



# Ευρετήρια

## Ευρετήρια



- Ένα **ευρετήριο (index)** είναι μια βοηθητική δομή αρχείου που κάνει πιο αποδοτική την αναζήτηση μιας εγγραφής σε ένα αρχείο
- Το ευρετήριο καθορίζεται (συνήθως) σε **ένα γνώρισμα** του αρχείου που καλείται **πεδίο ευρετηριοποίησης (indexing field)**

γνώρισμα	

Αρχείο Ευρετηρίου

γνώρισμα	υπόλοιπα γνωρίσματα

Αρχείο Δεδομένων

Εγγραφή στο ευρετήριο:

Τιμή Πεδίου Ευρετηριοποίησης	Δείκτης στο block της εγγραφής
------------------------------	--------------------------------



Στόχος: αποδοτικές λειτουργίες αναζήτησης

Οι λειτουργίες ενημέρωσης γίνονται γενικά πιο αργές, γιατί απαιτούν ενημέρωση και του ευρετηρίου

Διαφορετικού τύπου εγγραφές ανάλογα με το πεδίο ευρετηριοποίησης:

(α) πεδίο διάταξης του αρχείου ή όχι

(β) κλειδί ή όχι

- (πρωτεύον/δευτερεύον) - διαφορετικοί ορισμοί στα βιβλία



- **Πυκνό ευρετήριο**: μια καταχώρηση για κάθε εγγραφή του αρχείου

- **Μη πυκνό ευρετήριο**



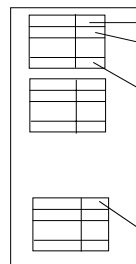
**Πρωτεύον ευρετήριο (primary index):** ορισμένο στο **κλειδί διάταξης** του αρχείου

Για κάθε **block** του **αρχείου** (μη πυκνό ευρετήριο) η εγγραφή  $i$  του ευρετηρίου είναι της μορφής  $\langle K(i), P(i) \rangle$  όπου:

- $K(i)$ : η τιμή του πρωτεύοντος κλειδιού της πρώτης εγγραφής του block (*άγκυρα* του block)
- $P(i)$ : δείκτης προς το block
- Το ευρετήριο στο πεδίο διάταξης (+ κλειδί) είναι ένα **μη πυκνό** ευρετήριο



Αρχείο  
Ευρετηρίου



Αρχείο  
Δεδομένων

*Ποιο είναι το μέγεθος του  
ευρετηρίου (πόσα blocks);*



- **Αναζήτηση**

**Διαδική αναζήτηση** στο πρωτεύον ευρετήριο  
Ανάγνωση του block από το αρχείο δεδομένων



- **Εισαγωγή εγγραφής**

αλλαγές και στο πρωτεύον ευρετήριο  
μη διατεταγμένο αρχείο υπερχείλισης  
συνδεδεμένη λίστα εγγραφών υπερχείλισης

- **Διαγραφή εγγραφής**

αλλαγές και στο πρωτεύον ευρετήριο  
χρήση σημαδιών διαγραφής



Παράδειγμα (υπολογισμός μεγέθους αρχείου ευρετηρίου)

Έστω διατεταγμένο αρχείο με  $r_A = 30.000$  εγγραφές, μέγεθος block  $B = 1024$  bytes, σταθερού μεγέθους εγγραφές μεγέθους  $R_A = 100$  bytes, όπου το πεδίο κλειδιού διάταξης έχει μέγεθος  $V_A = 9$  bytes, μη εκτεινόμενη καταχώρηση.

Κατασκευάζουμε πρωτεύον ευρετήριο, μέγεθος δείκτη block  $P = 6$  bytes

Μέγεθος αρχείου δεδομένων: 3.000 blocks

Μέγεθος αρχείου ευρετηρίου: 45 blocks



Παράδειγμα (υπολογισμός κόστους αναζήτησης)

Δεδομένα όπως πριν

(Έστω διατεταγμένο αρχείο με  $r_A = 30.000$  εγγραφές, μέγεθος block  $B = 1024$  bytes, σταθερού μεγέθους εγγραφές μεγέθους  $R_A = 100$  bytes, όπου το πεδίο κλειδιού διάταξης έχει μέγεθος  $V_A = 9$  bytes, μη εκτεινόμενη καταχώρηση. Κατασκευάζουμε πρωτεύον ευρετήριο, μέγεθος δείκτη block  $P = 6$  bytes)

$$bfr_A = 10$$

$$bfr_E = 68$$

Μέγεθος αρχείου δεδομένων: 3.000 blocks

Μέγεθος αρχείου ευρετηρίου: 45 blocks

Αναζήτηση χωρίς ευρετήριο:  $\lceil \log 3.000 \rceil = 12$  blocks

Αναζήτηση με ευρετήριο:  $\lceil \log 45 \rceil + 1 = 7$  blocks

block ευρετηρίου

block αρχείου

Διαδική γιατί το αρχείο ταξινομημένο



### Access paths (μονοπάτια προσπέλασης)

- Το ευρετήριο αρχείου είναι (πάντα) ένα **διατεταγμένο αρχείο** με σταθερού μήκους εγγραφές
- Το αρχείο ευρετηρίου καταλαμβάνει μικρότερο χώρο από το ίδιο το αρχείο δεδομένων (οι καταχωρήσεις είναι μικρότερες και λιγότερες)
- Κάνοντας **δυναμική αναζήτηση** στο ευρετήριο (γιατί το ευρετήριο είναι διατεταγμένο αρχείο) βρίσκουμε τον δείκτη στο block όπου αποθηκεύεται η εγγραφή που θέλουμε



**Ευρετήριο συστάδων (clustering index):** ορισμένο στο πεδίο διάταξης [το οποίο όμως δεν είναι κλειδί]

Υπάρχει μία εγγραφή για κάθε διακεκριμένη τιμή του πεδίου διάταξης (συστάδας) του αρχείου που περιέχει:

- την τιμή αυτή
- ένα δείκτη προς το πρώτο block του αρχείου δεδομένων που περιέχει μια εγγραφή με την τιμή αυτή στο πεδίο συστάδας
- Το ευρετήριο στο πεδίο διάταξης (+ όχι κλειδί) είναι ένα **μη πυκνό** ευρετήριο



▪ Ευρετήριο συστάδων ή συγκροτημένο ευρετήριο

Όταν η διάταξη του ευρετηρίου ακολουθεί αυτή του αρχείου δεδομένων



Παράδειγμα (υπολογισμός μεγέθους ευρετηρίου)

Έστω διατεταγμένο αρχείο με  $r_A = 30.000$  εγγραφές, μέγεθος block  $B = 1024$  bytes, σταθερού μεγέθους εγγραφές μεγέθους  $R_A = 100$  bytes, μη εκτεινόμενη καταχώρηση, όπου το πεδίο διάταξης έχει μέγεθος  $V_A = 9$  bytes και υπάρχουν 1000 διαφορετικές τιμές και οι εγγραφές είναι ομοιόμορφα κατανεμημένες ως προς τις τιμές αυτές. Υποθέτουμε ότι χρησιμοποιούνται άγκυρες block, κάθε νέα τιμή του πεδίου διάταξης αρχίζει στην αρχή ενός νέου block. Κατασκευάζουμε ευρετήριο συστάδων, μέγεθος δείκτη block  $P = 6$  bytes

Μέγεθος αρχείου δεδομένων: 3.000 blocks

Μέγεθος ευρετηρίου συστάδων: 15 blocks

$$bfr_A = 10$$

$$bfr_E = 68$$



- Αναζήτηση (όπως πριν)

Διαδική αναζήτηση στο ευρετήριο

Ανάγνωση blocks (τώρα μπορεί να είναι παραπάνω από ένα) από το αρχείο δεδομένων



### Παράδειγμα (υπολογισμός κόστους αναζήτησης)

(στοιχεία όπως πριν) Έστω διατεταγμένο αρχείο με  $r_A = 30.000$  εγγραφές, μέγεθος block  $B = 1024$  bytes, σταθερού μεγέθους εγγραφές μεγέθους  $R_A = 100$  bytes, μη εκτεινόμενη καταχώρηση, όπου το πεδίο διάταξης έχει μέγεθος  $V_A = 9$  bytes και υπάρχουν 1000 διαφορετικές τιμές και οι εγγραφές είναι ομοιόμορφα κατανεμημένες ως προς τις τιμές αυτές. Υποθέτουμε ότι χρησιμοποιούνται άγκυρες block, κάθε νέα τιμή του πεδίου διάταξης αρχίζει στην αρχή ενός νέου block. Κατασκευάζουμε ευρετήριο συστάδων, μέγεθος δείκτη block  $P = 6$  bytes

*Μέγεθος αρχείου δεδομένων: 3.000 blocks*

*Μέγεθος αρχείου ευρετηρίου: 15 blocks*

Αναζήτηση χωρίς ευρετήριο:  $\lceil \log 3.000 \rceil + \text{ταιριάσματα} (= 3) \approx 15$  blocks

Αναζήτηση με ευρετήριο:  $\lceil \log 15 \rceil + 3 = 7$  blocks





**Δευτερεύον ευρετήριο** (secondary index): ορισμένο σε πεδίο διαφορετικό του πεδίου διάταξης



Περίπτωση 1: Το πεδίο ευρετηριοποίησης είναι **κλειδί** (καλείται και *δευτερεύον κλειδί*)

Υπάρχει μία εγγραφή για κάθε εγγραφή του αρχείου που περιέχει:

- την τιμή του κλειδιού για αυτήν την εγγραφή
- ένα δείκτη προς το block (ή την εγγραφή) του αρχείου δεδομένων που περιέχει την εγγραφή με την τιμή αυτή
- Το ευρετήριο σε πεδίο ΟΧΙ διάταξης (+ κλειδί) είναι ένα **πυκνό** ευρετήριο



Παράδειγμα (υπολογισμός μεγέθους ευρετηρίου)

Έστω αρχείο με  $r_A = 30.000$  εγγραφές, μέγεθος block  $B = 1024$  bytes, σταθερού μεγέθους εγγραφές μεγέθους  $R_A = 100$  bytes, μη εκτεινόμενη καταχώρηση, όπου το πεδίο κλειδιού έχει μέγεθος  $V_A = 9$  bytes αλλά δεν είναι πεδίο διάταξης. Κατασκευάζουμε δευτερεύον ευρετήριο, μέγεθος δείκτη block  $P = 6$  bytes

Μέγεθος αρχείου δεδομένων: 3.000 blocks

Μέγεθος αρχείου ευρετηρίου: 442 blocks

45 για πρωτεύον



Παράδειγμα (υπολογισμός κόστους αναζήτησης)

Στοιχεία όπως πριν

(Έστω αρχείο με  $r_A = 30.000$  εγγραφές, μέγεθος block  $B = 1024$  bytes, σταθερού μεγέθους εγγραφές μεγέθους  $R_A = 100$  bytes, μη εκτεινόμενη καταχώρηση, όπου το πεδίο κλειδιού έχει μέγεθος  $V_A = 9$  bytes αλλά δεν είναι πεδίο διάταξης. Κατασκευάζουμε δευτερεύον ευρετήριο, μέγεθος δείκτη block  $P = 6$  bytes)

Μέγεθος αρχείου δεδομένων: 3.000 blocks

Μέγεθος αρχείου ευρετηρίου: 442 blocks

$bfr_A = 10$

$bfr_E = 68$

Αναζήτηση χωρίς ευρετήριο (σειριακή αναζήτηση, γιατί το αρχείο δεδομένων δεν είναι ταξινομημένο):  $3.000/2 = 1500$  blocks

Αναζήτηση με ευρετήριο:  $\lceil \log 442 \rceil + 1 = 10$  blocks

Για πρωτεύον ήταν 45 και 7 blocks αντίστοιχα



## Περίπτωση 2: Το πεδίο ευρετηριοποίησης **δεν είναι κλειδί**

1. Πυκνό ευρετήριο: μία καταχώρηση για κάθε εγγραφή
2. Μεταβλητού μήκους εγγραφές με ένα επαναλαμβανόμενο πεδίο για το δείκτη
3. *Μία εγγραφή ευρετηρίου για κάθε τιμή του πεδίου ευρετηριοποίησης + ένα ενδιάμεσο επίπεδο για την διαχείριση των πολλαπλών δεικτών*



### • Αναζήτηση

Διαδική αναζήτηση στο δευτερεύον ευρετήριο  
Ανάγνωση του block (ή των blocks) από το ενδιάμεσο επίπεδο  
Ανάγνωση των blocks (συνήθως τόσα όσες οι εγγραφές που ταιριάζουν) από το αρχείο δεδομένων

### • Εισαγωγή

Απλή αν δεν αφορά εισαγωγή νέας τιμής στο ευρετήριο



Παράδειγμα (υπολογισμός μεγέθους ευρετηρίου)

Έστω μη διατεταγμένο αρχείο (αρχείο σωρού) με  $r_A = 30.000$  εγγραφές, μέγεθος block  $B = 1024$  bytes, σταθερού μεγέθους εγγραφές μεγέθους  $R_A = 100$  bytes, μη εκτεινόμενη καταχώρηση, όπου το πεδίο ευρετηριοποίησης (δηλαδή, το πεδίο στο οποίο θα κατασκευάσουμε το ευρετήριο) έχει μέγεθος  $V_A = 9$  bytes. Υπάρχουν 100 διαφορετικές τιμές και οι εγγραφές είναι ομοιόμορφα κατανεμημένες ως προς τις τιμές αυτές. Κατασκευάζουμε ευρετήριο συστάδων χρησιμοποιώντας την επιλογή (3), μέγεθος δείκτη block  $P = 6$  bytes

Ευρετήριο  $bfr_E = 68$   $b_E = 2$

κόστος αναζήτησης:

Ενδιάμεσο επίπεδο -- Ποια είναι η οργάνωση του;

$bfr_{EE} = 170$   $b_{EE} = 177$  blocks



- Επιπρόσθετες δομές για την πιο αποδοτική εκτέλεση ερωτήσεων/αναζητήσεων - προκαλούν όμως επιβάρυνση στις τροποποιήσεις

- Εύκολη η λογική διάταξη των εγγραφών με βάση το πεδίο ευρετηριοποίησης

- Ανακτήσεις με σύνθετες συνθήκες, μπορεί να γίνουν χρησιμοποιώντας τα blocks του ευρετηρίου



Ιδέα:

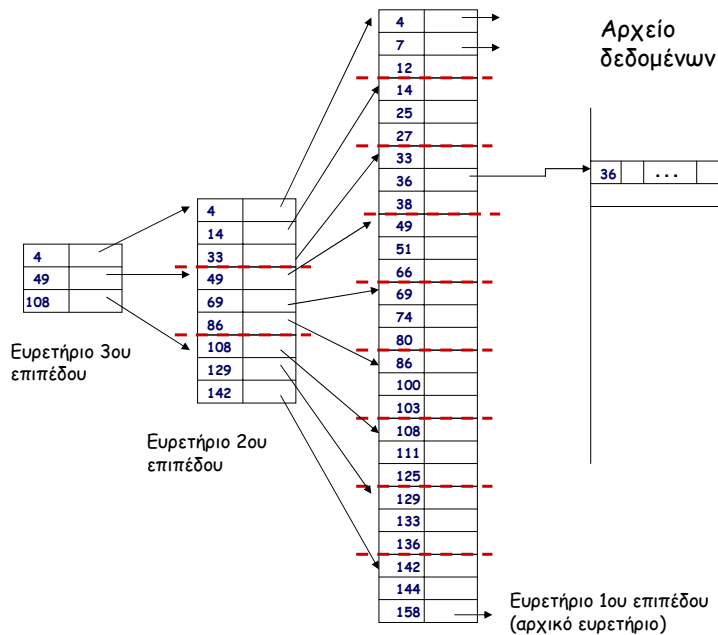
Τα ευρετήρια είναι αρχεία - χτίζουμε ευρετήρια πάνω στα αρχεία ευρετηρίου

Το αρχείο είναι **διατεταγμένο** και το πεδίο διάταξης είναι και κλειδί (άρα πρωτεύον ευρετήριο!)

*Υπενθύμιση (παράγοντας ομαδοποίησης: αριθμός εγγραφών ανά block)*

Παράγοντας ομαδοποίησης (blocking factor), όταν  $B \geq R$

$bfr = \lfloor B / R \rfloor$ , όπου B μέγεθος block σε byte και R μέγεθος εγγραφής σε bytes



## Ευρετήριο Πολλών Επιπέδων



- Έστω ότι το αρχείο ευρετηρίου είναι το **πρώτο ή βασικό επίπεδο**  
Έστω ότι ο παράγοντας ομαδοποίησης είναι  $f_0$  και ότι έχει  $r_1$  blocks  
Το αρχείο είναι διατεταγμένο και το πεδίο διάταξης είναι και κλειδί
- Δημιουργούμε ένα πρωτεύον ευρετήριο για το ευρετήριο πρώτου επιπέδου - **δεύτερο** επίπεδο

Παράγοντας ομαδοποίησης:  $f_0$  Αριθμός block  $\lceil r_1/f_0 \rceil$

- Δημιουργούμε ένα πρωτεύον ευρετήριο για το ευρετήριο δεύτερου επιπέδου - **τρίτο** επίπεδο

Παράγοντας ομαδοποίησης:  $f_0$  Αριθμός block  $\lceil r_1/(f_0)^2 \rceil$

## Ευρετήριο Πολλών Επιπέδων

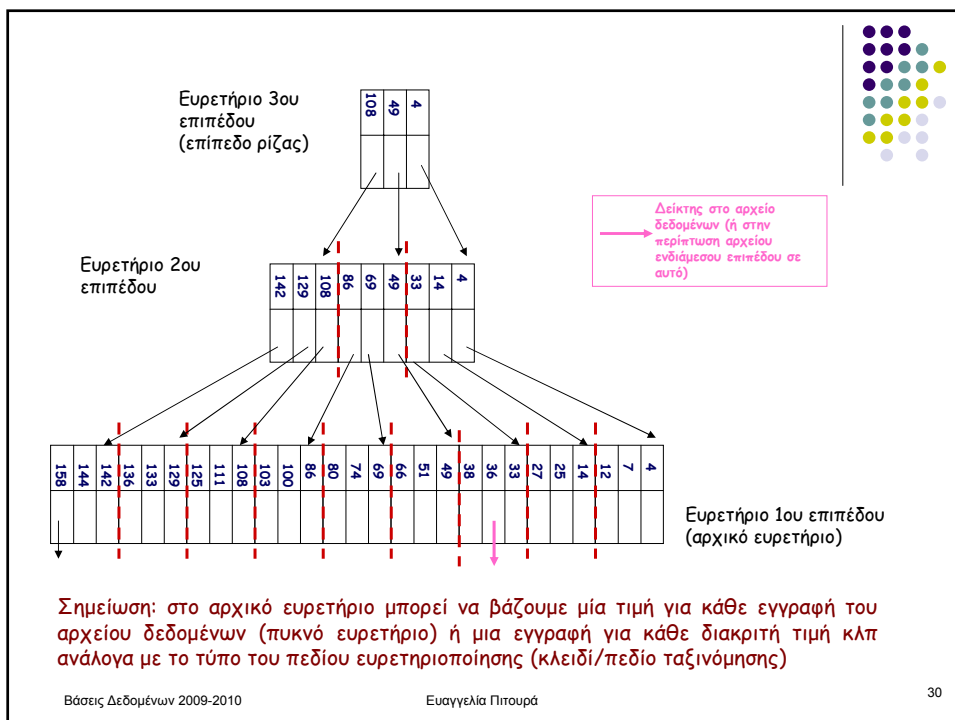
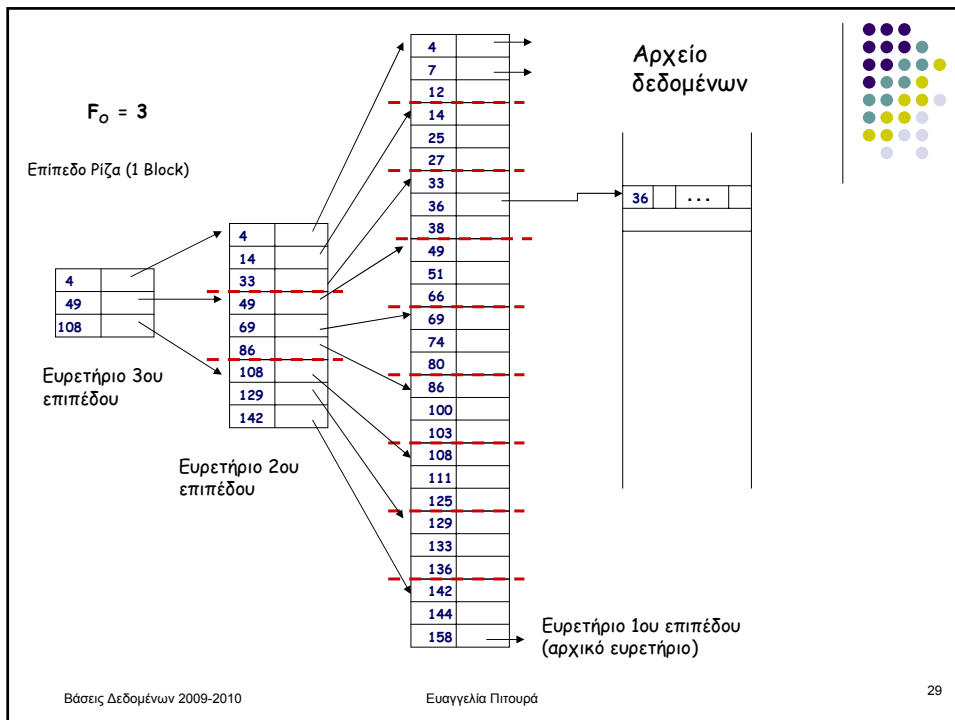


- Μέχρι πόσα επίπεδα:

Μέχρι όλες οι εγγραφές του ευρετηρίου να χωρούν σε ένα block

Έστω  $t$  κορυφαίο επίπεδο (top level)  $\lceil r_1/(f_0)^t \rceil = 1$

- Το  $f_0$  ονομάζεται και *παράγοντας διακλάδωσης* του ευρετηρίου





• Αναζήτηση

$p$  := διεύθυνση του block του κορυφαίου επιπέδου του ευρετηρίου

$t$  := αριθμός επιπέδων του ευρετηρίου

for  $j = t$  to 1 step -1 do

    read block με διεύθυνση  $p$  του ευρετηρίου στο επίπεδο  $j$

    αναζήτηση στο block  $p$  της εγγραφής  $i$  με τιμή  $K_j(i) \leq K < K_j(i+1)$

read το block του αρχείου δεδομένων με διεύθυνση  $p$

Αναζήτηση στο block  $p$  της εγγραφής  $i$  με τιμή  $K_j(i) \leq K < K_j(i+1)$

$F_0 = 3$

4	
49	
108	

Ευρετήριο 3ου επιπέδου (επίπεδο ρίζας)

36

4	
14	
33	
49	
69	
86	
108	
129	
142	

Ευρετήριο 2ου επιπέδου

4
7
12
14
25
27
33
36
38
49
51
66
69
74
80
86
100
103
108
111
125
129
133
136
142
144
158

Ευρετήριο 1ου επιπέδου (αρχικό ευρετήριο)

Αρχείο δεδομένων

36	...
----	-----







• Εισαγωγή/διαγραφή

τροποποιήσεις πολλαπλών ευρετηρίων

*Δυναμικό* πολυεπίπεδο ευρετήριο: B-δέντρα και B+-δέντρα



**Παράδειγμα (υπολογισμός μεγέθους ευρετηρίου)**

Έστω αρχείο με  $r_A = 30.000$  εγγραφές, μέγεθος block  $B = 1024$  bytes, σταθερού μεγέθους εγγραφές μεγέθους  $R_A = 100$  bytes, μη εκτεινόμενη καταχώρηση, όπου το πεδίο κλειδιού έχει μέγεθος  $V_A = 9$  bytes αλλά δεν είναι πεδίο διάταξης. Κατασκευάζουμε δευτερεύον ευρετήριο, μέγεθος δείκτη block  $P = 6$  bytes

$$f_0 = \lfloor (1024 / (9 + 6)) \rfloor = 68$$

Μέγεθος αρχείου δεδομένων: 3.000 blocks

Μέγεθος αρχείου ευρετηρίου *πρώτου* επιπέδου: 442 blocks

Μέγεθος αρχείου ευρετηρίου *δεύτερου* επιπέδου:  $\lceil (442 / 68) \rceil = 7$  blocks

Μέγεθος αρχείου ευρετηρίου *τρίτου* επιπέδου:  $\lceil (7 / 68) \rceil = 1$  block

Άρα  $t = 3$



### Παράδειγμα (υπολογισμός κόστους αναζήτησης)

Έστω αρχείο με  $r_A = 30.000$  εγγραφές, μέγεθος block  $B = 1024$  bytes, σταθερού μεγέθους εγγραφές μεγέθους  $R_A = 100$  bytes, μη εκτεινόμενη καταχώρηση, όπου το πεδίο κλειδιού έχει μέγεθος  $V_A = 9$  bytes αλλά δεν είναι πεδίο διάταξης,. Κατασκευάζουμε δευτερεύον ευρετήριο, μέγεθος δείκτη block  $P = 6$  bytes

Άρα  $t = 3$

Παράδειγμα

$t + 1 = 4$  προσπελάσεις

Για το δευτερεύον ήταν 10 και χωρίς ευρετήριο 1500



- Τα αρχεία ευρετηρίων είναι απλά αρχεία, άρα και σε αυτά μπορούν να οριστούν ευρετήρια
- Καταλήγουμε λοιπόν σε μια ιεραρχία δομών ευρετηρίων (πρώτο επίπεδο, δεύτερο επίπεδο, κλπ.)
- Κάθε επίπεδο του ευρετηρίου είναι ένα *διατεταγμένο* αρχείο, συνεπώς, εισαγωγές/διαγραφές εγγραφών απαιτούν επιπλέον δουλειά
- Ένα πολύ-επίπεδο ευρετήριο αποτελεί ένα *Δέντρο Αναζήτησης*  
Όπου κάθε κόμβος (block) έχει  $f_0$  δείκτες και  $f_0$  τιμές κλειδιού