8. UNITATEA DE MEMORIE

Memoria este partea sistemelor de calcul care se utilizează pentru păstrarea şi regăsirea ulterioară a datelor şi instrucțiunilor. Operațiile principale în care este implicată memoria sunt următoarele:

- Preluarea datelor de intrare în memorie;
- Păstrarea datelor până la prelucrarea lor de către UCP;
- Păstrarea datelor de ieşire până când ele vor putea fi preluate de echipamentele de ieşire;
- Transmiterea datelor din memorie la ieşire.

Sistemele de memorie influențează în mod critic performanțele calculatoarelor. Deoarece în memorie sunt păstrate atât datele, cât şi instrucțiunile, sistemul de memorie trebuie să satisfacă cererile simultane pentru prelucrarea datelor, execuția instrucțiunilor şi transferul între memorie şi exterior.

Într-un calculator de tip *von Neumann*, restricția principală impusă de sistemul de memorie este următoarea:

> Un singur modul de memorie cu structură convențională nu poate face acces la mai mult de un cuvânt în timpul fiecărui ciclu de memorie.

Există o mare varietate de tipuri, tehnologii, organizări, performanțe şi costuri ale memoriilor utilizate în sistemele de calcul. Nici una din acestea nu este optimă pentru satisfacerea tuturor cerințelor. Ca o consecință, sistemele de calcul sunt echipate cu o *ierarhie* de subsisteme de memorie, unele interne sistemului (accesibile direct de UCP), iar altele externe (accesibile prin UCP printr-un modul de I/E).

8.1. Caracteristicile sistemelor de memorie

Cele mai importante caracteristici sunt următoarele:

1) *Amplasarea:*
   - În cadrul UCP;
   - Memorii interne;
   - Memorii externe.

2) *Capacitatea:*

- Dimensiunea cuvântului;
- Numărul de cuvinte.

3) *Unitatea de transfer:*
- Cuvântul;
- Blocul.

4) *Metoda de acces:*

- Acces secvențial;
- Acces direct;
- Acces aleator;
- Acces asociativ.

5) *Performanțele:*

- Timpul de acces;
- Durata ciclului;
- Rata de transfer.

6) *Tipul memoriei:*

- Memorii semiconductoare;
- Memorii magnetice.

7) *Caracteristicile fizice:*

- Volatile / nevolatile;
- Cu / fără posibilitatea ştergerii.

8) *Organizarea.*

Amplasarea. Sistemele de calcul dispun de memorii *interne* şi *externe*. Memoria internă este considerată de cele mai multe ori ca *memorie principală*. Există însă şi alte forme de memorie internă. UCP necesită o memorie locală proprie, sub forma registrelor. Unitatea de comandă şi control din cadrul UCP poate necesita de asemenea o memorie proprie, în cazul unităților de comandă microprogramate. Memoria externă constă din dispozitivele periferice, ca discuri sau benzi magnetice, care sunt accesibile de către UCP prin controlere (module) de I/E.

Capacitatea. Se exprimă prin dimensiunea cuvântului de memorie (8, 16, 32, 64 sau 128 de biți) şi numărul de cuvinte (KB, MB, GB).

Unitatea de transfer. Pentru memoria internă, unitatea de transfer este egală cu numărul liniilor de date la şi de la modulul de memorie, deci cu numărul de biți transferați simultan. Unitatea de memorie nu trebuie să fie egală neapărat cu un cuvânt. Pentru memoria externă, datele sunt transferate de multe ori în unități mai mari decât un cuvânt, numite *blocuri*.

Metoda de acces. Există următoarele tipuri de acces la unitățile de date:

- *Acces secvențial*. Memoria este organizată în unități de date, numite *înregistrări*. Accesul trebuie realizat într-o secvență liniară. Se utilizează informații de adresare memorate pentru separarea înregistrărilor și pentru a permite regăsirea informațiilor. Timpul de acces la o anumită înregistrare arbitrară este variabil. Unitățile de bandă sunt echipamente cu acces secvențial.
- *Acces direct*. Blocurile sau înregistrările individuale au o adresă unică pe baza amplasării fizice a acestora. Timpul de acces este de asemenea variabil. Unitățile de disc sunt echipamente cu acces direct.
- *Acces aleator*. Fiecare locație adresabilă a memoriei are un mecanism de adresare încorporat. Timpul de acces a unei locații date este independent de secvențele accesurilor anterioare și este constant. Deci, fiecare locație poate fi selectată aleator, și poate fi adresată și accesată direct. Memoria principală este cu acces aleator.
- *Acces asociativ*. Memoria asociativă este un tip de memorie cu acces aleator, care permite compararea unor biți dintr-un cuvânt cu o anumită valoare specificată, și efectuarea acestei comparări în mod simultan pentru toate cuvintele. Deci, un cuvânt este regăsit pe baza unei părți a conținutului acestuia și nu pe baza adresei (memorie adresabilă prin conținut). Fiecare locație are propriul mecanism de adresare, iar timpul de regăsire este constant, independent de locație sau de secvențele accesurilor anterioare. Memoriile *cache* pot utiliza un acces asociativ.

Performanțele. Se utilizează trei parametri de performanță:

- *Timpul de acces*. Pentru memoria cu acces aleator, acesta este timpul necesar pentru execuția unei operații de citire sau scriere, deci timpul de la setarea adresei de memorie până când datele sunt disponibile pentru utilizare. Pentru alte tipuri de memorii (cu acces non-aleator), timpul de acces este timpul necesar poziționării corespunzătoare a capului de citire/scriere.
- *Durata ciclului de memorie*. Acest parametru este utilizat mai ales pentru memoria cu acces aleator și constă din timpul de acces plus timpul suplimentar necesar până când poate începe un nou acces. Acest timp suplimentar poate fi necesar pentru stabilizarea liniilor de semnal sau pentru a regenera datele dacă citirea acestora este distructivă.
- *Rata de transfer*. Este rata cu care datele pot fi transferate la sau de la unitatea de memorie. Pentru memoria cu acces aleator, rata de transfer este egală cu 1/(*Durata ciclului*). Pentru memoriile cu acces non-aleator, există următoarea relație:

$$T_N = T_A + \frac{N}{R} \tag{8.1}$$

unde:

T_N = Timpul mediu pentru citirea sau scrierea a N biți;

T_A = Timpul mediu de acces;
N = Numărul de biți;
R = Rata de transfer, în biți/s (bps).

Tipul memoriei. Cele mai utilizate tipuri de memorii sunt memoriile semiconductoare şi memoriile magnetice.

Caracteristicile fizice. În cazul memoriilor *volatile*, informațiile se pierd la întreruperea tensiunii de alimentare. La memoriile *nevolatile*, informațiile rămân nemodificate după înregistrarea lor, până la modificarea deliberată a acestora. Memoriile magnetice sunt nevolatile. Memoriile semiconductoare pot fi volatile sau nevolatile. Memoriile semiconductoare care nu pot fi şterse se numesc memorii de tip ROM (*Read Only Memory*). O asemenea memorie este de asemenea şi nevolatilă.

Organizarea. Prin organizarea memoriilor cu acces aleator se înțelege aranjarea fizică a biților pentru formarea cuvintelor.

8.2. Ierarhia de memorii

Principalele caracteristici de care trebuie să se țină cont la realizarea unui sistem de memorie sunt capacitatea şi performanțele memoriei, în special timpul de acces. Pe lângă acestea, trebuie să se ia în considerare şi costul memoriei. Aceste caracteristici sunt contradictorii. De exemplu, există în general următoarele relații între capacitatea, timpul de acces şi costul pe bit al diferitelor tehnologii utilizate pentru implementarea sistemelor de memorie:

- O capacitate mai mare implică un timp de acces mai mare;
- O capacitate mai mare implică un cost pe bit mai mic;
- Un timp de acces mai mic implică un cost pe bit mai mare.

Pe de o parte, trebuie utilizate tehnologii de memorie care asigură o capacitate ridicată, pentru că o asemenea capacitate este necesară, şi deoarece costul pe bit al acestor tehnologii este mai redus. Pe de altă parte, pentru a satisface cerințele de performanță, trebuie utilizate memorii cu un timp de acces redus, care au un cost ridicat şi o capacitate relativ redusă. Aceste cerințe contradictorii se pot asigura dacă se utilizează în cadrul unui sistem de calcul mai multe componente şi tehnologii de memorie, care formează o *ierarhie de memorii*. O ierarhie tipică este ilustrată în Figura 8.1.

Memoria internă principală a calculatorului este cea la care fac referire cele mai multe instrucțiuni şi date. Pentru operațiile interne ale UCP şi cele aritmetice şi logice, se utilizează *registrele*. Memoria principală este extinsă uneori cu o memorie mai rapidă, de dimensiuni mai mici, numită *memorie cache* sau memorie tampon rapidă.

Memoria principală şi extensiile sale sunt, în general, volatile. Datele sunt păstrate pe termen lung în *memorii externe* de masă, dintre care cele mai utilizate sunt discurile şi benzile magnetice. Acestea se utilizează pentru memorarea fişierelor de programe şi de date. Discurile se utilizează de asemenea pentru a asigura o extensie a memoriei principale, numită *memorie virtuală*.

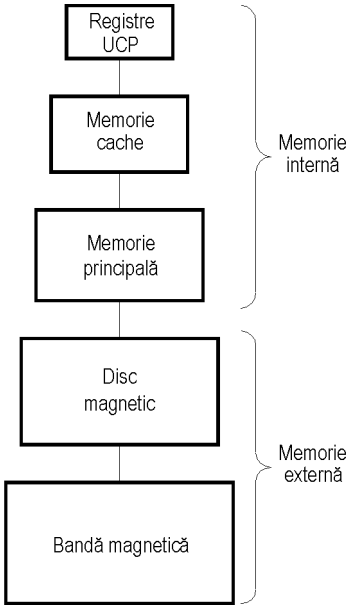

Figura 8.1. Ierarhie tipică a memoriilor.

Pe măsura deplasării din partea superioară a ierarhiei spre cea inferioară, se constată următoarele:

(a) Scade costul pe bit;
(b) Creşte capacitatea;
(c) Creşte timpul de acces;
(d) Scade frecvenţa de acces la memorie de către UCP.

Rezultă deci că memoriile rapide, cu un cost mai ridicat, sunt extinse prin memorii de dimensiuni mai mari, mai lente, dar mai ieftine. Dacă memoria poate fi organizată conform punctelor (a) - (c), şi dacă datele şi instrucţiunile pot fi distribuite în cadrul acestei memorii conform condiţiei (d), această organizare va reduce costurile globale, menţinând în acelaşi timp un anumit nivel al performanţelor.

Baza pentru validitatea condiţiei (d) o reprezintă principiul cunoscut sub numele de *localitate a referinţelor*. Referinţele la memorie efectuate de UCP, atât pentru date, cât şi pentru instrucţiuni, tind să se grupeze în anumite zone. Programele conţin de obicei un număr de bucle repetitive şi subrutine. După intrarea într-o asemenea buclă sau subrutină, vor exista referinţe repetate la un număr redus de instrucţiuni. În mod similar, operaţiile cu tablouri de date implică accesul la un set grupat de cuvinte de date. Într-o perioadă mai lungă de timp, zonele de memorie accesate se schimă, dar într-o perioadă scurtă de timp, UCP lucrează mai ales cu referinţe de memorie grupate.

În mod corespunzător, este posibilă organizarea datelor în cadrul ierarhiei astfel încât procentul de accesuri la nivelul imediat inferior este cu mult mai redus decât cel

la nivelul imediat superior. Considerăm, de exemplu, că memoria este organizată pe două nivele, nivelul 1 fiind nivelul superior. Nivelul 2 conține toate instrucțiunile şi datele programului. O parte a instrucțiunilor şi datelor pot fi plasate temporar în nivelul 1. Periodic, anumite zone din nivelul 1 sunt mutate în memoria de nivel 2, eliberând spațiu pentru alte zone din nivelul 2. În medie însă, majoritatea referințelor se vor efectua la instrucțiunile şi datele aflate în memoria de nivel 1.

8.3. Memorii semiconductoare

8.3.1. Tipuri de memorii semiconductoare

Tabelul 8.1 prezintă principalele tipuri de memorii semiconductoare.

Tabelul 8.1. Tipuri de memorii semiconductoare.

Tipul memoriei	Ştergere	Scriere	Volatilitate
RAM (*Random-Access Memory*)	Electrică	Electrică	Volatilă
ROM (*Read-Only Memory*)	Nu este posibilă	Prin măşti	Nevolatilă
PROM (*Programmable* ROM)		Electrică	Nevolatilă
EPROM (*Erasable* PROM)	Lumină UV	Electrică	Nevolatilă
EEPROM (*Electrically Erasable* PROM)	Electrică	Electrică	Nevolatilă

Memoriile RAM sunt numite şi memorii cu acces aleator. Această denumire este improprie, deoarece toate categoriile de memorii din tabel sunt cu acces aleator, deci cuvintele individuale ale memoriilor pot fi accesate în mod direct. Caracteristica principală a memoriilor RAM este că ele pot fi atât citite, cât şi înscrise în mod simplu, prin semnale electrice. O altă caracteristică a acestora este volatilitatea, conținutul memoriilor fiind pierdut la întreruperea tensiunii de alimentare. Memoriile RAM pot fi utilizate deci doar ca memorii temporare.

Există două tehnologii de memorii RAM: *dinamice* şi *statice*. Memoriile *dinamice* sunt realizate din celule care memorează datele ca sarcini capacitive, utilizând condensatoare. Prezența sarcinii unui condensator este interpretată ca valoarea binară 1, iar absența acesteia ca valoarea binară 0. Deoarece condensatoarele au tendința de a se descărca în timp, memoriile RAM dinamice necesită reîmprospătarea periodică a sarcinilor capacitive pentru a-şi păstra conținutul. Se utilizează circuite speciale de reîmprospătare.

În cazul memoriilor *statice*, valorile binare sunt memorate utilizând bistabile realizate din porți logice. Aceste memorii îşi păstrează conținutul atât timp cât se păstrează tensiunea de alimentare, fără a fi necesară reîmprospătarea lor.

O celulă de memorie dinamică este mai simplă şi deci de dimensiuni mai mici decât o celulă de memorie statică. De aceea o memorie RAM dinamică are o densitate mai mare şi un cost mai redus decât o memorie RAM statică echivalentă. Pe de altă par-

te, o memorie RAM dinamică necesită un circuit de reîmprospătare. Pentru memorii de dimensiuni mari, costul fix al circuitului de reîmprospătare este compensat de costul variabil mai redus al celulelor RAM dinamice. De aceea, memoriile RAM dinamice sunt mai avantajoase atunci când cerinţele de memorie sunt mari. Memoriile RAM statice sunt în general în oarecare măsură mai rapide decât cele dinamice.

Memoriile ROM pot fi doar citite, conţinutul acestora fiind fixat în timpul procesului de fabricaţie. Ca aplicaţii ale acestor memorii se amintesc păstrarea microprogramelor la unităţile de comandă microprogramate, păstrarea programelor de sistem, a unor subrutine utilizate mai frecvent, a unor tabele de funcţii. Avantajul acestor memorii este că ele sunt nevolatile, programele şi datele sunt în permanenţă în memoria principală, şi nu este necesară încărcarea lor de pe un suport extern. Dezavantajul este că operaţia de înscriere a conţinutului în timpul fabricaţiei implică costuri fixe mari, care nu se justifică la serii mici de producţie.

Memoriile PROM sunt similare cu memoriile ROM: ele sunt nevolatile şi pot fi înscrise (programate) o singură dată. În acest caz însă procesul de înscriere este electric, şi poate fi realizat de un furnizor sau utilizator în funcţie de necesităţi, după încheierea procesului de fabricaţie. Pentru înscriere este necesar un echipament special. Avantajul acestor memorii este flexibilitatea utilizării lor, şi costurile mai reduse atunci când este necesar un număr mai redus de memorii cu un anumit conţinut.

Memoriile EPROM sunt citite şi înscrise prin metode electrice, ca şi memoriile PROM. Spre deosebire de acestea, memoriile EPROM pot fi înscrise de mai multe ori, dacă este necesară modificarea conţinutului acestora. Înaintea unei operaţii de scriere, celulele de memorie trebuie şterse prin expunerea circuitului la o lumină ultravioletă. Aceste memorii sunt de asemenea nevolatile.

Memoriile EEPROM reprezintă cea mai avantajoasă formă de memorie nevolatilă. Modificarea conţinutului se poate realiza în orice moment, fără a fi necesară ştergerea vechiului conţinut. Operaţia de scriere necesită un timp considerabil mai lung decât cea de citire. Aceste memorii combină avantajul de a fi nevolatile cu posibilitatea modificării conţinutului lor, fără a fi necesar un echipament specializat, utilizându-se semnalele de adrese, date şi control ale magistralei calculatorului. Preţul acestor memorii este ceva mai ridicat.

8.5. Memoria cache

8.5.1. Principiul memoriei cache

Viteza UCP este superioară vitezei memoriilor, astfel că după iniţierea unui ciclu de acces la memorie, UCP trebuie să rămână inactivă un timp, aşteptând răspunsul acesteia.

Memoriile rapide sunt realizabile din punct de vedere tehnologic, dar costul lor este ridicat. Sunt cunoscute însă tehnici pentru combinarea unei memorii rapide de dimensiuni mici cu o memorie mai lentă de dimensiuni mai mari, pentru a se obţine aproximativ viteza memoriei rapide şi capacitatea mare a memoriei lente, la un preţ moderat.

Memoria rapidă de dimensiune mică se numeşte *memorie cache* (din limba franceză: *cacher* - a ascunde).

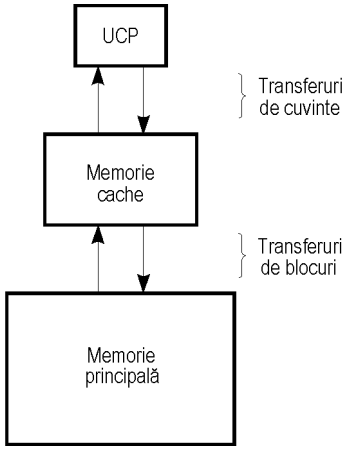

Figura 8.10. Principiul memoriei *cache*.

Principiul memoriei *cache* este ilustrat în Figura 8.10. Există o memorie principală de dimensiuni relativ mari, dar mai lentă, şi o memorie *cache* mai redusă, dar mai rapidă. Memoria *cache* conține o copie a unor părți din memoria principală. Atunci când UCP încearcă citirea unui cuvânt din memorie, se testează dacă respectivul cuvânt se află în memoria *cache*. În caz afirmativ, cuvântul este furnizat unității centrale. În caz contrar, se încarcă în memoria *cache* un bloc al memoriei principale, constând dintr-un număr fix de cuvinte, iar apoi cuvântul este returnat unității centrale.

Se cunoaşte că programele nu fac acces la memorie în mod complet aleator. Dacă se face o referire la o anumită adresă, este probabil că următoarea referire la memorie va fi în vecinătatea acestei adrese. În spațiul adreselor de memorie, câteva regiuni au o probabilitate ridicată de a fi accesate, câteva au o probabilitate moderată, iar celelalte au o probabilitate foarte mică de a fi accesate în viitorul apropiat.

O regiune care are o probabilitate înaltă este cea corespunzătoare contorului de program actual, deoarece este probabil să se execute următoarea instrucțiune din secvența de instrucțiuni. Alte regiuni care au o probabilitate mare de a fi accesate sunt cele care conțin datele active, procedurile şi punctul de întoarcere dintr-o procedură. Dacă programul este scris într-un limbaj structurat pe blocuri, ca de exemplu *Pascal*, zona de stivă pentru variabile locale şi parametri este o altă zonă cu probabilitate ridicată de acces.

Observația că referințele la memorie efectuate într-un interval scurt de timp utilizează o mică porțiune a memoriei reprezintă *principiul localității*, şi formează baza sistemelor de memorie *cache*. Atunci când este adresat un cuvânt, acesta este transferat din memoria lentă în memoria *cache*, astfel încât la următoarea utilizare va putea fi accesat în mod rapid.

În Figura 8.11 se prezintă structura unui sistem de memorie format din memoria principală şi o memorie *cache*.

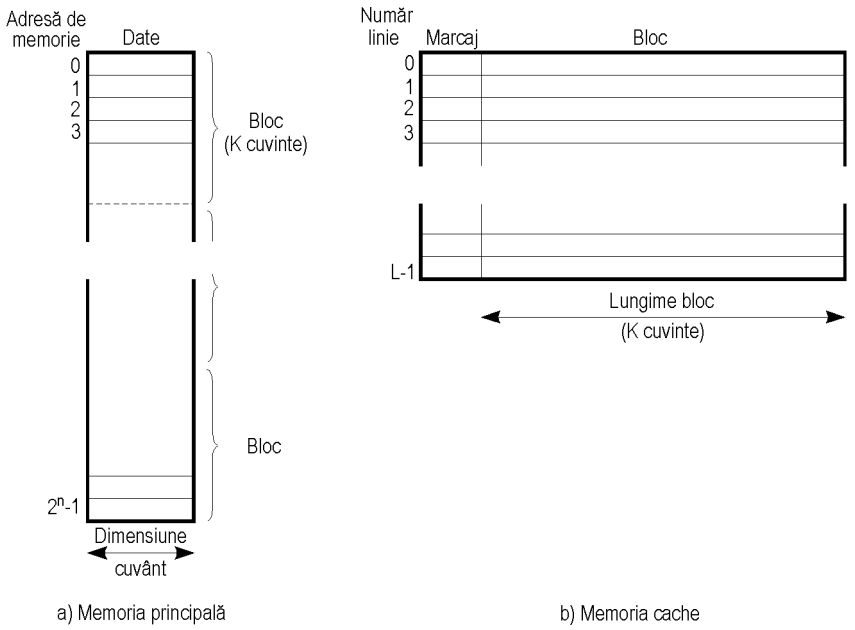

Figura 8.11. Structura unui sistem de memorie.

Memoria principală constă din 2^n cuvinte adresabile, fiecare cuvânt având o adresă unică de n biţi. Se consideră că această memorie este formată dintr-un număr de blocuri de lungime fixă de K cuvinte fiecare. Există deci $2^n/K$ blocuri. Memoria *cache* constă din blocuri de câte K cuvinte fiecare, un asemenea bloc fiind numit *linie*. Există L linii în memoria *cache*, numărul de linii fiind mult mai mic decât numărul blocurilor din memoria principală ($L << K$).

În orice moment, o anumită parte a blocurilor de memorie se află în liniile memoriei *cache*. Dacă se citeşte un cuvânt al unui bloc din memoria principală, blocul respectiv este transferat într-una din liniile memoriei *cache*. Deoarece există mai multe blocuri decât linii, o anumită linie nu poate fi dedicată în mod unic şi permanent unui anumit bloc. De aceea, fiecare linie conţine un *marcaj* care identifică blocul pe care îl conţine linia respectivă. Marcajul este de obicei o parte a adresei din memoria principală.

În Figura 8.12 se ilustrează operaţia de citire. UCP generează adresa unui cuvânt care trebuie citit (A). Dacă acest cuvânt se află în memoria *cache*, este transferat unităţii centrale. În caz contrar, blocul care conţine acest cuvânt este încărcat într-o linie a memoriei *cache*, iar cuvântul este transferat unităţii centrale.

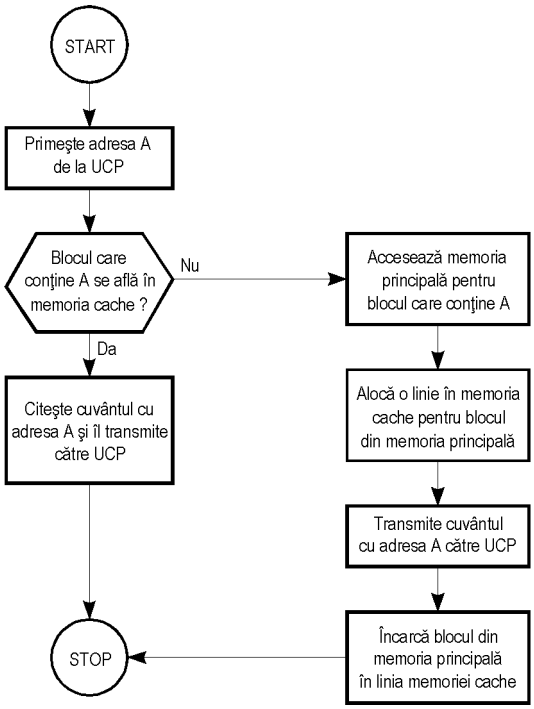

Figura 8.12. Operaţia de citire a memoriei *cache*.

Considerăm parametrii principali ai memoriei *cache*. *Costul mediu pe bit* C_S al sistemului de memorie format din memoria principală şi memoria *cache* este dat de:

$$C_S = \frac{C_C D_C + C_M D_M}{D_C + D_M} \tag{8.2}$$

unde:

C_C = costul mediu pe bit al memoriei *cache*;
C_M = costul mediu pe bit al memoriei principale;
D_C = dimensiunea memoriei *cache*;
D_M = dimensiunea memoriei principale.

Este de dorit ca pentru memoria *cache* costul C_C să fie aproximativ egal cu C_M. Deoarece $C_C >> C_M$, aceasta necesită ca $D_C << D_M$.

Timpul de acces al sistemului de memorie nu depinde numai de viteza memoriei *cache* şi a memoriei principale, ci şi de probabilitatea ca un anumit cuvânt adresat de UCP să fie găsit în memoria *cache*. Această probabilitate este numită *rată de succes*. Avem:

$$T_S = P_S\, T_C + (1 - P_S)\, T_M \tag{8.3}$$

unde:

T_S = timpul mediu de acces al sistemului de memorie;
T_C = timpul de acces al memoriei *cache*;
T_M = timpul de acces al memoriei principale;
P_S = rata de succes.

Este de dorit ca $T_S \approx T_C$. Deoarece $T_C << T_M$, este necesară o rată de succes apropiată de 1.

Studiile au arătat că o dimensiune relativ redusă a memoriei *cache*, de 64 KB sau 128 KB, este în general adecvată, obținându-se o rată de succes de peste 0,75, indiferent de dimensiunea memoriei principale.