# 3. Speichersubsysteme

## 3.1. Grundbegriffe

In den derzeitigen Computersystemen können wir zwei Arten von Speichern unterscheiden (Abbildungen 3.1 bis 3.3): Direktzugriffsspeicher und Massenspeicher.

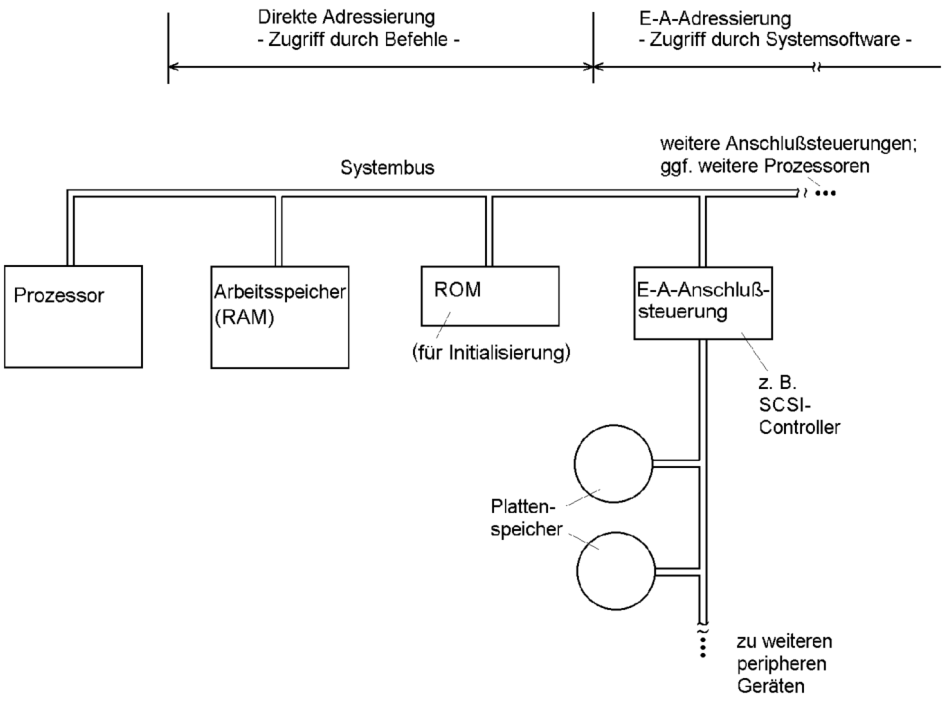



***Abbildung 3.1*** Ein typisches Speichersubsystem

*Direktzugriffsspeicher*
Das sind die Speichereinrichtungen, zu denen der Prozessor unmittelbar zugreift. Direktzugriffsspeicher werden *adressiert*, das heißt, der Prozessor (oder auch eine andere Einrichtung im System) liefert eine *Adresse*, die gleichsam eine binär codierte laufende Nummer der betreffenden *Speicherposition* (Speicherzelle) darstellt. Jede Speicherposition enthält eine feste Anzahl von Bitpositionen (deren Anzahl wird als *Zugriffs- bzw. Aufrufbreite* bezeichnet). Die geringste Aufrufbreite beträgt üblicherweise 8 Bits (= 1 Byte). In Abhängigkeit von der Ausgestaltung des Speichers, des Prozessors und des Systems können auch Worte, Doppelworte und andere Informationsstrukturen adressiert werden (typische Zugriffsbreiten sind 16, 32, 64, 128, 256 Bits usw.). Direktzugriffsspeicher funktionieren rein elektronisch, sie sind mit Schaltkreisen aufgebaut und haben keine mechanisch bewegten Teile.

*Massenspeicher*
Es handelt sich um Disketten- und Festplattenlaufwerke, Magnetbandgeräte usw. Sie können *nicht* unmittelbar über Adressen angesprochen werden. Die Massenspeicher enthalten keine einzeln zugänglichen Bytes, Worte usw., sondern Dateien. Alle Zugriffe werden über *Dateizugriffs- und Gerätesteuerprogramme* geführt, die zum *Betriebssystem* gehören. Solche Speicher beruhen zumeist auf mechanischen Prinzipien (um das Speichermedium zu bewegen und den gewünschten Speicherinhalt auszuwählen).

*Der Unterschied aus der Sicht des Praktikers:*

- die Direktzugriffsspeicher gehören zur Plattform. Sie sind typischerweise in unmittelbarer Nähe des Prozessors angeordnet (in PCs: auf dem Motherboard),
- die Massenspeicher gehören zur Peripherie. Sie werden wie andere periphere Geräte in den PC eingebaut oder außen angeschlossen.

*Hinweise:*

1. Auch manche Laufwerke werden in der Fachliteratur gelegentlich als Direktzugriffsspeicher bezeichnet (genauer: als DASD = Direct Access Storage Devices). Das hat aber tiefere Gründe.
2. In modernen Computersystemen gehört es zum Stand der Technik, den Arbeitsspeicher im Verbund mit Massenspeichern so zu betreiben, daß sich die Anordnung dem Nutzer (bzw. dem Anwendungsprogramm) gegenüber wie ein einziger adressierbarer Speicher sehr großer Kapazität verhält. Der Fachbegriff: *virtueller Speicher*. Die einzelnen Funktionen werden durch Zusammenwirken von Hard- und Software verwirklicht.

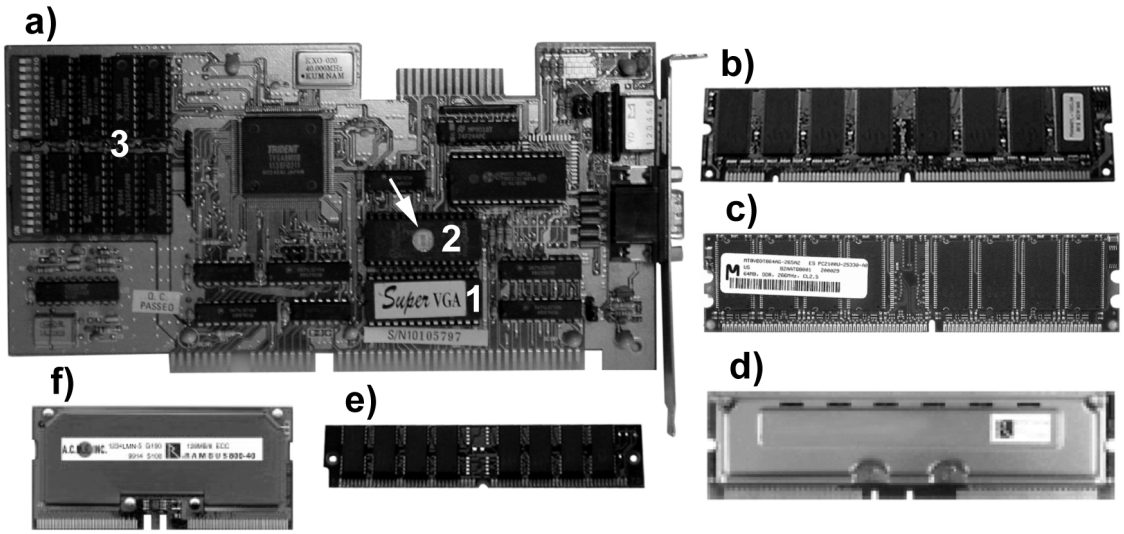



***Abbildung 3.2*** Direktzugiffsspeicher (Ausführungsbeispiele)

*Erklärung:*
a) - Steckkarte mit herkömmlichen Speicherschaltkreisen auf Fassungen; b) bis f) - Speichermoduln. b) - SDRAM-DIMM (SDR); c) - DDR-DIMM; d) - RIMM; e) - SIMM mit 72

Anschlüssen; f) - SO-RIMM. 1, 2 - Festwertspeicher (EPROMs). Der Pfeil zeigt auf das Quarzglasfenster, durch das der Speicherinhalt mittels UV-Licht gelöscht werden kann. 3 - Bildspeicher (DRAMs).

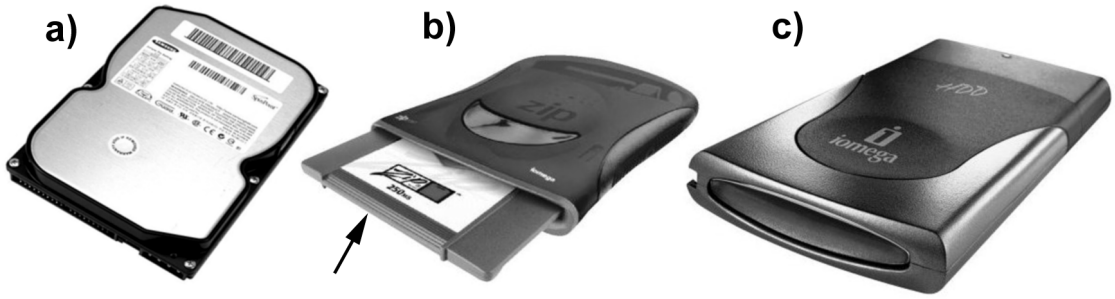



***Abbildung 3.3*** Massenspeicher (Ausführungsbeispiele)

*Erklärung:*
a) - interner Massenspeicher (Festplatte); b) und c) - externe Massenspeicher. b) - Massenspeicher (ZIP-Laufwerk) mit wechselbarem Datenträger (Pfeil); c) - externe Festplatte. Externe Laufwerke werden in mehr oder weniger extravagant gestalteten Gehäusen untergebracht.

Im folgenden geht es ausschließlich um Direktzugriffsspeicher.

**Wie funktioniert ein Direktzugriffsspeicher?**
An sich ist es ganz einfach (Abbildung 3.4). Ein solcher Speicher muß eigentlich nur drei Grundfunktionen ausführen können: Schreiben, Lesen und Daten behalten.

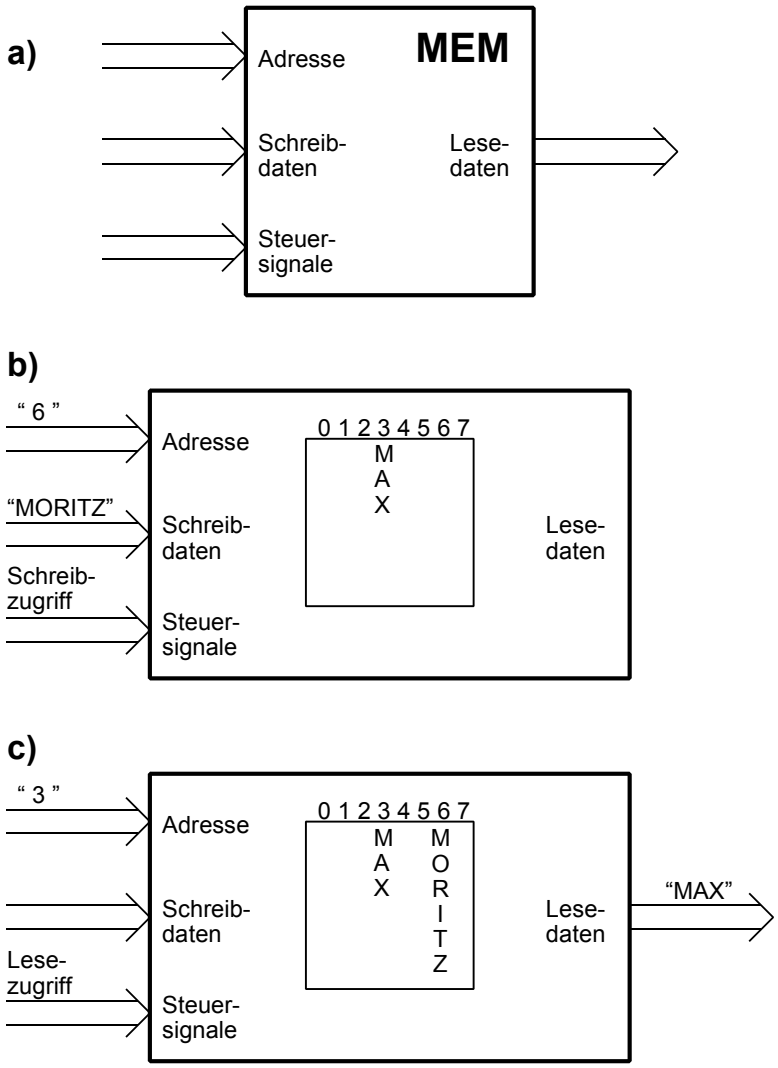



***Abbildung 3.4*** Zur Wirkungsweise des Direktzugriffsspeichers

*Erklärung:*

a) das allgemeine Schaltsymbol (MEM steht für Memory (= Speicher)). Der Speicher hat Eingänge für die Adresse, für die zu schreibenden Daten und für die Steuersignale (die ihm u. a. mitteilen, ob geschrieben oder gelesen werden soll). Zum Abliefern der gelesenen Daten sind entsprechende Ausgänge vorgesehen. In der Praxis haben viele Speicher nur einen Datenweg, der sowohl zum Lesen als auch zum Schreiben verwendet wird. Vgl. den bidirektionalen Datenweg der Prozessoren (Abschnitt 2.3).

b) ein Schreibzugriff. Es ist weiterhin ein gegebener Speicherinhalt dargestellt. Der Speicher hat Speicherpositionen (Speicherzellen), die fortlaufend durchnumeriert sind. Die laufende Nummer ist die Adresse der jeweiligen Speicherzelle. Wir schreiben hier auf Speicherzelle Nr. 6.

c) ein Lesezugriff. Es ist weiterhin der Speicherinhalt dargestellt, wie er sich nach dem vorangegangenen Schreiben ergeben hat. Wir lesen von Speicherzelle Nr. 3 und erhalten somit deren Belegung.

# 3.2. Speicherschaltkreise im Überblick

## 3.2.1. Grundlagen

*Speicherprinzip und Speicherinterface*
Die einzelnen Speicherarten (ROMs, SRAMs, DRAMs usw.) unterscheiden sich vor allem im Aufbau der Speicherzellen. Die einzelnen Speicherzellen sind in sog. Speichermatrizen zusammengefaßt. Der Speicherschaltkreis ist ein Verbund aus der Speichermatrix (manchmal aus mehreren Speichermatrizen)*) und dem jeweiligen Interface (Abbildung 3.5). Es gibt asynchrone und synchrone Speicherinterfaces. Heutzutage sind die Hersteller in der Lage, jede Speichertechnologie mit jedem Interface zu fertigen.

*): der Fachbegriff: 1 Matrix einschließlich Adreßdecoder, Leseverstärker usw. = 1 Zugriffsweg = 1 Bank (sprich: Bänk). Demgemäß spricht man von Schaltkreisen mit beispielsweise 1, 2, 4, 8, 16 oder 32 Banks. Beachten Sie aber, daß „Bank" in der Speicherwissenschaft mehrere verschiedene Bedeutungen hat.

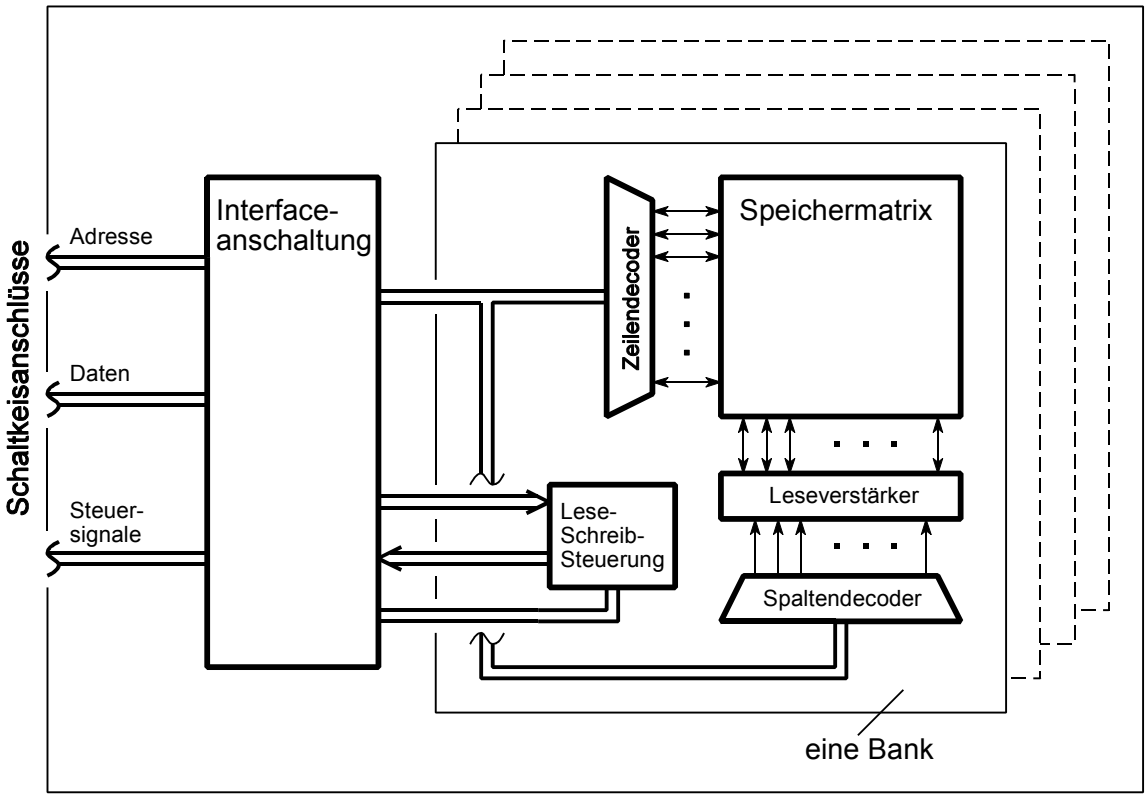



***Abbildung 3.5*** Die grundsätzliche Struktur eines Speicherschaltkreises

*Asynchrone Speicherschaltkreise* werden direkt gesteuert. Die Speicheranordnung wird von den Adreßeingängen aus unmittelbar adressiert, und der Datenweg ist praktisch eine einfache Durchreiche zu den Datenanschlüssen.

*Synchrone Speicherschaltkreise* enthalten besondere Zwischenspeicher (Register) für Adressen,

Daten und Steuersignale. Alle Register werden mit einem gemeinsamen Taktsignal angesteuert. Hierdurch werden alle Übertragungsvorgänge auf einen gemeinsamen Takt bezogen (vgl. Abbildung 2.5). Somit kann man die Speicher in eine „vollsynchrone" Taktorganisation einbeziehen. Das bringt folgende Vorteile:

- günstigere Zeitverhältnisse der Speicheransteuerung: die Signalbelegungen sind nur in Bezug auf die Taktflanken von Bedeutung (Prinzip: „wie die Signale in der restlichen Zeit aussehen, ist egal."). Demgegenüber müssen beim asynchronen Speicher die Adreß- und Steuersignale praktisch während des gesamten Speicherzyklus stabil gehalten werden.
- in die Speicherschaltkreise können autonome Adressierungsschaltungen (Adreßzähler) eingebaut werden, um zusammenhängende Datenblöcke lückenlos zu lesen oder zu speichern (Burst-Betrieb).

## 3.2.2. Festwertspeicher (ROM)

Ein Festwertspeicher wird einmalig mit seinem Inhalt versehen (programmiert). Der Speicherinhalt kann dann nur noch gelesen werden; ein Ändern ist typischerweise nicht mehr möglich (siehe aber auch Tabelle 3.1 und den darauf folgenden Text). Festwertspeicher enthalten die Programme, die beim Starten des Computers, z. B. nach dem Einschalten, benötigt werden, um die Hardware zu überprüfen (Anfangstest) und um den normalen Betrieb einzuleiten. Eine weitere Anwendung (von den Stückzahlen her sicherlich die wichtigste) ist die Speicherung von Programmen in Embedded Systems - vom Toaster über Waschmaschine und Faxgerät bis hin zum Auto. Derartige Programme müssen sofort nach dem Einschalten verfügbar sein (stellen Sie sich vor, Sie drehen den Zündschlüssel Ihres Autos - und ehe der Motor anspringt, rollen zunächst einmal Anzeigen jener Art über das Armaturenbrett, wie sie der PC nach dem Einschalten von sich gibt....).

*Wie werden ROMs programmiert?*
Hier wollen wir uns mit einem Überblick über die Programmierverfahren begnügen (Tabelle 3.1).

In der PC-Technik finden wir ROMs auf Motherboards und Steckkarten, aber auch in Laufwerken und anderen Geräten. Dem Zug der Zeit entsprechend werden die Speicherkapazitäten immer gewaltiger und die damit realisierten Funktionen immer komplizierter - und kaum ein Hersteller kann der Verlockung (oder: dem Zwang?) widerstehen, mit Neuheiten so schnell wie möglich am Markt zu sein. Infolgedessen steigt natürlich die Fehlerrate. Deshalb bevorzugt man heutzutage Festwertspeicher, deren Inalt auch noch beim Anwender geändert werden kann. So kann man das BIOS der Motherboards und die ROM-Inhalte von Steckkarten, Laufwerken usw. über Diskette oder gar über das Internet umprogrammieren, um die Firmware zu ändern bzw. auf einen neueren Stand zu bringen (sprich: appzudeeten...).

*Die klassische Form des Update:* der Schaltkreistausch. Hierzu muß der ROM-Schaltkreis auf einer Steckfassung angeordnet sein. Beispiele finden wir u. a. auf herkömmlichen Motherboards und Steckkarten (vgl. Abbildung 3.2a).

| Programmierverfahren | Bezeichnung | Änderbarkeit | Bemerkungen |
|---|---|---|---|
| Maskenprogrammierung | Mask ROM | nicht änderbar | nur bei extremen Stückzahlen von praktischer Bedeutung |
| Durchschmelzprinzip (Fuse) | PROM | nicht änderbar | alle Verbindungen sind vorgefertigt, die nicht benötigten werden beim Programmieren getrennt |
| Ladungsspeicherung mit UV-Löschung | EPROM, OTP | durch Löschen und Neuprogrammieren | Löschen durch UV-Licht. Erfordert Quarzglasfenster imSchaltkreis*). Es gibt auch preisgünstige Ausführungen ohne Fenster. Diese kann man nicht mehr löschen (OTP) |
| Ladungsspeicherung mit elektrischer Löschung | EEPROM, Flash ROM | durch Löschen und Neuprogrammieren (auch in der Anwendungsumgebung (In System Programming)) | Löschen durch elektrische Impulse |

PROM = Programmable Read Only Memory; EPROM = Erasable Programmable Read Only Memory; OTP = One Time Programmable; EEPROM = Electrically Erasable Read Only Memory. *): vgl. Position 2 (Pfeil) in Abbildung 3.2

***Tabelle 3.1*** Programmierverfahren für Festwertspeicher (Überblick)

*Die moderne Form des Update:* das Umprogrammieren des ROM-Schaltkreises. Typischerweise werden sog. Flash-ROMs eingesetzt. Das „Flash" (sprich: Fläsch) = Blitz, Aufflammen usw. steht als bildhafte Beschreibung des Löschvorgangs. Der Speicherinhalt wird typischerweise auf einen Schlag gelöscht, so daß der Schaltkreis anschließend erneut programmiert werden kann. Solche Schaltkreise sind zumeist fest eingelötet.

Diese - für die Hersteller - so vorteilhafte Lösung ist aber mit Problemen verbunden:

- ist der Speicherinhalt erst einmal gelöscht, aber beim erneuten Programmieren etwas schiefgegangen, so funktioniert anschließend gar nichts mehr,
- auch ROM-Inhalte sind - angesichts der allgemeinen Vernetzung - nicht mehr sicher: es gibt z. B. tatsächlich Viren, die den Inhalt von BIOS-ROMs verändern.

*Hinweis:*
Die Schaltkreishersteller sind Profis und haben natürlich von Anfang an eine gewisse Vorsorge getroffen - durch einzelne Schreibzugriffe (die ja auch irrtümlich ausgelöst werden könnten, z. B. infolge eines Defekts) kann ein ROM-Inhalt nicht verändert werden. Typische Schutzvorkehrungen auf Motherboards, in Geräten usw.:

- hardwareseitiger Schreibschutz über Steckbrücke (Jumper),
- nichtlöschbare Bereiche im ROM (im einfachsten Fall ein sog. Boot Block - ein Bereich, der die ganz elementare Firmware enthält - einschließlich der Notprogramme zum Wiederherstellen des Speicherinhalts),

- zwei ROMs, ein änderbarer und ein fester. Der feste enthält eine Art Not-BIOS, das einen elementaren Systembetrieb unterstützt (manchmal bis hin zum Hochfahren von Windows).

E gibt verschiedene Stufen des Komforts, den die Hersteller vorsehen:

- beigegebene Software zum Erstellen von Update- und Wiederherstellungs-Disketten, ergänzt durch gutgemeinte Hinweise, wie man z. B. das BIOS nach einem mißglückten Update-Versuch wiederherstellen kann (Fachbegriff: BIOS Recovery),
- fertige Notfalldisketten mit Update- und Wiederherstellungssoftware,
- Update über Windows und Internet (wenn es ums BIOS geht, sollte das Motherboard ein unzerstörbares Notfall-BIOS haben - sonst wäre es ein wirkliches Glücksspiel...).

## 3.2.3. Speicher mit wahlfreiem Zugriff (RAM)

Computersysteme können mehrere Speicheranordnungen mit wahlfreiem Zugriff enthalten. Sie werden mit RAM-Schaltkreisen bestückt. Somit können beliebige Speicherpositionen (Bytes, Worte, Doppelworte usw.) in ihrem Inhalt verändert werden. Mit dem Abschalten der Speisespannung geht der Speicherinhalt verloren. Im PC-Bereich sind vor allem drei derartige Speichersubsysteme von Bedeutung: der Arbeitsspeicher, die Caches und der Bildspeicher.

*Statische RAMs (SRAMs)*
Diese Schaltkreise halten die gespeicherten Daten so lange, wie die Speisespannung anliegt. Die Speicherzellen der SRAMs sind allerdings recht aufwendig (man benötigt z. B. 4 bis 6 Transistoren je Bit), so daß sich vergleichsweise hohe Kosten ergeben. SRAM-Arbeitsspeicher kommen deshalb nur in (kleineren) Embedded Systems zum Einsatz. Typische Speicherkapazitäten von SRAM-Schaltkreisen liegen zwischen 8 kBits und 16 MBits. Die größten Stückzahlen fließen in die Fertigung von Telekommunikationsgeräten (Mobiltelefone, Faxgeräte usw.).

*Dynamische RAMs (DRAMs)*
Nach dem DRAM-Prinzip läßt sich die einzelne Speicherzelle mit einem einzigen Transistor aufbauen. DRAMs können deshalb mit extremen Speicherkapazitäten gefertigt werden (dem Stand der Technik entsprechen 64 MBits bis 4 GBits). Die Nachteile:

- der Speicherinhalt klingt mit der Zeit ab (auch bei anliegender Betriebsspannung). Um die gespeicherten Daten zu erhalten, sind fortlaufend Zugriffe erforderlich, die die Daten in den Speicherzellen gleichsam auffrischen (Refresh-Zugriffe).
- die Ansteuerung von DRAMs ist komplizierter als die von SRAMs.

## 3.2.4. DRAM-Schaltkreise im Überblick

Alle DRAM-Schaltkreise beruhen auf dem gleichen Speicherprinzip. Die Auslegung der Speichermatrix bestimmt letzten Endes die Zugriffszeit.

*Wir merken uns:*
Die eigentlichen Speicherfunktionen (Schreiben, Lesen, Datenerhalt) beruhen in allen modernen DRAM-Schaltkreisen - gleich welcher Bauart - auf den gleichen Prinzipien. Alle in vergleichbaren Technologien gefertigten Speichermatrizen haben nahezu gleiche Zeitkennwerte, gleichgültig ob sie sich in einem EDO-DRAM, in einem SDR- oder DDR-SDRAM oder in einem RDRAM befinden. Die einzelnen Typen unterscheiden sich nur in der Anzahl der Banks und im Interface. Asynchrone DRAMs haben typischerweise eine einzige Bank, synchrone DRAMs hingegen zwei bis 32.

Die Überlegenheit der modernen (synchronen) DRAMs beruht nicht auf grundsätzlich besseren Speicherzellen, sondern auf der Auslegung der Schaltkreise und Speichersubsysteme:

- da es mehrere Banks gibt, kann auf entsprechend viele Bits gleichzeitig zugegriffen werden,
- die synchronen Interfaces gewährleisten eine sehr schnelle Datenübertragung,
- die synchronen Interfaces untersützen die zeitlich überlappte Ausführung mehrerer unabhängiger Zugriffe.

Die modernen DRAMs sind aber nur dann überlegen, wenn bestimmte Voraussetzungen gegeben sind:

- es wird vorwiegend auf längere Datenblöcke zugegriffen,
- es gibt tatsächlich immer wieder mehrere unabhängige Zugriffe.

Gäbe es immer nur wahlfreie Zugriffe auf einzelne Bytes mit beliebig verteilten Adressen, so wären die modernen Speichersubsysteme den altmodischen sogar unterlegen, da zu den „eigentlichen" Zugriffen (auf die Speichermatrix) noch die Eintaktierungszeiten der synchronen Schnittstellen (Speicherschaltkreis, Speichersteuerung, Brücke, Prozessorbus) hinzukommen.

*FPM*
FPM = Fast Page Mode. Die ursprüngliche (= älteste) Auslegung asynchroner DRAMs.

*EDO*
EDO = Extended Data Out. Eine nicht ganz so alte Auslegung asynchroner DRAMs.

FPM und EDO unterscheiden sich in gewissen Spitzfindigkeiten der Belegung des Datenbus. Wichtig ist zunächst, daß es diese beiden Ausführungen gibt und daß sie sich nicht immer untereinander vertragen.

*SDRAM*
SDRAM = Synchronous DRAM. Ein DRAM mit einem taktgesteuerten (synchronen) Interface. Beim „klassischen" SDRAM wird jeweils nur eine Taktflanke zur Datenübertragung ausgenutzt (vgl. Abbildung 2.5a).

*SDR-SDRAM*
SDR = Single Data Rate. Eine andere Bezeichnung für die „gewöhnlichen" SDRAMs.

*DDR-SDRAM*
DDR = Double Data Rate. DDR-SDRAMs sind SDRAMs mit doppelter Datenrate. Beide Flanken des Taktsignals werden zur Datenübertragung ausgenutzt (vgl. Abbildung 2.5b).

*SGRAM*
SGRAM = Synchronous Graphics Memory. Ein synchroner DRAM mit zusätzlichen Funktionen, der eigens für Bildspeicher vorgesehen ist (gibt es mit einfacher und mit doppelter Datenrate (SDR/DDR)).

*GDDR-DRAM*
Ein vorzugsweise für Bildspeicher vorgesehener synchroner DRAM mit doppelter Datenrate.

*RDRAM*
RDRAM = Rambus-DRAM. Ein DRAM mit Rambus-Interface. Nach und nach wurden verschiedene Ausführungen entwickelt: das ursprüngliche Rambus-Interface, Concurrent Rambus, DirectRambus. Im PC-Bereich ist nur DirectRambus von Bedeutung.

*Datenrate und Frequenz als Kennwert*
Hier heißt es aufpassen - im Datenmaterial der modernen DRAMs ist oft von Frequenzen die Rede, wenn eigentlich Datenraten gemeint sind. Der typische Kennwert heißt „Datenrate je Anschluß": MBits/s/p bzw. GBits/s/p (Bits je Sekunde und Pin), E-A-Frequenz (I/O Frequency) oder System Bus Speed. Er betrifft die Anzahl der Datenübertragungsvorgänge je Sekunde. Da beide Taktflanken zur Datenübertragung ausgenutzt werden, ist die eigentliche Taktfrequenz jeweils nur halb so hoch wie dieser Kennwert. Beispiele: 500 MHz E-A-Frequenz ≙ 250 MHz Takt, 800 MHz E-A-Frequenz ≙ 400 MHz Takt.

*DDR-Schaltkreise*
DDR-SDRAMs beruhen auf herkömmlichen Grundlagen (Weiterentwicklung der SDR-SDRAMs). Der typische Zeitkennwert (System Bus Speed) entspricht der doppelten Taktfrequenz.

*Hinweis:*
Nur die eigentliche Datenübertragung läuft mit der doppelten Taktfrequenz ab; Kommandos und Adressen werden nur mit einer Taktflanke übertragen (wie SDR).

*DDR (DDR-I)*
DDR-I bezeichnet den bisherigen Stand der Technik. Typische Taktfrequenzen: 100 MHz, 133 MHz, 166 Hz und 200 MHz, womit sich die jeweils doppelte Datenrate je Anschluß ergibt (die entsprechenden Schaltkreisbezeichnungen: DDR-200, DDR-266, DDR-333, DDR-400). Die Schaltkreise haben typischerweise 4 Banks.

*DDR2 (DDR-II)*
Die nächste Entwicklungsstufe. Doppelte Datenrate gegenüber DDR-I. Typische Taktfrequenzen: 200 MHz, 266 MHz, 333 MHz. Die doppelte Datenrate je Anschluß führt zu den Schaltkreisbezeichnungen DDR2-400, DDR2-533 und DDR2-667. Die Schaltkreise haben 4 oder (von 1 GBits an) 8 Banks.

*Rambus-Schaltkreise*
Die Rambus-Prinzipen wurden entwickelt, um hochleistungsfähige Speichersubsysteme besonders kostengünstig zu fertigen. Ursprünglich ging es vor allem um den Einsatz in Geräten, die für den Massenmarkt bestimmt sind.

Rambus-Speicherubsysteme beruhen auf vergleichsweise schmalen, mit extremen Taktfrequenzen betriebenen Signalwegen (Abbildung 3.6). Grundlage der Datenspeicherung sind aber nach wie vor herkömmliche DRAM-Speichermatrizen (der Vorteil: Kostensenkung durch Nutzung bewährter Technologien). Es handelt sich also um eine evolutionäre Weiterentwicklung*). Und das sind die wesentlichen Ansätze:

- Taktsteuerung statt Strobe-Steuerung (synchrone Arbeitsweise),
- Kommando-Steuerung statt Steuerung über Einzelsignale,
- Speicherschaltkreise mit vielen Banks (16 oder 32 sind typisch),
- paketweise Übertragung statt Übertragung einzelner Adreß- und Datenworte (man spricht von „paketorientierten" - Packed Based - DRAMs),
- Schaltkreisauswahl über Adreßvergleich statt über besondere Auswahlsignale,
- Betrieb mit extremen Taktfrequenzen (400 MHz und mehr); Nutzung beider Taktflanken, besondere Signalpegel (Rambus Signaling Levels RSL).

*): die Rambus-Entwicklung begann Anfang der 90er Jahre (8 Bits Datenwegbreite, 250 MHz).

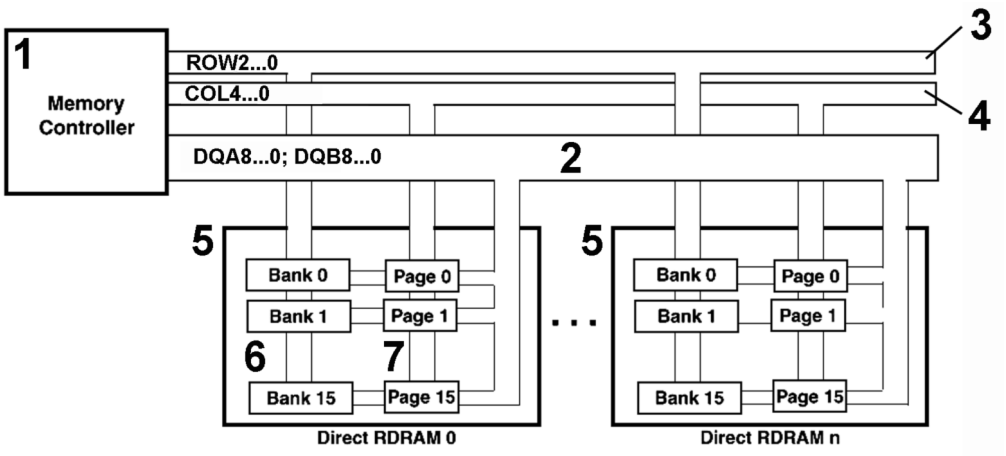



***Abbildung 3.6*** Struktur eines DirectRambus-Speichersubsystems (Rambus Inc.)

*Erklärung:*
Die Abbildung zeigt einen sog. Rambus-Kanal (RDRAM Channel). 1 - Speichersteuerung; 2 - Datenweg; 3, 4 - zusätzliche Signalwege zur Kommando- und Adreßübertragung (3 bzw. 5 Bits breit); 5 - Speicherschaltkreis mit 16 DRAM-Speicherbanks.

Die extreme Taktfrequenz ist das Entscheidende - hierdurch allein ist es möglich, kostengünstige, aus nur wenigen Schaltkreisen bestehende Speichersubsysteme für hohe Datenraten und

annehmbare Latenzzeiten auszulegen[1]. Je höher die Taktfrequenz, desto weniger Anschlüsse braucht man, um eine bestimmte Datenrate zu gewährleisten. DirectRambus-Speichersubsysteme kommen deshalb mit 2 Bytes breiten Datenwegen aus. Zur Übertragung von Kommandos und Adressen sind unabhängige Signalwege vorgesehen (ROW: für „Zeilenpakete", COL: für „Spaltenpakete")[2]. Die Pakete enthalten u. a. die Zeilen- bzw. Spaltenadressen.

Typische Taktfrequenzen: 400 MHz, 533 MHz, 600 MHz. Damit ergeben sich Datenübertragungsfrequenzen von 800, 1066 und 1200 MHz (I/O Frequency). Dementsprechend werden auch die Schaltkreise bezeichnet (RDRAM-800 oder PC800, RDRAM-1066 oder PC1066, RDRAM-1200 oder PC1200).

*Anmerkungen:*

1) vgl. Abschnitt 2.3.8. Es sei nochmals darauf hingewiesen, daß bei aufeinanderfolgenden datenabhängigen Zugriffen die Latenzzeiten durchaus ein Problem sein können. (Das zeigt sich in der Praxis daran, daß mit RDRAMs bestückte Computer in Tests nicht immer überzeugen - es kommt sehr darauf an, welche Arten von Zugriffen in den jeweiligen Programmen besonders häufig vorkommen.)
2) parallel zu laufenden Datenübertragungen können hierüber bereits weitere Adressen und Zugriffskommandos an die RDRAMs geschickt werden,

*Noch höhere Datenraten: XDR*
Eine neue Schnittstelle, die von den Entwicklern des Rambus kommt (Abbildung 3.7). Typische Merkmale im Überblick:

- Datenweg: Einzelanschluß der DRAMs an den Steuerschaltkreis (kein Bus, sondern eine Punkt-zu-Punkt-Verbindung),
- Datenwegbreite: im Prinzip von 1 bis 32 Bits. Typisch sind zwei Bytes.
- Übertragungsverfahren: differentielle Signalübertragung (zwei Leitungen je Bitposition) mit extrem niedrigen Pegeln (Differential Rambus Signaling Levels DRSL mit nur 0,2 V Signalhub) und sehr hoher Übertragungsfrequenz (3,2 GHz),
- Adreß- und Kommandoübertragung: über besonderen Bus (12 Informationsleitungen), der an alle Speicherschaltkreise geführt ist. Übertragungsverfahren: RSL (Pegel und Takte wie beim Rambus).
- Taktsystem: Grundtakt 400 MHz (wie Rambus). Datenübertragung mit achtfacher Datenrate (Octal Data Rate Signaling ODR). Der Original-Rambus-Takt wird vervierfacht, und es werden beide Taktflanken zur Datenübertragung ausgenutzt. Übergang auf 800 MHz Grundtakt vorgesehen.
- Datenrate eines Speichersubsystems mit 16 Bits breiten Datenwegen: 6,4 GBytes/s.

*Eine kurze Zusammenfassung:*

- asynchrone Speicher = Schaltkreise ohne Takt. Der wichtigste Zeitkennwert ist die Zugriffszeit (so spricht man z. B. von einem 60-ns-DRAM).
- synchrone Speicher = Schaltkreise mit Takt. Der wichtigste Zeitkennwert ist die (höchstzulässige) Taktfrequenz (so spricht man z. B. von einem 133-MHz-Speichermodul).
- Rambus-Speicher = Schaltkeise mit (extrem schnellem) Takt und paketorientiertem (schmalem) Interface. Hier ist eine Freqenzangabe, die die Datenrate betrifft, der wichtigste Zeitkennwert (I/O Frequency). Typische Werte: 800, 1066, 1200 MHz. Die eigentliche Taktfrequenz beträgt jeweils die Hälfte (400, 533, 600 MHz)).
- DDR-Speicher = synchrone Speicher mit verbessertem herkömmlichen Interface und Ausnutzung beider Taktflanken zur Datenübertragung. Der wichtigste Zeitkennwert ist eine Frequenzangabe, die gelegentlich als System Bus Speed bezeichnet wird. Sie betrifft - wie die I/O Frequency des Rambus - die Datenrate je Anschluß. Es gibt zwei Entwicklungsstufen: DDR und DDR2. Typische Werte DDR: 200 MHz, 266 MHz, 333 MHz, 400 MHz. Typische Werte DDR2: 400 MHz, 533 MHz, 667 MHz.. Die Taktfrequenz (Clock Frequency) beträgt jeweils die Hälfte (100 MHz, 133 MHz, 166 MHz; 200 MHz, 266 MHz, 333 MHz).
- *Schaltkreisbezeichnungen* für DDR und Rambus betreffen dieDatenübertragungsfrequenz (System Bus Speed, I/O Frequency). Beispiele: DDR-200, DDR-333, DDR2-533, RDRAM 1066.

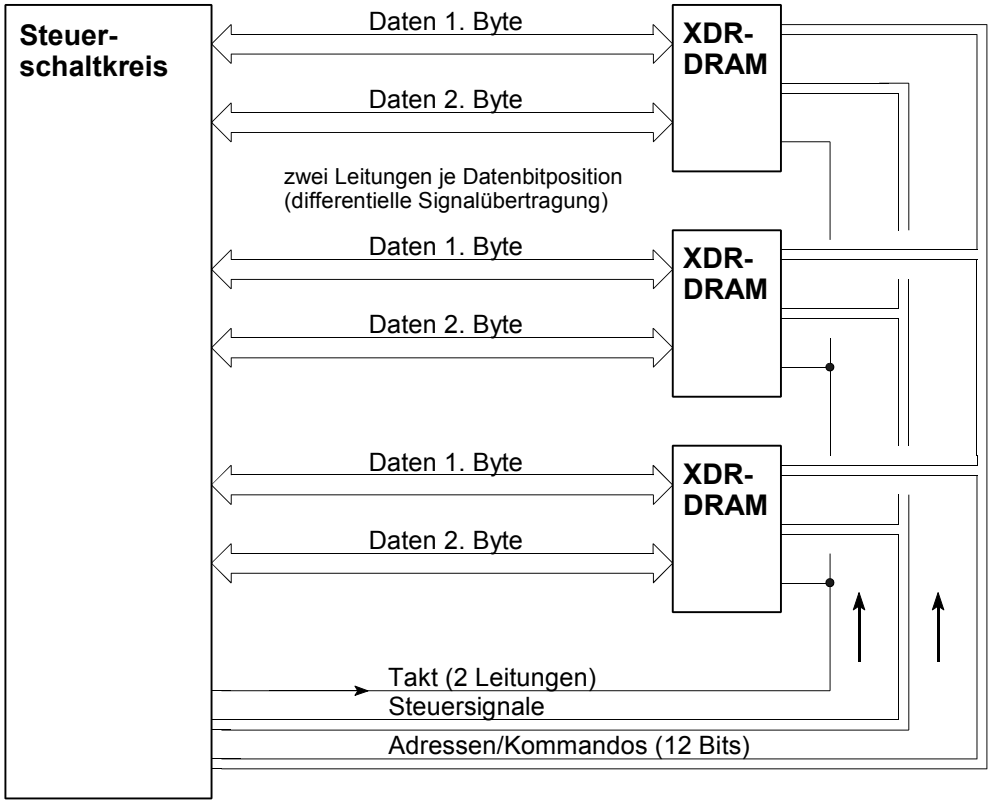


***Abbildung 3.7*** Struktur eines XDR-Speichersubsystems

*Speicherkapazitäten und Organisationsformen*
Die Speicherkapazität des einzelnen Schaltkreises wird üblicherweise in Bits angegeben. Oft kennzeichnet man die Organisationsform durch eine Kapazitätsangabe der Form „n adressierbare Speicherzellen · b Bits" (auf die gleichzeitig zugegriffen werden kann). Tabelle 3.2 nennt einige Beispiele.

*Hinweis:* Schaltkreise mit geringerer Datenwegbreite lassen sich kostengünstiger fertigen (weniger Anschlüsse, kleinere Gehäuse). Deshalb bevorzugt man für größere Speicherkapazitäten (Arbeitsspeicher) die Organisationsformen · 4 und · 8.

| Bezeichnung | Speicherkapazität | Zugriffs- bzw. Datenwegbreite*) |
|---|---|---|
| 256k · 1 | 256 kBits | 1 Bit |
| 256k · 4 | 1 MBits | 4 Bits |
| 1 M · 4 | 4 MBits | 4 Bits |
| 2 M · 8 | 16 MBits | 8 Bits |
| 2 M · 32 | 64 MBits | 32 Bits |

*): jeweils so viele Bits werden gleichzeitig geschrieben bzw. gelesen

***Tabelle 3.2*** Organisationsformen von Speicherschaltkreisen (ausgewählte Beispiele)

*Entwicklungstendenzen*
Abbildung 3.8 veranschaulicht die Entwicklung der Speicherkapazität von DRAM-Schaltkreisen seit den 80er Jahren. Die DRAM-Fertigung ist derzeit durch Typen von 64 MBits bis 4 GBits gekennzeichnet. Was gefertigt wird, hängt weniger von der Beherrschung der Technologien ab als vielmehr von kommerziellen Erwägungen (die Halbleiterfirmen liefern das, wofür der Kunde zahlt...).

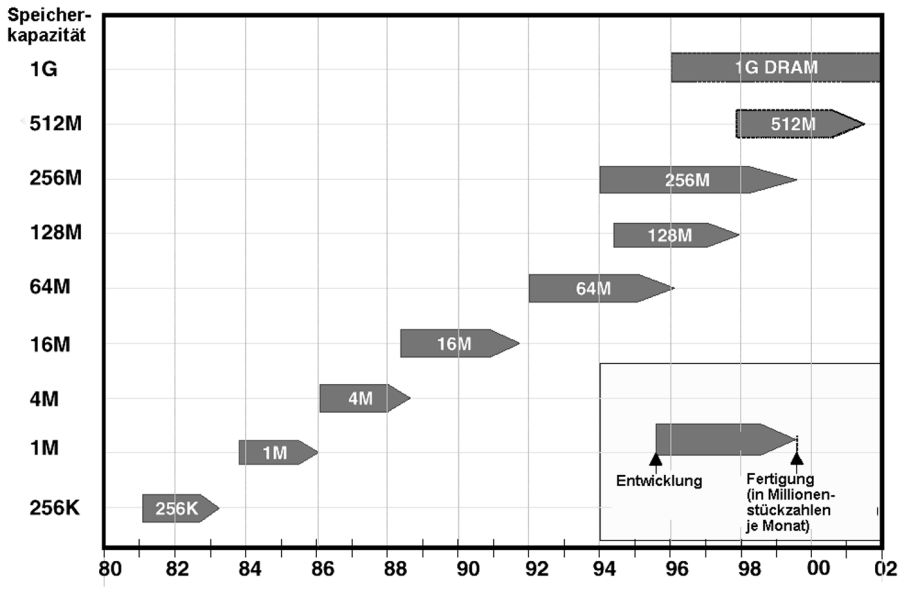



***Abbildung 3.8*** Zur Entwicklung der Speicherkapazität von DRAM-Schaltkreisen. Ein historischer Rückblick (Advanced Memory International, Inc.). Die Spitzen bezeichnen den

Beginn der eigentlichen Massenfertigung

# 3.2.5. RAMs mit Datenerhalt

Gelegentlich ist es erforderlich, Speicherinhalte, die während des Betriebs verändert werden können, auch dann zu erhalten, wenn die Speisespannung abgeschaltet ist. Fachbegriff: Nonvolatile RAM („nichtflüchtiger RAM"). Die funktionellen Anforderungen „Erhalt der gespeicherten Daten bei ausgeschalteter Speisespannung" und „schnelle, wahlfreie Lese- und Schreibzugriffe" lassen sich nur mit vergleichsweise beträchtlichen Aufwendungen erfüllen.

**RAMs mit Batteriestützung**
Es liegt nahe, RAM-Anordnungen aus einer Batterie zu speisen. SRAMs kann man recht einfach zwischen Normalbetrieb und Datenerhalt umschalten; in DRAM-Anordnungen sind hingegen zusätzliche Aufwendungen erforderlich (Refresh-Funktion). Die Speicherdauer wird von der Stromentnahme und von der Kapazität der Batterie bestimmt. Kleine SRAMs in CMOS-Technologie brauchen so wenig Strom, daß sie mit einer Batterie auch über Jahre hinweg gespeist werden können.

*NV- oder Zeropower-RAMs* sind Bauelemente, die einen SRAM, die Speisespannungsumschaltung sowie die Stützbatterie in einem gemeinsamen Gehäuse enthalten. Da im Schaltkreisgehäuse die Batterie untergebracht werden muß, kommen nur vergleichsweise große Gehäusebauformen in Frage, und die Bauhöhe übertrifft typischerweise die der üblichen SRAM-Schaltkreise (DIL-Gehäuse: ca. 10 mm, PLCC-Gehäuse: ca. 5 mm). Zugesichert wird ein Datenerhalt über wenigstens 10 Jahre (die Anzahl der Schreibzugriffe ist dabei vollkommen beliebig).

*Hinweise:*

1. Aufpasssen. Die Batterie kann auch durch unsachgemäßes Lagern oder unachtsamen Umgang (Kurzschluß) entladen werden!
2. Bei modernen Schaltkreisen ist die Batterie zunächst nicht zugeschaltet; sie wird erst dann wirksam, wenn die Speisespannung erstmals einen bestimmten Schwellwert überschreitet (mit anderen Worten: die Batterie eines solchen fabrikneuen NV-SRAMs wird durch bloßes Lagern praktisch nicht entladen). Trotzdem aufpassen!

**NOVRAMs**
Dies sind SRAMs mit untergelegten EEPROM-Zellen. In dieser Technologie sind aber nur vergleichsweise geringe Speicherkapazitäten realisierbar (einige kBytes), und der Datenaustausch zwischen SRAM und EEPROM erfordert eine gewisse Zeit (EEPROM → SRAM (Recall): einige µs; SRAM → EEPROM (Store): einige ms). Der Vorteil: weil man keine Batterie mitverpacken muß, kann man NOVRAMs auch in kleinen und flachen Gehäusen liefern. Zugesicherte Eigenschaften: Datenerhalt > 100 Jahre, wenigstens 1 Million EEPROM-Schreibzugriffe.

**ROM-Schaltkreise**
Ist es erforderlich, wahlfreie Schreibzugriffe auszuführen, scheiden alle ROM-Technologien aus. Im System änderbare ROMs eignen sich aber dazu, gelegentlich (z. B. vor dem Ausschalten) bestimmte Daten zu retten. Grundsätzlich kommen nur EEPROM- oder Flash-Typen in Frage.

Die Anzahl der Schreib- bzw. Programmiervorgänge ist begrenzt. Man geht aber davon aus, daß diese Anzahl - über die Lebensdauer der Hardware gesehen - kaum ausgeschöpft werden kann (z. B. würden 10 000 Programmierzyklen rund 27 Jahren entsprechen, wenn jeden Tag ein solcher Vorgang stattfindet).

**Die Auslagerung auf Massenspeicher**
Dieses Verfahren, in Rechenzentren seit Jahrzehnten üblich, erfordert allerdings eine gewisse „Infrastruktur", da während des Auslagerns praktisch die gesamte Hardware noch betriebsfähig sein muß. (Eine Uralt-Lösung: die gesamte Installation wird über einen Motorgenerator gespeist. Bei Netzausfall läuft dieser über seine Schwungmasse noch einige Sekunden weiter. Diese Zeit genügt, um den Speicherinhalt automatisch zu retten.)

**Nichtflüchtige Speicher im PC**
Im PC-Bereich finden wir praktisch alle erläuterten Prinzipien vor.

*Der CMOS-RAM*
Es handelt sich um eine kleine RAM-Anordnung, die bei ausgeschaltetem PC durch eine Batterie gespeist wird. Sie enthält die Konfigurationsdaten des Systems (Setup-Angaben). Dieser RAM ist üblicherweise mit der Tageszeituhr des PCs in einem Schaltkreis vereinigt.

*EEPROMs zum Aufzeichnen von Betriebsdaten und Fehlerzuständen*
Vor allem Laufwerke und Geräte sind mit (seriellen) EEPROMs ausgestattet, in denen die jeweilige Firmware Betriebsdaten, Konfigurationseinstellungen und Fehlermeldungen aufzeichnet (die Auswertung dieser Daten wird leider von der gängigen Systemsoftware nicht immer unterstützt). Auch die seriellen EEPROMs der modernen Speichermoduln (vgl. Abbildung 3.13) können entsprechend genutzt werden (bis hin zum Hinterlegen von Eigentümer-Kennzeichnungen, Inventarnummern usw.).

*Der Flash-ROM des BIOS*
Dieser wird gelegentlich dazu genutzt, veränderliche Konfigurationsdaten zu halten - eine im Grunde häßliche Sparlösung, die es erschwert, den einfachsten und wirksamsten Virenschutz einzusetzen - nämlich das hardwareseitige Sperren der Programmierzugriffe.

*Die Stützung des Arbeitsspeichers (Netzteil im Stromsparbetrieb, Batterie)*
Der PC arbeitet nicht mehr, kann aber die Arbeit jederzeit wieder aufnehmen. Manche Funktionseinheiten sind stillgesetzt, stehen aber unter Betriebsspannung, manche sind abgeschaltet (Stromsparen). Der Inhalt des Arbeitsspeichers bleibt erhalten. Diese Betriebsart ist typisch für moderne PCs, vor allem für portable. Fachbegriff: Suspend to RAM.

*Die Auslagerung des Arbeitsspeicherinhaltes auf die Festplatte*
Der PC arbeitet nicht mehr, kann aber die Arbeit jederzeit wieder aufnehmen. Nahezu die gesamte Hardware ist abgeschaltet (Stromsparen). Vor dem Abschalten wurde aber der Inhalt des Arbeitsspeichers auf die Festplatte geschrieben. Soll die Arbeit wieder aufgenommen werden, wird der Speicherinhalt in den Arbeitsspeicher zurückgeschafft. Diese Betriebsart ist typisch für moderne PCs, vor allem für portable. Fachbegriff: Suspend to Disk.

Die eben erläuterten Vorkehrungen (Suspend to RAM/to Disk) kommen auch für ortsfeste PCs

in Mode (Stromsparen, Umgehen des langwiergen Startvorgangs nach dem Einschalten). Der grundlegende Standard: ACPI (Advanced Configuration and Power Interface).

## 3.2.6. Arbeitsspeicher

Der Arbeitsspeicher nimmt Daten, Programme usw. auf, mit denen der Prozessor während des normalen Betriebs arbeitet. Viele Embedded Systems kommen mit einem vergleichsweise kleinen Arbeitsspeicher aus, da die Programme typischerweise direkt aus dem ROM abgearbeitet werden (Firmware). Für einfache Anwendungen genügen sogar nur wenige Bytes. Viele Mikrocontroller haben eine entsprechende RAM-Kapazität eingebaut (Größenordnung: von 32 Bytes an aufwärts, z. B. 256 Bytes, 2 kBytes usw. bis zu 32 kBytes und mehr). „Richtige" Computer brauchen hingegen geradezu jede Menge an Arbeitsspeicherkapazität (Faustregel: der Arbeitsspeicher ist immer zu klein). Moderne Arbeitsspeicher-Subsysteme werden typischerweise mit steckbaren Speichermoduln (Abschnitt 3.3.3) aufgebaut.

*Der Arbeitsspeicher als Speichersubsystem*
Das „klassische" Speichersubsystem beruht auf einem parallelen Bus, der für jede Signalart (Adressen, Daten, Steuersignale) eigene Leitungen enthält und dessen Datenwegbreite der jeweils geforderten Zugriffsbreite entspricht (z. B. 16, 32 oder 64 Bits). Im PC-Bereich hat man diesem Bus typischerweise so ausgelegt, daß sich die Speicherschaltkreise direkt anschließen lassen; nahezu alle Leitungen sind 1:1-Verbinungen zwischen dem Steuerschaltkreis (North Bridge, Hauptverteiler) und den Speicherschaltkreisen (Abbildung 3.9). Diese Prinzip wurde von den ersten FPM-DRAMs bis hin zu den SDR- und DDR-DRAMs beibehalten.

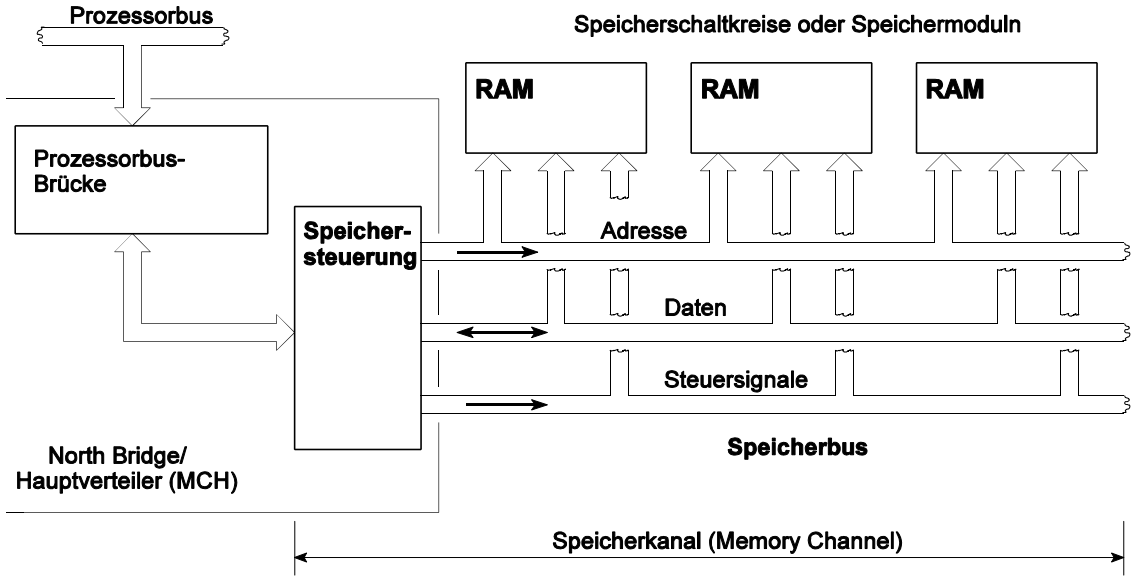



***Abbildung 3.9*** Blockschaltbild eines herkömmlichen Speichersubsystems

*Der Speicherkanal (Memory oder DRAM Channel)*
Das ist eine übliche Allgemeinbezeichnung für den Verbund aus Speicherbus und den zugehörigen Steuerschaltungen.

*Extreme Anforderungen*

Geht es um große Speicherkapazitäten und hohe Datenraten, so liegt es nahe, zwei Auslegungsprinzipen anzuwenden (und zwar einzeln oder zusammen):

- mehrere Bussysteme mit angeschlossenen Speicherschaltkreisen (mit anderen Worten: mehrere Speicherkanäle),
- Bussysteme (Speicherkanäle) mit sehr breiten Datenwegen (z. B. 128, 256 oder 512 Bits).

*Der Arbeitsspeicher im PC*
Im PC-Bereich war es von Anfang an üblich (Kostenfrage), sich mit einem einzigen Speicherkanal zu begnügen, dessen Datenwegbreite der Zugriffsbreite des Prozessors entspricht. In den oberen Leistungsklassen gibt es aber auch Systeme mit mehreren und/oder mit sehr breiten Speicherkanälen (vgl. die Abbildungen 1.33 und 1.35).

*DDR-SDRAM-Speicherkanäle*
Sie entsprechen noch dem herkömmlichen Prinzip, arbeiten aber mit geringeren Signalpegeln. Die Datenwegbreite: 64 oder (mit Fehlerkorrektur (ECC)) 72 Bits. DDR-SDRAM-Speicherkanäle haben nach wie vor ein typisches Kennzeichen der herkömmlichen Bussysteme: die Stichleitungen (Stubs) vom Bus auf dem Motherboard über den Steckverbinder zu den Schaltkreisen auf dem Speichermodul (Abbildung 3.10a). Solche Stichleitungen sind sog. Inhomogenitäten im Signalweg, die sich um so mehr störend bemerkbar machen, je höher die Übertragungsfrequenz ist. Um damit zurechtzukommen, hat man eigens eine besondere Auslegung der Signalwege entwickelt (Buskoppelstufen, Signalpegel, Leitungsabschluß usw.). Der allgemeine Fachbegriff: SSTL = Stub Series Teminated Logic.

*Rambus-Speicherkanäle*
Sie haben eine Datenwegbreite von nur 18 Bits. Der Rambus-Kanal ist an sich ein Bus (mit durchgehenden Signalleitungen, an die alle Speicherschaltkreise angeschlossen sind). Es gibt aber keine Stichleitungen. Sie werden durch eine besondere Leitungsführung vermieden (Abbildung 3.10b). Die Rambus-Speichermoduln sind gleichsam in Reihe geschaltet; die Signale sind von Modul zu Modul geführt. Sie treten über eine erste Reihe von Anschlüssen in das Modul ein und werden über eine zweite Reihe zum nächsten weitergereicht (Daisy-Chain-Prinzip). Die Speicherschaltkreise sitzen direkt über den Signalleitungen, mit denen sie über Anschlüsse im Gehäuseboden verbunden sind. Diese Kette setzt sich im System von Modul zu Modul fort. Steckfassungen ohne Modul unterbrechen somit den Signalweg. In nicht bestückte Rambus-Steckfassungen sind deshalb Blindmoduln einzusetzen. Das sind Leiterplatten in RIMM- oder SO-RIMM-Form ohne Schaltkreise, die einfach die ankommenden mit den abgehenden Signalen verbinden (vgl. Abbildung 3.22).

*Mehr als ein Speicherkanal (1): Rambus*
Weil der einzelne Rambus-Kanal (vgl. Abbildung 3.6) so schmal ist, kann man es sich leisten, mehrere davon vorzusehen (Abbildung 3.11), und zwar auch auf Motherboards, die für Mainstream-PCs bestimmt sind (zwei Kanäle sind typisch).

*Mehr als ein Speicherkanal (2): DDR*
Es gibt Schaltkreissätze, die mehrere DDR-DRAM-Kanäle unterstützen. Im Bereich der Mainstream-PCs sind zwei Kanäle üblich[*)], in den oberen Leistungsbereichen auch mehr

(beispielsweise vier oder acht; vgl. die Abbildungen 1.33 und 1.35).

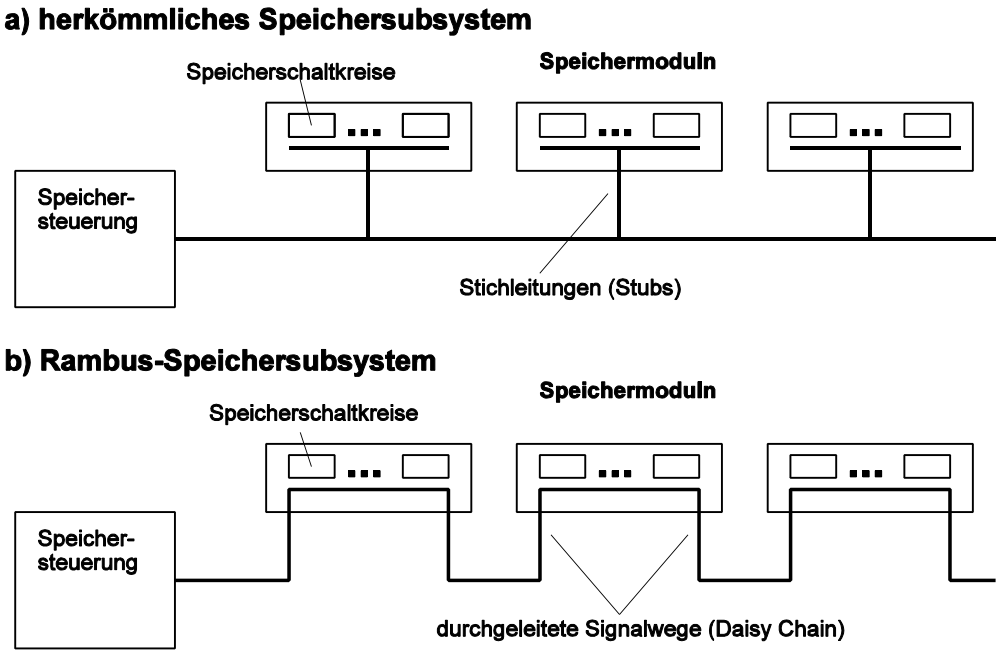



***Abbildung 3.10*** Busstrukturen im Vergleich

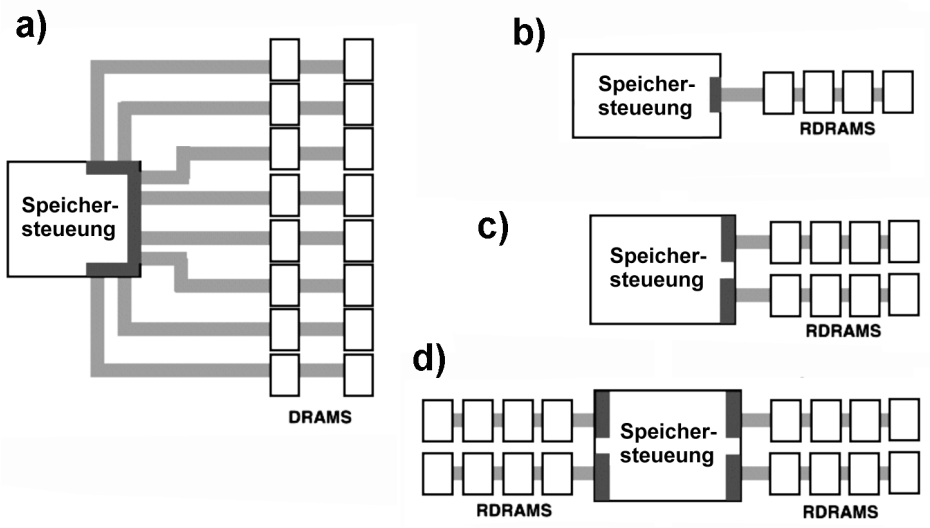



***Abbildung 3.11*** Speichersubsysteme im Vergleich (nach: Rambus, inc.)

*Erklärung:*
a) - herkömmliches Speichersubsystem (z. B. mit DDR-DRAM). Breiter Datenweg (64 oder 72 Bits), viele Leitungen. b), c), d) - Rambus-Speichersubsysteme mit 1, 2 und 4 Kanälen.

*): Beispiel: Intels Dual Channel Architecture (Abbildung 3.12).

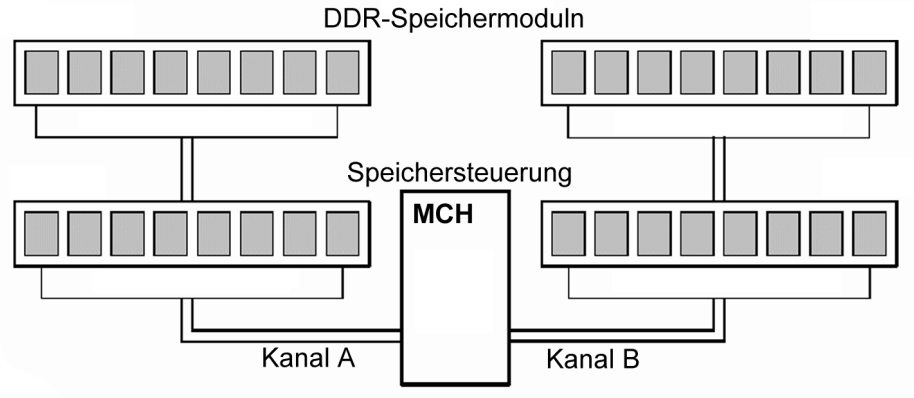



***Abbildung 3.12*** DDR-Speichersubsystem mit zwei Zugriffswegen (Dual Channel Architecture)

*Wandlung zwischen verschiedenen Speicherkanälen*
Eine typische Ausführung (vgl. Abbildung 1.35): der Speichersteuerschaltkreis (MCH, SNC o. dergl.) hat Rambus-Kanäle, die aber nicht zu Speichermoduln, sondern zu weiteren Verteilerschaltkreisen*) geführt sind, an die ihrerseits DDR-Speicherkanäle angeschlossen sind. Typischerweise ist ein Rambus-Kanal mit zwei DDR-Kanälen verbunden. Weshalb so eine Lösung?

- der Rambus-Kanal belegt deutlich weniger Anschlüsse als ein DDR-Kanal. Er eignet sich deshalb gut als Hochleistungsschnittstelle des Speichersteuerschaltkreises (vor allem dann, wenn mehrere solcher Schnittstellen vorzusehen sind).
- DDR-Speichermoduln sind kostengünstiger als Rambus-Moduln.

*): der Intel-Fachbegriff: DDR Memory Hub (DMH).

*Eine typische Ausführung:*
Die Rambus-Känal enden auf dem Motherboard an Slot-Steckverbindern für spezielle Speichersteckkarten*). Solche Karten gibt es sowohl mit Rambus-Moduln als auch mit Verteilerschaltkreisen und DDR-Moduln.

*): der Intel-Fachbegriff: Memory Expansion Card (MEC).

*Rambus oder DDR?*
Ein Rambus-Speichersubsystem braucht wenigstens zwei Kanäle, um hinsichtlich der Datenrate einem DDR-Speichersubsystem mit nur einem Kanal überlegen zu sein. Beispiel:

- ein Kanal DDR-400: mit 400 MHz werden 8 Bytes übertragen: 3 200 MBytes/s,
- ein Kanal Rambus-1066: mit 1066 MHz werden 2 Bytes übertragen: 2132 MBytes/s.

Die Praxis zeigt aber, daß „rohe" Datenrate nicht alles ist. Viele Programme laufen auf vergleichbaren Systemen (Prozessor, Speicherkapazität usw.) nahezu gleich schnell. Im Einzelfall ist manchmal Rambus, manchmal DDR überlegen.

Es kommt darauf an, in welchem Maße das Zugriffsverhalten des Programms den Eigenheiten der Speicherschaltkreise und des Speichersubsystems entgegenkommt. Beispiel: Rambus unterstützt mehr gleichzeitige, unabhängige Zugriffe als DDR. Deshalb ist zu vermuten, daß in typischen Server-Anwendungen (wo immer wieder unabhängige Zugriffe auszuführen sind) eine Rambus-Maschine überlegen ist, in Workstation-Anwendungen aber nicht (hier könnte sogar eine - ansonsten gleichartig ausgestattete - DDR-Maschine besser sein). Das ist aber reine Theorie, denn es kommt auch auf den Schaltkreissatz an*). Nicht alle Schaltkreissätze unterstützen alles, was die Speicherschaltkreise können (so gibt es u. a. Einschränkungen hinsichtlich der Banks, die gleichzeitig aktiv sein dürfen), und es sind auch nicht immer alle Funktionen optimal realisiert (Entwurfsschwächen). Also: testen und vergleichen...

*): und auf die Kosten. Was ist besser: 2 GBytes Rambus oder 4 GBytes DDR? - Vor allem im Serverbereich zieht man typischerweise den größeren Speicher vor, auch wenn er etwas langsamer ist.

## 3.2.7. Bildspeicher

Er enthält die Bilddarstellung, die auf dem Bildschirm des PCs erscheint. Um Bildanzeigen aufzubauen und darzustellen, sind wenigstens zwei verschiedenartige Zugriffe auszuführen - und zwar so, daß sie sich nicht gegenseitig stören: ein Schreiben seitens des Prozessors (programmseitiger Bildaufbau) und ein zyklisches (immer wiederholtes) Lesen zwecks Darstellung auf dem Bildschirm. Die Forderungen an Speicherkapazität und Datenrate werden immer höher (Videowiedergabe, Abspielen von DVDs, neumodische Computerspiele mit bewegten Bildern in 3 D und „wie echt" aussehenden Farben). Um diese Anforderungen zu erfüllen, hat man spezielle Bildspeicher-Schaltkreise entwickelt (Video-RAMs, Window-RAMs, SGRAMs, GDDR-DRAMs usw.). Meist bevorzugt man aber schnelle DRAM-Schaltkreise aus der Masssenfertigung (z. B. DDR oder Rambus) und erfüllt die zeitlichen Anforderungen durch entsprechende Auslegung der Steuer- und Adressierungsschaltungen.

## 3.2.8. Schnellspeicher (Caches)

Ob ein Speicher „schnell" oder eher langsam genannt werden kann, hängt von zwei Angaben ab, die man näherungsweise gleichsetzen kann: von der Zugriffs und der Zykluszeit (Access bzw. Cycle Time). In der hier angemessenen - nicht ganz exakten - Betrachtungsweise ist dies die Zeit vom Bereitstellen der Adresse bis zum Liefern der Daten beim Lesen oder bis zum Abnehmen der Daten beim Schreiben. „Schnellspeicher" im eigentlichen Sinne sind RAM-Anordnungen mit besonders kurzer Zugriffszeit. Es liegt nahe, in solchen Speichern häufig gebrauchte Programme und Daten zu halten, so daß man ohne oder mit nur wenigen Wartezuständen darauf zugreifen kann.

In modernen Computersystemen legt man die Schnellzugriffsspeicher typischerweise als *transparente Caches* aus. Dem Begriff „Cache" haben wir bereits in Abschnitt 2.2.4 erklärt. Kennzeichnend für einen *transparenten* Cache ist, daß neben dem eigentlichen Schnellzugriffsspeicher eine Cache-Steuerung (Cache Controller) vorgesehen ist, die bewirkt, daß Teile des Arbeitsspeicherinhaltes im Cache bereitgehalten werden und dort unter den jeweiligen Arbeitsspeicheradressen zugänglich sind.

*Der Zweck:* Die Anordnung aus dem Schnellspeicher (dem Cache) und dem eigentlichen Arbeitsspeicher soll sich aus der Sicht des Prozessors so verhalten wie ein einziger Speicher mit der (geringen) Zugriffszeit des Schnellspeichers und mit der (großen) Speicherkapazität des Arbeitsspeichers.

Die Funktionsweise sei anhand eines Lesezugriffs kurz beschrieben: Der Prozessor greift mit einer Arbeitsspeicheradresse auf den Verbund aus Cache und Arbeitsspeicher zu. Die Cache-Steuerung sieht nach, ob sich die Daten im Cache befinden oder nicht*). Befinden sie sich im Cache, so wird gar nicht erst auf den Arbeitsspeicher zugegriffen. Vielmehr werden die Daten gleich aus dem Cache an den Prozessor geliefert. Nur dann, wenn sie sich nicht im Cache befinden, ist ein Arbeitsspeicherzugriff erforderlich (der deutlich länger dauert). Die Daten werden dabei nicht nur zum Prozessor transportiert, sondern auch noch vorsorglicherweise (in Erwartung kommender weiterer Zugriffe) in den Cache geschafft.

*): hierzu müssen im Cache neben den Daten die jeweils zugehörige Adreßangaben mitgespeichert werden. Neben dem eigentliche Datenspeicher wird deshalb ein sog. Kennzeichnungsspeicher (TAG RAM) vorgesehen.

Kann ein Zugriff aus den Cache bedient werden (das ist dann der Fall, wenn sich der betreffende Speicherinhalt im Cache befindet), so spricht man von einem „Treffer" bzw. *Cache Hit*, andernfalls von einem *Cache Miss*. Der Quotient aus der Anzahl der Treffer und der Anzahl aller Zugriffe in einem gegebenen Zeitintervall heißt „Trefferrate" (Hit Ratio; sprich: Hitt Reeschjo) und ist ein wichtiges Maß für die Effektivität eines transparenten Caches.

$$\text{Trefferrate} = \frac{\text{Anzahl der Treffer}}{\text{Anzahl aller Zugriffe}} \cdot 100\%$$

Alle Funktionen (Adreßvergleich, Nachladen usw.) laufen völlig autonom ab; sie erfordern kein Eingreifen der Software. Die Bezeichnung „Cache" (= Versteck) erklärt sich aus dieser Funktionsweise: es geht darum, einen langsamen Arbeitsspeicher hinter dem Schnellspeicher gleichsam zu verstecken. Die Software merkt davon nichts (deshalb die Bezeichnung „transparent" = durchscheinend).

*Wir merken uns:*

- transparente Caches sind eine sozusagen statistische Angelegenheit - sie wirken nur mit einer gewissen Wahrscheinlichkeit. Der einschlägige Kennwert ist die Trefferrate. Die Trefferraten moderner Caches sind typischerweise größer als 90% (d. h., mehr als 90% aller Zugriffe können aus dem Cache bedient werden).
- transparente Caches haben den Vorteil, daß die Software von der jeweiligen Auslegung des Caches vollkommen unabhängig und auch bei fehlendem Cache lauffähig ist. Das Problem: wegen der statistischen Wirksamkeit (Stichwort: Trefferrate) sind die Programmlaufzeiten nicht exakt vorhersehbar.

# 3.3. Speicherbestückung und Speichererweiterung

Viele Embedded Systems haben eine feste Speicherbestückung. Für PCs und andere „richtige" Computer ist hingegen eine veränderliche Speicherausstattung kennzeichnend*).

*): das gilt auch für manche Steckkarten (z. B. Graphikkarten) und Geräte (z. B. Drucker).

## 3.3.1. Speichererweiterung mit steckbaren Schaltkreisen

Für jeden einzelnen Speicherschaltkreis ist eine Steckfassung vorgesehen. Das war in den 80er Jahren üblich und ist auch noch in Hardware aus den 90er Jahren zu finden. Beispiele:

- auf dem Motherboard zwecks Erweiterung des Arbeitsspeichers mit DRAM-Schaltkreisen (vgl. die Abbildungen 1.6 bis 1.8),
- auf Graphikkarten zwecks Erweiterung des Bildspeichers (typischerweise mit DRAMs; vgl. Position 3 in Abbildung 3.2a),
- zur Erweiterung des externen Caches (mit asynchronen SRAMs).

## 3.3.2. Speichererweiterung mit Steckkarten

Das war bis Anfang der 90er Jahre die einzige Möglichkeit, PCs mit einer (seinerzeit) wirklich überdurchschnittlichen Speicherkapazität auszurüsten (der Platz auf dem Motherboard war beschränkt - also mußte man auf Steckkarten ausweichen). Es wurden Steckkarten für die seinerzeit üblichen Bussysteme (ISA, EISA, MCA) gefertigt, aber auch für herstellerspezifische Lokalbus-Schnittstellen. Manche Speichersteckkarten wurden bereits mit Steckfassungen für Speichermoduln versehen. Auch heutzutage verfallen die Entwickler gelegentlich auf solche Lösungen (vgl. Abbildung 1.24).

## 3.3.3. Speichermoduln

Moderne Speicherschaltkreise sollten so schnell wie möglich arbeiten (geringste Zugriffszeiten, höchste Taktfrequenzen) und so kostengünstig wie möglich herzustellen sein. Um beide Anforderungen zu erfüllen, ist es notwendig, die Gehäuse so klein wie möglich zu bauen.

Für derartige Speicherschaltkreise kann man aber keine kostengünstige Steckfassungen fertigen (die Abstände zwischen den Schaltkreisanschlüssen betragen beispielsweise nur noch 0,5 mm!). Andererseits brauchen herkömmliche Steckkarten ziemlich viel Platz und sind auch recht teuer. Der Ausweg: sehr kleine Steckkarten, die mit Speicherschaltkreisen bestückt sind, mit anderen Worten: Speichermoduln (Abbildung 3.13).

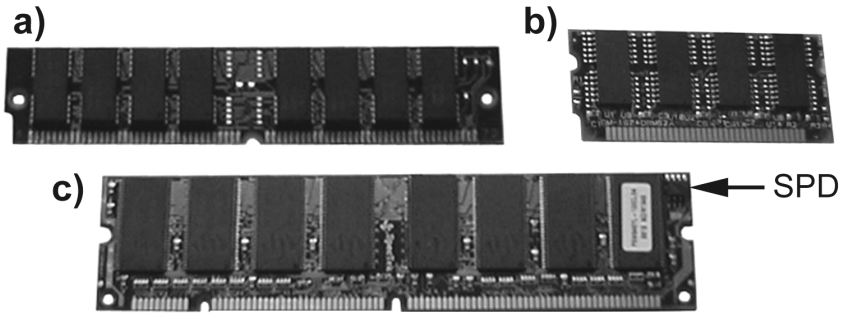



***Abbildung 3.13*** Herkömmliche Speichermoduln (eine kleine Auswahl)

*Erklärung:*
a) - SIMM mit 72 Anschlüssen; b) - SODIMM mit 72 Anschlüssen; c) - DIMM mit 168 Anschlüssen. Der Pfeil zeigt auf einen kleinen Festwertspeicher (einen sog. seriellen EEPROM), der die Konfigurationsangaben enthält (SPD = Serial Presence Detect).

**Grundbegriffe**

*SIMM*
Ein Speichermodul mit *einer* Reihe von Steckkontakten (SIMM = Single In-Line Memory Module). Auch wenn wir auf beiden Seiten der Leiterplatte Steckkontakte erkennen - sie sind zusammengeschaltet (bilden also nur *eine* Kontaktreihe).

*DIMM*
Ein Speichermodul mit 2 Reihen von Steckkontakten (einer auf jeder Seite der Leiterplatte; DIMM = Dual In-Line Memory Module).

*SODIMM* (auch: SO-DIMM)
Ein DIMM in besonders kleiner Bauform (SODIMM = Small Outline Dual In-Line Memory Module).

*RIMM*
Ein Speichermodul, das mit DirectRambus-Schaltkreisen (RDRAMs) bestückt ist (RIMM = Rambus In-Line Memory Module).

*SO-RIMM*
Ein RIMM in besonders kleiner Bauform (SO = Small Outline).

*Parity (Paritätskontrolle)*
Ein Verfahren zur Fehlererkennung. Es erfordert, daß je Byte ein zusätzliches Bit (das Paritätsbit) mitgespeichert wird.

*ECC*
Ein Verfahren zur Fehlererkennung und Fehlerkorrektur (ECC = Error Correcting Code). Es erfordert, daß (wir beschreiben hier die in PCs übliche Auslegung) zu je 8 Bytes ein zusätzliches Korrekturbyte mitgespeichert wird (mit anderen Worten: 64 Datenbits + 8 ECC-Bits).

*Organisationsform*
Diese wird typischerweise als „Speicherkapazität · Datenweg- bzw. Zugriffsbreite (in Bits)" angegeben. Dabei werden Paritäts- und ECC-Bits mitgezählt. Beispiele:

- 1 M · 9,
- 4 M · 36,
- 8 M · 32,
- 8 M · 64,
- 16 M · 72.

In Kurzform spricht man von Moduln · 9, · 32, · 64 usw. Die Organisationsformen · 9, · 36, · 72 sind für Speichersubsysteme mit Paritätskontrolle oder Fehlerkorrektur (ECC) vorgesehen.

*Presence Detect*
Wörtlich übersetzt: Erkennung der Anwesenheit. Viele Speichermoduln haben eigens Anschlüsse, über die der Typ des Moduls abfragbar ist. In der einfachsten Form wird über diese Anschlüsse ein festes Bitmuster geliefert. Moderne Speichermoduln haben hingegen einen seriellen EEPROM, der recht ausführliche Konfigurationsangaben enthält (Serial Presence Detect (SPD); vgl. den Pfeil in Abbildung 3.13). Das betrifft u. a. alle mit Moduln, die mit SDRAMs, DDR-DRAMs und Rambus-DRAMs bestückt sind.

**Die Typenvielfalt der Moduln**
Es gibt eine Vielzahl von Ausführungsformen. Viele Modul-Typen sind in JEDEC-Standards spezifiziert. Hinzu kommen herstellerspezifische Ausführungen (mit denen z. B. in Laserdruckern zu rechnen ist). Im folgenden wollen wir uns auf einen knappen Überblick über jene Ausführungen beschränken, die in PCs des Massenmarktes von vorrangiger Bedeutung sind oder waren.

**SIMMs mit 30 Anschlüssen**
Derartige Speichermoduln waren die ersten, die in PCs eingesetzt wurden. Sie sind auch noch in 486-Systemen zu finden, die Anfang der 90er Jahre auf den Markt kamen. Die Moduln sind typischerweise mit FPM-DRAMs bestückt.

*Grundabmessungen (gerundet):*
Länge = 89 mm, Breite 20...22 mm, Anschlußabstand 2,54 mm.

*Organisationsform*
Man hat seinerzeit die Paritätsprüfung bevorzugt, so daß vorzugsweise Typen in · 9-Organisation gefertigt wurden.

*Real Parity*
Diese Moduln haben in der 9. Bitposition anstelle eines DRAMs einen Paritätsgenerator (preisgünstiger). *Achtung:* es kann sein, daß hochwertige PCs die Paritätsprüfung auf Funktionsfähigkeit testen (im Rahmen des Anfangstests (POST) oder mittels spezifischer Prüfsoftware). Real-Parity-SIMMs sind dann nicht ohne weiteres einsetzbar (Szenarium: beispielsweise ein älterer Netzwerk-Server soll weiter am Leben gehalten werden).

*True Parity*
Dies ist die typische Katalogbezeichnung für Moduln, die in der 9. Bitposition tatsächlich einen DRAM haben.

*Aufbrauchen vorhandener Moduln*
Es gibt Zwischenadapter, um vier derartige Moduln anstelle eines 72-poligen SIMMs einsetzen zu können (Abbildung 3.14). Das Praxisproblem wird aber oft die *Bauhöhe* einer solchen Anordnung sein.

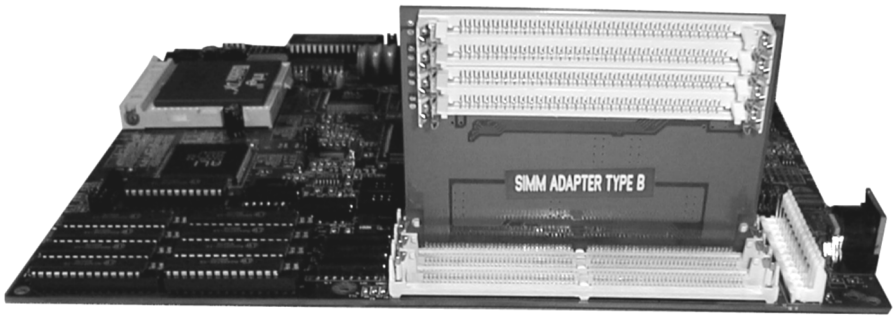



***Abbildung 3.14*** Zwischenadapter für SIMMs

**SIMMs mit 72 Anschlüssen**
Derartige SIMMs bilden in den Organisationsformen · 32 oder · 36 die allgemein übliche Speicherbestückung der PCs, die - rund gerechnet - zwischen 1994 und 1997/98 gefertigt wurden. Die Moduln (vgl. Abbildung 3.13a) sind mit asynchronen DRAMs bestückt.

*Grundabmessungen (gerundet):*
Länge = 108 mm, Breite ca. 26 mm, Anschlußabstand 1,27 mm.

*Organisationsformen*
Wir unterscheiden 72-polige SIMMs in folgenden Organisationsformen:

- · 32-Organisation,
- · 36-Organisation, vorgesehen für Paritätsprüfung (Parity SIMMs),
- · 36-Organisation, vorgesehen für Fehlerkorrektur (ECC SIMMs),
- · 40/36-Organisation, vorgesehen für Fehlerkorrektur (40/36 Bit ECC SIMMs).

SIMMs der ersten drei Organisationsformen haben eine im wesentlichen gleichartige Anschlußbelegung, 40/36-Bit-SIMMs sind hingegen durch eine besondere Anschlußbelegung gekennzeichnet. (In üblichen PCs kommen sie praktisch nicht vor.)

*Spannungsversorgung und Speisespannungskennung*
72-polige SIMMs sind für 5 V oder für 3,3 V Speisespannung ausgelegt. Die jeweilige Speisespannung wird durch die Lage der Kerbe gekennzeichnet (Abbildung 3.15).

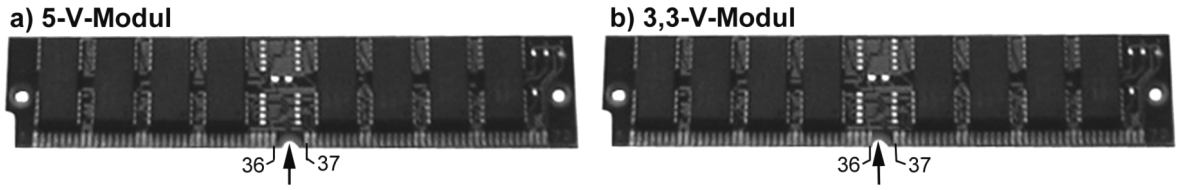



***Abbildung 3.15*** Speisespannungscodierung (Voltage Keying) 72-poliger SIMMs

*Erklärung:*
a) - Kerbe in der Mitte: 5-V-Modul; b) - Kerbe nach links versetzt (näher an Kontakt Nr. 36): 3,3-V-Modul.

*Die Vorzugsbestückung der PCs:* 5-V-Moduln in · 32-Organisation.

*Weitere Unterscheidungen (Auswahl):*

- wieviele Seiten des Speichermoduls sind bestückt? - es gibt ein- und beidseitig bestückte Moduln,
- welche DRAM-Typen sind bestückt? - es gibt FPM- und EDO-SIMMs,
- wieviele Adreßsignale werden zur Zeilen- und Spaltenadressierung verwendet?

**SODIMMs mit 72 Anschlüssen**
Es handelt sich um miniaturisierte Moduln (vgl. Abbildung 3.13b), die an sich den 72-poligen SIMMs entsprechen, deren Anschlüsse aber in 2 Reihen angeordnet sind (eine Reihe auf jeder Seite der Leiterplatte). Solche Moduln werden vor allem in portablen PCs eingesetzt. Die Moduln sind mit asynchronen DRAMs (FPM oder EDO) bestückt und typischerweise für eine Speisespannung von 3,3 V ausgelegt.

*Grundabmessungen (gerundet):*
Länge 57 mm, Höhe ca. 26 mm, Anschlußabstand 1,27 mm.

**DIMMs und SODIMMs zur Erweiterung externer Caches**
Cache-Moduln waren in der zweiten Hälfte der 90er Jahre einige Zeit in Mode. Zwei Arten von Steckverbindern gelten praktisch als Industriestandard (Tabelle 3.4). PCs des Massenmarktes wurden vorwiegend für Cache-Moduln gemäß der (von Intel herausgegebenen) COASt-Spezifikation ausgelegt (COASt = Cache on a Stick). Erweiterbarkeit: typischerweise bis zu 512 kBytes (ein COASt-Modul enthält zumeist 256 kBytes).

| Ausführung | Anschlüsse | Anschlußabstand | Leiterplatten-Abmessungen | Anwendung |
|---|---|---|---|---|
| SODIMM | 144 | 0,8 mm | ≈ 68 mm · 25 mm | portable PCs |
| DIMM (COASt) | 160 | 1,27 mm | ≈ 110 mm · 29...33 mm | ortsfeste PCs |

***Tabelle 3.3*** Cache-Moduln (Übersicht)

**DIMMs mit 168 Anschlüssen**
Seit Ende der 90er Jahre werden die PCs mit derartigen Moduln bestückt (Abbildung 3.16; vgl. auch Abbildung 3.13c). An sich gibt es solche Moduln sowohl mit asynchronen als auch mit synchronen DRAMs. Moderne PCs werden jedoch nahezu ausnahmslos mit *synchronen* DRAMs bestückt (SDRAM-Moduln).

*Grundabmessungen (gerundet):*
Länge = 134 mm, Breite ca. 32 mm, Anschlußabstand 1,27 mm.

*Typische Organisationsformen:* · 64 (keine Fehlerkontrollvorkehrungen) bzw. · 72 (ECC).

*Speisespannung:* 3,3 V.

*Taktfrequenzen:* 66, 100, 133 MHz.

Die Moduln haben einen seriellen EEPROM, der die Konfigurationsdaten enthält (Serial Presence Detect SPD).

*Ungepufferte Moduln (Unbuffered SDRAM DIMMs)*
Die Speicherschaltkreise sind direkt mit den Anschlüssen des Moduls verbunden. Es sind die preisgünstigen Speichermoduln für PCs des Massen-Marktes. Typische Speicherkapazitäten: 8, 16, 32, 64, 128, 256, 512 MBytes (1, 2, 4, 8, 16, 32, 64 M · 64 oder 72).

*Moduln mit Registerpufferung (Registered SDRAM DIMMs)*
Diese Moduln enthalten Pufferregister für die Adreß- und Steuersignale. Hierdurch wird es möglich, die Speicherkapazität je Modul zu erhöhen und die Zeittoleranzen zu vermindern. Die Speicherkapazität kann bis zu 1 GBytes betragen (bis zu 128 M · 64 oder 72).

Diese Moduln können wie herkömmliche ungepufferte Typen eingesetzt werden (sie haben hierzu einen Steuereingang, der vom Motherboard aus entsprechend belegt wird).

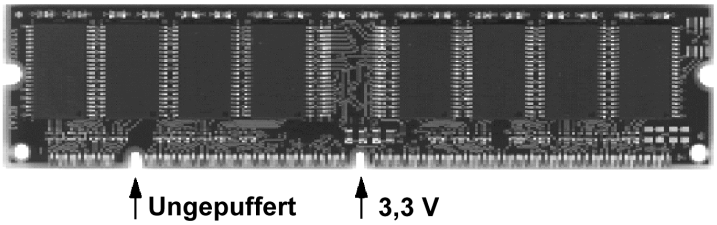



***Abbildung 3.16*** DIMM mit 168 Anschlüssen. Beispiel eines ungepufferten Moduls

**SODIMMs mit 144 Anschlüssen**
Es handelt sich um miniaturisierte Moduln mit einer Zugriffsbreite von 64 oder 72 Bits. Solche Moduln (Abbildung 3.17a) werden in portablen PCs, in Graphik-Subsystemen usw. eingesetzt. Neben SDRAM-Moduln gibt es auch SGRAM-Typen zum Bestücken von Bildspeichern. Abgesehen von der Anschlußbelegung und der Bauform unterscheiden sich SDRAM-SODIMMs praktisch nicht von den ungepufferten 168-poligen SDRAM-DIMMs.

*Speisespannung:* vorzugsweise 3,3 V.

*Grundabmessungen (gerundet):*
Länge 68 mm, Höhe ca. 26 mm, Anschlußabstand 0,8 mm.

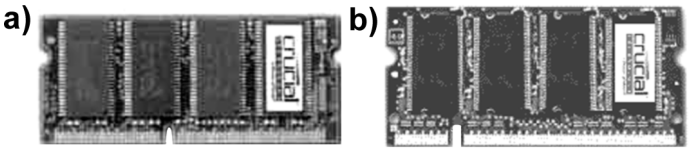



***Abbildung 3.17*** SODIMMs. a) SDR mit 144 Anschlüssen; b) DDR mit 240 Anschlüssen (Crucial Technology)

**DDR-SDRAM-DIMMs mit 184 Anschlüssen**
Es gibt gepufferte (Registered) und ungepufferte (Unbuffered) Typen.

*Grundabmessungen (gerundet):*
Länge 134 mm, Höhe ca. 32 mm, Anschlußabstand 1,27 mm (Abbildung 3.18). Die Außenabmessungen entsprechen denen der 168-poligen DIMMs, die Kontaktreihen sind aber länger.

*Typische Organisationsformen:* · 64 (keine Fehlerkontrollvorkehrungen) bzw. · 72 (ECC).

*Speisespannung:* 2,5 V.

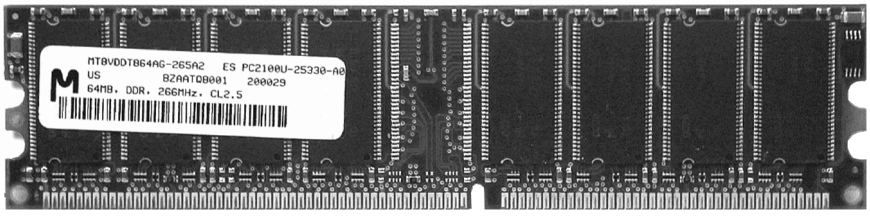



***Abbildung 3.18*** DIMMs mit 184 oder 240 Anschlüssen (DDR/DDR2)

**DDR2-SDRAM-DIMMs mit 240 Anschlüssen**
Es gibt gepufferte (Registered) und ungepufferte (Unbuffered) Typen.

*Grundabmessungen (gerundet):*
Länge 134 mm, Höhe ca. 30 mm, Anschlußabstand 1 mm. (Außenabmessungen entsprechen denen der 184-poligen DIMMs, aber mehr Kontakte in engerem Abstand. Vgl. Abbildung 3.18.)

*Typische Organisationsformen:* · 64 (keine Fehlerkontrollvorkehrungen) bzw. · 72 (ECC).

*Speisespannung:* 1,8 V.

**DDR-SDRAM-SODIMMs mit 240 Anschlüssen**
Diese Moduln sehen ähnlich aus wie die 144-poligen, sind aber etwas länger (vgl. Abbildung 3.17b). Ansonsten entsprechen sie weitgehend den 184-poligen DDR-SDRAM-DIMMs.

*Grundabmessungen (gerundet):*
Länge = 76,2 mm, Höhe ca. 26 oder 29 mm., Anschlußabstand 0,8 mm.

**AIMMs**
AIMM = AGP Inline Memory Module. Diese Moduln (Abbildung 3.19) haben 66 Anschlüsse und passen in AGP-Slots.

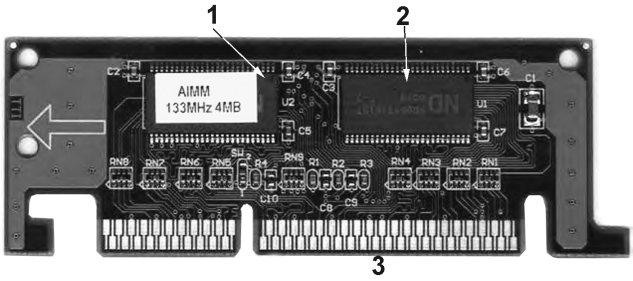



***Abbildung 3.19*** AIMM (Spezifikation: Intel, Ausführungsbeispiel: Asus). 1, 2 - Speicherschaltkreise; 3 - AGP-Steckkontakte

*Grundabmessungen (gerundet):*
Länge = 90 mm, Höhe ca. 17 mm, Anschlußabstand 1 mm (AGP-Kontaktschema; vgl. Abbildung 4.15).

*Organisationsform und Speicherkapazität:*
1 M · 32 = 4 MBytes. Bestückung mit 133-MHz-SDRAMs.

*Anwendung:*
Auf Motherboards mit eingebauter Videohardware und AGP-Slot (Stichwort: Three Load AGP; vgl. Abschnitt 4.4.4). Das Modul wird in den (an sich ungenutzten) AGP-Slot gesteckt und stellt so einen Bildpuffer-Cache von 4 MBytes bereit. Der Trick: wenn ein solches Modul in einem AGP-Slot steckt, so werden die auf dem Motherboard vorhandenen AGP-Signalwege nicht als AGP-Interface betrieben, sondern als SDRAM-Schnittstelle.

*Hinweis:*
Ein AIMM funktioniert nicht in allen AGP-Slots; seine Nutzung muß vom Schaltkreissatz untersützt werden.

**RIMMs**
Diese Moduln (Abbildung 3.20) sind mit Direct-Rambus-DRAMs (RDRAMs) bestückt. Sie haben 184 Anschlüsse. Der Betrieb mit extremen Taktfrequenzen hat zur Folge, daß die Speicherschaltkreise richtig heiß werden. Deshalb haben Rambus-Moduln eine wärmeableitende Metallverkapselung (Heat Spreader).

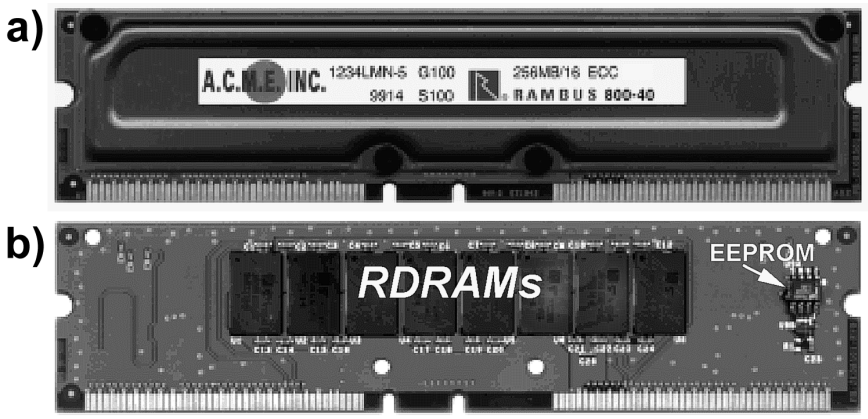



***Abbildung 3.20*** Rambus-Modul (RIMM). a) mit, b) ohne Heat Spreader

*Grundabmessungen (gerundet):*
Länge 134 mm, Höhe ca. 32 mm, Anschlußabstand 1 mm.

*Typische Organisationsformen:* · 16 und · 18. Die beiden zusätzlichen Bits waren ursprünglich zur Paritätskontrolle gedacht, können aber im Systementwurf freizügig verwendet werden (die typische Nutzung: als Fehlerkorrekturbits (ECC-Bits)).

*Typische Speicherkapazitäten:* 64, 96, 128, 256, 512 usw. MBytes in · 16-Organisation; 72, 108, 144, 288, 576 usw. MBytes in · 18-Organisation (wenn alle Bits zur Datenspeicherung ausgenutzt werden).

*Speisespannung:* 2,5 V (sowie gesonderte Versorgung für CMOS-Interfaceschaltungen und SPD-EEPROM).

**SO-RIMMs**
Diese miniaturisierten Moduln (Abbildung 3.21) haben 160 Anschlüsse.

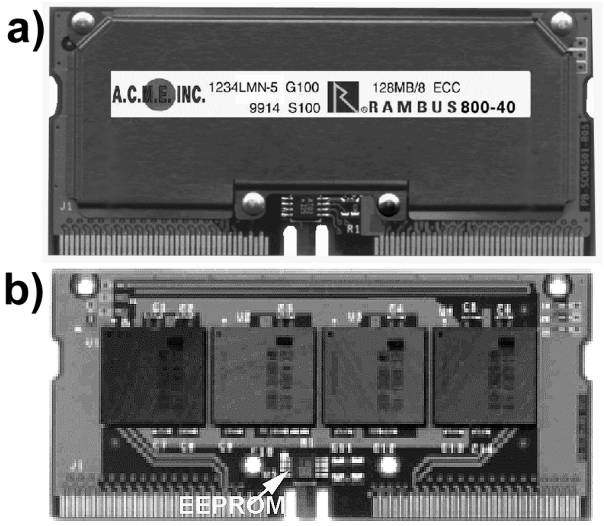



***Abbildung 3.21*** SO-RIMM. a) mit, b) ohne Heat Spreader

*Grundabmessungen (gerundet):*
Länge = 67 mm, Höhe ca. 32 mm, Anschlußabstand 0,65 mm.

*Typische Organisationsformen:* · 16 bzw. · 18.

*Typische Speicherkapazitäten:* 16, 32, 64, 96, 128 usw. MBytes in · 16-Organisation; 18, 36, 72, 108, 144 usw. MBytes usw. in · 18-Organisation (wenn alle Bits zur Datenspeicherung ausgenutzt werden).

*Speisespannung:* 2,5 V (sowie gesonderte Versorgung für CMOS-Interfaceschaltungen und SPD-EEPROM).

**32/36-Bit-RIMMs**
Diese Moduln sind zum Betrieb an zwei unabhängigen Rambus-Kanälen vorgesehen. Sie haben 232 Anschlüsse. Die Außenabmessungen entsprechen denen der „gewöhnlichen" RIMMs.

**Rambus-Blindmoduln (RIMM/SO-RIMM Continuity Modules)**
In nicht bestückte Rambus-Steckfassungen sind Blindmoduln einzusetzen. Das sind Leiterplatten in RIMM-oder SO-RIMM-Form ohne Schaltkreise, die einfach die ankommenden mit den abgehenden Signalen verbinden (Abbildung 3.22).

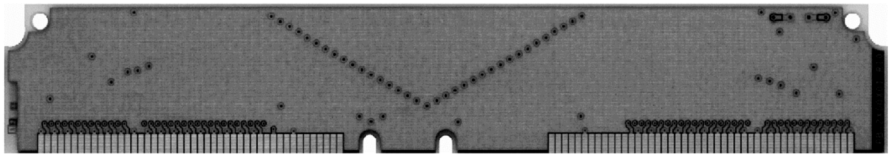


***Abbildung 3.22*** Rambus-Blindmodul (RIMM Continuity Module)

**Zur Bezeichnung moderner Speichermoduln im PC-Bereich**

*SDRAM-Moduln (SDR)*
Die Bezeichnung beginnt mit den Buchstaben *PC*, denen eine Frequenzangabe folgt. Diese kennzeichnet die maximale Taktfrequenz des Speicherbus. Folgende Typen sind üblich:

- PC66: 66 MHz Bustakt,
- PC100: 100 MHz Bustakt,
- PC133: 133 MHz Bustakt.

*DDR-Moduln*
Die Bezeichnung beginnt mit den Buchstaben *PC*, denen eine Frequenzangabe folgt. Diese kennzeichnet die Gesamt-Datenrate in MBytes/s (über den 8 Bytes breiten Datenbus). Rechengang: Gesamtdatenrate = Taktfrequenz · 16 oder Datenratenkennwert (System Bus Speed) des Speicherschaltkreises · 8.

*DDR:*

- PC-1600: Bustakt 100 MHz (Speicherschaltkreise DDR-200),
- PC-2100: Bustakt 133 MHz (Speicherschaltkreise DDR-266),
- PC-2700: Bustakt 166 MHz (Speicherschaltkreise DDR-333),
- PC-3200: Bustakt 200 MHz (Speicherschaltkreise DDR-400).

*DDR2:*

- PC2-3200: Bustakt 200 MHz (Speicherschaltkreise DDR2-400),
- PC2-4300: Bustakt 266 MHz (Speicherschaltkreise DDR2-533),
- PC2-5300: Bustakt 333 MHz (Speicherschaltkreise DDR2-667).

*Rambus-Moduln*
Die Bezeichnung beginnt mit den Buchstaben *RIMM*, denen eine Frequenzangabe folgt. Diese kennzeichnet die Gesamt-Datenrate in MBytes/s (über die jeweilige Anzahl Rambus-Kanäle). Rechengang: Datenratenkennwert (I/O Frequency) des Speicherschaltkreises · 2 · Anzahl der Rambus-Kanäle. Beispiele:

- RIMM1600: 1,6 GBytes/s (Speicherschaltkreise PC800, 1 Rambus-Kanal),
- RIMM2100: 2,1 GBytes/s (Speicherschaltkreise PC1066, 1 Rambus-Kanal),
- RIMM4200: 4,2 GBytes/s (Speicherschaltkreise PC1066, 2 Rambus-Kanäle (32/36-Bit-RIMMs)),
- RIMM4800: 4,8 GBytes/s (Speicherschaltkreise PC1200, 2 Rambus-Kanäle (32/36-Bit-RIMMs)),
- RIMM9600: 9,6 GBytes/s (Speicherschaltkreise PC1200, 4 Rambus-Kanäle (64/72-Bit-RIMMs (in Planung)).

*Wir merken uns:*
SDRAM-Moduln haben dreistellige PC-Kennummern (Bustakt), DDR-Moduln haben vierstellige PC-Kennummern, Rambus-Moduln haben vierstellige RIMM-Kennumern (Gesamtdatenrate in MBytes/s).

## 3.3.4. Grundlagen der Speicherbestückung

Welche Speichermoduln passen auf das Motherboard? Welche Spitzfindigkeiten sind zu beachten? Woran könnte es liegen, wenn es offensichtlich nicht funktioniert? - Das sind gelegentlich knifflige Fragen; die Speicherbestückung ist auch heutzutage nicht immer eine Sache des „Plug and Play".

*Der ultimative Praxistip:* Das Handbuch des betreffenden Motherboards. Falls nicht zu beschaffen, die Dokumentation des jeweiligen Schaltkreissatzes. Suchet, so werdet Ihr finden (Internet)...

*Was meistens auf Anhieb funktioniert:*

- ein (halbwegs) modernes Motherboard (ab Ende der 90er Jahre),
- SDR- oder DDR-SDRAM,
- ein einziges Speichermodul in einer x-beliebigen Fassung.

*Wir merken uns:*
Moderne Motherboards können an sich in beliebiger Weise bestückt werden (keine ZwangsReihenfolge in der Belegung der Fassungen, kein Zwang zur Einheitlichkeit der Speichermoduln). Es kann aber sein, daß eine solche Bestückung nicht optimal ist (Datenrate zu gering, installierte Speicherkapazität nicht voll ausgenutzt).

*Eine kleine Auswahl typischer Problemstellen:*

- 64 oder 72 Bits? - Manche Boards unterstützen keine Fehlerkorrektur. Dann nur 64-Bit-Moduln einsetzen.
- gepuffert oder ungepuffert? - Die für den Massenmarkt bestimmten Boards unterstützen oft nur ungepufferte Moduln.
- ein- oder zweiseitig bestückte Moduln? Kann sein, daß das Board nur einseitig bestückte (Single Sided) Moduln unterstützt. Manchmal kommt es auf die Steckfassung an. Beispiel: zwei Fassungen können nur einseitig bestückte Moduln aufnehmen. In die dritte Fassung paßt aber auch ein zweiseitig bestücktes (Dual Sided) Modul.
- Leistungseinschränkungen? - Manche Boards sind in manchen Betriebsarten auf bestimmte Datenraten beschränkt. Beispiel: bei Nutzung von On-Board-VGA nur maximal 333 MHz, sonst 400 MHz. Wird On-Board-VGA verwendet, lohnt es sich nicht, die schnelleren (= teureren) Moduln zu bestücken.
- ältere PCs: bestimmte Belegungsreihenfolge der Steckfassungen, mehrere gleichartige Moduln gemäß Datenwegbreite (z . B. 4 Stück zu 9 Bits für einen 32-Bit-Datenweg, 2 Stück zu 32 Bits für einen 64-Bit-Datenweg usw.),
- Rambus-Subsysteme mit zwei Kanälen: paarweise Bestückung beider Kanäle. Blindmoduln nicht vergessen!

*Zwei DDR-Speicherkanäle (Dual Channel Architecture)*
Eine sozusagen rücksichtslose Bestückung ist auch hier möglich, nur kommt dann das Leistungsvermögen des Speichersubsystems nicht in vollem Maße zur Wirkung. Schlimmstenfalls bringt es nicht mehr Systemleistung als ein einziger herkömmlicher 64-Bit-Kanal.

*Wir merken uns:*
Ein Dual-Channel-Speichersubsystem erreicht nur dann seine volle Leistung (Zugriffszeiten, Datenraten), wenn es wenigstens paarweise mit gleichartigen Moduln bestückt wird. Am besten: Vollbestückung mit gleichartigen Moduln.

*Grenzen der Erweiterbarkeit*
Es ist klar, daß sich eine Speicherausstattung nicht unbegrenzt erweitern läßt. Naheliegenderweise bestimmen folgende Gegebenheiten die jeweils größte installierbare Speicherkapazität:

- die Anzahl und Art der verfügbaren Steckpositionen auf dem Motherboard,
- die maximale Speicherkapazität der jeweils bestückbaren Schaltkreise bzw. Moduln.

Zudem kommt es gelegentlich auf die Auslegung der Speichersteuerung und Speicheradressierung an. Das kann vor allem dann zum Problem werden, wenn ein älteres Motherboard mit moderneren Speichermoduln bestückt werden soll (ältere Steuerschaltkreise unterstützen nicht immer die auf heutigen Moduln verfügbaren Speicherkapazitäten).

# 3.4. Die Speicherausstattung üblicher PCs

## 3.4.1. ROM

Auf herkömmlichen AT-kompatiblen Motherboards hatte man üblicherweise 128 kBytes ROM vorgesehen. Damit kommt man heutzutage natürlich nicht mehr aus. Moderne PCs sind typischerweise mit 512 kBytes bzw. 1 MBytes (Flash-) ROM ausgestattet. Steckkarten können zusätzliche ROM-Schaltkreise enthalten.

*Zur Organisation der ROM-Anordnung auf dem Motherboard*
Die Auslegungen unterscheiden sich im Typ der ROMs und in der Organisation der Datenwege. Herkömmliche (ältere) PCs haben „gewöhnliche" (nicht änderbare) ROMs (zumeist auf Steckfassung). Es ist oft nur ein ROM vorhanden (zwei bei 16-Bit-Datenweg). Auf (sehr) alten Motherboards können hingegen bis zu 4 Steckfassungen vorgesehen sein. Update: durch Auswechseln. Moderne PCs haben typischerweise nur einen (fest eingelöteten) Flash-ROM. Update: mittels Hilfssoftware.

## 3.4.2. RAM mit Datenerhalt (CMOS-RAM)

Üblicherweise ist der Funktionsumfang des Schaltkreises 1287 oder MC146818 vorgesehen (entweder als wirklicher Chip oder innerhalb eines hochintegrierten Steuerschaltkreises). Im 1287 sind 64 Bytes adressierbar. Davon stehen 50 für Konfigurationsangaben (Setup) zur Verfügung.

## 3.4.3. Caches

Die Ausstattung reicht von 16...32 kBytes (PCs mit 386-Prozessor, Anfang der 90er Jahre) bis zu mehreren MBytes. Moderne PCs des Massen-Marktes haben Caches von einigen hundert kBytes. Die Cache-Kapazität ist typischerweise auf zwei L1-Caches und einen L2-Cache aufgeteilt. Einige Hochleistungsprozessoren (z. B. Xeon und Itanium) haben darüber hinaus einen L3-Cache. Manchmal baut man besonders kostenoptimierte Prozessoren ohne L2-Cache (z. B. einige Celeron-Typen). Man verläßt sich dann auf die L1-Caches und auf das Leistungsvermögen des Arbeitsspeichersubsystems.

## 3.4.4. RAMs

Im PC sind RAMs sowohl für den eigentlichen Arbeitsspeicher als auch für den Bildspeicher des Videoadapters vorgesehen. Zum wünschenswerten Umfang der Speicherausstattung gilt die alte Programmierer-Weisheit, daß der Speicher immer zu klein ist. Eine unbeschränkte Erweiterung ist naturgemäß nicht möglich. Aber auch wenn wir nur bezahlbare Speichergrößen betrachten, stellt uns der IBM-kompatible PC vor grundsätzliche Probleme.

**Der einheitliche lineare Adreßraum**
Dieser Begriff bezeichnet ein einfaches und elegantes Prinzip: in der Prozessor-Architektur werden für die Speicheradresse so viele Bits vorgesehen, daß der Adreßraum technisch nicht „ausschöpfbar" ist (der gesamte Adreßraum muß eine viel größere Speicherkapazität ansprechen können als technisch überhaupt realisiert werden kann). Sonderfunktionen, wie z. B. die Bildspeicherung für die Bildschirm-Anzeige, werden in einen beliebigen Ausschnitt dieses Adreßraumes gelegt. (Ist der Adreßraum groß genug, so hat man hierbei weitgehende Narrenfreiheit.)

Heutzutage bildet die 32-Bit-Adresse den Stand der Technik. Die damit adressierbaren 4 GBytes (über vier Milliarden Bytes!) erschienen vor wenigen Jahren tatsächlich (als RAM) kaum realisierbar. Heute kann man solche Speicher ohne weiteres bauen - und auch die Kosten kommen mit der Zeit in annehmbare Größenordnungen.

(Übung: Sehen Sie nach, was die gängigen Speichermoduln momentan kosten (Schaufenster der Händler, Kataloge, Anzeigen, Internet) und berechnen Sie damit die Kosten einer Speicherausstattung von 4 GBytes.)

*Hinweis:*
Intel hat die IA-32-Prozesoren auf 36-Bit-Adressierung eweitert. Das einzelne Anwendungsprogramm kann aber nach wie vor nur mit 32-Bit-Adressen arbeiten. Die Nutzung der 36-Bit-Adressierung ist Sache der Systemsoftware.

**Die Erblast des IBM-PCs**
Hätte IBM 1981 nur eine 32-Bit-Adresse vorgesehen! Aber damals erschien schon der Adreßraum des 8086/8088-Prozessors (20 Adreßbits ≙ 1 MBytes) als extrem groß*), und man hat sowohl die Hard- als auch die Software auf diesen Adreßraum hin ausgelegt. Dabei wurde für den Bildspeicher des Videoadapters noch ein Ausschnitt von 64 kBytes fest zugeordnet - was seinerzeit auch sehr viel war. Abzüglich der Speicherkapazitäten für Bildspeicher, für ROMs und für Erweiterungs-Reserven (Steckkarten) verblieben von dem einen Megabyte schließlich maximal 640 kBytes, um Daten und Programme aufzunehmen. Und diese Beschränkungen sorgen auch heute noch ab und an für Ärger...

*): die meisten der seinerzeit gängigen „persönlichen" Computer hatten einen 16-Bit-Adreßbus und konnten somit nur 64 kBytes adressieren (und die waren meistens nicht einmal voll bestückt...).

**Die Entwicklungstendenz**
Noch längere Adressen. Typische Adreßlängen: 36, 40, 48 und 64 Bits (vgl. Abschnitt 2.2.1).

Der große praktische Vorteil des linearen Adreßraumes: es kostet nicht allzu viel, die Adressierungsschaltungen und Adreßwege entsprechend „breit" auszulegen. Dieser vergleichsweise geringe Mehraufwand ermöglicht es, *jede beliebige Speicherkapazität ohne weiteres zu installieren*, sobald man sie bezahlen kann.

Seit Beginn der Computertechnik hat aber das Prinzip der Sparasamkeit gegolten. Die Adressierungsschaltungen und Adreßwege werden nur für so viele Adreßbits ausgelegt, wie dies zum Adressieren technisch sinnfälliger Speicherkapazitäten erforderlich ist. (Wenn beispielsweise ein Speicher von 32 kBytes das obere Ende des gerade noch Bezahlbaren darstellt, so genügt eine 15-Bit-Adresse.) Es sei allerdings bemerkt, daß die Ingenieure der 50er...70er Jahre gar nicht anders konnten - Computer mit z. B. 32-Bit-Adressierung wären seinerzeit viel zu groß und zu teuer geraten.

**Installierbare Speicherkapazität**
Das vom „klassischen" PC vorgegebene Megabyte gehört seit Ende der 80er Jahre zum Stand der Technik. In den 90er Jahren nahm die Grundausstattung an Speicherkapazität auch bei den PCs des Massenmarktes nach und nach zu: 4, 16, 32, 64, 128 usw. MBytes (Abbildung 3.23). Auf vielen der heutigen Motherboards kann man mehrere GBytes installieren.

*Hinweis:*
Die Motherboards der Mainstream-PCs unterstützen nicht immer den gesamten 32-Bit-Adreßraum (4 GBytes). Etwas ältere Modelle können oft nur mit einigen hundert MBytes bestückt werden. Computer des oberen Leistungbercichs (Workstations, Server) kann man hingegen mit Arbeitsspeicherkapazitäten von mehreren GBytes ausstatten (bis hin zu 64 GBytes, also bis zur Obergrenze dessen, was sich mit der 36-Bit-Adresse der Intel-Prozessoren (s. oben) ansprechen läßt).

Der Bildspeicher der Graphikkarte ist zwar über einen Ausschnitt des Speicheradreßraumes zugänglich, aber aus technischer Sicht vom Arbeitsspeicher völlig getrennt*). Die Entwicklung begann mit Speicherkapazitäten von 256 kBytes, 512 kBytes, 1 MBytes, 2 MBytes, 4 MBytes usw. Aber vor allem dann, wenn es ums Spielen geht (3D-Karten), ist mit 128 oder 256 MBytes sicherlich noch nicht die Grenze erreicht...

*): es gibt auch Graphik-Subsysteme, die einen besonderen Bereich des Arbeitsspeichers als Bildspeicher nutzen.

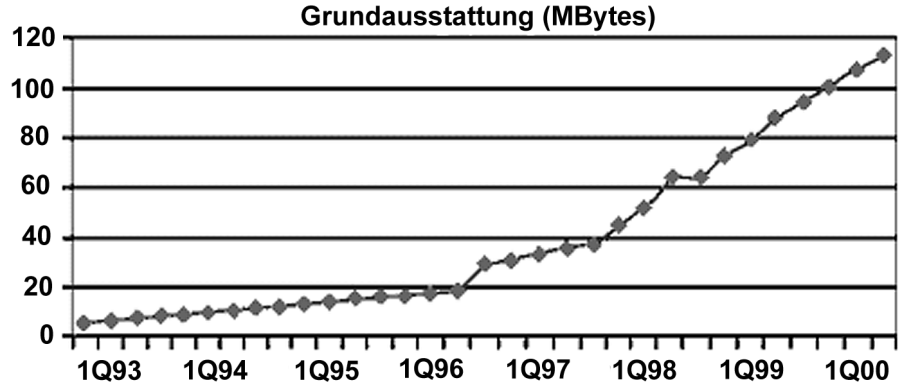



***Abbildung 3.23*** Die Entwicklung der Speicherausstattung von PCs des Massenmarktes - ein historischer Rückblick (Crucial Technology)

*Hinweis:*
Vergleichen Sie die deutlich erkennbaren Sprünge mit den Zeitpunkten, zu denen jeweils neue Windows-Betriebssysteme eingeführt wurden. Ersichtlich ist auch, daß es immer weiter bergauf geht...

**Nutzung der installierten Speicherkapazität**
Die installierte Speicherkapazität voll zu nutzen, hängt weniger von der Hardware als vielmehr von der Software, namentlich vom Betriebssystem ab. Moderne Betriebssysteme verwenden wenigstens 32-Bit-Adressen, so daß praktisch ein beliebig großer Speicher - wie oben beschrieben - ohne weiteres nutzbar ist[*)].

Wenn wir es hingegen mit dem „uralten" Betriebssystem DOS zu tun haben, so helfen nur noch Tricks.

*): bei Windows 3.x/95/98 ist mit Einschränkungen zu rechnen.

## 3.4.5. Die Speicherbereiche des kompatiblen PCs - ein Überblick

In Abbildung 3.24 sind die verschiedenen Speicherbereiche und Speicherarten zusammengestellt, die wir in kompatiblen PCs vorfinden. Diese Aufteilung des Speicheradreßraums - mit ihren Besonderheiten und Spitzfindigkeiten - ist nur dann von Bedeutung, wenn der PC unter Steuerung eines DOS-Betriebssystems „gefahren" wird (wir können aber durchaus auch noch unter Windows 95/98 davon betroffen sein).

Eine solche Speicheraufteilung ist an sich nichts Ungewöhnliches. Jeder Entwickler, der mit Mikrocontrollern und Mikroprozessoren arbeitet, geht so vor: er besieht sich den verfügbaren Speicheradreßraum und überlegt sich eine Aufteilung in die Bereiche, die er benötigt (für ROM, Arbeitsspeicher, Bildspeicher, Speicherplätze in E-A-Einrichtungen usw.). Die Besonderheit der Aufteilung gemäß Abbildung 3.24 besteht nur darin, daß sie infolge der massenhaften Verbreitung der PCs buchstäblich über Jahrzehnte hinweg zum Industriestandard geworden ist - und uns auch heutzutage gelegentlich noch beschäftigt.

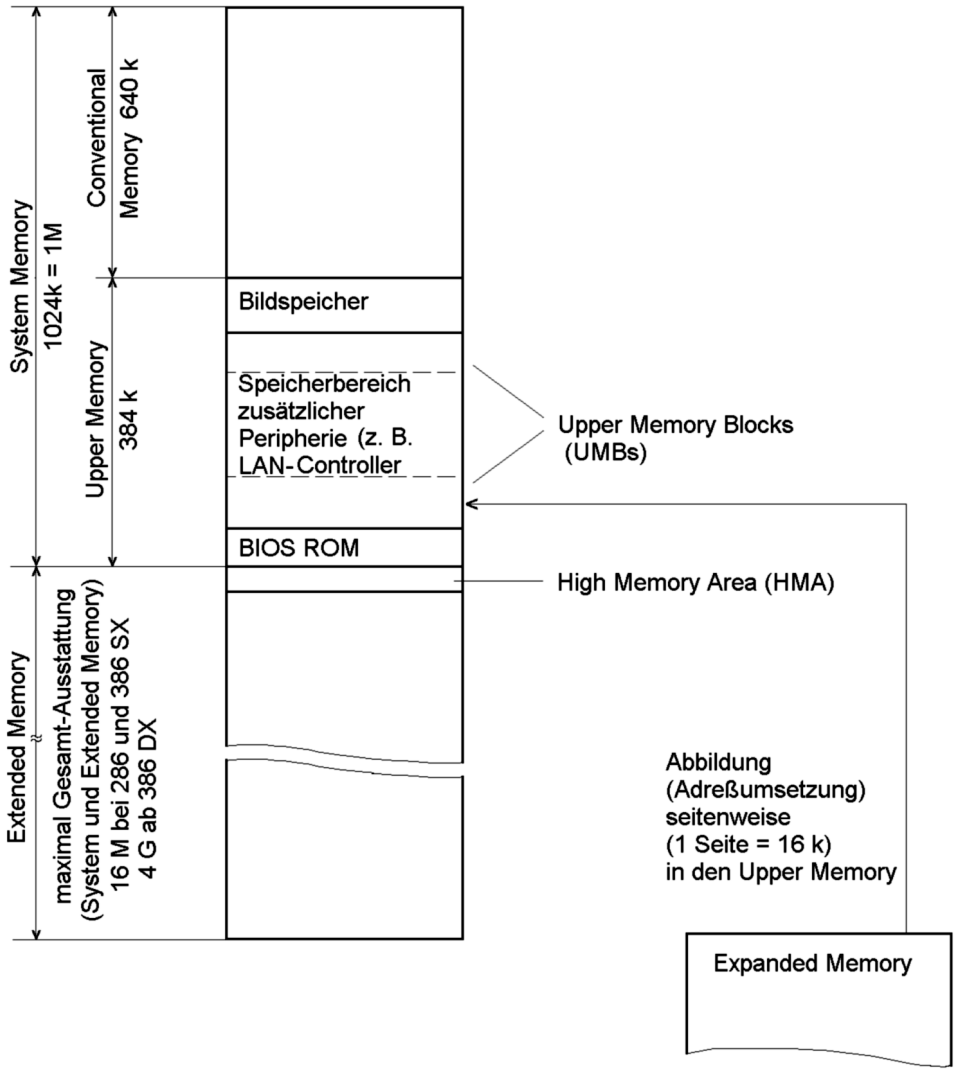



***Abbildung 3.24*** Speicherbereiche des kompatiblen PCs

**System Memory**
Das ist die Allgemeinbezeichnung des „untersten" Megabytes. Es entspricht genau dem Speicheradreßraum der ursprünglichen x86-Prozessoren.

**Conventional Memory**
Die untersten 640 kBytes wurden als „eigentlicher" Programmspeicher reserviert.

**Upper Memory**
Dieser „obere" Speicher ist praktisch der Rest: 1 MBytes - 640 kBytes = 384 kBytes). Dieser Teil des Adreßraumes ist mit ROMs, mit dem Bildspeicher des Videoadapters und manchmal mit Speicherbereichen anderer Steckkarten belegt (z. B. mit Pufferbereichen von Netzwerkadaptern).

**Upper Memory Blocks (UMBs)**
Die Bereiche, die im Upper Memory nicht von ROMs, Steckkarten usw. belegt sind, können als Arbeitsspeicher genutzt werden.

**Extended Memory**
Das ist die Allgemeinbezeichnung des Speicheradreßraums oberhalb des 1. MBytes. Er ist nur von 286- und IA-32-Prozessoren aus direkt zugänglich.

**High Memory Area (HMA)**
Das ist ein Bereich oberhalb des 1. MBytes, zu dem 286- und IA-32-Prozessoren auch in der elementaren Betriebsart (dem Realmodus) noch zugreifen können. Dessen Größe: 65 520 Bytes (64 kBytes - 16 Bytes).

**Expanded Memory**
Auch als es noch keine 286- und IA-32-Prozessoren gab, haben ehrgeizige Entwickler versucht, die Speicherkapazität über 1 MBytes hinaus zu erhöhen. Das wurde durch ein hardwareseitiges Adreßumsetzungsverfahren gelöst. Der hierüber zugängliche Speicher heißt Expanded Memory.

**Hochladen**
Eine Wissenschaft für sich, deren Ziel darin besteht, den eigentlichen Anwendungsprogrammen möglichst viel Platz im Conventional Memory zur Verfügung zu stellen. Hierzu werden Systemprogramme, Gerätetreiber usw. in alle möglichen Ecken (UMBs, HMA) gestopft.

# 4. Bussysteme

Das grobe Blockschaltbild nahezu jeder Anordnung, die auf Mikroprozessoren beruht, zeigt einen *Systembus*, der alle Funktionseinheiten (Prozessor(en), Speicher, E-A-Einrichtungen usw.) untereinander verbindet (vgl. Abbildung 1.2). Er dient dazu, Daten zu transportieren und Zugriffe auf Speicher- und E-A-Einrichtungen zu ermöglichen.

*Wodurch ist ein Bussystem gekennzeichnet?*

- durch die Struktur, d. h. die Anzahl und Art der vorgesehenen Signalwege,
- durch die Funktionsprinzipien, d. h. die vorgesehenen Signalspiele, Adressierungsverfahren usw.
- durch die elektrische Auslegung (Signalpegel, Signalflanken, Anforderungen an die Buskoppelstufen),
- durch die mechanische Auslegung (Steckverbinder, Abmessungen von Steckkarten, zulässige Leitungslängen usw.),
- durch Bedingungen der „Infrastruktur" (Stromversorgung, Kühlung usw.).

Zu den besonders augenfälligen Merkmalen gehören die Zugriffsbreite (so spricht man von einem 8-Bit-Bus, einem 16-Bit-Bus usw.) und die maximal erreichbare Datenrate (die in der Werbung gern herausgestellt wird). Merkmale, die sozusagen auf den zweiten Blick auffallen, sind u. a. das Adressierungsvermögen*), die Auslegung der Steckverbinder (Slots) und der Steckkarten sowie die Anforderungen an die Stromversorgung. Um ein Bussystem genau zu dokumentieren, müssen aber *sämtliche* Merkmale bis in die Einzelheiten dargestellt sein. Sie werden es sich denken können, daß dies bei einem modernen Bussystem nicht mit ein paar Seiten abgetan sein wird.

*): es ist offensichtlich ein Unterschied, ob ein Bussystem 16, 20, 24, 32 oder 64 Bits breite Adressen übertragen kann...

## 4.1. Die Bussysteme der PCs

Die ersten „persönlichen" Computer hatten tatsächlich ein universelles Bussystem ähnlich Abbildung 1.2. Oft hat man es sich ganz einfach gemacht und den Bus des jeweiligen Prozessors als Systembus verwendet. In modernen PCs hat man hingegen den einzigen Bus des groben Blockschaltbildes in mehrere Bussysteme aufgelöst (vgl. Abbildung 1.4). Der Praktiker interessiert sich naturgemäß besonders für jene Bussysteme, an die *Steckkarten* angeschlossen werden können (Abbildung 4.1). Während alle anderen Informationswege, z. B. Direktverbindungen zwischen Prozessor und Speicher, ohne weiteres beliebig ausgelegt werden können (jeder Entwickler hat hier - zumindest vom Grundsatz her - freie Gestaltungsmöglichkeiten, die er nutzen kann, um seine Leistungs- bzw. Kostenziele zu erreichen), ist es wichtig, daß ein Systembus allgemein anerkannten *Standards* genügt. Denn nur so ist es möglich, das Angebot an Steckkarten zu nutzen, den PC zu erweitern, zu modifizieren usw. In der Entwicklungsgeschichte der IBM-kompatiblen PCs haben mehrere Standards Bedeutung erlangt (Tabelle 4.1):

1. der ISA-Bus bzw. AT-Bus (ISA=Industry Standard Architecture),

2. der Mikrokanal (MCA = Microchannel Architecture),
3. der EISA-Bus (EISA = Extended Industry Standard Architecture),
4. verschiedene Lokalbussysteme.

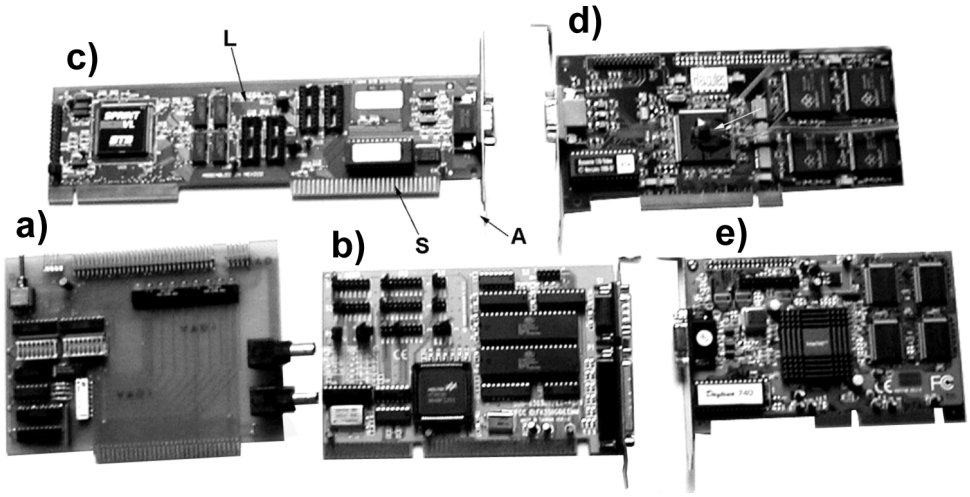



***Abbildung 4.1*** Steckkarten für PC-Bussysteme

*Erklärung:*
a) - ISA, 8 Bits; b) - ISA , 16 Bits; c) - VESA; d) - PCI; e) - AGP.

*Um welche Steckkarten handelt es sich?*
a) - ein Diagnoseadapter; b) - eine Schnittstellenkarte; c), d), e) - Graphikkarten.

*Zum grundsätzlichen Aufbau der Steckkarten:*
L - Leiterplatte mit Schaltkreisen; S - (direkter) Slot-Steckverbinder; A - Slot-Abdeckblech (Bracket). Am Slot-Abdeckblech sind die Steckverbinder für ggf. anzuschließende Interfacekabel montiert. (Der Diagnoseadapter hat kein Abdeckblech, da er nicht zum Verbleib im PC bestimmt ist.) Achten Sie auch darauf, daß die PCI- und AGP-Karten „anders herum" bestückt sind als die ISA- und VESA-Karten.

**Zum Aufbau eines Systembus**
Im Grunde ist ein Systembus genauso aufgebaut wie ein Prozessor-Bus (den wir in Abschnitt 2.2. kennengelernt haben). Abbildung 4.2 zeigt den grundsätzlichen Aufbau eines Systembus. Die Ähnlichkeit zum Prozessorbus wird klar, wenn Sie Abbildung 4.2 mit den Abbildungen 2.1 und 2.2 vergleichen. Auch in funktioneller Hinsicht gibt es keine grundsätzlichen Unterschiede.

| Merkmale | ISA | | EISA | MCA | Lokalbus | |
|---|---|---|---|---|---|---|
| | XT-Bus | AT-Bus | | | VESA | PCI |
| Datenleitungen | 8 | 16 | 32 | 32 | 32 | 32 oder 64 (Multiplex-betrieb) |
| Adreßleitungen | 20 | 24 | 32 | 32 | 32 | |
| maximale Datenrate (MBytes/s)*) | 2 | 8 | > 30 | > 30 | 130 | 132, 264 oder 528 |
| Bus-Takt (MHz) | 4,77 | 8...10 (typisch: 8,33) | 8,33 | 10 | Prozessortakt | 33 oder 66 |

*): gerundet

***Tabelle 4.1*** Technische Daten von PC-Bussystemen

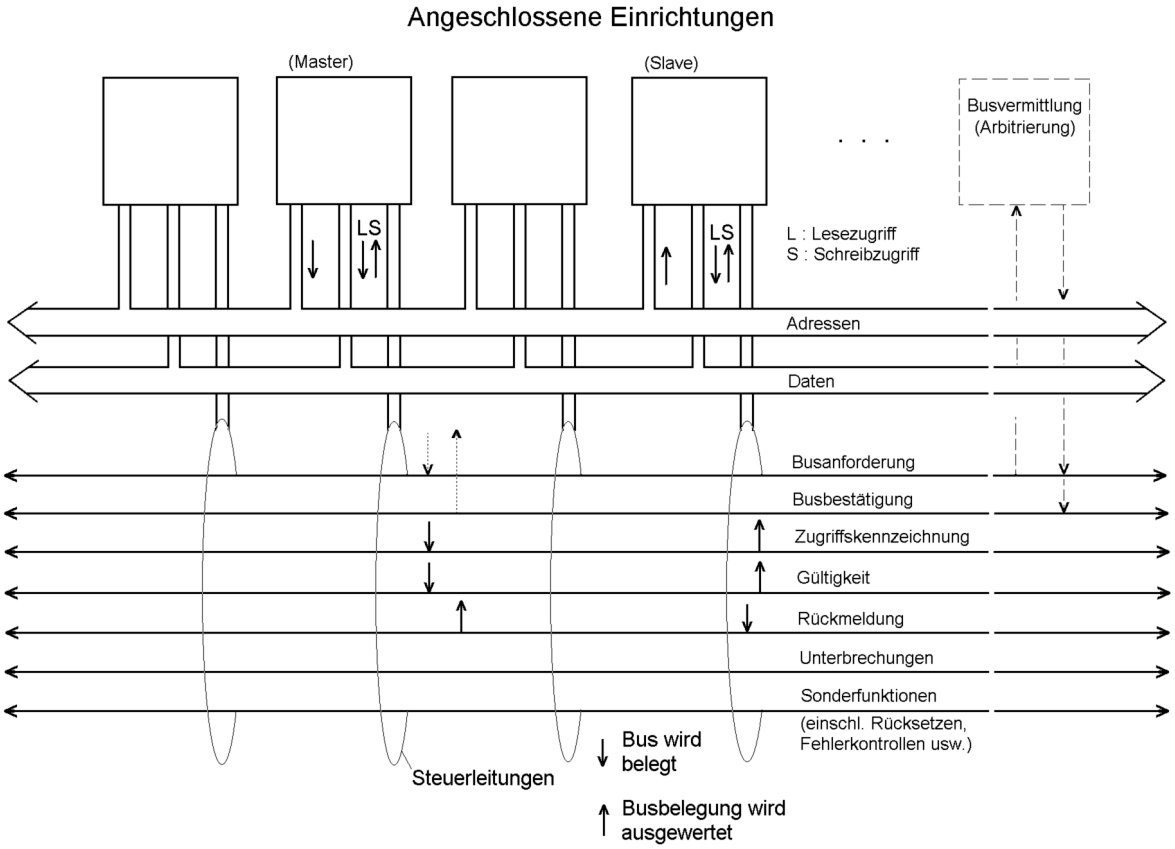



***Abbildung 4.2*** Der grundsätzliche Aufbau eines Systembus

**Mehrfachnutzung von Leitungen**

Die Anzahl der Signale hat wesentlichen Einfluß auf die Kosten eines Bussystems (Größe und Bauart der Steckverbinder, von den Leiterzügen belegte Fläche auf dem Motherboard, Kontakte an den Steuer- und Koppelschaltkreisen). Deshalb hat man immer wieder Bussysteme entwickelt, die mit vergleichsweise wenigen Leitungen auskommen. Die extreme Lösung: ein einziger Signalweg (serielle Informationsübertragung). Dieses Prinzip war aber lange Zeit nicht leistungsfähig genug. Deshalb hat man sich darauf beschränkt, verschiedene Leitungen eines

„breiten2 (0 parallelen) Bussystems mehrfach auszunutzen (Fachbegriff: Multiplex-Betrieb). Die wichtigste Form dieser Mehrfachnutzung ist die Übertragung von Adressen und Daten über dieselben Leitungen. Das ist an sich ohne Schwierigkeiten möglich, da in jedem Zugriff zuerst die Adresse benötigt wird. Beim Lesen kann die adressierte Einrichtung ihre Daten ohnehin erst eine gewisse Zeit später an den Prozessor liefern, so daß man in der Zwischenzeit die Leitungen für den Einwärtstransport freischalten kann. Auch beim Schreiben kann man die Daten der Adresse nachschicken. Anwendungsbeispiele: PCI und AGP. Die nächste Stufe: die Leitungsanzahl ist geringer als die Zugriffsbreite, so daß man die Daten in mehreren aufeinanderfolgenden Zeitabschnitten übertragen muß (serienparallele Übertragung). Beispiel: LPC. Die letzte Stufe: bitserielle Übertragung mit extremen Datenraten. Beispiel: PCI Express.

**Master und Slave, Initiator und Target**
Eine Einrichtung, die den aktuellen Buszyklus auslöst und bestimmt, welche Art von Zugriff mit welcher Adresse ausgeführt wird (die, wie man sagt, die Kontrolle über den Bus bzw. die Busherrschaft hat), heißt allgemein (Bus-) *Master* (sprich: Maaster) oder *Initiator*. Die jeweils ausgewählte Einrichtung (zu der zugegriffen wird) heißt *Slave* (sprich: Sleef) oder *Target*[*)].

*): sprich: Targett oder auch Torgitt (nicht aber etwa Tardschett!)).

**Multi-Master-Betrieb**
Multi-Master-Betrieb bedeutet, daß mehrere Einrichtungen Master werden (also die Kontrolle über den Bus ausüben) können. Natürlich kann in einem Buszyklus nur eine Einrichtung Master am Bus sein. Streben mehrere Einrichtungen gleichzeitig die Kontrolle über den Bus an, so muß es irgendeine „Schiedsrichter"-Funktion geben: Jede Einrichtung, die als Master aktiv werden will, stellt eine Anforderung (Bus Request). Die Einrichtung, die beim nächsten Buszyklus jeweils „dran" ist, erhält den Bus zugesprochen (Bus Grant). Die entsprechende „Schiedsrichter"-Funktion heißt *Busvermittlung* oder Arbitrierung (Arbitration).

**Robustheit**
Hiermit meinen wir im wesentlichen eine Gebrauchseigenschaft auf *elektrischer* Ebene: es muß möglich sein, eine Vielzahl von Steckkarten unterschiedlicher Hersteller problemlos anschließen zu können; der Bus muß hinreichend belastbar und unempfindlich gegen Störungen sein.

**Einfache Erweiterbarkeit**
Steckkarten sollten sich ohne besondere Mühe einbauen bzw. austauschen lassen: Gehäuse auf → Steckkarte rein → Gehäuse wieder zu → fertig. Dieser Wunschvorstellung kommt man heutzutage schon recht nahe - aber ab und zu macht es doch noch Mühe[*)]. Der zugehörige Allerwelts-Begriff: *Plug and Play* (auch: Plug&Play oder PnP; sprich: Plack änd Plee). Die Praktiker haben hierzu folgende Interpretation gefunden: „Plug" heißt „Stecken" und „Play" heißt „Spielen", also müssen wir solange stecken, bis es spielt... Tabelle 4.2 gibt einen Überblick über derartige Vorkehrungen.

*): was richtig Arbeit machen kann: (1) ISA-Steckkarten, (2) die Installation der Gerätetreiber (das ist aber ein *Software*-Problem).

| Bussystem | Plug&Play-Vorkehrungen |
|---|---|
| ISA, herkömmlich (Legacy Devices) | keine |
| ISA gemäß Microsoft-Spezifikationen zu Windows 95 (seit etwa 1995/96) | gemäß Plug and Play ISA Specification (Microsoft, 1994). Konfigurationsdaten in Festwertspeicher auf Steckkarte |
| EISA und MCA (veraltet) | Konfigurationsangaben in Dateiform; manche EISA-Karten sind auch gemäß Plug and Play ISA Specification ausgelegt (Konfigurationsdaten in Festwertspeicher auf Steckkarte) |
| PCI und AGP | Konfigurationsdaten in Festwertspeicher auf Steckkarte; bedarfsweise zusätzliche Konfigurationsdateien |

***Tabelle 4.2*** Plug&Play-Vorkehrungen (Überblick)

*Wir merken uns:*
PCI-, AGP- und PnP-ISA-Karten haben die Konfigurationsdaten sozusagen stets bei sich (bloßes Stecken genügt).

**Ganz neumodisch - Hot Plugging**
Noch einfacher dürfte es wirklich nicht gehen - wir müssen den PC nicht einmal ausschalten, um eine Steckkarte zu tauschen oder einzusetzen. Aber Vorsicht - keines der Busysteme in üblichen PCs ist dafür ausgelegt. (Also: zum Tauschen oder Einbauen von Steckkarten den PC ausschalten!)

*Hinweise:*

1. Es gibt Schnittstellen, die von vornherein für diese Nutzungsweise vorgesehen sind (z. B. das PC-Card-Interface und die modernen seriellen Bussysteme (USB, Firewire usw.)),
2. Auch PCI-Systeme können entsprechend ausgelegt sein. Hiermit ist vor allem in Servern und im industriellen Bereich zu rechnen. Trotzdem geht es nicht um ein wildes Stecken und Ziehen - der Servicetechniker muß sich schon an einige Dienstvorschriften halten.

**Unabhängigkeit vom Prozessor**
Diese Forderung erscheint heutzutage selbstverständlich. Bei den ersten Personalcomputern konnte man sie aber (aus Kostengründen) nicht verwirklichen. Vielmehr blieb gar nichts anderes übrig, als die Auslegung des Systembus an die des Prozessorbus anzupassen. In modernen PCs sind es die Motherboard-Schaltkreise, die für die Unabhängigkeit vom Prozessor sorgen. Die entsprechenden Funktionseinheiten sind die Bus-Brücken (Bridges).

**Bus-Brücken**
Eine Brücke (Bridge) ist eine Einrichtung, die zwei Bussysteme miteinander koppelt. Brücken unterscheiden sich von einfachen Busverlängerungen dadurch, daß sie für jedes der beiden Bussysteme aktive Steuerschaltungen enthalten. Im allgemeinen Fall kann die Brücke an jedem der Bussysteme als Master oder als Target arbeiten (Abbildung 4.3).

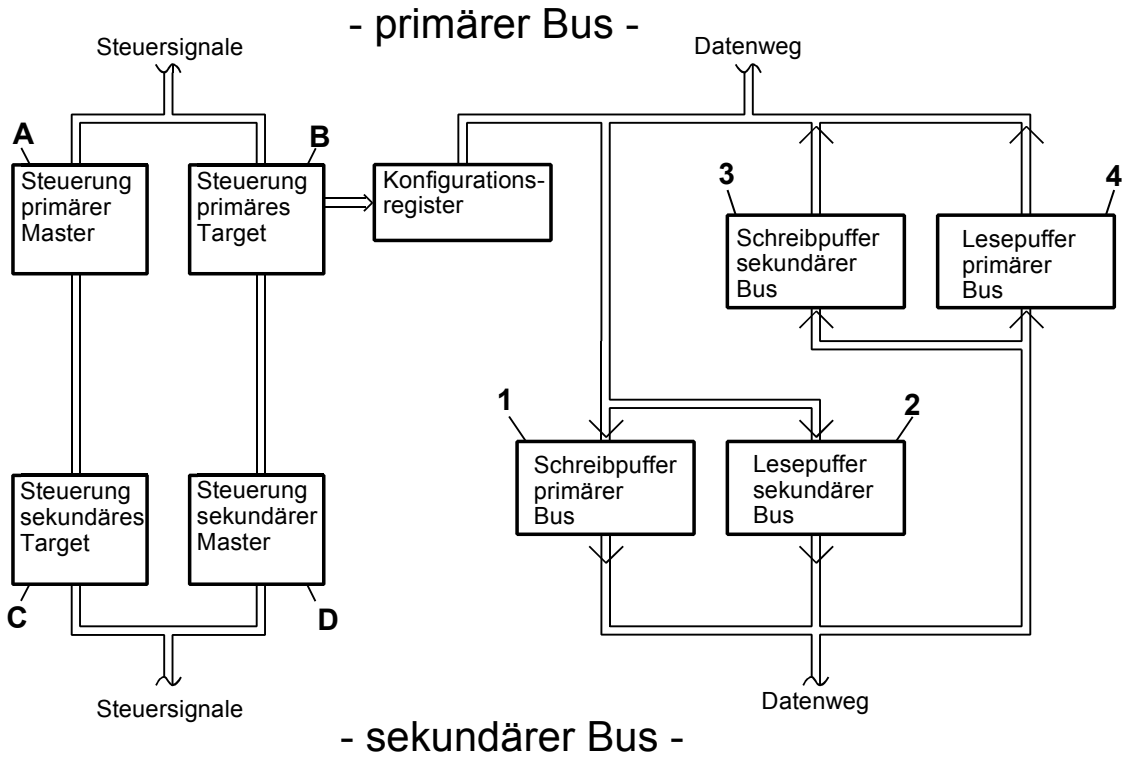



***Abbildung 4.3*** Aufbau einer Brücke

*Erklärung:*
Die primäre Master-Steuerung A ist mit der sekundären Target-Steuerung C verbunden, die primäre Target-Steuerung B mit der sekundären Master-Steuerung D. Die Datenwege beider Bussysteme sind über Schreib- und Lesepuffer 1...4 verbunden.

*Zugriffsabläufe (1): Master am primären, Target am sekundären Bus*
Der Master adressiert die Brücke als Target (Steuerung B). Die Brücke wird am sekundären Bus als Master aktiv (Steuerung D).

- Schreibzugriff: der Master spricht die Brücke mit einem Schreibkommando an und schreibt in Schreibpuffer 1. Die Brücke führt ihrerseits einen Schreibzugriff auf das gewünschte Target aus. Die Schreibdaten stammen aus Schreibpuffer 1.
- Lesezugriff: der Master spricht die Brücke mit einem Lesekommando an. Die Brücke führt ihrerseits einen Lesezugriff auf das gewünschte Target aus. Die gelesenen Daten werden in Lesepuffer 4 abgelegt. Der Master liest die Daten aus Lesepuffer 4.

*Zugriffsabläufe (2): Master am sekundären, Target am primären Bus*
Der Master adressiert die Brücke als Target (Steuerung C). Die Brücke wird am primären Bus als Master aktiv (Steuerung A).

- Schreibzugriff: der Master spricht die Brücke mit einem Schreibkommando an und schreibt in Schreibpuffer 3. Die Brücke führt ihrerseits einen Schreibzugriff auf das gewünschte Target aus. Die Schreibdaten stammen aus Schreibpuffer 3.

- Lesezugriff: der Master spricht die Brücke mit einem Lesekommando an. Die Brücke führt ihrerseits einen Lesezugriff auf das gewünschte Target aus. Die gelesenen Daten werden im Lesepuffer 2 abgelegt. Der Master liest die Daten aus Lesepuffer 2.

Moderne Brückenschaltungen sind in der Lage, Bussysteme und Interfaces zu verbinden, die sich in der Anzahl der Signalleitungen, in den Funktionsprinzipien, in den Datenraten usw. beträchtlich voneinander unterscheiden. Sie sind zudem auf Leistung optimiert. Die interne Wirkungsweise ist deshalb ziemlich kompliziert. So ist beispielsweise ein Zugriff über einen „breiten" Bus in mehrere Zugriffe über einen „schmaleren" Bus aufzulösen. (Es ist offensichtlich, daß ein 64-Bit-Zugriff wenigstens zwei Zugriffe über einen 32-Bit-Bus erfordert.) Darüber hinaus wird aber noch mehr getan. So werden z. B. mehrere aufeinanderfolgende Zugriffe mit 8, 16 oder 32 Bits zu einem 64-Bit-Zugriff zusammengefaßt, und bei manchen Lesezugriffen liest die Brücke von sich aus weiter (vorbeugendes Lesen), wobei die gelesenene Daten in der Brücke zwischengespeichert werden (in der Annahme, daß sie vom jeweils anderen Bus demnächst ohnehin abgefordert werden).

**Systembus und Lokalbus**
Nach der sozusagen klassischen Auffassung ist ein Systembus universell, robust und - was die Anzahl der Steckpositionen betrifft - „großzügig" auszulegen. Das hat aber seinen Preis. Daß Schaltkreise, die auch mal eine Überlastung vertragen, daß hochwertige Steckverbinder usw. Geld kosten, ist klar. Aber auch das Leistungsvermögen (vor allem: hinsichtlich der Datenrate) ist grundsätzlich beschränkt: jede Standardisierung, jede Vorkehrung für Robustheit in elektrischer Hinsicht, kostet zusätzliche Schaltmittel bzw. erfordert längere Signalwege, die von den informationstragenden Signalen durchlaufen werden müssen, kostet also Geschwindigkeit bzw. Datendurchsatz.

Wir kennen aber bereits den vom Prinzip her einfachsten Ausweg, das Leistungsvermögen des Prozessors voll zur Wirkung kommen zu lassen und trotzdem kompatibel zu bleiben: wir ordnen die betreffende Hardware auf dem Motherboard an und haben somit volle Narrenfreiheit in der Schaltungsauslegung (vgl. Abschnitt 1.7.1). Ein richtiger Systembus ist dann nur noch notwendig, um „Wald- und Wiesen-„ Peripherie anschließen zu können. Aber auch die Nachteile dieses Prinzips sind nicht zu übersehen: das Motherboard wird teuer, und die unmittelbar mit dem Prozessor gekoppelte Hochleistungs-Hardware ist weder erweiter- noch austauschbar. Einige Hersteller haben dafür schon in den 80er Jahren einen Ausweg gefunden: besondere Steckpositionen für einen leistungsoptimierten Bus. Dafür hat sich die Bezeichnung Lokalbus (Local Bus) durchgesetzt.

*Typische Merkmale eines herkömmlichen Lokalbus:*

- nur wenige Steckpositionen (typischerweise 1...3),
- enge Kopplung zum Prozessorbus,
- Anordnung der Slots in unmittelbarer Nähe des Prozessors (kürzeste Signalwege),
- Verzicht auf „Robustheit" in der elektrischen Auslegung - die Steckkarten sind oft schon in der Handhabung recht empfindlich. Stichwort: ESD (s. die Anregungen zur praktischen Selbstbetätigung).

**Standardisierte Lokalbussysteme**
Zunächst war der Lokalbus eine herstellerspezifische Angelegenheit. Man hat sich aber auch hier um Standardisierung bemüht. Bis Mitte der 90er Jahre standen zwei Systeme im Wettbewerb:

- der VL- bzw. VESA-Bus (VESA = Video Electronics Standard Association),
- der PCI-Bus (PCI = Peripheral Component Interconnect).

Der VESA-Bus war eine „Schnell- und Billiglösung", die vor allem von Graphikkartenherstellern gefördert wurde. Hierfür hat man den 32-Bit-Bus der Prozessoren 386 und 486 als Grundlage genommen und den Bus so entworfen, daß die Hardware mit weithin verfügbaren Schaltkreisen aufgebaut werden konnte. (Was man sich hierdurch eingehandelt hat, waren allerdings erhebliche Kompatibilitäts- und Zuverlässigkeitsprobleme.)

Die PCI-Entwicklung wurde ursprünglich vor allem von Intel angeregt. PCI war von Anfang an als wirklich prozessor-unabhängiger Hochleistungsbus gedacht. (Intel konnte es sich - als erfahrener Schaltkreishersteller - leisten, kompliziertere Wirkprinzipien zu wählen und auf die Nutzung von Wald- und Wiesen-Schaltkreisen zu verzichten.) Zu den kennzeichnenden Merkmalen des PCI-Bus gehören die Kopplung über Brücken sowie die sorgfältige Spezifikation der elektrischen Kennwerte.

**Nostalgie: MCA, EISA, VESA**
Diese Bussysteme sind veraltet. MCA (Microchannel Architecture) ist das Bussystem der PS/2-Modelle von IBM. Nur wenige Hersteller haben Lizenzen genommen. IBM selbst hat MCA 1995 zugunsten von PCI aufgegeben. EISA (Extended ISA Architecture) war die Konkurrenzentwicklung zu MCA. Der VESA-Lokalbus ist in 386- und 486-PCs aus der 1. Hälfte der 90er Jahre zu finden.

# 4.2. Der ISA-Bus

Der ISA-Bus (auch: AT-Bus) ist der sozusagen klassische Systembus der AT-kompatiblen PCs. Es handelt sich um eine Einfachlösung mit - heutzutage - sehr beschränktem Leistungsvermögen und praktisch nicht vorhandenem Komfort in Hinblick auf das Einbauen von Steckkarten.

Dieses Bussystem, in der 1. Hälfte der 80er Jahre eingeführt, wird uns in der Praxis sicherlich noch einige Zeit erhalten bleiben. Die wesentlichen Gründe hierfür sind:

- die Anzahl, Verbreitung und Vielfalt der vorhandenen Steckkarten (einschließlich der vielen Spezialitäten, z. B. auf dem Gebiet der Meßtechnik und der „industriellen" Schnittstellen),
- die Einfachheit des Bussystems (ISA-Steckkarten kann man, was die Buskoppelhardware angeht, buchstäblich am Küchentisch entwickeln und bauen).

*Das Verschwinden des ISA-Bus*
Dieses Ziel verfolgen vor allem Microsoft und Intel seit den 90er Jahren (Legacy Free PC[*)]). Ganz so schnell ging es jedoch nicht. Seit der Jahrtausendwende sind die ISA-Slots aus den PCs des Massenmarktes nach und nach verschwunden. Der ISA-Bus verbleibt aber noch einige Zeit im industriellen und Embedded-Bereich[**)], und auch Privatanwender, die gern am und mit dem PC basteln, dürften ISA-Slots weiterhin nachfragen.

*): Microsoft und Intel bezeichnen die Hardware, die noch mitzuschleppen bzw. zu unterstützen ist, als Legacy (= Erblast). Im allgemeinen Sinne ist jede ISA-Steckkarte eine sog. Legacy Device, im engeren Sinne sind es u. a. jene ISA-Steckkarten, die nicht der Plug&Play-ISA-Spezifikation entsprechen.

**): wenn auch typischerweise in anderen Formfaktoren.

*Abhilfe*
Manchmal möchte man vorhandene Steckkarten weiternutzen, manchmal geht es um ein einfaches Hardware-Interface. Der ISA-Bus ist sicherlich alles andere als ideal. Er ist aber im PC-Bereich praktisch das einzige standardisierte Bussystem, das es ermöglicht, Steckkarten mit einfachsten Mitteln aufzubauen.

*Naheliegende Auswege:*

- wir suchen nach älterer Hardware*),
- wir greifen auf das Angebot der „industriellen" PC-Technik zurück.

*): *Praxistip:* brauchbare Motherboards mit ISA-Slots nicht voreilig entsorgen.

Zudem kommt es vor, daß manche Hersteller in der Lösung dieses Problems eine Marktnische sehen und z. B. extern anschließbare ISA-Zusätze oder auch - Microsoft hin, Intel her - nach wie vor Motherboards *mit* ISA-Slots anbieten. Als Beispiel siehe Abbildung 1.16.

*Hinweis:*
Werden ISA-Slots gewünscht, so ist zu fragen, unter welchem Betriebssystem die Hardware laufen soll. Einfach-Interfaces sind nicht nutzbar, wenn die Systemsoftware den Zugriff zu ihnen verhindert (wie beispielsweise Windows 2000 oder Windows XP). Genauer: mit Zugriffsbefehlen im Anwendungsprogramm kann man eine solche Steckkarte nicht mehr ansprechen. Das gelingt nur noch über einen zum System passenden Gerätetreiber.

## 4.2.1. Signale und Slots

Abbildung 4.4 zeigt die Signalbelegungen. Die Anordnung der Slots auf dem Motherboard ist aus den Abbildungen 1.8 und 1.10 ersichtlich. Der 8-Bit-Slot enthält nur 8 Daten- und 20 Adreßleitungen. Beim 16-Bit-Slot handelt es sich um einen 8-Bit-Slot, der durch einen angefügten Steckverbinder erweitert ist (16-Bit-Datenübertragung, 24-Bit-Adressierung).

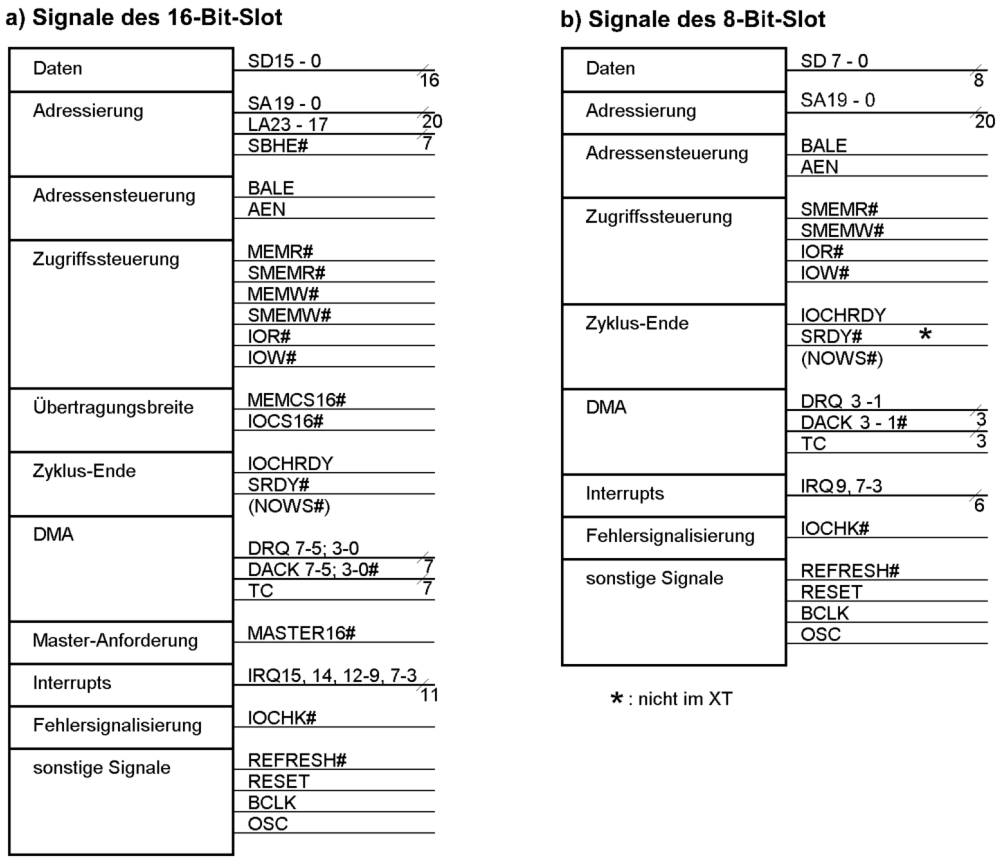



***Abbildung 4.4*** Die Signale des ISA-Bus

*Übersicht:*

- Einsatzbereich: Steckkarten-Erweiterungsbus auf Motherboards,
- Buskopplung: ursprünglich über Schaltkreise der Baureihe „Low Power Schottky TTL"*),
- funktionelle Auslegung: zentralgesteuertes Bussystem mit einem Master (dem Prozessor) und zusätzlichen DMA-Vorkehrungen (über die auch ein einfacher Multi-Master-Betrieb verwirklicht werden kann),
- Adressierung: 20 bzw. 24 Adreßbits,
- Datenwegbreite: 8 bzw. 16 Bits,
- Adreßräume: 2 (Speicher, Ein- und Ausgabe),
- Bustakt: 8...10, typisch 8,33 MHz. Steckkarten können am Bus praktisch asynchron arbeiten, müssen also den Bustakt nicht unbedingt auswerten.

*): eine ausgesprochen kostengünstige Schaltkreisfamilie.

*Datenraten*
Tabelle 4.3 gibt einen Überblick über die am ISA-Bus maximal erreichbaren Datenraten.

| Buszyklus | Datenrate | |
|---|---|---|
| | ohne Wartezustände | Normalbetrieb |
| 8-Bit-Speicherzugriff | 2,08 MBytes/s | 1,39 MBytes/s |
| 8-Bit-E-A-Zugriff | 2,78 MBytes/s | 1,39 MBytes/s |
| 8-Bit-DMA-Zugriff | - | 1,04 MBytes/s |
| 16-Bit-Speicherzugriff | 8,33 MBytes/s | 5,56 MBytes/s |
| 16-Bit-E-A-Zugriff | - | 5,56 MBytes/s |
| 16-Bit-DMA-Zugriff | - | 2,08 MBytes/s |

***Tabelle 4.3*** Maximale Datenraten am ISA-Bus (Bustakt: 8,33 MHz)

## 4.2.2. Steckkarten

ISA-Slots sind direkte Steckverbinder mit einem Anschlußabstand von 2,54 mm. Entsprechend den Slots gibt es 8-Bit- und 16-Bit-Steckkarten (vgl. Abbildung 4.1a, b).

*Wie groß darf eine Steckkarte sein?*
Für ISA/EISA-Systeme gibt es einen charakteristischen Formfaktor: eine Steckkarte darf maximal rund 339 mm lang und - ohne Steckverbinder - rund 100 mm hoch sein (Abbildung 4.5).

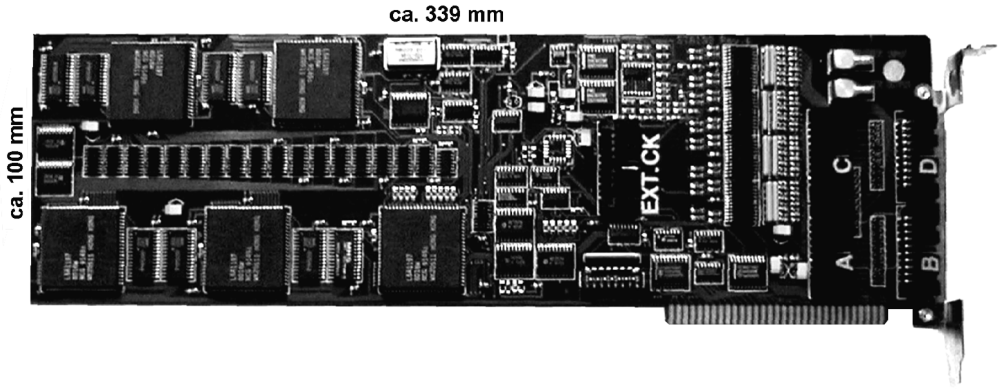



***Abbildung 4.5*** Eine Steckkarte, die den ISA-Formfaktor voll ausnutzt

*Speisespannungen*
In den ISA-Slots werden 4 Speisespannungen geführt: + 5 V, + 12 V, - 5 V, - 12 V. Von besonderer Bedeutung ist die „Logik"-Speisespannung (+ 5 V). Wir merken uns: ISA-Karten sind 5-V-Karten. Die typische Stromaufnahme (+ 5 V): bis zu 2 A je Karte.

## 4.2.3. Konfigurationspraxis

Eine Wissenschaft für sich! Herkömmliche ISA-Karten wollen von Hand konfiguriert sein. Zur Konfigurationseinstellung sind typischerweise Jumper (Steckbrücken) vorgesehen. Einzustellen sind:

- Adreßbereiche (für Speicher- und E-A-Zugriffe),
- DMA-Kanäle,
- Interruptleitungen.

*Wir merken uns:* jeder Adreßbereich, jeder DMA-Kanal, jede Interruptleitung darf grundsätzlich nur von *einer* Einrichtung (Schaltkreis auf dem Motherboard oder Karte) genutzt werden.

Die meisten Karten sind voreingestellt und funktionieren auf Anhieb. Probleme entstehen dann, wenn wir die ISA-Slots mit Karten vollstopfen. Vor allem Soundkarten, Netzwerkkarten und Karten mit zusätzlichen Schnittstellen bereiten gelegentlich Mühe.

*Plug&Play-Karten*
Moderne ISA-Karten haben typischerweise Plug&Play-Vorkehrungen. Die einschlägige Spezifikation ist eine Entwicklung der Fa. Microsoft. Sie ermöglicht es, die Konfiguration der Karten über Software einzustellen. Dafür hat man besondere Zugriffsfolgen definiert. Die Konfigurationsdaten werden auf den Karten in einem Festwertspeicher gehalten. Trotzdem läßt sich nicht jeder x-beliebige Konflikt lösen.

## 4.2.4. Der Einfachbus auf dem Motherboard: der X-Bus

Der X-Bus ist herkömmlicherweise eine Fortsetzung des ISA-Bus, an den die PC-typische Standard-Peripherie (Diskettenlaufwerks-Controller, Tastatur-Controller, Realzeituhr, Schnittstellen-Controller usw.) sowie der BIOS-ROM angeschlossen werden. Vgl. hierzu die Abbildungen 1.26 und 1.27.

Der X-Bus ist praktisch ein abgezweigter ISA-Bus mit folgenden Merkmalen:

- Datenwegbreite: 8 Bits,
- Adresse: bis zu 24 Bits (16-MByte-Speicheradreßraum),
- Funktionsweise: wie ISA (bei Beschränkung auf 8-Bit-Zugriffe),
- maximale Datenrate: knapp über 1 MBytes/s,
- keine Steckkarten-Slots; alle am X-Bus angeschlossenen Einrichtungen sind fest auf dem Motherboard angeordnet.

## 4.2.5. Der Nachfolger auf dem Motherboard: LPC

LPC = Low Pin Count Interface Specification. Dieser Interfacestandard wurde von Intel entwickelt, um damit den X-Bus auf dem Motherboard abzulösen (Abbildung 4.6). Typische Merkmale:

- der Bus ist auf das Motherboard beschränkt,
- wie beim X-Bus gibt es keinen Steckkarten-Slots,
- die Datenrate muß nicht höher sein als jene des X-Bus (1...2 MBytes/s),
- Unterstützung des gesamten linearen Speicheradreßraums (4 GBytes),
- anschließbare Einrichtungen: wie beim X-Bus (Peripherie-Controller, BIOS-ROM usw.). Darüber hinaus können Systemverwaltungsvorkehrungen, auf dem Motherboard untergebrachte Audio-Einrichtungen usw. angeschlossen werden.

LPC wurde als synchrones Bussystem in Anlehnung an PCI ausgelegt. Es arbeitet mit dem gleichen Bustakt (33 MHz). Die Grundausstattung umfaßt - neben dem Takt - nur 6 Signale, die bedarfsweise um weitere Signale (maximal 5) ergänzt werden können. Kennzeichnend ist eine „serienparallele" Übertragung von Zugriffskommandos, Adressen und Daten in 4 Bits breiten Abschnitten (das kann man sich leisten, weil die Datenrate nicht höher sein muß als die des herkömmlichen X-Bus).

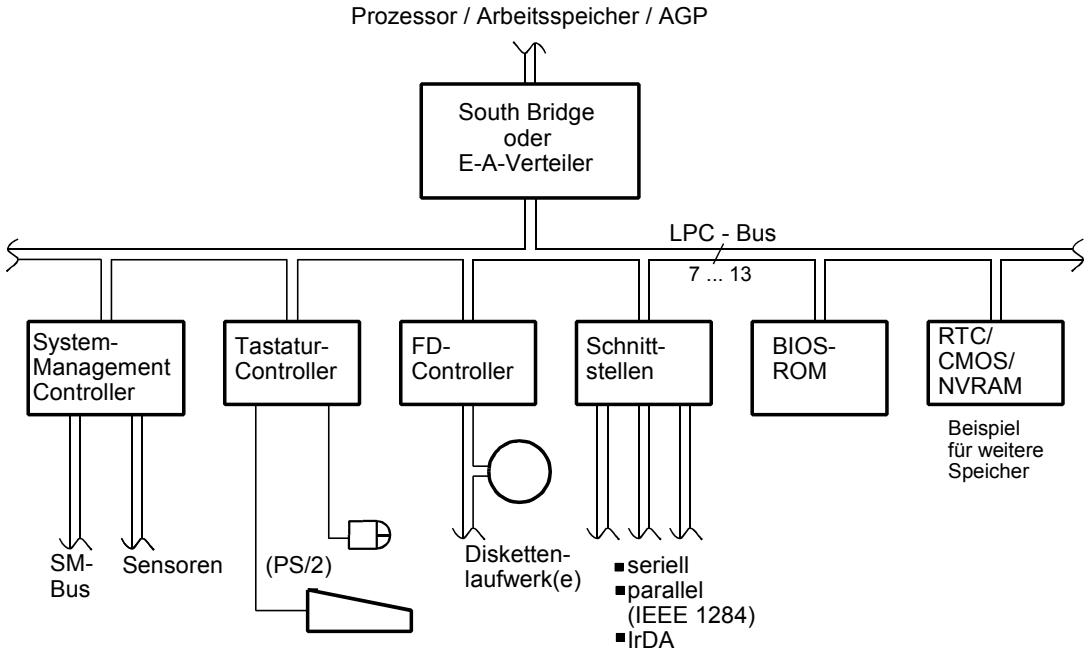



***Abbildung 4.6*** Motherboard mit LPC-Bus (Ausschnitt)

# 4.3. PCI

Der PCI-Bus wurde ursprünglich (Anfang der 90er Jahre) als reiner Lokalbus für Motherboards kompatibler PCs entwickelt. Er ist vorzugsweise zum Anschließen peripherer Einrichtungen vorgesehen (PCI = Peripheral Components Interconnect). Solche Einrichtungen können sowohl als Schaltkreise auf dem Motherboard als auch in Form von Steckkarten ausgeführt sein. Die ursprünglichen Entwicklungsziele lassen sich folgendermaßen beschreiben:

- der Bus sollte leistungsfähiger sein (höhere Datenraten, kürzere Latenzzeiten) als die seinerzeit gängigen Bussysteme (ISA, EISA, MCA),
- der Bus sollte moderne Systemkonzepte unterstützen (32-Bit-Adressierung, Caches, Multi-Master-Betrieb, Burst-Zugriffe),

- automatische, durch Software steuerbare Konfiguration (Plug and Play),
- es sollte ein echter, sowohl in funktioneller als auch in elektrischer Hinsicht wohldefinierter Standard geschaffen werden,
- Abwärtskompatibilität zu den herkömmlichen PC-Systemen,
- Kostenoptimierung auf den Einsatzzweck hin (Lokalbus in üblichen PCs) - also Verzicht auf eine „total universelle" Auslegung.

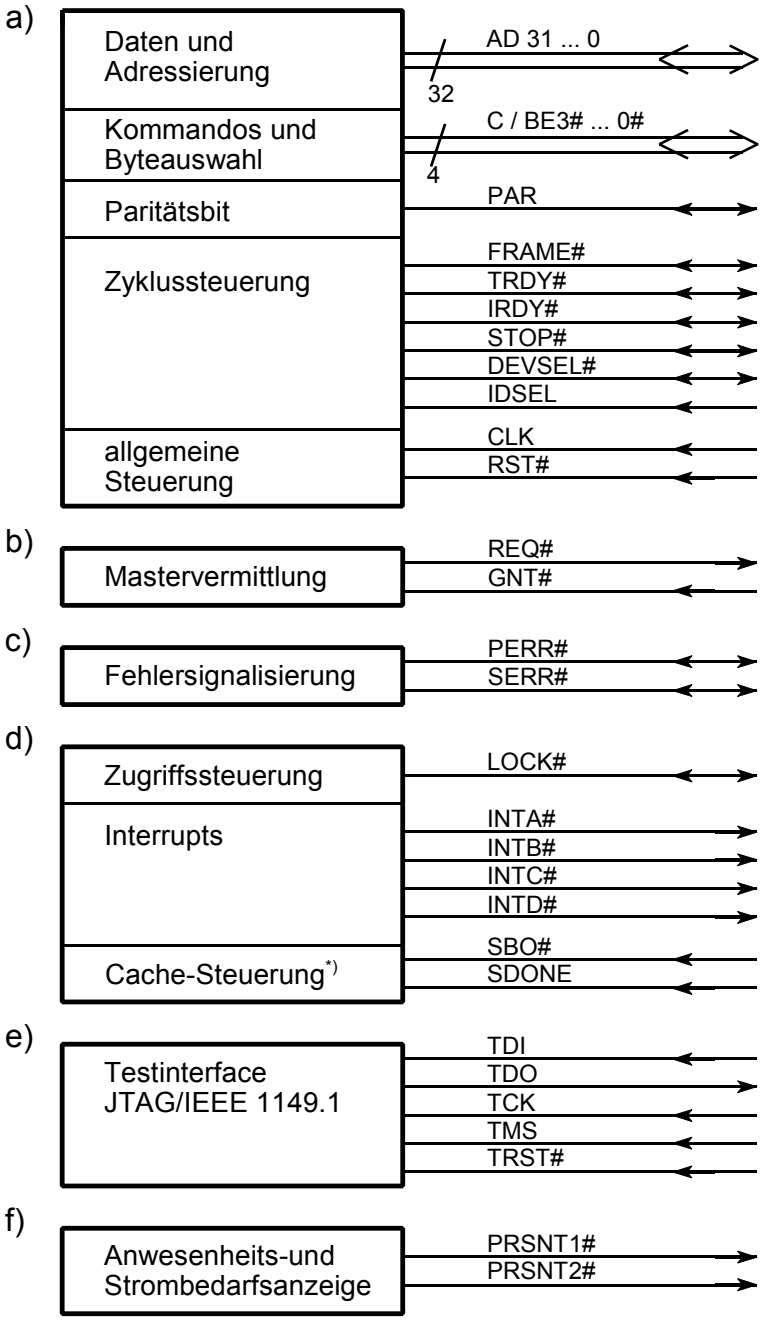
Der PCI-Bus ist heutzutage *der* Industriestandard bei den Bussystemen. Im Laufe der Zeit ist er in anderen Bereichen (Telekommunikation, Automatisierungstechnik) eingeführt worden. Für bestimmte Einsatzgebiete wurden besondere Formfaktoren entwickelt (Beispiele: CardBus, CompactPCI, Mini PCI, Small PCI).

## 4.3.1. Signale und Slots

Abbildung 4.13 zeigt die Signalbelegungen. Beispiel für PCI-Slots finden Sie u. a. in den Abbildungen 1.1, 1.10 und 1.15.

*Übersicht:*

- Einsatzbereich (ursprüngliche Bestimmung): Motherboards und Motherboard-Steckkarten-Kombinationen,
- Buskopplung: kostengünstiger Direktanschluß ohne besondere zwischenzuschaltende Buskoppelschaltkreise. Oftmals ist eine PCI-Einrichtung ein einziger Schaltkreis.
- funktionelle Auslegung: synchrones, zentral gesteuertes Multimaster-Bussystem mit Zeitmultiplexübertrag der Adressen und Daten,
- Adressierung: 32 Adreßbits, erweiterbar auf 64-Bit-Adressierung,
- Datenwegbreite: 32 Bits, erweiterbar auf 64 Bits,
- Adreßräume: 3 (Speicher, Ein- und Ausgabe, Konfiguration),
- Bustakt: 33 oder 66 MHz,
- Master-Auswahl (Arbitrierung): Prinzip der unabhängigen Anforderungen. Zentrale Vermittlung auf dem Motherboard.
- anschließbare Einrichtungen: das können sowohl Schaltkreise auf dem Motherboard als auch Steckkarten sein. Beschränkung auf 10 Buslasten. 1 Buslast = 1 Schaltkreis bzw. 1 Steckverbinder (Slot). Ein Slot mit Steckkarte entspricht also 2 Lasten.
- Erweiterung des Bussystems: über Brücken (PCI-to-PCI-Bridges).




*): mit Spezifikation 2.2 entfallen

***Abbildung 4.7*** Die Signale des PCI-Bus (32-Bit-Bus gemäß Spezifikation 2.1)

*Datenraten*
Tabelle 4.4 gibt einen Überblick über die am PCI-Bus maximal erreichbaren Datenraten.

| Datenwegbreite | Bustakt | maximale Datenrate |
|---|---|---|
| 32 Bits | 33 MHz | 132 MBytes/s |
| | 66 MHz | 264 MBytes/s |
| 64 Bits | 33 MHz | 264 MBytes/s |
| | 66 MHz | 528 MBytes/s |

***Tabelle 4.4*** Maximale Datenraten am PCI-Bus

*PCI-Einrichtungen*
PCI-Einrichtungen können grundsätzlich auf zweierlei Weise realisiert sein:

- als (fest eingelötete) Schaltkreise auf dem Motherboard,
- als Steckkarten (Slot-Implementierung).

Allen Ausführungsformen ist gemeinsam, daß (meistens) die eigentliche PCI-Einrichtung aus einem einzigen Schaltkreis besteht oder über einen einzigen Schaltkreis mit dem PCI-Bus gekoppelt ist (Abbildung 4.8).

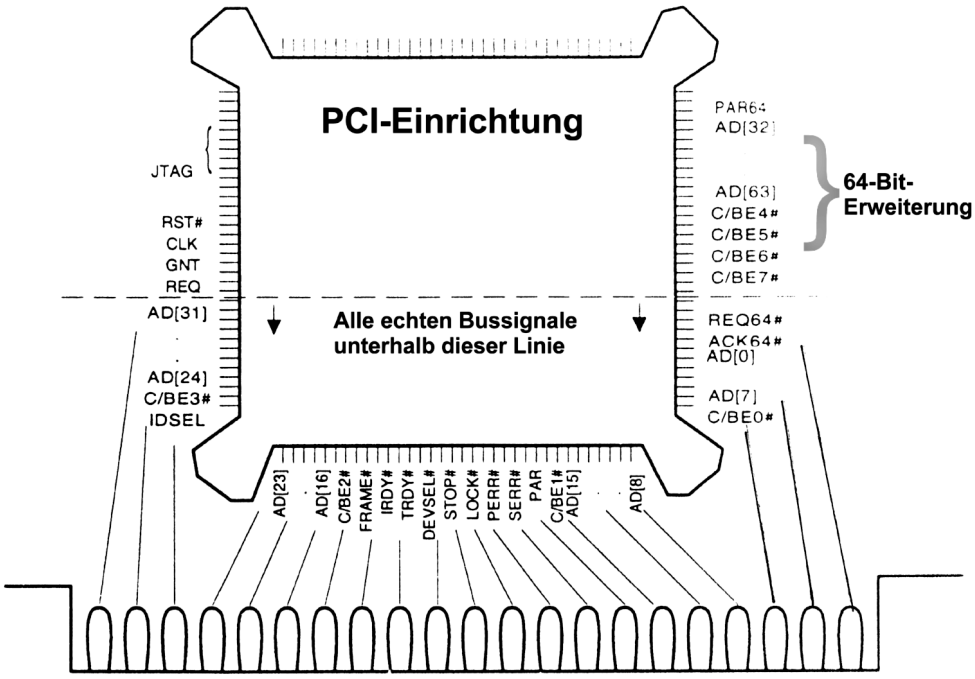



***Abbildung 4.8*** PCI-Einrichtung als Schaltkreis (PCI SIG)

Die Abbildung zeigt den Schaltkreis auf einer Steckkarte. Auf dem Motherboard fest eingelötete PCI-Schaltkreise sehen ähnlich aus. Wichtig ist, daß es keine besonderen Treiber,

Buskoppelschaltkreise o. dergl. gibt, sondern daß Treiber, Empfänger und Bussteuerschaltungen in einem einzigen Schaltkreis untergebracht sind.

*Hinweis:*
Die elektrischen Anforderungen der PCI-Spezifikation sind so scharf, daß man sie auch nur auf diese Weise erfüllen kann. Eine PCI-Einrichtung läßt sich praktisch nicht „diskret" (also mit einzelnen Bustreibern, Gattern, Registern usw.) aufbauen.

## 4.3.2. Steckkarten

*Steckverbinder*
Es handelt sich um direkte Leiterplatten-Steckverbinder mit einem Anschlußabstand von 1,27 mm (0,05"). Anzahl der Steckkontakte: für den 32-Bit-Bus 120, für den 64-Bit-Bus 184. Weitere Einzelheiten im folgenden Abschnitt 4.3.3.

*Steckkartenabmessungen*
PCI-Steckkarten sind so spezifiziert, daß sie u. a. zusammen mit ISA-Karten eingesetzt werden können. Der maximal ausnutzbare Formfaktor entspricht Abbildung 4.5. Die weitaus meisten Karten sind aber erheblich kleiner (Abbildung 4.9).

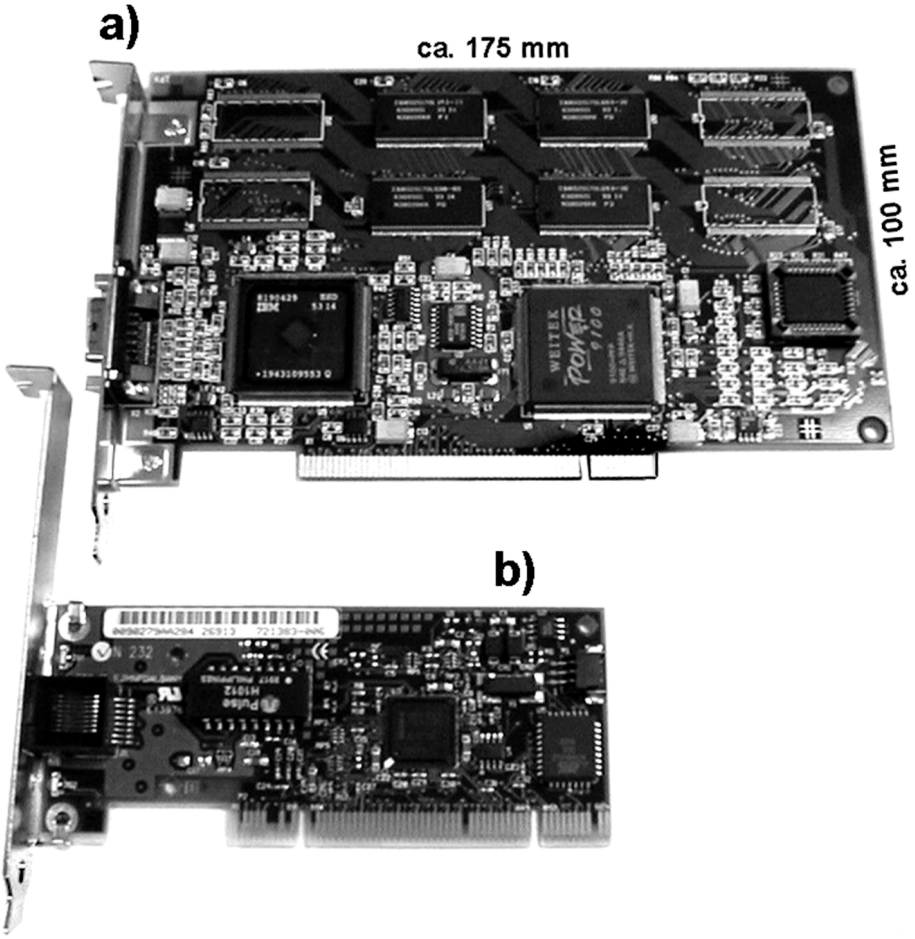



***Abbildung 4.9*** PCI-Steckkarten

*Erklärung zu Abbildung 4.9:*
a) - eine Karte, die den PCI-Formfaktor für sog. kurze Karten weitgehend ausnutzt; b) - eine sehr sparsam bemessene Karte.

## 4.3.3. PCI-Konfigurationen

*PCI-Systeme unterscheiden sich:*

- in den Signalpegeln: 5-V-PCI und 3,3-V-PCI,
- in der Zugriffsbreite: 32-Bit-PCI und 64-Bit-PCI,
- in der Bustaktfrequenz: 33-MHz-PCI und 66-MHz-PCI.

Tabelle 4.5 nennt die zulässigen Kombinationen.

| **Signalpegel: 5 V** | **Signalpegel: 3,3 V** | |
|---|---|---|
| ▪ 32 Bits, 33 MHz,<br>▪ 64 Bits, 33 MHz | ▪ 32 Bits, 33 MHz,<br>▪ 64 Bits, 33 MHz | ▪ 32 Bits, 66 MHz,<br>▪ 64 Bits, 66 MHz |

***Tabelle 4.5*** Zulässige PCI-Konfigurationen

*Zur Steckpraxis*
Was die Signalpegel (5 V oder 3,3 V) und die Zugriffsbreiten (32 oder 64 Bits) betrifft, so bestimmen die Steckverbinder - durch ihre Formgebung - was zusammenpaßt und was nicht. Die Taktfrequenz hingegen wird über ein besonderes Bussignal (M66EN) gleichsam ausgehandelt:

1. 66-MHz-Einrichtungen können an einem 33-MHz-Bus betrieben werden,
2. ein Bussystem, das an sich für 66 MHz ausgelegt ist, wird auf 33 MHz zurückgeschaltet, sobald eine 33-MHz-Einrichtung angeschlossen ist.

*Die Signalpegel*
Es gibt 2 Arten von PCI-Systemen: 5 V und 3,3 V. Diese Bezeichnungen erinnern an typische Speisespannungen, stehen aber für bestimmte Spezifikationen der Signalpegel. Tatsächlich liegen den Bezeichnungen gewisse „Vorzugslösungen" der Spannungsversorgung zugrunde:

- „5 V" = vorzugsweise 5 V Speisespannung, Signalpegel gemäß TTL-Spezifikation,
- „3,3 V" = vorzugsweise 3,3 V Speisespannung, Signalpegel gemäß CMOS-Spezifikation.

Jede PCI-Konfiguration (Schaltkreise auf dem Motherboard + Slots) muß für jeweils eine dieser Spezifikationen ausgelegt sein (gemischte Konfigurationen sind unzulässig).

Wenn es sich um die Bestückung des Motherboards handelt, so ist es Sache des Herstellers, die passenden Schaltkreistypen auszuwählen.

Steckkarten sind durch mechanische Verriegelung (genauer: durch Sperren im Slot-Steckverbinder und Kerben in der Leiterplatte) gegen fälschliches Stecken abgesichert (Abbildung 4.10).

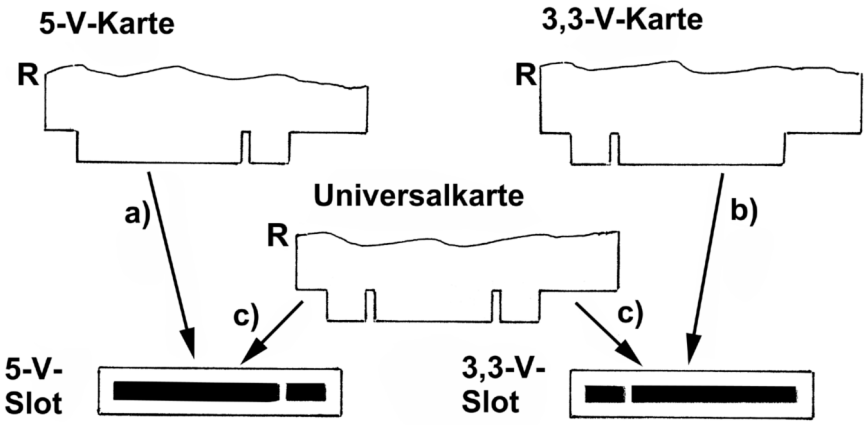



***Abbildung 4.10*** PCI-Steckkarten und Signalpegel

*Erklärung:*

a) 5-V-Karten passen nur in 5-V-Slots (Kerbe/Sperre vorn[1)]),
b) 3,3-V-Karten passen nur in 3,3-V-Slots (Kerbe/Sperre hinten[2)]),
c) Universalkarten (Dual Voltage Signaling Boards) passen in beide Arten von Slots (Kerben vorn und hinten). Sie stellen sich automatisch auf die jeweilige Umgebung (5 V oder 3,3 V) ein.

*Anmerkungen:*

1) „vorn" = der PC-Rückseite bzw. dem Slot-Abdeckblech (Bracket) der Karte abgewandt,
2) „hinten" = der PC-Rückseite bzw. dem Slot-Abdeckblech (Bracket) der Karte zugewandt.

*Speisespannungen*
In den Slots sind folgende Speisespannungen vorgesehen: + 3,3 V, + 5 V, + 12 V, - 12 V. Eine Steckkarte darf eine maximale Verlustleistung von 25 W haben.

*Zur Konfigurationspraxis - wir merken uns:*

- 3,3-V-PCI: Kerbe hinten (nahe dem Slot-Abdeckblech),
- 5-V-PCI: Kerbe vorn,
- *Universalkarten* haben Kerben an beiden Stellen,
- die meisten PCs haben nach wie vor einen 5-V-Bus mit 32 Bits und 33 MHz Bustakt,
- 32-Bit-PCI-Karten passen in 64-Bit-Slots und umgekehrt. Der Betrieb von 64-Bit-Karten in 32-Bit-Slots ist allerdings mit einer - u. U. beträchtlichen - Leistungsminderung verbunden.

- PCI ist von Grund auf ein Plug&Play-Bus. PCI-Karten haben keine Jumper o. dergl. zur Konfigurationseinstellung. Es genügt typischerweise, die Karte zu stecken und den PC einzuschalten (oftmals sind aber noch *Gerätetreiber* (= Software) zu installieren).

## 4.3.4. PCI-X

PCI-X ist eine Weiterentwicklung des PCI-Bus, die vor allem für Systeme des oberen Leistungsbereichs in Frage kommt (32 oder 64 Bits Zugriffsbreite, 66 bis 133 MHz Bustakt). Die Weiterentwicklung besteht vor allem in einem Übertragungsprotokoll, das geteilte Übertragungen (Split Transactions; vgl. Abschnitt 2.3.7) unterstützt.

*Signale und Slots*
Sowohl die Signale als auch die Slots entsprechen der PCI-Spezifikation. In PCI-X-Slots passen auch gewöhnliche PCI-Karten (nur wird dann der gesamte Bus mit der Taktfrequenz der „langsamsten" Karte betrieben).

*Einsatz*
Vor allem in Hochleistungssystemen. Dabei bevorzugt man die 64-Bit-Ausführung (vgl. die Abbildungen 1.20 bis 1.23 sowie 1.32 und 1.33).

*Taktfrequenzen und Datenraten*
Es geht knapp zu. Die 133 MHz sind typischerweise nur dann nutzbar, wenn ein einziger Slot bestückt ist (vgl. die Erklärung zu Abbildung 1.39). Maximale Datenrate bei 133 MHz: 1,06 GBytes/s.

*Unterscheidung zwischen PCI- und PCI-X-Karten*
Daß eine PCI-X-Karte im Slot steckt, erkennt das Motherboard anhand eines zusätzlichen Signals. PCI-Karten belegen dieses Kontaktposition im Slot mit Masse, PCI-X-Karten nicht.

## 4.3.5. PCI Express

PCI Express (PCI XP) ist die Weiterentwicklung zum bitseriellen Hochleistungsinterface. Die Übertragungsfunktionen entsprechen PCI, nur werden die Kommandos, Adressen, Daten usw. nicht über eine Vielzahl von Busleitungen, ssondern bitseriell übertragen. Die PCI-XP-Einrichtungen (Steckkarten oder Schaltkreise) sind einzeln an Schaltverteiler angeschlossen (Punkt-zu-Punkt-Verbindungen; Abbildung 4.11).

*Übertragungsprinzip:*
Für jede Übertragungsrichtung ist ein Signalleitungspaar vorgesehen (differentielle Signalübertragung). Die Übertragungsrate: jeweils 2,5 GBits/s. Da je Byte 10 Bits übertragen werden (8B/10B-Codierung), ergibt sich eine Datenrate von rund 250 MBytes/s (= zweimal 33-MHz-PCI). Es gibt keinen Bustakt*). Statt dessen muß jeder Empfänger seinen eigenen Takt mit den ankommenden Signalen synchronsieren (Taktrückgewinnung).

*): im System steht aber ein Bezugstakt von 100 MHz zur Verfügung, den die PCI-XP-Einrichtungen als Grundlage der Taktaufbereitung ausnutzen können.

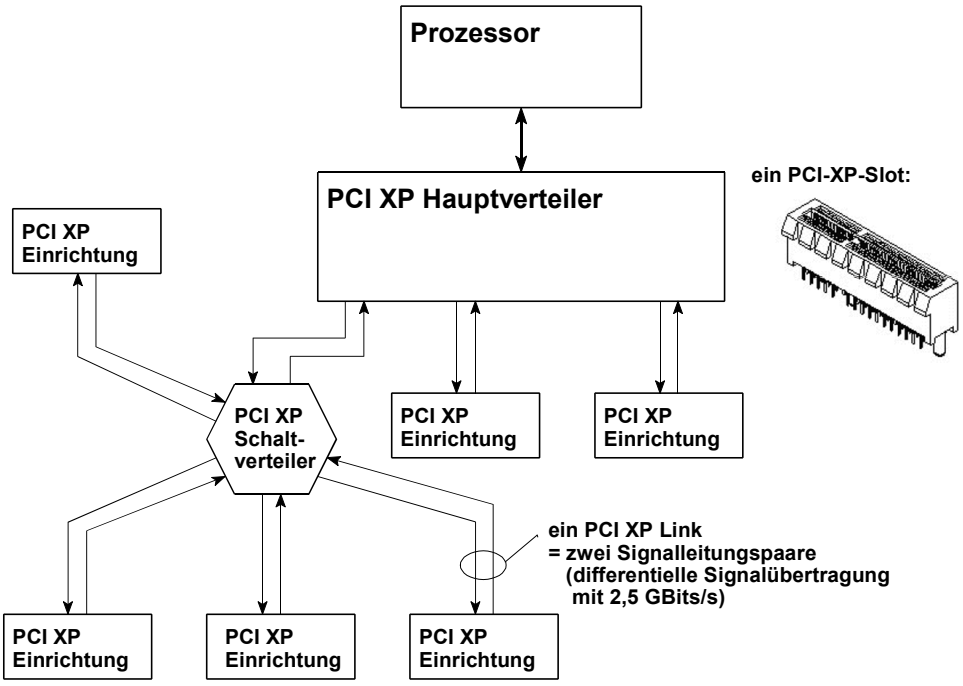



***Abbildung 4.11*** Beispiel einer PCI-XP-Konfiguration

*Extreme Datenraten*
Es ist möglich, mehrere Signalwege gleichzeitig zu nutzen. Ein Signalweg (= 2 Signalleitungspaare) heißt eine Lane (sprich: Lehn). Eine PCI-XP-Verbindung (Link) kann aus 1, 2, 4, 8, 12, 16 oder 32 Lanes aufgebaut werden.

*Slot-Steckverbinder*
Sie sehen ähnlich aus wie die des PCI-Bus. Ihre Länge hängt von der Anzahl der angeschlossenen Lanes ab. Die Speisespannungen: + 3,3 V und + 12 V.

# 4.4. AGP

AGP (= Accelerated Graphics Port) ist ein von Intel entwickeltes Interface zum Anschließen hochleistungsfähiger Videoadapter. AGP ist kein Bus (an den sich mehrere Einrichtungen anschließen lassen), sondern ein Punkt-zu-Punkt-Interface zum Betreiben eines Videoadapters in einer PC-Plattform (Abbildung 4.12).

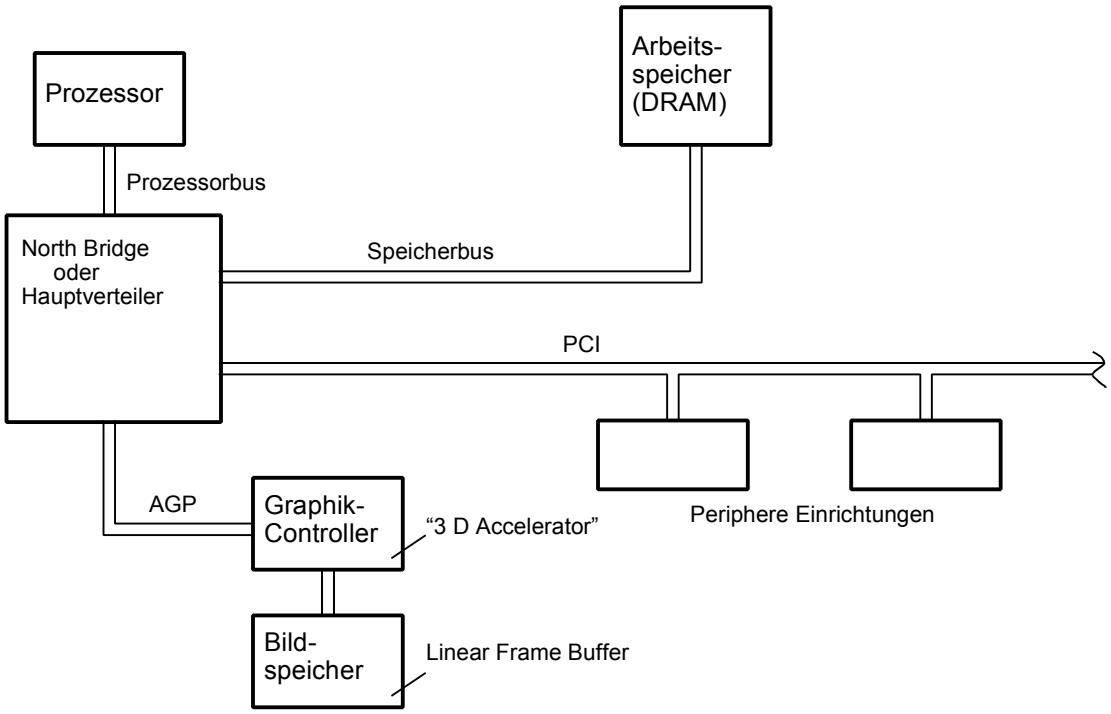



***Abbildung 4.12*** Das AGP-Interface auf einem Motherboard

*Zur „Philosophie"*
Zu den wichtigsten Entscheidungen bei der Auslegung eines Graphik-Subsystem gehört die Arbeitsteilung zwischen Prozessor und Videoadapter oder Graphik-Controller:

- soll der Prozessor „alles" allein tun, vereinfacht sich der Videoadapter zu einer Anordnung, die den Bildspeicher zwecks Darstellung ausliest, Synchronsignale erzeugt und aus den Pixel-Daten Videosignale bildet. Bei Forderung nach höchstem Leistungsvermögen ist es am besten, den Bildspeicher als Teil des Arbeitsspeichers anzuordnen - allerdings nicht im Sinne einer Sparlösung, sondern wirklich auf Leistung orientiert (der Prozessor darf nicht merken, daß er eigentlich auf den Bildspeicher zugreift (auch nicht durch längere Wartezustände)).
- soll der Prozessor weitgehend entlastet werden, brauchen wir einen „intelligenten" Videoadapter mit eigenem Prozessor, genug Speicher usw. Da dann eigentlich nur noch Aufträge zu übermitteln sind, ist es nahezu gleichgültig, an welchen Bus man den Videoadapter anschließt (PCI genügt bei weitem).

Nun sind beide Ansätze in reiner Form weniger geeignet, wenn ganz wichtige Anwendungen realisiert werden sollen, nämlich Spiele mit bewegten Bildern in 3D: um alles allein zu, tun wären auch moderne Prozessoren nicht leistungsfähig genug, und mit „Intelligenz" ist nicht viel anzufangen, wenn es sich um das Wandeln von Pixel-Haufen handelt. (Beispielsweise läßt sich ein Kreis durch Mittelpunktskoordinaten, Radius und Füllfarbe vollständig beschreiben. Wie beschreibt man aber eine Landschaft, in der sich muskelbepackte Prügler, Schießer und Messerstecher tummeln? - Das gelingt praktisch nur, indem man von einer als Pixel-Muster gegebenen Szenerie ausgeht, die fortlaufend umgeformt wird.) Als kostengünstig hat sich folgende Arbeitsteilung erwiesen:

- der Prozessor erledigt die eigentlichen Rechenoperationen (wozu er Gleitkomma- und „Multimedia"-Befehle (MMX, 3DNow, SSE) zur Verfügung hat),
- der Graphikcontroller erledigt die Oberflächengestaltung der darzustellenden Objekte (so daß sie möglichst realistisch aussehen). Die Operationen sind an sich recht einfach, erfordern aber das Bewegen großer Datenmengen.

Der Graphikcontroller ist hierbei nicht nur ein passiver Verwalter des Bildspeichers, sondern eine aktive Einrichtung, die mit eigenen Schaltmitteln auf den Bildinhalt einwirkt. Man spricht deshalb auch vom Graphik*beschleuniger* (3D Accelerator).

AGP wurde als Interface für Graphikbeschleuniger in typischen PCs entwickelt. Dabei ging es nicht nur um die Geschwindigkeit (schneller als PCI), sondern auch um die Kostensenkung. Um kostengünstige Graphikhardware bauen zu können, sollte der Graphikbeschleuniger Teile des Arbeitsspeichers mitbenutzen. Hierzu muß der Graphikbeschleuniger aber selbst als Master wirksam werden.

*AGP oder 66-MHz-PCI?*
Weshalb hat man überhaupt AGP entwickelt und nicht einfach kostengünstige Motherboard-Schaltkreise herausgebracht, die den 66-MHz-Betrieb des PCI-Bus unterstützen?

Nun werden wir wohl mit einer Antwort nicht danebentreffen: weil man sich hierdurch höhere Marktanteile, mehr Gewinn usw. verspricht.

Es gibt aber auch handfeste Begründungen technischer Art:

- die 66-MHz-Spezifikation des PCI-Bus sieht bereits recht harte Einschränkungen vor (Anzahl der Slots, Leitungslängen usw.),
- infolge der elektrischen Spezifikation als „echter" Bus ist eine weitere Leistungssteigerung kaum noch möglich,
- die PCI-Übertragungsprotokolle betreffen sozusagen „klassische" Buszyklen als Einheit von Anforderung (Adressierung, Kommandoübertragung) und Datentransport (ungeteilte Übertragung).

Ein neuer Ansatz (nämlich AGP) erlaubt es hingegen, 2 potentielle Vorteile wirklich auszunutzen:

- ein Punkt-zu-Punkt-Interface hat günstigere elektrische Eigenschaften als ein „echter" Bus[1)] (66-MHz-Betrieb wird eher praktikabel, ja es ist sogar ein Übergang auf 133 MHz möglich),
- man kann leistungsfähigere Übertragungsprotokolle implementieren[2)].

*32 Bits oder 64 Bits?*
Wenn die Datenraten schon nicht hoch genug sein können - weshalb hat man dann kein 64-Bit-Interface vorgesehen?

Auch dem stehen handfeste Tatsachen technischer Art entgegen:

- ein 64-Bit-Interface ist teurer und braucht mehr Platz (Siliziumfläche, Schaltkreisanschlüsse, Schaltkreisgehäuse, Leiterzüge auf dem Motherboard, Slot-Steckverbinder),
- es schalten nahezu doppelt so viele Signale gleichzeitig; demgemäß hat man es mit intensiveren Störungen zu tun (gegenseitige Störbeeinflussung, Störabstrahlung),
- typische Pixel-Daten sind 32-Bit-Strukturen.

Wir finden also auch hier den Ansatz, der u. a. auch den modernen Speicher-Interfaces zugrunde liegt (vgl. Abschnitt 3.2.4):

- Verzicht auf extrem breite Datenwege,
- statt dessen Wahl der höchstmöglichen (d. h. noch beherrschbaren Taktfrequenz),
- Implementierung effektiver Übertragungsprotokolle (Split Transactions, Signalisieren neuer Anforderungen parallel zu laufenden Datenübertragungen über gesonderte „seriell-parallele" Signalwege (Paketprinzip)).

*AGP 1, AGP 2, AGP 3*
Die Spezifikation wurde in diesen Schritten weiterentwickelt (Tabelle 4.6). Die meisten PCs haben AGP 2 oder (von 2002/2003 an) AGP 3.

| | **AGP 1** | **AGP 2** | **AGP 3** |
|---|---|---|---|
| Wirksamkeit | ab 1998 | ab 2000 | ab 2002 |
| Signalisierung | 3,3 V | 1,5 V | 0,8 V (TMS)*) |
| Geschwindigkeiten**) | 1X, 2X | 1X, 2X, 4X | 4X, 8X |
| Steckverbinder | 3,3 V | 1,5 V oder universell | 1,5 V oder universell |

*): mit minimaler Anzahl an Schaltvorgängen (Transition Minimized Signaling); **): 1 X = 264 MBytes/s

***Tabelle 4.6*** Die Entwicklungsschritte der AGP-Spezifikation

*Weitere AGP-Entwicklungen?*
Abwarten. Man spricht vom Übergang auf PCI Express...

## 4.4.1. Signale und Slots

AGP ist ein synchrones zentral gesteuertes Punkt-zu-Punkt-Interface mit zeitmultiplexer Übertragung von Adressen, Kommandos und Daten (Abbildungen 4.13, 4.14). Das Leitungssystem entspricht im wesentlichen dem des PCI-Bus*). Grundlage der elektrischen Spezifikation bildet die 66-MHz-Auslegung des PCI-Bus.

*): die Signale entsprechen (Feinheiten außer acht gelassen) der PCI-Spezifikation. Es handelt sich aber nicht um einen „durchgeschleiften" PCI-Bus, sondern um gesonderte Leitungen.

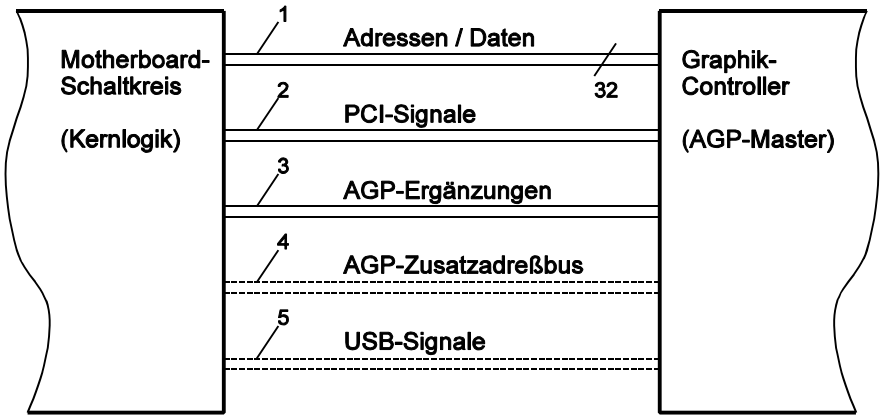



***Abbildung 4.13*** Übersicht über die AGP-Signalwege

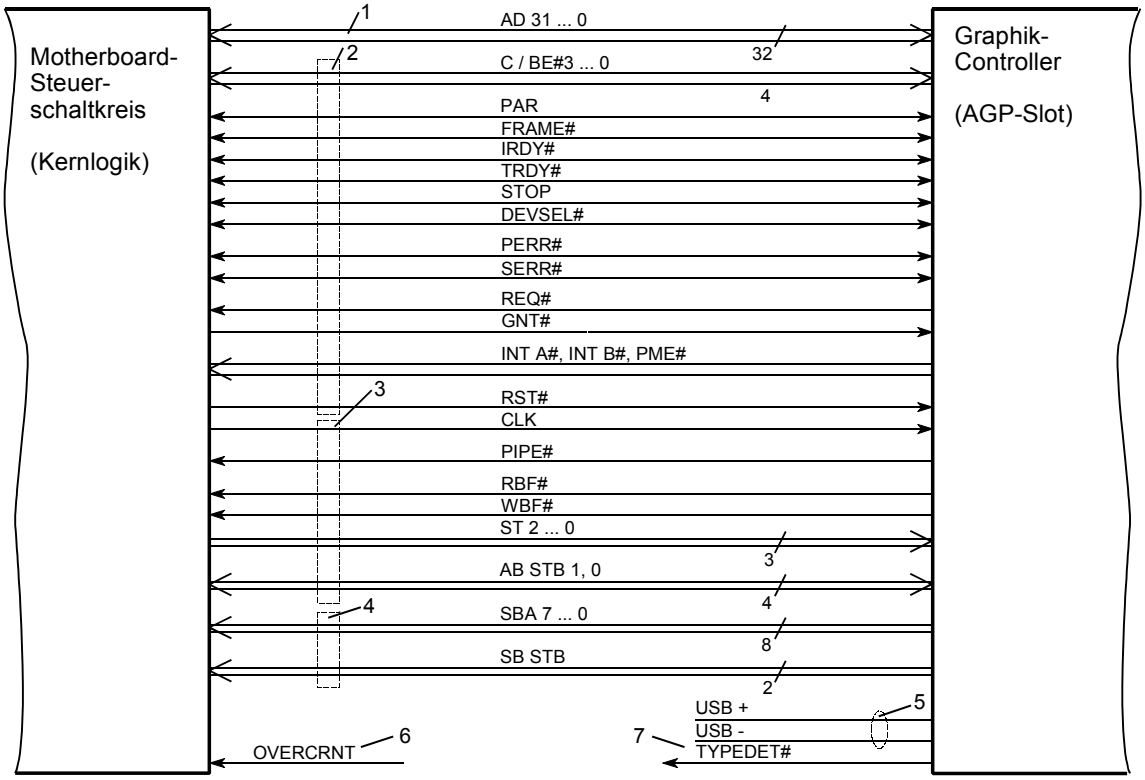



***Abbildung 4.14*** Das AGP-Interface (AGP 2)

*Erklärung (zu beiden Abbildungen):*
Abbildung 4.13 gibt einen Überblick, Abbildung 4.14 zeigt die Signale im einzelnen.

1) Adreß- und Datensignale. Wie beim PCI-Bus ist ein zeitmultiplex genutzter 32-Bit-Signalweg für Adreß- und Datenübertragungen vorgesehen.
2) PCI-Signale. Es sind die wesentlichen Steuersignale gemäß PCI-Spezifikation vorgesehen.
3) AGP-Ergänzungen. Diese Signale sind vorgesehen, um im Verbund mit den Signalen (1) und (2) die AGP-typischen Übertragungsprotokolle zu implementieren.

4) AGP-Zusatzadreßweg (Sideband Address Bus). Über diesen Signalweg kann der AGP-Master gleichzeitig zu laufenden Datenübertragungen neue Adressen und Kommandos übertragen (und zwar auf „serienparallele" Weise in aufeinanderfolgenden Abschnitten zu jeweils 8 Bits).
5) USB-Signale. Diese haben an sich nichts mit dem AGP-Interface zu tun. Sie werden lediglich mitgeführt, um die Bereitstellung eines USB-Anschlusses am Video-Ausgang zu unterstützen.
6) USB-Überstrommeldung,
7) Anzeige des AGP-Signalpegels (3,3 V oder 1,5 V).

*Übertragungsprinzip*
Um höchste Datenraten zu erreichen, hat man bei den AGP-Transaktionen die Adressierung von der Datenübertragung getrennt (Split Transactions). Der Graphikcontroller erteilt zunächst Aufträge (Kommando + Adreßangaben), während die eigentliche Datenübertragung mit maximal möglicher Geschwindigkeit jeweils dann ausgeführt wird, wenn die Daten tatsächlich bereitstehen. Damit ist das Interface während der Zugriffszeiten für anderweitige Nutzung frei. Kommandos und Adressen können wahlweise über die AD-Leitungen oder über den Zusatzbus (die SBA-Leitungen) übertragen werden. Die Nutzung des Zusatzbus ermöglicht es, neue Kommandos und Adressen parallel zum laufenden Datentransport zu übertragen.

*Datenübertragungsgeschwindigkeiten*
AGP-Subsysteme können für 4 Übertragungsgeschwindigkeiten ausgelegt werden:

1. einfache Geschwindigkeit (AGP 1X): Datenübernahme mit der Low-High-Flanke des Bustaktes (CLK). Taktfrequenz: 66 MHz. Maximale Datenrate: 264 MBytes/s.
2. doppelte Geschwindigkeit (AGP 2X): Datenübernahme mit beiden Flanken der Strobe-Signale AD_STB1, 0. Schaltfrequenz: 66 MHz (entsprechend Bustakt). Maximale Datenrate: 528 MBytes/s. Die Strobe-Signale wirken hier gegen Masse („single ended"); die invertierten Strobes werden nicht ausgenutzt.
3. vierfache Geschwindigkeit (AGP 4X): Die Strobe-Signale AD_STB1, 0 bzw. SB_STB schalten mit doppelter Geschwindigkeit (2 · Bustakt), und es werden beide Flanken ausgenutzt. Maximale Datenrate: 1,056 GBytes/s. Dabei werden auch die invertierten Strobes verwendet (differentielle Signalübertragung). Diese Betriebsweise ist nur bei Logikpegeln von 1,5 V nutzbar.
4. achtfache Geschwindigkeit (AGP 8X, AGP 3): Die Strobe-Signale AD_STB1, 0 bzw. SB_STB schalten mit vierfacher Geschwindigkeit (4 · Bustakt). Es werden beide Flanken ausgenutzt. Maximale Datenrate: 2,112 GBytes/s. Diese Betriebsart wurde mit der Spezifikation AGP 3 eingeführt (0,8-V-Signalisierung, zusätzliche Signale zur Änderungsanzeige (Transition Minimized Signaling TMS)). AGP 3 ist abwärtskompatibel zu AGP 4X.

AGP ist vor allem dazu vorgesehen, dem Graphikcontroller schnellen Zugriff auf den Arbeitsspeicher zu ermöglichen. Hierzu wurden folgende Vorkehrungen getroffen:

- im AGP-Betrieb arbeitet der Graphikcontroller stets als Master,

- Anforderung und Datenübertragung sind voneinander getrennt (Split Transactions),
- es können mehrere Übertragungsvorgänge gleichzeitig in Arbeit sein,
- der Master darf weitere Übertragungen anfordern, ohne die Erledigung zuvor angeforderter Übertragungen (Outstanding Transactions) abzuwarten,
- AGP-typische Datenübertragungen betreffen nur den Arbeitsspeicher und kümmern sich nicht um die Cache-Kohärenz.

*AGP Pro*
AGP Pro ist keine Abwandlung des eigentlichen AGP-Interfaces, sondern eine zusätzliche Spezifikation der mechanisch-konstruktiven Auslegung. Hiermit soll (1) mehr Platz geschaffen werden (größere zulässige Bestückungs-Bauhöhe, breitere Slot-Abdeckungen usw.), und es soll (2) möglich sein, mehr Strom zu entnehmen (eine AGP-Pro-Einrichtung darf bis zu 110 W Verlustleistung haben). Der Zweck besteht darin, auf Grundlage des AGP-Interfaces überdurchschnittlich leistungsfähige Graphikhardware bauen zu können (Intel spricht von „advanced workstation graphics"). Es gibt AGP-Pro-Steckkarten und AGP-Pro-Motherboards (vgl. die Abbildungen 1.14 und 1.20).

## 4.4.2. Steckkarten

*Anordnung auf dem Motherboard*
Es sind nur kurze Signalwege zulässig. Der AGP-Slot befindet sich deshalb unmittelbar in der Nähe der North Bridge oder des Hauptverteilers (vgl. die Abbildungen 1.10 bis 1.20).

*Steckverbinder*
Es werden direkte Steckverbinder eingesetzt. Die Kontaktflächen auf der Leiterplatte sind abwechselnd nach oben und unten versetzt (Abbildung 4.15). Der Anschlußabstand (Pin Pitch) beträgt 1 mm. Es sind Ausführungen mit 124 Anschlüssen und mit 132 Anschlüssen (Universal-slots ohne Kerben) vorgesehen. Anhand der versetzten Kontakte lassen sich AGP-Karten auf den ersten Blick von PCI-Karten unterscheiden.

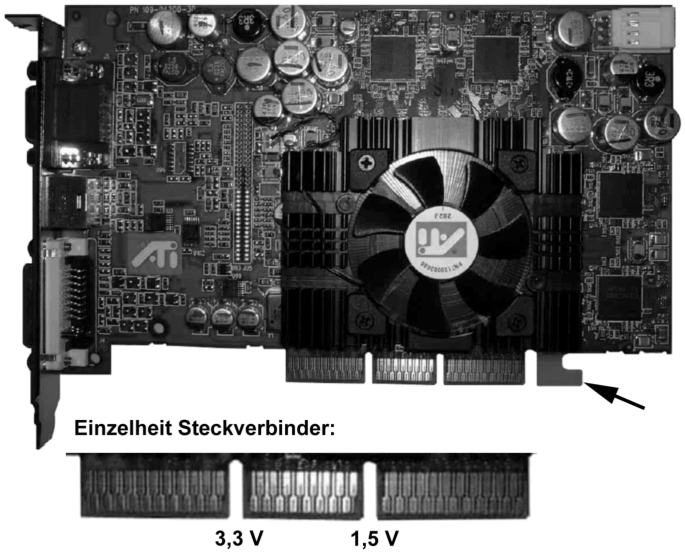


***Abbildung 4.15*** Gaphikkarte mit AGP-Steckverbinder. Die Nase (Pfeil) dient zur Arretierung im Slot

*Speisespannungen*
In einem AGP-Slot sind folgende Speisespannungen vorgesehen: + 3,3 V, + 5 V, + 12 V und Vddq (3,3 V oder 1,5 V). Die Verlustleistung einer AGP-Karte sollte 25 W nicht übersteigen.

*AGP-Pro-Karten*
Der wesentliche Unterschied zu gewöhnlichen AGP-Karten liegt in der extrem breiten rückseitigen Abdeckung (Abbildung 4.16). Hierfür sind 2 Ausführungen spezifiziert, die als High Power Cards und Low Power Cards bezeichnet werden (wir wollen sie kurz „dick" und „dünn" nennen).

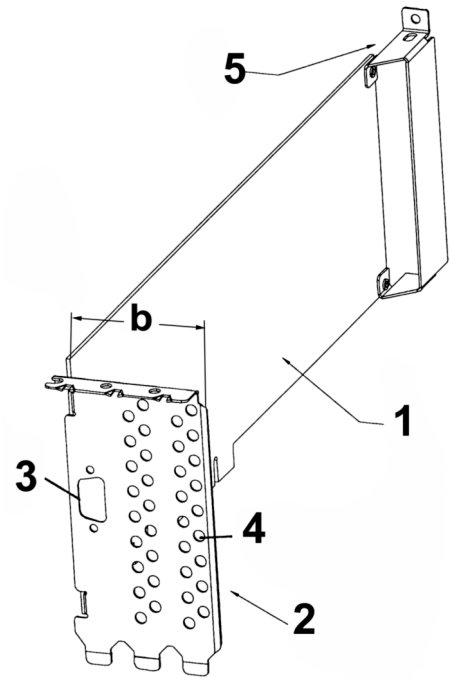



***Abbildung 4.16*** AGP-Pro-Steckkarte (dick), unbestückt mit Befestigungsteilen. Ansicht von hinten auf Bestückungsseite (nach: Intel)

*Erklärung:*
1 - Steckkarte; 2 - rückseitige Abdeckung (I/O Bracket); 3 - Durchbruch für Interface-Steckverbinder; 4 - Luftaustrittsöffnungen; 5 - vordere Halterung (End Retainer Bracket); b - diese Bauhöhe (rund 55 mm) darf zum Bestücken der Karte ausgenutzt werden (für Speichermoduln, für Schaltkreise mit Kühlkörpern und Lüftern usw.).

Der AGP-Slot liegt unmittelbar neben den PCI-Slots. *Dicke Karten* überdecken zwei dem AGP-Pro-Slot benachbarte PCI-Slots, dünne Karten nur einen (hierdurch ergibt sich eine ausnutzbare Bauhöhe von rund 34 mm).

In ihren Grundabmessungen entsprechen AGP-Pro-Karten dem herkömmlichen ATX-Formfaktor. Der AGP-Pro-Steckverbinder ist an sich ein normaler AGP-Steckverbinder, der an beiden Seiten um zusätzliche Stromversorgungskontakte erweitert ist (AGP Pro unterscheidet sich von AGP nur in der erweiterten Speisespannungszuführung).

*Die Strombelastbarkeit von AGP-Pro-Slots*
Die zusätzlichen Kontakte in diesen Slots sind dazu vorgesehen, aus der 3,3-V- und der 12-V-Versorgung höhere Ströme entnehmen zu können: + 3,3 V: maximal 7,6 A, + 12 V: maximal 9,2 A. Die Obergrenze der Stromaufnahme ist durch eine maximal zulässige Verlustleistung von 110 W gegeben.

*AGP-Karten in AGP-Pro-Slots*
Diese Kombination ist möglich (eine AGP-Karte läßt sich in ein AGP-Pro-Motherboard stecken).

## 4.4.3. AGP-Konfigurationen

Sowohl die Motherboards als auch die Steckkarten unterscheiden sich in der Signalisierung (3,3 V, 1,5 V, 0,8 V) und in der maximalen Datenübertragungsgeschwindigkeit (1X, 2X, 4X, 8X). Ähnlich wie beim PCI-Bus wird weitgehend für eine flexible Kombinierbarkeit und gegenseitige Paßfähigkeit gesorgt, und es wird auf mechanische Weise (über Kerben und Sperren in den Steckverbindern) gewährleistet, daß sich Kombinationen, die nicht zusammenpassen, gar nicht erst stecken lassen.Aber Vorsicht - bei AGP 3 gilt das nicht immer (vgl. den Warnhinweis in Anschluß an Tabelle 4.7).

*3,3-V-Auslegung*
Der Slot-Steckverbinder auf einem 3,3-V-Motherboard hat im hinteren Drittel seiner Länge eine Sperre. Eine 3,3-V-Karte hat an der entsprechenden Position eine Kerbe.

*1,5-V-Auslegung*
Der Slot-Steckverbinder auf einem 1,5-V-Motherboard hat im vorderen Drittel seiner Länge eine Sperre. Eine 3,3-V-Karte hat an der entsprechenden Position eine Kerbe. (Es handelt sich im Grunde um den gleichen Steckverbinder. Er ist gegenüber der 3,3-V-Auslegung lediglich um 180° gedreht.)

*Universalauslegung*
Es gibt Universalkarten (mit zwei Kerben) und Universalmotherboards. Diese tragen einen Universalsteckverbinder, der gar keine Kerben hat, so daß sich Karten aller Art stecken lassen. Das Motherboard wertet die Erkennungssignale der Karte aus und stellt demgemäß das AGP-Interface ein.

*Hinweis:*
In der ursprünglichen AGP-Spezifikation waren keine Universalkarten, sondern nur Universalmotherboards vorgesehen. Es wurden aber Universalkarten gefertigt, die von Hand (Jumper) auf 3,3 V oder 1,5 V einzustellen sind.

Die Tabellen 4.7 und 4.8 geben einen Überblick über die verschiedenen Varianten von Motherboards und Steckkarten.

*Achtung:*
AGP3-Motherboards können Schaden nehmen, wenn man versucht, eine 3,3-V-Karte zu betreiben. Solche Karten werden typischerweise nicht unterstützt - auch nicht auf Boards des

Typs Universal AGP3.x. Die einschlägigen Warnhinweise in der Dokumentation beachten!

*Problemstellen:*

- mit einem Universalslot bestückte Motherboards. Da keine Sperre vorhanden ist, wird das Stecken von 3,3-V-Karten nicht verhindert.
- ältere Universalkarten mit zwei Kerben, die in Slots aller Art gesteckt werden können, aber von Hand einzustellen sind.

| Motherboard-Typ | Steckverbinder (Slot) | typische Merkmale |
|---|---|---|
| AGP 3,3 V | Sperre hinten (3,3 V) | nur 3,3-V-Signalisierung; nur 1X und 2X |
| AGP 1,5 V | Sperre vorn (1,5 V) | nur 1,5-V-Signalisierung; 1X, 2X, 4X |
| Universal AGP (UAGP) | keine Sperre | 3,3- oder 1,5-V-Signalisierung; 1X, 2X, 4X |
| AGP 3.x | Sperre vorn (1,5 V) | nur 0,8-VSignalisierung (AGP 3); nur 4X und 8X |
| Universal AGP3.x | Sperre vorn (1,5 V) | 1,5-V- oder AGP3-Signalisierung (0,8 V). Im AGP2-Betrieb 1X, 2X, 4X; im AGP3-Betrieb 4X, 8X. |

***Tabelle 4.7*** Motherboards mit AGP-Slots

| Graphikkarten-Typ | in welchen Slot paßt die Karte? | typische Merkmale |
|---|---|---|
| AGP 3,3 V | 3,3 V und Universal AGP (UAGP) | nur 3,3-V-Signalisierung; nur 1X und 2X |
| AGP 1,5 V | 1,5 V und Universal AGP (UAGP) | nur 1,5-V-Signalisierung; 1X, 2X, 4X |
| Universal AGP (UAGP) | in jeden | 3,3-V oder 1,5-V-Signalisierung; 1X, 2X, 4X |
| AGP 3.x | 1,5 V und Universal AGP3 | nur AGP3-Signalisierung; nur 4X und 8X |
| Universal AGP3.x | 1,5 V und Universal AGP3 | 1,5-V- oder AGP3-Signalisierung. Im AGP2-Betrieb 1X, 2X, 4X; im AGP3-Betrieb 4X, 8X. |

***Tabelle 4.8*** AGP-Karten

## 4.4.4. AGP mit 2 Graphik-Einrichtungen

An den Steuerschaltkreis sind über das AGP-Interface 2 Graphikcontroller angeschlossen, einer direkt auf dem Motherboard und einer auf einer Graphikkarte (Abbildung 4.17). Der Intel-Fachbegriff: *Three Load AGP*.

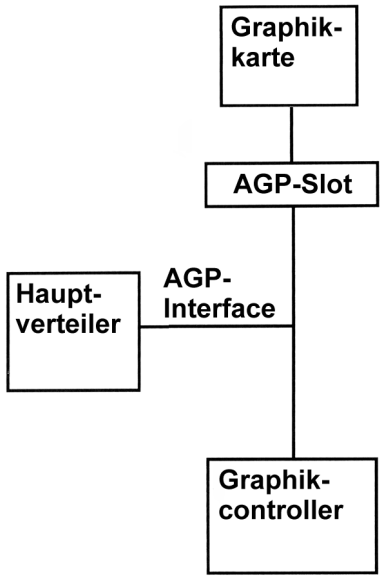



***Abbildung 4.17*** 2 Graphik-Einrichtungen am AGP-Interface (nach: Intel)

Beide Graphikcontroller sind AGP-Master. Da es nur einen Master geben kann, AGP aber keine Arbitrierung vorsieht, muß einer der Graphikcontroller stets außer Betrieb gesetzt sein. Die Anordnung ähnelt den altbekannten Konfigurationen mit einem VGA-Controller auf dem Motherboard, der außer Betrieb gesetzt wird, wenn eine Graphikkarte steckt. Auch der Zweck ist ähnlich: man möchte kostengünstige Motherboards anbieten, die „alles" enthalten (also u. a. auch ein komplettes Video-Subsystem), aber die spätere Erweiterung nicht ausschließen.

*Hinweise:*

1. Die Dokumentation entsprechender Motherboards genau ansehen (vor allem darauf hin, wie man den eingebauten Graphikcontroller außer Betrieb bekommt (z. B. automatisch oder über Setup-Einstellung oder über Jumper).
2. Anordnungen gemäß Abbildung 4.17 eignen sich *nicht* dazu, 2 Bildschirme gleichzeitig zu betreiben.