Κατακερματισμός

Οργάνωση Αρχείων (σύνοψη)

Οργάνωση αρχείων: πως είναι τοποθετημένες οι εγγραφές ενός αρχείου όταν αποθηκεύονται στο δίσκο

1. Αρχεία Σωρού
2. Ταξινομημένα Αρχεία

   Φυσική διάταξη των εγγραφών ενός αρχείου με βάση την τιμή ενός από τα πεδία του το οποίο λέγεται **πεδίο διάταξης** (ordering field)

Οργάνωση Αρχείων

- Αρχεία Κατακερματισμού

Βασική ιδέα: η τοποθέτηση των εγγραφών στα blocks του αρχείου γίνεται εφαρμόζοντας μια συνάρτηση κατακερματισμού σε κάποιο από τα πεδία της εγγραφής

Εσωτερικός Κατακερματισμός

Εσωτερικός Κατακερματισμός (τα δεδομένα είναι στη μνήμη, όπως στις δομές δεδομένων)

Πίνακας κατακερματισμού με M θέσεις - κάδους (buckets)

h: *συνάρτηση κατακερματισμού*

h(k) = i ← Σε ποιο κάδο - τιμή από 0 έως M-1

Πεδίο αναζήτησης - Πεδίο κατακερματισμού

Αρχεία Κατακερματισμού

Εξωτερικός Κατακερματισμός (εφαρμογή σε δεδομένα αποθηκευμένα σε αρχεία)

Στόχος

h(k) = i ← Διεύθυνση (αριθμός) block του αρχείου που είναι αποθηκευμένη

Τιμή του πεδίου κατακερματισμού

Η εγγραφή με τιμή στο πεδίο κατακερματισμού k αποθηκεύεται στο i-οστο block (κάδο) του αρχείου

Κατακερματισμός

h: *συνάρτηση κατακερματισμού*

Ομοιόμορφη κατανομή των κλειδιών στους κάδους (blocks)

- Συνηθισμένη συνάρτηση κατακερματισμού:

$$h(k) = k \bmod M$$

Συχνά M πρώτος

• **Σύγκρουση (collision)**: όταν μια νέα εγγραφή κατακερματίζεται σε μία ήδη γεμάτη θέση

• **Καλή συνάρτηση κατακερματισμού**: κατανέμει τις εγγραφές ομοιόμορφα στο χώρο των διευθύνσεων (ελαχιστοποίηση συγκρούσεων και λίγες αχρησιμοποίητες θέσεις)

• **Ευριστικοί**:

-- αν r εγγραφές, πρέπει να επιλέξουμε το M ώστε το r/M να είναι μεταξύ του 0.7 και 0.9

-- όταν χρησιμοποιείται η mod τότε είναι καλύτερα το M να είναι πρώτος

Επίλυση Συγκρούσεων

1. *Ανοιχτή Διευθυνσιοδότηση* (open addressing): χρησιμοποίησε την επόμενη κενή θέση

2. *Αλυσιδωτή Σύνδεση* (chaining): για κάθε θέση μια συνδεδεμένη λίστα με εγγραφές υπερχείλισης

3. *Πολλαπλός Κατακερματισμός* (multiple hashing): εφαρμογή μιας δεύτερης συνάρτησης κατακερματισμού

Κάδος: μια συστάδα από συνεχόμενα blocks του αρχείου

h(k) = i ← Σχετική διεύθυνση του κάδου (ποιος κάδος του αρχείου)

Τιμή του πεδίου κατακερματισμού

Ο κατακερματισμός είναι πολύ αποδοτικός για επιλογές (ερωτήσεις) ισότητας

Ένας πίνακας που αποθηκεύεται στην επικεφαλίδα του αρχείου μετατρέπει τον αριθμό κάδου στην αντίστοιχη διεύθυνση block

0	διεύθυνση 1ου block του κάδου στο δίσκο
1	διεύθυνση 1ου block του κάδου στο δίσκο
2	διεύθυνση 1ου block του κάδου στο δίσκο
…	…
M-1	διεύθυνση 1ου block του κάδου στο δίσκο

Εξωτερικός Κατακερματισμός

Συγκρούσεις - αλυσιδωτή σύνδεση - εγγραφές υπερχείλισης ανά κάδο

1. **Ανάγνωση όλου του αρχείου (scan)**

   Έστω ότι διατηρούμε κάθε κάδο γεμάτο κατά 80% άρα ένα αρχείο με μέγεθος B blocks χρειάζεται 1.25 B blocks

   $1.25 * B * (T_D + R * T_C)$

2. **Αναζήτηση**

   Συνθήκη ισότητας και μόνο ένα block ανά κάδο: $T_D + R * C$

   Αν συνθήκη περιοχής (διαστήματος): scan!

Οργάνωση Αρχείων

Κόστος: μεταφορά blocks (I/O)

	Σωρός	Ταξινομημένο	Κατακερματισμένο
Ανάγνωση του αρχείου	B	B	1.25B
Αναζήτηση με συνθήκη ισότητας	0.5 B	logB	1
Αναζήτηση με συνθήκη περιοχής	B	logB + ταιριάσματα	1.25 B
Εισαγωγή	2	αναζήτηση + B	2
Διαγραφή	αναζήτηση + 1	αναζήτηση + B	αναζήτηση + 1

Πρόβλημα: Στατικός Κατακερματισμός

Έστω M κάδους και r εγγραφές ανά κάδο - το πολύ M * r εγγραφές (αλλιώς μεγάλες αλυσίδες υπερχείλισης)

Δυναμικός Κατακερματισμός

- Επεκτατός
- Γραμμικός

Δυναμικός Εξωτερικός Κατακερματισμός

- Δυαδική αναπαράσταση του αποτελέσματος της συνάρτησης κατακερματισμού, δηλαδή ως μια ακολουθίας δυαδικών ψηφίων
- Κατανομή εγγραφών με βάση την τιμή των αρχικών *(ή τελικών)* ψηφίων

Δυναμικός Εξωτερικός Κατακερματισμός

- Το αρχείο ξεκινά με ένα μόνο κάδο

- Μόλις γεμίσει ένας κάδος διασπάται σε δύο κάδους με βάση την τιμή του 1ου (ή τελευταίου) δυαδικού ψηφίου των τιμών κατακερματισμού -- δηλαδή οι εγγραφές που το πρώτο (τελευταίο) ψηφίο της τιμής κατακερματισμού τους είναι 1 τοποθετούνται σε ένα κάδο και οι άλλες (με 0) στον άλλο

- Νέα υπερχείλιση ενός κάδου οδηγεί σε διάσπαση του με βάση το αμέσως επόμενο δυαδικό ψηφίο κοκ

Δυναμικός Εξωτερικός Κατακερματισμός

Έτσι δημιουργείται μια δυαδική δενδρική δομή που λέγεται **κατάλογος** (dirtectory) ή **ευρετήριο** (index) με δύο ειδών κόμβους

- εσωτερικούς: που καθοδηγούν την αναζήτηση
- εξωτερικούς: που δείχνουν σε ένα κάδο

Δυναμικός Εξωτερικός Κατακερματισμός (Παράδειγμα)

Χρήση των τελευταίων bits της δυαδικής αναπαράστασης

Αποτέλεσμα συνάρτησης κατακερματισμού	
1	000001
4	000100
5	000101
7	000111
10	001010
12	001100
15	001111
16	010000
19	010011
21	010101
32	100000
13	001101
20	010100

4 εγγραφές ανά κάδο

Δυναμικός Εξωτερικός Κατακερματισμός

Αλγόριθμος αναζήτησης

```
h := τιμή κατακερματισμού
t  := ρίζα του δέντρου
i := 1
while (t εσωτερικός κόμβος)
        if (i-οστό bit του h είναι 0)
                t := αριστερά του t
        else t := δεξιά του t
     i := i +1
```

Δυναμικός Εξωτερικός Κατακερματισμός

- *Που αποθηκεύεται ο κατάλογος*

  στη μνήμη, εκτός αν είναι πολύ μεγάλος

  τότε στο δίσκο – οπότε θα απαιτούνται επιπρόσθετες προσπελάσεις

- Δυναμική επέκταση αλλά *μέγιστος αριθμός* επιπέδων (το πλήθος των δυαδικών ψηφίων της συνάρτησης κατακερματισμού)

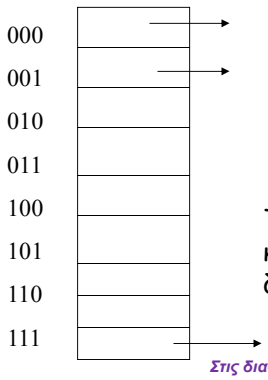
- Ισοζύγιση

- Συνένωση κάδων (δυναμική συρρίκνωση)

Επεκτατός Εξωτερικός Κατακερματισμός

Extendible hashing

Ο **κατάλογος** είναι ένας πίνακας με 2^d διευθύνσεις κάδων (**d: *ολικό βάθος* του καταλόγου**)



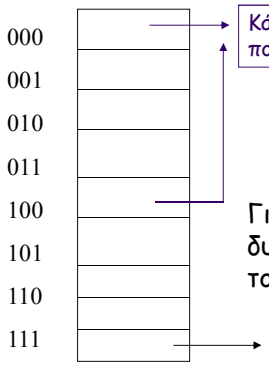
Κάδος για τις εγγραφές με τιμές κατακερματισμού που τελειώνουν σε 000

Τα τελευταία d ψηφία της τιμής κατακερματισμού χρησιμοποιούνται ως δείκτης στον πίνακα

Στις διαφάνειες, χρησιμοποιούμε τα τελευταία bits της δυαδικής αναπαράστασης

Επεκτατός Εξωτερικός Κατακερματισμός

Δε χρειάζεται ένας διαφορετικός κάδος για κάθε μία από τις 2^d θέσεις - μπορεί η θέση του πίνακα να δείχνει στη διεύθυνση του ίδιου κάδου αν αυτές χωράνε σε ένα κάδο



Κάδος για τις εγγραφές με τιμές κατακερματισμού που τελειώνουν από 00

Για κάθε κάδο, **τοπικό βάθος d'** ο αριθμός των δυαδικών ψηφίων στα οποία βασίζεται η χρήση του κάδου

Παράδειγμα: 2 εγγραφές ανά κάδο

εισαγωγή 2, 4, 3, 10, 7, 9

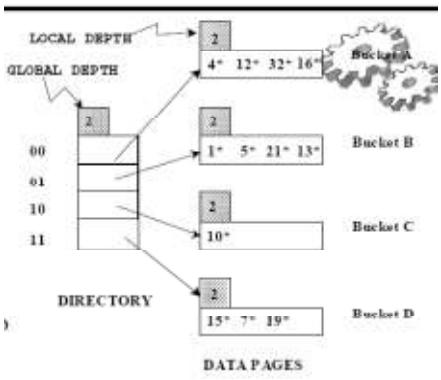
Επεκτατός Εξωτερικός Κατακερματισμός (Παράδειγμα)

Χρήση των τελευταίων bits της δυαδικής αναπαράστασης

1	000001
4	000100
5	000101
7	000111
10	001010
12	001100
15	001111
16	010000
19	010011
21	010101
32	100000
13	001101

4 εγγραφές ανά κάδο

Επεκτατός Εξωτερικός Κατακερματισμός (Παράδειγμα)



Χρήση των τελευταίων bits της δυαδικής αναπαράστασης

1	000001
4	000100
5	000101
7	000111
10	001010
12	001100
15	001111
16	010000
19	010011
21	010101
32	100000
13	001101

Επεκτατός Εξωτερικός Κατακερματισμός

Η τιμή του d μπορεί να αυξάνεται (μέχρι $2^κ$, κ: αριθμός δυαδικών ψηφίων της τιμής κατακερματισμού) ή να μειώνεται

- **Αύξηση της τιμής του d**

  Όταν ένας κάδος με τιμή d' = d υπερχειλίσει

  Διπλασιασμός του πίνακα

  Δε χρειάζεται rehash (επανα-κερματισμό), διασπάμε κάθε κάδο

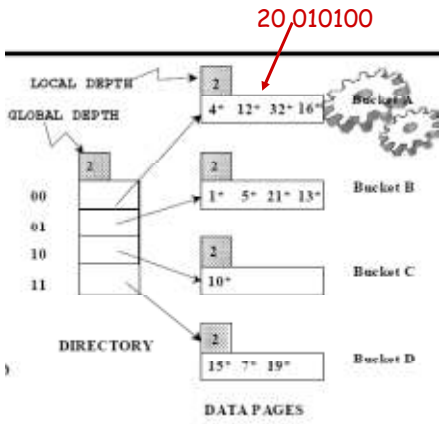
  Επίσης, κάθε φορά μόνο τον κάδο που υπερχείλισε

- **Μείωση της τιμής του d**

  Όταν για όλους τους κάδους d' < d

  Μείωση του μεγέθους του πίνακα στο μισό

Επεκτατός Εξωτερικός Κατακερματισμός (Παράδειγμα)



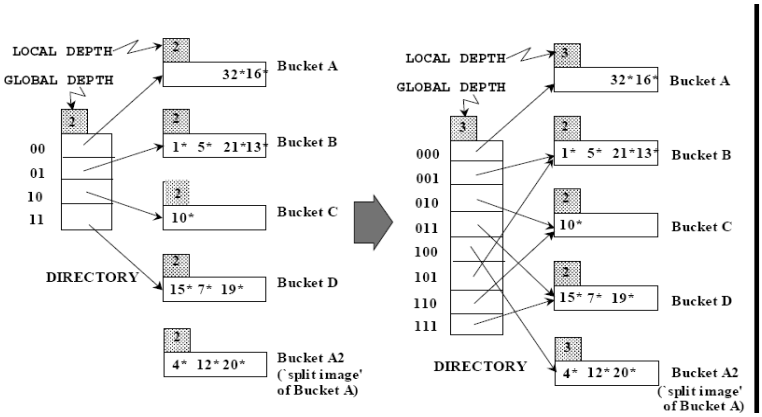
1	000001
4	000100
5	000101
7	000111
10	001010
12	001100
15	001111
16	010000
19	010011
21	010101
32	100000
13	001101

Διάσπαση -> Ολικό βάθος 3

Επεκτατός Εξωτερικός Κατακερματισμός

1	000001
4	000100
5	000101
7	000111
10	001010
12	001100
15	001111
16	010000
19	010011
21	010101
32	100000
13	001101
20	010100

4 12 32 16 20 -> διάσπαση



ΠΡΟΣΟΧΗ – ΓΕΝΙΚΗ ΠΑΡΑΤΗΡΗΣΗ

Τι αποθηκεύουμε στους κάδους;

Στα παραδείγματα δείχνουμε μόνο την τιμή του πεδίου κατακερματισμού

- Την ίδια την εγγραφή; (οργάνωση αρχείου)
  - μέγεθος κάδου -> 1 block (ή συστοιχία από συνεχόμενα blocks)
- Τιμή του πεδίου κατακερματισμού (+δείκτη στο υπόλοιπο της εγγραφής);

Τι γίνεται αν το πεδίο κατακερματισμού δεν είναι κλειδί (παραπάνω από μια εγγραφή με την ίδια τιμή)

Γραμμικός Κατακερματισμός

Θέλουμε να αποφύγουμε τη χρήση καταλόγου +

Διπλασιασμό μεγέθους του καταλόγου

Προσοχή! Αυτή η μέθοδος:

- Διατηρεί λίστες υπερχείλισης
- Δε χρησιμοποιεί τη δυαδική αναπαράσταση

Γραμμικός Εξωτερικός Κατακερματισμός

Χρησιμοποιεί μια οικογένεια από συναρτήσεις κατακερματισμού

$h_0(k)$, $h_1(k)$, …, $h_d(k)$

Κάθε συνάρτηση *διπλάσιους κάδους* από την προηγούμενη:

$h_0(k) = k \bmod M$, $h_1(k) = k \bmod 2M$, $h_2(k) = k \bmod 4M$, …,

$h_j(k) = k \bmod 2^j M$

Όταν συμβαίνει η πρώτη υπερχείλιση ενός κάδου, πάμε στην επόμενη συνάρτηση μέχρι να διασπαστούν όλοι οι κάδοι με αυτήν τη συνάρτηση

ΠΡΟΣΟΧΗ: δε διασπάμε τον κάδο που υπερχειλίζει, αλλά έναν-έναν τον κάδο με τη σειρά!

Γραμμικός Εξωτερικός Κατακερματισμός

Βασικά σημεία

- Πολλές συναρτήσεις κατακερματισμού (άλλη σε κάθε βήμα)
- Οι κάδοι σε κάθε βήμα διασπώνται με τη σειρά (ο ένας μετά τον άλλο – ανεξάρτητα αν έχουν ή όχι υπερχειλίσει)
- *Επίσης, υποθέτουμε ότι κάθε υπερχείλιση, οδηγεί σε διάσπαση*

Γραμμικός Εξωτερικός Κατακερματισμός

Αρχικά:

Βήμα Διάσπασης (ποια συνάρτηση χρησιμοποιούμε) αρχικά **j = 0**:

Πλήθος Διασπάσεων (στο τρέχον βήμα) αρχικά **n = 0,**

j -> ποια συνάρτηση χρησιμοποιούμε
n -> ποιο κάδο διασπάμε

Έστω αρχικά M κάδους αριθμημένους από 0 έως M - 1 και

αρχική συνάρτηση κατακερματισμού

$$h_0(k) = k \bmod M$$

Γραμμικός Εξωτερικός Κατακερματισμός

j -> ποια συνάρτηση χρησιμοποιούμε
n -> ποιο κάδο διασπάμε

Όταν συμβεί μια υπερχείλιση σε έναν οποιοδήποτε κάδο, **ο κάδος 0** χωρίζεται σε δύο κάδους: τον αρχικό κάδο 0 και ένα νέο κάδο M στο τέλος του αρχείου με βάση την συνάρτηση $h_1(k) = k \bmod 2M$

Βήμα Διάσπασης (ποια συνάρτηση χρησιμοποιούμε) j = 1

Πλήθος Διασπάσεων n = 1

Συνεχίζουμε γραμμικά, διασπώντας με τη σειρά τους κάδους 1, 2, 3, ...

μέχρι να διασπαστούν όλοι οι «παλιοί» κάδοι

η μεταβλητή n («Πλήθος Διασπάσεων») κρατάει ποιος κάδος έχει σειρά για διάσπαση

Γραμμικός Εξωτερικός Κατακερματισμός

Βήμα διάσπασης (ποια συνάρτηση χρησιμοποιούμε) j = 1:

Πλήθος Διασπάσεων **n = m -1** :

Όταν συμβεί μια υπερχείλιση σε έναν οποιοδήποτε κάδο,

ο κάδος m -1 χωρίζεται σε δύο κάδους: τον αρχικό κάδο m - 1 και ένα νέο κάδο m + k - 1 στο τέλος του αρχείου με βάση την συνάρτηση $\mathbf{h_1(k) = k \bmod 2M}$

Δηλαδή, σε κάθε υπερχείλιση χωρίζουμε όλους τους κάδους με τη σειρά ξεκινώντας από τον πρώτο κάδο

Γραμμικός Εξωτερικός Κατακερματισμός

Συνεχίζουμε ...

Όλοι οι κάδοι έχουν διασπαστεί όταν: n = M

Τότε έχουμε 2M κάδους

Όταν **n = M**,

μηδενίζουμε το n, n = 0

και για οποιαδήποτε νέα διάσπαση εφαρμόζουμε την

$\mathbf{h_2(k) = k \bmod 4M}$

Διασπώντας πάλι τον κάδο 0, 1, ... κ.τ.λ

Γραμμικός Εξωτερικός Κατακερματισμός

Γενικά βήμα διάσπασης **j** (j = 0, 1, 2, …)

$h_j(k) = k \bmod 2^j M$,
και την $h_{j+1}(k)$ για διασπάσεις

Γραμμικός Εξωτερικός Κατακερματισμός

32
9
44
31
25
5
35
7
36
14
18
10
11
30

Κάθε κάδος 4 εγγραφές

Αρχικά 4 κάδους (M = 4)

ΠΡΟΣΟΧΗ: **Δε** χρησιμοποιούμε τη δυαδική αναπαράσταση

Γραμμικός Εξωτερικός Κατακερματισμός (παράδειγμα)

$h_0(k) = k \bmod 4$

$h_1(k) = k \bmod 8$

Για μη διασπασμένους κάδους: παλιά συνάρτηση

Για διασπασμένους κάδους: νέα συνάρτηση

Διασπάμε τον πρώτο κάδο

22

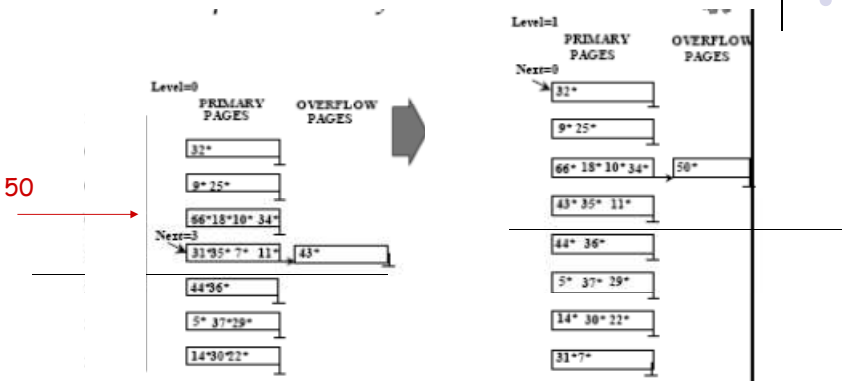
66

34

Βήμα διάσπασης 0 (χρήση h_0)

Πλήθος διασπάσεων = 0

Γραμμικός Εξωτερικός Κατακερματισμός (παράδειγμα)



Βήμα διάσπασης 0 (χρήση h_0)

Πλήθος διασπάσεων = 0

Αναζήτηση Εγγραφής (γενικά)

Τι χρειάζεται να ξέρουμε για να βρεθεί ο κάδος της εγγραφής k που ψάχνουμε;

- ποια συνάρτηση χρησιμοποιούμε (δηλαδή, το j)
- σε ποια διάσπαση βρισκόμαστε (δηλαδή το n)

Έστω ότι είμαστε στο βήμα j,

Τότε θα πρέπει να κοιτάξουμε είτε το

$h_j(k)$ αν ο κάδος δεν έχει διασπαστεί

ή το

$h_{j+1}(k)$ αν έχει διασπαστεί

Πως θα ελέγξουμε αν ο κάδος έχει διασπαστεί ή όχι

Αναζήτηση Εγγραφής

Δύο περιπτώσεις ο κάδος στον οποίο είναι (1) έχει ή (2) δεν έχει διασπαστεί

Κρατάμε μια μεταβλητή το πλήθος n των διασπάσεων

Έστω n ο αριθμός διασπάσεων και ότι αναζητούμε το k,

βρίσκεται στον κάδο $\mathbf{h_0(k)}$

τότε αν $\mathbf{n \leq h_0(k)}$ ο κάδος δεν έχει διασπαστεί

ενώ αν $\mathbf{n > h_0(k)}$ ο κάδος έχει διασπαστεί και εφαρμόζουμε την $h_1(k)$

Αλγόριθμος Αναζήτησης

j : βήμα διάσπασης n : πλήθος διασπάσεων στο βήμα j

```
if (n = 0)
    then m := h_j(k);
else {
     m := h_j(k);
     if (m < n) then m := h_{j+1}(k)
}
```

σημαίνει ότι ο κάδος έχει διασπαστεί