5 Rechnerarchitektur

Computer bestehen aus ein oder mehreren Zentraleinheiten (engl. *Central Processing Units*, kurz *CPUs*), einem Arbeitsspeicher (engl. *Random Access Memory*, kurz *RAM*) und Peripheriegeräten. Alle diese Teile sind hochkomplexe Schaltkreise, die hauptsächlich aus *Transistoren* aufgebaut sind. Transistoren werden hier als elektrisch gesteuerte Ein-Aus-Schalter eingesetzt. Durch geschickte Kombination vieler solcher Schalter entstehen Schaltkreise, die jedes gewünschte Verhalten realisieren können. Die *boolesche Algebra*, die wir in der ersten Hälfte dieses Kapitels kennen lernen, erlaubt es uns, zu einer beliebigen Schaltaufgabe einen entsprechenden Schaltkreis auszurechnen. Damit ausgerüstet zeigen wir, wie die wichtigsten Bauelemente eines Rechners, nämlich ALU (engl. *Arithmetic Logic Unit*, kurz *ALU*) und Speicher, aus einfacheren Schaltkreisen aufgebaut werden können. Aus diesen konstruieren wir danach eine mikroprogrammierte CPU und vollziehen damit den Übergang von der Hard- zur Software. Wir verfolgen diesen bis zum Maschinencode und Assembler und diskutieren anschließend noch RISC (engl. *Reduced Instruction Set Computer)* als alternative CPU-Architektur.

Dieses Kapitel erläutert also prinzipiell, wie durch geschickte Kombination von Transistoren ein komplexes Gerät wie ein PC entsteht. Wenn man wollte, könnte man Transistoren auch durch optische Schalter ersetzen und mit den gleichen Prinzipien einen optischen Computer konstruieren. Durch die schnelleren Umschaltzeiten optischer Bauteile darf man sich einen erheblichen Geschwindigkeitsgewinn erhoffen. Allerdings sind optische Schalter heute noch nicht so einfach zu realisieren wie Transistoren. Insbesondere ist eine technische Lösung für die Zusammenfassung (Integration) von Tausenden oder gar Millionen optischer Bauelemente auf einem Chip noch in weiter Ferne. Heute ist die CMOS-Technik in der Realisierung von Transistorschaltungen führend. Wir werden lernen, wie sich in dieser Technik besonders leistungsfähige Bauelemente entwerfen und realisieren lassen. Auch einige Aspekte der Herstellung elektronischer Chips wollen wir in diesem Kapitel beleuchten.

5.1 Vom Transistor zum Chip

Das für uns wichtigste elektronische Bauelement ist der so genannte *MOS-Transistor*. MOS ist die Abkürzung für den englischen Begriff *metal oxide semiconductor* (Metalloxid-Halbleiter). Es gibt verschiedene Arten von MOS-Transistoren, alle sind, wie auch in der folgenden Abbildung zu sehen, aus mehreren Materialschichten aufgebaut. Ausgangspunkt ist kristallines Silizium, das durch Einbringung von Fremdatomen *dotiert* (verunreinigt) ist. Man unterscheidet zwischen n-dotiertem und p-dotiertem Silizium. Im ersten Fall entsteht durch die

Fremdatome ein Elektronenüberschuss und damit freie negative Ladungsträger, im Falle von p-dotiertem Silizium ein Mangel an Elektronen, was man als freie positive Ladungsträger interpretieren kann.

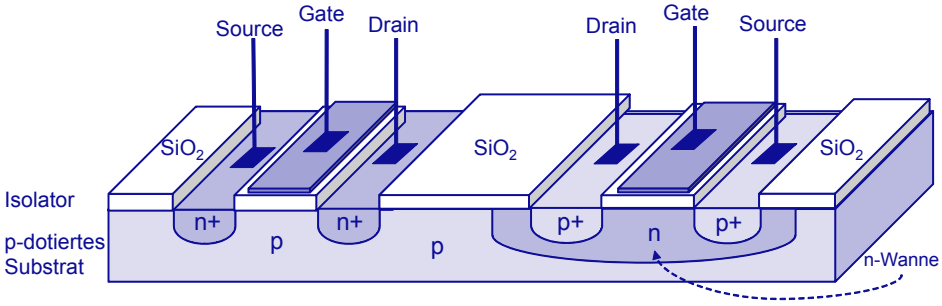

Abb. 5.1: n-MOS-Transistor und p-MOS Transistor auf gemeinsamem p-Substrat

Zwischen den p- und n-dotierten Bereichen bilden sich Grenzschichten aus, in denen der Elektronenüberschuss der n-Schicht den Elektronenmangel der angrenzenden p-Schicht ausgleicht. Dadurch entsteht eine neutrale Zone, in der keine freien Ladungsträger vorhanden sind, so dass diese als Isolator wirkt. Eine zwischen den mit *Source* und *Drain* bezeichneten stark dotierten Bereichen (n+, bzw. p+) angelegte Spannung bewirkt daher keinen Stromfluss.

Hier kommt der in der obigen Figur als *Gate* bezeichnete metallische Kontakt ins Spiel, der durch einen Isolator (SiO) von der schwach p-dotierten Schicht (p) getrennt ist. Legt man eine positive Spannung zwischen Gate und Source an, so lädt sich das Gate positiv auf. Wird dabei ein gewisser Schwellwert überschritten, so wird in dem p-dotierten Bereich unter dem Gate ein elektrisch leitender Kanal aus negativen Ladungsträgern induziert. Eine ausreichende Spannung am Gate schaltet somit eine elektrische Verbindung zwischen Source und Drain; fällt diese Spannung unter einen Schwellwert, so wird die Verbindung wieder unterbrochen.

Wenn man in der obigen Erklärung die n-dotierten und p-dotierten Bereiche austauscht, erhält man einen p-MOS Transistor. Eine negative Spannung am Gate induziert im n-Substrat einen Kanal positiver Ladungsträger. Da sowohl n-MOS als auch p-MOS Transistoren auf dem gleichen Substrat aufgebracht werden müssen, fertigt man zunächst eine in das p-Substrat eingelassene n-Wanne, in der man dann den p-MOS Transistor aufbaut. Im Schaltbild wird der p-MOS Transistor durch einen kleinen Kreis am Gate kenntlich gemacht. Den Grund dafür werden wir in Abschnitt 5.4.1 erfahren.

Abb. 5.2: Schaltbilder für n-MOS und p-MOS Transistoren

Für uns ist einstweilen nur diese Schalterwirkung der Transistoren von Interesse. Dies soll auch in den symbolischen Schaltbildern zum Ausdruck kommen. Steigt die Spannung zwischen Gate und Source über einen Schwellwert, dann schaltet Source zu Drain durch. Ein Abfallen der Spannung unterbricht diese Verbindung.

5.1.1 Chips

Ein *Chip* ist ein dünnes Silizium-Scheibchen, auf das die Transistorschaltung beim Herstellungsprozess aufgebracht wird. Da auf einer daumennagelgroßen Fläche eine sehr große Anzahl von Schaltgliedern zu einem Schaltkreis zusammengefasst werden, nennt man das entstandene Bauteil auch *Integrated Circuit* (*IC*). Mit den Jahren wuchs die Anzahl der Bauelemente auf einem einzigen Chip um mehrere Größenordnungen, entsprechend wandelte sich auch der Name über *LSI (large scale IC)* zu *VLSI (very large scale IC)*. Heutige CPU-Chips enthalten einige Milliarden Transistoren auf einer Fläche von weniger als 100 mm^2. Speicherchips können aufgrund ihrer regelmäßigeren Struktur noch höher integriert werden.

Die Dicke eines Chips beträgt nur etwa 1/10 mm, die der *aktiven Schicht* ist noch erheblich geringer. In der aktiven Schicht finden sich die Transistoren, Dioden, Widerstände und die Leitungen. Der Chip ist in ein Gehäuse aus Kunststoff oder Keramik eingebettet, das erheblich größer ist als das Silizium-Scheibchen. Die Verbindungen von dem inneren Silizium-Scheibchen zu den Außenkontakten des Chip-Gehäuses werden mithilfe hauchdünner Golddrähtchen hergestellt. Klassische Chips sind in einem rechteckigen Gehäuse mit zwei Reihen seitlich angebrachter Anschlussdrähte, den *Beinchen* (engl. *pin*), untergebracht. Die Anzahl der Außenverbindungen ist bei solchen Chips auf etwa 64 beschränkt. Chips mit mehr Anschlüssen (bis zu etwa 100) setzt man oft in ein quadratisches Gehäuse mit Anschlussdrähten an allen vier Seiten. Noch mehr Außenverbindungen schafft man durch Anbringung der Beinchen unter dem Chip. Durch diese *Pin Grid Array* (PGA) genannte Technik lassen sich Chips bauen, die mehrere hundert Verbindungen aufweisen können. Diese Technik wurde weiterentwickelt; heute üblich sind *Land Grid Arrays* (LGA). Die Anschlüsse sind auf einem *Sockel* angeordnet. Dieser hat federnde Kontaktstifte, das Prozessorgehäuse nur mehr Kontaktflächen, sogenannte Lands. Der erste Intel Pentium Prozessor hatte ein PGA mit 273 Pins, die ersten Versionen des Pentium-4 kamen auf 423 Pins. Der neueste Prozessor aus der Intel x86 Serie, der Core i7 3770, hat ein LGA mit 1155 Pins.

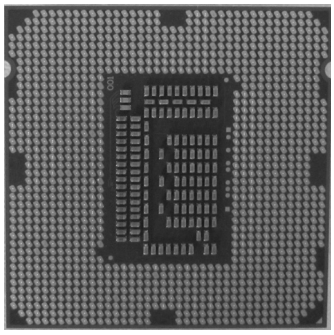

Abb. 5.3: Ansicht eines LGA. Beispiel Unterseite des Intel i7 3770

Eine Leiterplatte ist in der Regel mit Chips unterschiedlicher Bauart bestückt und enthält zusätzlich einzelne klassische Bauelemente wie Kondensatoren oder Widerstände. Die Verdrahtung erfolgt meist in mehreren, mindestens jedoch zwei Verdrahtungsebenen. Diese stehen auf der Leiterplatte zur Verfügung, sind untereinander isoliert und haben Querverbindungen zu den anderen Ebenen. Werden mehrere Leiterplatten benötigt, sind diese meist senkrecht in eine Systemplatine (engl. *motherboard*) eingesteckt, die die Verbindungen enthält. Mit Anschlussbuchsen für genormte, mehrpolige Stecker kann ein Anschluss zu Netzteilen, externen Geräten etc. erfolgen.

In heutigen Computern finden sich meist eine oder mehrere Leiterplatten mit weniger als 50 Chips. In unmittelbarer Zukunft wird man durch höhere Integration die Anzahl der Chips in einem Rechner auf weniger als zehn reduzieren und zur selben Zeit die Leistung der Geräte um mehrere Größenordnungen steigern können.

5.1.2 Chipherstellung

Für die Herstellung eines Chips wird zunächst gereinigtes Silizium (Quarzsand) auf über tausend Grad erhitzt, bis es flüssig wird. Aus dieser Schmelze werden so genannte Einkristalle gezogen, die bis zu 2 m lang sein können und einen Durchmesser von etwa 20 bis 30 cm haben. Sie werden nach dem Erkalten in dünne Scheiben gesägt und poliert. Diese Scheiben sind das Ausgangsmaterial für den Herstellungsprozess, im Laufe dessen auf jeder einzelnen hunderte von Chips in einem Arbeitsgang entstehen.

Komplexe Chips erfordern mehrere hundert Herstellungsschritte. Sie können viele Millionen individueller Transistoren enthalten. Für jeden Schritt kommt, in jeweils abgewandelter Form, ein fotolithografisches Grundverfahren zur Anwendung. Dabei wird jedesmal zunächst eine Materialschicht aufgetragen und mit Fotolack überzogen. Dieser wird mithilfe einer Maske, auf der die Chipstrukturen ausgespart sind, belichtet. Nach der Entwicklung werden die unbelichteten Stellen bearbeitet, das heißt entweder weggeätzt, dotiert oder mit Kontakten versehen. Dann wird der restliche Fotolack entfernt.

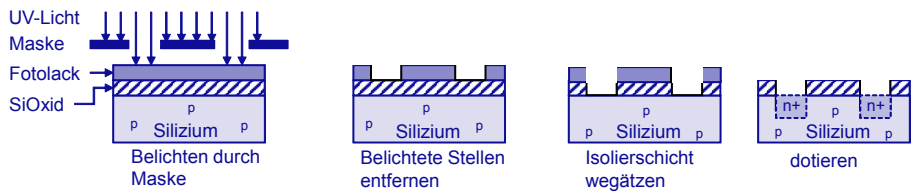

Abb. 5.4: Photolithographische Chipbearbeitung

Nach dem Aufbringen der Transistoren und Leiterbahnen entsteht auf den rechteckigen Siliziumscheiben, in einem durch die verwendeten Masken definierten Gebiet, ein waffelartiges Muster einzelner Chips. Daher werden die Siliziumscheiben auch *Wafer* genannt. Sie werden zersägt, in die Gehäuse eingebaut und mit den Anschlussdrähten verbunden (engl.: *bonding*). Das Gehäuse wird endgültig verschlossen – und fertig ist der Chip.

Gegenwärtig ist Silizium der Rohstoff der Wahl für die Fertigung von Chips. Es ist billig und einfacher zu bearbeiten, als der alternative Rohstoff Galliumarsenid, aus dem man Chips mit erheblich kürzeren Schaltzeiten fertigen kann, die in Supercomputern und anderen kritischen Anwendungen gelegentlich eingesetzt werden.

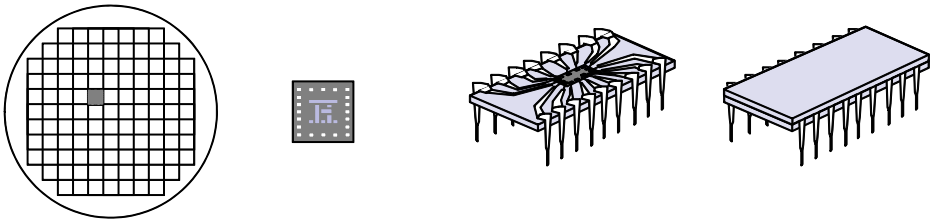

Abb. 5.5: Wafer mit Chips, ausgesägter Chip, bonding und gekapselter Chip

5.1.3 Kleinste Chip-Strukturen

Ein wesentlicher Parameter bei der Chip-Herstellung ist die Größe der kleinsten Strukturen. Dabei handelt es sich um Leitungen, den Abstand zwischen zwei Leitungen oder um die Größe von Transistorzellen. Die kleinsten erzeugbaren Strukturen lagen lange Zeit im Bereich von 100 nm bis 1µ (1µ = 1 Mikrometer = 10^{-6} m, 1nm = 1 Nanometer = 10^{-9}m). Zurzeit werden Strukturen von 22 nm bis 100 nm verwendet. Ein weiteres Absenken der kleinsten Strukturen auf 16 nm, 14 nm usw. ist für die nächsten Jahre geplant. Die Schwierigkeiten beim Verkleinern der Chip-Strukturen bestehen im Herstellen geeigneter Masken für die verschiedenen fotolithografischen Prozesse, in der exakten Positionierung der Masken, in Belichtungsproblemen, wenn die Wellenlänge des für die Belichtung verwendeten Lichts erreicht wird, und in mikroskopischen Ungenauigkeiten beim Ätzen, Beschichten etc.

Diese Schwierigkeiten konnten bisher immer wieder bewältigt werden. Meist waren dafür jedoch langwierige Forschungs- und Entwicklungsarbeiten erforderlich, so dass die kleinsten beherrschbaren Strukturen nur relativ langsam von 2 µ auf 1 µ und dann schrittweise auf 22 nm verkleinert werden konnten. Die Verkleinerung auf Werte in der Größenordnung von 5 bis 15 nm wird weitere technische Innovationen erfordern. Als konsequente Fortsetzung der optischen Lithografie hin zu kürzeren Wellenlängen gilt z.B. die *EUV*-Lithografie (*Extreme Ultra Violet*). Dabei werden Wellenlängen im Bereich 13,5 nm genutzt, um Strukturen zwischen 45 nm und 22 nm und kleiner zu erzeugen.

5.1.4 Chipfläche und Anzahl der Transistoren

Die Herstellung von Chips ist ein langwieriger und fehleranfälliger Prozess. Der Anteil von funktionsfähigen Chips betrug daher vor einigen Jahren nur etwa 5 bis 50%, bezogen auf die Gesamtproduktion, je nach der bereits gewonnenen Produktionserfahrung mit einem bestimmten Herstellungsprozess. Mittlerweile werden die Fertigungsprozesse besser beherrscht; es wird eine funktionsfähige Ausbeute von bis zu 80% erreicht. Die Fehlerrate bei den einzelnen Chips ist von der Fläche des produzierten Chips abhängig. Um die Produk-

tion wirtschaftlich zu machen, versucht man, die Chipfläche auf ein vertretbares Minimum zu reduzieren. Nur wenn es nicht anders geht, erhöht man die Chipfläche, um die Anzahl der Transistoren zu erhöhen. Gegenwärtig ändert sich die effektiv ausgenutzte Chipfläche von ca. 100 bis 250 mm^2 nur wenig, da die Herstellungsprozesse so häufig verbessert werden, dass eine Vergrößerung der Chipfläche kaum notwendig ist.

Der Prozessor des Core i7-3370 wird seit April 2012 gefertigt und verfügt über 1,4 Milliarden Transistorfunktionen auf einer Fläche von nur 160 mm^2, gefertigt wird er mit einem 22 nm Prozess. Ein Vorgängermodell, der Core i7-3820 wurde mit einem 32 nm Prozess hergestellt und besitzt 1,27 Milliarden Transistorfunktionen auf einer Fläche von 294 mm^2. Beide Prozessoren haben jeweils vier CPU-Kerne und etwa 10 MB Cache. Das neuere Modell hat zusätzlich einen integrierten Grafikprozessor (GPU). In beiden Modellen benötigen die Prozessorkerne vermutlich jeweils etwa 50 Millionen Transistorfunktionen. Bei zukünftigen Generationen wird die Anzahl der Transistorfunktionen vermutlich weiter steigen – und für eine größere Zahl von Prozessorkernen bzw. für noch mehr Cache-Speicher genutzt werden.

5.1.5 Weitere Chip-Parameter

Je geringer die kleinsten Strukturen auf einem Chip sind, desto geringer sind die Schaltverzögerungen pro Transistor und der Energieverbrauch pro Schaltvorgang. Wenn dieser Energieverbrauch, der gegenwärtig ca. 1 pJ (Picojoule) beträgt, nicht um eine ganze Größenordnung gesenkt werden könnte, wäre eine Erhöhung der Transistorzahl gar nicht möglich – die Chips würden zu heiß werden.

Die Schaltverzögerung von modernen MOS-Transistoren beträgt weniger als 0,1 ns (NanoSekunden). Die Schnelligkeit einer ganzen Leiterplatte wird nicht nur durch die Geschwindigkeit der Transistoren in den Chips bestimmt, sondern auch durch die Zahl und die Länge der Verbindungen der verschiedenen Chips untereinander. Je mehr Transistoren in einem Chip untergebracht werden können, desto weniger Inter-Chip-Verbindungen sind erforderlich – um so schneller ist die Leiterplatte.

5.1.6 Speicherbausteine

Auch der Speicher eines Rechners ist aus Chips aufgebaut, den so genannten RAM-Chips. *RAM* ist die Abkürzung für den englischen Begriff *Random Access Memory* – zu deutsch: Speicher mit wahlfreiem Zugriff. Verwendet man die Ladung auf dem Gate eines Transistors zur Speicherung eines Bit, kommt man, zusammen mit der Adressierlogik, auf Speicherbausteine mit weniger als 1,5 Transistoren pro Bit. Allerdings verlieren diese *dynamischen* Speicherbausteine (*DRAM*) nach kurzer Zeit ihre Ladung wieder. Jedes Bit muss innerhalb einer bestimmten Zeit, die im Nanosekundenbereich liegt, wieder aufgefrischt, also gelesen und neu geschrieben werden. Eine Alternative ist die Verwendung *statischer* Speicherbausteine (*SRAM*). Diese müssen zwar nicht ständig aufgefrischt werden, benötigen aber mehrere Transistoren pro Bit. Sowohl dynamische als auch statische RAM-Chips verlieren die gespeicherte Information, wenn kein Strom vorhanden ist. Dies kann durch bestimmte, aufwändige Schaltungen oder durch Verwendung von Akku-Puffern verhindert werden. Heute werden dynamische RAM-Chips mit 1, 2 und 4 GBit Speicherkapazität gefertigt – in absehbarer Zeit wird es voraussichtlich auch 8 und 16