

ORGANIZACE PAMĚŤOVÉHO SYSTÉMU POČÍTAČE

HIERARCHIE PAMĚŤOVÉHO SYSTÉMU

je několikaúrovňové uspořádání pamětí různých kapacit a rychlostí s cílem dosáhnout výhodného poměru výkonnosti a ceny.

Cena paměti je: -přímě úměrná kapacitě
-přibližně nepřímě úměrná době přístupu

Paměťová hierarchie

typ paměti	typická realizace	doba přístupu	kapacita
registry	klopné obvody	jednotky ns	desítky-stovky B
vyrovnávací paměť	statická RAM	10-20ns	stovky kB-jednotky MB
hlavní paměť	dynamická RAM	50-70ns	desítky-stovky MB
vnější paměť	pevný disk	5-15ms	jednotky-desítky GB
záložní paměť	optický disk mag. páska	stovky ms- stovky s	stovky GB- jednotky TB

Hierarchické uspořádání pamětí řeší konflikt mezi požadavky na rychlost paměti a na její kapacitu.

VIRTUÁLNÍ PAMĚŤ

je systém několika pamětí s různými parametry (kapacita, rychlost), který je řízen tak, aby vytvářel adresové prostory potřebné velikosti pro programy a pro data.

Virtuální paměť umožňuje:

- realizaci jednoho nebo několika logických (virtuálních) adresových prostorů, z nichž každý může být větší než je skutečná kapacita hlavní paměti označovaná jako fyzický adresový prostor. Logické adresové prostory jsou ve skutečnosti realizovány ve vnější paměti. Části programu a položky dat jsou automaticky přesouvány do hlavní paměti požaduje-li k nim procesor přístup.
- úsporné využití hlavní paměti tím, že v ní jsou přítomny jen ty části programu a dat, se kterými procesor právě pracuje.
- přemístění částí programu v hlavní paměti bez nutnosti je znovu překládat.
- vzájemnou ochranu jednotlivých programů v paměti a ochranu dat před neoprávněným přístupem a modifikací.

V režimu virtuální paměti pracuje program vždy s logickými (virtuálními) adresami.

Hlavní paměť se adresuje fyzickými adresami.

Příklad logických adres na fyzické zajišťuje mechanismus virtuální paměti.

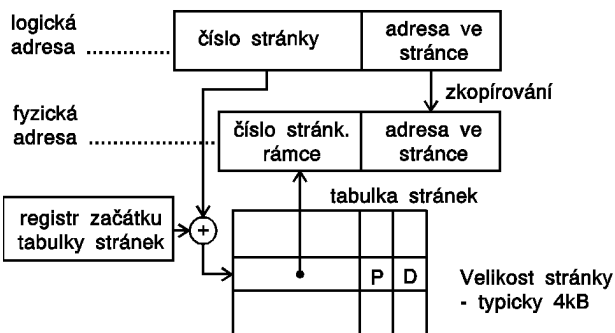
STRÁNKOVÁNÍ

je zaměřeno na realizaci velkého logického (virtuálního) adresového prostoru při malé kapacitě hlavní paměti.

Logický adresový prostor je rozdělen na úseky pevné délky nazývané stránky nebo také logické stránky.

Fyzický adresový prostor je rozdělen na stejně velké úseky nazývané stránkové rámce nebo také fyzické stránky.

Logický adresový prostor je realizován ve vnější paměti. Data (úseky programu) se přesouvají do hlavní paměti po jednotlivých stránkách tehdy, jsou-li v průběhu výpočtu požadována a pokud se již příslušná stránka v hlavní paměti nenachází. Stránkovací mechanismus pracuje s datovou strukturou - tabulkou stránek - uloženou též v hlavní paměti. Každé stránce odpovídá 1 položka v tabulce stránek. Ta obsahuje informaci o tom, zda se kopie příslušné stránky nachází v hlavní paměti a pokud ano tak ve kterém stránkovém rámci.



Problém: Tabulka stránek musí obsahovat 1 položku pro každou stránku v logickém adresovém prostoru i když tato stránka není použita. Je-li velký poměr mezi velikostí logického a fyzického adresového prostoru, může tabulka stránek zabírat podstatnou část hlavní paměti.

Příklad:

virt. paměť	4GB 32b
hlavní paměť	4MB 22b
velikost stránky	4kB 12b

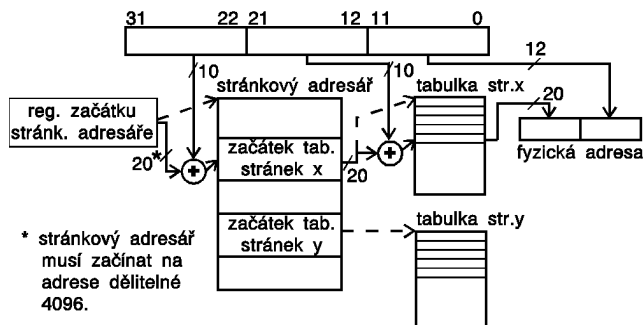
LA 20 12

FA 10 12

Tabulka stránek má 2^{20} položek
1 položka ... 10 bitů horní část fyzické adresy } 2B pro
1 bit platnost } 1 položku
⇒ tabulka stránek zabírá 2MB - nepřijatelné.

Řešení:

Dvouúrovňová organizace tabulky stránek (použito v 80386 a vyšších)



Tabulka stránek musí obsahovat:

- horní část fyzické adresy (10 bitů)
- příznak přítomnosti stránky v hlavní paměti
- příznak indikující, že do stránky bylo po dobu přítomnosti v hlavní paměti zapisováno (Dirty bit)
- další příznaky - viz literatura

Funkce stránkovacího mechanismu

- je-li stránka přítomna v hlavní paměti (viz příznak přítomnosti stránky), přeloží se logická adresa na fyzickou
- není-li stránka přítomna, vyvolá se přerušení. Jeho obslužný program vyvolá načtení požadované stránky z vnější paměti.
- pokud při načítání nové stránky není volný žádný stránkový rámec v hlavní paměti, je třeba některý uvolnit. Toho se dosáhne přesunem vhodné stránky (např. nejdéle nepoužité) zpět do vnější paměti. Nebylo-li však po dobu přítomnosti této stránky v hlavní paměti do ní zapisováno (viz příznak), není třeba ji ukládat a stačí ji pouze přepsat nově načtenou stránkou.

SEGMENTACE

umožňuje dosáhnout úspory kapacity hlavní paměti tím, že se program do HP zavádí po částech.

Segmenty jsou funkčně samostatné části programu proměnné délky, které lze zavádět do hlavní paměti v případě potřeby.

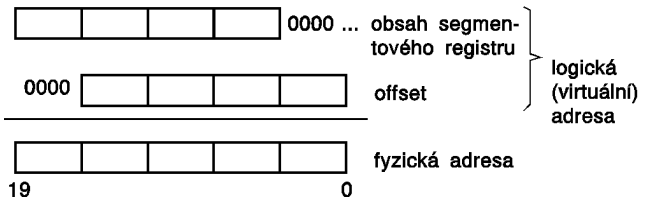
Adresy v segmentu jsou relativní vůči začátku (bázi) segmentu => přemístitelnost segmentů v hlavní paměti.

Logická adresa se skládá z obsahu segmentového registru a offsetu (posunutí).

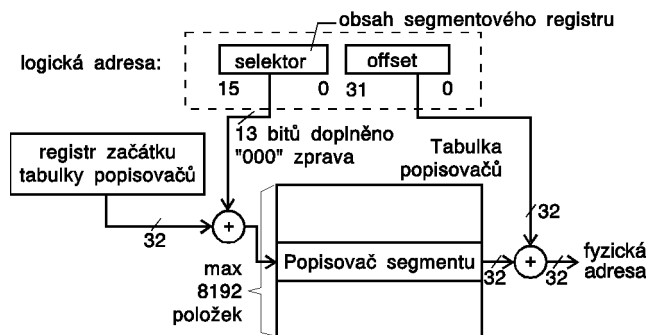
Báze segmentu může být uložena

a) v segmentových registrech

- např. 8086 a další procesory řady 80x86 v reálném módu (opakování)



b) v tabulkách adresovaných segmentovými registry
- v chráněném režimu procesoru 80386 a vyšších



Popisovač segmentu obsahuje (u 80386 a vyšších, délka 8B)

- 32-bitovou bázi segmentu
- 20-bitový limit segmentu (max. délka)
- tzv. bit G (Granularity), udávající zda limit segmentu je vyjádřen ve slabkách nebo v blocích à 4096 slabik (4kB)
- přístupová práva

Výhody řešení ad b)

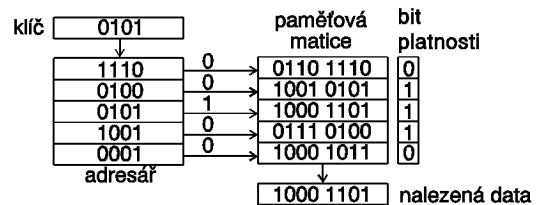
- fyzický adresový prostor (velikost hlavní paměti) až 4GB
- logický adresový prostor max 8192 segmentů à max 4GB
- umožňuje přiřadit segmentům přístupová práva - vzájemná ochrana jednotlivých běžících programů a dat
- hodnota selektoru není v přímém vztahu k fyzické adrese. Slouží jako index do předem připravené tabulky popisovačů jednotlivých segmentů.

RYCHLÁ VYROVNÁVACÍ PAMĚŤ (Cache)

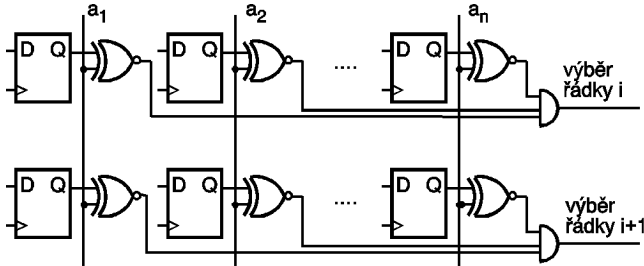
- rychlá paměť s malou kapacitou (desítky-stovky kB) zařazená mezi procesor a hlavní paměť
- obsahuje kopie nejčastěji používaných položek z hlavní paměti
- realizuje se jako statická paměť RAM - doba přístupu typicky 10-15ns
- pro její realizaci je nutná asociativní paměť - paměť adresovaná obsahem

PLNĚ ASOCIATIVNÍ PAMĚŤ

- paměť typu RAM se speciální strukturou liší se od adresovatelné paměti RAM
- adresuje se částí datové položky, která se má vyhledat - tzv. klíčem
- na rozdíl od adresovatelné paměti RAM má místo adresového dekodéru tzv. adresář



Struktura adresáře asociativní paměti



Každá položka adresáře obsahuje logické obvody umožňující najednou prohledat všechny položky adresáře.

POUŽITÍ PLNĚ ASOCIATIVNÍ PAMĚTI JAKO RYCHLÉ VYROVNÁVACÍ PAMĚTI

- data zapsaná v paměťové matici asociativní paměti budou kopie "často" používaných položek dat z hlavní paměti
- klíčem bude adresa, která každou položku jednoznačně identifikuje

Jak bude probíhat čtení ?

- pokusíme se současně o čtení z cache a z hlavní paměti. Pokud se položka v cache nalezne, použije se a cyklus hlavní paměti se nedokončí. V opačném případě se data přečtou z hlavní paměti (a zpravidla se současně uloží do cache)

... a zápis ?

- pokud položka v cache není, zapíše se (zpravidla) jen do hlavní paměti
- pokud tam je, lze postupovat dvěma způsoby:
 - zapsat novou hodnotu současně do cache a do hlavní paměti - tzv. průběžný zápis (write through)
 - zapsat novou hodnotu jen do cache - tzv. odložený zápis (write back)

Metoda odloženého zápisu je složitější v tom, že při uvolňování položky cache musíme někdy její původní obsah zapsat zpět do hlavní paměti (tehdy byla-li položka v cache modifikována).

POUŽITÍ PAMĚTI S OMEZENÝM STUPNĚM ASOCIATIVITY

Nevýhody plně asociativní paměti:

- pro konstrukci adresáře nelze použít standardní obvody RAM
- při stejné kapacitě cca trojnásobná plocha čipu (a tudíž i cena) oproti standardní RAM

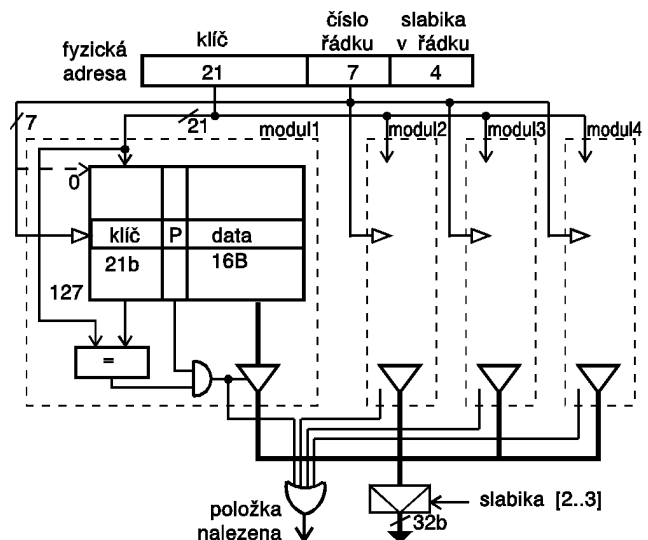
Řešení: Paměť s omezeným stupněm asociativity

Každé položce z HP je podle její adresy přiděleno jedno pevné místo (případně několik* míst), kde se v cache může nacházet. Adresář cache lze pak realizovat běžnou pamětí RAM, přítomnost položky se zjistí porovnáním požadovaného klíče s klíčem uloženým v adresáři (případně s několika* klíči).

Poznámka: Pro zvýšení efektivity se data z HP do cache nepřenesají po jednotlivých slabikách nebo slovech, ale po delších blocích zvaných též řádky o délce např. 16 slabik.

*počet míst na nichž se konkrétní položka může v cache nacházet se nazývá stupeň asociativity

Řešení Interní cache v procesoru 80486



Funkce

Při pokusu o čtení z cache se v každém modulu porovná klíč uložený v příslušném řádku (udávající adresu bloku dat, jehož kopie se v tomto řádku nachází) s klíčem odvozeným z adresy hledané položky dat. Shoda klíčů znamená, že hledaná položka byla v příslušném modulu cache nalezena.