Fakultät für Elektrotechnik und Informationstechnik

Lehrstuhl für Integrierte Schaltungen
der Technischen Universität München

Entwicklung neuer physikalischer Optimierungs- und Regelungsverfahren für thermische Reaktoren in der Halbleiterprozessierung

Thomas Schafbauer

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

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende:	Univ.-Prof. Dr. rer.nat. D. Schmitt-Landsiedel
Prüfer der Dissertation:	1. Univ.-Prof. Dr.-Ing. I. Ruge
	2. Univ.-Prof. F. Koch, Ph.D.

Die Dissertation wurde am 10.08.2000 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 02.04.2001 angenommen.

Zusammenfassung

Thema der Dissertation ist der Entwurf und die experimentelle Umsetzung von Optimierungs- und Regelverfahren in thermischen Halbleiterprozessen.
Die in dieser Arbeit angewandte Methodik beruht auf einer Aufspaltung des Regelungsproblems in eine detaillierte simulationstechnische Analyse des Prozesses und eine Optimierung, die einen geringen Satz von Charakterisierungsmessungen benötigt.
Auf Basis der Simulationen werden Verbesserungen an den Kammerbedingungen entwickelt. Ziel ist die Vereinfachung des Regelproblems bereits im Vorfeld.
Der zweite Schritt besteht im Entwurf eines reduzierten Modells mit numerischen und experimentell gewonnen Parametern. Nach der Parameterextraktion wird mit einem Optimierungsverfahren eine homogene Einstellung der Prozeßparameter gewonnen.
Ein weiterer wichtiger Punkt besteht in dem Entwurf von modellbasierten Steuer- und Regelstrategien für thermische Einzelscheibenprozesse. Anwendung findet dies bei der Analyse der Reglerperformance, sowie bei der Regelung von Wafern mit variierenden optischen Eigenschaften und bei der Regelung unter der Nebenbedingung der Minimierung des thermischen Budgets.
Mit Hilfe dieses Ansatzes konnte unter anderem die Optimierung der Temperaturuniformität in einer Oxidationskammer erreicht werden.

Abstract

The dissertation focuses on the development and experimental realization of optimization- and model-based-control methods for thermal semiconductor processes. The new method proposed in this work uses a split approach of the underlying control problem into a detailed simulation-based analysis of the process and an optimization using only a small set of measurements.
To reduce the burden of the control problem improvements on the process chamber are employed using equipment simulation in an early stage. In a second step a numerically reduced model is developed which contains both – simulation- and experiment-based – parameters. Optimization strategies are used then after the parameter-extraction to achieve a homogenous configuration of the target process parameters.
Of equal importance is the development of control techniques with and without feedback in thermal, single-wafer process steps. Fields of application are the analysis of controller performance, the feedback control of wafers with varying optical properties, and process control under the constraint of the thermal budget. Using this approach one of the achievements is the optimization of temperature uniformity in an rapid thermal oxidation chamber.

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Anforderungen in der Halbleiterentwicklung

Die Halbleiterindustrie ist ein seit Jahren stetig expandierender Wirtschaftssektor mit jährlichen Produktivitätssteigerungen von durchschnittlich 30%. Die stetige technische und wissenschafliche Weiterentwicklung und der daraus resultiernde Kostendruck führt zu einer Steigerung der Speicherkapazität und CPU-Rechenleistung um einen Faktor vier alle drei Jahre bei stabilen oder sogar fallenden Preisen. Dieser aus Marktanalysen in den 60er und 70er Jahren gefundene Zusammenhang (Moore'sches Gesetz) scheint seit Mitte der 90er Jahr sogar noch übertroffen zu werden. Außer auf das eigene Umsatzvolumen wirkt die Mikroelektronik auch als Entwicklungsmotor auf nahezu alle anderen Bereichen der technischen Industrie – von der Medizintechnik über die Automobilindustrie bis hin zum Flugzeugbau.

Einhergehend mit der Expansion des Halbleitermarktes kommt es zu umfangreichen Investitionen im Bereich der Halbleiterherstellung und Technologieentwicklung. Nur auf diese Weise kann durch fortschreitende Miniaturisierung der Produktivitätszuwachs aufrecht erhalten werden. Der weltweite Umsatz von Halbleiterequipment hat sich nach Marktanalysen von SEMI von 3.6 Mrd. US-Dollar 1992 auf 21.4 Mrd. US-Dollar 1996 mehr als verfünffacht [3]. An der Spitze liegt hier das Front-End, also die eigentliche Waferbearbeitung, mit einem Marktanteil von 70%, deutlich vor dem Packaging (7.5%) und den Testvorrichtungen (14%).

Der Grund für die Umsatzsteigerung liegt zum einen im Wachstum der Halbleiterindustrie insgesamt (Abbildung 1.1), zum anderen an der Verteuerung der

Maschinen infolge der die ansteigenden Anforderungen an die Prozesse (Tabelle 1.1).

Für die IC-Hersteller bedeutet dies einen steigenden Kapitaleinsatz für den Aufbau neuer Fabriken. Beliefen sich die Kosten für den Aufbau einer Halbleiterfabrik im Jahre 1992 noch auf durchschnittlich etwa 750 Millionen US-$, so waren es im Jahr 1997 zwischen 1.5 und 2 Milliarden US-$. Mit einem weiteren Anstieg der Investitionen ist zu rechnen: zur Fertigung neuer Technologiegenerationen steigen die Anforderungen an die Homogenität während des Prozesses und an die Ausbeute der Produktion überproportional an. Größere Waferdurchmesser senken zwar die Kosten der Produktion insgesamt, erhöhen aber die Kosten für den einzelnen Reaktor. Gleichzeitig muß die Zeit zum Einfahren des Equipments konstant gehalten oder sogar gesenkt werden.

Die Hauptprodukte der Halbleitertechnologie bestehen in der Anfertigung von Speicher- und Mikroelektronikbauelemente (ICs) auf Basis von Silizium. ZADementsprechend konzentriert sich die vorliegende Arbeit vornehmlich auf die Prozessierung von Siliziumscheiben, ist aber in analoger Weise auf andere Materialien anwendbar.

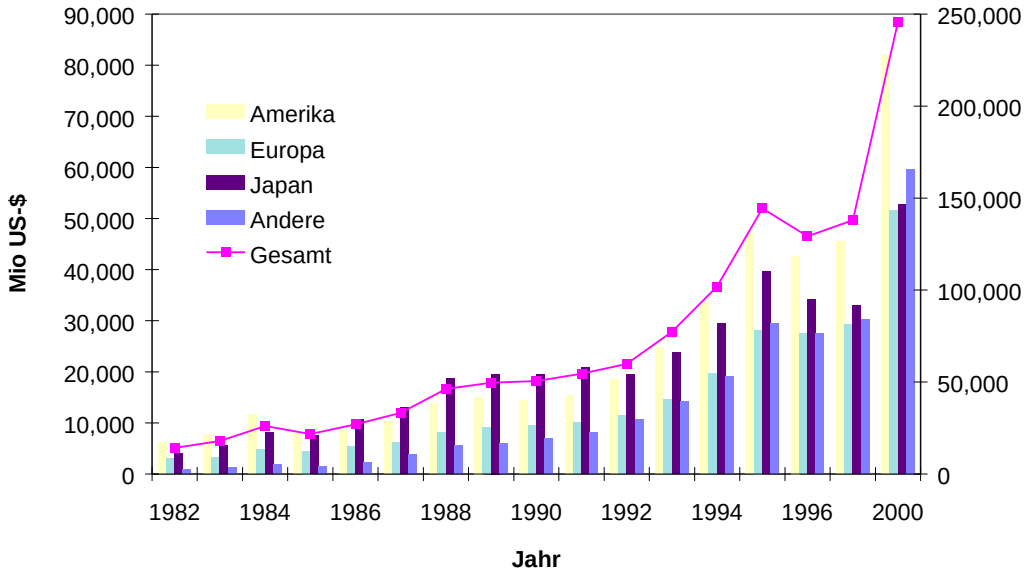


Abbildung 1.1: Entwicklung des Halbleiterumsatzes weltweit (rechte Achse) und aufgeteilt nach Regionen (Balken, linke Achse). Das durchschnittliche Wachstum der Branche von 1959 bis heute beträgt 17% jährlich. [6]

	1995	1997	1999	2001	2003	2006	2009
Technologie Generation	350nm	250nm	180nm	150nm	130nm	100nm	70nm
CPU Frequenz (MHz)	500	750	1200	1400	1600	2000	2500
DRAM Chipfläche (mm^2)	190	280	420	640	560	790	1120
Kosten/Bit (microcent)	170	120	60	30	15	5.3	1.9
Waferdurchmesser (mm)	200	200	300(!)	300(!)	300(!)	300(!)	450(!!)
Kantenausschluß (mm)	5	3	2(!)	2(!)	2(!)	1(!!)	1(!!)
Reaktorauswahl (Monate)		10(!!)	10(!!)	9(!!)	8(!!)	7(!!)	7(!!)
T_{ox} equiv. (nm)	4-5	3-4	2-3	2-3	1.5-2	< 1.5	
T_{ox} Kontrolle (3σ)	4%	4%	4% (!)	4%-6% (!)	4%-8% (!!)	4%-8% (!!)	
Modellierung/Simulation							
Reaktor scale Konz.		Trends	Trends	50%(!)	50%(!)	20%(!!)	20%(!!)
Feature scale Konz.		Trends	Trends	50%(!)	50%(!)	20%(!!)	20%(!!)
Ätzen/Abscheidung		Trends(!)	100%(!)	50%(!)	50%(!)	20%(!!)	20%(!!)
Δ Gate-Oxid Dicke (nm)		0.45(!)	0.35(!)	0.30(!!)	1 layer(!!)	1 layer(!!)	1 layer(!!)

Tabelle 1.1: Auszug aus der SIA Roadmap für die in dieser Arbeit relevanten Entwicklungen. Aufgrund der Wettbewerbssituation wird ein rasches Aufeinanderfolgen der Technologiegenerationen erwartet, welches das Moore'sche Gesetz (Verdopplung der Speicherkapazität und Rechenleistung etwa alle 18-20 Monate) noch übertrifft. Gleichzeitig steigen die Anforderungen an die genaue Einhaltung der Spezifikationen und an die Herabsetzung der Kosten durch Erhöhung der Packungsdichte und Verringerung der Einfahrzeiten. Für die Felder mit `!` werden derzeit Lösungen entwickelt; für die Felder mit `(!!)` zeichnen sich bisher noch keine praktikablen Lösungsansätze ab.

1.2 Bedeutung der Modellierung als Entwicklungswerkzeug

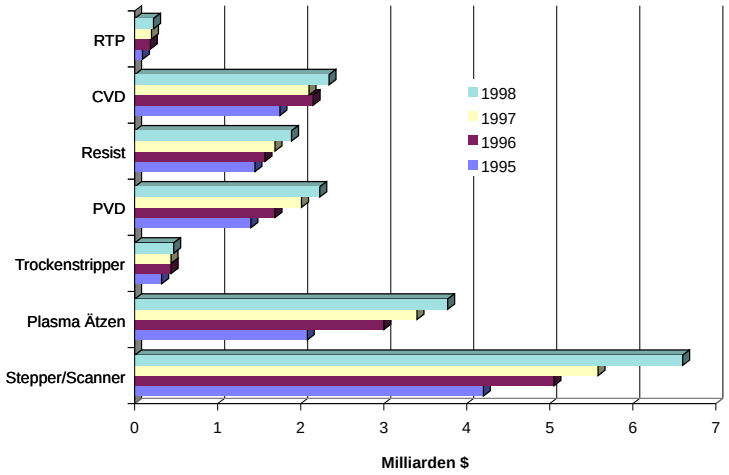


Abbildung 1.2: Weltweite Aufwendungen für Equipment für die wichtigsten Prozeßschritte im Front-End. Der Marktanteil von Rapid-Thermal-Processing und Chemical-Vapour-Depostition ist demgegenüber gering. Die Steigerungsraten von RTP liegen aber mit 12.4% hinter denen für die Lithographie/Stepper (18.4%) und vor den anderen Prozessen. Der Anteil von RTP bei Hochtemperaturprozessen liegt bei etwa 20%.

Empirische Verfahren zur Konfiguration und Regelung des Reaktors stoßen bei kritischen Prozeßschritten angesichts der oben beschriebenen Anforderungen an ihre Grenzen. Die Simulation wurde deshalb von der Semiconductor Industry Association als wichtiges Hilfsmittel zum Erreichen der Anforderungen erkannt. Daher enthält die Ausgabe der Roadmap von 1997 erstmals Anforderungen an die Genauigkeit der Simulationsergebnisse. In der vorliegenden Arbeit soll gezeigt werden, bei welchen Fragestellungen Modellierung und Simulation hinreichend genaue Aussagen liefern können, um Prozeßverbesserungen zu erzielen.

Einer der Schwerpunkte dieser Arbeit liegt auf der Optimierung von RTP-Prozessen[1]. Zwar ist das mit diesen Prozeßschritten verbundene Kosten- und Investitionsvolumen im Verhältnis zu anderen Frontend-Prozessen vergleichsweise gering (Abbildung 1.2), jedoch sind sie Schlüsselprozesse in der modernen

[1]Eine genaue Beschreibung der RTP-Reaktoren findet sich im folgenden Kapitel

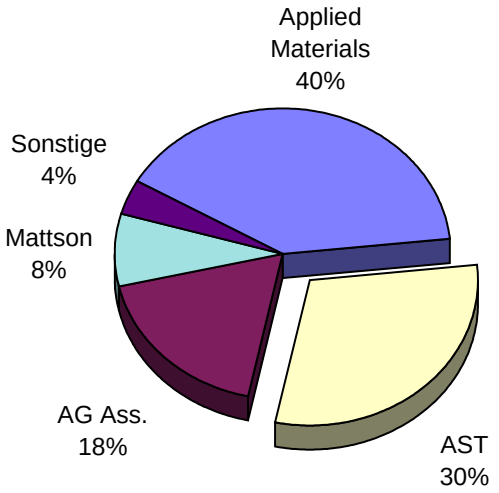


Abbildung 1.3: Marktanteile der RTP-Hersteller 1997 (geschätzt [5])

Halbleiterindustrie. Aufgrund der schwierig zu erreichenden Temperaturuniformität ist ihr Einsatz problematisch und sie stellen damit einen wichtigen Einsatzbereich für Regel- und Optimierungsverfahren.

Aus zwei Gründen ist mit einem weiter verstärkten Einsatz von RTP zu rechnen: Zum einen ist der induzierte Gravitationsstreß bei der Waferauflage in herkömmlichen Ofenreaktoren insbesondere bei Waferdurchmessern von 300 Millimeter und mehr derart hoch, daß Versetzungslinien im Siliziumkristall verstärkt auftreten; in RTP-Systemen kann die Waferauflage anders positioniert werden. Bei Ofenprozessen muß zum anderen bei steigendem Waferduchmesser die Aufheizzeit erhöht werden, um eine gleichmäßige Temperatur auf der Scheibe zu gewährleisten. Dies führt zu einer Erhöhung des thermischen Budgets (siehe Kapitel 4.5), also der Aktivierung sekundärer Transportprozesse im Wafer, z.B. des Ausschmierens von Implantationsprofilen. Kleinere Strukturen erfordern Dotierungsprofile mit geringer und reproduzierbarer Tiefenverteilung im Silizium. Dies wiederum erfordert ein reduziertes thermisches Budget. Auch Verfahren zur Qualitätssicherung wie in-situ Monitoring sind bei Einzelscheibenverfahren wie RTP leichter einzusetzen.

Die in den folgenden Kapiteln vorgestellte Methodik beschränkt sich nicht auf RTP, sondern läßt sich vielmehr in abgewandelter Weise auch auf andere Prozesse, wie z.B. CVD und Plasma erweitern.

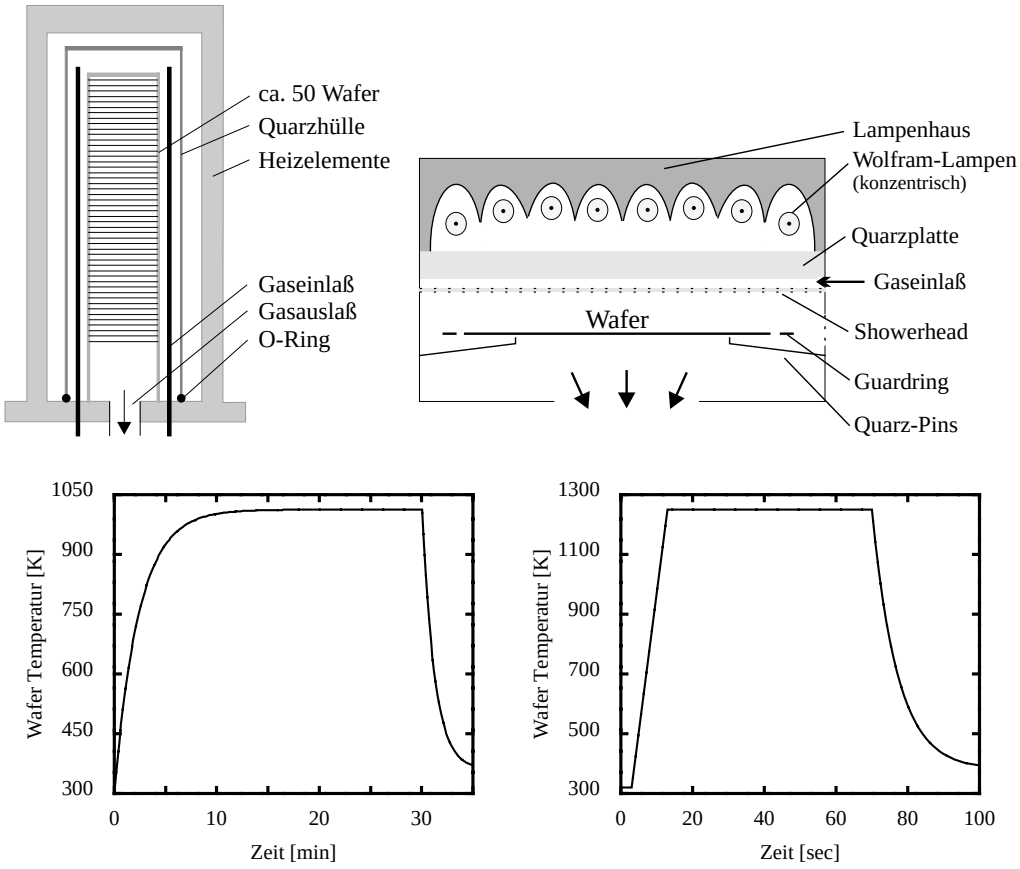


Abbildung 1.4: Aufbau eines vertikalen LPCVD Batch-Reaktors (links) und eines typischen RTP-Reaktors, darunter die typischen Prozeßzeiten. Durch die rasche Aufheizphase und die kurze Verweildauer bieten RTP-Systeme Vorteile in bezug auf die Diffusion von Dotierstoffprofilen (siehe auch Kapitel 4.5).

1.3 Ergebnisse dieser Arbeit

Kern dieser Arbeit war die Entwicklung eines physikalischen Optimierungs- und Regelungsverfahrens für RTP-Systeme. Davon ausgehend wurden im Rahmen der Promotion Strategien für die Optimierung und Regelung von thermischen Halbleiterprozessen entwickelt und erfolgreich umgesetzt. Im einzelnen umfaßt dies:

1. Detaillierte, dreidimensionale Simulationen einer für die Produktion von EEPROM eingesetzten Kammer. Der hierfür verwendete Simulator wurde in wesentlichen Teilen, insbesondere für die Strahlungssimulation auf unregelmäßigen Gittern und für die Analyse geregelter Prozesse erweitert
2. Analyse des transienten Reglerverhaltens eines PID-gesteuerten Suszeptorsystems
3. Verfahren zur Minimierung des thermischen Struktureffekts durch ein Regelverfahren
4. Entwicklung eines reduzierten Modells für Einzelscheibensysteme, das zur Echtzeitregelung und zur Kalibration an Messungen geeignet ist. Die Gültigkeit des Modells wird sowohl in der Simulation wie auch im Experiment nachgewiesen.
5. Entwicklung eines neuen Verfahrens zur Optimierung der Wafertemperatur auf der Basis des reduzierten Modelles bei Vorgabe eines Satzes von Thermoelementmessungen
6. Erfolgreiche experimentelle Durchführung des Verfahrens an einer RTO-Kammer
7. Entwurf und Programmierung eines Verfahrens zur Minimierung des thermischen Budgets bei Einzelprozessen und Überprüfung durch Prozeßsimulationen

1.4 Aufbau der Dissertation

Der Ausgangspunkt des Verfahrens ist eine detaillierte Modellierung des Systems (Kapitel 2). Zunächst werden die relevanten physikalischen Gleichungen

erläutert, die in dem Simulator PHOENICS implementiert sind und für die Simulation der untersuchten Reaktoren von Bedeutung sind. In der Simulation werden auf Basis dieser Gleichung Sensitivitätsstudien des Reaktorverhaltens durchgeführt, um die kritischen Einflußgrößen auf Prozeßergebnis und Uniformität zu identifizieren.

Mit Hilfe dieser Erkenntnisse wird ein reduziertes Modell (Kapitel 3.3) abgeleitet, das die wesentlichen physikalischen Einflußfaktoren enthält, gleichzeitig aber für die Parameter, die nicht mit genügend hoher Genauigkeit modellierbar sind, eine Anpaßbarkeit an Messungen vorsieht. Bei der Formulierung des Modells wird darauf geachtet, daß die Parameter mit einem einfachen Meßverfahren ermittelbar sind.

Der experimentelle Schritt liefert die nötigen Parameter, die dann in der ex-situ Optimierung (Kapitel 3) Anwendung finden. Durch die Parameterextraktion mit hoher Statistik ist dies in einem Arbeitsbereich und nicht nur an einem Arbeitspunkt möglich. Die erfolgreiche Anwendung sowie die Stabilität dieses Verfahren wird sowohl numerisch (Abschnitt 3.5) wie auch experimentell (Abschnitt 3.6) bei der Uniformitätsoptimierung einer RTO-Kammer gezeigt.

Mit den im Simulationsabschnitt gewonnen Kenntnissen bleibt für die eigentliche Regelung (Kapitel 4) dann die deutlich einfachere Aufgabe des Konstanthaltens im stationären Zustand. Die Simulation ist auch bei der Effektivitätsanalyse von Regelungen von Bedeutung. Dies wird in Abschnitt 4.2 am Beispiel eines RTA-Prozesses demonstriert, der unterschiedliche Prozeßergebnisse je nach Position des Wafers in der Bearbeitung aufweist.

Weitergehende Regelungsfragen erfordern je nach Meßbarkeit und Fragestellung unterschiedlich komplexe physikalische Modellierungen. Dazu zählt z.B. die Leistungssteuerung bei variierenden optischen Eigenschaften (Abschnitt 4.3) mit einer einfachen Abschätzung oder die optimale Trajektorienberechnung bei Minimierung des thermischen Budgets (Abschnitt 4.5) mit sehr detaillierten Simulationen.

Im letzten Abschnitt werden Möglichkeiten und Grenzen der Methodik bei der Anwendung auf andere Einzelprozesse in der Halbleiterbearbeitung diskutiert.

1.5 Motivation: Fertigungstoleranzen für die Halbleiterprozessierung

Grundlegendes Ziel eines Optimierungs- und Regelungsansatzes ist es, einen Zielwert möglichst genau einzuhalten. In diesem Abschnitt soll zunächst geklärt

werden, welche Genauigkeitsanforderungen für Sub-Mikron Halbleiterprozesse erfüllt werden müssen.

Ausgehend vom eigentlichem Ziel – der bestmöglichen Funktion einer Halbleiterschaltung – wird versucht, Spezifikationen für die Einzelprozesse festzulegen. Von grundlegenden Produktanforderungen ausgehend wird anhand von Beispielschaltungen auf die tolerierbaren Schwankugen von Transistoren geschlossen, woraus sich schließlich Anforderungen an die Toleranzen für die thermischen Prozeßschritte ergeben.

Die Überlegungen stellen indes allenfalls eine Motivation, keine stringente Ableitung dar. Genauere Analysen sind den Veröffentlichungen in [10] zu entnehmen.

1.5.1 Qualitätsskriterien für Halbleiterschaltungen und -produkte

Das Produkt- und Anwendungsspektrum von Halbleiterschaltungen umfaßt so unterschiedliche Bereiche wie z.B. Datenerfassung, Datenverarbeitung, Telekommunikation, Automobilsteuerung und Chipkarten.

Jede diese Anwendungen stellt unterschiedliche Anforderungen an die dahinterstehenden Schaltungen. Eine konsequente Ableitung der Anforderungen für jeden Herstellungsschritt ist daher nicht eindeutig möglich, da die "bestmögliche Funktionen" produktspezifisch sind.

Die folgenden Betrachtungen werden daher auf drei relevante Qualitätskriterien beschränkt, die einem Großteil der Produkte gemein sind:

1. Geringe Herstellungskosten/Yield
   Der treibende Faktor der Miniatisierung ist der Produktivitätsgewinn. Wie bereits in der Einleitung dargestellt, vervierfacht sich die Komplexität der Schaltungen etwa alle 3 bis 4 Jahre. Wichtig für niedrige Produktionskosten ist ein hoher "Yield", d.h. der Prozentsatz von Chips pro Wafer, die die Poduktspezifikationen erfüllen, muß maximiert werden.

2. Hohe Verarbeitungsgeschwindigkeit (Performance) P_{Chip}
   Erst mit einer hinreichenden Schaltgeschwindigkeit werden viele Anwendungen möglich. Dies gilt vor allem für Mikroprozessoren in Computern und für Telekommunikation (Handy, Glasfaser). Die Prozessorleistung verdoppelt sich etwa alle 4 bis 5 Jahre, was zu schnellerer Bewältigung von Rechenanwendungen und neuen Einsatzgebieten führt.

3. Geringe Verlustleistung $W_{Verlust}$
   Aufgrund der hohen Packungsdichte (bei heutigen Halbleiterprodukten mehr als 40 Millionen Bauelemente pro Quadratzentimeter) summiert sich der Leistungsverbrauch von Prozessorchips in 0.18μ Technologien auf mehr als 40 Watt. Bis zu 100 Watt Aufnahmeleistung wird fuer die 0.13μm Mikroprozessorgeneration erwartet. Die Wärmeabfuhr über das wenige Quadratzentimeter große Prozessorgehäusse begrenzt den tolerierbaren Stromverbrauch. Für batteriebetriebene Produkte sind die Verlustanforderungen noch weitaus höher, um die Gebrauchszeiten zwischen den Batterieaufladungen (stand-by) zu maximieren.

Es zeigt sich, daß diese drei Produktanforderungen gegenläufig sind. Komplexe Schaltungen mit hoher Performance weisen zumeist eine hohe Verlustleistung im Betrieb und/oder im stand-by auf. Zudem ist ihr Flächenbedarf auf dem Wafer hoch, was die Fertigungskosten erhöht und zumeist den Yield verringert.

Jeder Prozeßschritt unterliegt Fertigungstoleranzen, die zu einer Verteilung der Chips bzgl. der Parameter 2. und 3. führt. Die Standardabweichungen dieser Verteilungen seien durch

$$\sigma_P := \langle P_{Chip} - \hbar P_{Chip} \rangle \, \sigma_W := \langle W_{Verlust} - \hbar W_{Verlust} \rangle \tag{1.1}$$

gegeben, wobei die überstrichenen Werte den Mittelwert der Verteilung darstellen.

Eine Schaltung muß so entworfen werden, daß sie in einem Fertigungsfenster noch operabel ist. Dieses Fertigungsfenster F läßt sich durch Faktoren x_P und x_W beschreiben

$$\mathcal{F}_P = \pm x_P \times \sigma_P \mathcal{F}_W = \pm x_W \times \sigma_W \tag{1.2}$$

Ist das Fenster asymmetrisch bezüglich $\hbar P_{Chip}$ oder $\hbar W_{Verlust}$, so bezeichnen x_P und x_W den kleineren der beiden Abstände zwischen Mittelwert und oberer bzw. unterer Fensterrand.

Verschärft man nun die Spezifikationen bzgl. der Punkte 2. und 3., so verringert sich bei vorgegebenen Fertigungsschwankungen die Toleranzfaktoren x und somit der Yield.

Da die Anzahl der Prozeßschritte für ein Halpleiterprodukt groß ist – typischerweise sind mehr als 300 Prozessierungsschritte notwendig –, ist die Annahme

einer Normalverteilung der Chips bzgl. der Parameter 2. und 3. zumeist gerechtfertigt [11]. Hieraus ergeben sich für verschiedene Faktoren x die folgende Yieldtabelle:

x	Yieldverlust (1 - Yield)
0.5	
1	31.73%
2	4.55%
3	0.27%
4	63 ppm
5	0.57 ppm
6	0.00197 ppm

Die Annahme einer Normalverteilung ist aber aufgrund von Bedienungsfehlern und Begrenzungen der Zuverlässigkeit in der Praxis nur in einer Umgebung von weniger als 3 σ richtig. Die Wahrscheinlichkeit, daß ein Wert von 5 σ durch Fehlprozessierung auftritt ist deutlich höher als 1 in einer Million.

1.5.2 Einzelprozeßschwankungen

Jeder Bearbeitungsschritt unterliegt Fertigungstoleranzen, die sich auf die elektrischen Eigenschaften des Bauelements und der Schaltung auswirken. Die Auswirkung dieser Schwankugen wird hier anhand der Lithografieschwankung bei der Belichtung des Polysiliziumsgates gezeigt, da diese zumeist den größten Beitrag zu den Fertigungsschwankungen liefert.

In Abbildung 1.5 ist die Verteilung der elektrischen Kanallänge über 10 Lose mit jeweils etwa 6 Wafern und 12 gemessenen Chips pro Wafer für eine 0.18μm Technologie dargestellt. Die Lose wurden innerhalb eines Zeitraums von etwa 4 Monaten gefertigt und zeigen eine Verteilung, die einer Normalverteilung ähnlich ist. Aus der Verteilung ergibt sich eine Standardabweichung σ von etwa 10 Nanometern. Die elektrische Kanallänge wurde mit dem Shift-and-Ratio Verfahren ermittelt, da eine optische Messung der Gatelänge zu aufwendig und auch ungenauer ist.

1.5.3 Bauelementedesign

Die Aufgabe des Bauelementedesigns besteht darin, durch Festlegung der Prozeßbedingungen der Einzelprozesse (Implantationen, Temperaturschritte,

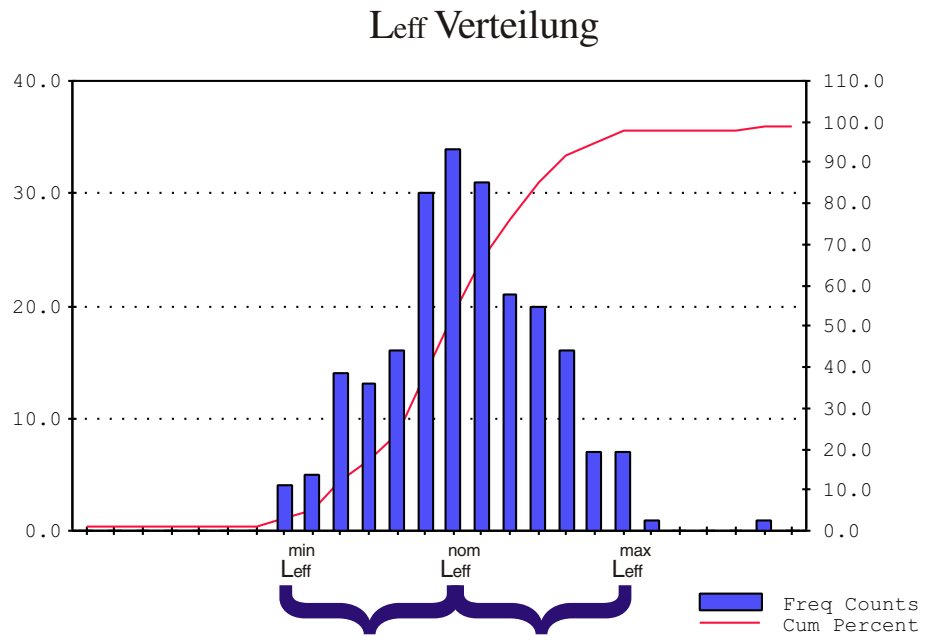


Abbildung 1.5: Verteilung der elektrischen Kanallänge von nMOS Transistoren in einer 0.18μm Generation. Die Verteilung entspricht in guter Näherung einer Normalverteilung.

Schichtdicken etc.) eine Optimierung bzgl. der drei Qualitätskriterien zu erzielen. Im folgenden werden die Überlegungen auf Basis von MOSFETs, dem am weitesten eingesetzten Bauelement durchgeführt. Diese lassen sich mit kleineren Änderungen auch auf andere Bauelemente übertragen.

Bezogen auf den MOSFET lassen sich die Qualitätskriterien auf folgende Transistoreigenschaften zurückführen

1. Eine möglichst kleine Zahl von Prozessierungsschritten, insbesondere Masken- und Belichtungsschritten
2. Maximierung des Anstroms I_{on} des Transistors (d.h. für den Fall Gate- und Drainspannung auf auf Versorgungsspannung, Source- und Wannenspannung auf 0 Volt)
3. Minimierung des Ausschaltstrom I_{off} des Transistors (d.h. für den Fall Drainspannung auf Versorgung, alle anderen Anschlüsse auf 0 Volt)

Aus den Punkten 2. und 3. ergibt sich, daß die Kurve Anstrom gegen Ausstrom bestimmend für die Qualität eines Bauelements ist. Das grundsätzliche Vorgehen des Bauelementedesigns ist in Abbildung 1.6 dargestellt: Ausgehend von der

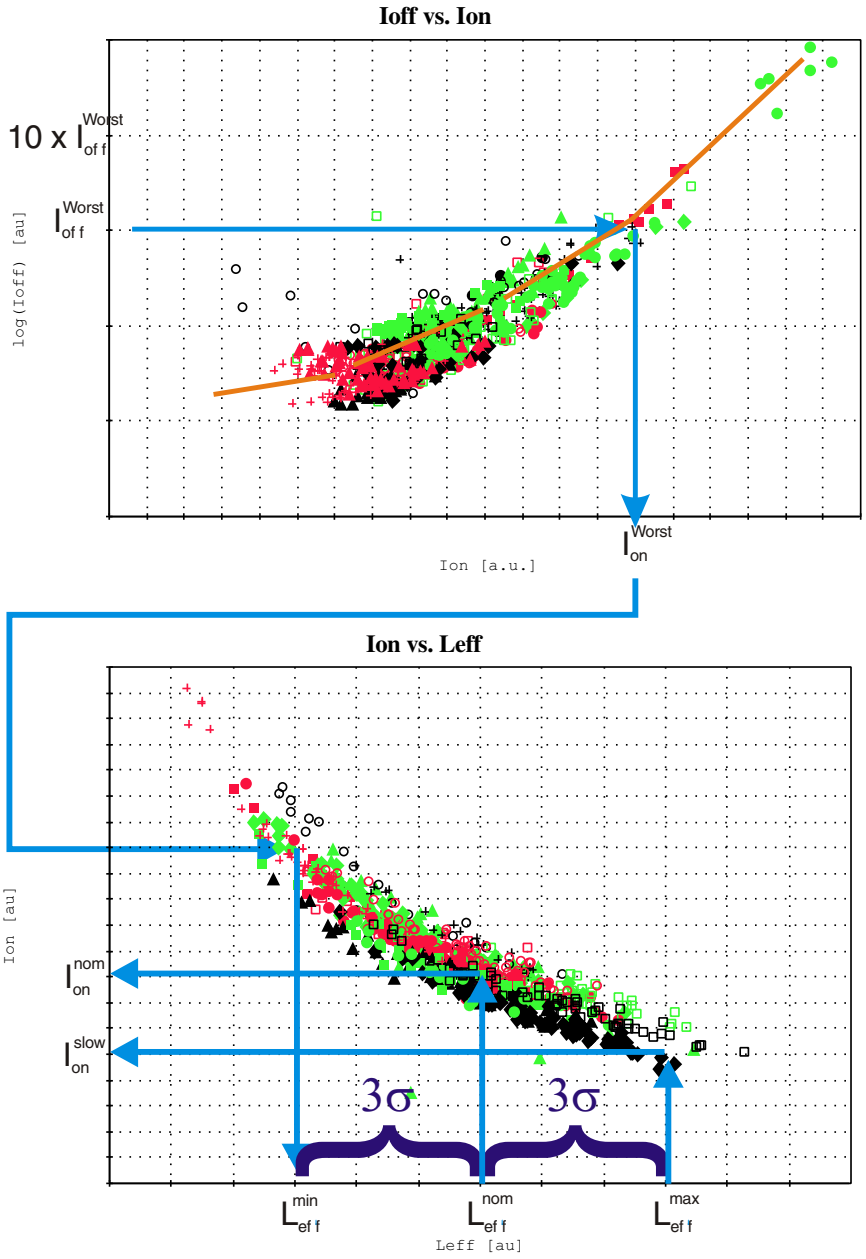


Abbildung 1.6: Grundsätzliches Vorgehen beim Bauelementedesign. Ausgehend vom tolerierbaren Worst-Case ergeben sich zusammen mit den Prozeßschwankungen der Nominalfall und der Fall mit geringstem Anstrom ("slow-case").

I_{on} zu I_{off} Kurve wird aus den Produktanforderungen der Punkt mit noch tolerablen Ausschaltstrom I_{off}^{Worst} ermittelt. Aus dem zugehörigen Anstrom wird in der Auftragung Anstrom gegen Kanallänge das gesamte Prozeßfenster zwischen I_{on}^{Worst} und I_{on}^{slow} ermittelt. In Abbildung 1.6 wurde hierbei eine Schwankung von 3σ angenommen. Für größere Schwankungen von 4σ und mehr ergeben sich entsprechend größere Prozeßfenster.

Aus Gleichung 1.2 ergibt sich, daß die Faktoren x_P und x_W um so größer gewählt werden könnten, je kleiner die Technologieschwankugen σ_P und σ_W sind.

Die Parameter σ_P und σ_W resultieren aus den Schwankungen der Bauelemente, aus denen die Schaltung konstruiert wird. Ziel des Bauelementedesigns ist es, eine hohe Nominalperformance P bzw. niedrige Verlustleistung W durch geschickte Prozessierung zu erzielen.

1.5.4 Schaltungsentwurf

Das Schaltungsdesign wird so durchgeführt, daß die Schaltung auch bei einer Performance von $P = -x_P \times \sigma_P$ funktioniert.

Bei vorgegebenen Schwankungen σ gilt es für den Schaltungsentwurf abzuwägen, ob man ein schwankungstolerantes Design für den Fall geringer Performance whlt, um einen möglichst hohen Yield zu erzielen, oder ein performancekritisches Design wählt, das hohen Ausschuß in der Fertigung zur Folge hat.

Dies ist nun der Startpunkt der folgenden Überlegungen: wie groß sollten x_P und x_W gewählt werden?

Als konkreter Anwendungsfall wird ein typisches digitales Schaltungselement, der Ringoszillator mit Invertern[12], untersucht. Als Kenngröße gilt hier die Verzögerung eines Invertergliedes τ_{stage}, die die Schwingungsfrequenz des Ringoszillators bestimmt. Eine Fertigungstechnologie ist um so performanter, je geringer diese Gatterlaufzeit τ ist.

In Abbildung 1.8 ist die Verteilung der Gatterverzgerung in einer 0.18μm Technologie aufgetragen. Das 3-Sigma Fenster über einige Hundert Wafer beträgt einige Pikosekunden.

Abbildung 1.7 zeigt eine Korrelation zwischen Ringoszillatorfrequenz und dem Anstrom eines Transistors. Je geringer der On-Strom des nMOS des Transistors desto geringer die Schaltgeschwindigkeit des Ringoszillators. Zusammen mit Abbildung 1.6 ergibt sich: je höher die Performance desto größer der Verluststrom, was ein Abwägen zwischen den Größen x_P und x_W erfordert.

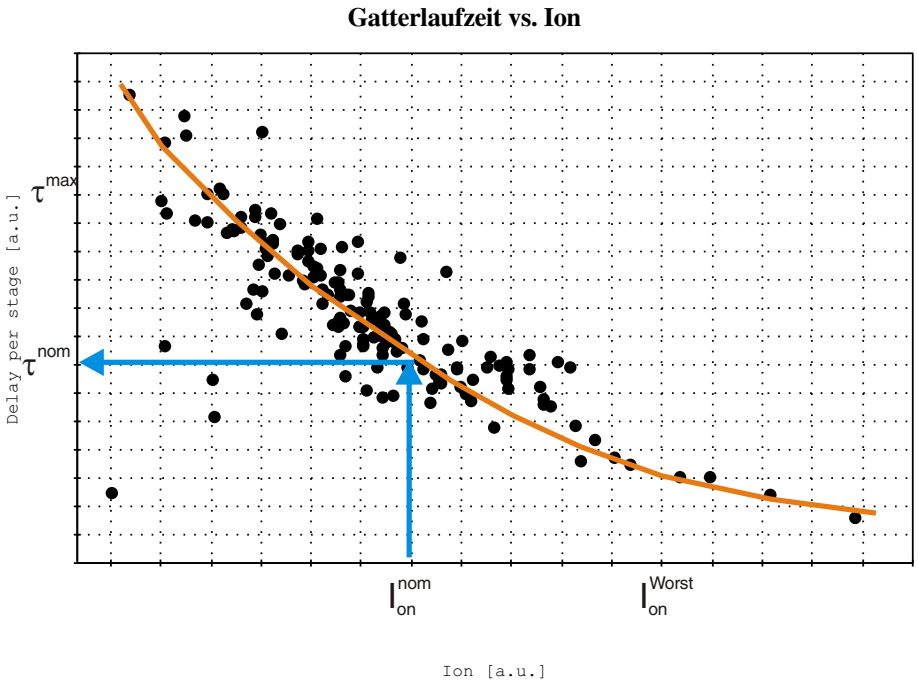


Abbildung 1.7: Abhängigkeit der Gatterlaufzeit von dem Anstrom I_{on} in einer 0.18μm Technologiegeneration.

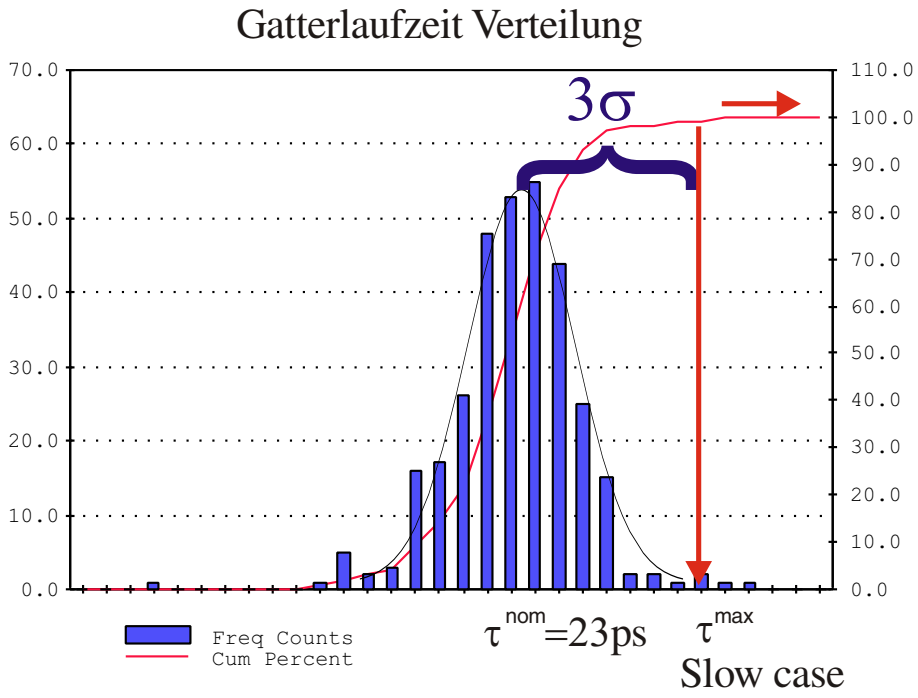


Abbildung 1.8: Statistische Verteilung der Gatterlaufzeit eines Inverterringoszillators über 10 Lose in einer 0.18μm Technologiegeneration.

Gleiches gilt auch für die Abwägung Yield zu Performance: wählt man ein schwankungstolerantes Design für große Toleranzfaktoren x, um einen hohen Yield zu erzielen, so muß die Schaltung auch bei kleinen Anströmen funktionieren, was eine kleine Performance bedeutet.

Damit ist ein Zusammenhang zwischen der Schwankung des Einzelprozesses und den Faktoren x zum Yield hergestellt. Im folgenden wird nun die Wahl des Toleranzfaktors x für den Einzelprozeß diskutiert.

1.5.5 Wahl des Schwankungsfensters für den Einzelprozeß

In den vorhergehenden Kapitel wurde der Zusammenhang zwischen Einzelprozeß- schwankung und Produktqualitätskriterien skizziert. Als tolerierbare Schwankung wurde in den vergangenen Kapiteln ein Wert von $\pm 3\sigma$, d.h. ein $x = 3$ in Gleichung 1.2, angenommen. Die Wahl der Parameter x_P und x_W erfolgt nach Abwägen von Yield gegen Performance.

- Für performancekritische Schaltungen wird man kleine Werte $x_P < 2$ wählen, dann aber nur einen geringen Yield erzielen.
- Für kostenkritische Schaltungen empfehlen sich Werte $x_P \approx 3$.
- Höhere Werte für x_P sind nur wenig sinnvoll, da die Verteilung der Prozeßparameter nur um den Nominalwert in guter Näherung einer Normalverteilung entspricht. Eine gaußförmige Korrelation zwischen $x_P \times \sigma_P$ und Yield gilt folglich nur für kleine x_P. Ausbeuten im parts-per-million Bereich mit großen x_P sind nicht erzielbar.

Für zahlreiche Einzelprozesse ist aber die Annahme einer Normalverteilung nicht gerechtfertigt. Vielmehr spielen hier systematische Abhängigkeiten, z.B. eine ortsabhängige Verteilung auf dem Wafer eine große Rolle. Auch ist der Zusammenhang zwischen der primären Prozeßgröse (Temperatur, Ätzrate, optische Auflösung etc.) und den Werten für Performance und Verlusstleistung nicht-linear und schwer zu ermitteln. Ein einfach ablesbarer Zusammenhang zwischen Yield und Schwankung ist dann nicht mehr gegeben. Ein Beispiel dafür ist z.B. der Zusammenhang zwischen elektrischer Kanallänge und Temperatur (Abbildung 1.9). Niedrigere Temperaturen führen zu unzureichender thermischer Aktivierung der Dopanden und damit zu deutlich größeren Kanallängen.

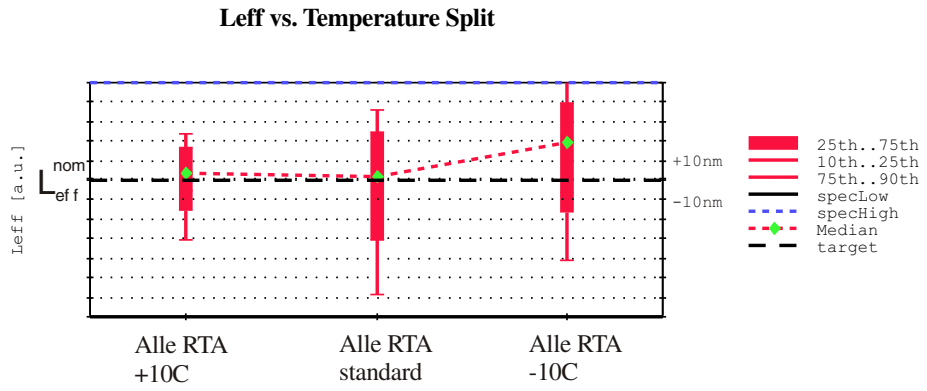


Abbildung 1.9: Veränderung der effektiven, elektrischen Kanallänge von nMOS Transistoren in einer 0.18μm Technologie bei einer gezielten Auslenkung der RTA-Temperaturen um $\pm$10°C.

In dieser Arbeit werden daher zwei Werte als Ergebnis der Optimierung angegeben: die Standardabweichung σ und die Spanne zwischen Minimal- und Maximalwert, die die systematischen Schwankungen beinhalten. Durch Angabe von Maximal- und Minimalwert läßt sich z.B. durch Prozeß-/Devicesimulation ermitteln, ob die Bauelemente und Schaltungen untern diesen Schwankungen noch funktionsfähig sind, bzw. welcher Prozentsatz des Wafers außerhalb des tolerablen Bereichs liegt.

1.5.6 Einfluß der Schwankungen von thermischen Prozessen auf das Bauelementverhalten

Als abschließende Betrachtung in diesem Kapitel soll ein Zusammenhang zwischen den Prozeßschwankungen der thermischen Prozessierungsschritte und den Schalteigenschaften von MOS-Transistoren hergestellt werden. Dazu wurden in der oben dargestellten 0.18μm Technologie alle RTA Schritte für Implantationsausheilungen nach Unterdiffusions- und pn-Übergangsimplantation um 10 Grad nach oben und unten ausgeheilt.

Die Ergebnisse sind den Abbildungen 1.10 und 1.11 dargestellt. Eine Schwankung von $\pm$10°C führt zu erheblichen Schwankungen der Transistorparameter. Daher läßt sich für diese Technologie eine Obergrenze von etwa $\pm$7°C angeben (in den Experimenten wurden *alle* RTA Schritte zugleich ausgelenkt, was außerhalb der Schwankungen eines Prozesses liegt).

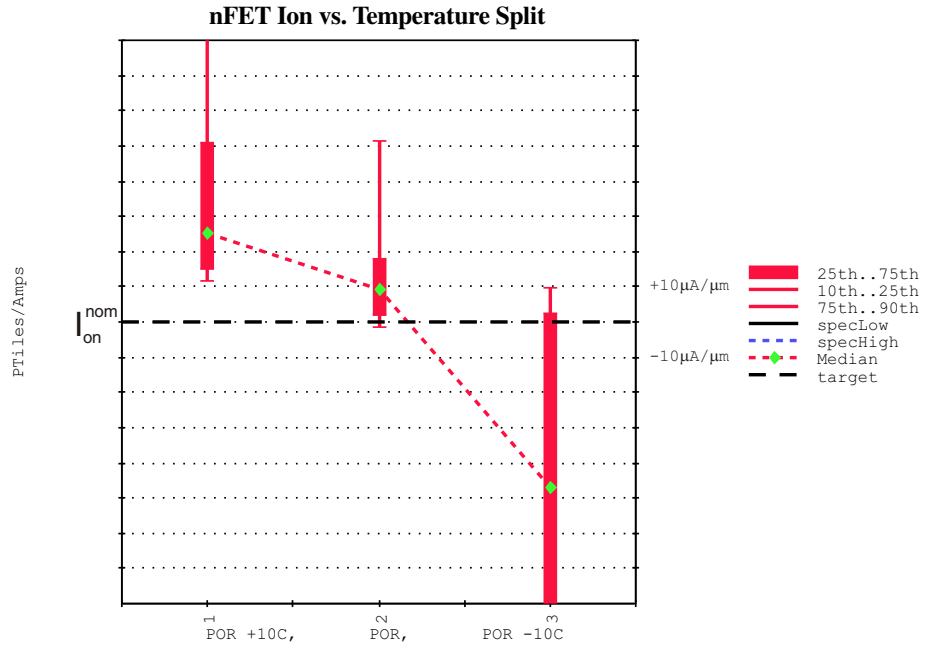


Abbildung 1.10: Veränderung des Anstroms I_{on} von nMOS Transistoren in einer 0.18μm Technologie bei einer gezielten Auslenkung der RTA-Temperaturen um $\pm 10°$C.

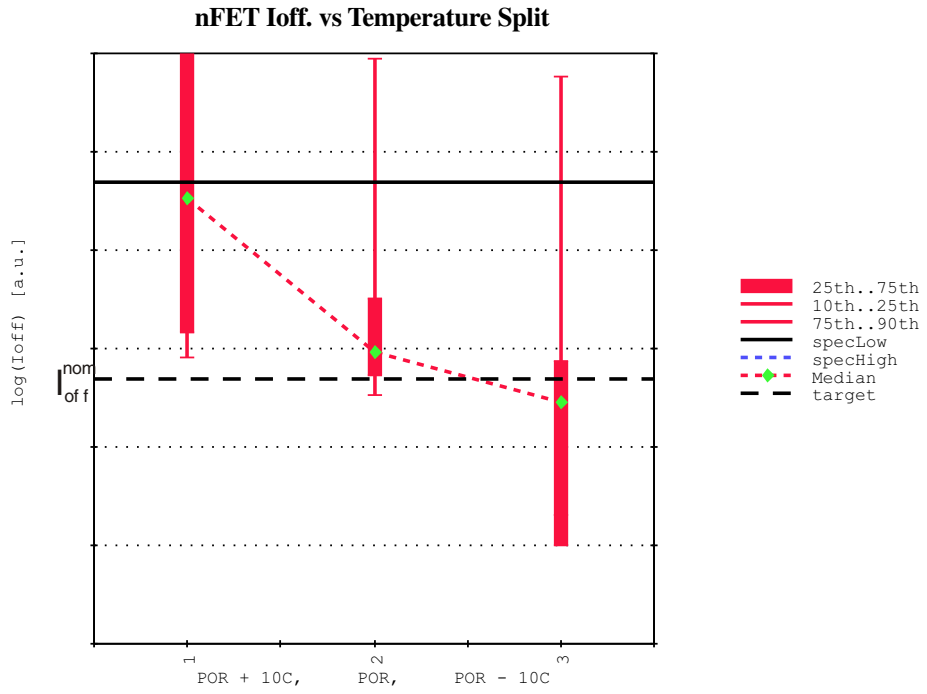


Abbildung 1.11: Veränderung des Ausschaltleckstroms I_{off} von nMOS Transistoren in einer 0.18μm Technologie bei einer gezielten Auslenkung der RTA-Temperaturen um $\pm$10°C.

Kapitel 2

Simulationsgestützte Optimierung einer RTO-Kammer

2.1 Problemstellung

Gegenstand dieser Arbeit ist die Optimierung des Verhaltens von Reaktoren in der Halbleiterherstellung mit Hilfe physikalischer Modelle. Im folgenden wird gezeigt, inwieweit die Simulation des Reaktorverhaltens Aufschluß über Schwachstellen im Reaktordesign bringen kann.

Die folgenden Untersuchungen werden für ein Rapid-Thermal-Oxidation (RTO) Prozeß für das Tunneloxid eines EEPROM-Floating-Gatestacks[1] durchgeführt. Bei diesem Prozeß sind die Anforderungen an die Uniformität des Oxids über den Wafer sehr hoch. Auf Basis der Simulationsergebnisse konnten hier Verbesserungen am Reaktordesign erzielt werden. Des weiteren dienen die Simulationen als Hilfsmittel für die Sensitivitätsanalyse und die Parametergewinnung für die spätere Optimierung und Regelung.

Der Simulation kommt aufgrund der beschränkten meßtechnischen Zugangsmöglichkeiten während des Prozesses eine große Bedeutung zu: Vor dem Einsatz weitergehender Optimierungs- und Regelansätze müssen die relevanten Einflußfaktoren auf das Prozeßergebnis identifiziert werden (hier vor allem das

[1] Bei **E**lectrically **E**rasable and **P**rogrammable **R**ead **O**nly **M**emory handelt es sich um Bausteine für nichtflüchtige Speicher, die z.B. in Chipkarten eingesetzt werden.

langsame Aufheizen der Quarze in der Kammer, sowie einer eventuell zu geringen Lampenleistung zum Erreichen der geforderten Temperaturuniformität und Konvektionsbewegungen in der Kammer).

Ein Versuch, die oben genannten Störfaktoren durch eine geschickte Regelungsstrategie zu kompensieren, schlägt fehl. So sind z.B. die Konvektionsbewegungen weder während des Prozesses meßbar noch ausreichend genau berechenbar. Konvektive Effekte müssen daher durch eine Veränderung der Prozeßbedingungen oder des Kammeraufbaus minimiert werden. Die Schwierigkeit des anschließenden Meß- und Optimierungsproblems wird so drastisch reduziert.

2.2 Beschreibung der RTO-Kammer und des Prozesses

Gegenstand der folgenden Simulationen ist eine RTO-Kammer in einer Cluster-Anlage (Abbildung 2.1) zur Prozessierung des Gate-Stacks für EEPROM-Zellen (Abbildung 2.2). Die vorliegende Ausführung arbeitet auf Basis eines nichtkontaktierten Gates, auf das durch Anlegen einer Spannung zwischen Source und Programmiergate Elektronen durch Fowler-Nordheim-Tunneln vom Leitungsband des Silizium in das Leitungsband des Tunneloxids gelangen und so auf dem floating gate dauerhaft gespeichert werden. Die Ladung auf dem floating gate beeinflußt nun die Schalteigenschaften des Transistors. Die Einhaltung einer vorgegebenen Oxiddicke ist notwendig, da der Potentialabfall über das Oxid und damit der Fowler-Nordheim-Strom von der Dicke des Tunneloxids abhängt. Die Fowler-Nordheim-Stromdichte ergibt sich zu [28]

$$J_{FN} = A_{FN}\mathcal{E}_x^2 \exp(-\frac{B_{FN}}{\mathcal{E}_x}) \tag{2.1}$$

mit dem elektrischen Feld

$$\mathcal{E}_x \approx \frac{U_{floatox}}{x_{Ox}}, \tag{2.2}$$

der Oxiddicke x_{ox}, dem Potentialabfall über das Oxid $U_{floatox}$ und den Materialkonstanten A_{FN} und B_{FN}, die von der Potentialdifferenz zwischen Oxid und Silizium und den effektiven Massen abhängen.

Der Aufbau der Kammer ist in Abbildung 2.3 dargestellt. Der Wafer wird in der axialsymmterischen, polierten Stahlkammer durch zwei Lampenhäuser mit Goldreflektoren beheizt. Zum einen durch 6 Ringe im oberen Lampenhaus mit

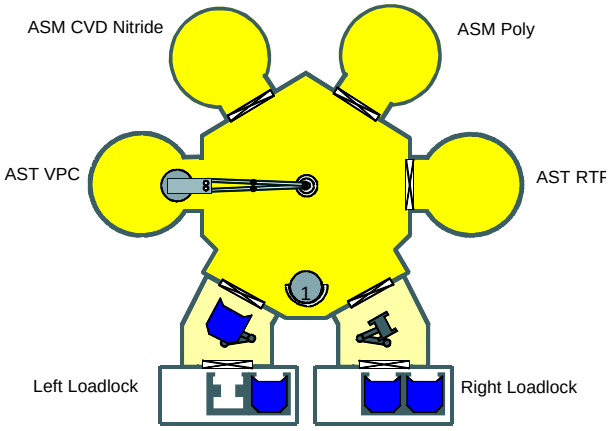


Abbildung 2.1: Clusteranlage zur Herstellung des EEPROM-Gatestacks in 0.5μ Technologie mit HF-Reinigungsmodul, Nitridier-, Polyabscheidungs- und Oxidationskammer (von links).

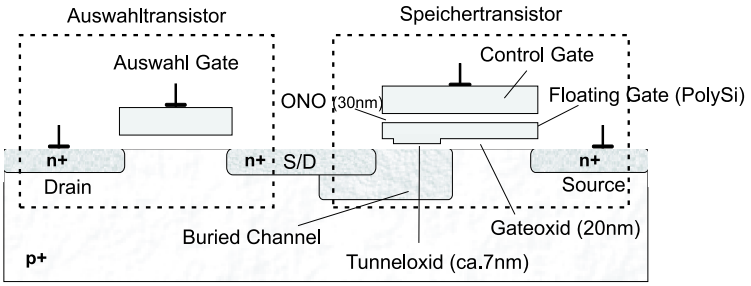


Abbildung 2.2: Das Cluster-Tool wird für die Produktion von EEPROM Zellen eingesetzt. Das Tunneloxid des Programmiertransistor weist eine Dicke von etwa 7nm auf.

etwa 100 Einzelhalogenstrahlern mit je 250 Watt und 16 bzw. 18 Stabhalogenlampen á 1.5kW. Der Wafer wird durch eine Türöffnung mit ungefähr 45° Öffnungswinkel in die Kammer gebracht und auf Quarzpins abgelegt. Die Reaktorkammer ist von den Lampenhäusern durch zwei, etwa 2cm dicke Quarzplatten separiert, die von der Lampenhausseite luftgekühlt werden. Die Wafertemperatur wird pyrometrisch während des Prozesses an bis zu drei Stellen gemessen. Der im folgende betrachtete Prozeß ist eine Oxidation in reinem Sauerstoff bei Drücken zwischen 250 und 760 Torr und Temperaturen zwischen 1000°C und 1150°C.

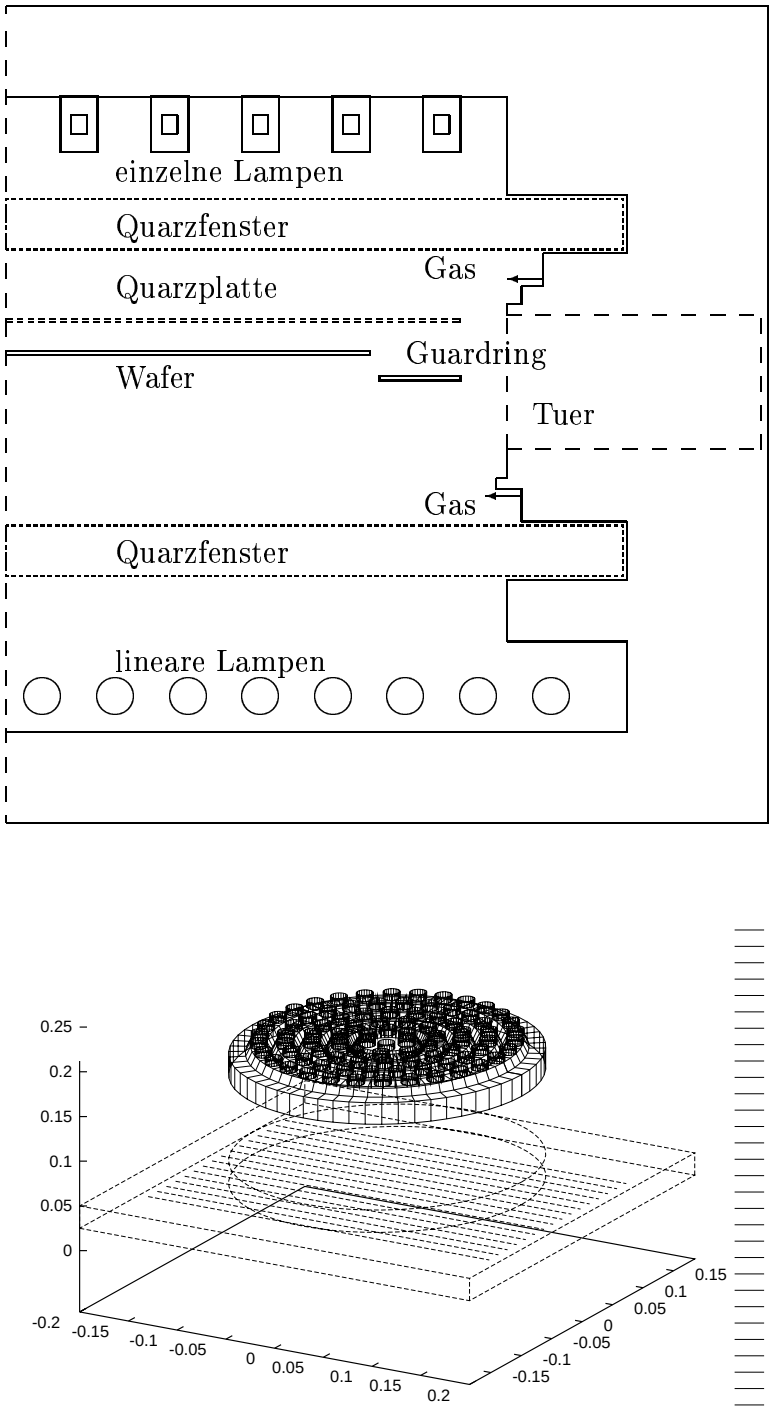


Abbildung 2.3: Schema des Reaktors im Querschnitt (oben). Die Scheibe wird von zwei Lampenhäusern (unten) beheizt. An beiden Seiten der Kammer befinden sich jeweils Öffnungen für Wafereinlaß und Pumpport, so daß die Kammer nicht axialsymmetrisch ist.

2.3 Formulierung der physikalischen Gleichungen

In den folgenden Unterabschnitten werden die grundlegenden physikalischen Energietransportgleichungen diskutiert. Sie dienen als Grundlage für die sich anschließende Regelbarkeits- und Stabilitätsanalyse.

2.3.1 Der Equipmentsimulator PHOENICS/CVD

Als Simulationswerkzeug wurde in dieser Arbeit der Fluiddynamik- und Strahlungssimulator PHOENICS/CVD [29] verwendet. PHOENICS/CVD ist das in der Verfahrenssimulation derzeit am weitesten entwickelte, kommerziell erhältliche Werkzeug. Es umfaßt iterative Lösungsverfahren für die zwei- und drei-

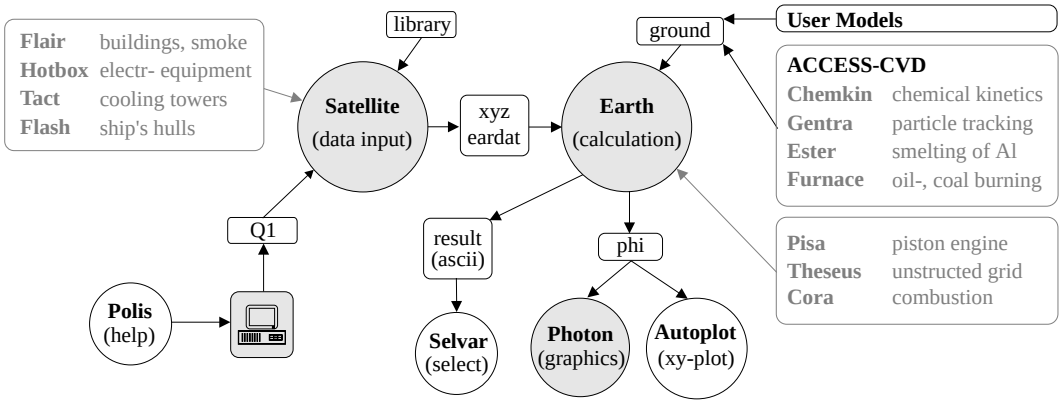


Abbildung 2.4: Aufbau des Simulatorprogramms PHOENICS. Vom Anwender können Modell für Transportkoeffizienten und Quellterme an den Gauß-Seidel-Solver angekoppelt werden.

dimensionale Berechnung von Wärme- und Strahlungstransport, chemischer Gasphasen- und Oberflächenreaktionen über eine Chemiedatenbank für Prozeßschritte wie Silan- und TEOS-Abscheidung, laminare und turbulente Strömungen in einem weiten Druckbereich, Mischungsabhängigkeit der thermodynamischen Transportgrößen, Modelle für Knudsen-Transport über Spalte wie z.B. bei Suszeptorauflagen. Dies wurde in einem Vergleichstest in bezug auf Rechenzeit und Stabilität zwischen verschiedenen Simulatoren bestätigt [30]. Die hier verwendete Version ist gegenüber den Benchmark-Ergebnissen, vor allem bei der Strahlungs- und Chemiemodellierung, noch deutlich erweitert worden.

2.3.2 Strömungsmodellierung

Für die Modellierung der RTO-Prozesse können die folgenden Annahmen gemacht werden:

1. Da die freie Weglänge der Moleküle ℓ viel kleiner als die charakteristische Reaktordimension $\mathcal{L}$ ist, d.h. die Knudsen-Zahl $Kn := \ell/\mathcal{L}$ ist klein gegen 1, kann das Gas als kontinuierliches Medium betrachtet werden. Bei Drücken um 50 Torr und Temperaturen von 700 K ist die freie Weglänge der CVD-Gase von der Größenordnung $\geq 10^{-5}$m, während die Reaktorgröße 10^{-2}m nicht unterschreitet. Damit genügt eine hydrodynamische Näherung.

2. Die Gase können näherungsweise als ideale Gase betrachtet werden. Die Transportgrößen des Gases werden dabei temperaturabhängig behandelt, Innere Reibungsterme werden durch temperatur- und druckabhängige thermodynamische Größen, wie z.B. Wärmekapazität und Leitfähigkeit, berücksichtigt.

3. Bei den Gasflüssen mit der typischen Gasgeschwindgkeit V und der Viskosität ν ist die zugehörige Reynolds-Zahl $\mathrm{Re} := V\mathcal{L}/\nu \approx 10^{-2} - 10^{2}$ klein. Diese liegt unter dem turbulenten Strömungsregime, das etwa bei $\mathrm{Re} = 2 \times 10^{3}$ beginnt[31].

4. Das Gas unterliegt keiner Strahlungswechselwirkung. (Die infrarote Absorptivität bei manchen reaktiven Gasen wird also nicht berücksichtigt).

Damit lassen sich die grundlegenden Gleichungen formulieren:
Kontinuitätsgleichung

$$\frac{\partial \rho}{\partial t} = -\boldsymbol{\nabla} \cdot (\rho \mathbf{v}) \tag{2.3}$$

Impulserhaltung (Navier-Stokes)

$$\underbrace{\frac{\partial \rho \mathbf{v}}{\partial t}}_{\text{Impulsänd./Volumen}} = -\underbrace{\boldsymbol{\nabla} \cdot (\rho \mathbf{v}\mathbf{v})}_{\text{Konvektion}} + \underbrace{\boldsymbol{\nabla} \cdot \overleftrightarrow{\boldsymbol{\tau}}}_{\text{Viskosität}} - \underbrace{\boldsymbol{\nabla} p}_{\text{Druckgradient}} + \underbrace{\rho\, \mathbf{g}}_{\text{Gravitation}} \tag{2.4}$$

mit dem Viskositätstensor für newtonsche Flüssigkeiten

$$\overleftrightarrow{\boldsymbol{\tau}} = \mu(\nabla v + (\nabla v)^T) + (\kappa - \frac{2}{3}\mu)(\boldsymbol{\nabla} \cdot \mathbf{v}) \cdot \overleftrightarrow{1} \tag{2.5}$$

wobei $\mathbf{g}$ den Gravitationsvektor, μ in $\frac{\mathrm{m}^2}{\mathrm{s}}$ die dynamische und κ die Volumenviskosität bezeichnet. Die Volumenviskosität κ verschwindet für monoatomare Gase bei geringer Dichte und ist klein bei dichten Gasen [38, S. 503f].

Zentraler Bestandteil der folgenden Betrachtungen ist die Energiegleichung des Gassystems:

$$\underbrace{c_p \frac{\partial \rho T}{\partial t}}_{\substack{\text{Energiegewinn} \\ \text{pro Volumen}}} = \frac{Dp}{Dt} - \underbrace{c_p \boldsymbol{\nabla} \cdot (\rho T \mathbf{v})}_{\substack{\text{Konvektion} \\ \text{Kompression}}} + \underbrace{\boldsymbol{\nabla} \cdot (\lambda \boldsymbol{\nabla} T)}_{\text{Konduktion}} + \underbrace{\mu \Phi_v}_{\substack{\text{Viskosität} \\ = \overleftrightarrow{\tau} : \boldsymbol{\nabla}_{\mathbf{v}}}} \tag{2.6}$$

mit der Dissipationsfunktion Φ_v, die die (irreversible) Umwandlung in innere Energie beschreibt. Ihr Wert ist aber für CVD bedeutsame Strömungsgeschwindigkeiten klein.

Spielen ferner Mehrphasensysteme und chemische Reaktionen eine Rolle, so treten weitere Ausdrücke wie Dufour- und Interdiffusionsterme hinzu [31, S. 566ff.]. Der Einfluß auf die Wafertemperatur wird bei den hier betrachteten Prozessen als vernachlässigbar angenommen.

2.3.3 Strahlungsmodellierung

Folgende Annahmen werden für den Strahlungstransport in den folgenden Simulationen gemacht:

1. Polarisations- und Photonenstreueffekte, sowie die endliche Lichtgeschwindigkeit werden nicht berücksichtigt
2. Die Kohärenzlänge des Lichts ist klein gegenüber den Ausmessungen der Kammer, aber groß gegenüber den Schichtdicken auf dem Wafer.
3. Die Strahlungstransport zwischen den strahlenden Körpern kann durch einen effektiven Transport zwischen den Oberflächen ersetzt werden. Vielfachreflexionen werden durch die McMahon-Approximation beschrieben [34].
4. Das Gas in der Kammer absorbiert nicht. Für alle in dieser Arbeit betrachteten Prozesse ist diese Annahme erfüllt.

Ausgehend von der Planck'schen Energieverteilung erhält man für die spektrale Intensitätsverteilung eines schwarzen Körpers der Temperatur $T(x)$

$$I^{\mathrm{SK}}_{\nu,T}(x, \vec{k}) = \frac{2h}{c^2} \left| \vec{n}(x) \cdot \vec{k} \right| \frac{\nu^3}{e^{h\nu/k_B T(x)} - 1} \tag{2.7}$$

mit der Frequenz ν, dem Normalvektor auf die Oberfläche $n(x)$ und dem Richtungsvektor $\vec{k}$ der Strahlen. Die Leuchtdichte eines schwarzen Strahlers ist unabhängig von der Beobachtungsrichtung, woraus sich die Lambert'sche Kosinusverteilung von I ergibt; die Strahlungsdichte für nicht-ideale Lambert'sche Strahler weist bei Nichtmetallen zumeist eine Erhöhung der Verteilung in Normalrichtung auf.

Daraus erhält man die Energiedichte, die von einer Oberfläche Ω_+ abgegeben wird

$$\int_0^\infty \int_{\Omega^+} I_{\nu,T}^{SK}(x,\vec{k}) d\vec{k} d\nu = \sigma T(x)^4 \tag{2.8}$$

Der Transport zwischen den Festkörpern wird auf den Strahlungstransport zwischen den Oberflächen reduziert. Eine Ableitung der Transportkoeffizienten ist [34] zu entnehmen. Alle Festkörper seien bereits numerisch diskretisiert. Dann läßt sich der Austausch über Strahlung als Lösung eines Transportproblems beschreiben.

Die zugrundeliegende Transportgleichung ergibt sich zu

$$\ddot{a} I_\nu(x,\vec{k}) = \Theta(n(x)\cdot\omega) e_{\nu,T}(x,\vec{k}) I_{\nu,T}^{\mathrm{SK}}(x,\vec{k}) + \int_\Omega \zeta_{\nu,T}(x,\vec{k}' \to \vec{k}) I_\nu(x,\vec{k}') d\vec{k}' \tag{2.9}$$

mit der Transportfunktion

$$\zeta_{\nu,T}(x,\vec{k}' \to \vec{k}) := \begin{cases} \rho_{\nu,T}(x,\vec{k}' \to \vec{k}), & \text{für } \vec{k}' \in \Omega^- \text{Reflektivität} \\ \tau_{\nu,T}(x,\vec{k}' \to \vec{k}), & \vec{k}' \in \Omega^+ \text{Transmissivität} \end{cases} \tag{2.10}$$

und der von der Frequenz ν, der Temperatur T und dem Richtungsvektor $\vec{k}$ abhängigen Emissivität e, sowie der Heavysidefunktion Θ, die die Abstrahlung nur auf die Richtung des äußeren Normalenvektors beschränkt. Es existieren verschiedene Verfahren zur Lösung von Gleichung 2.9. Die Approximation diffuser Reflexionen führt auf das Viewfaktorverfahren. Dieses Verfahren kann allerdings nur bedingt semitransparente Materialien und spiegelnde Oberflächen berücksichtigen. Exakt diese Bedingungen liegen aber in thermischen Reaktoren vor. Daher basiert diese Arbeit auf dem Monte-Carlo Lösungsverfahren [34].

In der McMahon Approximation ergibt sich für die frequenzabhängige Intensitätsverteilung auf der Oberfläche

$$I_\nu(x,\omega) = \Theta_x^\omega e_{\nu,T}(x,\omega) \left| n(x)\cdot\omega \right| c i_{\nu,T}^{SK}(x) + \int_\Omega \zeta_{\nu,T}(x,\vec{w}' \to \vec{w}) I_\nu(x - s\omega', \omega') d\omega' \tag{2.11}$$

Die emittierte Intensität $E(x)$ ergibt sich als Integral des ersten Summanden über den Raumwinkel ω und über das ganze Frequenzspektrum ν. Die absorbierte Strahlung, die von einer anderen Fläche D_j ausgesandt wird und am Ort x auf der Fläche D_i absorbiert wird, ergibt sich zu

$$A(x) = \int_0^\infty \int_{\Omega_y} \delta(x - X(y,\omega'))a_{\nu.T}(x,\omega\prime)\,|n(x)\cdot\omega|\,ci_\nu(y,\omega\prime)d\omega\prime dy \tag{2.12}$$

Dabei enthält $X(y,\omega\prime)$ das Ziel eines Strahles aus Richtung y mit Richtung ω' und enthält damit auch die Sichtblockierung zwischen der Zielfläche und der Ursprungsfläche D. In einer numerischen Diskretisierung läßt sich somit ein Strahlungsaustausch zwischen zwei Flächen definieren. Hierbei ist zu beachten, daß im Integranden $i_\nu(y,\omega\prime)$ steht und nicht die Graukörperradianz. Man summiert daher über alle Vielfachreflexionen auf.

Die gesamte absorbierte Strahlung eines Oberflächenelements i, die vom Flächenelement j ausgesendet wird, ist schließlich

$$K_{T^*}(i,j) = \overbrace{\int_{D_i} dx \int_{D_j} dy}^{\text{gemittelt über Oberflächen}} \quad \overbrace{\int_0^\infty d\nu}^{\text{spektral gemittelt}} \quad \overbrace{\sum_{k=0}^{\infty}}^{\text{Vielfach-reflektionen}} \overbrace{\int_{[\Omega]^{k+1}} d\omega_k \ldots d\omega_0 \int_{[D]^k} dy_k \ldots dy_1}^{\text{Mittel über Zwischenflächen}} \times$$

$$\times \quad \underbrace{B^{(k)}(x,y,y_1,\ldots,y_k,\omega_0,\omega_1,\ldots,\omega_k,\nu)}_{\text{Wahrscheinchlichkeit des Weges}} \quad \underbrace{e_{\nu,T^*}(y) I^{SK}_{\nu,T^*}(y)\frac{\cos\theta}{\pi}}_{\text{Wahrscheinlichkeit für Emission}} \tag{2.13}$$

Dabei ist T^* ein Temperaturarbeitspunkt, der zur Berechnung der optischen Eigenschaften verwendet wird.

Im folgenden wird häufig der Begriff der Strahlungkopplungsmatrix verwendet. Dabei wird die Gebhard-Matrix definiert als

$$G_{T^*}(i,j) = \frac{1}{e^{eff}_{T^*_j}\sigma\left(T^*_j\right)^4} K_{T^*}(i,j) \tag{2.14}$$

Die komplette Strahlungsflußmatrix ergibt sich schließlich zu

$$\Phi_{T^*}(i,j) = G_{T^*}(i,j) - \delta_{ij} \tag{2.15}$$

Für die Berechnung des eigentlichen Strahlungsaustausches als Quellterm in der Wärmetransportgleichung wird dann mit der tatsächlichen Emissivität und

Temperatur multipliziert. Dies setzt voraus, daß man sich bei der Berechnung von $G_{T^*}(i,j)$ nahe an der endgültigen Temperaturverteilung befindet.

Im folgenden bezeichnet die Kopplungsmatrix Lampen $\Rightarrow$ Wafer die Matrix der geometrischen Sichtfaktoren jedes Gitterelements der Waferoberfläche zu jeder Lampe; die Waferstrahlungsmatrix bezeichnet die Strahlungsflußmatrix der Waferelemente untereinander. Sonstige Kopplungsmatrizen zwischen zwei Festkörpern A und B bezeichnen die Kopplung aller Oberflächenelemente von A zu allen Oberflächenelementen von B.

Für die Richtungsverteilung der Lampen wurde eine an Experimente angepaßte Strahlungsverteilung verwendet. Ein Anpassen der Ausgangsverteilung für die Monte Carlo Simulation anstelle eine direkten Modellierung der Filamente ist aus mehreren Gründen sinnvoll. Zum einen liegt dies an der Verwendung von Quarzbirnen in den Kammern, die einen kaum modellierbaren Abschmelzpunkt – verursacht durch die Lampenherstellung – aufweisen. Zum anderen ist ein detailliertes Auflösen der Lampenwendel und des Lampenglases rechentechnisch aufwendig und mit der nötigen Detailtreue nahezu unmöglich. Daher wird in den folgenden Simulationen ein an Messungen angepaßtes Abstrahlungsprofil der Wendel angenommen (Bild 2.5). Die in den Monte Carlo Rechnungen verwendete räumliche Verteilung (nach A. Kersch) zeigt eine sehr gute Übereinstimmung mit Intensitätsmessungen.

Es bleibt die Frage zu klären, ob entlang der Wendel, insbesondere bei Halogenstablampen, ein nennenswerter Temperaturabfall auftritt, oder ob das Filament mit konstanter Temperatur angenommen werden kann. Zu diesem Zweck wurde das Filament getrennt simuliert (Abbildung 2.6). Die Abkühlung am Rand der Wendel wird zum kleineren Teil durch Wärmeableitung durch den Anschlußdraht, zum größeren Teil durch die Abstrahlung der äußersten Wicklung aufgrund der fehlenden Nachbarwicklung verursacht. Allerdings ist der Temperaturabfall auf ein sehr kleines Segment des Filaments beschränkt, so daß der Temperaturgradient in den folgenden Rechnungen vernachlässigt werden kann.

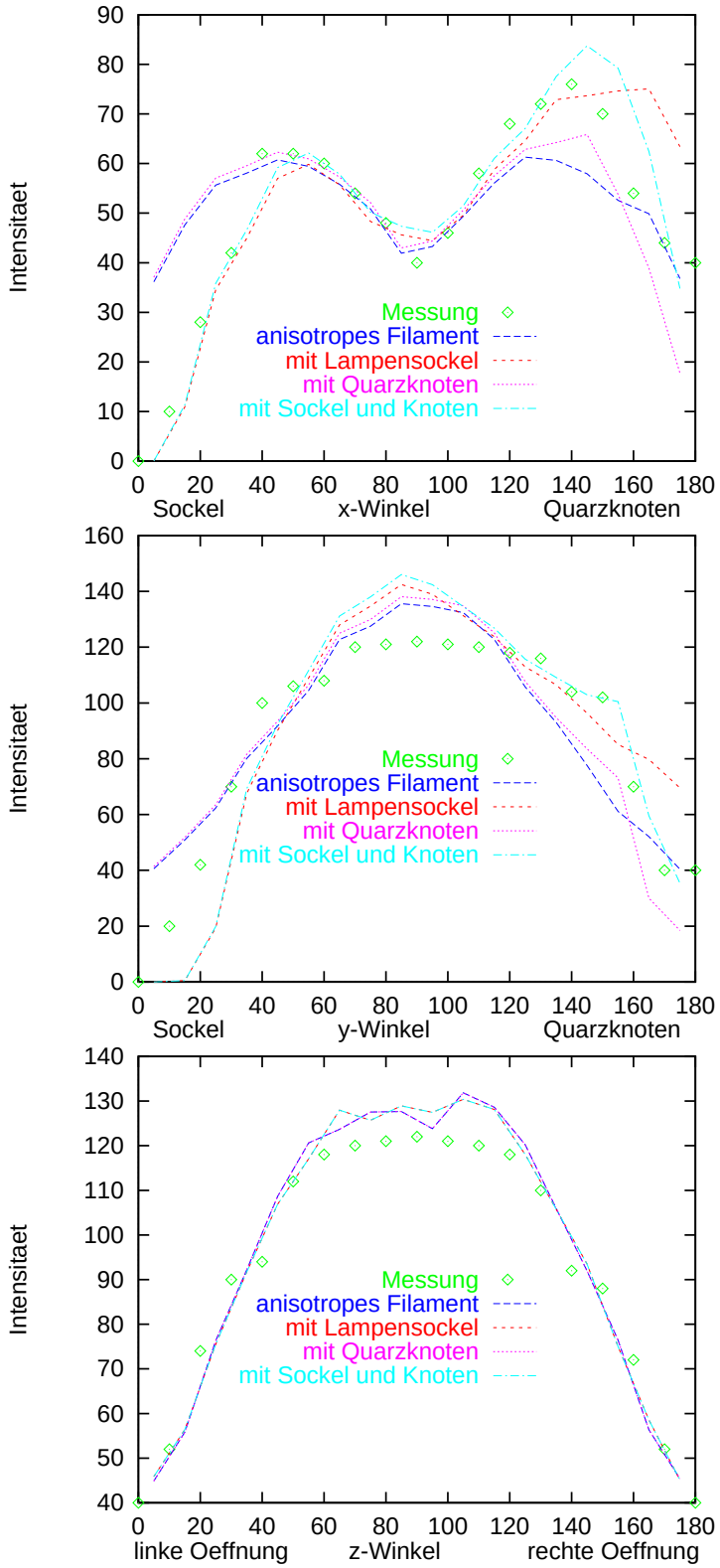
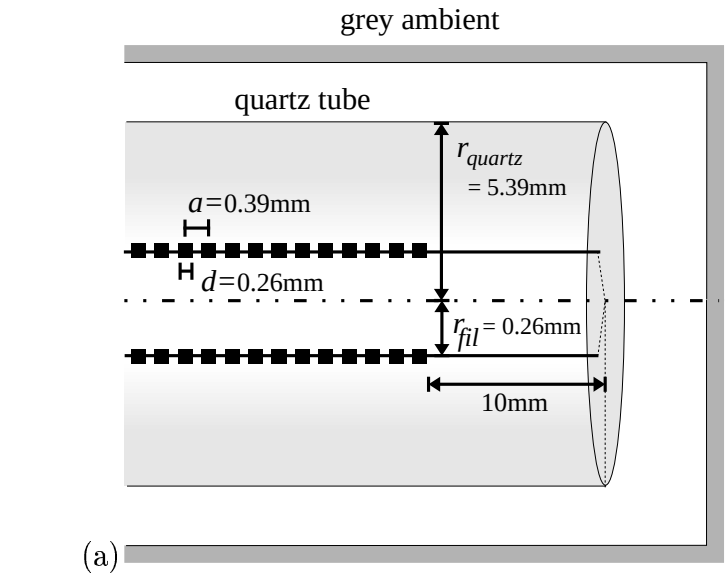
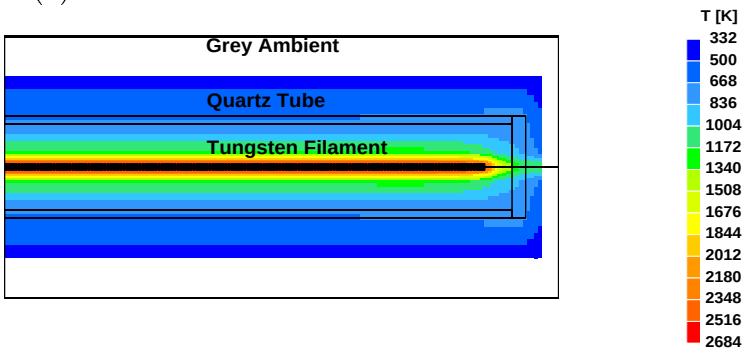


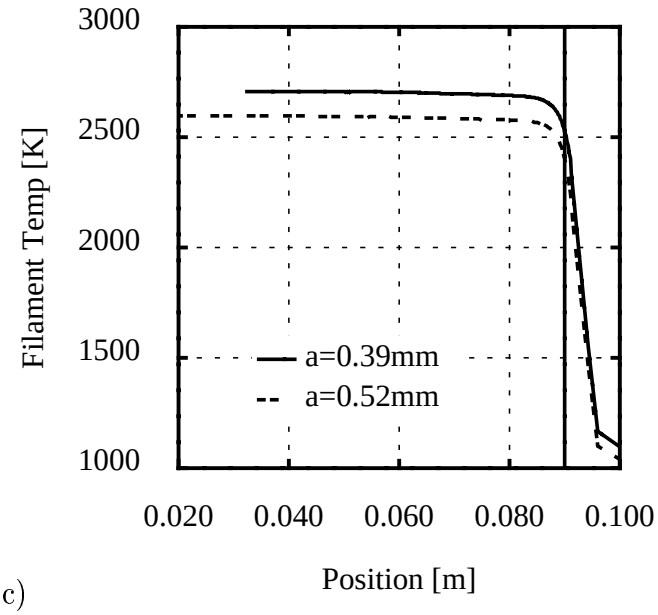
Abbildung 2.5: Intensitätsverteilung der in der Kammer (Abschnitt 3) verwendeten Halogenpunktstrahler, dargestellt in drei Schnitten. Gut erkennbar sind die Charakteristik des Abschmelzpunktes und das Minimum beim "Hindurchschauen" durch die Wendel.[26]



(a)



(b)



(c)

Abbildung 2.6: Simulation der Temperaturverteilung einer Halogenstablampe. (a) Aufbau in der Simulation, (b) Temperaturverteilung, (c) Temperaturprofil des Filaments.

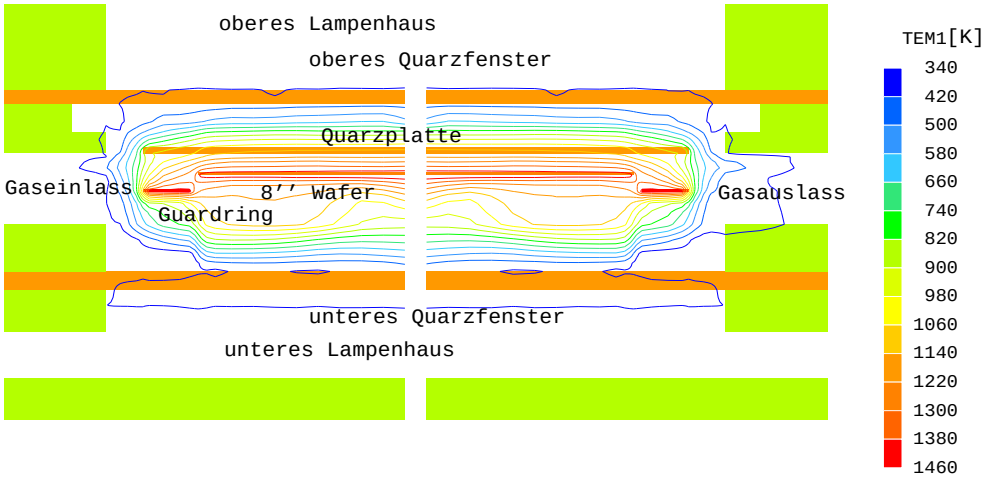


Abbildung 2.7: Temperaturverteilung in der Kammer nach einer Aufheizzeit von 120s. Mit dünner Quarzplatte.

2.4 Sensitivität der Wafertemperatur auf die Prozeßbedingungen

2.4.1 Temperaturverteilung in der Kammer

Abbildung (2.7) zeigt die simulierte Temperaturverteilung in der Kammer am Ende einer dreißigsekündigen Aufheizung in Vakuum.

Der Wafer hat seine Endtemperatur nahezu erreicht, während oberes und unteres Quarzfenster noch nicht aufgeheizt sind. In der Abbildung ist des weiteren eine dünne Quarzplatte erkennbar, die, wie unten gezeigt wird, zur Verminderung von Konvektionseffekten eingesetzt wird.

Der positive Einfluß des sogenannten Guardrings – eines dünnen Siliziumkarbidrings um den Wafer – wird in der Abbildung deutlich: der radiale Gradient des Temperaturfeldes wird abgeschwächt, so daß der konduktive Wärmeverlust auf der Waferfläche durch den vertikalen Gradienten bestimmt wird.

2.4.2 Temperaturuniformität auf der Scheibe während des Aufheizens

Die zeitliche Veränderung der Temperatur des Wafers während des Betriebs findet auf zwei verschiedenen Zeitskalen statt. Die rasche Aufheizung und

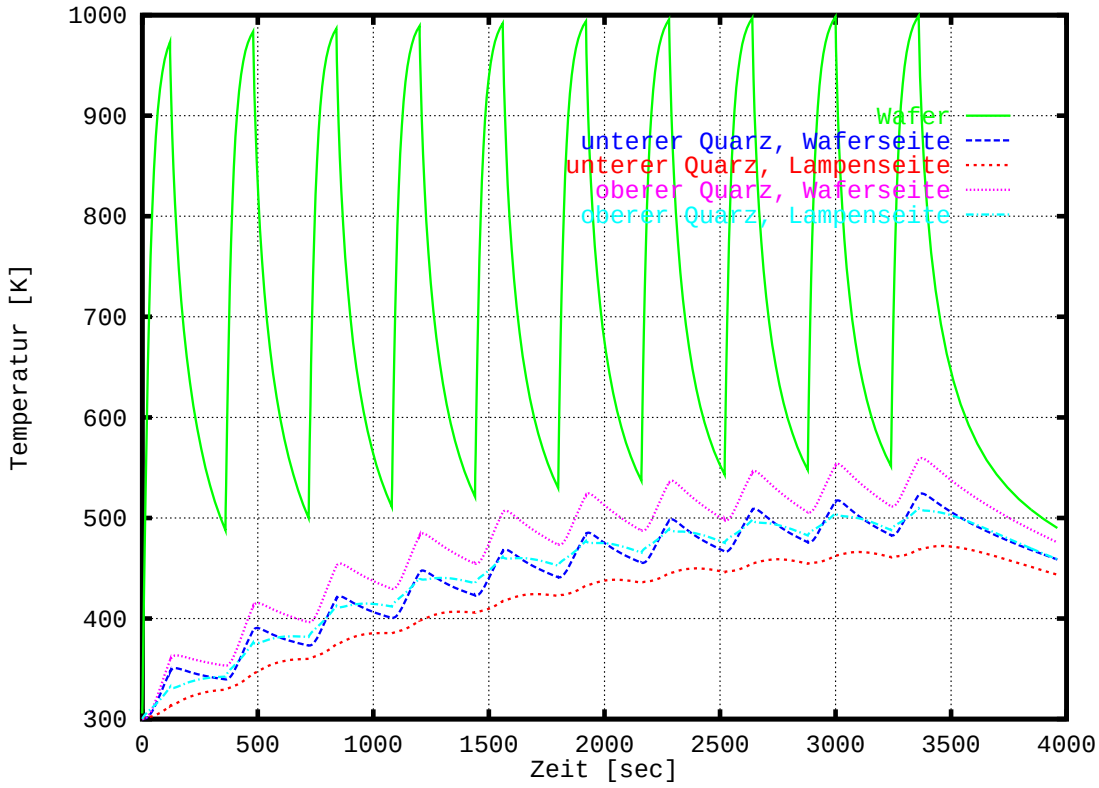


Abbildung 2.8: Zyklische Prozessierung von mehreren Wafern (Losprozessierung) mit konstanter Maximalleistung.

Abkühlung werden durch die Wärmekapazität des Wafers und die Strahlungsankopplung an Lampen und Reaktorwände bestimmt. Die typische Zeitskala liegt hierbei im Bereich von wenigen Sekunden.

Während eines längeren Prozesses oder aufgrund der Prozessierung mehrerer Wafer in Folge heizt sich auch die Reaktorumgebung auf. Dies gilt insbesondere für die Quarzfenster. Die Zeitskala der daraus resultierenden Drift der Wafertemperatur liegt im Bereich von mehreren Minuten.

In Abbildung 2.8 ist die Simulation eines zyklischen Aufheizen eines Wafers dargestellt. Für jeweils 120 Sekunden werden die Lampen mit konstanter Leistung angeschaltet, so daß sich der Wafer auf etwa 1000°C aufheizt; danach werden die Lampen für 120 Sekunden abgeschaltet. In Abbildung 2.9 ist die Differenz zwischen Wafermitte und Waferrand für das zyklische Aufheizen mit konstantem Rezept dargestellt. Der Unterschied zwischen dem ersten und zehnten Zyklus in der Temperaturverteilung im stationären Zustand liegt bei 5 Grad, da die Wafermitte vom aufgeheizten Quarz mehr beheizt wird als der Rand.

Abbildung 2.9 zeigt ferner die gute Grundhomogenität der Kammer mit nur

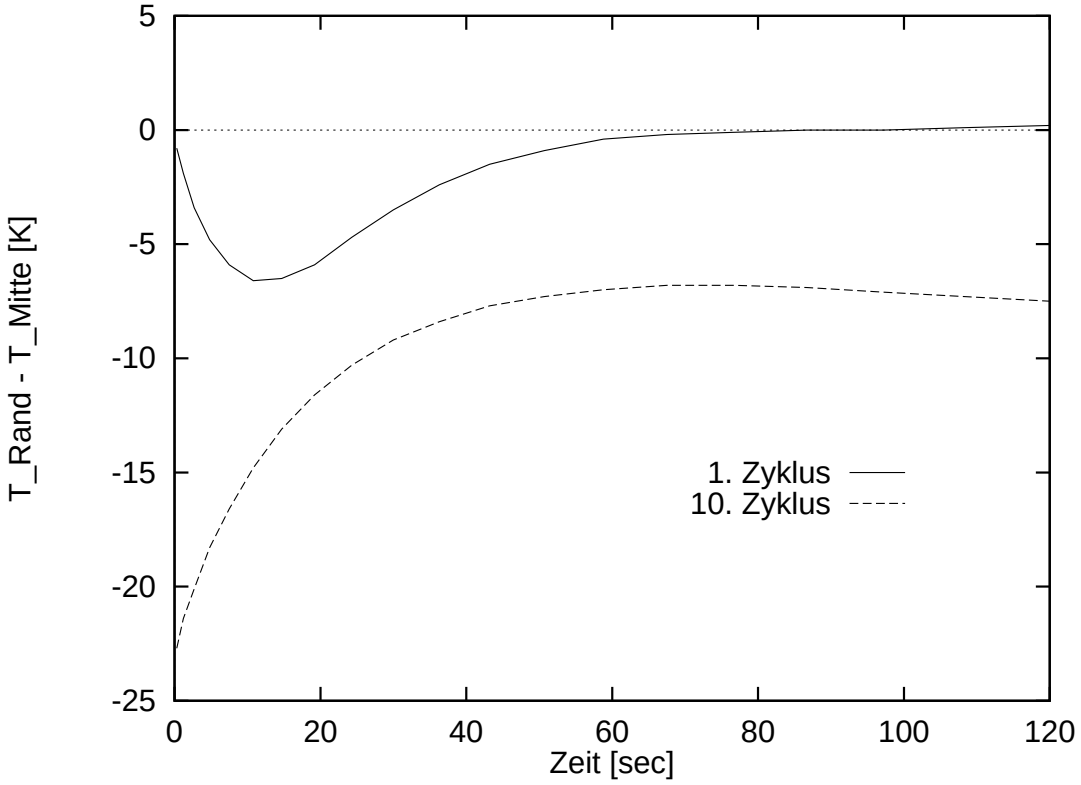


Abbildung 2.9: Temperaturinhomogenität während des Aufheizens für den ersten und zehnten Wafer bei der Losprozessierung

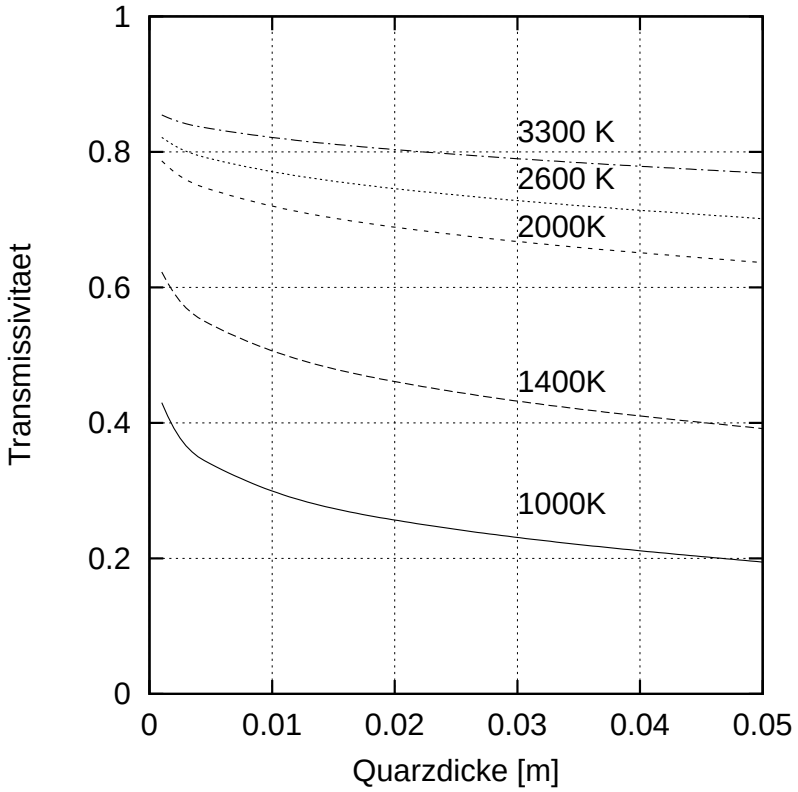


Abbildung 2.10: Transmissivität durch ein Quarzfenster bei verschiedenen Schwarzkörperstrahlungen.

geringer Mitte-Rand Differenz während des Aufheizens. Einer der Hauptgründe dafür ist die starke Absorptivität der Quarze im infraroten Bereich. So gelangt nur ein geringer Teil der emittierten Waferstrahlung zurück auf den Wafer und führt zu den in Metallkammern beobachteten Inhomogenitäten [37]. Dennoch bedarf es einer Korrektur der Lampenleistungen, um ein von der Waferposition im Los unabhängiges Prozeßergebnis zu erzielen.

2.4.3 Abhängigkeit von der Kammerreflektivität

Die Wafertemperatur zeigt eine starke Abhängigkeit von der Kammerreflektivität. Zwei Mechanismen sind dafür verantwortlich. Zum einen sinkt die reabsorbierte Strahlung des Wafers, zum anderen gelangt weniger Strahlungsenergie von den Lampen zum Wafer. Aus der Simulation läßt sich die Größe des Effekts abschätzen.

Um Größe und Ursache des Einflusses der Kammerreflektivität auf den Wafer zu untersuchen, wurde ein Modell für die Wellenlängenabhängigkeit der Wandabsorptivität entwickelt. Als Grenzfälle dienen dabei glänzendpolierter Stahl mit Daten aus [40] und mattes Wolfram. Die spektrale Abhängigkeit der Absorptivität ist bei den meisten Metallen sehr ähnlich, so daß der entscheidende Einfluß von der Reflektivität im Infraroten zu erwarten ist.

Analysiert man die Strahlungsflüsse aller Objekte in der Kammer auf den Wafer, so ergeben sich folgende Hauptstrahlungswege:

1. Direkte Aufheizung durch direkte Bestrahlung durch die Lampen und andere heiße Flächen mit der Leistung

$$I^{direkt} = g^{\text{Lampen}\to\text{Wafer}} P_{\text{Lampen}} \tag{2.16}$$

   und dem geometrischen Sichtfaktor $g^{\text{Lampen}\to\text{Wafer}}$

2. Indirekte Aufheizen durch an den Wänden reflektierte Lampenstrahlung mit der Intensität

$$I^{refl} = g^{\text{Lampen}\to\text{Wand}\to\text{Wafer}}(1 - a^{\text{eff}}_{\text{Wand}}(T_{\text{Lampen}}))P_{\text{Lampen}} \tag{2.17}$$

3. "Selbstaufheizung" durch reabsorbierte Waferstrahlung, die von den Wänden reflektiert wurde mit der Intensität

$$I^{reabs} = g^{\text{Wafer}\to\text{Wand}}(1 - a^{\text{eff}}_{\text{Wand}}(T_{\text{Wafer}}))g^{\text{Wand}\to\text{Wafer}} a^{\text{eff}}_{\text{Wafer}}(T_{\text{Wafer}})P_{\text{Wafer}} \tag{2.18}$$

Eine wesentliche Frage ist, ob der Einfluß der Lampen (Punkte 1 und 2) über die Selbstsicht (Punkt 3) dominiert. Ist dies nicht der Fall, so bedarf es einer sehr genauen Modellierung des Reabsorptionsterms. Eine einfache Abschätzung ergibt sich aus der Bilanz der Strahlungsflüsse

$$e^{\text{eff}}_{\text{Wafer}}\sigma T^4_{\text{Wafer}} = P_{\text{Lampen}} \times \frac{g^{\text{Lampen}\to\text{Wafer}} + g^{\text{Lampen}\to\text{Wand}\to\text{Wafer}}(1 - a^{\text{eff}}_{\text{Wand}}(T_{\text{Lampen}}))}{g^{\text{Wafer}\to\text{Quarz}} - g^{\text{Wafer}\to\text{Wand}}(1 - a^{\text{eff}}_{\text{Wand}}(T_{\text{Wafer}}))g^{\text{Wand}\to\text{Wafer}} a^{\text{eff}}_{\text{Wafer}}(T^{\text{Wafer}})} \tag{2.19}$$

Ist der Bruch auf der rechten Seite in Gleichung 2.19 groß, so führt eine kleine Änderung der Lampenleistung zu einer großen Temperaturerhöhung, die Kammerreflektivität dominiert über die Abstrahlungsverluste. Der genaue Wert des Verhältnisses hängt allerdings von der Reaktorgeometrie ab; im vorliegenden

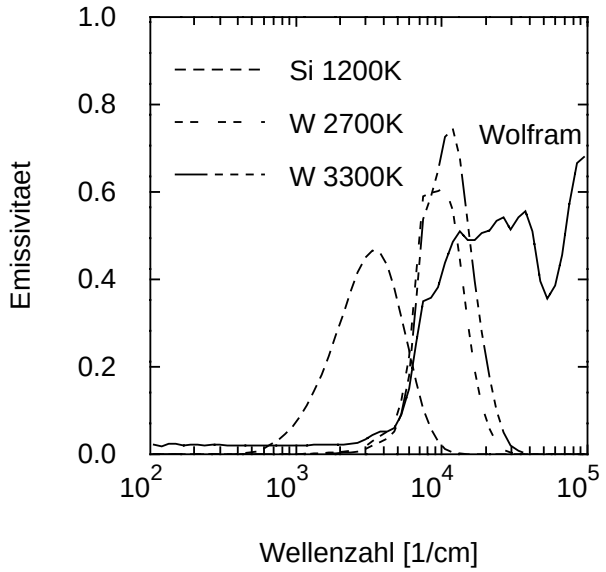


Abbildung 2.11: Emissivitätsspektrum der Wolframhalogenlampen und des Wafers.

Entwurf mit dicken Quarzplatten wird wenig vom Wafer ausgesandte Strahlung wieder auf den Wafer zurückreflektiert. Experimentell läßt sich bei voller Lampenleistung (etwa 50kW) eine Wafertemperatur von etwa 1500K erreichen. Damit ergibt sich für den Bruch ein Wert von etwa 4.0.

Abbildung 2.12 zeigt die Wafernominaltemperatur als Funktion der Reflektivität der Wand bei Wellenlängen von 1μm und 3μm. Die ausgesandte Strahlung des Wafers hat ihr Maximum in der Nähe von 3μm, die der Lampen bei etwa 1μm (Abb. 2.11). Deutlich ist zu erkennen, daß die Wafertemperatur stark von der Kammerreflektivität abhängt. Der Strahlungsverlust durch die Absorptivität der Wand (unteres Bild) sinkt von 14% links auf 2% rechts, die reflektierte Lampenleistung an der Wand steigt von 34% auf 88%.

Aus den Diagrammen ist zu entnehmen, daß eine Variation der Lampenleistung stets die gleiche Temperaturänderung hervorruft, nahezu unabhängig von der Wandreflektivität bei 3 μm. Daraus ergibt sich, daß der Bruch in Gleichung 2.19 stets einen nahezu konstanten Wert aufweist, wenig beeinflußt von der Kammerreflektivität.

Dies läßt sich auch aus den simulierten Werten für die Strahlungswerte erkennen: Zum einen trägt die reflektierte Waferstrahlung in der simulierten Kammer

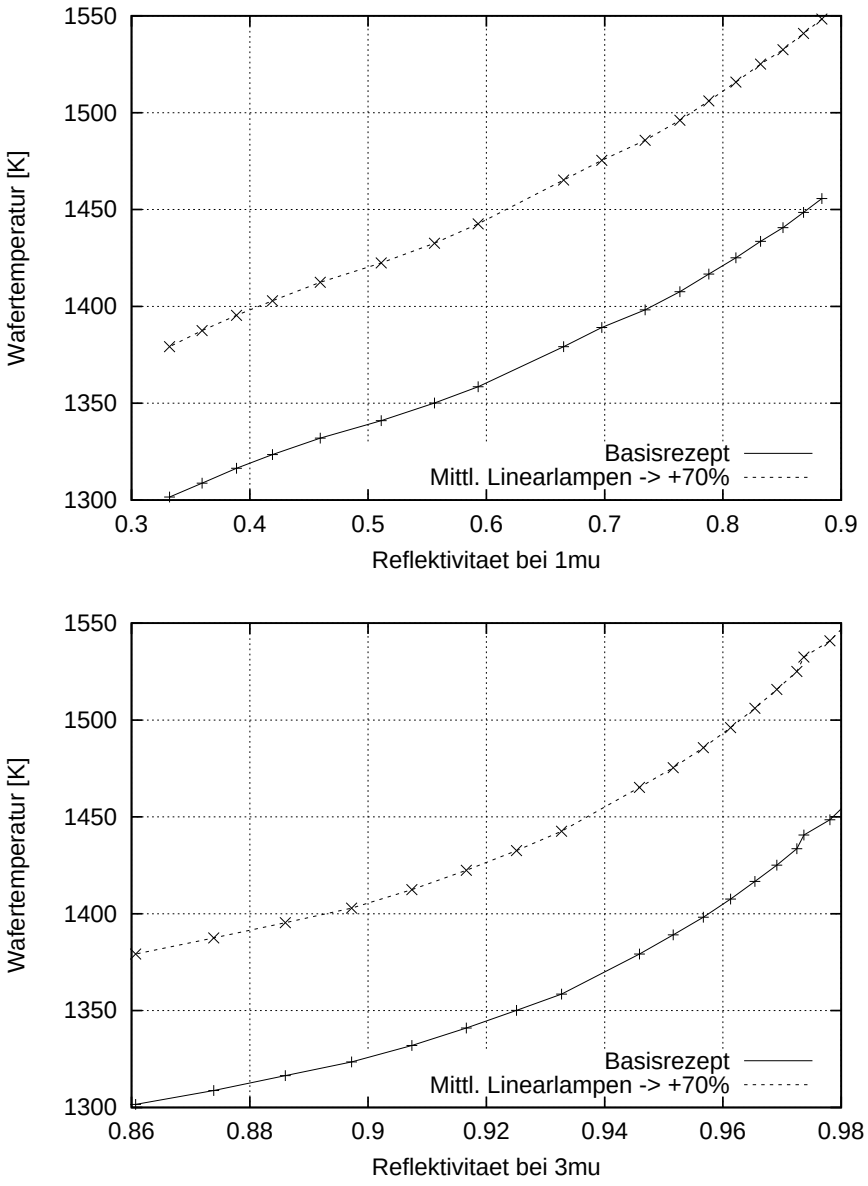


Abbildung 2.12: Auswirkung einer Leistungserhöhung in den unteren Linearlampen bei Schwankungen der Wandreflektivität im Wellenlängenbereich $1\mu m$ (oben) und $3\mu m$ (unten).

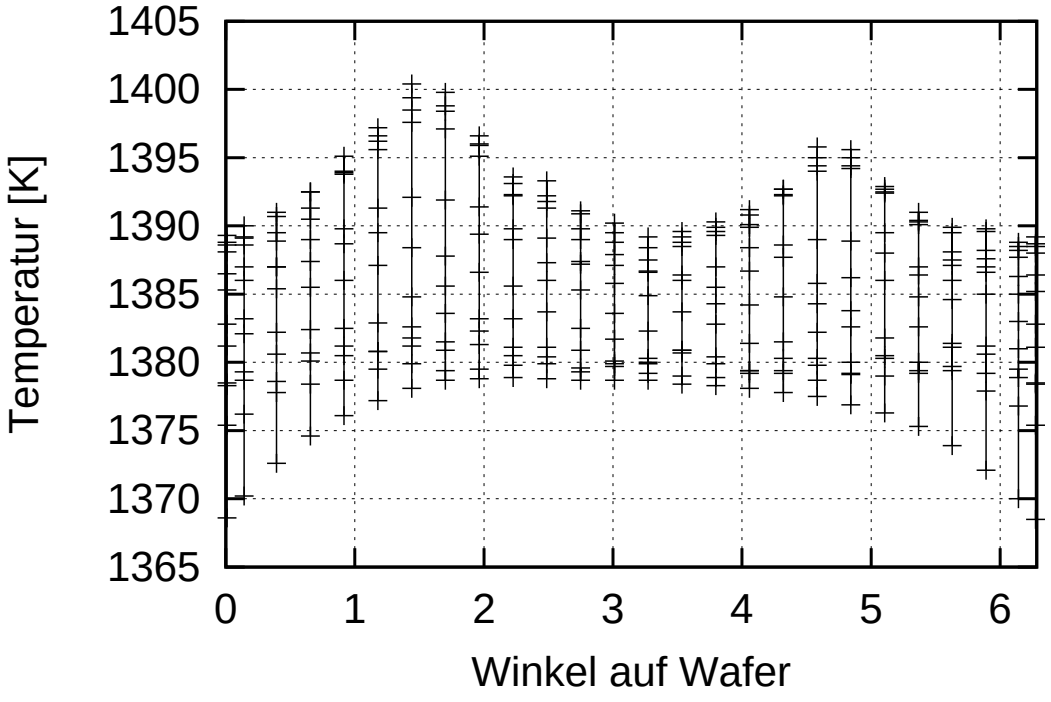


Abbildung 2.13: Temperaturschwankungen entlang verschiedener Schnitte von Wafermitte zum Waferrand (Winkel 0 bei der Tür). Die effektive Reflektivität der Kammerwand beträgt auf der einen Seite 0.10, auf der anderen 0.80.

nur mit wenigen Prozent zum Gesamtstrahlungshaushalt des Wafers bei, zum anderen ist die Änderung der Reflektivität bei 1μm zwischen dem polierten und unpolierten zu gering, um die großen Temperaturdifferenzen zu erklären.

Die Wafertemperatur steigt um 80K bei einer Variation der Gesamtleistung um etwa 15% (mittlere Stablampen 70% mehr Leistung). Um die gleiche Temperaturdifferenz durch Variation der Kammerabsorptivität zu erreichen, muß die Wandabsorptivität um 300% von $a = 1 - r = 0.02$ auf $a = 0.06$ erhöht werden.

Da die optimale Lampeneinstellung von den jeweiligen Herstellungsbedingungen der Kammer und des genauen Strahlungstransports von den Lampen zum Wafer abhängt, sind die Ergebnisse der Optimierung der Lampenbankleistungen allein aufgrund eines Modells (Abschnitt 2.5) nicht auf den realen Reaktor übertragbar. Abhilfe schafft hier das in Kapitel 3 vorgestellte Verfahren, das eine Kalibrierung eines reduzierten Modells zur Grundlage hat.

Zudem können die Lampeneinstellungen eines Reaktors nur dann auf ein Modell gleicher Bauart übertragen werden, wenn die Wände der beiden Kammern gleich vorbehandelt wurden. Bei dem für diese Arbeit untersuchten Reaktor wurde eine links-rechts Asymmetrie in der Temperaturverteilung bei symmetrischen Rezepten beobachtet. Als Grund dafür kommt die elektrolytisch Polierung der Wände in Betracht (Abbildung 2.21 unten), die in zwei Polierungen, für jede Reaktorhälfte getrennt, durchgeführt wird. Bei Kammern mit handpolierten Stahlwänden trat diese Unsymmetrie nicht auf. Eine Simulation der Temperaturverteilung bei einer unsymmetrisch bearbeiteten Reaktorkammer zeigt Abbildung 2.13.

2.4.4 Einfluß der Strömungsbewegungen in der Kammer

Rapid Thermal Oxidation Schritte werden typischerweise bei Drücken zwischen 100 und 750 Torr durchgeführt.

Die Kammerwände in einem RTP-System weisen eine deutlich niedrigere Temperatur als die Scheibe auf. Neben der erzwungenen Konvektion durch die Gaseinleitung kommt es im oben erwähnten Druckbereich daher zu natürlichen Konvektionsbewegungen. Das Gas wird über dem Wafer erhitzt, steigt aufgrund geringer werdender Dichte zum Quarz auf, wo es kühlt dort ab. Dadurch entstehen Walzenbewegungen, die zu Kühlungsmustern auf dem Wafer führen. In Abbildung 2.14 sind Simulationen dieser Gasbewegungen dargestellt.

Im Falle der Boussinesq-Näherung, d.h. für inkompressible, homogene, Newtonsche Flüssigkeiten gilt für die Kontinuitäts- und die Impulsgleichung in dimen-

sionsloser Formulierung:

$$\begin{aligned} \nabla \cdot \underline{v} &= 0 \\ \frac{\partial \underline{v}}{\partial t} + (\underline{v}\dot{\nabla})\underline{v} &= \mathrm{Pr}\left(-\nabla p + \nabla^2 \underline{v}\right) + \mathrm{Gr}\, T \underline{e_g} \end{aligned} \tag{2.20}$$

Als dimensionslose Kennzahl treten hier die Grashof- und Prandtl-Zahlen auf:

$$\mathrm{Gr} = \frac{\alpha \Delta T g l^3}{\nu^2} = \frac{\alpha \Delta T g l^3 \rho^2}{\eta^2}, \qquad \mathrm{Pr} = \frac{c_p \eta}{\lambda} \tag{2.21}$$

dabei ist $\alpha = \frac{\partial \rho}{\partial T}$ der Ausdehnungskoeffizient des Gases, η die dynamische Viskosität, $\nu := \frac{\eta}{\rho}$ die kinematische Viskosität, λ die Wärmeleitfähigkeit, c_p die spezifische Wärme, ΔT die typisch auftretende Temperaturdifferenz, $\underline{g} = g\underline{e_g}$ die Erdbeschleunigung und l eine charakteristische Länge des Systems.

Die Reynoldszahl $Re = \frac{vl}{\nu}$, die als Verhältnis zwischen Trägheitskraft und innerer Reibung zu verstehen ist, liegt im hier untersuchten Fall für 1300K bei maximal $\mathrm{Re} \approx \frac{0.3\mathrm{m/s}0.1\mathrm{m}0.27\mathrm{kg/m^3}}{510^{-5}\mathrm{kg/(ms)}} = 162$, also laminar und noch unterhalb des turbulenten Regimes.

Die Grashof-Zahl Gr hingegen beschreibt das Verhältnis von thermischer Auftriebs- zu viskoser Kraft und liegt bei einem Quarzabstand von 0.03 Metern und einer Temperaturdifferenz von 800K bei $\mathrm{Gr} = 6.2 \times 10^6$ also schon nahe dem Bereich des turbulenten Gastransportes.

Die Simulation kann jedoch qualitative Aussagen liefern, wenn die Diskretisierung fein genug gewählt wird. Nimmt man die Gitterdiskretisierung von etwa 3mm als Größenordnung für die charakteristische Länge in einer Gitterzelle so liegt die daraus errechnete "lokale" Grashof-Zahl um etwa drei Größenordnungen unter der Grashof-Zahl des Reaktors.

Aus Messungen und Simulationen von turbulenten Strömungen ist bekannt [23], daß die entstehenden Konvektionswalzen ein Aspektverhältnis von etwa eins haben. Allerdings ist ihre Zahl selbst in einfachen Geometrien zeitlich nicht konstant, sobald eine kritische Grashofzahl überschritten wird. So existieren Lösungen der Gleichungen 2.21 mit einer um zwei erhöhten oder erniedrigten Zahl von Walzen, eine Walze "teilt" sich auf in drei. Ein zeitliches Oszillieren der Walzenanzahl, Wanderungsbewegungen der Walzen und Kühlungseffekte turbulenter Strömungen wurde in der Arbeit von Kessler [23] untersucht. Es stellt sich heraus, daß für eine vorgegebene Geometrie eine kritische Grashof-Zahl existiert, ab der die Temperaturoszillationen durch die sich verändernden Walzen stark zunehmen. Ähnliches ist auch in den Messungen des simulierten

Reaktors zu beobachten. In Abbildung 2.15 sind die Temperaturschwankungen eines mit Thermoelementen bestückten Wafers bei verschiedenen Drücken gezeigt. Sinkt der Druck und damit die Dichte, so lassen die Oszillationen in den Thermoelementen nach.

Das nichtstationäre Verhalten der Konvektionskühlung ist durch eine Regelung nicht kompensierbar. Die Längenskala, auf der die Konvektionsmuster auftreten, liegt bei etwa ein bis vier Zentimetern und ist nicht modellmäßig erfaßbar. Der Wafer müßte während des Prozesses mit dieser Ortsauflösung abgetastet werden, was etwa 300 Meßpunkten entspricht.

Ziel muß es also sein, diesen Effekt nicht durch Regelung, sondern durch ein verändertes Kammerdesign zu vermindern. Aus Gleichung 2.21 ergeben sich folgende Einflußmöglichkeiten

1. Erhöhung der Flußrate in der Kammer, so daß die erzwungene Strömung dominiert
2. Verringerung der Temperaturdifferenz zwischen Wafer und Quarz
3. Herabsetzung des Drucks und damit der kinematischen Viskosität
4. Verringerung des Abstands zum Quarz

Eine Verringerung der Temperaturdifferenz zum Quarz ist aus designtechnischen Gründen nicht möglich und scheidet daher aus. Die notwendige Flußmenge, die zur Dominanz der Strömungsverhältnisse in der Kammer nötig wäre, läßt sich abschätzen durch den Fall einer homogenen Staupunktströmung auf den Wafer, z.B. durch einen Showerhead. Bei den obigen Strömungen wäre bei Strömungen ab $F = A_{wafer} v \approx 200$ Standardliter/Minute die Strömungsmenge größer als die natürliche Konvektion. Eine stabilisierender Effekt ist ab etwa 40 Standardlitern zu erwarten. Da derart hohe Strömungen nicht erreichbar sind und die konvektiven Effekte nicht verschwinden, sondern nur stabilisiert werden, ist dieser Weg nicht sinnvoll.

Wie in Abbildung 2.15 experimentell nachgewiesen wird, ist eine Druckreduktion eine mögliche Lösung. Wie im Abschnitt A gezeigt wird, ist allerdings die Oxidationsrate in etwa proportional zu $(p/1\text{atm})^{0.5}$, d.h. die Oxidationszeit steigt beim Übergang auf 500mbar etwa um 40% an, was den Durchsatz der Anlage herabsetzt.

Eine Alternative zur Herabsetzung des Kammerdrucks ist eine Verringerung des Abstands zum oberen Quarz (Abbildung 2.18), z.B. durch ein Hinaufsetzen

der Waferposition. Allerdings werden dadurch neben dem erschwerten Einlegen des Wafers die Konvektionsbewegungen in der unteren Kammerhälfte verstärkt. Als Ausweg bietet sich das Einlegen einer weiteren Quarzplatte in die Kammer an. Bei einer Höhe von weniger als 15mm über dem Wafer werden die Effekte aufgrund der l^3 Abhängigkeit auf ein Minimum reduziert.

Die weitere Quarzplatte hat selbst auch Einfluß auf die Wafertemperatur. Zum einen wird das Licht der Lampen an der Quarzplatte gebrochen, was zu einem Parallelversatz des Lichtkegels führen kann; die Lampe wirkt damit so, als wäre sie näher am Wafer (siehe Abbildung 2.16). Die virtuelle Höhendifferenz ergibt sich nach kurzer Rechnung zu

$$\Delta h = d_{\text{Quarz}}(1 - \frac{1}{n_{\text{Quarz}}}) \approx 0.36 d_{\text{Quarz}}. \tag{2.22}$$

Aus Gleichung 2.22 und den Reaktorsimulationen ist ersichtlich, daß bei Dicken der Quarzplatte bis zu 3.5 Millimeter keine nennenswerten Veränderungen in der Temperaturverteilung auf dem Wafer zu erwarten sind, so daß eine Neukalibrierung des Reglers, z.B. bei dem Verfahren in Abschnitt 3.3, durch das Einfügen der Quarzplatte nicht nötig ist.

Da die Platte aus Gründen der Reaktorgeometrie nur wenig über den Wafer hinausragen kann, ist die Strahlungskopplung zwischen Platte und Wafer am Rand kleiner als in der Mitte. Die Temperaturdrift in der Wafermitte ist damit etwas höher als am Rand. Dies führt bei einer Prozeßzeit von etwa 2 Minuten zu einem leichten Temperaturabfall des Waferrandes gegenüber der Mitte von einigen Grad.

Des weiteren ist die thermische Masse einer etwa 3mm dicken Konvektionsplatte deutlich geringer als die des Quarzfensters. Die Drift der Wafertemperatur während des Prozesses wird demnach deutlich höher und muß durch den Regelalgorithmus kompensiert werden. Der Vorteil aber besteht darin, daß die nicht voraussagbaren Temperaturschwankungen der Konvektion durch eine berechenbare und kontrollierbare Störung ersetzt werden.

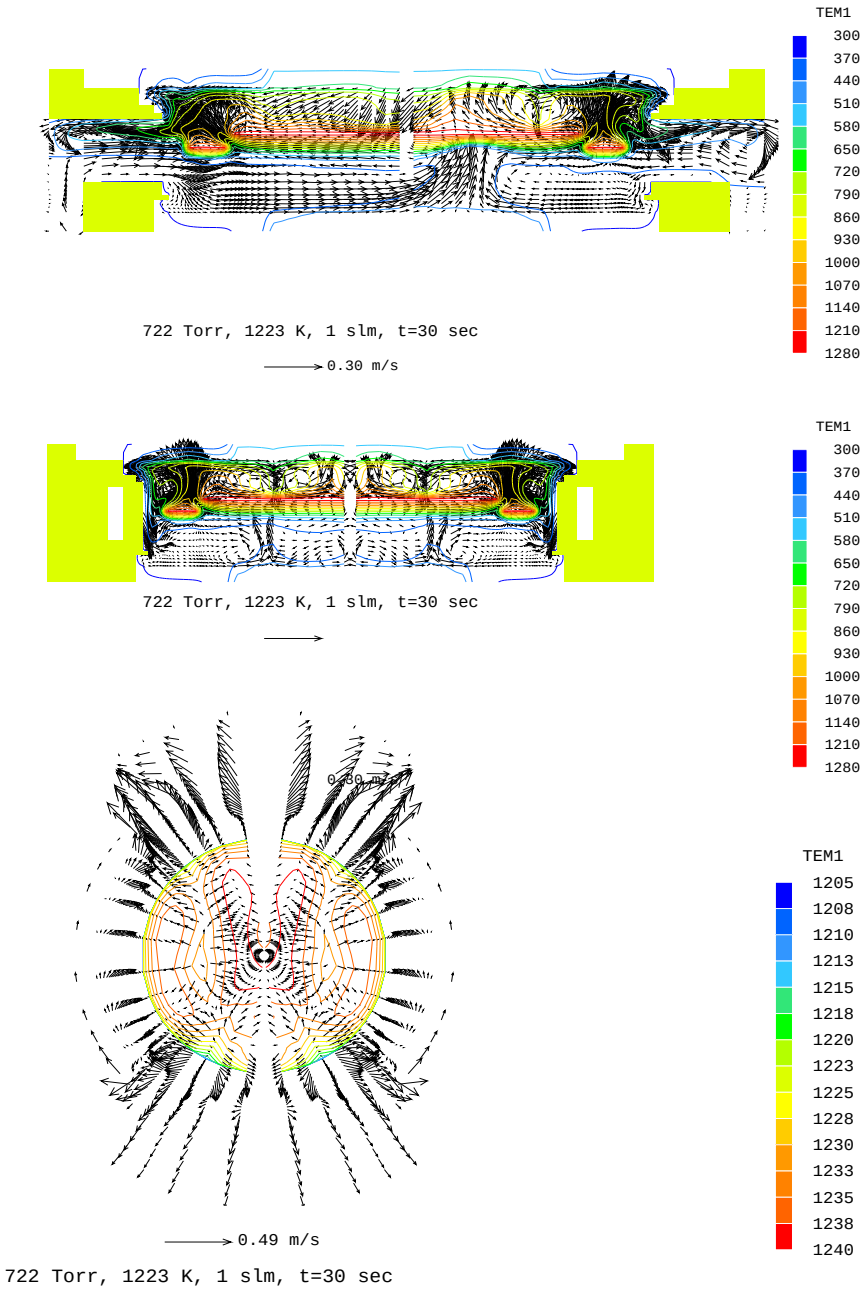


Abbildung 2.14: Konvektionsbewegungen in der Kammer [26]. Die Gasgeschwindigkeiten durch die aufsteigenden Gase sind sehr hoch und können durch die Einlaßdüsen erst durch hohe Flußraten von mehr als 10 Standardliter/Minute beeinflußt werden.

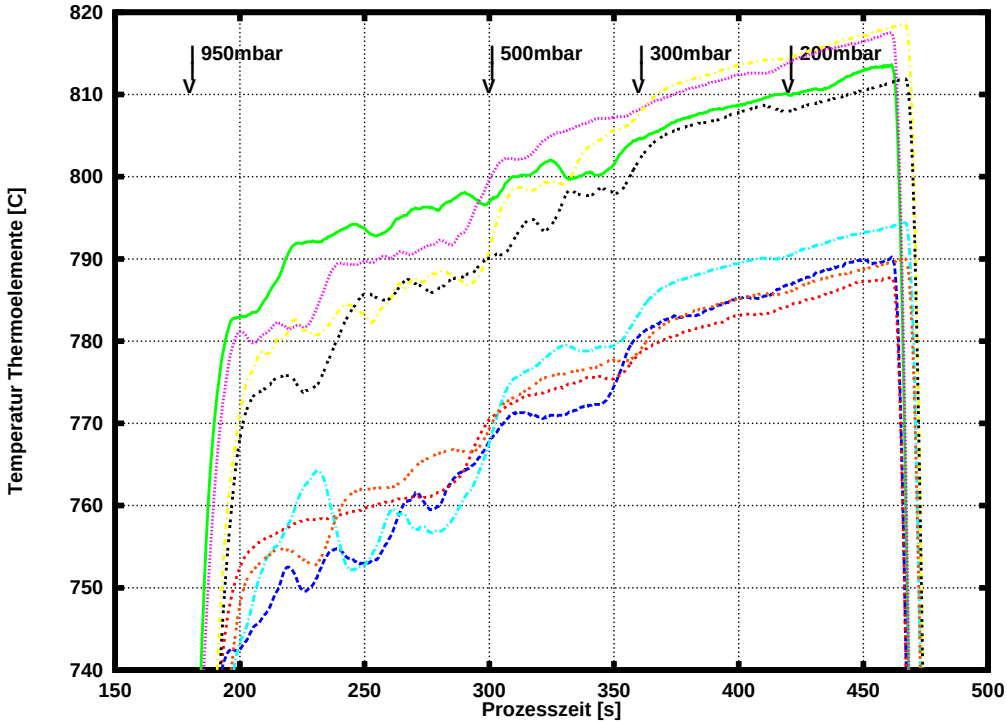


Abbildung 2.15: Messungen der Wafertemperatur mit Thermoelementen in Stikstoffatmosphäre bei verschiedenen Kammerdrücken. Die großen Schwankungen der Thermoelemente liegen in der von der Simulation für die Konvektion vorhergesagten Größenordnung von ± 4 Grad bei 1000mbar.

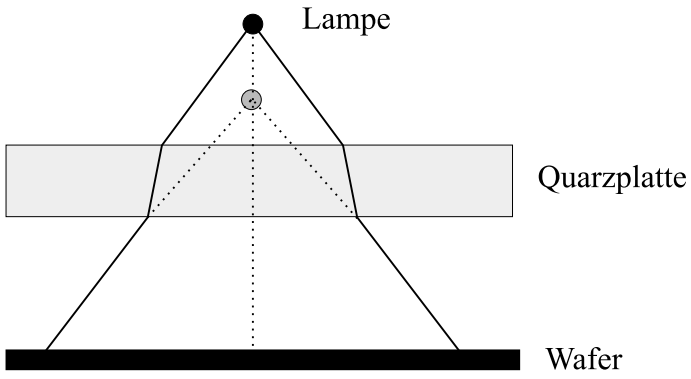


Abbildung 2.16: Parallelversatz des Strahlenkegel der Lampen durch einen Quarzliner.

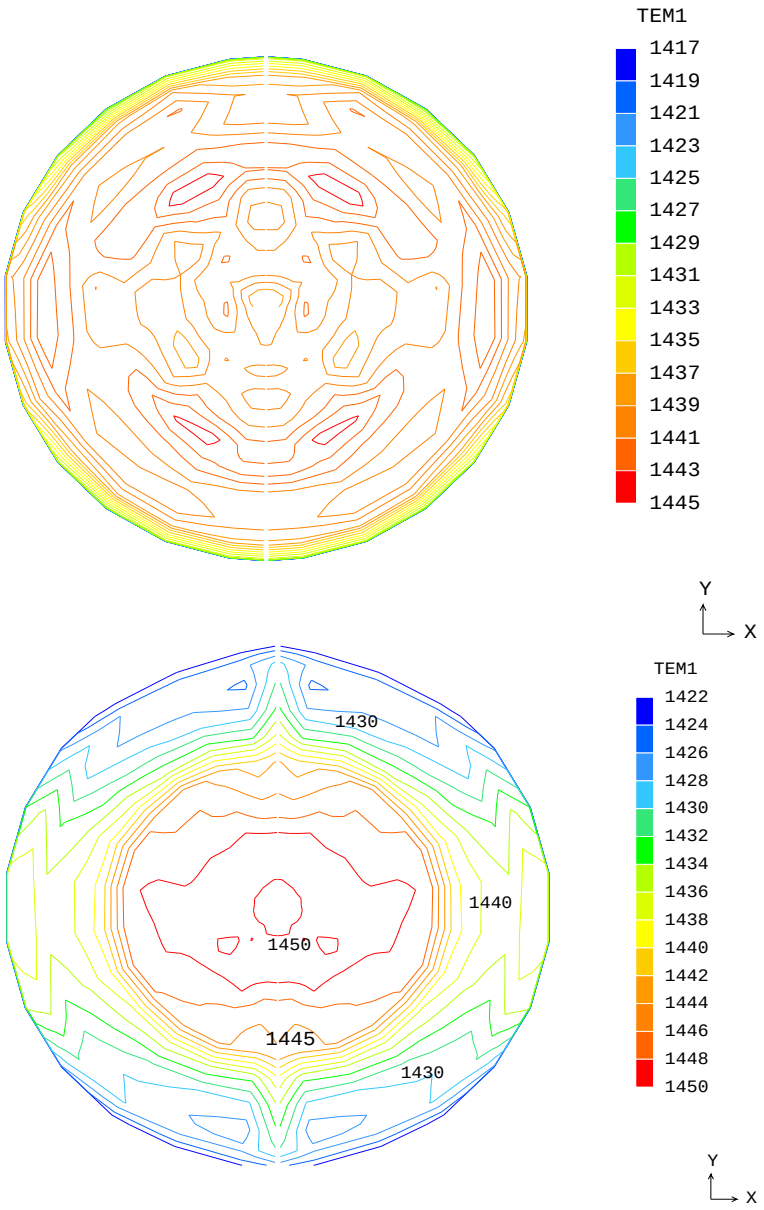


Abbildung 2.17: Simulierte Temperaturverteilungen mit 1mm (oben) bzw. 5mm (unten) Liner mit einem nicht-optimierten Basisrezept nach 180 Sekunden Prozeßzeit. Der Abfall zum Rand verringert sich bei näherliegendem Liner.

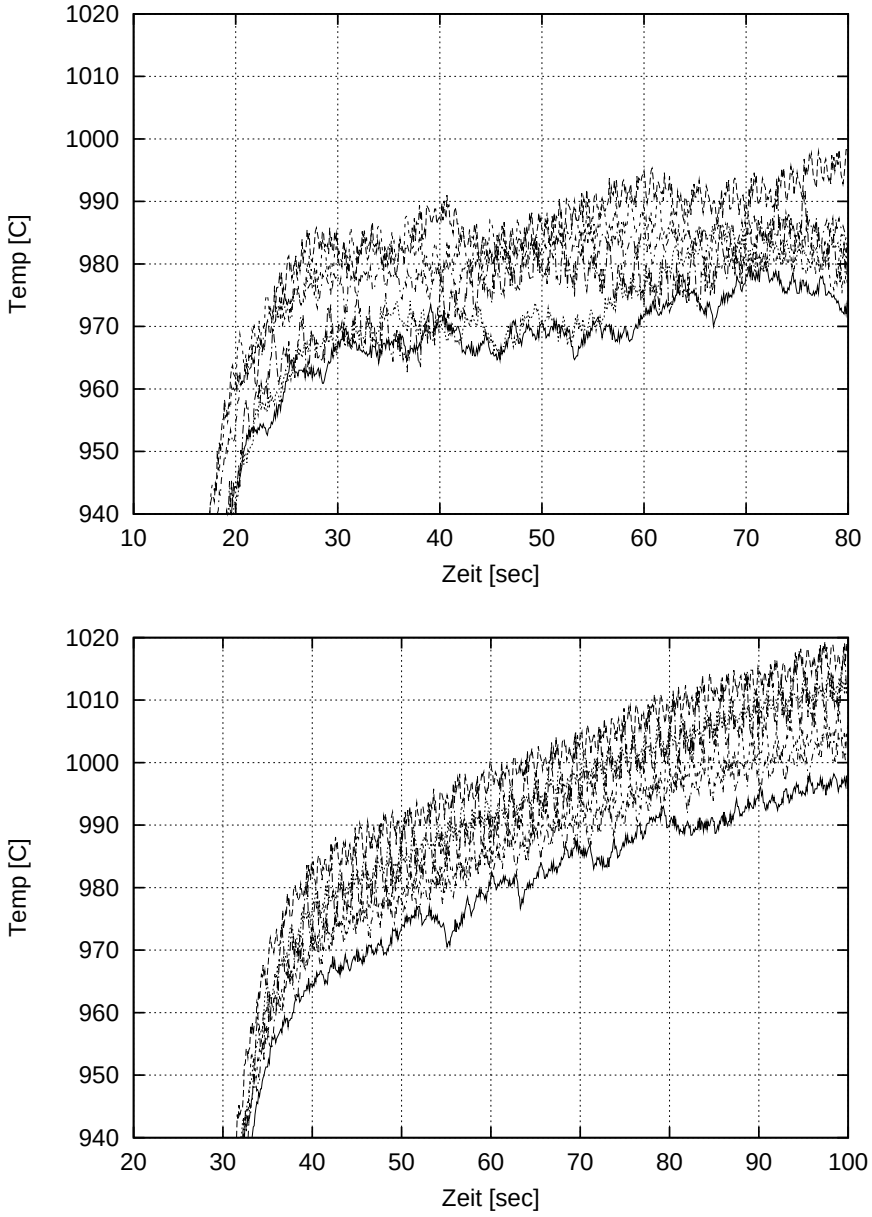


Abbildung 2.18: Messungen der Wafertemperatur mit Thermoelementen vor (oben) und nach Einfügen der Konvektionquarzplatte (unten) bei 950mbar Stickstoffatmosphäre [27]. Die Schwankungen der Thermoelemente werden durch die Konvektionsplatte auf die elektrischen Störungen reduziert.

2.5 Optimierung der Temperaturuniformität auf Basis der Simulationen

Das im nächsten Kapitel näher erläuterte Verfahren zur Optimierung der Temperaturhomogenität kann direkt mit der errechneten Strahlungsmatrix verwendet werden. Dies entspricht dem Fall einer exakt meßbaren Strahlungsmatrix. Unter Anwendung des Optimierungsverfahrens läßt sich die maximal erreichbare Temperaturhomogenität des Reaktors analysieren.

Der Einsatz weiterer oder stärkerer Lampen im unteren und auch im oberen Lampenhaus ist also notwendig für das Erzielen einer verbesserten Temperaturuniformität.

Eine wichtige Frage ist, inwiefern die Simulation als Basis für die Optimierung der Uniformität ausreicht. Das auf Basis der Simulationen ermittelte uniforme Lampenrezept wurde an der Kammer eingestellt und anschließend die Oxiddicke auf dem Wafer gemessen.

2.5.1 Optimierungsergebnisse

Zunächst ist zu klären, ob die Sensitivität des Wafers auf Änderungen der Lampenleistungen in der Simulation hinreichend beschrieben wird. In den Abbildungen 2.19 und 2.20 ist die Temperaturerhöhung der Wafertemperatur bei Erhöhung der Leistung in jeweils einer Bank dargestellt. Die Übereinstimmung in der örtlichen Verteilung zwischen den Simulationen (Bild oben) und den Thermoelementmessungen (Bild unten) ist sehr gut, d.h. die simulierte Verteilung der Lampenintensität auf dem Wafer entspricht der des realen RTP-Systems.

Im nächsten Kapitel wird ein reduziertes Modell und ein Extraktionsverfahren zur Ermittlung der Kopplungsparameter zwischen Lampen und Wafer vorgestellt. In das reduzierte Modell lassen sich aber auch direkt die Kopplungsparameter aus der Equipmentsimulation einsetzen und das Optimierungsverfahren zur Ermittlung der bestmöglichen Lampenleistungskonfiguration anwenden. Dieses Vorgehen eliminiert eventuelle Fehler, die durch das Extraktionsverfahren induziert werden.

Im reduzierten Modell ergibt sich für die Zieltemperatur eine erreichbare Uniformität von $\pm 4°C$ über den ganzen Wafer. In der vollen Equipmentsimulation ergibt sich die gleiche Temperaturuniformität, lediglich das Temperaturniveau unterscheidet sich vom reduzierten Modell um wenige Grad.

Es bleibt die Frage zu beantworten, ob das mit der errechneten Kopplungsmatrix optimierte Rezept im realen System zu einer ausreichenden Uniformität führt (Abbildung 2.22). Wie in Kapitel 3.3.3 erläutert, liefert das Oxidmodell etwas zu geringe Oxiddicken, die Temperaturabhängigkeit ist nicht so stark wie experimentell beobachtet. Die Inhomogenität ist in den dargestellten Abbildungen daher eher als überschätzt zu bewerten.

Mit dem optimierten Rezept ergibt sich in der Simulation eine Uniformität von $\pm 4°C$, im Experiment jedoch von mehr als $\pm 10°C$. Hierbei sind vor allem numerische Gründe anzuführen, da die Optimierung der Uniformität ein inverses Problem darstellt, das im allgemeinen schlecht konditioniert ist. Fehler in der Strahlungsmatrix wirken sich so besonders stark aus.

Obwohl die Intensitätsverteilung einer Lampe der Verteilung im Experiment entspricht, addieren sich die Modellierungsfehler der einzelnen Lampen auf. Vergleicht man die Temperaturverteilung in der Simulation mit dem Experiment (Abbildung 2.21), so treten die Temperaturmaxima an unterschiedlichen Stellen in Simulation und Experiment auf.

Wie im nächsten Kapitel gezeigt wird, liefert aber eine experimentell extrahierte Matrix verläßlichere Ergebnisse.

2.6 Zusammenfassung

Im vorausgegangen Kapitel wurden Einflußfaktoren auf die Wafertemperatur analysiert, um experimentell schwer auffindbare Effekte vorauszusagen und einfache, effektive Designverbesserungen anzubringen.

Die dominanten Effekte sind nicht ohne weiteres aus Messungen zu gewinnen; besonders das experimentelle Auffinden der Konvektionswalzen ist hierbei schwierig. Die Simulation und Modellierung hat wesentlich zum Verständnis und zur Abhilfe der Mechanismen beigetragen. Dies ist auch ein wesentlicher Punkt beim Entwurf eines modellbasierten Reglers: aufgrund der begrenzten Meßinformationen während des Prozesses müssen Störeffekte, die nicht quantitativ modelliert werden können, vermieden werden.

Eine Optimierung der Uniformität allein mit Hilfe der Simulation liefert keine ausreichend guten Ergebnisse. Die Verwendung von experimentell ermittelten Parametern ist unverzichtbar.

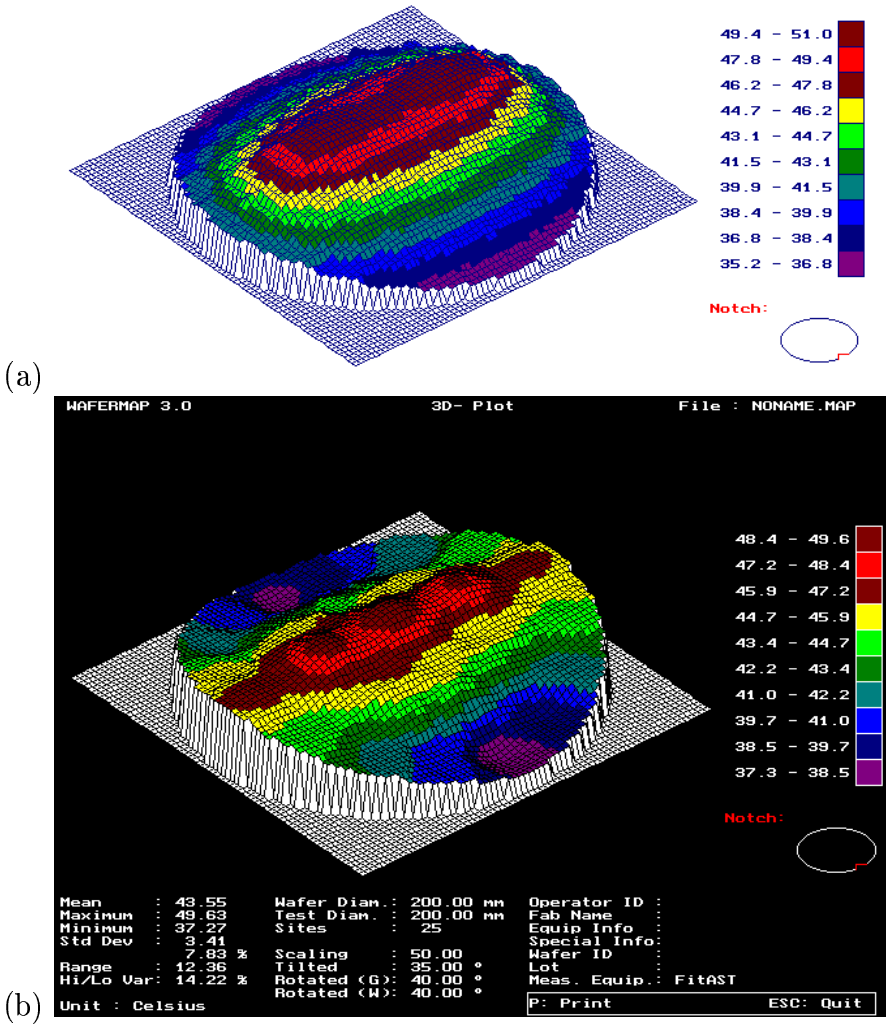
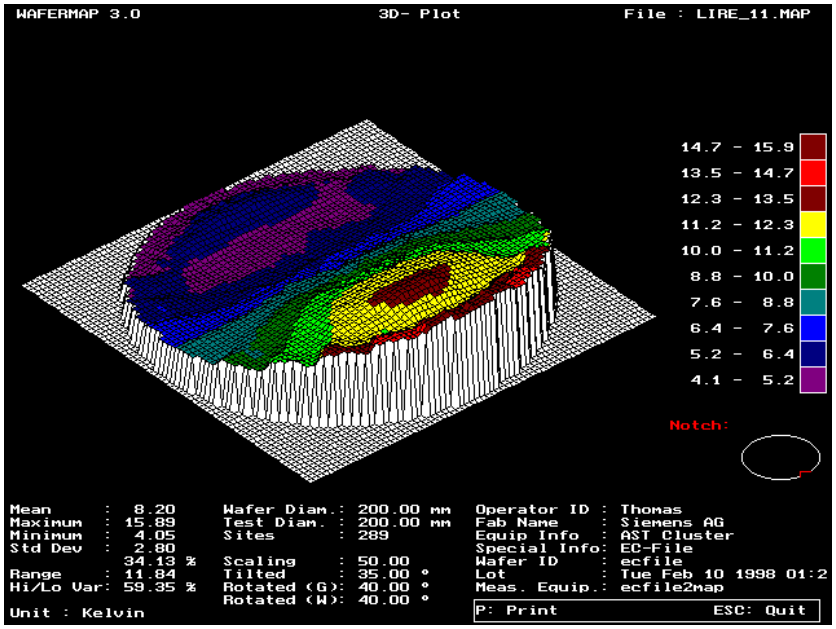


Abbildung 2.19: Änderung der Wafertemperatur bei Erhöhung der Lampenleistung in den mittleren 6 Linearlampen von 33% auf 100%. Simulation (oben) und Messung (unten) unterscheiden sich von Niveau und Verteilung nur geringfügig. Die unterschiedliche Form ist auf die Gitterinterpolation zurückzuführen.

(a)



(b)

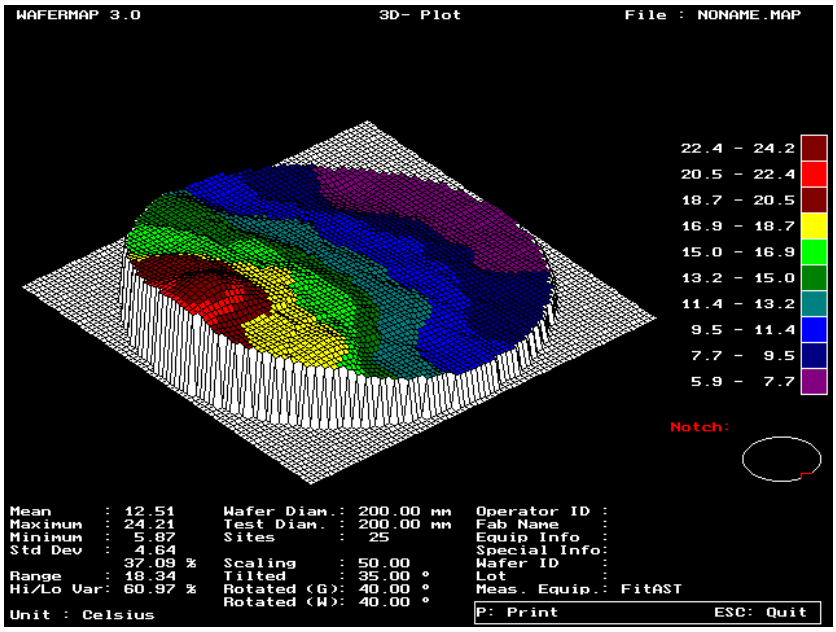


Abbildung 2.20: Ausgangspunkt wie in Bild 2.19, nur daß hier die Leistung in der rechten Hälfte des 4. Rings von 28% auf 100% erhöht wurde (Simulation oben/Messung unten). *Die Simulationen sind aus Darstellungsgründen um 90 Grad verdreht.*

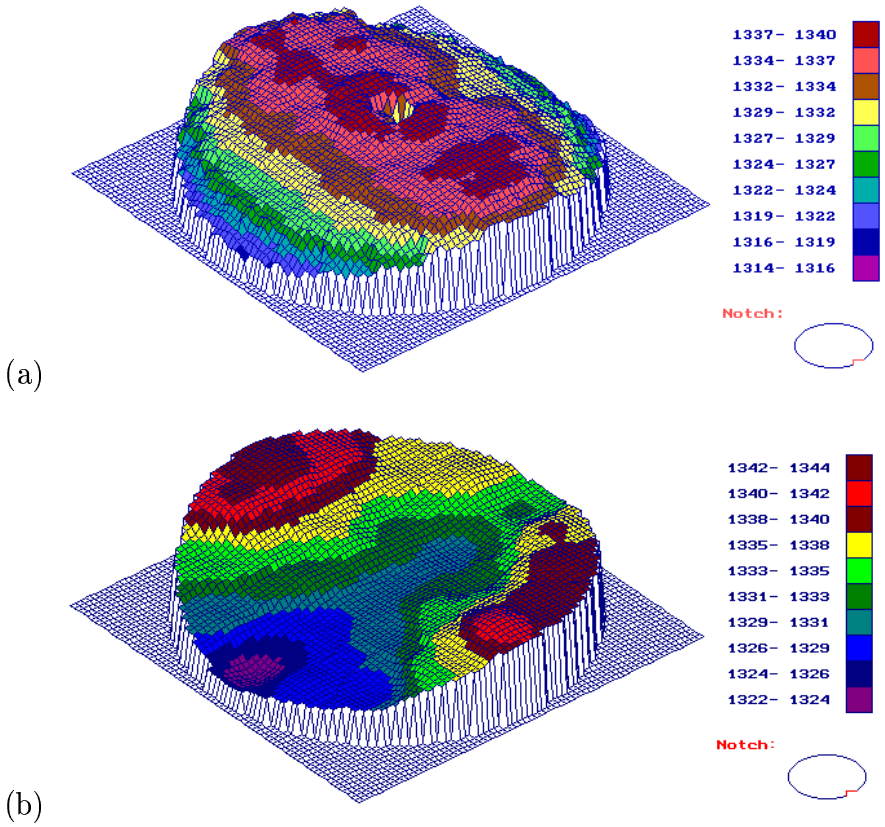


Abbildung 2.21: Vergleich der Simulationsergebnisse (Bild oben) mit Thermoelementmessungen (Bild unten) bei 500mbar. *Die Simulationen sind aus Darstellungsgründen um 90 Grad verdreht.* Für die Thermoelementmessungen wurde die in Kapitel 3.6 Konfiguration gewählt (Simulation oben/Messung unten). In den Simulationen ist eine Asymmetrie in der Wandreflektivität nicht berücksichtigt.

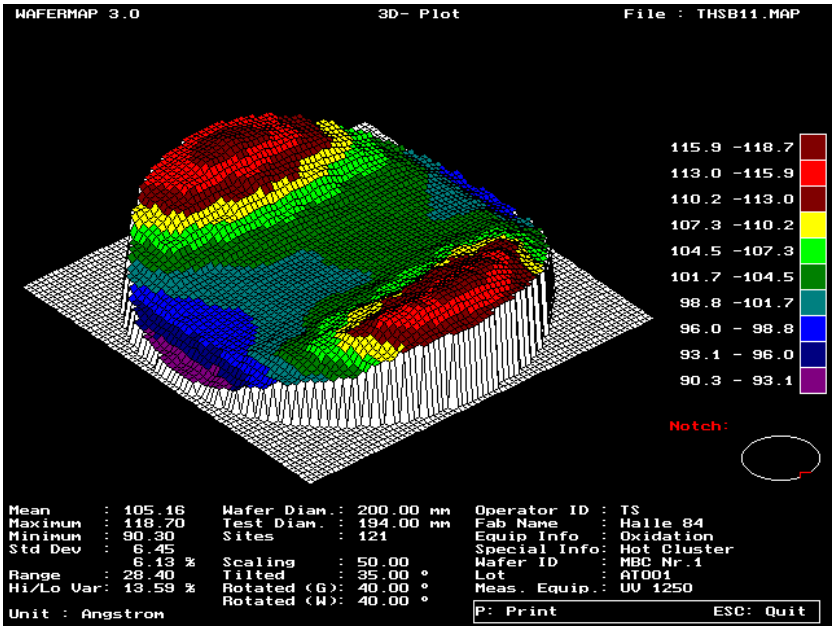


Abbildung 2.22: Oxidmessungen für das auf Basis der Simulation optimierte Rezept für einen 120 Sekunden Oxidationsprozeß bei 500mbar. Da die simulierten Verteilungen nicht genau mit den Messungen übereinstimmen, ergibt sich beim inversen Problem eine noch stärkere Inhomogenität.

Kapitel 3

Verfahren zur Optimierung der Temperaturuniformität in einer RTO-Kammer

3.1 Problemstellung

Im vorangegangenen Kapitel wurde aufgezeigt, welche Verbesserungen an einer RTO-Kammer mit Hilfe der Equipmentsimulation erzielt werden können. Die bisher diskutierten Fragestellungen betrafen vor allem das grundlegende Verhalten des Reaktors.

Zur bestmöglichen Steuerung eines Reaktors erscheint dies nicht als ausreichend. Dort spielen Fragestellungen wie Uniformität über den Wafer eine große Rolle. Im folgenden wird gezeigt, daß für eine Optimierung der Temperaturuniformität die Genauigkeit und der Rechenzeitbedarf der Equipmentsimulation nicht ausreicht. Da andererseits auch eine Charakterisierung des Systemverhaltens allein durch Messungen nicht durchführbar ist, wurde ein neues Extraktionsverfahren basierend auf einem reduzierten Reaktormodell entwickelt.

Die Vorteile des neuen Verfahrens liegen im Gewinn physikalischer Kopplungsgrößen unter Verwendung eines überschaubaren Satzes von Messungen. Dadurch lassen sich optimierte Rezepte über einen Arbeitspunkt der Kammer hinaus auch bei anderen Bedingungen ermitteln. Zum anderen läßt sich die Zeit zum Auffinden der optimalen Kammereinstellungen deutlich reduzieren. Als Anwendungsschwerpunkt dient die im vorausgehenden Kapitel diskutierte Rapid-Thermal-Oxidation Kammer.

Zunächst wird nun das Konzept des Verfahrens und das dazu entwickelte reduzierte Modell vorgestellt. Im Abschnitt 3.5.4 wird die Stabilität des Verfahrens mit Hilfe des Equipmentsimulators als Referenzsystem analysiert. Im Unterkapitel 3.6 wird schließlich die erfolgreiche Anwendung des Verfahrens an einer acht Zoll RTO-Kammer demonstriert. Zum Abschluß werden weitere mögliche Anwendungsbereiche dieser Methodik diskutiert.

3.2 Konzept des Extraktions- und Optimierungsverfahrens

An die in dieser Arbeit betrachteten Oxidationsschritte werden im Fertigungsprozeß Uniformitätsanforderungen von etwa 2% Oxiddickenvariation über den Wafer gestellt. Dies entspricht einer Temperaturschwankung von weniger als etwa $\pm 4°C$ bei einer Prozeßtemperatur von 1050°C. Bei der Vielzahl von Bänken und Einflußgrößen ist eine Optimierung von Hand durch ein Versuch-und-Fehler Verfahren sehr zeit- und kostenaufwendig.

Aus der Regelungs- und Systemtheorie sind hingegen eine Reihe von Ansätzen zur Steuerung/Regelung von Mehrparametersystemen bekannt, die im folgenden auf ihre Anwendbarkeit auf das vorliegende Problem untersucht werden.

3.2.1 Optimierung aufgrund detaillierter Modelle

Eines der grundlegenden Verfahren ist die Optimierung auf der Basis eines Modells. Der Vorteil dieses Verfahrens ist seine Anwendbarkeit bei allen Prozeßbedingungen, für die das entwickelte Modell gültig ist. So ist man nicht auf eine Arbeitstemperatur festgelegt.

Die Anforderungen an eine Temperaturuniformität von weniger als 3 Promille stellen allerdings nichterfüllbare Anforderungen an die Modellgleichungen. Im vorangegangenen Kapitel wurde ein detaillierter Simulator für thermische Reaktoren vorgestellt, bei dem besonderer Wert auf die quantitative Erfassung aller physkalischen Effekte gelegt wurde. Dennoch sind einige physikalische Größen, wie z.B. die exakte räumliche Intensitätsverteilung der Einzellampen oder die Reflektivität der Kammerwände, nicht mit hinreichender Genauigkeit modellierbar. Ein Einbringen von Parametermessungen ist demnach erforderlich. Die dafür nötigen Kalibrationsmessungen, z.B. die wellenlängenabhängige Messung der Kammerwandreflektivität, ist aber nicht praktikabel; auch die Messung der Abstrahlungscharakteristik der Lampen ist Einschränkungen unterworfen.

3.2.2 Systemidentifikation mittels Sprung- und Impulsantwort

Die Prozeßidentifikation durch Messen der Antwort des Systems auf Änderungen in den Stellgrößen ist eines der typischen Verfahren zur Modellierung des Systemverhaltens, wenn nur eine unzureichende theoretische Modellierung möglich ist. Basis der Systemidentifikation ist zumeist das Festlegen auf eine Transferfunktion, also z.B. die Annahme, daß es sich um ein System erster Ordnung mit Totzeit handelt.

Die richtige Auswahl ist schon dadurch erschwert, weil verschiedene Systeme ein ähnliches Verhalten bei der Sprungantwort zeigen, insbesondere werden Systeme mit steigender Ordnung immer ähnlicher zu Totzeitsystemen [51].

Betrachtet man nun ein System erster Ordnung mit Totzeit

$$\tau\dot{x} = -x + ku(t - t_0) \tag{3.1}$$

so erhält man bei einer Änderung der Aktuatorgröße u um den Wert A einen Signalverlauf der Form

$$x(t) = \begin{cases} 0 & \text{für} t < t_0 \\ Ak\,(1 - \exp(-(t - t_0)/\tau)) & \text{für} t \geq t_0 \end{cases} \tag{3.2}$$

Aus einer Serie von Messungen lassen sich so die Parameter K, τ und t_0 extrahieren.

Das betrachtete System der Wafertemperatur ist erster Ordnung (siehe Abschnitt 3.3), so daß die Größen k durch Variation der Lampenleistungen meßbar sind [52]. Jedoch entsteht dadurch eine Temperaturinhomogenität auf dem Wafer, die durch Wärmeleitung ausgeglichen wird und das Meßergebnis verfälscht. Außerdem ist das System stark nichtlinear aufgrund der T^4-Abstrahlung des Wafers, so daß zum einen die Signaländerungen klein und damit schwer meßbar sind, zum anderen der Gültigkeitsbereich der linearen Approximation sehr klein ist. Die Systemidentifikation hochdimensionaler Systeme ist schwierig und erfordert einen immensen Meßaufwand.

Der Vorteil der Identifikation per Sprungantwort besteht in der Information über die Zeitkonstanten des Systems. Gerade aber der zeitliche Verlauf ist bei den zur Verfügung stehenden Meßverfahren mit Thermoelementen nicht zuverlässig möglich. Die Zeitkonstante der Thermoelemente, die von den optischen Eigenschaften ihrer Keramikeinbettung abhängt, ist von der gleichen Größenordnung wie die des Wafers.

Es wird ferner in Gleichung 3.1 vorausgesetzt, daß sich das System vor Beginn des Schrittwechsels im Gleichgewichtszustand befindet. Durch die Aufheizung der Quarze kommt es jedoch zu einer Drift der Wafertemperatur von etwa zehn Grad pro Minute im hier analysierten System.

3.2.3 Statistische Optimierungsverfahren

Von Davis [48] wurde ein Verfahren vorgestellt, das die Uniformität einer Silanabscheidung mit RTCVD unter festgelegten Prozßbedingungen verbessern soll. Im Grenzfall einer hohen Anzahl von gemessenen Verteilungen und vielen Korrelationsparametern liefert das Verfahren gute Ergebnisse. Allerdings ist dieses Verfahren auch dann auf einen engen Arbeitsbereich beschränkt. Dies ist dann nicht praktikabel, wenn wie im vorliegenden Fall bei verschiedenen Temperaturen und Prozeßzeiten gearbeitet werden soll.

Ein ähnliches Verfahren von Knutson [49] erreicht eine Temperaturuniformität von 950°C mit mehr als 14 Grad Temperaturvariation nach mehreren Optimierungszyklen. Für den hier vorliegenden Prozeß ist dies nicht ausreichend, da bei höheren Temperaturen die Inhomogenität i.d.R. stark ansteigt.

3.2.4 Hybridverfahren

Von Tillmann [50] wurde eine iteratives Optimierungsverfahren vorgestellt, welches ausgehend von einem Prozeßergebnis Variationen der Lampenleistungen aufgrund einer simulierten Leistung errechnet. Als Eingabeverteilung dient z.B. eine Oxiddickenmessung und eine Strahlungssimulation der Lampen, welche eine bessere Uniformität erreichen soll. Die Oxidverteilung mit den so errechneten Einstellungen dient dann als neuer Startwert.

Die Methode entspricht somit einem Newtonverfahren zum Auffinden eines Optimums. Das Verfahren wird mit großem Erfolg bei solchen System angewendet, bei denen eine hinreichend gute Simulation der Lampenverteilungen möglich ist, d.h. wenn die Verteilung der einzelnen Lampen keine komplizierte räumliche und damit schwer modellierbare Verteilung aufweist. Wie im vorigen Kapitel beschrieben, trifft dies für das vorliegende System nur bedingt zu. Auch ist das Verfahren bei Veränderungen der Prozeßbedingungen zu wiederholen.

Im Rahmen dieser Arbeit wurde daher ein neues Verfahren entwickelt, das auch für Systeme mit lokalisierten Verteilungen und damit hoher Regelbarkeit anwendbar ist. Basis des Verfahrens ist ein detailliertes physikalisches Modell,

in dem alle wesentlichen Einflüsse auf die Wafertemperatur mit der benötigten Genauigkeit erfaßt werden. Die Parameter dieses Modells werden je nach Berechenbarkeit und Größe ihres Einflusses entweder analytisch oder numerisch berechnet oder aus Messungen extrahiert.

Man gewinnt so ein physikalisches Modell, das Gültigkeit in einem endlichen Bereich von Prozeßbedingungen – und nicht nur an einem Arbeitspunkt – hat. Ausserdem sind die benötigten Meßdaten durch wenige Messungen, hier unter Verwendung eines Thermoelement-Wafers, bestimmbar. Besonderer Wert wurde auf eine zuverlässige Auswertung der Signale gelegt. Da die Zeitkonstanten des Wafers und der anderen Flächen berechenbar sind, ist nur die Temperaturverteilung im eingeschwungenen oder nahezu eingeschwungenen Zustand als Eingabegröße erforderlich.

Im folgenden Kapitel werden zunächst die entwickelten reduzierten Modelle beschrieben.

3.3 Entwicklung eines reduzierten Modells

Der Entwurf eines reduzierten Modells des Systems ist aus mehreren Gründen notwendig. Die in Kaptel 2 angesprochenen Simulationen benötigen zur Berechnung einer Temperaturverteilung bei bekannter Strahlungsmatrix etwa 30 Minuten auf einem MIPS 8000 Prozessor. Zur Optimierung der Lampeneinstellung zu einer Temperatur werden mindestens $4 * n$ Schritte benötigt, wobei n die Anzahl der zu optimierenden Lampenleistungen ist. Bei einer Zahl von 30 Lampenzonen wie im vorliegenden Fall ist daher die benötigte Rechenzeit nicht akzeptabel.

Zum anderen ist die Kalibration der Simulationen schwierig. Wie im vorigen Kapitel gezeigt, sind Details in der Kammergeometrie und der Filamentform für Vorhersagen im Gradbereich ausschlaggebend, was die Entwicklung einer Methodik zur Extraktion relevanter Parameter nötig macht.

Grundidee des reduzierten Modells ist die Beschränkung auf die Modellierung der Waferdynamik. Das zeitliche Aufheizen der anderen Teile der Kammer geschieht ebenfalls über effektive Modelle. Die Lösung des vollständigen Satzes von Differentialgleichung wird ersetzt durch die Lösung eines Differentialgleichungssystems pro Waferzelle i mit deutlich geringerer Steifigkeit, da alle Zellen nun eine ähnliche Kopplung zu den Nachbarzellen und ähnliche Wärmekapazitäten aufweisen:

$$\rho V_i c_p(T)\frac{dT_i}{dt} = A_i\left(q_i^{lampen} + q_i^{rad.waf} + q_i^{wand} + q_i^{gas}\right) + q_i^{cond.waf} \tag{3.3}$$

dabei ist V_i das Volumen der Waferzelle i, A_i dessen Fläche. Die Wärmeflüsse $\underline{q}$ beschreiben den Wärmegewinn und -verlust der Zelle aufgrund der absorbierten Strahlung von den Lampen, der emittierten und reabsorbierten Strahlung des Wafers, die absorbierte Strahlung von den Wänden, der konduktive und konvektive Verlust durch das Gas und schließlich die Wärmediffusion in der Scheibe. Auf die Modellierung der einzelnen Terme wird nun kurz eingegangen.

3.3.1 Beschreibung des Strahlungsaustausches

Der Wärmetransport in lampengeheizten Systemen geschieht in erster Linie über die Strahlungskopplung zwischen den Oberflächen der Festkörper. In Gleichung 3.3 wird nur der Strahlungstransport auf und vom Wafer verwendet. Für die einzelnen Strahlungsterme ergibt sich mit den in Kapitel 2.3.3 beschriebenen Strahlungsaustauschmatrizen

1. **Aufheizungsterm von den Lampen**

$$q_i^{\text{lampen}} = \sum_{l=1}^{L} L_{il} P_l \tag{3.4}$$

Dabei enthält die Matrix $\mathbf{L}$ neben den geometrischen Sichtfaktoren die spektral gefalteten Emissivitäten der Lampen für die typische Lampenstrahlungstemperatur von 3000K und die spektral gemittelten Absorptivität des Wafers, $\mathbf{P}$ enthält die in die Lampen eingekoppelte elektrische Leistung. Um z.B. die Semitransparenz des Wafers bei Temperaturen unter 700°C zu berücksichtigen, können unterschiedliche Matrizen je nach Wafertemperatur verwendet werden oder eine temperaturabhängige effektive Absorptivität verwendet werden.

2. **Waferkopplungsterm**

$$q_i^{\text{rad.waf}} = \sum_{j=1}^{I} R_{ij} \cdot \sigma e_j^{\text{wafer}} T_j^{\text{wafer}4} \tag{3.5}$$

Die Matrix $\mathbf{R}$ enthält dabei in der Diagonale nicht nur die Selbstsichtfaktoren für alle Waferdiskretisierungselement $i, j = 1, \ldots, I$, sondern auch die Abstrahlungsverluste der Waferzellen. Die Matrix ist ebenfalls von

der Wafertemperatur abhängig, da hier zwar wegen $a(\nu,T) = e(\nu,T)$ weniger die optischen Eigenschaften des Wafers eine Rolle spielen, wohl aber der optische Weg für niedrige ($< 700°$) und hohe Temperaturen $\lambda > 900°$ unabhängig ist: die Quarze absorbieren im nahen Infraroten, daher ist die Sicht des Wafers auf die verspiegelten Lampenhäuser bei niedrigen Temperaturen höher. Da die spektrale Abhängigkeit der Quarzabsorption nahezu eine Rechtecksform aufweist, kann mit zwei Matrizen gearbeitet werden, die eine im nahen Infraroten $\lambda < 4\mu$m, die anderen im fernen $\lambda > 4\mu$m. Die Matrizen werden dann je nach Wafertemperatur und Strahlungsleistung des Wafers in diesen zwei Bändern gewichtet und addiert.

3. **Kopplung von Wänden/Quarz auf den Wafer**

$$q_i^{\text{wand}} = \sum_{k=1}^{K} W_{ik} \cdot \sigma e_k^{\text{wand}} T_k^{\text{wand}\,4} \tag{3.6}$$

Da die Stahlwände auf etwa Zimmertemperatur $T^{\text{wand}} \approx 300\text{K}$ gehalten werden, ist hauptsächlich die Strahlungswärme des Quarzes der bedeutende Term. Der sich aufheizende Quarz führt hauptsächlich über diesen Strahlungsterm zu der Langzeitdrift der Wafertemperatur. Dieser Term läßt sich aufgrund der zumeist planparallelen Anbringung der Quarze zum Wafer mit hoher Genauigkeit berechnen. $\mathbf{W}$ ist hier die Kopplungsmatrix zwischen den Waferzellen i und allen Wanddiskretisierungselementen $k = 1, \ldots, K$.

Die Matrizen $\mathbf{L}$, $\mathbf{R}$ und $\mathbf{W}$ sind Teilmatrizen der Strahlungsmatrix $\mathbf{G}$ (siehe Kapitel 2.3.3)

$$\begin{pmatrix} \mathbf{q}^{\text{rad},\rightarrow\text{wafer}} \\ \mathbf{q}^{\text{rad},\rightarrow\text{lamps}} \\ \mathbf{q}^{\text{rad},\rightarrow\text{ambient}} \end{pmatrix} = \mathbf{G}\,\sigma\boldsymbol{\epsilon}^{\text{eff}}\mathbf{T}^4 \tag{3.7}$$

$$= \begin{array}{c} N\{ \\ L\{ \\ K\{ \end{array} \left(\begin{array}{c|c|c} \overbrace{R_{ij} - \delta_{ij}}^{N} & \overbrace{L_{il}}^{L} & \overbrace{W_{ik}}^{K} \\ \hline & & \\ \hline & & \end{array} \right) \begin{pmatrix} \epsilon_j^{\text{eff}}\sigma T_j^4 \\ P_l \\ \epsilon_k^{\text{eff}}\sigma T_k^4 \end{pmatrix}$$

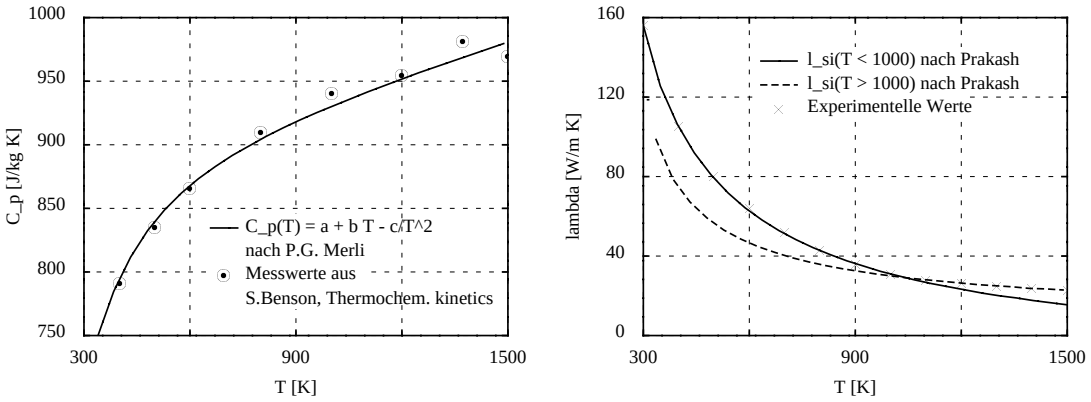


Abbildung 3.1: Wärmekapazität und Wärmeleitfähigkeit von Silizium in Abhängigkeit der Temperatur nach [19] [61].

3.3.2 Modell für den konduktiven und konvektiven Wärmeverlust

Das Modell für den konduktiven und konvektiven Wärmeverlust wird im folgenden nur kurz diskutiert. In [61] und [60] ist die Vorgehensweise bereits genauer erläutert worden, hier folgt nur eine kurze Zusammenfassung der zentralen Gleichungen.

Der konduktive Wärmetransport wird aufgespaltet in die Wärmeleitung innerhalb des Wafers und durch das Gas, was aufgrund des Verhältnisses der Wärmeleitfähigkeiten von $\frac{\lambda_{\mathrm{Si}}}{\lambda_{\mathrm{O}_2}} \approx \frac{25\mathrm{W/mK}}{0.025W/mK}$ gerechtfertigt ist.

Für den Wärmetransport im Wafer in Zylinderkoordinaten ergibt sich daher [14]

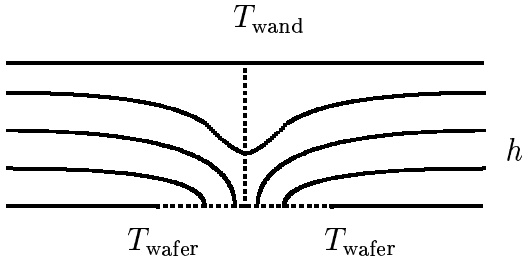
$$q_i^{cond.waf} = \frac{1}{r}\frac{\partial}{\partial r}\left(\lambda_{\mathrm{Si}} r \frac{\partial T}{\partial r}\right) + \frac{1}{r^2}\frac{\partial}{\partial \theta}\left(\lambda_{\mathrm{Si}}(T)\frac{\partial T}{\partial \theta}\right) \tag{3.8}$$

mit den Randbedingungen:

$$\begin{aligned} \lambda_{\mathrm{Si}}\frac{\partial T}{\partial r} &= 0 \text{ in der Wafermitte } r = 0 \\ \lambda_{\mathrm{Si}}\frac{\partial T}{\partial r} &= q(\theta, r_{\mathrm{Wafer}}, z) \text{ am Rand für } r = r_{\mathrm{wafer}} \end{aligned} \tag{3.9}$$

Daraus ergibt sich im zweidimensionalen Fall eine pentadiagonale Kopplungsmatrix in der Gitterdiskretisierung.

Für den Wärmentransport durch das Gas läßt sich eine Lösung der stationären Wärmeleitungsgleichung im axialsymmetrischen Fall angeben [61], [60].



$$q_i^{gas} = \frac{p\Delta T}{h} \frac{\sinh \frac{\pi}{h}x}{\sqrt{(p(\cosh \frac{\pi}{h}x - 1) - 1)^2 - 1}}, \quad p = \frac{2}{\cosh \frac{\pi d}{h} - 1}$$

Abbildung 3.2: Analytische Lösung der Wärmetransportgleichung für den Fall eines Guardrings auf gleicher Höhe neben dem Wafer.

Die Axialsymmetrie, wie sie in [60] für die Lösung der Poissongleichung gefordert wird, ist im vorliegenden Reaktor aufgrund der Türöffnungen nicht voll erfüllt. In [61] wurde ein Gebietszerlegungsverfahren vorgeschlagen, das in ähnlicher Weise auch hier angewendet wird und den Modellfehler reduziert. Auch wird die verstärkte Abkühlung an der Kante mehr durch den Abstand zur darüber- bzw. darunterliegenden Wand bestimmt.

3.3.3 Modell für das Oxidwachstum

Zur Bestimmung der Temperatur aus der Oxiddicke ist die Verwendung eines Oxidwachstumsmodells notwendig. Der hier betrachtete Prozeßschritt ist eine Oxidation von $< 100 >$ Silizium mit Dotierungen von $10^{15} cm^{-3}$ bis $10^{17} cm^{-3}$ bei Temperaturen zwischen 1000°C und 1150°C in reiner Sauerstoffatmosphäre, N_2O und NO mit Drücken zwischen 0.1 und 1 Atmosphäre. Eine detaillierte Beschreibung der Modelle befindet sich in Anhang A.

Zur Berechnung der Oxiddicke auf dem Wafer nach einem Prozeß erhält man somit

$$d_{Ox}(r, \phi) = d_{native} + \int_{t=0}^{t=t_{Proc}} \frac{dx_{ox}(T(t', r, \phi), x_{ox}, p)}{dt'} dt'. \tag{3.10}$$

Hierbei ist d_{native} die Anfangsoxiddicke, die bei unter Normalbedingungen gelagerten Wafern bei etwa $15 - 20$Å liegt, und t_{Proc} die Dauer der Oxidation.

Bei bekannter zeitlicher Drift der Wafertemperatur, die über ein Pyrometer während des Prozesses aufgezeichnet wird, läßt sich das Auffinden der Waferno-

minaltemperatur als Inversion des Integrals (3.10) formulieren. Die Temperatur erhält man so durch Lösen eines eindimensionalen Minimierungsproblems der Form

$$T(0, r_i, \phi_i) = \min \|d_{Ox}(r_i, \phi_i) - d_{meas}(r_i, \phi_i)\| \tag{3.11}$$

wobei $d_{meas}(r_i, \phi_i)$ die gemessenen Oxiddicken an den Meßpunkten r_i, ϕ_i sind.

3.4 Erläuterung des neu entwickelten Verfahrens

In den vorigen Kapiteln wurde die Lampeneinstrahlung als der dominante und am schwierigsten zu modellierende Kopplungsfaktor identifiziert. Diese Größe muß daher aus Messungen extrahiert werden.

Annahme ist, daß der Hauptteil der eingefangen Strahlung am Wafer direkt von den Lampen ohne Reflexion auf dem Wafer eintrifft. Die Gültigkeit dieser Annahme läßt sich über den Sichtwinkel leicht verifizieren. Ferner sind die Reflexionen im Lampenhaus und an den Wänden nicht von der Frequenz des Lichts abhängig, d.h. der optische Weg wird kaum durch die Strahlungstemperatur beeinflußt. Dies ist gerechtfertigt, da die metallischen Wände eine über das Spektrum nahezu konstante Reflektivität aufweisen.

Jede Lampe verursacht dann auf dem Wafer eine Flußverteilung der Form:

$$q^{\text{Lampe}}(r,\phi) = g(r,\phi)r(P)\int_0^\infty e_{\text{Lampe}}(\nu,T)\alpha_{\text{Quarz}}(\nu,T_{\text{Quarz}})I_{SK}(\nu,T)d\nu \quad (3.12)$$

wobei $g(r,\phi)$ einen geometrischen Kopplungsfaktor, $r(P)$ den Anteil des in Strahlung umgewandelten Stroms durch die Lampe, α_{Quarz} den mittleren absorbierten Anteil der Strahlung durch die Quarze und e_{Lampe} die Emissivität der Wolframwendel angibt.

In Kapitel 2 wurde bereits diskutiert, daß $r(P)$ für Leistungen von größer ca. 5% der Maximalleistung sehr nahe an eins liegen muß, d.h. die elektrische Leistung, die durch die Glühwendel fließt, wird vollständig in Licht oder Wärmestrahlung umgewandelt, Randeffekte durch Wärmeableitung durch den Draht finden nicht statt. Die typische Lampentemperatur bei Maximalleistung liegt bei etwa 3200 Kelvin, bei 10% ihrer Leistung also minimal bei 1800K.

Das Extraktionsverfahren hat die Ermittlung der Verteilungsfunktion in Gleichung 3.12 zum Ziel. Dazu wird von einem Grundrezept ausgegangen, das z.B. mit dem rein simulationsgestützten Optimierungsverfahren bestimmt wurde (Abschnitt 2.5). Das genaue Erreichen der Zieltemperatur ist nicht ausschlaggebend. Die Gültigkeit und Anwendbarkeit des Verfahrens zeigt in der praktischen Realisation einen Anwendbarkeitsbereich von etwa ± 125°C. Die Temperaturverteilung dieses Basisrezepts wird z.B. mit einem Thermoelementwafer bestimmt.

Da Thermoelemente nur bis auf einige Grad absolut kalibriert sind, empfiehlt sich eine Referenzmessung z.B. mittels eines Oxidationsprozesses zur Kalibration der Thermoelemente untereinander.

Von diesem Basisrezept ausgehend werden dann die Leistungen in den einzelnen Lampenbänken verändert und die neue Temperaturverteilung aufgezeichnet. Dabei wird die zeitliche Entwicklung der Temperatur, die im allgemeinen schwer meßbar ist nicht benötigt, sondern nur die neuen Endtemperaturen nach der Leistungserhöhung/-erniedrigung.

Ausgehend vom reduzierten Modell (Gleichung 3.3) läßt sich aus dem Satz von Messungen die Strahlungsmatrix der Lampen zum Wafer extrahieren, da die anderen Wärmetransportterme hinreichend gut modelliert werden:

Es ist ersichtlich, dass zumindest so viele Lampeneinstellungen vermessen werden wie unterschiedlich ansteuerbare Lampenbänke in der Kammer vorhanden sind. Eine höhere Anzahl von Messungen verringert jedoch den Einfluß von Meßstörungen.

Schreibt man Gleichung 3.3 und 3.4 um, so erhält man für jede der $m = 1, \ldots, M$ Messungen in N Wafermeßpunkten nach Umsortierung ein Gleichungssystem der Form

$$\begin{pmatrix} -(q_1^{rad.waf} + q_1^{wand} + q_i^{cond.waf} + q_1^{gas}) \\ -(q_2^{rad.waf} + q_2^{wand} + q_2^{cond.waf} + q_2^{gas}) \\ \vdots \\ \vdots \\ -(q_N^{zonen} + q_N^{wand} + q_N^{cond.waf} + q_N^{gas}) \end{pmatrix}_m = \begin{pmatrix} L_{11} & \cdots & L_{1L} \\ L_{21} & & L_{2L} \\ \vdots & \ddots & \vdots \\ \vdots & & \vdots \\ L_{N1} & \cdots & L_{NL} \end{pmatrix} \begin{pmatrix} P_1 \\ \vdots \\ P_L \end{pmatrix}_m \tag{3.13}$$

Die Matrixelemente L_{il} lassen sich mittels eines Verfahrens der kleinsten Fehlerquadrate anpassen.

Die Annahme, die Ankopplung der Lampen zum Wafer sei linear, ist nur näherungsweise richtig. Durch Veränderung der Lampenleistung und damit der Filamenttemperatur kommt es zu einer Verschiebung des Emissionsspektrums der Lampen. Die Wandreflektivität nimmt mit bei höheren Wellenzahlen ab. Der Transmissivität des Quarzes nimmt leicht zu.

Dies kann entweder durch einen Korrekturterm in Gleichung 3.12 oder durch eine Bestimmung der Parameter in der Nähe des Arbeitspunktes erzielt werden.

3.5 Analyse des Verfahrens mit Hilfe der Equipmentsimulation

Im ersten Schritt wird das Verfahren mit Hilfe der Equipmentsimulation auf seine Funktion hin untersucht. Der Strahlungs/CFD-Simulator PHOENICS dient hierbei als Referenzsystem, an dem die "Versuchsserie" durchgeführt wird.

Der Simulator verhält sich ähnlich wie ein RTP-System. Zwar stimmen die Simulationsergebnisse mit dem realen Equipment nicht exakt überein, das physikalische Verhalten einer typischen Kammer wird aber hinreichend wiedergegeben. Dadurch lassen sich Analysen über die Stabilität des Verfahrens gegenüber Störungen der Thermoelemente analysieren.

3.5.1 Beschreibung der Simulationen

Als Grundlage der Parameterextraktion dient ein Satz von insgesamt 28 Messungen. Es wird von einem Basisrezept der Form

$$\mathcal{M}_0 := (P_0^1, P_0^2, \ldots, P_0^L) \tag{3.14}$$

mit den Lampenleistungen P für alle Lampenbänke 1 bis L ausgegangen. Die aus dem Basisrezept resultierende Temperaturverteilung $T(r, \phi)$ sollte innerhalb des typischen Arbeitsbereichs der Kammer liegen, im vorliegenden Fall etwa 1050°C. Die Grundeinstellung der Lampen kann in der Regel geschätzt werden, so daß eine akzeptable Uniformität erreicht wird. In der Praxis ist dies von Vorteil, um den thermischen Streß auf dem Meßwafer gering zu halten.

Ausgehend von diesem Basisrezept werden die Lampenleistungen der einzelnen Zonen nach der axialen Variationsmethode [46] einzeln verändert. Der Satz von Messungen ergibt sich zu:

$$\mathcal{M} := \bigcup_m P_m, P_m = (P_0^1, \ldots, P_0^{(m-1)}, \alpha_m P_0^m, P_0^{(m+1)}, \ldots, P_0^L)\} \tag{3.15}$$

mit

$$0 \leq \alpha_i \leq \frac{P_{\max}^i}{P_0^i} \tag{3.16}$$

Die axiale Variationsmethode aus dem Bereich des Experimental Designs erweist sich im vorliegenden Fall als günstig: Zum einen ist die Anzahl der Aktuatorgrößen klein genug (maximal etwa 40), um alle Parameter auch in mehreren

Stufen variieren zu können. Zum anderen erlaubt die gezielte Variation einzelner Lampe eine gute Kontrolle der Messungen, die bei Zufallsverfahren in der Regel nicht möglich ist. Die Kopplung zwischen den Aktuatorgrößen ist im vorliegenden Fall zu vernachlässigen, so daß eine vollständige Abtastung des Aktuatorraums wie bei der zweidimensionalen Boxmethode

$$\mathcal{M} := \bigcup_m \bigcup_{n \neq m} (P_0^1, \ldots, P_0^{(m-1)}, \alpha_m P_0^m, P_0^{(m+1)}, \ldots, P_0^{(n-1)}, \alpha_n P_0^n, P_0^{(n+1)}, P_0^L) \tag{3.17}$$

mit

$$0 \leq \alpha_i \leq \frac{P^i_{\max}}{P_0^i} \tag{3.18}$$

nicht notwendig ist.

Die Wahl der Parametervariationen muß einerseits groß genug sein, um die numerische Stabilität des Verfahrens zu gewährleisten. Entscheidend hierfür ist eine kleine Konditionszahl

$$K(\mathcal{M}) := \left| \frac{\lambda_{max}(\mathcal{M})}{\lambda_{min}(\mathcal{M})} \right| \tag{3.19}$$

der Matrix $\mathcal{M}$ in Gleichung 3.17, wobei λ_{max} und λ_{min} die betragsmäßig größten bzw. kleinsten Eigenwerte sind. Sie sinkt mit steigender Variation, da die Matrix der Lampenleistungen diagonaldominant wird. Das Verfahren wird so stabiler. Bei ungünstig positionierten Thermoelementen wie in Kapitel 3.6 können andererseits zu große Variationen Schwierigkeiten bei der Interpolation verursachen. Vorteilhaft ist hier eine Anordnung wie in Abbildung 3.3 dargestellt.

Prinzipiell lassen sich auch andere Verfahren mit verbessertem statistischen Verhalten, wie z.B. das Latin-Hypercube Verfahren verwenden. Hier werden die Messungen statisch gezogen, wobei jede Leistung durch eine Zufallszahl zwischen 0 und der Gesamtzahl der durchzuführenden Messungen ermittelt wird:

$$\mathcal{M} := \bigcup_m (P_m^1, \ldots, P_m^L) \tag{3.20}$$

$$j_m^i = \mathrm{ran}(0, M-1), j_m^i \neq j_{m-1}^i \neq \ldots \neq j_1^i \tag{3.21}$$

$$P_m^i = P^i_{\min} + \frac{j_m^i}{M-1}(P^i_{\max} - P^i_{\min}) \tag{3.22}$$

Bei M Messungen wird also jeder Parameter in jedem Unterbereich mit der Granularität $M-1$ einmal ausgewertet. Die Ergebnisse sind dann aber nicht

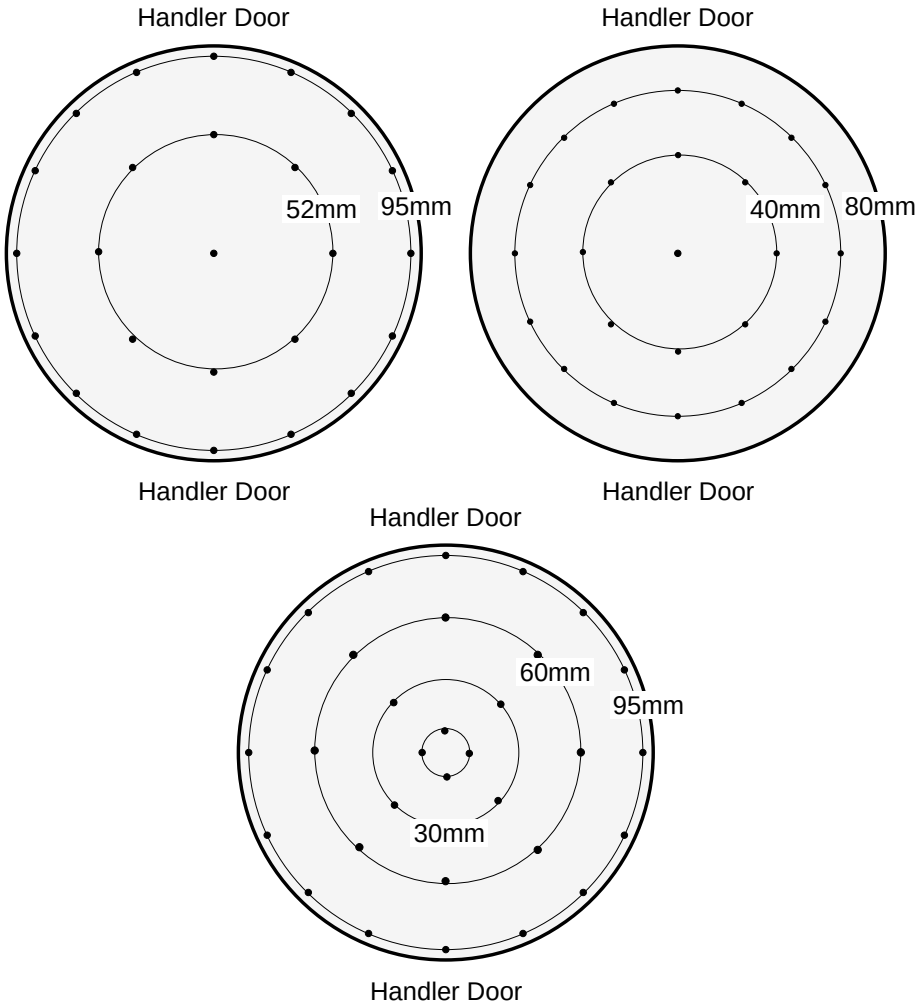


Abbildung 3.3: Anordnung der Thermoelemente auf dem Wafer. Oben: Anordnung für die Studie des Verfahrens in der Equipmentsimulation. Mitte: Experimentelle Anordnung mit nach innen verrückten äußerstem Ring. Unten: Vorschlag für eine an das Verfahren angepaßte Anordnung.

mehr manuell auf ihre Korrektheit überprüfbar und es treten große Temperaturschwankungen auf dem Wafer auf. Vorteilhaft ist hier das Entstehen einer Zufallsmatrix in Gleichung 3.21, die im allgemeinen gut konditioniert ist.

Die Lampenhäuser wurden in insgesamt 28 Bänke aufgeteilt, im unteren Lampenhaus wurden die Lampen paarweise, also zu insgesamt 9 Bänken zusammengefaßt. Wie noch gezeigt wird, ist diese ad hoc Zusammenfassung gut genug, um eine akzeptable Uniformität zu erreichen. Bei einer Veränderung der Zusammensetzung wurde keine Verbesserung der Homogenität beobachtet.

Die Lampenleistungen aus Gleichung 3.15 werden nacheinander simuliert, die Temperaturen an den 25 Thermoelementpunkten extrahiert und daraus mit einem Verfahren der kleinsten Fehlerquadrate die Strahlungsmatrix extrahiert. Die Anordnung der Thermoelemente ist die übliche Standardkonfiguration von 8-Zoll-TC-Wafern.

3.5.2 Verifikation des reduzierten Modells

Die simulierten "Messungen" wurden bei Temperaturen von etwa 1050°C durchgeführt. Mit dem oben beschrieben Vorgehen werden die eingegebenen Meßwerte gut reproduziert, da die Parameter an diese Messungen angepaßt wurden. Um die Gültigkeit des reduzierten Modells aber auch bei anderen Temperaturen zu testen, wurde eine Lampeneinstellung mit einer deutlich höheren Gesamtleistung simuliert.

Zur Analyse wird hierzu zunächst eine hohe Bedeckung des Wafers mit Thermoelementen angenommen, um Effekte aufgrund der geringen Zahl von Thermoelementen getrennt beurteilen zu können. In Abbildung 3.4 ist der Vergleich des reduzierten Modells mit den PHOENICS Simulationen in zwei Schnitten, einmal entlang der Verbindungslinie der beiden Türen und einmal senkrecht dazu, dargestellt. Das reduzierte Modell gibt also Temperaturverteilung und -niveau sehr gut wieder.

Ein genauerer Vergleich des Wafermodells mit Reaktorsimulationen ist [60] zu entnehmen.

3.5.3 Optimierungsergebnisse

Mit der Standardanordnung der 25 Thermoelemente und der daraus gewonnen Strahlungsmatrix wurden nun die Lampenleistungen so optimiert, daß eine

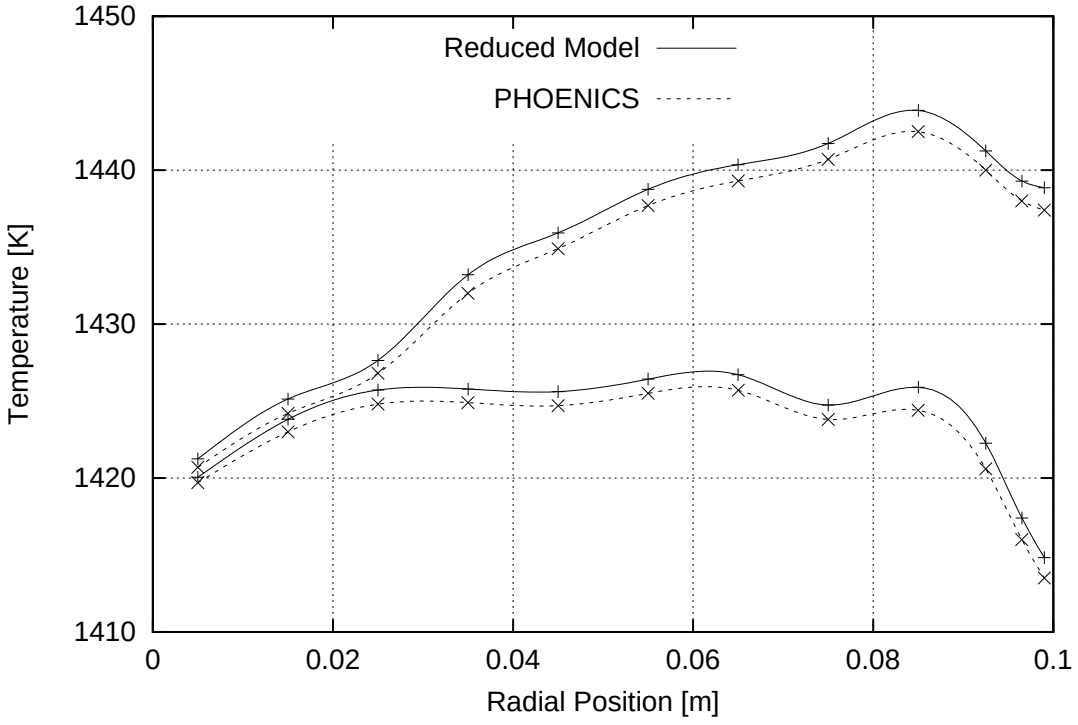


Abbildung 3.4: Vergleich des reduzierten Modells mit der vollständigen Lösung der dreidimensionalen Navier-Stokes-Gleichung. Die Zeichnung zeigt zwei Schnitte, einmal entlang der Verbindungslinie (obere Kurve) und einmal senkrecht dazu.

vorgegebene Wafertemperatur über den ganzen Wafer in der numerischen Diskretisierung erzielt wird. Das dazu verwendete Zweischrittverfahren ist in [60] und [61] näher erläutert. Die Grundidee im ersten Schritt besteht in dem Ausgleich der Wärmeflußbilanz für jedes Waferelement durch Variation der Lampenleistungen P_l fr alle Lampenbänke $l = 1, \ldots, L$. Daraus ergibt sich ein Arbeitspunkt $\mathbf{P}_0$ für die Lampenleistungen. Im zweiten Schritt wird dann der verbleibende Temperaturgradient auf dem Wafer durch Variation der Lampenleistungen weiter minimiert.

Als Zieltemperatur in Abbildung 3.5 wurde 1150°C angegeben. Die berechneten, optimierten Lampenleistungen wurden mit PHOENICS simuliert, was der Verifikation am realen Equipment entspricht. Die Ergebnisse sind in Abbildung 3.5 dargestellt. Trotz der hohen Temperatur wird eine gute Standardabweichung von 2.3°C und eine maximale Temperaturdifferenz – definiert durch $(T_{\max} - T_{\min})/2$ – von 4.5°C erzielt. Hierbei handelt es sich um die Temperaturhomogenität über den *gesamten* Wafer bis zum Rand, nicht nur zwischen den Meßpunkten, an denen ursprünglich die Matrix extrahiert wurde.

Nimmt man einen Randausschlußbereich von 3mm an, so verbessert sich die Standardabweichung zu etwa 2°C und die maximale Temperaturschwankung zu 4°C.

Grund für die begrenzte Temperaturuniformität sind – wie in Kapitel 2 diskutiert – die geringen Maximalleistungen der Lampen. Dies wird insbesondere deutlich, wenn die Optimierung auf eine Untermenge von Lampen begrenzt wird.

3.5.4 Stabilitätsanalyse gegen Meßfehler

Bisher wurde angenommen, daß die Meßsignale der Thermoelemente – von numerischen Schwankungen der PHOENICS Rechnungen abgesehen – frei von Meßfehlern sind. Das Verfahren ist zwar so angelegt, daß nur der zeitliche Mittelwert der Thermoelementsignale ausgewertet wird, jedoch können auch hier in der Praxis z.B. durch elektrische Einstreuung Schwankungen von einigen Grad auftreten.

Um die Sensitivität der Verfahrensweise auf derartige Fehler zu untersuchen wurde ein normalverteiltes Rauschen zu den simulierten Verteilungen addiert und die Extraktions- und Optimierungsverfahren erneut durchgeführt.

Die extrahierte Strahlungsmatrix erhält durch die addierten Störungen statistische Fehler, was auch die Temperaturhomogenität am berechneten Optimum

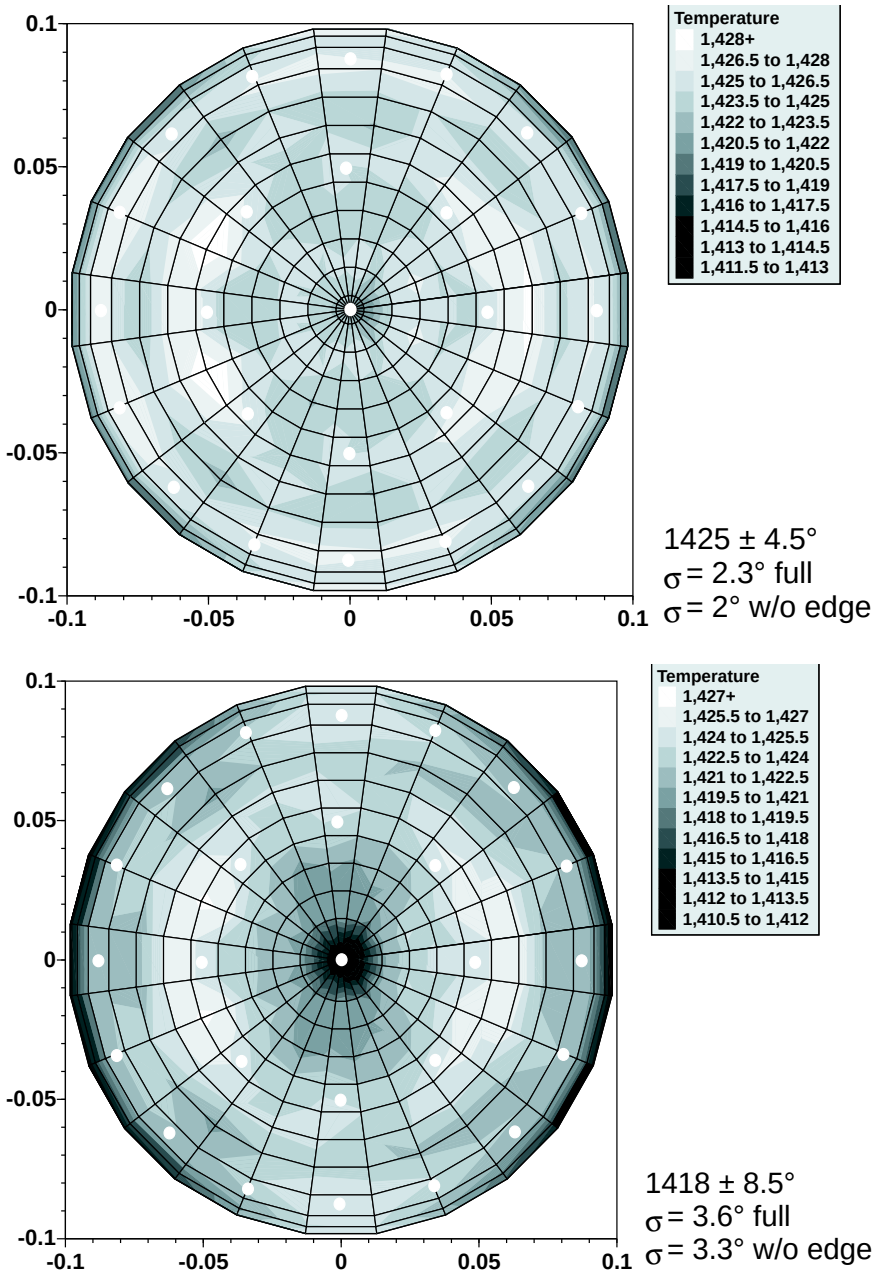


Abbildung 3.5: Optimierungsergebnisse für 1150°C mit der extrahierten Strahlungsmatrix. Oben: Alle Lampen wurden bei der Optimierung verwendet, der Abstand zwischen maximaler und minimaler Temperatur beträgt 9°C. Unten: Ohne die äußersten Linearlampen und die Zentrallampe im oberen Lampenhaus verschlechtert sich die Uniformität auf ±17°C.

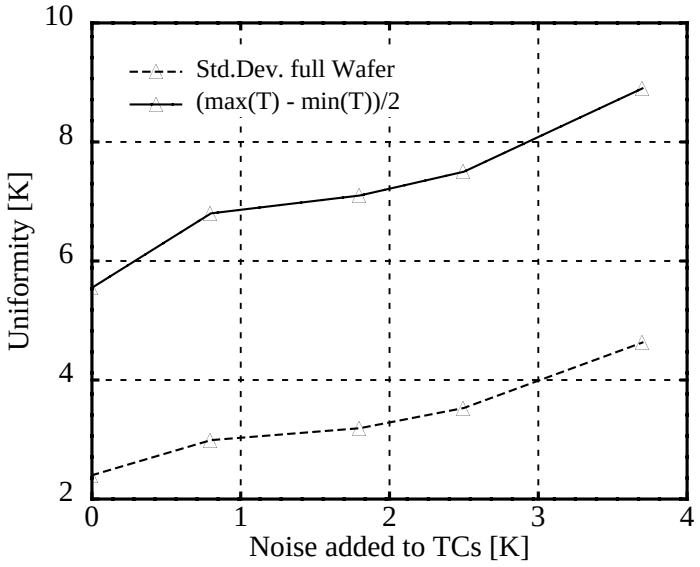


Abbildung 3.6: Zunahme der Standardbreite der Temperaturhomogenität in Abhängigkeit der Störungen der Thermoelemente. Für die zu erwartenden Meßsignalfehler von etwa $\pm 1°C$ ist die Zunahme nur geringfügig.

verschlechtert (Abbildung 3.6). Allerdings ist die Verschlechterung bei kleinen Fehlern in den Thermoelementsignalen gering. Die Lampenleistungen für unterschiedlich stark verrauschte Sensorsignale sind in Abbildung 3.7 dargestellt. Die Verteilung in den unteren Bänken (linker Teil der Balken) weisen große Ähnlichkeiten auf, die Leistungen für Bänke mit wenigen Lampen zeigen leichte Unterschiede, da eine Leistungsänderung dort nur eine geringe Temperaturänderung bewirkt.

Wie in Kapitel 3.6 nachgewiesen wird, sind statistische Schwankungen der Thermoelemente von besser als $\pm 1°C$ erreichbar. Bei besonders gut kalibrierten Thermoelementen wird auch schon von einer Reproduzierbarkeit von besser als $0.5°C$ berichtet [45].

3.6 Experimentelle Durchführung des Verfahrens

Das oben beschriebene Verfahren wurde analog zum Vorgehen bei der Simulationsstudie meßtechnisch durchgeführt.

Die Messungen wurden bei geringem Druck in Argonatmosphäre mit Hilfe eines

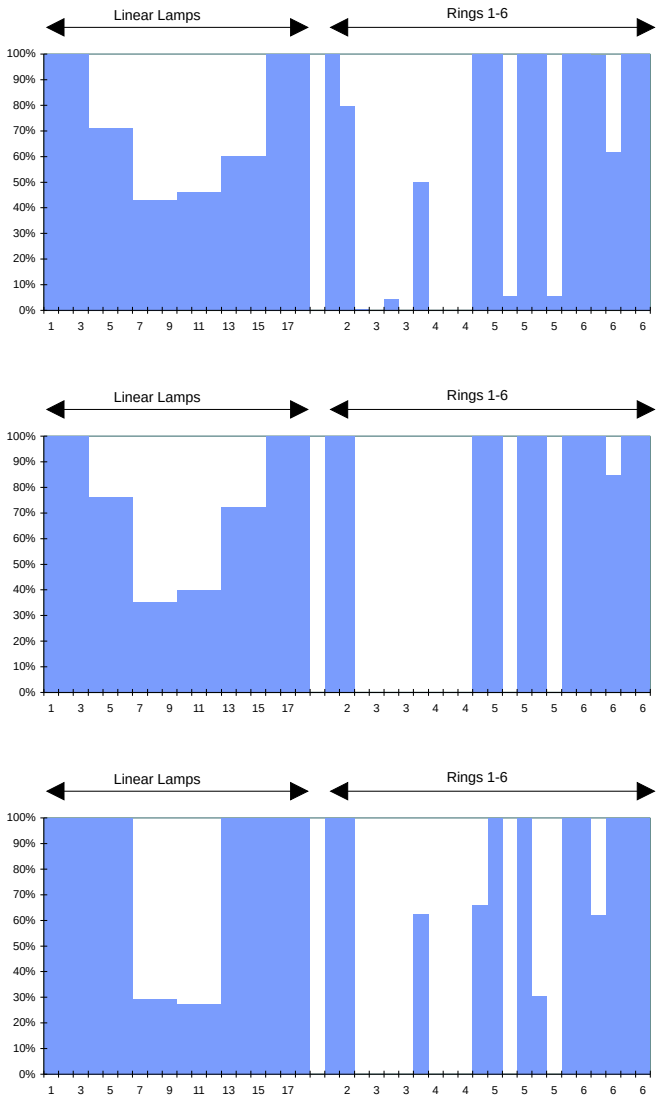


Abbildung 3.7: Leistungsverteilungen in den Bänken
(a) Optimales Rezept mit 288 Meßpunkten, was dem Grenzfall sehr vieler Thermoelemente entspricht (“echtes” Optimum)
(b) Optimum mit der aus 25 TCs extrahierten Matrix an den Standardpositionen, aber ohne Signalstörung
(c) Optimum mit der aus 25 TCs extrahierten Matrix mit einem Rauschen von $\pm 2°C$

Thermoelementwafers mit 25 Thermoelementmeßpunkten der Firma Sensarray vorgenommen. Vorteil dieses Meßverfahrens ist die rasche Durchführbarkeit der Meßserie. Im vorliegenden Fall wurden etwa 30 verschiedene Kombinationen der Lampenleistungen verwendet. Einschließlich Aufbau und Justage erfordert dies wenige Stunden Meßzeit an der Kammer. Die Auswertung und Optimierung verursacht keinen weiteren Ausfallzeiten der Kammer.

Die Installation und der Anschluß der Thermoelemente erfordert einige Sorgfalt. So ist der Einsatz von Rausch- und Integrationsfiltern sowie eine gemeinsame, konstante Referenztemperatur für die Thermoelemente unverzichtbar. Bei der Drahtführung im Kammerinneren ist auf eine möglichst geringe Abschattung anderer Thermoelemente zu achten, die zu Ergebnisverfälschungen führen könnten. Auch sollten die Meßdrähte nah am Wafer geführt werden, um einen starken Wärmegradient im Draht zu verhindern, der den Meßwert des Thermoelements beeinflussen kann.

Die Methodik verlangt nicht zwangsläufig den Einsatz von Thermoelementwafern. Es wurden andere ortsaufgelöste Verfahren über kompensierte Pyrometer [44] oder über akkustische Ausbreitungsgeschwindigkeiten im Wafer [42] [43] zur Wafertemperaturmessung vorgestellt. Ihr Vorteil liegt in der wenn auch limitierten Einsetzbarkeit auch während des Prozessierens selbst. Die Anzahl der gewonnen Meßpunkte ist allerdings klein (maximal 5 bis 10 bei 8-Zoll Wafern) und die Temperaturgenauigkeit nicht ausreichend (etwa 5 bis 10 Grad). Auch sind umfangreiche Umbaumaßnahmen an der Kammer notwendig.

Ebenso könnte über eine thermische Oxidation über die aufgewachsene Oxidschichtdicke die Temperaturverteilung bestimmt werden. Vorteil wäre die deutlich längere Integrationszeit über die Temperatur, die hohe Informationsdichte und höhere Ausfallsicherheit. Allerdings sind Aufwand und Waferkosten deutlich höher. Nachteil der hohen Integrationszeit ist der Verlust einer zeitlichen Auflösung, so daß Drifteffekte, wie in Kapitel 2 diskutiert, anderweitig bestimmt und herausgerechnet werden müssen. Die Oxidationsmessung eignet sich daher als Temperaturnormale zur Eichung der Thermoelemente.

3.6.1 Thermoelementmessungen und Regelbarkeit des Systems

Wichtig bei der Thermoelementauswertung ist die Signalextraktion aus den Meßwerten. Da Thermoelemente nach mehrfacher thermischer Belastung durch Rampen und Abkühlvorgängen altern und daher zu verschiedenartigen Defekten, wurde ein fehlertoleranter Auswertealgorithmus entworfen.

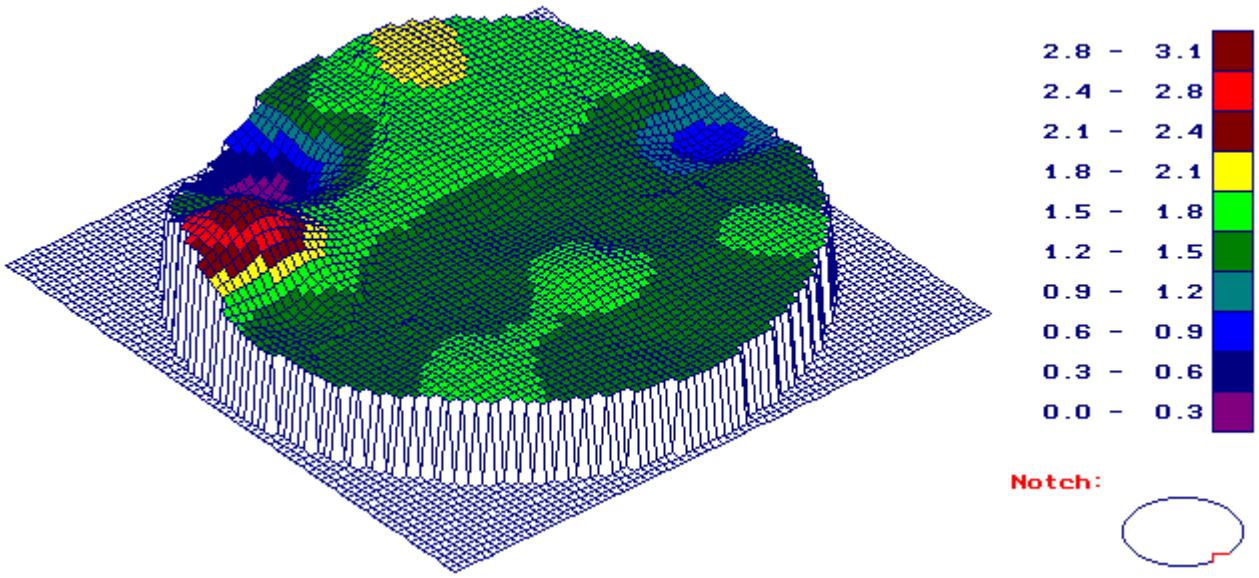


Abbildung 3.8: Reproduzierbarkeit der Thermoelementmessungen im Abstand von einem Tag.

Bei Totalausfall eines Thermoelements kann durch Symmetrieüberlegungen gegebenenfalls ein äquivalentes Signal verwendet werden. Symmetriebedingungen können überdies zu Varianzreduktion eingesetzt werden. Die Stabilität des Interpolationsverfahrens wird freilich durch zunehmende Ausfälle beeinträchtigt.

Da die Anzahl der Thermoelemente gering ist, erkennt der Algorithmus temporäre Aufälle und extrahiert zu jedem Prozeßschrittwechsel die mit einigen Bandfiltern extrahierten Werte.

Als Ausgangspunkt wurde ein Basisrezept bei etwa 1070°C gewählt und die Einstellungen analog zu der simulationsgestützten Extraktion durchgeführt.

Bis auf zwei stärker schwankende Thermoelemente (Abbildung 3.8) sind die extrahierten Werte reproduzierbar. Eine Differenz zwischen zwei Messungen wird durch unterschiedliche Quarztemperaturen hervorgerufen. Dieser Effekt wird aber in dem reduzierten Modell 3.3 berücksichtigt.

3.6.2 Optimierungsergebnisse

Analog zum Vorgehen in der Simulation wurde nach Extraktion der Strahlungsmatrix eine Optimierung durchgeführt. Die Ergebnisse für zwei Temperaturen

sind den Abbildungen 3.9 und 3.10 zu entnehmen. Die Nominaltemperatur wurde bei der Messung in etwa erreicht, für die Uniformität ergibt sich ein sehr guter Wert von nur etwa $\pm 3°C$ bei 1050°C bei einem Randausschluß von $3mm$. Unter Hinzunahme des Randes verschlechtert sich die Uniformität auf $\pm 6°C$.

Da es sich hier um das inverse Problem zur Simulation handelt, wirken sich Fehler – wie in der Optimierung allein aus der Simulation heraus – besonders stark aus.

In den open-loop – also ohne Regelung der nominellen Wafertemperatur gefahrenen – Messungen wurden die Zieltemperaturen nicht voll erreicht. Eine mögliche Ursache liegt in unterschiedlichen optischen Eigenschaften zwischen dem Thermoelement- und dem Prozeßwafer. Der Thermoelementwafer weist ein dickeres Oxid auf; seine effektive Absorptivität ist damit höher als die von reinen Silizium. Ist die effektive Absorptivität des Wafers bekannt, so kann sie aber in Gleichung 3.3 bei der Optimierung eingeführt werden.

Die Uniformität der gemessenen Oxiddicken liegt in dem aus der Simulation vorhergesagten Bereich und erfüllt die an das Equipment gestellten Anforderungen bei Temperaturen um 1050°C, unter Randausschluß auch bei den höheren Temperaturen.

Aufgrund der ungünstigen Anordnung der Thermoelemente im Fall der Messungen sind kaum Meßinformationen über die Randbereiche des Wafers verfügbar. Folge davon ist, daß die Kalibration des reduzierten Modells aufgrund der notwendigen Extrapolation numerisch schlechter konditioniert ist. Die Wafertemperatur am Rand fällt deutlich stärker ab als vom reduzierten Modell prognostiziert. Abhilfe schafft hier ein verbesserte Anordnung der Thermoelemente (Abbildung 3.3) oder ein numerisch aufwendigeres Extraktionsverfahren (Kapitel 3.7).

3.7 Entwurf für ein erweitertes Extraktionsverfahren

Das in den vorangegangenen Abschnitten durchgeführte Verfahren liefert eine gute Uniformität bei Temperaturen von 1050 bis 1150 Grad. Jedoch ist ein Temperaturabfall innerhalb der letzten 10 Millimeter noch vorhanden, der auf die ungünstigen Positionen der Thermoelemente für die Umrechnung auf das numerische Gitter im vorliegenden Fall zurückzuführen ist.

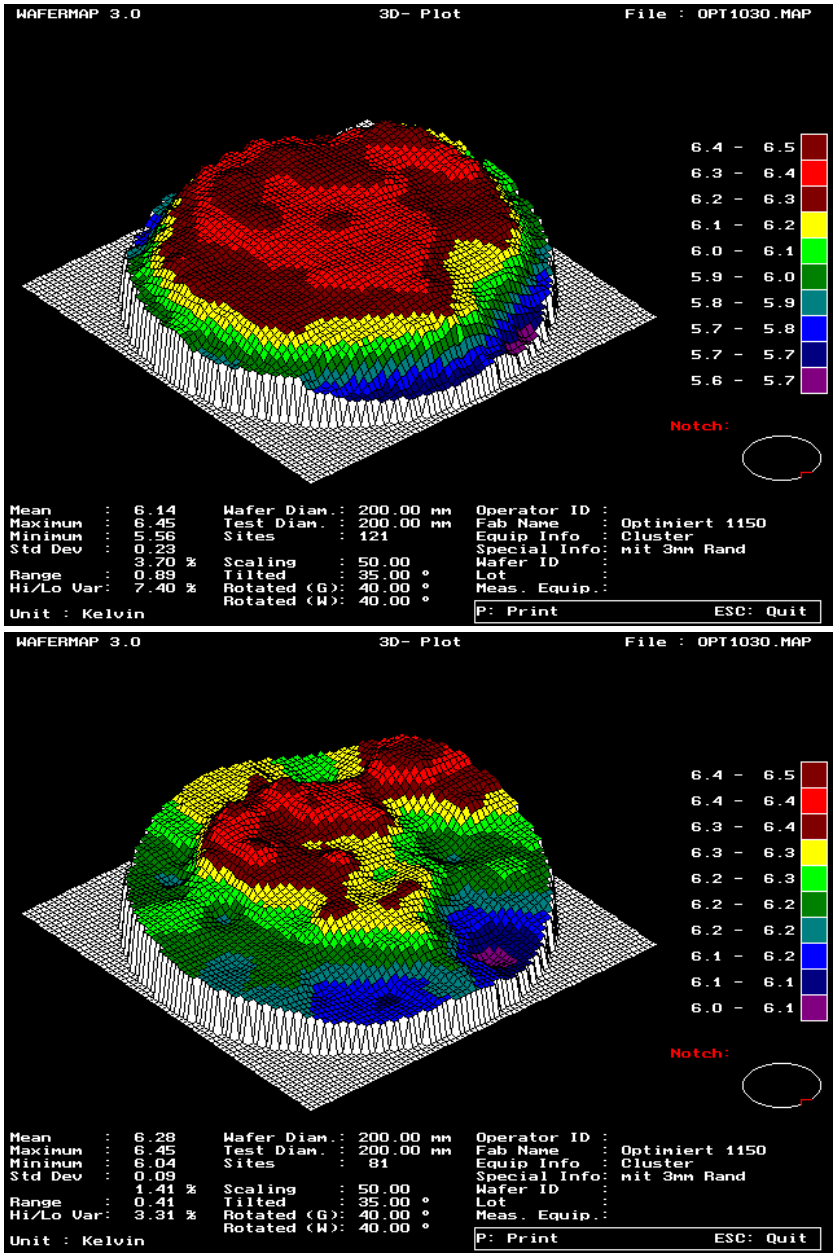


Abbildung 3.9: Oben: Oxiddickenverteilung einer 120-Punkt-Messung bei den optimierten Ergebnissen für 1050°. Die erreichte Nominaltemperatur ergibt sich mit dem in Kapitel 3.3.3 beschriebenen Han/Helms Modell zu 1040°C. Unten: Mit einem Randausschluß des 3mm-Ringes ergibt sich die Temperaturvariation zu ±3°C. Einschließlich der Randzone ergibt sich ±6°C.

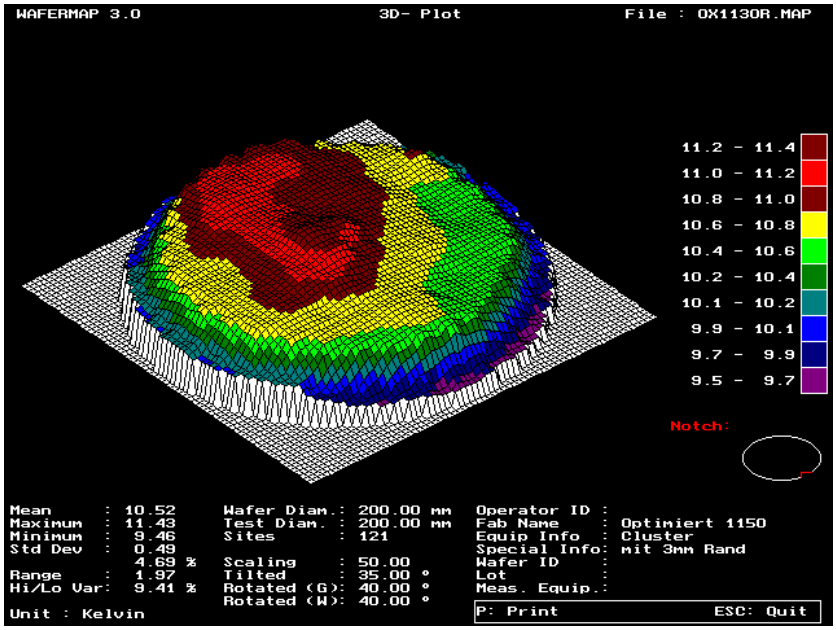


Abbildung 3.10: Oben:Oxiddickenverteilung einer 120-Punkt-Messung bei den optimierten Ergebnissen für 1150°. Die erreichte Nominaltemperatur ist – mit dem in Kapitel 3.3.3 erläuterten Verfahren – 1135°C. Die Temperaturschwankung beträgt ±6°C im Waferinneren, mit Rand ±11°C (Unten).

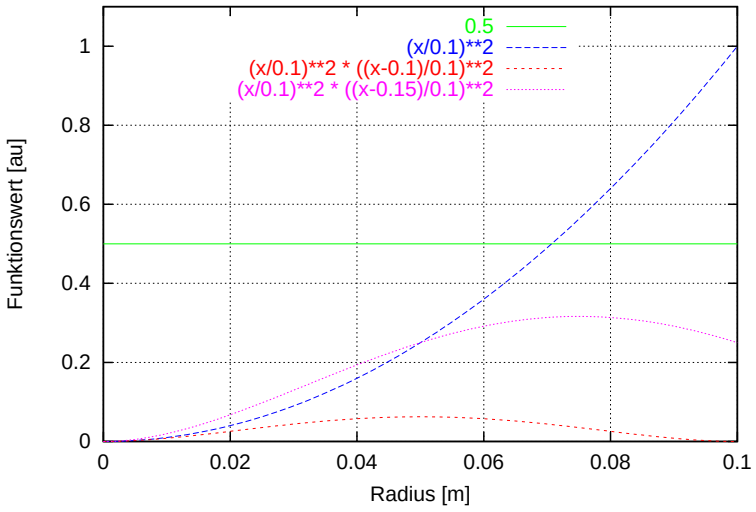


Abbildung 3.11: Radialer Schnitt durch vier der Entwicklungsfunktionen, die für das erweiterte Parameterextraktionsverfahren verwendet wurden.

Im folgenden wird deshalb ein Verfahren vorgestellt, das auf Kosten eines erhöhten numerischen Aufwands die Lampenverteilungen extrahiert. Die Grundidee bleibt die Extraktion der wichtigen physikalischen Größe der Lampenverteilungen aus Messungen.

Im vorangegangen Kapitel wurde dies im Ortsraum durchgeführt. Da nicht an allen Punkten Messungen vorlagen, wurde die Annahme einer im Mittel bilinearen Verteilung in den zwei Koordinaten der Temperatur gemacht. Bei Punkten, an denen extrapoliert werden muß, sind diese Verfahren aber bekanntermaßen numerisch nicht mehr stabil [100]. Zur Verdeutlichung dieses Sachverhalts nehme man einen Meßpunkt auf einem Drittel und einen auf Zweidritteln des Waferradius an. Beide Punkte seien anfangs auf gleichem Temperaturniveau, daraus ergibt sich in diesem eindimensionalen Fall eine homogene Temperaturverteilung. Der äußere Meßpunkt werde nun um 1 Grad gestört. Zwischen den Meßpunkten entsteht dadurch ein Fehler < 1 Grad, am Waferrand durch die Extrapolation verstärkt sich der Fehler jedoch auf 2 Grad. Der Fehler wirkt sich allerdings im oberen Verfahren aufgrund der statistischen Auswertung weniger dramatisch aus.

Die Interpolationsmethode ist für Systeme vorteilhaft, in denen die Lampenverteilungen nicht bekannt sind. In Wirklichkeit existiert – aus der Anschauung oder aus Simulationen – bereits eine Vorstellung über die Verteilung.

Daher wird im folgenden die Entwicklung der Lampenverteilungen auf dem

Wafer in einem Funktionenraum diskutiert. Die Kopplung einer Lampenbank l zum Wafer wird als Linearkombination von N_l Funktionen

$$L_l(r,\phi) = \sum_i^{N_l} \xi_{li} f_{li}(r,\phi). \tag{3.23}$$

dargestellt. Als Basisfunktionen $f_{li}(r,\phi)$ kommen zum einen die simulierten oder mit dem Standardverfahren extrahiertem Strahlungsprofile sowie analytische Funktionen in Betracht (Abbildung 3.11). Die Anzahl der Entwicklungskoeffizienten ist durch die zur Verfügung stehende Anzahl von Meßpunkten beschränkt. Da die Anzahl der Meßpunkte und damit N_l klein ist, ist eine Aufspaltung der Form $f_{li}(r,\phi) = g_l(r)h_i(\phi)$, z.B. mit sphärischen Besselfunktionen, nicht möglich. Geeigneter ist der Verzicht auf einen orthonormalen Basissatz und die Anwendung von Funktionen, die die simulierten Lampenintensitätsprofile annähern.

Die Koeffizienten in Gleichung 3.23 werden durch ein nichtlineares Minimierungsproblem der Form

$$\min_{\substack{\xi_{lj} \\ l=1,\ldots,L, \quad j=1,\ldots,N_l}} \delta T \tag{3.24}$$

mit

$$\delta T = \sum_m \sum_{p_m} ||T^{\mathrm{sim}}(r(p_m),\phi(p_m),\mathbf{L},\mathbf{P}_m) - T^{\mathrm{exp}}_{p_m}|| \tag{3.25}$$

ermittelt; dabei gehen die Summen über alle zur Verfügung stehenden Messungen m und die dazu gehörenden Meßpunkte p_m mit den Meßwerten T_{p_m} zu den Lampeneinstellungen P_m. Summiert wird über die Differenz der Meßwerte zu den simulierten Temperaturen mit der Kopplungsmatrix für alle Diskretisierungselemente k mit Oberfläche A_k

$$\mathcal{L}_{kl} = \oint_{A_k} \sum_i^{N_l} \xi_{li} f_{li}(r,\phi) d\sigma_k \tag{3.26}$$

Die Interpolationsproblematik insbesondere bei ausgefallenen Meßpunkten tritt hier nicht mehr auf, da nur über die tatsächlich aufgezeichneten Werte einer Messung summiert wird. Die Anzahl der zu bestimmenden Koeffizienten ist in der Regel hoch (ca. 10 bis 15 pro Bank).

Der numerische Aufwand der Optimierung wird durch eine mehrstufige Optimierung verringert:

1. Skalierung der Gesamtmatrix, $\min_\lambda \delta T$, $\mathcal{L}_{kl} = \lambda \mathcal{L}^0_{kl}$, Dimension des Optimierungsproblems: eins

2. Skalierung der Bänke, $\min_{\lambda_l} \delta T$, $\mathcal{L}_{kl} = \lambda_l \mathcal{L}^0_{kl}$, Dimension: l

3. Skalierung der Verteilung innerhalb einer Bank, $min_{\lambda_{j_l}} \delta T$, l Optimierungen mit Dimensionen λ_{j_l}

4. Vollständiges Minimierungsproblem (Gleichung 3.25) der Dimension $d = \sum_l j_l$. Typischerweise ist $d > 200$, so daß dies nur mit obigen Vorkonditionierungen numerisch lösbar ist

Als Optimierungsverfahren eignet sich das Verfahren von Brent [99] (Anhang B.1.2).

Ferner ist eine Berechnung der stationären Lösung des reduzierten Modell für alle Messungen erforderlich. Ein Newtonverfahren erfordert die Inversion der Jacobi-Matrix, die aufgrund der Strahlungskopplung voll besetzt ist. Eine Lösung mittels Fixpunktiteration und linearisierten Quelltermen ist daher vorzuziehen. Die daraus resultierende, schwach besetzte Kopplungsmatrix ist aufgrund der Abstrahlungsterme stark diagonaldominant und läßt sich effizient mit dem Verfahren der stabilisierten bikonjugierten Gradienten (BiCGSTAB) [102] und einem Jacobi-Präkonditionierer lösen.

3.8 Zusammenfassung

Wie in den vorigen Abschnitten gezeigt, führt nur eine Kombination aus meßtechnischen und simulationsgestützten Verfahren zum Erfolg. Die Simulation enthält nicht alle Material- und Geometrieparameter mit hinreichender Genauigkeit, um die dominanten Wärmetransporte und damit die Wafertemperatur mit einer Genauigkeit von wenigen Promille zu berechnen. Die Meßtechnik ohne Korrekturen durch geeignete Modelle eignet sich aufgrund meßtechnischer Schwierigkeiten und eingeschränktem Gültigkeitsbereich auch nur bedingt. Erst die sorgfältige Kombination in einem semiempirschen Verfahren liefert ein den Genauigkeitsanforderungen genügendes Verfahren.

Einige Konzepte, wie ein ähnliches Vorgehen auch in anderen Systemen zum Erfolg führen könnte, sind in Kapitel 5 zusammengestellt.

Kapitel 4

Modellbasierte Steuer- und Regelverfahren

4.1 Problemstellung

Die in den vorhergehenden Kapiteln erzielten Verbesserungen durch Anwendung der physikalischen Optimierung zielten zum einen auf die Verbesserung des Gerätes (Kapitel 2), zum anderen auf die Optimierung des Prozeßergebnisses (Kapitel 3) ab.

Der Ansatz, durch konsequente physikalische Modellbildung den Prozeß zu optimieren, wird in diesem Abschnitt fortgeschrieben. Im folgenden wird gezeigt, daß die physikalische Modellierung auch beim Reglerentwurf und der Auswertung des Sensorsignals Anwendung findet. Auch hier bedarf es je nach Anwendung den Entwurf eines reduzierten physikalischen Modells.

Im Rahmen dieser Promotion wurde das Verfahren der Regleradaption auf zwei Reaktortypen angewendet: Zum einen auf einen Reaktor zur Silizidierung von Titan (Rapid Thermal Annealing). In dem untersuchten Reaktor tritt das Problem des sogenannten "first-wafer" Effekts auf, also eine Abhängigkeit des Prozeßergebnisses von der Waferposition innerhalb eines Loses. Mit Hilfe der Simulation des Gesamtsystems, also Reaktor und Regler, läßt sich der auftretende Effekt quantitativ verstehen und das Reglerverhalten optimieren. Aus den experimentellen Daten allein ist eine Differenzierung der Einflüsse nicht möglich.

Den zweiten Teil bildet die Reduktion des Struktureffekts ("pattern-effect"). Dabei handelt es sich um das Auftreten einer Temperaturinhomogenität und damit Ergebnisschwankung, die vornehmlich in lampengeheizten Systemen auf-

tritt und auf der Größenordnung der Bauelemente aufgrund örtlich variierender optischer Eigenschaften des Wafers liegt. Entscheidend ist hier, daß zwar Veränderungen an der Kammer möglich sind, um ähnlich wie in Kapitel 2 den Regleraufwand zu verringern, jedoch negative Auswirkungen auf das Prozeßergebnis durch die Modifikationen zu erwarten sind. Daher wird in der Equipmentsimulation und mit Hilfe eines vereinfachten Modells gezeigt, wie eine modellgestützte Regleradaption ("model scheduled control") den Struktureffekt verringern kann.

Als Ausblick auf die Anwendung optimaler Regelungsstrategien wird die Trajektorienermittlung unter Berücksichtigung des thermischen Budgets als Mehrzieloptimierung diskutiert.

4.2 Analyse des Reglerverhaltens in einem Suszeptorsystem

Einzelscheibenprozesse, bei denen der Wafer auf einem resistiv geheizten Suseptor aufliegt, versprechen aufgrund der gleichmäßigen Beheizung eine gute Temperaturuniformität und eine geringere Abhängigkeit der Wafertemperatur von den optischen Eigenschaften. Die Uniformität der Wafertemperatur hängt in erster Linie von der Uniformität der Suszeptortemperatur ab [35], die leichter zu erzielen ist als die Uniformität auf der Scheibe in lampengeheizten Systemen.

Diese Reaktoren eignen sich daher zum Einsatz in der Silizidierung und zum Ausheilen von Defekten nach einer Ionenimplantation im Bereich mittlerer Temperaturen von 600°C bis 800°C. Bei höheren Temperaturen ist das Erreichen einer homogenen Suszeptortemperatur nur durch umfangreiche Isolierungen und Mehrzonenheizer möglich. Im vorliegenden Fall wird er zur Titansilizidierung für den Gatekontakt bei DRAMs verwendet. Bei zu niedriger bzw. zu hoher Temperatur ist der Schichtwiderstand zu hoch bzw. zu niedrig, was die Schalteigenschaften der Transistoren nicht reproduzierbar macht. Auf eine Einhaltung der Temperatur ist hier also genau wie im Fall der RTO-Schritte zu achten. Die Prozeßzeit liegt zwischen 40 und 60 Sekunden bei Temperaturen um 750°C und Drücken um 2 Torr [54].

Die Aufheizung des Wafers geschieht auf einer deutlich langsameren Zeitskala als bei Lampensystemen. Die Wafertemperatur erreicht ihre Endtemperatur erst nach mehreren Minuten.

In dem vorliegenden Reaktortyp (Abbildung 4.1) tritt das Problem des sogenannten First-Wafer-Effekts auf. Die ersten fünf bis zehn Wafer eines zu pro-

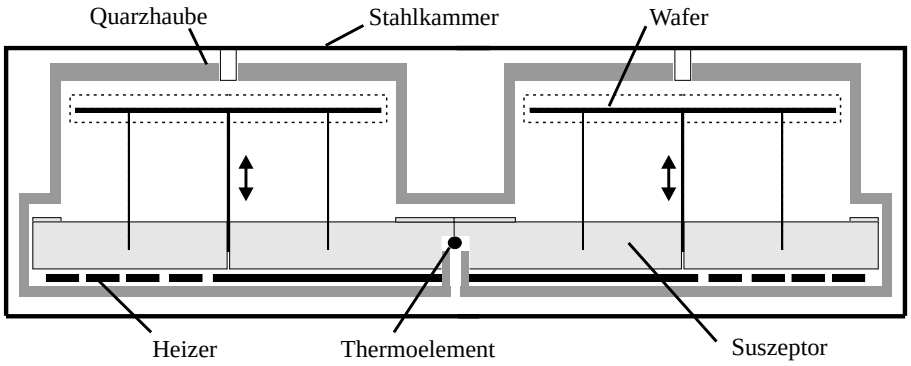


Abbildung 4.1: Skizze der in diesem Kapitel untersuchten Reaktorkammer. Der Wafer wird in die Kammer eingelegt und auf den Suszeptor abgesenkt. Die Wafertemperatur nähert sich ihrem Endwert nur langsam an (siehe Bild 4.2).

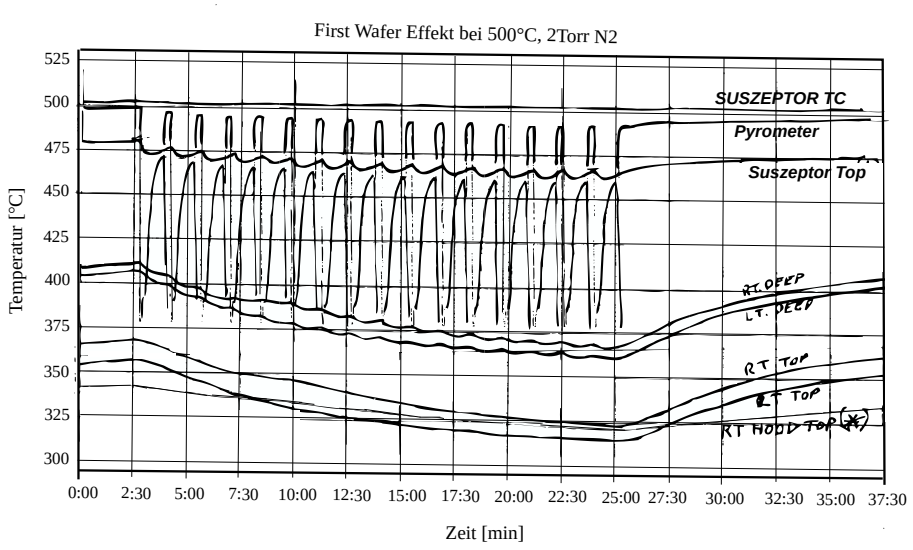


Abbildung 4.2: TC-Messung des zeitlichen Temperaturverlaufs an verschiedenen Positionen im Reaktor bei der Prozessierung von 16 Wafern [65].

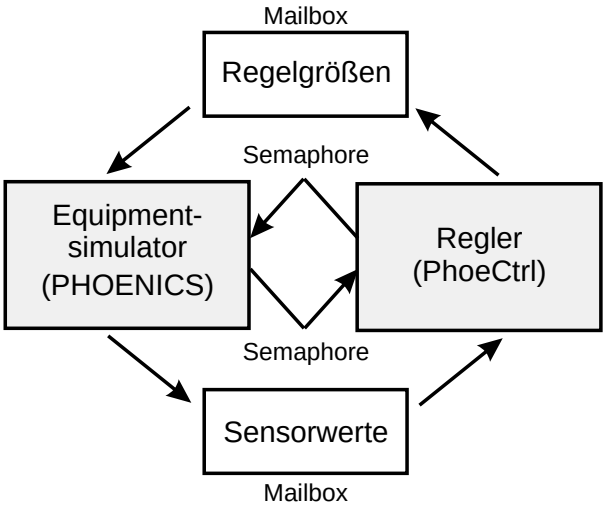


Abbildung 4.3: Kommunikation zwischen der Reglersimulation (PhoeCtrl) und dem Equipmentsimulator. Die errechneten Zellwerte an den Sensorpunkten werden nach jedem Zeitschritt des Simulators über Mailboxen an den im Hintergrund laufenden Regler übergeben, der dann die Aktuatorsignale zurückschickt. Diese Stellgrößen werden dann über entsprechende Quellterme im Simulator eigekoppelt.

zessierenden Loses weisen trotz eines vorhandenen PID-Reglers eine deutliche höhere Temperatur zum Rest des Loses auf (Abbildung 4.2).

Der Effekt ist nicht ad hoc aus den Messungen erklärbar, da das im Suszeptor eingebaute Thermoelement die Temperatur nachzuregeln scheint. Simulation und Messungen zeigen ein Absinken der Temperatur auf der Suszeptoroberseite und einen Temperaturabfall der Quarzhaube. Aus den obigen Beobachtungen folgen zwei mögliche physikalische Ursachen:

1. durch Einbringen des kalten Wafers in die Kammer bzw. auf den Suszeptor wird Wärmeenergie entzogen, wodurch sich die Suszeptoroberseite von Wafer zu Wafer abkühlt. Die Suszeptorunterseite wird dabei durch den Heizer konstant auf 1000°C gehalten.

2. der Wafer wirkt als Strahlungsschild zwischen Suszeptor und Quarzhaube; dadurch gelangt weniger Strahlung auf den Quarz, der sich abkühlt, was wiederum die Temperatur auf dem Wafer verringert (bei den geringen Drücken fast ausschließlich Strahlung).

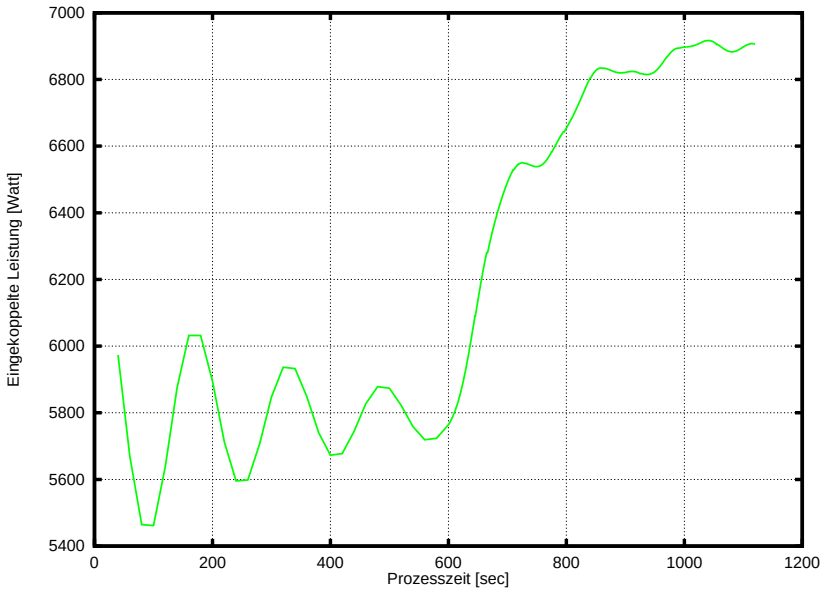


Abbildung 4.4: Simulation der in den Heizer eingekoppelte Stromleistung nach Einlegen des Wafers. Die Regelung der Nominaltemperatur erfolgte durch einen PID-Regler, der an den Equipmentsimulator PHOENICS angekoppelt wurde.

In der vollständigen Equipmentsimulation zeigt sich, daß diese beiden Erklärungen für die experimentell beobachtete Größe des First-Wafer-Effekts nicht ausreichen. Hinzu kommt, daß der Effekt für jede Kammer trotz gleicher Bauart unterschiedlich ist und durch Verwenden eines Pyrometers an der Waferoberseite zur Regelung umgekehrt werden kann, also die ersten Wafer zu kalt werden.

Aus diesem Grund wurde ein Reglersystem an den Simulator PHOENICS angekoppelt (Abbildung 4.3). Das Programm stellt eine Sensorsimulation dar: ein Meßwert – hier die Temperatur an der Thermoelementposition – wird vom Equipmentsimulator berechnet und an den Regler übergeben, der daraus eine oder mehrere Aktuatorgrößen – im vorliegenden Fall die Stromleistung des Heizers – an PHOENICS zurückgibt, was dort als Quellterm eingekoppelt wird. Damit stellt sich die Aufgabe, die tatsächliche Temperatur des Wafers abzuschätzen.

Mit Hilfe der Simulation des Systems Regler/Equipment konnte eine dritte Ursache für den First-Wafer-Effekt identifiziert werden: Das Thermoelement wird durch andere strahlende Körper außer dem Suszeptor gestört. Bei der Anbringung an der Suszeptorunterseite findet die Störung durch die Heizung und bei der Anbringung an der Waferoberseite durch die Quarzhaube statt. Damit kann auch die Umkehrung des First-Wafer-Effektes erkärt werden.

Mit Hilfe des in PHOENICS realisierten Reglers konnten zwei Meßpunkte als mögliche Eingabegrößen für eine modellprädiktive Regelung verwendet werden, zum einen am Suszeptor und zum anderen an der Quarzhaube, woraus dann die Temperatur unter dem Wafer abgeschätzt werden kann.

In dem beschriebenen Reaktor wurde daraufhin die Rückkopplung der Quarztemperatur mit einem vereinfachten Modell mit Erfolg realisiert. Durch diese Maßnahme konnte der First-Wafer Effekt auf wenige Grad reduziert werden.

Die Koeffizienten für den Regler wurden aus Messungen gewonnen.

4.3 Verringerung des Struktureffekts durch modellbasierte Steuerung

Die Aufheizung von Lampen-RTP-Systemen geschieht im allgemeinen mit Wolfram-Halogen-Lampen, die eine deutlich höhere Temperatur (ca. 3300K) als der Wafer (ca. 1300K) annehmen. Dadurch ist im allgemeinen die effektive Emissivität ungleich der effektiven Absorptivität. Sind auf dem Wafer zusätzlich Bereiche unterschiedlicher optischer Eigenschaften vorhanden, z.B. metallisierte oder oxidierte Gebiete und reines Silizium, so heizen sich die einzelnen Bereiche unterschiedlich auf und zeigen in Abhängigkeit der Vorder- und Rückseitenemissivität auch eine unterschiedliche stationäre Temperatur. Dies wird als Struktureffekt oder pattern-effect"bezeichnet und führt im Einsatz zu ungewollten Prozeßschwankungen auf einer Scheibe, Temperaturdifferenzen von bis zu mehreren zehn Grad wurden gemessen [66].

In der Siliziumprozessierung tritt der Struktureffekt vornehmlich dann auf, wenn große Bereiche eines Dies (z.B. bei Embedded DRAMs) eine deutlich andere Struktur aufweisen, oder unterschiedliche Produkte auf einem Wafer gefertigt werden, z.B. in MiniFabs. Der mikroskopische Struktureffekt auf Devicegröße ist aufgrund der hohen Wärmleitfähigkeit von Silizium vernachlässigbar. Dies zeigen auch Rechnungen von [67].

Die Wafertemperatur ist aber nicht mit einer örtlichen Auflösung meßbar, um während des Prozesses die Temperaturentwicklung zu beobachten. Im Prinzip ist es durch den Einsatz einer sehr großen Anzahl von Lampen und dem Verfahren aus Kapitel 3 möglich, eine optimale Lampeneinstellung zu finden, doch wäre dazu eine örtlich sehr gut bekannte Strahlungsmatrix notwendig.

Der folgende Ansatz versucht, durch Einführen einer zusätzlichen Steuergröße, nämlich der relativen Leistungsverteilung zwischen den Lampenhäusern, das Problem zu lösen.

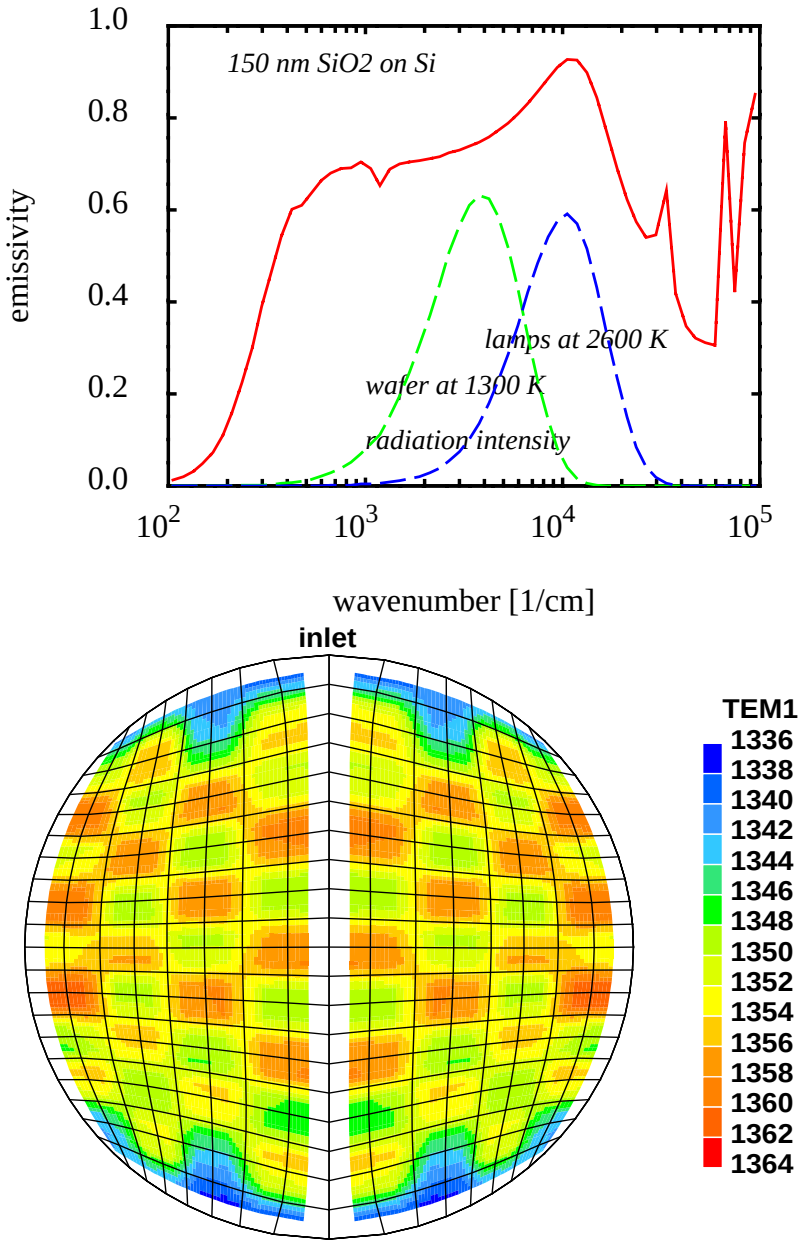


Abbildung 4.5: Ursache für den Patterneffekt ist eine Differenz zwischen effektiver Absorptivität und effektiver Emissivität [62], die zu Temperaturschwankungen auf dem Wafer führten. Simulationen der Temperatur analog zur Meßanordnung von [66].

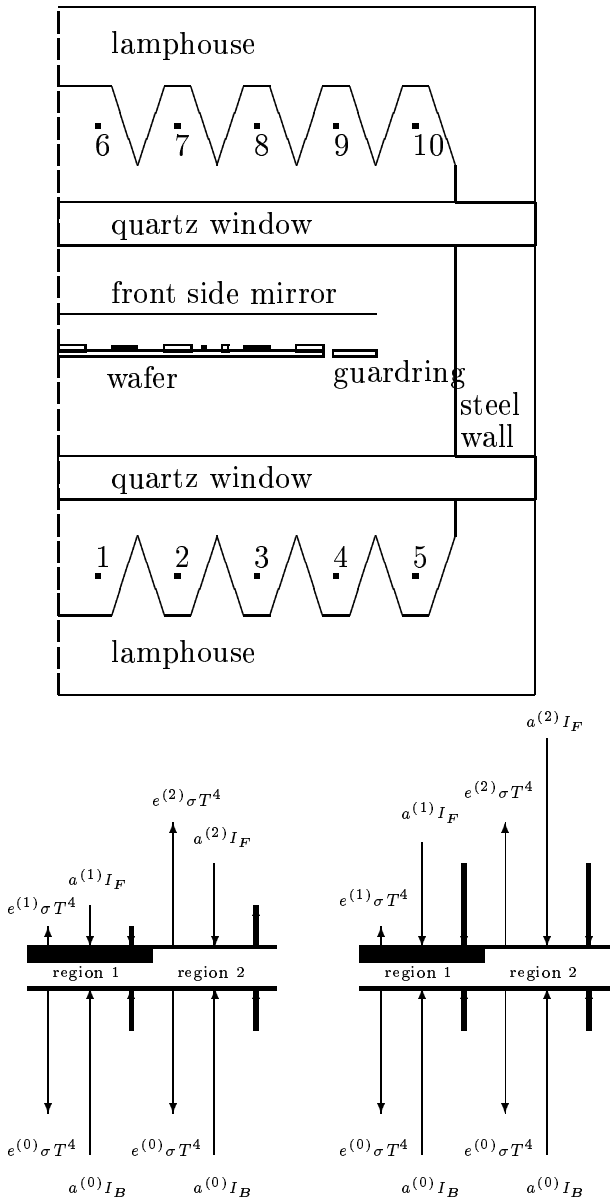


Abbildung 4.6: Oben: Axialsymmetrischer Reaktor mit strukturiertem Wafer (Strukturierungen überhöht dargestellt). Die untersuchten Variationen sind Vorder- und Rückseitenbeheizung und das Anbringen von Spiegeln. Unten: Wärmeflußbilanz (fette Pfeile) für schwächere und stärkere Vorderseitenbestrahlung. Bei der Einstellung auf der rechten Seite haben die beiden Bereiche trotz unterschiedlicher optischer Eigenschaften die gleiche Flußbilanz.

Aus der Flußbilanzskizze in Abbildung 4.6 ergibt sich – unter Vernachlässigung der lateralen Kopplung der Zellen – die Wärmeflußbilanz in den beiden Bereichen zu

$$a^{(0)}I_B + a^{(1)}I_F - e^{(0)}\sigma T^4 - e_{(1)}\sigma T^4 = C_1 \quad (4.1)$$

$$a^{(0)}I_B + a^{(2)}I_F - e^{(0)}\sigma T^4 - e_{(2)}\sigma T^4 = C_2 \quad (4.2)$$

wobei $a^{(0)}$ die effektive Absorptivität der Rückseite, I_B und I_F die Rückseiten- bzw. Vorderseitenintensität und $a^{(1)}$ und $a^{(2)}$ die effektiven Absorptivitäten der beiden Bereiche sind. Der Term C enthält in erster Linie die Aufheizrate $C_{1,2} = h\rho c_p \partial T_{1,2}/\partial t$. Da der Fehler, der durch das Nichtberücksichtigen der anderen Wärmeleitungsmechanismen gemacht wird, vernachlässigbar ist, kann $C_{1,2}$ im folgenden als konstant angenommen werden.

Bei vorgegebenen Intensitäten läßt sich aus Gleichung 4.2 das unterschiedliche Aufheizverhalten der beiden Bereiche berechnen. Gleichung 4.2 kann aber auch als Gleichungssystem für I_B und I_F aufgefaßt werden. Setzt man zusätzlich $C_1 = C_2$, so erreicht man ein gleichmäßiges Aufheizen beider Bereiche, also ein Verschwinden des Struktureffekts.

Das vorgeschlagene Verfahren zeigt sich anderen Ansätzen überlegen. Z.B. wurde in [53] das Einbringen einer spiegelnden Fläche über dem Wafer vorgeschlagen. Idee ist, jedem strukturiertem Bereich i mit der Emissivität e_i die emittierte Strahlung wieder zurückzusenden. Aufgrund des Kirchhoff'schen Gesetzes gilt $e(\nu, T) = a(\nu, T)$, wodurch bei einem beliebig nahen, perfekten Spiegel kein Patterneffekt auftritt. Nachteil der Idee ist jedoch zum einen, daß mit dem Spiegel ein nach mehreren Waferdurchläufen kontaminierter Gegenstand in die Kammer eingeführt wird, zum anderen das durch seine thermische Masse die Ramprate des Systems verringert wird, was prozeßtechnisch unerwünscht ist.

Weist der Spiegel einen endlichen Abstand zum Wafer auf, so gelangt Licht von den Lampen über Reflexionen in den Zwischenraum und beeinträchtig die Wirksamkeit des Spiegels (Abbildung 4.8). Wenn der Rand dieses Zwischenraums zur optimalen Nutzung ebenfalls verschlossen wird, bleibt nur noch ein zu vernachlässigender Struktureffekt aufgrund des nicht-perfekten Spiegels übrig, falls die Spiegeltemperatur von der Wafertemperatur verschieden ist.

Um die Realisierbarkeit des Verfahrens zu zeigen, wurde ein mit 1cm und 0.25cm breiten Wolfram und SiO_2-Streifen versehener Silizium-Wafer simuliert. Für hohe Temperaturen ergibt sich $e^{(2)} = 0.04, a^{(2)} = 0.28$ für Wolfram und $a^{(0)} = a^{(1)} = e^{(1)} = 0.68$ für Silizium und damit aus Gleichung 4.2 ein Verhältnis von $I_F : I_B = 4 : 1$. Im Falle eines SiO_2/Si Wafers mit $e^{(2)} = 0.75$ und $a^{(2)} = 0.77$ ein Verhältnis von $I_F : I_B = 2 : 1$.

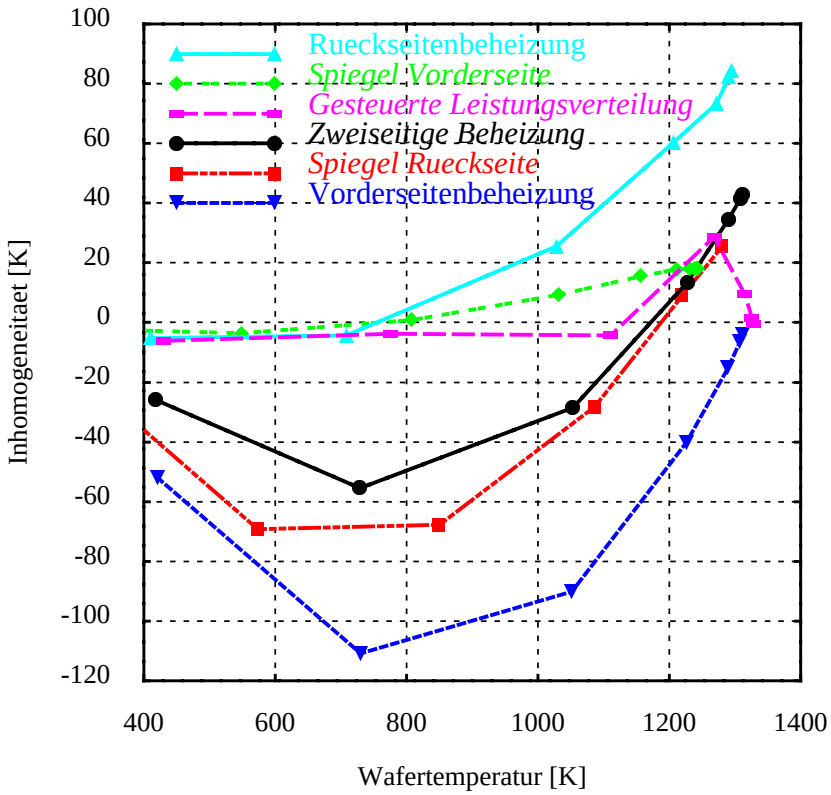


Abbildung 4.7: Temperaturinhomogenität auf dem Wafer bei den in der Literatur erwähnten Verfahren und dem Ansatz der geregelten Leistungsverteilung in dieser Arbeit (gesteuertes Leistungsverhältnis). Mit Hilfe einer modellbasierten Steuerung läßt sich der Struktrueffekte über einen weiten Temperaturbereich auf Nahe 0 reduzieren (Simulationen aus [35]).

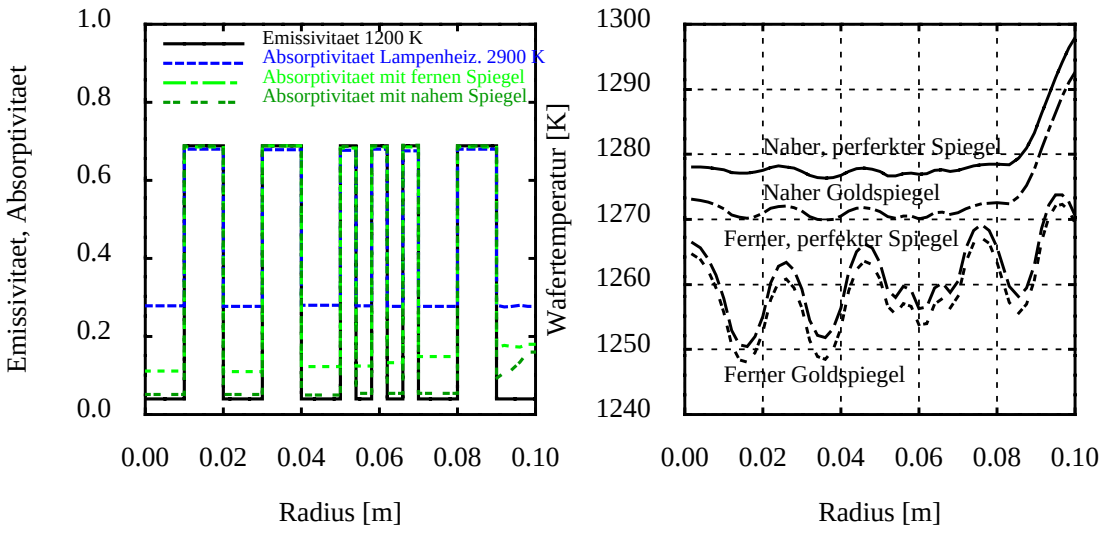


Abbildung 4.8: Links: Emissivität und Absorptivitätsstruktur für die simulierte Teststruktur und stationäre Temperaturuniformität für verschiedene Vorderseitenspiegel im Abstand 2mm und 2cm, 100% Reflektivität und Reflektivität von Gold.

Die Rampe wurde durch Einschalten der Lampen auf ihre Nominalleistung simuliert. Wird die Lampenleistung verändert, so variieren auch die optischen Eigenschaften und das Leistungsverhältnis muß entsprechend korrigiert werden. Die Ergebnisse für Wolfram und Silizium sind in Abbildung 4.7 dargestellt.

Die Begrenzung dieser einfachen Form des Verfahrens ist, daß nicht die Intensitäten der Lampenhäuser aus Gleichung 4.2 berechnet werden, sondern die Intensitäten auf der Waferoberfläche. In der beschriebenen Geometrie gelangt Strahlung vom Lampenhaus über Reflexionen an die gegenüberliegende Waferseite, was in der Formel zu berücksichtigen ist.

Durch modellbasierte Steuerung ("model-based scheduled control") ist es also möglich im Fall zweier unterschiedlich beschichteter Waferbereiche den Struktureffekt reduzieren zu können. Voraussetzung natürlich ist, daß eine homogene Wafertemperatur auch dann möglich ist, wenn nur eins der beiden Lampenhäuser verwendet wird.

4.4 Regelung der Temperaturkurve in der Aufheizphase

Im vorausgegangenen Kapitel wurde mit einer konstanten Lampenleistung gerampt, die Aufteilung der Lampenleistungen aber durch ein vorherberechnetes Modell bestimmt. Die weitere Aufgabe einer Regelung ist aber die Verarbeitung von in Echtzeit gemessenen Werten und die Bestimmung einer optimalen Lampeneinstellung während des Prozesses. Dabei ist insbesondere eine Aufheizkurve vorgegeben, die unter der Randbedingung einer möglichst homogenen Temperatur nachgefahren werden soll.

Hier wird nun eine Vorgehensweise vorgestellt, die aufbauend auf den Parameterextraktion aus dem Kapitel 3 in der transienten Aufheizphase die Temperatur regelt. Es wurde bereits im Simulationskapitel gezeigt, daß im quasi-stationären Zustand nur noch die Drift der Quarzfenster eine leichte Veränderung der Homogenität verursacht. Auf eine genauere Formulierung wird am Ende dieses Abschnitts eingegangen.

Analog zum Optimierungsproblem im stationären Zustand wird nun eine lokaloptimale transiente Regelung entwickelt. Da es während des transienten Aufheizens in vielen Reaktoren zum Vorlauf der Waferkante in der Temperatur kommen kann, muß mit einer transienten Mehrzonenregelung gearbeitet werden. Das Optimierungsproblem für jeden Zeitschritt $t_i, i = 0, \ldots, n$ finde $\mathbf{P}(0), \mathbf{P}(1), \ldots, \mathbf{P}(M)$, die das Optimierungsproblem

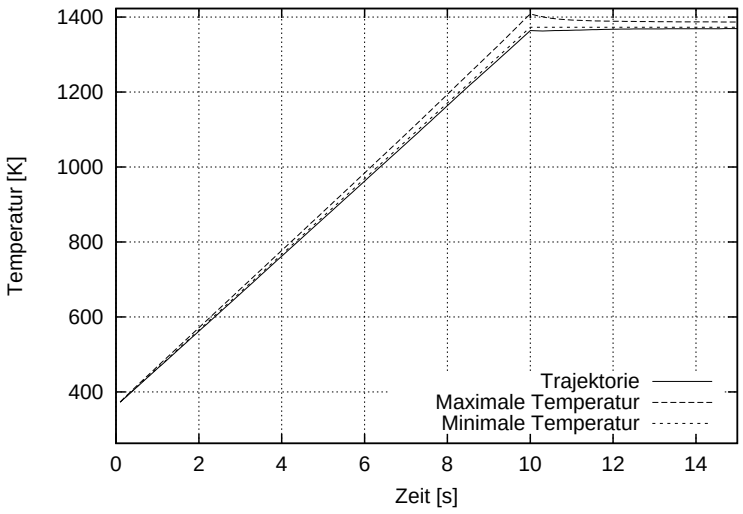


Abbildung 4.9: Trajektorie für die transiente Optimierung mit Ramprate 100°C/sek.

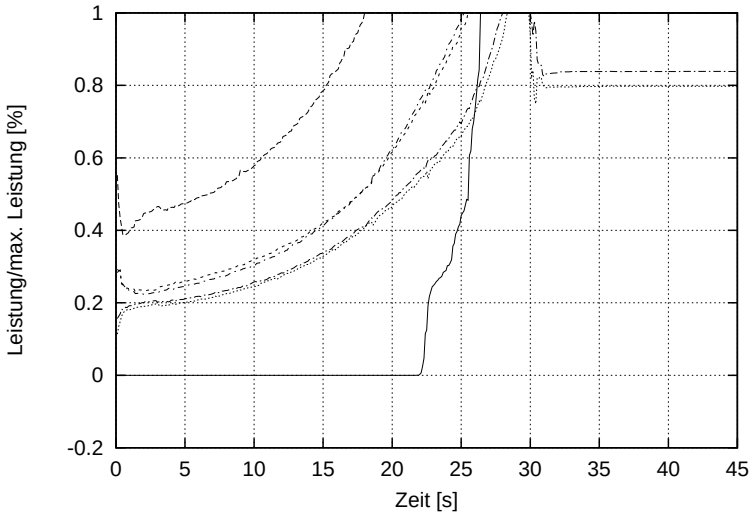


Abbildung 4.10: Leistungseinkopplung in die Linearlampen bei einer Ramprate von 30°C/sek.

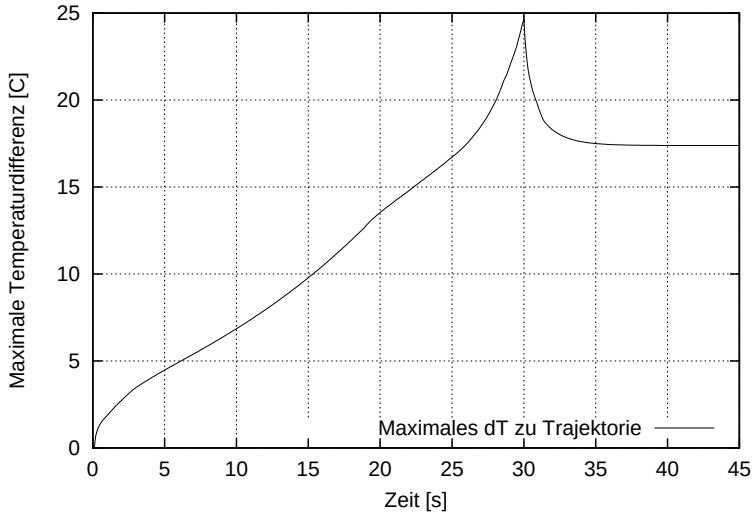


Abbildung 4.11: Temperaturinhomogenität bei einer Ramprate von 30°C/sek.

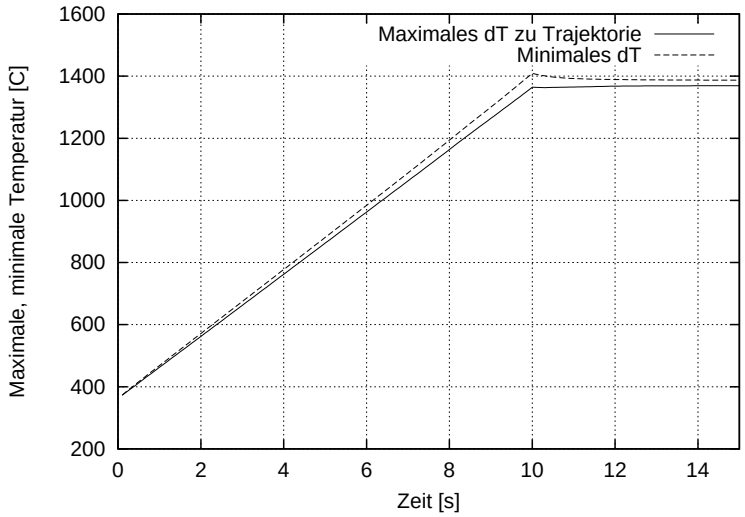


Abbildung 4.12: Leistungseinkopplung in die Linearlampen bei einer Ramprate von 100°C/sek.

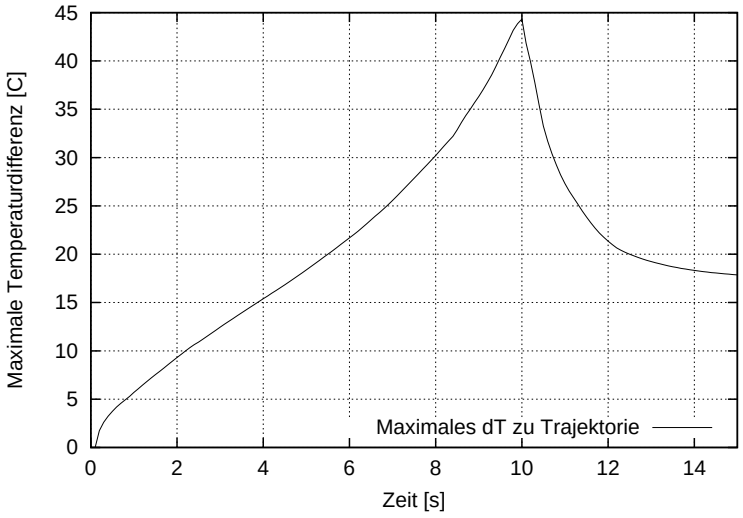


Abbildung 4.13: Temperaturinhomogenität bei einer Ramprate von 100°C/sek.

$$\max_{m=1,\ldots,M} ||\mathbf{E}(m)||, \quad \mathbf{E}(m) := \mathbf{T}(m) - \mathbf{T}^{\text{ref}}(m) \tag{4.3}$$

$$\text{mit } 0 \leq \mathbf{P}(m) \leq \mathbf{P}^{\max}$$

erfüllen [61], wobei $\mathbf{T}^{\text{ref}}(m)$ die geforderte Temperatur zum Zeitschritt m bezeichnet.

Dabei wird von einer Ausgangsverteilung gestartet. Typischerweise ist dies die Einführungstemperatur des Wafers. Danach wird für jeden Zeitschritt das Optimierungsproblem 4.3 gelöst. Es genügen wenige Iteration, da die Leistungsverteilung des vorigen Zeitschritts ein guter Ausgangspunkt für die Minimierung sind.

Die Optimierungsergebnisse sind in den Abbildungen 4.10 und 4.12 für verschiedene Rampraten dargestellt. Bei der geringen Leistungsreserve der Kammer erweist sich das genaue Nachfahren der Trajektorie bei hohen Rampraten als besonders schwierig. Ist die Temperaturinhomogenität bei 30°C/sec noch maximal 20 Grad (Abbildung 4.11), so ist die nahezu volle Leistungskraft aller Lampen notwendig, um eine Rampe mit 100°C/sek erzielen zu können. Daraus resultiert die schlechte Uniformität auf dem Wafer (Abbildung 4.13). Nicht dargestellt ist das Rampen mit Lampen, die die doppelte Leistung der real verwendeten Lampen haben. Dann läßt sich die transiente Uniformität detulich verbessern. Dies gilt für das RTP-System in [62] und den im Ausblick (Abbildung 5.1) dargestellten Reaktor.

4.5 Optimale Regelung unter Minimierung des thermischen Budgets

Eines der Hauptprobleme bei der Verarbeitung von Wafern ist die (ungewollte) Diffusion von Dopanden im Silizium durch thermische Bearbeitungsschritte nach Implantation der Dotierstoffe. Die klassische Definition des thermischen Budgets über [68]

$$\Theta_1(T) = \int_0^{t_p} T(\tau)d\tau = t_p\overline{T} \tag{4.4}$$

oder

$$\Theta_2(T) = \int_0^{t_p} e^{\frac{T(\tau)}{T_0}}d\tau \tag{4.5}$$

sind entworfen worden, um alle negativen Effekte wie das Verbreitern von Dotierungsprofilen, die Aktivierung von Dopanden etc. bei Erhöhung der Anzahl der thermischen Prozesse zu subsummieren. Dabei ist t_p die Summe über alle Zeiten der einzelnen Prozesse.

Die Definition der Form 4.4 beinhaltet, daß ein Prozeß mit doppelter Länge und halber Temperatur das gleiche thermische Budget wie der Ausgangsprozeß hat. Mikroskopische Effekte gehorchen aber selbst im einfachsten Fall Arrhenius-Abhängigkeiten.

In [69] wurde experimentell gezeigt, daß die Definition des thermischen Budgets über das Integral der $T(t)$ Kurve bei Anwendung auf Einzelprozesse an ihre Grenzen stößt. Untersucht wurden Bordotierungsprofile mit verschiedenen Temperatur-Zeit-Profilen. Zwar nimmt natürlich die Verbreiterung des Profils mit steigender Temperatur zu, wie auch von 4.4 vorhergesagt, jedoch stimmt dies bei Variation der Aufheizrate nicht mehr.

In dieser Arbeit wird daher von einer allgemeineren Definitiondes thermischen Budgets ausgegangen: $P_j(t)$ sei das zu erzielende primäre Prozeßergebnis bei einem RTO-Schritt, z.B. Oxiddicken oder Abscheideschichten. Weiterhin bezeichne $S_i(T,t) = \int \sigma_i(T,t)dt$ einen Satz von $i = 1,\ldots,I$ sekundären Prozeßgrößen, die sich mit den Raten $\sigma_i(T,t)$ ändern. Ziel nach der Prozeßabfolge die möglichst geringe Abweichung der $S_i(T,t)$ von ihrem Startwert $S_i(0)$. Dies trifft z.B. auf eine gewünschte Dopandenprofilform oder die Position eines pn-Übergangs zu. Dann wird

$$\Theta := \left(\sum_i g_i\right)^{-1} \sum_i g_i \frac{1}{||S_i(0)||} ||S_i(0) - \sum_p \int_0^{t_p} \sigma_i(T_p(\tau),\tau,S)d\tau|| \tag{4.6}$$

als das thermische Budget einer Abfolge von Prozeßschritten p mit Prozeßzeit t_p und Temperaturkurve $T_p(t)$ definiert. $S_i(0)$ bezeichnet die gewünschten Werte der Störfaktoren, die g_i sind Gewichtsfaktoren. Die Formulierung hat den Vorteil der generellen Anwendbarkeit und stellt eine erweiterte Formulierung der erstgenannten Definitionen für bestimmte σ_i dar.

Die Aufgabe zur Minimierung des thermischen Budgets kann nun in zwei Aufgaben gespalten werden:

1. Minimierung durch Veränderung des Gesamtprozesses. Dies beginnt beim Einfügen von einzelnen Ausheilschritten, der Variation der Implantationsbedingungen und geht bis zum Einsatz anderer Materialien

2. Optimierung des thermischen Budgets beim Einzelprozeß durch Variation der Temperaturkurve

Punkt 1 verspricht die größten Verbesserungen und ist eine der Hauptfragestellungen der Prozeßsimulation. Das Hinzufügen oder Weglassen von Prozeßschritten erfordert eine genaue Kenntnis der Zielsetzungen und physikalischen Abläufe. Eine algorithmische Optimierung scheidet aufgrund des zu großen Parameterraums und mangelnder Modellgenauigkeit aus.

Daher wird im folgenden der Fall *einer* Primärgröße P, also ein Einzelprozeß, betrachtet. Die Aufgabe besteht nun in der Minimierung der Gleichung 4.6 zu

$$\min_{T(t)} \Theta \quad \text{unter der Nebenbedingung} \quad \int \pi_j(T,\tau,P)d\tau = P_0 \text{ für alle } j \qquad (4.7)$$

durch Variation der Temperaturtrajektorie[1].

In den folgenden Abschnitten werden nun Lösungen des Minimierungsproblems 4.7 diskutiert. Für einfache Temperaturabhängigkeiten der Raten σ und π lassen sich einfache Abhängigkeiten angeben. Der allgemeinere Fall für zeitabhängige Prozesse, wie z.B. transient enhanced diffusion und oxidation enhanced diffusion erfordert physikalische Simulation und wird in Abschnitt 4.5.2 am Beispiel des RTO demonstriert.

4.5.1 Fall mit zeitunabhängigen Reaktionsraten

In diesem Abschnitt werden folgende Voraussetzungen an den Prozeß gestellt:

[1]Auch andere Größen, z.B. Ätztiefe als Primärgröße, Plasma-Damage als Sekundärgröße und Ionendichte als Regelgröße kämen hier in Betracht, erfordern aber andere Modelle. Dabei ist $\pi_j(T,t)$ die totale Ableitung von P_j nach der Zeit, z.B. eine Reaktionsrate, dabei durchläuft j alle primären Pozeßergebnisse. Auch Kostenargumente für verlängerte Prozeßzeiten sind potentielle Sekundärgrößen.

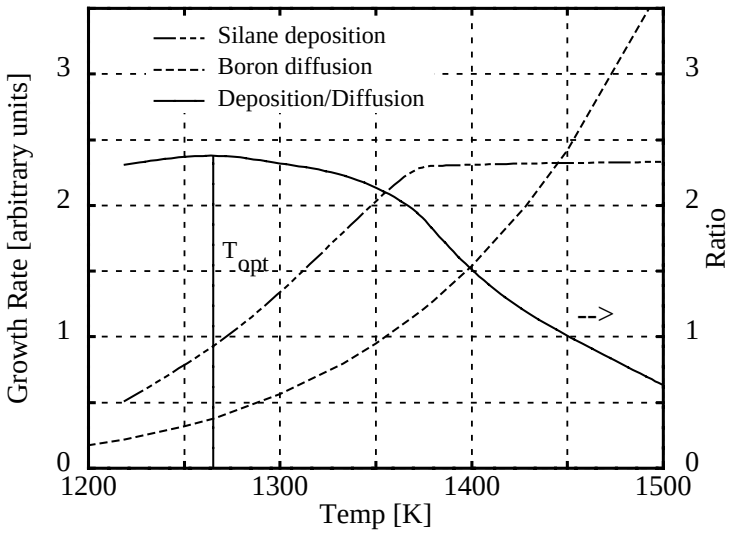


Abbildung 4.14: Bestimmung der optimalen Abscheidetemperatur für Silan unter der Nebenbedingung des minimalen thermischen Budgets.

1. Primär- und Sekundärgröße (Gleichungen 4.7 und 4.6) sind durch Variation der Prozeßzeit und/oder der Prozeßtemperatur veränderbar.

2. Die Reaktionsrate der Primärgröße sei nur temperaturabhängig der Form

$$\pi(T, t, P) = \pi(T). \tag{4.8}$$

   Die Primärgröße des Prozesses aus Gleichung steige bei konstanter Temperatur monoton mit der Prozeßzeit (bei monotonem Fallen wähle man $p' = p_0 - p$). Die Annahme ist bei vielen Wachstumsvorgängen erfüllt, z.B. bei CVD oder beim Wachstum sehr dünner Oxide.

3. Für die Reaktionsrate der Sekundärgröße gelte

$$\sigma(T, t, S) = \sigma(T)g(t) \tag{4.9}$$

Die Einschränkung (2) gilt z.B. für alle Arrhenius-artigen Prozesse ($\sigma = \sigma_0 exp(-\frac{E_A}{kT})$) und Einschränkung (3), wie unten gezeigt wird, für das Zerfließen von gaußförmigen Dotierstoffprofilen ($\sigma(T,t) = \frac{d}{dt}\left[\sqrt{D_0 \exp(-\frac{E_A}{kT})t}\right]$). Nicht berücksichtigt sind hier einige transiente Prozesse, wie z.B. die transiente Diffusion nach einer Implantation oder auch die verstärkte Diffusion während der Oxidation, hier ist die Reaktionsrate σ über die Diffusion der Interstitials von der Grenzfläche zum Dotierstoff zeitabhängig und nicht geschlossen berechenbar.

Punkt 1 ist notwendig, um überhaupt Einfluß auf das Prozeßergebnis durch Variation der Temperaturkurve nehmen zu können. Zu Punkt 2 sei angemerkt, daß auch eine Zeitabhängigkeit der Form $\pi(T,t,P) = \pi(T)f(t)$ mit bekanntem $f(t)$ im folgenden durch einen Variationsansatz berücksichtigt werden kann.

Sind Voraussetzungen 2 und 3 erfüllt, so kann eine analytische Abschätzung gefunden werden. Im nächsten Unterkapitel wird danach der allgemeine Fall ohne diese Einschränkung behandelt.

Einer anschauliche Darstellung von Gleichung 4.6 wird nun gegenüber der abstrakten Formulierung der Vorzug gegeben: Der Prozeß werde in n äquidistante Temperaturschritte mit konstanter Temperatur $T_i, i = 1, \ldots, n$ zerlegt.

Dann soll für die Primärgröße am Ende des Prozesses der vorgegebene Wert erreicht werden

$$P_0 = \sum_{i=1}^{n} \pi(T_i)\Delta t \tag{4.10}$$

Die Breite der Schritte Δt wird nun so gewählt, daß P_0 erreicht wird, dann gilt für die Sekundärgröße am Prozeßende

$$S = \frac{P_0 \sum\limits_i \sigma(T_i)}{\sum\limits_i \pi(T_i)} \tag{4.11}$$

Sei nun $\overline{T}$ die Temperatur, so daß $\frac{\sigma(T)}{\pi(T)}$ im vorgegebenen Temperaturbereich minimal ist, dann folgt

$$S \geq P_0 \frac{\sigma(\overline{T})}{\pi(\overline{T})} \tag{4.12}$$

Das Gleichheitszeichen gilt, wenn in allen Prozeßschritten die Temperatur $\overline{T}$ eingehalten wird. Daraus ergibt sich, daß es unter der Voraussetzung zeitlich nicht konstanter Reaktionsraten eine optimale Temperatur gibt, bei der der störende Einfluß der Sekundärgröße am besten unterdrückt werden kann. Für streng Arrhenius-förmige Prozesse ist dies entweder die minimal oder die maximal mögliche Prozeßtemperatur (Abbildung 4.15).

Diese Betrachtung kann auch auf das Verfließen von Dotierstoffprofilen erweitert werden: Die Gaußverteilung ist die Greensfunktion der Fick'schen Diffusionsgleichung. Daher beschränken wir uns auf die Verbreiterung eines gaußförmigen Dotierprofils. Man sieht leicht, daß für die Breite des Gaußprofils gilt:

$$G(T,t) = \sqrt{D(T)t + G_0^2} = \sqrt{D(T)(t - t_\infty)} \tag{4.13}$$

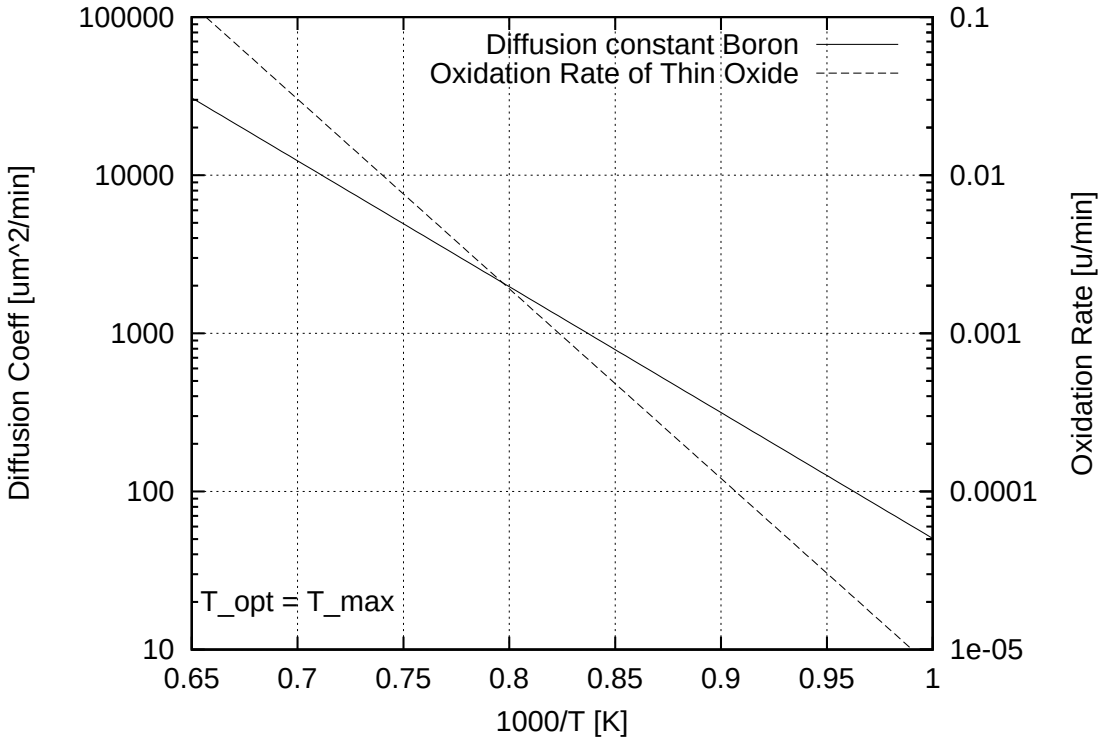


Abbildung 4.15: Optimale Temperaturbestimmung bei Prozessen mit bekannter Zeitabhängigkeit. Oxidationsrate und Diffusionskoeffizient zeigen beide eine Arrheniusabhängigkeit, daher ist die optimale Temperatur entweder die maximal oder minimal mögliche.

mit der Anfangsbreite G_0 und der temperaturabhängigen Diffusionskonstante $D(T)$. Da $G(T,t)$ stets positiv ist, läßt sich die Aufgabe der Funktionsminimierung auf die Minimierung des Quadrats der Profilbreite $G(T,t)^2$ umschreiben. Die obige Argumentation für Arrheniusprozesse gilt daher analog auch für Gaußprofile.

Die Gaußkurve ist als grobe Abschätzung des Diffusionsverhaltens geeignet. Sie genügt aber im allgemeinen nicht, um die komplexe Diffusionschemie ausreichend zu beschreiben. Hierbei ist zusätzlich zumindest noch die Entstehung und Rekombination von Punktdefekten zu berücksichtigen.

4.5.2 Allgemeiner Fall mit zeitabhängiger Reaktionsrate

Die oberen Überlegungen gelten im allgemeinen nicht für Prozesse, bei denen die Sekundärgröße nicht konstant ist. Eine allgemeine Lösung des Optimierungsproblems für Gleichung 4.6 muß hier numerisch aufgrund eines komplexeren Modells gefunden werden.

Zur Demonstration der Anwendbarkeit des Verfahrens wird nun die Diffusion von Bor während eines Oxidationsvorgangs betrachtet. Der Transport von Bor und andere Dotierstoffe in Silizium geschieht nahezu ausschließlich über Kristalldefekte ("Paardiffusion"). Dazu gehören zum einen die langreichweitigen Fehler wie Versetzungen, zum anderen Punktdefekte, also Zwischengitteratome ("Interstitials") I und Fehlstellen im Gitter ("Vakanzen") V. Je nach mikroskopischer Anordnung liegen Punktdefekte unterschiedlicher Ladung vor $(I^+, I^0, I^-, I^{2-}, V^+, V^0, V^-, V^{2-})$.

Das Siliziumsubstrat in der Halbleitertechnologie ist zumeist einkristallin, Versetzungen werden durch thermische oder mechanische Verspannungen hervorgerufen und können durch geeignete Lagerung und Homogenitätsregelung der thermischen Prozesse minimiert werden. Punktdefekte dagegen werden thermisch angeregt oder durch Verschiebungen von Interfaces generiert.

Für die Generation von Diffusionspaaren eines Dotierstoffs A der Ladung $q_A := d * e$ sind folgende Reaktionen zu berücksichtigen

$$A^d + V^n \rightleftharpoons AV^{d+n} \tag{4.14}$$
$$A^d + I^m \rightleftharpoons AI^{d+m} \tag{4.15}$$
$$AV^{d+n} + I^m \rightleftharpoons A - (m+n)e^- \tag{4.16}$$
$$AI^{d+m} + V^n \rightleftharpoons A - (m+n)e^- \tag{4.17}$$
$$I^m + V^n \rightleftharpoons -(m+n)e^- \tag{4.18}$$
$$I^m + e \rightleftharpoons I^{m-1} \tag{4.19}$$
$$V^n + e \rightleftharpoons V^{n-1} \tag{4.20}$$

Die elektronischen Reaktionen 4.19 und 4.20 finden auf einer im Vergleich zur Prozeßzeit und den anderen Reaktionen vernachlässigbar kleinen Zeitskala statt und sind im Gleichgewicht. Dann kann die Konzentration für geladene Punktdefekte durch die Konzentration neutraler Punktdefekte ausgedrückt werden:

$$C_{P^k} = R_{P^k} \left(\frac{n_e}{n_{e,intrin}} \right)^{-k} C_{P^0} \tag{4.21}$$

mit einer Gleichgewichtskonstanten R_{P^k} und der Elektronenkonzentration n_e.

Reaktionen zwischen Dotierstoff/Punktdefekt-Paaren werden aufgrund der hier betrachteten, niedrigen Dotierungen vernachlässigt. Alle Dotierstoffatome im Silizium sind in obigem Modell vollständig ionisiert. Die Paarungsreaktionen werden für die folgenden Rechnungen als im Gleichgewicht angenommen ("3-Spezies-Modell" mit $C_A = C_{A^d} + C_{AI} + C_{AV}, C_I$ und C_V). Diese Annahme ist für Hochtemperaturprozesse über 800°C gerechtfertigt (Reaktionszeit etwa 25×10^{-6}sec, für Temperaturen um 600°C jedoch bei nahezu 30 Minuten [71]). Es zeigt sich, daß im Niedertemperaturbereich eine möglichst schneller Rampe – analog zu den Überlegungen im vorigen Unterabschnitt – das Dotierprofil am geringsten beeinflußt.

Die Diffusion einer Konzentration C_A eines Dotierstoffs A ergibt sich als Summe der verschieden geladenen Interstitials- J_I und Vakanzendiffusionsströme J_V :

$$\frac{\partial C_A}{\partial t} = -\nabla \cdot (\vec{J_V} + \vec{J_I}) \tag{4.22}$$

$$\vec{J_I} = -\left(\sum_m D_I^m \left(\frac{n_e}{n_{e,intrin}}\right)^{-m}\right)\left(\nabla\left(C_{A,mobil}\frac{M}{M'}\right) - \left(C_{A,aktiv}\frac{C_I}{C_I'}\right)\frac{q_A\vec{\varepsilon}}{kT}\right) \tag{4.23}$$

$$\begin{aligned}\vec{J_V} &= -\left(\sum_n D_V^n \left(\frac{n_e}{n_{e,intrin}}\right)^{-m}\right) \\ &\quad \left(\nabla\left(C_{A,mobil}\frac{N}{N'}\right) - \left(C_{A,aktiv}\frac{C_V}{C_V'}\right)\frac{q_A\vec{\varepsilon}}{kT}\right)\end{aligned} \tag{4.24}$$

$$\vec{\varepsilon}|_{\mathrm{Si}} = -\frac{kT}{e}\frac{\nabla n}{n} \quad \text{elektrisches Feld} \tag{4.25}$$

$$n_{e,intrin}|_{\mathrm{Si}} = 3.87 \times 10^{16}\mathrm{cm}^{-3}\left(\frac{T}{1\mathrm{K}}\right)^{3/2}\exp\left(-0.605\mathrm{eV}/kT\right) \tag{4.26}$$

dabei bezeichnet $C_{A,mobil}$ die bewegliche und $C_{A,aktiv}$ die aktive Konzentration an Dotierstoff. $\frac{C_I}{C_I'}$ und $\frac{C_V}{C_V'}$ bezeichnen die Überhöhung der Interstitial- bzw. Vakanzenkonzentration gegüber dem thermodynamischen Gleichgewicht, $C_I' = 1.25 \times 10^{29}\mathrm{cm}^{-3}\exp(-3.26\mathrm{eV}/kT)$, $C_V' = 1.25 \times 10^{29}\mathrm{cm}^{-3}\exp(-3.26\mathrm{eV}/kT)$ [73]. Die Summen laufen jeweils über alle Ladungszustände der Punktdefekte.

Während des Oxidationsprozesses kommt es an der Grenzschicht zwischen Silizium und Siliziumdioxid zur Injektion von Interstitials [76]. Aus Diffusionsexperimenten bei Oxidation und Nitridierung ist bekannt, daß Bor und Phosphor bevorzugt mit Interstitial-Paaren diffundieren, während Arsen stärker mit Vakanzen transportiert wird. Für die Transportgleichungen der Punktgleichungen

sind die Reaktionsgleichungen 4.18 als Senken einzubinden. Am Si/SiO_2 Interface, das lokal mit der Geschwindigkeit v wächst, ergibt sich [73]

$$\frac{\partial C_I}{\partial t} = \nabla\cdot(D_I\nabla\, C_I) - r_{IV}(C_I C_V - C_I' C_V') - r_I^{\text{surf}}(C_I - C_I') + \rho_{\text{Si}}\theta_I \left(\frac{v}{\frac{\partial x_{ox}}{\partial t}}\right)^{\iota} \quad (4.27)$$

$$\frac{\partial\, C_V}{\partial t} = \nabla\cdot(D_V\nabla\, C_V) - r_{IV}(C_I C_V - C_I' C_V') - r_V^{\text{surf}} \left(1 + \left(\frac{v}{\frac{\partial x_{ox}}{\partial t}}\right)^{0.5}\right)(C_V - C_V') \quad (4.28)$$

$$\begin{aligned} D_I = D_V &= 3.65\times 10^{-4}\frac{\text{cm}^2}{\text{sec}}\exp(-1.58\text{eV}/kT) && (4.29)\\ r_{IV} &= 1.0\times 10^{-21}\text{sec}^{-1}\exp -1.0\text{eV}/kT && (4.30)\\ \iota &= -0.7 && (4.31)\\ \theta_I &= 0.01 && (4.32) \end{aligned}$$

Eine genauere Beschreibung der Modelle ist [73] zu entnehmen. Umfangreiche Parameterstudien finden sich in [72], mikroskopische Analysen der Transportprozesse in [74] [75].

Im folgenden wird nun das Beispiel einer Bordiffusion für eine typische Implantation eines NMOS-Kanalprofils (Dosis 10^{15}cm^{-3}, Energie 20keV) mit anschließendem Aufbringen eines Gateoxids der Dicke 10nm betrachtet.

Ziel ist es nun, unter Einhaltung der Oxiddicke die Temperaturkurve zu variieren. Sekundärgröße ist die Eindringtiefe des Profils, also die Konzentrationsgrenze von 10^{17}cm^{-3}. Als Minimierungsalgorithmus wurde ein simulated annealing Verfahren [104] [101] gewählt. Dieses stochastische Verfahren ist toleranter gegenüber numerischen Schwankungen bei der Auswertung des Funktionals 4.7 und vermeidet die Konvergenz zu lokalen Minima.

Der Prozeß wird in zwei Temperaturschritte mit dazwischenliegender Rampe unterteilt. Als Optimierungswerte dienen das Verhältnis der Zeiten der beiden Schritte, sowie die Temperaturwerte. Zur Vereinfachung wird angenommen, daß die Heiz- und Abkühlrampen vorgegeben sind. Die Gesamtzeit der beiden Prozeßschritte wird solange skaliert, bis die Zieloxiddicke erreicht ist.

Zur Reduktion der Rechenzeit wird hierzu als Auswertefunktion zunächst das Irene-Plummer-Modell (siehe Abschnitt A) verwendet. Zur Berechnung der Ausgangsprofile wird in der Optimierungsschleife der Prozeßsimulator TMA

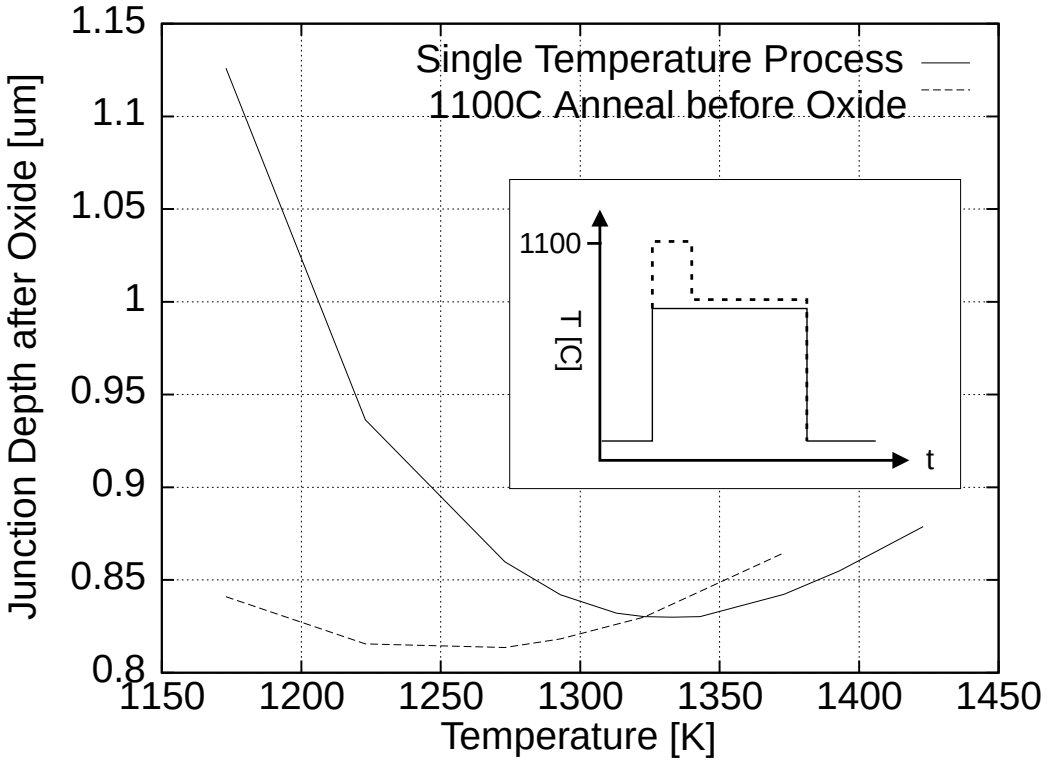


Abbildung 4.16: Ergebnis der transienten Optimierung mit einem Zweistufenprozeß. Der zusammengesetzte Prozeß liefert eine leicht geringere Eindringtiefe als ein Einschrittprozeß mit optimaler Temperatur.

TSuprem in der Version 6.5 aufgerufen und aus dem diffundierten Profil die Eindringtiefe, bei der die Dotierstoffkonzentration unter dem Wert von 10^{16} absinkt, extrahiert. Als Simulationsparameter werden die oben aufgeführten Literaturkonstanten verwendet. Im Falle der Anwendung auf einen realen Prozeßfluß ist gegebenenfalls eine Kalibration des Implantationsprofils und der Diffusionsparameter notwendig.

In Abbildung 4.16 wird das Ergebnis aus dieser Zweitemperatur-Optimierung mit einem Eintemperatur-Optimierung verglichen. Der Zweitemperaturprozeß liefert eine geringere Eindringtiefe als ein Prozeßschritt alleine. Die Temperaturtrajektorie liefert damit ein Werkzeug zur Erzeugung flacherer, definierter p/n-Übergänge.

Als Erweiterung dieses Ansatzes sind auch komplexere Temperaturkurven verwendbar. Hierzu eignet sich z.B. eine Linearkombination von Polynomen geringer Ordnung mit Nebenbedingungen.

4.6 Zusammenfassung

Verschiedene Problemstellungen bei thermischen Prozessierungsschritten wurden in den vorausgegangen Abschnitten durch unterschiedlich komplexe Modellierungs- und Regelungsstrategien analysiert. Die zur Problemlösung angewandten Verfahren beginnen bei der simulationsgestützten Verhaltensanalyse eines klassischen PID-Reglers und reichen bis zur nichtlinearen optimalen Regelung unter Nebenbedingungen. Gemeinsam ist allen Punkten, daß weniger das Einbringen eines ausgefeilten Reglers die beste Lösung hervorbringt – alle Systeme sind erster Ordnung in der Zeit mit Zeitkonstanten, die ein Schwingen des Systems wenig wahrscheinlich machen – sondern vielmehr die Notwendigkeit einer genauen physikalischen Modellierung.

Kapitel 5

Zusammenfassung und Ausblick

5.1 Zusammenfassung der Ergebnisse

In dieser Arbeit wurde der erfolgreiche Einsatz von physikalisch basierten Optimierungs- und Regelungsverfahren in der Halbleiterverarbeitung gezeigt. Trotz verschiedener Einsatzbereiche bei thermischen Reaktoren war die zugrundeliegende Methodik stets ähnlich:

- Simulationen und physikalische Modellierungen wurden so weit wie möglich eingesetzt, um das Systemverhalten zu verstehen und um Stabilitätsverbesserungen zu erzielen. Dadurch wird der transiente Regelbedarf verringert.
- ein reduziertes Modell wurde entwickelt, das die dominanten Einflußfaktoren enthält, gleichzeitig aber einen deutlich niedrigeren numerischen Berechnungsaufwand aufweist.
- da Messungen mit der benötigten Genauigkeit und in der notwendigen Anzahl während des Prozesses nicht zur Verfügung stehen, ist eine Optimierung – hier der Temperaturuniformiät – auf Basis der reduzierten Modelle notwendig. Da ein physikalisches Modell verwendet wird, ist es in einem breiten Bereich anwendbar.
- die Simulation ist nicht genau genug, um einen optimalen Arbeitspunkt festzulegen. Daher werden kritische Parameter an das System angepaßt – hier die Änderung der Temperatur als Funktion der Lampenleistungen.

- dem eigentlichen Regler kommt die deutlich einfachere Arbeit zu, den Arbeitspunkt zu halten oder eine berechnete Trajektorie zu verwenden. So kann die Ordnung des Reglers, also die Zahl der zu regelnden Parameter, herabgesetzt werden.

Der erfolgreiche Einsatz dieses Verfahrens wurde in der Simulation und in Experimenten gezeigt.

5.2 Zukünftige Anwendungen und Grenzen der Methodik

Das in den vorhergehenden Kapiteln vorgestellte Vorgehen läßt sich in angepaßter Weise auch auf andere Reaktortypen in der Halbleiterverarbeitung übertragen. Die Methodik kann dann angewandt werden, wenn

1. ein physikalisches Modell entwickelt werden kann, das alle wesentlichen physikalischen Abhängigkeiten quantitativ erfaßt (im vorigen Abschnitt waren dies das reduzierte Wafer- und das Oxidmodell) und einen geringen numerischen Rechenaufwand hat
2. ein hinreichend genaues Meßverfahren – vorzugsweise in-situ – zur Verfügung steht (im vorigen Fall die Thermoelementmessungen)
3. die Anzahl der meßtechnisch zu bestimmenden Modellparameter aufgrund des Modellverständnisses gering gehalten werden kann und die Parameter untereinander nur gering gekoppelt sind (im vorigen Fall die Kopplungsmatrix zwischen Lampen und Wafer)

Den meisten Prozeßschritten in der Halbleiterproduktion ist gemein, daß nur wenige störungsfreie Meßverfahren während des Prozesses zur Verfügung stehen. Als Beispiele seien hier die Abscheidung aus der Gasphase (*C*hemical *V*apour *D*eposition) und Plasmaätz- und -abscheideprozesse genannt:

Im ersten Fall ist das Erreichen einer vorgegebenen Schichtdicke, z.B. bei einer Siliziumschicht mit Hilfe von SiH_4 oder Si_2H_6 von Interesse. Eine Messung des Schichtwachstums durch Mehrwellenlängenellipsometrie an mehreren Punkten in der Kammer ist schwer durchführbar.

Im Fall des Plasmaätzens ist zum einen das Erzielen einer homogenen Ätzrate an der Waferoberseite von Interesse, zum anderen die Erkennung des Ätzendpunktes, also der Zeitpunkt, an dem die zu entfernende Schicht abgetragen ist und der Ätzprozeß zu beenden ist, um ein Überätzen in darunterliegende Strukturen zu verhindern.

Im letzten Fall ist eine Modellfindung sehr erschwert; eine genaue Kenntnis der tatsächlichen Schichtdicken auf jedem Wafer, z.B. ex-situ vor dem Prozeß gemessen, wäre nötig. Daher konzentrieren sich hier die Forschungsaktivitäten auf eine optische Analyse der Restgase mit einer Erkennung einer veränderten Gaszusammensetzung durch neuronale Netze [82]. Hier ist also Anwendung stark regelungstechnischer Verfahren auf Basis einer an/aus-Steuerung (bang-bang-control) von Vorteil.

Durch die konsequente Zurückführung der Equipmentsimulation auf einfachere Gleichungen, die dennoch die wesentlichen physikalischen Zusammenhänge beinhalten, könnten auch in obigen Fällen Verbesserungen und eine schnellere Optimierung der Kammerbedingungen erzielt werden.

5.2.1 Lampenoptimierung für einen 12-Zoll RTP-Reaktor

Das Verfahren läßt sich natürlich mit wenigen Änderungen auf andere RTP-Systeme übertragen; so wurde das Optimierungsverfahren kürzlich von [77] auf eine 12 Zoll Kammer mit Linearlampen angewendet (Abbildung 5.1), die für Oxidations- und Ausheilschritte bei der Speicherproduktion eingesetzt werden soll. Zum Vergleich verschiedener Kammerkonfigurationen kann nur die optimale Lampeneinstellung herangezogen werden. Gleichzeitig will man eine möglichst gleichmäßige Auslastung der Lampen erreichen, um Reserven für schnelle Aufheizphasen zu haben.

Zur Parameterextraktion für das reduzierte Modell bei großen Waferdurchmesser eignet sich insbesondere die in Kapitel 3.7 entwickelte Methode, da hier auch bei einer geringeren Meßpunktdichte mit Hilfe von simulierten Verteilungen eine gute Charakterisierung des Systems erzielt werden kann.

5.2.2 Regelungsstrategien für Plasmareaktoren

Voraussetzung für das Vorgehen ist die effiziente, aber physikalisch genaue Modellierung. Dies bestimmt auch die Grenzen dieses Ansatzes. Prozesse, bei denen eine Modellierung mit genügender Genauigkeit nur unzureichend möglich

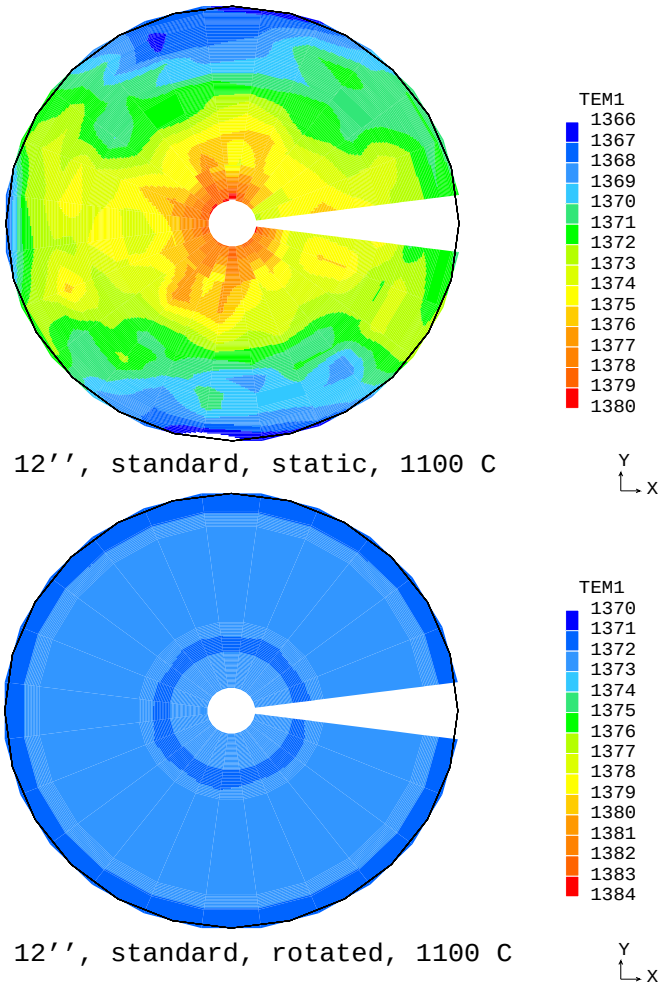


Abbildung 5.1: Optimierungsergebnisse für eine 300 Millimeter RTP-Kammer mit dem in Kapitel 3 erläuterten Verfahren. Die Optimierung wurde auf Basis von Equipmentsimulationen durchgeführt [77]. Links: Bestmögliche Temperaturverteilung für einen unrotiertern Wafer bei 1100° mit einem $\Delta T = T_{max} - T_{min}$ von 12°C. Das System weist in dieser Konfiguration eine geringe Regelbarkeit auf; es gibt keine Lampen, die ausschließlich auf den oberen und unteren Rand strahlen. Die Regelbarkeit des Systems ist deutlich geringer, die Grunduniformität aber höher als bei der Kammer in Kapitel 3. Rechts: Optimierung im rotierten Fall. Hier kann die Temperaturuniformität auf etwa $\pm$1C verbessert werden. Es genügt nicht, das optimierte Rezept des unrotierten Falls zu nehmen: Lampen, die vornehmlich auf den linken und rechten Rand strahlen, können jetzt vom Optimierer zur Kompensation verwendet werden, daher die deutliche Verbesserung.

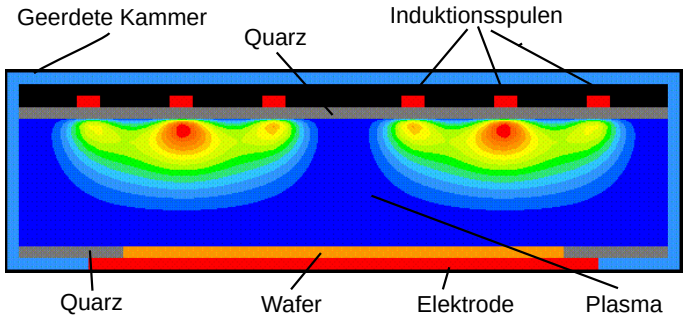


Abbildung 5.2: Konzept des induktiv geheizten Multizonen-Plasmareaktors (links) und Berechnungen der induktiv eingekoppelten Leistung bei einer mittleren Elektronendichten von $n_e = 10^{10} cm^{-3}$ von [78]. Der Strom in den Spulen ist hierbei in allen drei Ringen konstant, verursacht aber aufgrund der Randbedingungen eine unterschiedliche Leistungseinkopplung.

ist, müssen durch herkömmliche Identifizierungsmethoden wie Antwortfunktionen charakterisiert werden. Dazu zählen gegenwärtig z.B. Plasma-Reaktoren.

Von Werner und Brinkmann wurde ein Reaktorkonzept zur Multizonenregelung von Plasmareaktoren vorgestellt [79]. Erfindungsidee ist dabei die unabhängige Ansteuerung mehrerer Spulen in einer induktiv geheizten Plasmakammer. Durch Variation des Stroms durch die einzelnen Spulen (siehe Abbildung 5.2) läßt sich die Ionisation durch induktive Heizung der Elektronen lokal beeinflussen. Ziel ist das Erreichen einer homogenen Ätzrate durch einen uniformen Ionenfluß auf der Waferoberseite.

Allerdings gilt hier im Gegensatz zu dem Wärmeflußmodell bei thermischen Prozessen keine Superposition. Der Einfluß einer Spule auf die Elektronen- und Ionendichte in der Nähe des Wafers ist eine Funktion der Skintiefe und damit der Elektronendichte selbst sowie der mittleren freien Weglänge der Elektronen und damit des Drucks. Der Einfluß einer Stromänderung durch eine der Spulen hängt also von den Strömen der anderen Spulen ab, selbst wenn die induktive Kopplung zwischen den Spulen durch ein Schaltungsnetzwerk kompensiert wird. Eine Kalibrierung eines reduzierten Modells ist somit gegenüber einem RTP-System erschwert.

Der Einsatz der Simulationen zur prinzipiellen Regelbarkeitsanalyse ist aber auch hier zumindest qualitativ möglich und hilft so, den Meßaufwand zu begrenzen. Eine Berechnung des Einflusses der Spulen ist jedoch derzeit nicht mit einer Genauigkeit von besser 30% und einem Rechenaufwand möglich, der dies im Rahmen der Prozeßzeit von wenigen Minuten erlaubt. Modellierungen

finden daher bei der Auswertung der Sensorsignale, z.B. von Langmuir-Proben, eine effektivere Anwendung.

Einige Problemstellungen, wie z.B. die Endpunkterkennung zur Vermeidung des Überätzens, sind indes durch modellbasierte Verfahren nicht adäquat zu behandeln. Hier handelt es sich zumeist um die Problematik der Mustererkennung, wo aber durch Einsatz neuronaler Netze gute Ergebnisse erzielt werden können.

5.2.3 Modellbasierte Regelung für CVD-Systeme

Bei der Regelung von CVD-Prozessen ist das Ziel eine gleichmäßige Abscheidung über den Wafer. An jedem Punkt soll am Ende der Prozeßzeit t_J die Dicke x_J erreicht werden. Ist die Wachstumsrate über die Temperatur regelbar, so ist für jeden Zeitschritt j die zu erreichende Temperatur T_{j+1} so zu wählen, daß der prognostizierte Endwert dem Prozeßziel entspricht:

$$\min_{T_{j+1}^{Modell}} |x_J - x_j + \int_{t_j}^{t_J} r(x(t), T_{j+1}^{Modell}) dt| \tag{5.1}$$

Dies unterscheidet sich von klassischen Regelverfahren, da hier *innerhalb einer vorgegebenen Zeit* das Ergebnis erzielt werden soll.

Linearisiert man $r(x,T) \approx r_0(x_j, T_j) + \left.\frac{\partial r}{\partial T}\right|_{x_j,T_j} (T - T_j)$ so erhält man

$$T_{j+1}^{Modell} = T_j + \frac{x_j + r_0(x_j, T_j)(T_J - T_j)}{\left.\frac{\partial r}{\partial T}\right|_{x_j,T_j} (T_J - T_j)} \tag{5.2}$$

Weitere Linearisierungsverfahren und die Brücke zu klassischen Regelverfahren sind in [51] zu finden.

Die x_j können sowohl über das Reaktionsmodell errechnet oder auch z.B. mittels Schichtdickenmeßverfahren wie Ellipsometrie ermittelt werden.

Zur Stabilisierung des Verfahrens kann eine Gewichtsfunktion $g(t)$ eingeführt werden und die Temperaturänderung ΔT_j von einem Schritt j zum Schritt $j+1$ bestimmt werden zu

$$\Delta T_j = g(t_j)(T_{set} - T_j) + (1 - g(t_j))(T_{j+1}^{Modell} - T_j) \tag{5.3}$$

mit der Prozeßzieltemperatur T_{set}. $g(t)$ wird am Anfang des Prozesses auf 1 gesetzt, um die Zieltemperatur zu erhalten und wird im Laufe des Prozesses

kontinuierlich auf 0 verändert. Vorausgesetzt wird dabei, daß die Wafertemperatur stets homogen bei jeder Nominaltemperatur eingestellt werden kann.

Eine vereinfachte, effektive Beschreibung der Reaktionschemie ist zumeist ausreichend, da der Prozeß in einem beschränkten Temperatur- und Druckbereich gehalten wird.

Als Meßgrößen für ein reduziertes Reaktionsmodell können außerdem z.B. ein Gasflußsensor oder eine Restgasanalyse verwendet werden. Diese Daten sind in vielen Reaktortypen ohnedies vorhanden.

Ist das Einhalten der genauen Prozeßzeit nicht gefordert, so läßt sich das Regelungsverfahren deutlich vereinfachen, indem – ähnlich zur Endpunktregelung bei Plasmareaktoren – die Prozeßzeit vom Regler verlängert oder verkürzt wird.

Der obige Ansatz läßt sich auch mit der Homogenitätsregelung der vorausgegangenen Kapitel kombinieren, dann ist für jede Waferzelle k der Diskretisierung ein ΔT_j^k nach Gleichung 5.3 zu berechnen und zu regeln.

5.2.4 Regelungsverfahren für Vertikalöfen

Die Aufheizung in Vertikalöfen mit bis zu 150 Wafern geschieht über Widerstandsheizungen in den Reaktorwänden. Der Waferstapel wird daher vom Rand her aufgeheizt. Zwei Aufgaben sind während des Rampens des Ofens zu beachten: Da die Aufheizung der Wafer vom Rand her erfolgt, entsteht ein Temperaturgradient über die Scheibe in Abhängigkeit der Aufheizrate. Der Gradient hängt von der Beladung der Kammer ab, die zwischen 30 und 150 Wafern schwanken kann. Des weiteren ist die Aufheizung der Wafer in der Mitte des Stapels langsamer als an den Enden, was eine vertikale Mehrzonenheizung – und damit eine Mehrzonenregelung erforderlich macht.

Da Silizium in dem Aufheizbereich von 20°C bis 800°C semitransparent ist, muß ein modellbasiertes Regelverfahren die steigende Absorptivität berücksichtigen. Ein System, das mittels Umschalten von Regelungsparametern für die unterschiedlichen Temperaturbereiche arbeitet, wurde von Semitool entwickelt und zeigt gute Ergebnisse bei unterschiedlicher Beladung.

Allerdings sind durch eine detaillierte Modellierung Verbesserungen (insbesondere im Übergangsbereich zwischen den Modellen mit unterschiedlichem Gültigkeitsbereich) gegenüber dem Umschalten der Regelparameter erreichbar: Da die Aufheizzeiten im Bereich mehrerer Minuten anstatt Sekunden wie in RTP-Systemen liegen, sind die Anforderungen an die effiziente Auswertung eines

reduzierten Modells deutlich geringer. Das Wafermodell kann durch ein axialsymmetrisches Lösungsverfahren der Wärmediffusionsgleichung ersetzt werden, das dann iterativ gelöst wird. So kann im Fall zu steiler Temperaturgradienten über den Waferradius die Aufheizrate des Gesamtsystems gedrosselt werden.

Da in Vertikalöfen an mehreren Stellen Thermoelemente in-situ vorliegen, kann ein adaptives Regelverfahren zur Echtzeitabschätzung der Absorptivität der Wafer eingesetzt werden.

Als Regelungsverfahren eignen sich die in dieser Arbeit verwendeten Ansätze, insbesondere da in Vertikalöfen z.T. Mehrzonenheizungen vorgesehen sind.

Wie in diesem Kapitel kurz angedeutet, sind die Anwendungsmöglichkeiten physikalisch motivierter Regelungsverfahren vielfältig. Für die Investitionen in ein physikalisches Modell erhält man robuste Regelungsverfahren mit einer Gültigkeit unter vielen Prozeßbedingungen.

Anhang A

Oxidwachstumsmodelle

In Kapitel 3 wurden Oxidwachstumsmodelle zur Umrechnung der Oxiddickenmessungen verwendet. Jedes der erwähnten Modelle wird nachfolgend kurz zusammengefaßt. Für eine detaillierte Darstellung sei auf die Literatur [97] verwiesen.

Da in den in dieser Dissertation betrachteten Reaktoren nur reiner Sauerstoff verwendet wurde, wird auf eine Darstellung der Kinetik anderer Oxidanten verzichtet. Modelle hierzu finden sich z.B. in [92].

A.1 Allgemeine Bemerkungen zur Modellauswahl

Ho [87] konnte an mit Phosphor dotierten Substraten zeigen, daß das Oxidwachstum erst ab hohen Dotierungen $\geq 10^{20} cm^{-3}$ von der Dotierstoffkonzentration abhängt. Bei hohen Temperaturen von etwa 1100°C ist zudem aufgrund der hohen intrinsischen Ladungsträgerdichte keine Abhängigkeit von der Dotierung mehr zu beobachten [95]. Bei dem vorliegenden Tunneloxid über dem Programmiergate des EEPROM mit moderaten Kanaldotierung ist diese Abhängigkeit daher zu vernachlässigen.

Der genaue Mechanismus der Oxidation ist Gegenstand zahlreicher Untersuchungen. Ein allgemein akzeptiertes Modell existiert bis dato nicht. Insbesondere der mikroskopische Mechanismus des Oxidwachstums dünner Oxide ist nicht bekannt. Folgende Punkte sind die Haupteinflußfaktoren auf die Kinetik des Oxidwachstums [93] [90] [94]

1. Gasphasentransport des Sauerstoffs O_2 und anderer Sauerstoffmoleküle wie O_2^- zur SiO_2 Oberfläche
2. Diffusion der Sauerstoffmoleküle durch den SiO_2 Bulkbereich
3. Reaktionskinetik an der SiO_2/Si Grenzfläche (eine Diffusion von Sauerstoff tiefer in den Si-Bulkbereich wird nicht experimentell beobachtet)

A.2 Modell von Deal/Grove und Irene/Plummer

Für Oxide > 50nm liefert das Standardmodell von Deal/Grove [87], das auf Basis von Stetigkeitsüberlegungen für den Transport des Sauerstoffs durch das SiO_2 beruht, einen Ausdruck für die Oxiddicke der Form

$$x_{ox}^2 + Ax_{ox} = B(t + t_0) \tag{A.1}$$

$$x_{ox}^2 \approx Bt \quad \text{diffusionsdominiertes Regime} \tag{A.2}$$

$$\text{für } x_{ox} \gg A \approx 0.4\mu m \text{ bei } 900^\circ\text{C} \tag{A.3}$$

$$x_{ox} \approx B/A(t + t_0) \text{reaktionsdominiertes Regime}$$

$$\text{für } x_{ox} \ll A$$

Dabei ist x_{ox} die Dicke des Oxids. Die Zeit t_0 berücksichtigt ein anfänglich vorhandenes Oxid.

Irene/Plummer [90] verwenden zur Beschreibung der Oxidation eine Erweiterung des Oxidationsmodell von Deal und Grove mit zwei Korrekturtermen für das Anfangswachstum dünner Oxide < 30nm. Differenziert man Gleichung A.1 so ergibt sich mit den beiden Korrekturtermen:

$$\frac{dx_{ox}}{dt} = \frac{B}{A + 2x_{ox}} + \left(c_{thin1} e^{-\frac{E_{thin1}}{kT}} e^{-\frac{x_{ox}}{L_{thin1}}} + c_{thin2} e^{-\frac{E_{thin2}}{kT}} e^{-\frac{x_{ox}}{L_{thin2}}} \right) \left(\frac{p_{O_2}}{1atm} \right)^{\alpha} \tag{A.4}$$

mit der linearen und der parabolischen Ratenkonstante

$$B/A = c_{linear}\, e^{-\frac{E_{linear}}{kT}} \left(\frac{p_{O_2}}{1atm} \right)^{\rho(T)} \tag{A.5}$$

$$B = c_{parabol}\, e^{-\frac{E_{parabol}}{kT}} \left(\frac{p_{O_2}}{1atm} \right). \tag{A.6}$$

Die Aktivierungsenergien

$$E_{thin1}(<100>) = 2.24\text{eV}, \quad (A.7)$$
$$E_{thin2}(<100>) = 2.33\text{eV}, \quad (A.8)$$
$$E_{thin2}(<111> = 1.80\text{eV}), \quad (A.9)$$
$$E_{linear}(<100>) = -2.0\text{eV und} \quad (A.10)$$
$$E_{parabol}(<100>) = -1.23\text{eV} \quad (A.11)$$

sowie die charakteristischen Längen $L_{thin1}(<100>) = 12.4\text{Å}(<111> 9\text{Å})$ und $L_{thin2}(<100>) = 78\text{Å}(<111> 60\text{Å})$ sind abhängig von der Substratorientierung [73].

Die Abhängigkeit der Wachstumsrate vom Sauerstoffpartialdruck verhält sich, wie den Daten von Massoud [89] zu entnehmen ist, bei dünnen Oxiden nach einem Potenzgesetz der Form $\frac{dx_{ox}}{dt} \propto (\frac{p}{1atm})^{0.55}$. Die lineare Oxidrate B/A geht nach Daten von Hu [88] mit $\frac{dx_{ox}}{dt} \propto (\frac{p}{1atm})^{0.55+0.32\frac{T-900K}{400K}}$.

A.3 Oxidationsmodell von Wolters

Zur Erklärung des raschen Anfangswachstums wurde von Wolters et al. [94] basierend auf einem Potentialtopfmodell für den Transport von Oxidionen durch das Oxid zur Si/SiO_2 Grenzfläche ein Ausdruck für die Oxidationsrate der Form

$$\frac{dx_{ox}}{dt} = \frac{J_{ion}}{N_{ox}q} = \frac{\nabla\mu_{O_2}}{N_{ox}zq^2}\sigma_{Ion} \quad (A.12)$$
$$\sigma_{Ion} = \frac{\sigma_0}{2}\exp\left(-\frac{W + zqaU_{ox}/x_{ox}}{kT}\right)\left(\frac{x_{ox}}{x_{ox}^0} + 1\right)^{-\delta} \quad (A.13)$$
$$\Rightarrow \frac{dx_{ox}}{dt} \approx Cx_{ox}^{-\beta} \text{ für } x_{ox} \gg x_{ox}^0 \quad (A.14)$$
$$x_{ox} = (C(1+\beta))^{\frac{1}{1+\beta}}t^{\frac{1}{1+\beta}} \quad (A.15)$$

abgeleitet, das auf den Ionentransport im Oxid abzielt.

Dabei ist W die Höhe und a die räumliche Ausdehnung der Potentialbarrieren für den Transport der mit zq geladenen Sauerstoffionen im Oxid, σ_0 die Leitfähigkeit der Ionen bei hohen Temperaturen, U_{ox} das durch feste Oxidladungen nahe der Si/SiO_2 und Raumladungen hervorgerufene Feld, N_{ox} die Anzahl von SiO_2 Molekülen pro Volumen und $\nabla\mu_{O_2}$ der Gradient des thermodynamischen Potentials. $\delta \approx 0.76$ enthält einen Ausdruck, der die Abhängigkeit

der Potentialbarriere von den Ladungen im Oxid und der thermischen Energie der wandernden Ionen beschreibt; x_{ox}^0 entspricht dem mittleren Abstand von Raumladungen im Oxid. Die Reaktionen an den Grenzflächen werden nicht modelliert.

Die Koeffizienten C und β wurden für zahlreiche veröffentlichte Oxidmessungen angepaßt [96], sind allerdings sehr empfindlich von den Prozeßbedingungen abhängig. Für sehr dünne Oxide von $\leq 3nm$ kann mit dem Modell von Wolters eine sehr gute Übereinstimmung mit Messungen durch geeignete Wahl von β auch bei sehr dünnen Oxiden $<$ 5nm erzielt werden. Das rasche Wachstum innerhalb der ersten Sekunden beschreibt dagegen das Irene/Plummer Modell nur unzureichend.

In [47] wurde für den untersuchten Reaktortyp Werte von $\alpha := \frac{1}{1+\beta} = 0.67$ bei einem Druck von $0.5atm$ und $T = 1000°\mathrm{C}$ gefunden, allerdings ist die Temperaturabhängigkeit entgegengesetzt zu den Werten von [95].

A.4 Modell von Han und Helms

Han und Helms [86] [84] führen das rasche Wachstum dünner Oxide auf die Reaktionskinetik zweier diffundierer Spezies – O_2^- und nach außen diffundierende Vakanzen – zurück. Die Oxidation verläuft über die Schritte Adsorption, Einbringung in SiO_2, Diffusion zur Grenzfläche, Dissoziation an der Grenzfläche und Reaktion $Si + 2[O] \rightarrow SiO_2$. Nach Auflösen der Gleichungen nach der Generationsrate an der Si-Grenzfläche erhält man:

$$\frac{dx_{ox}}{dt} = \left(-G\left(1+\frac{x_{ox}}{\ell_{thin}}\right) + \sqrt{G^2\left(1+\frac{x_{ox}}{\ell_{thin}}\right)^2 + Kp}\right)\left(1+\frac{\beta}{2x_{ox}}\right) \quad \text{(A.16)}$$

$$\text{für } x_{ox} \ll L, \text{ mit } L(900°\mathrm{C}) = 1450\text{Å}, L(1200°\mathrm{C}) = 450\text{Å folgt}$$

$$\approx \left(-1+\sqrt{1+\frac{Kp}{G^2}}\right)\frac{G\beta}{2x_{ox}} \quad \text{(A.17)}$$

mit an experimentelle Daten angepaßten Konstanten $G(T)$, $\ell_{thin}(T)$ und $K(T)$, die Verhältnisse der Reaktionsraten der beteiligten Transport- und Reaktionsprozesse beinhalten. $\beta(T)/x_{ox}$ beschreibt hier den Einfluß der diffundierenden Vakanzen. Für eine Reihe von in der Literatur verfügbaren Messungen liefert das Modell eine Übereinstimmung bis auf einige Prozent, insbesondere bei der

Oxidation von HF-vorbehandelten Wafern, die auch im in dieser Arbeit betrachteten Clustertool vorliegen.

Für alle Modelle gilt, daß die Anpaßbarkeit an vorhandene Meßdaten groß ist, da nahezu alle Parameter arrheniusartigen Temperaturabhängigkeiten gehorchen und damit die Anzahl der anpaßbaren Parameter groß ist. Keines der diskutierten Modelle erklärt den Oxidationsprozeß für alle Prozeßbedingungen konsistent. Eine mikroskopische Simulation, z.B. mit Molekulardynamikverfahren, könnte Aufschluß über die dominanten Prozesse liefern.

A.5 Vergleich der Modelle mit Messungen

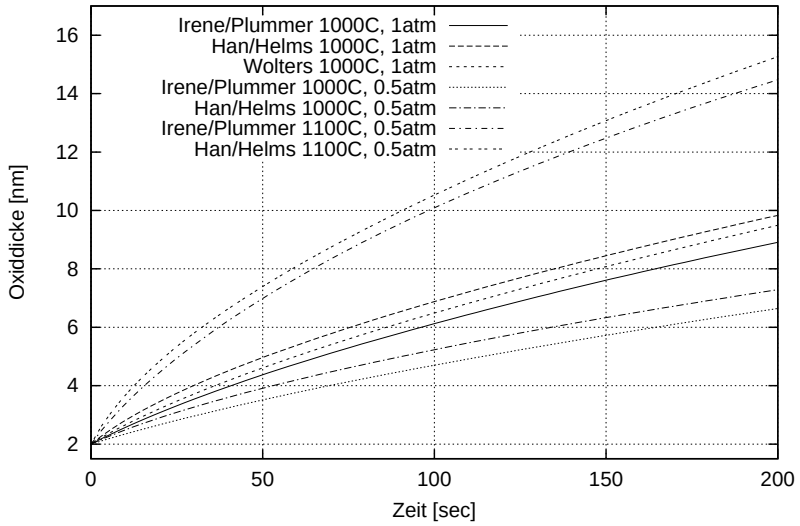


Abbildung A.1: Vergleich der drei diskutierten Oxidmodelle von Irene/Plummer, Wolters/van-Duynhofen und Han/Helms.

Für den betrachteten Prozeßbereich über 3nm Oxiddicke ergeben alle Modelle ähnliche Werte (siehe Abbildung A.1) mit einer Differenz von etwa 10%. Bei der Auswertung der Oxidmessungen in dieser Arbeit wurde dem Modell von Han/Helms wegen der besser verstandenen Druckabhängigkeit der Vorzug gegenüber dem Wolters-Modell gegeben. Als Koeffizienten wurden die in der Literatur angegebenen Standardwerte verwendet.

Die oben genannten Modelle sind an Oxidmessungen aus Ofenprozessen angepaßt worden. Bei RTP-Prozessen liefern die hier genannten Modelle etwas zu geringe Oxidationsraten. In Abbildung A.2 wird ein Vergleich zwischen

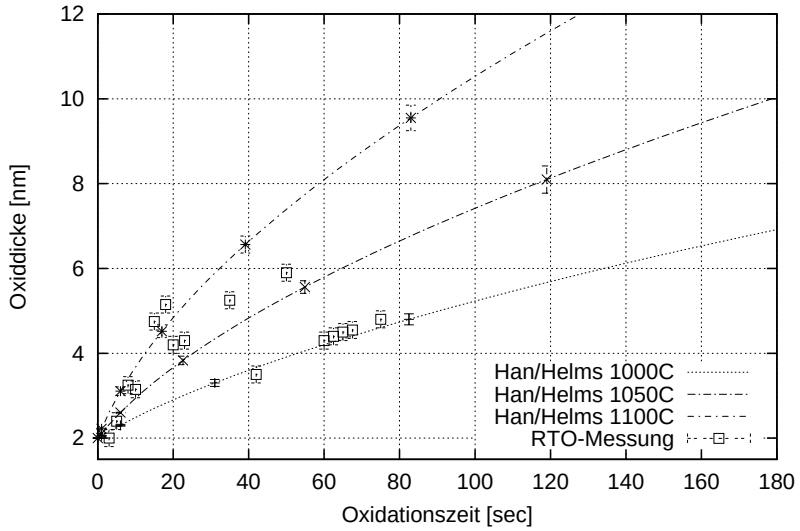


Abbildung A.2: Vergleich des in dieser Arbeit favorisierten Oxidmodells von Han/Helms mit Messungen an der Reaktorkammer von [47]. Die durchgezogenen Linien sind die simulierten Oxiddicken. Im Fehlerbalken der Messungen ist die typische Reproduzierbarkeit der Werte wiedergegeben. Die Fehlerbalken der simulierten Kurven stellen die Änderung der Oxiddicke bei einer Variation der Temperatur um 5°C dar, was der Genauigkeit der Pyrometermessungen entspricht.

Han/Helms Modells mit Standardparametern und RTO-Messungen [47] gezogen. Bei den Messungen ist allerdings zusätzlich eine Stabilisierungsphase sowie eine Aufheiz- und Abkühlphase von mehreren Sekunden vorhanden, so daß die Oxiddicke leicht nach unten zu korrigieren ist. Hinzu kommt die Meßungenauigkeit des Pyrometers bei der Temperaturbestimmung und die sehr kurze Prozeßzeit von nur wenigen Sekunden bei hohen Temperaturen.

Ein möglicher Grund für die Differenz liegt an den unterschiedlichen Prozeßbedingungen in Ofen- und RTP-Prozessen. RTP-Systeme sind lampengeheizt mit einem Strahlungsmaximum der Lampen bei etwa 1μm. Das Strahlungsspektrum ist gegenüber Ofenprozessen zu kürzeren Wellenlängen verschoben, was zum einen die Dissoziation von $O_2^- + h\nu \rightarrow O^- + O$ begünstigt, zum anderen zur Anreicherung von photonisch erzeugten Elektronen an der Si/SiO_2-Grenzschicht und damit zu einer erleichterten Diffusion der Sauerstoffionen führt. In Gleichung A.13 erniedrigt sich so das Potential U_{ox}.

Anhang B

Numerische Methoden

Dieses Kapitel gibt einen groben Überblick über die verwendeten Routinen. Eine detailliertere Übersicht ist der angegebenen Literatur zu entnehmen. Die Methoden wurden den Anforderungen der einzelnen Regelungsprobleme angepaßt und erweitert.

B.1 Optimierungsverfahren für Funktionen mehrerer Veränderlicher

Eine der Kernfragen der Regelungsverfahren in den vorausgegangenen Kapiteln war das Ermitteln eines Optimalwerts einer Zielfunktion, z.B. der Homogenität der Wafertemperatur oder des thermischen Budgets. In den erstellten Optimierungsprogrammen wurden je nach Problemstellung und verfügbaren Informationen unterschiedliche Optimierungsverfahren eingesetzt.

Im folgenden werden die beiden Klassen der Minimierungsalgorithmen kurz dargestellt, für detaillierte Darstellungen sei auf die Literatur verwiesen.

B.1.1 Minimierung bei bekannten Gradienten

Sind die Richtungsableitungen der zu minimierenden Funktion bekannt, wie z.B. beim Optimierungsverfahren der Wafertemperatur zeigt ein konjugierte Gradientenverfahren die besten Konvergenzeigenschaften. Die zu minimierende Funktion $f(\mathbf{x})$ läßt sich durch eine quadratische Form approximieren

$$f(\mathbf{x}) \approx \mathbf{c} - \mathbf{b} \cdot \mathbf{x} + \frac{\mathbf{1}}{\mathbf{2}} \mathbf{x} \cdot \mathbf{A} \cdot \mathbf{x} \tag{B.1}$$

$$\nabla f(\mathbf{x}) \approx \mathbf{A} \cdot \mathbf{x} - \mathbf{b} \tag{B.2}$$

Die Berechnung und Inversion der Hessematrix $\mathbf{A}$ ist numerisch zu aufwendig. Strategie aller nicht-stochastischer multidimensionaler Minimierungsverfahren ist die Richtungsminimierung einer Funktion entlang von Richtungsvektoren, also zu jedem gegebenem Vektor $\mathbf{h}$ löse

$$\min_{\lambda} f(\lambda \mathbf{h}) \tag{B.3}$$

Konjugierte Gradientenverfahren konstruieren die Sequenz von Richtungsvektoren so, daß jeder Richtungsvektor senkrecht zu allen bisherigen Minimierungsrichtungen ist.

Der Algorithmus lautet

	Konjugierte Gradienten Minimierung
1	Wähle Startwert $\mathbf{x_0}$, $\mathbf{g}_0 = \mathbf{h}_0 = -\nabla f(\mathbf{x_i})$, $i = 1$
2	Bestimme $\mathbf{x}_i = \mathbf{x}_{i-1} + \kappa \mathbf{h}_{i-1}$ mit $\kappa = \min_\lambda f(\lambda \mathbf{h}_{i-1})$
3	falls $\lvert f(\mathbf{x}_i) - f(\mathbf{x}_{i-1}) \rvert < tol$, gehe zu 6
4	$\mathbf{g}_i = -\nabla f(\mathbf{x}_i)$, $\mathbf{h}_i = \dfrac{(\mathbf{g}_i - \mathbf{g}_{i-1}) \cdot g_i}{g_{i-1} \cdot g_{i-1}} h_{i-1} - g_i$
5	$i \leftarrow i + 1$ und gehe zu 2
6	$\mathbf{x}_i$ ist Minimum

Für einen Beweis der Orthogonalität der $\mathbf{g}_i$ siehe [101].

Das konjugierte Gradientenverfahren wurde für die Zweischrittoptimierung des optimalen Temperaturprofils eingesetzt. Für diese Problemstellung benötigt das Verfahren in etwa die gleiche Anzahl von Schritten zur Minimierung wie quasi-Newton Methoden mit erhöhtem Speicherbedarf.

B.1.2 Minimierung ohne Gradienteninformation

Für Verfahren ohne Gradienteninformationen, wie sie z.B. bei dem ersten Schritt der uniformen Lampeneinstellung notwendig sind, eignet sich das Powell-Verfahren. Ebenso wie das konjugierte Gradienten Verfahren im vorigen Abschnitt basiert die Minimierung auf der Konstruktion von konjugierten Richtungen, d.h. für für zwei aufeinanderfolgende Richtungsminimierungen entlang h_i und h_{i+1} soll

$$\mathbf{h}_i \cdot \mathbf{A} \cdot \mathbf{h}_{i+1} \equiv 0 \tag{B.4}$$

gelten.

Das Verfahren von Powell verwendet dazu folgenden Algorithmus zum Erreichen quadratischer Konvergenz

	Powell-Brent Minimierung
1	Startwert $\mathbf{x}_0$, $u_i = \mathbf{e}_i$, $i = 1, \ldots, N$
2	$\phi = f(\mathbf{x}_0)$
3	Für $i = 0, \ldots, N$: $x_i = x_{i-1} + \kappa \mathbf{h}_i$ mit $\kappa = \min_\lambda f(\lambda \mathbf{h}_i)$
4	Lasse Richtung u_1 fallen, für $i = 1, \ldots, N$: $h_i \leftarrow h_{i+1}$
5	$h_N = x_N - x_0$
6	$x_0 = x_N + \kappa \mathbf{h}_N$ mit $\kappa = \min_\lambda f(\lambda \mathbf{h}_N)$
7	Falls $\lvert f(x_0) - \phi \rvert < tol$ Ende
8	alle N Iterationen führe Singulärwertzerlegung der h_i, $i = 1, \ldots, N$, aus und lasse Richtungen mit zu kleinen Singulärwerten fallen
9	gehe zu 2

Der numerische Aufwand ist ungleich höher als bei Verfahren mit Gradienteninformation, da jeweils N Richtungsminimierungen nötig sind.

B.1.3 Einbringen von Nebenbedingungen

Für die Optimierung der Temperaturhomogenität war die Beachtung der Leistungsgrenzen der Lampen erforderlich. Diese läßt sich durch eine Ungleichheitsnebenbedingung der Form

$$P_{min} \leq P_l \leq P_{max} \text{ für alle } l \tag{B.5}$$

formulieren.

Zwei Verfahrensweisen zum Einbringen der Nebenbedingungen wurden in dieser Arbeit verwendet. Bei einfachen Schranken wie in Gleichung B.5 wird nach einem Zulässigkeitsverfahren ("trust region") vorgegangen.

In dem Richtungsminimierungsproblem in Gleichung B.3 wird der Skalierungsparameter λ nur in einem Bereich variiert, bei dem keine der Nebenbedingungen aus Gleichung B.5 verletzt wird. Das Konvergenzverhalten der Minimierung

wird aber im allgemeinen schlechter, da die Orthogonalität der Richtungsvektoren beim knojugierten Gradientenverfahren nicht mehr garantiert ist. Bei komplizierteren Nebenbedingungen eignet sich die Addition einer Straffunktion zu der zu minimierenden Funktion, im obigen Fall wäre das

$$p(P, \mathbf{a}_n) = \sum_l a_{n,l} \left((\max(0, P_{min} - P_l))^2 + (\max(0, P_l - P_{max}))^2 \right) \qquad \text{(B.6)}$$

Es werden nacheinander N Optimierungen mit wachsendem $a_{n,l}, n = 1, \ldots, N$ durchgeführt, falls Nebenbedingungen verletzt werden. Die Minimum der einzelnen Optimierungen konvergieren dann gegen einen Wert am Rand des zulässigen Gebiets. Die Quadrate garantieren Differenzierbarkeit der zu minimierenden Funktion. Ist die Funktion außerhalb des zulässigen Bereichs nicht auswertbar, so kann ein Folge mit gegen Null gehenden Koeffizienten a_n mit sogenannten Grenzfunktion ("barrier functions") verwendet werden. Dann ist:

$$p(P, \mathbf{a}_n) = -\sum_p a_{n,l} \left(\log(P_{min} - P_p + s_{min}) + \log(P_p - P_{max} + s_{max}) \right) \qquad \text{(B.7)}$$

Dabei sind s_{min} und s_{max} Verschiebungen, für die die Funktion noch auswertbar ist und die ebenfalls in der Minimierungsfolge gegen Null gehen.

Bei Gleichheitsnebenbedingungen, wie sie z.B. bei der Einhaltung einer vorgegebenen Schichtdicke gefordert werden, eignen sich entweder die Addition über Straffunktionen oder die Einführung von Lagrange-Faktoren, die auf die Kuhn-Tucker Gleichungen führen. Lösungsalgorithmen, wie z.B. sequentielle quadratische Programmierung sind gegebenenfalls in [103] zu finden.

B.1.4 Stochastische Minimierungsverfahren

Die in den vorausgegangenen Kapiteln beschriebenen Optimierungsverfahren neigen zur Konvergenz in Nebenminima und reagieren bei Annäherung an das Optimum empfindlich auf Fehler in den Funktionswerten, wie sie z.B. beim numerischen Rauschen von Simulationslösungen auftreten.

Für Verfahren wie in Kapitel 4.5 eignet sich daher ein statistisches Verfahren wie das Simulated Annealing besser.

B.2 Lösungsverfahren für die Waferdifferentialgleichung

Auf die Lösung der Strahlungstransportgleichung nach dem Verfahren von Kersch und Morokoff wurde bereits in [61] eingegangen. Dabei wird ein Monte

Carlo-Verfahren mit Varianzreduktion über Quasizufallszahlen und Importance-Sampling eingesetzt.

Der Verfahrenssimulator PHOENICS setzt zur Lösung der partiellen Differentialgleichungen für die Diffusion das iterative Verfahren nach Gauß-Seidel ein. Dabei ergibt sich ausgehend von einer Startlösung $\mathbf{x}^{(k)}$ für die nächsten Lösung $\mathbf{x}^{(k+1)}$ die Iterationsvorschrift

$$x_i^{(k+1)} = (1-\omega)x_i^{(k)} + \frac{\omega}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + \sum_{j=1}^{i+1} a_{ij}x_j^{(k)}\right) \tag{B.8}$$

wobei die a_{ij} die Elemente der Matrix in der zu lösenden Gleichung $\mathbf{Ax} = \mathbf{b}$ sind und ω ein Relaxationsfaktor zur Konvergenzbeschleunigung ist.

Die Simulationsdomäne wird in xy-Schichten, sog. Slabs, durchlaufen.In jeder Schicht wird zunächst Konvergenz erreicht und dann zur nächsten Schicht weitergegangen. Dieses Durchlaufen der Schichten ("Sweep") wird bis zur entgültigen Konvergenz wiederholt. Die nichtlinearen Quellterme, wie z.B. für die Strahlung, werden für jeden Durchlauf erneut berechnet.

Für die Lösung der Waferdifferentialgleichung erwies sich das PHOENICS- Verfahren als nicht optimal. Eine Lösung mit einem schnelleren iterativen Solver, dem stabilisierte bikonjugierte Gradientenverfahren [102], ist um etwa eine Faktor 8-10 effektiver.

Die bikonjugierte Methode wurde eingesetzt, um auch eine unsymmetrische Kopplung zwischen Nachbarzellen berücksichtigen zu können. Das bikonjugierte Verfahren sucht die ungefähre Lösung des Gleichungssystems in einem m-dimensionalen Unterraum $\mathcal{N} = span\left\{\mathbf{h}_1, \mathbf{A}\cdot\mathbf{h}_1, \ldots, \mathbf{A}^{m-1}\cdot\mathbf{h}_1\right\}$ mit $\mathbf{h}_1 = \frac{\mathbf{r_0}}{||\mathbf{r_0}||}$ und $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\cdot\mathbf{x}_0$ mit Vektoren aus einem m-dimensionalen Unterraum $\mathcal{M} = span\left\{\mathbf{g}_1, \mathbf{A}^T\cdot\mathbf{g}_1, \ldots, \mathbf{A}^{Tm-1}\cdot\mathbf{g}_1\right\}$ [1].

Ohne Ableitung sind hier die Algorithmen des konjugierte und bikonjugierte Gradientenverfahren angegeben. Dabei ist $\mathbf{P}$ eine leicht invertierbare Matrix zu Präkonditionierung.

[1] Für das einfache Gauß-Seidel Verfahren gilt $\mathcal{M} = \mathcal{N} = span(\mathbf{e}_i)$, $i = 1, \ldots, N$, da nacheinander die Gleichung für jede Zelle erfüllt wird

	Konjugierte Gradienten	Stabilisierte bikonjugierte Gradienten
1	$\mathbf{r}_0 = \mathbf{b} - \mathbf{A} \cdot \mathbf{x_0}$, $j = 0$	
	$\mathbf{z}_0 = \mathbf{P}^{-1}\mathbf{r}_0$, $\mathbf{p}_0 = \mathbf{z}_0$	$\mathbf{r}_0^*$ beliebig, $\mathbf{p}_0 = \mathbf{r}_0$, $\mathbf{z}_0 = \mathbf{P}^{-1}\mathbf{r}_0$
2	$\alpha_j = \dfrac{\mathbf{r}_j \cdot \mathbf{z}_j}{\mathbf{p}_j^T \cdot \mathbf{A} \cdot \mathbf{p}_j}$	$\alpha_j = \dfrac{\mathbf{r}_0^* \cdot \mathbf{r}_j}{\mathbf{r}_0^* \cdot \mathbf{A} \cdot \mathbf{z}_j}$
3		$\mathbf{s}_j = \mathbf{r}_j - \alpha_j \mathbf{A} \cdot \mathbf{z}_j$
4		$\hat{\mathbf{s}}_j = \mathbf{M}^{-1}\mathbf{s}_j$, $\mathbf{s}_j^* = \mathbf{A}\hat{\mathbf{s}}_j$
5		$\omega_j = \dfrac{\mathbf{s}_j^T \cdot \mathbf{s}_j^*}{\mathbf{s}_j^{*T} \cdot \mathbf{s}_j^*}$
6	$\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j \mathbf{p}_j$	$\mathbf{x}_{j+1} = \mathbf{x}_j + \alpha_j \mathbf{z}_j + \omega_j \hat{\mathbf{s}}_j$
7	$\mathbf{r}_{j+1} = \mathbf{r}_j - \alpha_j \mathbf{A} \cdot \mathbf{p}_j$	$\mathbf{r}_{j+1} = \mathbf{s}_j - \omega_j \mathbf{s}_j^*$
8	$\mathbf{z}_{j+1} = \mathbf{P}^{-1} \cdot \mathbf{r}_{j+1}$	$\mathbf{z}_{j+1} = \mathbf{P}^{-1} \cdot \mathbf{p}_{j+1}$
9	$\beta_j = \dfrac{\mathbf{z}_{j+1}^T \cdot \mathbf{r}_{j+1}}{\mathbf{z}_j^T \cdot \mathbf{r}_j}$	$\beta_j = \dfrac{\mathbf{r}_0^* \cdot \mathbf{r}_{j+1}}{\mathbf{r}_0^* \cdot \mathbf{r}_j} \dfrac{\alpha_j}{\omega_j}$
10	$\mathbf{p}_{j+1} = \mathbf{z}_{j+1} + \beta_j \mathbf{p}_j$	$\mathbf{p}_{j+1} = \mathbf{r}_{j+1} + \beta_j (\mathbf{p}_j - \omega_j \mathbf{A} \cdot \mathbf{z}_j)$
11	falls nicht konvergiert, $j \leftarrow j + 1$ und gehe zu 2	

In der Diskretisierung von Wafer und Guardring und nach Integration über ein Zellvolumen und Linearisierung der Strahlungsterme ergibt sich für die zu lösende Gleichung für den Zeitschritt $t_h \to t_{h+1}$ für ein Zellvolumen ij im Wafer

$$\begin{aligned}
A_{ij}q_{ij}^{in} &= \frac{V_{ij}c_p(T_{ij})}{t_{h+1}-t_h}(T_{ij}^{(h+1)} - T_{ij}^{(h)}) \\
&\quad -\rho_{i+1j}(T_{i+1j}^{(h+1)} - T_{ij}^{(h+1)}) \\
&\quad -\rho_{ij}(T_{i-1j}^{(h+1)} - T_{ij}^{(h+1)}) \\
&\quad -\theta_{ij+1}(T_{ij+1}^{h+1} - T_{ij}^{(h+1)}) \\
&\quad -\theta_{ij}(T_{ij-1}^{h+1} - T_{ij}^{(h+1)}) \\
&\quad +8A_{ij}\sigma e_{ij}^{eff}(1-r_{ij})\left(T_{ij}^{(h)}\right)^3 (T_{ij}^{(h+1)} - T_{ij}^{(h)}) \\
&\quad -\sum_{m,n} W_{m,n\to i,j}\left(T_{mn}^{(h)}\right)^3 (T_{mn}^{(h+1)} - T_{ij}^{(h)}) \qquad \text{(B.9)} \\
\rho_{ij} &= (\lambda_{i+1j}^{-1} + \lambda_{ij}^{-1})^{-1} h_{wafer}(\phi_j - \phi_{j-1})\frac{1}{2}\frac{r_{i+1}+r_i}{r_{i+1}-r_i} \qquad \text{(B.10)} \\
\theta_{ij} &= (\lambda_{ij+1}^{-1} + \lambda_{ij}^{-1})^{-1} 2h_{wafer}\frac{r_{i+1}-r_i}{(\phi_{j+1}-\phi_{j-1})(r_{i+1}+r_i)} \qquad \text{(B.11)} \\
q_{ij}^{in} &= \sum_l L_{ijl}P_l + 2\sigma e_{ij}^{eff}(1-r_{ij})\left(T_{ij}^{(h)}\right)^4 + q_{ij}^{ext} \qquad \text{(B.12)}
\end{aligned}$$

wobei q_{ij}^{ext} die von der Wafer- und der Guardringtemperatur unabhängigen Term von Lampen, Wänden etc. einschließt.

Am Waferrand und in der Wafermitte verschwinden die Transportterme für die Wärmeleitung nach außen bzw. innen, bei der Randzelle ist im Abstrahlungs- und Absorptionsterm die Seitenfläche mitzuberücksichtigen. In Winkelrichtung ϕ gelten zyklische Randbedingungen.

Aufgrund des linearisierten Strahlungsterms ist die Kopplungsmatrix voll besetzt. Da aber die Temperatur der Waferzellen dominant durch die Absorption der Lampenstrahlung sowie durch die emittierte und vom Quarz absorbierte Strahlung bestimmt wird, ist die nichtlineare Kopplung der Zellen untereinander eher gering.

Häufig verwendete Zeichen

$a_{\nu,T}$	Richtungsabhängige Absorptivität
c	Lichtgeschwindigkeit $3 \times 10^8\,[\mathrm{m\,s^{-1}}]$
c_p	molare Wärmekapazität $[\mathrm{J\,kg^{-1}\,K^{-1}}]$
d_{wafer}	Dicke des Wafers $= 0.65\mathrm{mm}$
$e_{\nu,T}$	Richtungsabhängige Emissivität
$\mathbf{G}$	Gebhart-Matrix $[\mathrm{m^{-2}}]$
h	Planck Konstante $6.6210^{-34}\,[\mathrm{J\,s}]$
k_B	Boltzmann Konstante $1.3807 \times 10^{-23}\,[\mathrm{W\,K^{-1}}]$
ℓ	Mittlere freie Weglänge [m]
$\mathcal{L}$	Charakteristische Länge des Systems [m]
$\mathbf{L}$	Strahlungsaustauschmatrix Lampen $\rightarrow$ Wafer $[\mathrm{m^{-2}}]$
m_s	Molare Masse der Spezies s $[\mathrm{kg\,mol^{-1}}]$
$\mathbf{P}$	Strahlungsleistung der Lampen [W]
q	Flußdichte $[\mathrm{W\,m^{-2}}]$
Q	Wärme- oder Strahlungsfluß [W]
$\mathcal{Q}^{\mathrm{shower}}$	Gasfluß in den Reaktor in $[\mathrm{kg\,s^{-1}}]$
t	Zeit [s]
$t_{\nu,T}$	Transmissivität eines semitransparenten Mediums
T	Temperatur [K]
$\mathbf{v}$	Geschwindigkeitsvektor $[\mathrm{m\,s^{-1}}]$
$\mathbf{W}$	Strahlungsaustauschmatrix Umgebung $\rightarrow$ Wafer $[\mathrm{K^{-4}}]$
α_ν, T	Absorptionskoeffizient des Mediums $[\mathrm{m^{-1}}]$
$\Lambda(x)$	Geometriefunktion der Wärmeleitung
λ	Wärmeleitfähigkeit $[\mathrm{W\,m^{-1}\,K^{-1}}]$
σ_B	Stefan-Boltzmann Konstante $5.6705 \times 10^{-8}\,[\mathrm{W\,m^{-2}\,K^{-4}}]$
μ	Dynamische Viskosität $[\mathrm{kg\,m^{-1}\,s^{-1}}]$
ν	Frequenz der Strahlung $[\mathrm{s^{-1}}]$
ρ	Dichte $[\mathrm{kg\,m^{-3}}]$

Literaturverzeichnis

[1] D.E. Hicks, *Evolving Complexity and Cost Dynamics in the Semiconductor Industry.* IEEE Trans. Semic. Man., Vol. 9, 3, 1996

[2] J.D. Meindl, *Limits and opportunities for gigascale integration (GSI)*, SISPAD Proc. 1995

[3] Solid State Technology, Oct. 1996, Tulsa, USA

[4] *World News.* Solid State Technology, Jan. 1996, Tulsa, USA

[5] A. Weinberg, S. Bahl, Applied Materials

[6] Semiconductor Industry Association, *World Semiconductor Trade Statistics.* San Jose, 1998

[7] A.R. Alvarez, *Process Requirements Through 2001*, Proc. RTP '94

[8] P. Gargini et.al., *The SIA's 1997 National Technolog Roadmap for Semiconductors.* Sol.St.Techn. 1/98

[9] R.P.S. Thakur et.al., *Rapid Thermal Processing - Manufacturing Perspective.* Mat. Res. Soc. Symp. Proc. Vol. 387, p. 187

[10] Semiconductor Industry Association, *International Technology Roadmap for Semiconductors: 1999*, Austin, SEMATECH, 1999

[11] O. Prigge, M. Suetake, M. Miura-Mattausch, *Worst/Best Device and Circuit Performances for MOSFETs Determined from Process Fluctuations.* IEICE Transactions 1999, Vol.E82-C No.6

[12] P. Horrowitz, W. Hill, *The Art of Electronics.* Cambirdge Univ. Press, NY 1990

[13] R. Mahnkopf, T. Schafbauer et al., *"System on a chip" Technology Platform for* 0.18μ *Digital, Mixed Signal & eDRAM Applications.* Proc. IEDM 1999

[14] H.A. Lord, *Thermal and Stress Analysis of Semiconductor Wafers in a Rapid Thermal Processing Oven*, IEEE Transactions on Semicon. Manuf., vol.1, no.3, S.105-114, Aug 1988.

[15] R. Kakoschke, E. Bußmann, H. Föll, *Modeling of Wafer Heating during Rapid Thermal Processing*, Appl. Phys. A, vol.50, no.2, S.141-150, Feb 1990.

[16] J-M. Dilhac, N. Nothier, *Thermal Model for Rapid Thermal Processors: Theory and Applications*, RTP '93

[17] R.S. Gyurcsik, T.J. Riley, F.Y. Sorrell, *A Model for Rapid Thermal Processing: Achieving Uniformity through Lamp Control*, IEEE Trans. Semicon. Manuf., vol.4, no.1, S.9-13, Feb 1991.

[18] S.A. Norman, *Optimization of Transient Temperature Uniformity in RTP Systems*, IEEE Trans. Electron. Dev. , vol.39, no.1, S.205-207, Jan 1992.

[19] C. Prakash, *Thermal Conductivity Variation of Silicon with Temperature*, Microel. Reliab., vol.18, S.333, 1978.

[20] E. Truckenbrodt, *Fluidmechanik, Bd. 1*, 2. Aufl., Springer, 1980.

[21] E. Truckenbrodt, *Fluidmechanik, Bd. 2*, 2. Aufl., Springer, 1980.

[22] S.W. Benson, *Thermochemical Kinetics*, J. Wiley, 1976

[23] R. Kessler, *Oszillatorische Konvektion.* DFVLR-Forschungsbericht 84-14, Göttingen 1984

[24] G.A. Sod, *Numerical Methods in Fluid Dynamics*, Cambridge Univ. Press, 1985

[25] R.A. Svehla, *Estimated Viscosities And Thermal Conductivities of Gases at High Temperatures*, Lewis Research Center Report R-132, Cleveland

[26] A. Kersch, T. Schafbauer, *Simulation der RTO-Kammer von AST des Cluster Tools im Rahmen des CIC-DIP*, Bericht 1996

[27] Messungen zur Verfügung gestellt von R. Strohmaier, AST elektronik, Dornstadt

[28] S.M. Sze, *Physics of Semiconductor Devices, 2nd. Ed.*, J. Wiley 1981

[29] D.B. Spalding, *The* PHOENICS *User's Guide*, Wimbledon 1991

[30] *SEMATECH Technical Report: Benchmarking of Commercial CFD-Software*, USA 1993/94

[31] R.B. Bird, W.E. Stewart, E.N. Lightfoot. *Transport Phenomena*. Wiley, NY 1960

[32] R.C. Ried, J.M. Prausnitz, T.K. Sherwood, *The Properties of Gases and Liquids*, McGraw Hll, 1977.

[33] C.R. Kleijn, C. Werner. *Modeling of Chemical Vapour Deposition of Tungsten Films*. Birkhäuser, Germany, 1993

[34] A. Kersch, W.J. Morokoff, *Transport Simulation in Microelectronics*, Birkhäuser 1995.

[35] A. Kersch, *Benefits and Limitations of Radiatively Heated Suszeptors*, MRS Symposium Proceedings Vol. 470, p. 159-173, 1997

[36] R.J. Kee, F.M. Rupley, J.A.Miller. *The Chemkin Thermodynamic Data Base*, Sandia Report, NM 1990.

[37] Referenz Photonbox

[38] J.O. Hirschfelder, C.F. Curtiss, R.B. Bird, *The Molecular Theory of Gases and Liquids*, Wiley, NY 1954

[39] V.E. Borisenko, P.J. Hesketh, *Rapid Thermal Processing of Semiconductors*, Plenum Press, London 1997

[40] E.D. Palik, *Handbook of Optical Constants of Solids*, Academic Press, New York, 1985

[41] Prof.R. Fritz, Prof. Schroer, FH Ulm, Persönliche Kommunikation

[42] F.L. Degertekin et. al., *In Situ Ultrasonic Wafer Temperature Sensor for RTP*, Proc. RTP' 95

[43] P. Dankoski et. al., *Toward RTP Control Using Ultrasonic Sensors*, Proc. RTP' 95

[44] Z. Wang et. al., *A New Fiber Optic System for Wafer Temperature Measurement in a Multi-zone and Multiphase RTP Furnace*, Proc. RTP' 95

[45] B. Lojek, *Difference in Measurement of Temperature at the Wafer Level and Die Level*, Proc. RTP' 96

[46] D.S. Boning, P.K. Mozumder, *DOE/Opt: A system for design of experiments, response surface modeling, and optimization using process and device simulation*, Texas Instruments Report 1993

[47] M. Beichele, *Elektrische Charakterisierung ultradünner nitridierter Oxide aufgewachsen mittels RTP auf in-line konditioniertem Silicium*, Diplomarbeit Fakultät Physik, Erlangen 1997

[48] J.C. Davis et al, *Improved Within-Wafer Uniformity Modeling through the Use of Maximum Likelihood Estimation of the Mean and Covariance Surfaces*, Proceedings ECS'95, Reno, Nevada

[49] K. Knutson, T.L. Cooper, *Response Surface Technique for Ex-Situ Process Uniformity Optimization in a Multi-Zone RTP-System*, Proc. MRS 1996

[50] A. Tillmann, *Model Based Temperature Uniformity Control during Rapid Thermal Processing*, Proceedings RTP'96

[51] B.A. Ogunnaike, *Process Dynamics, Modeling, and Control.* Oxford Press 1994

[52] J.D. Stuber, T.F. Edgar, T. Breedijk, *Model Based Control of Rapid Thermal Processes*, Proc. ECS Vol. 95-4, 1995

[53] P.J. Timans, *Temperature Measurement Strategies for Rapid Thermal Processing in Semiconductor Manufacturing*, Proc. RTP 96, 1996

[54] A. Kersch, T. Schafbauer, *Simulation des 'First Wafer Effektes' beim Titansilizidprozeß im MATTSON RTP*, Siemens Bericht, 1997

[55] A. Kersch, T. Schafbauer, H.J. Timme, A. Ajmera, *Equipment Simulation for Open-loop Rapid Thermal Processing*, Proc. RTP 95, Amsterdam 1995

[56] A. Kersch, T. Schafbauer, L. Deutschmann, Strategies for the reduction of pattern effects, Proc. MRS, Spring 1996

[57] T. Schafbauer, A. Kersch, *Patent Nr. DE 19514083 C2*

[58] A. Kersch, T. Schafbauer, *Patent Nr. DE 19711702 C1*

[59] A. Kersch, T. Schafbauer, *Patentameldung DE 19747164.1*

[60] T. Schafbauer, A. Kersch, *Temperature control in RTP using reduction of equipment models*, Proc. ECS 95, Reno 1995

[61] T. Schafbauer, *Simulation und modellbasierte Regelung von RTP-Reaktoren*. Diplomarbeit TU-München 1995

[62] A. Kersch, T. Schafbauer, *3D Simulation and Optimization of an RTO Chamber with Monte Carlo Heat Transfer in Comparison with Experiments*, Proc. RTP 96

[63] T. Schafbauer, A. Kersch, *Optimization of wafer temperature uniformity with application to a RTO-chamber*. Proc. RTP 97, New Orleans 1997

[64] A. Kersch, Th. Schafbauer, *Development of Model Based Control with Simulation*, in: Semiconductor Equipment and Materials beyond JESSI - 300 mm and Single Wafer Processing, Productronica 95, München

[65] Kürner, SIMEC Dresden, persönliche Kommunikation

[66] Z. Nenyei, A. Gschwandtner, S. Marcus, How to manage the pattern challenge?. Proc. RTP '95

[67] J.P. Hebb, K.F. Jensen, *Length scales and pattern effects in RTP heat transfer*. Proc. RTP '95

[68] R. Ditchfield, E.G. Seebauer, *General Kinetic Rules For Rapid Thermal Processing*. Mat.Res. Soc. Proc. Vol. 429, 1996

[69] R. Ditchfield, E.G. Seebauer, *Problems with the Concept of Thermal Budget: Experimental Demonstrations*. Mat.Res. Soc. Proc. Vol. 470, 1997

[70] P.A. Stolk et al., *Physical mechanisms of transient enhanced dopant diffusion in ion-implanted silicon*. J.Appl.Phys. 81(9), 1997

[71] I. Bork, A. v.Schwerin, *The importance of pairing reactions for the modeling of defect-dopant interactions in silicon*. Proc. MRS., Spring 1998

[72] M. Uematsu, *Simulation of boron, phosphorous, and arsenic diffusion in silicon based on an integrated diffusion model, and the anomalous phosphorous diffusion mechanism.* J. Appl. Phys.82(5), Sep. 1997

[73] TMA Associates, *TSUPREM-4 User's Manual.* Sunnyvale, USA 1997

[74] P.M. Fahey, P.B. Griffin, J.D. Plummer, *Point defects and dopant diffusion in silicon.* Rev. Mod. Physics, Vol. 61/2, 1989

[75] J. Zhu, *Ab initio presudopotential calculations of dopant diffusion in Si.* Proc. MRS 1997, Spring 1997

[76] S.M. Hu, *On Interstitial and Vacancy Concentrations in Presence of Injections.* J. Appl. Phys., Vol. 57, 1985

[77] Simulationen für ein 300mm RTP-System von A. Kersch, persönliche Kommunikation, 1998

[78] R.P. Brinkmann, *Effictient two-dimensional simulation of electronegative RF-discharges.* Proc. of the 44th AVS 97, San Jose

[79] C. Werner, R.P. Brinkmann, *Patent Nr. DE 44 43 608 C1.* 1994

[80] M.A. Liebermann, A.J. Lichtenberg. *Priciples of plasma discharges and materials processing.* J.Wiley 1994

[81] O.D. Patterson, P.P. Khargonekar, *Reduction of loading effect in reactive ion erching using real-time closed-loop control*, J. Electrochem. Soc., Vol. 144, Aug. 1997

[82] H.Maynard, *Plasma etching endpointing by monitoring RF-power systems with an artificial neural network.* Proc. ECS 95, Reno 1995

[83] A. Theodoropoulou et al., *Model Reduction for Optimization of Rapid Thermal Chemical Vapor Deposition Systems.* IEEE Trans. Semi. Manuf., Feb 1998

[84] J.M. Delarios, C.R. Helms et al., *Parallel Oxidation Model for Si Including Both Molecular and Atomic Oxygen Mechanisms.* Appl.Surf.Sci.39, 1989

[85] S.F. Devyatova, V.G. Erkov, E.L. Molodtsova, *Growth Kinetics of thermal silicon dioxide at low oxygen pressure.* Russ. Microel.26(3), 1997

[86] C.J. Han, C.R. Helms, *Parallel Oxidation Mechanism for Si Oxidation in Dry O_2.* J.Electrochem.Soc., 1987, S.1297ff.

[87] C.P. Ho, J.D. Plummer, J.Electrochem.Soc.126 p. 1523, 1979

[88] S.M. Hu, *Thermal oxidation of silicon: Chemisorption and linear rate constant.* J.Appl.Physics.55, 1984, S.4095

[89] H.Z. Massoud, J.D. Plummer, E.A. Irene, *Thermal Oxidation of Silicon in Dry Oxygen: Growth-Rate Enhancement in the Thin Regime.* J. Electrochem. Soc., 11/1985, S.2693ff

[90] E.A. Irene, *Silicon Oxidation Studies: Some Apstects of the Initial Oxidation Regime.* J.Electrochem.Soc.125 p. 1708, 1978

[91] A. Kazor, *Space-charge oxidant diffusion model for rapid thermal oxidation of silicon.* J.Appl.Phys.77 (4), 1995, S.1477

[92] J. Kuehne, S. Hattangady, *Kinetics, Chemical Composition and Reoxidation Kinetics of Rapid Thermal N_2O Oxynitride Growth.* Proceedings RTP 96

[93] V. Murali, S.P. Murarka, *Kinetics of ultrathin SiO_2 growth.* J.Appl.Phys.60, 1986

[94] D.R. Wolters, A.T.A. Zegers-van Duynhofen, *Kinetics of dry oxidation of silicon: Space-charge-limited growth.* J.Appl.Phys.65 (12), 1989

[95] D.R. Wolters, A.T.A. Zegers-van Duynhofen, *Kinetics of dry oxidation of silicon: Conditions affecting the growth.* J.Appl.Phys.65 (12), 1989

[96] D.R. Wolters, A.T.A. Zegers-van Duynhofen, *Silicon oxidation and fixed oxide charge*, J. Electrochem. Soc. Vol. 139/1, 1992

[97] J.F. Verwey, E.A. Amerasekera, J. Bisschop, *The physics of SiO_2 Growth*, Rep. Prog. Phys. 53 (1990)

[98] J.P. Zöllner, V. Cimalla, J. Pezoldt, *RTP-temperature monitoring by means of oxidation*, J. Non.Cryst. Sol. 187 (1995)

[99] R.P. Brent, *Algorithms for Minimization without Derivatives*, Prentice Hall, 1973

[100] J. Stoer, R. Bulirsch, *Numerische Mathematik*, 2 Bände, Springer 1990

[101] W.H. Press et. al., *Numerical Recipes in C*, Cambridge 1992

[102] G.H. Golub, C.F. van Loan, *Matrix Computations*. J. Hopkins, Maryland 1996

[103] A. Grace, *MATLAB Optimization Toolbox*, MathWorks, Massachusetts 1994

[104] L. Ingber, B. Rosen, *Genetic algorithms and very fast simulated reannealing: a comparison*, Mat. and Comp. Modelling, 16(11) 1992

[105] P.J.M. v.Laarhoven, E.H.L. Aarts, *Simulated Annealing: Theory and applications*, D.Reidel, Dordrecht 1987

Danksagung

Für ihre Mithilfe für den Erfolg dieser Promotionsarbeit geht mein besonderer Dank an:

Dr. Alfred Kersch, Siemens AG, Neuperlach,
für die hervorragenden Ideen und Vorschläge zu dieser Arbeit insbesondere auf dem Gebiet der Simulation, die Grundlagen einer Vielzahl der Ergebnisse dieser Arbeit sind.

Prof. Dr. Ingolf Ruge und **Dr. Walter Stechele**
vom Lehrstuhl für Integrierte Schaltungen der TU München für zahlreiche Ratschläge und die universitäre Betreuung des interdisziplinären Promotionsthemas

Prof. Dr. Frederik Koch für die offene Bereitschaft, das Zweitgutachten für diese Arbeit zu übernehmen, sowie die fachlichen Diskussionen an seinem Lehrstuhl.

Prof. Krishna Saraswat, Prof. Jim McVittie, Len Booth, Stanford University, für die der Experimente während meines Aufenthalts am Center for Integrated Systems und ihren Beiträgen zur Regelung von RTP-Systemen

Hartwig Bierhenke, Dr. Andreas Spitzer und Dr. Christoph Werner, Siemens AG, für die inhaltliche und finanzielle Unterstützung der Doktorarbeit von Siemens-Seite

sowie allen Mitarbeitern und Doktoranden bei Siemens ZT ME 4, Neuperlach, für das angenehme Arbeitsklima

Rainer Strohmaier, Dr. Frederique Glowacki, Dr. Barbara Fröschle, Thomas Knarr und Nicole Sacher, STEAG-AST elektronik, sowie **Dr. Alexander Gschwandtner**, Siemens HL,
für ihre Hilfe bei der Durchführung der Experimente am Cluster-Tool.

meinen Eltern für die finanzielle und moralische Unterstützung während meines Phyisikstudiums und meiner Doktorarbeit

Die vorliegende Arbeit wurde am ..... an der Technischen Universität München an der Fakultät für Physik und am Lehrstuhl für Integrierte Schaltungen eingereicht.

Ich erkläre hiermit, daß ich die vorliegende Dissertation selbstständig verfaßt und noch nicht anderweitig zu Prüfungszwecken vorgelegt habe. Sämtliche Quellen und Hilfsmittel sind angegeben, wörtliche und sinngemäße Zitate sind als solche gekennzeichnet.

München, den 21.6.1999 .....

Thomas Schafbauer
Spitzwegstr. 52b
85521 Ottobrunn

Abbildungsverzeichnis