Paměti počítačů

Studijní materiál pro předmět
Architektury počítačů a paralelních systémů

Ing. Petr Olivka, Ph.D.
katedra informatiky FEI VŠB-TU Ostrava
email: petr.olivka@vsb.cz

Ostrava, 2020

1 Paměti počítačů

1.1 Základní historický přehled

- V roce 1955 fungovala **feritová paměť** na pricipu zmagnetizovaných feritových jader.
- V **bubnových pamětech** byl magnetický materiál nanesen na nemagnetický buben, který se otáčel vysokou rychlostí.
- **Bublinové paměti** byly magnetické paměti, které jsou založeny na využití velkokapacitních magnetických posuvných registrů.
- **Polovodičové paměti** byly jednobitové a vícebitové posuvné registry.
- V roce 1960 byla vyvinuta **polovodičová technologie MOS**.
- V roce 1970 byly představeny **DRAM** a v roce 1971 **SRAM** paměti.

1.2 Dělení

Podle typu přístupu mohou být paměti rozděleny na:

- RAM (Random Access Memory) - paměti s libovolným přístupem.
- SAM (Serail Access Memory) - paměti se sériovým přístupem.
- Paměti se speciálními způsoby přístupu - asociativní paměť, paměť typu fronta, paměť typu zásobník, vícebránové paměti, paměti s kombinovaným řízením.

Podle možnosti zápisu/čtení mohou být paměti rozděleny na:

- RWM (Read Write Memory) - paměti pro zápis i čtení.
- ROM (Read Only Memory) - paměti pouze pro čtení.
- Kombinované paměti
  - NVRAM (Non Volatile RAM) - kombinace RWM a E^2PROM.
  - WOM (Write Only Memory) - paměť, do které lze pouze zapisovat.
  - WORM (Write Once-Read many times Memory) - optické disky CD ROM.

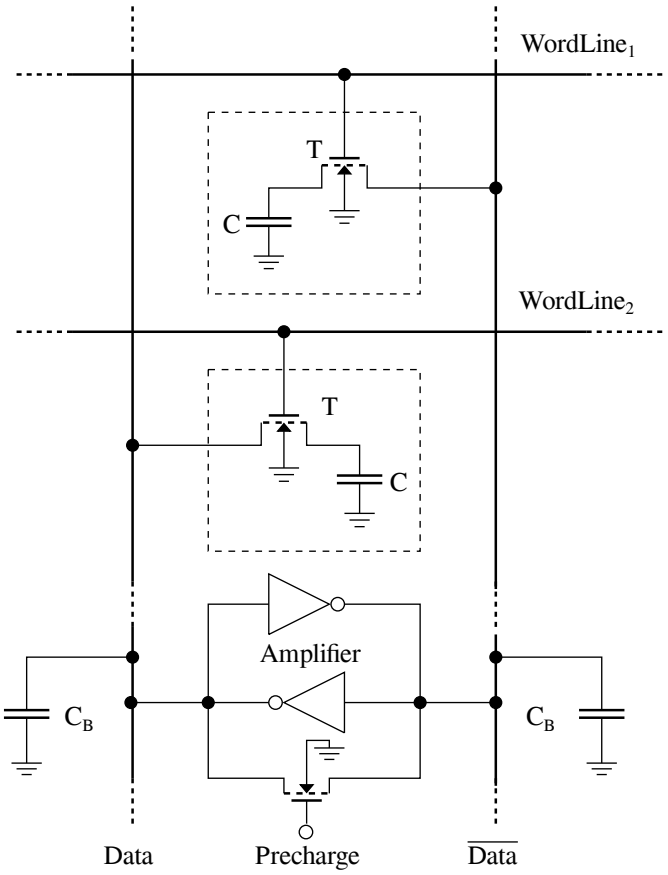
Podle principu elementární buňky mohou být paměti rozděleny na:

- SRAM statické paměti.
- DRAM dynamické paměti.
- PROM, EPROM, EEPROM, FLASH - programovatelné paměti.

Podle uchování informace po odpojení napájení:

- Non Volatile se označují ty paměti, které informaci uchovávání i po odpojení napájení, jako jsou různé typy pamětí xxROM,
- Volatile - paměti ztrácející uloženou informaci po odpojení napájení, jako jsou paměti DRAM a SRAM.

Otázkou může v této chvíli být, kam zařadit hlavní paměť počítače, které se hovorově říká „ramka“. Z uvedeného dělení pamětí je zřejmé, že toto označení je neúplné a tedy nepřesné. Hlavní paměť počítače je určena pro čtení a zápis, musí umožňovat náhodný přístup a informace je uložena v kondenzátoru a po odpojení napájení se uložené informace ztratí. Úplné označení by tedy dle předchozího dělení být Volatile RWM DRAM.

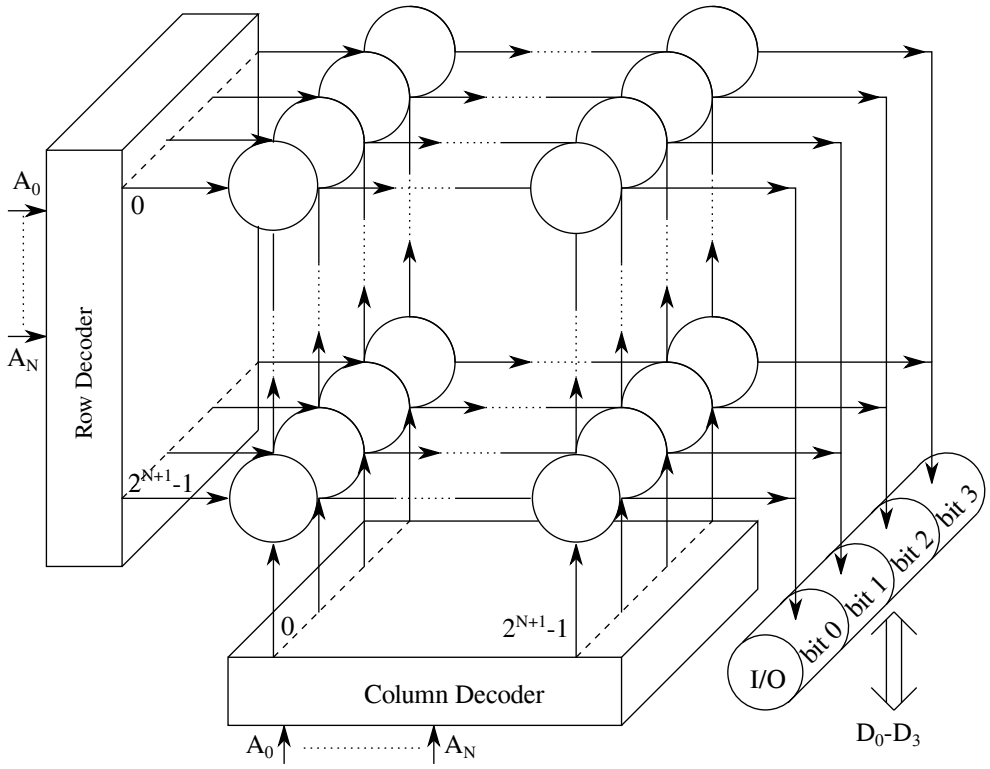


Obrázek 1: Paměťové buňky DRAM

2 Dynamické paměti

V dynamických pamětech je informace uložena ve formě náboje v kondenzátoru. Kondenzátor může být buď nabitý (logická 1) nebo vybitý (logická 0).

Na obrázku 1 jsou znázorněny dvě jednotranzistorové paměťové buňky dynamické paměti. Jelikož kondenzátory jsou extrémně miniaturní, jsou to jen parazitní kapacity na tranzistotu, jejich kapacita je velmi malá - jednotky až desítky fF (femtoFarad). Kondenzátor si tak není schopen zachovat uloženou informaci po neomezenou dobu, i velmi malý proud tekoucí do nebo z tohoto kondenzátoru, vyvolá velké změny jeho napětí v krátkém čase. Je proto nutné často obnovovat napětí kondenzátoru - tato procedura je označovaná jako občerstvení neboli refresh. Pro tento proces jsou dnes na čipu implementovány speciální obvody, těm stačí pouze pravidelně přečíst libovolnou buňku z každé řady a následovně bude občerstvena celá tato řada.



Obrázek 2: Organizace paměti DRAM

Dynamické paměti zapomenou všechna svá uložená data během přibližně 10 ms. Více informací lze najít v části 2.3.

Obrázek 2 ukazuje zjednodušenou organizaci paměti DRAM, kde každé kolečko reprezentuje jednu paměťovou buňku. Paměťové buňky jsou umístěny ve čtvercové matici v jedné nebo více vrstvách. Výběr buňky tak musí být proveden ve dvou krocích pomocí řádkového (Row) a sloupcového (Column) dekodéru. Vodiče vedoucí z dekodérů řádku a sloupce slouží k výběru paměťových buněk a vodiče vedoucí z paměťových buněk slouží k přenosu uložených dat do I/O bufferu.

Rozdělení adresování na dva dekodéry přináší výhodu při adresování. Například pro adresování 2^{20} (1 Mbit) lineárně uspořádaných paměťových buněk, je potřeba 20 adresových vodičů $A_0 \div A_{19}$. Pokud se však paměťové buňky uspořádají do matice, stačí řádkovému i sloupcového dekodéru pro adresování stejné kapacity 1 Mbit ($2^{20} = 2^{10} \times 2^{10}$) jen polovina adresních vodičů $A_0 \div A_9$. Během přístupu do paměti je nejprve dodána adresa řádku a na stejných vodičích následně adresa sloupce vybrané buňky. Díky tomu je počet vodičů zredukován a obsahová kapacita zvětšena. Pro výběr dekodéru, kterému je právě adresa určena, však musí být navíc přidány dva řídící

signály RAS a CAS. Podrobněji je činnost popsána v dalším textu.

2.1 Konvenční DRAM

2.1.1 Organizace čipu DRAM

S neustálým vývojem paměťových čipů s většími kapacitami byly zavedeny různé formy organizace. 1 Mbit čip s jedním datovým vývodem má organizaci 1 Mword na 1 bit. To znamená, že paměťový čip se skládá z 1 M slov o šířce 1 bit na každý vývod. Další široce užívaná forma organizace pro 1 Mbit čip je 256 Kword $\times$ 4 bitová organizace. Tyto čipy mají tedy 256 Kwords se šířkou čtyři bity, takže mají čtyři vývody. Kapacita paměti je také 1 Mbit. První číslo vždy znamená počet slov (words) a druhé počet bitů na slovo. Hlavní vlastností je počet datových vývodů, to jest šířka, ve které slovo může být vloženo na vstupu nebo obdrženo na výstupu během přístupu do paměti.

Dnešní paměťové čipy mají mnohem větší kapacity, uvedený příklad je uveden jako jednoduchá ukázka.

2.1.2 Princip činnosti DRAM

Podle adresy, kterou poskytlo CPU přijme adresový buffer adresu paměti jako výstup externího paměťového kontroléru. Z tohoto důvodu je adresa rozdělena na dvě části, adresu řádku a adresu sloupce. Tyto dvě adresy jsou čteny do adresového bufferu jedna za druhou. Tento proces se nazývá mutliplexing. Důvod tohoto dělení je zřejmý: na adresování jedné buňky v 4 Mb čipu s 2048 řádky a 2048 sloupci by bylo třeba celkem 22 adresových bitů (11 pro řádek a 11 pro sloupec). Na obrázku 2 tomuto příkladu odpovídá $N = 10$. Pokud by měly být přesunuty všechny adresové bity najednou, bylo by třeba 22 adresových vývodů. Potom by musel být pouzdro čipu velmi velký.

Proto je lepší přesunout adresu paměti ve dvou částech. Obvykle adresový buffer nejprve čte adresu řádku a potom adresu sloupce. Tento adresní multiplexing je kontrolován RAS (Row Address Strobe) a CAS (Column Address Strobe) řídícími signály. Pokud paměťový kontrolér pošle adresu řádku, tak zároveň aktivuje RAS signál. RAS informuje čip DRAM, že dodaná adresa je adresa řádku. Nyní kontrolér DRAM aktivuje adresový buffer k získání adresy a přesune ji do dekodéru řádku, který ji dekóduje. Pokud později paměťový kontrolér poskytne adresu sloupce, potom aktivuje CAS signal. Tak kontrolér DRAM pozná, že tentokrát je přesunována adresa sloupce a aktivuje znovu adresový buffer. Adresový buffer přijme poskytnutou adresu a přesune ji do dekodéru sloupce.

Paměťová buňka adresovaná tímto způsobem předá na výstup uložená data, která jsou zesílena čtecími zesilovači a přesunuta do I/O bufferu. Buffer

nakonec poskytne informace jako výstupní data *Dout* přes datové vývody paměťového čipu.

Pokud mají být data zapsána, paměťový kontrolér aktivuje WE (Write Enable) signál a přesune zapisovaná data *Din* do I/O bufferu. Pomocí čtecích zesilovačů je informace zesílena, přesunuta do adresované paměťové buňky a v ní uložena.

Paměťový kontrolér počítače tedy řeší 3 různé úkoly: rozdělení adresy získané z CPU na adresu řádku a sloupce, které jsou přesunuty do paměti jedna po druhé; správně aktivuje RAS, CAS, WE a READ signály; přesunuje uložená data a přijímá data k zapsání do paměti. Neupravené adresové a datové signály z CPU nejsou vhodné pro paměť, proto je paměťový kontrolér nezbytnou součástí počítačového paměťového subsystému.

2.1.3 Čtení a zápis dat

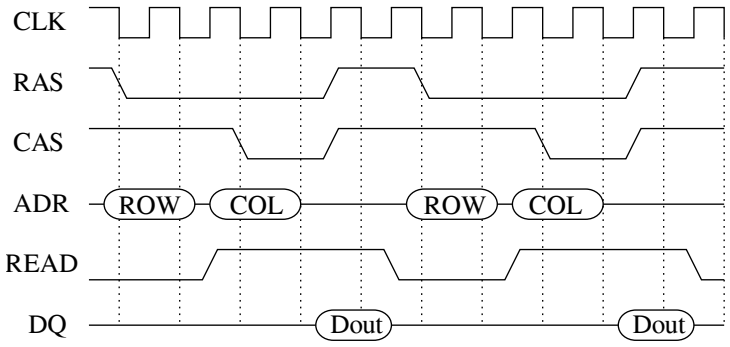
Paměťová buňka na obrázku 1 má kondenzátor, který udržuje data ve formě elektrického náboje, a přístupový tranzistor, který slouží jako přepínač pro výběr kondenzátoru. Báze (gate) tranzistoru je připojena na adresový vodič. Pole paměťových buněk obsahuje jeden adresový vodič, číslovaný *0* až $2^{N+1} - 1$, na každou zformovanou řadu.

Pole paměťových buněk, kromě adresových vodičů, také obsahuje takzvané páry datových vodičů DATA a $\overline{\text{DATA}}$. Existuje právě jeden pár datových vodičů na každý sloupec v poli paměťových buněk. Datové vodiče jsou střídavě připojeny k emitorům přístupových tranzistorů. Jádrem celé buňky je kondenzátor, který představuje paměťový element jedné buňky. Jedna z jeho elektrod je připojena na kolektor odpovídajícího přístupového tranzistoru a druhá je uzemněna.

Paměťový kontrolér adresující paměťovou buňku na čipu, nejprve poskytne signál adresy řádku a aktivuje odpovídající adresový vodič. Všechny přístupové tranzistory připojené k tomuto adresovému vodiči se zapnou. Náboje všech kondenzátorů z adresovaného řádku protečou do odpovídajících datových vodičů a do kondenzátorů C_B a dále přes čtecí zesilovače do I/O bufferu. Dekodér sloupce dekóduje přivedený signál adresy sloupce a aktivuje právě jeden datový vodič. I/O buffer zesílí znovu signál dat a předá jej jako výstupní data *Dout*. Jelikož přístupové tranzistory zůstávají zapnuty, přečtená data se zapíší zpět do paměťových buněk jednoho řádku pomocí kondenzátorů C_B, ve kterých jsou původní přečtená data. Přečtení jedné paměťové buňky tudíž současně vede k občerstvení celého řádku.

Obrázek 3 ukazuje chování nejdůležitějších paměťových signálů během vykonávání procesu čtení dat.

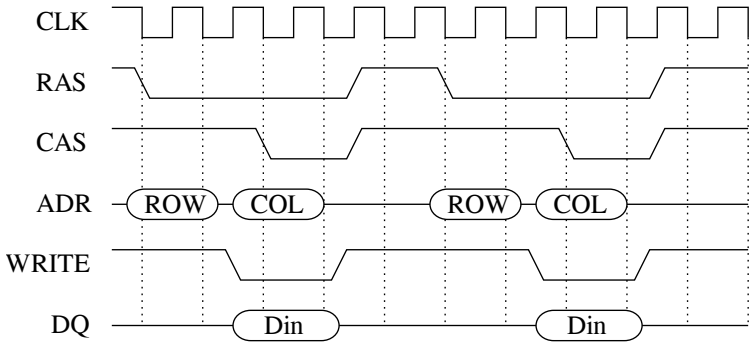
Zápis dat je proveden téměř stejným způsobem jako čtení dat. Nejprve paměťový kontrolér poskytne signál adresy řádku, poté aktivuje RAS adresní signál. Ve stejnou chvíli také aktivuje WE řídící signál. Zapisovaná data (*Din*) jsou poslána do vstupního datového bufferu, zesílena a přesunuta



Obrázek 3: Časové schéma čtení z DRAM

do I/O bufferu. Dekodér řádku dekóduje signál adresy řádku a aktivuje odpovídající adresový vodič. Přístupové tranzistory se sepnou a přesunou uložené náboje z kondenzátorů do párů datových vodičů DATA, $\overline{\text{DATA}}$. Potom paměťový kontrolér aktivuje CAS signál a poskytne adresu sloupce do dekodéru sloupce. Data jsou přesunuta z I/O vodičů do odpovídajícího čtecího zesilovače. Potenciály datových vodičů jsou přesunuty zpět do kondenzátorů jako odpovídající náboje.

Obrázek 4 ukazuje chování nejdůležitějších paměťových signálů během vykonávání procesu zápisu dat.

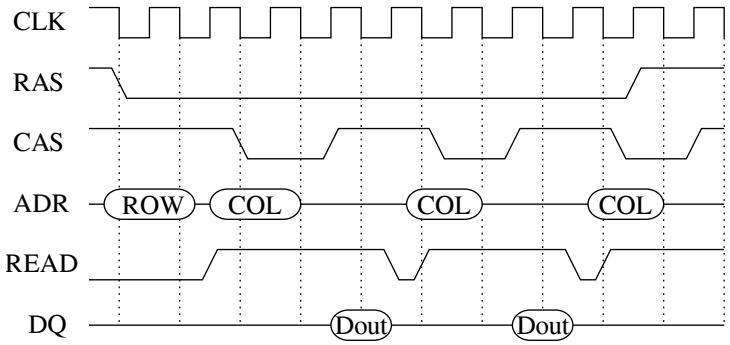


Obrázek 4: Časové schéma zápisu do DRAM

2.1.4 Další operační módy

Minulá sekce popisovala normální mód DRAM. Paměťové čipy mohou také vykonávat jeden nebo více jiných sloupcových módů pro snížení přístupové doby. Ten nejznámější je stránkový mód (Page Mode). Obrázky 5 až 8 ukazují chování nejdůležitějších paměťových signálů během vykonávání procesu čtení dat během jednoho z těchto vysokorychlostních módů.

Stránkový mód (Page Mode) Sekce 2.1.3, *Čtení a zápis dat*, se zmiňuje, že v průběhu přístupu do paměťové buňky je nejprve zadána adresa řádku s aktivním RAS signálem a potom adresa sloupce s aktivním CAS signálem. Pokud se další přístup do paměti vztahuje na stejný řádek a pouze jiný sloupec (to znamená, že adresa řádku zůstala stejná a jen adresa sloupce je změněna), není nutné vkládat a dekódovat znovu adresu řádku (ve stránkovém módu je změněna pouze adresa sloupce, ale adresa řádku zůstala zachována). Takže jedna stránka koresponduje právě s jedním řádkem v poli paměťových buněk. Průběh signálů ve stránkovém módu je na obrázku 5.



Obrázek 5: Časové schéma čtení z FP DRAM

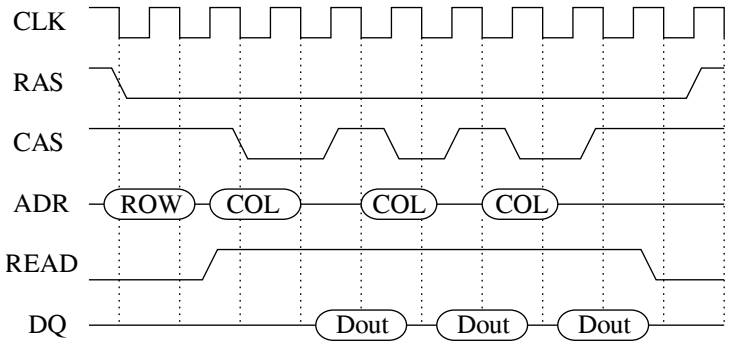
Během stránkového módu při přístupu do paměťové buňky ve stejném řádku paměťový kontrolér nedeaktivuje RAS signál. Pouze CAS signál je deaktivován na krátkou dobu a poté znovu aktivován. Všechny přístupové tranzistory připojené na adresovém vodiči adresovaného řádku proto zůstanou sepnuty a všechna přečtená data jsou na konci datových vodičů. Nová adresa sloupce je dekódována v dekodéru sloupce. Ve stránkovém módu je přístupová doba o 50% (a doba cyklu až o 70%) kratší, než v normálním módu. Toto samozřejmě platí pro druhý i každý další přístup.

EDO mód (Extended Data Out) V EDO módu časová vzdálenost mezi dvěma následnými CAS aktivacemi je kratší, než u stránkového módu (viz. obrázek 6). Adresy sloupců jsou přesunovány rychleji a přístupová doba je výrazně kratší (až o 30% ve srovnání se stránkovým módem) a tudíž přenosová rychlost je proto větší. V EDO módu musí CAS signál být deaktivován před poskytnutím nové adresy sloupce.

2.2 Vylepšené typy DRAM pamětí

2.2.1 SDRAM

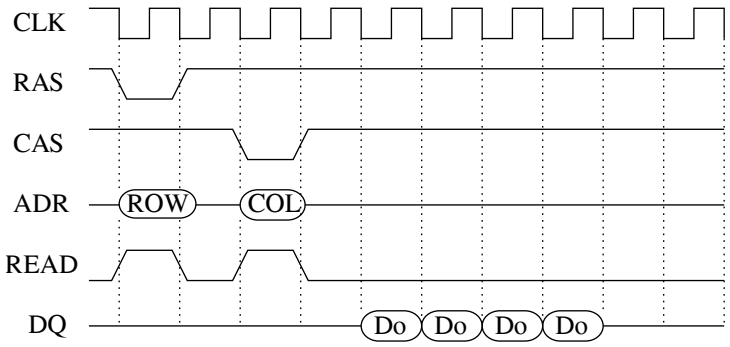
SDRAM znamená synchronní DRAM, což by se nemělo zaměňovat s SRAM (statické RAM), které jsou popsány v sekci 3. SDRAM mají typickou přístu-



Obrázek 6: Časové schéma čtení z EDO DRAM

povou dobu pouze 8 až 15 ns a mohou fungovat synchronně se systémovou taktovací frekvencí. Ta může být 66 MHz, 133 MHz nebo i více. EDO DRAM mají přístupovou dobu 50 až 60 ns. V praxi není rozdíl mezi SDRAM a EDO DRAM podstatný. Jeden z důvodů je paměť L2 cache, která zvedá výkon slabších EDO čipů o nějaký stupeň. Při porovnání s EDO DRAM je patrný větší výkon pamětí SDRAM pokud je systémová taktovací frekvence větší než 100 MHz, což je případ drtivé většiny dnešních systémů.

SDRAM pracují v burst módu a se synchronní taktovací frekvencí, ne s různým RAS a CAS časováním jak tomu je u jiných RAM čipů. SDRAM také používají odpovídající signály RAS, CAS, WE a CE, ale používají je k přesunutí příkazů jako zápis, čtení a burst stop. Signály RAS a CAS jsou zkombinovány pro vytvoření příkazové sběrnice, jak je patrné časovém schématu (obrázek 7).

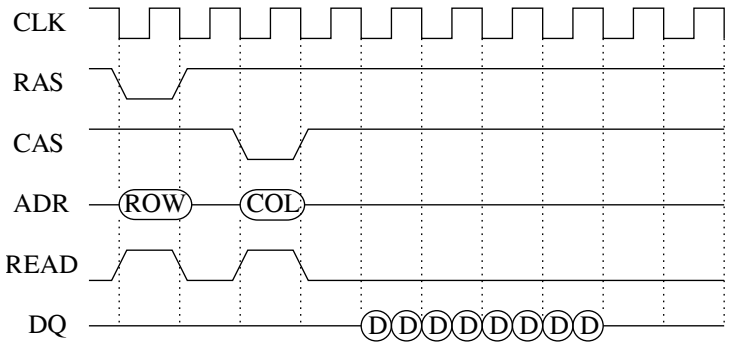


Obrázek 7: Časové schéma čtení z SDRAM

SDRAM používá principu podobnému prokládání paměťových polí tak, že zatímco s jedním pracuje (je z něj čteno), druhé se připravuje na následující přístup.

2.2.2 DDR SDRAM

Přenosová rychlost dat může být zdvojnásobena, pokud jsou data přenášena nejen na náběžné hraně hodinového pulsu, ale také na sestupné hraně hodinového pulsu. Přesně tento princip používají paměti double data rate DRAM (DDR-RAM). Je to zpětně kompatibilní typ paměti, který vedl k modulům PC-266 a jehož vývoj nadále pokračuje. Časové schéma čtení je zobrazeno na obrázku 8.



Obrázek 8: Časové schéma čtení z DDR SDRAM

2.3 Občerstvování DRAM (Refresh)

Postupem času se kondenzátory vybíjejí přes přístupový tranzistor a jeho dielektrické vrstvy. Díky tomu dochází k vybití uložených nábojů a tím i ke ztrátě dat. Kondenzátor musí být pravidelně občerstvován. V průběhu čtení z paměti jsou paměťové buňky adresovaného řádku automaticky občerstveny, protože proces čtení je destruktivní. Normální DRAM musí být občerstvena zhruba každých 10 ms. V současnosti se používají tři občerstvovací metody: RAS-only refresh, CAS-before-RAS refresh a Hidden refresh.

RAS-only refresh Nejjednodušší a nejvíce používaná metoda pro občerstvování paměťové buňky je vykonání předstíraného cyklu čtení. Během tohoto cyklu je aktivován RAS signál a DRAM se poskytne adresa řádku (adresa občerstvení), zatímco CAS signál zůstává neaktivní. K občerstvení celé paměti je potřeba, aby externí obvod nebo sám procesor poskytl DRAM adresy řádků přesně jak jdou po sobě.

CAS-before-RAS refresh Pro tento typ občerstvení má DRAM čip svou vlastní občerstvovací logiku s adresním počítadlem. Během CAS-before-RAS refresh je CAS udržován na nízké úrovni po jistou dobu, než RAS klesne na nízkou úroveň (proto CAS-before-RAS). Vnitřní občerstvovací logika je tím aktivována a vykoná automatické vnitřní občerstvení.

Hidden refresh Zde je cyklus občerstvování "skryt" za normálním přístupem pro čtení. Během skrytého občerstvování je CAS signál udržován na nízké úrovni a pouze RAS signál je přepnut. Protože čas potřebný pro cyklus občerstvování je většinou kratší než cyklus čtení, tento způsob občerstvování šetří čas.

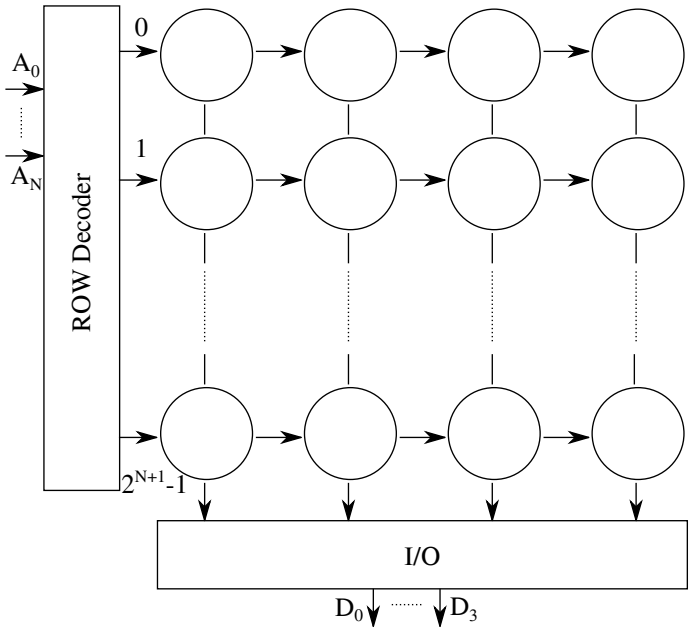
2.4 Paměťové moduly

Paměťové čipy jsou nazývány DIP, což znamená Dual Inline Package. Jsou to integrované obvody s vývody na obou stranách. Pro snadnější instalaci pamětí jsou tyto DIP čipy umístěny na modulu. Dnes se používají kompaktní moduly jako SIMM a DIMM než jednotlivé čipy. Pro dosažení patřičné paměťové kapacity je na modul instalován odpovídající počet čipů. Moduly musí být vloženy do soketů, které jsou pro ně umístěny na základní desce.

SIMM moduly Single Inline Memory Module. Mohou mít DIP čipy na jedné nebo obou stranách a to se 30 nebo 72 piny. Normálně jsou dostupné ve verzi se 72 piny, které podporují 32 bitový přesun dat mezi procesorem a pamětí.

DIMM moduly Double Inline Memory Module. 168 pinové DIMM mají vždy šířku 64 bitů. Jako DIMM se používají hlavně paměti SDRAM a DDR-RAM.

RIMM moduly Rambus Inline Memory Module. Mají 184 kontaktů a jsou dostupné v kapacitách 64, 128 a 256 Mb. Podle jejich popisu pracují s maximální taktovací frekvencí 400 MHz, což často vede k hodnotě "800 MHz", ačkoliv stejně jako v případě DDR-RAM je přenos dat uskutečněn na obou hranách hodinového pulsu. Intel používá odlišnou implementaci, Direct RAMBus, s datovou sběrnicí o šířce 16 bitů. Tento text nepopisuje paměti RAMBus, které byly instalovány na tento typ modulů. Tyto moduly se krátce používaly v počítačích s procesory Intel P4 kolem roku 2000.



Obrázek 9: Organizace paměti SRAM

3 Statické paměti

Ve statických pamětech je informace uložena stavem klopného obvodu. Realizace klopného obvodu je možná pomocí 4 nebo 6 tranzistorů. Obě varianty jsou na obrázcích 10 a 11. Podrobněji budou popsány dále.

Obrázek 9 zobrazuje zjednodušenou organizaci paměti SRAM, kde každé kolečko reprezentuje jednu paměťovou buňku, a ty jsou jsou organizovány jen do 2D mřížky, kde jeden řádek tvoří jedno datové slovo (word). Vodorovné čáry jsou adresové vodiče, které vybírá signál dekodéru řádku (ROW) na základě vstupní adresy $A_0 \div A_N$. Na těchto vodičích jsou připojeny přístupové tranzistory paměťových buněk, což bude vysvětleno dále. Svislé čáry jsou datové vodiče, které přenáší uloženou informaci vybraného řádku do výstupního bufferu I/0.

Paměťovému kontroléru SRAM čipů jsou adresy řádků poskytovány jako jedna informace. A jelikož zde chybí adresní multiplexing, je zapotřebí více pinů a SRAM čipy jsou větší než DRAM čipy. Vnitřní adresování paměťových buněk je tím ale jednodušší a SRAM čipy jsou tak rychlejší než DRAM. A to nejen při adresování, ale i při čtení a zápisu informace.

Díky statickému charakteru paměti není potřeba občerstvování (refresh). Mezi dvěma stavy se klopné obvody přepínají pomocí vnějšího signálu a stav klopných obvodů paměti je udržován tak dlouho, dokud je SRAM čip

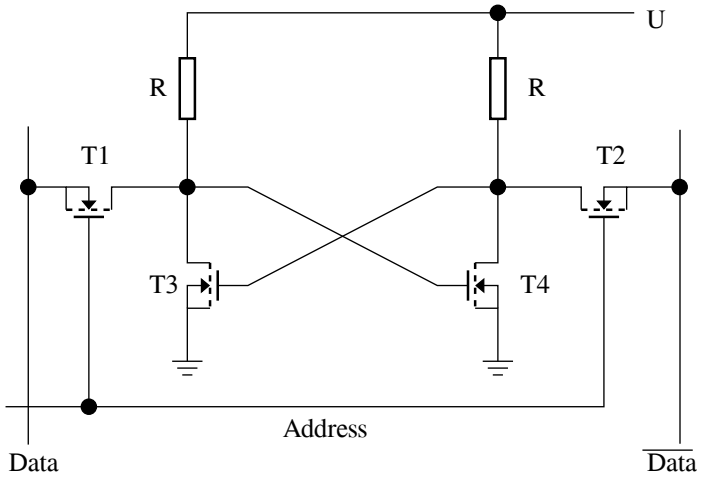
napájen.

SRAM čip je dražší a může obecně pojmou méně dat než DRAM čip kvůli menší hustotě paměťových buněk na jednotku místa. Integrační hustota DRAM čipů je asi čtyřikrát až šestkrát větší než u SRAM čipů za použití stejné technologie. Z tohoto důvodu se SRAM čipy hlavně používají pro malé a rychlé cache paměti (popsané v sekci 5.3), zatímco DRAM čipy pro velké a relativně pomalé hlavní paměti (RAM).

3.1 SRAM - Static Random Access Memory

3.1.1 Paměťová buňka SRAM

Jak již bylo zmíněno, paměťová buňka SRAM se může skládat ze 4 nebo 6 tranzistorů. Jednotlivá řešení jsou na obrázku 10 a 11.

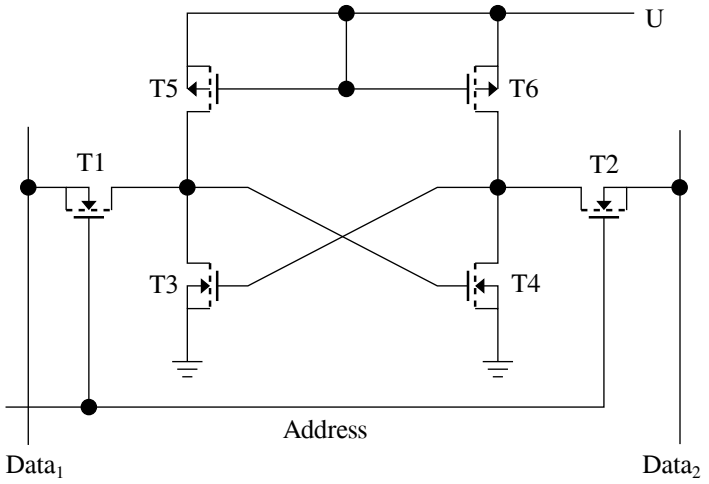


Obrázek 10: Paměťová buňka SRAM se 4 tranzistory

V obou případech se buňka SRAM skládá ze dvou NMOS přístupových tranzistorů T1 a T2 a klopného obvodu se dvěma NMOS paměťovými tranzistory T3 a T4. Dalších dva elementy jsou dvou odpory, nebo dva PMOS tranzistory T5 a T6, které vedle s tranzistory T3 a T4 tvoří CMOS dvojici.

V SRAM jsou paměťové buňky uspořádány v mřížce s řádky a sloupci, viz obrázek 9, které se vybírají dekodéry řádku. Báze (gate) přístupových tranzistorů T1 a T2 jsou připojeny na adresové vodiče a emitory (source) na páry datových vodičů DATA, $\overline{\text{DATA}}$. .

Paměti SRAM lze vyrábět i pomocí technologie TTL. Buňka takové paměti funguje na podobném principu jako paměťová buňka používající technologii MOS, ale skládá se pouze ze dvou PNP tranzistorů a dvou odporů.



Obrázek 11: Paměťová buňka SRAM se 6 tranzistory

3.1.2 Čtení a zápis dat

Protože SRAM nevyužívá adresní multiplexing, je adresa do čipu přenesena jako jedna informace. Při čtení i zápisu dat z/do paměťové buňky SRAM aktivuje dekodér řádku odpovídající adresový vodič.

Při čtení se dva přístupové tranzistory T1 a T2 zapnou a propojí klopný obvod paměti s datovými vodiči DATA a $\overline{\text{DATA}}$. Po stabilizaci dat vybere dekodér odpovídající sloupec (to jest odpovídající datové vodiče DATA, $\overline{\text{DATA}}$) a předá na výstupu data do I/O bufferu a tím do vnějších obvodů.

Zápis dat probíhá opačným způsobem. Přes vstupní datový buffer se zapisovaná data zavedou na odpovídající čtecí zesilovač. Ve stejnou dobu aktivuje dekodér řádku adresní vodič a zapne přístupový tranzistor T1. Stejně jako v procesu čtení dat se klopný obvod snaží předat uložená data na datové vodiče DATA, $\overline{\text{DATA}}$. Nicméně čtecí zesilovač je silnější než paměťový tranzistor T3 a poskytne datovým vodičům DATA, $\overline{\text{DATA}}$ signál, který odpovídá zapisovaným datům. Proto se klopný obvod přepne podle nově zapisovaných dat nebo si udržuje již uloženou hodnotu v závislosti na tom, zda se zapisovaná data shodují s uloženými daty nebo ne.

3.2 Async, Sync a PB SRAM

Asynchronní SRAM Tato paměť existuje od dob procesoru 386 a stále se nachází v pamětech cache L2 mnoha PC. Nazývá se asynchronní, protože není synchronizovaná se systémovými hodinami, a proto CPU musí na data vyžádaná z paměti cache L2 čekat (ne však tak dlouho, jako u DRAM).

Synchronní Burst SRAM Synchronní dávková paměť SRAM. Podobně jako SDRAM je synchronní SRAM synchronizovaná se systémovými hodinami.

Pipeline Burst SRAM Zřetězená dávková SRAM. Použitím dávkové technologie lze požadavky na SRAM zřetězit, neboli shromáždit je tak, že požadavky v dávce se vykonávají téměř okamžitě. PB SRAM používá zřetězení, a ačkoliv mírně zaostává za systémovými synchronizačními frekvencemi, představuje zlepšení proti synchronní SRAM, protože je navržena pro spolupráci se sběrnicemi.

4 Paměti s trvalým obsahem

Schémata k těmto technologiím jsou uvedena ve studijním materiálu k tématu Technologie výroby číslicových obvodů.

Nevýhodou paměťových modulů popsaných v předchozích sekcích je jejich neschopnost udržet si uložená data i po odpojení napájení. DRAM a SRAM čipy nejsou vhodné pro startovací proces PC, protože ve chvíli, kdy nejsou zásobovány elektřinou, je jejich obsah zapomenut. Místo toho se používají ROM čipy. Data jsou zde uložena jednou energeticky nezávislým způsobem, takže jsou udržována po dlouhou dobu, i když jsou ROM čipy odpojeny od zdroje energie.

Ve funkci paměťových prvků pamětí ROM se v historii počítačů vystřídaly všechny základní pasivní i aktivní elektronické prvky. Byly tak použity odpory, indukčnosti, feritová jádra, kondenzátory, diody, tranzistory unipolární i bipolární.

Hlavním úkolem těchto pamětí je pamatovat si data v době, kdy je odpojeno napájení. Z tohoto důvodu se používají například pro uchování BIOSu.

4.1 ROM

ROM znamená Read Only Memory, tedy paměť pouze pro čtení. Buňka paměti je představována elektrickým odporem nebo pojistkou. Výrobce některé z nich elektronicky přepálí. Neporušené prvky pak vedou proud, je v nich minimální napětí - tzn. nesou logickou 0. Přepálené prvky proud nevedou, je v nich maximální napětí - nesou informaci o logické 1. Informaci do nich zapisuje výrobce. Doba pamatování není ohraničena.

4.2 PROM

U programovatelných ROM čipů vypaluje informace do paměti uživatel pomocí programátoru. Data jsou zapsána použitím elektrického pulsu. Jednou z metod je přepálení pojistky mezi adresovým a datovým vodičem. Tyto pojistky jsou vyrobeny z niklu a chromu nebo křemíku. Stejně jako do ROM tak ani do PROM není po naprogramování možný zápis.

Jiná možná realizace paměti PROM je za pomoci bipolárních multi-emitorových tranzistorů. Tento typ PROM se skládá z jednoho multi-emitorového tranzistoru na každý adresový vodič. Pokud se má z paměti číst, potom se na jisté adresové vodiče přivede logická 1, multi-emitorový tranzistor se otevře a ve směru kolektor-emitor prochází proud. Pokud pojistka nebyla přepálena, potom proud otevře tranzistor, který je připojený jako invertor a na výstup se přečte logická 0. Pokud pojistka byla přepálena, potom se tranzistor neotevře a na výstup se přečte logická 1.

4.3 EPROM

Erasable Programmable Read Only Memory - patří mezi paměti, do nichž je možné opakovatelně zapisovat. Paměťová informace se uchovává pomocí elektrického náboje. Ten je kvalitně izolovaný, a tak udrží svoji hodnotu i po odpojení elektrického napětí. K naprogramování je potřeba pulz trvající 50 ms o napětí + 5 V (u některých typů až + 12 V). Také EPROM se programuje pomocí speciálního programátoru. Je ji možné vymazat pomocí ultrafialového záření a po vymazání do ní opět zapsat nová data. EPROM lze poznat podle okénka na pouzdře, kterým vstupuje do paměti mazací ultrafialové záření. Doba pamatování je omezena na 10 až 20 let. Používá se kapacita izolovaného hradla tranzistoru MOS.

4.4 EEPROM

Okénko v pouzdře a UV lampa pro mazání dat jsou komplikovaná a taky drahá vybavení pro mazání EPROM čipů. Bylo by lepší a o hodně jednodušší, kdyby mohl být čip vymazán stejným způsobem jako byl naprogramován, tedy elektrickým pulzem. To je příklad EEPROM - elektricky vymazatelné PROM. Programování paměťové buňky se provádí stejným způsobem jako u EPROM, to jest relativně dlouhým (50 ms) pulzem o napětí + 5 V (nebo + 12 V). Pro vymazání se pouze obrátí polarita pulzu. Počet zápisů a mazání do EEPROM je ohraničený, doba pamatování uložené informace je omezena na 10 až 20 let.

4.5 Flash paměti

V posledních letech se využívá typ EEPROM paměti, schopný si uchovat informace i po odpojení zdroje energie, jako náhrada za diskety a pevné disky. Jde o tak zvanou flash paměť. Její největší výhodou je možnost ji rychle naprogramovat přímo v počítači. Doba uchování uložené informace je nejméně deset let a většinou kolem sto let.

Struktura jejich paměťových buněk je v základě stejná jako ta u pamětí EEPROM. Pro mazání a programování je potřeba pulz trvající jen 10 μs a méně a napájecí napětí. Vymazání celé paměti je velmi rychlé. Je možno vykonat bez problémů 10 000 a více programovacích a mazacích cyklů, dle typu paměti.

Adresový buffer přijímá signály adres a přesunuje je do dekodéru řádku a sloupce. Flash paměti, stejně jako SRAM čipy, nevykonávají adresní multiplexing. Dekodéry řádku a sloupce vyberou jeden adresní vodič a jeden nebo více datových vodičů tak jako v běžném čipu. Přečtená data jsou předána na výstup přes vstupně/výstupní datový buffer anebo v případě zápisu jsou zapsána do adresované paměťové buňky tímto bufferem přes I/O bránu.

Proces zápisu je trochu složitější. Je možné zapsat “0”, ale není možné zapsat “1” normálním způsobem. K tomu je potřeba zkopírovat celý sektor

do RAM paměti a vymazat ho z flash paměti. V RAM paměti se “1” zapíše do daného řádku a ten se potom celý zapíše zpět do flash paměti na jeho původní místo.

5 Další typy pamětí

5.1 Video paměti

5.1.1 Video RAM (VRAM)

Běžné DRAM paměti použité pro video karty většinou nemají dostatečně široké přenosové pásmo pro udržení velkého rozlišení a barevné hloubky při akceptovatelné obnovovací frekvenci. Kvůli tomu byl vyvinut nový typ paměti nazvaný video RAM neboli VRAM. Tato paměť je dvouportová : má dva přístupové porty pro paměťové buňky, jeden se používá pro neustálé obnovování obrazu, druhý pro změnu dat, která se mají zobrazovat. Tyto dva porty tedy znamenají zdvojnásobení kapacity přenosového pásma a v důsledku toho vyšší grafický výkon.

5.1.2 WRAM

Podobně jako VRAM je také WRAM dvouportovým typem paměti RAM a používá se výlučně pro zvýšení grafického výkonu. WRAM je svým fungováním podobná VRAM, nabízí však širší celkové přenosové pásmo (zhruba o 25%) a několik grafických funkcí, jež mohou využít tvůrci aplikací. K nim patří systém dvojitého vyrovnávání dat (tzv. double-buffering), několikanásobně rychlejší než vyrovnávací systém VRAM, což vede k podstatně vyšším obnovovacím frekvencím zobrazení.

5.1.3 SGRAM

Synchronní Grafická RAM v podstatě funguje jako SDRAM. Nejdůležitější rozdíly mezi nimi jsou, že SDRAM je optimalizována pro nejvyšší možnou paměťovou kapacitu a SGRAM je optimalizována pro nejvyšší možný přenos dat.

5.2 FIFO paměti

Paměti FIFO se realizují buďto přímo v mikroprocesoru, nebo jsou k dispozici jako stavební členy s různou organizací. V zásadě je můžeme rozdělit na typy:

- bez přesouvání obsahu - zápis a čtení z fronty se řídí dvojicí registrů - čte se podle obsahu registru začátku fronty, zapisuje se podle obsahu registru konce fronty. Řídící obvody též musejí vytvářet dva důležité stavové signály - fronta prázdná a fronta plná (k předejití podtečení a přetečení).
- s přesouváním obsahu - obvodové realizace fronty s přesouváním obsahu při čtení a při zápisu jsou stejně složité; mají jeden přídavný

registr. Fronta s probubláváním posouvá asynchronně každou položku po zápisu až do posledního volného místa, přečtením jedné položky ze začátku fronty se jedno místo uvolní, načež je položky stejným mechanismem obsadí. Princip vyžaduje, aby u každého paměťového místa existoval indikátor obsazenosti (klopný obvod).

5.3 Cache paměti

S pamětí cache se v celé počítačové architektuře setkáme často. Je to jakýsi mezisklad dat mezi různě rychlými komponentami počítače. Jeho účelem je vzájemné přizpůsobení rychlostí - rychlejší komponenta čte data z cache a nemusí čekat na komponentu pomalejší (u které si cache data již načetla).

5.3.1 Paměti L1 cache

Tato paměť je integrována přímo na procesoru. Slouží k zásobování procesoru daty ze sběrnice. Cache přečte více dat ze sběrnice, které potom čekají v tomto meziskladě. Jakmile je procesor potřebuje, přečte si je z cache. Protože cache pracuje rychleji než sběrnice, nemusí procesor čekat, jak by tomu bylo v případě odebírání dat přímo ze sběrnice.

5.3.2 Paměti L2 cache

Paměť cache L2 je umístěna mezi mikroprocesorem a operační pamětí, takže všechna data, která putují mezi těmito dvěma díly, v cache uvíznou a pokud je bude mikroprocesor znovu potřebovat, přečte si je z rychlejší cache. Navíc je cache ovládána speciálním řadičem, který se snaží předpovědět, která data bude asi mikroprocesor v nejbližší době požadovat. Pracuje také jako víceportová paměť, kde mohou být data zároveň zapisována i čtena.

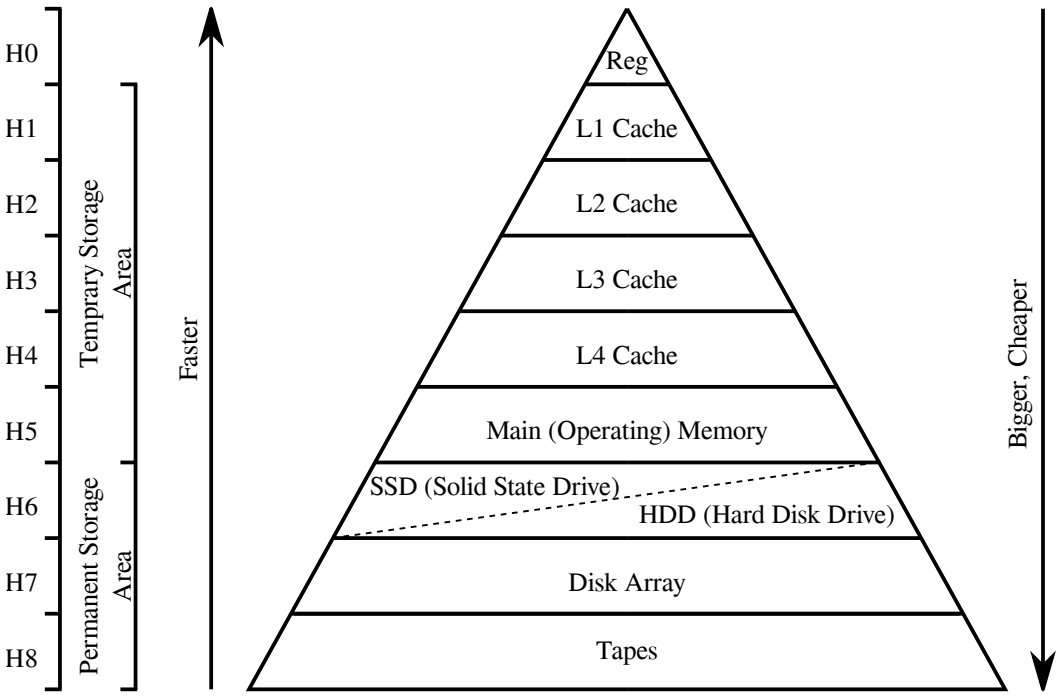
Cache paměti používají 3 režimy:

- Write-Through - (zápis skrz cache; přímý zápis) je nejstarší a nejpomalejší způsob, typický pro mikroprocesory 486. Data ukládaná do cache zapisuje současně i do operační paměti. Při čtení pak porovná řadič cache požadované adresy operační paměti s adresami již uloženými. Pokud jsou potřebná data v cache nalezena, jsou z ní přečtena.

- Write-Back - (opožděný zápis) je novější a rychlejší metodou používanou u Pentií a některých rychlejších řad 486. Data jsou zapisována pouze do cache a teprve při odstranění z cache jsou zapsána do operační paměti. Než se data do operační paměti dostanou, mohou v cache několikrát změnit svoji hodnotu. V tomto režimu se tedy šetří čas, potřebný na opakované zápisy do pomalejší operační paměti.

- Pipeline Burst - je nejnovější a nejrychlejší systém práce, nyní běžně používaný. Pracuje tak, že provede více operací zřetězeně - pokud čte

z určité adresy informaci, přečte zároveň informace i z následujících adres (což by pravděpodobně dělal za chvíli). Přístupová doba k datům se pohybuje mezi 9 až 15 ns.



Obrázek 12: Paměťová hierarchie

6 Hierarchie pamětí v počítači

Velké množství pamětí, zmíněné v předchozím textu, si výrobci počítačů nevyvíjeli a nevyrábí jen proto, aby byla na trhu pestrá nabídka technologií. Důvody jsou čistě praktické, ekonomické a samozřejmě i technické. Každá technologie má své vlastnosti a výrobní cenu za jednotku kapacity a to ji předurčuje, kde a v jaké velikosti je ekonomicky rentabilní tu kterou paměť použít.

Kdyby byla k dispozici technologie, která umožňuje rychlý přístup k datům, uchovává si svůj obsah i po odpojení napájení a je levná, tak by měly počítače pouze jednu paměť a jejich konstrukce by se zjednodušila. Zatím ale taková technologie není k dispozici a stávající konstrukce počítačů je kompromisem mezi cenou a rychlostí a kapacitou.

Již v jednom z předchozích témat - Monolitické počítače - bylo zmiňováno, že paměti jsou v počítači uspořádány do více vrstev a to platí pro všechny počítače. Tomuto uspořádání pamětí se obvykle říká „Paměťová hierarchie“.

Příklad paměťové hierarchie je uveden na obrázku 12. Tato ukázka je pouze ilustrativní, protože uspořádání pamětí se bude vždy lišit mezi růz-

nými typy počítačů. Jinak to bude u minipočítače, např. Raspberry Pi, jinou konfiguraci bude mít běžný pracovní notebook a něco jiného nalezneme ve výkonnějším serveru, kterému by právě konfigurace na obrázku 12 mohla odpovídat.

Bez ohledu na konkrétní počítač však můžeme hierarchické uspořádání pamětí v počítačích shrnout do několika společných bodů:

- Paměti jsou v počítači uspořádány podle rychlosti, ceny a kapacity. V obrázku 12 je to znázorněno šipkami na levé a pravé straně pyramidy. Nejrychlejší a nejdražší jsou paměti v CPU, tedy registry na vrcholu pyramidy.
- Každá vyšší úroveň paměti (v pyramidě) tvoří vyrovnávací paměť (cache) pro úroveň pod sebou.

  Např. hlavní paměť tvoří vyrovnávací paměť pro pevný disk a data hlavní paměti jsou částečně uložena v Cache L4, atd.
- Paměti se dělí na oblast pro ukládání dočasných dat a dat trvalých.
- Dříve se používalo dělení pamětí na vnitřní a vnější paměti a dodnes je možno v některých textech tuto informaci nalézt. Za dělící rovinu mezi těmito dvěma typy pamětí se považoval přechod mezi H6 a H5.

  Dnes je vhodnější dělení na paměti čistě polovodičové a paměti s mechanickými součástkami. Toto dělení je naznačeno čárkovanou čarou v úrovni H6. Všechny paměti nad touto čarou jsou čistě polovodičové a paměti pod čarou obsahují nějaké mechanické díly. Toto dělení je velmi důležité z hlediska rychlosti paměti. Všechny paměti s mechanickými díly jsou i o několik řádů pomalejší, než paměti polovodičové.

Dále je možno charakterizovat jednotlivé použité technologie, jejich parametry a kapacity v paměťové hierarchii:

H0 - registry procesoru. Rychlost odpovídá rychlosti CPU a pohybuje se v jednotkách ns, nebo desetinách ns. Kapacita se může pohybovat ve stovkách bytů na jádro, takže celkově i v jednotkách kB.

H1 - Cache L1, SRAM, rychlost odpovídá vnitřním sběrnicím CPU, obvykle v nižších jednotkách ns. Kapacita desítky až stovky kB na jádro.

H2 - Cache L2, SRAM, rychlost obvykle od jednotek do 10 ns. Kapacita stovky kB na jádro.

H3 - Cache L3, SRAM, rychlost v nižších desítkách ns, kapacita v jednotkách až desítkách MB.

H4 - Cache L4, SRAM, rychlost v desítkách ns, kapacita desítky až stovky MB.

H5 - Hlavní paměť, někdy nazývaná operační paměť, DRAM, rychlost v desítkách ns, kapacita jednotky až stovky GB,

H6 - SSD disky, Flash, rychlost v desetinách či jednotkách ms, kapacita stovky GB až jednotky TB. Pevné disky, rychlost v jednotkách až desítkách ms, kapacita v jednotkách TB.

H7 - Disková pole, parametry odvozeny od pevných disků, kapacita od desítek TB až po jednotky PB.

H8 - Páskové jednotky, rychlosti dle použitého řešeni od desítek minut až po hodiny.

7 Chyby pamětí, jejich detekce a oprava

7.1 Chyby pamětí

Paměť je elektronické uložiště a všechna tato zařízení mají potenciál vracet nesprávné informace odlišné od těch, které byly původně uloženy. DRAM paměti jsou díky svojí charakteristice náchylné občas vracet paměťové chyby. DRAM ukládá jedničky a nuly jako náboje v malých kondenzátorech, které musí být neustále občerstvovány, aby se předešlo ztrátě dat. DRAM je méně spolehlivá než statické úložiště používané SRAM.

Existují dva typy chyb, které mohou nastat v systémové paměti. První se jmenují opakující se nebo tvrdé chyby. V této situaci je část hardware rozbitá a bude neustále vracet chybný výsledek. Je relativně jednoduché takovéto chyby rozpoznat a opravit, protože jsou konzistentní a opakující se.

Druhý typ chyb se nazývá přechodné nebo měkké chyby. Nastávají ve chvíli, když se jednou zpět přečte bit se špatnou hodnotou, ale následně už funguje správně. Tyto problémy jsou pochopitelně hůře rozpoznatelné a také, bohužel, běžnější. Příležitostně se měkká chyba většinou zopakuje, ale to se může stát kdykoliv mezi pár minutami až po několik let.

Jediná skutečná ochrana proti chybám paměti je použít nějaký typ jejich detekce a opravných metod. Některé metody mohou pouze detekovat chyby v jednom bitu z bytu (osm bitů); jiné mohou automaticky detekovat chyby ve více než jednom bitu. Další mohou jak detekovat tak i opravovat chyby.

7.2 Parita

Parita se dělí na fyzickou a logickou. Fyzická parita znamená, že paritní bity jsou poskytnuty paměťovému kontroléru během procesu zápisu dat a jsou uloženy v paměťovém modulu. Během čtení paměťový modul předá na výstup informaci o uložené paritě. Když je použita logická parita, paměťový kontrolér generuje paritní bity a poskytne je paměťovému modulu během zápisu, ale modul si tyto paritní bity neuloží. Během čtení jednoduchý obvod v modulu generuje paritní informace z bitů uložených dat. Proto nemůže nikdy nastat paritní chyba; paritní informace poskytované modulem nejsou důležité. Takové moduly ušetří více než 10% (4 až 36 bitů) paměťové kapacity a jsou proto levnější.

Paměťové moduly bývají tradičně dostupné ve dvou typech: s paritou a bez parity. Běžná paměť je bez parity - obsahuje přesně jeden bit paměti pro každý bit uložených dat. Osm bitů je třeba k uložení každého bytu dat. Paměti s paritou přidávají navíc jeden bit pro kažých osm bitů dat, který se používá pouze pro detekování chyb. Devět bitů je třeba k uložení každého bytu. Paměti s paritou mohou použít kontrolu parity, jednoduchou formu detekce chyb. Paměti bez parity neposkytují žádné možnosti detekce chyb.

7.2.1 Kontrola parity

Kontrola parity je jednoduchý způsob pro detekci jednobitových chyb v systémové paměti. Každý byte dat uložených v paměti obsahuje osm bitů realných dat, buď nulu nebo jedničku. Je možné spočítat počet nul nebo jedniček v bytu. Např. byte 10110011 má tři nuly a pět jedniček. Některé byty budou mít sudý počet a některé lichý počet jedniček.

Když se do paměti zapíše byte, logický obvod nazvaný generátor/kontrolér parity vyhodnotí tento byte a určí, zda má sudý nebo lichý počet jedniček. Pokud má sudý počet jedniček, potom je devátý paritní bit nastaven na jedna, jinak je nastaven na nula. Výsledkem je, že bez ohledu na to, kolik bylo jedniček v původním bytu, v celých devíti bitech jich bude vždy lichý počet. Tomuto se říká lichá parita.

Během čtení dat z paměti slouží paritní obvod pro kontrolu. Čte všech devět bitů a znovu zjišťuje, jestli je tam sudý nebo lichý počet jedniček. Pokud je počet jedniček sudý, musela být v jednom z bitů chyba, protože při ukládání byl nastaven paritní bit, takže by počet jedniček měl být vždy lichý. Tímto způsobem funguje detekce chyb u paměti s paritou. Systém ví, že v jednom bitu je chyba, ačkoliv neví ve kterém. V případě detekování paritní chyby paritní obvod vygeneruje "nemaskovatelné přerušení" neboli "NMI", které se většinou používá pro okamžité zastavení procesoru.

Kontrola parity má své meze. Řekněme, že byte dat "00100100" je uložen jako "001001001 1" včetně paritního bitu. Nyní řekněme, že se zpětně přečte jako "01100000 1". Zde došlo k prohození dvou bitů, nicméně počet jedniček zůstal lichý. Jak lze vidět, parita neumožňuje ochranu proti chybám ve dvou bitech.

7.3 ECC paměti

Kontrola parity umožňuje detekovat jednobitové chyby paměti, ale není schopen detekovat vícebitové chyby a neposkytuje žádný způsob opravy paměťových chyb. Z tohoto důvodu byl vynalezen zdokonalený protokol pro detekci a opravu chyb s názvem ECC (Error Correction Code). Tento protokol nejenže detekuje jednobitové a více bitové chyby, ale je schopen opravit jednobitové chyby během čtení.

ECC používá speciální algoritmus pro kódování informací do bloku bitů, které obsahuje dostatek detailů pro obnovu chyby jednoho bitu. Na rozdíl od parity, která používá jeden bit jako ochranu pro osm bitů, ECC používá skupiny bitů: sedm bitů pro ochranu 32 bitů nebo osm bitů pro ochranu 64 bitů.

ECC může detekovat chyby dvou, tří a dokonce i čtyř bitů, ale nemůže je už opravovat. ECC paměť se s těmito vícebitovými chybami vypořádá stejně jako paměť s paritou pomocí nemaskovatelného přerušení (NMI). Více bitové chyby jsou ale v pamětech velice vzácné.

Na rozdíl od kontroly parity ECC způsobí nepatrné zpomalení systémových operací. Důvodem je, že algoritmus ECC je více komplikovanější a chvíli trvá, než ECC opraví nalezené chyby. Z toho důvodu je potřeba vložit jeden čekací stav (wait state) během čtení z paměti. Reálně to znamená snížení výkonu přibližně o 2-3%.

8 Kontrolní otázky

1. Dle jakých kritérií či vlastností se dělí paměti počítačů?
2. Jak je v dynamických pamětech ukládána informace a jak je udržována?
3. Jaká je vnitřní organizace dynamických pamětí?
4. Popište stručně historii vývoje dynamických pamětí.
5. Jak je ve statických pamětech ukládána informace a jak je udržována?
6. Jak je organizována vnitřně statická paměť?
7. Jaké typy pamětí si udržují svůj obsah i po odpojení napájení?
8. Paměti s trvalým obsahem umožňují svůj obsah přepsat. Jak se přepis u jednotlivých typů provádí?
9. Jaké speciální typy pamětí se používají?
10. Jak se u pamětí detekují a opravují chyby?