Oliver Freyd

Parallele Hardware-Realisation eines Neuronalen Netzes mit Hilfe einer analogen Halbleiterkopplungsstruktur

– 2001 –

Experimentelle Physik

Parallele Hardware-Realisation eines Neuronalen Netzes mit Hilfe einer analogen Halbleiterkopplungsstruktur

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften im Fachbereich Physik
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

vorgelegt von

Oliver Freyd
aus Aachen

– 2001 –

Dekan:	Prof. Dr. W. Lange
Erster Gutachter:	Priv.-Doz. Dr. M. Bode
Zweiter Gutachter:	Prof. Dr. H.-G. Purwins
Tag der mündlichen Prüfungen:	07.06.2002
Tag der Promotion:	18.07.2002

Zusammenfassung

In dieser Arbeit wird die parallele Implementation eines selbstorganisierenden, topologieerhaltenden Systems mit Hilfe analogelektronischer Neuronen und einer neuartigen Kopplungsstruktur untersucht. Die Eigenschaften der selbstorganisierenden Karte resultieren aus einem Lernverfahren, das kompetitive und kooperative Elemente verbindet. Die Implementation dieser beiden gegensätzlichen Elemente wird durch die vorgestellte Kopplungsstruktur mit einem minimalen Aufwand an Verbindungen erreicht. Als Kopplungsstruktur werden Thyristorstrukturen, also Halbleiterstrukturen mit pnpn-Schichtfolge, untersucht. Eine solche Struktur stellt ein bistabiles, nichtlineares Medium dar, das lokal gezündet werden kann und mit der Propagation einer Schaltfront reagiert. Die Fähigkeit zur unabhängigen lokalen Zündung erlaubt die Implementation der Kompetition, während die Frontausbreitung den Kooperationsprozeß, eine Nachbarschaftskopplung variabler Stärke und Reichweite, vermittelt. Es werden Kopplungsstrukturen unterschiedlicher Geometrien experimentell untersucht, schließlich wird mit einem Hardware-Prototyp die praktische Anwendbarkeit des vorgestellten Hardware-Konzepts demonstriert. Durch Simulationen werden die Konsequenzen einer fortschreitenden Miniaturisierung der Kopplungsstruktur untersucht, und Möglichkeiten für eine analogintegrierte Implementation des vorgestellten Hardware-Konzepts aufgezeigt.

Inhaltsverzeichnis

Einleitung

Schon seit Jahrhunderten fasziniert die Idee einer intelligenten, denkenden Maschine die Menschheit. Das Ziel einer Maschine, die der strengen mathematischen Logik folgt, und die somit mathematische Algorithmen ausführen kann, führte zur Computertechnologie, deren Entwicklung nicht nur Auswirkungen auf die Mathematik und die Physik hatte, sondern die gesamte menschliche Gesellschaft von Grund auf veränderte.

Jedoch erwiesen sich einige Probleme, die viele Lebewesen und insbesondere der Mensch mit Leichtigkeit bewältigen, für herkömmliche, also regelbasierte, Computerprogramme als sehr schwierig zu lösen. Es handelt sich um Probleme wie die Erkennung von Objekten, die Bewertung von Situationen, die Steuerung von Bewegungen, oder ganz allgemein Probleme bei denen sich keine festen Regeln aufstellen lassen, die eine Lösung mit Hilfe logischer Schlussfolgerung zulassen.

Seit etwa einem halben Jahrhundert werden *künstliche neuronale Netze* entwickelt, lernfähige Systeme, die einen grundsätzlich anderen Ansatz verfolgen. Ein neuronales Netz besteht, wie das natürliche Vorbild, aus einer mehr oder weniger großen Zahl von Neuronen, die eine Anzahl Eingänge mit einer Anzahl Ausgänge koppeln. Die Kopplungen zwischen den Neuronen, die Synapsen, bestimmen dabei die Abbildung der Eingänge auf die Ausgänge. Das erste Neuronenmodell – Perzeptron genannt – wurde von McCulloch und Pitts 1943 [McC43] vorgestellt. Es modelliert die Funktion eines einzelnen biologischen Neurons auf stark vereinfachte Art und Weise. Es besitzt eine Anzahl Eingänge, die bei biologischen Neuronen als Dentriten bezeichnet werden, und einen Ausgang, der dem Axon des biologischen Vorbilds entspricht. Die Ausgangsaktivität wird aus einer Linearkombination der Eingangsaktivitäten bestimmt, deren Wertebereich durch eine sigmoide *Aktivierungsfunktion* begrenzt wird. Diese ist im einfachsten Fall binärwertig, das Neuron kann also nur aktiv oder inaktiv sein. Die Aufgabe eines solchen Neurons ist die Trennung von Mustervektoren in zwei Klassen, die es jeweils durch Aktivität oder Inaktivität anzeigt. Die Fähigkeiten des einfachen Perzeptrons sind jedoch sehr beschränkt: Es ist nur in der Lage, linear separierbare Vektoren eines Musterraum zu trennen, wie Minsky und Papert 1969 feststellten [Min69]. Es definiert

gewissermaßen eine trennende Hyperebene durch diesen Raum, und feuert nur für Muster, die auf einer Seite der Ebene liegen. Schon einfache logische Probleme sind mit dem Perzeptron nicht lösbar, eine einfache Exklusiv-Oder-Verknüpfung übersteigt seine Fähigkeiten. Diese Feststellung blockierte die Entwicklung dieser Art von neuronalen Netzen für mehr als ein Jahrzehnt. Es war zwar bekannt, daß man viele Perzeptrons zu einem Netzwerk kombinieren kann, aber es fehlte eine Lernregel, mit der man die Gewichte der Neuronen so trainieren kann, daß das Netz je nach angelegtem Mustervektor die gewünschte Aktivität zeigt.

Ein solches mehrschichtiges Netz bezeichnet man als Multilayer-Perzeptron, oder Feed-Forward-Netzwerk. Es besteht aus mehreren Schichten von Perzeptrons, wobei Verbindungen nur zwischen den Neuronen einer Schicht und der nächsten bestehen. Jedes Neuron bildet das Skalarprodukt des Eingangsvektors mit seinem sog. *Gewichtsvektor*, welches dann von einer sigmoiden Aktivierungsfunktion auf einen Wertebereich von beispielsweise $[-1, 1]$ begrenzt wird. Das Netz bildet somit einen Eingabe-Vektorraum auf einen Ausgabe-Vektorraum ab.

Die Aufgabe des Netzes besteht darin, einen Eingangs-Mustervektor auf ein Ausgangsaktivitätsmuster abzubilden. Die Abbildung ist durch die Gewichte der Neuronen parametrisiert. Durch Lernen sollen die Gewichte der Neuronen so angepaßt werden, das des Netz möglichst gut die Abbildung einer Trainingsmustermenge auf vorgegebene Soll-Ausgangsaktivierungen approximiert. Es handelt sich also um ein *überwachtes Lernverfahren* (supervised learning).

Rumelhart et al. stellten 1986 mit dem Backpropagation-Algorithmus [Rum86, Wer88] ein solches Lernverfahren für das Multilayer-Perzeptron vor. Es optimiert die Gewichte der Neuronen in einem Trainingsprozeß so, daß der mittlere Fehler der Ausgangsaktivität, als die Abweichung des tatsächlichen vom Soll-Ausgangszustand minimiert wird. Die Voraussetzung des überwachten Lernens ist, daß für das Training eines solchen Netzes eine Mustermenge mit zugeordneten Ausgangsvektoren erforderlich ist. Die Mustervektoren müssen also gegebenenfalls von Hand klassifiziert werden, bevor das Netz diese lernen kann. Dadurch kann die Erzeugung großer Trainingsdatensätze sehr aufwendig werden. Wichtig für die praktische Anwendung ist die Fähigkeit zur Generalisierung, also die Eigenschaft, daß aus der trainierten Mustermenge auf die korrekte Abbildung noch „unbekannter“ Mustervektoren auf eine möglichst sinnvolle Ausgangsaktivität führen. Bei einer großen Zahl von Parametern ist das Netz zu einer feinen Approximation der trainierten Abbildung fähig, kann aber nur schlecht generalisieren, bei relativ wenigen Parametern wird nur eine grobe, aber glatte und gut interpolierende Approximation erreicht.

Im Gegensatz dazu steht das *unüberwachte Lernen*, bei dem das neuronale Netz selbständig eine Menge von Trainingsmustern ordnen soll, ohne daß eine gewünsch-

te Ausgabe vorgegeben wird. Der Sinn des Lernvorgangs ist hier die Erzeugung einer effizient reduzierten Repräsentation des Eingangsraumes. Diese kann zur Datenkompression verwendet werden, zur Clusteranalyse, um Häufungen von Mustern in einem Datensatz zu finden, oder zur Visualisierung von hochdimensionalen Zusammenhängen.

Ein lineares Verfahren dieser Art ist die *Principal Component Analysis* [Hay99], die mit Hilfe einer linearen Transformation eine effiziente Darstellung des Eingangsraumes sucht. Dazu wird die Richtung maximaler statistischer Varianz der Mustermenge gesucht, dann der Musterraum durch eine Koordinatentransformation so gedreht, daß eine Achse parallel zu der ermittelten Richtung liegt. Mit dem verbleibenden Unterraum wird dieser Vorgang wiederholt. Damit findet die PCA eine Koordinatentransformation, die die Varianz der Mustermenge in möglichst wenigen Komponenten konzentriert. Die einzelnen Komponenten sind nach der Transformation mit fallender Varianz geordnet, die ersten Komponenten haben die größte Varianz, sind also gewissermaßen die wichtigsten. Somit lassen sich die „hinteren Komponenten“ mit geringem Kompressionsfehler weglassen. Die PCA erlaubt jedoch nur dann eine sinnvolle Kompression, wenn es über den gesamten Musterraum gemittelt lineare Korrelationen zwischen den Komponenten des Musterraumes gibt, die Mustervektoren sich also gewissenmaßen auf einer Hyperebene des Raumes konzentrieren.

Ein nichtlineares, unüberwachtes Lernverfahren ist der Vektorquantisierer (VQ) [Lin80]. Seine Aufgabe ist die Zuordnung der Mustervektoren zu einem von n Prototypen, gewöhnlich demjenigen, der dem Muster am nächsten liegt. Jedem Prototypen wird damit ein Einzugsgebiet des Musterraumes zugeordnet. Diese Einzugsgebiete werden als *Voronoi-Zellen* bezeichnet. Mustervektoren werden quantisiert bzw. codiert, indem jeder Mustervektor durch den *Index* des zugehörigen Prototypen ersetzt wird. Bei der Decodierung wird dann der entsprechende Prototyp statt des Original-Mustervektors eingesetzt, wodurch sich ein Quantisierungsfehler entsprechend dem Abstand des Mustervektors vom zugeordneten Prototypen ergibt. Ein Vektorquantisierer ermöglicht also eine verlustbehaftete Datenkompression.

Man kann den Vektorquantisierer als ein neuronales Netz betrachten, wobei die Gewichte der Neuronen die Prototypen darstellen, und jedes Neuron den Abstand zwischen seinem Prototypen und dem Mustervektor bestimmt. Durch einen nicht näher bestimmten inhibierenden Konkurrenzprozeß wird nur der Ausgang eines Neurons aktiviert, derjenige des Neurons mit dem kleinsten Abstand von allen.

Auch für Vektorquantisierer existieren Lernregeln, mit deren Hilfe die Prototypen an eine Menge von Mustervektoren angepasst werden können. Die als Linde-Buzo-Gray-Algorithmus [Lin80] bekannte Lernregel minimiert dabei den mittleren Quantisierungsfehler des VQ. Ein Vektorquantisierer erlaubt eine besonders effek-

tive Datenkompression, wenn die Muster sich in einem hochdimensionalen Vektorraum auf einen niedrigdimensionalen Subraum konzentrieren, oder beispielsweise in Clustern oder auf andere Weise sehr ungleichmäßig angeordnet sind. Dies ist beispielsweise bei Sprachsignalen und in noch größerem Maße bei Videosignalen der Fall, die stets eine große Redundanz beinhalten. Oftmals ergibt sich eine solche Konzentration der Musterverteilung auch erst nach einer geeigneten Vorverarbeitung der Signale, etwa durch eine Fourier- oder Wavelettransformation.

Auf dem Vektorquantisierer baut die von Kohonen 1982 [Koh82] eingeführte selbstorganisierende Karte (Self-Organizing Map, SOM) auf. Sind die Neuronen beim Vektorquantisierer ungeordnet, so erhalten sie in der SOM eine Nachbarschaftsbeziehung. Sie sind in einem *Kortex* angeordnet, der in der Regel ein zweidimensionales Gitter ist. Wird beim VQ nur der Gewinner einer Musterpräsentation bei einem Lernschritt angepaßt, so wird beim SOM der Gewinner und dessen *Nachbarschaft* dem jeweiligen Mustervektor angenähert. Diese Nachbarschaftskopplung bewirkt, daß benachbarte Neuronen auch korrelierte Lernschritte durchführen. So erklärt sich die Tendenz der SOM, eine topologisch geordnete Struktur im Musterraum anzunehmen. Daher lassen sich selbstorganisierende Karten, oder Kohonen-Netze, wie sie nach ihrem Erfinder auch genannt wird, gut zur *Visualisierung* der Struktur hochdimensionaler Vektorräume anwenden.

Sowohl der VQ als auch die SOM sind prinzipiell hochparallele Algorithmen, die jedoch gewöhnlich auf seriellen Computern ausgeführt werden. Dadurch wird insbesondere der Vergleich des Mustervektors mit allen Prototypen zu einem zeitintensiven Schritt, der große Netze dieser Art ineffektiv macht. Eine parallele Ausführung auf spezieller Hardware ermöglicht daher enorme Geschwindigkeitssteigerungen. Die Kopplung der Neuronen ist dabei ein zentrales Problem neuronaler Hardware. Feed-Forward-Netze sind üblicherweise *vollständig* gekoppelt, d.h. jedes Neuron einer Schicht ist mit jedem Neuron der nächsten Schicht verbunden. Dies führt zu einer quadratisch mit der Anzahl der Neuronen wachsenden Zahl der Verbindungen. Daher existieren nur relativ kleine Hardware-Implementationen von Feed-Forward-Netzwerken. Das Kohonen-Netz erfordert im Gegensatz dazu eine globale inhibierende Wechselwirkung der Neuronen zur Gewinnersuche, sowie eine lokale Nachbarschaftskopplung zur Implementation der kooperativen Lernregel. Dies sollte eine verbindungs-effiziente Implementation des Kohonen-Algorithmus erlauben, so daß die Größe der selbstorganisierenden Karte prinzipiell unbeschränkt ist [Ruw98].

Hier bietet sich die Betrachtung physikalischer Systeme an, in denen vielfältige Strukturbildungsphänomene durch Wechselwirkung aktivierender und inhibierender Kopplungen entstehen. Solche nichtlinearen Medien werden durch Reaktions-Diffusions-Gleichungen beschrieben. Ihren Ursprung haben die RD-Systeme in che-

mischen Systemen, wo die Konzentrationen der einzelnen Reagentien die Komponenten des entsprechenden RD-Systems darstellen. Die Reaktionsdynamik wird durch eine Reaktionsfunktion beschrieben, und Diffusionsterme beschreiben die eventuell für jede Komponente unterschiedliche Diffusion. Entgegen der Annahme, Diffusion führe immer zur Glättung eventuell vorhandener Strukturen, können solche Systeme Strukturbildung in verschiedenen Formen zeigen. Turing [Tur52] beschrieb 1952 im Zusammenhang mit Morphogenese die Strukturbildung in zweikomponentigen Aktivator-Inhibitor-Systemen. Im einfachsten Fall entstehen durch geeignete Wahl der Diffusionskonstanten von Aktivator und Inhibitor räumlich periodische Strukturen.

Schon im Jahr 1906 stellte Luther [Lut06] ein chemisches System vor, in dem eine Reaktionsfront durch einen äußeren Reiz getriggert werden kann, die sich daraufhin durch das Medium fortpflanzt. Diese Entdeckung führte zu der Vermutung, die Signalfortpfanzung in Nervenfasern könne auf ähnliche Weise ablaufen. Dies bestätigte sich mit dem Modell für die Signalpropagation auf Nervenfasern von Hodgkin und Huxley [Hod52]. Die Nervenmembran wird hier durch ein zweikomponentiges System modelliert, dessen Aktivator zwei stabile Zustände besitzt. Der Inhibitor baut einen dieser Zustände auf einer längeren Zeitskala wieder ab, so daß das System lokal anregbar wird: Durch einen Reiz wird es in den angeregten Zustand geschaltet, um dann nach Ablauf einer gewissen Zeit wieder in den Grundzustand zurückzufallen. Die Diffusion entlang der Nervenfaser führt zu einer Propagation dieser Aktivitätspulse. Die Breite und Form der Pulse wird dabei von den Konstanten des Systems bestimmt, nicht von der Art der Anregung.

Eine große Vielfalt an Strukturen, von periodischen Filamenten über Spiral- und Autowellen [Ast98] bis hin zu lokalisierten Strukturen, die zu Kollisionen, Erzeugungs- und Vernichtungsprozessen fähig sind, beobachtet man in Gasentladungssystemen [Ast01]. Modellrechnungen zeigen lokalisierte Strukturen mit vielfältigen Wechselwirkungen in dreikomponentigen RD-Systemen [Sch00, Bod01b].

Einkomponentige, bistabile RD-Systeme zeigen Strukturbildung auf eine besonders einfache Art und Weise [Fif79]: Eine Front verbindet zwei Domänen, in denen jeweils einer der beiden stabilen Zustände vorherrscht. Im allgemeinen propagiert diese Front, so daß eine Domäne auf Kosten des anderen wächst. In der Halbleiterelektronik sind solche Frontausbreitungsprozesse als Zündfronten bei Thyristoren [Ger79] von Bedeutung. Es handelt sich dabei um pnpn-Strukturen, die als effiziente Schalter für höchste Spannungen und Ströme Verwendung finden. Die wechselseitige, regenerative Injektion von Ladungsträgern aus zwei pn-Übergängen bewirkt ihre Bistabilität. Großflächig ausgedehnte pnpn-Strukturen lassen sich als RD-Systeme betrachten[Var70]. Die laterale Kopplung des Systems erfolgt dabei durch die Schichtleitfähigkeit der Basisschichten. Wird die Struktur lokal gezün-

det, also in den niederohmigen Zustand geschaltet, so breitet sich dieser Zustand als Front durch die gesamte Struktur hindurch aus. Der Zünd- und Frontausbreitungsprozeß im Thyristor kann nun auf geschickte Weise dazu genutzt werden, die Nachbarschaftskopplung beim Lernvorgang des Kohonen-Algorithmus zu implementieren. Dazu benötigt jedes Neuron nur eine Verbindung mit dem Thyristor-Medium, über die es sowohl die Zündung auslösen, als auch die entstehende Front detektieren kann.

Das Ziel dieser Arbeit ist, eine Implementation des Kohonen-Algorithmus mit analoger Hardware vorzustellen, die das Problem der Nachbarschaftskopplung mit Hilfe eines thyristorbasierten, aktiven Mediums löst. Das erste Kapitel stellt die Grundlagen von Vektorquantisierern und selbstorganisierenden Karten vor, und erläutert einige interessante Anwendungen dieser selbstorganisierenden Abbildungen.

Das zweite Kapitel stellt das Prinzip der Hardware-Implementation des Kohonen-Algorithmus vor, und beschreibt dann den Aufbau der im Verlauf dieser Arbeit konstruierten Demonstrations-Hardware [Bod01a]. Die Funktionsfähigkeit der Hardware wird anhand von Eichmessungen der verschiedenen Untereinheiten und anhand von Klassifikations- und Lernversuchen der kompletten neuronalen Hardware dargestellt. Dann wird auf Probleme und Verbesserungsmöglichkeiten eingegangen und Wege zur Konstruktion einer integrierten Version der Kohonen-Hardware aufgezeigt.

Im dritten Kapitel werden die pnpn-Strukturen vorgestellt, die im Verlauf dieser Arbeit präpariert wurden, mit dem Zweck, sie als aktives Medium zur Kopplung der Neuronen der SOM-Hardware einzusetzen. Die Strukturen besitzen eine Anzahl Gatekontakte, jeweils in unterschiedlichen Anordnungen, an denen Zündfronten sowohl gestartet als auch detektiert werden können. Optische und elektrische Messungen der Schaltfrontausbreitung auf diesen Proben werden vorgestellt. Eine Bestrahlung der Thyristorproben mit energiereichen Elektronen wird zur Verkleinerung der typischen Frontbreite untersucht.

Im vierten Kapitel werden Simulationen der Frontausbreitung in pnpn-Strukturen vorgestellt mit dem Ziel der Miniaturisierung des aktiven Mediums und dessen Integration mit den einzelnen Neuronen. Es werden, ausgehend von einer 0.75 mm dicken Struktur, deren Dotierungsprofil den experimentell untersuchten Strukturen entspricht, die Auswirkungen einer Schichtdickenverkleinerung der Vierschichtstruktur vorgestellt. Die Resultate zeigen Möglichkeiten auf, eine aktive Kopplungsstruktur auf pnpn-Basis in eine analog aufgebaute SOM-Hardware zu integrieren, was den Weg zu einem kompakteren Aufbau der Hardware mit einer wesentlich größeren Neuronenzahl öffnet [Nie01].

Kapitel 1

Vektorquantisierer und selbstorganisierende Karten

In diesem Kapitel werden der Vektorquantisierer und darauf aufbauend die selbstorganisierende Karte als unüberwachte Lernverfahren vorgestellt. Nach der Einführung der Algorithmen werden Zusammenhänge mit biologischen Systemen erläutert. Abschließend folgt eine Übersicht über bekannte Anwendungsgebiete von selbstorganisierenden Karten.

1.1 Lernregel des Vektorquantisierers

Ein Vektorquantisierer ordnet den Vektoren $\mathbf{X}$ eines Mustermenge jeweils einen von n Prototypen $\mathbf{W}_j$ zu, gewöhnlich denjenigen, der den Abstand $d = ||\mathbf{X} - \mathbf{W_j}(\mathbf{x})||$ minimiert. Der VQ kann zur komprimierten Übertragung von vektoriellen Daten genutzt werden, indem statt der Mustervektoren $\mathbf{X}$ jeweils der Index w des vom VQ ermittelten *Gewinnerprototypen* übertragen wird. Auf der Empfängerseite wird dann aus dem Index w der Prototyp $\mathbf{W}_w$ rekonstuiert. Somit ergibt sich ein Quantisierungsfehler, der gewöhnlich als quadratisches Abstandsmittel von Mustervektor und zugeordnetem Prototypen ausgedrückt wird.

Nun sollen die präsentierten Mustervektoren $\mathbf{X}_i$ eine bestimmte Verteilung im Musterraum besitzen, ausgedrückt durch die Wahrscheinlichkeitsdichte $p(\mathbf{X})$. Der mittlere Quantisierungsfehler D hängt dann von Lage der Prototypen im Musterraum ab:

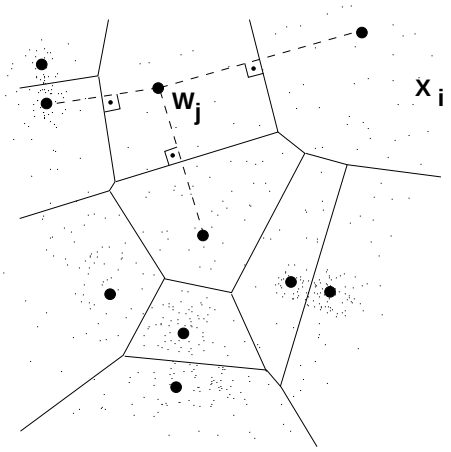


Abbildung 1.1: Vektorquantisierer mit Mustervektoren $\mathbf{X}_i$ und Prototypen $\mathbf{W}_j$. Die Einzugsgebiete der Prototypen sind die Voronoizellen.

$$D = \int_{-\infty}^{\infty} p(\mathbf{X}) \, ||\mathbf{X} - \mathbf{W}(\mathbf{X})||^2 \, d\mathbf{X} \tag{1.1}$$

Der *k-means* Algorithmus [Lin80] ist eine Lernregel, die eine optimale Anordnung einer festen Zahl von Prototypen findet, so daß der mittleren Quantisierungsfehler minimiert wird. Die kontinuierliche Dichteverteilung $p(\mathbf{x})$ wird durch wird durch eine endliche Trainings-Mustermenge $\mathbf{X}_i$ dargestellt. Der Algorithmus arbeitet folgendermaßen:

1. Initialisiere die Prototypen $\mathbf{W}_j$, z.B. durch zufällige Auswahl aus den Elementen der Mustermenge.

2. Ordne jedem Mustervektor $\mathbf{X}_i$ den Prototypen $\mathbf{W}_w$ zu, der den kleinsten Abstand zum Mustervektor $\mathbf{X}_i$ hat:

$$w = \arg \min_j(||\mathbf{X}_i - \mathbf{W}_j||) \tag{1.2}$$

3. Aktualisiere die Prototypen. Bestimme das Mittel (Centroid) der Mustervektoren, die im vorigen Schritt dem Prototypen $\mathbf{W}_j$ zugeordnet wurden, und weise es diesem Prototypen zu.

4. Iteriere die letzten beiden Schritte, bis die Prototypen stationär werden.

Charakteristisch für den Vektorquantisierer ist die Modellierung der Musterdichteverteilung durch die Prototypen: Gebiete mit hoher Musterdichte tragen einen

großen Anteil zum Gesamt-Quantisierungsfehler bei, so daß sie dichter mit Prototypen besetzt werden, während Gebiete geringerer Dichte einen höheren lokalen Quantisierungsfehler erlauben, so daß sich dort ein großer Abstand zwischen den Prototypen einstellt. Diese Eigenschaft erlaubt die effiziente Quantisierung ungleichmäßiger Musterverteilungen.

Der k-means-Algorithmus ist als *Batch*-Algorithmus formuliert, es wird also jeweils die gesamte Mustermenge durchiteriert, bevor die Prototypen verändert, bzw. neu berechnet werden. Dies ist von Nachteil, wenn die zu lernenden Mustervektoren nicht aus einer endlichen Menge stammen, sondern z.B. kontinuierlich neu erzeugt werden.

In diesem Falle wird die *Online*-Version des Vektorquantisierers angewendet:

1. Initialisiere die n Prototypen $\mathbf{W}_j$, z.B. durch zufällige Auswahl aus den Elementen der Mustermenge.
2. Iteriere die folgenden Schritte für jedes Element der Mustermenge $\mathbf{X}_i$.
3. Bestimme den Index w des Prototypen $\mathbf{W}_w$, der den kleinsten Abstand zum Mustervektor $\mathbf{X}_i$ hat (den *Gewinnerprototypen*).
4. Verändere den Gewinner nach folgender Regel, so daß er sich an den präsentierten Mustervektor annähert:

$$\mathbf{W}_w^{t+1} = \mathbf{W}_w^t + \eta(t)(\mathbf{X}^t - \mathbf{W}_w^t) \tag{1.3}$$

Auch hier werden in einem stochastischen Prozeß die Prototypen an die Mustermenge angepaßt. Der Lernvorgang kann fortgesetzt werden, während das Quantisierungsergebnis, die Gewinnerindizes w, bereits genutzt werden können.

Beide Versionen des Vektorquantisierers können im Verlauf der Optimierung in einem lokalen Minimum hängenbleiben, obwohl dies im Falle des online-Lernens durch die stochastischen Veränderungen der Prototypen während des Lernvorgangs durch die zufälligen Auswahl der präsentierten Mustervektoren weniger wahrscheinlich als bei der *batch*-Version ist.

Existieren beispielsweise mehrere getrennte *Cluster* hoher Musterdichte im Musterraum, so können die Prototypen praktisch nicht zwischen diesen Clustern umverteilt werden, da immer nur der Gewinnerprototyp am Lernvorgang beteiligt ist. Die Optimierung geschieht gewissermaßen nur lokal. Ein ähnliches Problem sind *tote* Prototypen. Ist ein Prototyp in einer ungünstigen Position, so daß er für keinen Mustervektor Gewinner wird, so wird seine Position nie verändert, so daß

dieser Prototyp für die Quantisierung verloren ist. Der einfache Vektorquantisierer ist also von einer sinnvollen Initialisierung der Prototypen abhängig, um eine gute Quantisierung zu erreichen. Erweiterungen der einfachen Lernregel umgehen diese Beschränkung z.B. durch dynamisches Einfügen und Entfernen von Prototypen, etwa indem tote Prototypen entfernt und in schlecht aufgelösten Gebieten des Musterraums neue Prototypen eingefügt werden.

Eine andere dem VQ mangelnde Eigenschaft ist eine Ordnung unter den Prototypen. Die Indizierung der Prototypen des Vektorquantisierers ist willkürlich und hat keine Beziehung zu deren Anordnung im Musterraum. Um einen Mustervektor zu quantisieren, müssen daher in jedem Fall alle Prototypen durchsucht werden, um den günstigsten zu finden. Dies ist umso zeitaufwendiger, je größer das Codebuch, also die Menge der Prototypen ist.

1.2 Lernregel der selbstorganisierenden Karte

Die selbstorganisierende Karte[1] ist eine Erweiterung des Vektorquantisierers und erweitert diesen um einige sinnvolle Eigenschaften. Die Prototypen der SOM besitzen eine Anordnung in einem *Kortex*, und die SOM-Lernregel dient dazu, die topologische Ordnung des Musterraums auf den Kortex zu übertragen.

Die Prototypen bilden nach erfolgtem Training eine *Karte* des Musterraumes, die eventuell entsprechend der Musterdichte der trainierten Muster verzerrt ist: Gebiete mit hoher Musterdichte werden hoch aufgelöst, während Gebiete mit geringer Musterdichte nur niedrig aufgelöst werden. Diese Karte kann zur Klassifizierung von Mustervektoren dienen, aber auch zur Visualisierung komplizierter Zusammenhänge in hochdimensionalen Vektorräumen. Die Ordnung der Neuronen erlaubt auch eine Interpolation zwischen benachbarten Mustervektoren.

Solche topologieerhaltenden Karten werden auch bei vielen biologischen Vorbildern gefunden, wie z.B. den *somatosensorischen Karten* im menschlichen Kortex, dem auditiven Kortex der Fledermäuse, dem visuellen Kortex der Katze, und anderen.

Im folgenden nehmen wir den Musterraum als einen n-dimensionalen Vektorraum an, und betrachten eine SOM aus m Neuronen. Die Neuronen sind auf einem M-dimensionalen Gitter auf den Gitterplätzen $\mathbf{r_i} \in \mathbb{R}^M$ angeordnet. In diesem Gitter – dem *Kortex* – besitzen die Neuronen eine Nachbarschaftsbeziehung. Der Kortex kann beispielsweise die Topologie einer Kette ($M = 1$), eines quadratischen oder hexagonalen Gitters ($M = 2$) oder eine höherdimensionale Topologie haben. Jedes

[1] Self-Organizing Map, kurz SOM

der m Neuronen stellt in Form seiner Gewichte $\mathbf{W} \in \mathbb{R}^n$ einen *Prototypen* des Musterraums dar, nicht anders als der Vektorquantisierer.

Die Kohonen-Lernregel besteht wie die Lernregel des VQ aus 2 Teilen, der Klassifikation der Mustervektoren und dem Lernvorgang. Die Klassifikation ordnet einem Mustervektor einen Prototypen zu, dies geschieht nicht anders als beim Vektorquantisierer durch Suche des Prototypen, der den kleinsten Abstand vom Mustervektor besitzt. Der Lernvorgang aktualisiert die Prototypen in Wechselwirkung mit den präsentierten Mustervektoren, und sorgt für die Strukturierung der Karte. Im Unterschied zum VQ wird hier nicht nur der Gewinnerprototyp aktualisiert, sondern auch dessen Nachbarn im Kortex.

1.2.1 Klassifikation

Das Kohonen-Netz ordnet einem Mustervektor $\mathbf{X}$ den Prototypen mit dem kleinstmöglichen Abstand d zu.

$$d_i = ||\mathbf{X} - \mathbf{W}_i||_p \tag{1.4}$$

ZUr Abstandsbestimmung wird gewöhnlich die euklidische Norm ($p = 2$) verwendet, mit gewissen Einschränkungen kann aber auch die Manhattan-Norm ($p = 1$) verwendet werden. Die Manhattan-Norm hat den Vorteil der einfacheren Berechnung, was insbesondere für eine Hardware-Realisation von Vorteil ist, jedoch den Nachteil, daß die dabei entstehenden Voronoi-Zellen nicht mehr konvex sind.

Im Bild der Vektorquantisierer wird der Gewinnerprototyp einfach durch eine sequentielle Suche ermittelt. Sieht man das Kohonen-Netz als neuronales Netz an, so schaltet gewissermaßen das Gewinnerneuron seinen Ausgang auf 1, alle anderen Neuronen aber auf 0. Diese gegenseitige Inhibition der Neuronen kann man sich durch einen Konkurrenzprozeß vermittelt vorstellen, etwa in der Weise, daß die Summe der Ausgangsaktivitäten aller Neuronen auf 1 begrenzt bzw. normiert ist. Jedes Neuron erzeugt dann eine umso höhere Aktivität, je kleiner sein Abstand zum Mustervektor ist. Die Normierung unterdrückt dann alle Neuronen außer den Gewinner. Allerdings führt ein solches Vorgehen zu einer mehr oder weniger „weichen“ Konkurrenz der Neuronen, die Ausgänge sind also nicht exakt 0 oder 1, sondern können Zwischenwerte annehmen.

1.2.2 Lernvorgang

Der Lernvorgang der SOM läuft folgendermaßen ab: Dem Netz werden zufällig aus dem Musterraum, bzw. der Mustermenge ausgewählte Muster präsentiert. Für

jedes Muster $\mathbf{X}$ wird das Gewinnerneuron w ermittelt, dann wird dieses *und seine kortikale Nachbarschaft* dem präsentierten Muster $\mathbf{X}$ angenähert.

$$\mathbf{W}_j(t+1) = \mathbf{W}_j(t) + \eta(t) h_{wj}(t)[\mathbf{X}(t) - \mathbf{W}_j(t)] \tag{1.5}$$

Hierbei ist $\eta(t)$ die Lernrate, die üblicherweise monoton fallend gewählt wird, um eine schnelle Anfangskonvergenz zu erreichen, und die Bewegungen der Prototypen mit fortschreitendem Lernvorgang immer mehr zu verlangsamen. $h_{wj}(t)$ ist die *Nachbarschaftsfunktion*, die gewöhnlich als Funktion des kortikalen Abstands des betrachteten Neurons j vom Gewinnerneuron w definiert wird: $h(d_{wj}, t)$. Sie soll monoton nach aussen hin abnehmen, und bei einem bestimmten Lernradius r gleich null werden. Somit werden nur Neuronen, deren kortikaler Abstand zum Gewinnerneuron kleiner als r ist, am Lernvorgang beteiligt.

Für $h(d)$ werden von Kohonen zwei verschiedene Möglichkeiten vorgeschlagen: die *Stufenfunktion*

$$h(d) = \begin{cases} 1 & : \quad 0 > d > r \\ 0 & : \quad d > r \end{cases} \tag{1.6}$$

oder eine kontinuierliche Funktion z.B. die Gauss-Funktion,

$$h(d) = exp(-\frac{d^2}{2\sigma^2}) \tag{1.7}$$

wobei σ die Breite der Nachbarschaft darstellt. Bei der Software-Implementation kann $h(d)$ für große Abstände $d > d_{\max}$ abgeschnitten werden, um Rechenzeit für Lernschritte mit sehr kleiner Lernrate zu sparen.

Die Einführung der Nachbarschaftsfunktion, und somit das Mitlernen einer Umgebung des Gewinnerneurons unterscheidet das Kohonen-Netz vom Vektorquantisierer. Sie bewirkt, daß benachbarte Neuronen stark korrelierte Lernschritte ausführen, und somit mit hoher Wahrscheinlichkeit auch deren Prototypen im Musterraum benachbart sind. Dies führt zunächst zu einer lokalen topologischen Ordnung des Kortex, wobei die Reichweite der Ordnung durch die Breite der Nachbarschaftsfunktion bestimmt wird.

Vor Beginn des Lernvorgangs müssen die Werte der Prototypen initialisiert werden. Dies kann mit völlig ungeordneten, zufällig gewählten Werten geschehen, oder mit geordneten Prototypen, beispielsweise in einer regulären Gitteranordnung im Musterraum.

Im Falle der zufälligen Initialisierung ordnet sich das das Netz in der Anfangsphase des Lernvorgangs. Um eine *globale* Ordnung des Netzes zu erreichen, wird

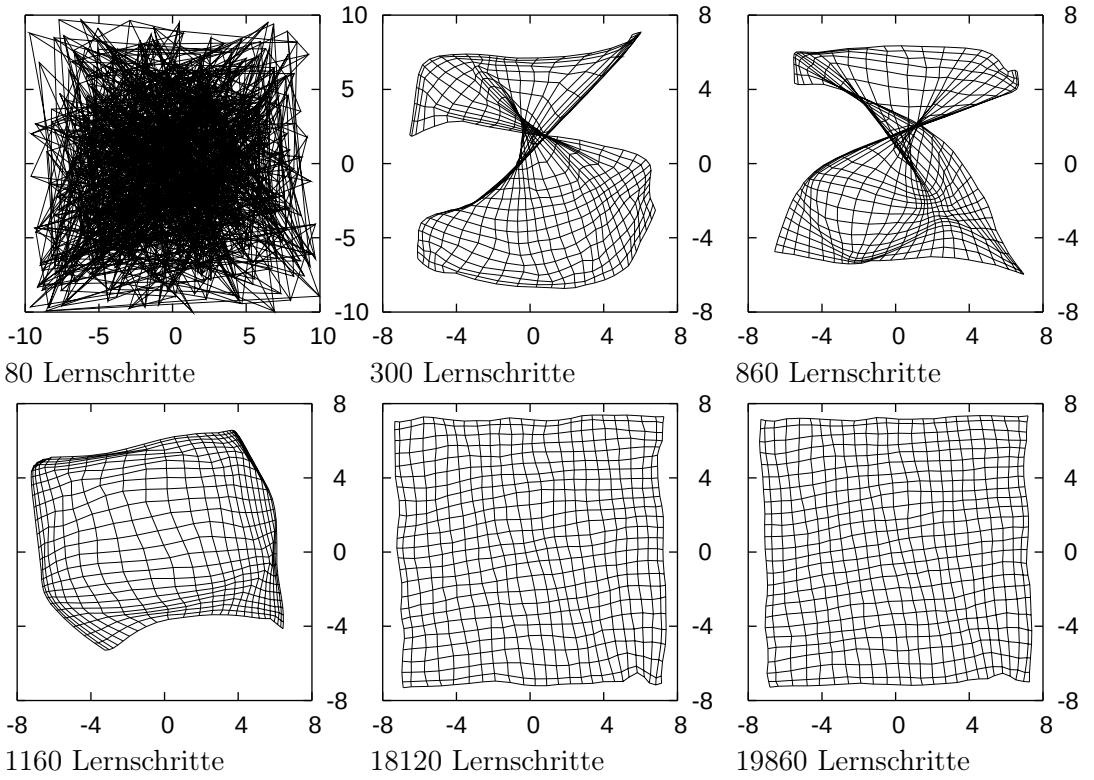


Abbildung 1.2: Lernvorgang einer SOM. Musterraum: $\mathbb{R}^2$, die Muster liegen innerhalb des Quadrates $-8 < x_1 < 8, -8 < x_2 < 8$. Größe der Karte: 25×25 Neuronen. Lernparameter: Gauss'sche Nachbarschaftsfunktion, für die ersten 1000 Lernschritte $\eta = 0.5$, $\sigma = 5$, $d_{\max} = 25$. Danach wird die Lernrate abgesenkt bis auf: $\eta = 0.05$, $\sigma = 1$, $d_{\max} = 3$.

der Lernvorgang mit einer großen Nachbarschaftsbreite begonnen. Bei zu geringer Nachbarschaftsbreite gelingt die Entfaltung des Netzes nicht vollständig und es s können topologische Defekte wie Knoten und Verwicklungen bestehen bleiben.

Nach dieser Ordnungsphase werden dann die Breite der Nachbarschaftsfunktion und die Lernrate gesenkt. Nach erfolgter Konvergenz bei nun geringer Lernrate führen die Prototypen nur noch statistische Schwankungen um eine Gleichgewichtslage aus. In Abbildung 1.2 ist dieser Lernprozeß anhand eines zweidimensionalen Musterraums und einem zweidimensionalen Kortex dargestellt. Die Mustervektoren werden gleichverteilt aus einem quadratischen Gebiet gezogen. Des Netz wird zufällig initialisiert, daher die anfängliche Unordnung, die dann in einen lokal geordneten Zustand mit topologischen Defekten übergeht. Bei genügend großem Radius der lernenden Nachbarschaft verschwinden die Verdrehungen, und die Karte paßt sich der Verteilung der Muster im Musterraum an.

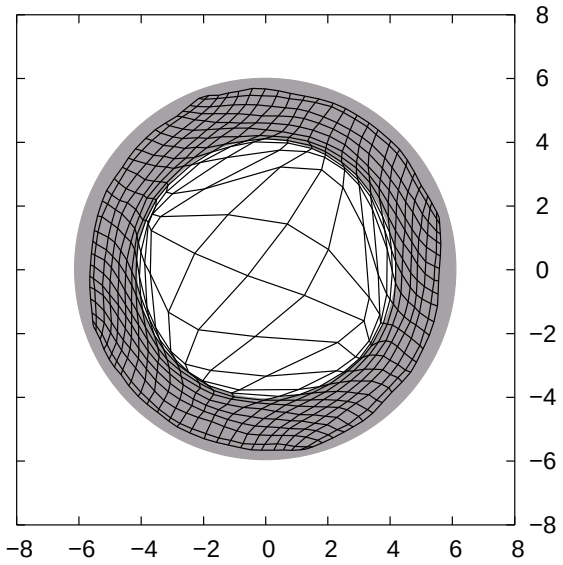


Abbildung 1.3: Quadratische SOM auf einer ringförmigen Musterverteilung. Die Muster liegen innerhalb des Rings mit: $4 < r < 6$. Größe der Karte 25×25 Neuronen. 15080 Lernschritte, Lernparameter wie in Abb. 1.2. Die Neuronen in der Mitte des Rings sind „tot“.

Durch die Eigenschaft der lokalen Topologieerhaltung kann die SOM die Dichteverteilung einer Mustermenge unter Umständen nicht so gut nachbilden wie ein Vektorquantisierer; das Nachbarschaftslernen kann Neuronen auf Positionen des Musterraumes zwingen, an denen sie ungünstig liegen und selten oder nie Gewinner werden, wie in Abbildung 1.3. Dort werden zweidimensionale Muster aus einem ringförmigen Gebiet trainiert. Die Karte paßt sich der Topologie lokal an, einige Prototypen werden aber in den Innenraum des Ringes gezogen, der von Mustervektoren frei ist. Die entstehende Karte ist gewissenmaßen ein guter Kompromiß zwischen der festgelegten Netztopologie der Karte und der Ringtopologie des Musterraumes.

Eine interessante selbstorganisierte Ordnung entsteht, wenn der Musterraum eine höhere Dimensionalität aufweist als der Kortex. In diesem Fall faltet sich das Netz, um den hochdimensionalen Raum möglichst vollständig auszufüllen. In Abbildung 1.4 ist dieser Fall am Beispiel eines zweidimensionalen Musterraums und einem eindimensionalen Kortex mit 400 Prototypen dargestellt. Wieder werden die Muster zufällig aus einem quadratischen Grundgebiet gezogen. Der Lernvorgang wird mit einer großen Breite der Nachbarschaftsfunktion begonnen, so daß die Prototypen sehr stark untereinander gekoppelt sind. Nach und nach wird die Breite der Nachbarschaftsfunktion verringert, die Kopplung der Neuronen nimmt ab. Daher bildet das Netz immer feinere Mäanderstrukturen aus, um den Musterraum besser auszufüllen. Entsprechend bildet eine zweidimensionale Karte in einem drei- oder mehrdimensionalen Raum Falten aus, um diesen optimal zu erfassen.

Abbildung 1.5 zeigt ein praktisches Beispiel, wie eine SOM zur Klassifikation und Dimensionsreduktion eines hochdimensionalen Vektorraumes dienen kann. Der

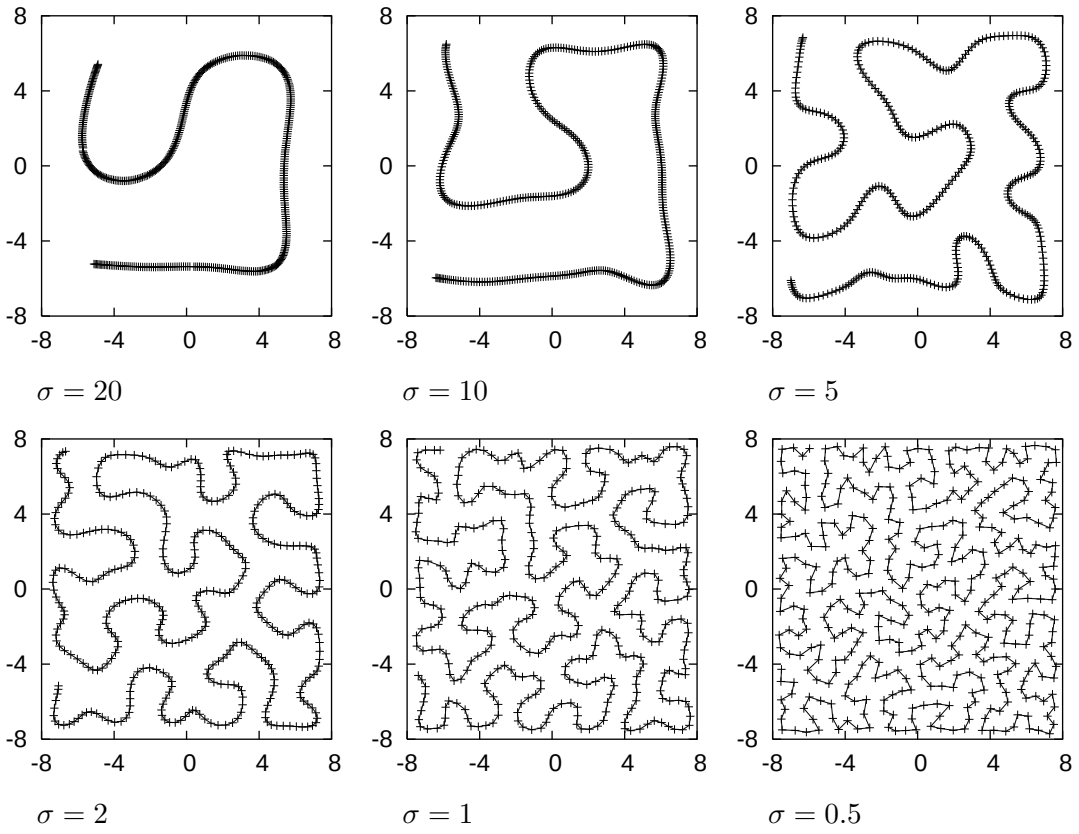
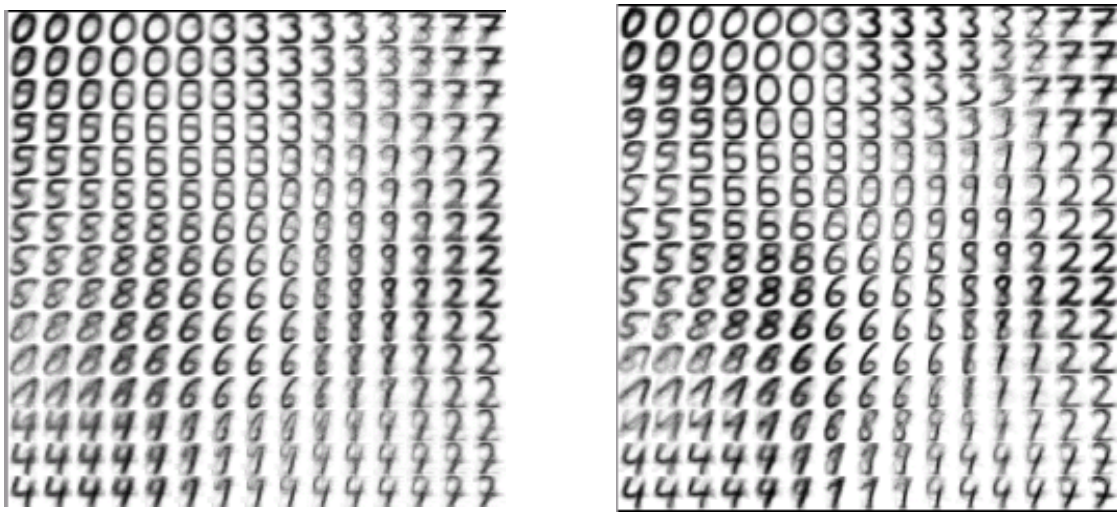


Abbildung 1.4: Dimensionsreduktion mittels SOM. Musterraum: $\mathbb{R}^2$, die Muster liegen innerhalb des Quadrates $-8 < x_1 < 8, -8 < x_2 < 8$. Eindimensionales SOM mit 400 Neuronen. Lernparameter: Gauss'sche Nachbarschaftsfunktion, $\eta = 0.02$, $d_{\max} = 50$, $\sigma = 20 \cdots 0.5$. Die Nachbarschaftsbreite σ wurde schrittweise reduziert, und jeweils bis zur Einstellung eines „stationären" Zustandes gewartet. Das Netz füllt den zweidimensionalen Raum in mäanderförmiger Weise aus, wobei seine Struktur mit abnehmender Nachbarschaftsbreite nach Art eines Fraktals immer feiner wird.

Musterraum besteht hier aus Bildern einer Größe von 16×16 Pixeln, die als 256-dimensionaler Mustervektor einer SOM mit einem zweidimensionalen Kortex präsentiert werden. Die Trainingsmustermenge besteht aus 4000 handgeschriebenen Ziffern, die eingescannt und skaliert wurden, so daß sie eine Bitmap von 16×16 Pixeln ausfüllen. Die Mustervektoren füllen nur einen kleinen Unterraum des vollen 256-dimensionalen Bildraumes aus, so daß die Abbildung auf eine zweidimensionale Karte möglich erscheint. Das der Karte zugrundeliegende Distanzmaß ist eine einfache euklidische Distanz, die pixelweise bestimmt wird. Die SOM ordnet die Prototypen nach pixelweiser Ähnlichkeit bzw. Unähnlichkeit. Es entstehen *Cluster*



Nachbarschaftsbreite 2 Nachbarschaftsbreite 1

Abbildung 1.5: Visualisierung eines hochdimensionalen Musterraums. Trainingsmustermenge: 4000 handgeschriebene Ziffern, gescannt und skaliert auf Bitmaps von 16×16 Pixeln. Größe der SOM: 15×15 Neuronen.

von nahezu gleichen Bildern (Ziffern), und dazwischen Übergänge von einer Ziffer zur anderen, wobei diese so angeordnet sind, daß sich an den Übergängen möglichst wenigen Pixel eines Prototypen verändern werden. Die entstehende Merkmalskarte zeigt die Ähnlichkeit benachbarter Prototypen bezogen auf das verwendete Distanzmaß. Bemerkenswert ist dabei, daß eine so simple Metrik wie ein pixelweiser Vergleich von Bildern zu einer intuitiv einleuchtenden Anordnung der Prototypen führt. Allerdings trägt die Vorverarbeitung der Muster, in diesem Fall die Zentrierung und Skalierung der Ziffernbilder, zum Erfolg des pixelweisen Vergleiches bei. Würden der Karte um einen oder mehrere Pixel verschobene Bilder präsentiert, würden diese kaum den richtigen Prototypen zugeordnet werden. Erst die Einführung eines verschiebungsinvarianten Abstandsmaßes könnte für Abhilfe sorgen, und die Bewertung der „Abstände“ von Ziffern oder sonstigen Symbolen der menschlich-intuitiven Wertung ähnlicher machen.

1.3 Weiterentwicklungen und Abwandlungen der SOM

Seit der Entwicklung der SOM sind einige erweiterte, bzw. abgewandelte selbstorganisierende Netzwerke entstanden, die eine flexiblere Struktur als die feste vorgegebene Netztopologie des Kohonen-Netzes erlauben.

Der *Neural-Gas*-Algorithmus [Mar93] ist eher eine Variation des klassischen Vektorquantisierers als eine Variation der SOM. Wie beim VQ gibt es keine festgelegte Nachbarschaftsbeziehung zwischen den Neuronen, es gibt keinen Kortex. Der Lernvorgang läuft ähnlich wie beim VQ ab, es wird jeweils das *Gewinnerneuron* dem präsentierten Mustervektor angenähert, aber zusätzlich lernt auch das *zweit-*, *dritt*nächste Neuron usw. mit jeweils geringerer Lernrate. Die Gewinnersuche läuft also so ab, daß die Prototypen nach dem Abstand zum Mustervektor sortiert werden, und je nach dem erreichten Rang unterschiedlich stark am Lernvorgang beteiligt werden. Diese unterschiedlichen Lernraten werden wie bei Kohonens SOM durch eine Nachbarschaftsfunktion ausgedrückt, doch bezieht sie sich auf die Nachbarschaft im Musterraum. Auch hier nimmt die Kopplung der Nachbarn mit der Zeit ab, so daß die Beteiligung der Nachbarn am Lernvorgang eher der Vermeidung toter Neuronen und der Konvergenzverbesserung des Lernvorgangs dient als der Ausbildung einer topologieerhaltenden Abbildung.

Eine weitere Entwicklungsmöglichkeit aller bisher erwähnter unüberwachten Lernverfahren ist die Einführung dynamisch wachsender Strukturen. Dazu gehört das *Growing Neural Gas* [Fri95], das dem *Neural Gas* einen Algorithmus hinzufügt, der an möglichst „lohnenden“ Stellen Neuronen hinzufügt oder wegnimmt. Dazu speichert jedes Neuron ein Fehlermaß, zu dem es, sofern es zum Gewinnerneuron wird, den quadrierten Abstand zum Mustervektor hinzuaddiert. Nach einer bestimmten Anzahl präsentierter Muster wird ein zusätzliches Neuron in der Nähe des Neurons mit dem größten akkumulierten Fehler eingefügt. Andererseits wird ein Neuron, das zu lange gar nicht Gewinner wurde, als „totes“ Neuron angesehen und entfernt.

Auf ähnliche Weise kann auch Kohonens SOM zu einer wachsenden Struktur erweitert werden (*Growing Grid*). Man fügt dann von Zeit zu Zeit eine ganze Zeile oder Spalte von Neuronen ein, deren Werte durch Interpolation ihrer Nachbarn (im Kortex) initialisiert werden. Auch hier kann die jeweils günstigste Lage der einzufügenden Zeile bzw. Spalte durch das Akkumulieren eines Fehlermaßes bestimmt werden. Durch das allmähliche Wachsen des Netzes wird die Ordnungsphase, in der sich die Grobstruktur ausbildet, sehr effizient ausgeführt: Anstatt eine große Zahl Neuronen stark korrelierte Lernschritte ausführen zu lassen, lernt anfangs nur ein kleines Netz, welches durch Interpolation nach und nach erweitert wird.

Eine interessante Modifikation, die sowohl auf den VQ als auch auf die SOM anwendbar ist, ist DeSienos *conscience*-Mechanismus [DeS88]. Die Neuronen haben dabei ein „Gewissen“ in Form eines „Handicap“, das auf die Distanz zum Mustervektor aufgeschlagen wird, bevor der Gewinner bestimmt wird. Ein hohes Handicap läßt somit das Einzugsgebiet eines Neurons schrumpfen. Das Handicap wird erhöht, wenn ein Neuron oft gewinnt, und abgesenkt im umgekehrten Fall. Somit findet ein Ausgleich der Gewinnhäufigkeit der Neuronen statt. Der Ausdruck „Gewissen“ rührt von der Idee her, ein Neuron bekäme ein schlechtes Gewissen, wenn es zu oft gewinnt, und hielte sich in Zukunft entsprechend zurück. Im Grenzfall führt dieser Ansatz zu einer näherungsweise gleichen Gewinnhäufigkeit aller Neuronen, also zu Entropiemaximierung.

Während die dynamisch wachsenden kompetitiven Lernverfahren sich wegen ihrer flexiblen Struktur nur schlecht für eine gewissermaßen fest verdrahtete Hardware-Implementation eignen, sollten Neural-Gas-ähnliche Modifikationen und der Conscience-Mechanismus sich leicht zu einem parallelen Hardwarekonzept hinzufügen lassen.

1.4 Der Kortex als Motivation für selbstorganisierende Karten

Neurobiologische Untersuchen zeigen, daß sich das Gehirn in viele Bereiche unterschiedlicher Funktion einteilen läßt. Die komplexesten und stammesgeschichtlich jüngsten Zentren liegen auf der Hirnrinde, dem *Kortex*. Die grobe Einteilung dieser Bereiche wurde oftmals durch die Untersuchung von Funktionsausfällen bei Patienten mit räumlich lokalisierten Läsionen gewonnen. Viele dieser Funktionszentren haben eine topologische Ordnung. Ein gutes Beispiel für diese Eigenschaft sind die sensorischen und motorischen Rindenfelder des menschlichen Kortex. Die Körperoberfläche ist dort in mehr oder weniger zusammenhängender Weise auf den Kortex abgebildet, wie bereits 1937 von Penfield und Boldrey [Pen37] durch elektrische Stimulation der Kortexoberfläche gezeigt wurde. Diese Zuordnung, die in Abbildung 1.7 skizziert ist, ist im Kleinen recht zusammenhängend, im Großen aber auch zerstückelt. Beispielsweise sind Daumen, Hand, Arm, Rumpf, und Bein zusammenhängend in dieser Reihenfolge auf das sensorische Rindenfeld abgebildet. Das Gesicht ist einem größeren, getrennten Bereich zugeordnet. Aus diesem wiederum sind Gaumen, Kiefer und Zunge herausgetrennt. In den Rindenfeldern des Cortex gibt es also eine Topologieerhaltung ähnlich der von Kohonens SOM.

Die Karte ist keine flächentreue Abbildung der Körperoberfläche, sondern die Größenverhältnisse in der Karte entsprechen der „Bandbreite“ der Sinneswahrneh-

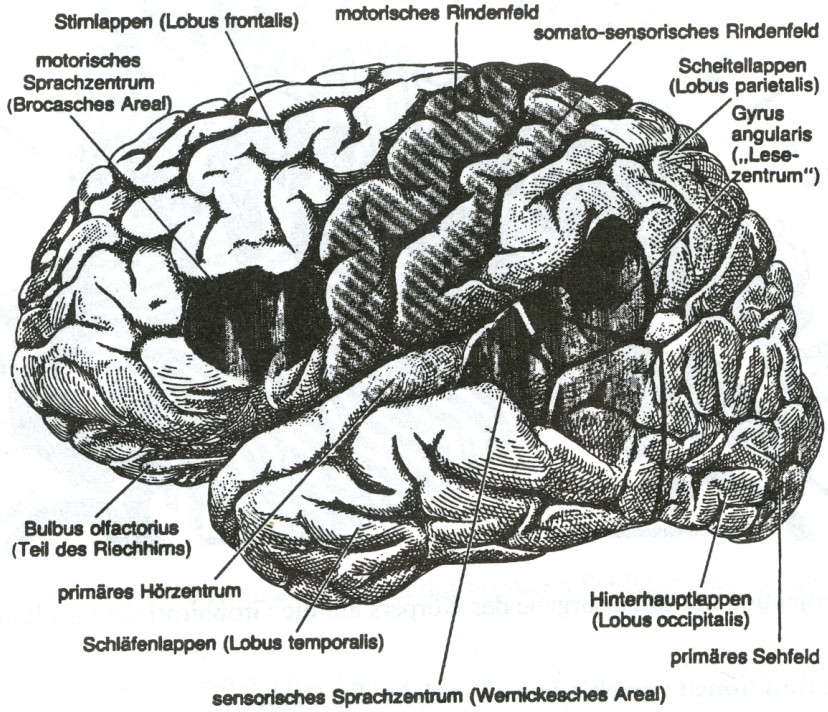


Abbildung 1.6: Motorische und sensorische Rindenfelder im menschlichen Cortex (aus [Ges92]).

mung, also der Dichte der Sinneszellen bzw. Nerverfaser, die aus einem Hautareal kommen, bzw. der Genauigkeit der motorischen Steuerung, die ein Körpergebiet benötigt. Beispielsweise sind die motorischen und sensorischen Felder für Gesicht und Hände besonders groß. Auch dies entspricht dem Verhalten des SOM, die Dichte der Prototypen an die Dichte der zum Lernen verwendeten Mustervektoren anzupassen.

Andere prominente Beispiele für kortikale Karten sind die *tonotopischen* Karten im auditorischen Kortex, wo die räumliche Ordnung der Neuronen der Tonhöhe bzw. Frequenz der wahrgenommenen Töne entspricht. Im auditorischen Kortex von Fledermäusen gibt es bemerkenswerterweise eine topologische Ordnung nach Empfangsfrequenz und Echolaufzeit [Sug79], was genau den Erfordernissen der Echoortung entspricht, Zielabstände über die Laufzeit und Relativgeschwindigkeit mittels des Dopplereffektes über die Frequenz zu bestimmen.

Man mag einwenden, in all diesen Fällen entstehe diese Ordnung nicht dynamisch, sondern könne schon genetisch vordefiniert sein. Einige Experimente zeigen jedoch, daß sich die Zuordnung einer sensorischen Karte verändern kann, wenn beispiels-

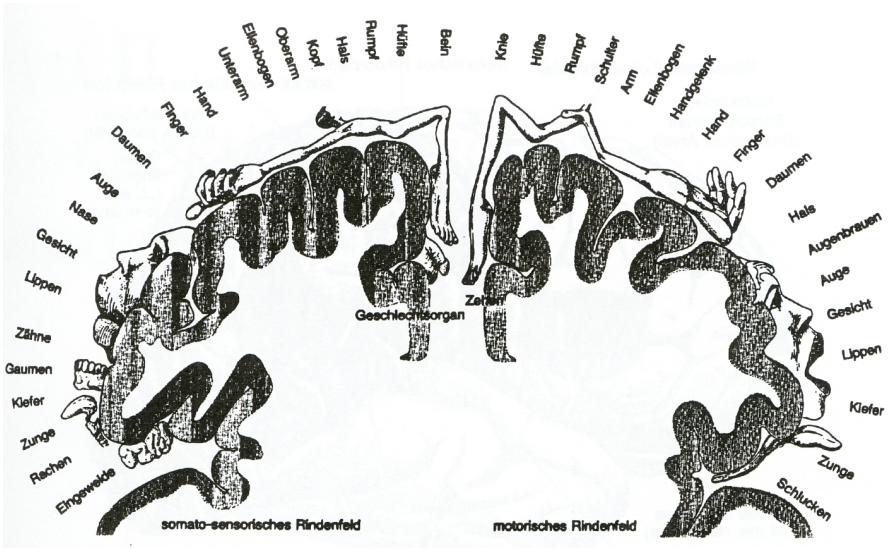


Abbildung 1.7: Querschnitt des sensorischen bzw. motorischen Cortex mit den zugeordneten Körperarealen (aus [Ges92]).

weise ein Gebiet der Körperoberfläche nicht mehr sensorisch gereizt wird. Das entsprechende Areal im Kortex wird nach und nach von den benachbarten Gebieten eingenommen.

Weiterhin ist bekannt, daß die kortikalen Karten von Menschen und Säugetieren zur korrekten Entwicklung ein entsprechendes Training während der Wachstumsphase braucht. Der visuelle Kortex der Katze ist in dieser Hinsicht besonders genau untersucht worden, er besitzt eine komplexe Ordnung hinsichtlich Augendominanz und Orientierungsempfindlichkeit (d.h. Sensibilität auf die Orientierung eines streifenförmigen Lichtreizes) der Neuronen [Hub62]. Die Entwicklung dieser Strukturen wird behindert, wenn eine Katze die ersten Lebenswochen im Dunkelheit verbringt[Hub70]. Ähnliche Effekte hat das Heranwachsen in einer horizontal- bzw. vertikal gestreiften Umgebung.

Die Struktur des Gehirns scheint sich also nur in Wechselwirkung mit den verarbeiteten Informationen zu organisieren. Das Modell von Kohonen weist eine ähnliche Fähigkeit zur Selbstorganisation auf. Die Winner-take-all-Lernregel erscheint jedoch biologisch unplausibel, so daß die SOM nur ein sehr vereinfachtes Modell für die Strukturierung kortikaler Karten sein kann.

1.5 Applikationen für selbstorganisierende Karten

Seit der Entwicklung der selbstorganisierenden Karte in den 80‘er Jahren wurde eine Vielzahl von sinnvollen Anwendungen gefunden.

Klassische Beispiele sind die schon von Kohonen [Koh89] selbst untersuchten Probleme der Sprach- bzw. Phonemerkennung [Bea93], z.B. zum Zweck der Sprachsteuerung. Ähnliche Problemstellungen sind Sprechererkennung [And94] und die Analyse von Sprachakzenten [Cha94] und Sprachfehlern [Cal99]. Bei allen praktischen Anwendungen ist die Vorverarbeitung der zu analysierenden Daten entscheidend. Auf welche Weise die Mustervektoren aus den Rohdaten gewonnen werden, hat einen großen Einfluß auf die sinnvolle Anwendung des Kohonen-Algorithmus.

Bei der Sprachanalyse wird gewöhlich das Sprachsignal zunächst in Blöcke von etwa 10-20 ms Länge zerlegt, aus diesen wird per Fourieranalyse ein Kurzzeitspektrum gewonnen. Dessen Frequenzskala wird oft auf eine dem menschlichen Hörempfinden entsprechende, nichtlineare Frequenzskala transformiert, und in relativ wenige Kanäle zusammengefaßt. Damit reichen etwa 15 Kanäle, die jeweils etwa eine drittel Oktave umfassen, zur Codierung des sprachrelevanten Spektrums aus. Diese Codierung enthält nur noch einen kleinen Teil der ursprünglich vorhandenen Information, das Sprachsignal kann also nur noch unvollständig rekonstruiert werden. Sie enthält jedoch eine zeitlich variable Grobstruktur des Sprachspektrums. Die Phoneme unterscheiden sich ja gerade durch die unterschiedlichen Resonanzen, die durch die unterschiedliche Stellung der Artikulatoren im Mund- und Rachenraum entstehen. Die Feinstruktur des Sprachspektrums kann eher der Anregung durch Stimmbänder bzw. turbulenten Strömungen zugerechnet werden. Die Intensitäten dieser 15 Kanäle können nun als Mustervektor für eine SOM dienen, die so die dem System präsentierten Spektren nach ihrer Ähnlichkeit anordnen kann. Am besten lassen sich so die Phoneme klassifizieren, die nur wenig zeitlich variabel sind, wie Vokale, während Plosive (z.B. /p/ und /t/) hauptsächlich durch rasche zeitliche Änderung ihres Spektrums bzw. ihrer Intensität charakterisiert sind und somit in einer solchen statischen Phonemkarte nicht erfasst werden. Solche Phonemkarten können zur effizienten Codierung von Sprache bei reduzierter Bitrate eingesetzt werden. Zum Zweck der Sprechererkennung lassen sich Phonemkarten verschiedener Sprecher ähnlich wie Fingerabdrücke vergleichen.

Eine weitere Anwendung ist die Erkennung von Ziffern und Schriftzeichen als Teil einer automatischen Texterkennung. Ebenso wurden SOM zur Analyse medizinischer Bilder und Daten verwendet, etwa zur Erkennung von abnormen Zellen oder zur Klassifikation von EEG-Daten [Elo92]. Die Erkennung von Massenspektren wird in [Bel97] untersucht. In eine ähnliche Richtung zielt die Auswertung der Daten von Multisensor-Systemen, die als künstliche Nase Verwendung finden

[Nat97]. Auf ähnliche Weise kann eine selbstorganisierende Karte zur Visualisierung von System- und Maschinenzuständen in der Industrie dienen, besonders zur raschen Erkennung von Fehlzuständen.

Auch in der Robotik gibt es Probleme, die mit Hilfe der SOM gelöst werden können, wie etwa die visuomotorische Koordination, also die Kontrolle der Bewegungen eines Roboterarmes durch ein visuelles System. Ganz allgemein kann die SOM dazu verwendet werden, die nichtlineare Abbildung zwischen einem sensorischen bzw. visuellen Koordinatensystem und einem motorischen Koordinatensystem zu lernen.

Eine hardwaremäßige Realisierung des SOM-Algorithmus könnte besonders für schnelle Echtzeitanwendungen lohnend sein, wie etwa Sprach- und Videokompression etwa im Mobilfunk. Im Bereich der Hochenergiephysik wird neuronale Hardware (meist Feedforward-Netze) bereits zur Vorklassifikation von Events eingesetzt, da dort riesige Datenmengen in Echtzeit zu bewältigen sind. Baldanza [Bal95] beschreibt den Einsatz den Einsatz einer Neurohardware, die auf einem kommerziellen analogen Feedforward-Chip [eta92] basiert, zur Vorsortierung von Detektordaten auf der Suche nach Elementarteilchen.

Diese Übersicht über Anwendungen der selbstorganisierenden Karten ist keinesfalls vollständig. Sie soll einen Einblick in mögliche Anwendungsgebiete geben und eine Motivation für die Entwicklung neuronaler Hardware geben.

Kapitel 2

Parallele Implementation der selbstorganisierenden Karte

In diesem Kapitel wird die im Rahmen dieser Dissertation entwickelte parallele Architektur der selbstorganisierenden Karte vorgestellt. Zunächst wird die Funktionsweise der analogelektronischen Bausteine beschrieben, die die einzelnen Neuronen der SOM bilden. Frontpropagation auf einem aktiven Medium, in diesem Fall einem Thyristor, dient zur Kopplung der Neuronen untereinander. Nach der Darlegung des Schaltungskonzepts wird als konkrete Anwendung eine Demonstrationshardware mit 5 analogen Neuronen in einem eindimensionalen Kortex vorgestellt. Anhand dieser Hardware werden Stärken und Schwächen der Architektur, Integrationsmöglichkeiten und alternative Ansätze diskutiert.

2.1 Prinzipieller Aufbau der SOM-Hardware

Die selbstorganisierende Karte ist – wie neuronale Netze im allgemeinen – ein hochparalleler Algorithmus. Dies ist die Motivation zur Entwicklung einer parallelen Implementation der SOM, die eine enorme Geschwindigkeitssteigung verspricht, sollte es gelingen, sowohl den Konkurrenzprozeß zur Gewinnersuche als auch die Kooperation des Lernvorgangs zu parallelisieren. Beide Aufgaben werden in der hier vorgestellten Implementation von einem bistabilen, aktiven Medium übernommen, indem laufende Fronten die Nachbarschaftskopplung der Neuronen bewirken.

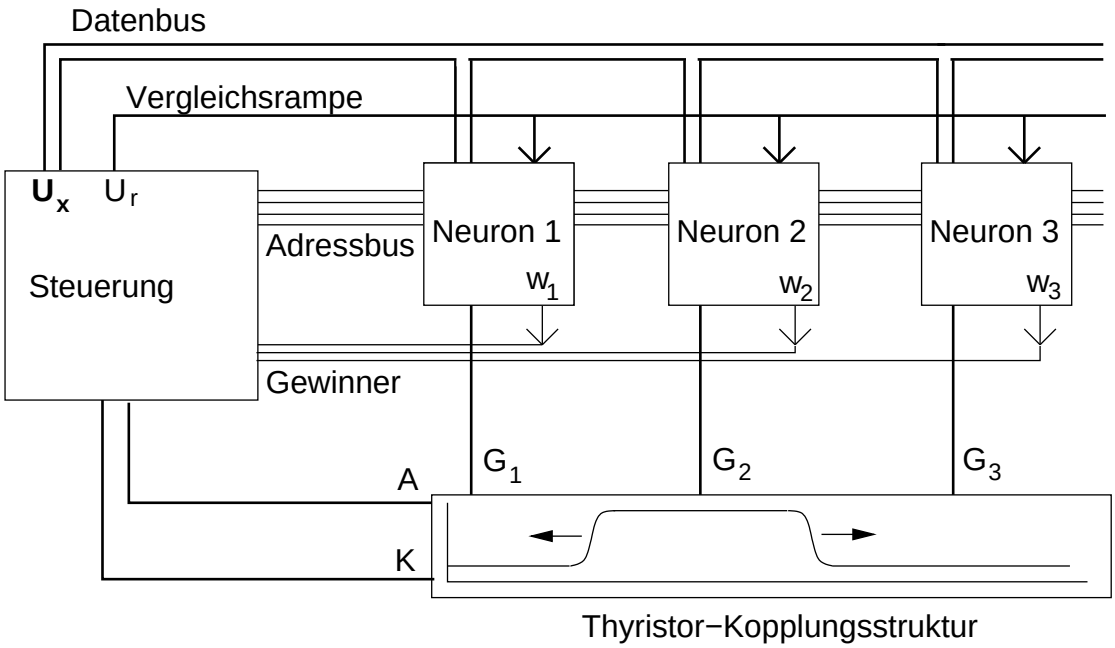


Abbildung 2.1: Modularer Aufbau der SOM-Hardware

Dem parallelen Aufbau der SOM entsprechend ist die Hardware modular aufgebaut (s. Abb. 2.1). Sie besteht aus einer zentralen Steuerungseinheit, für jedes Neuron eine Neuroneneinheite, und dem aktiven Medium, das die Nachbarschaftsbeziehung zwischen den Neuronen vermittelt.

Die Steuereinheit dient zur Aufbereitung der dem Netz zu präsentierenden Mustervektoren und zur zeitlichen Ablaufsteuerung von Klassifikation und Lernvorgang. Ein Bussystem versorgt alle Neuronen mit dem jeweils angelegten Mustervektor, und erlaubt das Schreiben und Lesen der Prototypen der einzelnen Neuronen über einen Adressbus.

Entscheidend ist die Frage, ob das System in Anolog- oder Digitaltechnik aufgebaut werden soll. Eine parallele digitale Implementation auf der Basis von vernetzten PIC-Microcontrollern wird in [Ruw98] beschieben. Eine analoge Implementation hat zwar den Nachteil einer geringeren Rechengenauigkeit, dafür lassen sich die Elemente des Kohonen-Algorithmus in Analogtechnik mit bestechend einfachen Mitteln realisieren. So wurde hier der Weg einer analogen Repräsentation der Mustervektoren und Prototypen eingeschlagen.

Die Steuereinheit kommuniziert mit dem Hostcomputer, von dem sie die zu ler-

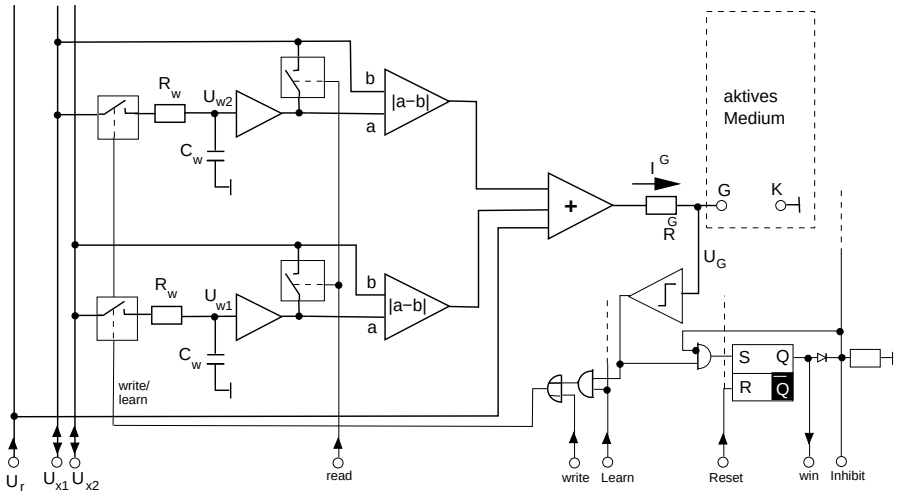


Abbildung 2.2: Aufbau eines einzelnen Neurons, mit den Speicherkondensatoren C_w, den Lernwiderständen R_w, der analogen Distanzrechenschaltung, und dem Zünddetektor. Großgeschriebene Signale sind allen Neuronen gemeinsam, was auch durch gestrichelt fortgesetzte Leitungen angedeutet wird.

nenden Mustervektoren erhält. Die n-dimensionalen Mustervektoren verlangen ein Bussystem aus n Leitungen,auf denen je eine Komponente X_i des Mustervektors durch eine Spannung U_x^i codiert wird. Mittels Analogschaltern kann die Steuereinheit einen Mustervektor auf den Bus legen, so daß alle Neuronen ihn simultan mit ihren Prototypen vergleichen können (Broadcasting).

Die Speicherung der Prototypen U_w^i in den einzelnen Neuronen basiert auf Sample-and-Hold-Gliedern, also Speicherkondensatoren, die durch Pufferverstärker möglichst von der Außenwelt entkoppelt sind (s. Schaltskizze der einzelnen Neuronen in Abb. 2.2). Auch hier ermöglichen Analogschalter, den Prototypen eines Neurons auf den am Bus anliegenden Vektor zu setzen, also den Prototypen zu schreiben. Ebenso dient der Bus zum Lesen des Prototyps. Zum Lesen und Schreiben müssen die Neuronen einzeln adressiert werden, was ein Adressierungssystem erfordert.

Weiter enthält jedes Neuron eine auf Operationsverstärkern basierende Rechenschaltung, die den Abstand des gespeicherten Prototypen vom am Bus anliegenden Mustervektor bestimmt. Der Musterabstand wird dabei nach der Manhattan-Norm $d = \sum_{0<i<n} |(x_i - w_i)|$ bestimmt, da diese nur die Addition von Beträgen erfordert. Die Betragsbildung läßt sich elektronisch durch Gleichrichtung realisieren, etwa mit Hilfe einer Diode. Liegt diese Diode in der Gegenkopplung eines Operationsverstärkers, so wird der Einfluß des Spannungsabfalls an der Diode kompensiert, und man erhält einen nahezu idealen Gleichrichter. So erzeugt jedes Neuron eine zu seinem Abstand d proportionale Spannung U_d.

Die Gewinnersuche wird nun parallel ausgeführt, indem jedes Neuron mit Hilfe eines Komparators seinen Abstandswert U_d^j mit einem globalen Vergleichswert U_r vergleicht, der linear mit der Zeit anwächst: $U_r(t) = m \cdot t$. Sobald der Komparator des ersten Neurons umschaltet, also für ein Neuron gilt: $U_r > U_d^j$, ist der Gewinner gefunden. Dazu ist ein Komparator je Neuron erforderlich. Der Vergleichswert d_g wird in Form einer Spannungsrampe U_r von einem Integrator in der Steuerungseinheit erzeugt. Ein Flip-Flop je Neuron dient zur Speicherung des Gewinnerstatus des Neurons bis zur späteren Auswertung. Eine allen Neuronen gemeinsame digitale Busleitung dient zur Inhibition dieses Flip-Flops. Das erste eingeschaltete Flip-Flop setzt diese Inhibitionsleitung (`inhibit` in der Schaltskizze), so daß nie mehr als ein Flip-Flop gesetzt sein kann.

Ein aktives Medium, in diesem Fall eine Thyristorstruktur, dient zur Kopplung der Neuronen, gewissermaßen zur Kommunikation. Es stellt den Kortex des Systems dar. Jedes Neuron j kann den Zustand des Mediums an seiner Position $\mathbf{x}_j$ überwachen (lesen) und beeinflussen (schreiben). Das Medium ist bistabil, kann sich also in zwei Zuständen (0 und 1) befinden. Im Ruhezustand ist es homogen im Zustand 0, und kann von jedem Neuron *gezündet* werden, also in den Zustand 1 versetzt werden. Dies geschieht zunächst lokal, dann breitet sich der Zustand 1 über das gesamte Medium in Form einer Front mit einer definierten Frontgeschwindigkeit v aus. Die Topologie des Mediums bestimmt die Netztopologie: Das Medium könnte etwa eine lineare, eindimensionale Topologie haben, oder eine zweidimensionale Topologie mit kreisförmigen Fronten. Die Frontlaufzeiten bestimmen die Nachbarschaftskopplung der Neuronen.

Der Zeitaufwand für die Gewinnersuche hängt beim Vergleich mit einer linearen Rampe nur von Anstiegsrate m des Vergleichswertes U_r ab. Diese Anstiegsrate bestimmt die kleinste Abstandsdifferenz, die das System noch erkennt, also die Vergleichsgenauigkeit. Dieser Zusammenhang ergibt sich folgendermaßen: Nach dem Schalten des ersten Komparators beim Abstand d_1 vergeht eine bestimmte Schaltzeit t_s, bis die Gewinnersuche beendet werden kann und das getriggerte Neuron als Gewinner ausgewiesen wird. In dieser Zeit steigt die Rampe noch um $\Delta d = mt_s$ an. Ist nun der „zweitbeste“ Abstand d_2 nicht mindestens um Δd größer, so schaltet in dieser Zeit auch der zweite Komparator, und es werden zwei Neuronen als Gewinner erkannt. So können zwei oder mehrere „fast“ gleich weit vom Mustervektor entfernte Neuronen ansprechen, was im Kohonen-Algorithmus nicht vorgesehen ist. Sofern das Netz schon eine geordnete Topologie besitzt, liegen die im Musterraum benachbarten Neuronen auch im Kortex benachbart, sodaß die Auswirkungen des nicht eindeutigen Gewinners auf den Lernvorgang gering sind. Lediglich die Klassifikation des Mustervektors ist dann nicht mehr eindeutig möglich.

Eine effizientere Möglichkeit des Parallelvergleiches wäre eine Intervallschachtelungsmethode nach Art der sukzessiven Approximation: Der mögliche Wertebereich der zu vergleichenden Abstände wird halbiert. Dieser erste Wert wird als globaler Vergleichswert an das System angelegt, alle Neuronen vergleichen ihn mit dem „eigenen“ Wert. Falls (mindestens) ein Komparator triggert, ist der Wert zu groß und die unteren Hälfte des Restintervalls wird erneut halbiert. Fall kein Komparator triggert, ist der Wert zu klein und man fährt mit der oberen Hälfe des Intervalls fort. Ein solches Verfahren benötigt nur eine logarithmisch von der Vergleichsgenauigkeit abhängige Zeit.

Die Steuerung des Lernverfahrens ist jedoch im Fall der linearen Vergleichsrampe einfacher, da so in jedem Fall nur ein Neuron zugleich einschaltet und somit direkt die Frontausbreitung im aktiven Medium zünden kann. Bei der Intervallschachtelung muß die Zündung der Front solange unterdrückt werden, bis der Gewinner eindeutig feststeht.

2.2 Lernvorgang

Nach Bestimmung des Gewinnerneurons lernt dieses und seine Nachbarschaft, indem sein Prototyp dem präsentierten Mustervektor angenähert wird, entsprechend der Lernregel Gl. 1.5 (s. Abb. 2.3).

Dieser Lernvorgang wird sehr einfach realisiert: Zum Lernen werden die Speicherkondensatoren C_w eines Neurons über einen Widerstand R_w mit dem von außen angelegten Muster U_x für eine kurze Zeit t verbunden. Während dieser Zeit findet eine exponentielle Annäherung (Gl. 2.1) des Prototypen U_w an U_x statt.

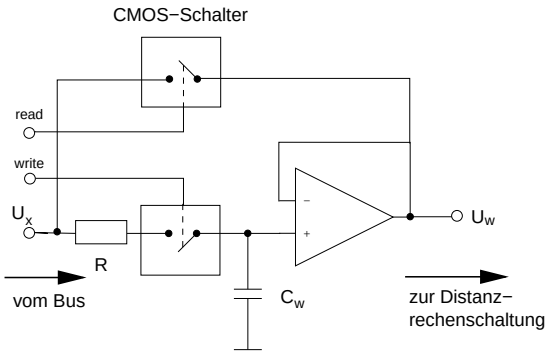


Abbildung 2.3: Analoge Speicherzelle mit Impedanzwandler am Ausgang

$$U_w(t) = U_x + (U_w(0) - U_x)\exp(-t/\tau) \quad \text{mit} \quad \tau = RC \tag{2.1}$$

Durch Vergleich mit Kohonens Lernregel ergibt sich eine mit steigender Lernzeit t wachsende Lernrate:

$$\eta\, h_{wj} = 1 - \exp(-t/\tau) \tag{2.2}$$

Läßt man nur das Gewinnerneuron für eine bestimmte Lernzeit t lernen, so erhält man eine Implementation eines Vektorquantisierers. In einer SOM müssen nun die Nachbarn des Gewinners am Lernvorgang beteiligt werden, mit einer mit wachsendem Abstand zum Gewinner sinkenden Lernrate. Dies läßt sich Kopplung des Lernvorgangs an den Frontausbreitungsprozeß auf dem aktiven Medium erreichen:

- $t = 0$: Das Gewinnerneuron startet eine Front im aktiven Medium.
- Die Front breitet sich mit der Geschwindigkeit v aus.
- Jedes Neuron beginnt den Lernvorgang, wenn es von der Front erreicht wird, also das Medium am Ort des Neurons eingeschaltet wird.
- Zum Zeitpunkt $t = T$ wird des aktive Medium ausgeschaltet, der Lernvorgang endet für alle Neuronen gleichzeitig. T ist dabei eine vorgewählte Gesamtlernzeit, und wird von der Steuerungseinheit bestimmt.

Die Lernzeit t jedes Neurons nimmt mit zunehmendem Abstand vom Gewinnerneuron ab: $t_j = T - v \cdot d(j, w)$, wobei der Abstand der Neuronen durch ihre Anordnung auf dem aktiven Medium bestimmt wird. Diese Anordnung bestimmt so die Topologie des Netzes.

Die Zeitkonstante $\tau = RC$ der Speicherglieder muß der maximalen Lernzeit T angepaßt sein, damit sinnvolle Lernraten und Nachbarschaftsbreiten entstehen. Im allgemeinen wird im Verlauf des Trainings einer Kohonen-Karte die Lernrate und die Nachbarschaftsbreite reduziert, was erreicht werden kann, wenn T nach und nach reduziert wird.

Fordert man unabhängige Einstellmöglichkeiten für die Lernrate η und die Nachbarschaftsfunktion h_d, so muß auch τ justierbar ausgeführt werden, z.B. durch Verwendung verschiedener Ladewiderstände R, die mittels Analogschaltern umgeschaltet werden.

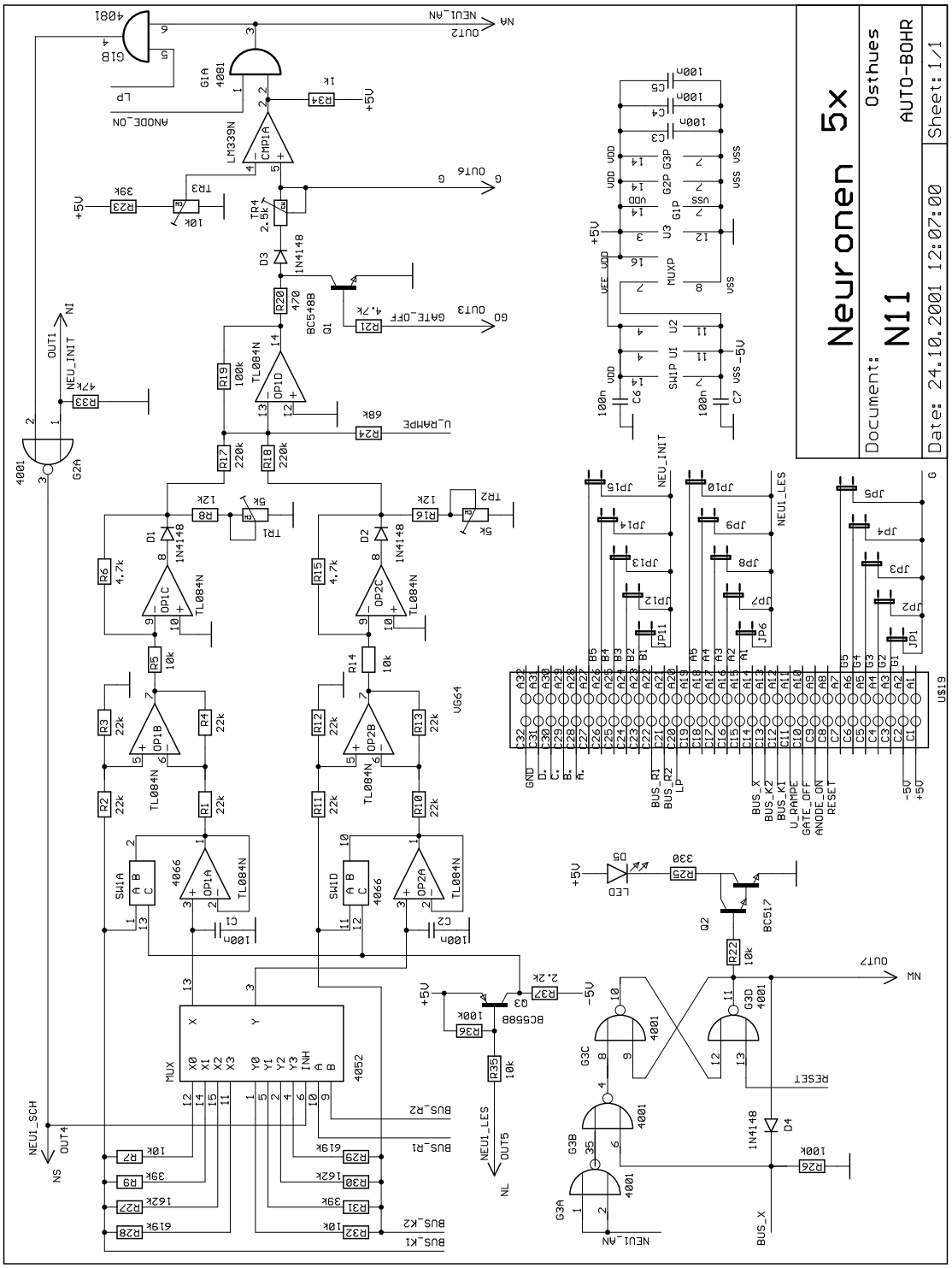


Abbildung 2.4: Schaltplan eines Neurons der Demonstrationshardware

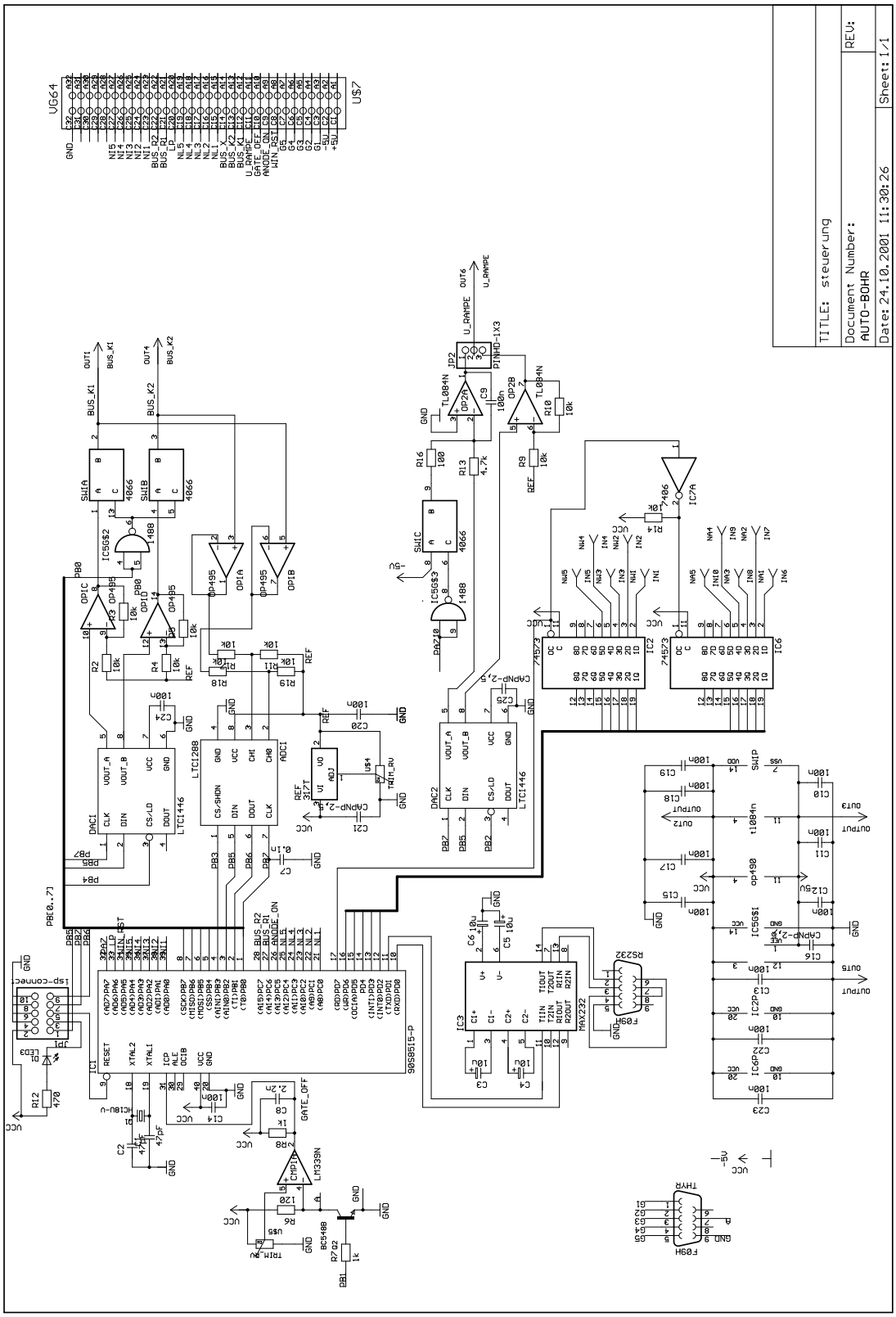


Abbildung 2.5: Schaltplan der Steuerplatine der Demonstrationshardware

2.3 Die Demonstrationshardware

Nach den oben beschriebenen Prinzipien wurde ein Demonstrationsaufbau realisiert. Er implementiert ein analog aufgebautes Kohonennetz mit 5 Neuronen auf einem linearen Kortex mit einem zweidimensionalen Musterraum. Der Aufbau besteht aus 5 Neuronenplatinen und einer Steuerplatine, die über ein Bussystem miteinander kommunizieren. Die Kopplung der Neuronen geschieht mittels Frontausbreitung auf einer Thyristorprobe. Zunächst wurde Probe 98-1D eingesetzt, die Ergebnisse in Abschnitt 2.4 wurden jedoch mit Probe 99/4 gewonnen, da diese durch die Art ihrer Herstellung bessere kleinere Kontaktwiderstände und eine bessere Homogenität der Zündschwellen der einzelnen Gatekontakte aufweist. Die Eigenschaften der Thyristorstrukturen werden im einzelnen im Kapitel 3 dargestellt. Der Demonstrator wurde im Zusammenhang mit der Diplomarbeit von Jens Fischer [Fis00] konstruiert, und in der Folge schaltungstechnisch optimiert und umfassend getestet.

2.3.1 Steuerplatine

Die Steuerplatine (Abbildung 2.5) enthält als zentrales Element einen Microcontroller `IC1, 90AT8515` aus der AVR-Serie der Firma Atmel. Die Taktfrequenz beträgt 8 MHz, der Prozessor ist in Harvard-Architektur aufgebaut mit 8kB Flash-EEPROM für den Programmcode und 512 Bytes RAM als Datenspeicher. Das Programm, das auf dem Controller läuft, ist in C geschrieben und mit dem freien C-Compiler `avr-gcc` auf einem unter Linux laufendem Host compiliert[1]. Dessen Verfügbarkeit ist ein wichtiger Grund für die Wahl eines Controllers aus der AVR-Serie, wie auch die verschiedenen integrierten Schnittstellen, und die Möglichkeit, den Programmcode des Controllers *in Circuit* programmieren zu können, also ohne den Controller aus der Schaltung zu nehmen, was die Fehlersuche und -beseitigung erleichtert.

Der Controller besitzt eine eingebaute RS-232-Schnittstelle, die verwendet wird, um mit dem Hostrechner zu kommunizieren. Ein 2-Kanal D/A- und ein A/D-Wandler (`DAC1,ADC1`)werden als Interface zur eigentlichen analogen Hardware, also den 5 Neuronen-Platinen, eingesetzt. Die Wandler haben 12 bit Auflösung und eine serielle Schnittstelle namens SPI, die vom AVR-Controller hardwaremäßig unterstützt wird. Dazu dienen die Port-Pins `SCK, MOSI, MISO`. Der SPI-Bus besitzt eine Takt- und zwei Datenleitungen – je eine für Hin- und Rückrichtung – und arbeitet wie ein Schieberegister: Datenbytes werden Bit für Bit übertragen, jedes

[1] Der auf die AVR-Architektur portierte GNU-C-Compiler `gcc`, eine kurze Anleitung ist unter [Nes00] zu finden.

Bit wird synchron mit einer steigenden Taktflanke auf den Datenbus gelegt, mit der fallenden Taktflanke wird es dann vom Empfänger gelesen. Dank der 2 Datenleitungen ist eine bidirektionale Übertragung möglich. An diesem Bus sind die Wandler DAC1, DAC2 und ADC1 angeschlossen, wobei jeweils eines dieser Bauteile durch ein *Chip Select* ausgewählt wird.

Die Taktfrequenz des SPI-Bus wurde auf 125 kHz gesetzt, da sie die Arbeitsfrequenz des AD-Wandlers bestimmt, und dieser bei höheren Frequenzen nicht ordnungsgemäß arbeitet. Die seriellen 2-Kanal-Wandler der Firma Linear Technologies wurden Wandlern mit parallelem Interface vorgezogen, obwohl diese eine größere Wandlungsgeschwindigkeit bieten. Die nötige Anzahl IO-Leitungen für die parallelen Wandler hätte die Steuerplatine wegen der erforderlichen Latches erheblich aufwendiger gemacht.

Ein zweiter D/A-Wandler (DAC2) am gleichen SPI-Bus erzeugt eine Spannung, aus der ein Integrator (IC OP2A) die zur Gewinnerbestimmung nötige Rampenspannung gewinnt. Die Rampe wird über den CMOS-Schalter SW1C gestartet bzw. gestoppt. Somit kann programmgesteuert die Anstiegsgeschwindigkeit der Rampe in weiten Grenzen eingestellt werden.

Die Rampenspannung kann auch direkt vom zweiten Kanal des D/A-Wandlers DAC2 erzeugt werden, so daß programmgesteuert ein beliebiger Verlauf der Rampe erzeugt werden kann. Dazu muß der Jumper JP2 umgeschaltet werden.

Die restlichen freien Pins des Controllers steuern einzelne digitale Steuerleitungen an, mit denen der Ablauf des Lernvorgangs gesteuert werden. Im einzelnen sind dies:

- WR schaltet die D/A-Wandler der Steuerplatine auf den Bus, in Vorbereitung eines Schreib- oder Lernvorgangs.
- NI1-5 (Neuron Init) läßt das jeweilige Neuron das am Bus anliegende Muster lernen. Bei ausreichender Dauer des Schreibvorgangs erreicht das Neuron den präsentierten Zustand.
- NL1-5 (Neuron Lesen) das Neurons legt seinen Prototypen an den Bus, der Controller kann ihn mit Hilfe von ADC1 lesen.
- Der Lernvorgang wird folgendermaßen gesteuert: Zunächst wird der Thyristor mit dem Signal TH-AUS gelöscht. Die Rampenspannung wird durch Abschalten von Rampe-Reset RR gestartet, wodurch die einzelnen Neuronen ihre Gateströme kontinuierlich erhöhen. Der Komparator CMP1A detektiert

den Zündzeitpunkt des Thyristors. Er gibt das Signal `GATE-OFF` an die Neuronen, die dann ihren Gatestrom abschalten. Der Controller startet zum Zeitpunkt der Zündung des Thyristors einen Timer. Nach Ablauf der *Lernzeit* wird der Thyristor abgeschaltet und der Lernvorgang beendet. Während des Lernvorgangs lernt jedes Neuron, sobald die Zündfront „seinen“ Gatekontakt erreiht hat.

- Lern-Phase `LP` dient dazu, den Lernvorgang ganz zu unterdrücken und nur die Gewinnerermittlung durchzuführen, also um ein austrainiertes Netz als Klassifikator einzusetzen.

- Die Signale `NW1-5` (Neuron Winner) und `NA1-5` (Neuron aktiv) werden über 2 Latches auf dieselben Pins des Controllers gemultiplext, da nicht genügend Pins für beide Signale am Controller vorhanden sind. Dabei zeigen die `NW1-5` (Neuron-Winner) das Gewinner-Neuron nach erfolgter Gewinnersuche an, während die `NA1-5` anzeigen, daß ein Neuron momentan lernt. Fragt man die `NA`-Signale während des Lernvorgangs ab, so läßt sich die momentane Breite der am Lernvorgang beteiligten Nachbarschaft in Erfahrung bringen.

- Der Thyristor ist mit Anode, Kathode und 5 Gate-Anschlüssen an die Steuerplatine angeschlossen. Die Gate-Anschlüsse werden über die Busplatine auf die 5 Neuronen verteilt.

Digital-Analog-Interface

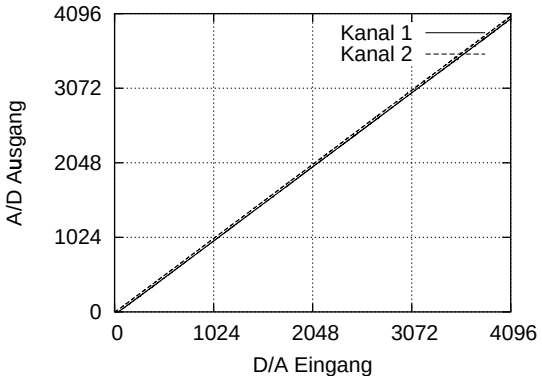
Die Wandler `DAC1` und `ADC1` verbinden die eigentliche analoge neuronale Hardware mit dem digitalen Teil, dem Controller und letzlich dem Hostcomputer. Der D/A-Wandler hat eine Auflösung von 12 Bit, also 4096 Quantisierungsstufen, mit einem Ausgangsspannungsintervall von $[0, 4.096]$ V. `OP1C` und `OP1D` verstärken die Ausgangsspannung um den Faktor 2 und transformieren sie in das Intervall $[-4.096, 4.096]$ V. Diese wird über die Analogschalter `SW1A` und `SW1B` an den analogen Bus gelegt.

`OP1A` und `OP1B` puffern die Spannung am Bus, die nachfolgenden Spannungsteiler transformieren sie wieder in das Intervall $[0, 4.096]$ V. Der 12-Bit-A/D-Wandler `ADC1` erzeugt dann einen digitalen Ausgangswert..

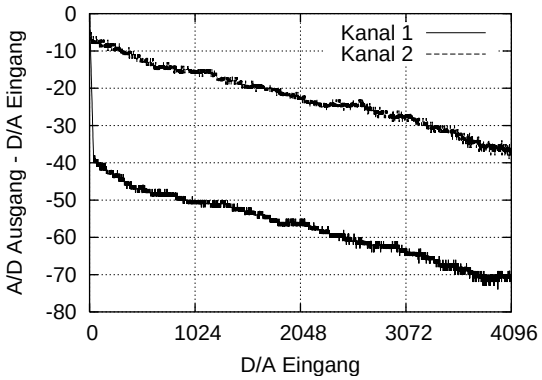
In Abbildung 2.6 sind die Ergebnisse der Kalibrierung des Digital-Analog-Interfaces zu sehen. Es sind links oben der digitale Eingangswert X_{in} und der digitale Ausgangswert X_{out} gegeneinander aufgetragen, die idealerweise gleich sein sollten. Die vorhandene Abweichung $X_{\text{out}} - X_{\text{in}}$ ist oben rechts zu sehen. Der größte Teil dieses Fehlers ist eine lineare Abweichung, die durch die unvermeidlichen

Toleranzen der Widerstände entsteht, die die Verstärkung der beteiligten Operationsverstärker bestimmen.

Die durch einen Fit bestimmte linearisierte Übertragungsfunktion des Interfaces ist unten links zu sehen, der Restfehler nach ihrer Subtraktion unten links. Diese linearisierte Übertragungsfunktion wird während des Betriebs der SOM-Hardware auch zur Korrektur der von der Hardware gelesenen Mustervektoren eingesetzt. Abgesehen von einem schmalen Bereich an den Rändern des Wertebereichs, wo Clipping auftritt, ist der Restfehler recht klein, im Mittel ca. 1 LSB und in jedem Fall unter 4 LSB.



Übertragungsfunktion X_{out} gegen X_{in}:
$X_{\text{in}} \rightarrow$ DAC $\rightarrow$Analogbus$\rightarrow$ ADC $\rightarrow X_{\text{out}}$

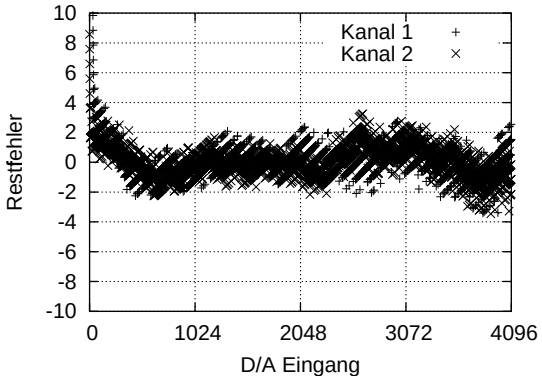


Übertragungsfehler $X_{\text{out}} - X_{\text{in}}$

linearisierte Übertragungsfunktion
$X_{\text{out}} = H(X_{\text{in}}) = aX_{\text{in}} + b$

Fitparameter:

Kanal	a	b
1	0.993	-42.6
2	0.993	-8.5



Restfehler nach linearer Korrektur
$X_{\text{out}} - \text{H}(X_{\text{in}})$

Abbildung 2.6: Übertragungsfunktion des Digital-Analog-Interfaces

2.3.2 Bussystem

Eine 32-adrige Busplatine verbindet die Steuerplatine mit den 5 Neuronenplatinen. Der Bus enthält den analogen Mustervektor-Bus, über den sowohl die Muster den Neuronen präsentiert, als auch Prototypen gelesen und geschrieben werden können. Weiterhin verteilt der Bus die Gateanschlüsse des Thyristors auf die 5 Neuronen, ebenso wie die Schreib- und Lesesignale `NI1-5`, `NL1-5`. Diese stellen den Adressbus dar. Auf die Binärcodierung der Adressen wurde wegen der kleinen Neuronenzahl verzichtet.

2.3.3 Neuronenplatine

Die Neuronen sind auf 5 gleichen Platinen in analoger Bauweise auf Operationsverstärkerbasis realisiert. Jedes Neuron (Abbildung 2.4) besteht aus den folgenden Komponenten:

- 2 Analogspeicher, bestehend aus einem Speicherkondensator, einem CMOS-Schalter(`MUX`) und einem Pufferverstärker(`OP1A, OP2A`). Zum Schreiben und Lernen verbindet der CMOS-Schalter den Kondensator mit dem analogen Bus. Durch die Ausführung des Schalters als Vierfach-Multiplexer können vier verschiedene RC-Konstanten für den Lernvorgang gewählt werden.
- der Distanzrechenschaltung, bestehend aus 2 Subtrahierern (`OP1B, OP2B`), 2 Gleichrichtern (`OP1C, OP2C`), und einem Addierer(`OP1D`). Dessen Ausgang liefert eine Spannung proportional zur negativen Distanz zwischen Mustervektor und Prototyp, zuzüglich der Rampenspannung. Daraus wird der Gatestrom gebildet, den jedes Neuron in das ihm zugeordnete Gate des Thyristors einspeist. So erzeugt das Neuron mit dem kleinsten Abstand den größten Gatestrom, wodurch schließlich zum Gewinnerneuron wird.
- einem Komparator (`CMP1A`) zur Detektion der Schaltfront auf dem Thyristor. Der Komparator vergleicht das Gatepotential mit einem festen Schwellwert. Oberhalb dieses Schwellwertes gilt der Thyristor als lokal gezündet, was das Neuron zum Lernen veranlaßt.
- Der dem Komparator nachgeschalteten Logik, die über den CMOS-Schalter `MUX` den Lernvorgang steuert.
- Dem Gewinnererkennungs- oder WTA-Flip-Flop[2] (Abb. 2.7). Durch eine Inhibitor-Leitung hemmen sich die WTA-Flip-Flops der einzelnen Neuronen gegenseitig.

[2]WTA: Winner-Take-All

Analogspeicher

Ein Lernschritt geschieht durch kurzzeitiges Einschalten des CMOS-Schalters, der den Speicherkondensator über den Lernwiderstand mit der Busleitung verbindet. Dabei findet eine exponentielle Annäherung an am Bus anliegende Spannung statt. Die RC-Zeitkonstante, die ebenso wie die Lernzeit die Lernrate bestimmt, läßt sich umschalten, dafür sind 4 verschiedene Lernwiderstände vorhanden, von denen jeweils einer aktiv ist. Abbildung 2.8 demonstriert den zeitlichen Verlauf des Lernvorgangs, der zur Messung der RC-Zeitkonstanten der einzelnen Speicherglieder aufgenommen wurde.

Dabei wurde nach Initialisierung des Analogspeichers durch einen langen Lernschritt (mehrere RC-Konstanten) jeweils der gleiche Wert durch kurze Lernschritte geschrieben. Es findet als ein exponentieller Ladevorgang statt, der ständig unterbrochen wird. Nach jedem Lernschritt wird die Spannung des Speicherkondensators über den Bus ausgelesen. Es ergab sich eine gute Übereinstimmung mit dem erwarteten exponentiellen Verlauf der Ladekurve. Die vom Timer des Microcontrollers erzeugten Ladezeiten sind sehr konstant, der Zeitfehler (Jitter) liegt unter 1 µs. Die Zeitkonstanten der 10 Analogspeicher zeigen Abweichungen von ca. ±10%, was auf die großen Kapazitätstoleranzen der Kondensatoren zurückzuführen ist. Die somit unterschiedlichen Lernraten der einzelnen Neuronen haben aber keinen negativen Einfluß auf den Ablauf des Kohonenalgorithmus.

Die Drift des Analogspeichers bei offenem CMOS-Schalter erweist sich als relativ klein: Mit den verwendeten 100 nF-Kondensatoren verliert das schlechteste Neuron 0.75 mV/s auf praktisch lineare Weise. Gemessen in den Einheiten des AD-Wandlers driftet ein Neuron etwa 1 LSB in 2.5 s, und 16 LSB in 45 s, so daß dann von der 12 Bit Auflösung des Wandlers noch 8 Bit gültig sind.

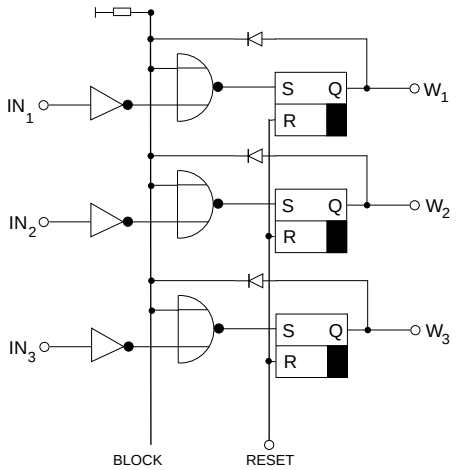
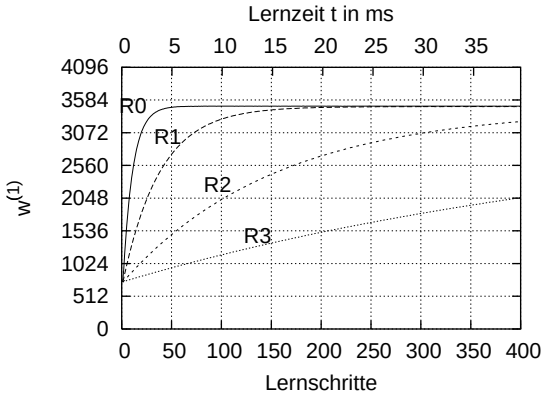


Abbildung 2.7: Gewinnerdetektor-Logik. Zu jedem Neuron gehört ein WTA-Flip-Flop, wobei durch die `inhibit`-Leitung nur jeweils ein Flip-Flop gesetzt werden kann.



R in kΩ	C in nF	τ (steps)	τ in ms
10	100	9.8	0.98
39	100	37.4	3.71
162	100	153.1	15.19
619	100	552.3	54.80

Abbildung 2.8: Demonstration unterschiedlicher Lernraten durch schaltbare Lernwiderstände. Ein Lernschritt dauerte 99.2 µs, es wurden 400 Lernschritte ausgeführt. Die Zeitkonstanten wurden bestimmt durch einen Fit der Funktion $U(t) = (U_0 - U_\infty)\exp(-t/\tau) + U_\infty$ an die Meßdaten. Gezeigte Messung an Neuron 0, Kanal 0.

Distanzrechenschaltung

Die Distanzrechenschaltung bestimmt den Abstand[3] des in den Kondensatoren gespeicherten Prototypen vom am Bus angelegten Mustervektor. Zunächst bildet je Kanal ein Subtrahierer die Differenz der Spannungen am Bus und im Speicher. Ein Gleichrichter bildet davon den Betrag, und der Addierer `OP1d` addiert die Beträge aller Kanäle, und die Rampenspannung U_R, die den Abstandsvergleich steuert. Der Ausgang der Addierers `OP1d` ist über einen Vorwiderstand R_g mit einem Gatekontakt des Thyristors verbunden. Die Ausgänge der Gleichrichter haben immer positives Vorzeichen, so daß der Addierer im Ruhezustand eine negative Spannung zeigt, und kein Gatestrom fließt. Die Rampenspannung U_r ist im Ruhezustand ebenfalls positiv und fällt im Verlauf des Vergleiches, so daß die Addierer-Ausgangsspannungen ansteigen. Sobald diese positiv werden, fließt ein entsprechender Gatestrom.

Der Gleichrichter bzw. Betragsbildner (Abb. 2.9) arbeitet mit einer Diode in der Gegenkopplung eines Operationsverstärkers. Im Fall $U_{\text{in}} < 0$ ist die Ausgangsspannung des OPs positiv, die Diode leitet, und man erhält einen invertierenden Verstärker:

$$U_{\text{out}} = -U_{\text{in}}\frac{R_2}{R_1} \tag{2.3}$$

Für $U_{\text{in}} > 0$ wird der Ausgang des OPs negativ, die sperrende Diode koppelt ihn

[3] Die Manhattan-Distanz wurde implementiert.

von U_{out} ab, und $R_1 + R_2$ und R_3 verhalten sich wie ein Spannungsteiler:

$$U_{\text{out}} = U_{\text{in}} \frac{R_3}{R_1 + R_2 + R_3} \tag{2.4}$$

Wählt man R_1, R_2 und R_3 richtig, so sind die beiden Verstärkungsfaktoren bis auf das Vorzeichen gleich. Mit $R_1 = 2R_2, R_3 = 3R_2$ ergibt sich der Verstärkungsfaktor $1/2$, also

$$U_{\text{out}} = 1/2|U_{\text{in}}| \tag{2.5}$$

Der Ausgangswiderstand des Gleichrichters ist in beiden Zuständen unterschiedlich, er ist nur dann niederohmig, wenn die Diode leitet und der OP aktiv ist, sonst wird er durch den Spannungsteiler bestimmt. Eine Belastung des Gleichrichterausgangs führt zu einem Fehler. Der nachfolgende Addierer bildet jedoch einen festen Eingangswiderstand nach Masse, so daß man diesen mit in den Spannnungsteiler einbeziehen kann, und so den Belastungsfehler kompensieren kann.

Abbildung 2.10 zeigt die gemessene Übertragungsfunktion eines Kanals der Distanzrechenschaltung. Die Ausgangsspannung des Gleichrichters wurde als Funktion der beiden Eingangsspannungen des Subtrahierers gemessen und aufgetragen. Man erkennt eine gute Übereinstimmung des theoretischen Sollwertes mit dem gemessenen Ergebnis, abgesehen von einem Skalenfaktor, die Gesamtverstärkung ist etwa 10% geringer als 1/2. Bei großen Differenzwerten fällt auf, daß Clipping einsetzt und der Ausgangswert des Gleichrichters begrenzt wird. Dies geschieht im vorgeschalteten Subtrahierer, wenn die Differenz der Eingangswerte zu groß wird und der Ausgang des Subtrahierers in Sättigung geht. Eine leichte Asymmetrie zwischen den beiden Ästen, also unterschiedliche Verstärkungen für positive und negative Eingangswerte der Gleichrichters, beruht auf den Toleranzen der Widerstände im Gleichrichter. Die Funktion der SOM-Hardware wird durch diese Abweichungen nicht wesentlich beeinflußt.

Thyristorinterface

Jedes Neuron ist mit einem Gatekontakt auf dem Thyristor verbunden. Der Thyristor ist gewissermaßen der Kortex des Systems und gibt den Neuronen die Nachbarschaftsbeziehung. Die Wechselwirkung der Neuronen mit dem Thyristor ist bidirektional: Ein Neuron kann den Zustand des Thyristors beobachten, oder ihn beeinflussen.

Die Neuronen können einen Gatestrom in „ihren“ Gatekontakt einprägen und den Thyristor damit lokal zur Zündung bringen. Dieser Gatestrom wird von der Distanzrechenschaltung erzeugt. Zur ermittelten Distanz wird dort vom Addierer

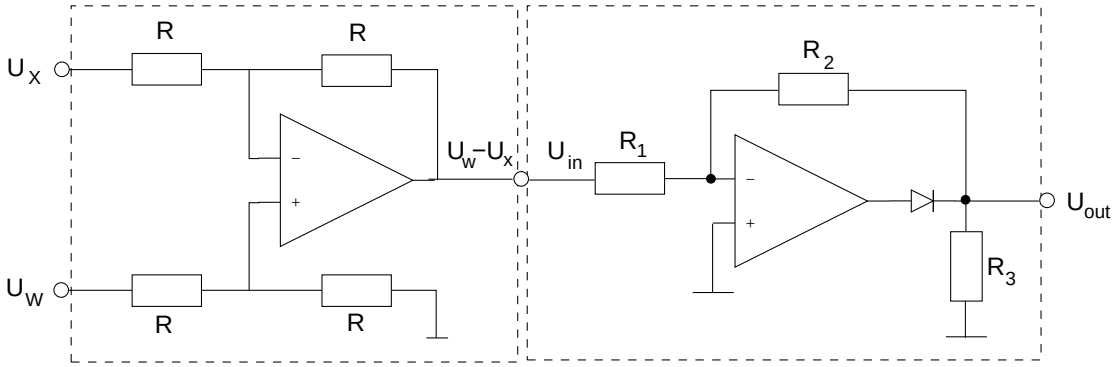


Abbildung 2.9: Distanzrechenschaltung aus Subtrahierer und Gleichrichter. Die Widerstände im Gleichrichter haben folgende Werte: $R_1 = 10\ \mathrm{k\Omega}$, $R_2 = 5\ \mathrm{k\Omega}$, $R_3 = 15\ \mathrm{k\Omega}$, damit ergibt sich: $U_{\mathrm{out}} = 1/2|U_{\mathrm{in}}|$.

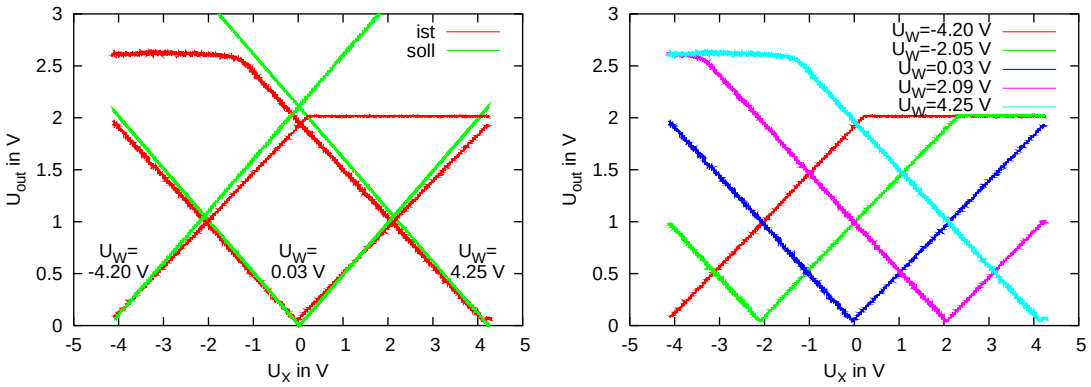


Abbildung 2.10: Distanz-Rechenschaltung, Übertragungsfunktion für einen Kanal. Aufgetragen ist die Ausgangsspannung U_{out} gegen die Eingangsspannung U_X (Mustervektor), bei jeweils unterschiedlichen Werten der Eingangsspannung U_W (Prototyp). Die Ausgangsspannung stimmt bis auf einen Skalierungsfaktor gut mit dem erwarteten Sollwert überein. Dies und die Abweichung der Steigungen des positiven und negativen Astes von 1/2 ist durch die Toleranz der verwendeten Widerstände bedingt. Das Clipping bei großen Abständen tritt im vorgeschalteten Subtrahierer durch Eintreten der Sättigung auf.

OP1D die „Rampenspannung“ U_R addiert, die zentral von der Steuerung erzeugt wird. Die Zündung wird ausgelöst, wenn der Gatestrom des ersten Gatekontakts die Zündschwelle überschreitet.

Jedes Neuron beobachtet den Zustand des ihm zugeordneten Gatekontakts mit einem Komparator. Dieser ist mit der Steuerlogik für den Lernvorgang verbunden. So wird die Nachbarschaftssteuerung der Lernrate erzeugt.

2.3.4 Prozessablauf und Timing

Der Klassifikations- und Lernvorgang wird vom Controller folgendermaßen gesteuert:

- ein Mustervektor wird mittels DAC1 an den analogen Bus gelegt.
- Die Neuronen bestimmen (praktisch instantan) ihre Distanzen zum Mustervektor.
- Die Vergleichsrampe wird gestartet, indem *Rampe-Reset* abgeschaltet wird. U_r liegt im Ruhezustand in der positiven Sättigung bei ca. 4 V und fällt nun mit einer Rate, die durch die Ausgangsspannung des Wandlers DAC2 bestimmt wird. Entsprechend steigen die Gateströme der einzelnen Neuronen nun an.
- Der Controller wartet auf die Zündung des Thyristors, der vom Komparator CMP1A über Einbruch der Anodenspannung detektiert wird.
- Sobald der Gatestrom an einem Gatekontakt die Zündschwelle überschritten hat, zündet der Thyristor. Der im Controller integrierte Timer wird gestartet, um den Lernvorgang nach Ablauf der vorgewählten Lernzeit beenden zu können. Das Signal Anode-on wird gesetzt, was den Neuronen das Lernen erlaubt, wenn ein Neuron von der Front erreicht wird, und gleichzeitig die Gateströme abschaltet.[4] Abb. 2.11 zeigt den Ablauf von der Zündung der Thyristors bis zum Ende des Lernvorgangs.
- Nach Ablauf des Timers wird Anode-on wieder abgeschaltet, damit beenden alle Neuronen den Lernvorgang.

[4] Es würde im Prinzip ausreichen, nur das Gatepotential am jeweiligen Gatekontakt zum Steuern des Lernvorgangs zu verwenden, die ersten präparierten Thyristorproben wiesen jedoch relativ hochohmige Gatekontakte auf, so daß das Gatepotential vor der Zündung durch die Einprägung des Gatestromes schon Werte oberhalb der Detektorschwelle annehmen konnte, was den Lernvorgang schon vor der Zündung gestartet hätte.

- Vor dem letzten Schritt können die Signale `NA1-5` ausgelesen werden, um die Ausdehnung der Nachbarschaftsfunktion zu bestimmen.

- Dann werden die Ausgänge der WTA-Flipflops `NW1-5` gelesen und so das Gewinnerneuron bestimmt.

- Die Rampe wird zurückgesetzt und der Thyristor gelöscht.

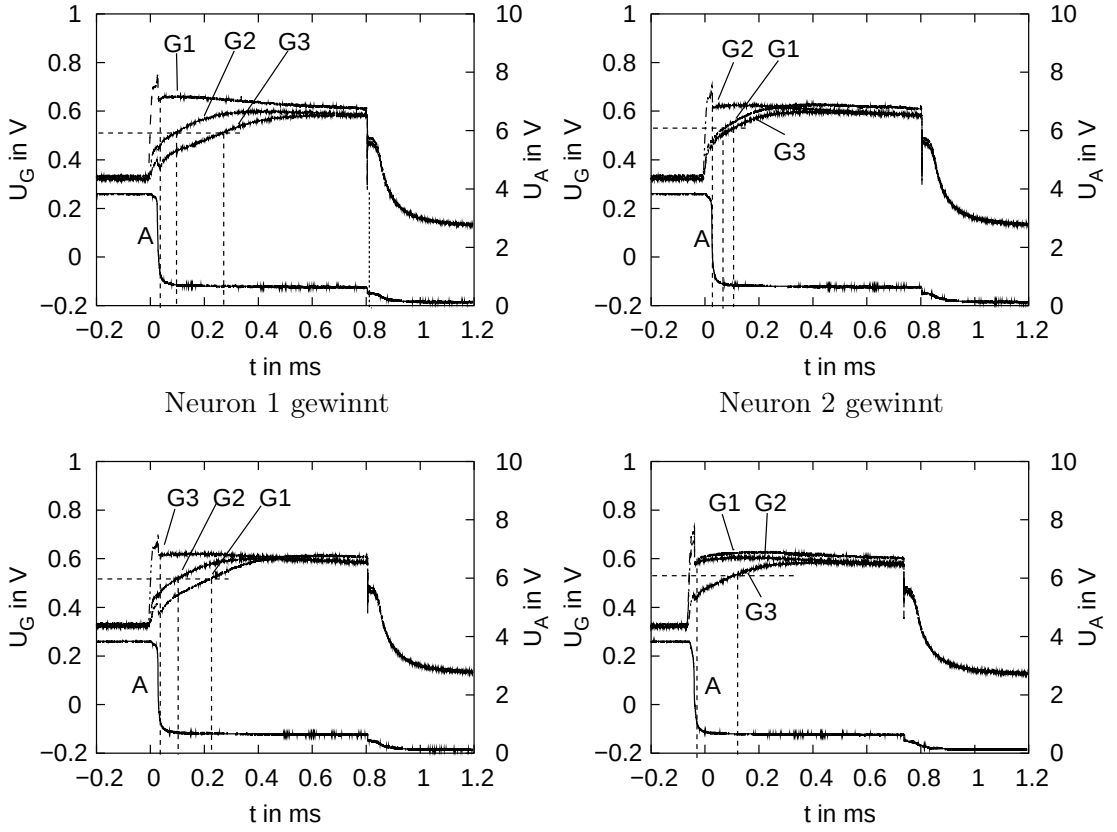


Abbildung 2.11: Timing des Lernvorgangs bei jeweils unterschiedlichem Gewinnerneuron. Die Mustervektoren wurden dazu jeweils passend gewählt. Die dargestellte Zündung des Thyristors geschieht nachdem die Rampenspannung soweit gefallen ist, daß der erste Komparator (`OP1D`) der Zündstrom für einen Gatekontakt einschaltet. Die Zündung erfolgt nach einer Zündverzugszeit von ca. 30 µs. Der Spannungsanstieg an den anderen Gatekontakten folgt nach einer entsprechenden Frontlaufzeit. Der Lernvorgang der Neuronen beginnt jeweils mit dem Anstieg der Gatespannung über die Schaltschwelle von $= 0.55$ V, und endet mit dem Abschalten des Thyristors. Parameter: $R_{17} = \infty$, d.h. die Komparatorfunktion liegt bei `OP1D`. Rampenanstiegsrate $dU_r/dt = 0.5\text{V/ms}$.

Die Gesamtzeit für einen Lernvorgang setzt sich hauptsächlich aus zwei Anteilen zusammen: Die Rampenlaufzeit (je nach gewählter Rampengeschwindigkeit ca. 10 ms – 0.7 ms) und die Lernzeit (ca. 64 µs – 500 µs bei einer Nachbarschaftsbreite von 0 – 4 Neuronen). Die Laufzeit der Rampe muß daher wesentlich größer als die Zündverzugszeit des Thyristors (ca. 30 µs) sein, um eine präzise Klassifikation zu ermöglichen. Die Lernzeit hängt von der gewünschten Nachbarschaftsbreite ab, die bei fortschreitendem Lernvorgang reduziert wird. Beide Zeiten sind aufgrund der Eigenschaften der verwendeten Thyristorprobe relativ lang, sollten sich aber durch Miniaturisierung der Thyristorstruktur erheblich reduzieren lassen.

Der Controller dient auch als Interface zwischen dem Host-PC und dem analogen Bus. Jede Aktion der Hardware, wie Schreiben und Lesen eines Neurons, oder Ausführung eines Lernschritts, wird durch einen Befehl vom Host eingeleitet, der über die serielle Schnittstelle empfangen wird. Die Befehle werden dabei durch einzelne ASCII-Zeichen codiert, worauf die erforderlichen Parameter folgen. Nach Ausführung der Aktion wird das Ergebnis wieder über zum Hostrechner übertragen. Interessant ist die Feststellung, daß das Lesen des analogen Busses über `ADC1` nur zum Auslesen der Prototypen notwendig ist, es dient also nur zur Kontrolle bzw. zur Darstellung des Netzzustandes. Für Klassifikation und Lernvorgang ist das Auslesen der Neuronen nicht notwendig.

2.4 Experimentelle Ergebnisse

Hier werden einige Experimente vorgestellt, die die Funktionsfähigkeit der parallelen Hardware darlegen. Zunächst wird die Klassifikation von Mustervektoren bei vorgegebenen Prototypen getestet, dann der eigentliche Lernvorgang, also die Adaption der Prototypen an eine vorgegebene Mustermenge überprüft.

2.4.1 Klassifikation

Zur Demonstration der Klassifikation der SOM-Hardware wird ein Versuch zur Bestimmung von Voronoizellen vorgestellt. Das Experiment wird durch ein auf dem Hostrechner laufendes Steuerprogramm namens `voronoi` gesteuert, das mit der SOM-Hardware kommuniziert.

Dazu werden zunächst die Neuronen mit vorgegebenen Prototypen initialisiert. Der zweidimensionale Musterraum wird in ein Gitter eingeteilt, die SOM-Hardware bestimmt dann für jeden Mustervektor jeweils das Gewinnerneuron. In einer graphischen Darstellung des Musterraums werden die zu den einzelnen Neuronen gehörenden Gebiete – die Voronoizellen – durch eine Farbcodierung dargestellt. Zum

Vergleich werden die Voronoizellen auf die gleiche Weise numerisch bestimmt. Als Abstandsmaß wird wie in der Hardware die Manhattan-Norm benutzt.

Durch den Vergleich dieser experimentellen und theoretischen Voronoi-Diagramme läßt sich die Genauigkeit der Klassifikation der analogen Hardware abschätzen, und es lassen sich verschiedene Fehlereinflüsse erkennen.

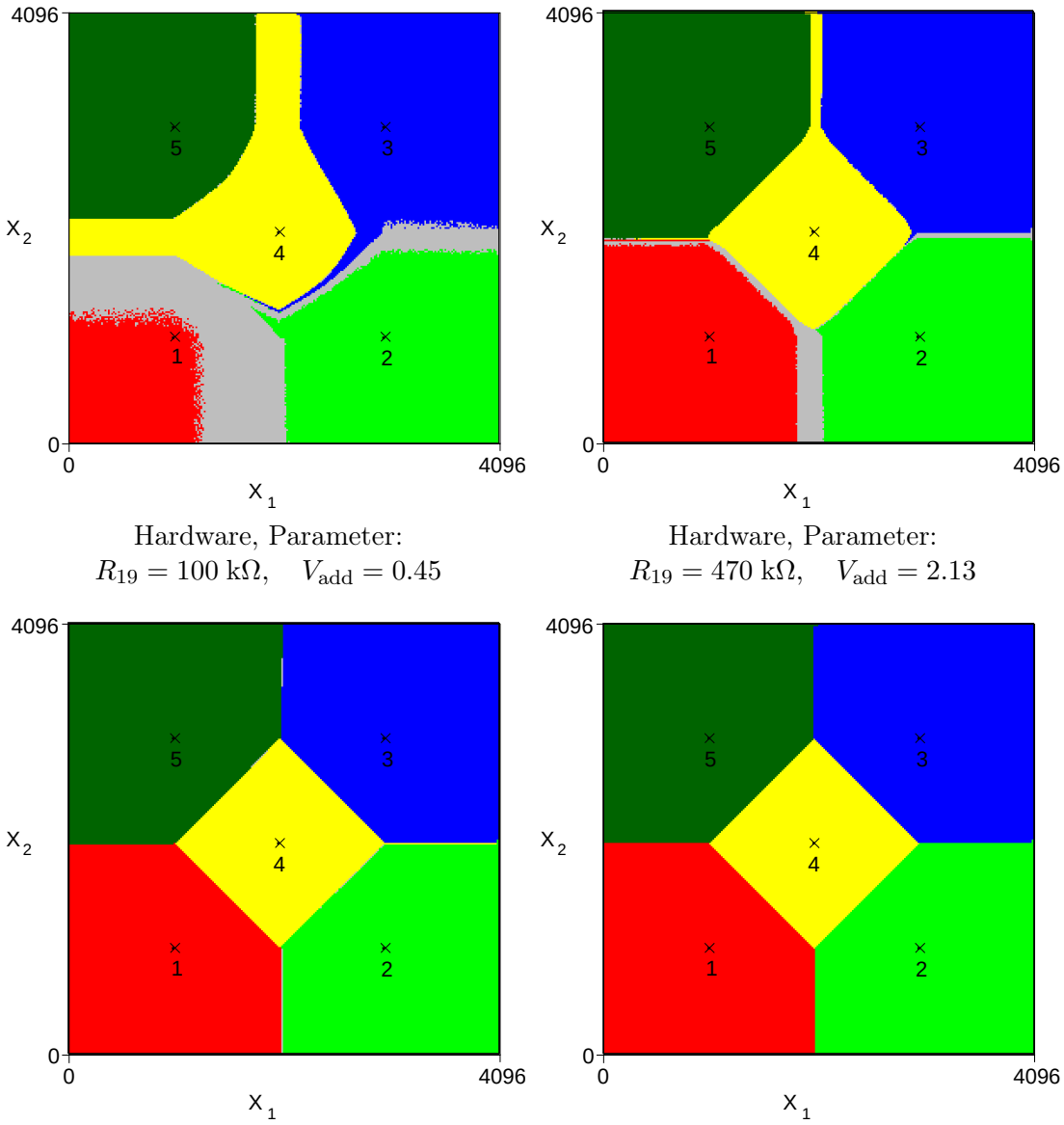
Zur Rasterung des zweidimensionalen Musterraums, aufgeteilt in $N \times N$ Gitterpunkte, benötigt man $O(N^2)$ Klassifikationen. Die neuronale Hardware ist relativ langsam, in der derzeitigen Ausführung erreicht sie maximal 50 Klassifikationen pro Sekunde. Zur Steigerung der Auflösung wird der Musterraum zunächst grob gerastert, dann werden die Grenzen der Einzugsgebiete geortet, und in der Umgebung dieser Grenzen erneut mit hoher Auflösung gerastert. So wird die Auflösung der Darstellung gezielt im Bereich der Zellengrenzen erhöht. Nur so kann eine feine Darstellung der Voronoizellen in vertretbarer Zeit gewonnen werden.

Die Experimente zeigten zwei hauptsächliche Fehlerquellen, deren Einfuß auf das Klassifikationsergebnis hier erläutert werden soll. Die erste besteht in der unterschiedlichen Zündempfindlichkeit der verschiedenen Gatekontakte der Thyristorprobe. Die Gewinnersuche geschieht ja durch den Vergleich des Gatestroms jedes Neurons I_{G}^{i} mit der Zündschwelle I_{th}^{i}. Der Gatestrom wird dabei vom Addierer `OP1D` aus den vom Neuron bestimmten Abstandsspannungen und der Rampenspannung gebildet.

$$I_g \approx -R_g \cdot (\frac{R_{19}}{R_{17}} U_d + \frac{R_{24}}{R_{17}} U_r) \tag{2.6}$$

Dabei ist R_g der Gatevorwiderstand, die Widerstände R_{19}, R_{17} und R_{24} bestimmen die Verstärkung der verschiedenen Eingänge des Addierers, und somit in Verbindung mit der Rampenspannung U_r die Anstiegsrate der Gateströme. Die Gateströme steigen, bezogen auf die Reaktionszeit der Thyristorstruktur, „langsam“ an, bis die Zündung ausgelöst wird. Sind nun die Zündschwellen der Gatekontakte unterschiedlich, so verändern sich die Größen der jeweiligen Einzugsgebiete der Neuronen entsprechend. Die Voronoizelle eines Neurons mit übermäßig empfindlichem Gatekontakt dehnt sich auf Kosten der anderen Voronoizellen aus.

Ein anderer Fehler entsteht, wenn zwei Neuronen ungefähr den gleichen Gatestrom in ihre Gatekontakte einprägen, wenn sich also der Mustervektor an der Grenze zweier Voronoizellen befindet. Die Zündschwelle ist dann durch eine Wechselwirkung der Gateströme vermindert, außerdem kann die Zündung fälschlicherweise *zwischen* den betreffenden Gatekontakten geschehen und detektiert werden. So kann beispielsweise ein Gebiet auf der Grenze von Neuron 1 und Neuron 3 fälschlicherweise Neuron 2 zugeordnet werden. Diese Effekte hängen mit der relativ



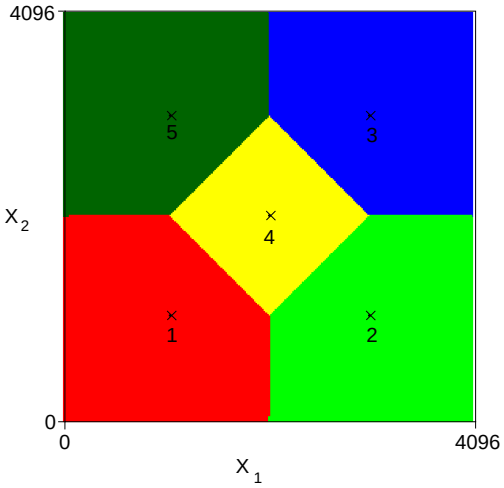
Hardware, Parameter:
$R_{19} = 100\ \mathrm{k\Omega}, \quad V_{\mathrm{add}} = 0.45$

Hardware, Parameter:
$R_{19} = 470\ \mathrm{k\Omega}, \quad V_{\mathrm{add}} = 2.13$

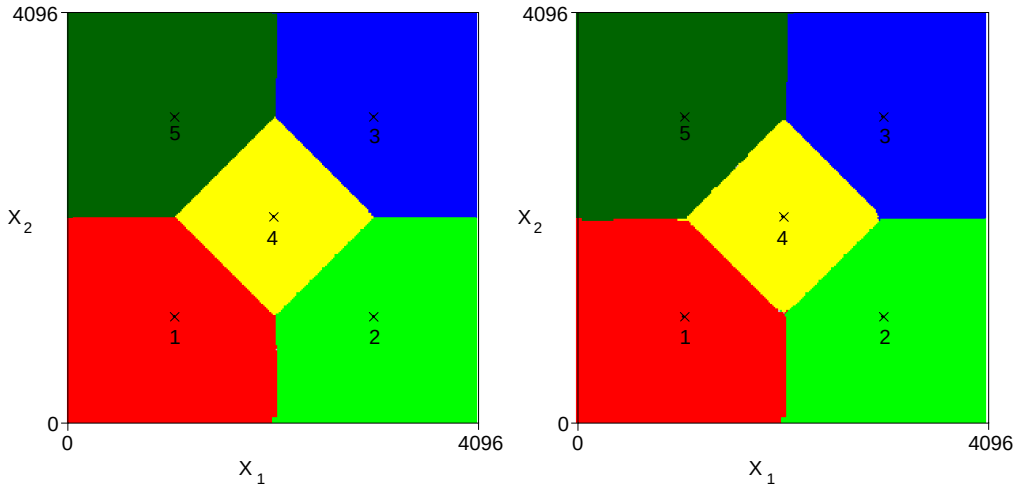
Hardware, Parameter:
$R_{19} = \infty\ \mathrm{k\Omega}, \quad V_{\mathrm{add}} = \infty$

Exakte Klassifikation

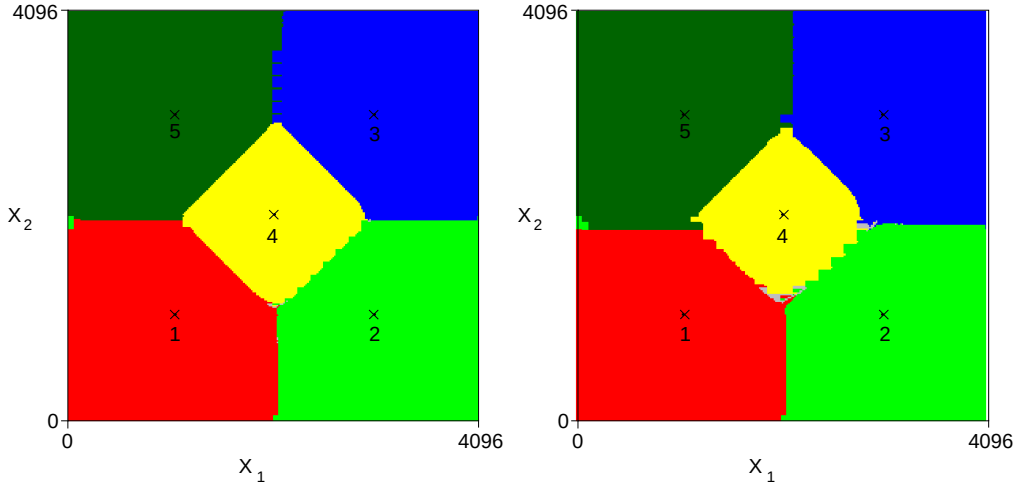
Abbildung 2.12: Einfluß der Verstärkung V_{add} des Addierers `OP1D` auf das Klassifikationsergebnis der SOM-Hardware. Grau dargestellt sind Gebiete, in denen kein oder mehrere Neuronen auf einen Mustervektor ansprechen.



Klassifikation exakt



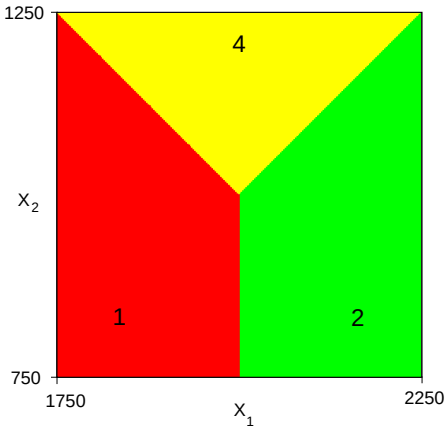
$dU_r/dt = 0.5$ V/ms $dU_r/dt = 2.0$ V/ms



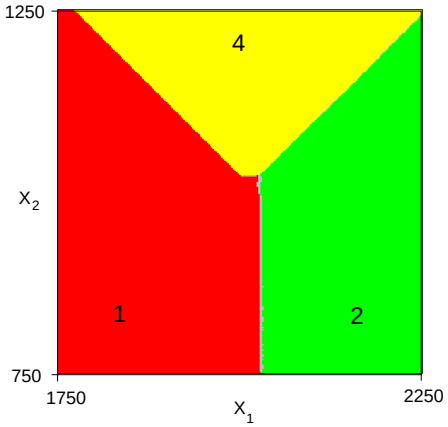
$dU_r/dt = 4.0$ V/ms $dU_r/dt = 8.0$ V/ms

Klassifikation SOM-Hardware

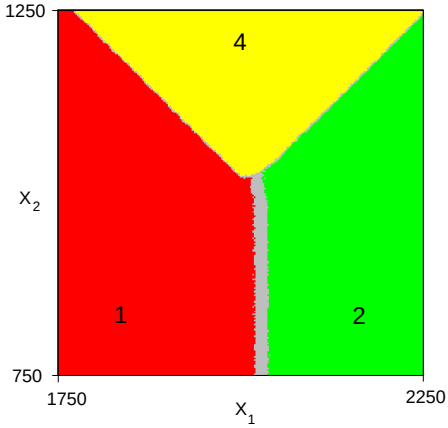
Abbildung 2.13: Einfluß der Rampenanstiegsrate dU_r/dt auf die Funktion der SOM-Hardware als Klassifizierer. Die Verstärkung des Addierers V_{add} wurde auf ∞ festgelegt.



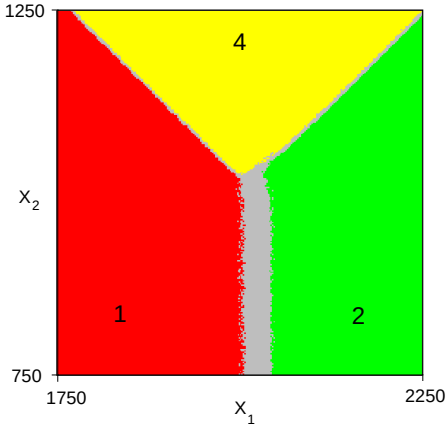
Klassifikation exakt



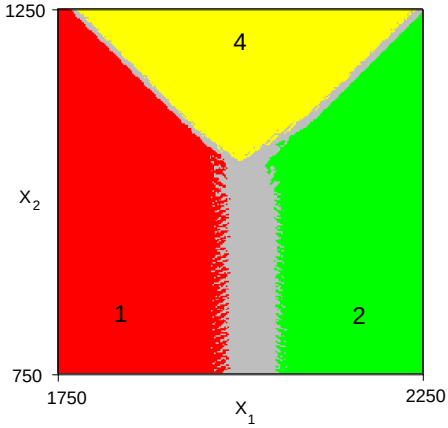
$dU_r/dt = 0.12$ V/ms



$dU_r/dt = 0.5$ V/ms



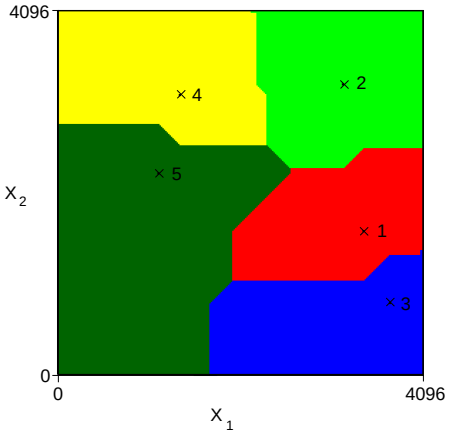
$dU_r/dt = 1.0$ V/ms



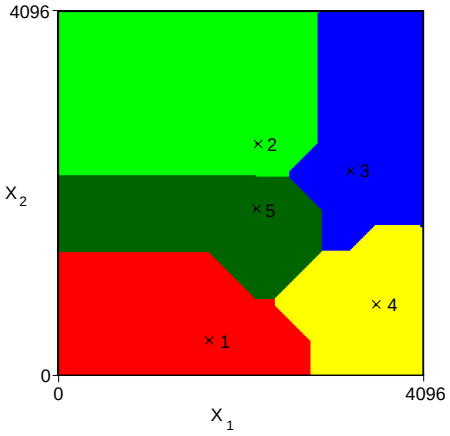
$dU_r/dt = 2.0$ V/ms

Klassifikation SOM-Hardware

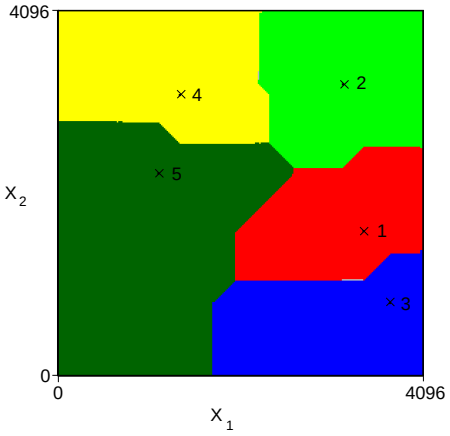
Abbildung 2.14: Ausschnittvergrößerung von Abb. 2.13, Vergrößerung 8:1. Die Artefakte zwischen den korrekt erkannten Gebieten sind deutlich zu erkennen, insbesondere schmale Bereiche, in denen 2 Prototypen als Gewinner erkannt werden (graue Bereiche). Diese fehlerhaft erkannten Gebiete nehmen mit der Anstiegsgeschwindigkeit der Vergleichsrampe zu.



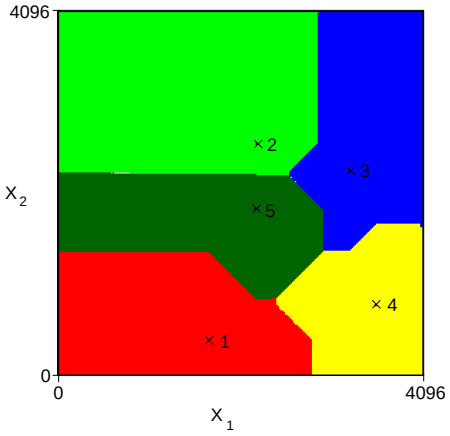
exakt



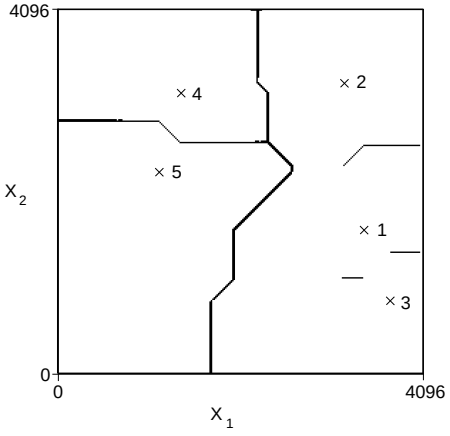
exakt



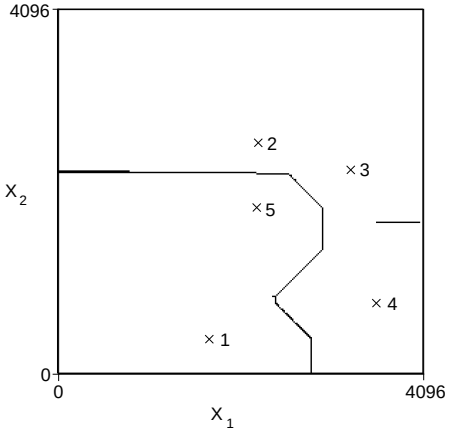
Hardware



Hardware



Differenz



Differenz

Abbildung 2.15: Numerisch und mittels SOM-Hardware bestimmte Anordnungen von Voronoizellen, mit zwei verschiedenen, unregelmäßigen Anordnungen der Prototypen.

großen Breite der Zündfronten zusammen, die sie in der Thyristorprobe ausbreiten. Um sie zu vermeiden, müssen die Abstände der Gatekontakte groß gegenüber der Frontbreite sein.[5]

Die Verwendung der Zündschwelle der hier untersuchten Thyristorproben als Komparator ist also problematisch, insbesondere die gleichzeitige Einprägung mehrerer Gatestöme sollte vermieden werden. Ein weiterer Effekt des bisher beschriebenen Klassifikationsmodus ist die relativ große Zündverzugszeit, die entsteht wenn der Thyristor mit einem nur knapp überschwelligen Strom gezündet wird. Dies macht den Vergleich besonders langsam, und die Zündung des Thyristors anfällig gegen Einkopplung von Störungen.

Diese Fehler lassen sich vermeiden, indem die Verstärkung des Addierers `OP1D` erhöht wird, indem R_{17} vergrößert wird. Im Extremfall wird $R_{17} = \infty$ gesetzt, also der Widerstand ganz entfernt, so daß `OP1D` mit Leerlaufverstärkung arbeitet. In diesem Fall übernimmt `OP1D` die Komparatorfunktion, und schaltet den Gatestrom I_G bei Überschreiten einer Schaltschwelle abrupt ein. In diesem Modus wird zumindest bei genügend langsamer Anstiegsrate der Rampenspannung jeweils nur ein Gatekontakt mit Strom versorgt. Das erste Neuron, dessen Komparator schaltet, zündet den Thyristor und wird Gewinner der Klassifikation. Die Zündverzugszeit ist durch den stark überschwelligen Zündstrom wesentlich kürzer.

Abb. 2.12 zeigt die mit einer regelmäßigen Anordnung von Prototypen gewonnene Einteilung des Musterraumes in Voronoizellen. Die Bilder wurden mit verschiedenen Einstellungen der Verstärkung des Addierers gewonnen. Mit der kleinsten verwendeten Verstärkung, im Bild oben links zu sehen, zeigen sich die Klassifikationsfehler am deutlichsten. Deutlich zeigen sich breite Zonen, in denen entweder zwei Neuronen als Gewinner erkannt werden (graue Bereiche), oder ein *falsches* Neuron Gewinner wird. Bei erhöhter Verstärkung des Addierers verringern sich die Fehler, da der Einfluß der Gate-Zündschwellen nun geringer ist. Praktisch ganz verschwinden die Klassifikationfehler, wenn der Addierer mit maximaler Verstärkung betrieben wird.

[5] Vergleiche Abschn. 3.8.5 und 3.8.6.

Tabelle 2.1: Trennschärfe der Klassifikation. Angegeben ist die maximale Breite der nicht eindeutig klassifizierten Grenzbereiche zwischen den Voronoizellen in Abb. 2.14. Gemessen wurde die Breite des grauen Bereiches zwischen den Gebieten von Neuron 1 und Neuron 2. Die Werte sind auf das analoge Invervall $[-4.1\ \mathrm{V}, 4.1\ \mathrm{V}]$ bezogen, das auf 12-Bit-Ganzzahlen abgebildet wird, also 4096 Quantisierungsstufen.

dU_r/dV V/ms	Breite LSB	Breite mV
1024	77	155
512	41	82
256	18	36
128	9	18
64	4	8

Die Zündverzugszeit stellt die Reaktionszeit des Klassifikators dar. Die Genauigkeit der Klassifikation hängt daher insbesondere von der Anstiegsgeschwindigkeit der Rampenspannung dU_r/dt ab. Die Voronoidiagramme in Abbildung 2.13 zeigen das Klassifikationsergebnis in Abhängigkeit von der Rampenanstiegsgeschwindigkeit dU_r/dt, wobei die oben diskutierte Verstärkung des Addierers auf ∞ festgelegt wurde, R_{17} also entfernt wurde. Man erkennt, daß bei geringer Anstiegsgeschwindigkeit die Voronoizellen praktisch mit dem Sollzustand übereinstimmen, während die Abweichungen bei höherer Geschwindigkeit zunehmen. Deutlicher ist dies in der Ausschnittvergrößerung in Abb. 2.14 zu erkennen. Je größer die Rampengeschwindigkeit gewählt wird, desto größer werden die grau markierten Bereiche, in denen kein eindeutiger Gewinner erkannt wird.

Die Breite dieser nicht eindeutig klassifizierten bereiche ist ein Maß für die erzielte „Trennschärfe“ der Klassifikation. Sie gibt die Abstandsdifferenz an, die zwei Prototypen mindestens haben müssen, damit einer eindeutig als Gewinner erkannt wird. Nimmt man eine linear ansteigende Vergleichsrampe an, so ergibt sich die minimale Abstandsdifferenz als der Anstieg der Vergleichsrampe, der innerhalb der Reaktionszeit des Komparators geschieht. Tatsächlich nimmt die Breite der nicht eindeutig klassifizierten Bereiche linear mit der Rampengeschwindigkeit zu (s. Tabelle 2.1).

Eine begrenzte Trennschärfe bei der Gewinnererkennung kann aber auch sinnvoll sein: Da gemessene Signale nie beliebig genau sind, sind auch die Bereichsgrenzen zwischen zwei Musterklassen nicht beliebig genau zu definieren. Eine Randzone bestimmter Breite kann als verbotene Zone definiert werden, Muster in dieser Randzone werden dann keiner Klasse eindeutig zugeordnet.

2.4.2 Lernvorgang

Um die Funktion der SOM-Hardware entsprechend dem Kohonenalgorithmus auszutesten, wurde ein einfaches Steuerprogramm namens `neurogui` mit graphischer Benutzeroberfläche entwickelt. Damit kann der Trainingsvorgang interaktiv gesteuert und in Echtzeit beobachtet werden. Der Hostrechner übermittelt dabei nur die Mustervektoren an die SOM-Hardware, wo der gesamte Lernvorgang stattfindet.

Es lassen sich Trainingsmusterverteilungen vorgeben, die dem zweidimensionalen Musterraum entsprechend als Bilddatei codiert werden: Grauwerte codieren die Wahrscheinlichkeit, mit der ein Mustervektor an einer bestimmten Stelle des Musterraumes gezogen wird. Somit kann die Trainingsmenge den gesamten Musterraum oder Teile davon abdecken.

Die Parameter des Netzes wie Lernzeit T, RC-Zeitkonstante τ (durch die Auswahl eines von vier Lernwiderständen), Rampengeschwindigkeit dU_r/dt können während

des Lernvorganges verändert werden. Sie bestimmen die Lernrate und die Ausdehnung der Nachbarschaftsfunktion. Die Lage der Neuronenkette im Musterraum wird zusammen mit den bisher präsentierten Mustervektoren ständig dargestellt. Ausgehend von der aktuellen Anordnung der Neuronen kann ein Voronoidiagramm durch Rasterung des Musterraumes erzeugt werden.

Abbildung 2.16 zeigt den Endzustand der Neuronenkette nach vielen Lernschritten. In Abbildung 2.17 wird der Verlauf eines Lernvorgangs der Neuronenkette dargestellt. Die Neuronen sind zu Beginn auf Werte nahe 0 initialisiert, der kleinstmögliche „Lernwiderstand“ wurde gewählt, also die größtmögliche Lernrate. Die Lernzeit wurde so gewählt, daß neben dem Gewinner nur die nächsten Nachbarn am Lernvorgang beteiligt sind. Die zum Training verwendete Musterverteilung besteht aus drei Clustern, die im ersten und letzten Plot durch drei Kreise markiert sind. Wie erwartet, ziehen die 3 Cluster die Prototypen an. Zunächst nähern sich die Prototypen nur 2 Clustern an, bis ein Ende der Kette vom dritten Cluster angezogen wird, und die Neuronen sich auf alle 3 Cluster verteilen. Die jeweils zwischen den Clustern liegenden „toten“ Neuronen sind eine Folge der Nachbarschaftskopplung.

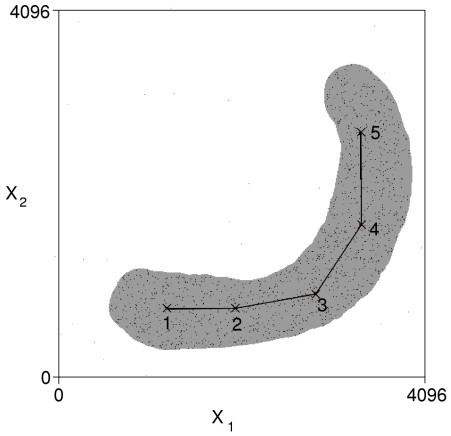


Abbildung 2.16: Zustand der Neuronenkette nach einigen tausend Trainingsschritten. Die Muster werden gleichverteilt aus der grau unterlegten Zone gezogen. Lernparameter (nach Konvergenz): Rampenanstiegsrate $dU_r/dt = 0.25$ V/ms, Lernzeit $T = 64$ µs, Lernwiderstand $R_w = 10$ kΩ, Lernzeitkonstante $R_w C_w = 1$ ms

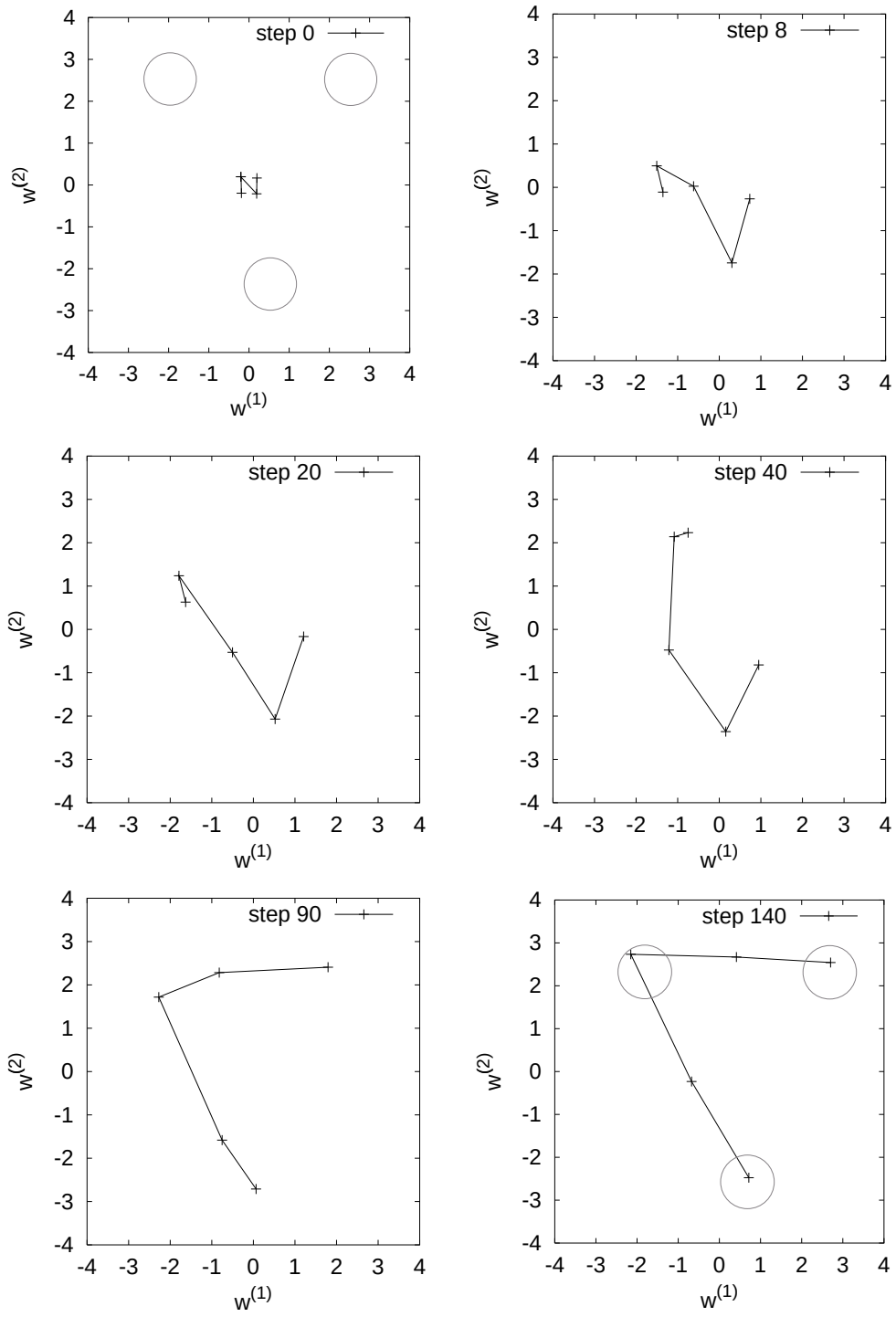


Abbildung 2.17: Ablauf eines Lernvorgangs der neuronalen Hardware, trainiert mit einer aus 3 Clustern bestehenden Mustermenge. Dargestellt sind jeweils die Prototypvektoren im Musterraum, entsprechend der linearen Netztopologie mit einer Linie verbunden. Lernparameter: $C_w = 100$ nF, $R_w = 10$ kΩ, $T = 256$ µs

2.5 Möglichkeiten zur Integration der Kohonen-Hardware

Der hier aufgebaute Prototyp zeigt die prinzipielle Funktionsfähigkeit einer analogen Kohonenhardware. Zur Integration auf einem Chip müssen zunächst die einzelnen Komponenten so weit wie möglich vereinfacht werden.

Das analoge Speicherglied, im Grunde genommen ein Sample-and-Hold-Glied, kann in MOS-Technik einfach aufgebaut werden. Der Speicherkondensator wird dabei durch Gatekapazität eines MOS-Transistors gebildet, der als Sourcefolger gleichzeitig den nachfolgenden Pufferverstärker darstellt. Der Analogschalter wird durch ein CMOS-Transistorpaar oder einen einzelnen Transistor implementiert. Ein mögliches Problem kann der unvermeidliche Leckstrom sein, der durch den gesperrten Analogschalter abfließt. Die Drift des gespeicherten Prototypen wird sich durch die verkleinerten Speicherkapazitäten beschleunigen. Bei den nach diesem Prinzip aufgebauten dynamischen RAM-Zellen verlangt die Drift einen regelmäßigen Refresh nach maximal einigen 10 ms, mit entsprechend großflächigen Speicherkapazitäten sollte eine Speicherzeit im Sekundenbereich machbar sein. Ohne Refresh eignet sich die kapazitive Speicherung der Prototypen also eher für Anwendungen, bei denen ständig nachtrainiert wird, so daß Driftverluste durch Training ausgeglichen werden.

Die Distanzrechenschaltung könnte in vereinfachter Form auf Differenzverstärkern basieren, oder gar auf Einzeltransistoren, etwa in CMOS oder auch in Bipolartechnik. Eine Differenz-Gleichrichterschaltung auf Basis eines Differenzverstärkers ist in Abbildung 2.18 gezeigt (nach Tietze-Schenk [Tie78]). Dabei wird die Eingangsspannungsdifferenz in eine proportionale Stromdifferenz der beiden Kollektorströme durch T1 und T2 umgesetzt. Die beiden Emitterfolger T3 und T4 folgen dem höheren Potential an der Basis, so daß die Ausgangsspannung bis auf einen Offset dem Betrag der Eingangsspannungen entspricht. Auch diese Schaltung kann prinzipiell auf die gleiche Weise in MOS-Technik aufgebaut werden, sogar mit dem Vorteil, daß ein einzelner MOS-Transistor mit kurzgeschlossener Gate-Source-Strecke als Stromquelle genutzt werden kann.

Ein weiterer wichtiger Bestandteil der hier untersuchten Implementation des Kohonenalgorithmus ist die Gewinnersuche. Sowohl zur Gewinnersuche, als auch zur Implementation der Nachbarschaftskopplung wurde hier das aktive Medium benutzt. Zur Gewinnersuche wird dazu die Schaltschwelle des aktiven Mediums als Komparator genutzt. Daher muß die Schaltschwelle mit einer Rampe erreicht werden, da nach Zündung des Mediums die Frontausbreitung einsetzt. Die Zündschwelle muß also, bezogen auf die Zeitskala des Zündprozesses, langsam erreicht werden.

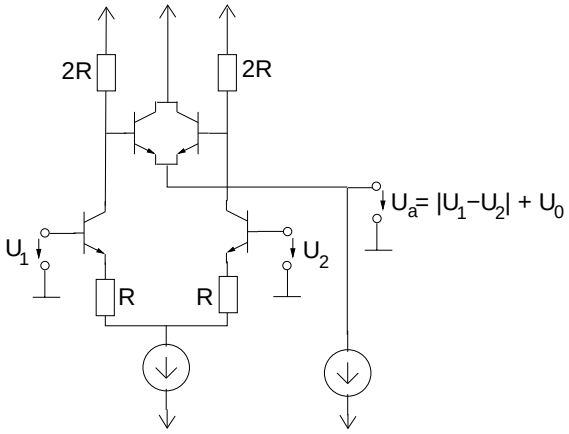


Abbildung 2.18: Differenzgleichrichter auf Basis eines Differenzverstärkers.

Dies erschwert auch die Verwendung eines Intervallschachtelungs-Algorithmus zur Gewinnersuche.

Eine Abwandlung eines Differenzverstärkers eignet sich jedoch als direkte Winner-Take-All-Schaltung oder Maximumdetektor (Abbildung 2.19). Die Schaltung beruht auf der Kopplung der Emitter und dem festen Emittergesamtstrom. Die exponentielle I_c-U_{be}-Kennlinie bewirkt, daß praktisch der gesamte Emitterstrom von dem Transistor getragen wird, der die größte Basisspannung erhält, sofern die Differenzen der Eingangsspannungen ausreichend groß sind. Da der Kollektorstrom um den Faktor e zunimmt, wenn U_{be} um $U_T = kT/q$ [6] zunimmt, gibt U_T das Maß für die erforderlichen Eingangsdifferenzen an. Ist eine größere Empfindlichkeit gefordert, könnte man die Schaltung kaskadieren, jedoch können Abweichungen der Basisspannungen der einzelnen Transistoren zu einem Offset führen, ganz so wie man es bei Operationsverstärkern feststellt.

Diese WTA-Schaltung alleine ohne die Kopplung durch ein bistabiles Medium kann zum Bau eines analogen Vektorquantisierers dienen, indem die WTA-Ausgänge direkt den Lernvorgang der Neuronen steuern.

Eine integrierte WTA-Schaltung in CMOS-Technik wird u.a. [Mos96] beschrieben. Sie dient als Teil der Auswerteelektronik eines PET-Detektors[7], und kann innerhalb von 50 ns den „Gewinner“ von 16 Eingangskanälen bestimmen, sofern eine Spannungsdifferenz von mindestens 20 mV zum nächst kleineren Eingang vorhanden ist.

Durch Begrenzung der Kollektorströme kann die WTA-Schaltung so erweitert werden, daß nicht nur der größte, sondern die n größten Eingänge selektiert werden.

[6] $U_T \approx 25mV$ bei $T = 300K$

[7] Positronen-Emissions-Tomographie

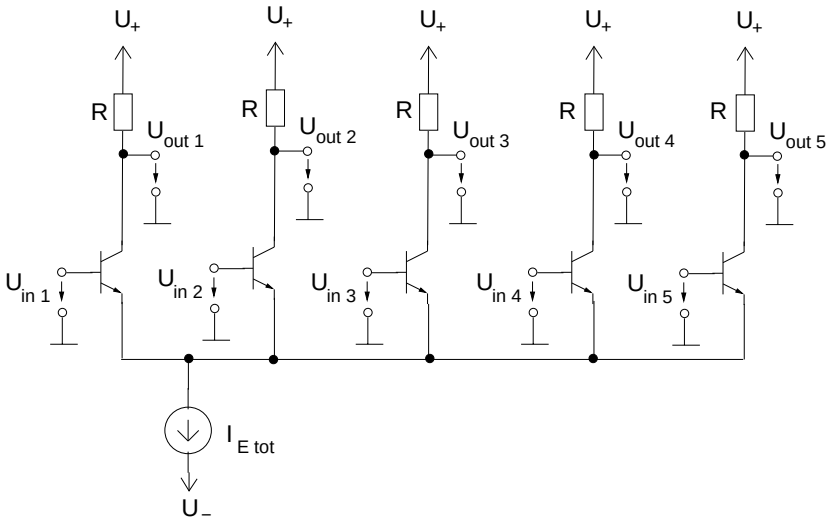


Abbildung 2.19: Winner-Take-All-Schaltung.

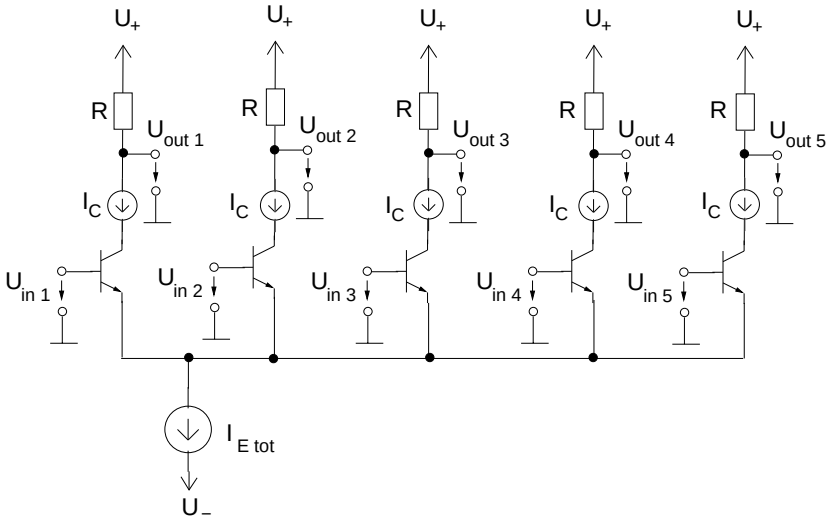


Abbildung 2.20: modifizierte WTA-Schaltung zur Detektion der n größten Inputs.

Dazu werden die Kollektorströme der einzelnen Transistoren durch je eine Stromquelle begrenzt, und der Gesamtstrom $I_{\mathrm{E\,tot}}$ auf $n \cdot I_C$ festgelegt, so daß n Transistoren durchschalten müssen, um den Gesamtstrom zur Verfügung zu stellen. Diese „n-WTA-Schaltung“ kann zur Implementation des Neural-Gas-Algorithmus dienen, da bei diesem nicht nur das Gewinnerneuron, sondern auch im Musterraum benachbarte Neuronen lernen.

Es gibt jedoch auch bei dieser Schaltung keine Kopplung der Neuronen im Kortex. Durch Kombination der obigen WTA-Schaltung mit einem bistabilen Medium, das von dem einzigen „eingeschalteten“ WTA-Ausgang gezündet wird, läßt sich eine SOM-Hardware aufbauen, die ganz auf den rampengesteuerten Vergleichsvorgang verzichten kann. Die Zeit zur Gewinnersuche hängt nur von den RC-Zeitkonstanten in der WTA-Schaltung ab, während der Lernvorgang mit der abstandsabhängigen Lernrate durch Frontausbreitung gesteuert wird.

Die Miniaturisierung der Thyristorstruktur mit ihrer doppelten Funktion als Schwellwertdetektor und bistabilem Medium mit Frontausbreitung wird in Abschnitt 4 untersucht. Zur Vermeidung überflüssiger Verbindungen müßte die Kopplungsstruktur in Form eines Gitters realisiert werden, die Neuronen selbst liegen dann in den Lücken des Gitters, so daß jedes Neuron direkt mit einem Gatekontakt auf dem Gitter verbunden ist. So entspricht das Chiplayout direkt der Topologie einer zweidimensionalen SOM.

2.6 Vorteile analoger neuronaler Hardware

Die hier beschriebene analoge Implementation eines neuronalen Netzes bietet einige Vorteile: Neuronale Netze lassen sich in analog-integrierter Technik platz- und leistungssparender realisieren als auf digitale Weise [Mea89]. Jedes Neuron besteht aus wenigen Transistoren, Signale können durch Spannungen oder Ströme codiert werden, wobei die Stromdarstellung sich besonders zur analogen Addition eignet. Lernfähige Gewichte können durch Sample-and-Hold-Glieder, also Speicherkondensatoren, für langfristige Speicherung eventuell mit Floating-Gate-Transistoren realisiert werden. Die Größe der Lernschritte ist nicht durch Quantisierungseffekte begrenzt, was in einer digitalen Implementation bei Verwendung von Integerarithmetik zu geringer Bittiefe dazu führen kann, daß kleine Lernraten unmöglich werden.

Jedoch ist die Rechengenauigkeit der Neuronen bei Analoger Implementation geringer (im Bereich einiger Prozent [Bal95, eta92]). Die in der SOM nötige Abstandsrechenschaltung weist durch die Toleranzen der verwendeten Widerstände eine Ungenauigkeit auf. Relative Unterschiede der einzelnen Neuronen führen zu

unterschiedlicher Emfindlichkeit bei der Gewinnersuche, dadurch können die Einzugsgebiete der einzelnen Neuronen von der idealen Größe abweichen. Beispielsweise kann ein Neuron, das seinen Abstand zum Mustervektor zu klein einschätzt, eine größere Voronoizelle beanspruchen. Solche Fehler wurden in der diskret aufgebauten SOM-Hardware beobachtet (s. Abschnitt 2.4.1), sollten aber wegen der guten Homogenität heutiger Fertigungsprozesse integrierter Schaltungen in einer integrierten Version nicht zusehr ins Gewicht fallen.

Andererseits zeichnen sich gerade die neuronalen Netze zeichnen sich jedoch durch große Fehlertoleranz aus, insbesondere im Grenzfall sehr vieler Neuronen. Sogar der Ausfall eines Neurons der SOM würde ohne große Folgen bleiben, die jeweiligen Nachbarn würden dessen rezeptives Feld im Musterraum unter sich aufteilen. Sollte die Frontausbreitung an einem Neuron lokal behindert werden (Pinning), so kann die Front zumindest in einem zweidimensionalen Kortex die Störstelle umlaufen. Somit sollte eine große selbstorganisierende Karte selbst auf den Ausfall einzelner Neuronen unkritisch reagieren.

Eine gute Analogie ist das Nervensystem von Tier und Mensch, das zuverlässig die gelernten Aufgaben erfüllt, obwohl die einzelnen Neuronen im Prinzip sehr unzuverlässige Bauteile sind, von denen immer wieder einige ausfallen, ohne das die Gesamtfunktion wesentlich beeinträchtigt wird.

Ein nicht von der Hand zu weisender Nachteil einer analogen Implementation ist die Drift der analogen Speicher. Diese dürfte sich in einer integrierten Version wegen der drastisch verkleinerten Speicherkapazität noch verstärken. Zur Behebung des Driftproblems sind analoge Speicherzellen mit periodischem Refresh und quantisierten Werten entwickelt worden [Mac90]. Sie arbeiten auf folgende Weise: Angenommen, der Speicherinhalt sinke durch die Drift kontinuierlich ab. Der Wert einer Zelle wird nun von Zeit zu Zeit gelesen, dabei quantisiert, wobei zum nächsthöheren Wert aufgerundet wird, und zurückgeschrieben. Geschieht dies schnell genug, so wird immer wieder der ursprüngliche Wert zurückgewonnen.

Für die langfristige Speicherung von Prototypen und Gewichten kann ein Floating-Gate-Transistor verwendet werden, also eine EEPROM-artige Speicherzelle. Die Ladung auf dem Floating-Gate bestimmt die Schwellenspannung dieses Transistors. Die gute Isolation des Floating-Gates ermöglicht Speicherzeiten von mehreren Jahren, der Schreibvorgang geschieht durch Überwindung der Oxidisolation mit Hilfe des Fowler-Nordheim-Tunnelns. Solche EEPROM-Zellen sind neben der üblichen Anwendung als digitale nichtflüchtige Speicher auch schon als Analogspeicher [Har98] verwendet worden, etwa um in integrierten Analogschaltungen externe Abgleichwiderstände zu ersetzen, und eben auch als adaptive Synapsen in neuronaler Hardware [Vit91]. Allerdings ist der Schreibvorgang extrem nichtlinear, so daß die Implementation des Lernvorgangs erheblich komplizierter wird.

Berücksichtigt man alle diese Gesichtspunkte, findet man, daß die Implementation eines Kohonen-Netzes mit Hilfe digitaler Logik möglich ist [Rüp95], der analoge Aufbau eines solchen Netzes jedoch wegen der Anwendung einer Vielzahl von physikalischen Prinzipien zur Verkörperung der Neuronen wesentlich bestechender, aber auch wegen der Vielzahl von Störeinflüssen um einiges schwieriger und ungewöhnlicher ist.

Kapitel 3

Frontausbreitung in Thyristorstrukturen

Dieses Kapitel widmet sich den experimentellen und numerischen Untersuchungen von Frontausbreitungsphänomenen in großflächigen Thyristoren. Nach einer theoretischen Betrachtung der Grundlagen der Bistabilität und der Schaltfrontausbreitung in diesen Bauelementen werden die Ergebnisse von elektrischen und infrarot-optischen Messungen an verschiedenen Thyristorproben vorgestellt. Die gefundenen Schaltfronten erlauben den Einsatz einer solchen Struktur zur Kopplung eines neuronalen Netzes. Zuletzt werden Erfordernisse für eine Miniaturisierung der Kopplungsstruktur erläutert.

3.1 Struktur eines Thyristors

Ein Thyristor ist ein Halbleiterbauelement mit einer pnpn-Struktur, die gewöhnlich durch mehrere Diffusionsschritte aus einer Siliziumscheibe hergestellt wird. Die äußeren p- und n-Schichten bilden die Anode und Kathode, die mittleren Schichten werden als n- und p-Basis bezeichnet. Die Herstellung der pnpn-Struktur geschieht in zwei Diffusionsschritten: Zunächst werden Anode und p-Basis durch eine p-Diffusion von beiden Seiten der Halbleiterscheibe her erzeugt; dies kann in einem Diffusionsschritt geschehen. Dann wird die Kathode durch eine flache n-Diffusion erzeugt. Aus diesem Herstellungsprozeß resultiert die Asymmetrie der Thyristorstruktur: Die n-Basis stellt das Gebiet zwischen den eindiffundierten p-Zonen dar, und ist damit relativ dick (einige 100 µm) und niedrig dotiert. Die p-Basis hingegen weist eine höhere und nach innen abfallende Dotierungskonzentration auf, bei wesentlich geringerer Dicke.

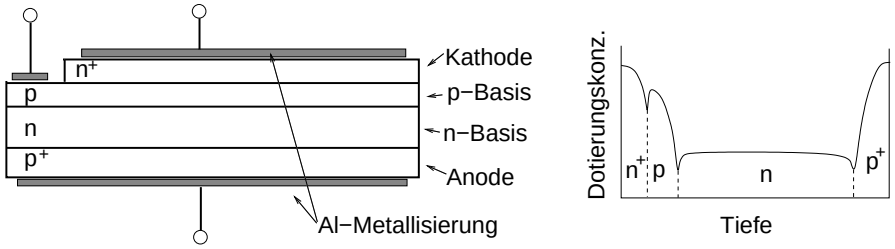


Abbildung 3.1: Profilskizze eines großflächigen Thyristors.

Anode und Kathode werden großflächig mit einer Metallschicht (z.B. Al) kontaktiert. Die p-Basis wird lokal kontaktiert, indem eine Vertiefung in die Kathodenschicht geätzt wird, oder indem in Planartechnik die Basiskontaktfläche bei der Kathodendiffusion ausgespart wird. Dieser Basis- oder Gatekontakt dient als Steueranschluß.

Thyristoren dienen in der Halbleitertechnik als Leistungsschalter. Der Laststrom fließt dabei über Anode und Kathode, während das Gate zur Steuerung dient. Der Thyristor verhält sich wie eine Diode, wobei er jedoch in Flußrichtung gepolt zunächst hochohmig ist. Erst durch einen Gatestromimpuls wird er in den niederohmigen Zustand geschaltet. Danach ist der Gateanschluß wirkungslos, erst nachdem der Laststrom unter einen Schwellenwert sinkt oder die Anodenspannung umgepolt wird, schaltet der Thyristor wieder in den hochohmigen Zustand. Ein Bauteil kann typischerweise eine Fläche von wenigen mm^2 bis zu einem ganzen Wafer einnehmen. Deswegen werden Thyristoren als Leistungsschalter für höchste Leistungen eingesetzt, für Spannungen bis ca. 6000 V und Ströme von mehreren 1000 A.

Hier soll jedoch ein anderer Aspekt betrachtet werden, nämlich der Zündvorgang selbst, insbesondere die Ausbreitung der Schaltfront, die zum Zeitpunkt des Gatestromimpulses am Ort des Gatekontakts beginnt und sich dann über die gesamte aktive Fläche ausbreitet und sie in den leitenden Zustand schaltet.

3.2 Die statische I-U-Kennlinie

Wir betrachten zunächst das Verhalten einer quasi eindimensionalen pnpn-Struktur, vernachlässigen also Effekte, die mit der lateralen Ausdehnung der Struktur zusammenhängen.

Die in Abbildung 3.2 gezeigte statische Kennlinie läßt sich folgendermaßen erklären [Ger79]:

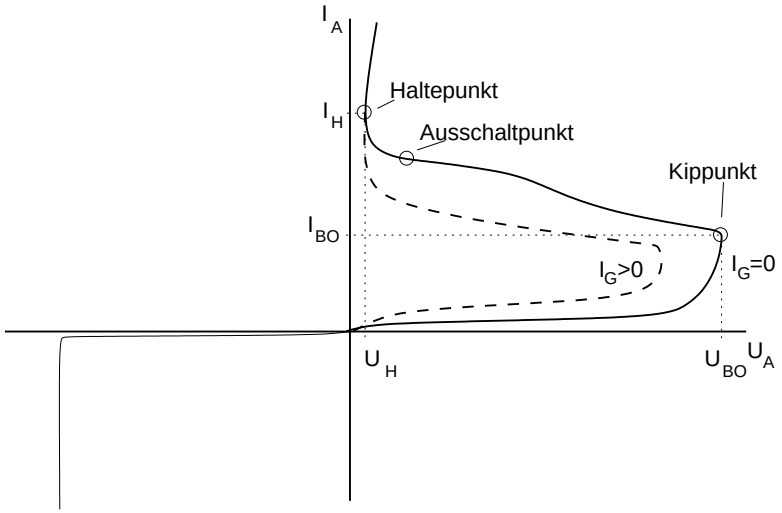


Abbildung 3.2: Kennlinie eines Thyristors mit ausgezeichneten Punkten.

Im 3. Quadranten der Kennlinie, also bei negativ gepolter Anode, werden die beiden äußeren pn-Übergänge J1 und J3 in Sperrichtung belastet, J2 ist in Flußrichtung gepolt. Da die p-Basis relativ hochdotiert ist, hat J3 eine relativ niedrige Durchbruchspannung U_{BR3} von wenigen Volt. Somit trägt J1 nahezu die gesamte Sperrspannung. Die Sperrkennlinie wird durch den Transistor T2 mit offener Basis bestimmt. Bei steigender Sperrspannung wächst sowohl die Größe der Raumladungszone von J1 als auch die maximale Feldstärke in dieser Raumladungszone. Der Stromanstieg beim Erreichen der Durchbruchspannung entsteht entweder durch Lawinendurchbruch in der Raumladungszone von J1, oder durch den Punch-Through-Effekt: Die Raumladungszone von J1 füllt die ganze n-Basis aus und erreicht die p-Basis. Zur Erzielung einer hohen Sperrspannung muß also zumindest eine Basis niedrig dotiert sein,um den Lawinendurchbruch zu vermeiden, und die Basisweite muß zur Unterbringung der Raumladungszone groß genug sein. Dies ist der Grund für die breiten und schwach dotierten n-Basen von Hochspannungsthyristoren.

Im 1. Quadranten hat die Kennlinie einen S-förmigen Verlauf. Sie wird an den Punkten mit vertikaler Steigung unterteilt in die Sperrkennlinie bei niedrigen Strömen und hohen Spannungen, einen Übergangsbereich mit negativ differentiellem Widerstand, und die Durchlaßkennlinie bei hohen Strömen und niedriger Spannung. Für eine Ohm'sche Last ergeben sich entsprechend der Lastgeraden-Konstruktion zwei stabile Zustände im Sperr- und Durchlaßbereich, sowie ein instabiler Zustand im Übergangsbereich.

Plausibel machen läßt sich diese Bistabilität durch Shockleys Zwei-Transistor-

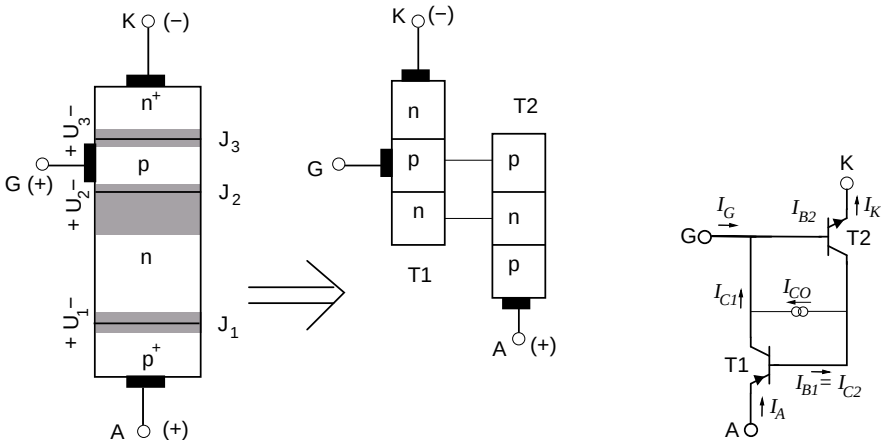


Abbildung 3.3: Shockleys Zwei-Transistor-Analogon.

Analogon (s. Abbildungen 3.3 und 3.4): Im sperrenden Zustand sind die Emitter J1 und J3 in Durchlaßrichtung gepolt, der Kollektor J2 ist in Sperrichtung gepolt. Die Transistoren T1 und T2 haben die Stromverstärkungsfaktoren α_1 und α_2. Die Kirchhoff'schen Regeln ergeben dann die Strombilanz:

$$I_A = \alpha_1 I_A + \alpha_2 I_K + I_{CO2} \tag{3.1}$$

Dabei ist I_{CO2} der von U_2 abhängige Sperrstrom von J2. Mit $I_K = I_G + I_A$ ergibt sich folgender Zusammenhang:

$$I_A = \frac{\alpha_2 I_G + I_{CO2}}{1 - (\alpha_1 + \alpha_2)} \tag{3.2}$$

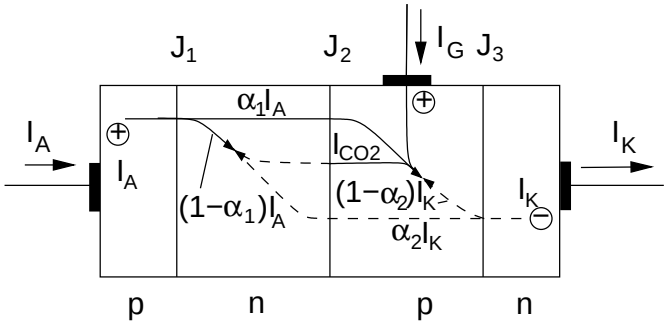


Abbildung 3.4: Einzelströme im Thyristor.

Die Stromverstärkungen $\alpha_{1,2}$ sind nun selbst vom Emitterstrom abhängig. Die Stromverstärkung α ist ja definiert als $\alpha = I_C/I_E$, α ist also der Anteil der vom Emitter emittierten Ladungsträger, die die Basis durchqueren und den Kollektor erreichen. Dieser Anteil liegt bei ausreichendem Strom nahe 1, sinkt aber bei kleinen Strömen durch Rekombinationsverluste in der Basisschicht ab. Sofern der Anodenstrom so gering ist, so daß $\alpha_1(I_A)+\alpha_2(I_A) < 1$ erfüllt ist, bleibt der Thyristor im Sperrzustand. Wird der Anodenstrom erhöht, so steigen die Stromverstärkungen an, bis $\alpha_1 + \alpha_2 = 1$ erreicht wird. Damit divergiert I_A in Gl. 3.2, und der sperrende Zustand wird instabil.

Diese Zündbedingung wird erreicht, wenn die Anodenspannung die *Kippspannung* U_{BO} überschreitet. Bei Erhöhung von U_A wächst der Sperrstrom I_{CO2} an, wodurch die Stromverstärkungen $\alpha_{1,2}$ steigen bis die Zündbedingung erfüllt ist. Die Einprägung eines Gatestromes I_G trägt – verstärkt durch den Transistor T2 – ebenfalls zur Erhöhung des Gesamtstromes bei, so daß sich auch auf diese Art die Zündung herbeiführen läßt.

Die Zündung führt durch die regenerative Stromverstärkung der beiden Transistoren zum Anwachsen des Anodenstromes, während die Anodenspannung durch den Spannungsabfall am Lastwiderstand bis auf einen Wert von etwa 1 V fällt. Dann sind beide Transistoren im Sättigungsbetrieb, so daß ihre Stromverstärkung nun begrenzt wird. Der Kollektor J2 ist in Durchlaßrichtung gepolt, und die Anodenspannung ist die Summe der Spannungsabfälle an J1, J2 und J3. U_2 ist nun negativ, so daß die Summe U_A nur wenig größer als der Spannungsabfall an einer Diode ist. Die beiden Basen sind mit Minoritätsträgern überschwemmt, was bei hoher Stromdichte zumindest für die schwach dotierte n-Basis zum Fall *starker Injektion* führt, d.h. die Elektronen- und Löcherkonzentration übersteigen die Grunddotierung, so daß wegen der Neutralitätsbedingung deren Konzentrationen praktisch gleich werden. Dann sind die Dotierungen der Basen unerheblich gegen die injizierten Ladungsträger, und die Struktur verhält sich wie eine pin-Diode. Sinkt der Anodenstrom nun wieder, so sinken auch die Stromverstärkungen, bis $\alpha_1 + \alpha_2 < 1$ gilt, und der eingeschaltete Zustand instabil wird.

Für die Bistabilität des Thyristors ist also entscheidend, daß für kleine Anodenströme $\alpha_1+\alpha_2 < 1$ gilt, für größere jedoch $\alpha_1+\alpha_2 > 1$. Eine stark stromabhängige Stromverstärkung der beiden Teiltransistoren ist also für das Schaltverhalten nötig.

3.3 Schaltvorgänge

Durch den S-förmigen Verlauf der I-U-Kennlinie ergeben sich für eine Ohm'sche Last gewöhnlich 2 stabile Fixpunkte und ein instabiler Fixpunkt (s. Abb. 3.5, Lastgerade zu U_{V2}).

Befindet sich das System auf dem Niedrigstromast der Kennlinie, so kann dieser Zustand durch Erhöhung der Versorgungsspannung destabilisiert werden. Die Lastgerade verschiebt sich nach rechts, bis sie die Kennlinie nicht mehr berührt. Hier findet eine Sattel-Knoten-Bifurkation statt, der stabile Zustand auf dem Niedrigstromast der Kennlinie verschwindet, und entsprechend der Reaktionsfunktion wechselt das System auf den einzigen verbliebenen stabilen Fixpunkt auf dem Hochstromast der Kennlinie.

Wir nun die Versorgungsspannung gesenkt, wandert das System den Hochstromast hinunter, bis die Lastgerade von der Kennlinie „abreißt", und das System auf den verbliebenen Fixpunkt auf dem Nidrigstromast der Kennlinie wechselt.

Der Einschaltvorgang kann ebenso durch Einprägung eines ausreichend hohen Gatestromes ausgelöst werden, wodurch sich der Kipppunkt der Kennlinie nach links verschiebt.

Der Schaltvorgang weist den typischen *bottleneck* einer Sattel-Knoten-Bifurkation auf [Str94]: Die Schaltzeit wird umso länger, je näher die Lastgerade der stabile

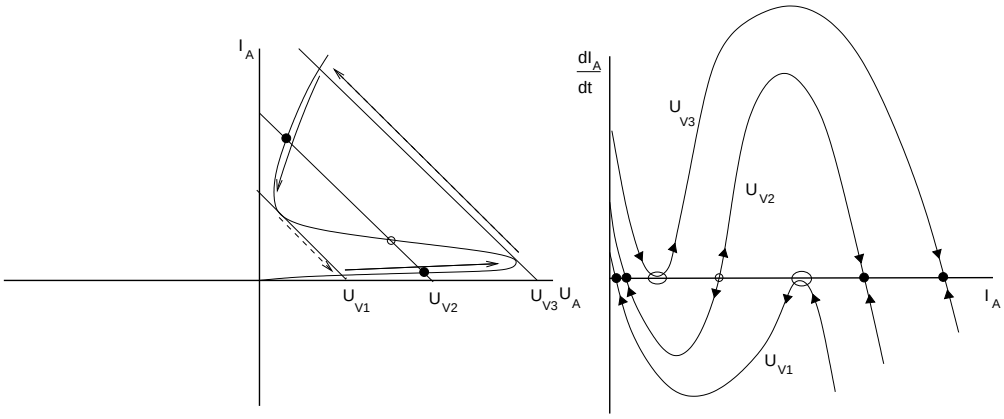


Abbildung 3.5: U-I-Kennlinie mit Lastgerade links, Zustandsdynamik für verschiedene Lastgeraden rechts. Stabile Punkte sind mit gefüllten Kreisen gekennzeichnet, instabile mit offenen Kreisen. Die ovalen Markierungen zeigen die Sattel-Knoten-Bifurkationen der Schaltpunkte an. Die Dynamik an diesen Punkten weist die typische kritische Verlangsamung eines *Bottleneck* auf.

Kennlinie nach der Veränderung der Parameter noch ist. Ein großer Zündstrom führt also i.A. zu schnellem Durchschalten. Ein gerade ausreichender Zündstrom führt umgekehrt zu einem sehr langen *Zündverzug*.

Die Zeitskala, auf der das Umschalten stattfindet, hängt hauptsächlich von der Laufzeit der Träger von der einen zur anderen Basis ab. Es findet ja eine wechselseitige Injektion von Trägern statt, die für den Transfer zur anderen Basis jeweils eine Laufzeit benötigen. So nimmt die Schaltzeit mit steigender Schichtdicke zu, und mit steigender Stromdichte, also kleinerem Lastwiderstand ab.

Beim Ausschaltvorgang müssen die in den Basen vorhandenen Minoritätsträger abgebaut werden, was sowohl durch Abtransport als auch durch Rekombination geschieht. Somit hat die Lebensdauer der Minoritätsträger in den Basen einen entscheidenden Einfluß auf die Abschaltzeit. Zur Verkürzung der Abschaltzeit werden Thyristoren mit energiereichen Teilchen bestrahlt, oder es wird Gold eindiffundiert. Diese Maßnahmen erzeugen zusätzliche Rekombinationszentren, so daß die Lebensdauer sinkt. Der Nachteil dieser Maßnahme ist jedoch die Erhöhung des Spannungsabfalles im eingeschalteten Zustandes, die erhöhte Rekombination läßt ja im eingeschalteten Zustand auch die injizierte Ladungsträgerdichte in den Basen sinken.

3.4 Ausgedehnte pnpn-Strukturen

Bei großflächigen Thyristoren ist die laterale Ausdehnung des Bauteils gegenüber der Dicke nicht mehr vernachlässigbar. Der Gatestrom wird lokal an einem Kontaktpunkt in die p-Basis eingeprägt, und fließt in der Umgebung dieses Kontakts über den pn-Übergang J3 zur Kathode ab. Es ergibt sich eine Gatestromverteilung mit einer typischen Ausdehnung im mm-Bereich, abhängig von dem zur Zündung nötigen Gatestrom und der Querleitfähigkeit der p-Basisschicht. Bei ausreichendem Gatestrom zündet der Thyristor daher zunächst lokal am Ort des Gatekontakts.

Im gezündeten Gebiet ist das Potential in der p-Basis erhöht gegenüber dem Ruhezustand, so daß an der Grenze zum ungezündeten Gebiet ein Querfeld in der Basis herrscht. Dieses Feld treibt einen Basisstrom in das noch ungezündete Gebiet, der dort nach einer Schaltzeit ebenfalls die Zündung auslöst (Feldmodell). So breitet sich die Zündung als Schaltfront über die bauteilfläche aus. Die Stärke der lateralen Kopplung wird bestimmt durch die Leitfähigkeit der Basisschichten und die zur Zündung notwendige Stromdichte. Im Normalfall hat die p-Basis durch ihre höhere Dotierung eine größere Leitfähigkeit als die n-Basis, so daß sie den größeren Einfluß auf die Frontgeschwindigkeit und -breite haben sollte.

Weiterhin ist mindestens die n-Basis im gezündeten Zustand mit Ladungsträgern beider Polaritäten überschwemmt, man kann also von einem Plasmazustand sprechen. Die Trägerdichten sind gegenüber dem ausgeschalteten Zustand stark erhöht. Diese Träger diffundieren nun entsprechend ihrem Dichtegradienten in den noch ungezündeten Bereich (Diffusionsmodell). Auch dies ist ein Modell zur Erklärung der Schaltfrontausbreitung. Nach [D'y77] dominiert das Feldmodell die Frontausbreitung, der Einfluß des Diffusionsmodells sei dagegen vernachlässigbar.

In der Leistungselektronik führt insbesondere eine zu langsame Zündfrontausbreitung zu Problemen: Sie begrenzt die nach der Zündung zulässige Stromanstiegsgeschwindigkeit dI/dt. Steigt der Laststrom zu rasch an, so kann die noch kleine eingeschaltete Teilfläche des Thyristors überlastet werden, was zur Zerstörung führen kann. Daher soll die Ausbreitung der Zündfront möglichst rasch und gleichmäßig geschehen. Zu diesem Zweck wurden z.B. fingerartig über den Thyristorwafer verteilte Gatekontakte entwickelt, so daß der von der Front zurückzulegende Weg minimiert wird. Der nun erhöhte Bedarf an Zündstrom wird von einem Hilfsthyristor gedeckt [Ger79].

Für die in dieser Arbeit diskutierte Anwendung als Kopplungsstruktur für neuronale Netze ist die Frontausbreitung entscheidend. Wichtig ist dabei ein möglichst scharfer Übergang zwischen den beiden Zuständen, also eine geringe Frontbreite, sowie eine dem Timing des neuronalen Netzes angemessene Ausbreitungsgeschwindigkeit.

3.5 Modell der Zündausbreitung auf Basis eines RD-Systems

Die Propagation der Zündfront in einem Thyristor kann mit diversen Modellen unterschiedlicher Komplexität beschrieben werden, angefangen bei einem rigorosen Drift-Diffusions-Modell mit detailliert definiertem Dotierungsprofil, das sich nur numerisch simulieren läßt. Dieser Weg wird in Kapitel 4 eingeschlagen, um die Frontausbreitung in konkret definierten Thyristorstrukturen für die Anwendung in neuronalen Netzen zu simulieren und quantitative Aussagen über Frontgeschwindigkeiten und Frontbreiten machen zu können.

Ein einfacheres Modell auf der Basis eines einkomponentigen Reaktions-Diffusions-Systems erlaubt eher qualitative Aussagen über die Form und Geschwindigkeit der Schaltfronten, es kann aber wesentlich einfacher simuliert werden. Die hier verwendeten Modellgleichungen werden in [Mei97, Mei98a, Mei98b] zur Untersuchung von gategesteuerten Zündfronten in Thyristoren eingesetzt. in solches RD-System hat

neben den trivialen homogenen Lösungen Frontlösungen, wobei diese sich mit einer bestimmten Geschwindigkeit bewegen können, so daß einer der beiden stabilen Zustände sich auf Kosten des anderen ausdehnt [Bod93, Fif79].

$$\tau \frac{\partial V(x,t)}{\partial t} = l^2 \frac{\partial^2 V(x,t)}{\partial x} + f(V, U_A, U_G) \tag{3.3}$$

$$f(V) = -\alpha V + \exp(V) - \beta \exp(2V - U_A) + \gamma U_A + \kappa U_G \tag{3.4}$$

$$I_A(x,t) = I_s(e^{V(x,t)} - 1) \tag{3.5}$$

Das Ersatzschaltbild in Abb. 3.6 zeigt das Prinzip des Modells, das eine zweidimensionale pnpn-Struktur als eine Kette von eindimensionalen nichtlinearen Elementen betrachtet, die durch Ohm'sche Widerstände gekoppelt sind. Die einzige Variable des Modells ist der Spannungsabfall V zwischen p-Basis und Kathode. Die Dynamik der Einzelelemente wird durch die Reaktionsfunktion $f(V, U_A)$ und die Zeitableitung $\tau \frac{\partial V(x,t)}{\partial t}$ behandelt: Die Reaktionsfunktion beschreibt den bei einer bestimmten Basisspannung V entstehenden Stromüberschuss der p-Basis. Es fließt ja einerseits Strom zur Kathode hin ab, andererseits aber auch über den pnp-Transistor von der Anode nach. Abhängig von V ist diese Strombilanz positiv oder negativ. Die Kapazität der p-Basis bewirkt dann eine zu f proportionale Änderung $\frac{\partial V(x,t)}{\partial t}$.

Die einzelnen Terme der Reaktionsfunktion lassen sich folgendermaßen plausibel machen: Der Term $\exp(V)$ ist proportional zum Basisstrom des npn-Transistors T2. Dieser Basisstrom erzeugt über die beiden Transistoren T2 und T1 einen Netto-Stromzufluß von der Anode zur p-Basis. Diese Betrachtung geht von konstanten Stromverstärkungen $\alpha_1 + \alpha_2 > 1$ und einer exponentiellen I-U-Kennlinie der Basis-

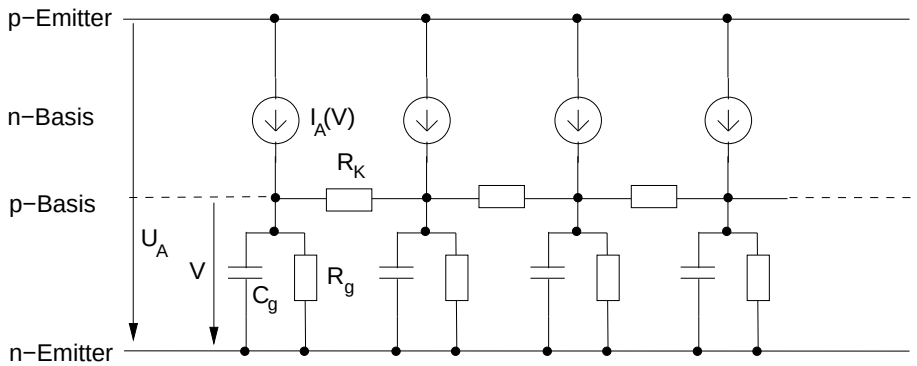
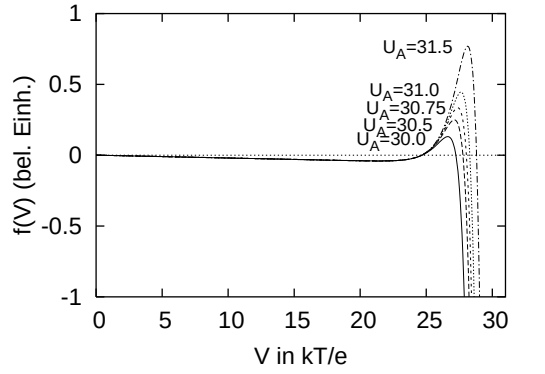
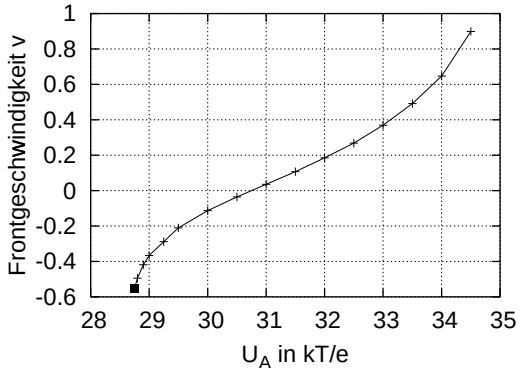


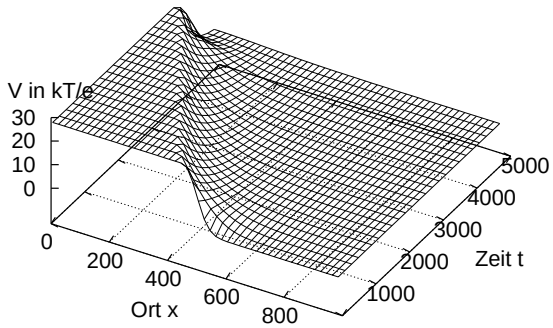
Abbildung 3.6: Ersatzschaltbild des einkomponentigen Thyristormodells



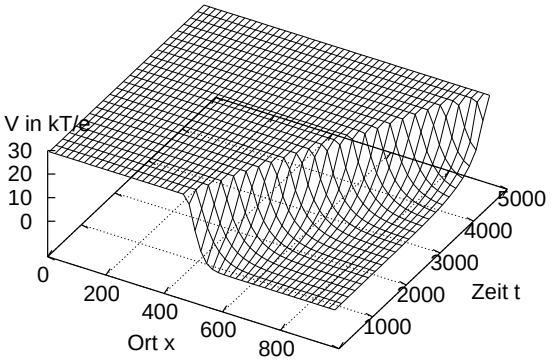
Reaktionsfunktion $f(V, U_A)$



Frontgeschwindigkeit $v(U_A)$



Rückwärts laufende Front
$U_A = 30.0$

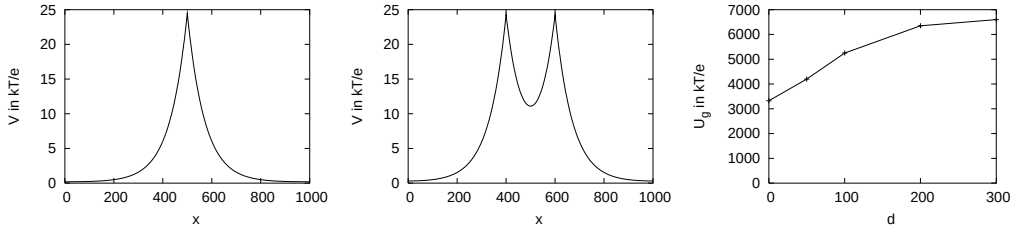


Vorwärts laufende Front
$U_A = 31.75$

Parameter:

α	$2 \cdot 10^{-3}$
β	$1.45 \cdot 10^{-11}$
γ	$1 \cdot 10^{-5}$
κ	$1 \cdot 10^{-3}$
U_g	0
l^2	10
τ	1

Abbildung 3.7: Fronten im Thyristor-RD-Modell.Oben links ist die Reaktionsfunktion für verschiedene Anodenspannungen U_A gezeigt. Bei $U_A = 30.75$ sind die Flächen unter und über der Kurve gleich, die Frontgeschwindigkeit v wird Null. Unterhalb von $U_A = 29.8$ besitzt $f(V)$ nur eine Nullstellen, daher verschwindet dort die Frontlösung, und der Graph der Frontgeschwindigkeit endet dort.



Ein Gatekontakt, krit. Zündstrom, $d = 0$

Zwei Gatekontakte, krit. Zündstrom, $d = 200$

Kritischer Reiz U_g gegen Kontaktabstand d

Abbildung 3.8: Überlappung der Gatepotentiale erzeugt durch Zündstrom an zwei Gatekontakten, mit Absenkung des kritischen Zündstromes. Dargestellt ist jeweils die Potentialverteilung $V(x)$, wobei U_G an einem, bzw. zwei Punkten erhöht wurde, bis die Zündung einsetzt. Rechts ist die Abhängigkeit des kritischen Gatestroms vom Abstand der Einprägepunkte dargestellt.

Emitter-Diode J1 aus. Um den ausgeschalteten Zustand zu stabilisieren, sind zusätzliche Verlustterme nötig. Der lineare Term $-\alpha V$ beschreibt Ohm'sche Verluste in Form eines Leckwiderstandes R_b. γU_A beschreibt den Leckstrom des Kollektors J2 durch einen Ohm'schen Leckwiderstand. $-\beta \exp(2V - U_A)$ stellt den exponentiell wachsenden Strom über den in Flußrichtung gepolten Kollektorübergang dar, der den Strom im eingeschalteten Zustand begrenzt. κU_G beschreibt einen externen Gatestrom durch eine orts- und zeitabhängige Steuerspannung U_G. Die laterale Kopplung der Einzelelemente wird durch den Diffusionsterm beschrieben.

Die hier verwendete Reaktionsfunktion hat bei ausreichend großer Anodenspannung U_A drei Nullstellen, so daß das System zwei stabile Zustände hat und Frontlösungen existieren. Die *Equal-Area-Rule* bestimmt, welcher Zustand dominant ist, also in welche Richtung sich die Front bewegt (s. Anhang A).

Die laufenden Fronten in Abb. 3.7 wurden numerisch mit dem in Anhang A beschriebenen impliziten Verfahren berechnet. Das eindimensionale Grundgebiet wurde gleichmäßig diskretisiert, Dirichlet'sche Randbedingungen ($\partial V/\partial x = 0$) entsprechen einem offenen Abschluß der Basisschicht. Die Simulation beginnt mit einen stufenförmigen Verlauf des Gatepotentials $V(x)$, das nach wenigen Zeitschritten zur Gleichgewichtsform der Front relaxiert. Die asymmetrische Form der Front entsteht durch die Asymmetrie der Reaktionsfunktion, deren positiver Teil wesentlich schmaler und zugleich höher ist als der negative. Die regenerative Verstärkung des Hochstromzustandes geschieht also nur in einem kleinen Bereich nahe dem oberen Ende der Front. Die Verluste, die den ausgeschalteten Zustand stabilisieren, sind dagegen über einen breiten Bereich verteilt. Gerät die Front im Verlauf der Simulation in die Nähe des Randes des Grundgebiets, so macht sich eine anziehen-

de Wechselwirkung bemerkbar: Die Bewegung der Front wird beschleunigt. Der Grund dafür ist die Dirichlet'sche Randbedingung, die dem Gatepotential $V(x)$ am Rand eine horizontale Steigung aufzwingt. Damit wird $V(x)$ am Rand im Falle einer Einschaltfront angehoben, die Beiträge der Reaktionsfunktion steigen. Die Dirichlet'sche Randbedingung entspricht einer achsensymmetrisch am Rand gespiegelten Front, so daß zwei aufeinander zu laufende Fronten die gleiche Anziehung aufweisen, wenn sie sich einander nähern.

Zur Bestimmung der Frontgeschwindigkeit v wird die Simulation über viele Zeitschritte mit konstanten Parametern fortgesetzt. Die Frontbewegung wird dabei durch eine entsprechende Translation des Grundgebietes kompensiert, so daß die Front in der Mitte des Grundgebietes fixiert wird. Die so bestimmte Abhängigkeit der Frontgeschwindigkeit von der Anodenspannung ist in Abbildung 3.7 dargestellt. Mit steigender Anodenspannung U_A wächst die Fläche unter dem positiven Teil der Reaktionsfunktion, so daß die Frontgeschwindigkeit zunimmt. Bei $U_A = 30.75$ ist $\int_{V_1}^{V_3} f(V) = 0$, die Flächen unter und über dem Graphen sind also gleich groß. Hier wird die Front stationär, eingeschalteter und ausgeschalteter Zustand stehen im Gleichgewicht. Bei noch kleinerer Anodenspannung ist der ausgeschaltete Zustand dominant, die Front läuft rückwärts. Unterhalb eines Schwellenwertes von ca. $U_A = 29.8$ verschwindet der eingeschaltete Zustand in einer Sattel-Knoten-Bifurkation, da $f(V)$ darunter nur noch eine Nullstelle besitzt. Daher endet dort der Graph der Frontgeschwindigkeit.

In Abbildung 3.8 ist der Zündvorgang des Systems dargestellt, wenn es durch lokale Einprägung von Zündstrom zum Schalten gebracht wird. Dazu wurde U_G überall gleich Null gesetzt, außer an einzelnen Punkten, was einer punktförmigen Einprägung von Zündstrom in die Gateschicht entspricht. Es wurde entweder an einem Punkt, oder an zwei Punkten im Abstand d „Strom eingeprägt", d.h. $U_G > 0$ gesetzt. Diese lokale Steuerspannung U_G wurde manuell erhöht, bis der Schaltvorgang einsetzte. Die Diagramme zeigen einen gerade noch subkritischen Zustand, in dem die Zündung eben noch nicht einsetzt. Man erkennt, wie sich die Potentialverteilungen bei Stromeinprägung in 2 Einprägepunkten überlappen. Dadurch sinkt der kritische Zündstrom ab, im Extremfall beim Abstand $d = 0$ bis auf die Hälfte. Eine solche Wechselwirkung der Zündströme wird auch experimentell beobachtet (s. Abschnitt 3.8.6).

Das obige RD-System modelliert die Zünd- und Frontausbreitungsprozesse in einer ausgedehnten Thyristorprobe qualitativ richtig. Detailliertere Simulationsrechnungen von Thyristorstrukturen mit einem kommerziellen Halbleiter-Simulationstool werden in Kapitel 4 beschrieben.

3.6 Präparation der Thyristorstrukturen

Im Verlauf der Untersuchungen wurden verschiedene Proben von Thyristorstrukturen angefertigt. Alle untersuchten Proben wurden aus Wafern von Infineon präpariert, die zur selben Charge gehören und das T96 genannte Dotierungsprofil besitzen (Abb. 3.9).

Dieses Profil wurde ausgehend von einer konstanten n-Grunddotierung (Ph) mit Hilfe von zwei Diffusionsschritten erzeugt: Eine tiefe p-Diffusion (Ga) beidseitig zur Erzeugung von p-Basis und Anode, und eine flache n-Diffusion (Ph) zur Erzeugung der Kathode.

3.6.1 Proben Typ I

Zunächst wurden aus dem Grundmaterial Thyristorproben vereinzelt und strukturiert. Dabei müssen Zugänge zur p-Basis geschaffen werden, und Kontakte für Anode, Kathode und p-Basis erzeugt werden.

Aus dem Wafer wurden durch Anritzen mit einem Diamantstift mehrere ca. 20×30 mm große Stücke gebrochen. Mittels Photolithographie wurden Gatezugänge definiert, und dann mittels Naßätzung geätzt. Die Gatezugänge sind dabei Vertiefungen von 1 mm Durchmesser, die im Abstand von 5 mm angebracht sind.

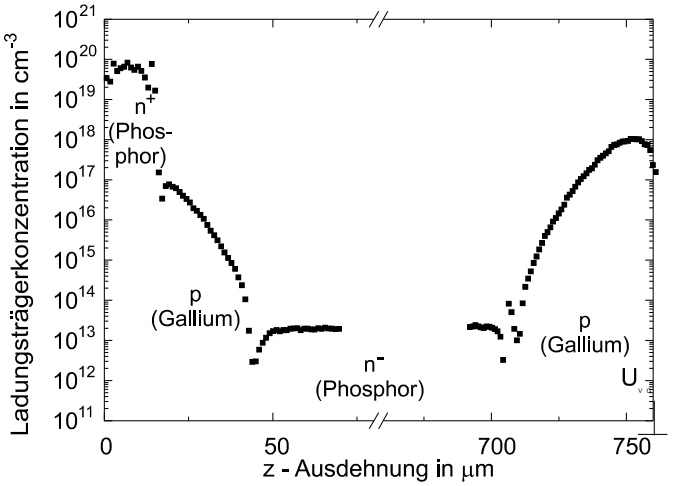
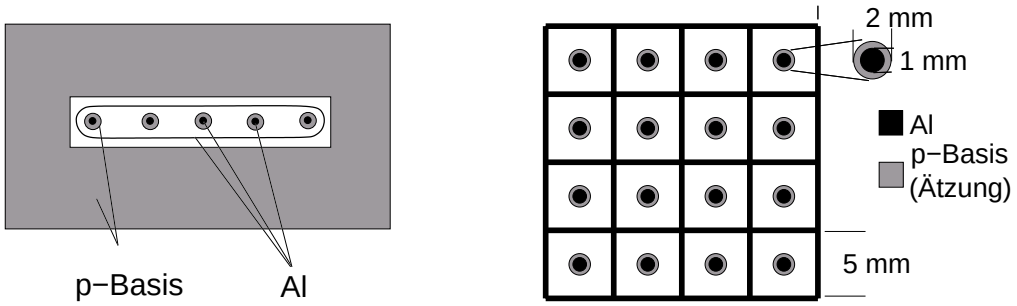


Abbildung 3.9: Dotierungsprofil der T96-Thyristorstrukturen, bestimmt durch Spreading-Resistance Verfahren (aus [Pla97]).



Probe 98-1D Probe 98-2D

Abbildung 3.10: Zwei Geometrien von selbst präparierten Thyristorproben.

Um den Zugang zur Gateschicht zu erhalten, mußte eine Ätztiefe von ca. $20-30$ µm erreicht werden, was einige Anforderungen an die Haltbarkeit des verwendeten Photoresists im Säurebad stellte. Die Härte und Haltbarkeit des Resists läßt sich durch ein Aushärten bei 80°C - 150 °C zwar verbessern, jedoch lößt sich der Lack trotzdem nach einigen Minuten vom Wafer ab, was durch Gasblasen verursacht wird, die bei der Ätzung auch unterhalb des Lackes entstehen. Die erforderlich Ätztiefe konnte jedoch erreicht werden.

Geätzt wurde mit einem Gemisch aus Salpetersäure, Flußsäure, Essig- und Phosphorsäure (CP6-Ätze). Dabei sind HNO_3 und HF die Hauptreagenzien: HNO_3 oxidiert die Silizium-Oberfläche, HF löst das gebildete Oxid auf[Rug84]. Die Ätzrate ist stark von Temperatur und Durchmischung der Ätze abhängig, daher die Ätztiefe nur schwer zu kontrollieren. Die Reaktion ist exotherm und die Säure erwärmt sich während der Ätzung. Sie ätzt auch Glas, weshalb sie in einem Teflongefäß gehandhabt wird. Dieses ist zwecks Temperaturregelung und Kühlung doppelwandig, trotzdem kann die Temperatur während der Ätzung nicht konstant gehalten werden, da die Wärmeleitung der Teflonwände nicht ausreichend ist. Daher wird die Temperatur vor der Ätzung auf 20°C geregelt, und steigt während des Ätzvorgangs um einige Grad an. Bei Ätzzeiten von 3-5 Minuten wird ein Abtrag von $20-30$ µm erreicht, so daß der Zugang zur p-Basis sicher erreicht wird.

Nun wird Aluminium zur Kontaktierung aufgedampft und per Lift-Off-Technik strukturiert: Die Oberseite (Kathodenseite) der Probe wird mit Photoresist überzogen, mit einer zweiten Maske belichtet und entwickelt. Dann wird die Probe im Hochvakuum mit Al bedampft. Der Lack verhindert an den beschichteten Stellen die Metallisierung des Siliziums. Dann wird der Lack mitsamt Al-Beschichtung im Ultraschallbad mit Hilfe eines Lösungsmittels abgelöst. An den ungeschützten Stellen ist das Si mit Al überzogen. Die Unterseite der Probe (Anodenseite) wird nicht maskiert und im gleichen Prozeß vollständig metallisiert. Zuletzt erfolgt eine

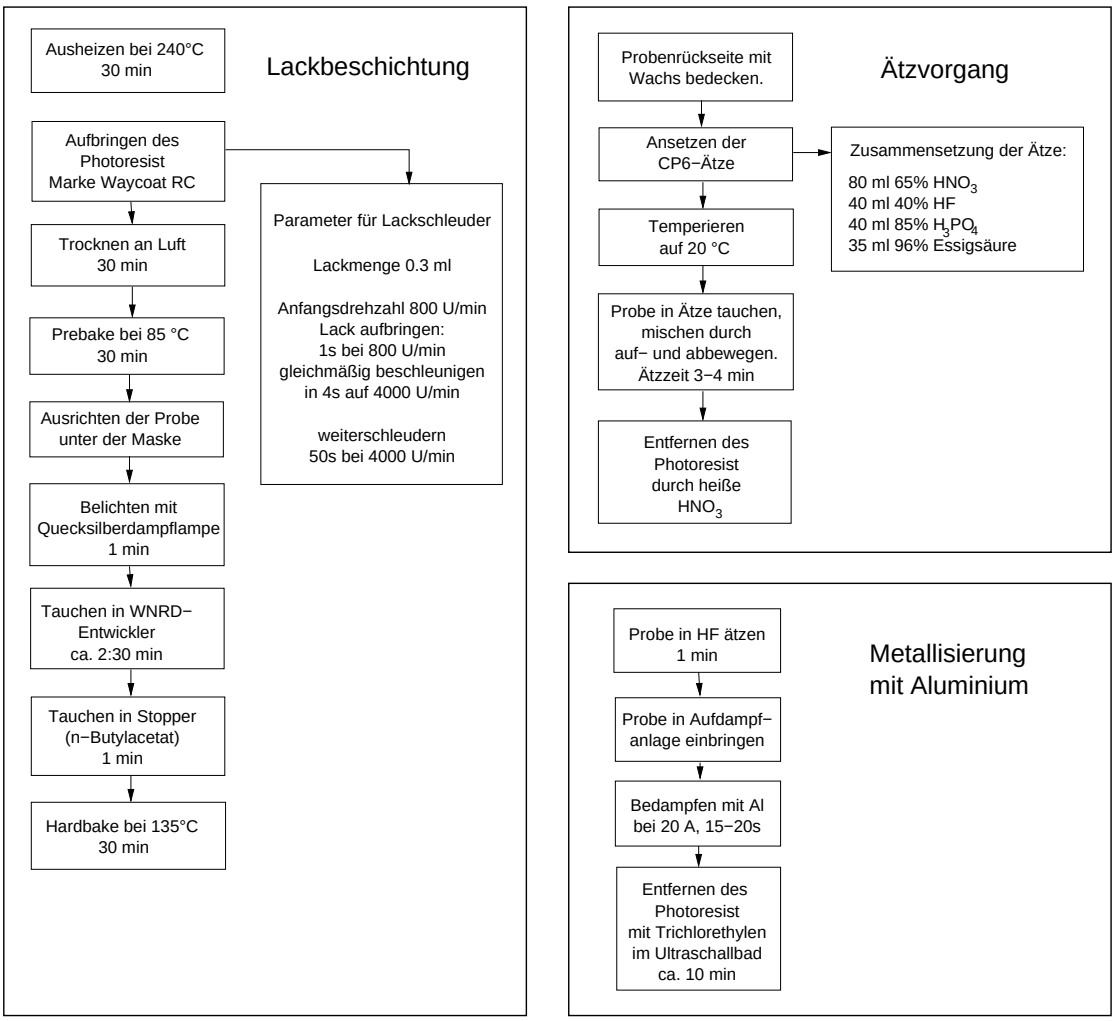


Abbildung 3.11: Präparationsschritte der selbstpräparierten Thyristorproben.

Sinterung bei 450°C für ca. 1 h, um den elektrischen Kontakt zwischen Al und Si zu stabilisieren.

Die so präparierte Probe 98-1D besitzt 5 Gatekontakte in linearer Anordnung, im Abstand von jeweils 5 mm, auf einem aktiven Grundgebiet von 25×5mm. Die Kathode wurde mit einem Aluminiumring kontaktiert. Sie läßt sich an jedem der Gatekontakte zünden, und zeigt sowohl in elektrischen wie auch in optischen Beobachtungen Zündfronten.

Eine zweite so präparierte Probe (Probe 98-2D) besitzt ein zweidimensionales, quadratisches Gitter von Gatekontakten mit 5 mm Rastermaß. Die Kathode ist

durch ein Gitter aus Aluminiumstreifen kontaktiert. Sie wurde nicht vollständig metallisiert, um die Rekombinationsstrahlung – und somit die Zündfronten – auch optisch beobachten zu können. An dieser Probe wurden zweidimensionale Fronten beobachtet, die sich von jedem Gatekontakt aus zünden ließen und die sich konzentrisch von diesem Gate ausgehend über die gesamte Probe ausbreiteten.

3.6.2 Proben Typ II

Proben mit größeren Kontaktflächen und einer größeren Anzahl von Gatekontakten wurden in Zusammenarbeit mit Infineon präpariert. Auch hier wurde eine Ätzung und die Metallisierung durchgeführt. Die Ätzung wurde hier mit dem Verfahren der *Sprühätzung* durchgeführt. Dabei wird ständig frische Ätzlösung auf den Wafer gesprüht, so daß Reaktionsprodukte abtransportiert werden. Damit ist die Ätzrate wesentlich besser kontrollierbar, und die geätzten Oberflächen werden glatter und homogener, da die Gasbildung bei der Ätzung nicht mehr zu Blasenbildung führt.

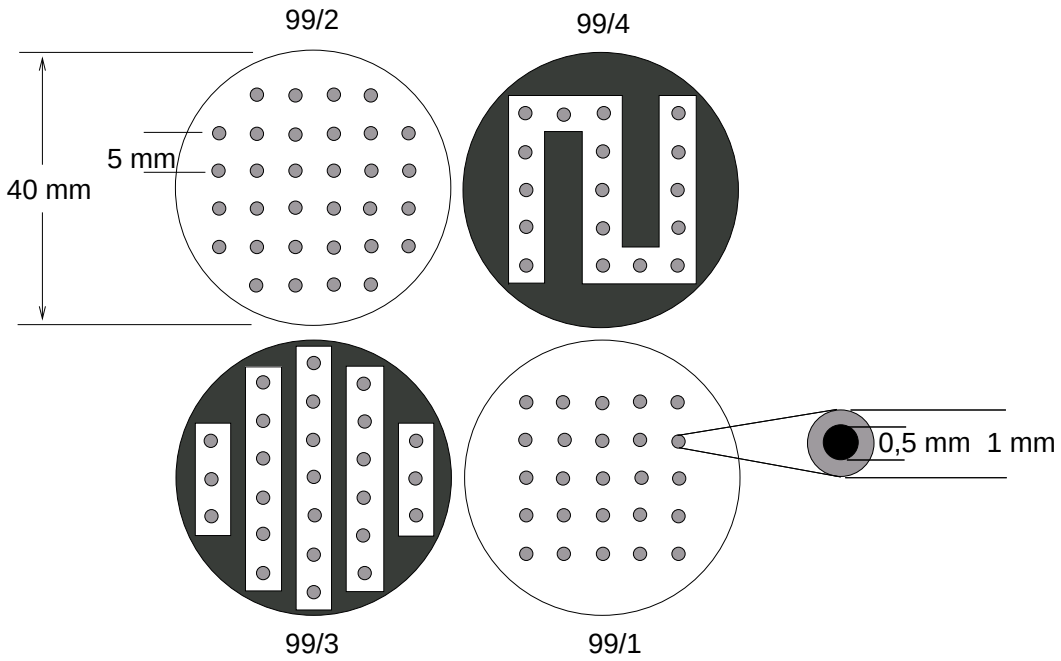


Abbildung 3.12: Geometrie der 4 bei Infineon gefertigten Thyristorproben. Die Proben wurden aus einem einzigen 4-Zoll-Wafer geschnitten, die Gatevertiefungen (hellgrau) sind bis zur p-Basis abgeätzt, die dunkelgrauen Zonen bis zur n-Basis. Die Kathode ist ganz mit Al metallisiert, wobei ein Netz von feinen Löchern die Beobachtung der Rekombinationsstrahlung ermöglicht. Die Gatevertiefungen sind ebenfalls mit Al metallisiert.

Die Ätzung wurde in zwei Stufen ausgeführt, so daß zwei verschiedene Ätztiefen erreicht werden konnten. Damit konnte nun ein *Grundgebiet* definiert werden, indem der Thyristor außerhalb desselben bis zur n-Basis abgeätzt wurde. Dazu sind ca. 80 µm Ätztiefe erforderlich, was mit der Spühätzung möglich war.

Die Metallisierung wird nicht mehr durch Lift-Off-Technik, sondern ebenfalls mit einer Ätztechnik strukturiert. Es wird zuerst Al aufgedampft, dann mit Resist beschichtet, dieser wird belichtet und entwickelt. Dann wird das Aluminium mit einer speziellen Ätze an den ungeschützten Stellen abgetragen.

Insgesamt vier Proben wurden auf einem Wafer präpariert, mit jeweils unterschiedlichen Anordnungen der Grundgebiete und Gatekontakte (siehe Abb. 3.12). Dabei wurden die Vertiefungen der Gatekontakte auf ca. 25µm Tiefe abgeätzt, die Grundgebiete wie oben beschrieben auf 80µm Tiefe.

Die Al-Metallisierung der Kathode bedeckt die gesamte Kathodenoberfläche, mit Ausnahme von einem Gitter von Beobachtungslöchern, durch die die Rekombinationsstrahlung austreten kann.

Das kreisförmige Grundgebiet der Proben 99/1 und 99/2 dient zur Untersuchung zweidimensionaler Zündfronten, während Probe 99/3 ein mehrere eindimensionales Grundgebiete besitzt. Das mäanderförmige Grundgebiet von Probe 99/4 sollte längere, eindimensionale Kopplungsstrukturen ermöglichen.

3.6.3 Elektronenbestrahlung zur Absenkung der Trägerlebensdauer

Zwei der vier bei Infineon gefertigten Proben[1] wurden mit hochenergetischen Elektronen bestrahlt, um ihre Minoritätsträger-Lebensdauer zu senken.

Die Rekombination geschieht in Si hauptsächlich über Rekombinationszentren, die tiefe Störstellen innnerhalb der Bandlücke erzeugen. Diese Shockley-Read-Hall-Rekombination [Hal52, Sho52] bestimmt hauptsächlich die Lebensdauer der Minoritätsträger.

Um die Lebensdauer abzusenken, müssen zusätzliche Störstellen im Halbleiter erzeugt werden. Dazu werden entweder Störatome in den Halbleiter eindiffundiert, oder es werden Gitterfehler wie Leerstellen und interstitielle Atome durch Bestrahlung mit energiereichen Partikeln erzeugt. Zur Lebensdauerdotierung werden häufig Gold und Platin verwendet, da sie ein Niveau nahe der Bandmitte von Si besitzen (s. Abb. 3.13).

[1] Proben 99/1 und 99/3

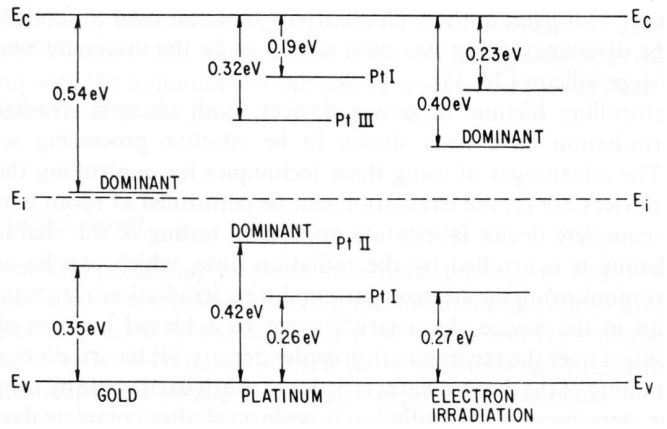


Abbildung 3.13: Energieniveaus von tiefen Störstellen, die durch Gold- und Platindotierung und Elektronenbestrahlung erzeugt werden (aus [Bal87])

Die Bestrahlung mit Elektronen hat einen ähnlichen Effekt, läßt sich aber besser dosieren, und sie kann nach der Metallisierung als letzer Bearbeitungsschritt erfolgen. Die Bestrahlung erzeugt interstitielle Atome und Leerstellen. Die Leerstellen sind nach [Bal87] auch bei niedrigen Temperaturen mobil, so daß man nach der Bestrahlung hauptsächlich an Dotier- oder Sauerstoffatome gebundene Vakanzen, und paarweise Vakanzen, sogenannte Divakanzen vorfindet.

In [Haz99] werden die in neutronendotiertem n-Silizium durch Bestrahlung entstehenden Störstellen quantitativ untersucht, und die dort angegebenen Werte erlauben die Abschätzung der Lebensdauerreduktion, die mit einer bestimmten Bestrahlungsdosis erzielt werden kann. Demzufolge findet man nach Bestrahlung mit 4MeV-Elektronen 2 Defekttypen: ein Vakanz-Sauerstoff-Zentrum, und ein Divakanz-Zentrum. Hazdra und Vobecky geben folgende Werte für die entstehende Störstellendichte und deren Wirkungsquerschnitt an:

Defekttyp	N_T/Fluenz in cm^3/cm^2	σ_n in cm^2
VO	0.17	$1.0 \cdot 10^{-14}$
V2(2-/-)	0.018	$8.0 \cdot 10^{-16}$
V2(-/o)	0.019	$1.0 \cdot 10^{-15}$

Die Lebensdauer kann für den Fall schwacher Injektion nach folgender Näherungsformel abgeschätzt werden:

$$\tau = 1/(N_T \sigma_n v_{\mathrm{th}}) \tag{3.6}$$

Mit einer thermischen Geschwindigkeit von $v_{\mathrm{th}} = 1 \cdot 10^7$ cm/s ergeben sich folgende Werte für die bestrahlten Proben, wobei nur des dominante VO-Zentrum berücksichtigt wurde.

Probe	Fluenz in cm^{-2}	τ in µs
99/1 + 99/3	$5.0 \cdot 10^{12}$	11.7
99/1	$7.5 \cdot 10^{12}$	7.84
99/3	$1.0 \cdot 10^{13}$	5.88

Diese Werte stimmen relativ gut mit denen überein, die ein *Abschaltexperiment* ergab, bei dem nach Umpolung des Thyristors vom Durchlass- in den Sperrzustand die Zeitkonstante des exponentiellen Abfalls des Sperrstroms bestimmt wird (s. Abschnitt 3.8.2).

Die Störstellen können oberhalb von 400 °C ausheilen, so daß der Effekt der Bestrahlung durch Temperung zumindest teilweise rückgängig gemacht werden kann.

Die verkleinerte Lebensdauer hat folgenden Effekt auf die Vorgänge im Thyristor: Die verstärkte Rekombination in einem bestrahlten Thyristor senkt die Stromverstärkung α der beiden Teiltransistoren, was die Zünd- und Halteströme des Thyristors anhebt. Die Frontbreite wird dadurch verkleinert. Dies läßt sich folgendermaßen begründen: Im Bereich der Front muß der Zündstrom lateral in der p-Basis vom ein- in das ausgeschaltete Gebiet fließen. Dieser Strom wird durch den Potentialgradienten in der p-Basis getrieben. Ein größerer Zündstrom erfordert also einen größeren Potentialgradienten, und somit einen schmaleren Übergang zwischen ein- und ausgeschaltetem Gebiet. Diese kleineren Frontbreiten in Verbindung mit höheren Zünd- und Halteströmen wurden auch tatsächlich beobachtet.

3.7 Optische Beobachtungen des Zündvorganges

Die Dynamik des Zündvorgangs wurde mit Hilfe der Rekombinationsstrahlung untersucht, die ein stromdurchflossener pn-Übergang erzeugt. Die Wellenlänge der Strahlung liegt bei Si bei ca. $\lambda = 1.10$ µs, entsprechend der Bandlücke von $W_g = 1.12$ eV, die Effizienz der Strahlungserzeugung ist jedoch gering, da Si ein indirekter Halbleiter ist. Die Intensität der emittierten Strahlung ist proportional zur Rekombinationsrate im pn-Übergang, die mit einer gewissen Verzögerung der Stromdichte durch den pn-Übergang folgt. Die Aufnahmen bilden also die Stromdichteverteilung in den Proben ab.

Daher wurden die Beobachtungen mit einer infrarotempfindlichen, bildverstärkenden Kamera der Fa. Hamamatsu vorgenommen. Es handelt sich dabei um ein System aus IR-empfindlicher Photokathode, einer Multichannelplate (MCP), einem

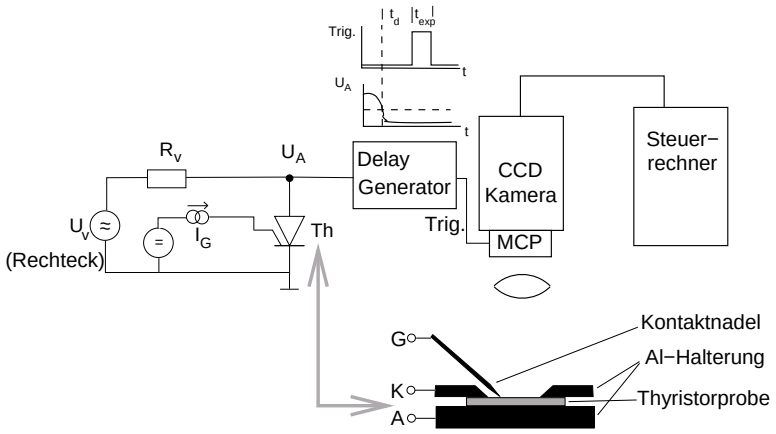


Abbildung 3.14: Versuchsaufbau zur optischen zeitaufgelösten Beobachtung von Zündfronten in Thyristorstrukturen. Die Thyristorprobe wird mit Rechteckspannung versorgt, wobei die Zündung durch einen konstanten Gatestrom erreicht wird. Die Multichannelplate wird synchron mit der Zündung von einem Delaygenerator angesteuert. Da ein Einzelbild nicht lichtstark genug ist, werden die Bilder vieler Perioden aufsummiert. Durch Variation der Verzögerung t_d wird eine Bilderserie vom Zündvorgang aufgenommen.

Leuchtstoff und einer CCD-Kamera. Die Photokathode wird durch ein Peltierelement gekühlt, um den Dunkelstrom thermisch emittierter Elektronen zu verringern. Die MCP dient als Sekundärelektronenvervielfacher, der den Elektronenstrom der Photokathode verstärkt. Der verstärkte Elektronenstrom regt dann den Leuchtstoff zum Leuchten an, und dieses sichtbare Licht wird auf die CCD-Kamera abgebildet. Die MCP besteht aus einer Vielzahl von kleinen Glasröhren, von denen jede als miniaturisierter Sekundärelektronenvervielfacher dient. So erhält man eine räumliche Auflösung, die etwa dem typischen Abstand der Röhrchen entspricht. Die Versorgungsspannung der MCP kann über einen externen Triggereingang geschaltet werden, was eine zeitaufgelöste Beobachtung der Vorgänge im Halbleiter ermöglicht.

Die Kamera erlaubt den Dauerbetrieb als auch den Impulsbetrieb mit zeitlicher Auflösung von ca. 100 ns , mit einer maximalen Pulsdauer von ca. 0.1 ms und einer Wiederholrate von ca. 100 Hz. Der Steuerrechner nimmt die Bilder der CCD-Kamera auf, und erlaubt die Integration vieler Einzelbilder über eine längere Zeit, zwecks Verbesserung des Signal-Rausch-Verhältnisses.

Die Probe 98-1D zeigt in einer zeitaufgelösten Messung deutlich die Ausbreitung einer Zündfront. Die MCP der Kamera wird zu diesen Messungen mit dem Zündzeitpunkt der Probe synchronisiert, wobei ein Delaygenerator eine definierte Zeitverzögerung zwischen Zündung des Thyristors und „Belichtung“ der Kamera erlaubt.

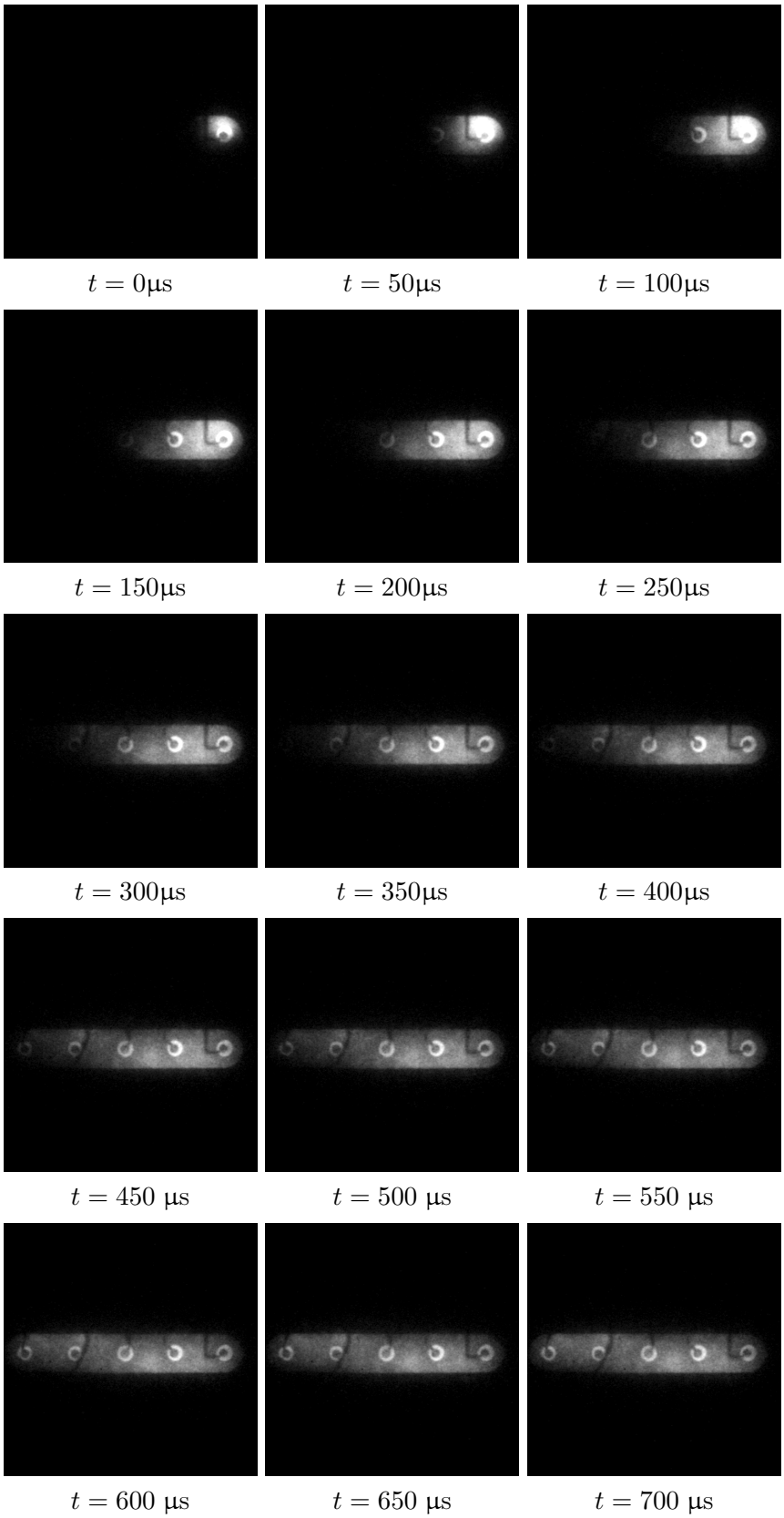
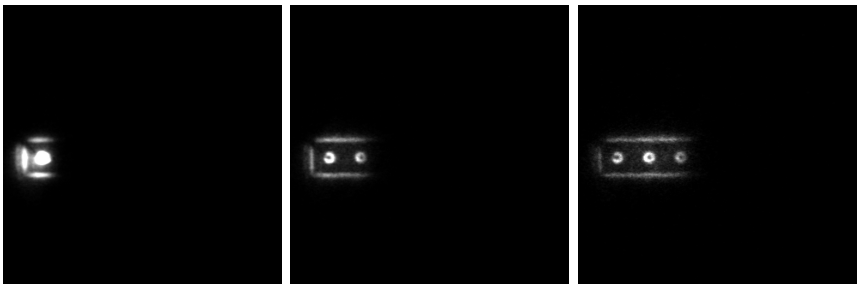
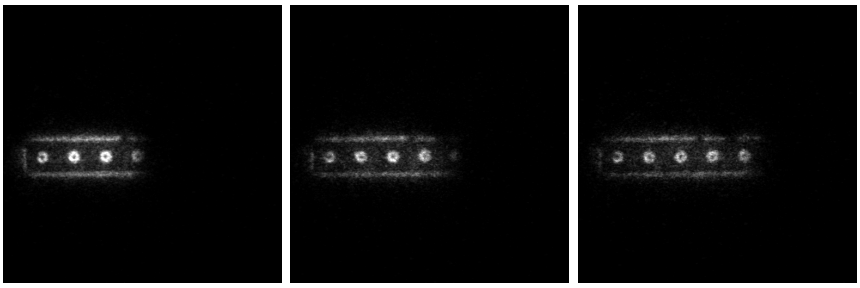


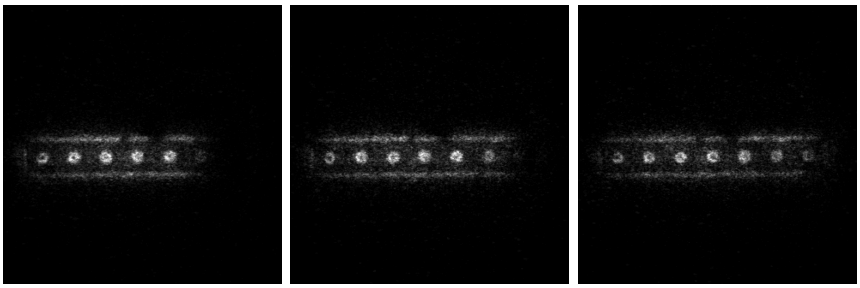
Abbildung 3.15: Zündfront in Thyristorprobe 98-1D, beobachtet mit bildverstärkter Infrarotkamera. Parameter: $U_V = 10$ V, $R_V = 20\ \Omega$, $I_A = 0.417$ A, Belichtungszeit $t_{\mathrm{exp}} = 40\ \mu s$



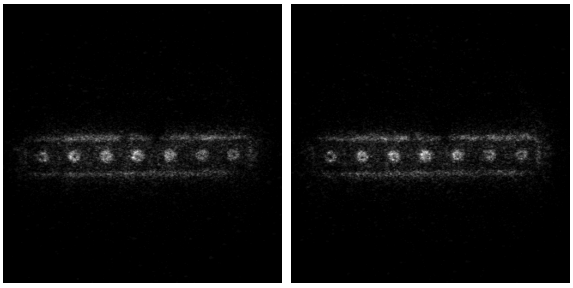
$t = 0\mu s$ $t = 200\mu s$ $t = 400\mu s$



$t = 600\mu s$ $t = 800\mu s$ $t = 1000\mu s$



$t = 1200\mu s$ $t = 1400\mu s$ $t = 1600\mu s$



$t = 1800\ \mu s$ $t = 2000\ \mu s$

Abbildung 3.16: Zündfront in Thyristorprobe 3/99, nach Bestrahlung mit $5 \times 10^{12} e^-/cm^2$. Deutlich ist der praktisch stufenförmige Verlauf der Front zu erkennen, mit nahezu konstanter Lichtintensität im eingeschalteten Gebiet. Mit wachsender Hochstromdomäne sinkt die Lichtintensität, da sich der durch den Vorwiderstand begrenzte Strom auf eine wachsende Fläche verteilt. Parameter: $U_V = 8.16$ V, $R_V = 10\ \Omega$, $I_A = 0.56$ A, Belichtungszeit $t_{\text{exp}} = 100\ \mu s$

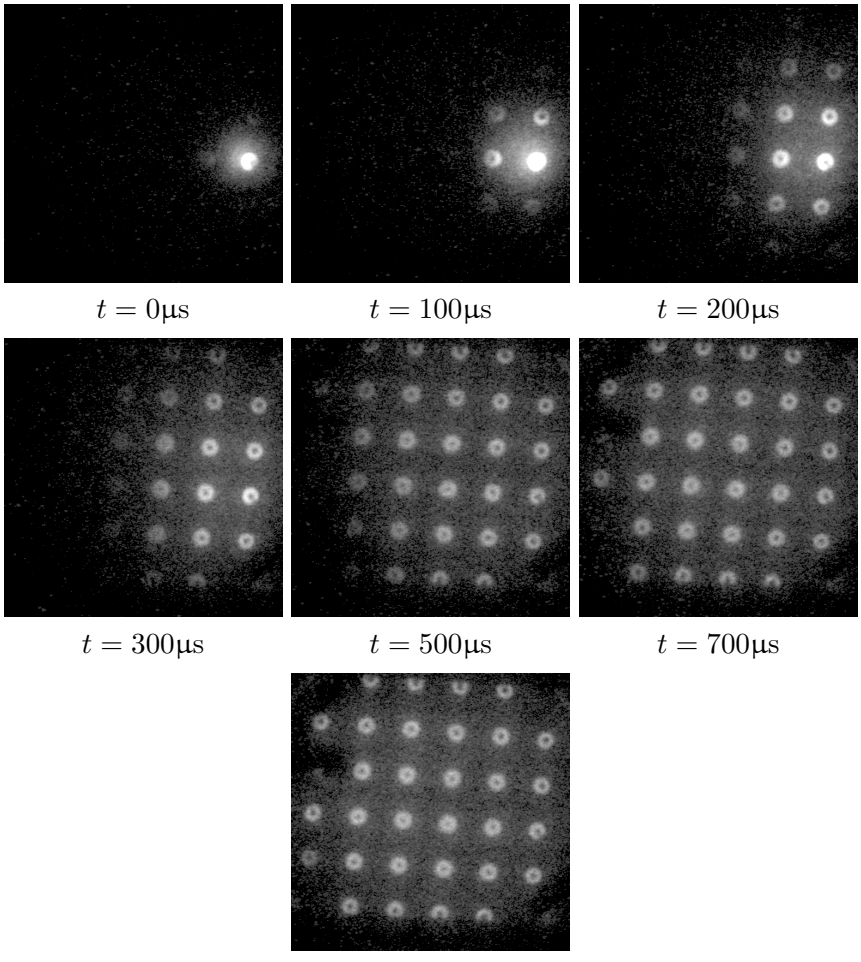


Abbildung 3.17: Zündfront in Thyristorprobe 1/99 (rund), unbestrahlt. Man erkennt die sich konzentrisch vom Zündkontakt ausbreitende Hochstromdömane. Parameter: $U_V = 3.0$ V, $R_V = 10\ \Omega$, $I_A = 0.23$ A, $I_G = 0.54$ mA, Belichtungszeit $t_{\exp} = 50\ \mu$s

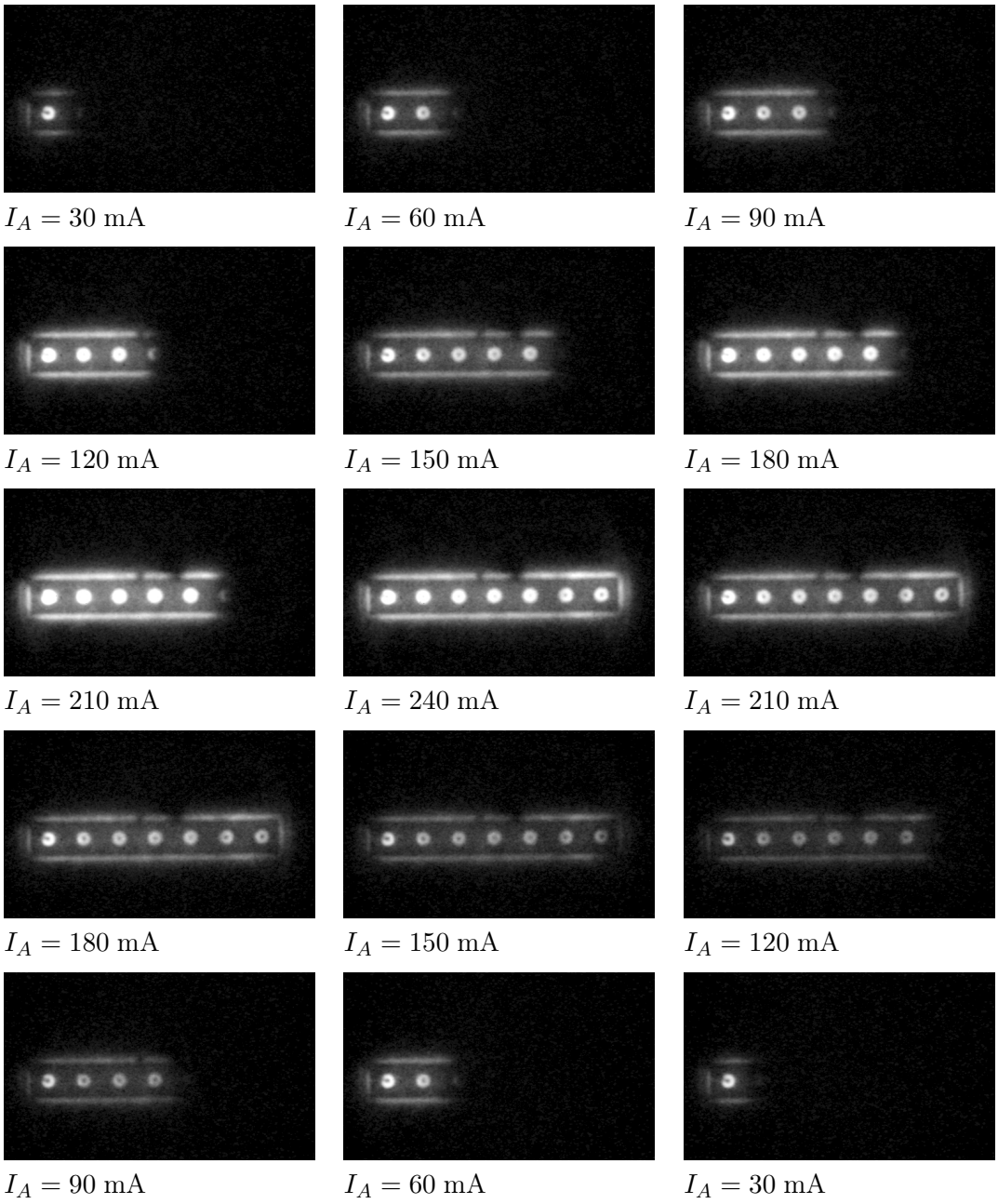


Abbildung 3.18: Quasistationäre Front. Der Anodenstrom wird langsam variiert, das eingeschaltete Gebiet dehnt sich mit steigendem Strom aus, und schrumpft mit fallendem Strom. Hysterese ist bei 120 mA und wieder bei 210 mA zu erkennen: Die Leuchtdichte des Filaments wächst, bis es offensichtlich einen Widerstand überwindet und sich weiter ausdehnt. Bei sinkendem Strom ist das Filament zwischen 210 und 90 mA breiter und lichtschwächer als bei steigendem Strom, es bewegt sich erst unterhalb einer bestimmten Leuchtdichte (d.h. Stromdichte) weiter. Parameter: Probe 3/99, bestrahlt mit $5 \cdot 10^{12} \mathrm{e}^-/\mathrm{cm}^2$. $R_V = 50\ \Omega$, $I_A = 30 - 240$ mA, $I_G = 0.58$ mA, Belichtungszeit $t_{\mathrm{exp}} = 10$ s

Die Probe wird hierbei mit einer Rechteckspannung von 100 Hz betrieben, und somit alle 10 ms erneut gezündet. Da in der Belichtungszeit eines Einzelbildes von typisch 100 µs nur wenige Photonen vom CCD registriert werden, müssen viele Einzelbilder aufintegriert werden. In einer typischen Meßzeit von 1 min sind dies 6000 Einzelbilder. Die Verzögerung t_d wird von Aufnahme zu Aufnahme variiert, um eine Bilderserie vom Frontausbreitungsprozeß zu erhalten.

Die Beobachtungen (s. Abb. 3.15) zeigen interessante Eigenschaften der Zündfronten: Wie erwartet, zündet die Probe in unmittelbarer Umgebung des Gatekontaktes, an dem ein Gatestrom eingeprägt wird. An jedem der fünf Gatekontakte der Probe lässt sich eine Front zünden. Die Front breitet sich dann mehr und mehr über das Grundgebiet aus, wobei sie augenscheinlich „flacher“ wird, d.h. die Breite des Übergangs zwischen hellem und dunklen Gebiet nimmt zu. Die Ursache dafür ist die globale Gegenkopplung durch den Vorwiderstand: Wenn die Versorgungsspannung wesentlich größer als die Anodenspannung im eingeschalteten Zustand ist, so ist der Spannungsabfall am Vorwiderstand fast konstant, und somit auch der Anodenstrom. Die Stromdichte im eingeschalteten Teil der Probe sinkt daher während der Ausbreitung der Front, so daß die Frontgeschwindigkeit abnimmt und die Frontbreite zunimmt.

Die beobachteten Fronten haben durch diese fortlaufende Abflachung wenig Ähnlichkeit mit den Frontlösungen eines RD-Systems, die mit konstanter Form eine reine Translationsbewegung ausführen. Jedoch könnte dies daher rühren, das die laterale Ausdehnung der Probe zu klein verglichen mit der charakteristischen Frontbreite im Gleichgewichtszustand ist, so daß die Frontbreite nach der Zündung wächst, bis der Rand des Grundgebietes erreicht wird.

Proben, die mit Elektronen bestrahlt wurden, zeigen Fronten mit wesentlich schärferen Kanten, die ein nahezu gleichmäßig strahlendes eingeschaltetes Gebiet von einem dunklen ausgeschalteten Gebiet trennen (Abb. 3.16). Die Frontgeschwindigkeit ist gegenüber den unbestrahlten Proben verringert. Diese Verbesserung der Frontform lässt sich, wie in Abschnitt 3.6.3 beschrieben, durch die verringerte Trägerlebensdauer in diesen Proben und die damit erhöhten Lateralströme erklären.

Mit der bestrahlten Probe 99/3 (streifenförmige Geometrie) gelang auch die Beobachtung einer stationären Front (Abb. 3.18). Dazu wurde die Probe mit einem konstanten Anodenstrom I_A versorgt. Die Zündung wurde mit einem kurzen Impuls an einem der Gatekontakte ausgelöst. Die Beobachtung mit dem Hot-Electron-Analyser zeigte, daß sich eine stationäre Front ausbildet, die den eingeschalteten Bereich vom ausgeschalteten Bereich trennt. Der eingeschaltete Bereich dehnt sich bei Erhöhung des Stromes I_A aus, und schrumpft bei fallendem Strom wieder. Die Leuchtstärke, und damit die Stromdichte im eingeschalteten Gebiet bleibt da-

bei etwa konstant. Allerdings ist die Stromdichte bei steigendem Strom größer als bei fallendem Strom, und die Front bewegt sich bei Stromänderungen nicht kontinuierlich, sondern sprunghaft. Sie bleibt gewissenmaßen an den Gatekontakten hängen. Bei Erhöhung des Stromes wächst die Leuchtstärke – also die Stromdichte – zunächst, bis die Front einen Gatekontakt weiter springt und die Leuchtstärke wieder absinkt. Bei sinkendem Gesamtstrom geschieht das gleiche, jedoch sinkt nun die Leuchtstärke auf einen tieferen Wert, bis die Front einen Gatekontakt zurückspringt. Dieses *Pinning* an den Gatekontakten schlägt sich auch in der Kennlinie der Probe (Abbildung 3.22) nieder. Man findet eine mehrfache Hysterese mit Sprüngen, die auftreten, wenn die Front jeweils einen Gatekontakt überwindet.

3.8 Elektrische Messungen

3.8.1 Kennlinien

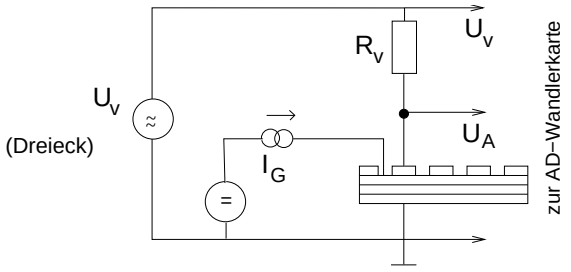
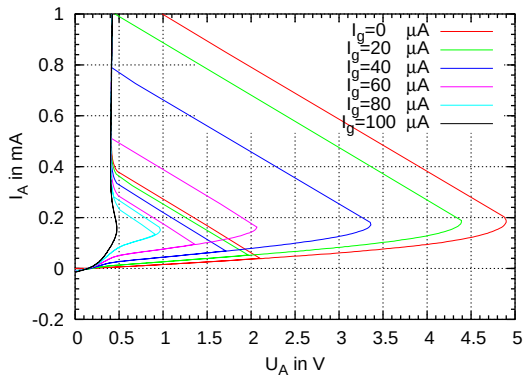
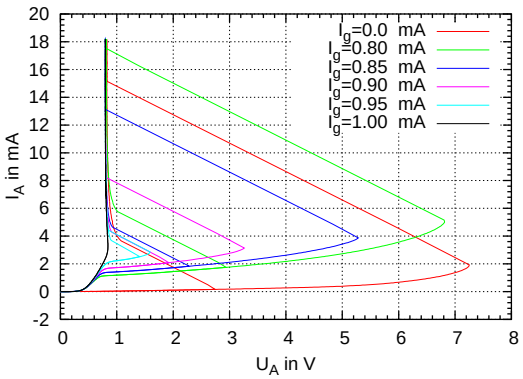


Abbildung 3.19: Schaltung zur Kennlinienmessung: Versorgung mit Dreieckspannung niedriger Frequenz: $f = 50$ mHz entspr. $T = 20$ s. Abtastrate $f_s = 500$ Hz, zwecks Rauschunterdrückung werden je 10 aufeinanderfolgende Meßwerte gemittelt. Gatestromeinprägung mittels spannungsgesteuerter Stromquelle.

Die Bistabilität der Thyristorproben wurde durch Bestimmung ihrer Kennlinie charakterisiert. Dazu wurde folgender Beschaltung (s. Abb. 3.19) verwendet: Die Anode der Thyristorprobe wird über einen Vorwiderstand mit Wechselspannung von einem Funktionsgenerator versorgt. Anodenspannung und -strom werden mit einem Speicheroszilloskop (Nicolet Pro 92) oder PC-gestützt mit einer Meßwerterfassungskarte (National Intruments PCI-MIO-16E-1) aufgenommen. Die Frequenz der Versorgungsspannung wurde so niedrig gewählt, daß das Bauteil den Strom- und Spannungsänderungen praktisch verzögerungsfrei folgen konnte. Der Gatekontakt wird mit einem regelbaren Gleichstrom I_G versorgt, so daß die Kennlinie $I_A(U_A)$ als Funktion von I_G gewonnen werden kann. Die Bestimmung der Kennlinie dient als erster Test, ob eine präparierte Probe überhaupt Schaltverhalten zeigt, und wenn ja, in welcher Größenordnung Zünd- und Halteströme und die Kippspannung liegen.



Probe 99/2, unbestrahlt, $R_v = 5\ \mathrm{k\Omega}$



Probe 99/1, bestrahlt, $R_v = 500\ \Omega$

Abbildung 3.20: Kennlinien von Proben des Typ 2, bestrahlt und unbestrahlt, mit variablem Gatestrom I_G. Die Bestrahlung mit $7.5 \cdot 10^{12}\ e^-/\mathrm{cm}^2$ erhöht den Haltestrom als auch die Ströme im Blockierbereich um etwa eine Größenordnung.

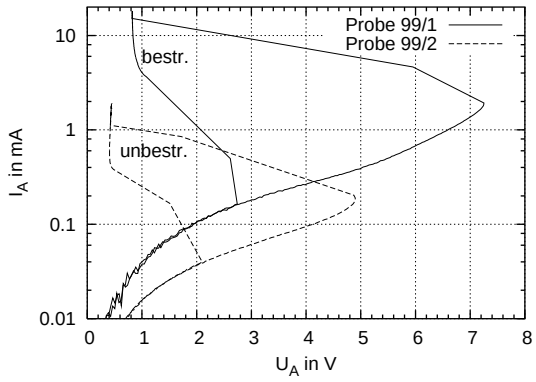


Abbildung 3.21: Kennlinien in logarithmischer Darstellung, ohne Gatestrom, mit und ohne Bestrahlung. Parameter wie in Abb. 3.20.

Abbildung 3.20 stellt das Schaltverhalten von zwei Proben vom Typ II dar. Die Kippspannung liegt selbst ohne Gatestrom ($I_G = 0$) unter 10 V, obwohl das Thyristormaterial für Hochspannungsthyristoren vorgesehen ist. Auch die Sperrströme erscheinen relativ hoch. Die vermutliche Ursache dafür sind Leckströme des gesperrten pn-Übergangs J2, die an den Rändern der Proben entstehen, wo die Raumladungszone von J2 freiliegt (Bei der Präparation fand keine besondere Passivierung oder Politur des Probenrandes statt). Ein Indiz dafür ist die Tatsache, daß bei Überhöhung der Anodenspannung die Zündung am Rand geschieht, wie die optische Beobachtung zeigt. Für die Anwendung in der SOM-Hardware bedeutet dies aber keine Einschränkung.

Mit steigendem Gatestrom sinkt die Kippspannung wie erwartet. Ebenso sinkt der Haltestrom, der Gatestrom unterstützt die Stabilisierung des eingeschalteten Zustands und das Ausschalten verschiebt sich zu kleineren Strömen hin.

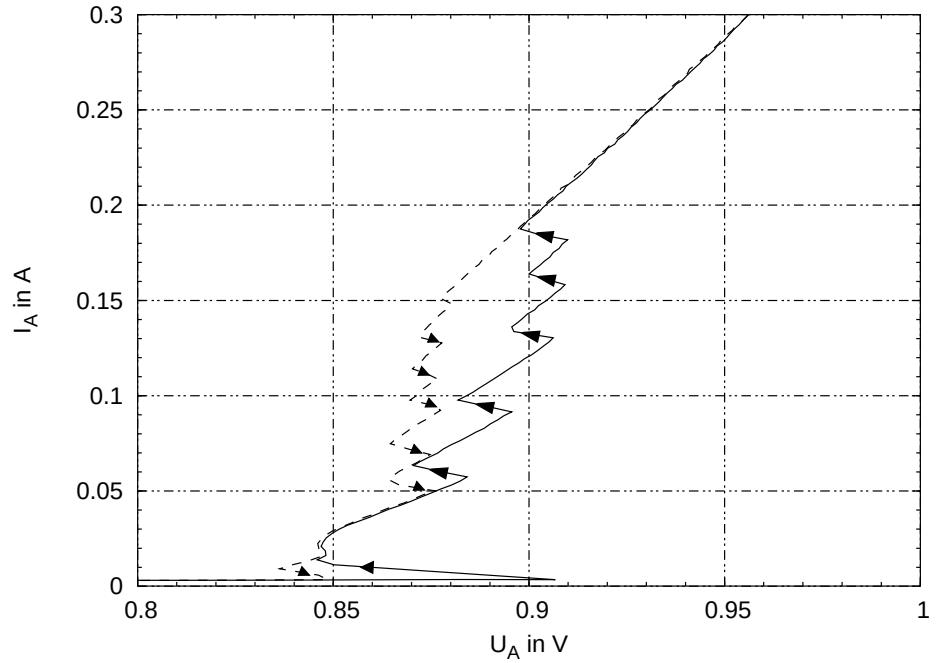


Abbildung 3.22: Kennlinienausschnitt von Probe 99/3, nach Bestrahlung. Bei Anodenströmen unter 0.19 A ist die streifenförmige Thyristorprobe nicht vollständig durchgeschaltet. Mit steigendem Strom gelangt eine immer größere Fläche in den eingeschalteten Zustand. Die mit Pfeilen markierten Sprünge beim Durchlaufen der Kennlinie entstehen durch Behinderung der Frontbewegung durch die 7 Gatekontakte der Probe. Die die Frontbewegung antreibende Anodenspannung muß erst eine Schwelle überschreiten, bis die Front einen Kontakt weiterspringen kann. Der Rücksprung geschieht erst bei einer geringeren Spannung, was die auftretende Hysterese erklärt.

Die mit Elektronen bestrahlten Proben zeigen folgende Veränderungen: Ihre Durchlaßkennlinien verschieben sich nach rechts, der Spannungsabfall im Durchlaßzustand wächst. Der Haltestrom I_H und der Zündstrom $I_G(U_A)$ nehmen zu. Dies läßt sich wie bereits erwähnt durch die verringerte Trägerlebendauer in den Basen erklären, wodurch bei gleicher Stromstärke durch erhöhte Rekombination die Trägerdichte abnimmt, und somit der Ohm'sche Spannungsabfall wächst. Die erhöhten Rekombinationsverluste heben Zünd- und Haltestrom an. Die Veränderungen sind gut in der logarithmischen Darstellung von Abb. 3.21 zu erkennen.

Optische Beobachtungen der bestrahlten Proben zeigten stationäre Frontzustände (s. Abb. 3.18), in denen nur ein Teil der Probe eingeschaltet ist. Die Fläche dieses Bereiches variiert mit dem Anodenstrom. Die Kennlinie dieser Probe zeigt einen entsprechenden Effekt (Abb. 3.22). Nach der Zündung schaltet zunächst nur

ein Teil der Probe ein, die Front bleibt an einem Gatekontakt hängen (*Pinning*). Wenn die Front einen Gatekontakt überspringt, entsteht ein Rücksprung in der Kennlinie. In umgekehrter Richtung treten die Rücksprünge bei jeweils kleineren Anodenspannungen auf, es tritt Hysterese ein.

3.8.2 Lebensdauerbestimmung durch Abschaltexperiment

Wie weiter oben ausgeführt, sollte die Bestrahlung der Thyristorproben die Lebensdauer der Minoritätsträger in den Basen stark absenken. Der Zweck dieser Maßnahme ist die Verkleinerung der Stromverstärkungen $\alpha_{1,2}$, mit der Konsequenz einer erhöhten Haltestromdichte und der daraus folgenden schmaleren Fronten. Die Lebensdauer wurde durch ein dynamisches Abschaltexperiment bestimmt. Der so erhaltene Wert dient als Vergleichswert für die Simulationsrechnungen in Kapitel 4.

Die Thyristorprobe wird dabei mit einem Vorwiderstand an Gleichspannung in Flußrichtung betrieben, und dann abrupt umgepolt. Durch die in den Basen verbliebenen Minoritätsträger entsteht sofort nach dem Kommutieren der Stromrichtung ein großer Sperrstrom. Während die Minoritätsträger aus den Basen ausgeräumt werden, klingt der Sperrstrom bis auf den stationären Sperrstrom ab.

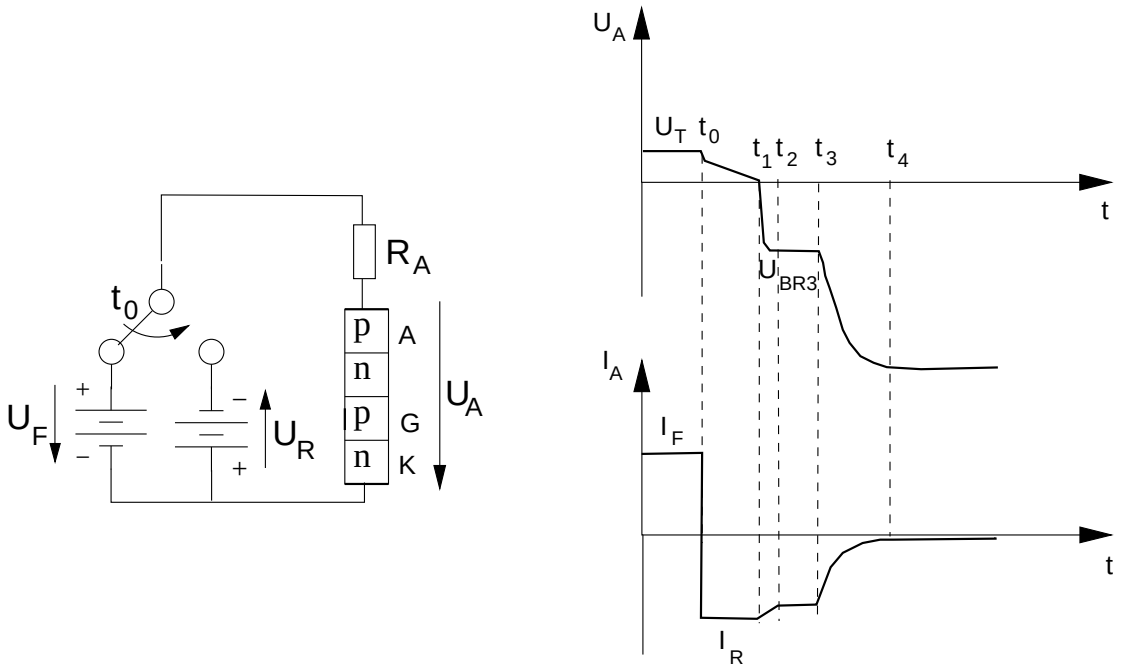
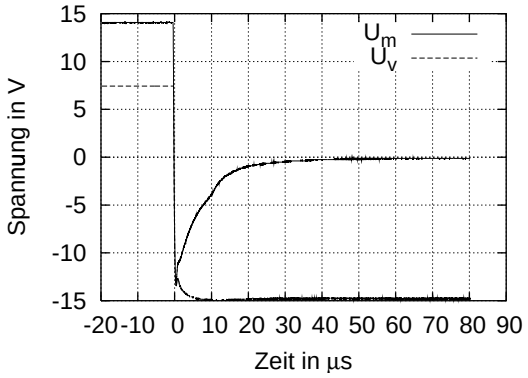


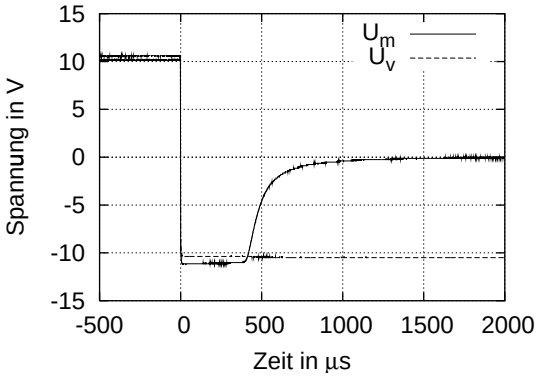
Abbildung 3.23: Abschaltvorgang bei abrupter Stromkommutierung.



Probe 99/1, bestrahlt mit $7.5 \cdot 10^{12}$ e$^-$/cm^2
Fitintervall: $[15, 50]$ µs

R_m in Ω	Abklingzeit τ in µs
500	8.20
100	8.71
50	8.09

$\tau_{\text{Mittel}} = 8.33 \pm 0.26$ µs



Probe 99/2, unbestrahlt
Fitintervall: $[800, 2000]$ µs

R_m in Ω	Abklingzeit τ in µs
500	305.6
100	301.0
50	311.7

$\tau_{\text{Mittel}} = 305 \pm 20$ µs

Abbildung 3.24: Abschaltvorgang mit abklingendem Sperrstrom im Vergleich für eine unbestrahlte und eine bestrahlte Thyristorprobe, und daraus bestimmte Lebensdauern.

Nach [Ger79] läßt sich dieser Abklingvorgang in vier Zeitabschnitte einteilen (s. Abb. 3.23): Die erste Speicherzeit t_{s1} von t_0 bis t_1, die erste Abfallzeit t_{f1} von t_1 bis t_2, die zweite Speicherzeit t_{s2} von t_2 bis t_3 und die zweite Abfallzeit t_{f2} von t_3 bis t_4.

Die erste Abfallzeit ist der Zeitraum, in dem die Anodenspannung noch auf dem Niveau der Durchlaßspannung liegt, also noch positiv ist. Der Sperrstrom I_R ist praktisch durch die von außen angelegte Spannung U_R festgelegt. Dieser Rückstrom bleibt praktisch konstant, bis sich die rückströmenden Elektronen in der p-Basis am n-Emitter (Kathode) zu erschöpfen beginnen, und der pn-Übergang J3 Sperrspannung aufnimmt. J2 injiziert dabei Löcher in die n-Basis, aber kaum Elektronen in die p-Basis, so daß die Elektronen in der p-Basis schneller erschöpft sind als die Löcher in der n-Basis. Dazu trägt auch die kleinere Basisweite der p-Basis bei.

Dann geschieht ein rascher Abfall der Anodenspannung bis zur Durchbruchspannung U_{BR3} von J3, wonach sich die zweite Speicherzeit anschließt. Diese dauert so lange an, bis die injizierten Minoritätsträger vor dem Übergang J1 abgebaut sind und J1 Sperrspannung aufnimmt. Das entsprechende Plateau ist in den hier dargestellten Abklingkurven nicht zu sehen, da die Durchbruchspannung U_{BR3} nicht erreicht wurde.

Entscheidend für die Bestimmung der Lebensdauer ist der letzte Abfall des Sperrstromes bis hin zum stationären Sperrstrom, also die zweite Abfallzeit. Nun ist nur noch in der relativ dicken n-Basis eine erhöhte Minoritätsträgerkonzentration vorhanden. Die Ladungsträger werden durch den geringen Strom kaum noch abtransportiert, so daß nur noch die Rekombination zum Abbau der Träger beiträgt, der Sperrstrom fällt nun exponentiell ab. In der n-Basis kann zu Beginn der 2. Abfallzeit noch starke Injektion vorliegen, abhängig von der Dotierungskonzentration und der Stromdichte. Erst wenn die Trägerkonzentration unter den Betrag der Grunddotierung fällt (hier $1.5 \cdot 10^{13}$ cm^{-3}), bleibt die Elektronenkonzentration praktisch konstant und die Löcherkonzentration fällt weiter ab. Aufgrund der sehr niedrigen Grunddotierung ist zu erwarten, daß der Übergang zur schwachen Injektion erst bei sehr geringen Strömen zu beobachten ist, und somit bei der im folgenden beschriebenen Messung nicht in Erscheinung tritt. Die Zeitkonstante des exponentiellen Abfalls entspricht somit der ambipolaren Lebensdauer $\tau_p = \tau_n$ bei starker Injektion in der n-Basis.

Ausgeführt wurde die Messung, indem die Probe, mit einem Lastwiderstand in Reihe geschaltet, von einem Rechteckgenerator nebst HF-Verstärker getrieben wurde. Ein 4-Kanal-Speicheroszilloskop (Nicolet Pro 92) nimmt die Versorgungsspannung U_V und die Spannung am Lastwiderstand U_M auf, die direkt dem Sperrstrom I_A proportional ist.

Das Abklingen des Sperrstroms hat, wie oben beschrieben, nach einer Anfangsphase einen exponentiellen Verlauf. Die Zeitkonstante dieses Abfalls wurde durch einen Fit des Anodenstromes mit der Funktion $I_A(t) = I_1 \cdot \exp(t/\tau) + I_2$ bestimmt, dabei entspricht I_2 dem asymptotisch erreichten stationären Sperrstrom, während $I_1 - I_2$ dem Sperrstrom bei $t = 0$ entspricht. Da der exponentielle Abfall erst nach einer Anfangsphase einsetzt, wurde der Fit nur mit Daten aus einem entsprechenden Zeitintervall durchgeführt, so daß die Anfangsphase nicht zum Ergebnis beiträgt.

Die Messung der Abklingzeitkonstante τ wurde mit verschiedenen Lastwiderständen wiederholt, wobei sich die Abklingkurven nur in der Anfangsphase unterscheiden, die Endphase jedoch exponentiell mit praktisch gleichen Zeitkonstanten verläuft. Dies zeigt, daß es sich bei dem exponentiellen Abklingen des Sperrstromes nicht um den Ladevorgang eines RC-Gliedes handeln kann, sondern die Zeitkonstante des Abfalls mit der Lebensdauer in Verbindung steht.

Der Mittelwert von drei Messungen mit unterschiedlichen Lastwiderständen wird als Ergebnis dieser experimentellen Lebensdauerbestimmung angesehen.

Dieses Ergebnis stimmt für die bestrahlte Probe recht gut mit den aus der Bestrahlungsdosis geschlossenen Werten in Abschnitt 3.6.3 überein. Wegen der komplizierten Dynamik des Abschaltvorganges und der Abhängigkeit der Trägerlebensdauer

von Dotierungskonzentrationen, Injektionsverhältnissen und einigen anderen Faktoren können die gewonnenen Werte jedoch nicht als sehr genau angesehen werden.

3.8.3 Zündverzug

Wird bei fester Versorgungsspannung U_V und festem Lastwiderstand R_A zum Zeitpunkt $t = 0$ ein Gatestrom I_G eingeschaltet, so schaltet die Thyristorprobe in den eingeschalteten Zustand, vorausgesetzt der Gatestrom überschreitet einen Schwellenstrom I_{GT}. Bis zum Erreichen des eingeschalteten Zustand vergeht eine bestimmte Zeit. Diese Zündzeit wird gewöhnlich in die Zündverzugszeit t_{d} und die Schaltzeit t_s eingeteilt, wobei in der Zündverzugszeit die kritische Speicherladung in den Basen aufgebaut wird, bevor dann die regenerative Phase des Zündvorgangs beginnt. Diese Phase, in der die Anodenspannung rasch einbricht, kann wesentlich kürzer sein.

Die Zündverzugszeit nimmt mit zunehmendem Gatestrom ab, da die kritische Speicherladung entsprechend schneller aufgebaut werden kann. Umgekehrt divergiert die Zündverzugszeit, wenn der Gatestrom sich der Zündschwelle asymptotisch annähert.

In Abb. 3.25 ist die Zündverzugszeit in Abhängigkeit zum Gatestrom aufgetragen, wie sie für zwei Proben vom Typ II gemessen wurde. Die Anode wurde dabei über einen Vorwiderstand mit Gleichspannung versorgt, der Gatestrom wurde durch

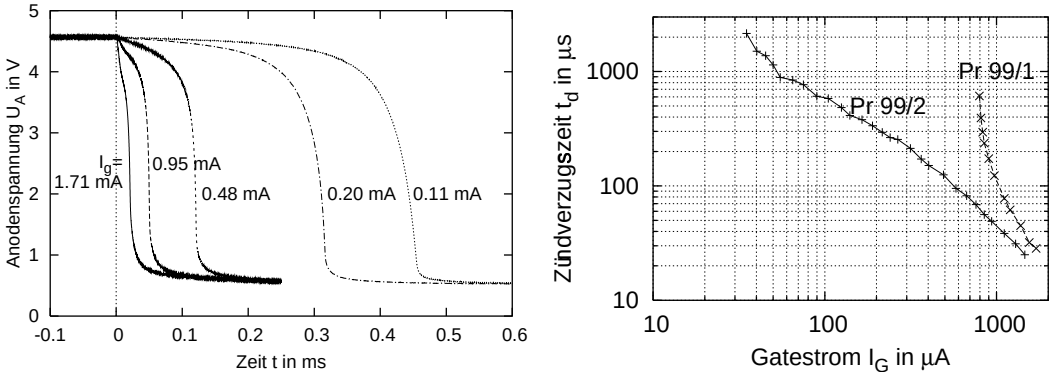


Abbildung 3.25: Links: Zündvorgang bei Einschalten eines Gatestromes zum Zeitpunkts $t = 0$. Probe 99/2 (unbestrahlt). Rechts: Zündverzug t_d in Abhängigkeit vom Gatestrom I_G, gemessen an Probe 99/1 (bestrahlt mit $7.5 \cdot 10^{12}$ $\mathrm{e}^-/\mathrm{cm}^2$) und Probe 99/2 (unbestrahlt). Triggerschwelle für Zünddetektion: 2.5 V, $U_V = 4.7$ V, $R_V = 200\ \Omega$.

eine spannungsgesteuerte Stromquelle erzeugt, die mit einem Rechteckimpuls versorgt wird. So wird der betreffende Gatekontakt mit einem sauber geschalteten Stromimpuls versorgt.

Nach Gerlach [Ger79] ist der Zusammenhang zwischen Zündverzugszeit t_d und Gatestrom I_G in erster Näherung eine umgekehrte Proportionalität:

$$t_d = t_0 + \frac{Q_{\mathrm{cr}}}{(I_G - I_{\mathrm{GT}})} \tag{3.7}$$

Dies folgt aus dem vereinfachten Modell, nach dem der den Schwellenstrom überschreitende Teil des Gatestromes $(I_G - I_{\mathrm{GT}})$ in den Basen eine kritische Speicherladung Q_{cr} aufbaut, nach deren Erreichen die Zündung selbständig und relativ rasch vollendet wird. Ein Fit der gemessenen Daten bestätigt ein solche umgekehrte Proportionalität recht gut.

Die Zündverzugszeit ist bedeutsam für die Anwendung der Thyristorstruktur zur Kopplung der SOM-Hardware, da sie die Reaktionszeit der Gewinnerdetektion begrenzt. Sie sollte möglichst kurz gehalten werden, was die Verwendung von Zündströmen nahe der Zündschwelle ungünstig macht.

3.8.4 Ortsaufgelöste Beobachtung der Frontausbreitung

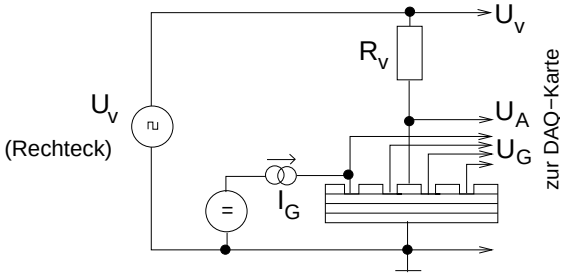
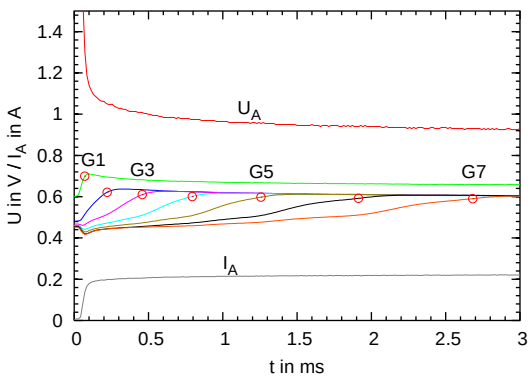
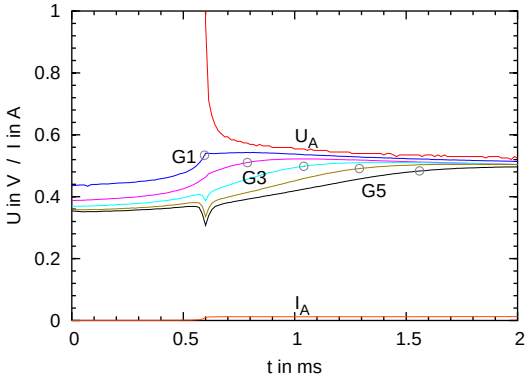


Abbildung 3.26: Schaltung zur elektrischen Frontdetektion.

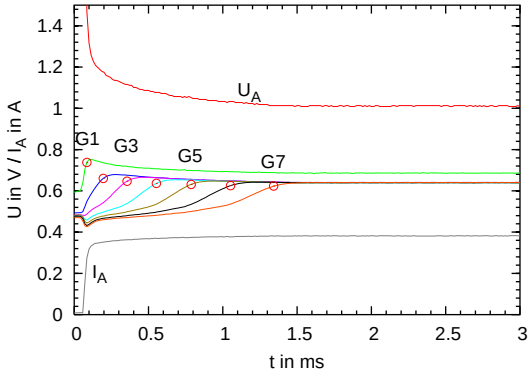
Der raumzeitliche Ablauf der Zündung wird mit folgender Beschaltung untersucht (s. Abb. 3.26): Ein konstanter, regelbarer Gatestrom wird an einem Gatekontakt eingeprägt, während die Anode mit einer Rechteckspannung versorgt wird. Der Anodenstrom wird dabei von einem Vorwiderstand begrenzt. Der zeitliche Verlauf von Versorgungsspannung, Anodenspannung den an mehreren Kontakten abgegriffenen Gatespannungen wird mittels eines Mehrkanaloszilloskops oder PC-gestützt mittels einer Meßwerterfassungskarte erfaßt. Die verwendete Karte erlaubt maximal 16 Känale mit einer Abtastrate von bis zu 1.5 Mhz je Kanal abzutasten; die Abtastrate wird auf die Kanäle aufgeteilt, da die Karte intern nur einen AD-Wandler besitzt, den sich die Kanäle über einen Multiplexer teilen. Die Datenerfassung wurde unter Linux über ein Treibermodul namens Comedi [Sch01] realisiert. Damit wurde bei 9 verwendeten Kanälen eine Abtastrate von 70 kHz realisiert.



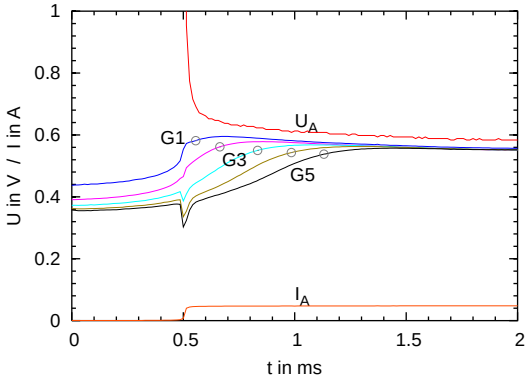
$U_V = 2.0$ V



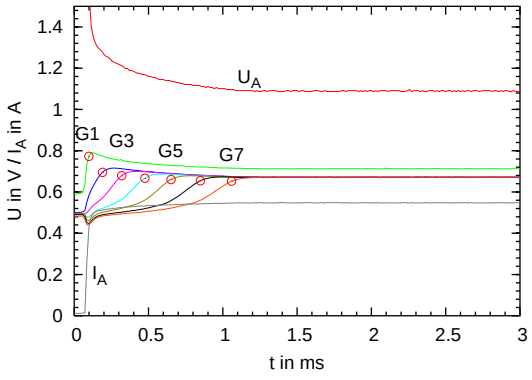
$R_V = 200\ \Omega$



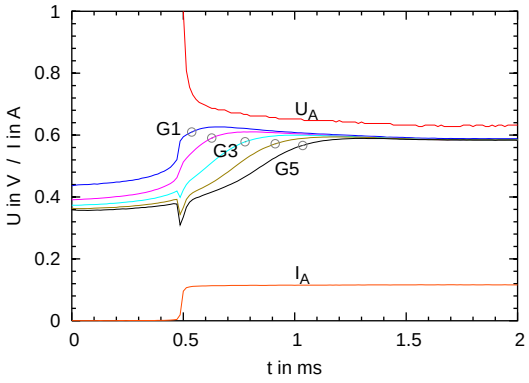
$U_V = 3.0$ V



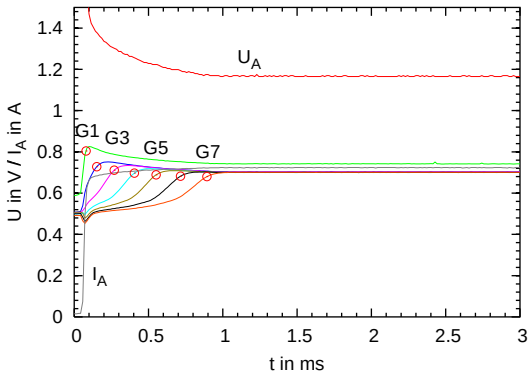
$R_V = 50\ \Omega$



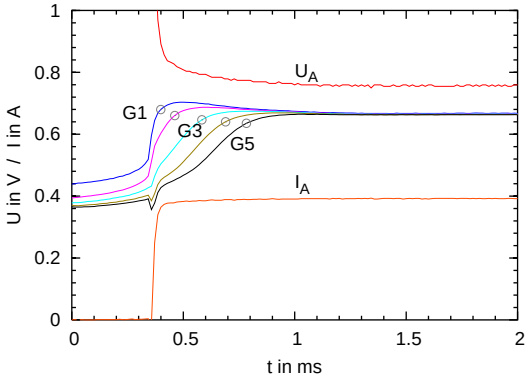
$U_V = 4.0$ V



$R_V = 20\ \Omega$



$U_V = 5.0$ V
$R_V = 5.0\ \Omega$



$R_V = 5\ \Omega$
$U_V = 3.0$ V

Abbildung 3.27: Zündfronten an Probe 99/3 (bestrahlt, links) und Probe 99/4 (unbestrahlt, rechts) , aufgenommen mittels DAQ-Karte und PC, Abtastrate 70 kHz mit 9 Kanälen. Von oben nach unten wächst der Anodenstrom, wobei dies links durch Variation der Versorgungsspannung U_V geschieht, rechts bei konstanter Versorgungsspannung durch Variation des Lastwiderstandes R_V.

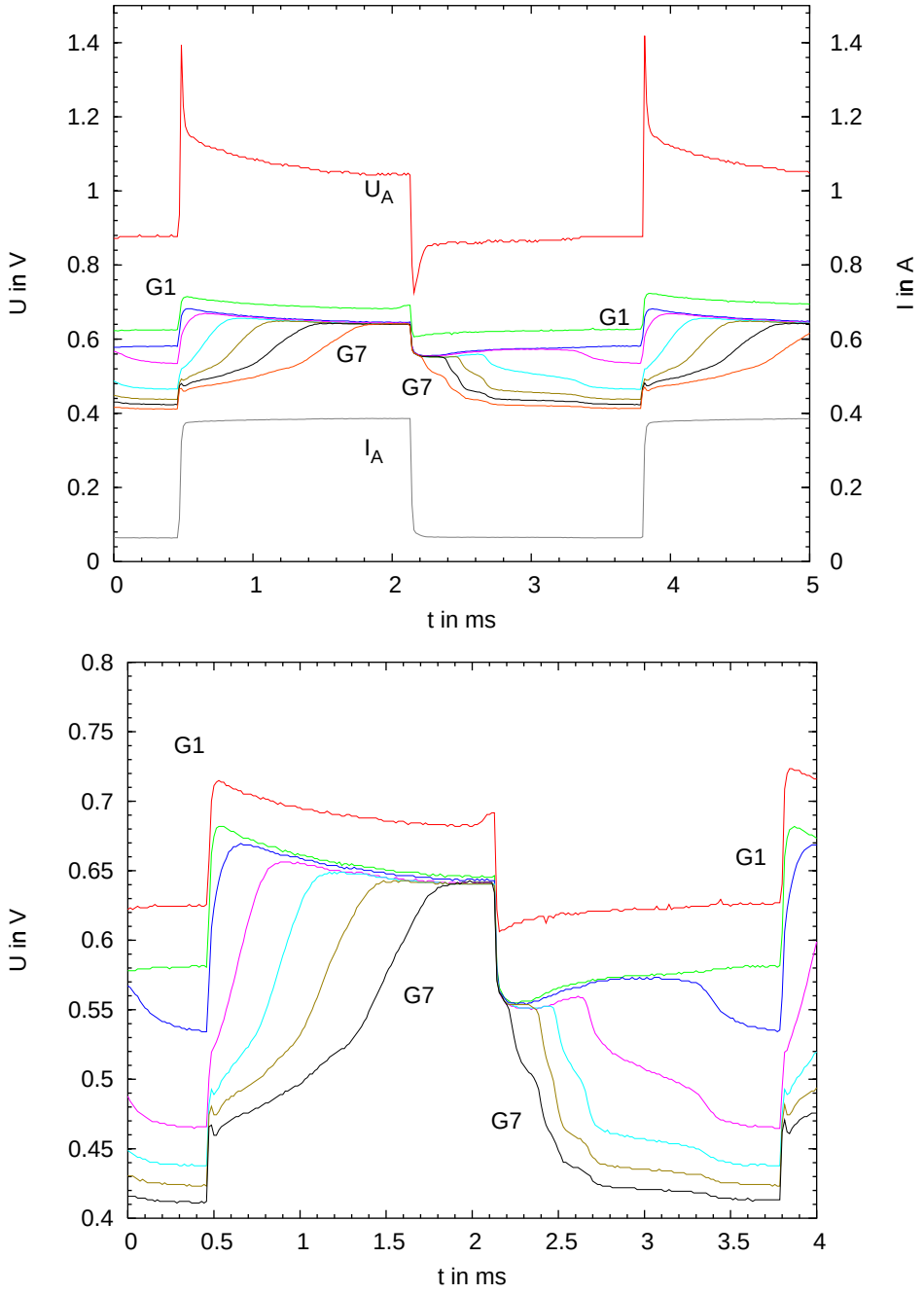


Abbildung 3.28: Frontausbreitung auf Probe 99/3, mit Elektronen bestrahlt. Die Probe wurde zwischen einem hohen und einem niedrigen Anodenstrom hin- und hergeschaltet. Dabei erkennt man Fronten, die entsprechend der Ausdehnung der Hochstromdomäne jeweils vor und zurück laufen. Dies zeigt, daß auch negative Frontgeschwindigkeiten möglich sind, daß also der Ausschaltvorgang genauso durch Frontausbreitung ablaufen kann wie der Einschaltvorgang. Allerdings kann der Abschaltvorgang auch homogen stattfinden, wenn der die vorhandene Stromdichte abrupt unter die Haltestromdichte fällt. An unbestrahlten Proben konnten keine rückwärts laufenden Fronten beobachtet werden.

Abhängig von Gatestrom I_G und Versorgungsspannung U_V, beginnt der Zündvorgang nach dem Nulldurchgang der Versorgungsspannung. Die Anodenspannung bricht nach Ablauf der Zündverzugszeit ein, während der Anodenstrom in einer Anstiegszeit von etwa 100 µs bis fast auf seinen Maximalwert ansteigt. Dabei beschränkt sich der Stromfluß zunächst auf einen leitenden Kanal in der Umgebung des Gatekontaktes, an dem die Zündung ausgelöst wurde. Danach breitet sich die leitende Zone über die Fläche des Thyristors aus. Der lokale Zustand der Thyristors, also die lokale Stromdichte j, spiegelt sich im Potential der Gatekontakte wieder. Die Gatepotentiale steigen nacheinander auf einen erhöhten Wert, je nach fleißender Stromdichte ca. 0.7 V. Dieser Anstieg erfolgt umso später, je weiter der betrachtete Kontakt zum Zündgate entfernt liegt. Dies gilt unabhängig davon, an welchem Gate der Zündstrom eingeprägt wird. Man beobachtet also tatsächlich laufende Fronten, die von dem Gatekontakt ausgehen, an dem der Zündstrom eingeprägt wird.

Der Anstieg der Gatepotentiale (s. Abb. 3.27) wird mit wachsendem Abstand zum Zündgate immer flacher. Dies wurde schon bei den optischen Beobachtungen festgestellt. Die Ursache für die Abnahme der Frontgeschwindigkeit ist die Gegenkopplung durch den (globalen) Vorwiderstand R_A. Der Spannungsabfall in diesem Vorwiderstand läßt die Anodenspannung im Verlauf der Zündung absinken, während sich der Strom erhöht. Dies läßt auch die Frontgeschwindigkeit sinken, im Extremfall bis auf Null. Die ideale Frontausbreitung mit konstanter Frontgeschwindigkeit und konstanter Anodenspannung war auch ohne Vorwiderstand nicht zu realisieren, da selbst kleine Kontaktwiderstände störend wirken, und der differentielle Widerstand des durchgeschaltete Thyristors sehr niedrig ist. Eine kleine Änderung der Anodenspannung (wenige mV) hat bereits einen großen Einfluss auf die sich einstellende Stromdichte und das Verhalten der Zündfront.

Während der Potentialanstieg bei der unbestrahlten Probe praktisch an allen Gatekontakten gleichzeitig beginnt und nur mit unterschiedlicher Steigung abläuft, zeigt die bestrahlte Probe ein deutlicheres Schaltverhalten. Das Gatepotential bleibt praktisch unverändert, bis die Front einen Kontakt erreicht, um dann relativ rasch anzusteigen. Die Frontbreite scheint also stark vermindert zu sein. Abb. 3.28 zeigt, daß auch rückwärts laufende Fronten möglich sind, indem der Anodenstrom plötzlich verringert wird. Dadurch wird die Stromdichte einer stationären Front unterschritten, so daß das eingeschaltete Gebiet schrumpfen muß, bis sich ein neues Gleichgewicht einstellt.

3.8.5 Bestimmung der Frontgeschwindigkeit

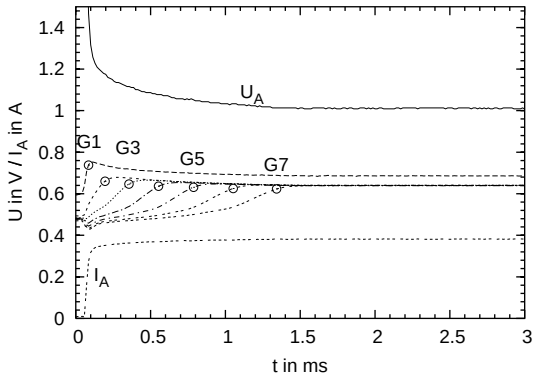
Die Geschwindigkeit der Zündfronten lässt sich wegen der relativ großen und variablen Frontbreite nur mit einer relativ großen Unsicherheit angeben. Dazu muß ein Zeitpunkt festgelegt werden, den man als Durchlaufzeit der Front an einem bestimmten Gatekontakt definiert. Dazu kann man das Überschreiten eines (willkürlichen) Schwellwerts verwenden.

Wegen des asymmetrischen Potentialverlaufs mit einem wagen Frontbeginn und einem scharf begrenzen Ende der Front wird der Frontdurchlauf nahe dem Erreichen des eingeschalteten Zustands detektiert. Der Schwellwert wird auf 90 % des maximal erreichten Gatepotentials festgelegt. Abbildung 3.29 zeigt den Potentialverlauf an 7 Gatekontakten einer Probe mit Markierung der Durchlaufzeiten, der auf diese Weise erfaßt wurde.

Abbildung 3.30 zeigt die Bewegung der Front in reduzierter Form, bei Zündung der Front an verschiedenen Positionen. Der Durchlaufzeitpunkt t_i ist gegen die Position x_i des betreffenden Kontaktes i aufgetragen. Es ist zu erkennen, wie die Geschwindigkeit der Fronten mit zunehmender Ausbreitung abnimmt. Ansonsten erscheint die Darstellung praktisch translationsinvariant.

Aus den wie oben beschriebenen Durchlaufzeiten der Front an den einzelnen Kontakten wird die Frontgeschwindigkeit berechnet. Da sie im Verlauf der Frontausbreitung nicht konstant ist, lassen sich verschiedene Geschwindigkeiten bestimmen: Die „momentane“ Geschwindigkeit, die sich aus der Zeitdifferenz des Durchlaufs durch zwei benachbarte Kontakte schätzen läßt, und die mittlere Geschwindigkeit, die aus der Gesamtlaufzeit bestimmt wird.

Abbildung 3.29: Erkennung des Frontdurchlaufs mittels Schwellwert, Probe 99/3, Parameter: $U_V = 3.0$ V, $R_v = 5.0\ \Omega$



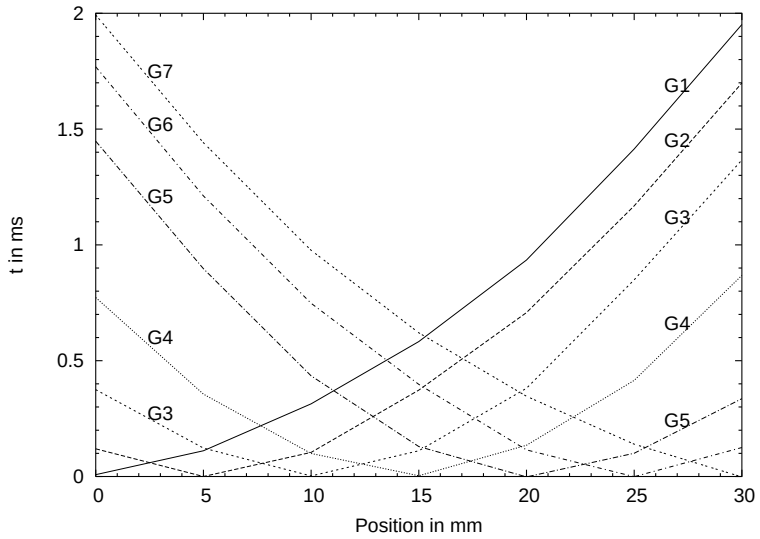


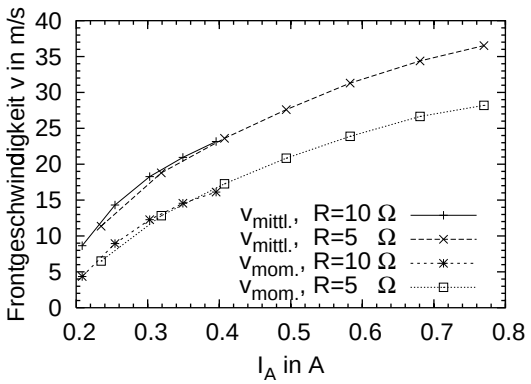
Abbildung 3.30: Ausbreitung von Fronten ausgehend von verschiedenen Gatekontakten. Dargestellt ist jeweils der Zeitpunkt des Frontdurchlaufs als Funktion des Ortes. Probe 99/3, bestrahlt , Parameter: $U_V = 4.1$ V, $R_V = 10.0\ \Omega$, Kontaktabstand 5 mm.

In Abbildung 3.31 sind die mittlere und momentane Frontgeschwindigkeit $v_{\text{mittl.}}$ und $v_{\text{mom.}}$ gegen den Anodenstrom I_A aufgetragen. Die momentane Geschwindigkeit wird hier zwischen dem vorletzten und letzten Gatekontakt bestimmt.

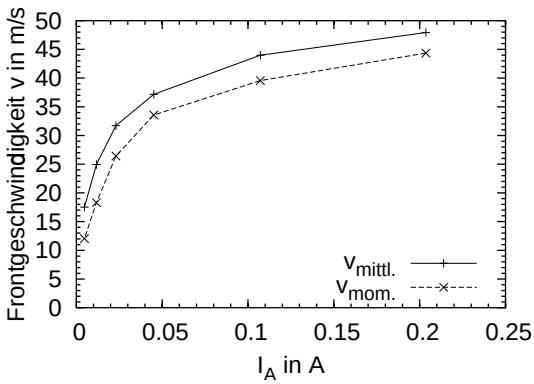
Wie zu erwarten nimmt die Frontgeschwindigkeit mit I_A zu, jedoch ist mit zunehmendem Strom ein Sättigungseffekt zu erkennen. Die beiden Äste des Graphen, die mit unterschiedlichen Vorwiderständen und dem gleichen Intervall der Versorgungsspannung U_V entstanden sind, zeigen, daß die Frontgeschwindigkeit praktisch allein vom Anodenstrom I_A, bzw. der Stromdichte j abhängt, und praktisch unabhängig von der Wahl des Vorwiderstands ist.

Die Frontbreite wurde für unbestrahlte und bestrahlte Proben folgendermaßen näherungsweise bestimmt: Für einen Gatekontakt wird die Schaltzeit t_s bestimmt als die Zeit, in der das Kontaktpotential von 50% bis auf 90% zunimmt. Dann wird die Frontbreite durch Multiplikation mit der vorher bestimmten Momentangeschwindigkeit bestimmt, also $w = t_s \cdot v_{\text{mom.}}$. Für die unbestrahlten Probe 99/4 ergab sich eine Frontbreite von $9-10$ mm, für die bestrahlten Proben ein Wert von $3.5 - 4.5$ mm. Diese Werte sind kaum vom eingeprägten Anodenstrom abhängig. Die Bestrahlung verkleinerte die Frontbreite also etwa um den Faktor 2.

Frontgeschwindigkeit und Frontbreite sind streng genommen nur für den Fall einer Front definiert, die eine reine Translationsbewegung ausführt. Da die gemessenen



Probe 99/3, bestrahlt
Parameter:
$U_V = 2 - 5$ V, $R_v = 5.0, 10.0\ \Omega$



Probe 99/4,unbestrahlt
Parameter:
$U_V = 3$ V, $R_v = 10.0 - 500.0\ \Omega$

Abbildung 3.31: Frontgeschwindigkeit in Abhängigkeit von I_A, im Vergleich an einer bestrahlten und einer unbestrahlten Probe.

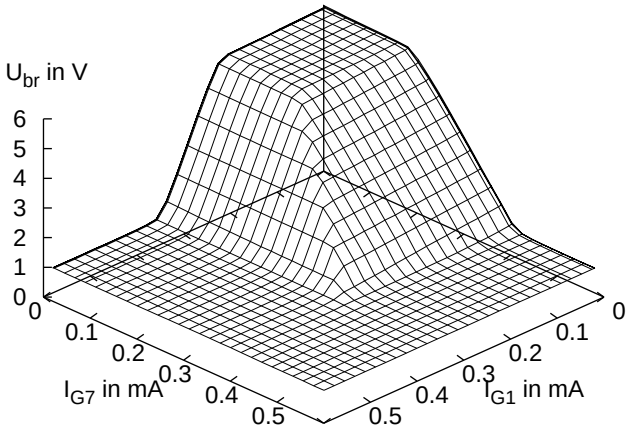
Schaltfronten sich während der Ausbreitung stark verbreitern, sind die mit der oben beschriebenen Schwellwertmethode ermittelten Frontgeschwindigkeiten und Frontbreiten mit großen Fehlern behaftet. Dies sind allerdings systematische Fehler, die bei allen Messungen gleich auftreten, so daß sie den relativen Vergleich von Geschwindigkeiten bzw. Frontbreiten zulassen.

3.8.6 Wechselwirkungen zwischen zwei Gatekontakten

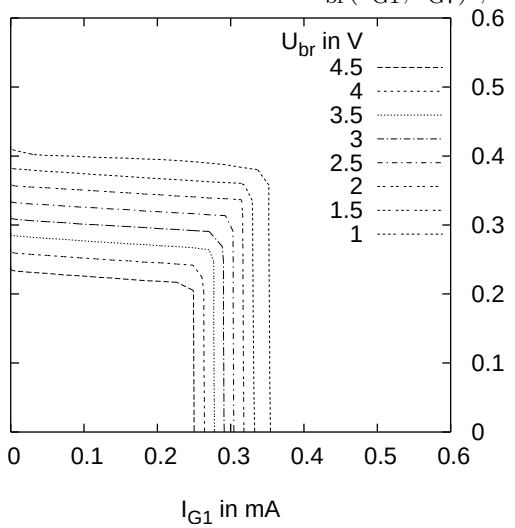
Für die Anwendung als Kopplungsstruktur eines neuronalen Netzes ist die Unabhängigkeit der Gatekontakte voneinander entscheidend: Die Zündung soll eingeleitet werden, sobald der Zündstrom an einem Gate den Schwellen-Zündstrom I_{th} überschreitet, unabhängig von „unterschwelligen“ Strömen an anderen Gates.

In der Realität ist dies nicht der Fall, wird ein kleiner Strom an einem Gatekontakt eingeprägt, so senkt dies die Zündschwelle an einem benachbarten Kontakt, umso mehr, je näher die beiden Kontakte benachbart sind. Dies rührt offenbar daher, daß der über einen Punkt in die p-Basis eingeprägte Strom sich über einen gewissen Radius verteilt, um dann in den n-Emitter abzufließen.

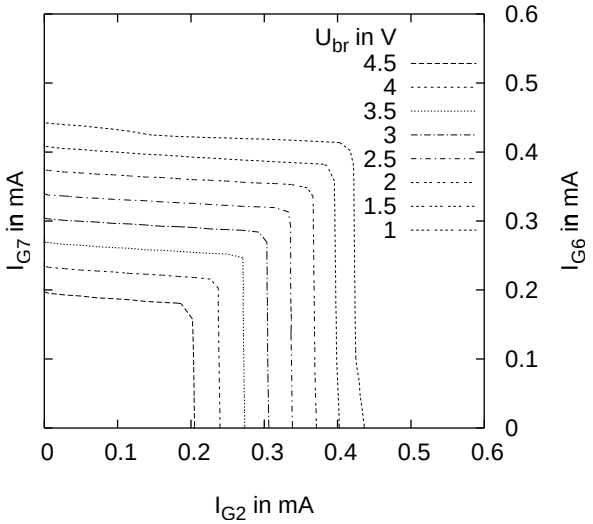
Dieses Verhalten ist offensichtlich für die Anwendung zur Gewinnersuche ungünstig, da das Übersprechen von Gateströmen dazu führen kann, daß nicht der größte Gatestrom zur Zündung führt: Prägt man z.B. drei Gateströme ein, wobei zwei Gatekontakte eng benachbart, der dritte aber weiter entfernt ist, so können die



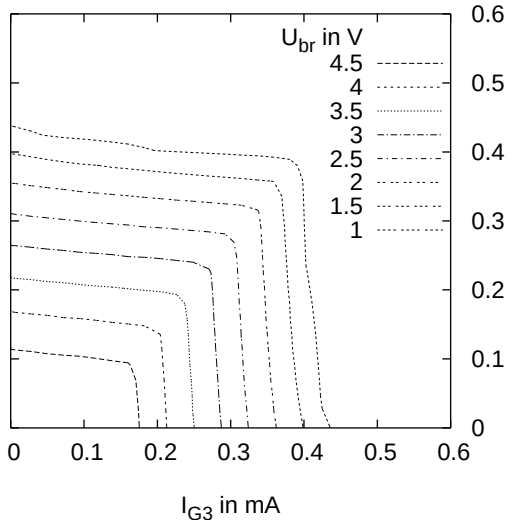
$U_{\mathrm{br}}(I_{\mathrm{G1}}, I_{\mathrm{G7}})$, Gateabstand $d = 30$ mm



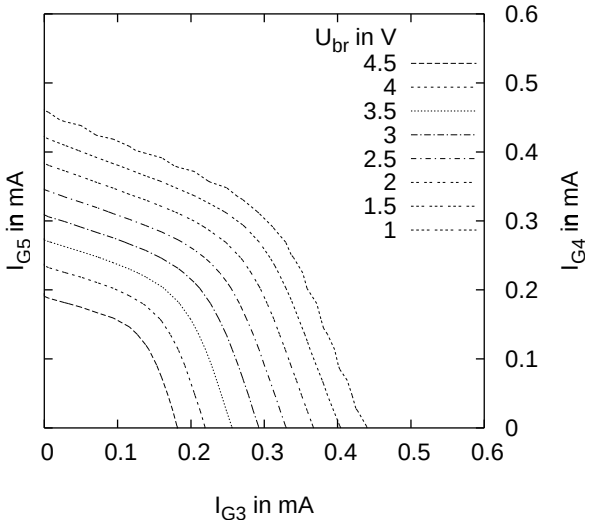
$U_{\mathrm{br}}(I_{\mathrm{G1}}, I_{\mathrm{G7}})$, $d = 30$ mm



$U_{\mathrm{br}}(I_{\mathrm{G2}}, I_{\mathrm{G6}})$, $d = 20$ mm



$U_{\mathrm{br}}(I_{\mathrm{G3}}, I_{\mathrm{G5}})$, $d = 10$ mm



$U_{\mathrm{br}}(I_{\mathrm{G3}}, I_{\mathrm{G4}})$, $d = 5$ mm

Abbildung 3.32: Wechselwirkung zwischen mehreren Gateströmen, dargestellt als Konturplots der Kippspannung U_{br} als Funktion der an 2 Gatekontakten eingeprägten Ströme. Bei großem Abstand d der Gatekontakte wird die Kippspannung praktisch nur vom größeren der beiden Gateströme bestimmt. Bei kleinem Abstand der Gatekontakte trägt jeder Gatestrom auch einen Anteil zur Zündung am jeweils anderen Gatekontakt bei.

G1	G2	d in mm	a/b
3	4	5	0.323
3	5	10	0.0778
3	6	15	0.0285
2	5	15	0.0275
2	6	20	0.0158
1	6	25	0.0135
1	7	30	0.0107

Probe 99/3, bestrahlt

G1	G2	d in mm	a/b
4	5	5	0.476
3	5	10	0.147
2	5	15	0.094
2	6	20	0.055
1	7	30	0.018

Probe 99/3, unbestrahlt

Tabelle 3.1: Übersprechen von einem Gatekontakt auf den anderen in Abhängigkeit vom Kontaktabstand. Gemessen an Probe 99/3 (streifenförmige Geometrie), einmal unbestrahlt, dann nach der Bestrahlung mit $1.0 \cdot 10^{13}$ $e^- cm^{-2}$.

Gateströme 1 und 2 zusammen eher die Zündung auslösen als der dritte, obwohl beide kleiner sind.

Die Größenordnung der Wechselwirkung zwischen den Gatekontakten wurde nun untersucht, indem die Kippspannung U_{BR} als Funktion zweier Gateströme I_{G1} und I_{G2} bestimmt wurde.

Die Messung wurde mit Hilfe einer DAQ-Karte und eines PCs automatisiert. I_{G1} und I_{G2} von der DAQ-Karte mit Hilfe von spannungsgesteuerten Stromquellen erzeugt. Beide Gateströme werden innerhalb eines vorgegebenen Intervalls variiert, es wird also die von I_{G1} und I_{G2} gebildete Parameterfläche abgerastert. Die Anode des Thyristors wird über einen Vorwiderstand mit einer Dreieckspannung versorgt ($f = 50$ Hz). Ein AD-Wandler nimmt die Anodenspannung U_A mit ausreichender Abtastrate auf, die Kippspannung U_{BR} wird dann bei jeder Einzelmessung als der höchste erreichte Wert von U_A bestimmt.

Dann läßt sich die Kippspannung als Funktion der beiden Gateströme in Form einer dreidimensionalen Grafik oder in Form von Isolinien mit gleichem U_{BR} in der $(I_{\mathrm{G1}}, I_{\mathrm{G2}})$-Ebene darstellen. In Abbildung 3.32 ist diese Darstellung zu sehen. Man erkennt die Wechselwirkung der Gateströme gemessen bei verschiedenen Abständen der beiden Gatekontakte. Für große Abstände findet man nahezu ideales Verhalten der Proben: Die Isolinien sind praktisch horizontal bzw. vertikal mit einem scharfen Knick in der Diagonalen $I_{\mathrm{G1}} = I_{\mathrm{G2}}$. Die Zündung wird also stets vom größeren Gatestrom ausgelöst, der kleinere Gatestrom hat praktisch keinen Einfluß auf die Zündung.

Bei kleineren Abständen steigt das Übersprechen des kleineren zum größeren der beiden Gateströme, die Isolinien weichen von der Horizontalen bzw. Vertikalen ab, und der Übergang in der Diagonalen wird abgerundet.

Als Maß für die Stärke des Übersprechens wird die Steigung der Isolinien in einem Bereich fernab von der Diagonalen verwendet, die über einen linearen Fit bestimmt wird. $U_{\mathrm{BR}}(I_{\mathrm{G1}}, I_{\mathrm{G2}})$ läßt sich recht gut durch eine lineare Funktion der Form $aI_{\mathrm{G1}} + bI_{\mathrm{G2}} + c$ approximieren, wenn jeweils nur einer der beiden Teilbereiche verwendet wird, die in Abb. 3.32 in der dreidimensionalen Grafik gut in Form von geneigten Ebenen zu erkennen sind. Das Verhältnis a/b (b/a) gibt nun im Bereich vertikaler (horizontaler) Isolinien das Verhältnis der Einflüsse des kleineren und des größeren Zündstromes auf die Kippspannung an. Dieses Maß ist in Tabelle 3.1 für verschiedene Abstände der Gatekontakte angegeben. Die Bestrahlung der Proben führt zu einer Verringerung des Übersprechens, verbessert also die Unabhängigkeit der Gatekontakte. Trotzdem ist die Wechselwirkung beim kleinsten Abstand von 5 mm noch beträchtlich.

3.9 Thyristorstrukturen zur Kopplung neuronaler Hardware

Einige Eigenschaften der Schaltfronten auf Thyristorstrukturen sind für den Zweck der Kopplung einer Hardwareimplementation der SOM besonders entscheidend. Dazu gehören die Frontbreite, die Zündwechselwirkung, die Frontgeschwindigkeit und die Haltestromdichte.

Die Frontbreite der Potentialverteilung in der p-Basis bestimmt die Genauigkeit, mit der der Frontdurchgang detektiert werden kann. Bei den vorhandenen Proben liegt diese Frontbreite in der Größenordnung 5 mm, so daß ein verkleinerter Kontaktabstand z.B. im Submillimeterbereich kaum sinnvoll wäre.

Auch die Wechselwirkung benachbarter Gatekontakte hat einen negativen Einfluß auf die Funktion des Kohonen-Algorithmus, insbesondere auf die Gewinnerdetektion. Die Gate-Zündschwelle wird ja als Komparator bei der Gewinnerdetektion verwendet. Wechselwirken nun zwei benachbarte Gates, so wird die Zündschwelle bei gleichzeitiger Erregung zweier Gates erniedrigt. Es ist also möglich, dass zwei benachbarte ähnliche Prototypen sich zusammen gegen einen einzelnen Prototypen durchsetzen, der „eigentlich“ dem Mustervektor näher liegt. Dadurch kann sowohl der Lernvorgang als auch die Klassifikation im ausgelernten Zustand beeinträchtigt werden.

Die Zündverzugszeit begrenzt die für die Klassifikation eines Mustervektors benötigte Zeit, es ist also wünschenswert, daß der Zündvorgang möglichst schnell abläuft. Die Verwendung möglichst großer Zündströme verringert zwar die Zündverzugszeit, erhöht aber den Leistungsbedarf des Systems.

In den vorherigen Abschnitten wurde gezeigt, daß eine Bestrahlung der Thyristorstukturen durch Absenkung der Trägerlebensdauer zur Verringerung der Frontbreite führt, was die Detektion des Frontdurchlaufs erleichtert. Jedoch erhöht sich dabei die Haltestromdichte unhältnismäßig, so daß der Leistungsbedarf der Thyristorstrukturen stark zunimmt.

Abhilfe kann nur eine Veränderung der Thyristorstruktur schaffen, insbesondere eine Minaturisierung der Struktur mit verkleinerten Schichtdicken im µm-Bereich, um eine mit integrierten Schaltungen kompatible Struktur zu erhalten. Wege in diese Richtung zu weisen ist das Ziel des nächsten Abschnitts, in dem Ergebnisse von Simulationen vorgestellt werden, die Frontausbreitungsprozesse in miniaturisierten Thyristorstrukturen modellieren.

Kapitel 4

Simulationen von Thyristorschaltfronten

In diesem Kapitel werden Simulationen vorgestellt, die mit dem kommerziellen Halbleitersimulator `ISE TCAD` unternommen wurden, um Möglichkeiten zur Optimierung der Frontausbreitung in Thyristorstrukturen zu untersuchen, insbesondere im Hinblick auf eine Miniaturisierung und Integration mit den restlichen Komponenten eines analogen neuronalen Prozessors. Die Ziele der Optimierung betreffen insbesondere eine möglichst kleine Frontbreite, also einen scharfen Übergang zwischen ein- und ausgeschaltetem Zustand, und eine hohe Frontgeschwindigkeit, mit der Nebenbedingung einer möglichst geringen Stromdichte.

4.1 Physikalisches Modell der Halbleiterbauelemente

In diesem Abschnitt werden die Modellgleichungen vorgestellt, die der Device-Simulator `dessis` verwendet, um die Vorgänge in einem Halbleiterbauelement zu simulieren [des00]. Die folgende Darstellung beschreibt dabei nur das Drift-Diffusions-Modell, das stets in den Simulationen der Vierschichtstrukturen eingesetzt wurde. Im Zentrum des Modells stehen die Poissongleichung und die Kontinuitätsgleichungen für Elektronen und Löcher.

ψ	Elektrostatisches Potential
p	Löcherkonzentration
n	Elektronenkonzentration
N_D	Donatorenkonzentration
N_A	Akzeptorenkonzentration
$\vec{J}_n$	Elektronenstromdichte
$\vec{J}_p$	Löcherstromdichte
R	Rekombinationsrate
μ_n	Elektronenbeweglichkeit
μ_p	Löcherbeweglichkeit
ϕ_n	Quasifermipotential der Elektronen
ϕ_p	Quasifermipotential der Löcher
$n_{i,\text{eff}}$	Effektive intrinsische Trägerkonzentration

Tabelle 4.1: Erläuterung der Symbole in den Modellgleichungen von `dessis`

$$\nabla\epsilon \cdot \nabla\psi = -q(p - n + N_D - N_A) \tag{4.1}$$

$$\nabla \cdot \vec{J}_n = qR + q\frac{\partial n}{\partial t} \tag{4.2}$$

$$-\nabla \cdot \vec{J}_p = qR + q\frac{\partial p}{\partial t} \tag{4.3}$$

Die Poissongleichung beschreibt die Entstehung elektrischer Felder, bzw. Potentiale durch eine Verteilung elektrischer Ladungen. Die Kontinuitätsgleichungen beschreiben die Ladungserhaltung. Die Bauteildefinition enthält die Verteilung der ionisierten Donatoren und Akzeptoren N_D und N_A, also das Dotierungsprofil. Das Drift-Diffusions-Modell beschreibt den Transport der Ladungsträger aufgrund von Feldern (Drift) und Dichtegradienten (Diffusion).

$$\vec{J}_n = -nq\mu_n\nabla\phi_n \tag{4.4}$$

$$\vec{J}_p = -pq\mu_p\nabla\phi_p \tag{4.5}$$

ϕ_n und ϕ_p sind die Quasi-Fermipotentiale der Löcher und Elektronen, die mit den Trägerdichten n und p entsprechend der Fermi- oder Boltzmannstatistik in Zusammenhang stehen. Die Boltzmannstatistik reicht für den Fall nichtentarteter Bänder aus, also bei nicht zu hoher Dotierung, und ergibt folgende Gleichungen:

$$n = n_{i,\text{eff}} \cdot \exp(\frac{-q(\phi_n - \psi)}{kT}) \tag{4.6}$$

$$p = n_{i,\text{eff}} \cdot \exp(\frac{q(\phi_p - \psi)}{kT}) \tag{4.7}$$

μ_n und μ_p sind die Beweglichkeiten der beiden Trägerspezies. Durch Streuung der Träger mit Phononen und mit den Dotieratomen entsteht eine Abhängigkeit der Beweglichkeiten von Temperatur und Dotierung, die von verschiedenen Modellen näherungsweise beschrieben wird. Hier wurde das *Philips Unified Mobility Model* [Kla92] verwendet.

Ein weiterer wichtiger Parameter ist die Rekombinationsrate R. Sie wird nach *Shockley-Read-Hall* als Rekombination an tiefen Störstellen modelliert:

$$R_{\text{net}}^{\text{SRH}} = \frac{np - n_{i,\text{eff}}^2}{\tau_p(n + n_1) + \tau_n(p + p_1)} \tag{4.8}$$

mit

$$n_1 = n_{i,\text{eff}} e^{\frac{E_{\text{trap}}}{kT}} \tag{4.9}$$

$$p_1 = n_{i,\text{eff}} e^{\frac{-E_{\text{trap}}}{kT}} \tag{4.10}$$

Dabei wurde das Energieniveau der Störstellen als $E_{trap} = 0$ gesetzt. Die Minoritätsträgerlebensdauern τ_n und τ_p werden als Produkt eines dotierungsabhängigen Teils, eines feldabhängigen Teils und eines temperaturabhängigen Teils modelliert:

$$\tau_c = \tau_{\text{dop}} \frac{f(T)}{1 + g_c(F)}, \; c = n, p \tag{4.11}$$

Die Dotierungsabhängigkeit wird mittels der empirischen *Scharfetter*-Relation modelliert:

$$\tau_{\text{dop}} = \tau_{\text{min}} + \frac{\tau_{\text{max}} - \tau_{\text{min}}}{1 + (N_i/N_{\text{ref}})^\gamma} \tag{4.12}$$

Dies führt für Dotierungskonzentrationen oberhalb von N_{ref} zu einer fortschreitenden Reduktion der Lebensdauer. Die Lebensdauer in einer Silizium-Probe kann

jedoch je nach Präparation stark variieren, da sie von vielen Faktoren beeinflußt wird. Fossum [Fos76, Fos82] stellt ein theoretisches Modell für die Dotierungsabhängigkeit der Lebensdauer auf, das von der Annahme ausgeht, die entscheidenden Rekombinationszentren seien Divakanzen, deren Löslichkeit im Silizium durch die Dotierung erhöht wird. Das Modell deutet auch an, daß die Konzentration von Störstellen stark durch den Herstellungsprozeß des verwendeten Materials, insbesondere durch Tempervorgänge, beeinflußt wird. Die Scharfetter-Relation kann daher nur eine empirische Näherung angesehen werden, die sich als brauchbar erwiesen hat. `Dessis` setzt die Scharfetter-Parameter auf folgende Standardwerte:

Parameter	Elektronen	Löcher	Einheit
$\tau_{\min}$	0	0	s
$\tau_{\max}$	$1 \cdot 10^{-5}$	$3 \cdot 10^{-6}$	s
N_{ref}	$1 \cdot 10^{16}$	$1 \cdot 10^{16}$	cm^3
γ	1	1	1

Diese Werte wurden in einigen Simulationsrechnungen verändert, um der Einfluß der Lebensdauer auf Kennlinien und Frontausbreitung zu untersuchen.

4.2 Ablauf der Simulationen

Die Simulation von Halbleiterbauelementen mit `ISE TCAD` läuft in mehreren Schritten ab. Zunächst wird eine Bauteilgeometrie definiert, mit den Kontaktflächen an

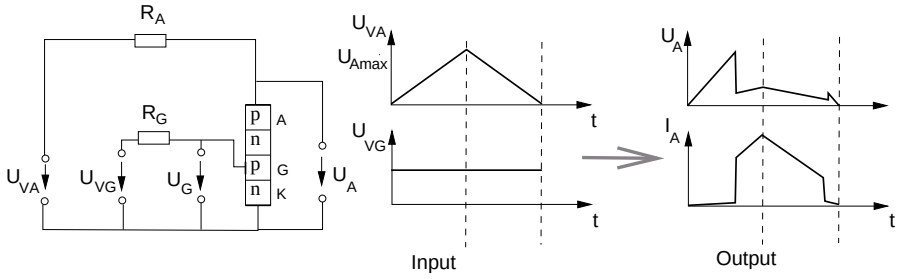


Abbildung 4.1: Simulation einer I-U-Kennlinie mit `ISE TCAD`. Die Probe ist quasieindimensional, R_G und U_{VG} werden groß gewählt, um das Gate in Stromeinprägung zu betreiben. Die die simulierte Zeit t, in der die Kennlinie durchlaufen wird, ist lang gegenüber der typischen Schaltzeit. Simulationsverlauf: Gateversorgungsspannung U_{VG} ist konstant, U_{VA} wächst zunächst linear an und fällt dann wieder bis auf Null ab. Abhängig von I_G zündet das Bauteil in der ersten Phase und geht bei Unterschreiten des Haltestromes wieder in den ausgeschalteten Zustand über.

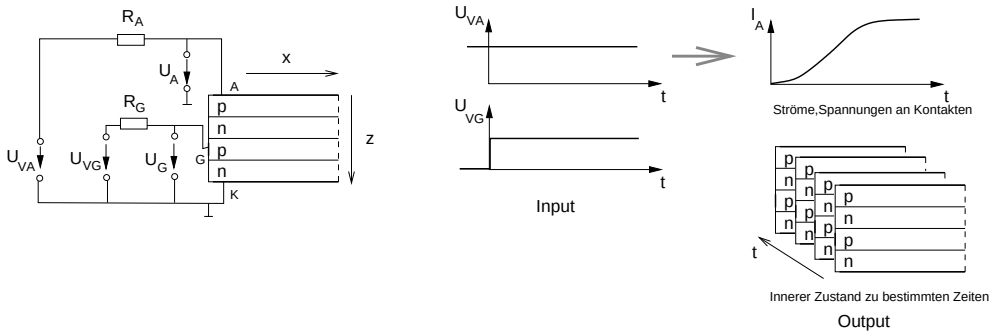


Abbildung 4.2: Simulation der Frontausbreitung. Die Probe ist zweidimensional, R_G und U_{VG} werden groß gewählt, um das Gate mit Stromeinprägung zu betreiben, $R_A = 0$, so daß U_A = const (Spannungseinprägung). Die simulierte Zeit t entspricht etwa der Laufzeit der Front. Simulationsverlauf: U_{VG} und U_A sind konstant, die Zündung setzt am Gatekontakt bei $x = 0$ ein und breitet sich über die Probe aus.

den Rändern und dem Dotierungsprofil, das die Verteilung der Dotierstoffe im Bauteilvolumen beschreibt. Diese Definitionen werden in einer symbolischen Kommandosprache abgefaßt und vom Gittergenerator `mesh` zu einem diskreten Gittermodell verarbeitet. In den hier beschriebenen Modellrechnungen wurden zunächst nur rechteckige, zweidimensionale Geometrien simuliert, die z-Koordinate des Grundgebietes entspricht dabei der Tiefendimension, und die x-Koordinate der lateralen Ausdehnung. Die Dotierung ist im einfachsten Fall nur abhängig von der Tiefe z. Anoden- und Kathodenkontakt sind der obere und untere Rand des Grundgebiets, der Laststrom fließt vertikal in z-Richtung. Die Modelle werden zweidimensional simuliert, aber zur Bestimmung der Kontaktflächen wird eine „Dicke“ in y-Richtung von 1 µm impliziert, die auch verändert werden kann. Der Mesher erzeugt ein Gitter mit variabler Dichte, wobei die Dichte an den Dotierungsgradienten angepaßt wird: Gebiete mit starkem Gradienten, wie die Umgebung von pn-Übergängen, werden feiner diskretisiert. Außerdem können von Hand Regionen definiert werden, in denen von vornherein mit größerer Dichte diskretisiert wird.

Das Simulationstool `dessis` simuliert dann die diskretisierten Modellgleichungen für das vorher definierte Bauteil. Dazu werden in einer Kommandodatei die elektrischen Randbedingungen, im einfachsten Fall die an den Kontaktflächen anliegende Spannungen und die Kontaktwiderstände definiert. Ein zeitlicher Verlauf der angelegten Spannungen kann definiert werden. In einer sogenannten *mixed-mode*-Simulation kann ein externer Schaltkreis mit diskreten Bauteilen an ein oder mehrere detailliert simulierte Bauelemente angekoppelt werden. Die zur Definition des Schaltkreises verwendete Syntax lehnt sich an die des Schaltungssimulators `SPICE` an, die diskreten Bauteile werden durch einzelne Parameter beschrieben,

die Verbindungen der Bauteile durch eine Netzliste. So kann ein zu untersuchendes „virtuelles“ Bauelement im Zusammenhang einer externen Schaltung simuliert werden.

Die oben beschriebenen Halbleitergleichungen können quasistationär gelöst werden, wobei die Zeitableitungen in den Modellgleichungen gleich Null gesetzt werden. Parameter wie etwa angelegte Spannungen können dann variiert werden, so daß etwa eine Kennlinie des Bauteils bestimmt werden kann.

Im einer dynamischen Simulation werden die Modellgleichungen räumlich und zeitlich diskretisiert. Dies führt zu einem nichtlinearen Gleichungssystem, das für jeden Zeitschritt gelöst werden muß. `dessis` verwendet zur zeitlichen Diskretisierung eine Methode zweiter Ordnung namens TR-BDF2 (Trapezoidal Rule / Backward-Differentiation-Formula [Ban85]). Die für jeden Zeitschritt notwendige Berechnung der Lösung eines nichtlinearen Gleichungssystem geschieht durch ein Newton'sches Iterationsverfahren. Die Iteration kann versagen, also nicht konvergieren, wenn der Anfangszustand des Systems zu weit von der gesuchten Lösung entfernt liegt. In diesem Fall wird die Zeitschrittweite verkleinert, gewöhnlich um den Faktor 2, und der entsprechende Zeitschritt erneut berechnet. Weiterhin wird als Nebenprodukt des TR-BDF2-Verfahrens eine Fehlerabschätzung gewonnen, die zur Adaption der Schrittweite benutzt wird.

Das Ziel der Simulationen ist zunächst, Zündfronten in Vierschichtstrukturen zu beobachten. Da die Simulation der Zündung und Frontausbreitung einer zweidimensionalen Vierschichtstruktur recht zeitaufwendig ist, und deren Erfolg von Geometrie und Dotierungsprofil der Struktur und weiteren Parametern wie der angelegten Anodenspannung abhängt, wird zunächst die I-U-Kennlinie der fraglichen Struktur an einer „schmalen“, quasi-eindimensionalen Geometrie getestet (s. Abb. 4.1). Die Breite der Probe soll also klein gegenüber der erwarteten Frontbreite sein, so daß der Zustand der Probe in jedem Fall in x-Richtung homogen bleibt.

Falls sich eine S-förmige Kennlinie ergibt, kann man aus Strom und Spannung am *Haltepunkt* die minimale Anodenspannung zur Frontzündung und die zu erwartende Stromdichte schließen . Dann wird auf einem „breiten“, zweidimensionalen Grundgebiet eine Serie von Frontzündungen (s. Abb. 4.2) simuliert, wobei die Anodenspannung U_A variiert wird. Die nötige Breite, um Frontlösungen zu erhalten, ist zunächst nicht bekannt und muß jeweils erprobt werden.

Die dafür nötige Rechenzeit ist um ein Vielfaches größer als die für die Bestimmung der Kennlinien. Aus den Ergebnissen dieser Simulationen werden schließlich die Frontgeschwindigkeit und die Frontbreite als Funktion der Anodenspannung U_A und eventuell anderer Parameter extrahiert.

Die Simulationen wurden auf zwei handelsüblichen PCs ausgeführt (ein AMD K6 II 500 Mhz und ein Pentium II 400 MHz), die Rechenzeit für eine Simulation liegt bei einigen Minuten (für eine Kennliniensimulation) bis hin zu einigen Stunden (für die Frontsimulation). Schließlich wurden die Geometrien und Dotierungsprofile nach und nach miniaturisiert, und die Konsequenzen dieser Strukturverkleinerungen untersucht.

4.2.1 Extraktion von Frontgeschwindigkeit und -breite

Die Simulation eines Zündvorganges liefert eine Serie von Zustandsdatensätzen zu bestimmten Zeitpunkten, im Normalfall in konstanten Zeitintervallen. Diese Datensätze enthalten bei der Simulation berechnete Variablen, wie Potential, Stromdichte, Elektronen- , Löcherdichten etc. an jedem Gitterpunkt. Aus jedem Datensatz wird nun das elektr. Potential entlang einer horizontalen Schnittgeraden in der Tiefe der p-Basisschicht extrahiert, entsprechend der Tatsache, daß die Fronten im Experiment genauso über das Potential in der p-Basis erfaßt wurden. In dieser eindimensionalen Darstellung läßt sich die raumzeitliche Dynamik der Frontausbreitung gut darstellen.

Um später die Frontgeschwindigkeit und Frontbreite für unterschiedliche Thyristorstrukturen vergleichen zu können, muß zuerst ein Maß für diese Größen definiert werden. Dazu wird die Position der Front als der Ort festgelegt, an dem das Potential den Mittelwert aus seinem Maximal- und Minimalwert (auf dem gesamten eindimensionalen Schnitt) annimmt.

$$V(x_{\mathrm{Front}}) = 1/2(V_{\max} + V_{\min}) \tag{4.13}$$

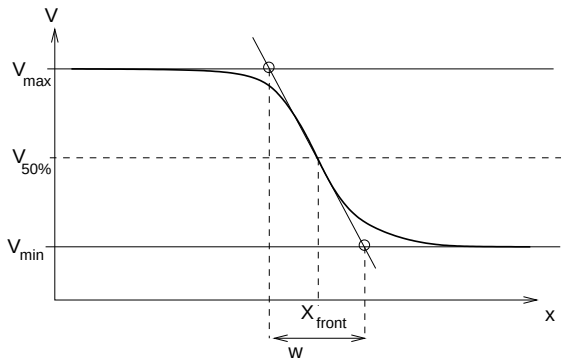


Abbildung 4.3: Maß für die Frontbreite einer sigmoiden Front

Die Festlegung dieses Schwellwertes ist willkürlich, sofern die Front jedoch eine reine Translationsbewegung ausführt, ist die Frontgeschwindigkeit v davon unabhängig. Dies ist der Fall, da der Zündvorgang mit fester Anodenspannung und ohne Vorwiderstand simuliert wird. Die Frontgeschwindigkeit wird dann aus den Frontpositionen zu aufeinanderfolgenden Zeitpunkten auf triviale Weise bestimmt.

Ein Maß für die Frontbreite ist etwa der Abstand zwischen den Positionen, an denen das Potential jeweils einen Schwellwert von 10% bzw. 90% des gesamten Potentialsprungs erreicht. Jedoch sind diese Schwellwerte willkürlich gewählt. Eine solche Schwellwert-Methode wurde bei der elektrischen Beobachtung der Frontausbreitung gewählt. Hier wird der s-förmigen Verlauf der Front approximiert durch eine Tangente an ihren Wendepunkt, also die Tangente mit der maximalen Steigung, und die beiden Horizontalen $V = V_{\max}$ und $V = V_{\min}$. Der Abstand w der beiden Schnittpunkte der Tangente mit den Horizontalen ist ein Maß für die Frontbreite (s. Abb. 4.3).

4.3 Thy750: Simulation und Vergleich mit der T96-Struktur

Ausgangspunkt der Simulationen ist ein Dotierungsprofil, das möglichst naturgetreu den experimentell untersuchten T96-Thyristorstrukturen entspricht. Die Struktur (Thy750 im folgenden) ist 750 µm dick, die Dotierungen der 4 Schichten sind als eine Kombination von konstanten Dotierungen und Gaussprofilen definiert, die Dotierparameter sind in Tabelle 4.2 aufgeführt. Im einzelnen ist auf der gesamten Struktur die konstante Grunddotierung definiert, zu der 3 normalverteilte Dotierungsprofile hinzuaddiert werden. Jede dieser Gaussverteilungen wird durch die Spezies (hier Ph oder Ga), die maximale Konzentration N_{peak}, die Position (Tiefe) des Peaks z_{peak} und die Standardabweichung σ definiert. Statt σ kann auch die Konzentration in einer bestimmten Tiefe angegeben werden, woraus

Schicht	Dotierung	N_{peak}, cm^{-3}	z_{peak}, µm	σ, µm	Tiefe z_{junction}
Grunddotierung:					
n-Basis	n(Ph)	$1.5 \cdot 10^{13}$ konst.	Dicke: 750 µm		
Gauss-Profile:					
Anode	p (Ga)	$1 \cdot 10^{19}$	750	7.60	40
p-Basis	p(Ga)	$5 \cdot 10^{18}$	0	7.80	40
Kathode	n(Ph)	$1 \cdot 10^{20}$	0	5.38	20

Tabelle 4.2: Dotierungsparameter des T750-Profils.

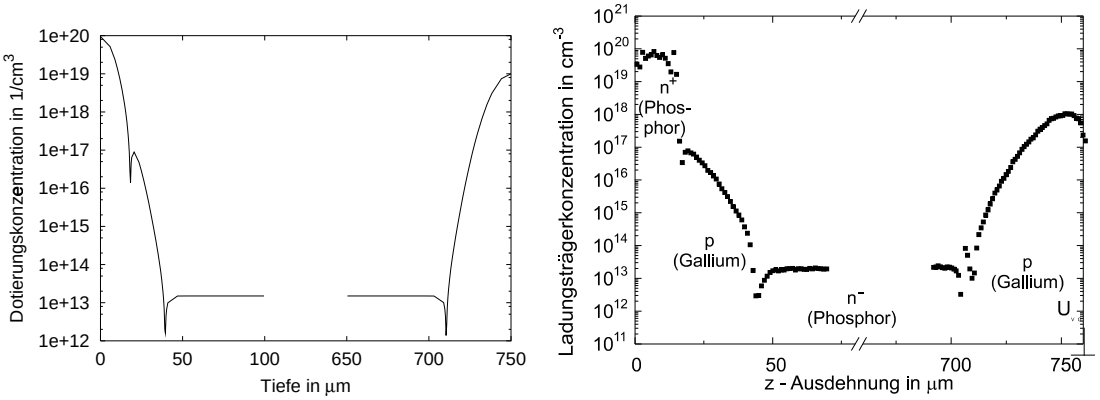


Abbildung 4.4: Dotierungsprofil der simulierten Thy750-Struktur (links) im Vergleich mit dem gemessenen Profil der T96-Struktur (vergl. Abschnitt 3.6).

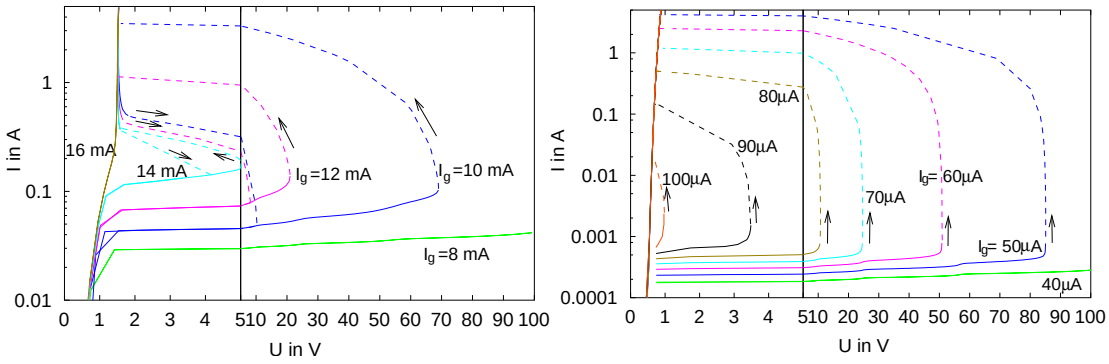
dann σ abgeleitet wird. So kann direkt die Tiefe des entstehenden pn-Übergangs angegeben werden.

Die beiden Basiszonen haben eine sehr unterschiedliche Dicke, nämlich 20 µm gegen 690 µm und auch eine sehr unterschiedliche Dotierungskonzentration.

Abbildung 4.4 zeigt, daß das den Simulationen zugrundeliegende Profil dem experimentell bestimmten Profil der T96-Proben relativ genau entspricht. Dabei sind Abweichungen des Dotierungsprofiles nicht die einzige Fehlerquelle, da die Lebensdauerverteilung in den Thyristorproben nicht bekannt ist und nur eine Schätzung des Maximalwertes der Lebensdauer möglich ist. Alle diese Parameter haben einen großen Einfluß sowohl auf I-U-Kennlinien und die Schaltschwellen der Struktur, als auch auf die Frontausbreitung. Trotzdem zeigen die im folgenden beschriebenen Simulationen eine bemerkenswerte Übereinstimmung mit dem beobachteten Verhalten der T96-Proben. Diese dienen zunächst dazu, die Vorgänge in den experimentell untersuchten Thyristorstrukturen besser zu verstehen, um dann Möglichkeiten zur Optimierung und Miniaturisierung der für die Neuronale Hardware benötigten Kopplungsstruktur zu finden.

4.3.1 Kennliniensimulation

Für die oben beschriebene Thyristorstruktur werden I-U-Kennlinien unter Variation des Gatestroms mit dem in Abschnitt 4.2 beschriebenen Verfahren bestimmt. Um den Einfluß der Elektronenbestrahlung auf die T96-Proben zu modellieren, werden Kennlinienscharen für unterschiedliche Minoritätsträger-Lebensdauern bestimmt. Die Lebensdauer wird dabei wie in Abschnitt 4.1 beschrieben durch die



$\tau = 10$ µs $\qquad\qquad$ $\tau = 100$ µs

Abbildung 4.5: Thy750: Kennlinien unter Variation des Gatestroms I_g, für 2 unterschiedliche Werte der Lebensdauer τ. Probenfläche: $A = 1$ cm^2.

Lebensdauer τ in µs	Kippspannung U_{BO} in V	Haltestrom j_H in A/cm^2
10	3078	3.77
15	2837	0.711
20	2595	0.25
25	2400	0.12
32	2167	0.059
100	842	0.0039

Tabelle 4.3: Thy750: Einfluß der Lebensdauer auf Kippspannung und Haltestrom.

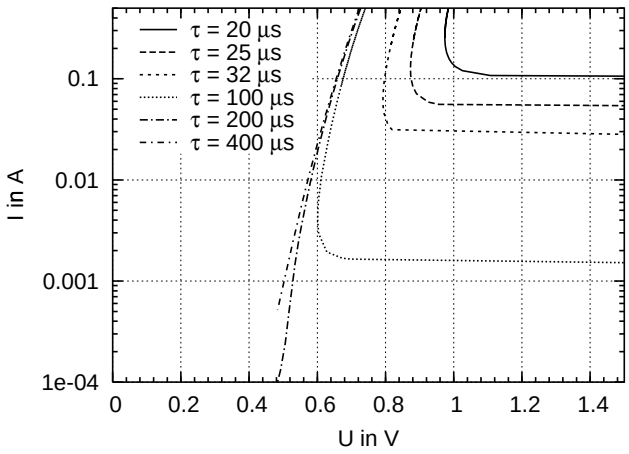


Abbildung 4.6: Thy750: Kennlinien bei Variation der Lebensdauer τ, gezeigt ist nur der Hochstrom-Ast, in fallender Richtung bis zum Ausschalten. Gatestrom $I_G = 0$. Probenfläche $A = 1$ cm^2

Scharfetter-Relation bestimmt, deren Parameter $\tau_{\mathrm{e,max}}$ bzw. $\tau_{\mathrm{p,max}}$ die maximale Lebensdauer angeben, die sie für Dotierungen weit unterhalb einer Schwelle N_{ref} ergibt. In den Simulationen wurde immer die maximale Eletronen-Lebensdauer $\tau_{\mathrm{e,max}}$ als Parameter variiert, wobei das Verhältnis von Löcher-Lebensdauer zu Elektronen-Lebensdauer $\tau_{\mathrm{p,max}} = 0.3\,\tau_{e\mathrm{max}}$ stets beibehalten wurde.

Abbildung 4.5 zeigt Kennlinienscharen der Thy750-Geometrie für zwei Werte von τ. Der Anodenstrom wurde dabei logarithmisch dargestellt, um auch die geringen Ströme im ausgeschalteten Zustand noch auflösen zu können. Die durchgezogenen Teile der Kennlinien werden kontrolliert durch die zeitl. Änderung der Versorgungsspannung „langsam“ durchlaufen, während die gestrichelten Teile die Transienten der Schaltvorgänge darstellen. Der Kennlinienteil mit negativ differentiellem Widerstand liegt zwischen der Ein- und Ausschalttransienten und ist so nicht zu bestimmen.

Sowohl die Zündströme als auch die Halteströme sind im Falle der kleineren Lebensdauer von 10 µs wesentlich höher, und es sowohl Einschalt- wie Ausschaltvorgang zu erkennen. Die Transienten liegen dabei auf Ohm'schen Lastgeraden, was in der logarithmischen Darstellung nicht erkennbar ist.

Ein weitere Simulation wurde ohne Einprägung eines Gatestromes durchgeführt, wobei die Versorgungsspannung U_{VA} einen Maximalwert von 5000 V erreichte. So wird die Zündung durch Überschreiten der Kippspannung U_{BO} erreicht, und die Vorwärtssperrfähigkeit der Struktur kann bestimmt werden. Durch die sehr hochohmige Versorgung der Anode bleibt der eingeschaltete Zustand mit fallendem Strom bis unter den Haltepunkt[1] stabil, so daß ein Teil der Kennlinie mit negativ differentiellem Widerstand bestimmt werden kann. Abbildung 4.6 zeigt den Hochstromast der Kennlinie für einige Werte der Lebensdauer τ. In Tabelle 4.3 sind die Ergebnisse dieser Simulation zusammengefaßt.

Die an quasi-eindimensionalen Proben gewonnenen Kennlinien lassen sich nicht direkt mit den an ausgedehnten Proben gemessenen Kennlinien vergleichen, da dort im eingeschalteten Zustand sich die homogene Stromverteilung destabilisieren kann und sich ein Stromfilament herausbildet. In diesem Zustand ist die Anodenspannung festgelegt, und mit dem Gesamtstrom variiert die Fläche des Hochstromgebietes, so daß die Kennlinie praktisch senkrecht verläuft.

Vernachlässigt man diesen Unterschied, so entspricht die Kennlinie für $\tau = 25$ µs am ehesten den bestrahlten Proben. Der simulierte Haltestrom von 0.12 A/cm^2 entpricht etwa dem Strom, bei dem die bestrahlte Probe 99/3 ganz im eingeschalteten Zustand ist (oberes Ende der Hysterese in Abb. 3.22) . Dieser beträgt 0.19 A, bezogen auf die Probenfläche von 3.5×0.5 cm^2 = 1.75 cm^2 also 0.108 A/cm^2.

[1] Der Punkt minimaler Spannung, an dem die Kennlinie vertikal verläuft

Der Haltestrom der unbestrahlten Probe 99/2 entspricht ungefähr der Simulation mit $\tau = 100$ µs. Die Sperrströme sind jedoch bei den gemessenen Kennlinien wesentlich größer als in der Simulation, und die Kippspannungen sind sehr klein (< 10 V). Dies scheint durch stark gestörte Oberflächen an den Rändern der Proben verursacht zu werden, wo der zweite pn-Übergang die Oberfläche des Wafers erreicht. Die aktiven Grundgebiete der Proben wurden ja durch Abätzung bis in die n-Basis definiert. Die nach der Ätzung nicht passivierte Oberfläche der Wafer könnte für die beobachteten großen Sperrströme verantwortlich sein, die dann zur verminderten Kippspannung führen. Dafür spricht auch, daß die optische Beobachtung der Zündung mittels Hot-Electron-Analyser zeigt, daß bei Überschreitung von U_{BO} ohne Gatestrom der Zündvorgang immer am Rand der Probe beginnt.

4.3.2 Frontsimulation

Die Ausbreitung von Zündfronten auf einer zweidimensionalen Thy750-Struktur wurde wie in Abb. 4.2 gezeigt simuliert. Die Struktur hat nun die Abmessungen 50 mm × 750 µm, mit einem p-Basiskontakt am linken Rand bei $x = 0$.

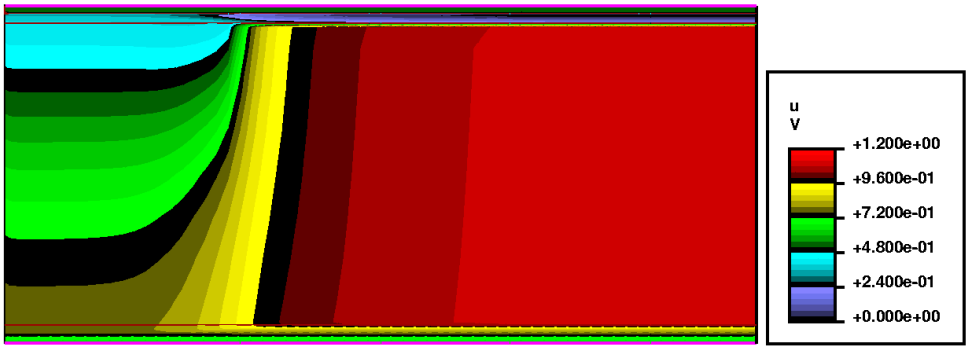
Wie weiter oben beschrieben, wird $R_A = 0$ gesetzt, um die Anodenspannung im Verlauf der Simulation konstant zu halten. U_A wird zwischen 0.6 V und 2 V variiert, um eine Beziehung zwischen Anodenspannung U_A, bzw. Stromdichte j im eingeschalteten Zustand einerseits, und der Frontgeschwindigkeit und Frontbreite andererseits zu gewinnen. Als zweiter Parameter wird wie im letzten Abschnitt die Lebensdauer τ variiert.

Die Simulation ergibt für jeden Punkt im Parameterraum je einen zweidimensionalen Datensatz für jeden Zeitschritt der Simulation, wobei man definieren kann, diese Datensätze in regelmäßigen Zeitintervallen auszugeben, unabhängig von den tatsächlichen Größe der simulierten Zeitschritte.

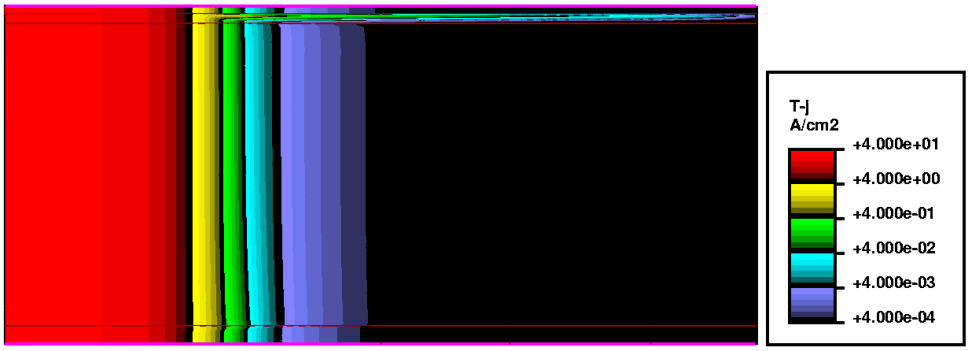
Aus einer solchen Serie von Datensätzen läßt sich eine „Filmsequenz“ generieren, die dann die Ausbreitung der Front anhand einer bestimmten Größe, z.B. dem Potential oder den Trägerkonzentrationen, darstellt. Momentaufnahmen einer solchen Sequenz sind in Abbildung 4.7 zu sehen. Man erkennt, daß der Thyristor im rechten Teil noch ausgeschaltet ist, am Kollektorübergang ist die Sperrschicht als Potentialsprung zu erkennen, während im linken Teil der Potentialsprung am Kollektor praktisch verschwunden ist und stattdessen ein Potentialgradient aufgrund des Ohm'schen Spannungsabfalls die gesamte n-Basis ausfüllt. In der Stromdichteverteilung erkennt man wie in der p-Basis eine Zone erhöhter Stromdichte vom ein- in den ausgeschalteten Bereich hineinragt. Es handelt sich um den Zündstrom, der vom eingeschalteten Gebiet zum noch ausgeschalteten Gebiet fließt

Anodenspannung	U_A	1.2 V
Anodenwiderstand	R_A	0 Ω
Grundgebiet Dicke	z	750 µm
Grundgebiet Länge	x	50 mm
Dotierungsprofil		Thy750
Lebensdauer	$\tau_{\mathrm{e,max}}$	32 µs
Gesamtzeit	t_{tot}	2 ms
Zeitschritt	Δt	0.2 ms

Simulationsparameter für die folgenden Abbildungen.

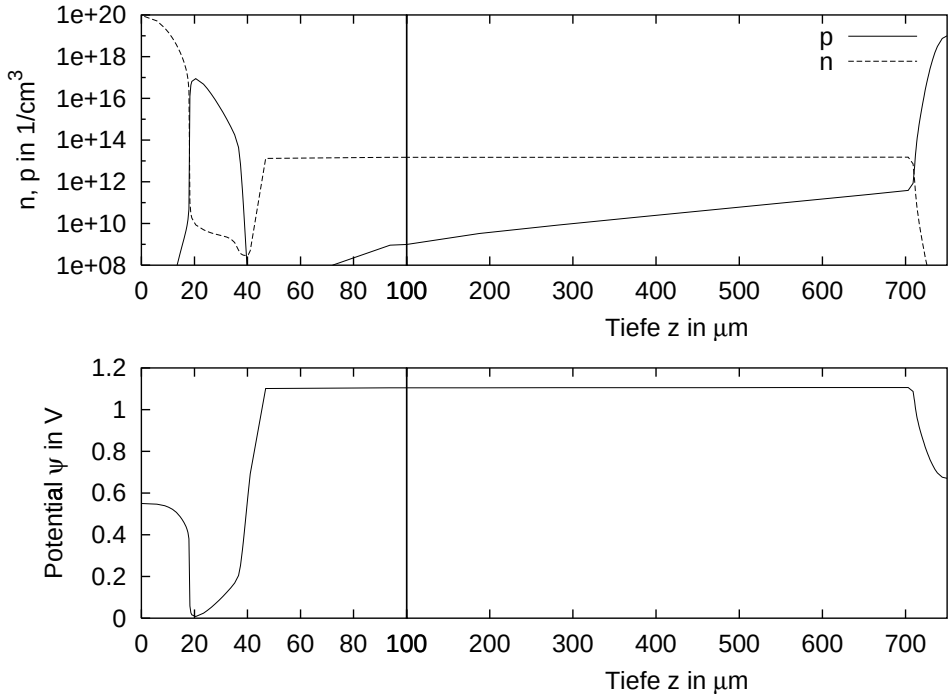


Potentialverteilung bei $t = 0.4$ms

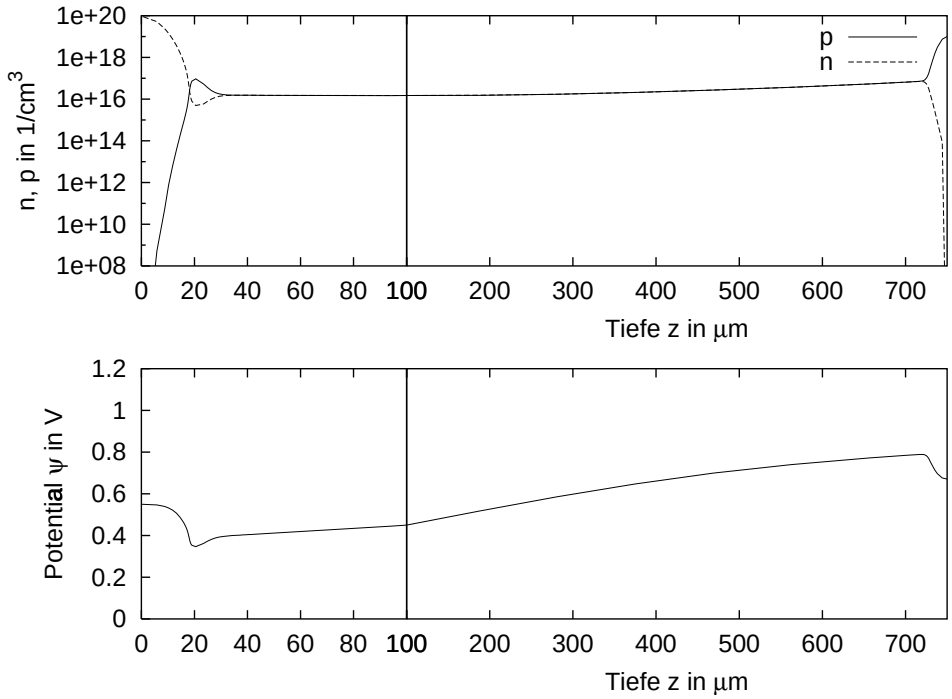


Stromdichteverteilung bei $t = 0.4$ms

Abbildung 4.7: Thy750: Momentaufnahme einer laufenden Front $t = 0.4$ ms nach der Zündung. Die Darstellung ist in z-Richtung (vertikal) um den Faktor 30 gedehnt. Mit Zeitschritt ist das Intervall zwischen 2 Schnappschüssen gemeint, der Simulationszeitschritt wird adaptiv gewählt.



AUS-Zustand



EIN-Zustand

Abbildung 4.8: Thy750: Elektronen- und Löcherkonzentrationen und Potentialverteilungen als Fkt. der Tiefe z jeweils für den aus- und den eingeschalteten Zustand. Es handelt sich um vertikale Schnitte aus dem in Abbildung 4.7 gezeigten Frontzustand, am linken (ON) und rechten (OFF) Rand des Grundgebiets gewonnen.

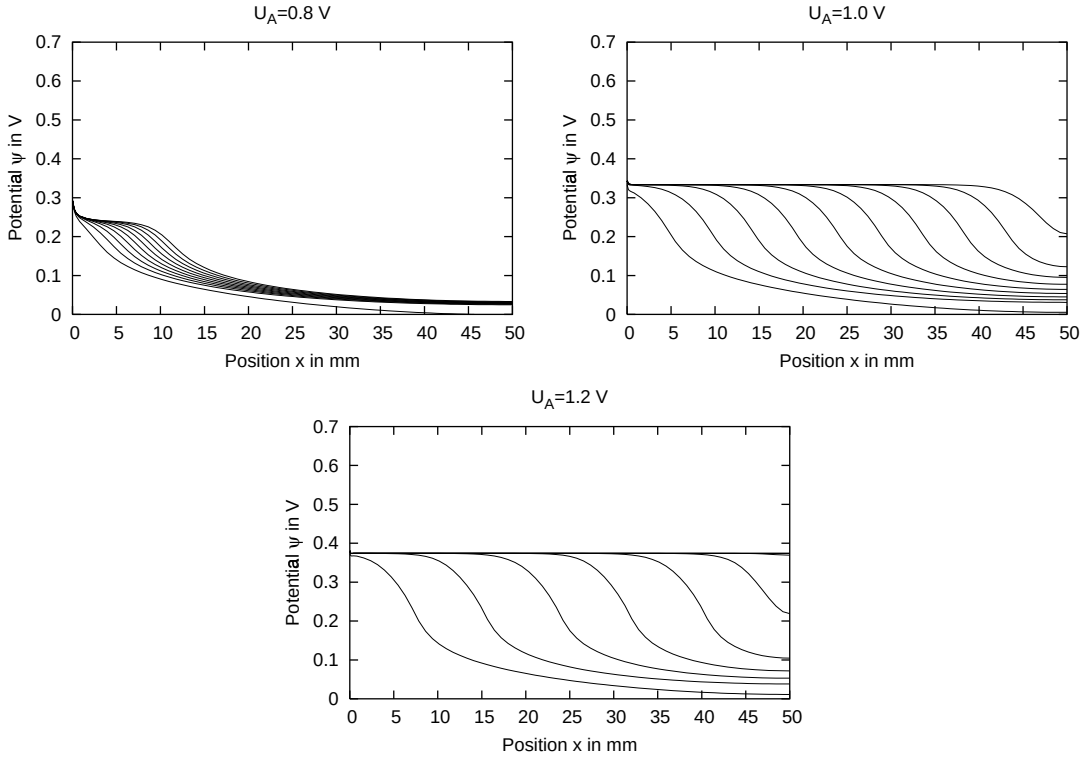


Abbildung 4.9: Thy750: Potentialverteilung in der p-Basis als Funktion des Ortes, in regelmäßigen Zeitintervallen nach Beginn der Zündung aufgetragen. Die einzelnen Graphen zeigen deutlich die Ausbreitung einer Front, deren Geschwindigkeit und Steilheit mit der Anodenspannung zunimmt. Parameter wie in Abb. 4.7, U_A variiert jeweils. Zeitintervall zwischen den einzelnen Plots: $\Delta t = 200$ µs

und die Front vorantreibt. Aufeinanderfolgende Datensätze des Systemzustandes zeigen die Translationsbewegung der Front.

Abbildung 4.8 zeigt die Potentialverteilung und die Ladungsträgerkonzentrationen quantitativ, in Form von zwei vertikalen Schnitten durch das Grundgebiet. Im ausgeschalteten Zustand nehmen Elektronen und Löcher praktisch ihre Gleichgewichtskonzentrationen an, nur in der Verarmungszone am Kollektorübergang sind die Konzentrationen abgesenkt. Im eingeschalteten Zustand hingegen sind beide Basen mit Minoritätsträgern angereichert, in der n-Basis herrscht ein Plasmazustand.

In Abbildung 4.9 ist die Verteilung der Potentials $\psi(x,t)$ in der p-Basis[2] zu aufeinanderfolgenden Zeitpunkten t dargestellt. Wie oben beschrieben kann man in

[2]Schnittiefe $z = 25$ µm

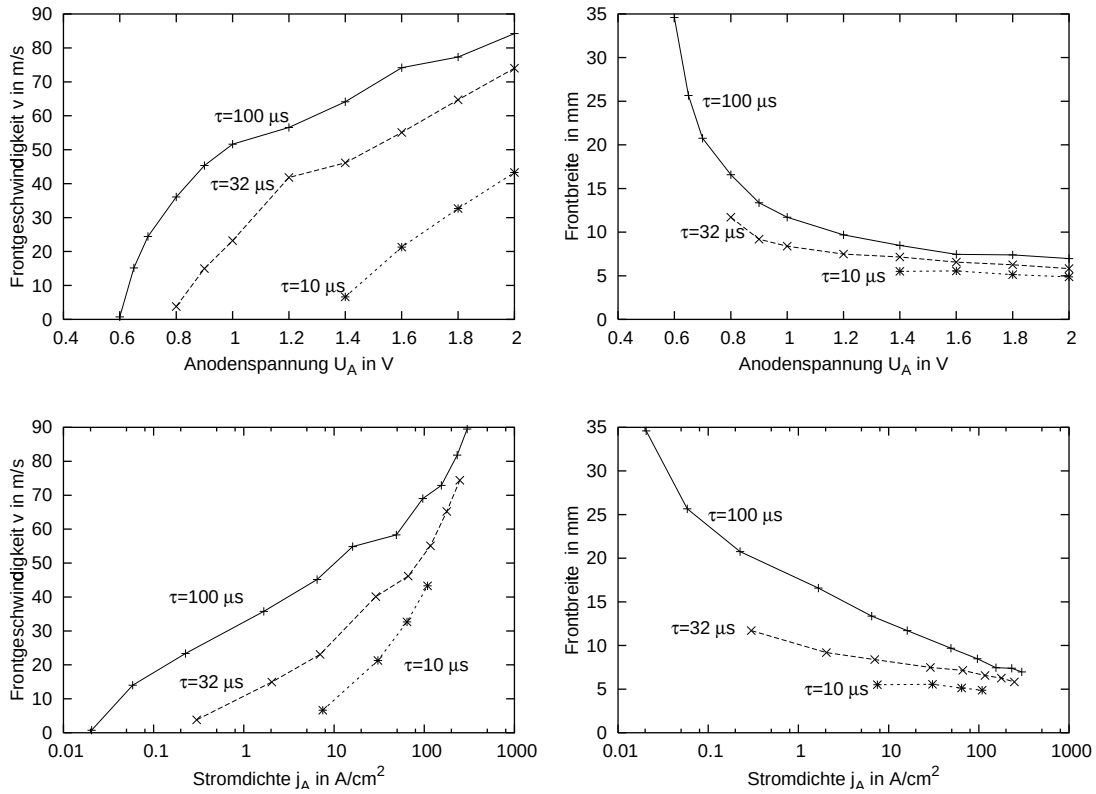


Abbildung 4.10: Thy750: Frontgeschwindigkeit und -breite in Abhängigkeit von Anodenspannung U_A und Trägerlebensdauer τ_n für die oben beschriebene Thyristorstruktur. Unten gleiche Abhängigkeit gegen die Stromdichte j im eingeschalteten Zustand aufgetragen.

dieser Darstellung kann die Frontform als auch deren Translationsbewegung gut erkennen, und nach dem in Abschnitt 4.2.1 angebenenen Verfahren Frontgeschwindigkeit und -breite extrahieren.

In Abbildung 4.10 sind Frontbreite w und Frontgeschwindigkeit v als Funktion der Anodenspannung U_A und der Lebensdauer τ dargestellt. Wie erwartet bewirkt die Absenkung der Lebensdauer eine geringere Frontbreite, aber auch eine geringere Geschwindigkeit. Die minimale Anodenspannung, die zu einer fortschreitenden Front führt, steigt rasch an, was auch die minimale Stromdichte entsprechend erhöht. Trägt man die Zielgrößen Frontgeschwindigkeit und -breite gegen die Stromdichte im eingeschalteten Zustand auf, so erkennt man, daß eine Verbesserung der beiden Größen durch einen exponentiellen Anstieg der Stromdichte erkauft wird.

4.3.3 Abgebremste Fronten durch globalen Anodenwiderstand

Da sowohl in den optischen als auch in den elektrischen Experimenten zur Frontausbreitung auf Thyristorproben immer ein Vorwiderstand den Anodenstrom begrenzte, selbst wenn es sich nur um unvermeidliche Kontaktwiderstände handelte, wird hier eine Simulation beschrieben, in der die Anode ebenfalls über einen Vorwiderstand R_A mit einer konstanten Versorgungsspannung U_V verbunden ist. Es ergeben sich wie im Experiment Fronten, die mit zunehmender Ausbreitung der eingeschalteten Domäne abgebremst werden. Sofern der Vorwiderstand groß genug gewählt ist, erreicht die Front den rechten Rand der Probe nicht. Es stellt sich also ein stationärer Frontzustand ein. Die logische Erklärung dafür ist die globale Gegenkopplung durch den Spannungsabfall am Vorwiderstand: Bei fortschreitender Front wächst der Anodenstrom, durch den Spannungsabfall an R_A sinkt die Anodenspannung, damit sinkt die Frontgeschwindigkeit v. Wird $v = 0$, so erreicht das System einen stationären Zustand.

Das hier simulierte Verhalten entspricht zumindest qualitativ den in Abbildung 3.27 dargestellten, experimentell an Thyristorproben gemessenen Verhalten. So bestätigt die Simulation die Annahme, daß tatsächlich die beobachtete Verlangsamung und Abflachung der Zündfronten mit zunehmender Ausbreitung vom Einfluß eines globalen Vorwiderstandes bedingt wird.

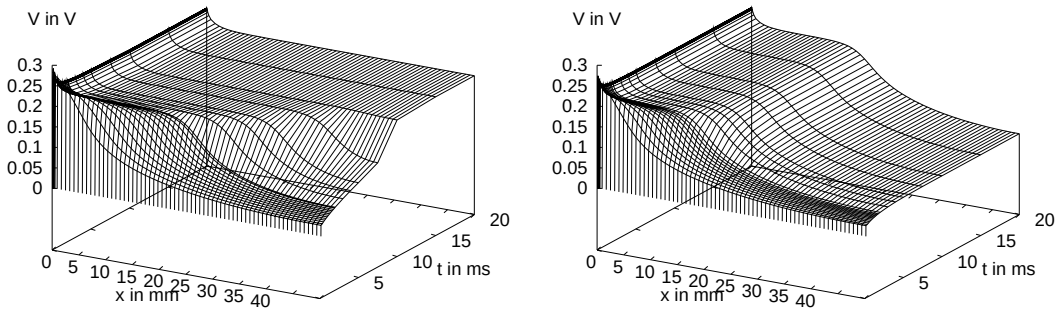


Abbildung 4.11: Thy750: Durch globalen Vorwiderstand R_A gebremste Frontausbreitung. Parameter wie in Abb. 4.7 außer (links, rechts): Probenfläche 50×0.1 mm^2, Versorgungsspannung $U_\mathrm{VA} = 10$ V, Vorwiderstand $R_\mathrm{A} = 8,16$ kΩ.

4.4 Miniaturisierung der Thyristorstruktur

Die im letzten Abschnitt an der Thy750-Struktur erhaltenen Frontbreiten lassen erkennen, daß eine Integration einer solchen Struktur auch einem „Neurochip“ nicht sehr sinnvoll wäre. Die Frontbreite läßt sich bestenfalls auf 5 mm bringen, das heißt ein wesentlich kleinerer Abstand der Gatekontakte wäre nicht sinnvoll, einerseits weil sich die Fronten dann nicht mehr eindeutig lokalisieren lassen, aber auch weil benachbarte Gatekontakte eine Wechselwirkung der Zündströme zeigen.

Daher sollte eine integrierte Version der Thyristorstruktur vor allem eine kleine Frontbreite haben, und einer „sinnvolle“, nicht zu kleine Frontgeschwindigkeit. Daher sollte vor allem die Schaltzeit der Vierschichtstruktur minimiert werden. Aus diesen Gründen wurden Vierschichtstrukturen mit reduzierten Schichtdicken simuliert.

Dabei ist allerdings zu bedenken, daß die Kennlinie und insbesondere die Bistabilität einer Vierschichtstruktur von den Stromverstärkungen der beiden Teiltransistoren abhängt, und diese wiederum von den Basisweiten w_1 und w_2. Bei verringerter Strukturbreite vergrößern sich die Stromverstärkungsfaktoren $\alpha_{1,2}$, mit der Konsequenz, daß die Struktur der sperrenden Zustand verlieren kann. Um dies zu verhindern, kann (1) die Rekombination in den Basen erhöht werden, um die Stabilität des sperrenden Zustandes wiederzugewinnen. Hierzu kann die Dotierung in den Basen erhöht werden, aber auch die im letzten Abschnitt untersuchte Absenkung der Lebensdauer τ wäre möglich.

Die Erhöhung der Dotierung erhöht jedoch auch die Querleitfähigkeit der Basisschichten, die ja den Kopplungsmechanismus darstellt. Während die Schaltzeit eines einzelnen „Thyristorelementes“ im Prinzig unabhängig von der Kopplung[3] der Elemente ist, skaliert die Frontgeschwindigkeit und -Breite proportional zur Wurzel der Kopplungsstärke.

Um schmalere Fronten zu erhalten, muß also Strom aus den Basen abgeführt werden, ohne die Querleitfähigkeit der Basis zu erhöhen. Hierzu können *Basis-Emitter-Kurzschlüsse* dienen, die auch bei Leistungsthyristoren eingesetzt werden, um die Sperrfähigkeit zu verbessern.

Eine andere Möglichkeit wäre die Verwendung einzelner „Thyristorelemente“, die mit Ohm’schen Widerständen verbunden werden. Damit ließen sich die Kopplungswiderstände unabhängig von dem Dotierungsprofil der Thyristorstruktur wählen.

Im folgenden werden nun die Ergebnisse verschiedener Simulationen von verkleinerten Strukturen, und Strukturvarianten vorgestellt.

[3] die Querleitfähigkeit der Basen spielt hier die Rolle der Diffusionskonstante in der RD-Gleichung von Abschnitt 3.5.

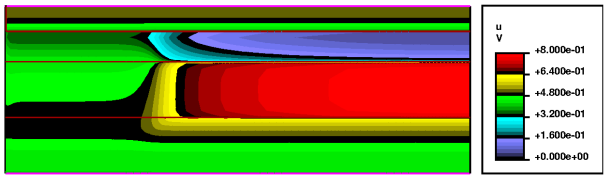
4.4.1 Thy120: Verkleinerung der n-Basis

Hier wurde gegenüber der Thy750-Struktur nur die Dicke des Grundgebiets auf 120 µm reduziert, und somit die Dicke der n-Basis von ca. 690 µm auf 40 µm reduziert, um die Schaltzeit der Struktur zu verkürzen.

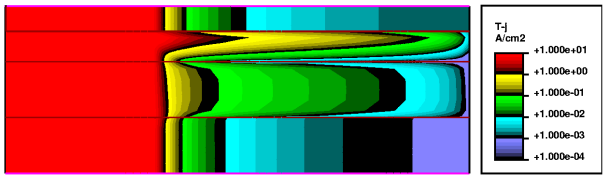
Zunächst wurde wieder die I-U-Kennlinie einer solchen Struktur in einer schmalen, quasi eindimensionalen Geometrie untersucht, wobei die Dotierungskonzentrationen der n- und p-Basis als Parameter variiert wurden. Damit konnten gültige Dotierungen ermittelt werden, bei denen die Struktur ein bistabiles Verhalten zeigt. Insbesondere mußte die Dotierung der n-Basis gegenüber der Thy750-Struktur von $1.5 \cdot 10^{13}$ cm^{-3} auf $1 \cdot 10^{15}$ cm^{-3} erhöht werden.

Mit einer solchen Struktur wurde dann die Frontausbreitung simuliert. Abb. 4.12 zeigt eine Momentaufnahme des Zündvorgangs 40 µs nach der Zündung. Abb. 4.13 stellt den Zündvorgang durch die zeitliche Entwicklung des Potentials in der Tiefe der p-Basis dar. Man erkennt, wie die Schaltfront mit steigender Versorgungsspannung schneller fortschreitet und auch steiler wird, also prinzipiell besser zu detektieren ist. In Abb. 4.14 ist der der Zusammenhang zwischen der Anodenspannung und der Frontgeschwindigkeit und -Breite dargestellt.

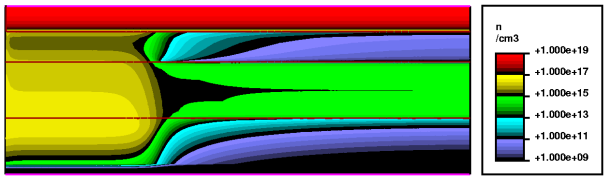
Die Frontbreite ist gegenüber der Thy750-Struktur wesentlich kleiner geworden; sie liegt nun, abhängig von U_A, zwischen 2 bis unter 1 mm. Die Frontgeschwindigkeit hat deutlich zugenommen. Für eine sinnvolle Integration in eine Neuro-Hardware sind die erzielten Frontbreiten aber noch zu groß, daher werden im folgenden wesentlich kleinere Strukturen untersucht.



Potential ψ

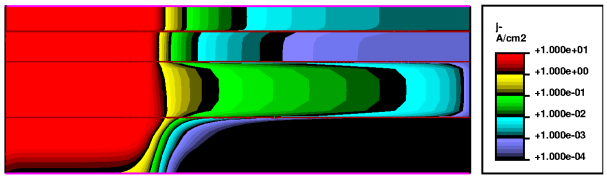


Stromdichte j

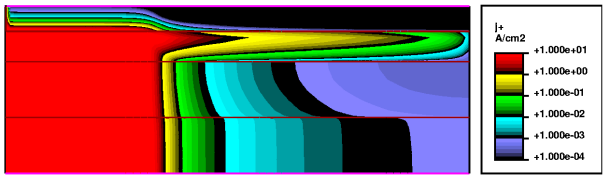


Elektronendichte n

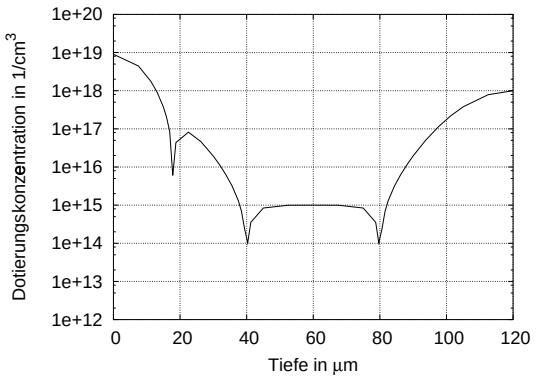
Löcherdichte p



Elektronenstromdichte j_n



Löcherstromdichte j_p



Dotierungsprofil $N(z)$

Anodenspannung	U_A	0.9 V
Anodenwiderstand	R_A	0 Ω
Grundgebiet Dicke	z	120 µm
Grundgebiet Länge	x	10 mm
Dotierungsprofil		Thy120
Dotierung p-Basis		$1 \cdot 10^{18}$ cm^{-3}
Dotierung n-Basis		$1 \cdot 10^{15}$ cm^{-3}
Lebensdauer	$\tau_{\mathrm{e,max}}$	3 µs
Simulierte Zeit	t_{tot}	40 µs
Zeitschritt	Δt	2 µs

Parameter:

Abbildung 4.12: Thy120: Momentaufnahme einer Zündfront, Darstellung in z-Richtung um Faktor 30 gedehnt. Aufnahme zum Zeitpunkt $t = 40$ µs. Von oben nach unten sind Kathode, p–Basis, n-Basis, und Anode zu sehen. In der eingeschalteten Zone sind die erhöhten Trägerkonzentrationen in den Basen zu erkennen, während die in lateraler Richtung vom ein- ins ausgeschaltete Gebiet fließenden Basisströme im Bild der e- bzw. h-Stromdichte zu erkennen sind.

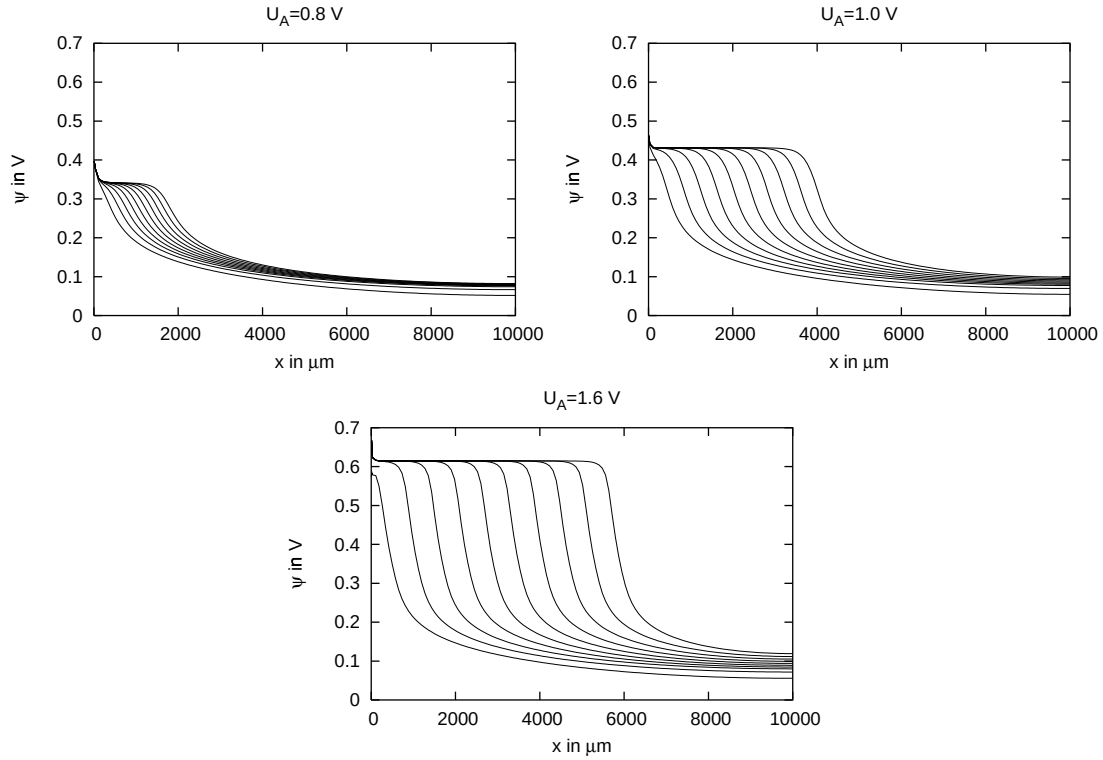


Abbildung 4.13: Thy120: Frontausbreitung dargestellt als Schnappschüsse des p-Basis-Potentials. Geschwindigkeit und Steilheit der Fronten ist gegenüber der Thy750-Struktur deutlich erhöht. Zeitintervall zwischen den einzelnen Schnappschüssen: $2\Delta t = 4$ µs

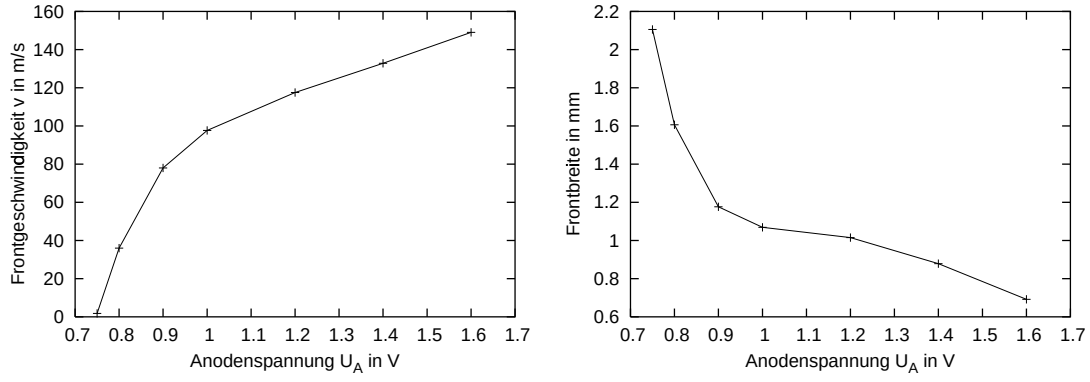


Abbildung 4.14: Thy120: Frontgeschwindigkeit und -breite in Abhängigkeit von der Anodenspannung U_A.

4.4.2 Mikrometer-pnpn-Strukturen

Die Struktur aus dem letzten Abschnitt wurde hier um den Faktor 10 verkleinert, die 4 Schichten sind nun also 12 µm dick. Die Lebensdauer wurde weiter verkürzt, um die Stabilität des ausgeschalteten Zustandes zu erhöhen.

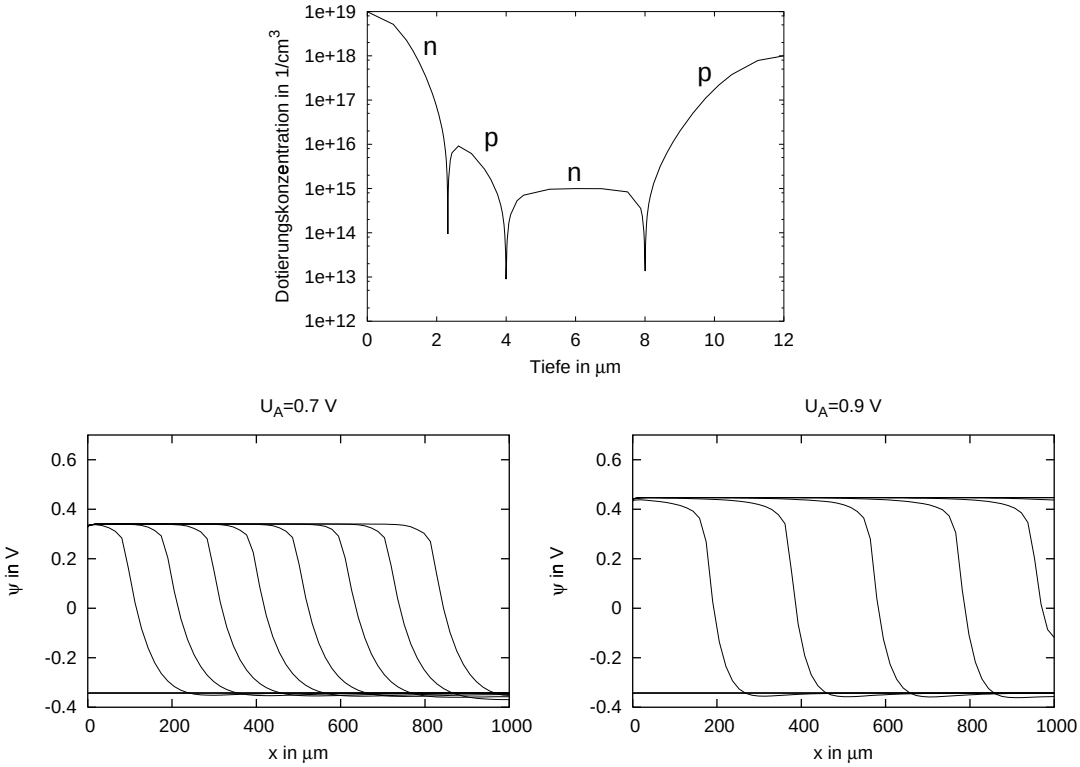
Das Vorgehen bei der Simulation entspricht den beiden vorangegangenen Beispielen. Die in Abb. 4.15 gezeigten Ergebnisse zeigen eine erhebliche Reduktion der Frontbreite sowie eine wesentliche Erhöhung der Frontgeschwindigkeit. Die Front durchläuft nun eine Strecke von 1 mm in weniger als 1 µs Die erzielte Frontbreite würde etwa 10 Gatekontakte bzw. Neuronen pro mm erlauben, die dann in 1 µs einen Lernschritt vollziehen. Der reduzierte Kontaktabstand würde es damit erlauben, eine Kohonenhardware in quadratischer Anordnung mit 10×10 Neuronen auf 1 mm^2 Chipfläche unterzubringen.

Basierend auf diesem Ergebnis wurde die Dicke der Struktur nun wieder um den Faktor 10, also auf 1.2 µm verkleinert. Die Schichtdicken der 4-Schichtstruktur liegen damit in einer für integrierte Schaltungen passenden Größenordnung. In Abbildung 4.16 sind die Kennlinien dieser Struktur namens Thy1.2 gezeigt, bei 2 unterschiedlichen Grunddotierungen. Die Kennlinien wurden mit einem Basis-Kathoden-Shunt bestimmt, der Strom aus der Basis ableitet und somit den gesperrten Zustand stabilisiert. Mit sinkendem Shunt erhöht sich Spannung und Strom am Kippunkt, ebenso der Haltestrom. Im Extremfall eines Basis-Kathoden-Kurzschlusses erhält man die Kennlinie des pnp-Teiltransistors, wobei der steile Anstieg oberhalb einer bestimmten Durchbruchsspannung durch den Punch-Through-Effekt entsteht. Durch die kleinen Basisweiten von 0.5 µm ist eine relativ große Dotierung nötig, um den Punch-Through erst bei Spannungen über 5 V eintreten zu lassen.

Die sehr hohen verwendeten Shuntwiderstandswerte zeigen die extreme Empfindlichkeit ein solchen dünnen pnpn-Struktur auf Basisstromverluste. Ein geringer Verlust scheint notwendig, um den ausgeschalteten Zustand stabil zu halten, zu hohe Verluste lassen jedoch den Haltestrom zu weit ansteigen.[4]

Die Ausbreitung einer Zündfront auf einer großflächigen, kontinuierlichen Struktur von 1.2 µm Dicke ist in Abb. 4.17 dargestellt. Im Unterschied zu den vorherigen Frontsimulationen besitzt die auf dem Thy1.2-2-Profil basierte Struktur im Abstand von 10 µm angebrachte Gatekontakte, die über Shunts mit Masse verbunden sind. Die Diagramme zeigen das Potential an den räumlich aufeinanderfolgenden Gatekontakten als Funktion der Zeit, während die vorherigen Darstellungen das

[4] Es zeigt sich, daß eine Erhöhung der p-Basisdotierung oder eine Absenkung der Trägerlebensdauer die gleiche Wirkung haben, so daß man auf die Shunts verzichten kann.

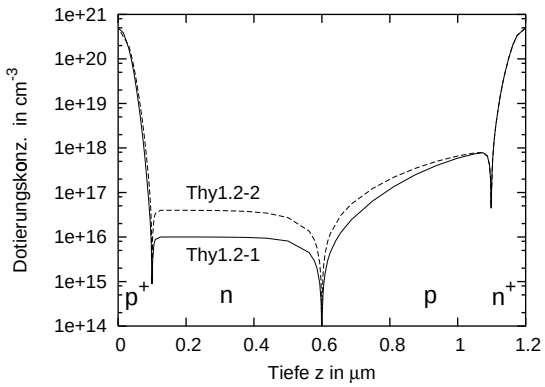


Anodenspannung U_A V	Frontgeschwindigkeit v m/s	Frontbreite µm
0.7	2112	79.0
0.75	2498	68.4
0.8	2898	63.2
0.9	3825	53.4

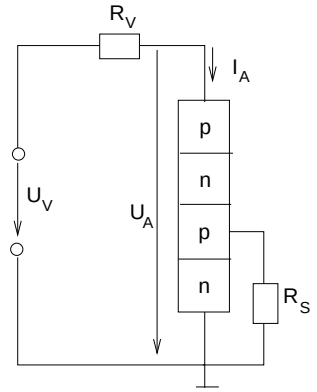
Simulationsparameter:

Grundgebiet Dicke	z	12 µm
Dotierung n-Basis		$1 \cdot 10^{15}\mathrm{cm}^{-3}$ const.
Dotierung p-Basis		$1 \cdot 10^{17}\mathrm{cm}^{-3}$ peak (gaussf.)
Grundgebiet Länge	x	1 mm
Lebensdauer	$\tau_{\mathrm{e,max}}$	1 µs
Simulierte Zeit	t_{tot}	500 ns
Zeitschritt	Δt	25 ns

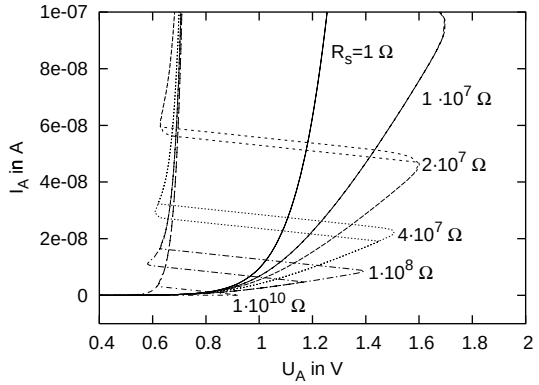
Abbildung 4.15: Thy12: Fronten auf einer 12 µm-pnpn-Struktur. Gezeigt ist zunächst das gewählte Dotierungsprofil, darunter die Ausbreitung der Zündfronten bei verschiedenen Anodenspannungen. Die Probe ist 1000 µm lang, die Simulationszeit beträgt 500 ns, die Zeitdifferenz zwischen den einzelnen „Schnappschüssen“ beträgt 25 ns.



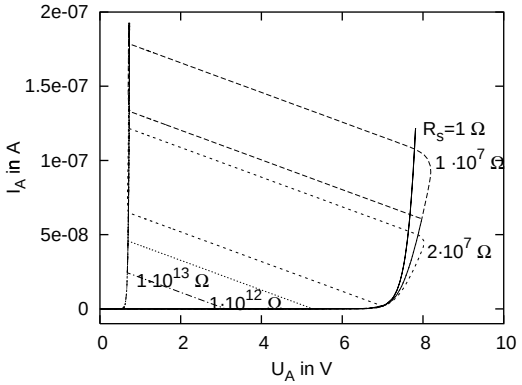
Dotierungsprofil



Kennlinienbestimmung mit Shunt



Thy1.2-1: Grunddotierung $1 \cdot 10^{16}$ cm^{-3}

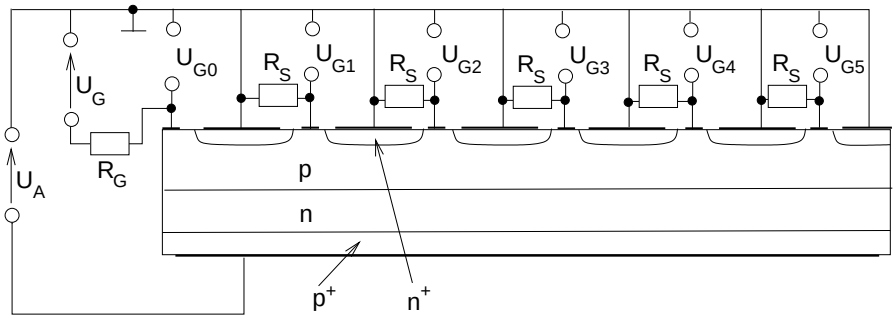


Thy1.2-2: Grunddotierung $4 \cdot 10^{16}$ cm^{-3}

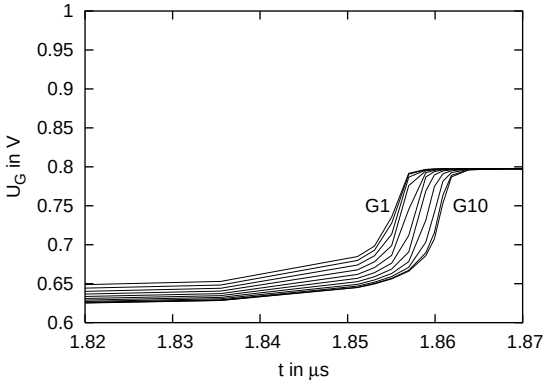
Simulationsparameter:

Grundgebiet Dicke	z	1.2 µm
Dotierung n-Basis		$1 \cdot 10^{16}, 4 \cdot 10^{16}$ cm^{-3} const.
Dotierung p-Basis		$1 \cdot 10^{18}$cm^{-3} peak (gaussf.)
Grundgebiet Länge	x	1.0 µm
Grundgebiet Tiefe	y	1.0 µm
Lebensdauer	$\tau_{\mathrm{e,max}}$	1 µs

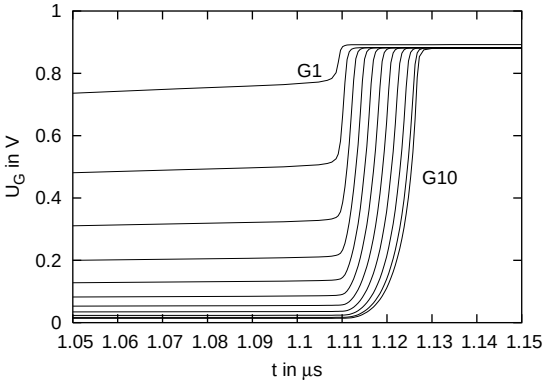
Abbildung 4.16: Thy1.2: Kennlinien der Strukturen Thy1.2-1 und Thy1.2-2. Beide sind 1.2 µm dick, die Dotierungsprofile sind oben links zu sehen. Sie unterscheiden sich nur in der n-Grunddotierung. Die Kennlinienscharen unten zeigen die Wirkung eines Basis-Shuntwiderstandes. Wird der Shunt verkleinert, steigen Kippspannung und Kippstrom, und der Haltestrom wächst. Wegen der geringen Basisweite der Struktur ist der Punch-Through-Effekt für den Durchbruch der Kollektorsperrschicht verantwortlich. Daher hat Thy1.2-2 die höhere Kippspannung.



Probengeometrie mit ext. Beschaltung



Zündvorgang ohne Shunts:



Zündvorgang mit Shunts:

Shuntwiderstand	R_S	$1 \cdot 10^{14}\ \Omega$
Anodenspannung	U_A	0.8 V
Stromdichte	j_{on}	120 A/cm^2
Anodenstrom	I_A	0.12 mA
Schaltzeit	t_s	2.7 ns
Laufzeit/Gate	t_l	0.52 ns
Frontbreite		52 µm
Frontgeschw.	v	19000 m/s

Shuntwiderstand	R_S	$1 \cdot 10^{5}\ \Omega$
Anodenspannung	U_A	0.9 V
Stromdichte	j_{on}	3800 A/cm^2
Anodenstrom	I_A	3.8 mA
Schaltzeit	t_s	4.39 ns
Laufzeit/Gate	t_l	1.62 ns
Frontbreite		27 µm
Frontgeschw.	v	6100 m/s

Grundgebiet Dicke	z	1.2 µm
Dotierung n-Basis		$4 \cdot 10^{16}$ cm^{-3} const.
Dotierung p-Basis		$1 \cdot 10^{18}$cm^{-3} peak (gaussf.)
Grundgebiet Länge	x	100 µm
Grundgebiet Tiefe	y	1.0 µm
Lebensdauer	$\tau_{\text{e,max}}$	1.0 µs
Gatekontaktabstand		10 µm
Gatekontakte		10

Abbildung 4.17: Thy1.2: Frontausbreitung in der pnpn-Struktur Thy1.2-2. Der erste Gatekontakt wird mit einer linear ansteigenden Spannung versorgt, bis die Zündschwelle überschritten wird. Gateshunts können zur Absenkung und Stabilisierung des Gatepotentials im ausgeschalteten Zustand dienen, erhöhen aber den Stromverbrauch erheblich.

Gatepotential als Funktion des Ortes zu verschiedenen Zeiten zeigten. Die Zündung wurde hier nicht durch einen Schaltimpuls, sondern durch ein langsames lineares Hochfahren der Steuerspannung für das erste Gate erreicht.

Wie man erwarten würde, bewirken kleinere Shuntwiderstände einen erhöhten Zündstrom, sowie verlängerte Schalt- und Laufzeiten der Front. Das Gatepotential des ausgeschalteten Zustands wird abgesenkt, was den Spannungshub der Front erhöht. Die Frontbreite verringert sich aber gemessen am erhöhten Strombedarf nur wenig.

Man erreicht etwa $50-25$ µm Frontbreite, womit man etwa bis zu 40 Gatekontakte pro mm unterbringen könnte, in einer zweidimensionalen Anordnung also 40×40 Neuronen. Eine Stromdichte von über 1000 A/cm^2 erscheint allerdings sehr hoch, und wäre allenfalls lokal in Strukturen von µm-Größe machbar. Eine Möglichkeit, den Gesamt-Strombedarf einer flächigen pnpn-Struktur zu verringern wäre etwa, sie als Gitter von schmalen Streifen auszubilden.

Ein weiteres Problem ist die stark nichtlineare j-U_A-Abhängigkeit, die die Stabilisierung der Versorgungsspannung erschwert. Man könnte dieses „Problem“ aber nutzbringend einsetzen, indem man nicht die Versorgungsspannung, sondern den Anodenstrom stabilisiert: Dann würde die Front sich nicht mit konstanter Geschwindigkeit ausbreiten, sondern die Größe des eingeschalteten Gebiets würde vom eingeprägten Strom festgelegt. Durch lineare Erhöhung des Anodenstromes kann die Bewegung der Front dann von außen gesteuert werden. Die Stromdichte im eingeschalteten Gebiet wäre relativ gering, in der Größenordnung der Haltestromdichte, wobei die so erreichte Frontgeschwindigkeit geringer wäre als im vorher beschriebenen Fall, da auch die im Gleichgewicht angenommene Anodenspannung geringer wäre.

Ein Problem bei dieser Betriebsart könnte *Pinning* der Front an inhomogenen Gatekontakten sein: Bei (langsamer) Erhöhung des eingeprägten Anodenstromes versucht das eingeschaltete Gebiet zu wachsen. In welche Richtung es sich dabei ausdehnt, sofern es nicht schon den Rand berührt, ist dabei zunächst unbestimmt. Die Gatekontakte stellen ein Hindernis für die Frontbewegung dar, das überwunden werden muß. Sind diese „Hindernisse“ nun unterschiedlich hoch, so könnte (auf einem eindimensionalen Grundgebiet) das eingeschaltete Gebiet auf einer Seite hängenbleiben und sich nur noch zur anderen Seite ausdehnen. Trägheitseffekte sollten dem jedoch entgegenwirken, so daß eine solche asymmetrische Frontausbreitung bei ausreichend hoher Frontgeschwindigkeit nicht mehr auftreten sollte.

4.4.3 Laterale Strukturierung

Die Zündfronten, die sich in den oben vorgestellten kontinuierlichen Kopplungsstrukturen ergeben, sind bezogen auf die Dicke der Struktur recht breit, die Frontbreite ist je nach den gewählten Parametern 10–50 mal größer als die Dicke der Struktur. Also sind zusätzliche Maßnahmen wünschenswert, um die Frontbreite weiter zu verringern. Dazu werden hier diskrete Strukturen aus pnpn-Einzelelementen vorgestellt, die mit Kopplungswiderständen an den p-Basen gekoppelt sind (Abb. 4.18).

Frontbreite und Ausbreitungsgeschwindigkeit hängen nun bei festgelegter lokaler Dynamik der pnpn-Strukur hauptsächlich von der lateralen Kopplung des Systems ab, also in erster Linie dem Schichtwiderstand der Basen. Dieser müßte erhöht werden, um die Frontbreite zu reduzieren. Die Dotierung der Basen kann aber bei dünnen Basisschichten nicht weiter reduziert werden, da entweder die Stromverstärkungen $\alpha_{1,2}$ der beiden Teiltransistoren zu groß werden, oder der Punch-Through-Effekt schon bei sehr kleiner Spannung einsetzt, so daß keine S-förmige Kennlinie mehr möglich ist, und die Struktur nicht mehr sperrt.

Dies geschieht bei der in Abbildung 4.16 gezeigten Struktur Thy1.2 bei einer Grunddotierung der n-Basis unter $1 \cdot 10^{16}$ cm^{-3}. Eine größere Dotierung von $4 \cdot 10^{16}$ cm^{-3} ist erforderlich, um die Kippspannung auf ca. 8 V zu erhöhen. Es ist also nicht einfach möglich, scharf begrenzte Fronten durch niedrig dotierte Basisschichten zu erzeugen, da diese 4-Schichtstrukturen nicht mehr bistabil wären.

Abgesehen von der Frontbreite ist ein anderer problematischer Aspekt einer großflächigen pnpn-Struktur die Spannungsversorgung. Die Stromdichte j im eingeschalteten Gebiet ist exponentiell von der Anodenspannung abhängig, so daß kleinste Änderungen der Anodenspannung zu großen Änderungen der Stromdichte und der Frontgeschwindigkeit führen. Ist die Versorgung der Anode nicht niederohmig genug, führt der Spannungsabfall an deren Innenwiderstand zu Stromgegenkopplung mit einer Abbremsung der Front während ihrer Ausbreitung. Hinzu kommt, daß die Struktur auf der gesamten Fläche Strom und damit Leistung verbraucht, aber nur an relativ weit entfernten Punkten, den Gatekontakten, eine Interaktion mit der Umgebung stattfindet.

Zur Lösung dieser Probleme bietet es sich an, die Thyristorstruktur in Einzelelemente zu zerlegen, sie gewissermaßen zu diskretisieren: Man erhält eine durch Widerstände gekoppelte Kette von Thyristorelementen (Abb. 4.18). In zwei Dimensionen kann dieselbe Schaltung als Widerstandsgitter mit Thyristorelementen an den Knotenpunkten aufgebaut werden. Die Widerstände koppeln die Basen der Thyristoren, was genau dem diskretisierten Reaktions-Diffusions-Modell der Frontausbreitung von Abschnitt 3.5 entspricht. Dabei genügt es, z.B. nur die p-Basen zu koppeln. Die Kathoden sind mit Masse verbunden, die Anoden über je

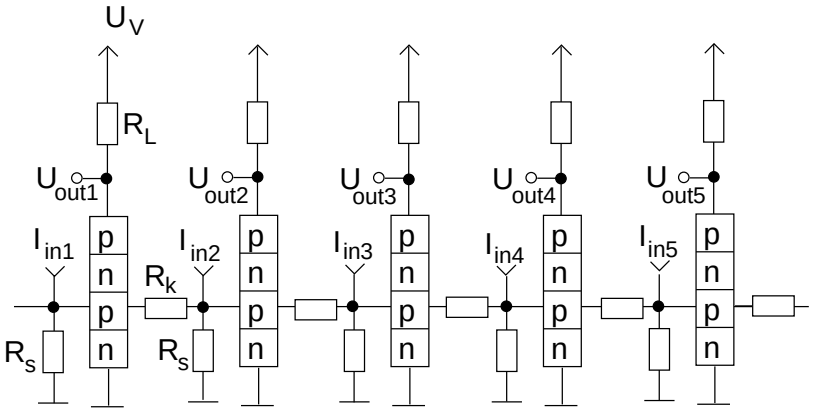
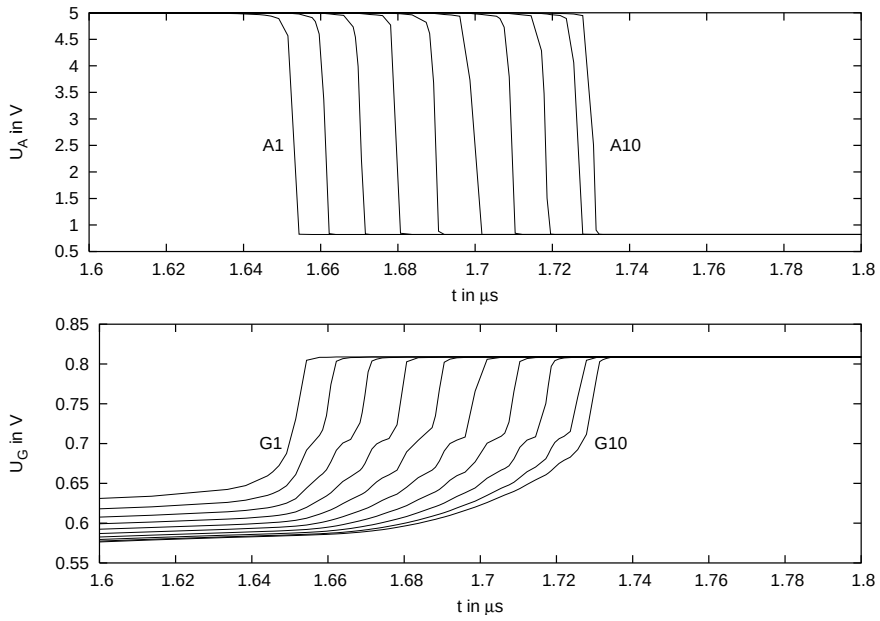


Abbildung 4.18: Diskret gekoppeltes aktives Medium.



Frontzündung in diskretem aktiven Medium

Elementquerschnitt	A	$A = 1\ \mu\text{m}^2$
Lastwiderstand	R_L	$1 \cdot 10^6 \Omega$
Kopplungswiderstand	R_K	$1 \cdot 10^6 \Omega$
Schaltzeit	t_s	2.1 ns
Laufzeit/Element	t_l	9.6 ns

Simulationsparameter

Abbildung 4.19: Zündfront in diskret gekoppeltem aktiven Medium. Das Gate G1 wurde mit einer linear ansteigenden Eingangsspannung versorgt, die schließlich die Zündung des ersten Thyristors auslöst. Dotierungsprofil der Einzelelemente: Thy1.2-2.

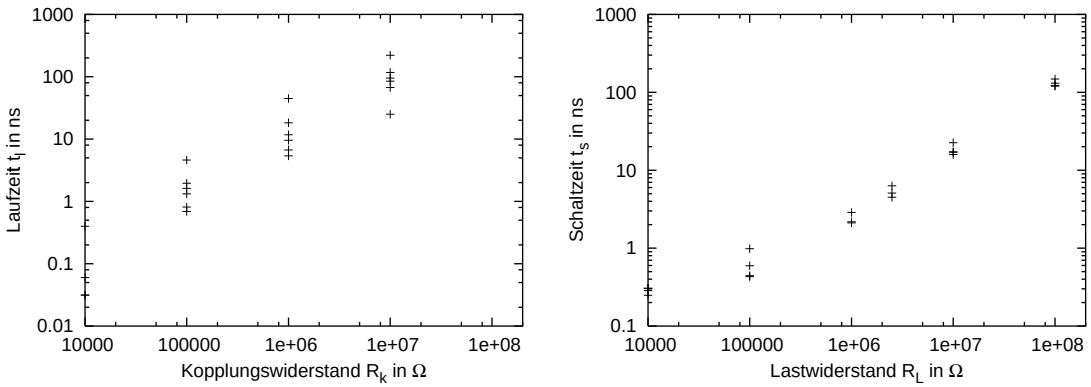


Abbildung 4.20: Zusammenhang zwischen den Parametern Last- und Kopplungswiderstand R_L und R_K und den Zielgrößen Schalt- und Frontlaufzeit im diskret gekoppelten aktiven Medium aus Abb. 4.18. Die Daten stammen aus Simulationen wie in Abb. 4.19. Der Laststrom bestimmt hauptsächlich die Schaltzeit der Elemente, während der Kopplungswiderstand hauptsächlich die Laufzeit der Front von einem Element zum nächsten bestimmt.

einen Lastwiderstand mit der Versorgungsspannung. An diesem Lastwiderstand kann dann ein Ausgangssignal gewonnen werden, daß direkt dem Anodenstrom proportional ist. Während die Gatespannungen schon vor der Zündung wegen der Querkopplung deutlich angehoben werden, was eventuell die Festlegung eines passenden Schwellwertes erschwert, ist dieser Effekt an der Anodenspannung nicht zu erkennen. Wegen der exponentiellen Beziehung zwischen I_A und U_G ist die Flankensteilheit des Anodensignals U_A wesentlich größer als die der Gatespannung U_G. Der Thyristor dient gewissermaßen selbst als Verstärker für seine Gatespannung. Die Amplitude der Ausgangsspannung U_A kann so je nach Versorgungsspannung mehrere Volt betragen, so daß das Ausgangssignal ohne weitere Verstärkung genutzt werden kann.

Der Lastwiderstand legt den Anodenstrom jedes Einzelelementes genauer fest als die exponentielle Diodenkennlinie es im Falle fehlender Lastwiderstände könnte, so daß sich die Stabilität der Versorgungsspannung weniger auf die Frontgeschwindigkeit auswirkt. Während die Schaltzeit des Einzelelementes (bei durch die Wahl der Vierschichtstruktur festgelegten Basiskapazitäten) durch die Stromdichte im eingeschalteten Zustand und damit durch den Lastwiderstand festgelegt wird, wird die Frontgeschwindigkeit durch den Kopplungswiderstand bestimmt.

In Abb. 4.19 wird gezeigt, wie sich eine Schaltfront in einem diskreten System aus 10 Einzelelementen ausbreitet. Die Einzelelemente haben das vertikale Dotierungsprofil Thy1.2-2, also eine Dicke von 1.2 µm. Die Schaltung wurde im *mixed-mode* simuliert, d.h. die Thyristorelemente werden als detailliertes 2D-Drift-Diffusions-Modell, die passiven Elemente durch einfache Modellgleichungen dargestellt.

Ein Parameterscan (s. Abb. 4.20) zeigt, daß die Schaltzeit der Einzelelemente praktisch nur vom Lastwiderstand abhängt, während die Laufzeit der Schaltfront, also die Verzögerung zwischen dem Einschalten zweier aufeinanderfolgender Elemente mit dem Kopplungswiderstand festgelegt wird. Durch Verwendung eines relativ großen Verhältnisses R_k/R_l erreicht man, daß die Front sich jeweils nur über ein Element erstreckt, daß also die Elemente einzeln nacheinander umschalten. Dies ist für die Anwendung für eine SOM-Hardware ein idealer Zustand, da die Front sehr präzise detektiert werden kann, und die Lernzeiten der Neuronen dadurch präzise monoton mit dem Abstand vom Gewinnerneuron abnehmen. Außerdem minimiert eine so geringe Kopplung die Wechselwirkung von Zündströmen an benachbarten Gatekontakten.

Das in dieser Simulation verwendete Thyristorelement ist als rein vertikale Vierschichtstruktur sehr einfach aufgebaut, in einem Siliziumwafer eindiffundiert würde sich etwa das Problem stellen, wie die untere Schicht, also die Anode, zu kontaktieren ist. Eine Alternative wäre der Aufbau des Thyristors aus einem vertikalen npn-Transistor und einem lateralen pnp-Transistor, die Anode würde also als flach eindiffundierte p-Zone in den n-Kollektor eines gewöhnlichen Planartransistors gesetzt. Außerdem berücksichtigt die hier gezeigte Simulation nur rein Ohm'sche Kopplungs- und Lastwiderstände, und vernachlässigt die Kapazitäten der realen Halbleiterwiderstände, seien es diffundierte Widerstände, oder FET-Strukturen. Somit werden die Schalt- und Frontlaufzeiten in der Realität größer sein.

Die Realisation von hochohmigen Last- und Kopplungswiderständen ist problematisch, da der Schichtwiderstand der p-Basisschicht der Thy1.2-2-Struktur nur etwa 2 kΩ/□ beträgt. Ein Lastwiderstand von 1 MΩ erfordert z. B. 500 „squares“ Chipfläche. Würde man die Querschnittsfläche der Thyristorelemente etwa um den Faktor 100 vergrößern, könnte man mit Widerständen um 10 kΩ auskommen, die sich leicht als diffundierter Widerstand herstellen läßt. Damit verspielt man allerdings die Chance, mit wenig Chipfläche und Strom für die Thyristorelemente auszukommen. Eine andere Möglichkeit wäre die Verwendung von MOS-Transistoren als aktive, hochohmige Last- und Kopplungswiderstände. Da die Linearität der Widerstände für die Funktion des aktiven Mediums nicht entscheidend ist, könnten MOS-Elemente verwendet werden, deren Sättigungsverhalten sollte jedenfalls nicht der Funktion der Schaltung schaden. Mit Last- und Kopplungswiderständen in der Größenordnung von $1 \cdot 10^7\ \Omega$ würde der Stromverbrauch auf etwa 0.4 μA pro Thyristorelement sinken, bei Schalt- und Laufzeiten von etwa 10 ns. So kann die Kopplungsstruktur flexibel in Form eines Gitters aufgebaut werden, so daß die eigentlichen Neuronen dazwischen Platz finden, der Stromverbrauch kann minimiert werden, und die Frontgeschwindigkeit kann durch die Wahl der Kopplungswiderstände über mehrere Größenordnungen variiert werden.

Zusammenfassung und Diskussion

In der vorliegenden Arbeit wurde die hardwaremäßige Implementation einer selbstorganisierenden Karte nach Kohonen untersucht. Das Hardwarekonzept basiert speziell auf der Kopplung einzelner Neuronen mit Hilfe von Schaltfronten, die sich auf einem aktiven Medium ausbreiten. Eine auf Analogelektronik basierende Implementation eines solchen neuronalen Netzes demonstriert die Funktionstüchtigkeit des vorgestellten Konzeptes. Simulationen mit einem kommerziellen Halbleiterstruktursimulator weisen einen Weg zu einer integrierten Version der SOM-Hardware.

Im ersten Kapitel werden zunächst selbstorganisierende neuronale Netze vorgestellt, die durch unüberwachtes Lernen trainiert werden. Nach der Einführung des Vektorquantisierers wird das Hauptaugenmerk auf die von Kohonen eingeführte selbstorganisierende Karte (SOM) gerichtet. Die SOM lernt auf selbstorganisierte Weise durch *Konkurrenz* und *Kooperation* die Abbildung eines hochdimensionalen Musterraumes auf einen niedrigdimensionalen Kortex. Anhand von Beispielen werden die wichtigen Eigenschaften dieser Abbildung vorgestellt, wie Dichtemodellierung der Häufigkeitsverteilung der trainierten Mustermenge und (lokale) Topologieerhaltung. Varianten und Abwandlungen des klassischen Kohonen-Algorithmus werden erläutert. Zuletzt werden die Applikationsmöglichkeiten, allem im Bereich der Mustererkennung, der Clusteranalyse und der Visualisierung hochdimensionaler Musterräume angesprochen. Beispiele aus der Spracherkennung, der medizinischen Bildanalyse, der Robotik u. a. werden vorgestellt.

Im zweiten Kapitel wird ein paralleles Hardware-Konzept für die SOM vorgestellt. Die Hardware besteht aus einer Anzahl analog aufgebauter Neuronen und einer Kopplungsstruktur. Jedes Neuron enthält ein Sample-and-Hold Glied, das die Komponenten des jeweiligen Prototypen speichert, und zur Implementation des Lernvorganges benutzt wird. Der Lernvorgang wird durch partielle Aufladung des Speicherkondensators auf den präsentierten Musterwert vorgenommen. Die Lernrate wird dabei durch die Ladezeit kontrolliert. Jedes Neuron enthält eine Abstandsrechenschaltung, die den Abstand des Prototypen vom Mustervektor bestimmt. Die Quantisierung eines Mustervektors, also die Zuordnung zu einem der

Neuronen, erfordert die Bestimmung des Neurons, das den minimalen Abstand zum Mustervektor aufweist. Diese Gewinnersuche wird parallelisiert, indem jedes Neuron den „eigenen“ Abstand mit einem globalen Referenzwert vergleicht, der kontinuierlich erhöht wird. Sobald das erste Neuron den Vergleich „gewinnt“, d.h. der Referenzwert größer geworden ist als sein Abstand, ist das Gewinnerneuron gefunden. Der zeitlich gesteuerte Lernvorgang, an dem die Nachbarschaft des Gewinnerneurons in monoton abnehmender Weise beteiligt wird, wird durch die Ausbreitung einer Schaltfront in einem aktiven Medium vermittelt. Dabei wird die Schaltfront durch das Gewinnerneuron selbst gezündet, der Lernvorgang jedes Neurons beginnt dann mit der Ankunft der Schaltfront an dessen Ort. Beendet wird der Lernvorgang für alle Neuronen gleichzeitig nach Ablauf einer vordefinierten Lernzeit. Die Anordnung der Neuronen auf dem aktiven Medium bestimmt dabei die Topologie des Kortex.

Nach dieser Einführung des Hardware-Konzeptes wird der im Rahmen dieser Arbeit gebaute Hardware-Prototyp erläutert. Er besteht aus 5 Neuronen in einem zweidimensionalen Musterraum, die von einer Thyristorstruktur mit einer linearen Anordnung von Gatekontakten gekoppelt werden. Dieses Halbleiterbauelement stellt das vorher beschriebene aktive Medium dar, das zur lokalen Zündung und Ausbreitung von Schaltfronten fähig ist. Die Neuronen koppeln jeweils an einen Gatekontakt der Thyristorstruktur, über den sie sowohl den lokalen Zustand der Probe (ein- oder ausgeschaltet) erfassen können, als auch die Zündung einleiten können. Weiterhin kann die Zündschwelle der Gatekontakte für die Gewinnersuche ausgenutzt werden. Zur Steuerung des Lernvorgangs sowie als Digital-Analog-Interface zur Übertragung der Mustervektoren an die Neuronen dient eine Steuerplatine auf Microcontroller-Basis. Nach den Erläuterungen zum Aufbau des Prototypen werden die einzelnen Baugruppen durch Testmessungen überprüft, dann werden Klassifikations- und Lernexperimente unter verschiedenen Bedingungen vorgestellt, die die einwandfreie Funktion des Hardwareaufbaus zeigen. Zuletzt werden Verbesserungsmöglichkeiten des Hardwarekonzepts erläutert, insbesondere im Hinblick auf eine hochintegrierte Version.

Das dritte Kapitel behandelt Thyristorstrukturen im Hinblick auf ihre Anwendung als aktives Medium der SOM-Hardware. Zunächst wird die Funktionsweise des Thyristors als Leistungsschalter erläutert, erst als quasi-eindimensionale pnpn-Struktur, dann als zweidimensionale flächenhafte Struktur, wie sie etwa industriell zum Schalten großer Spannungen und Ströme verwendet wird. Die Entstehung von Schaltfronten wird erklärt, und ein Reaktions-Diffusions-Modell wird eingeführt. Die Präparation der experimentell untersuchten Thyristorproben wird beschrieben, gefolgt von optischen und elektrischen Untersuchungen ihrer Eigenschaften. Optisch wurde die Ausbreitung von Zündfronten zeitlich aufgelöst direkt beobachtet. Die Bistabilität der Proben wurde durch Messung ihrer Kennlinien nach-

gewiesen. Die Zündung und Ausbreitung von Schaltfronten wurde durch elektrische Messungen untersucht. Die Frontbreite beträgt mehrere Millimeter, so daß ein entsprechend großer Gatekontakt-Abstand für eine SOM-Kopplungsstruktur erforderlich ist. Bestrahlung der Proben mit energiereichen Elektronen, um ihre Minoritätsträger-Lebensdauer zu senken, zeigen, daß dies zu einer Stabilisierung des ausgeschalteten Zustandes führt, mit der Konsequenz einer stark verminderten Frontbreite, aber auch verringerter Frontgeschwindigkeit und einer erhöhten kritischen Stromdichte. Dies ermöglicht eine bessere Detektion und Lokalisierung der Schaltfronten, ist aber mit erhöhtem Strombedarf verbunden. Weiterhin wurden stationäre Fronten beobachtet, wobei sich die Ausdehnung des eingeschalteten Gebiets durch den Anodenstrom steuern läßt. Dabei tritt eine Hysterese durch Pinning an Inhomogenitäten auf, die durch die eingeätzten Gatekontakte verursacht werden. Während erste von Hand präparierte Proben eine starke Streuung der Zündströme an den einzelnen Gatekontakten aufweisen, zeigen industriell präparierte Proben eine gute Homogenität.

Im letzten Kapitel werden nun Optimierungen der Thyristorstruktur im Hinblick auf Miniaturisierung und Integration mit den restlichen Bestandteilen des Hardware-Konzeptes vorgestellt, die mit Hilfe von Simulationsrechnungen gewonnen wurden. Zunächst wird das verwendete kommerzielle Simulationsprogramm `dessis` und dessen physikalische Modellgleichungen vorgestellt. Dann wird die Simulation einer Vierschichtstruktur erläutert, die den im dritten Kapitel vorgestellten Halbleiterproben entspricht. Sowohl die simulierten Kennlinien als auch die erhaltenen Frontgeschwindigkeiten und Frontbreiten entsprechen im Rahmen der zu erwartenden Genauigkeit den experimentellen Messungen an den Thyristorproben. Die Trägerlebensdauer wird als Parameter variiert, um den Effekt der lebensdauersenkenden Elektronenbestrahlung zu simulieren. Dies zeigt den großen Einfluß der Trägerlebensdauer sowohl auf die statischen Eigenschaften der pnpn-Struktur, insbesondere den Haltestrom und die Zündempfindlichkeit, als auch auf die Frontausbreitung. Daraufhin werden schrittweise verkleinerte Strukturen simuliert, bis hin zu einer Gesamtdicke im Mikrometerbereich, um die Einflüsse der Miniaturisierung auf den Frontzündungs- und Ausbreitungsprozeß zu analysieren. Wie erwartet, nimmt die Frontbreite nahezu proportional mit der Schichtdicke des Vierschichtsystems ab, während die Frontgeschwindigkeit beträchlich gesteigert werden kann. So steht die Möglichkeit einer sehr schnellen Hardware-Realisation des Kohonen-Algorithmus offen, sofern die Geschwindigkeit der restlichen Komponenten der SOM-Hardware ebenso gesteigert werden kann. Bei der Verkleinerung der Basisweiten der pnpn-Struktur müssen allerdings die Dotierungen der Basen angehoben werden, um die Bistabilität der Struktur zu erhalten. Dies erhöht die Querkopplung der Struktur, was Frontgeschwindigkeit und Frontbreite (relativ zur Strukturdicke) erhöht.

Um einerseits die Frontbreite relativ zur durch die Halbleiterpräparation begrenzten Strukturdicke weiter zu verkleinern, und andererseits den Strombedarf der Kopplungsstruktur zu senken, wurde ein alternativer Aufbau eines aktiven Mediums untersucht. Es handelt sich um eine Kette, bzw. in zwei Dimensionen ein Gitter von diskreten Thyristorelementen. Durch die Wahl von Kopplungs- und Lastwiderstand ergibt sich eine große Flexibilität bei der Festlegung von Frontbreite und Frontgeschwindigkeit. Der Aufbau der einzelnen Thyristorelemente mitsamt deren Lastwiderstand bestimmt die Schaltzeit der Einzelelemente, während die Wahl des Kopplungswiderstandes die Frontlaufzeit von einem Element zum nächsten bestimmt. Einzelemente in Mikrometerabmessungen wurden simuliert und zeigen wegen des kleinen Querschnittes sehr geringe Halteströme. So ist kann die vorher beschriebene Kopplungsstruktur für die SOM-Hardware wesentlich leistungssparender realisiert werden. Um die prinzipiell möglichen kleinen Ströme zu realisieren sind allerdings hochohmige Lastwiderstände oder etwa aktive Lasten, wie sie mit MOS-Transistoren realisiert werden können, erforderlich.

Die Thyristoreigenschaften wie Haltestrom und Kippspannung lassen sich durch das Dotierungsprofil und die Lebensdauerverteilung einstellen, wobei die gefundenen Abhängigkeiten so moderat sind, daß man mit den verfügbaren Herstellungsverfahren recht geringe Inhomogenitäten eines solchen aktiven Mediums erwarten kann. Es muß sichergestellt sein, daß bei der vorgesehenen Versorgungsspannung alle Thyristorelemente sicher bistabil sind, und möglichst gleiche Zündströme aufweisen. Die guten Erfahrungen mit den experimentell untersuchten Thyristorproben lassen auch für entsprechend miniaturisierte Proben eine gute Homogenität erwarten, so daß die Integration einer analogen selbstorganisierenden Karte nach dem in dieser Arbeit vorgestellten Konzept erfolgversprechend erscheint.

Anhang A

Lösungen der einkomponentigen Reaktionsdiffusionsgleichung

Im folgenden werden stationäre und dynamische Lösungen des einkomponentigen Reaktions-Diffusions-Systems (Gl. A.1) erläutert. Dabei soll das Grundgebiet zunächst unbegrenzt sein, und $y(x,t)$ im gesamten Definitionsbereich endlich bleiben.

$$\frac{\partial y(x,t)}{\partial t} = D\frac{\partial^2 y(x,t)}{\partial x} + f(y(x,t)) \tag{A.1}$$

Die Reaktionsfunktion $f(y)$ bestimmt das Verhalten des Systems. Zunächst betrachten wir die stationären Lösungen. Die Nullstellen der Reaktionsfunktion ergeben konstante, homogene Lösungen. Im Falle der hier skizzierten Reaktionsfunktion gibt es 3 Nullstellen, also 3 Fixpunkte, von denen y_1 und y_3 stabil sind, y_2 instabil.

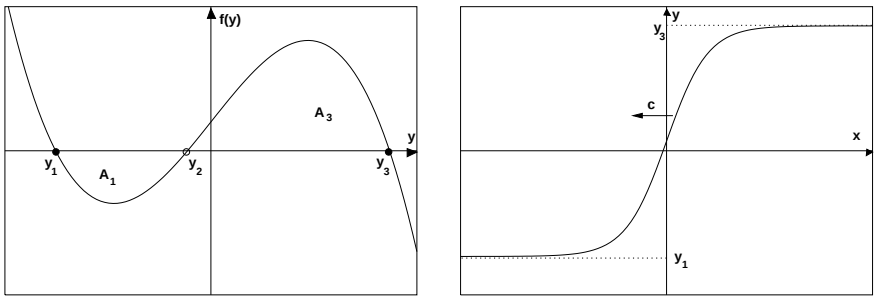


Abbildung A.1: Kubische Reaktionsfunktion $f(y)$ und daraus resultierende Front.

Präpariert man einen zweigeteilten Anfangszustand, wobei ein Teil im Zustand y_1 und der andere im Zustand y_3 ist, so relaxiert das System zu einer Frontlösung (s. Abb. A.1). Die Form und Bewegungsgeschwindigkeit dieser Front wird bestimmt durch das Wechselspiel zwischen dem linearen Diffusionsterm und dem nichtlinearen Reaktionsterm.

Die mit konstanter Geschwindigkeit laufende Front soll nun durch Transformation von Gl. A.1 in ein mitbewegtes Koordinatensystem gefunden werden.

$$x' = x - ct, t' = t \tag{A.2}$$

$$0 = D\frac{\partial^2 y(x')}{\partial x'} + c\frac{\partial y(x')}{\partial x'} + f(y(x')) \tag{A.3}$$

Dabei ist die Geschwindigkeit c eine Variable, die zusammen mit $y(x')$ zu bestimmen ist. Diese transformierte Gleichung hat die formal gleiche Form wie die Newton'sche Gleichung für die Bewegung eines Massenpunktes in einem Kraftfeld, wenn x durch t und y durch x ersetzt wird. In dieser Betrachtung stellt der Term erster Ordnung um c einen Reibungsterm dar.

Substituiert man $v = dy/dx'$ so erhält man eine Gleichung in y:

$$0 = Dv'(y)v(y) + cv(y) + f(y) \tag{A.4}$$

Für den Fall $c = 0$ läßt sich diese Gleichung von y_1 bis y_3 integrieren, wenn man beachtet, daß $v(y)v'(y) = \frac{dv^2(y)}{2dy}$ ist, und $v(y)$ bei y_1 und y_3 Null ist.

$$0 = \frac{D}{2}(v^2(y_3) - v^2(y_1)) = \int_{y_1}^{y^3} f(y)\,dy \tag{A.5}$$

Dies ist die Equal-Area-Rule:

$$\int_{y_1}^{y_3} f(y)\,dy = A = 0 \Longleftrightarrow c = 0 \tag{A.6}$$

Das Integral stellt die Gesamtfläche A unter dem Graphen der Reaktionsfunktion zwischen den beiden stabilen Nullstellen dar. Eine stationäre Front erhält man also nur für den Fall gleicher Flächen unter dem Graphen. Im anderen Fall ist der zur größeren Fläche gehörende Zustand dominant, und die Front bewegt sich so, daß sich dieser Zustand ausbreitet.

Die Frontgeschwindigkeit c und das dazugehörige $y(x')$ kann in diesem Fall nur numerisch bestimmt werden. Dazu kann A.3 oder A.4 durch numerische Integration näherungsweise gelöst werden, wobei letzteres vorteilhaft ist, da letztere eine gewöhnliche Differentialgleichung ODE 1. Ordnung ist, während erstere 2. Ordnung ist und außerdem von $-\infty$ bis ∞ integriert werden muß.

Um diese Schwierigkeiten zu umgehen, wurde die ursprüngliche, zeitabhängige partielle Differentialgleichung diskretisiert, so daß auch der zeitlichen Ablauf der Frontentstehung und -ausbreitung simuliert werden kann. Der Nachteil dieses Vorgehens ist die Beschränkung auf begrenzte Grundgebiete und die damit verbundenen Randeffekte, sowie die erhöhte Rechenzeit.

A.1 Numerische Integration der RD-Gleichung

Die ursprüngliche Reaktions-Diffusions-Gleichung (Gl. A.7) wird mit einem Finite-Differenzen-Verfahren in Ort und Zeit diskretisiert. Das endliche, eindimensionale Grundgebiet wird in eine bestimmte Anzahl Gitterpunkte unterteilt, wobei ein konstanter Gitterabstand Δx die Aufstellung der Differenzengleichung erleichtert. Die Zeitkoordinate wird in Form von gleichmäßigen Zeitschritten Δt diskretisiert. Die numerische Integration beginnt mit der Vorgabe eines Anfangszustandes und bestimmt durch Lösung der Differenzengleichung jeweils den Zustand zum nächsten Zeitschritt.

$$\frac{\partial y(x,t)}{\partial t} = D\frac{\partial^2 y(x,t)}{\partial x} + f(y(x,t)) \tag{A.7}$$

Das einfachste mögliche Diskretisierungsschema ist das folgende explizite Euler-Schema.

$$\frac{1}{\Delta t}(y_i^{n+1} - y_i^n) = \frac{D}{\Delta x^2}(y_{i-1}^n - 2y_i^n + y_{i+1}^n) + f(y_i^n) \tag{A.8}$$

Hierbei sind die y_i^n die Werte zum Zeitschritt n, aus denen die y_i^{n+1} bestimmt werden müssen. Es kann direkt nach y_i^{n+1} aufgelöst werden, da die räumlichen

Ableitungen zum Zeitschritt n gebildet werden. Das Verfahren ist jedoch nur dann stabil, wenn Δt klein genug gewählt wird [Pre92]:

$$\frac{2D\Delta t}{(\Delta x)^2} \leq 1 \tag{A.9}$$

Die kleinste erlaubte Zeitschrittweite wird bestimmt durch die charakteristische Diffusionszeit für Strukturen der Breite Δx. Moden hoher Raumfrequenz werden in der diskretisierten Gleichung als erste instabil, und begrenzen damit die Größe der Zeitschritte. Durch diese bedingte Stabilität benötigt ein explizites Schema übermäßig viel Rechenzeit und ist daher ineffizient.

Daher wurde für weitere Versuche ein implizites Euler-Schema 1. Ordnung implementiert.

$$\frac{1}{\Delta t}(y_i^{n+1} - y_i^n) = \frac{D}{\Delta x^2}(y_{i-1}^{n+1} - 2y_i^{n+1} + y_{i+1}^{n+1}) + f(y_i^{n+1}) \tag{A.10}$$

Hier werden die räumlichen Ableitungen und die Reaktionsfunktion zum (noch unbekannten) Zeitschritt $n+1$ bestimmt. Daher muß die Gleichung nach y_i^{n+1}aufgelöst werden. Vernachlässigt man den nichtlinearen Reaktionterm, muß für jedem Zeitschritt ein lineares Gleichungssystem gelöst werden. Handelt es sich, wie in unserem Fall, um ein 1-dimensionales System, so ist das entstehende System tridiagonal und somit ohne großen Aufwand lösbar.

Der nichtlineare Reaktionsterm $f(y_i^{n+1})$ macht ein Iterationsverfahren nach Art des Newton-Verfahrens zur Nullstellenbestimmung einer nichtlinearen Funktion notwendig. Der Term wird zunächst durch Taylor-Entwicklung in erster Ordnung linearisiert:

$$f(y_i^{n+1}) = f(y_i^n) + (y_i^{n+1} - y_i^n)f'(y_i^n) + \cdots \tag{A.11}$$

Nach Einsetzen in Gl. A.10 ergibt sich ein lineares Gleichungssystem. Dessen Lösung y_i^{n+1} ersetzt den Anfangswert y_i^n in der Linearisierung, so daß sich nun eine verbesserte Lösung ergibt. Diese Newton-Iteration konvergiert gewöhnlich schon nach wenigen Iterationsschritten. Das Newton-Verfahren ist empfindlich auf die Wahl des Anfangszustandes, mit dem die Iteration begonnen wird. Dieser darf nicht zu weit von der gesuchten Nullstelle entfernt liegen, damit die Konvergenz gesichert ist. Da hier der Zustand y_n zur Initialisierung dient, kann die Konvergenz immer durch Reduzierung der Zeitschrittweite Δt erreicht werden. Praktisch

gesehen ließen sich aber immernoch wesentlich größere Zeitschrittweiten als mit dem expliziten Verfahren erreichen. Abgesehen von der Newton-Iteration ist das Verfahren unbedingt stabil, es gibt also keine so gravierende Einschränkung der Zeitschrittweite wie beim expliziten Verfahren.

Basierend auf diesem impliziten Verfahren wurde ein interaktives Simulationsprogramm inklusive graphischer Benutzeroberfläche implementiert, mit dessen Hilfe man die zeitliche Entwicklung des Systemzustandes in Echtzeit verfolgen und durch Parameteränderungen beeinflussen kann. Als Anfangszustand wird ein Stufenfunktion verwendet, die nach dem Beginn der Simulation rasch zu einer glatten Frontlösung konvergiert. Damit die Front das Grundgebiet nicht verläßt, kann durch regelmäßige Translation der Lösung das Fortschreiten der Front kompensiert werden. Es wird einfach die gesamte Lösung um einen Gitterpunkt nach hinten verschoben, wenn sich die Front um mehr als einen Punkt fortbewegt hat. So kann auch die jeweilige Frontgeschwindigkeit der Front akkurat bestimmt werden, ohne daß Randeffekte störend werden.

Mit diesem Programm wurden die numerischen Simulationen zum einkomponentigen RD-Modell der Frontausbreitung aus Abschnitt 3.5 berechnet.

A.2 Skalierung der einkomponentigen RD-Gleichung

Es ist interessant zu betrachten, wie sich Änderungen der Parameter einer Differentialgleichung auf die Lösung auswirken. Nehmen wir also im folgenden an, die Lösung der folgenden RD-Gleichung sei bekannt, und es handele sich um eine sigmoide Funktion, sie habe also die typische Form einer Front. Die Zeitentwicklung der Lösung ist, wie im letzten Abschnitt gezeigt, eine einfache Translationsbewegung.

$$\tau \frac{\partial y(x,t)}{\partial t} = l^2 \frac{\partial^2 y(x,t)}{\partial x} + f(y(x,t)) \tag{A.12}$$

Nun führen wir eine Koordinatentransformation aus und skalieren die Orts- und Zeitkoordinaten, die bekannte Lösung $y(x,t)$ wird also zu $y(x',t') = y(ax,bt)$:

$$x' = ax, t' = bt \tag{A.13}$$

$$\frac{\partial y}{\partial x'} = a\frac{\partial y}{\partial x}, \frac{\partial y}{\partial t'} = b\frac{\partial y}{\partial t} \tag{A.14}$$

$$\tag{A.15}$$

Gleichung A.12 wird zu:

$$b\tau\frac{\partial y(x',t')}{\partial t'} = a^2l^2\frac{\partial^2 y(x',t')}{\partial x'} + f(y(x',t')) \tag{A.16}$$

Eine Skalierung von Raum- und Zeitskala führt also zu entsprechenden Parameteränderungen. Im einzelnen sind dies:

- Skalierung der Zeitskala.

  Sei $a = 1, b \neq 1$. Im skalierten Koordinatensystem erhält man die RD-Gleichung:

  $$b\tau\frac{\partial y(x',t')}{\partial t'} = l^2\frac{\partial^2 y(x',t')}{\partial x'} + f(y(x',t')) \tag{A.17}$$

  Die Skalierung der Zeitskala ist also zu einer entsprechende Änderung der Zeitkonstante τ äquivalent. Wird τ um den Faktor b erhöht, so läuft die Lösung um den Faktor b langsamer ab, die Frontgeschwindigkeit wird also: $v' = 1/bv$.

  Der räumliche Verlauf der Lösung bleibt erhalten.

- räumliche Skalierung.

  Sei $a \neq 1, b = 1$. Die Raumkoordinate x wird also um den Faktor a gestreckt. Frontbreite und Frontgeschwindigkeit werden um den Faktor a erhöht. Die RD-Gleichung wird zu:

  $$\tau\frac{\partial y(x',t')}{\partial t'} = a^2l^2\frac{\partial^2 y(x',t')}{\partial x'} + f(y(x',t')) \tag{A.18}$$

  Eine räumliche Skalierung um a ist also äquivalent zur Veränderung der Diffusionskonstante um a^2.

- Skalierung der Reaktionsfunktion.

  Sei $a^2 = b$. Die RD-Gleichung wird zu:

$$\tau \frac{\partial y(x', t')}{\partial t'} = l^2 \frac{\partial^2 y(x', t')}{\partial x'} + \frac{1}{b} f(y(x', t')) \tag{A.19}$$

  Die Skalierung der Reaktionsfunktion mit dem Faktor $1/b$ dehnt also die Zeitskala um den Faktor b und die räumliche Skala um $\sqrt{b}$. Dementsprechend ändert sich die Geschwindigkeit um den Faktor $1/\sqrt{b}$.

Literaturverzeichnis

[And94] T. R. Anderson und R. D. Patterson. Speaker recognition with the auditory image model and self-organizing feature maps: A comparison with traditional techniques. In: *ESCA Workshop on Automatic Speaker Recognition Identification and Verification*, 153–6. Armstrong Lab. , Wright Res. & Dev. Center, Wright-Patterson AFB, OH, USA, IDIAP, Martingny, Switzerland (1994).

[Ast98] Y. A. Astrov, I. Müller, E. Ammelt und H.-G. Purwins. Zigzag Destabilyzed Spirals and Targets. *Phys. Rev. Lett.*, **80**, 5341–5344 (1998).

[Ast01] Y. A. Astrov und H.-G. Purwins. Plasma spots in a gas discharge system: birth, scattering and formation of molecules. *Physics Letters A*, **238**, 349–354 (2001).

[Bal87] B. J. Baliga. *Modern Power Devices*. John Wiley & Sons (1987).

[Bal95] Baldanza. Results from an on-line non-leptonic neural trigger implemented in an experiment looking for beauty. *Nucl. Instrum. Methods A*, **361**, 506 (1995).

[Ban85] R. Bank, J. W. M. Coughran, W. Fichtner, E. H. Grosse, D. J. Rose und R. K. Smith. Transient simulation of silicon devices and circuits. *IEEE Transactions on Electron Devices*, **32** (10), 1992–2007 (1985).

[Bea93] L. Beauge, S. Durand und F. Alexandre. Plausible self-organizing maps for speech recognition. In: R. F. Albrecht, C. R. Reeves und N. C. Steele, Hg., *Artificial Neural Nets and Genetic Algorithms. Proceedings of the International Conference*, 221–6. CRIN-CNRS/INRIA Lorraine, Vandoeuvre-les-Nancy, France, Springer-Verlag, Berlin, Germany (1993).

[Bel97] I. Belic und L. Gyergyek. Neural network methodologies for mass spectra recognition. *Vacuum*, **48** (7–9), 633–7 (1997). (5th European Vacuum

Conference, EVC-5 Conf. Date: 23–27 Sept. 1996 Conf. Loc: Salamanca, Spain).

[Bod93] M. Bode. *Beschreibung strukturbildender Prozesse in eindimensionalen Reaktions-Diffusions-Systemen durch Reduktion auf Amplitudengleichungen und Elementarstrukturen*. Dissertation, Institut für Angewandte Physik, Westfälische Wilhelms-Universität Münster (1993).

[Bod01a] M. Bode, O. Freyd, J. Fischer, F.-J. Niedernostheide und H.-J. Schulze. Hybrid Hardware for a Highly Parallel Search in the Context of Learning Classifiers. *Artificial Intelligence*, **130** (1), 75–84 (2001).

[Bod01b] M. Bode, A. Liehr, C. Schenk und H.-G. Purwins. Interaction of dissipative Solitons: Particle-Like Behaviour of Localized Structures in a Three-Component Reaction-Diffusion System. *Physica D* (2001). Zur Veröffentlichung angenommen.

[Cal99] D. E. Callan, R. D. Kent, N. Roy und S. M. Tasko. Self-organizing map for the classification of normal and disordered female voices. *Journal of Speech, Language, and Hearing Research*, **42**, 355–66 (1999).

[Cha94] M. V. Chan, X. Feng, J. A. Heinen und R. J. Niederjohn. Classification of Speech Accents with Neural Networks. In: *Proc. ICNN'94, International Conference on Neural Networks*, 4483–4486. IEEE Service Center, Piscataway, NJ (1994).

[DeS88] DeSieno. Adding a conscience to competitive learning. In: *IEEE International Conference on Neural Networks*, Bd. 1, 117–124 (1988).

[des00] ISE Integrated Systems Engineering, Balgriststrasse 102, CH-8008 Zürich, Switzerland. *DESSIS-ISE Manual, ISE TCAD Release 6.1* (2000). URL http://www.ise.ch

[D'y77] M. I. D'yakonov und M. E. Levinshtein. Theory of propagation of the turned-on state in a thyristor. *Fiz. Tekh. Poluprovodn.*, **12**, 729–741 (1977).

[Elo92] P. Elo, J. Saarinen, A. Värri, H. Nieminen und K. Kaski. Classification of Epileptic EEG by Using Self-Organizing Maps. In: I. Aleksander und J. Taylor, Hg., *Artificial Neural Networks, 2*, Bd. II, 1147–1150. North-Holland, Amsterdam, Netherlands (1992).

[eta92] Intel Corp., 2250 Mission College Boulevard, Santa Clara, CA 95052-8125, USA. *80170NX Electrically Trainable Analog Neural Network Data Booklet* (1992).

[Fif79] P. Fife. *Mathematical aspects of reacting and diffusing systems*, Bd. 28 von *Lecture notes in biomathematics*. Springer-Verlag (1979).

[Fis00] J. Fischer. *Hardware-Realisation Neuronaler Netze mit Hilfe aktiver Medien auf Halbleiterbasis*. Diplomarbeit, Institut für Angewandte Physik, Westfälische Wilhelms-Universität Münster (2000).

[Fos76] J. Fossum. Computer-aided numerical analysis of Silicon solar cells. *Solid-State Electronics*, **19**, 269–277 (1976).

[Fos82] J. Fossum und D. S. Lee. A physical model for dependence of carrier lifetime on doping density in nondegenerate Silicon. *Solid State Electronics*, **25**, 741–747 (1982).

[Fri95] B. Fritzke. A growing neural gas network learns topologies. In: G. Tesauro, D. S. Touretzky und T. K. Lean, Hg., *Advances in Neural Information Processing Systems*, 625–632. MIT Press (1995).

[Ger79] W. Gerlach. *Thyristoren*. Halbleiter-Elektronik Bd. 12. Springer-Verlag (1979).

[Ges92] N. Geschwind. Die Großhirnrinde. In: *Spektrum der Wissenschaft: Gehirn und Nervensystem*, 76–85. Spektrum der Wissenschaft Verlag (1992).

[Hal52] R. Hall. Electron-hole recombination in germanium. *Physical Review*, **87**, 387 (1952).

[Har98] R. Harrison, P. Hasler und B. Minch. Floating-Gate CMOS Analog Memory Cell Array. In: *In Proceedings of the IEEE International Symposium on Circuits and Systems*, Bd. 2, 204–207 (1998).

[Hay99] S. Haykin. *Neural Networks: a comprehensive foundation*. Prentice-Hall (1999).

[Haz99] P. Hazdra und J. Vobecky. Nondestructive defect characterization and engineering in contemporary silicon power devices. *solid state phenomena*, **70**, 545–550 (1999).

[Hod52] A. L. Hodgkin und A. F. Huxley. A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve. *Journal of Physiology*, **117**, 500–544 (1952).

[Hub62] D. H. Hubel und T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, **160**, 106–154 (1962).

[Hub70] D. H. Hubel und T. N. Wiesel. The period of susceptibility to the physiological effects of unilateral eye closure in kittens. *Journal of Physiology*, **206**, 419–436 (1970).

[Kla92] D. B. M. Klaassen, J. Slotboom und H. de Graaf. Unified apparent bandgap narrowing in n- and p-type Silicon. *Solid State Electronics*, **35** (2), 125–129 (1992).

[Koh82] T. Kohonen. Analysis of a simple self-organizing process. *Biol. Cyb.*, **44** (2), 135–140 (1982).

[Koh89] T. Kohonen. The 'Neural' Phonetic Typewriter. In: *The Second European Seminar on Neural Networks, London, UK, February 16–17*. British Neural Networks Society, London, UK (1989).

[Lin80] Y. Linde, A. Buzo und R. M. Gray. An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications*, **COM-28**, 84–95 (1980).

[Lut06] R. Luther. Räumliche Fortpflanzung chemischer Reaktionen. *Zeitschrift für Elektrochemie*, **12** (32), 596–600 (1906).

[Mac90] D. Macq, J.-D. Legat und P. Jespers. Analog storage of adjustable weights. In: *Proc. SPIE Conf. Applications of Neural Networks*, 456–461 (1990).

[Mar93] T. Martinetz, S. berkovich und K. Schulten. Neural-gas network for vector quantization and its application to time series prediction. *IEEE Transactions on Neural Networks*, **4** (4), 558–589 (1993).

[McC43] W. McCulloch und W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, 115–133 (1943).

[Mea89] C. A. Mead. *Analog VLSI and neural systems*. Addison-Wesley (1989).

[Mei97] M. Meixner, P. Rodin und E. Schöll. Global Control of Front Propagation in Gate-Driven Multilayered Structures. *phys. stat. sol. (b)*, **204**, 493–496 (1997).

[Mei98a] M. Meixner, P. Rodin und E. Schöll. Accelerated, decelerated, and oscillationg fronts in globally coupled bistable semiconductor system. *Physical Review E*, **58** (3), 2796–2807 (1998).

[Mei98b] M. Meixner, P. Rodin und E. Schöll. Fronts in an bistable medium with two global constraints: Oscillatory instability and large-amplitude limit-cycle motion. *Physical Review E*, **58** (5), 5586–5591 (1998).

[Min69] M. Minsky und S. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press (1969).

[Mos96] W. W. Moses, E. Beuville und M. H. Ho. A „winner-take-all“ IC for determining the crystal of interaction in PET detectors. *IEEE Trans. Nucl. Sci.*, **NS-43**, 1615–1618 (1996).

[Nat97] C. D. Natale und A. D'Amico. Modelling and data analysis of multisensor systems with the self-organizing map: application to the electronic nose. In: *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4–6*, 14–19. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland (1997).

[Nes00] R. Neswold. *A GNU Development Environment for the AVR Microcontroller*. rneswold@enteract.com (2000).
URL http://www.enteract.com/~{}rneswold/avr/avr-lib.pdf

[Nie01] F.-J. Niedernostheide, H.-J. Schulze, O. Freyd, M. Bode und A. V. Gorbatyuk. Realization of a neural algorithm by using front-propagation in a thyristor-based hybrid system. *Chaos, Solitons, & Fractals* (2001). Zur Veröffentlichung eingereicht.

[Pen37] W. Penfield und E. Boldrey. Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. *Brain*, **60**, 389–443 (1937).

[Pla97] M. Plagge. *Präparation und experimentelle Untersuchung von Thyristorstrukturen für die Hardware-Realisierung von Kohonennetzen*. Diplomarbeit, Institut für Angewandte Physik, Westfälische Wilhelms-Universität Münster (1997).

[Pre92] W. H. Press, S. A. Teukolsky, W. T. Vetterling und B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press (1992).

[Rüp95] S. Rüping, U. Rücking und K. Goser. A chip for self-organizing feature maps. *IEEE MICRO*, **15** (3), 57–59 (1995).

[Rug84] I. Ruge. *Halbleiter-Technologie*. Springer-Verlag (1984).

[Rum86] D. Rumelhart, G. Hinton und R. Williams. Learning representations of back-propagation errors. *Nature*, **323**, 533–536 (1986).

[Ruw98] D. Ruwisch. *Physikalische Realisierung neuartiger Kopplungsmechanismen in neuronalen Netzwerken*. Dissertation, Institut für Angewandte Physik, Westfälische Wilhelms-Universität Münster (1998).

[Sch00] C. Schenk, A. W. Liehr, M. Bode und H.-G. Purwins. Quasi-Particles in a Three-Dimensional Three-Component Reaction-Diffusion System. In: W. J. E. Krause, Hg., *High Performance Computing in Science and Engineering 1999*, 354–364. Springer (2000).

[Sch01] D. Schleef et al. *Linux control and measurement device interface*. ds@schleef.com (2001).
URL http://stm.lbl.gov/comedi/

[Sho52] W. Shockley und W. T. Read. Statistics of the recombination of holes and electrons. *Physical Review*, **87**, 835–842 (1952).

[Str94] S. H. Strogatz. *Nonlinear dynamics and chaos*. Addison Wesley (1994).

[Sug79] N. Suga und W. E. O'Neill. Neural axis representing target range in the auditory cortex of the mustache bat. *Science*, **206**, 351–353 (1979).

[Tie78] U. Tietze und C. Schenk. *Halbleiter-Schaltungstechnik*. Springer Verlag, 4 Aufl. (1978).

[Tur52] A. M. Turing. The Chemical Basis of Morphogenesis. *Phil. Trans. Roy. Soc.*, **237**, 37 (1952).

[Var70] I. V. Varlamov, V. V. Osipov und E. A. Poltoratskii. Current filamentation in a four-layer structure. *Soviet Physics - Semiconductors*, **3** (8), 979–982 (1970).

[Vit91] E. Vittoz, H. Oguey, M. A. Maher, O. Nys, E. Dijkstra und M. Chevroulet. Analog Storage of Adjustable Synaptic Weights. In: U. Ramacher und U. Rückert, Hg., *VLSI Design of Neural Networks*, 47–63. Kluwer (1991).

[Wer88] P. J. Werbos. Backpropagation, Past and Future. In: *Proceedings of the International Conference on Neural Networks, I*, 343–353. IEEE Press (1988).

Danksagung

Diese Arbeit wurde im Institut für Angewandte Physik der Westfälischen Wilhelms-Universität angefertigt. Mein erster Dank gilt daher Herrn Prof. Dr. H.-G. Purwins für die Aufnahme in seine Arbeitsgruppe und die Förderung dieser Arbeit. Herrn Priv.-Doz. Dr. M. Bode und Herrn Priv.-Doz. Dr. F.-J. Niedernostheide danke ich für die maßgebliche Betreuung meiner Arbeit. Herrn Dr. Niedernostheide und der Fa. *Siemens* bzw. *Infineon* gilt besondere Anerkennung für die Beschaffung und Präparation der untersuchten Halbleiterproben. Viel Einblick in die Halbleiterphysik im speziellen und die russische Kultur im allgemeinen vermittelte mir Dr. A. Gorbatyuk.

Für ihre Anregungen und Ideen und die unzähligen interessanten Diskussionen danke ich Herrn Dr. J. Berkemeier, Andreas Burwick, Lars Stollenwerk, Andreas Liehr, Sandra Zuccaro, und Stefan Flothkötter.

Ein spezieller Dank gilt meinem Diplomanden und Mitstreiter Jens Fischer für seine tatkräftige Mitarbeit an der Demonstrator-Hardware. Die gesamte Arbeitsgruppe verdient ein Lob wegen der stets angenehmen Arbeitsatmosphäre.

Mein tief empfundener Dank gilt meinen Eltern, die mir das Studium ermöglichten, und nicht zuletzt meiner Freundin Vilma, für ihre liebevolle Unterstützung und Geduld, und allen anderen, denen ich in den letzten Monaten zuwenig Zeit widmete, und die es mir dennoch nicht verübelten.

Lebenslauf

Name:	Oliver Freyd
geboren am:	20.12.1969 in Aachen
Familienstand:	ledig
Namen der Eltern:	Martin und Sabine Freyd, geb. Mueller
Schulbildung:	Grundschule Saarstraße von 1976 bis 1980 in Aachen Couven–Gymnasium von 1980 bis 1989 in Aachen
Abitur:	3.5.1989 in Aachen
Studium:	von 1989 bis 1996 an der Rheinisch-Westfälischen Technischen Hochschule Aachen von 1992 bis 1993 einjähriger Auslandaufenthalt mit Erasmus-Stipendium an der Universität Trieste, Italien
Prüfungen:	Physik-Diplom am 5.12.1996
Zivildienst:	vom 1.1997 bis zum 1/1998 im Luisenhospital Aachen
Tätigkeiten:	1994/95 Wissenschaftliche Hilfskraft am 1. Physikalischen Institut der RWTH Aachen 1995/96 Wissenschaftliche Hilfskraft am Institut für Produktionstechnologie Wissenschaftlicher Mitarbeiter seit dem 1.2.1998 am Institut für Angewandte Physik der Westfälischen Wilhelms-Universität Münster
Promotionsstudium:	von SS 1998 bis SS 2001 an der Westfälischen Wilhelms-Universität Münster
Beginn der Dissertation:	1.2.1998 am Institut für Angewandte Physik der Westfälischen Wilhelms-Universität Münster Betreuer: Priv.-Doz. Dr. M. Bode