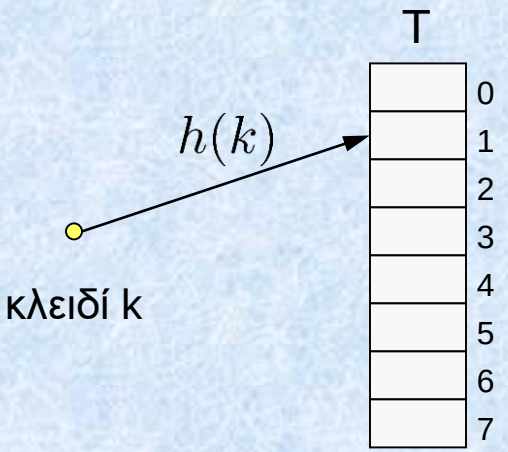
Πίνακες Διασποράς

Χρησιμοποιούμε ένα πίνακα διασποράς T και μια συνάρτηση διασποράς h

Ένα στοιχείο με κλειδί k αποθηκεύεται στη θέση $T[h(k)]$



U : χώρος πιθανών κλειδιών

T : πίνακας μεγέθους m

συνάρτηση διασποράς

$h : U \rightarrow \{0, 1, \ldots, m-1\}$

Καθολική Διασπορά

Για κάθε καθορισμένη συνάρτηση διασποράς μπορούμε να επιλέξουμε κλειδιά που θα δίνουν τη χειρότερη δυνατή επίδοση.

Προκειμένου να βελτιώσουμε την κατάσταση μπορούμε να βασιστούμε σε **τυχαιοκρατικούς αλγόριθμους**:

Ορίζουμε μία οικογένεια συναρτήσεων και κατά τη διάρκεια της εκτέλεσης επιλέγουμε τυχαία μία από αυτές τις συναρτήσεις ως συνάρτηση διασποράς.

⇨ Περιορίζουμε την πιθανότητα εμφάνισης παθολογικών περιπτώσεων.

Καθολική Διασπορά

$\mathcal{H}$ πεπερασμένη συλλογή συναρτήσεων

Η συλλογή $\mathcal{H}$ είναι καθολική αν για κάθε ζεύγος διαφορετικών κλειδιών $k, \ell \in U$ υπάρχουν το πολύ $|\mathcal{H}|/m$ συναρτήσεις $h \in \mathcal{H}$ με $h(k) = h(\ell)$

Καθολική Διασπορά

$\mathcal{H}$ πεπερασμένη συλλογή συναρτήσεων

Η συλλογή $\mathcal{H}$ είναι καθολική αν για κάθε ζεύγος διαφορετικών κλειδιών $k, \ell \in U$ υπάρχουν το πολύ $|\mathcal{H}|/m$ συναρτήσεις $h \in \mathcal{H}$ με $h(k) = h(\ell)$

Συνεπώς $\Pr_{h \in \mathcal{H}}[h(k) = h(\ell)] \leq 1/m$

Καθολική Διασπορά

$\mathcal{H}$ πεπερασμένη συλλογή συναρτήσεων

Η συλλογή $\mathcal{H}$ είναι καθολική αν για κάθε ζεύγος διαφορετικών κλειδιών $k, \ell \in U$ υπάρχουν το πολύ $|\mathcal{H}|/m$ συναρτήσεις $h \in \mathcal{H}$ με $h(k) = h(\ell)$

Συνεπώς $\Pr_{h \in \mathcal{H}}[h(k) = h(\ell)] \leq 1/m$

Πιθανότητα σύμπτωσης όταν $h(k)$ και $h(\ell)$ επιλέγονται **τυχαία** και **ανεξάρτητα**.

Καθολική Διασπορά

$\mathcal{H} =$ καθολική συλλογή συναρτήσεων διασποράς που αντιστοιχίζει το σύμπαν U στο σύνολο $\{0, 1, \ldots, m-1\}$

$K =$ αυθαίρετο υποσύνολο του U με $|K| = n \leq m$

$u =$ αυθαίρετο στοιχείο του U

$h =$ τυχαία επιλεγμένη συνάρτηση της $\mathcal{H}$

$X =$ αριθμός στοιχειών $k \in K$ με $h(k) = h(u)$ (τυχαία μεταβλητή)

Για κάθε $k \in K$ έχουμε δείκτρια τυχαία μεταβλητή $X_k = \begin{cases} 0, h(k) \neq h(u) \\ 1, h(k) = h(u) \end{cases}$

Έχουμε $X = \sum_{k \in K} X_k$

Άρα $\mathrm{E}[X] = \sum_{k \in K} \mathrm{E}[X_k] = \sum_{k \in K} \Pr[X_k = 1] \leq \sum_{k \in K} \frac{1}{m} = \frac{n}{m} \leq 1$

Καθολική Διασπορά

Θεώρημα Έστω ότι επιλέγουμε τυχαία $h \in \mathcal{H}$ και κάνουμε εισαγωγή n κλειδιών σε πίνακα T μεγέθους m. Έστω $i = h(k)$. Τότε

$$\mathrm{E}[n_i] \leq \begin{cases} \alpha, & k \notin T \\ 1+\alpha, & k \in T \end{cases}$$

όπου $\alpha = n/m$ ο συντελεστής πληρότητας.

Πόρισμα Ο αναμενόμενος χρόνος εκτέλεσης μίας ακολουθίας N πράξεων με $O(m)$ πράξεις εισαγωγής είναι $O(N)$

Συνεπάγεται από το παραπάνω θεώρημα και την παρατήρηση ότι

$$n = O(m) \Rightarrow \alpha = O(1)$$

Καθολικές Οικογένειες Συναρτήσεων Διασποράς

p πρώτος αριθμός, μεγαλύτερος από κάθε κλειδί $(p > m)$

$$\mathbf{Z}_p = \{0, 1, \ldots, p-1\}$$

$$\mathbf{Z}_p^* = \{1, 2, \ldots, p-1\}$$

Για $a \in \mathbf{Z}_p^*$ και $b \in \mathbf{Z}_p$

$$h_{a,b}(k) = [(ak + b) \bmod p] \bmod m$$

Παράδειγμα $p = 17, m = 6$

$$h_{3,4}(8) = [(3 \cdot 8 + 4) \bmod 17] \bmod 6 = (28 \bmod 17) \bmod 6 = 5$$

Καθολικές Οικογένειες Συναρτήσεων Διασποράς

p πρώτος αριθμός, μεγαλύτερος από κάθε κλειδί $(p > m)$

$$\mathbf{Z}_p = \{0, 1, \ldots, p-1\}$$

$$\mathbf{Z}_p^* = \{1, 2, \ldots, p-1\}$$

Για $a \in \mathbf{Z}_p^*$ και $b \in \mathbf{Z}_p$

$$h_{a,b}(k) = [(ak + b) \bmod p] \bmod m$$

$$\mathcal{H}_{p,m} = \{h_{a,b} \ : \ a \in \mathbf{Z}_p^*, b \in \mathbf{Z}_p\}$$

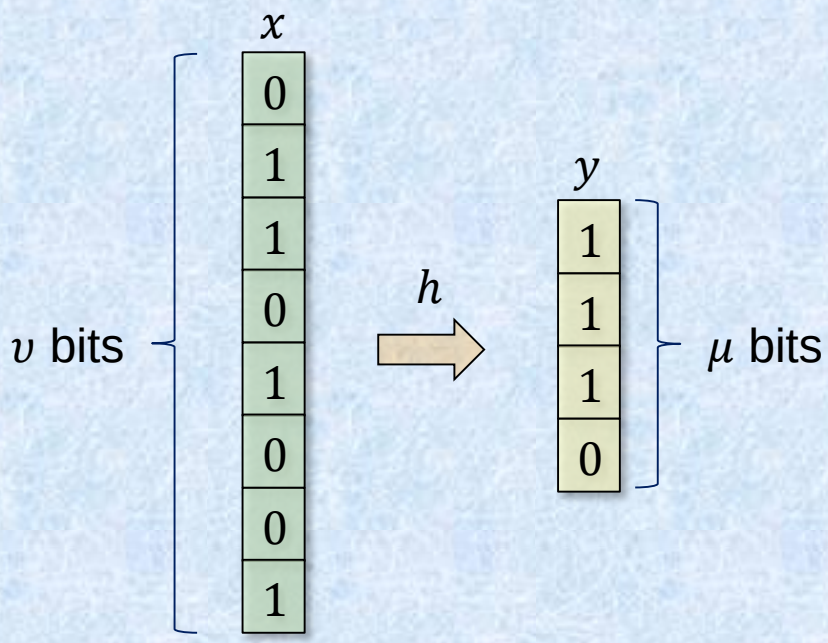
Περιλαμβάνει $p(p-1)$ συναρτήσεις

Καθολικές Οικογένειες Συναρτήσεων Διασποράς

Θα δώσουμε ένα ακόμα παράδειγμα καθολικής οικογένειας συναρτήσεων διασποράς $\mathcal{H}$ υποθέτοντας ότι :

- Το σύνολο των πιθανών κλειδιών περιλαμβάνει $|U| = 2^{\upsilon}$ τιμές.
- Οι συναρτήσεις $h \in \mathcal{H}$ αντιστοιχούν το U σε ένα σύνολο με $m = 2^{\mu}$ τιμές.

Άρα ουσιαστικά οι συναρτήσεις της $\mathcal{H}$ αντιστοιχούν διανύσματα x (ακολουθίες) των υ bits σε διανύσματα y των μ bits.

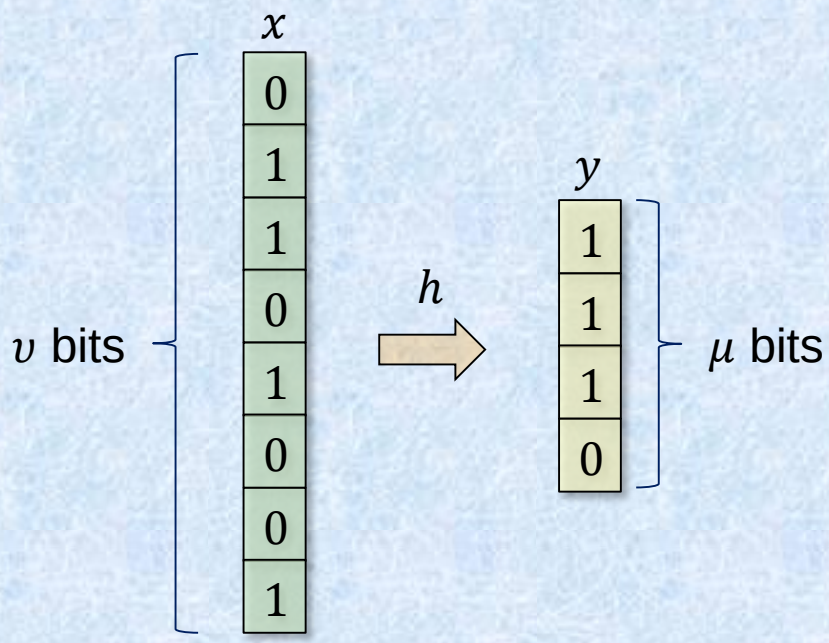


Καθολικές Οικογένειες Συναρτήσεων Διασποράς

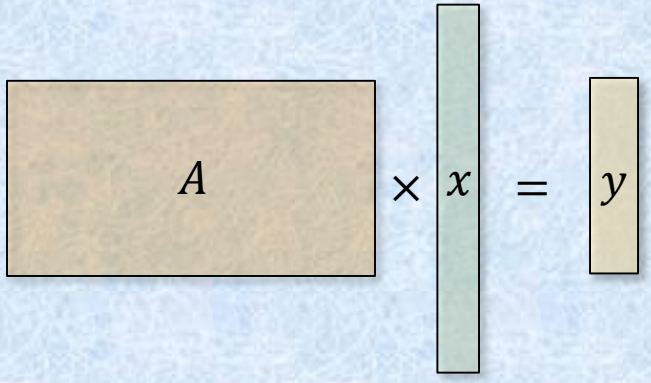
Θα δώσουμε ένα ακόμα παράδειγμα καθολικής οικογένειας συναρτήσεων διασποράς $\mathcal{H}$ υποθέτοντας ότι :

- Το σύνολο των πιθανών κλειδιών περιλαμβάνει $|U| = 2^{\upsilon}$ τιμές.
- Οι συναρτήσεις $h \in \mathcal{H}$ αντιστοιχούν το U σε ένα σύνολο με $m = 2^{\mu}$ τιμές.

Άρα ουσιαστικά οι συναρτήσεις της $\mathcal{H}$ αντιστοιχούν διανύσματα x (ακολουθίες) των υ bits σε διανύσματα y των μ bits.



Μπορούμε να λάβουμε μια τέτοια αντιστοιχία πολλαπλασιάζοντας το x με ένα $\mu \times \upsilon$ πίνακα A



Καθολικές Οικογένειες Συναρτήσεων Διασποράς

Θα δώσουμε ένα ακόμα παράδειγμα καθολικής οικογένειας συναρτήσεων διασποράς $\mathcal{H}$ υποθέτοντας ότι :

- Το σύνολο των πιθανών κλειδιών περιλαμβάνει $|U| = 2^{\upsilon}$ τιμές.
- Οι συναρτήσεις $h \in \mathcal{H}$ αντιστοιχούν το U σε ένα σύνολο με $m = 2^{\mu}$ τιμές.

Άρα ουσιαστικά οι συναρτήσεις της $\mathcal{H}$ αντιστοιχούν διανύσματα x (ακολουθίες) των υ bits σε διανύσματα y των μ bits.

Μπορούμε να λάβουμε μια τέτοια αντιστοιχία πολλαπλασιάζοντας το x με ένα Boolean $\mu \times \upsilon$ πίνακα A, δηλαδή έχουμε $y = h(x) = Ax$, όπου οι πράξεις γίνονται $\mathrm{modulo}\ 2$ (δηλαδή $0 + 0 = 1 + 1 = 0$ και $0 + 1 = 1 + 0 = 1$).

Η οικογένεια H ορίζεται από όλους τους $2^{\mu\upsilon}$ Boolean $\mu \times \upsilon$ πίνακες A.

Καθολικές Οικογένειες Συναρτήσεων Διασποράς

Η οικογένεια $\mathcal{H}$ ορίζεται από όλους τους $2^{\mu\upsilon}$ Boolean $\mu \times \upsilon$ πίνακες A.

Θεώρημα Η οικογένεια συναρτήσεων διασποράς $\mathcal{H}$ είναι καθολική.

Απόδειξη

Έστω A ένας τυχαίος Boolean $\mu \times \upsilon$ πίνακας και έστω $x \neq y$ διανύσματα του $\{0,1\}^{\upsilon}$. Θα δείξουμε ότι η πιθανότητα σύγκρουσης $h(x) = h(y)$ είναι το πολύ $1/m$, δηλαδή $\Pr_{h\in H}[h(x) = h(y)] \leq 1/m$.

Έστω $h(x) = h(y)$. Θέτουμε $z = x - y$. Έχουμε $Ax = Ay \Rightarrow A(x - y) = \mathbf{0} \Rightarrow Az = \mathbf{0}$, όπου $\mathbf{0} = (0\ 0\ \ldots 0)$ το διάνυσμα του $\{0,1\}^{\mu}$ με όλες τις μ συνιστώσες μηδέν.

Άρα θέλουμε να δείξουμε ότι $\Pr_{h\in H}[h(z) = \mathbf{0}] = \Pr_{h\in H}[Az = \mathbf{0}] \leq 1/m$.

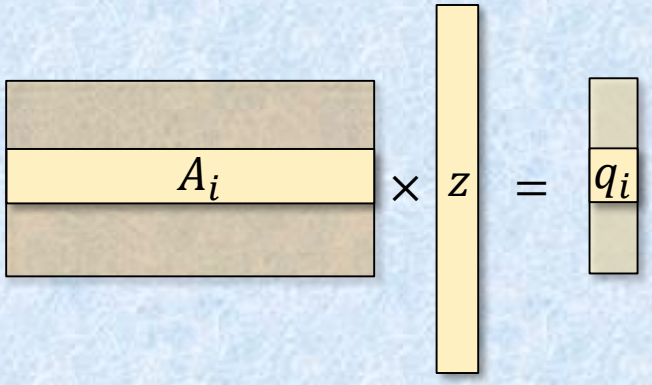
Καθολικές Οικογένειες Συναρτήσεων Διασποράς

Η οικογένεια $\mathcal{H}$ ορίζεται από όλους τους $2^{\mu\upsilon}$ Boolean $\mu \times \upsilon$ πίνακες A.

Θεώρημα Η οικογένεια συναρτήσεων διασποράς $\mathcal{H}$ είναι καθολική.

Απόδειξη

Άρα θέλουμε να δείξουμε ότι $\Pr_{h\in H}[h(z) = \mathbf{0}] = \Pr_{h\in H}[Az = \mathbf{0}] \leq 1/m$.
Έστω $q = (q_1, q_2, \ldots, q_\mu) = Az$. Η συνιστώσα q_i προκύπτει από το εσωτερικό γινόμενο της γραμμής i του A με το διάνυσμα z.



Εφόσον $z = x - y$ και $x \neq y$, το z έχει τουλάχιστον μία μη μηδενική συνιστώσα, έστω $z_k = 1$. Έχουμε $q_i = \sum_{j=1}^{\upsilon} A_{ij}z_j = A_{ik}z_k + \sum_{j\neq k} A_{ij}z_j = A_{ik} + \sum_{j\neq k} A_{ij}z_j$.

Καθολικές Οικογένειες Συναρτήσεων Διασποράς

Η οικογένεια $\mathcal{H}$ ορίζεται από όλους τους $2^{\mu\upsilon}$ Boolean $\mu \times \upsilon$ πίνακες A.

Θεώρημα Η οικογένεια συναρτήσεων διασποράς $\mathcal{H}$ είναι καθολική.

Απόδειξη

Άρα θέλουμε να δείξουμε ότι $\Pr_{h\in H}[h(z) = \mathbf{0}] = \Pr_{h\in H}[Az = \mathbf{0}] \leq 1/m$.
Έστω $q = (q_1, q_2, \dots, q_\mu) = Az$. Η συνιστώσα q_i προκύπτει από το εσωτερικό γινόμενο της γραμμής i του A με το διάνυσμα z.

Επομένως, $q_i = 0 \Rightarrow A_{ik} = \sum_{j\neq k} A_{ij} z_j$. Με τι πιθανότητα μπορεί να συμβεί αυτό;

Καθολικές Οικογένειες Συναρτήσεων Διασποράς

Η οικογένεια $\mathcal{H}$ ορίζεται από όλους τους $2^{\mu\upsilon}$ Boolean $\mu \times \upsilon$ πίνακες A.

Θεώρημα Η οικογένεια συναρτήσεων διασποράς $\mathcal{H}$ είναι καθολική.

Απόδειξη

Άρα θέλουμε να δείξουμε ότι $\Pr_{h\in H}[h(z) = \mathbf{0}] = \Pr_{h\in H}[Az = \mathbf{0}] \leq 1/m$.
Έστω $q = (q_1, q_2, \dots, q_\mu) = Az$. Η συνιστώσα q_i προκύπτει από το εσωτερικό γινόμενο της γραμμής i του A με το διάνυσμα z.

Επομένως, $q_i = 0 \Rightarrow A_{ik} = \sum_{j\neq k} A_{ij} z_j$. Με τι πιθανότητα μπορεί να συμβεί αυτό;

Αφού ο πίνακας A είναι τυχαίος, κάθε στοιχείο του επιλέγεται ανεξάρτητα. Επομένως, μπορούμε να υποθέσουμε ότι η τιμή (0 ή 1) του στοιχείου A_{ik} επιλέγεται τελευταία. Τη στιγμή που επιλέγουμε αυτήν την τιμή, το άθροισμα $c = \sum_{j\neq k} A_{ij} z_j$ είναι καθορισμένο, δηλαδή είναι μια σταθερά $c \in \{0,1\}$.

Συνεπώς η πιθανότητα να επιλέξουμε $A_{ik} = c$ είναι $1/2$.

Καθολικές Οικογένειες Συναρτήσεων Διασποράς

Η οικογένεια $\mathcal{H}$ ορίζεται από όλους τους $2^{\mu\upsilon}$ Boolean $\mu \times \upsilon$ πίνακες A.

Θεώρημα Η οικογένεια συναρτήσεων διασποράς $\mathcal{H}$ είναι καθολική.

Απόδειξη

Άρα θέλουμε να δείξουμε ότι $\Pr_{h\in H}[h(z) = \mathbf{0}] = \Pr_{h\in H}[Az = \mathbf{0}] \leq 1/m$.
Έστω $q = (q_1, q_2, \dots, q_\mu) = Az$. Η συνιστώσα q_i προκύπτει από το εσωτερικό γινόμενο της γραμμής i του A με το διάνυσμα z.

Επομένως, $q_i = 0 \Rightarrow A_{ik} = \sum_{j\neq k} A_{ij} z_j$. Η πιθανότητα να συμβεί αυτό είναι $1/2$.

Το ίδιο ισχύει για κάθε συνιστώσα $i \in \{1,2, \dots, \mu\}$, δηλαδή

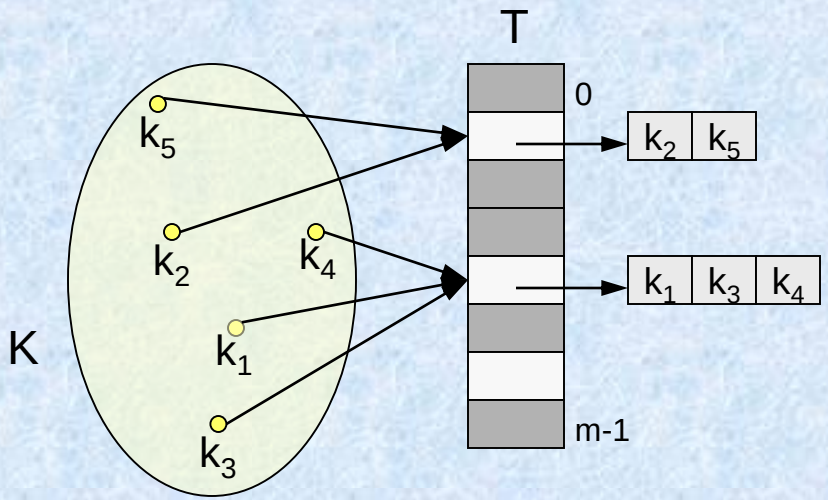
$$\Pr_{h\in H}[h(x) = h(y)] = \Pr_{h\in H}[Az = \mathbf{0}] = \left(\frac{1}{2}\right)^{\mu} = \frac{1}{m}$$

Πλήρης Διασπορά

Όταν το σύνολο των κλειδιών είναι **στατικό** μπορούμε να πετύχουμε άριστη επίδοση : Ο(1) χρόνο χειρότερης περίπτωσης ανά πράξη

Πλήρης Διασπορά

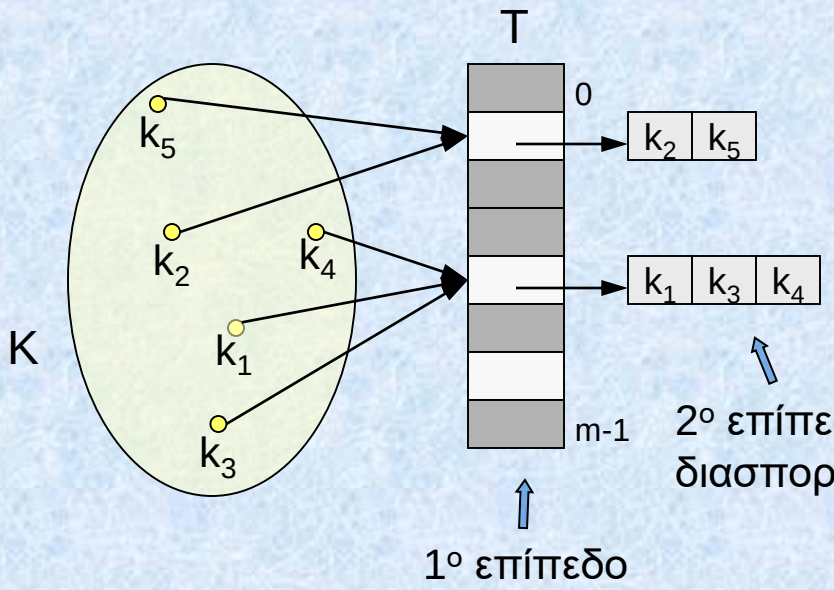
Όταν το σύνολο των κλειδιών είναι **στατικό** μπορούμε να πετύχουμε άριστη επίδοση : O(1) χρόνο χειρότερης περίπτωσης ανά πράξη



Ιδέα: Διασπορά δύο επιπέδων με χρήση καθολικής διασποράς ανά επίπεδο

Πλήρης Διασπορά

Όταν το σύνολο των κλειδιών είναι **στατικό** μπορούμε να πετύχουμε άριστη επίδοση : O(1) χρόνο χειρότερης περίπτωσης ανά πράξη

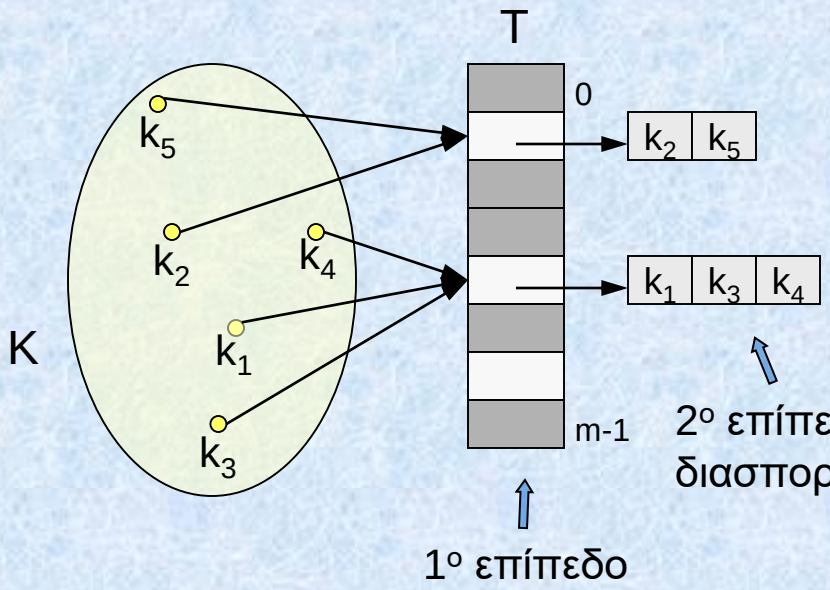


Ιδέα: Διασπορά δύο επιπέδων με χρήση καθολικής διασποράς ανά επίπεδο

2ο επίπεδο : δευτερογενής πίνακας διασποράς για κάθε T[i]

Πλήρης Διασπορά

Όταν το σύνολο των κλειδιών είναι **στατικό** μπορούμε να πετύχουμε άριστη επίδοση : O(1) χρόνο χειρότερης περίπτωσης ανά πράξη



Ιδέα: Διασπορά δύο επιπέδων με χρήση καθολικής διασποράς ανά επίπεδο

2ο επίπεδο : δευτερογενής πίνακας διασποράς για κάθε T[i]

Στο 2ο επίπεδο αποφεύγουμε τις συμπτώσεις

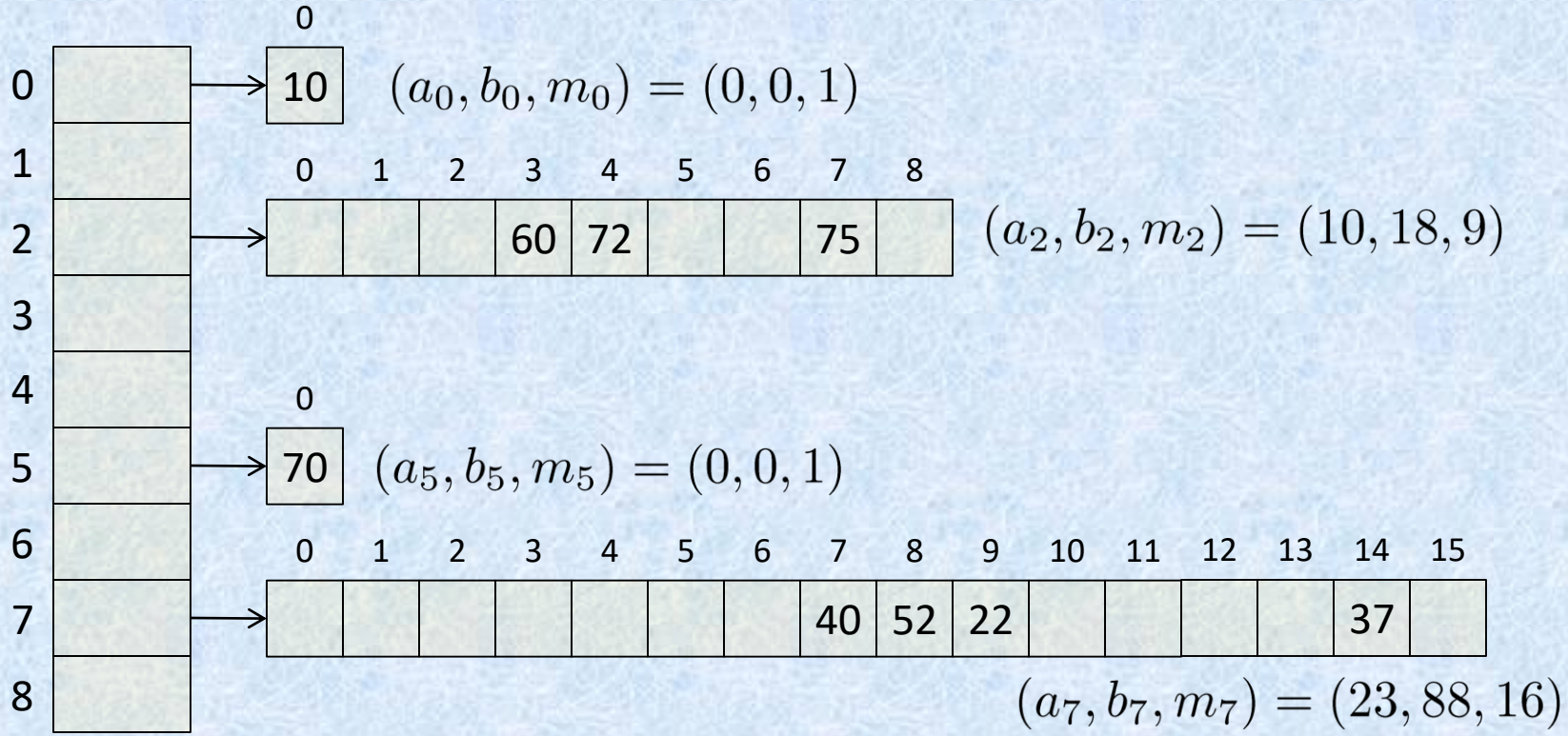
Πλήρης Διασπορά

Παράδειγμα $K = \{10, 22, 37, 40, 52, 60, 70, 72, 75\}$

Συνάρτηση διασποράς 1ου επιπέδου : $h(k) = [(a \cdot k + b) \bmod p] \bmod m$

$(a, b, p, m) = (3, 42, 101, 9)$

Συνάρτηση διασποράς 2ου επιπέδου : $h_j(k) = [(a_j \cdot k + b_j) \bmod p] \bmod m_j$



Πλήρης Διασπορά

$T =$ πρωτογενής πίνακας διασποράς με m θέσεις και συνάρτηση διασποράς h

$S_j =$ δευτερογενής πίνακας διασποράς με m_j θέσεις και συνάρτηση διασποράς h_j
$j = 0, 1, \ldots, m$

Έστω n ο συνολικός αριθμός κλειδιών και n_j ο αριθμός των κλειδιών που αποθηκεύονται στον S_j . $\left(\sum_{j=0}^{m-1} n_j = n\right)$

Για να αποφύγουμε τις συγκρούσεις στο δεύτερο επίπεδο θα επιλέξουμε $m_j = n_j^2$.

Θα δείξουμε ότι για κατάλληλα επιλεγμένη h ο αναμενόμενος χώρος που απαιτείται για τους δευτερογενείς πίνακες είναι

$$\mathbf{E}\left[\sum_{j=0}^{m-1} m_j\right] = O(n)$$

Πλήρης Διασπορά

Επιλέγουμε την πρωτογενή συνάρτηση διασποράς h από την καθολική οικογένεια $\mathcal{H}_{p,m}$.

Επιλέγουμε κάθε δευτερογενή συνάρτηση διασποράς h_j από την καθολική οικογένεια $\mathcal{H}_{p,m_j}$.

Πλήρης Διασπορά

Θεώρημα Έστω ότι επιλέγουμε τυχαία μια συνάρτηση διασποράς h από καθολική οικογένεια συναρτήσεων διασποράς. Αν αποθηκεύσουμε n κλειδιά και το μέγεθος του πίνακα διασποράς είναι $m = n^2$ τότε η πιθανότητα να υπάρχει σύγκρουση είναι $< 1/2$.

Ο αριθμός των πιθανών συγκρούσεων είναι $\binom{n}{2}$, όσα τα διαφορετικά ζεύγη κλειδιών.

Από τον τρόπο επιλογής της h προκύπτει ότι η πιθανότητα σύγκρουσης ενός ζεύγους κλειδιών είναι το πολύ $1/m$.

Άρα ο αναμενόμενος αριθμός συγκρούσεων X είναι

$$\mathbf{E}[X] \leq \frac{\binom{n}{2}}{m} = \frac{n(n-1)}{2n^2} < \frac{1}{2}$$

Από την ανισότητα του Markov έχουμε

$$\mathbf{Pr}[X \geq t] \leq \frac{\mathbf{E}[X]}{t} \;\Rightarrow\; \mathbf{Pr}[X \geq 1] < \frac{1}{2}$$

Πλήρης Διασπορά

Θεώρημα Έστω ότι επιλέγουμε τυχαία μια συνάρτηση διασποράς h από καθολική οικογένεια συναρτήσεων διασποράς. Αν αποθηκεύσουμε n κλειδιά και το μέγεθος του πίνακα διασποράς είναι $m = n^2$ τότε η πιθανότητα να υπάρχει σύγκρουση είναι $< 1/2$.

Το παραπάνω θεώρημα συνεπάγεται ότι για $m_j = n_j^2$ μπορούμε να βρούμε με λίγες δοκιμές μια συνάρτηση διασποράς h_j που να μην έχει συγκρούσεις για το σύνολο των κλειδιών που αποθηκεύεται στον S_j .

Πλήρης Διασπορά

Απομένει να επιλέξουμε μια πρωτογενή συνάρτηση διασποράς h έτσι ώστε

$$\mathbf{E}\Big[\sum_{j=0}^{m-1} m_j\Big] = \mathbf{E}\Big[\sum_{j=0}^{m-1} n_j^2\Big] = O(n)$$

Θεώρημα Έστω ότι επιλέγουμε τυχαία μια συνάρτηση διασποράς h από καθολική οικογένεια συναρτήσεων διασποράς. Αν αποθηκεύσουμε n κλειδιά και το μέγεθος του πίνακα διασποράς είναι $m = n$ τότε

$$\mathbf{E}\Big[\sum_{j=0}^{m-1} n_j^2\Big] < 2n$$

Πλήρης Διασπορά

Θεώρημα Έστω ότι επιλέγουμε τυχαία μια συνάρτηση διασποράς h από καθολική οικογένεια συναρτήσεων διασποράς. Αν αποθηκεύσουμε n κλειδιά και το μέγεθος του πίνακα διασποράς είναι $m = n$ τότε

$$\mathbf{E}\Big[\sum_{j=0}^{m-1} n_j^2\Big] < 2n$$

Για κάθε μη αρνητικό ακέραιο a ισχύει $a^2 = a + 2\binom{a}{2}$ άρα έχουμε

$$\mathbf{E}\Big[\sum_{j=0}^{m-1} n_j^2\Big] = \mathbf{E}\Big[\sum_{j=0}^{m-1} \Big(n_j + 2\binom{n_j}{2}\Big)\Big] = \mathbf{E}\Big[\sum_{j=0}^{m-1} n_j\Big] + 2 \cdot \mathbf{E}\Big[\sum_{j=0}^{m-1} \binom{n_j}{2}\Big]$$

Καθώς $\mathbf{E}\Big[\sum_{j=0}^{m-1} n_j\Big] = n$ απομένει να δείξουμε ότι $\mathbf{E}\Big[\sum_{j=0}^{m-1} \binom{n_j}{2}\Big] < \frac{n}{2}$

Πλήρης Διασπορά

Θεώρημα Έστω ότι επιλέγουμε τυχαία μια συνάρτηση διασποράς h από καθολική οικογένεια συναρτήσεων διασποράς. Αν αποθηκεύσουμε n κλειδιά και το μέγεθος του πίνακα διασποράς είναι $m = n$ τότε

$$\mathbf{E}\Big[\sum_{j=0}^{m-1} n_j^2\Big] < 2n$$

Παρατηρούμε ότι $\mathbf{E}\Big[\sum_{j=0}^{m-1} \binom{n_j}{2}\Big] =$ αναμενόμενο πλήθος συγκρούσεων, άρα από την επιλογή της h έχουμε $\mathbf{E}\Big[\sum_{j=0}^{m-1} \binom{n_j}{2}\Big] \leq \frac{\binom{n}{2}}{m} = \frac{n(n-1)}{2n} < \frac{n}{2}$

Πλήρης Διασπορά

Θεώρημα Έστω ότι επιλέγουμε τυχαία μια συνάρτηση διασποράς h από καθολική οικογένεια συναρτήσεων διασποράς. Αν αποθηκεύσουμε n κλειδιά και το μέγεθος του πίνακα διασποράς είναι $m = n$ τότε

$$\mathbf{E}\Big[\sum_{j=0}^{m-1} n_j^2\Big] < 2n$$

Άρα ο αναμενόμενος χώρος που απαιτείται για την αποθήκευση όλων των δευτερογενών πινάκων είναι

$$\mathbf{E}\Big[\sum_{j=0}^{m-1} m_j\Big] < 2n$$

Από την ανισότητα του Markov έχουμε

$$\mathbf{Pr}\Big[\sum_{j=0}^{m-1} m_j \geq 4n\Big] \leq \frac{\mathbf{E}\Big[\sum_{j=0}^{m-1} m_j\Big]}{4n} < \frac{1}{2}$$

Δηλαδή μπορούμε να βρούμε με λίγες δοκιμές μια κατάλληλη πρωτογενή συνάρτηση διασποράς.

Πλήρης Διασπορά

Η τεχνική της πλήρους διασποράς μπορεί να επεκταθεί και σε δυναμικά σύνολα [Dietzfelbinger, Karlin, Mehlhorn, Meyer auf der Heide, Rohnert, and Tarjan, 1994] :

$O(1)$ αντισταθμιστίκο αναμενόμενο κόστος ανά εισαγωγή και διαγραφή.

k-Καθολική Διασπορά

$\mathcal{H}$ = πεπερασμένη συλλογή συναρτήσεων διασποράς $U \to K$, $|U| > |K| = m$

Η συλλογή $\mathcal{H}$ είναι k-καθολική (ή k-ανεξάρτητη) αν για κάθε k διαφορετικά κλειδιά $x_1, x_2, \ldots, x_k$ και k τιμές $a_1, a_2, \ldots, a_k$ (όχι απαραίτητα διαφορετικές) ισχύει

$$\Pr_{h \in \mathcal{H}}[h(x_1) = a_1 \wedge h(x_2) = a_2 \wedge \ \ldots \wedge h(x_k) = a_k] \leq \frac{1}{m^k}$$

Ιδιότητες

- Αν η $\mathcal{H}$ είναι k-καθολική τότε είναι και $(k-1)$-καθολική (και καθολική)
- Για κάθε $x \in U$ και $a \in K$ ισχύει $\Pr_{h \in \mathcal{H}}[h(x) = a] = 1/m$

k-Καθολική Διασπορά

Η οικογένεια $\mathcal{H}$ ορίζεται από όλους τους $2^{\mu\upsilon}$ Boolean $\mu \times \upsilon$ πίνακες A.

Θεώρημα Η οικογένεια συναρτήσεων διασποράς $\mathcal{H}$ είναι καθολική.

Παρατήρηση Η παραπάνω οικογένεια δεν είναι 2-καθολική

Π.χ. για $x = \mathbf{0}$ έχουμε $h(x) = Ax = \mathbf{0}$ ανεξάρτητα από την επιλογή του A, δηλαδή $\Pr_{h\in\mathcal{H}}[h(x) = \mathbf{0}] = 1$.

Μπορούμε όμως να λάβουμε μια 2-καθολική οικογένεια αν θέσουμε

$$h(x) = Ax + b$$

όπου b ένα τυχαίο διάνυσμα του $\{0,1\}^{\mu}$.

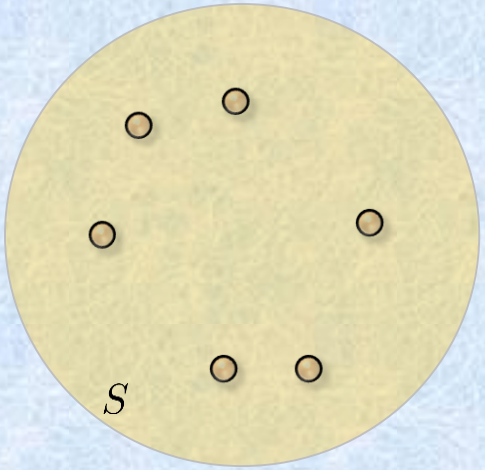
Φίλτρα Bloom [B. H. Bloom 1970]

Απλή δομή δεδομένων η οποία απαντά γρήγορα αν ένα στοιχείο ανήκει σε ένα σύνολο S

Βασικές λειτουργίες

$\text{ανήκει}(S, x)$: ελέγχει αν $x \in S$

$\text{εισαγωγή}(S, x)$: θέτει $S \leftarrow S \cup \{x\}$



Ένα φίλτρο Bloom επιτρέπει λανθασμένες θετικές απαντήσεις, δηλαδή μπορεί να επιστρέψει "$\text{ανήκει}(S, x) = \text{true}$" για $x \notin S$, αλλά με πολύ μικρή πιθανότητα.
Γι' αυτό και ονομάζεται «φίλτρο»! (Σε περίπτωση που λάβουμε θετική απάντηση μπορούμε να αναζητήσουμε το x σε μια άλλη δομή η οποία δίνει πάντα τη σωστή απάντηση αλλά μπορεί να είναι πιο αργή.)

Τα φίλτρα Bloom χρησιμοποιούνται σε πολλές εφαρμογές κυρίως σε δίκτυα (π.χ. αναφέρεται ότι τα χρησιμοποιεί το Google Chrome), αλλά και αλλού.

Φίλτρα Bloom

Η δομή αποτελείται από ένα Boolean πίνακα T μεγέθους m και k συναρτήσεις διασποράς $h_1, h_2, \dots, h_k : U \to \{0,1, \dots, m-1\}$

Για την ανάλυση υποθέτουμε ότι οι συναρτήσεις διασποράς ικανοποιούν την υπόθεση της απλής ομοιόμορφης διασποράς:

Υπόθεση **απλής ομοιόμορφης διασποράς**: κάθε νέο στοιχείο που εισαγάγουμε έχει ίση πιθανότητα να διασπαρεί σε οποιαδήποτε από τις m θέσεις του πίνακα T

Αρχικά θέτουμε $T[i] \leftarrow 0$ για $i = 0,1, \dots, m-1$

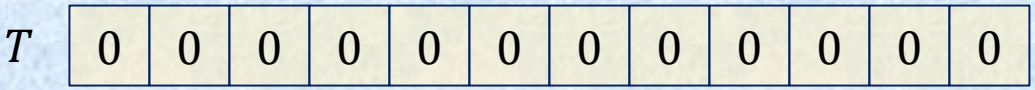
εισαγωγή(S, x): θέτουμε όλα τα bit $T[h_1(x)], T[h_2(x)], \dots, T[h_k(x)]$ να είναι 1

ανήκει(S, x): επιστρέφουμε "true" αν όλα τα bit $T[h_1(x)], T[h_2(x)], \dots, T[h_k(x)]$ είναι 1, διαφορετικά επιστρέφουμε "false"

Φίλτρα Bloom

Αρχικά θέτουμε $T[i] \leftarrow 0$ για $i = 0,1, \ldots, m-1$

εισαγωγή(S, x): θέτουμε όλα τα bit $T[h_1(x)], T[h_2(x)], \ldots, T[h_k(x)]$ να είναι 1

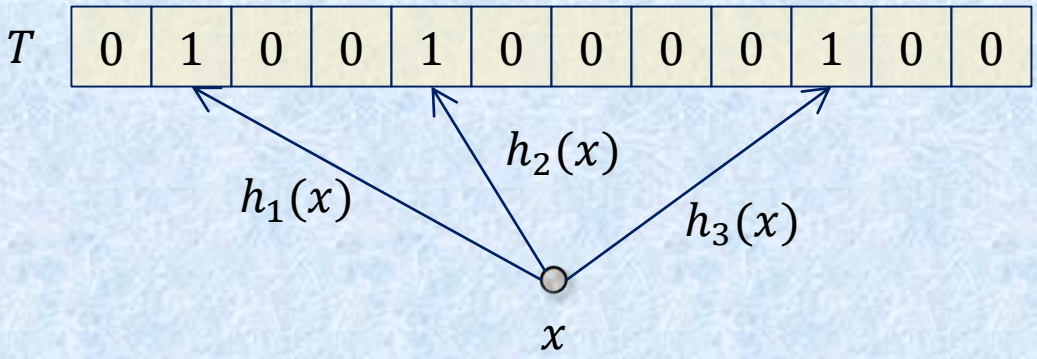


ανήκει(S, x): επιστρέφουμε "true" αν όλα τα bit $T[h_1(x)], T[h_2(x)], \ldots, T[h_k(x)]$ είναι 1, διαφορετικά επιστρέφουμε "false"

Φίλτρα Bloom

Αρχικά θέτουμε $T[i] \leftarrow 0$ για $i = 0,1, \ldots, m-1$

εισαγωγή(S, x): θέτουμε όλα τα bit $T[h_1(x)], T[h_2(x)], \ldots, T[h_k(x)]$ να είναι 1

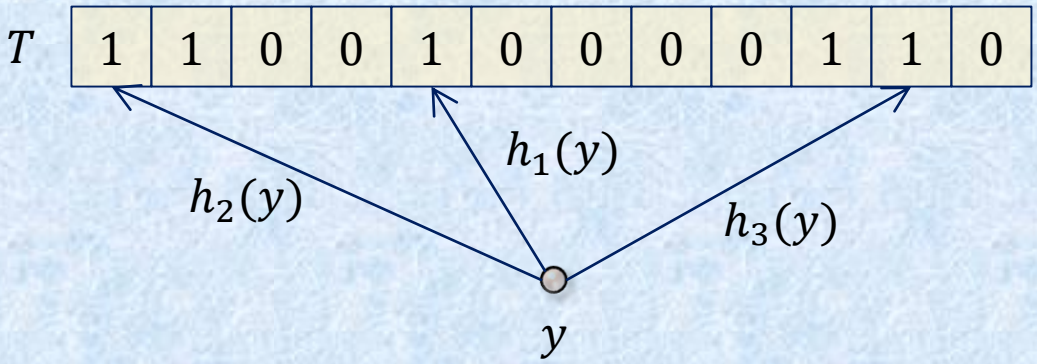


ανήκει(S, x): επιστρέφουμε "true" αν όλα τα bit $T[h_1(x)], T[h_2(x)], \ldots, T[h_k(x)]$ είναι 1, διαφορετικά επιστρέφουμε "false"

Φίλτρα Bloom

Αρχικά θέτουμε $T[i] \leftarrow 0$ για $i = 0,1, \ldots, m-1$

εισαγωγή(S, x): θέτουμε όλα τα bit $T[h_1(x)], T[h_2(x)], \ldots, T[h_k(x)]$ να είναι 1



ανήκει(S, x): επιστρέφουμε "true" αν όλα τα bit $T[h_1(x)], T[h_2(x)], \ldots, T[h_k(x)]$ είναι 1, διαφορετικά επιστρέφουμε "false"

Φίλτρα Bloom

Η δομή αποτελείται από ένα Boolean πίνακα T μεγέθους m και k συναρτήσεις διασποράς $h_1, h_2, \dots, h_k : U \to \{0,1, \dots, m-1\}$

Για την ανάλυση υποθέτουμε ότι οι συναρτήσεις διασποράς ικανοποιούν την υπόθεση της απλής ομοιόμορφης διασποράς:

Υπόθεση **απλής ομοιόμορφης διασποράς**: κάθε νέο στοιχείο που εισαγάγουμε έχει ίση πιθανότητα να διασπαρεί σε οποιαδήποτε από τις m θέσεις του πίνακα T

Έστω ένα αυθαίρετο στοιχείο $x \in U$ και θέση $\lambda \in \{0,1, \dots, m-1\}$
Τότε $\Pr[h_i(x) \neq \lambda] = 1 - 1/m.$

Άρα αν κάνουμε n εισαγωγές, η πιθανότητα να έχουμε $T[\lambda] = 0$ είναι
$p = (1 - 1/m)^{kn} \approx e^{-kn/m}$

Φίλτρα Bloom

Έστω ένα αυθαίρετο στοιχείο $x \in U$ και θέση $\lambda \in \{0,1,\dots,m-1\}$
Τότε $\Pr[h_i(x) \neq \lambda] = 1 - 1/m$.

Άρα αν κάνουμε n εισαγωγές, η πιθανότητα να έχουμε $T[\lambda] = 0$ είναι
$p = (1 - 1/m)^{kn} \approx e^{-kn/m}$

Υπολογίζουμε τώρα την πιθανότητα το φίλτρο Bloom να δώσει λάθος θετική απάντηση.

Έστω $x \notin S$ αλλά "ανήκει$(S, x) = \text{true}$". Τότε $T[h_1(x)] = \cdots = T[h_k(x)] = 1$
το οποίο συμβαίνει με πιθανότητα

$$(1-p)^k \approx \left(1 - e^{-kn/m}\right)^k$$

Π.χ. για $m = 8n$ και $k = 4$ έχουμε πιθανότητα σφάλματος $< 2{,}5\%$
Βέλτιστο $k = \ln 2 \cdot m/n$.
Αν θέλουμε πιθανότητα σφάλματος $< \varepsilon$ μπορούμε να θέσουμε $m \approx 1.44n \log(1/\varepsilon)$

Κατανομή Φόρτου Εργασίας

Έστω ότι έχουμε n εργασίες οι οποίες πρέπει να εκτελεστούν σε m υπολογιστές.

Πως μπορούμε να αναθέσουμε τις εργασίες στους υπολογιστές έτσι ώστε οι υπολογιστές να έχουν περίπου τον ίδιο φορτίο;

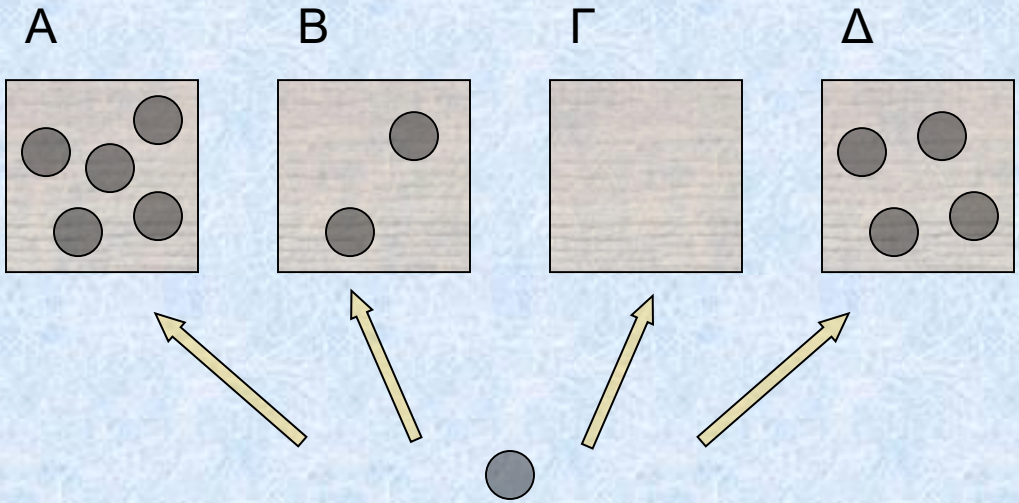
Αυτό επιτυγχάνεται εύκολα αν έχουμε μια κεντρική εργασία η οποία μπορεί να αναθέτει κάθε νέα εργασία σε κάποιον υπολογιστή, γνωρίζοντας το φορτίο κάθε υπολογιστή.

Αν δεν υπάρχει αυτή η δυνατότητα μπορούμε να χρησιμοποιήσουμε συναρτήσεις διασποράς οι οποίες αντιστοιχούν το σύνολο των εργασιών U στους m υπολογιστές.

Κατανομή Φόρτου Εργασίας

Ρίψη σφαιριδίων σε κάλπες

Έχουμε b κάλπες στις οποίες τοποθετούμε σφαιρίδια με τυχαίο τρόπο.



Έστω p η πιθανότητα ένα σφαιρίδιο να πάει σε μία συγκεκριμένη κάλπη.

Έχουμε $p = \frac{1}{b}$. Ρίχνουμε n σφαίρες. Ποια η πιθανότητα να πάνε ακριβώς k σε μία συγκεκριμένη κάλπη;

$$\Rightarrow \binom{n}{k} p^k (1-p)^{n-k}$$

διωνυμική κατανομή

Κατανομή Φόρτου Εργασίας

Ρίψη σφαιριδίων σε κάλπες

Έχουμε b κάλπες στις οποίες τοποθετούμε σφαιρίδια με τυχαίο τρόπο.

Έστω p η πιθανότητα ένα σφαιρίδιο να πάει σε μία συγκεκριμένη κάλπη.

Έχουμε $p = \frac{1}{b}$. Ρίχνουμε n σφαίρες. Ποια η πιθανότητα να πάνε ακριβώς k σε μία συγκεκριμένη κάλπη; ⇨ $\binom{n}{k} p^k (1-p)^{n-k}$

διωνυμική κατανομή

τυχαία μεταβλητή $X =$ αριθμός σφαιριδίων σε μία συγκεκριμένη κάλπη

Κατανομή Φόρτου Εργασίας

Ρίψη σφαιριδίων σε κάλπες

Έχουμε b κάλπες στις οποίες τοποθετούμε σφαιρίδια με τυχαίο τρόπο.

Έστω p η πιθανότητα ένα σφαιρίδιο να πάει σε μία συγκεκριμένη κάλπη.

Έχουμε $p = \frac{1}{b}$. Ρίχνουμε n σφαίρες. Ποια η πιθανότητα να πάνε ακριβώς k σε μία συγκεκριμένη κάλπη; ⇨ $\binom{n}{k} p^k (1-p)^{n-k}$

διωνυμική κατανομή

τυχαία μεταβλητή $X =$ αριθμός σφαιριδίων σε μία συγκεκριμένη κάλπη

$$\Pr[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

Κατανομή Φόρτου Εργασίας

Ρίψη σφαιριδίων σε κάλπες

Έχουμε b κάλπες στις οποίες τοποθετούμε σφαιρίδια με τυχαίο τρόπο.

Έστω p η πιθανότητα ένα σφαιρίδιο να πάει σε μία συγκεκριμένη κάλπη.

Έχουμε $p = \frac{1}{b}$. Ρίχνουμε n σφαίρες. Ποια η πιθανότητα να πάνε ακριβώς k σε μία συγκεκριμένη κάλπη; ⇨ $\binom{n}{k} p^k (1-p)^{n-k}$

διωνυμική κατανομή

τυχαία μεταβλητή $X =$ αριθμός σφαιριδίων σε μία συγκεκριμένη κάλπη

αναμενόμενος αριθμός σφαιριδίων σε μία κάλπη

$$\mathrm{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

Κατανομή Φόρτου Εργασίας

Ρίψη σφαιριδίων σε κάλπες

Έχουμε b κάλπες στις οποίες τοποθετούμε σφαιρίδια με τυχαίο τρόπο.

Έστω p η πιθανότητα ένα σφαιρίδιο να πάει σε μία συγκεκριμένη κάλπη.

Έχουμε $p = \frac{1}{b}$. Ρίχνουμε n σφαίρες.

τυχαία μεταβλητή $X =$ αριθμός σφαιριδίων σε μία συγκεκριμένη κάλπη

αναμενόμενος αριθμός σφαιριδίων σε μία κάλπη

$$\mathrm{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k}$$

$$= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{(n-1)-k} = np\big[p + (1-p)\big]^n = np$$

Κατανομή Φόρτου Εργασίας

Ρίψη σφαιριδίων σε κάλπες

Έχουμε b κάλπες στις οποίες τοποθετούμε σφαιρίδια με τυχαίο τρόπο.

Έστω p η πιθανότητα ένα σφαιρίδιο να πάει σε μία συγκεκριμένη κάλπη.

Έχουμε $p = \frac{1}{b}$. Ρίχνουμε n σφαίρες.

τυχαία μεταβλητή $X =$ αριθμός σφαιριδίων σε μία συγκεκριμένη κάλπη

αναμενόμενος αριθμός σφαιριδίων σε μία κάλπη

$$\mathrm{E}[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k}$$

$$= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{(n-1)-k} = np\big[p + (1-p)\big]^n = np$$

Ο υπολογισμός με τη βοήθεια δεικτριών τυχαίων μεταβλητών είναι πολύ πιο απλός!

Κατανομή Φόρτου Εργασίας

Ρίψη σφαιριδίων σε κάλπες

Έχουμε b κάλπες στις οποίες τοποθετούμε σφαιρίδια με τυχαίο τρόπο.

Έστω p η πιθανότητα ένα σφαιρίδιο να πάει σε μία συγκεκριμένη κάλπη.

Έχουμε $p = \frac{1}{b}$. Ρίχνουμε n σφαίρες.

τυχαία μεταβλητή $X =$ αριθμός σφαιριδίων σε μία συγκεκριμένη κάλπη

$$X_i = \begin{cases} 0 & \text{η i-οστη σφαίρα δεν πάει στο συγκεκριμένο κουτί} \\ 1 & \text{η i-οστη σφαίρα πάει στο συγκεκριμένο κουτί} \end{cases} \Rightarrow X = \sum_{i=1}^{n} X_i$$

Έχουμε $$\mathrm{E}[X] = \mathrm{E}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathrm{E}[X_i] = \sum_{i=1}^{n} \Pr[X_i = 1]$$

$$= \sum_{i=1}^{n} p = np = \frac{n}{b}$$

Κατανομή Φόρτου Εργασίας

Άρα αν τοποθετήσουμε n εργασίες σε n υπολογιστές με τυχαίο τρόπο τότε το αναμενόμενο φορτίο ανά υπολογιστή είναι 1 εργασία.

Ωστόσο μπορούμε να δείξουμε ότι

Θεώρημα Το μέγιστο φορτίο ενός υπολογιστή είναι $O\left(\frac{\log n}{\log \log n}\right)$ με πιθανότητα $\geq 1 - 1/n$

Σημείωση: Όταν η ανάθεση γίνεται με συνάρτηση διασποράς h, τότε για να ισχύει το παραπάνω θεώρημα πρέπει η h να προέρχεται από k-καθολική οικογένεια για $k = O\left(\frac{\log n}{\log \log n}\right)$

Μπορούμε να βελτιώσουμε το άνω φράγμα του θεωρήματος με τη βοήθεια περισσότερων συναρτήσεων διασποράς $h_1, \dots, h_d$

Κάθε εργασία x αντιστοιχείται σε $d \geq 2$ υπολογιστές $h_1(x), \dots, h_d(x)$ και τοποθετείται σε αυτόν με το ελάχιστο φορτίο.

Κατανομή Φόρτου Εργασίας

Χρησιμοποιούμε συναρτήσεις διασποράς $h_1, \dots, h_d$

Κάθε εργασία x αντιστοιχείται σε $d \geq 2$ υπολογιστές $h_1(x), \dots, h_d(x)$ και τοποθετείται σε αυτόν με το ελάχιστο φορτίο.

Θεώρημα Το μέγιστο φορτίο ενός υπολογιστή είναι $\approx \frac{\ln \ln n}{\ln d}$ με πιθανότητα $\geq 1 - 0(1/n)$

Διασπορά του Κούκου (Cuckoo Hashing)

Μέθοδος κατακερματισμού ανοικτής διευθυνσιοδότησης, η οποία χρησιμοποιεί την ιδέα των 2 επιλογών. Οι συγκρούσεις επιλύονται με «έξωση» του προηγούμενου κλειδιού.

Έχουμε δύο πίνακες διασποράς, T_1 και T_2, m θέσεων και αντίστοιχες συναρτήσεις διασποράς $h_1, h_2 : U \to \{0, \dots, m-1\}$, από κάποια οικογένεια $\mathcal{H}$.

Εισαγωγή κλειδιού x :

- Αν κάποια από τις θέσεις $T_1[h_1(x)]$ και $T_2[h_2(x)]$ είναι κενή, τοποθετούμε το εκεί το x.
- Διαφορετικά έστω ότι $T_1[h_1(x)] = y$. Κάνουμε «έξωση» στο κλειδί y και τοποθετούμε το x στη θέση του.
- Όταν γίνεται έξωση σε ένα κλειδί z από τη θέση $T_j[h_j(z)]$ (για κάποιο $j \in \{1,2\}$), τοποθετούμε το z στη θέση $T_{3-j}[h_{3-j}(z)]$, κάνοντας ενδεχομένως έξωση σε κάποιο άλλο κλειδί.
- Αν συμβούν $O(\log n)$ εξώσεις, επιλέγουμε διαφορετικές συναρτήσεις διασποράς και τοποθετούμε από την αρχή όλα τα αντικείμενα.

Διασπορά του Κούκου (Cuckoo Hashing)

Μέθοδος κατακερματισμού ανοικτής διευθυνσιοδότησης, η οποία χρησιμοποιεί την ιδέα των 2 επιλογών. Οι συγκρούσεις επιλύονται με «έξωση» του προηγούμενου κλειδιού.

Έχουμε δύο πίνακες διασποράς, T_1 και T_2, m θέσεων και αντίστοιχες συναρτήσεις διασποράς $h_1, h_2 : U \to \{0, \dots, m-1\}$, από κάποια οικογένεια $\mathcal{H}$.

Κάθε κλειδί x βρίσκεται σε μία από τις θέσεις $T_1[h_1(x)]$ και $T_2[h_2(x)]$, άρα η αναζήτηση και η διαγραφή γίνονται σε $O(1)$.

Για να έχουμε καλή απόδοση στις εισαγωγές πρέπει το m να είναι αρκετά μεγάλο.

Θεώρημα Για $m \geq 4n$, ο αναμενόμενος χρόνος εισαγωγής είναι $O(1)$