

Κατακερματισμός (Hashing)

Συνάρτηση Κατακερματισμού

- Μια συνάρτηση κατακερματισμού μπορεί να χρησιμοποιηθεί για να αντιστοιχίσει μια τιμή ενός πεδίου σε ένα μπλόκ αρχείου εγγραφών.
- Παράδειγμα. Για ένα πεδίο σε ένα αρχείο εγγραφών που μπορεί να παίρνει ακέραιες τιμές η συνάρτηση κατακερματισμού:
- $h(K) = K \bmod 8$ παράγει τους αριθμούς 0..7 οι οποίοι μπορεί να θεωρηθούν ότι είναι οι αριθμοί ή διευθύνσεις των μπλοκ που περιέχουν τις εγγραφές του πεδίου. Τα μπλόκ αναφέρονται και σαν κάδοι (buckets) κατακερματισμού.

Παράδειγμα κατακερματισμού με $h(K) = K \bmod 8$

5, 8, 10, 12, 24, 32, 37, 39, 41, 46, 51, 55, 60, 63, 66, 70

0	8
	24
	32

1	41

2	10
	66

3	51

4	12
	60

5	5
	13
	37

6	46
	70

7	55
	63

Πρόβλημα Συγκρούσεων (Collision)

- Τι κάνουμε όταν ο κάδος που δείχνει μια τιμή κατακερματισμού γεμίσει;
- Μπορούμε να χρησιμοποιήσουμε περιοχές υπερχείλισης οι οποίες μπορεί να περιέχουν τις εγγραφές και στις οποίες δείχνουν οι πρωτεύοντες κάδοι με κάποιο τρόπο.

0	8	8
	24	
	32	

8	40	

- Μπορεί να δημιουργηθεί μια αλυσίδα περιοχών υπερχείλισης αν οι εγγραφές που δημιουργούν συγκρούσεις αυξηθούν.

Πρόβλημα Συγκρούσεων (συν)

- Αν οι εγγραφές υπερχείλισης αυξηθούν τότε η αναζήτηση εγγραφών μέσω του κλειδιού μειώνεται σε απόδοση.
- Δεν μπορούμε να θεωρήσουμε ότι βρίσκουμε άμεσα μια εγγραφή βάση της τιμής κατακερματισμού (hash value). Θα πρέπει να προσπελάσουμε (κατά μέσο όρο περισσότερα από μια σελίδες για να εντοπίσουμε την εγγραφή).

Λύσεις στο Πρόβλημα συγκρούσεων / αύξησης του αρχείου

- Θα μπορούσαμε να αναδιοργανώσουμε το αρχείο χρησιμοποιώντας μια άλλη συνάρτηση κατακερματισμού.
- $H \bmod 8$ κατανέμει τις εγγραφές σε 8 κάδους
- $H \bmod 16$ κατανέμει τις εγγραφές σε 16 κάδους
- $H \bmod 128$ κατανέμει τις εγγραφές σε 128 κάδους.
- Κοκ
- Γενικά απλή αναδιοργάνωση μπορεί να είναι ακριβή διαδικασία γιατί όλες οι εγγραφές των κάδων πρέπει γενικά να διαβαστούν και να γραφούν σε νέους κάδους
- Υπάρχουν μέθοδοι που είναι πιο αποτελεσματικοί.

Επεκτατός Κατακερματισμός (Extendible Hashing)

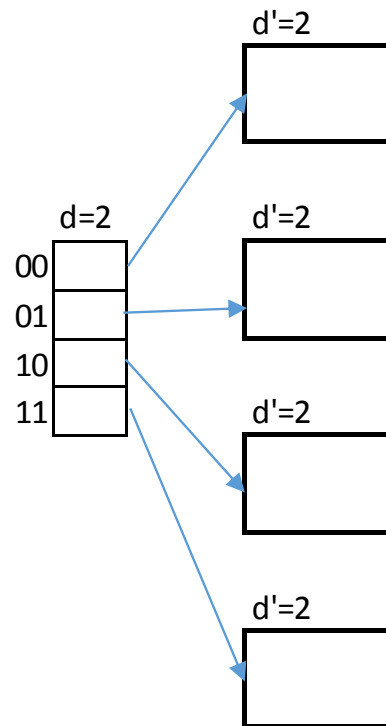
- Λύνει σταδιακά το πρόβλημα των συγκρούσεων δημιουργώντας νέους κάδους όταν είναι απαραίτητο και κάνοντας «τοπική» αναδιοργάνωση των εγγραφών.
- Χρησιμοποιεί ορισμένα bit της δυαδικής αναπαράστασης της τιμής κατακερματισμού για προσδιορισμό του κάδου. Α αριθμός των bit μπορεί να αυξάνεται σταδιακά.
- Χρησιμοποιείται ένας κατάλογο με τις διευθύνσεις όλων των υπαρχόντων κάδων κατακερματισμού. Ο αριθμός των θέσεων του καταλόγου προσδιορίζεται από τον αριθμό των bit που χρησιμοποιούνται για τον προσδιορισμό τους.

Παράδειγμα επεκτατού κατακερματισμού

- Θέλουμε να χρησιμοποιήσουμε την συνάρτηση κατακερματισμού **$K \bmod 128$** που παράγει τιμές από 0..127
- Αρχικά αναμένουμε ότι δεν θα έχουμε πολλές εγγραφές για να δικαιολογείται η δημιουργία 128 κάδων.
- Μπορούμε να χρησιμοποιήσουμε ένα αριθμό από τα bit των τιμών κατακερματισμού για τον προσδιορισμό των κάδων (**ολικό βάθος**).
- Θεωρούμε επίσης ότι οι κάδοι μπορούν να χωρέσουν μέχρι **δύο εγγραφές**.

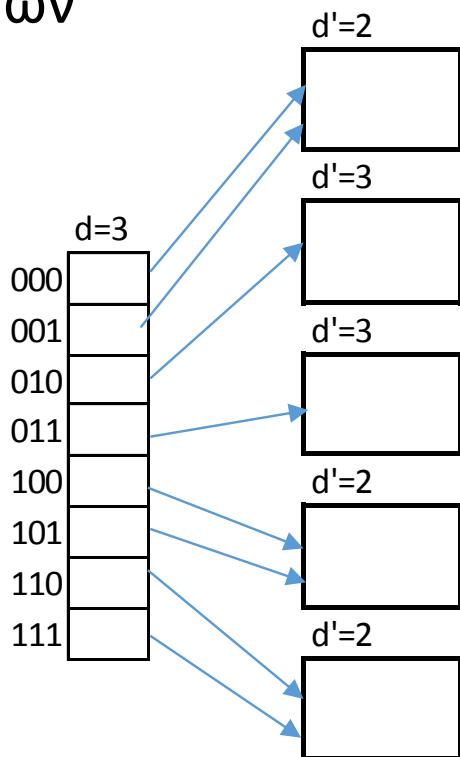
Παράδειγμα επεκτατού κατακερματισμού

Χρησιμοποιώντας ολικό βάθος 2, δηλαδή χρησιμοποιώντας 2 bit (τα πιο σημαντικά) μπορούμε να προσδιορίσουμε 4 κάδους (00, 01, 10, 11)



Παράδειγμα επεκτατού κατακερματισμού

Αν γεμίσει κάποιος κάδος (πχ ο δεύτερος) μπορούμε να επεκτείνουμε τον κατάλογο των κάδων διπλασιάζοντας τον κατάλογο και χρησιμοποιώντας 3 bit για τον προσδιορισμό των κάδων



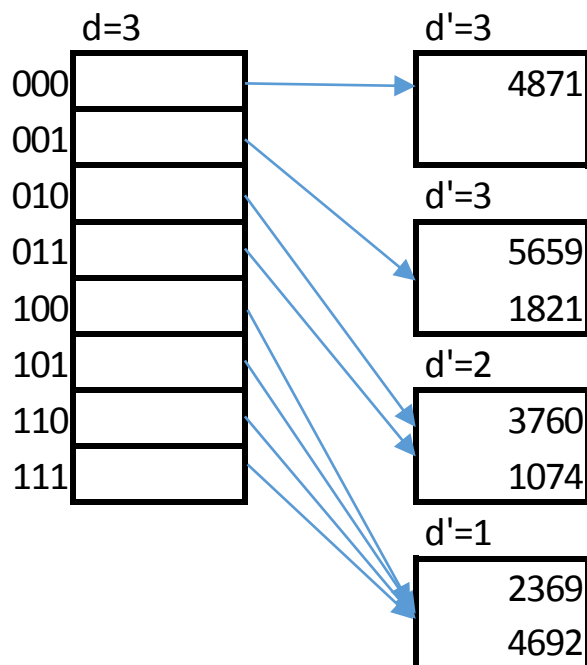
Προσθέτουμε όμως ένα νέο κάδο με εγγραφές.

Οι θέσεις που αντιστοιχούσαν στην τιμή 01 (με βάθος 2) αντιστοιχούν στις θέσεις 010 και 011 και δείχνουν σε δύο διαφορετικούς κάδους.

Οι άλλες θέσεις του καταλόγου δείχνουν στους κάδους που προσδιορίζονταν με τα δύο bit. Στους κάδους αυτούς το τοπικό βάθος παραμένει 2.

Παράδειγμα επεκτατού κατακερματισμού

Έστω ότι έχουμε εισάγει τις ακόλουθες εγγραφές με κωδικούς (part#) σε αρχείο επεκτατού κατακερματισμού

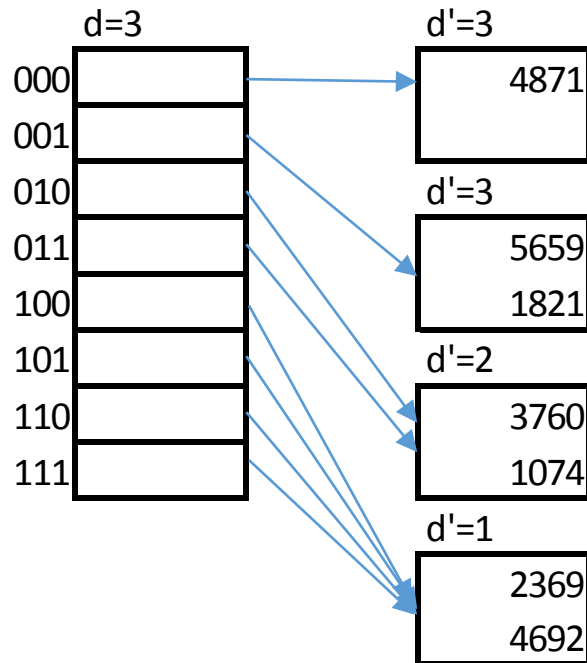


Part #	MOD 128	BINARY
2369	65	1000001
3760	48	0110000
4692	84	1010100
4871	7	0000111
5659	27	0011011
1821	29	0011101
1074	50	0110010
7115	75	1001011

Τι συμβαίνει αν αναζητηθούν οι εγγραφές που αναφέρονται στους πιο πάνω κωδικούς προϊόντων (part #) με χρήση $h(k) = k \bmod 8$;

Τι συμβαίνει αν προσπαθήσουμε να προσθέσουμε την εγγραφή 7115;

Παράδειγμα επεκτατού κατακερματισμού

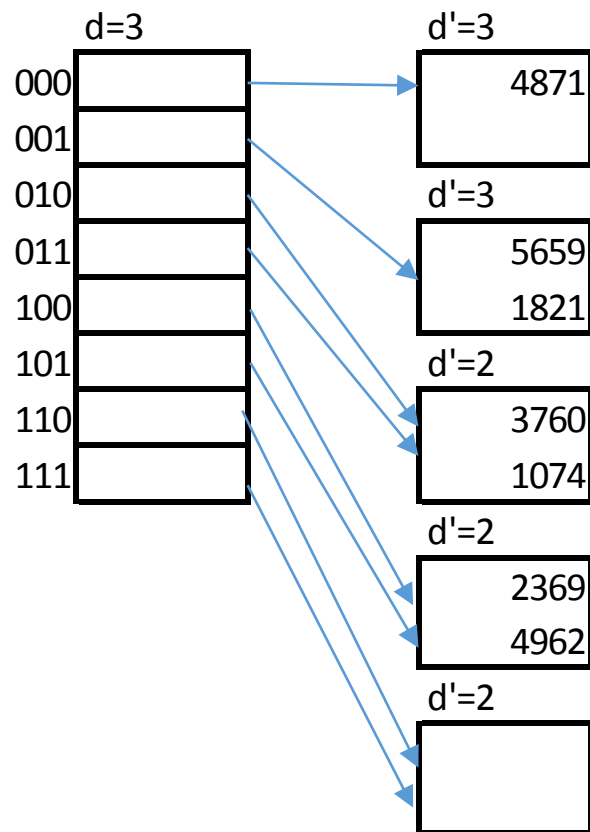


Η τιμή κατακερματισμού του κλειδιού 7115 είναι 75 με δυαδική αναπαράσταση : **1001011**

Θα πρέπει να τοποθετηθεί στο κάδο που δείχνει η θέση 100. Ο κάδος είναι γεμάτος. Επειδή έχει τοπικό βάθος $d'=1$, μπορούμε να τον αντικαταστήσουμε με δύο κάδους με τοπικό βάθος $d'=2$ και να ενημερώσουμε τους δείκτες στον κατάλογο των κάδων.

Στην συνέχεια αναδιατάσσουμε τις εγγραφές του κάδου για να τοποθετηθούν στον σωστό κάδο αν χρειαστεί.

Παράδειγμα επεκτατού κατακερματισμού

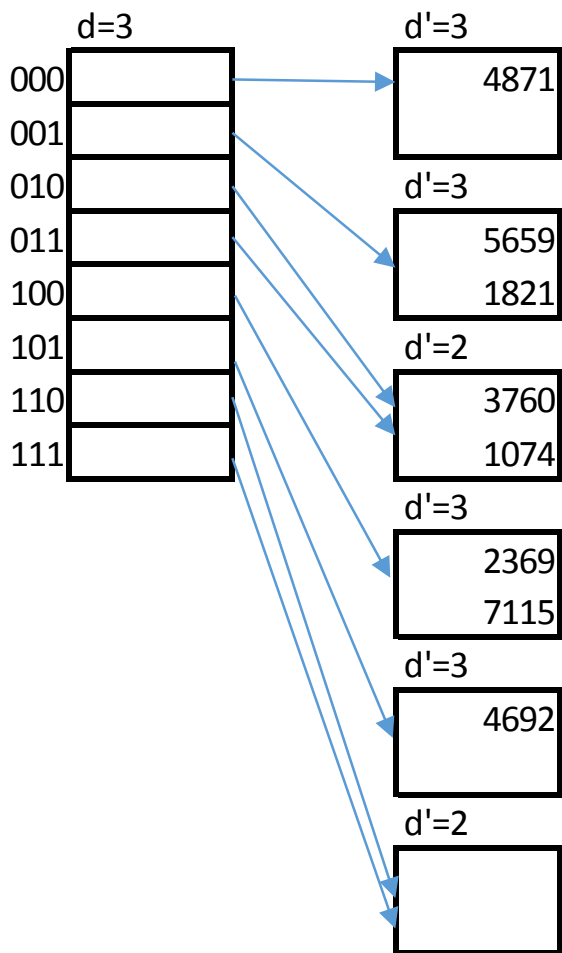


Part #	MOD 128	BINARY
2369	65	1000001
3760	48	0110000
4692	84	1010100
4871	7	0000111
5659	27	0011011
1821	29	0011101
1074	50	0110010
7115	75	1001011

Μετά την αλλαγή παρατηρούμε ότι οι εγγραφές 2369 και 4962 παραμένουν στον ίδιο κάδο με τοπικό βάθος 2 λόγω του ότι οι τιμές κατακερματισμού ξεκινούν με 10. Είναι ο κάδος που πρέπει να τοποθετηθεί και η 7115.

Μπορούμε να κάνουμε άλλη μια διάσπαση του κάδου αυξάνοντας το τοπικό βάθος σε 3 και να αναδιοργανώσουμε τις εγγραφές αν χρειαστεί.

Παράδειγμα επεκτατού κατακερματισμού



Part #	MOD 128	BINARY
2369	65	1000001
3760	48	0110000
4692	84	1010100
4871	7	0000111
5659	27	0011011
1821	29	0011101
1074	50	0110010
7115	75	1001011

Μετά την αναδιοργάνωση δημιουργήθηκε χώρος στον κάδο 100 όπου πρέπει να τοποθετηθεί η εγγραφή 7115

Θέματα με τον επεκτατό κατακερματισμό

- Χρησιμοποιεί χώρο για την δομή καταλόγου
- Όταν δημιουργείται επέκταση στον κατάλογο πρέπει να διπλασιαστεί ο χώρος του (χωρίς να απαιτείται οι νέες θέσεις να προσδιορίζουν διαφορετικούς κάδους).

Γραμμικός Κατακερματισμός (Linear Hashing)

- Θεωρείται ένας βέλτιστος τρόπος κατακερματισμού.
- Αποφεύγει την διατήρηση ενός καταλόγου κάρδων.
- Χρησιμοποιεί διαφορετικές συναρτήσεις κατακερματισμού για να επεκτείνεται ο χώρος των κάρδων που μπορεί να προσδιοριστούν.
- Κάθε φορά μπορεί να χρησιμοποιηθούν δύο διαφορετικές συναρτήσεις κατακερματισμού (h_i και h_{i+1}).
- Για παράδειγμα αν $h_i = K \bmod M$, $h_{i+j} = K \bmod (2^j M)$

Παράδειγμα Γραμμικού Κατακερματισμού

- Ας θεωρήσουμε ότι
 - χρησιμοποιούμε την συνάρτηση κατακερματισμού $h1 = K \bmod 2^1 = K \bmod 2$
 - Έχουμε κάδους 2 εγγραφών που περιέχουν από 2 εγγραφές ο καθένας (το μέγιστο)

	b[0]	4692	3760	
	b[1]	2369	4871	

Έρχεται μια νέα εγγραφή για καταχώρηση με κλειδί 5659.
Θα έπρεπε να μπει στον κάδο 1, αλλά είναι γεμάτος.
(Παρατηρήστε ότι είναι γεμάτος και ο 0 και θα είχαμε
παρόμοιο πρόβλημα αν προσπαθούσαμε να τοποθετήσουμε
και εκεί μια εγγραφή).

Παράδειγμα Γραμμικού Κατακερματισμού

- Η εγγραφή θα τοποθετηθεί σε ένα κάδο υπερχείλισης αλλά θα πυροδοτήσει μια διαδικασία διάσπασης κάδων.
- Για να αποφασισθεί ποιος κάδος θα διασπασθεί διατηρείται ένα μετρητής η με τον επόμενο κάδο που θα πρέπει να διασπασθεί. Αρχικά αυτός ο μετρητής είναι 0.
- Όταν ένας κάδος διασπασθεί θα πρέπει να χρησιμοποιηθεί μια νέα συνάρτηση κατακερματισμού για προσδιορισμό των εγγραφών αυτού του κάδου.
- Παρατηρήστε ότι ο κάδος που θα διασπασθεί δεν είναι πάντα αυτός που δημιούργησε κάποιο γεγονός υπερχείλισης. Η διάσπαση γίνεται με την σειρά.

Παράδειγμα Γραμμικού Κατακερματισμού

- Έτσι μετά την τελευταία εισαγωγή και την διάσπαση του κάδου 0 έχουμε την εξής κατάσταση

n=1	Ο επόμενος κάδος που μπορεί να διασπασθεί						
$h1 = K \bmod 2^1 = K \bmod 2$				για $h1(K) \geq n$			
$h2 = K \bmod 2^2 = K \bmod 4$				για $h1(K) < n$			
b[0]	4692	3760					
b[1]	2369	4871		5659			
b[2]							

	Part #	MOD 1	MOD 2	MOD 4
1	2369	0	1	1
2	3760	0	0	0
3	4692	0	0	0
4	4871	0	1	3
5	5659	0	1	3

Τι θα συμβεί αν προσπαθήσουμε να προσθέσουμε και την τιμή 1821

Παράδειγμα Γραμμικού Κατακερματισμού

- Το 1821 δεν χωράει στον πρωτεύοντα κάδο. Μπορούμε να το θεωρήσουμε γεγονός υπερχείλισης. Αυτό θα πυροδοτήσει μια νέα διάσπαση του κάδου 1 και αναδιάταξη των εγγραφών του. Σημείωση : Το τι ακριβώς θα πυροδοτήσει διάσπαση δεν είναι απόλυτο. Θα μπορούσαμε να θεωρήσουμε ότι θα πυροδοτηθεί διάσπαση όταν γεμίσει ο κάδος υπερχείλισης.

Μετά την διάσπαση αυτή όλοι οι κάδοι μπορούν να προσδιορισθούν από την συνάρτηση κατακερματισμού h_2

Η διαδικασία της διάσπασης αν χρειαστεί μπορεί να ξεκινήσει πάλι από τον κάδο 0 (το n γίνεται πάλι 0)

n=0			
$h_2 = K \bmod 2^2 = K \bmod 4$			
b[0]	4692	3760	
b[1]	2369	1821	
b[2]			
b[3]	4871	5659	