Fişiere cu organizare directă

10

Fișiere cu organizare directă

- *Organizarea directă* este o metodă de determinare a adresei unei înregistrări pe baza valorii unuia sau mai multor câmpuri (reprezentând, în general, cheia)

- Caz ideal: utilizarea unei funcții ce calculează adresa înregistrării (*funcție de dispersie*):

$$h: \{K_1, K_2, \ldots, K_n\} \rightarrow A,$$

$h(K_i)$ = adresa celei de-a i-lea înregistrări

(K_i e valoarea cheii înregistrării i,
A este mulțimea de adrese de memorie)

Fișiere cu organizare directă

- *În practică*: e greu de definit astfel de funcții:
  - trebuie cunoscute de la început toate valorile posibile ale cheilor
  - pentru tabele mari este aproape imposibil de menținut bijectivitatea

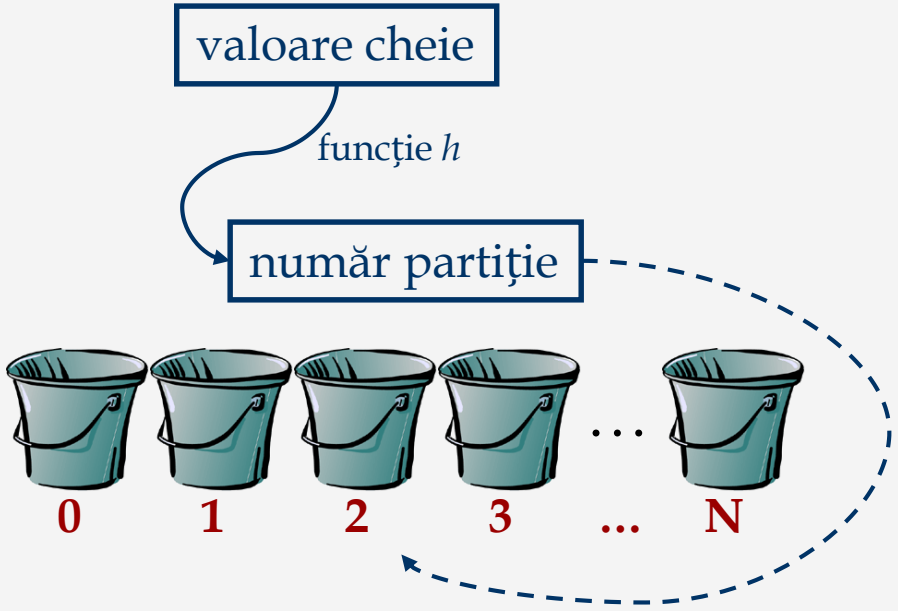
- Soluție → sunt permise coliziunile:

  $h(K_i) = h(K_j)$, pt. $i \neq j$ (*h – 'funcție de dispersie'*)

  - Înregistrările cu valorile pentru cheie K_i și K_j se numesc *sinonime*

Partiții (*buckets*)

- Toate înregistrările sinonime sunt stocate într-o partiție care începe la adresa returnată de *h*.
- O partiție este de obicei un bloc de memorie.



Provocări ale organizării directe

- Distribuirea – odată ales algoritmul de dispersie se pierde controlul asupra distribuirii înregistrărilor pe partiții.
- Gruparea – majoritatea înregistrărilor sunt plasate în aceeași partiție iar restul partițiilor conțin doar puține înregistrări.
- Depășirea – este posibilă depășirea spațiului de memorie alocat inițial unei partiții.

Definirea funcțiilor de dispersie

Cerințele pentru o funcție de dispersie ‘bună’:

- evaluare rapidă

- minimizează numărul de coliziuni

(dispersie uniformă a înregistrărilor pe partiții)

Definirea funcțiilor de dispersie

Fie 41 partiții goale unde adaugăm înregistrări:

- probabilitatea plasării înregistrării 1 într-o partiție goală e 41/41
- a doua înregistrare - 40/41,
- a treia înregistrare - 39/41, etc.

Probabilitatea plasării a 8 înregistrări în 8 partiții goale este :

$$(41/41)(40/41)(39/41)\ldots(34/41) = 0.482$$

$\Rightarrow$ sub 50%!!!

Definirea funcțiilor de dispersie

- Abordări:
  - Diviziune
  - Metoda pătratului (*mid-square)*
  - Împachetare (*folding*)
  - Multiplicare

- În general funcțiile de dispersie se aplică pe ***reprezentarea binară*** a cheii de căutare (de exemplu, pentru o cheie de căutare de tip text, se pot însuma reprezentările binare ale tuturor caracterelor acesteia iar rezultatul este transmis ca parametru funcției de dispersie).

Definirea funcțiilor de dispersie

- Diviziune:
  - Funcții de forma h(k) = k *mod* N
  - Garantează plasarea rezultatului h(k) în intervalul [0 ...N-1]
  - E ușor de calculat $\Rightarrow$ cea mai populară abordare
  - Dacă alegem $N = 2^d$, pentru un *d* număr natural oarecare, doar ultimii *d* biți ai lui *k* vor fi luați în considerare
  - S-a arătat că numerele prime sunt cele mai potrivite ca valori ale lui N

Definirea funcțiilor de dispersie

- Metoda pătratului (*mid-square*):
  - Se ridică numărul la pătrat apoi se extrag l cifre din mijlocul rezultatului.
  - $l = \text{lung}(k^2) - c_1 - c_2$, unde c_1 și c_2 sunt reprezintă numărul celor mai puțin semnificative, respectiv cele mai semnificative cifre eliminate din pătratul lui l
  - l si c_1 sunt fixate de la început

Definirea funcțiilor de dispersie

- Împachetare (*folding*):
  - Fiecare cheie *k* este partiționată în bucăți de lungimi egale (cu excepția ultimei) k_1, k_2, k_3,...,k_n
  - Funcția de dispersie e dată de formula

$$h(k) = (k_1 + k_2 + k_3 + ... + k_n) \bmod 10^l$$

unde *l* e numărul maxim de partiții

Definirea funcțiilor de dispersie

- Multiplicare
  - Se extrage partea fracționară a numărului Z * k (pentru un Z specific) și se înmulțește cu N (N fiind numărul partițiilor):

$$h(k) = \lfloor N * (Z*k - \lfloor Z*k \rfloor) \rfloor = \lfloor N * \{Z*k\} \rfloor$$

  - Cele mai bune rezultate se obțin pentru

$Z = (\sqrt{5} - 1)/2 = 0.61803...$ sau $Z = (3 - \sqrt{5})/2 = 0.38196...$

  - Pentru $Z = Z'/2^w$ și $N = 2^d$ (w: număr de biți într-un cuvânt)

$$h(k) = msb_d (Z'*k)$$

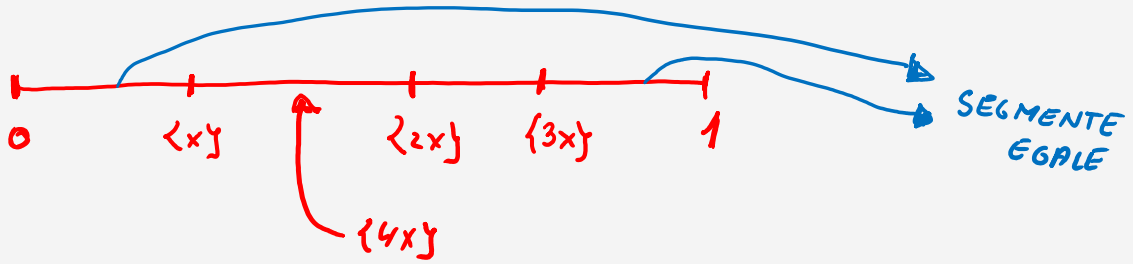
unde $msb_d(x)$ reprezintă cei mai semnificativi d biți ai lui x

(exemplu: 42 a reprez. binară 101010, $msb_3(42) = 5$ – binar 101)

Definirea funcțiilor de dispersie

Teoremă. Având un număr irațional x și plasând {x}, {2x},...,{nx} pe segmentul [0,1] se obțin n+1 segmente ce au cel mult 3 dimensiuni diferite.

De asemena, următoarea valoare, {(n+1)x}, va fi plasată pe unul dintre segmentele cu dimensiunea cea mai mare.



Exemple de funcții de dispersie

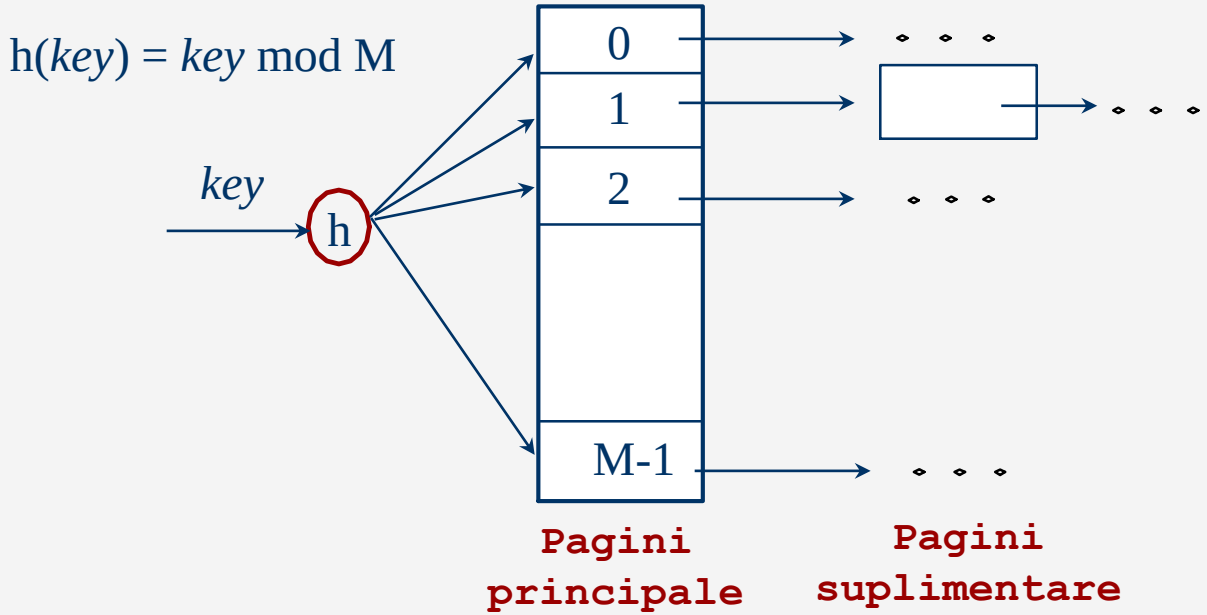
- Valoare cheie '*Toyota*'
  - Concatenând reprezentarea numerică a caracterelor de pe primele două poziții obținem: ***To***yota ⇒ | ***20*** | ***15*** |
- Funcții de dispersie posibile
  - *Diviziune*: mod by 97 → **2015** *mod* 97 = 75
  - *Metoda pătratului*: **2015²** = 4060225
  - *Împachetare*: **2015** → 20+15 = 35
  - *Multiplicare*: ⌊99*{**2015***0.61803}⌋ =32
- De ce nu se utilizează **2015** ca valoare de dispersie?
  - 4 cifre → 10000 valori posibile! → o bună parte a partițiilor vor fi goale
  - pentru cazul nostru 100 partiții sunt suficiente

Rezolvarea coliziunilor

- Noile înregistrări se memorează într-o zonă suplimentară
- Se aplică o a doua funcție de dispersie pentru a obține o adresă "*second choice*".
- Se memorează doar referințe (pointeri). La adresa returnată de funcția de dispersie avem:
  - Toți pointerii înregistrărilor sinonime – **partiție** de adrese.
  - Pointer către prima înregistrare. Aceasta va conține la rândul său un pointer către a doua înregistrare sinonimă, ș.a.m.d. – **listă înlănțuită**.
- Dispersie statică vs. dispersie dinamică

Dispersie statică

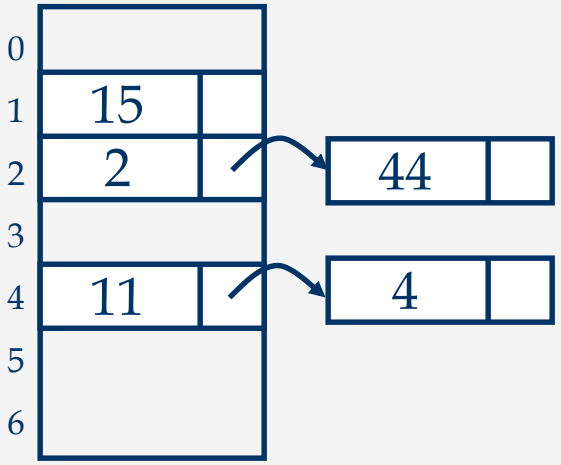
- Numărul paginilor/sloturilor *principale* e fix
- Paginile/sloturile sunt alocate secvențial, și nu sunt de-alocate; se pot utiliza pagini/sloturi *suplimentare* dacă este nevoie
- **h**(*k*) = *k* mod M = partiția căreia îi aparține intrarea cu cheia *k*. (M = număr de partiții)



Dispersie statică: Liste independente

- Toate sinonimele sunt stocate într-o listă înlănțuită specifică
- Fișierul conține o listă de M intrări reprezentând capul unei liste de sinonime (partiții)
- Ordinea sinonimelor în listă poate fi:
  - ordinea de inserare
  - ordine descrescătoare a frecvenței de căutare
  - ordinea crescătoare a cheii de căutare (căutările fără succes se vor termina mai rapid)

k	h(k) = k mod 7
11	4
2	2
44	2
4	4
15	1

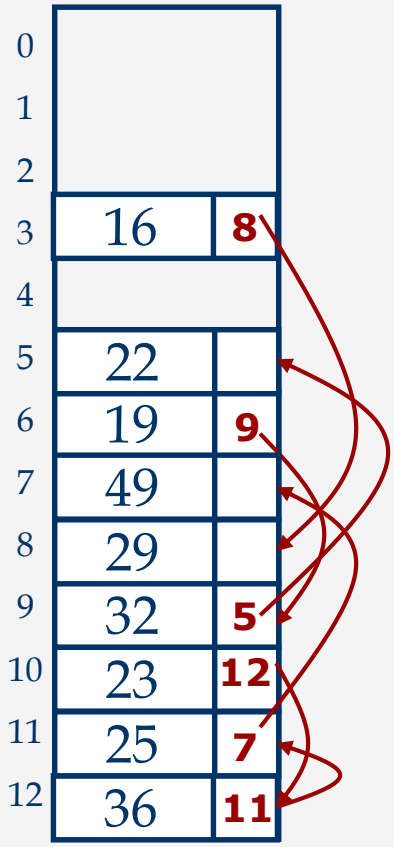


Dispersie statică: Liste întrețesute

- Nu se utilizează pagini/sloturi suplimentare
- La inserarea unei noi intrări cu valoare cheii K:
  - Dacă slotul de la adresa h(K) e gol: se memorează înregistrarea
  - Dacă slotul de la adresa h(K) e ocupat:
    - se ocupă primul slot liber (căutând de la baza fişierului)
    - slotul ocupat e inserat la finalul listei ce conţine slotul referit de h(*K*)

Exemplu:

k	h(k) = k mod 13
16	3
23	10
36	10
25	12
19	6
32	6
29	3
49	10
22	9



Dispersie statică: Liste întrețesute

- Ștergerea unei înregistrări cu valoarea cheii *K*:
  - Dacă slotul de la adresa h(K) e gol: mesaj de eroare
  - Dacă slotul h(K) e ocupat:

    - parcurgem lista pentru a căuta înregistrarea. Dacă aceasta nu e găsită: mesaj de eroare

      1. Dacă înregistrarea e găsită: se elimină
      2. Se caută, în continuarea listei, o înregistrare *r* cu $h(K_r) = h(K)$

      • dacă e găsită

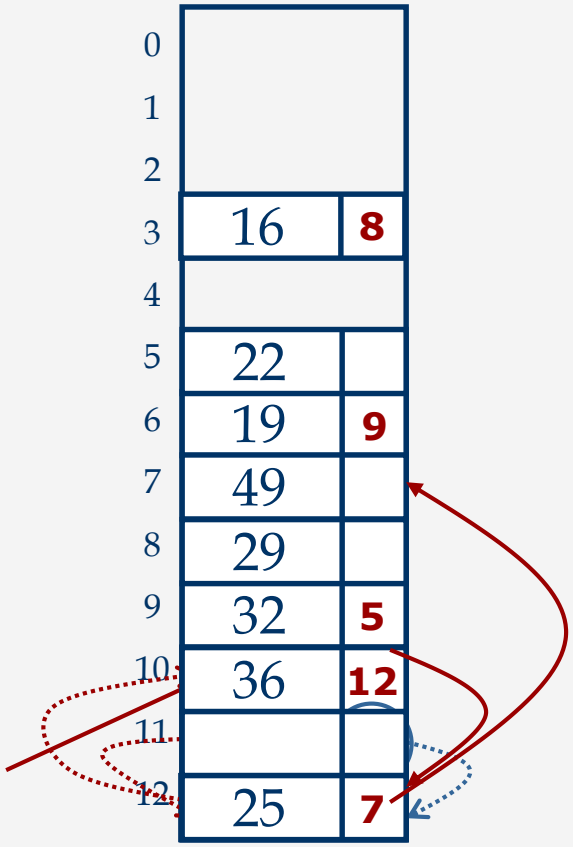
        • se mută înregistrarea în slotul curent.

        • se repetă pasul 2 pentru noul slot gol

      3. se copiază pointerul slotului gol în precedentul element din listă (dacă există)

Dispersie statică: Liste întrețesute

- Exemplu: eliminarea cheii cu valoarea cheii 23.



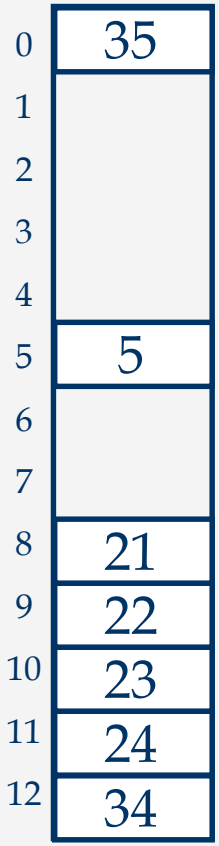
k	h(k) = k mod 13
16	3
23	10
36	10
25	12
19	6
32	6
29	3
49	10
22	9

Dispersie statică: Adresare deschisă

- Fișierul conține doare date (fără pointeri)
- Inserarea unei înregistrări cu valoarea cheii K:
  - Dacă slotul de la adresa h(K) e gol: se memorează înregistrarea
  - Dacă slotul de la adresa h(K) e ocupat se caută următorul slot liber la adresele h(K)+1, h(K)+2, ..., M-1, 0, ..., h(K)-1.
- Potrivit pentru ocupare de 75%

Exemplu:

k	h(k) = k mod 13
5	5
21	8
24	11
22	9
23	10
34	8
35	9



Dispersie statică: Adresare deschisă

- Ștergerea unei înregistrări cu valoarea cheii K:

  A. Înlocuirea valorii K cu un cod sau caracter special (vor trebui ajustați algoritmii de căutare și inserare pentru a interpreta corect valorile speciale)

  B. Se elimină înregistrarea și se fac transferuri de înregistrări astfel (*i*, *j*, *p* sunt adrese de memorie):
  - *i* e adresa înregistrării șterse,
  - nu sunt sloturi goale între *i* și *j*
  - înregistrarea stocată la *j* trebuie să fie stocată la *p*

i<j

i < p ≤ j : -

j ≤ p ≤ m-1 : înreg. din j se transferă la i

0 ≤ p ≤ i : înreg. din j se transferă la i

j <i

0< p ≤ j : -

j < p ≤ i înreg. din j se transferă la i

i < p ≤ i : -

Concluzii asupra dispersiei statice

- Partițiile conțin *înregistrări*.
- Dacă se adaugă multe înregistrări, se formează liste de pagini de memorie suplimentare (crește numărul de pagini transferate în memoria internă).
- Similar, dacă numărul înregistrărilor descrește semnificativ, este o risipă de spațiu de stocare (sloturile din partiții rămân nealocate).

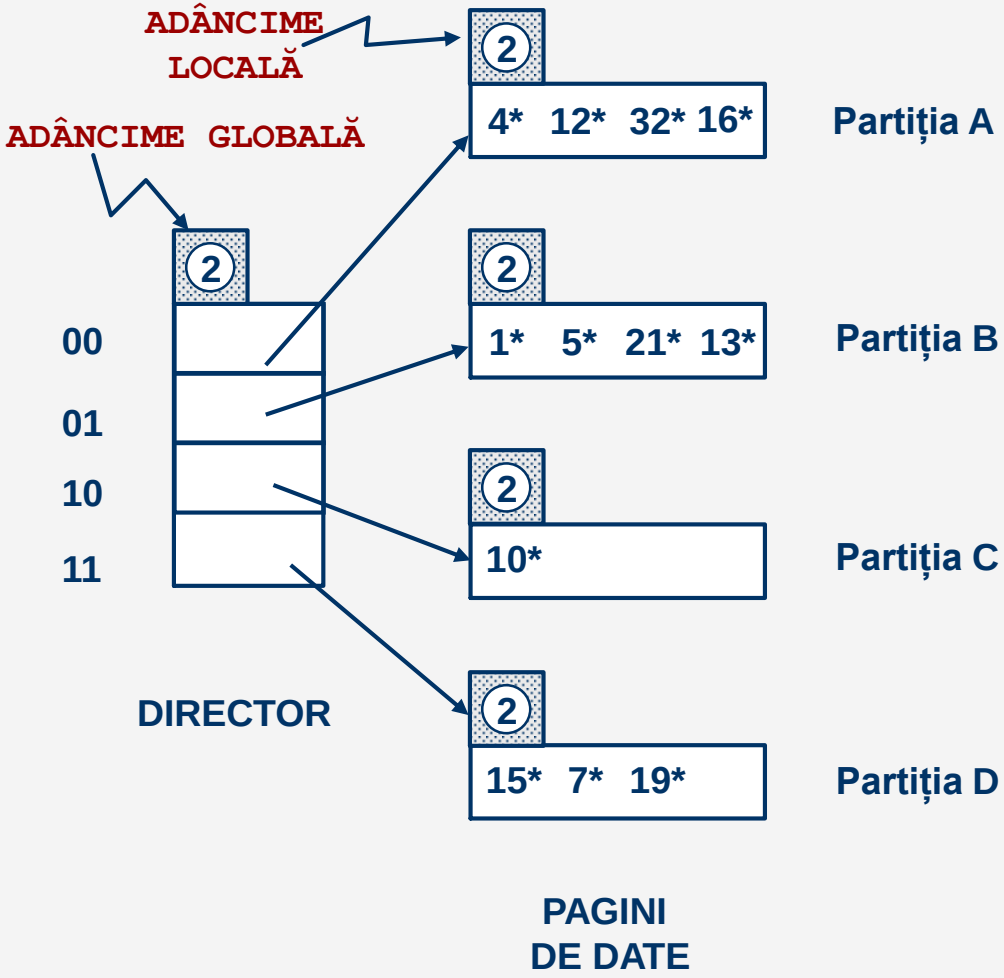
- *Dispersie extensibilă* și *liniară*: tehnici dinamice de abordare a dezavantajelor dispersiei statice.

Dispersie extensibilă

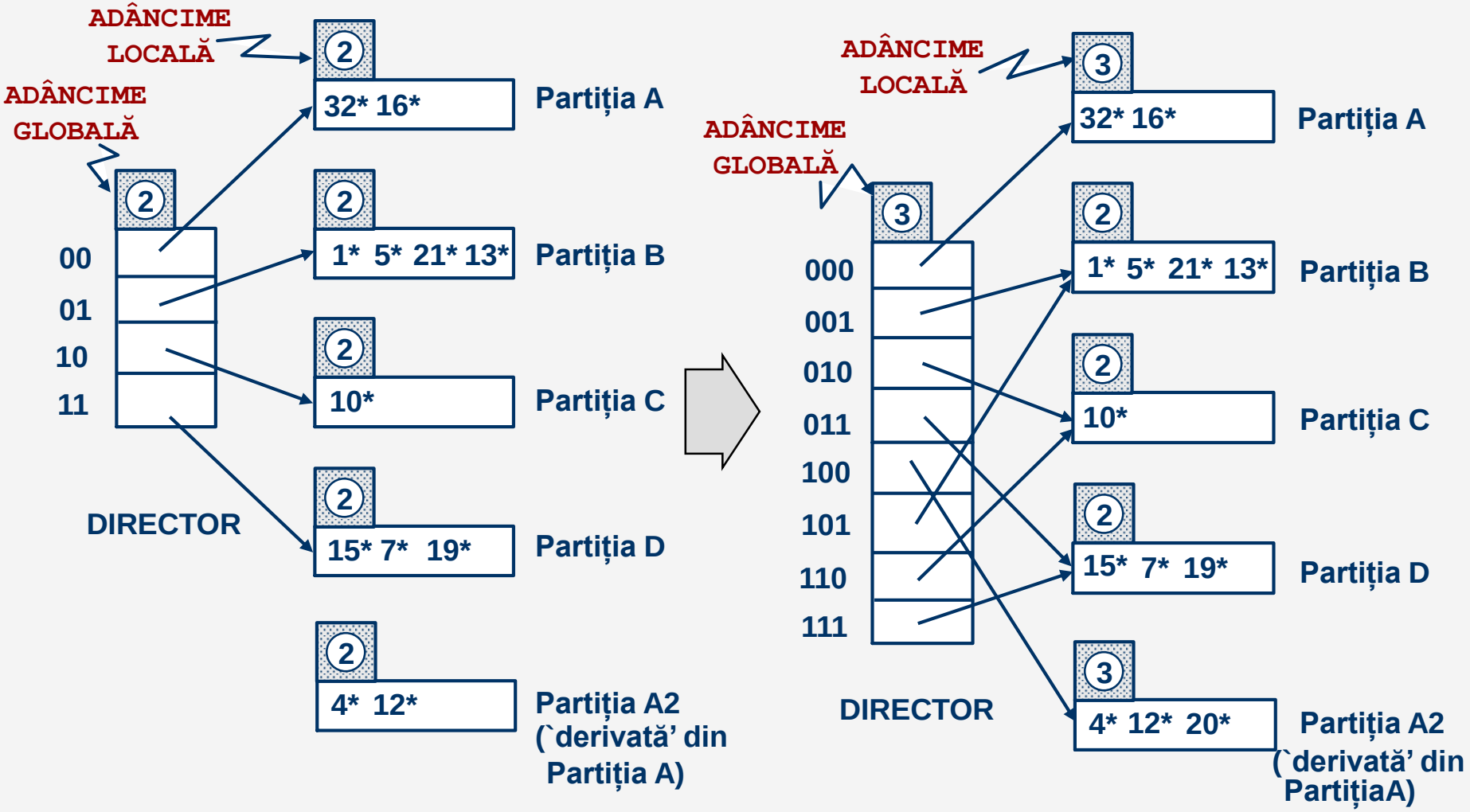
- Idee: Dacă partițiile primare devin neîncăpătoare, fișierul se reorganizează prin *dublarea* numărului de partiții
  - Fișierul va conține un *director de pointeri la partiții*,
  - Dublarea numărului de partiții se realizează prin *dublarea directorului,*
  - Se distribuie înregistrările doar din partiția ce a devenit neîncăpătoare
  - *Nu există pagini suplimentare*!
  - Funcția de dispersie va fi ajustată corespunzător

Exemplu

- Directorul are dimensiunea 4.
- Pentru a determina partiția lui *r*, se iau în considerare ultimii biți din reprezentarea binară a lui **h**(*r*) (numărul de biți luați în considerare e dat de `*adâncimea globală*').
  - Dacă **h**(*r*) = 5 = binar 101, se află în partiția referită de 01.
- **Inserarea**: Dacă partiția e plină, acesta se împarte în două (*se alocă o nouă pagină și se redistribuie înregistrări*).
- *Dacă e necesar*, directorul se dublează. (doar când *adâncimea globală* devine mai mică decât *adâncimea locală* a noii partiții)



Inserare k: h(k) = 20 → Dublare

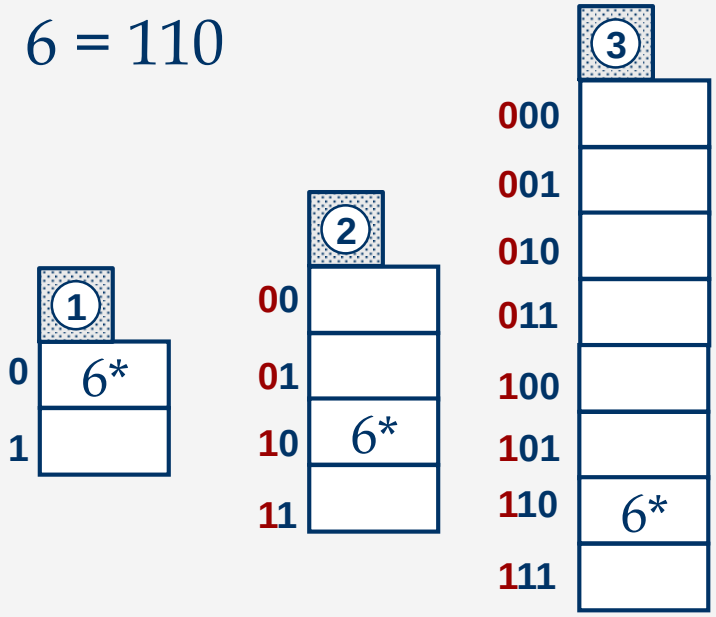


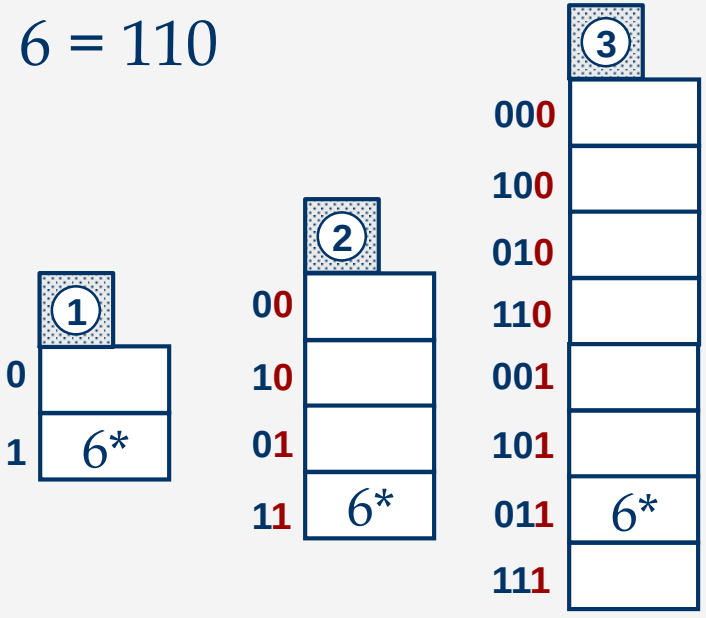
Dispersie extensibilă

- Inserare h(k) = 20 (binar 10100). Ultimii **2** biți (00) arată că *r* aparține lui A sau A2. E nevoie de ultimii **3** biți pentru a indetifica partiția potrivită.
  - *Adâncimea globală a directorului*: numărul maxim de biți necesari pentru a determina partiția căreia îi aparține o înregistrare.
  - *Adâncimea locală a partiției*: numărul de biți utilizați pentru a determina dacă o înregistrare aparține partiției.
- În ce condiții adăugarea unei noi partiții implică dublarea directorului?
  - Înaintea inserării, *adâncimea locală* partiției = *adâncimea globală*. Inserarea face ca *adâncimea locală* să devină > *adâncimea globală*; directorul se dublează prin copiere

Dublarea directorului

- Utilizarea celor mai puțin semnificativi biți → dublarea directorului se realizează prin copiere!





Cel mai puțin semnificativ vs. Cel mai semnificativ

Comentarii despre dispersia extensibilă

- Dacă directorul începe în memoria internă, interogările bazate pe egalități au nevoie de un singur acces la disc (suport extern);
  - Directorul crește brusc iar dacă distribuția valorilor funcției de dispersie nu e una normală, directorul se poate mări excesiv.
  - Sinonimele multiple pot crea probleme!
- **Ștergeri**: Dacă la eliminarea unei înregistrări o partiție se golește, aceasta poate fuziona cu perechea sa (ținând cont de valorile ultimilor biți ai adâncimii locale). Dacă toate adâncimile locale sunt mai mici decât adâncimea globală, directorul se poate înjumătăți.

Evaluarea fișierelor cu organizare directă

- Avantaje:
  - Acces foarte rapid
  - Cea mai bună soluție pentru *interogări cu egalități*

```
SELECT * FROM R WHERE A = k
```

deoarece (în general) indiferent de numărul de înregistrări adăugate sau șterse, răspunsul la interogare va fi dat de citirea unei singure pagini de memorie de pe suportul extern.

Evaluarea fișierelor cu organizare directă

- Dezavantaje:
  - Ordinea secvențială a înregistrărilor nu are nici o semnificație
  - Fișierele cu organizare directă sunt deseori fragmentate.
  - Nu suportă interogări cu intervale.
  - Nu sunt recomandate atunci când câmpul transmis ca parametru funcției de dispersie este actualizat frecvent