Duale Hochschule Manheim DHBW

Kurs TINF21AI1

Rechnerarchitekturen I

Speicher

Halbleiterspeicher

Die älteste Form der nicht flüchtigen Speicherung ist der ROM. ROM Bausteine haben lange Zeit in vielen Bereichen eine wichtige Rolle gespielt, die über das reine Speichern hinausgehen. Beispielwese konnte man mit ihnen elektronische Schaltwerke entwickeln, welche sich später weiter zu programmierbaren Logik Devices weiterentwickelt haben.

Die zweite Form ist der RAM welcher Computern einen kurzfristigen Speicher zur Verfügung stellt.

Eine dritte Speicherform sind die Cache Speicher, die sich aus RAM Bausteinen entwickelt haben um die Performance zu beschleunigen.

ROM (Read Only Memory - Nicht flüchtiger Speicher)

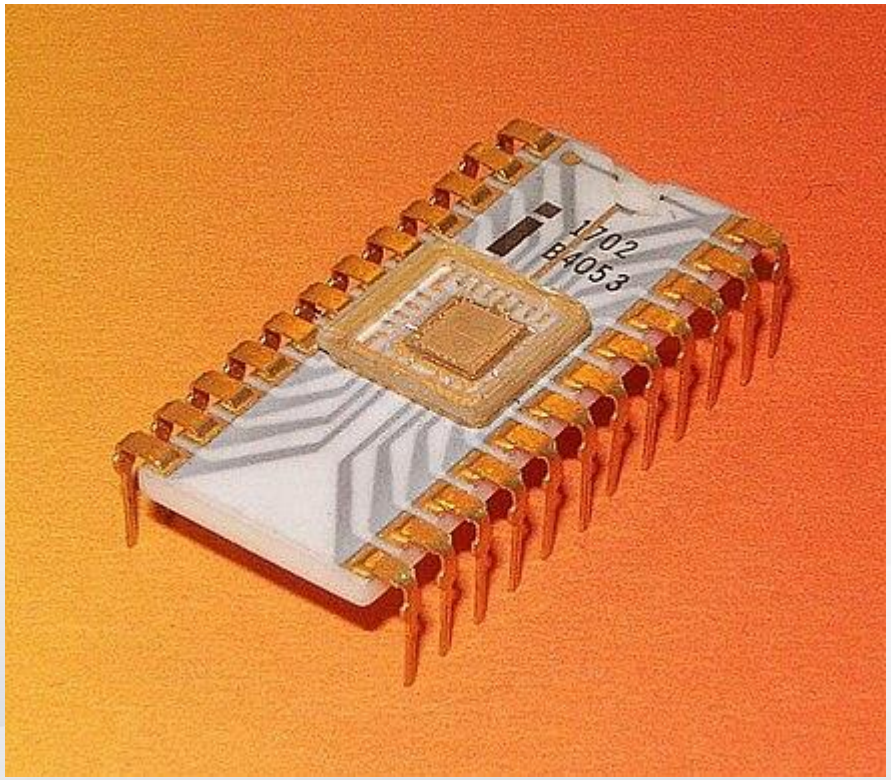
Das ***Masken*** ROM wird im Rahmen der Herstellung programmiert. Hierbei wird das Programm speziell für diesen Chip entwickelt. Nach der Herstellung sind keine Änderung mehr möglich. Diese Form hat sich früher als BIOS Chip oder Firmware auf den entsprechenden Platinen befunden. ROM Bausteine haben in der heutigen Zeit eine immer unbedeutendere Rolle, da schreibare ROM oder andere Speicherlösungen an deren Stelle getreten sind.

PROM (Programmable Read Only Memory)

In PROM ist nach Herstellung kein dediziertes Programm enthalten. Dieses wird erst mit einem speziellen Schreibgerät einprogrammiert. Einmal hochgeladen besteht dann keine Möglichkeit mehr eine Änderung vorzunehmen. Die Programmierung erfolgt dabei über eine Überspannung die an den Kreuzungspunkten angelegt wird. Diese Art der ROMs wird heutzutage kaum noch verwendet, da moderne Lösungen wie EPROM, SSD, PLA, Micro Controller, usw. weitaus flexiblere Lösung darstellen.

EPROM (Eraseable Programmable Read Only Memory)

Das erste PROM wurde bei Intel von Dov Frohman entwickelt. Es wird mit einem Programmer programmiert und kann per starkem UV Licht gelöscht werden. Die Programmierung erfolgt hier bei über eine Überspannung die an den Kreuzungspunkten angelegt wird. Das Fenster muss im Wirkbetrieb überklebt sein, damit keine Unwillkürliche Löschung erfolgt. Hierbei gilt zu beachten, das nach dem Löschen der Chip abkühlen muss, das sonst das Chipdie überbeansprucht wird. Es wird auch nur noch wenig verwendet, da es recht umständlich zu handhaben ist.



Christian Bassow Wiki

EEPROM (Electrically Erasable Programmable Read Only Memory)

Das EEPROM kann BIT Weise programmiert und gelöscht werden. Es hat sich in vielen Bereichen durchgesetzt und ist z.B. auf jedem Mainboard zu finden. Es spielt auch im MicroController Bereich eine wichtige Rolle, da die meisten Controller einen EEPROM enthalten als Programspeicher. Diese werden aber Zusehens von FLASH Speichern verdrängt.

Flash-Speicher

EEPROMs sind teuer und unhandlich. Eine moderne Alternative zu EEPROMS sind die von Toshiba und Intel entwickelten FLASH-Speicher. Dieser Speicher lässt sich schnell und in Blöcken auslesen.

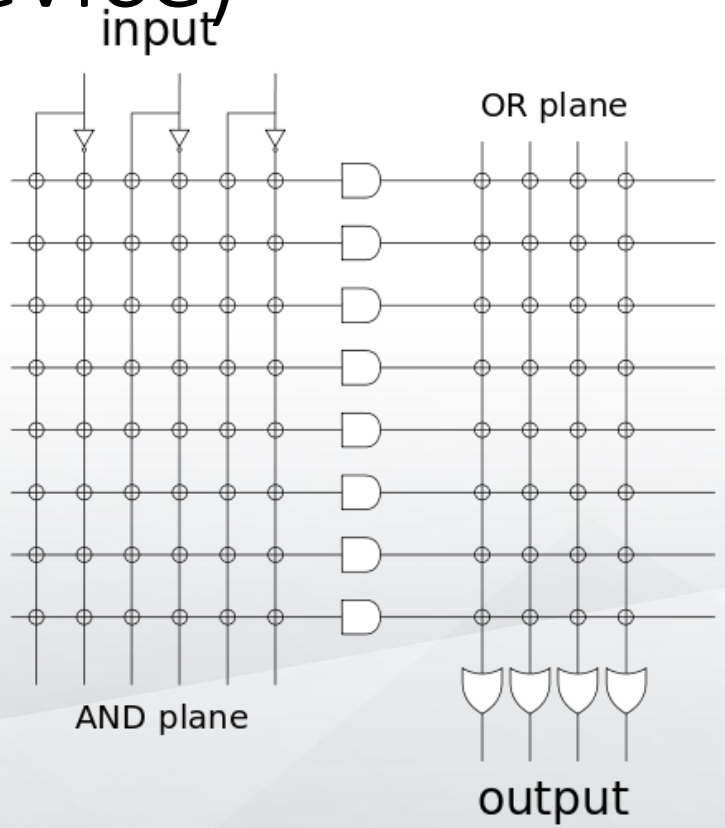
Hierbei wird in zwei Arten unterschieden: NAND und NOR Flash Speicher

- NAND ist nicht linear adressierbarer Speicher
- NOR ist linear adressierbarer Speicher

Eine negative Besonderheit stellt dabei die begrenzte Lebensdauer dieses Speichertyps dar.

Programmierbare logische Bauelemente (PLD) (Programmable Logic Device)

Aus dem Gedanken ROM auch für Programmierung einzusetzen sind PLAs entstanden. ROM kann man als UND und ODER Matrix verstehen. Wobei die UND-Matrix fest verbunden und der ODER Teil frei programmierbar ist. Die Matrix wird verwendet um entsprechende Gleichungen abzubilden. In PLDs wird dies dann weitergeführt. Dadurch ergibt sich dann die Möglichkeit komplexere Strukturen abzubilden und anzusteuern.



Ilia Kr. Wiki

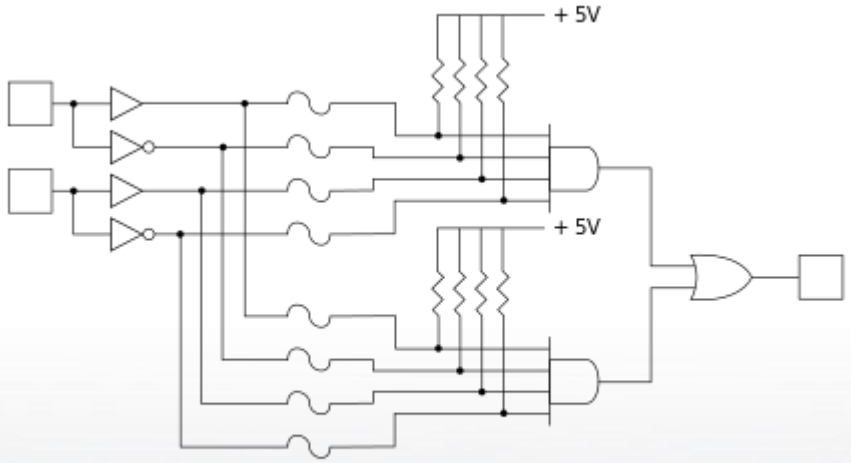
Programmable Arrays Logic

PALs waren die erste nur für diesen Zweck entwickelten Bausteine. Die Programmierung erfolgte über den UND-Teil, der ODER-Teil war fest. PALs konnten ebenfalls nur einmal, wie PROM—Bausteine, programmiert werden.

Die nächste Weiterentwicklung waren GAL (Generic Arrays Logic), die auch wiederbeschreibbar waren.

Bei PLAs (Programmable Logic Array) war auch der ODER-Teil programmierbar.

Auch diese Bausteine sind in modernen Systemen irrelevant geworden.



Simplified programmable logic device

DnetSvg at English Wikipedia

CPLD (Complex Programmable Logic Device)

CPLDs enthalten einen Ein- und Ausgabe Block in dem nach Bedarf komplexere Baugruppen und/oder Flipp Flops enthalten sind. Sie enthalten auch eine programmierbare UND- und ODER-Matrix, die mit einer programmierbaren Rückführungsmartix versehen ist.

FPGA (Field Programmable Gate Array)

FPGA sind deutlich komplexer als CPLDs. Sie bestehen aus vielen Logikelementen (Flip Flops, Logikschaltungen usw.). Darüber hinaus verfügen sie über komplexe Routing- und Speicherkonfigurationsoptionen. Außerdem haben sie die Möglichkeit mit PLL einen eignen Taks zu erzeugen oder anzupassen. Das Programm welches im CPLD enthalten ist kann ebenfalls verändert werden.

Diese Ansätze finden sich teils auch in modernen CPUs, was es möglich macht diese zu patchen.

Des weiteren finden diese Bausteine auch Verwendung in der Entwicklung von neuen CPUs, Grafikkarten, High-Speed-CPUs, Systems-on-chip etc.

Nach der Entwicklung wird der Code aus dem FPGA, dann in VHDL transformiert.

ASIC (Application-Specific Integrated Circuit)

Bei ASIC handelt es sich um Masken programmierte Logik Schaltungen, die vorangegangene Elemente enthalten.
Darüber hinaus können noch andere Elemente wie CPUs, Analoge Systeme usw. enthalten sein.
Vorteil der ASICs ist, das sie auf das Problem optimiert sind.
Es gibt auch gewisse Geschwindigkeitsvorteile.
Diese werden nach Kundenwunsch erstellt und sind nicht veränderbar.

Dies macht sie auch günstiger als eine CPU Lösung.

RAM (Random Access Memory - Flüchtiger Speicher)

RAM ist ein flüchtiger Speicher, der nach Stromverlust seine Daten verliert. Früher gab verschiedene Bauformen, wie z.B. FPM, EDO, usw. Die aktuell verwendete Bauform ist er synchrone DDR RAM.

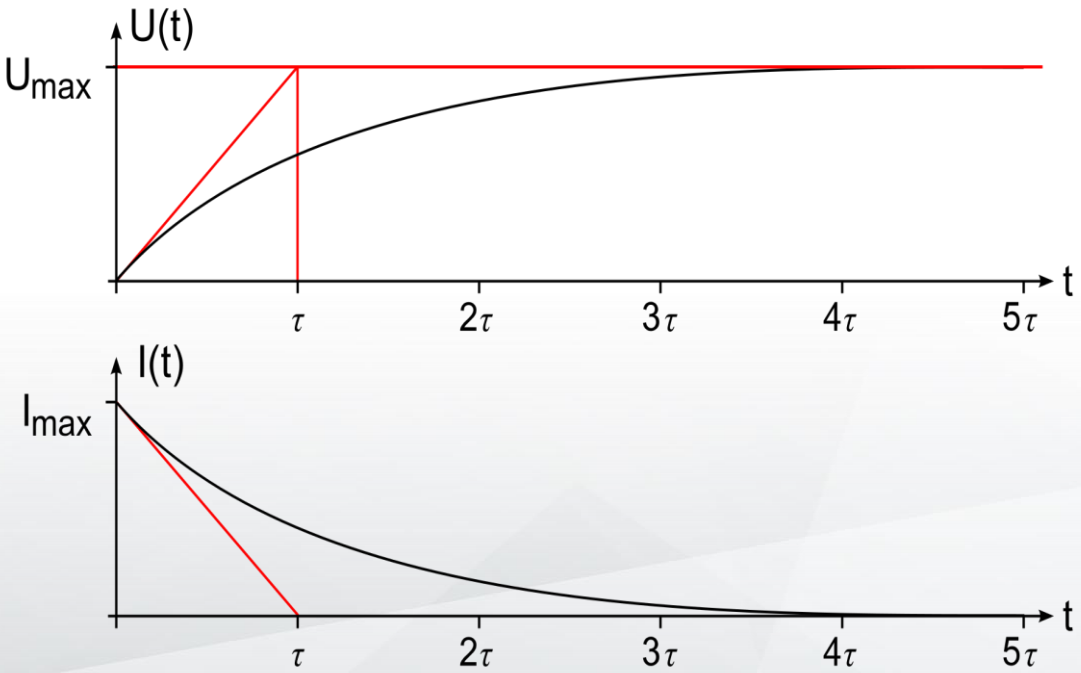
Synchron steht hierbei für: synchron zum Systemtakt (CPU). Zudem werden es noch Unterscheidungen zwischen dynamischen und statischen RAM getroffen:

- Statischer RAM verwendet ein D Flip Flop zur Speicherung
- Dynamischen RAM nutzt einen Kondensator zur Speicherung der Daten

Kondensator

Was ist ein Kondensator?

Bei anlegen einer Gleichspannung lädt sich eine Kondensator bis zu seinem technischen Maximum auf. Der Kondensator entlädt sich, wenn man an zwei seiner Enden einen beliebigen Widerstand anschließt. Der Entladevorgang benötigt dann eine gewisse Zeit. Ein Kondensator entlädt sich mit der Zeit, da die Ladung verloren gehen. Der Entladevorgang wird durch Wärme begünstigt.

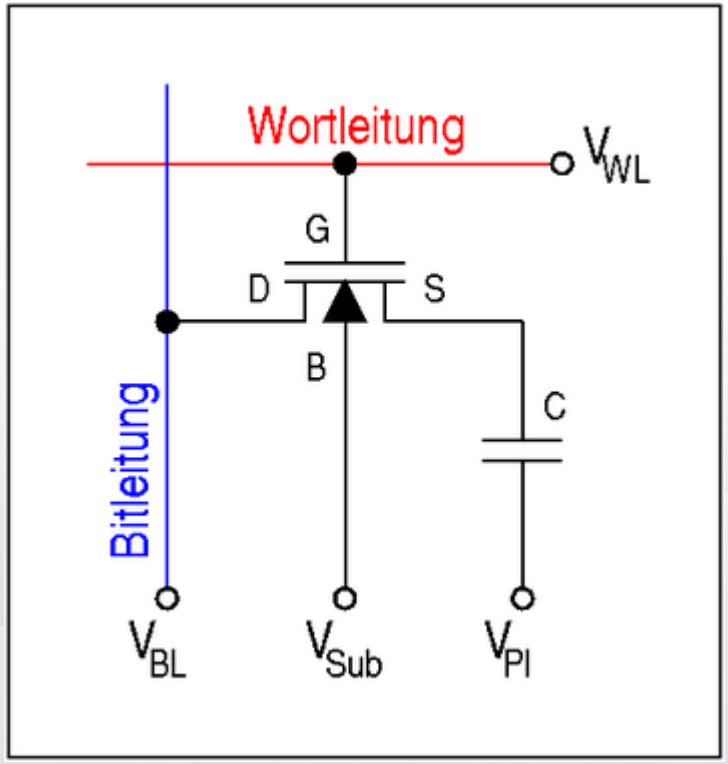


Von Honina.Frank Murmann at de:Wp - selbst vektorisiert, Vorlage: Bitmap von Template:Ud:Honina, Gemeinfrei, https://commons.wikimedia.org/w/index.php?curid=17971380

RAM Zelle

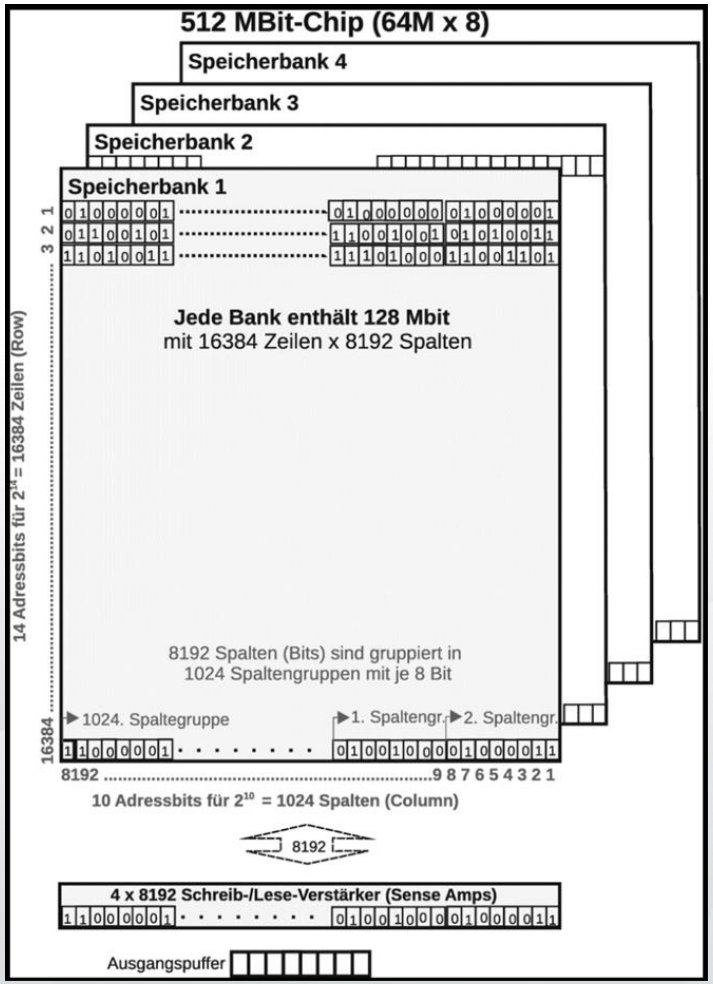
In diesem Bild ist eine 1T1C-Zelle zu sehen. Wobei T für einen Transistor und C für einen Kondensator steht.
Die 1 wird mit einem geladenen Kondensator dargestellt und die 0 mit einem leeren Kondensator. Wenn ~~ich~~ eine Information aus der Zelle ausgelesen wird, wird diese ebenfalls geleert. Daher muss ~~ich~~ diese Information wieder zurückgeschrieben werden. Aus diesem Grund ist „normaler“ RAM langsamer ist als statischer RAM.

Wie jedoch zu erkennen ist, wird nur ein einziger Transistor benötigt. In statischem RAM hingegen werden 6 bis 8 Transistoren benötigt.



Ranks

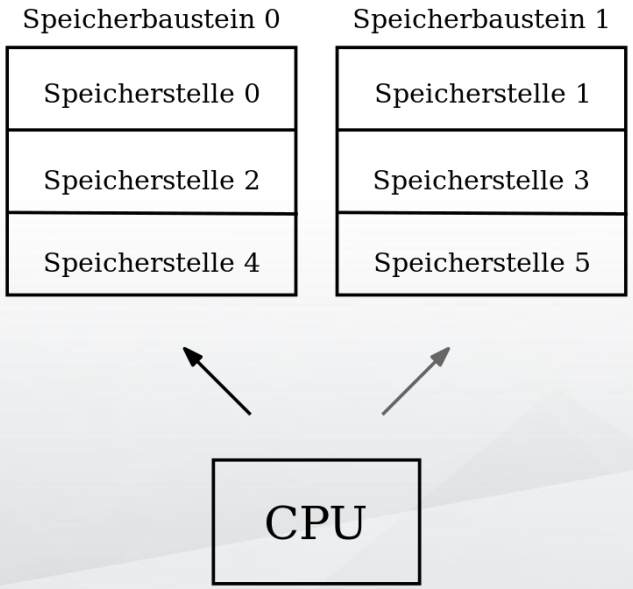
RAM Speicherbausteine sind in der Regel in Ranks organisiert. Das bedeutet, es stehen 4 und 8 Speicherbänke zu Verfügung, die ausgelesen werden können. In jeder Bank werden 4 (8) Bits bereitgestellt. Diese liegen dann im Schreib/Leseverstärker bzw. Sense Amps an und werden dann für die angeforderte Speicheradresse im Ausgangs Buffer abgelegt. Darüber hinaus muss der Chip die Zellen aktualisieren. Dies Erfolgt in der Regel im Intervall von ms. Beim Zugriff wird erst die Zeilen Adresse vom Speichercontroller übertragen und danach die Spaltenadresse. Dies ist ein großer Unterschied zu Statischen RAM.



Bits und Bytes in Mikrochips Klaus Brüderle

Memory interleaving

Um größere Bandbreiten im RAM Zugriff erreichen zu können die Speichermodule verschränkt werden. Es werden 4 Zyklen benötigt um die Daten aus dem RAM auszulesen. Dazu können die Speichermodule in Gerade und Ungerade aufgeteilt werden. Begonnen wird mit dem Lesen des geraden Moduls. Zwei Zyklen später, im 3. Zyklus, wird das ungerade Modul gelesen. Dieser Vorgang wird wiederholt, bis alle Daten verarbeitet sind. So wirkt es, als würden RAM Module in 2 Zyklen arbeiten. Dies setzt einen speziellen Speichercontroller voraus und wird meisten im Grafikkarten Bereich verwendet.



Von Biezl - Eigenes Werk, Gemeinfrei, https://commons.wikimedia.org/w/index.php?curid=10537053

Übersicht SDRAM (Synchronous Dynamic Random Access Memory) / DDR1

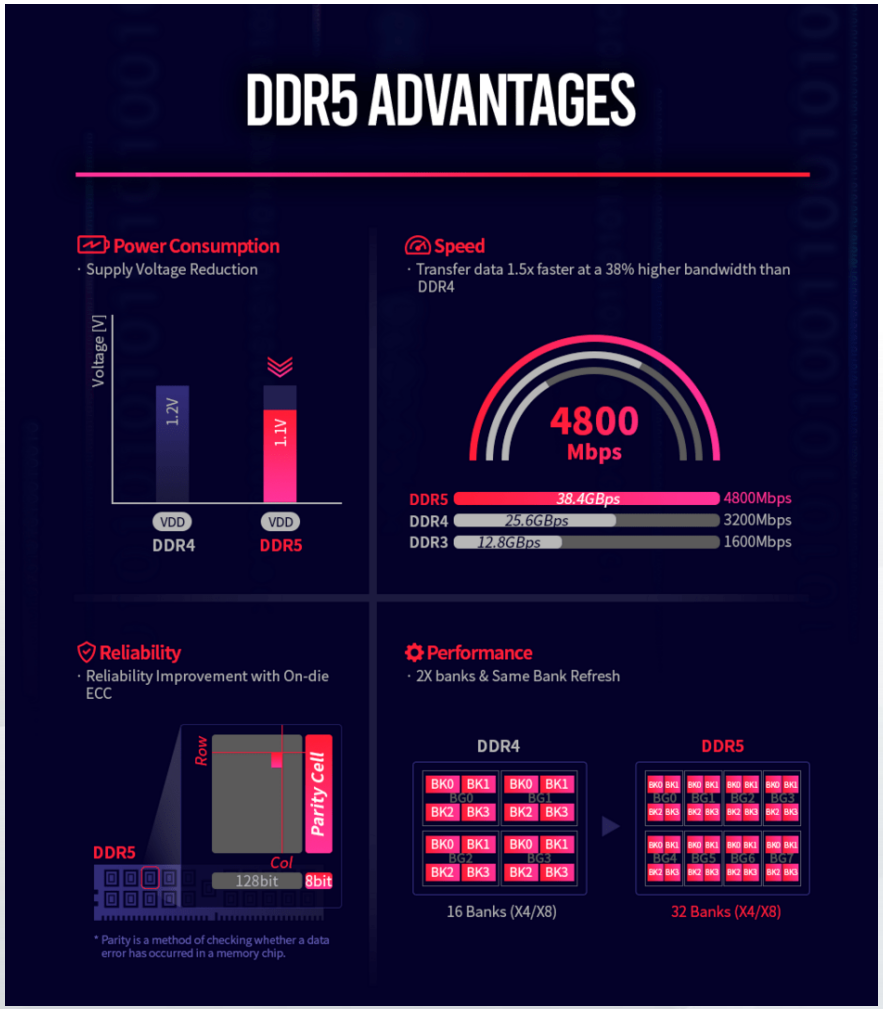
		Spericher-Taktfrequenz	Flanken	Prefetch	Takt		Bandbreite
PC66	SDR	66 MHz	1	1	66 MHz	64 Bit	0,50 GByte/s
PC100	SDR	100 MHz	1	1	100 MHz	64 Bit	0,75 GByte/s
PC133	SDR	133 MHz	1	1	133 MHz	64 Bit	0,99 GByte/s
PC150	SDR	150 MHz	1	1	150 MHz	64 Bit	1,12 GByte/s
PC166	SDR	166 MHz	1	1	166 MHz	64 Bit	1,24 GByte/s
PC1600 (PC200)	DDR1	100 MHz	2	2	200 MHz (DDR200)	64 Bit	1,49 GByte/s
PC2100 (PC266)	DDR2	133 MHz	2	4	266 MHz (DDR266)	64 Bit	1,98 GByte/s
PC2700 (PC333)	DDR3	166 MHz	2	8	333 MHz	64 Bit	~ 2,7 GByte/s
PC3200	DDR4	Ab 200 MHz	2	8	400 MHz (DDR400)	64 Bit	~ 3,2 GByte/s
PC3200	DDR5	Ab 200 Mhz	2	16	Siehe Tabelle	64Bit	>25,6 Gbyte/s

Speichermodul	Taktfrequenz	Transferrate
DDR5-3200	1,6 GHz	25,6 GByte/s
DDR5-3600	1,8 GHz	28,8 GByte/s
DDR5-4000	2,0 GHz	32,0 GByte/s
DDR5-4400	2,2 GHz	35,2 GByte/s
DDR5-4800	2,4 GHz	38,4 GByte/s
DDR5-5200	2,6 GHz	41,6 GByte/s
DDR5-5600	2,8 GHz	44,8 GByte/s
DDR5-6000	3,0 GHz	48,0 GByte/s
DDR5-6400	3,2 GHz	51,2 GByte/s

DDR4 hat keinen großen Sprung gemacht, aber hat neue Technologien eingeführt die ab DDR5 zu tragen kommen.
Die Speicherbandbreite konnte gesteigert werden, dies ist nun bei DDR5 angekommen. Mit den DDR wurde nicht mehr ein BIT je Zyklus übertragen sondern man hat diese auf 2 Bits und in DDR5 auf einen Prefetch von **16(32)** gesteigert. Die Bandbreite ist im normalen Betrieb nicht so stark zu spüren, das der CPU Cache dieses gut mit dem Cache kompensiert. Nur in speziellen Anwendungen und Szenarien wirkt sich diese Bandbreite stark aus. Moderne DDR Rams verfügen über einen eigenen Controller und ein SPD-EEPROM, das die Daten des Moduls vorhält. Dies kann vom Main Board und deren Anwendungen verwendet werden.

DDR5 ist die aktuellste Bauform

Mit DDR5 wurde die Spannung der Module gesenkt um Energiesparender zu sein und der Wärmeentwicklung entgegen zu wirken. Dies schont auch die Kondensatoren. Mit dem Anstieg der Speicherkapazität eines RAM Bausteins, musste auch dem Problem der RAM Fehler begegnet werden. Hierzu sind interne Parity Checks implementiert, der selbsttätig Fehler korrigiert. Es wurde auch ein erweitertes Power-Management implementiert um eine stabile Spannungsversorgung zu erreichen. Die Hersteller gehen dazu über die Chips zu stapeln um mehr Speicher unterzubringen.



Golem.de

Kennwerte

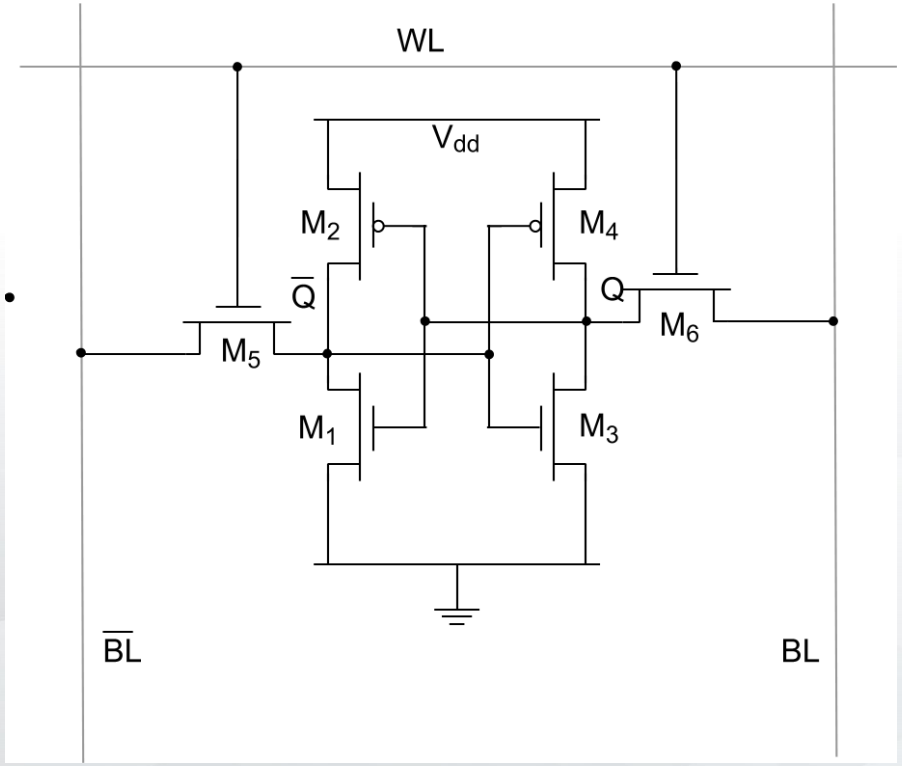
Column Access Strobe Latenz tCL	Senden des Befehls an Speicher und Begin der Antwort!
RAS precharge tRP	Minimale Zeit um auf die nächste Zelle zuzugreifen. Bei einem Fehler muss ein neuer Zyklus gestartet werden => tRP +tRCD + CL Zeit
RAS-to-CAS Delay tRCD	Die benötigte Zeit zum Lesen des Speichers
Row Active Time tRAS	Beschreibt die Zeit, die benötigt wird, bis die Daten im Ausgangspuffer vorliegen, um auf sie Daten zu zugreifen

ECC (Error Correction Code) Speicher Module für Server

In einem professionellen Einsatzgebiet, wie z.B. Servern, gibt es RAM Bausteine mit einer zusätzlichen Paritätsprüfung nach Hamming. Damit können Fehler frühzeitig erkannt und beseitigt werden. Dies ist in der Regel für ein Bit möglich. Mann kann diese ECC-Module an der ungeraden Anzahl an RAM Bausteinen erkennen.

SRAM (Static random-access memory)

SRAMs verbrauchen im Gegensatz zu DRAM deutlich mehr Transistoren. 6T-SRAM-Zelle sind die kleinsten mit einer Mindestzahl von 6 Transistoren. SDRAMs sehr schnell und finden überall dort Anwendung wo es auf Geschwindigkeit ankommt. Wie Buffer, Cache, Register, FIFO, BIOS (mit Batterie) u. a. Was sie auch schneller macht ist, das die Speicheradresse in einem Stück übertragen wird. Das ist ein großer Unterschied zu DDR-RAM.



Von Abelsson - Eigenes Werk, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=504547

Cachespeicher

Zur Steigerung der Performance wurden sehr schnelle in CPUs, Festplatten, RAM usw. schnelle RAMs, sog. Caches integriert. Ein grobe Einteilung hierzu:

- Kleiner L1 Cache: Schnelle Flop Flops die in einem Zyklus antworten können
- Mittelgroßer L2 Cache: SRAM.
- Größere L3 Caches
- Es gibt auch Systeme mit einem L4 Cache. Dieser steht nur für interne CPU Grafikkarten zur Verfügung.

Die L1, L2 und L3 Caches können in zwei Gruppen aufgeteilt werden.

- **Inklusiven Cache:** Alle Daten aus L1 sind auch in L2 und L3. L2 Daten sind auch in L3 vorhanden. Wenn die Daten in L1 verdrängt wurden, dann sind werden diese nicht in L2 verdrängt.
- **Exklusiven Cache:** Alle Daten sind einzigartig in allen drei Cache leveln vorhanden. Wird ein Datensatz in einem Level verdrängt muss diese im Level absteigen.

Speicher Hierarchie

Type	Größe	Bandbreite	Bandbreite
Register	16 bis 32	1 Zyklus	
L1-Cache	(32 kByte)	199.541 MByte/ s	ca. 200 GByte/ s
L2-Cache	(256 kByte)	49.067 MByte/	ca 49 GByte
L3-Cache (SmartCache)	3 MByte	36.952 MByte/ s	40 GByte/ s
Hauptspeicher	8 GByte	14.672 MByte/ s	15 GByte/ s
Festplatte	2 TByte	200 MB/ s	0,20 GByte/ s

Caches in CPUs

In CPUs ist der L1 Cache so implementiert, dass hier nur ein Wert gespeichert wird. Mittlerweile sind L1 Caches in einen Data- und Code Cache aufgeteilt, was für eine Harvard Architecture spricht, da Daten und Code getrennt sind.

L2 / L3 Caches halten eine Seite. Auch gibt es hier Unterschiede in der Art von Konsistenz.

Der L3 Cache befindet sich in einer Wechselbeziehung zum Speichercontroller und in Mulitcore Systemen, zu den anderen L2 Caches der anderen Kernen.

Für Programmierer ist es wichtig zu Wissen, dass Code den Erfolg des Caches negativ beeinflussen kann.
Moderne Programmiersprachen verwenden darüber hinaus meistens einen eigenen Speicher Manger, der versucht effizient auf den Speicher zuzugreifen.

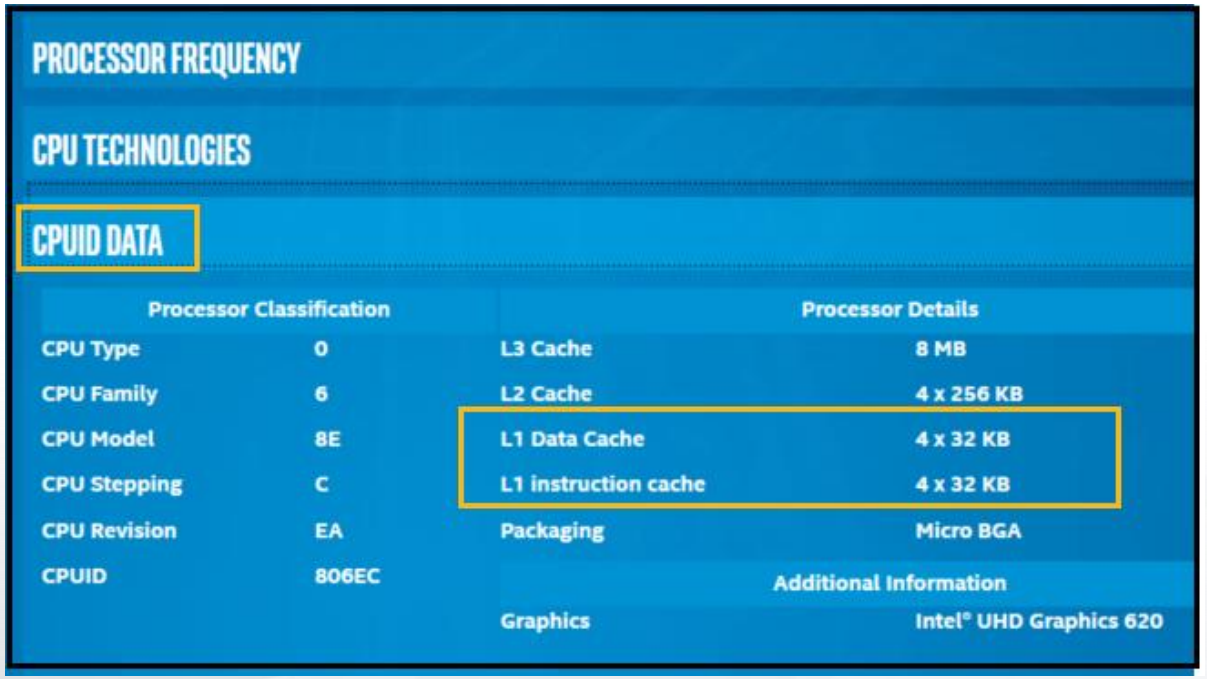


Bild Intel

Cachespeicher Bewertung des Zugriffes

Klassifikation von Cacheverhalten:
Hit Rate:

- Erfolgreiche Zugriff auf den Cache

Miss Rate:

- **Capacity Miss:** Daten wurden aus dem Cache verdrängt und werden wieder abgefragt
- **Conflict:** Der lesenden Datenblock ist nicht genug und man brauch weitere Daten. Es ist aber noch in anderen Bereichen des Caches Platz
- **Cold Start Miss (Compulsory):** Im Cache ist noch Platz, aber die Daten sind nicht da und müssen geladen werden. Dies beim erstmaligen Zugriff auf die Daten.

Cachespeicherverhalten beim Lesen und Schreiben

Man unterscheidet zwischen zwei Zugriffsarten beim Cache.

Verhalten des Caches beim schreiben:

- **Write-through:** Direktes schreiben auf das Speichermedium
- **Write-back:** Hält die Daten im Cache und schreibt diese später auf das Speichermedium

Dies kann man unter Unix Systemen mit dem Befehl sync erzwingen.

Verhalten des Caches beim schreiben (Verdrängungsverfahren):

- **First In First Out (Fifo)** Der älteste Eintrag wird verdrängt.
- **Least Recently Used (LRU)** Der Eintrag der am längsten nicht verwendet wird, wird verdrängt.
- **Least Frequently Used (LFU)** Der Eintrag, der am seltensten genutzt wird, wird verdrängt.
- **Random** Nach eine Zufallswert wird ein Element aus dem Cache entfernt. (Geringer Stromverbrauch)
- **CLOCK** Die Daten werden nach einer Miss und Zeitwert bewertet
- **Statistik** Mit dem Einzug der Statistik in der moderne Datenverarbeitung KI, usw. werden entsprechende Verfahren zur Cachesteuerung verwendet. Es werden hiermit spekulativ Daten geladen.

Speicher Controller

In den ersten CPUs war der Zugriff auf den Speicher relativ simpel gehalten. Mit steigender Komplexität hat sich eine eigene Chip Klasse entwickelt, der Speicher Controller.

Der Speicher Controller ist für das Management des Speichers der CPU zuständig.

Der Controller überwacht, das jeweils nur eine Funktionseinheit schreibt oder lesen kann. Der Controller steuert den RAM an um ihn zu refreshen, zu adressieren und die Daten zu übertragen read/write.
Der Controller managend den DMA Zugriff, damit mit anderen Einheiten und der CPU keine Kollisionen entstehen.

Im Controller können auch Einheiten enthalten sein die ECC auswerten oder weitere Funktionalitäten bereitstellen.
Der Speicher Controller ist heute in der Regel in der CPU integriert, da hiermit eine besser Perfomance erzielt wird.

DMA Controller

DMA Controller oder: Wie kommen meine Daten in den Speicher ohne CPU?

Wenn Daten durch den Prozessor geleitet werden, dann leidet meist die Performance darunter.
Um dieses Problem zu lösen, wurden DMA Controller entwickelt.
DMA Controller leiten Daten direkt in den Speicher, ohne die CPU zu belasten. Dieser liegt extern und meldet sich beim Speicher Controller an wenn er Daten sendet oder abfragen soll.
Der Speichercontroller gewährt den Zugriff auf den Speicher mit Hilfe des Arbiter Mechanismus.
Er steuert dann alles bis die Daten in den Speicher geschrieben sind und schaltet sich dann wieder ab.

Für das sammeln und weiterleiten der Daten, gibt es einige verschiedene Strategien.

Quellen

https://www.mikrocontroller.net/articles/Hauptseite

https://www.youtube.com/watch?v=lvRrCGeGC74

https://www.youtube.com/watch?v=0Ho4rDswOeE&list=PL0pU5hg9yniZ2ka-XBXROXNR0pAEAEFCB

https://en.wikipedia.org/wiki/Main_Page
https://de.wikipedia.org/wiki/Wikipedia:Hauptseite

Brüderle, Klaus Bits und Bytes in Microchios, Springer Fachmedien

Klaus Fricke, Digitaltechnik, SpringerView

Tietze, Schenk, Halbleiter-Schaltungstechnik, Springer-Verlag