Kapitel 2

High-κ-Dielektrika

Die treibenden Kräfte für die in der Halbleiterindustrie seit über drei Dekaden verfolgte Skalierung der elektronischen Bauelemente zu immer kleineren Strukturen sind hauptsächlich betriebswirtschaftlicher, aber auch technischer Natur. *Moore* stellte 1965 fest, dass sich der Komplexitätsgrad eines Chips (Packungsdichte der Komponenten), welcher bei feststehenden Herstellungskosten produziert werden kann, in einem charakteristischen Zeitintervall verdoppelt. Seine weitere Beobachtung, dass sich die minimalen Herstellungskosten pro Komponente auf einem integrierten Chip während des gleichen Zeitintervalls halbieren, bedeutet, dass die Kosten zur Herstellung eines Chips mit jeweils vergrößertem, optimiertem Komplexitätsgrad über die Zeit konstant sind [Hut05]. Dieser Funktionszugewinn bei gleichbleibenden Herstellungskosten eines Chips über die Produktionszyklen hinweg wird ökonomisch genutzt und erfordert die Skalierung der Chip-Komponenten zu kleineren Strukturen.

Die entscheidende Rolle aus technischer Sicht spielt die Ladung, welche von der mit einer angelegten Spannung versorgten Elektrode auf der gegenüberliegenden, durch einen Isolator getrennten Elektrode induziert wird. Sowohl bei Schaltelementen als auch bei Speicherelementen der CMOS-Technologie soll trotz abnehmender Strukturgrößen eine ausreichend große Ladungsmenge transportiert bzw. gehalten werden. Die in der Halbleitertechnik ausgenutzte Haupteigenschaft von Isolatoren mit hohen Dielektrizitätszahlen (*high-κ dielectrics*), eingesetzt in Kondensatoranordnungen zwischen leitenden oder halbleitenden Materialien, ist die Fähigkeit, bei einer vorgegebenen Spannung und Geometrie in solchen Bauelementen mehr Ladung pro Fläche speichern zu können.

$$\frac{Q}{A} = D = \varepsilon \cdot E = \varepsilon \frac{V}{d} \tag{2.1}$$

Dabei entspricht Q der Ladung, A der Kondensatorfläche, D der dielektrischen Verschiebungsdichte, ε der Permittivität des isolierenden Mediums (Produkt aus elektrischer Feldkonstante ε_0 und Dielektrizitätszahl ε_r), E dem elektrischen Feld, V der angelegten Spannung und d der Dicke des Dielektrikums. Die Eigenschaft der spannungsabhängigen Ladungsspeicherung wird mit der Größe der Kapazität $C = Q/V = \varepsilon A/d$ bewertet und einerseits für Speicherelemente, andererseits für Transistoren ausgenutzt.

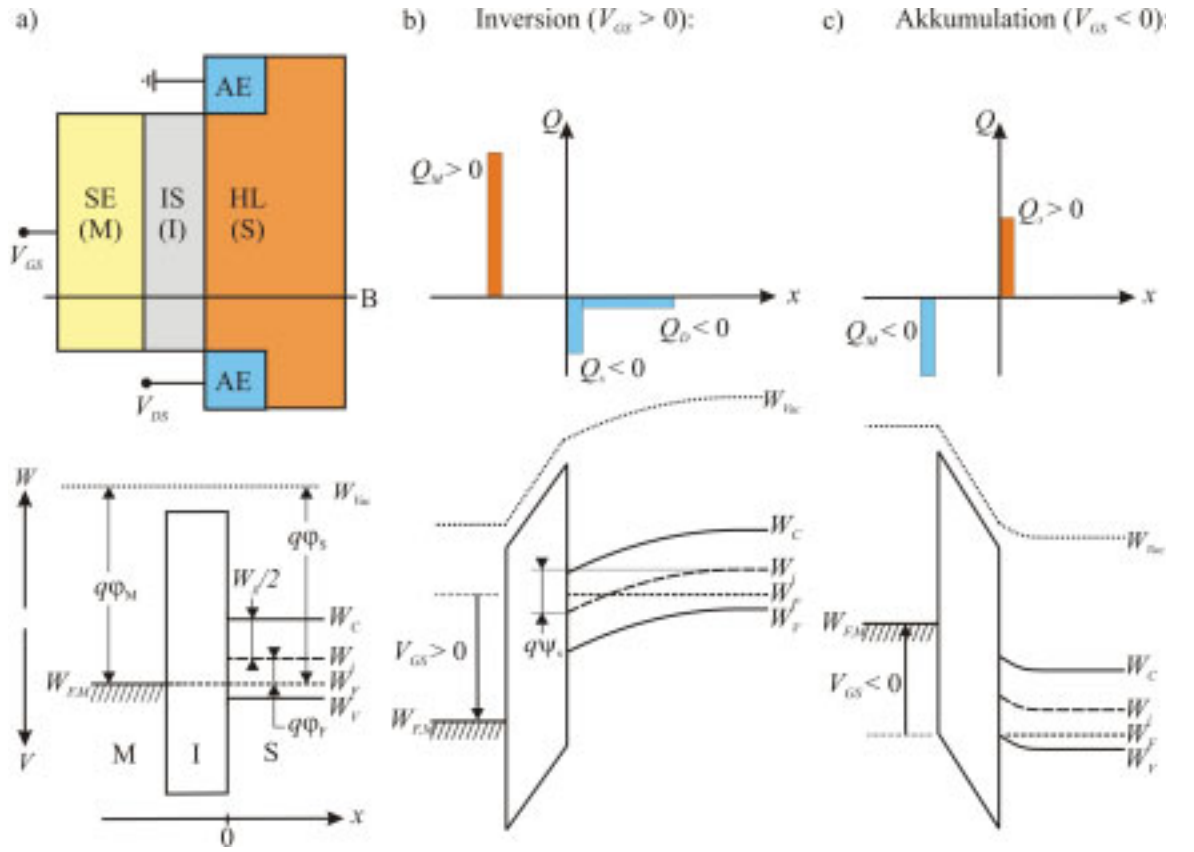

Abb. 2.1: Aufbau und Funktionsweise einer MIS-Struktur; (a) Schema eines MIS-Transistors mit Steuerelektrode (SE), Isolator (IS), p-Typ-Halbleitersubstrat (HL) sowie n-Typ-Anschlusselektroden (AE); darunter: Bandschema einer idealen MIS-Struktur entlang der Schnittgeraden B; (b) Ladungsverteilung in Inversion mit vereinfachtem, abruptem Profil (*depletion approximation*) und schematischer Potenzialverlauf; während dieses Betriebszustandes können negative Ladungsträger zwischen den Anschlusselektroden transportiert werden; (c) Ladungsverteilung in Akkumulation und zugehöriger, schematischer Potenzialverlauf; in diesem Zustand findet kein Ladungstransport zwischen den Anschlusselektroden statt[1].

Die zentrale Kondensatoranordnung in der CMOS-Technologie besteht aus einer Schichtfolge aus Metall, Isolator und Halbleiter (*Metal Insulator Semiconductor*, MIS). Über eine an der metallischen Elektrode angelegte, ausreichend hohe Spannung entsprechender Polung wird durch Ladungsumverteilung im Halbleiter eine Ladung an der Grenzfläche zum Isolator induziert, welche den Stromtransport zwischen zwei seitlich angeordneten Anschlusselektroden durch den Halbleiter ermöglicht. Durch Umpolung oder Ausschalten der Steuerspannung kann der Stromtransport zwischen den Anschlusselektroden unterbunden werden.

[1] Symbole: Elementarladung q, Spannung an der Steuerelektrode V_{GS}, Drainspannung V_{DS}, Ladung der Steuerelektrode Q_M, Oberflächenladung Q_s, Ladung in der Verarmungszone Q_D, Ferminiveau im Metall $W_{F,M}$, Austrittsarbeiten im Metall $q\varphi_M$ bzw. Halbleiter $q\varphi_S$, Halbleiterbandlücke W_g, Leitungsbandunterkante W_C, Valenzbandoberkante W_V, intrinsisches Niveau W_i, Ferminiveau W_F, Fermipotenzial φ_F, Oberflächenpotenzial φ_s, Vakuumniveau W_{Vac}.

In Abb. 2.1a ist schematisch eine ideale MIS-Schichtfolge für eine Transistorstruktur sowie das dazugehörige Banddiagramm entlang der Schnittlinie B dargestellt. Abb. 2.1b und Abb. 2.1c demonstrieren vereinfacht die Ladungsverhältnisse der MIS-Struktur für verschiedene Polungen der Steuerelektrode (Inversion und Akkumulation) sowie zugehörige Potenzialverläufe entlang der Schnittlinie B. Der Fall der Inversion, der erst bei ausreichend großer Spannungsdifferenz zwischen der Steuerelektrode[2] und dem Halbleiter auftritt, entspricht dem eingeschalteten Transistor, welcher den Stromtransport zwischen den Anschlusselektroden[3] ermöglicht.

Bereits aus Gleichung (2.1) wird ersichtlich, dass eine Erhöhung der Flächenladungsdichte bei vorgegebener Versorgungsspannung durch Reduktion der Dicke des Isolators oder alternativ durch Austausch des Isolators durch ein Dielektrikum mit größerer Permittivität möglich ist[4]. Da man mit der Dickenskalierung des konventionell verwendeten Siliziumdioxids (SiO_2) aufgrund von Leckströmen an technische Grenzen stößt, wurde in den letzten 10 Jahren sowohl in der öffentlichen als auch industriellen Forschung intensiv nach Dielektrika mit hoher Permittivität als Ersatz gesucht.

2.1 Grundlegende Eigenschaften

Die physikalische Ursache für die makroskopische Größe der Dielektrizitätszahl liegt an der mikroskopischen Polarisierbarkeit des Mediums. Der Zusammenhang wird durch die Clausius-Mosotti-Beziehung (2.2) beschrieben [Kop89]. Fällt die Summe der Polarisierbarkeiten α_i über alle Polarisationsarten größer aus, resultiert dies auch in einer höheren, relativen Dielektrizitätszahl ε_r.

$$\varepsilon_r = \frac{\varepsilon_0 + \frac{2}{3}\sum_j n_j \alpha_j}{\varepsilon_0 - \frac{1}{3}\sum_j n_j \alpha_j} \geq 1 \tag{2.2}$$

Dabei indiziert j die Polarisationsart und n_j entspricht der jeweiligen Konzentration der mikroskopischen Dipole im Medium. Von Bedeutung sind im Rahmen dieser Arbeit die elektronischen und ionischen Polarisationsanteile.

Der elektronische Anteil der Polarisierbarkeit entsteht durch die im äußeren elektrischen Feld bedingte, räumliche Umverteilung der Elektronen-Gleichgewichtsverteilung bezüglich der positiven Atomkerne. Dies geht einher mit der Erzeugung eines lokalen Dipolmoments.

[2] engl. *gate* (bei MIS-Strukturen)

[3] engl. *source* und *drain*

[4] Eine detailliertere Beschreibung der technologischen Zusammenhänge und Konsequenzen bei der Skalierung findet sich in Kapitel 2.2.

Die räumliche Umverteilung entspricht im Bild der Quantenmechanik einer elektronischen Anregung. Die Anregungshäufigkeit dieser Exzitonen ist dabei indirekt proportional zur Energielücke des Materials, welche dazu mindestens überwunden werden muss. Daher weisen Materialien mit einer geringeren Energielücke eine größere elektronische Polarisierbarkeit und damit eine erhöhte Dielektrizität auf.

Im Bereich niedriger Frequenzen bis zu ca. 1 THz, welche in der CMOS-Anwendung von Interesse sind, spielt jedoch der ionische Anteil der Polarisierbarkeit die dominierende Rolle. Die von einem äußeren Feld erzeugten Dipolmomente durch räumliche Umverteilung von Ionen in einem Kristall fallen größer aus, je schwächer deren Ionenbindung untereinander ist. Eine Schwächung der Ionenbindung im äußeren elektrischen Feld ist bei den Oxiden der Übergangsmetalle zu beobachten, welche im Vergleich zu SiO_2 höhere Dielektrizitätszahlen besitzen. Die Ursache für dieses Phänomen der Bindungsschwächung ist der Pseudo-Jahn-Teller-Effekt [Ber04]. Dieser fällt stärker aus, je geringer die Energielücke in den Metalloxiden ist, und führt zu einer erhöhten Polarisierbarkeit und Permittivität dieser Materialien. Dies erklärt den oft zitierten, empirisch festgestellten Zusammenhang, dass bei High-κ-Materialien die Bandlücke reziprok mit der Dielektrizitätszahl skaliert [Rob02].

Die bei äußeren, elektrischen Wechselfeldern induzierten Dipolmomente schwingen aufgrund ihrer Trägheit nicht in Phase mit dem angelegten Feld. Dadurch wird bei dielektrischen Materialien die für eine ideale Kondensatoranordnung theoretische Phasenverschiebung des Stroms von $\pi/2$ gegenüber dem angelegten Wechselfeld um den Winkel δ reduziert. Es entstehen dielektrische Verluste, die mit dem Dissipationsfaktor $\tan(\delta)$ bewertet werden. In der Halbleitertechnik soll jedoch jegliche Form von Energiedissipation vermieden werden. Zusätzlich stellen die Reaktionszeiten der sich ausrichtenden Dipole unter Umständen eine Begrenzung der Schaltgeschwindigkeiten für die Bauelemente dar. Daher sind bei der Auswahl des Isolatormaterials für die Halbleiterbauelemente auch geringe, dielektrische Relaxationszeiten bzw. geringe dielektrische Verluste von Bedeutung.

2.2 Anwendungsbeispiele in der Halbleitertechnologie

In der Halbleitertechnologie finden Isolatoren mit hoher Dielektrizitätszahl Anwendungen bei verschiedenen Bauelementen. Exemplarisch werden in diesem Kapitel die Vorteile der High-κ-Dielektrika in Metal-Oxid-Halbleiter-Feldeffekttransistoren und nicht flüchtigen Speicherelementen beschrieben (Abschnitte 2.2.1 und 2.2.2). Außerdem wird auf die spezielle, vorteilhafte Auswirkung von dicken High-κ-Gateoxiden beim Tunnel-Feldeffekttransistor eingegangen (2.2.3).

2.2.1 Metall-Oxid-Halbleiter-Feldeffekttransistor (MOSFET)

Aus technischer Sicht führt der Wunsch, über die Produktionszyklen hinweg die Rechenleistung von Prozessoren zu erhöhen, welche dabei einen möglichst geringen Energiekonsum aufweisen, zur Skalierung der zugrundeliegenden Bauelemente. Diese sind insbesondere Metall-Oxid-Halbleiter-Feldeffekttransistoren (*Metal Oxide Semiconductor Field Effect Transistor, MOSFET*). Die Verlustleistung P in der energiesparenden CMOS-Technologie wird durch das Power-Delay-Produkt beschrieben [Hol90]:

$$P = N \cdot f \cdot C_L \cdot V_{DS}^2 \tag{2.3}$$

Hier repräsentiert N die Zahl der aktiven CMOS-Inverter in einem integrierten Schaltkreis, f die Taktrate, C_L die Lastkapazität der Verschaltung mit dem nächsten Transistor und V_{DS} die Versorgungsspannung.

Eine Erhöhung der Rechenleistung – dargestellt durch das Produkt $N \cdot f$ – ist bei festgelegter Verlustleistung P nur durch Reduktion der Lastkapazität und der Versorgungsspannung möglich. Abgesehen von den kapazitiven Verlusten in den Verschaltungsleitungen wird die Lastkapazität von der Gate-Kapazität des angeschlossenen Transistors bestimmt. Diese kann über die geometrische Skalierung eines Transistors verringert werden. Durch die simultane Skalierung von Länge und Breite des Transistors kann der Drainstrom konstant gehalten werden, da für Letzteren im eingeschalteten Zustand des Transistors gilt[5]:

$$I_{DS} = \frac{w}{l} \mu_{eff} C'_{is} \left(\left(V_{GS} - V_T \right) \cdot V_{DS} - \frac{V_{DS}^2}{2} \right) \tag{2.4}$$

Der Drainstrom I_{DS} ist nach dieser Gleichung für den Fall starker Inversion proportional zum Verhältnis der Weite w und Länge l des Transistorkanalgebiets, proportional zur effektiven Beweglichkeit der Ladungsträger μ_{eff}, der flächennormierten Isolatorkapazität C'_{is} und abhängig von der Differenz aus Steuerspannung V_{GS} und Einsatzspannung V_T sowie der angelegten Drain-Source-Spannung V_{DS}.

Die Verkürzung der Kanallänge erfordert jedoch eine Skalierung der Versorgungsspannung V_{DS} bzw. V_{GS}, um bei steigender Stromdichte in den Zuleitungen die Elektromigration zu verhindern. Bei diesem Phänomen führt die hohe Stromdichte der Elektronen durch Stöße an den Metallionen zu einem lokalen Erhitzen des Metalls, welches die Diffusion von Material aus

[5] Die angegebene Formel gilt für einen idealen MIS-Transistor bei Vernachlässigung des Diffusionsstroms und des Leckstroms in Sperrrichtung, unter der Annahme einer konstanten Ladungsträgerbeweglichkeit im Inversionskanal mit homogener Dotierung sowie für den Fall, dass das elektrische Feld senkrecht zum Kanal wesentlich größer ist als das Feld entlang des Kanals (*Gradual Channel Approximation*). Dabei wird zur Vereinfachung zusätzlich angenommen, dass die Drainspannung V_{DS} wesentlich kleiner ist als das doppelte Fermipotenzial φ_F, welches der auf die Elementarladung normierten Energiedifferenz zwischen Ferminiveau aufgrund der Dotierung im Halbleiter und dem intrinsischen Niveau im undotierten Halbleiter entspricht [Sze81].

der Zuleitung begünstigt. Aufgrund der entstehenden, weiteren Verjüngung der Leitung wird der Effekt verstärkt und führt zur Zerstörung der Stromzuführung zur Gate-Elektrode. Das daher notwendige Herabsetzen der Versorgungsspannung V_{DS} bzw. V_{GS} erfordert jedoch entsprechend der Gleichung (2.4) eine geringere Einsatzspannung V_T, welche durch eine Erhöhung der auf die Fläche normierten Gateoxidkapazität C'_{is} erreicht werden kann [Sze81].

$$V_T = V_{FB} + 2\varphi_F + \frac{Q'_D}{C'_{is}}, \text{ wobei } C'_{is} = \frac{\varepsilon_0 \varepsilon_{r,is}}{d_{is}};\ Q'_D = \pm\sqrt{4\varphi_F \varepsilon_{Si} q N_{A/D}} \qquad (2.5)$$

Dabei beschreibt die Fermispannung φ_F den Potenzialunterschied des Ferminiveaus im dotierten Halbleiter gegenüber dem Ferminiveau im intrinsischen Halbleiter. Die Flachbandspannung V_{FB} einer idealen Metall-Isolator-Halbleiter-Anordnung entspricht der Differenz der auf die Elementarladung normierten Austrittsarbeiten von Metallelektrode φ_M und Silizium-Halbleiter φ_S. Die Ladung Q'_D in der Verarmungszone des Halbleiters ist eine Funktion der Fermispannung, der Permittivität ε_{Si} des Siliziumsubstrat, der Elementarladung q sowie der Akzeptoren- bzw. Donatorenkonzentration $N_{A/D.}$

Die notwendige Vergrößerung der flächennormierten Isolatorkapazität C'_{is} hat entsprechend (2.4) den weiteren Vorteil eines erhöhten Drain-Source-Stroms. Die Skalierung der physikalischen Parameter nach diesem Verfahren wird als klassische Skalierung bezeichnet und führt zum Erhalt des elektrischen Felds, welches das Gateoxid in den Kanal vermittelt. Bis vor Kurzem wurde die Erhöhung von C'_{is} durch Reduktion der Siliziumoxiddicke realisiert.

Bei Oxiddicken unter ca. 1,5 nm führen jedoch die aufgrund des Elektronentunnelns durch das Gateoxid hervorgerufenen Leckströme zu inakzeptabel hohen Leistungsverlusten in den Transistoren. Daher ist die öffentliche und industrielle Forschung bemüht, diese Skalierungsgrenze der flächennormierten Gateoxidkapazität mit dem Austausch des Oxidmaterials SiO_2 durch Metalloxide höherer Permittivität zu überwinden. Um die dielektrische Wirkungsweise von High-κ-Materialien im Vergleich zum konventionellen SiO_2 zu beschreiben, wird die Größe der äquivalenten Oxiddicke (*Equivalent Oxide Thickness, EOT*) eingeführt. Diese berechnet sich aus den Permittivitäten des Siliziumdioxids (SiO_2) und des High-κ-Ersatzmaterials (HK) sowie dessen physikalischer Dicke d_{HK}.

$$EOT = \frac{\varepsilon_{SiO_2}}{\varepsilon_{HK}} d_{HK} \qquad (2.6)$$

Die Reduktion von *EOT* führt zu der erwünschten Erhöhung der flächennormierten Gateoxidkapazität. Die Größe der äquivalenten Oxiddicke ist für zukünftige Skalierungsschritte in Form einer Meilensteinplanung durch die *International Technology Roadmap for Semiconductors* (ITRS) vorgegeben [PIDS07].

2.2.2 Nicht flüchtige Speicherelemente (NVRAM)

Die Technologie nicht flüchtiger Speicher mit wahlfreiem Zugriff (*Non Volatile Random Access Memory,* NVRAM) ermöglicht die elektronische Speicherung von Informationen ohne ständige Aufrechterhaltung der Energieversorgung. Neben anderen NVRAM-Konzepten ist die Flash-Speichertechnologie heutzutage von größter Relevanz. Darunter wird insbesondere der CTF-Technologie (*Charge Trap Flash*) Bedeutung beigemessen, da sie gegenüber der FGFET-Technologie (*Floating Gate Field Effect Transistor*) den Vorteil höherer Integrationsdichten aufweist. Die Funktionsweise entspricht dabei einem Feldeffekttransistor, dessen Einsatzspannung (2.5) über eine Veränderung der Flachbandspannung V_{FB} durch bewusste Manipulation der Isolator-Flächenladungsdichte Q'_{is} variiert wird.

$$V_{FB} = \varphi_M - \varphi_S - \frac{Q'_{is}}{C'_{is}} \tag{2.7}$$

Zwei Beladungszustände mit unterschiedlichen Oxid-Flächenladungsdichten drücken sich somit durch zwei abweichende Einsatzspannungen aus. Legt man zum Auslesen des Speicherelements eine Gatespannung an, die zwischen den beiden Einsatzspannungen liegt, kann damit der Beladungszustand im Gateoxid identifiziert werden.

Realisiert werden die Flash-Speicherelemente mit Hilfe eines Schichtstapels verschiedener Isolatoren als Gateoxidersatz, sodass die Speicherfunktion in der Steuerkapazität des Auswahltransistors integriert ist (vgl. Abb. 2.2a). Bei diesen MONOS-Strukturen (*Metal Oxide Nitride Oxide Semiconductor*) erfüllt das Nitrid die Funktion des Ladungsspeichers. Wird am Gate eine große Spannung angelegt, können Ladungsträger aus der Inversionsschicht im Halbleiter über einen Fowler-Nordheim-Tunnelprozess (FNT-Prozess) durch das Tunneloxid (TS) in die Speicherschicht (SS) injiziert werden (vgl. Abb. 2.2b). In dem Speichernitrid verbleiben die Ladungen nicht im Leitungsband, sondern relaxieren in energetisch niedrigere, freie und lokalisierte Zustände (Haftstellen, engl. *traps*). Dadurch wird ihre Beweglichkeit innerhalb der Speicherschicht erheblich reduziert. Dies ist der entscheidende Vorteil gegenüber den elektrisch leitenden Speicherschichten der FGFET-Technologie: Durch die Fixierung der Ladungsträger wird der Einfluss von benachbarten Speicherzellen deutlich reduziert.

Beim Löschvorgang wird die Gatespannung umgepolt. Die Speicherschicht kann über einen FNT-Prozess in das Substrat entladen werden. Die Barriereschicht (BS) hat während dieses Schritts die Aufgabe, Ladungsträger aus der Steuerelektrode zurückzuhalten. Abb. 2.2c zeigt ein berechnetes Banddiagramm[6], welches den Potenzialverlauf einer MONOS-Struktur während des Löschvorgangs darstellt, deren Schichtdicken, Materialien, sowie Gatespannungen den ITRS-Vorgaben für die 32 nm NAND-Flash Technologie entsprechen [PIDS07]. Die Dicke der Barriereschicht aus SiO_2 beträgt dabei 7 nm.

[6] Zur Berechnungsmethodik vgl. Anhang A.2. Die Substratdotierung wurde als Kompromiss zur Erzielung eines hohen I_{on}/I_{off}-Verhältnisses einerseits und technologisch begrenzter Zuverlässigkeit des Bauelements andererseits mit $N_A=1\cdot10^{18}$ cm^{-3} angesetzt [Kwa07].