiconductor Process Design: Representations, Tools, and Methodologies*. Dissertation, Massachusetts Institute of Technology, Cambridge, 1991.

Boning, D., Chung, J., *Statistical Metrology - Measurement and Modeling of Variation for Advanced Process Development and Design Rule Generation*, 1998 Int. Conference on Characterization and Metrology for ULSI Technology, Gaithersburg, MD, März 1998.

Bosch, K., 2000. *Elementare Einführung in die angewandet Statistik*, Verlag Friedr. Vieweg & Sohn, Braunschweig/Wiesbaden, 2000.

Gbondo-Tugbawa, T. E., *Chip-Scale Modeling of Pattern Dependencies in Copper Chemical Mechanical Polishing Processes*, Ph.D. Thesis, MIT Dept. of Electrical Engineering and Computer Science, Mai 2002.

Mehrota, J., 2001. *Modeling the Effects of Systematic Process Variation on Circuit Performance*, Ph. D. Thesis, MIT Dept. of Electrical Engineering and Computer Science, Mai 2001.

Park, T. H., 2002. *Characterization and Modeling of Pattern Dependencies in Copper Interconnects for Integrated Circuits*, Ph.D. Thesis, MIT Dept. of Electrical Engineering and Computer Science, Mai 2002.

Precht, M., Kraft, R., Bachmaier, M., 1999. *Angewandte Statistik 1*, Oldenbourg Wissenschaftsverlag, München, 1999.

Terrovitis M. T., 1996. *Process Variability and Device Mismatch*. Master's Thesis, University of California, Berkeley, 1996.

Terrovitis M. T., Spanos C. J., 1996. *Process Variability and Device Mismatch*. International Workshop on Statistical Metrology. Honolulu, Juni 1996.

Yu C., 1996. *Integrated Circuit Process Design for Manufacturability Using Statistical Metrology*. Dissertation, University of California, Berkeley, 1996.

Zhang H., 2002. *Causal Analysis of Systematic Spatial Variation in Optical Lithography*, Dissertation, University of California, Berkeley, 2002.

Ladungsinjektion

Wegmann G., Vittoz E. A., Rahali F., 1987. Charge Injection in Analog MOS Switches. *IEEE Journal of Solid-State Circuits*. Band 22, Nr. 6, Dezember 1987, S. 1091–1097.

Kapazitätsberechnung

Akcasu, O. E., et. al. 1995, *NET-AN a Full Three-Dimensional Parasitic Interconnect Distributed RLC Extractor for Large Full Chip Applications*, Proceedings International Electron Devices Meeting, 1995, S. 495–498.

Iverson, R. B., Le Coz, Y. L., 2001. *A floating random-walk algorithm for extracting electrical capacitance*. Mathematics and Computers in Simulation. Band 55, Nr. 1-3, Februar 2001, S. 59–66.

Janak, J. F., Ling, D. D., Huang, H.-M., 1989. *C3DSTAR: a 3D wiring capacitance calculator*. Proceedings of the 1989 IEEE International Conference on Computer-Aided Design. November 1989, S. 530–533.

Kapur, S., Long, D. E., 2000. *Large-Scale capacitance calculation.* Proceedings of the 37th IEEE Design Automation Conference. 2000, S. 744–749.

Magma Design Automation, Inc., 2004. *Quickcap 4.2 User Guide*, Oktober 2004.

Nabors, K., White, J., 1992. Multipole-Accelerated Capacitance Extraction Algorithms for 3-D Structures with Multiple Dielectrics. IEEE Transactions on Circuits and Systems. Band 39, Nr. 11, November 1992, S. 946–954.

Nowacka, E. B., 1996. *Hybrid Models for Parasitic Capacitances in Advanced VLSI Circuits*. Dissertation, Technische Universiteit Delft, 1996.

Kapazitätsmessung

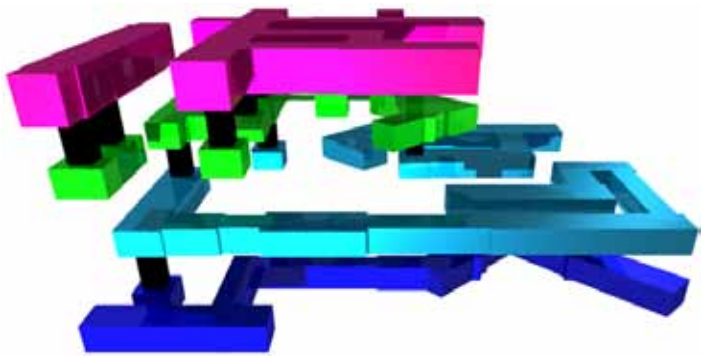
Krummenacher, F., 1985. *A High-Resolution Capacitance-to-Frequency Converter*. IEEE Journal of Solid-State Circuits. Band 20, Nr. 3, Juni 1985, S. 666–670.

* * *

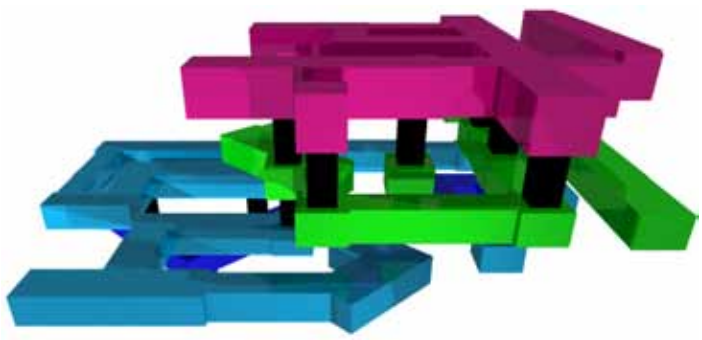
A2 Farbtafeln

(a) 3D-Schrägansicht, vorne.

(d) Entwurfsansicht.



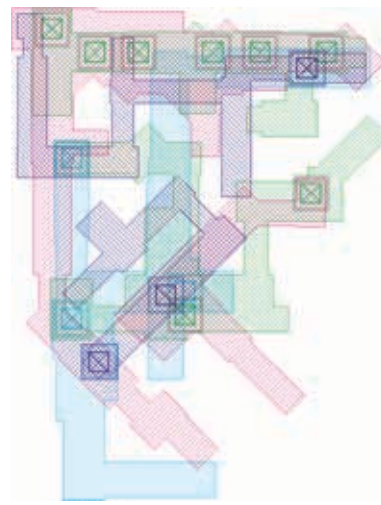
(b) 3D-Seitenansicht, links.



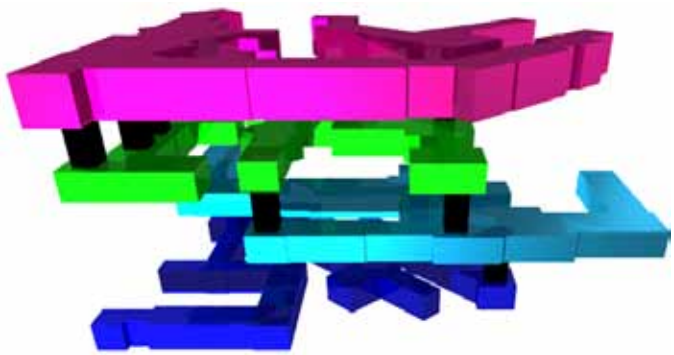
(c) 3D-Seitenansicht, rechts.

Farbtafel I. Cluster mit 5,52/6,30 fF (Quickcap bei 0,2%, typical-/worst-case), 5,25/6,49 fF (Assura-FS), 6,01/7,05 fF (Assura), 8,82 fF (Calibre, worst-case), 11,48 fF (Diva, worst-case).

(a) 3D-Schrägansicht, vorne.

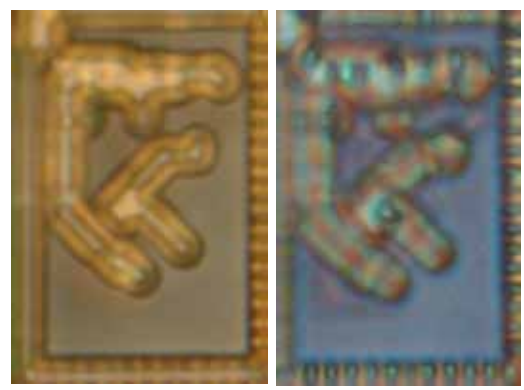


(d) Entwurfsansicht.



(b) 3D-Seitenansicht, links.

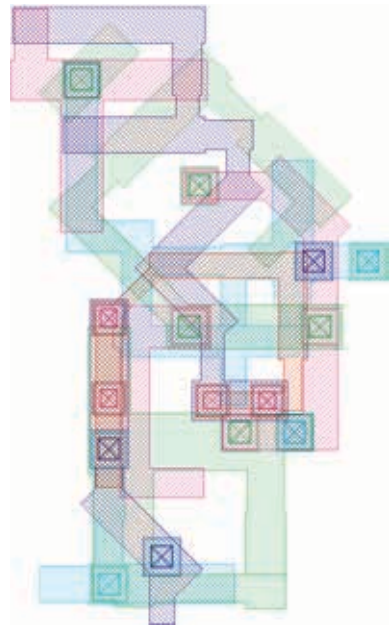
(c) 3D-Seitenansicht, rechts.



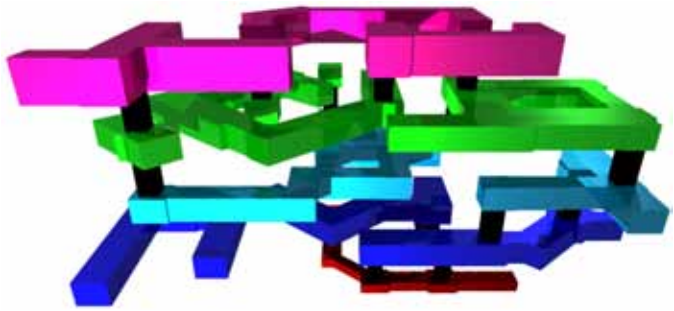
(e) Mikroskopische Aufnahme.

Farbtafel II. Cluster mit 6,21/7,43 fF (Quickcap bei 0,2%), 5,79/7,51 fF (Assura-FS), 6,50/7,99 fF (Assura), 9,12 fF (Calibre, worst-case), 12,57 fF (Diva, worst-case).

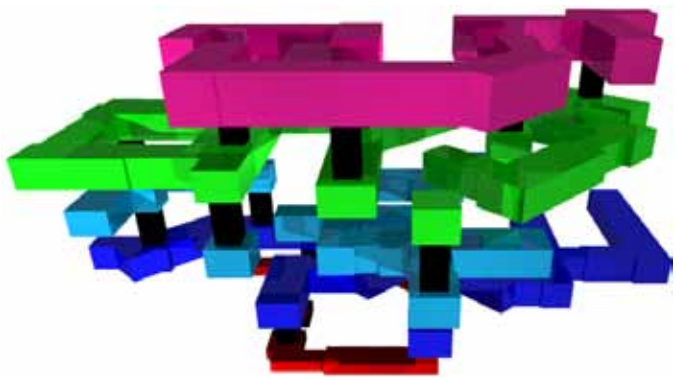
(a) 3D-Schrägansicht, vorne.



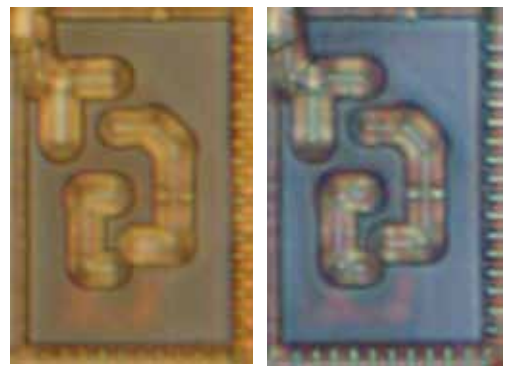
(d) Entwurfsansicht.



(b) 3D-Seitenansicht, links.

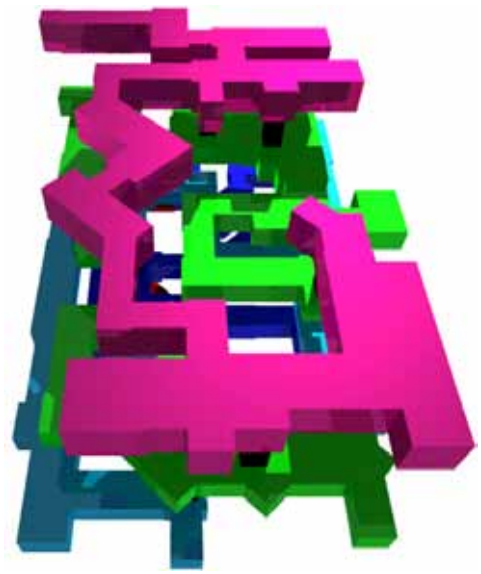


(c) 3D-Seitenansicht, rechts.



(e) Mikroskopische Aufnahme.

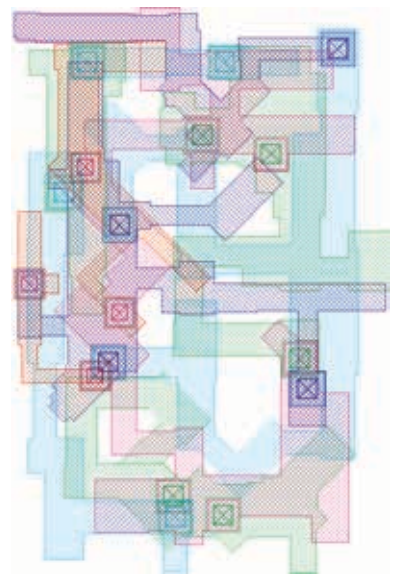
Farbtafel III. Cluster mit 9,07/11,08 fF (Quickcap bei 0,2%), 10,45/11,34 fF (Assura-FS), 9,00/11,29 fF (Assura), 12,30 fF (Calibre, worst-case), 15,61 fF (Diva, worst-case).



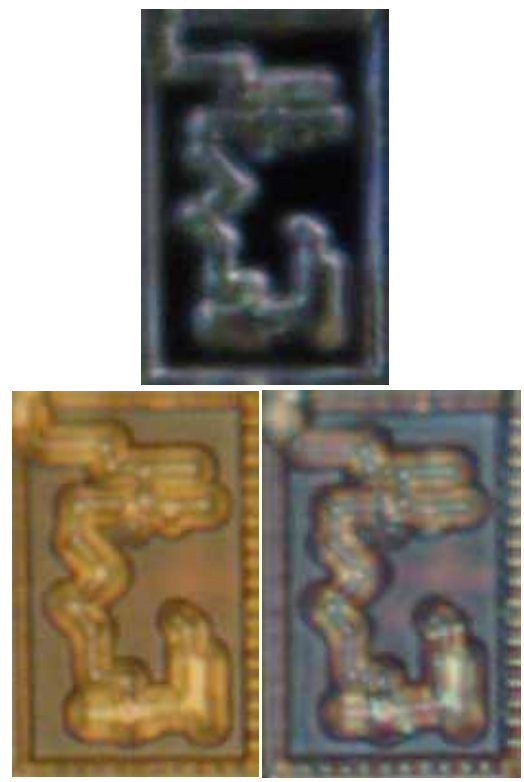
(a) 3D-Schrägansicht, vorne.

(b) 3D-Seitenansicht, links.

(c) 3D-Seitenansicht, rechts.



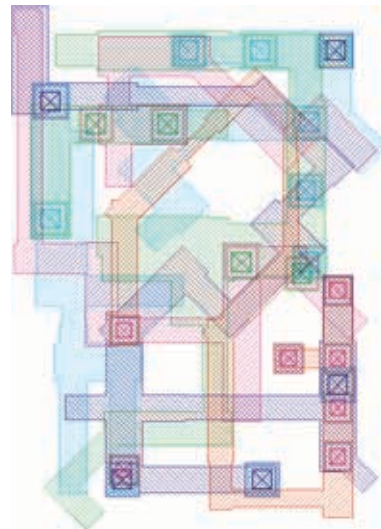
(d) Entwurfsansicht.



(e) Mikroskopische Aufnahme.

Farbtafel IV. Cluster mit 10,25/13,23 fF (Quickcap bei 0,2%), 10,22/13,03 fF (Assura-FS), 11,73/15,05 fF (Assura), 16,37 fF (Calibre, worst-case), 23,23 fF (Diva, worst-case).

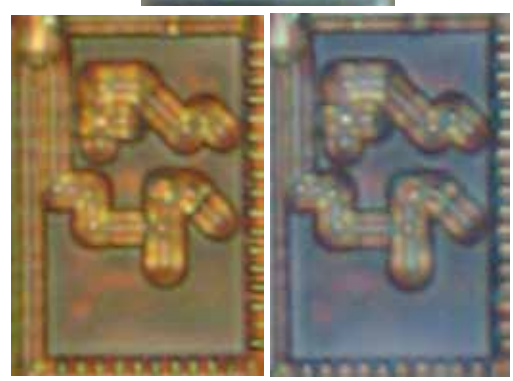
(a) 3D-Schrägansicht, vorne.



(d) Entwurfsansicht.

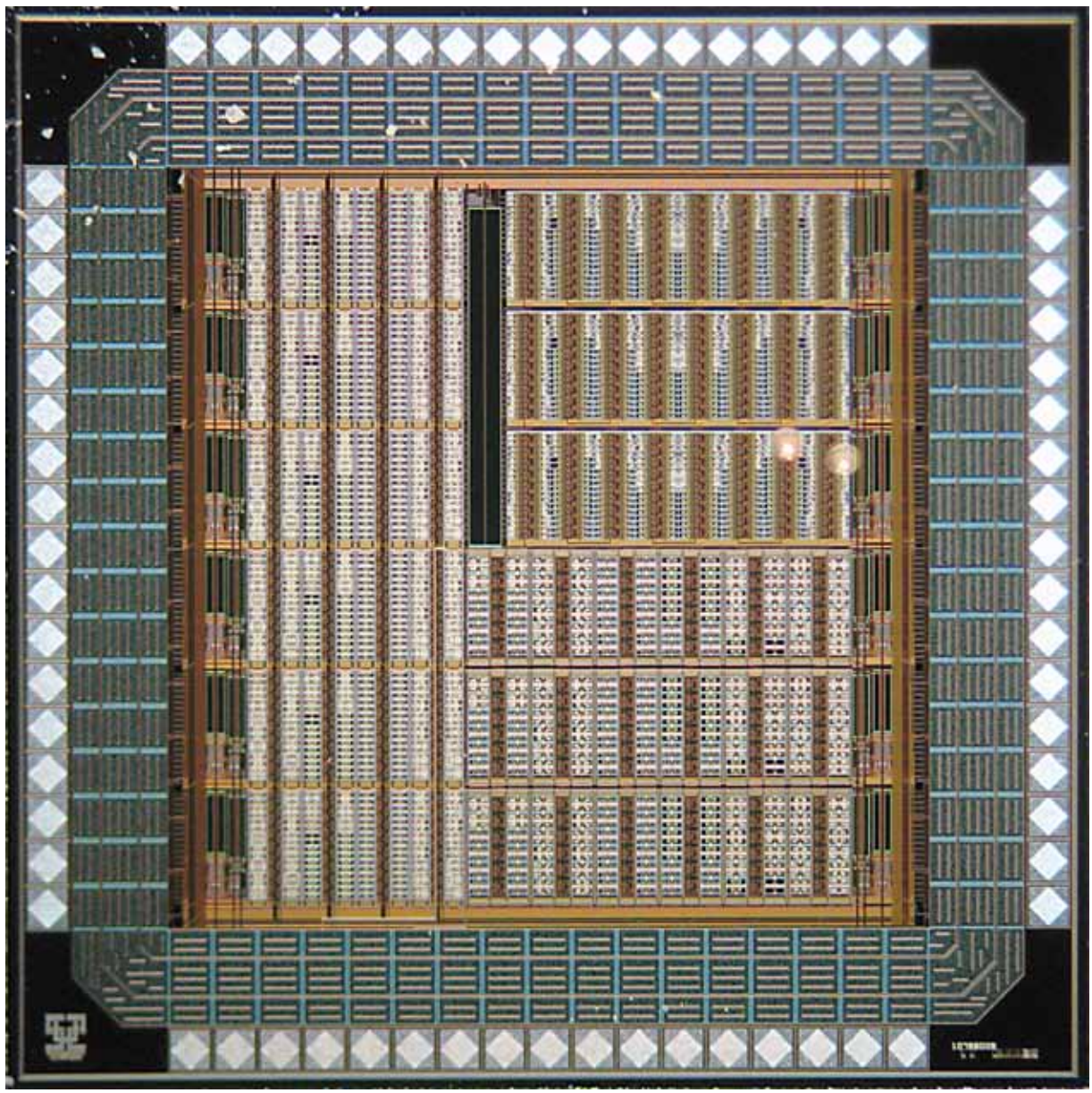
(b) 3D-Seitenansicht, links.

(c) 3D-Seitenansicht, rechts.

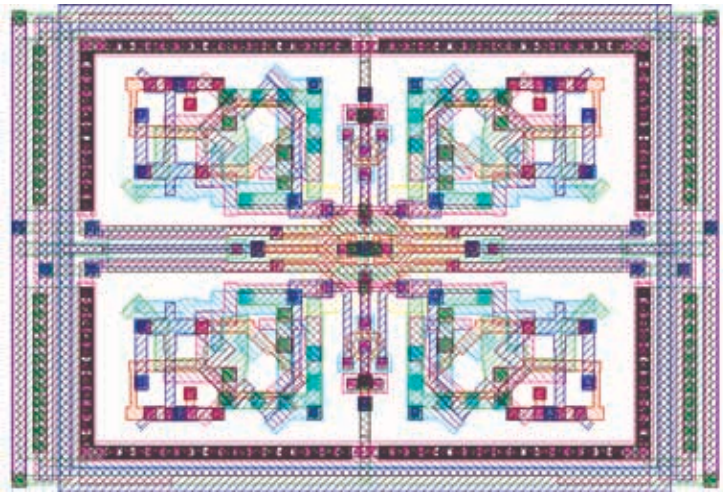


(e) Mikroskopische Aufnahme.

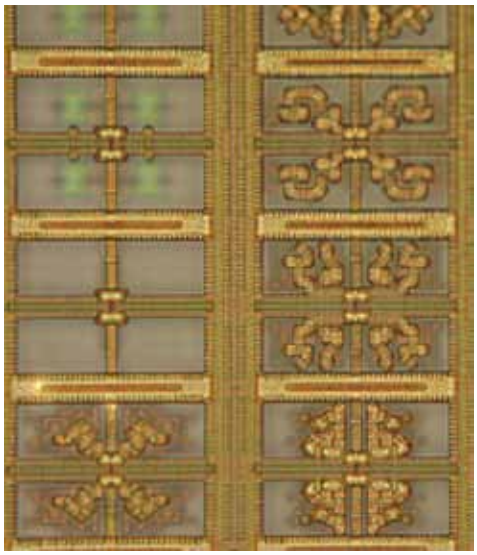
Farbtafel V. Cluster mit 12,57/15,02 fF (Quickcap bei 0,2%), 12,49/14,37 fF (Assura-FS), 13,59/16,47 fF (Assura), 17,66 fF (Calibre, worst-case), 22,87 fF (Diva, worst-case).



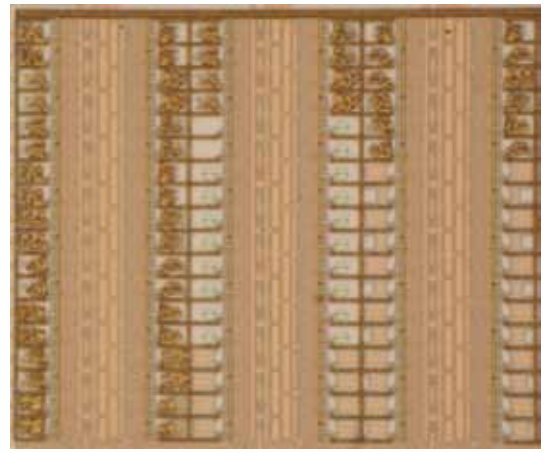
Farbtafel VI. Mikroskopische Aufnahme des Schüsselelektronik-Testchips.



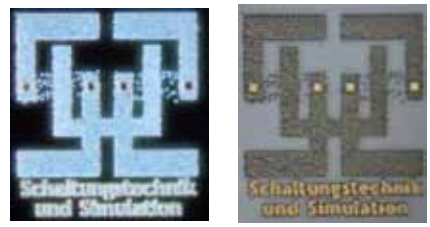
(a) Layout der Variante mit vier Pumpzweigen



(b) Mikroskopbild von 6 Zellen aus (a)



(c) Mikroskopbild einiger Spalten der Variante mit zwei Pumpzweigen



(d) Mikroskopbild des Lehrstuhl-Logos

Farbtafel VII. Layout und Mikroskopbild der Schaltungsvariante mit vier Pumpzweigen auf der linken Seite. Die Matrix in der Mitte rechts beinhaltet eine Schaltungsvariante, die das Kapazitätsverhältnis von jeweils zwei Clustern auswertet.

A3 Abbildungs- und Tabellenverzeichnis