

2.3 Prozessor-Design und Hardware-Implementierung

Für den elektronischen Entwurf eines Mikroprozessors wird zunächst das Design mit Hilfe spezieller Software (z.B. Schematic Entry, VHDL Entry) auf einem geeigneten Rechner erarbeitet. Die Verifikation der Schematics erfolgt in der anschließenden Simulation. Hier werden die Eingänge der Funktionseinheiten mit spezifischen Testmustern belegt und die Ausgangssignale ausgewertet. Bei der Simulation wird zwischen 'Functional'- und 'Timing'-Simulation unterschieden. Die 'Functional'-Simulation überprüft, ob die Schematics logisch richtig arbeiten, wobei für alle Verbindungen die gleiche Verzögerungszeit eingetragen wird. Die Verzögerungen der Gatter und FlipFlops werden experimentiell gemessen und als bekannt vorausgesetzt. Nach der Partitionierung des Gesamt-Designs, der Platzierung der Schaltkreise bzw. der Funktionseinheiten wird die Verdrahtung durchgeführt. Diese 3 Schritte nehmen moderne Entwurfssoftware-Systeme automatisch vor. Nach der Verdrahtung wird in einem Backannotation-Schritt die aktuelle Verzögerung jeder Verbindung in das Schematic eingetragen. Jetzt erst kann die 'Timing'-Simulation des Designs durchgeführt werden, weil je nach Beschaffenheit einer Verbindungsleitung die entsprechend vorher berechnete Verzögerung in die Simulation eingeht.

Nach der erfolgreichen Simulation des Prozessor-Designs erfolgt dessen Implementierung in Hardware. Mit Hilfe der Integrationstechnik werden heute Schaltkreise, Funktionseinheiten (Verknüpfungsglieder, Schaltnetze, Schaltwerke) hergestellt. In monolithisch integrierten Schaltungen entstehen alle benötigten Bauelemente aus einem Einkristall (z.B. Silizium), der gleichzeitig als gemeinsamer Träger (Substrat) dient. Die Bauelemente (Transistoren, Dioden, Widerstände) werden durch lokal begrenzte Zonen mit unterschiedlicher Dotierung erzeugt. Mittels Masken werden diejenigen Flächen abgedeckt, an denen keine Dotierung des Siliziums erfolgen soll. An den nicht maskierten Stellen diffundieren die Dotieratome in das Substrat, während alle anderen Bereiche vor dem Eindringen der Dotieratome geschützt sind. Die Masken werden aus Siliziumdioxid hergestellt, das für Dotieratome undurchdringlich ist. Die Kosten derartiger Maskensätze für moderne Mikroprozessoren liegen bei ca. 50.000 €.

Als wichtigster Baustein für alle Schaltkreise dient heute bei monolithischen Verfahren der Isolierschicht-Feld-Effekt-Transistor (MOS-FET). Abbildung 2.4 zeigt einen Schnitt durch einen MOS-FET. Bei diesem Typ ist das Gate durch eine dünne Oxidschicht vom Halbleitermaterial isoliert. Je nachdem, ob der Halbleiter, dessen Leitfähigkeit gesteuert wird, ein n- oder ein p-Halbleiter ist, spricht man von einem p-Kanal bzw. n-Kanal MOS-FET.

2 Technologische Grundlagen

In der Abbildung 2.4 handelt es sich um einen n-Kanal-MOS-FET, d.h. bei positiver Gate-Spannung relativ zum Substrat (threshold-Spannung) entsteht ein leitender Kanal aus freien Elektronen zwischen den beiden n-Zonen, d.h. der Transistor leitet. Im Fall negativer Gate-Spannung relativ zum Substrat sperrt der Transistor.

Der Grund dafür, warum sich MOS-Schaltkreise für integrierte Schaltungen durchgesetzt haben, besteht darin, dass sich bei Feld-Effekt-Transistoren die Isolierung im Herstellungsprozess (Lithographieverfahren) durch die Realisierung der Funktionseinheiten von selbst ergibt (Selbstisolation).

Die Entwicklung von modernen, leistungsfähigen Mikroprozessoren ist ein sehr kostenintensiver Prozess und setzt ein erfahrenes und gut geführtes Team von Entwurfsingenieuren voraus. Außer im Entwicklungslabor der Firma IBM Böblingen werden in Deutschland von der Firma Hyperstone Electronics Mikroprozessoren entwickelt.

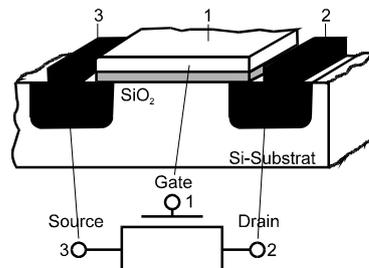


Abbildung 2.4 Schnitt durch einen MOSFET-Transistor

Für ein derartiges Entwicklungs-Projekt sind zunächst von einer Arbeitsgruppe Schaltkreise zu entwerfen. Diese werden zu Schaltkreisbibliotheken zusammengefaßt.

Die Abbildung 2.5 zeigt einen Invertierschaltkreis und einen invertierenden 2 Eingangs-UND- sowie einen invertierenden 2 Eingangs-ODER-Schalter.

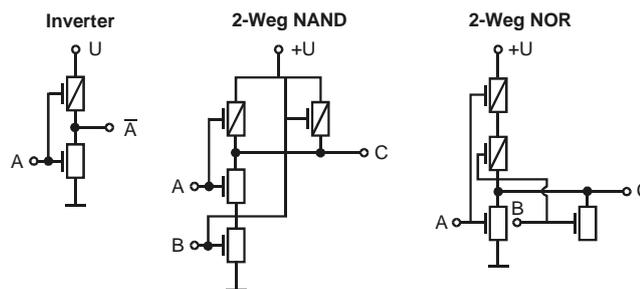


Abbildung 2.5 Schaltbilder einfacher Funktionen

Im Normalfall werden Bibliotheken von bis zu 50 verschiedenen Schaltkreistypen verwendet. Neben den schon erwähnten Schaltern beinhalten diese Bibliotheken FlipFlops der unterschiedlichsten Formen, komplizierte Latches, Treiberschaltkreise

usw. Eine weitere Arbeitsgruppe verwendet diese Bibliotheken, um die entsprechenden Funktionseinheiten (Multiplexer, Schaltwerke) zu entwerfen, auf dem Chip zu plazieren (Placement) sowie zu verdrahten (Routing). Nach Abschluss des Entwurfes kann der Chip produziert werden.

Der Querschnitt durch einen Prozessor-Chip ist in Abbildung 2.6 dargestellt. In der untersten Schicht befinden sich die Transistoren, die über die schraffiert gekennzeichneten Anschlüsse verbunden werden müssen. Die Verdrahtung erfolgt typischerweise in mehreren Metallisierungsebenen, wobei normalerweise mindestens 2 Ebenen verwendet werden. Mittlerweile wird die Verdrahtung auf deutlich mehr Metallisierungsebenen vorgenommen, was die Produktion dieser Chips sehr viel aufwendiger und damit teurer macht. Es ist auch durchaus normal, wenn in einem Chip von beispielsweise 700 Millionen Transistoren nur etwa 70% benutzt werden. Die restlichen 210 Millionen Transistoren sind physisch auf dem Chip vorhanden, sie sind aber nicht angeschlossen und haben folglich auch keine Funktion.

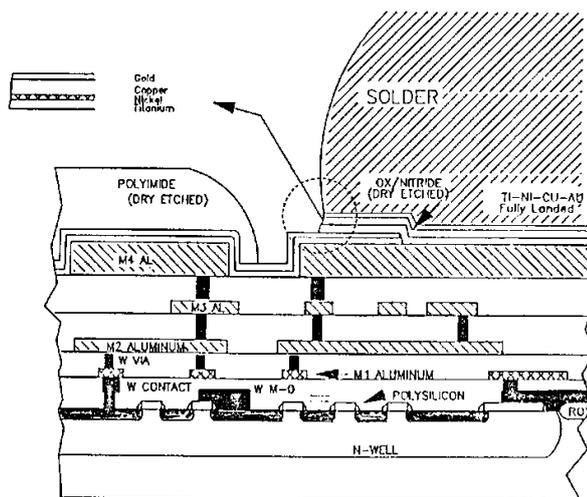


Abbildung 2.6 Schichtaufbau eines Prozessors

Bei der Betrachtung eines Chips von oben sieht man nicht etwa die Transistoren sondern nur das Verdrahtungsmuster (Abbildung 2.7).

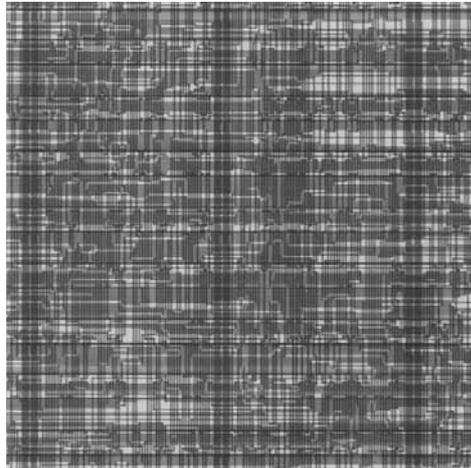


Abbildung 2.7 Verdrahtungsmuster eines Prozessors

Die Abbildung 2.8 stellt den Ausschnitt eines Verdrahtungsmusters, der mit Hilfe eines Plotters erzeugt wurde, dar. Die dickeren Linien sind Stromverteilungsschienen, die die Transistoren mit Spannung versorgen. An den Kreuzungspunkten, an denen sich in der Perspektive scheinbar die Verbindungsleitungen schneiden, wird mit einer der beiden Leitungen in eine andere Ebene aufgestiegen. In der Abbildung 2.9 ist ein kleiner Chip-Teilbereich mit 2 Metallisierungsebenen gezeigt. Als Größenvergleich zu einer Leiterbahn (Abbildung 2.9) dient die Stärke eines menschlichen Haares.

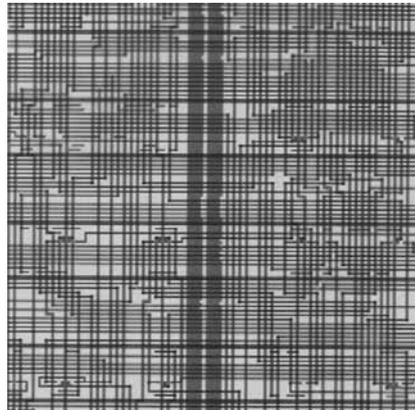


Abbildung 2.8 Vergrößerter Ausschnitt einer Prozessor-Verdrahtung

Bei der Verdrahtung von 10.000.000 Transistoren müssen ca. 100.000.000 Leitungszüge auf dem Chip untergebracht werden. Bei Abmessungen von etwa 2 cm Kantenlänge des Chips bilden diese Leitungen eine Entfernung von rund 100 m. Eine der

2.3 Prozessor-Design und Hardware-Implementierung

Hauptschwierigkeiten bei der Produktion bildet die saubere Ausbildung aller 100.000.000 Leiterbahnen. Bei einem einzigen Entwicklungsfehler ist der gesamte Chip nicht funktionsfähig. Eine Nachbesserung bei einem Fehler ist aus dem Grund ausgeschlossen, weil Eingriffe in den fertigen Chip unmöglich sind.

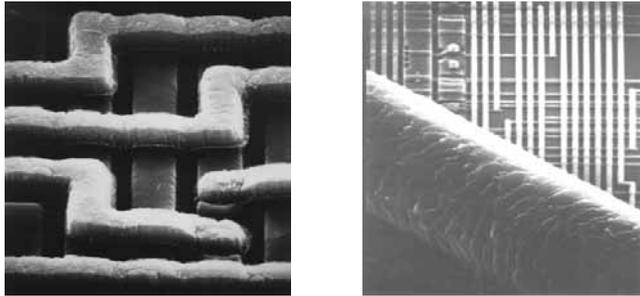


Abbildung 2.9 links: Mikroskopische Aufnahme der Leiterbahnen, rechts: Menschliches Haar als Vergleich zu den Leiterbahnen des Chips

Oft werden bei der Neuentwicklung eines Mikroprozessors gleichfalls unerprobte Herstellungsprozesse eingesetzt. Letztere haben die Eigenschaft, Chips zu produzieren, von denen 99 % defekt sind, da diese Prozesse in der Serienfertigung noch nicht beherrscht werden. Die Ausbeute wird im Nachhinein auf etwa 50 % angehoben. Bevor man jedoch eine weitere Effektivitätssteigerung erreicht, wird bereits auf die nächste Fertigungstechnologie umgestellt.

In der Abbildung 2.10 ist die vollständige Verdrahtung eines Chips gezeigt.

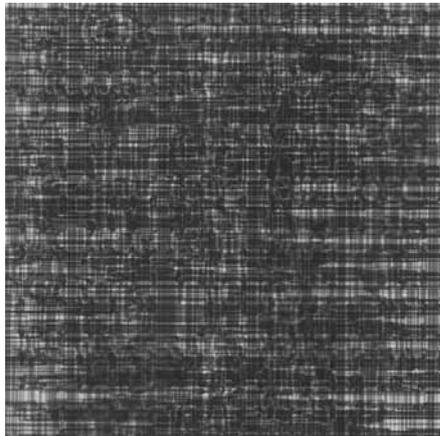


Abbildung 2.10 Eine vollständige Verdrahtung eines Chips

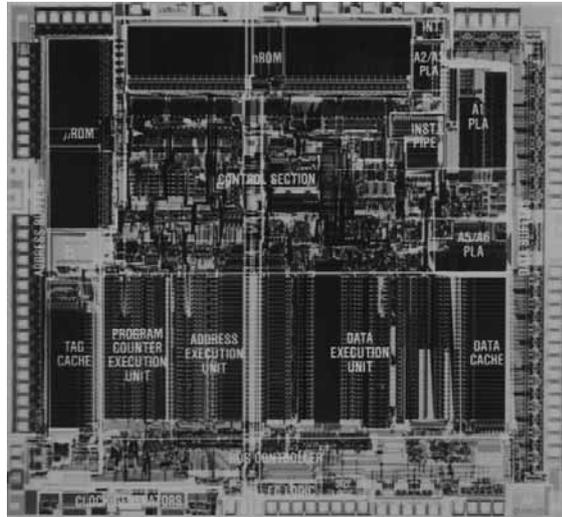


Abbildung 2.11 Prozessoraufbau des Motorola 68020

Die Abbildung 2.11 stellt den Prozessor-Chip 68020 von der Firma Motorola dar. Die einzelnen Strukturen sind noch per Hand erzeugt worden. Der Mikroprozessor PowerPC 601, eine Entwicklung von IBM/Motorola, zeigt in eine komplexere Struktur (Abbildung 2.12), in der alle notwendigen Schritte wie Partitionierung, Placement und Routing rechnergestützt vorgenommen wurden. Die Kantenlänge des Chips beträgt etwa 20 mm. Die einzelnen Funktionseinheiten wie Caches, Memory Management Unit (MMU), Fetch- und Branch-Unit, Load- und Store-Unit, Integer-Unit mit einem 32 Bit-Addierwerk, Mehrzweckregister und Floating Point Unit (FPU) verteilen sich auf einzelne Bereiche und sind nicht erkennbar.

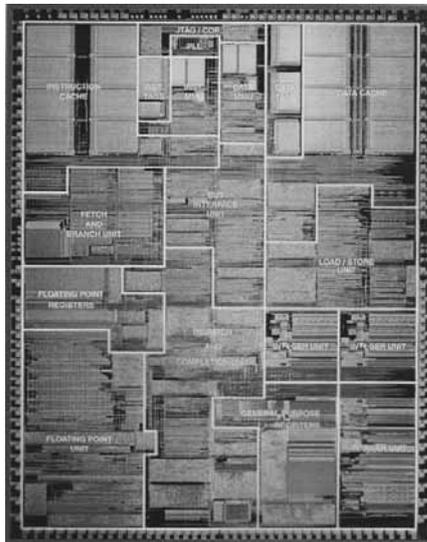


Abbildung 2.12 PowerPC Prozessor

Um einen Chip benutzen zu können, muss er einer Platine mit entsprechenden Leiterbahnen, auch gedruckte Schaltung genannt, aufgesetzt werden. Der Chip wird zunächst auf ein sogenanntes Keramik-Substrat (z.B. Bariumtitanat) aufgelötet, anschließend müssen die nach außen führenden Verbindungen des Chips mit entsprechend größeren Anschlüssen, die letztlich mit der gedruckten Schaltung in Kontakt stehen, verbunden werden (Packaging). Prinzipiell existieren zwei grundsätzlich verschiedene Verfahren, wie diese Verbindung hergestellt werden kann. Bei der sogenannten Dual-in-line-Methode, die in der Abbildung 2.13 gezeigt ist, werden über feine Drähte die Chip-Anschlüsse mit Lötballchen verschweißt. Diese stehen in direktem Kontakt mit den Pins des gesamten Moduls, wobei letzterer aus dem eigentlichen Chip, dem Bariumtitanat-Substrat, den einzelnen Verbindungsdrähten und den Pins einschließlich der Lötballchen besteht. Bei dieser Methode befinden sich alle Pins an der Außenseite des Moduls.

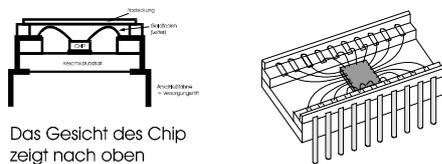


Abbildung 2.13 Dual-in-line-Methode

Bei modernen Mikroprozessoren sind das einige hundert Anschlüsse. Die zweite, modernere Flip-Chip-Methode dreht den Chip um, d.h. wenn die Oberfläche des Chips unter dem Elektronenmikroskop betrachtet wird, so sind keinerlei Strukturen

2 Technologische Grundlagen

sichtbar sondern nur reines Silizium. Bei der Dual-in-line-Methode wären sämtliche Muster erkennbar, weil in diesem Fall die entscheidende Seite oben liegt.

In der Abbildung 2.14 ist zu erkennen, dass der Chip wieder über Lötballchen mit einer gedruckten Schaltung verbunden ist. Als Basis dient auch hier das Keramiksubstrat, aus dem letztlich die Pins herausgeführt werden. Wird beim Flip-Chip die Schutzkappe abgenommen, so ist der vollständige Mikroprozessor mit etwa $15 * 15$ externen Anschlussstiften (Pins) sichtbar. Diese Pins haben einen Abstand von ca. 100 mil, wobei 1 mil ein tausendstel Inch beträgt.

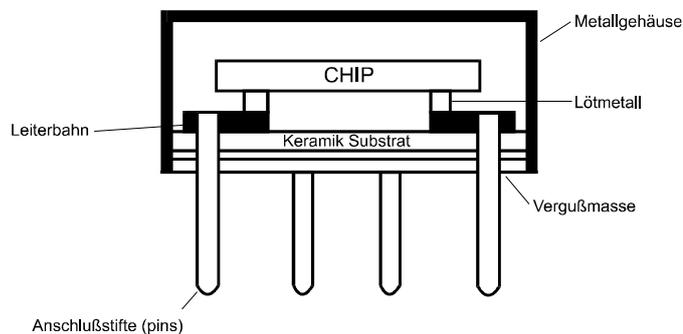


Abbildung 2.14 Flip-Chip-Methode

In der Abbildung 2.15 ist die Rückseite eines Dual-in-line Chips dargestellt, die entscheidende Fläche ist die zwischen dem Chip und der Oberfläche des Keramikmoduls. Der schraffierte Teil zeigt das Modul, während der Chip schwarz gezeichnet ist. Die Kreise stellen die Lötballchen dar, mit denen der Chip auf das Keramiksubstrat aufgelötet ist. Das Keramiksubstrat wiederum wird mittels spezifischer Verfahren mit der gedruckten Schaltung leitend verbunden.

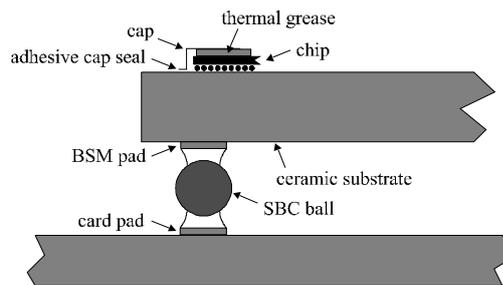


Abbildung 2.15 Solder Ball Connect (SBC) Darstellung

Bei der klassischen 'Pin-through-hole'-Methode werden die Pins durch speziell dafür gebohrte Löcher geführt und dann anschließend in einem Schwallötbad entweder auf der Ober- oder Unterseite oder auf beiden verlötet. Bei der 'Solder Ball Connect (SBC) '-Methode, einem moderneren Verfahren, wird über sehr kleine Lötballchen das

Keramiksubstrat mit der gedruckten Schaltung verbunden. In der Abbildung 2.16 sind beide Methoden gegenübergestellt.

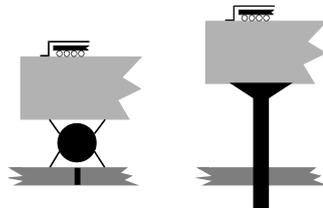


Abbildung 2.16 Das moderne SBC-Verfahren und das Pin-through-hole-Verfahren

Die Flip-Chip-Methode ist das einzig anwendbare Verfahren, wenn es sich um einige hundert Kontakte handelt, weil hier die Verbindungen der Außenanschlüsse des Chips mit den Pins automatisch hergestellt werden können. Dieser Arbeitsgang muss im Fall der Dual-in-line-Methode manuell erfolgen.

2.4 Energieprobleme in Rechnersystemen

Eine der Hauptschwierigkeiten, die der Rechnerarchitekt zu bewältigen hat, ist darauf zurückzuführen, dass die Rechenoperationen des Computers durch eine Vielzahl von Schaltern und Schaltvorgängen realisiert werden. Während vor einigen Jahrzehnten z.B. UND- bzw. ODER-Schalter noch in Form von Relais implementiert wurden, werden heute Transistoren benutzt.

Abbildung 2.17 zeigt das Prinzip des UND- sowie des ODER-Schalters. Der UND-Schalter wird durch zwei in Reihe angeordnete Schalter gebildet, beim Oder-Schalter liegen beide parallel.

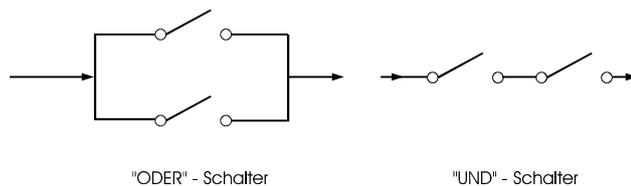


Abbildung 2.17 Darstellung des Prinzips eines ODER- und eines UND-Schalters

Immer dann, wenn ein Schalter geöffnet oder geschlossen wird, verbraucht dieser Energie. Der Hauptanteil der Energie entfällt auf Umladungsvorgänge infolge des Ohmschen Widerstandes $R=U/I$ der Leitungen. Hinzu kommen die Leitungskapazitäten der Verbindungsdrähte innerhalb eines Rechners. Beim Schalten werden die Leitungen entweder auf- oder entladen. Jeder dieser Lade- oder Entlade-Vorgänge verbraucht die elektrische Energie $\frac{1}{2} \cdot C \cdot U^2$, dabei bedeuten C die Leitungskapazität und U die anliegende Spannung. Bei genauerer Betrachtung gehen noch die Leitungskapazitäten der Transistor-Anschlüsse mit ein, sie sind aber im Verhältnis zu den Verbindungsleitungen vernachlässigbar. Die elektrische Energie wird in Wärmeenergie um-

gewandelt und an die Umgebung abgegeben. Die umgesetzte Energie wächst linear mit der Taktrate des Prozessors, weil die Schalter in einer festgelegten Zeit öfter schalten.

Zwei unterschiedliche Probleme ergeben sich aus der Existenz der zum Rechnen notwendigen Schaltvorgänge. Das erste Problem besteht darin, den Transistoren auf dem Chip bei zunehmender Integration genügend Energie zuzuführen. Durch den immer mehr abnehmenden Querschnitt der Leitungsdrähte wird die Stromdichte in den Leitungen sehr groß. Der Strom hat bei solch hohen Dichten die Eigenschaft, Atome aus dem Gitterverband herauszureißen, was nach einer bestimmten Betriebsdauer zur Folge hat, dass durch diese Elektronen-Migration keine Atome im Gitter mehr vorhanden sind. Das bedeutet den Ausfall des Mikroprozessors, d.h. die Betriebsdauer eines Mikroprozessors ist infolge der Elektron-Migration begrenzt. Da es sich hierbei um statistische Effekte handelt, kann trotzdem davon ausgegangen werden, dass ein heutiger Mikroprozessor das gesamte Rechnersystem überlebt. Der physische Verschleiß anderer Rechner Teile und natürlich der moralische Verschleiß der gesamten Anlage schreiten schneller fort als der Alterungsprozess des Mikroprozessors. Das zweite unmittelbare Problem während des Betriebs ist neben dem Heranführen der elektrischen Energie das Abführen der Wärmeenergie. Als Vergleich möge an dieser Stelle das Bügeleisen dienen. Die von diesem Gerät ausgestrahlte Energiedichte beträgt etwa 5 Watt / cm². In der Mikroprozessortechnik sind heute Werte um 50 Watt/cm² üblich. Einer der ersten Pentium 60 Mikroprozessoren hatte einen Wert von etwa 13 Watt, allerdings war der Chip größer als 1 cm², d.h. der Wert bewegte sich zwischen 7 und 9 Watt / cm². Der erste Alpha-Chip hatte bei ca. 2 cm² Chipoberfläche eine Energieabstrahlung von 30 Watt bzw. 15 Watt / cm². Der Prozessor des IBM-Großrechners der /390 Architektur hatte vor 10 Jahren einen Wert um 20 Watt/cm². Die Aufgabe besteht nun vorrangig darin, entsprechende Kühlmechanismen zu entwickeln, die verhindern, dass sich das ganze System soweit erhitzt, dass dadurch Bauelemente oder ganze Funktionseinheiten durchbrennen und die Rechenanlage ausfällt. Jeder Computer-Hersteller hat diesbezüglich seine eigene Technologie entwickelt. Die Lösung der Firma Motorola ist in Abbildung 2.18 gezeigt.

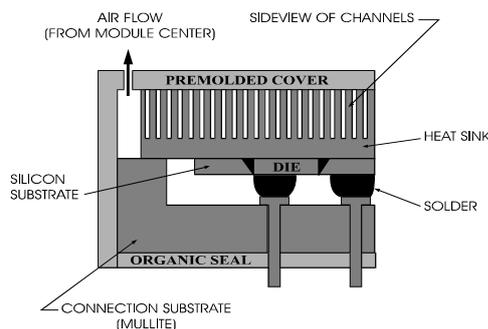


Abbildung 2.18 Kühlsystem von Motorola

Das Flip-Chip (DIE) ist über die Lötballchen mit den Pins verbunden. Letztere sind vom Keramiksubstrat umgeben. Um in diesem Fall die Wärmemengen abführen zu können, wird auf das Silizium-Substrat ein Kühlkörper aufgesetzt. Dieser Kühlkörper besitzt eine Menge von Lamellen, durch die Luft geblasen wird. Im Gegensatz dazu hat der Pentium-Prozessor anstatt dieses Kühlkörpers einen kleinen, speziellen Lüfter, der direkt auf dem Prozessorchip angebracht ist und dafür sorgt, dass die Wärmeenergie von 10 Watt abgeführt wird. Da der Lüfter für seinen Betrieb ebenfalls elektrische Energie verbraucht und davon auch einen Teil in Wärme umsetzt, wird sehr viel Aufwand betrieben, um diese Energiebilanzen zu optimieren.

Bei allen Maßnahmen der Energie-Zufuhr und -abfuhr stellt sich die Frage, ob es bei einem Schaltvorgang unbedingt zu einem Energieumwandlungsprozess kommen muss. Momentan werden weltweit große Anstrengungen unternommen (z.B. IBM-Forschungszentrum in Yorktown), um Antworten auf diese Fragen zu finden. Die derzeitig am weitesten verbreitete Meinung ist, dass es theoretisch möglich sein müsste, Rechner zu bauen, bei denen der Energieverbrauch pro Schaltvorgang vernachlässigbar klein ist. Die Praxis zeigt aber, dass zurzeit noch niemand weiß, wie derartige Schaltelemente aussehen sollen. Die derzeitige Lösung des Problems besteht darin, die Spannungspegel sämtlicher Transistoren immer weiter herabzusetzen. Beispielsweise arbeiteten die ersten Pentium 60-Chips noch mit 5 Volt Spannungspegel, die neuen Mikroprozessor-Chips wurden dagegen auf meist deutlich unter 1,5 Volt reduziert. Da die Spannung mit der 2. Potenz in die Gleichung für den Energieverbrauch bei einem Schaltvorgang eingeht, bedeutet die Herabsetzung auf 3,5 Volt bereits einen um den Faktor 2 geringeren Energieverlust. In den nächsten Jahren wird der Trend in Richtung noch niedrigerer Spannungen gehen, da durch die höhere Fertigungsdichte und die immer mehr in einem Chip untergebrachten Bauelemente eine immer höhere Verlustwärme produzieren. Es darf bei allem Optimismus aber nicht vergessen werden, dass mit dem Herabsetzen der Spannungspegel auch Schwierigkeiten auftreten. Dieser Maßnahme ist dadurch eine untere Grenze gesetzt, dass die Transistorkennlinien nicht genügend scharf abknicken, was Auswirkungen auf das Schaltverhalten dieser Bauelemente hat. Andere Überlegungen beschäftigen sich damit, die Transistoren nicht bei Zimmertemperatur sondern bei Temperaturen um 77°K, d.h. bei der Temperatur des flüssigen Stickstoffs, zu betreiben.

Der Umstand, warum die Temperatur des flüssigen Stickstoffs als Betriebstemperatur gewählt wird, hat nichts mit dem Kühlproblem zu tun. Es hat sich jedoch herausgestellt, dass der Knick in der Transistorkennlinie bei Temperaturen um 77 K (~ -200 °C) wesentlich schärfer ist und die Transistoren bei gleicher Geschwindigkeit mit weniger Energie zu betreiben sind bzw. die Transistoren bei gleichem Energieverbrauch schneller schalten. Es ist demzufolge vorstellbar, dass zukünftige Rechner nicht mehr bei Zimmertemperatur sondern bei der Temperatur des flüssigen Stickstoffs betrieben werden. Derartige Rechner sind nicht neu, sie sind schon gebaut worden aber nur mit mäßigem Erfolg. Auf Grund der immer mehr wachsenden Integration und der oben

erwähnten Probleme ist es durchaus vorstellbar, dass diese Idee wieder größere Beachtung erfährt.

Einen weiteren Schritt auf dem Weg zur Lösung des Energieproblems stellt die Supraleitung dar. Mit Supraleitung wird der Effekt bezeichnet, bei dem bei Temperaturen in der Nähe des absoluten Nullpunkts (~ -273 °C) der elektrische Strom widerstandslos die Materie passiert. Dieses Phänomen wurde erstmals von den beiden Physikern Müller und Bednorz in bestimmten Keramiken experimentell nachgewiesen. Der Gedanke, supraleitende Computer zu bauen, ist aber auf Grund von technologischen Schwierigkeiten noch weit von der Realität entfernt. Er könnte aber langfristig die Lösung der Energie-Probleme in Rechenanlagen bedeuten.

Das Energie-Problem wird noch prekärer infolge des Technologie-Trends, mehrere Chips auf ein Substrat aufzulöten. Ein Beispiel dafür bildete der Prozessor P6 der Firma Intel als Nachfolger des Pentium. Hier sind 2 Chips auf einem Modul aufgelötet. Der erste Chip ist der eigentliche Mikroprozessor, der zweite ist der sogenannte L2-Cache.

Auf Grund der Tatsache, dass die Chips jetzt noch enger aneinander rücken, wird die Wärmeabfuhr noch dringender. Während die Energiedichte der Wärmeabstrahlung bei der Technologie vor 2 oder 3 Jahren noch um $20 \text{ Watt} / \text{cm}^2$ lag, liegt diese jetzt bei 40 bis 50 Watt pro cm^2 . Das stellt erhöhte Anforderungen an die Wärmeableitung und es muss auf andere Methoden zurückgegriffen werden. Das am weitesten entwickelte Verfahren ist die TCM-(Thermal Conduction Module) Technologie der Firma IBM. Abbildung 2.19 zeigt ein derartiges TC-Modul, bei dem ein Rechteck ausgeschnitten ist. Einem keramischen Chip-Träger sind z.B. 35 Chips (z900) aufgelötet. Jedem Chip wird ein Aluminium-Stempel aufgesetzt, der mittels einer Feder im oberen Teil der Führung auf den Chip drückt und so den Kontakt herstellt. Der untere Teil des Chips und der Stempel sind von flüssigem Helium umgeben. Die Energie wird von dem Chip über den Stempel in den umgebenden Aluminium-Körper und von diesem in das sogenannte Cold Plate übertragen. Durch das letztere wird entweder eine Kühlflüssigkeit oder Luft gespült. Bei einer Luftkühlung sind die Öffnungen in dem Kühlkörper größer. Der Luftstrom muss mit hoher Geschwindigkeit durch die Kanäle gedrückt werden, um diese hohe Wärmemenge abzuführen. Es werden dabei ultraschall-ähnliche Geschwindigkeiten verwendet, so dass der Geräuschpegel Werte erreicht, die ein Arbeiten in der Nähe dieser Anlagen unmöglich machen. Im Gegensatz dazu kühlt die Firma CRAY ihre Zentraleinheit mit Schwefelhexafluorid (SF_6). Dabei wird diese Flüssigkeit mittels einer Pumpe in schnellem Umlauf gehalten; Schwefelhexafluorid hat eine bessere Wärmeleitfähigkeit als Wasser oder Luft, bei Wartungsarbeiten an den CRAY-Rechenanlagen wird die Kühlflüssigkeit in besondere Vorratsbehälter gepumpt. Ein Nachteil dieses Kühlmittels ist der hohe Preis für seine Entsorgung.

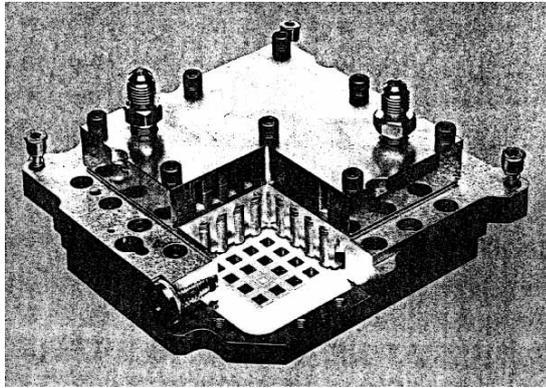


Abbildung 2.19 TCM-Technologie von IBM

Die TCM-Technologie ist heute die am weitesten verbreitete, da sie als einzige in der Lage ist, auf relativ einfache Weise sehr große Wärmemengen zu transportieren. Der Nachteil liegt in den hohen Herstellungskosten dieser Module. Ihr Vorteil besteht dagegen in der hohen Zuverlässigkeit. Nach Aussagen des Herstellers sind in der gesamten Zeit des Betriebes der TC-Module keinerlei Ausfälle registriert worden.

Eine weitere Schwierigkeit wird von der TCM-Technologie beseitigt. Die einzelnen Verbindungen der Kontakte müssen von den Chips zu den Pins nach außen geführt werden. Bei einem einzelnen Chip ist das noch mit einigem geometrischen Aufwand möglich. Befinden sich aber mehrere Chips auf einem Keramik-Modul (**Multi Chip Modul**), dann müssen die Verbindungen zwischen den verschiedenen Chips extrem kurz ausfallen. Als Folge davon werden keine gedruckten Schaltungen mehr benutzt, vielmehr sind für derartig komplizierte Verdrahtungsmuster spezielle mehrlagige Keramik-Module entwickelt worden. Als Substrat für die Mehrlagen-Module verwendete man Aluminiumoxid (Al_2O_3). Neuerdings wird anstatt Aluminiumoxid Keramik verwendet, dessen relative Dielektrizitätskonstante um $1/3$ kleiner ist und damit die Laufzeiten der Signalverbindungen verkürzt. Innerhalb der verschiedenen Schichten, deren Anzahl bei der IBM-MCM-Technologie bis zu 107 (z900) betragen kann, entstehen sehr komplexe Verdrahtungsmuster. Die senkrechten Verbindungen zwischen den verschiedenen Schichten bestehen aus leitenden Bohrungen, die wiederum innerhalb einer Schicht in horizontalen Leiterbahnen weitergeführt werden und an einer Bohrung zu einer darunter- oder darüberliegenden Schicht enden usw. Die Abbildung 2.20 zeigt ein ESA/9000 Mehrschichtkeramik TC-Modul.

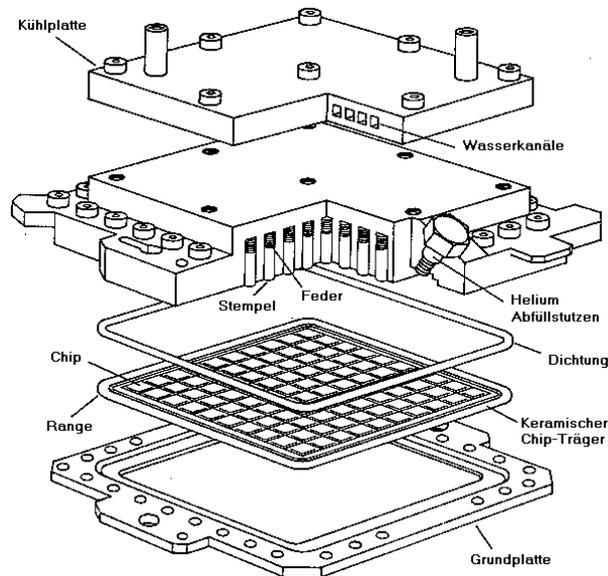


Abbildung 2.20 ESA/9000 Mehrschichtkeramik TC-Modul

Die Frage nach der Optimierung des Energieverbrauchs wird heute im Wesentlichen schon bei der Auswahl der verwendeten Transistor-Technologie geklärt. Vor ca. 10 Jahren wurden für die Implementierung von Mikroprozessoren der mittleren Leistungsklasse fast ausschließlich CMOS- (Complementary MOS) Schaltkreise verwendet. Für CMOS-Transistoren ist die Schaltzeit umgekehrt proportional zur angelegten Spannung, d.h. der Transistor schaltet umso schneller, je mehr Energie ihm zugeführt wird. Dieser lineare Zusammenhang zwischen angelegter Spannung und Schaltgeschwindigkeit gilt aber nur für einen bestimmten Bereich. Wird die Spannung weiter über diesen Grenzbereich hinaus gesteigert, erhöht sich die Schaltgeschwindigkeit nicht mehr. Die Grenze liegt momentan für CMOS-Schaltkreise bei etwa 2-5 Picosekunden. Ein ähnlich linearer Zusammenhang ergibt sich für ECL-(Emitter Coupled Logic) Schaltkreise. Der Unterschied zu CMOS-Schaltkreisen besteht darin, dass ein ECL-Schaltkreis bei gleicher Schaltgeschwindigkeit etwa das Hundertfache an Energie verbraucht. Die ECL-Schaltkreise haben dagegen den Vorteil eines signifikant größeren linearen Bereichs. Somit kann, auf Kosten des Energieverbrauches, die Schaltgeschwindigkeit weiter gesteigert werden. Letzteres hat zur Folge, dass in leistungsfähigen Großrechenanlagen bevorzugt bipolare Schaltkreise eingesetzt wurden, die entsprechend kostspielige Flüssigkeitskühlsysteme benötigten.

Inzwischen sind die Unterschiede zwischen Großrechnern und Workstations bezüglich der CPU-Nutzer-Zeit geringer geworden, d.h. die Energie-Probleme übertragen sich auf die PC- und Workstation-Generationen.

Obwohl die Bipolar- als auch die CMOS-Technik jedes Jahr weiter verbessert wurden, blieb der relative Abstand zwischen beiden um den Faktor 4-5 erhalten. Desweiteren konnten die Kosten eines Rechnersystems mit der schnelleren Bipolar-Technologie verringert werden, wenn anstatt Flüssigkeitskühlung eine Luftkühlung verwendet wurde. Damit verringerte sich natürlich die Rechengeschwindigkeit. Prognosen, die in Richtung Weiterentwicklung der Bipolar-Technologie und deren Einsatz in luftgekühlten Großrechenanlagen zielen, haben sich als falsch herausgestellt. Momentan zeichnet sich auf dem internationalen Computermarkt ab, dass das Marktvolumen für den Umsatz von Großrechnersystemen gleich geblieben ist. Die Preise dieser Produkte sind aber dramatisch gesunken. Dafür konnten die Hersteller in der gleichen Zeit höhere Stückzahlen verkaufen. Infolge ungeheurer gestiegener Kosten für die Technologieentwicklung haben alle großen Computerfirmen (SGI, Hitachi, Fujitsu, IBM, Intel) beschlossen, die Bipolar-Technologie, obwohl diese etwa bis zu einer halben Größenordnung schneller ist als die CMOS-Technik, nicht mehr weiterzuentwickeln. Man hat sich dafür auf den Kompromiß geeinigt, auf dem Silizium-Chip sowohl FET- als auch Bipolar-Transistoren unterzubringen. Die Strategie ist folgende: An bestimmten kritischen Stellen, an denen besondere Geschwindigkeitsgewinne zu erreichen sind, werden gezielt bipolare Transistoren eingesetzt, selbst wenn der Energieverbrauch um den Faktor 30 höher liegt als bei CMOS-Transistoren. Der Anteil der Bipolar-Technologie an der sogenannten BICMOS-Technologie liegt zwischen 2 und 4 %. Der Alpha-Mikroprozessor der Firma DEC stellt mit einem Anteil von bis zu 20 % eine Ausnahme dar. Dies ist auch der Grund dafür, dass der Prozessor sehr hohe Betriebstemperaturen entwickelt.

Eine weitere Möglichkeit zur Steigerung der Schaltgeschwindigkeit von Transistoren bildet die Gallium-Arsenid-Technologie. Gallium-Arsenid besitzt im kristallinen Zustand eine analoge Struktur zu Silizium, hat aber durch die asymmetrische Bindung eine größere Elektronenbeweglichkeit. Schaltkreise auf Gallium-Arsenid-Basis haben deutliche Geschwindigkeitsvorteile gegenüber der Silizium-Technologie. Die Entwicklung und die Produktion sind aber dermaßen kostenintensiv, dass Aufwand und Nutzen zurzeit in keinem vertretbaren Verhältnis stehen. Es existieren in den USA einige Firmen (Vitesse), die Gallium-Arsenid-Chips vertreiben. Die Firma Convex produziert mit ihrer C3-Serie ebenfalls Rechner, in denen Funktionseinheiten in Gallium-Arsenid-Technik implementiert werden. Vermutlich wird sie aber mittelfristig wieder zur Silizium-Technologie zurückkehren, weil der Markt einfach nicht bereit ist, diese teure Technik zu bezahlen.

2.5 SOI-Technologie

In den letzten zehn Jahren ist weltweit intensiv an der Entwicklung einer neuen Technologie für die Produktion von Mikroprozessoren gearbeitet worden. Die Firmen Motorola und IBM haben maßgeblichen Anteil daran, dass die SOI-Technologie (Silicon On Insulator) einen Stand erreicht hat, um vollfunktionsfähige Mikroprozessoren herzustellen (<http://www.chips.ibm.com/bluelogic/showcase/soi/>). SOI bildete auch