HASHING

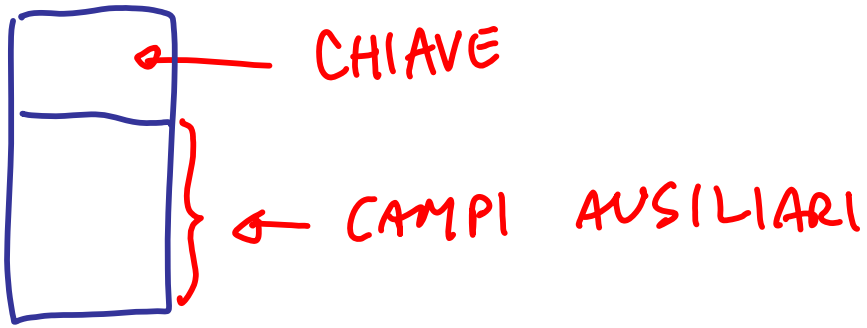
PROBLEMA: RAPPRESENTAZIONE DI INSIEMI DINAMICI CON SUPPORTO EFFICIENTE DI

- INSERIMENTI
- RICERCHE (MEDIANTE CHIAVI)
- CANCELLAZIONI

(DIZIONARIO)

ES. APPLICAZIONE: - MANTENIMENTO TABELLE DI SIMBOLI (IDENTIFICATORI)

NOTA: GLI ELEMENTI SONO RAPPRESENTATI MEDIANTE RECORD:



RAPPRESENTAZIONE MEDIANTE TAVOLE AD INDIRIZZAMENTO DIRETTO (ARRAY)

- $U = \{0, 1, \dots, m-1\}$ (UNIVERSO DELLE CHIAVI)

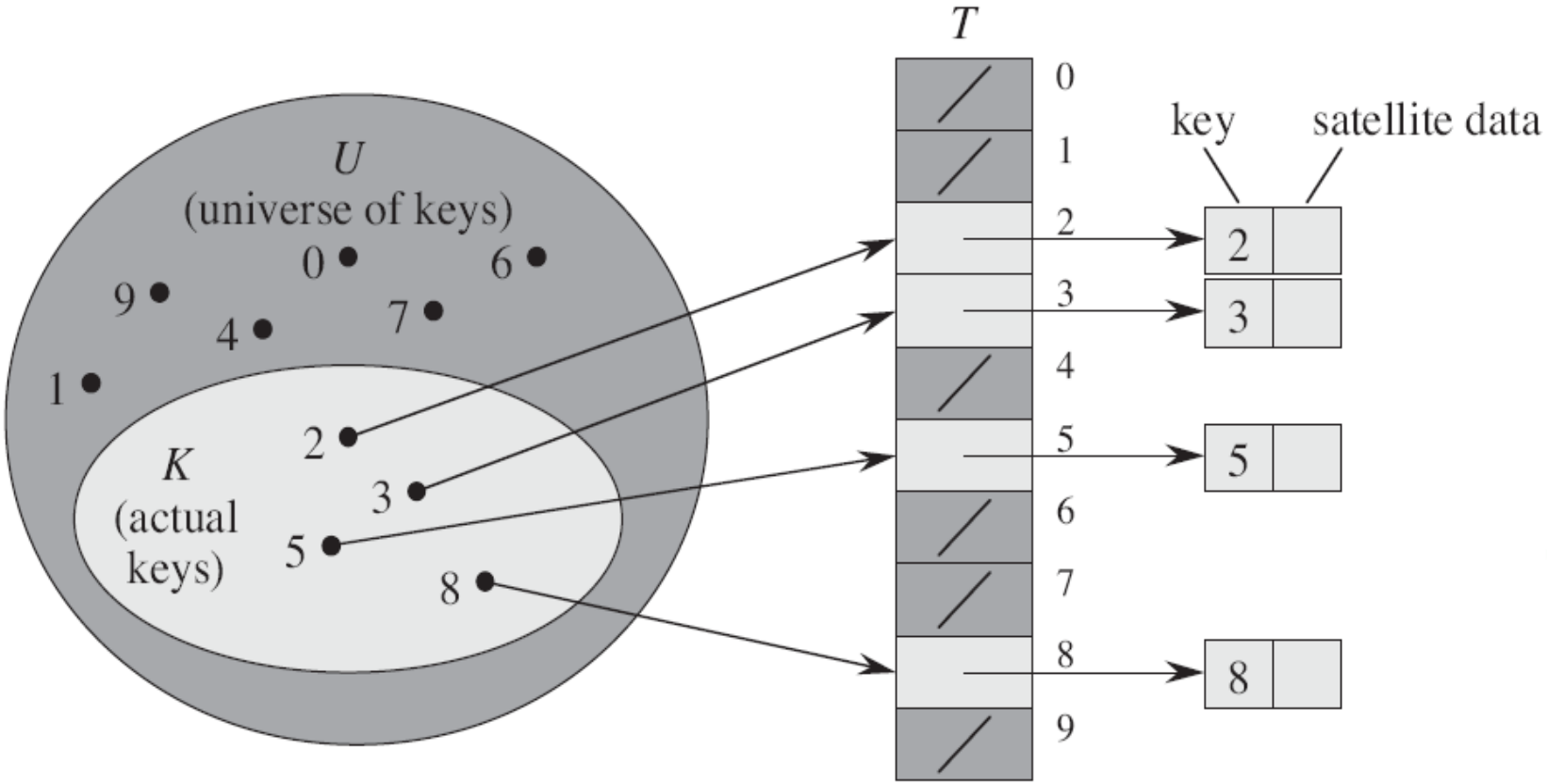
- RAPPRESENTIAMO UN INSIEME A DI ELEMENTI LE CUI CHIAVI SONO IN U MEDIANTE UN ARRAY $T_A[0..m-1]$ TALE CHE

$$x \in A \Rightarrow T_A[x.key] = x$$

$$x \notin A \Rightarrow T_A[x.key] = NIL$$

- NOTA: LE CHIAVI DEBBONO ESSERE A DUE A DUE DISTINTE

ESEMPIO


T
U
(universe of keys)
K
(actual
keys)
0
1
2
3
4
5
6
7
8
9
key
satellite data
2
3
5
8

IMPLEMENTAZIONE DELLE OPERAZIONI DI DIZIONARIO

COMPLESSITA'

```
DIRECT-ADDRESS-SEARCH(T, k)
    return T[k]
```

O(1)

```
DIRECT-ADDRESS-INSERT(T, x)
    T[x.key] = x
```

O(1)

```
DIRECT-ADDRESS-DELETE(T, x)
    T[x.key] = NIL
```

O(1)

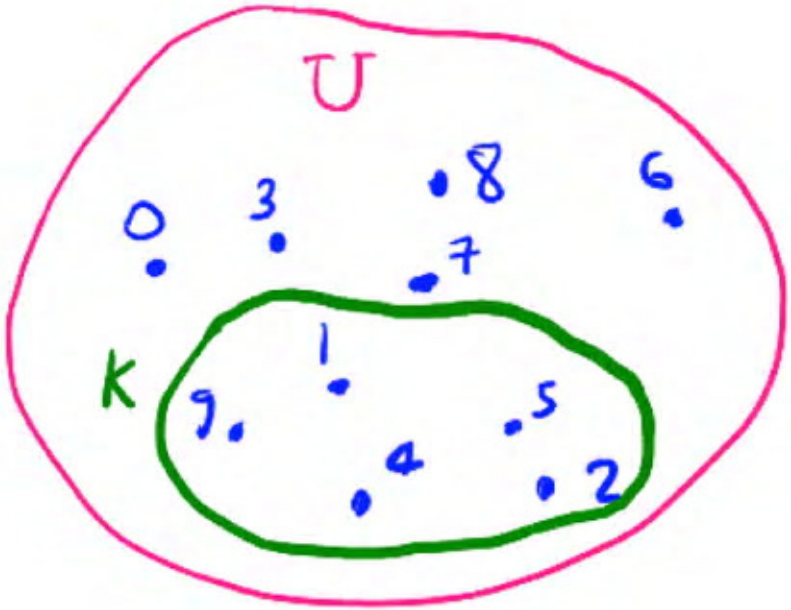
COMPLESSITA' SPAZIALE : $\Theta(m)$ (INDIPENDENTE DA $|A|$!!)

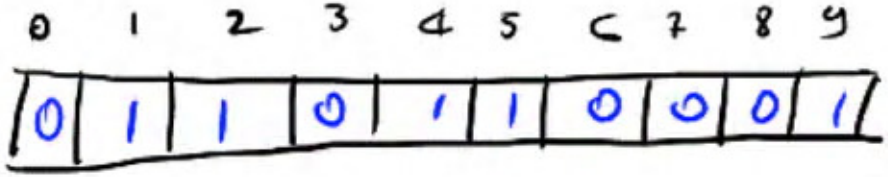
— DUNQUE $|U|$ NON DEVE ESSERE TROPPO GRANDE !

VETTORI DI BIT

- UN CASO SPECIALE SI HA NELLA RAPPRESENTAZIONE DI INSIEMI DI CHIAVI (SENZA DATI SATELLITI)
- IN TAL CASO SI POSSONO UTILIZZARE ARRAY $T[0..m-1]$ DI BIT

ESEMPIO

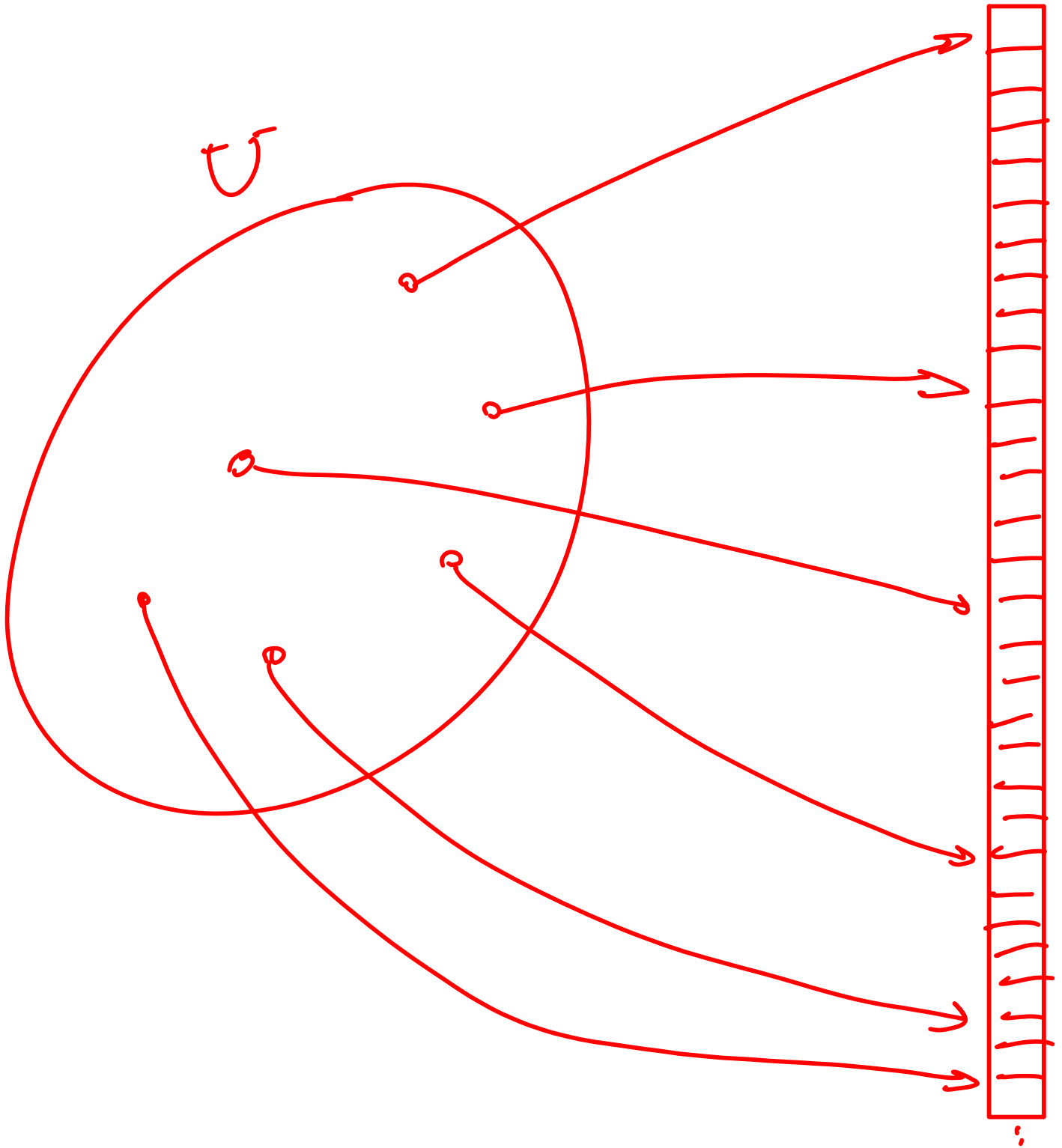


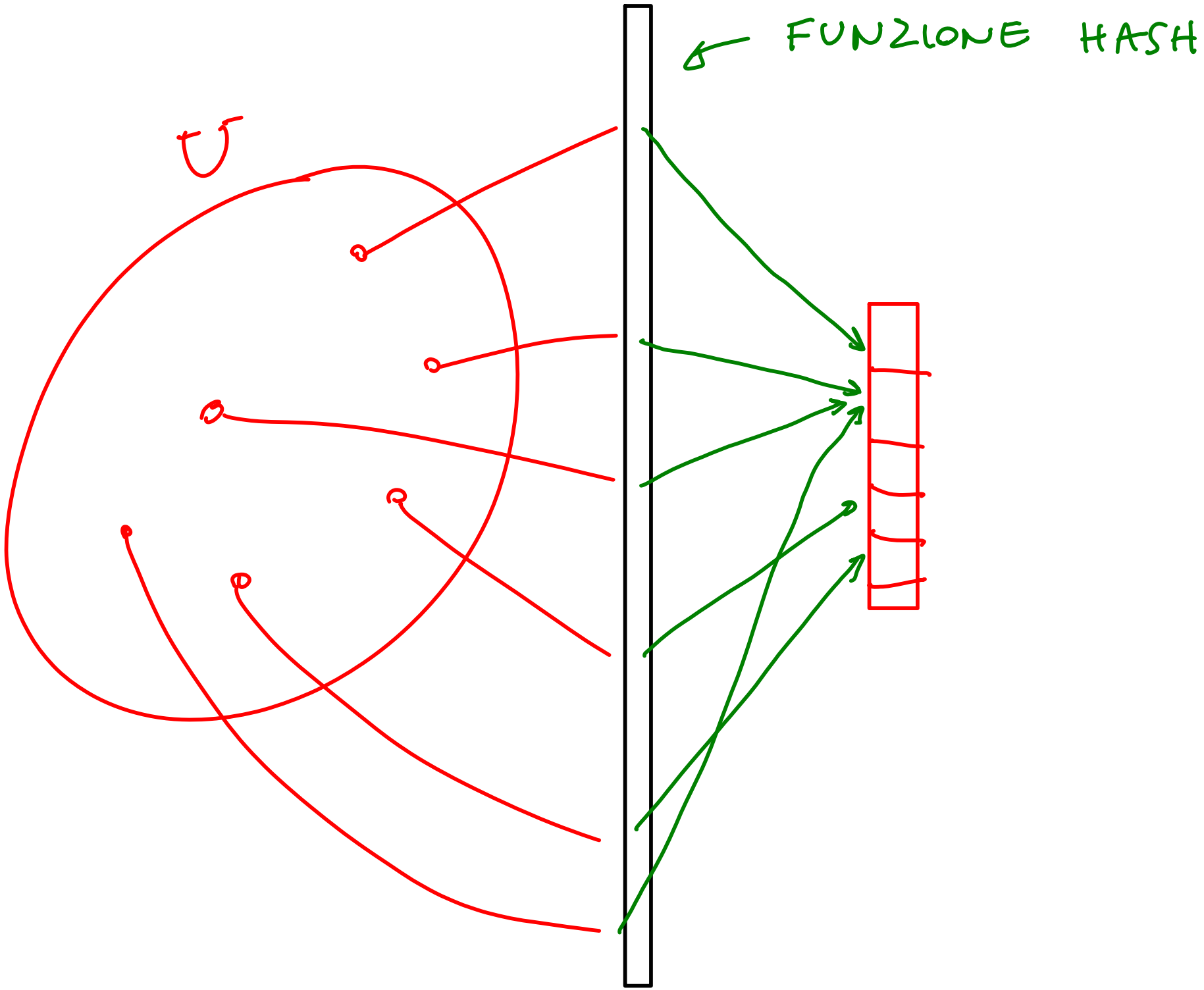


- TALE RAPPRESENTAZIONE SUPPORTA IN MANIERA ABBASTANZA EFFICIENTE ANCHE LE OPERAZIONI DI: -UNIONE, -INTERSEZIONE, -DIFFERENZA INSIEMISTICA

TABELLE HASH

- LE TABELLE HASH RISOLVONO IN PRATICA IL PROBLEMA DELLA RAPPRESENTAZIONE DI INSIEMI DINAMICI QUANDO
  - L'UNIVERSO U DELLE POSSIBILI CHIAVI È GRANDE (ANCHE INFINITO) E QUINDI RISULTA PROIBITIVO (SE NON IMPOSSIBILE) ALLOCARE UN ARRAY T DI $|U|$ COMPONENTI
  - LA DIMENSIONE DELL'INSIEME DA RAPPRESENTARE È PICCOLA

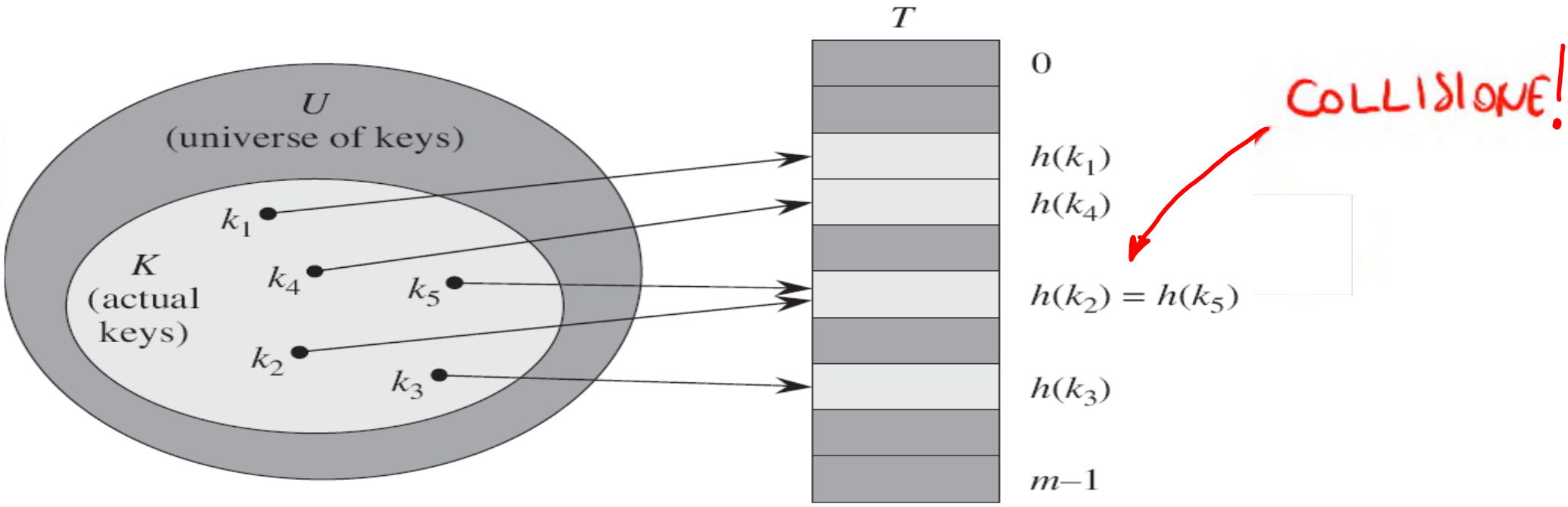

U


FUNZIONE HASH
U

- VIENE ALLOCATA UNA "TABELLA HASH" DI DIMENSIONE m CONFRONTABILE CON QUELLA DELL'INSIEME CHE SI INTENDE RAPPRESENTARE
- SI UTILIZZA UNA OPPORTUNA "FUNZIONE HASH"

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

- LA TABELLA HASH VIENE UTILIZZATA COME UNA TABELLA AD INDIRIZZAMENTO DIRETTO, FILTRANDO LE CHIAVI MEDIANTE LA FUNZIONE HASH

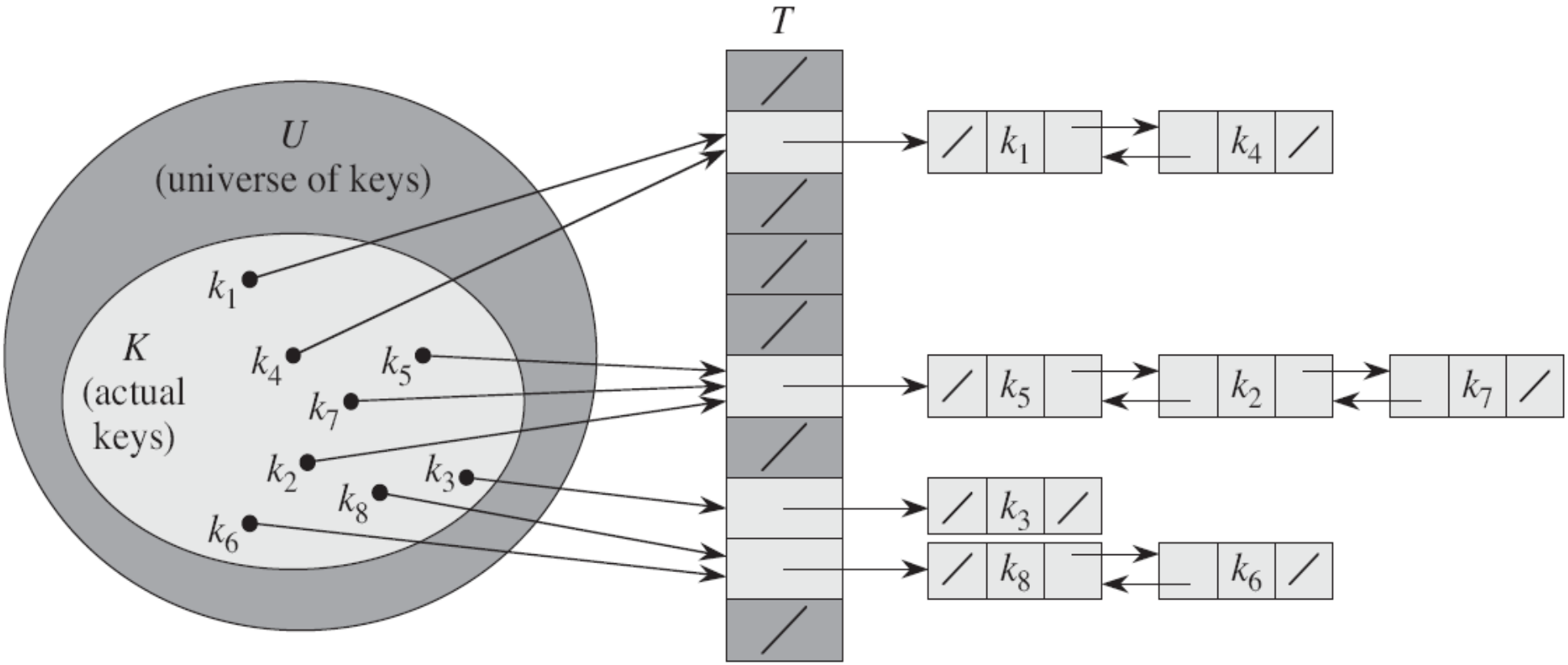


- SE $|U| > |T|$ E' INEVITABILE CHE SI POSSA VERIFICARE IL PROBLEMA DELLE COLLISIONI, CIOE' CHE VI SIANO CHIAVI $k_1 \neq k_2$ TALI CHE $h(k_1) = h(k_2)$

- VEDREMO DUE SOLUZIONI AL PROBLEMA DELLE COLLISIONI
  - TABELLE HASH CON CONCATENAMENTO
  - TABELLE HASH AD INDIRIZZAMENTO APERTO

TAVOLE HASH CON CONCATENAMENTO

- SI PONGONO TUTTI GLI ELEMENTI ASSOCIATI AD UNA STESSA CELLA IN UNA LISTA CONCATENATA (INSERIMENTO IN TESTA)



NOTA: LE LISTE DOPPIAMENTE CONCATENATE SEMPLIFICANO LE CANCELLAZIONI

OPERAZIONI

COMPLESSITÀ (CASO PESSIMO)

CHAINED-HASH-INSERT(T, x)
    insert x at the head of list $T[h(x.key)]$

$O(1)$ (SE E' NOTO CHE x NON SIA GIÀ PRESENTE)

CHAINED-HASH-SEARCH(T, k)
    search for an element with key k in list $T[h(k)]$

Θ(LUNGHEZZA LISTA)

CHAINED-HASH-DELETE(T, x)
    delete x from the list $T[h(x.key)]$

$O(1)$ (CON LISTE DOPPIE)

ANALISI PROBABILISTICA DELL'HASHING CON CONCATENAMENTO

- L'ANALISI SARÀ EFFETTUATA SOTTO LA SEGUENTE IPOTESI SULLA FUNZIONE HASH h

IPOTESI DI HASHING UNIFORME SEMPLICE

PER OGNI $i \in \{0, 1, \ldots, m-1\}$, $Pr\{h(x) = i\} = \frac{1}{m}$

- PER $j = 0, 1, \ldots, m-1$, SIA

$n_j =_{def}$ lunghezza lista $T[j]$

- SI HA: - $n = n_0 + n_1 + \ldots + n_{m-1}$

QUINDI: $n = E[n] = E[n_0] + E[n_1] + \ldots + E[n_{m-1}]$

E POICHE' $E[n_0] = E[n_1] = \ldots = E[n_{m-1}]$,

SI OTTIENE $E[n_j] = \frac{n}{m}$.

- CHIAMIAMO FATTORE DI CARICO DELLA TAVOLA T

IL RAPPORTO $\alpha = \frac{n}{m}$.

- FAREMO L'IPOTESI CHE $h(k)$ SI CALCOLI IN TEMPO $O(1)$

TEOREMA 1 IN UNA TAVOLA HASH CON CONCATENAMENTO, UNA RICERCA SENZA SUCCESSO RICHIEDE UN TEMPO $\Theta(1+\alpha)$ NEL CASO MEDIO, NELL'IPOTESI DI HASHING UNIFORME SEMPLICE.

DIM SIA k UNA CHIAVE NON IN T,
IL TEMPO ATTESO PER CERCARE k IN $T[h(k)]$ E'
PARI ALLA LUNGHEZZA ATTESA DI $T[h(k)]$,
CIOE' $E[n_{h(k)}] = \alpha$.
TENENDO ANCHE CONTO DEL TEMPO PER CALCOLARE $h(k)$,
SI OTTIENE UN TEMPO MEDIO COMPLESSIVO $\Theta(1+\alpha)$.

■

TEOREMA 2 IN UNA TAVOLA HASH CON CONCATENAMENTO, UNA RICERCA CON SUCCESSO RICHIEDE UN TEMPO $\Theta(1+\alpha)$ NEL CASO MEDIO, NELL'IPOTESI DI HASHING UNIFORME SEMPLICE.

DIM.

- SUPPONIAMO CHE TUTTE LE CHIAVI IN T ABBIANO LA STESSA PROBABILITÀ DI ESSERE RICERCATE
- SIANO $x_1, x_2, \dots, x_m$ GLI ELEMENTI NELL'ORDINE IN CUI SONO STATI INSERITI E SIA $k_i = x_i.key$, PER $i = 1, \dots, n$.
- PONIAMO $X_{ij} = I\{h(k_j) = h(k_i)\}$.
- SI HA: $E[X_{ij}] = Pr\{h(k_j) = h(k_i)\} = \frac{1}{m}$ $(j > i)$

(PER L'IPOTESI DI HASHING UNIFORME SEMPLICE)

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\left(1+\sum_{j=i+1}^{n}X_{ij}\right)\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(1+\sum_{j=i+1}^{n}E[X_{ij}]\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(1+\sum_{j=i+1}^{n}\frac{1}{m}\right) = 1+\frac{1}{nm}\sum_{i=1}^{n}(n-i)$$

$$= 1+\frac{1}{nm}\sum_{i=1}^{n-1} i$$

$$= 1+\frac{1}{nm}\cdot\frac{(n-1)n}{2} = 1+\frac{1}{m}\left(\frac{n}{2}-\frac{1}{2}\right)$$

$$= 1+\frac{\alpha}{2}-\frac{\alpha}{2n} = \textcircled{n}(1+\alpha)$$

■

INTERPRETAZIONE DELL'ANALISI DI COMPLESSITÀ

- SE $n = O(m)$, SI HA $\alpha = \frac{n}{m} = \frac{O(m)}{m} = O(1)$.

- QUINDI SE SI SCEGLIE m PROPORZIONALE AD n, LA RICERCA CON O SENZA SUCCESSO RICHIEDE IN MEDIA TEMPO $O(1)$.

ESERCIZI

11.2-1

Suppose we use a hash function h to hash n distinct keys into an array T of length m. Assuming simple uniform hashing, what is the expected number of collisions? More precisely, what is the expected cardinality of $\{\{k, l\} : k \neq l \text{ and } h(k) = h(l)\}$?

11.2-2

Demonstrate what happens when we insert the keys $5, 28, 19, 15, 20, 33, 12, 17, 10$ into a hash table with collisions resolved by chaining. Let the table have 9 slots, and let the hash function be $h(k) = k \bmod 9$.

11.2-3

Professor Marley hypothesizes that he can obtain substantial performance gains by modifying the chaining scheme to keep each list in sorted order. How does the professor's modification affect the running time for successful searches, unsuccessful searches, insertions, and deletions?

FUNZIONI HASH

- UNA BUONA FUNZIONE HASH DEVE SODDISFARE APPROSSIMATIVAMENTE L'IPOTESI DI HASHING UNIFORME SEMPLICE

- PER VERIFICARE TALE IPOTESI SAREBBE NECESSARIO CONOSCERE LA DISTRIBUZIONE DI PROBABILITA' DELLE CHIAVI

  INOLTRE LE ESTRAZIONI DELLE CHIAVI DOVREBBERO ESSERE INDIPENDENTI L'UNA DALL'ALTRA

ESEMPIO:

- SE LE CHIAVI SONO NUMERI CASUALI IN

$$U = \{ k \in \mathbb{R} : 0 \leq k < 1 \},$$

LA FUNZIONE HASH $h(k) = \lfloor k\,m \rfloor$ SODDISFA

L'IPOTESI DI HASHING UNIFORME SEMPLICE

- $m = 100$, $h(0.12576) = 12$

$h(0.576914) = 57$

$h(0.01147) = 1$

...

INTERPRETAZIONE DELLE CHIAVI COME NUMERI NATURALI

- NELLA MAGGIOR PARTE DELLE FUNZIONI HASH, VIENE ASSUNTO CHE $U = \mathbb{N} = \{0, 1, 2, \ldots\}$
- QUINDI, PER UTILIZZARE TALI FUNZIONI HASH OCCORRERA' MAPPARE U IN $\mathbb{N}$ QUALORA, $U \not\subseteq \mathbb{N}$

ESEMPIO

U = INSIEME DELLE STRINGHE FINITE DI CARATTERI ASCII A 7 BIT

$pt \longrightarrow (112, 116) \longrightarrow 112 \cdot 128 + 116 = 14452$

FUNZIONI HASH CON IL METODO DELLA DIVISIONE

- SIA m LA DIMENSIONE DELLA TABELLA HASH.
  SI PONE: $h(k) = k \bmod m$

- TALE METODO È MOLTO EFFICIENTE, MA OCCORRE AVERE CURA DI SCEGLIERE UN VALORE m CHE APPROSSIMI BENE L'IPOTESI DI HASHING UNIFORME SEMPLICE

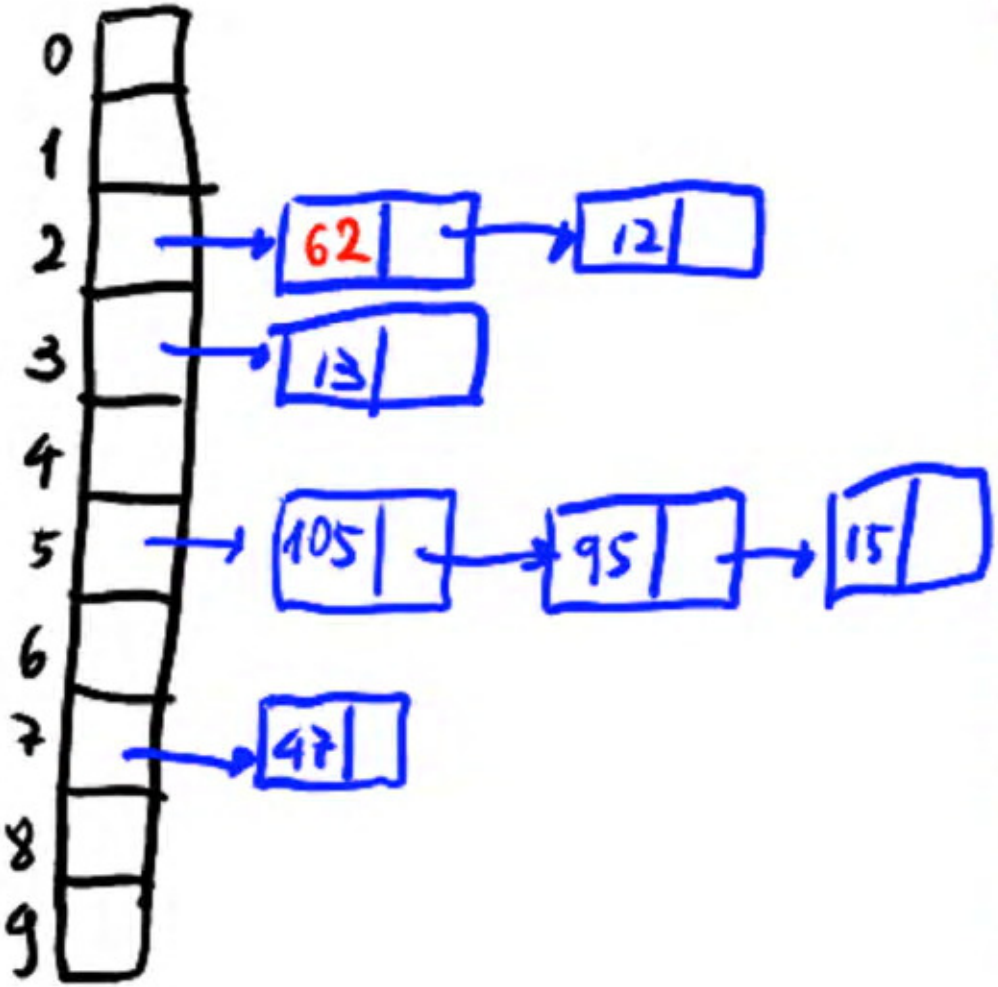
  ES. - SE $m = 2^p$, $h(k)$ DIPENDE DAI p BIT DI ORDINE INFERIORE

  - UNA BUONA SCELTA CONSISTE IN GENERE NEL SELEZIONARE PER m UN NUMERO PRIMO ABBASTANZA DISCOSTO DA POTENZE DI 2

  ES. $n = 2000 \longrightarrow m = 701 \longrightarrow \alpha \simeq 3$

ESEMPIO

$m = 10$, 47, 12, 15, 95, 62, 13 , 105



METODO DELLA DIVISIONE

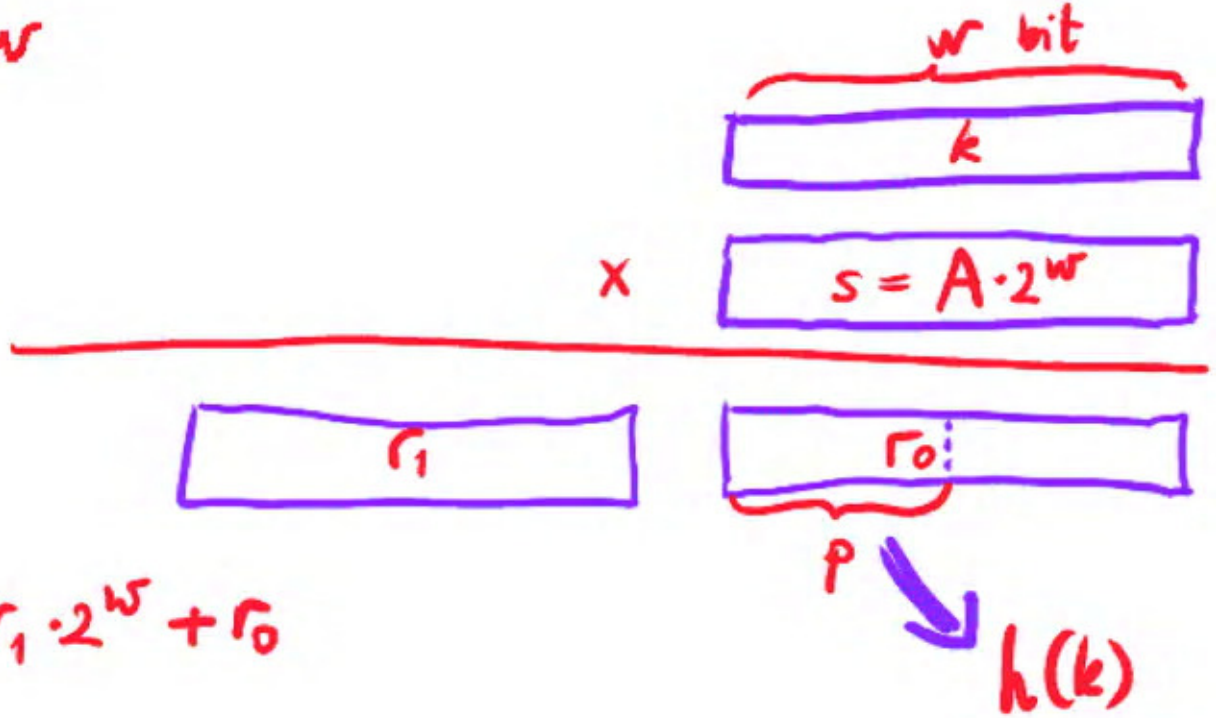
$h(x) = x \bmod 10$

$h(47) = 7$
$h(12) = 2$
$h(15) = 5$
$h(95) = 5$
$h(62) = 2$
$h(13) = 3$
$h(105) = 5$

FUNZIONI HASH CON IL METODO DELLA MOLTIPLICAZIONE

- SIA $0 < A < 1$ UNA COSTANTE FISSATA.
  SI PONE $h(k) = \lfloor m (kA \bmod 1) \rfloor$
  (DOVE $kA \bmod 1 = kA - \lfloor kA \rfloor$)
- LA SCELTA DI m NON E' CRITICA
- CONVIENE UTILIZZARE IL VALORE
  $A = (\sqrt{5} - 1)/2 \simeq 0.6180339887 \ldots$

- TIPICAMENTE SI SCEGLIE $m = 2^p$
- CIO' RENDE IL CALCOLO DI $h(k)$ PARTICOLARMENTE EFFICIENTE IN SITUAZIONI IN CUI, POSTO w LA DIMENSIONE DI UNA PAROLA, SI HA
  - $0 \leq k < 2^w$
  - $A = \frac{s}{2^w}$, CON $0 < s < 2^w$
  - $p < w$



$k \cdot s = r_1 \cdot 2^w + r_0$

ESEMPIO:

$k = 123456$

$p = 14$

$m = 2^{14} = 16384$

$w = 32$

$A = \frac{s}{2^{32}} = \frac{2654435761}{2^{32}} \simeq \frac{\sqrt{5}-1}{2}$

QUINDI:

$k \cdot s = 327706022297664 = (76300 \cdot 2^{32}) + 17612864$

$r_1 = 76300$

$r_0 = 17612864$

I 14 BIT PIÙ SIGNIFICATIVI DI r_0 FORNISCONO IL VALORE

$h(k) = 67$

HASHING UNIVERSALE

- AL FINE DI EVITARE CHE POSSANO ESISTERE INSIEMI DI CHIAVI CHE CAUSINO SEMPRE UN ALTO NUMERO DI COLLISIONI, E' STATO PROPOSTO LO SCHEMA DELL' HASHING UNIVERSALE, CHE PREVEDE CHE LA FUNZIONE HASH SIA SELEZIONATA IN MANIERA RANDOM DA UNA CERTA FAMIGLIA DI FUNZIONI HASH, CHE GODE DI OPPORTUNE PROPRIETA',

- SIA $\mathcal{H}$ UNA FAMIGLIA DI FUNZIONI HASH $h: U \to \{0..m-1\}$

DEFINIZIONE $\mathcal{H}$ SI DICE UNIVERSALE SE

$$(\forall x,y \in U)\left(x \neq y \longrightarrow |\{h \in \mathcal{H} : h(x) = h(y)\}| \leq \frac{|\mathcal{H}|}{m}\right) \quad \blacksquare$$

PROPRIETÀ SIA $\mathcal{H}$ UNIVERSALE E SIANO $x, y \in U$ TALI CHE $x \neq y$, ALLORA

$$Pr\{h \in \mathcal{H} : h(x) = h(y)\} = \frac{|\{h \in \mathcal{H} : h(x) = h(y)\}|}{|\mathcal{H}|} \leq \frac{1}{m}.$$

(CIOÈ, LA PROBABILITÀ DI AVERE UNA COLLISIONE SU DUE ELEMENTI x, y SELEZIONANDO h DA $\mathcal{H}$ È MINORE O UGUALE ALLA PROBABILITÀ DI OTTENERE UNA COLLISIONE SELEZIONANDO IN MANIERA RANDOM I DUE VALORI HASH SU x E y) $\blacksquare$

TEOREMA SUPPONIAMO CHE UNA FUNZIONE HASH h SIA SCELTA A CASO DA UNA FAMIGLIA UNIVERSALE $\mathcal{H}$ DI FUNZIONI HASH E CHE SIA UTILIZZATA PER INSERIRE n ELEMENTI IN UNA TAVOLA HASH T DI DIMENSIONE m (INIZIALMENTE VUOTA) OVE LE COLLISIONI SONO RISOLTE MEDIANTE CONCATENAMENTO.

SIA $n_i = |T[i]|$, PER $i = 0, 1, \ldots, m-1$.

ALLORA

$$E[n_{h(k)}] \leq \alpha = \frac{n}{m}, \quad \text{SE } k \notin T$$

$$E[n_{h(k)}] < 1 + \alpha, \quad \text{SE } k \in T$$

DIM. - PER OGNI $k, \ell \in U$, CON $k \neq \ell$, DEFINIAMO

$$X_{k\ell} = \mathbb{I}\{h(k) = h(\ell)\}.$$

SI HA $\quad E[X_{k\ell}] = Pr\{h(k) = h(\ell)\} \leq \frac{1}{m}$.

- PER OGNI $k \in U$, DEFINIAMO

$$Y_k = |\{\ell \in U : \ell \in T, h(\ell) = h(k), \ell \neq k\}|$$

SI HA $\quad Y_k = \sum_{\substack{\ell \in T \\ \ell \neq k}} X_{k\ell}$, $\quad$ E QUINDI:

$$E[Y_k] = E\left[\sum_{\substack{\ell \in T \\ \ell \neq k}} X_{k\ell}\right] = \sum_{\substack{\ell \in T \\ \ell \neq k}} E[X_{k\ell}] \leq \sum_{\substack{\ell \in T \\ \ell \neq k}} \frac{1}{m}$$

$$= \frac{1}{m} \cdot |\{\ell : \ell \in T, \ell \neq k\}|$$

$$E[Y_k] \leq \frac{1}{m} \cdot |\{l : l \in T, l \neq k\}| .$$

CASO : $k \notin T$

$$|\{l : l \in T, l \neq k\}| = |T| = n \quad ; \quad n_{h(k)} = Y_k$$

$$\Rightarrow E[Y_k] \leq \frac{1}{m} \cdot |\{l : l \in T, l \neq k\}| = \frac{n}{m} = \alpha$$

$$E[n_{h(k)}] = E[Y_k] \leq \alpha$$

CASO : $k \in T$

$$|\{l : l \in T, l \neq k\}| = |T| - 1 = n - 1 \quad ; \quad n_{h(k)} = Y_k + 1$$

$$\Rightarrow E[Y_k] \leq \frac{1}{m} \cdot |\{l : l \in T, l \neq k\}| = \frac{n-1}{m} < \alpha$$

$$E[n_{h(k)}] = E[Y_k] + 1 < 1 + \alpha$$

■

COROLLARIO UTILIZZANDO HASHING UNIVERSALE CON CONCATENAMENTO PER ESEGUIRE UNA SEQUENZA DI N OPERAZIONI INSERT, SEARCH E DELETE CONTENENTE $O(m)$ OPERAZIONI INSERT, DOVE m E' LA DIMENSIONE DI UNA TAVOLA INIZIALMENTE VUOTA, OCCORRE UN TEMPO MEDIO $\Theta(N)$.

DIM. SIA n LA CARDINALITA' MASSIMA DELLA TAVOLA T. ALLORA $n = O(m)$ E DUNQUE $\alpha \leq \frac{n}{m} = O(1)$ DURANTE L'ESECUZIONE DELLE N OPERAZIONI.

- CIASCUNA OPERAZIONE INSERT E DELETE RICHIEDE TEMPO $O(1)$ (CASO PEGGIORE)
- PER IL TEOREMA PRECEDENTE (E POICHE' $\alpha = O(1)$), CIASCUNA OPERAZIONE SEARCH RICHIEDE TEMPO ATTESO $O(1)$.
- QUINDI IL TEMPO ATTESO PER LE N OPERAZIONI E' $O(N)$.
- OVVIAMENTE, TALE TEMPO ATTESO E' ANCHE $\Omega(N)$. ■

COSTRUZIONE DI UNA CLASSE UNIVERSALE DI FUNZIONI HASH

- SIA p PRIMO TALE CHE $U \subseteq \mathbb{Z}_p$, CON

$$\mathbb{Z}_p = \{0, 1, \dots, p-1\}.$$

- PONIAMO $\mathbb{Z}_p^* = \{1, 2, \dots, p-1\}$

- OVVIAMENTE AVREMO ANCHE $p > m$, DOVE m È LA DIMENSIONE DELLA TAVOLA T.

- PER OGNI $a \in \mathbb{Z}_p^*$ E $b \in \mathbb{Z}_p$, PONIAMO

$$h_{ab}(k) = ((ak + b) \bmod p) \bmod m \qquad (h_{ab} : \mathbb{Z}_p \to \mathbb{Z}_m)$$

ES. PER $p = 17$, $m = 6$, $a = 3$, $b = 4$, SI HA

$$h_{3,4}(k) = ((3k + 4) \bmod 17) \bmod 6$$

E QUINDI

$$h_{3,4}(8) = ((3 \cdot 8 + 4) \bmod 17) \bmod 6 = (28 \bmod 17) \bmod 6$$
$$= 11 \bmod 6 = 5.$$

PONIAMO $\mathcal{H}_{pm} = \{ h_{ab} : a \in \mathbb{Z}_p^*, b \in \mathbb{Z}_p \}$.

$$h_{ab}(k) = ((ak + b) \bmod p) \bmod m \qquad (h_{ab} : \mathbb{Z}_p \to \mathbb{Z}_m)$$

NOTA: NON E' NECESSARIA ALCUNA IPOTESI PARTICOLARE SU m (TRANNE CHE $m < p$)

TEOREMA LA CLASSE $\mathcal{H}_{pm}$, CON $m < p$, E' UNIVERSALE.

TEOREMA LA CLASSE $\mathcal{H}_{pm}$, CON $m < p$, E' UNIVERSALE.

DIM - DATI $k, l \in \mathbb{Z}_p$ $(k \neq l)$

OBIETTIVO: $Pr\{h_{ab}(k) = h_{ab}(l)\} \leq \frac{1}{m}$

- SIA $h_{ab} \in \mathcal{H}_{pm}$

$$h_{ab}(k) = \underbrace{((ak+b) \bmod p)}_{r} \bmod m$$
$$h_{ab}(l) = \underbrace{((al+b) \bmod p)}_{s} \bmod m$$
$$\Rightarrow \quad r \neq s .$$

INFATTI: $r = s \Rightarrow ak + b \equiv al + b \pmod p$

$\Rightarrow ak \equiv al \pmod p$

$\Rightarrow a(k-l) \equiv 0 \pmod p$

$\Rightarrow k - l \equiv 0 \pmod p \Rightarrow k = l$

FISSATI $k, l \in \mathbb{Z}_p \quad (k \neq l)$

$a \in \mathbb{Z}_p^*, \; b \in \mathbb{Z}_p$

$$(a,b) \in \mathbb{Z}_p^* \times \mathbb{Z}_p \xmapsto{\alpha} (r,s) \in (\mathbb{Z}_p \times \mathbb{Z}_p) \setminus \Delta_p$$

(CON $\Delta_p = \{(t,t) : t \in \mathbb{Z}_p\}$)

VICEVERSA

$$(r,s) \in (\mathbb{Z}_p \times \mathbb{Z}_p) \setminus \Delta_p \xmapsto{\beta} (a,b) \in \mathbb{Z}_p^* \times \mathbb{Z}_p$$

$$a = \left((r-s)\left((k-l)^{-1} \bmod p\right)\right) \bmod p$$

$$b = (r - ak) \bmod p$$

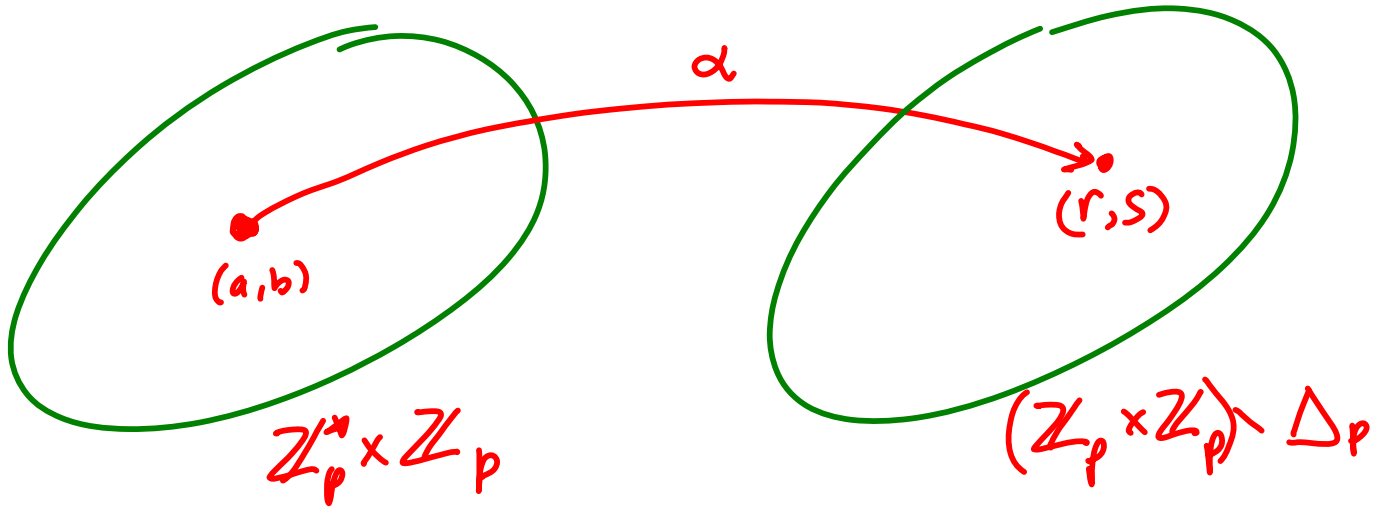
CON $\beta(\alpha(a,b)) = (a,b)$

QUINDI α E' UNA CORRISPONDENZA BIUNIVOCA TRA

$\mathbb{Z}_p^* \times \mathbb{Z}_p$ E $(\mathbb{Z}_p \times \mathbb{Z}_p) \setminus \Delta_p$

(ENTRAMBI DI CARDINALITA' $p(p-1)$)

- SE (a,b) VARIA IN $\mathbb{Z}_p^* \times \mathbb{Z}_p$ CON PROBABILITA' UNIFORME,
  ALTRETTANTO FARA' $(r,s) = \alpha(a,b)$ IN $(\mathbb{Z}_p \times \mathbb{Z}_p) \setminus \Delta_p$.



- QUINDI $Pr\{h_{ab}(k) = h_{ab}(\ell)\} = Pr\{r \equiv s \pmod m\}$

CALCOLO DI $Pr\{r \equiv s \pmod m\}$

- FISSATO r, VI SONO AL PIÙ $\left\lceil \frac{p}{m} \right\rceil - 1$ VALORI DI $s \neq r$ TALI CHE $r \equiv s \pmod m$.

$$\left\lceil \frac{p}{m} \right\rceil - 1 \leq \frac{p+m-1}{m} - 1 = \frac{p-1}{m},$$

- QUINDI $Pr\{r \equiv s \pmod m\} \leq \frac{p-1}{m} \cdot \frac{1}{p-1} = \frac{1}{m}$,

CIOÈ $Pr\{h_{ab}(k) = h_{ab}(l)\} \leq \frac{1}{m}$.

- PERTANTO $\mathcal{H}_{pm}$ È UNA FAMIGLIA UNIVERSALE DI FUNZIONI HASH.

■

TABELLE HASH AD INDIRIZZAMENTO APERTO

- NELLE TABELLE HASH CON CONCATENAMENTO PARTE DELLA MEMORIA E' IMPEGNATA CON PUNTATORI
- SI PUÒ EVITARE DI UTILIZZARE PUNTATORI MANTENENDO TUTTI I DATI ALL'INTERNO DELLA STESSA TABELLA, UTILIZZANDO LO SCHEMA DELL' INDIRIZZAMENTO APERTO
- IN TAL CASO SI AVRÀ $\alpha \leq 1$
- L'INSERIMENTO DI UN NUOVO ELEMENTO UTILIZZERÀ UNA SEQUENZA DI SCANSIONE DIPENDENTE DAL VALORE DELLA CHIAVE k
- LA MEDESIMA SEQUENZA DI SCANSIONE DOVRÀ ESSERE UTILIZZATA NELLA RICERCA

- PER GENERARE LE SEQUENZE DI SCANSIONE, SARANNO UTILIZZATE FUNZIONI HASH DEL TIPO:

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

(m È LA DIMENSIONE DELLA TABELLA)

- DATA LA CHIAVE k, SARÀ UTILIZZATA LA SEGUENTE SEQUENZA:

$$\langle h(k,0), h(k,1), \ldots, h(k,m-1) \rangle$$

- PER AUMENTARE L'EFFICIENZA DELLE TABELLE HASH AD INDIRIZZAMENTO APERTO È IMPORTANTE CHE LE SEQUENZE DI SCANSIONE SIANO PERMUTAZIONI DI $\langle 0, 1, \ldots, m-1 \rangle$

```
HASH-INSERT(T, k)
1  i = 0
2  repeat
3      j = h(k, i)
4      if T[j] == NIL
5          T[j] = k
6          return j
7      else i = i + 1
8  until i == m
9  error "hash table overflow"
```

```
HASH-SEARCH(T, k)
1  i = 0
2  repeat
3      j = h(k, i)
4      if T[j] == k
5          return j
6      i = i + 1
7  until T[j] == NIL or i == m
8  return NIL
```

NOTA: TALE SCHEMA NON SUPPORTA LA CANCELLAZIONE IN MANIERA IMMEDIATA

INDIRIZZAMENTO APERTO E CANCELLAZIONE

- PER NON INTERROMPERE LE SEQUENZE DI SCANSIONE IN CASO DI CANCELLAZIONE, SI POSSONO MARCARE COME DELETED LE CELLE DA CUI SONO STATE CANCELLATE LE CHIAVI

- L'ANALISI DI COMPLESSITA' DOVRA' TENERE CONTO DELLE CELLE MARCATE DELETED PER IL CALCOLO CORRETTO DEL FATTORE DI CARICO α,

```
HASH-INSERT'(T, k)
1  i = 0
2  repeat
3      j = h(k, i)
4      if T[j] == NIL   or T[j] == 'DELETED'
5          T[j] = k
6          return j
7      else i = i + 1
8  until i == m
9  error "hash table overflow"
```

- NELLA NOSTRA ANALISI DI COMPLESSITÀ FAREMO L'IPOTESI DI HASHING UNIFORME:

PER OGNI PERMUTAZIONE π DI $\langle 0, 1, \ldots, m-1 \rangle$ SI HA

$$Pr\{\langle h(k,0), h(k,1), \ldots, h(k,m-1)\rangle = \pi\} = \frac{1}{m!}$$

(CIOÈ TUTTE LE PERMUTAZIONI SONO EQUIPROBABILI)

- IN PRATICA, PERÒ, L'IPOTESI DI HASHING UNIFORME SARÀ SOLTANTO APPROSSIMATA

FUNZIONI HASH PER L'INDIRIZZAMENTO APERTO NELLA PRATICA

- ISPEZIONE LINEARE
- ISPEZIONE QUADRATICA
- DOPPIO HASHING

ISPEZIONE LINEARE

- SIA $h' : U \to \{0, 1, \ldots, m-1\}$ UNA FUNZIONE HASH AUSILIARIA
- PONIAMO $h(k, i) = (h'(k) + i) \bmod m$

VARIANTE:

$$h(k, i) = (h'(k) + c \cdot i) \bmod m$$

CON $1 \leq c < m$ COSTANTE PRIMA CON m

ESEMPIO:

CHIAVI: 10, 22, 31, 4, 15, 28, 17, 88, 59 $\quad m = 11$

$h'(k) = k \bmod 11$

$h(k,i) = ((k \bmod 11) + i) \bmod 11 = (k+i) \bmod 11$

0	1	2	3	4	5	6	7	8	9	10
22	88			4	15	28	17	59	31	10

$h(10,0) = (10+0) \bmod 11 = 10$

$h(22,0) = (22+0) \bmod 11 = 0$

$h(31,0) = (31+0) \bmod 11 = 9$

$h(4,0) = (4+0) \bmod 11 = 4$

$h(15,0) = (15+0) \bmod 11 = 4, \quad h(15,1) = 5$

$h(28,0) = (28+0) \bmod 11 = 6$

$h(17,0) = (17+0) \bmod 11 = 6, \quad h(17,1) = 7$

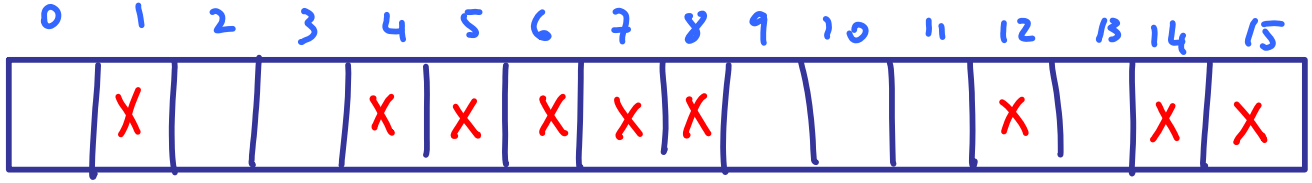
$h(88,0) = (88+0) \bmod 11 = 0, \quad h(88,1) = 1$

$h(59,0) = (59+0) \bmod 11 = 4, \; h(59,1) = 5, \; h(59,2) = 6, \; h(59,3) = 7, \; h(59,4) = 8$

- # SEQUENZE DI SCANSIONE = m << $m!$

- LA SCANSIONE LINEARE SOFFRE DEL PROBLEMA DELL'AGGLOMERAZIONE PRIMARIA: SI TENDONO A FORMARE LUNGHE SEQUENZE DI CELLE OCCUPATE CHE RALLENTANO LE OPERAZIONI DI RICERCA E DI INSERIMENTO

- INFATTI, LA PROBABILITA' CHE VENGA OCCUPATA UNA CELLA PRECEDUTA DA i CELLE OCCUPATE E' $\frac{i+1}{m}$

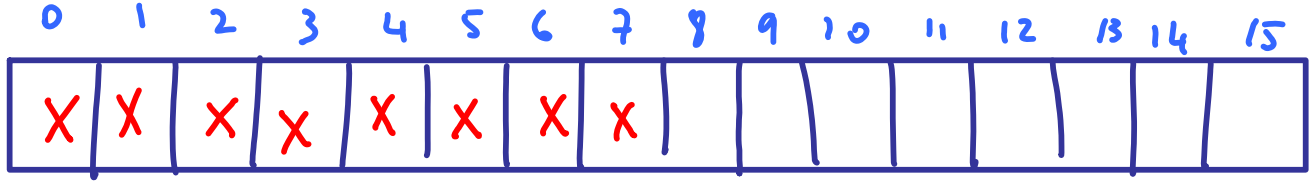


PROBABILITA' $\frac{3}{16}$ $\frac{2}{16}$ $\frac{1}{16}$ $\frac{6}{16}$ $\frac{1}{16}$ $\frac{1}{16}$ $\frac{2}{16}$

ESEMPIO (CASO PESSIMO)

m PARI, $n = \frac{m}{2}$, $\alpha = \frac{1}{2}$

UNICA AGGLOMERAZIONE DI DIMENSIONE n



ACCESSI $\frac{m}{2}+1$ $\frac{m}{2}$ $\frac{m}{2}-1$ ... ... 2 1 1 1 .... 1 1

MEDIO DI ACCESSI IN UNA RICERCA SENZA SUCCESSO

$$= \sum_{i=0}^{m-1} \#\,ACC[h'(k)=i] \cdot Pr\{h'(k)=i\}$$

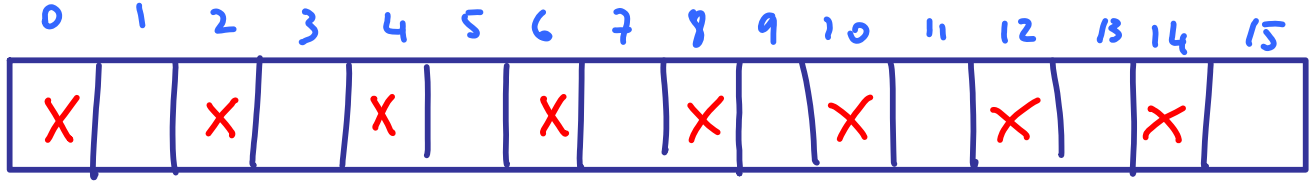
$$= \left(\frac{m}{2}+1\right)\frac{1}{m} + \frac{m}{2}\cdot\frac{1}{m} + \ldots + 2\cdot\frac{1}{m} + \frac{m}{2}\cdot\frac{1}{m}$$

$$= \frac{1}{m}\,\Theta(m^2) = \Theta(m)$$

ESEMPIO (CASO MIGLIORE)

m PARI, $n = \frac{m}{2}$, $\alpha = \frac{1}{2}$

NESSUNA AGGLOMERAZIONE

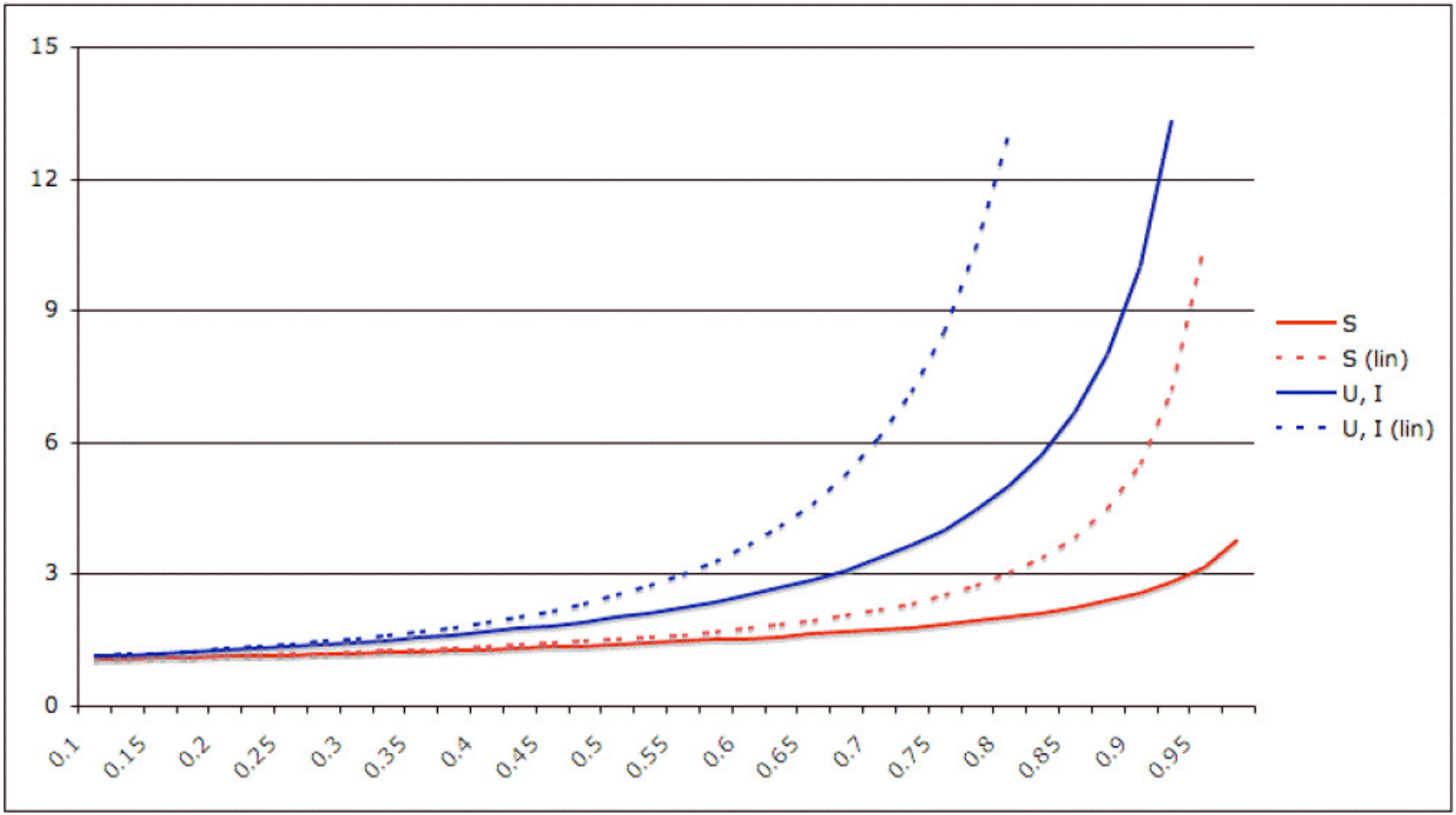


ACCESSI 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1

MEDIO DI ACCESSI IN UNA RICERCA SENZA SUCCESSO

$$= \sum_{i=0}^{m-1} \#\,ACC[h'(k) = i] \cdot Pr\{h'(k) = i\}$$

$$= \frac{m}{2}\left(2 \cdot \frac{1}{m}\right) + \frac{m}{2} \cdot \frac{1}{m} = \frac{3}{2}$$



RICERCA SENZA SUCCESSO $\frac{1}{2}\left(1+\frac{1}{(1-\alpha)^2}\right)^2$

RICERCA CON SUCCESSO $\frac{1}{2}\left(1+\frac{1}{1-\alpha}\right)$

ISPEZIONE QUADRATICA

- SIA $h' : U \longrightarrow \{0, 1, \ldots, m-1\}$ UNA FUNZIONE HASH AUSILIARIA
- PONIAMO $h(k, i) = (h'(k) + c_1 i + c_2 i^2) \bmod m$

  DOVE c_1, c_2 SONO COSTANTI CON $c_2 \neq 0$
- PERCHE' LE SEQUENZE DI SCANSIONE RISULTINO PERMUTAZIONI DI $\langle 0, 1, \ldots, m-1 \rangle$, I VALORI DI c_1, c_2 ED m DEBBONO ESSERE SCELTI CON ACCURATEZZA
- # SEQUENZE DI SCANSIONE $= m \ll m!$
- LA SCANSIONE QUADRATICA SOFFRE DEL PROBLEMA DELL'AGGLOMERAZIONE SECONDARIA, UNA FORMA PIÙ LIEVE DELL'AGGLOMERAZIONE PRIMARIA

COME SCEGLIERE m, c_1 E c_2

PONIAMO: $c_1 = c_2 = \frac{1}{2}$, $m = 2^r$

$$h(k,i) = \left(h'(k) + \frac{1}{2}i + \frac{1}{2}i^2\right) \bmod 2^r$$

VERIFICHIAMO CHE SE $i \neq j$, CON $0 \leq i, j < 2^r$, ALLORA

$$h(k,i) \neq h(k,j).$$

SE PER ASSURDO $h(k,i) = h(k,j)$, ALLORA

$$\left(h'(k) + \frac{1}{2}i + \frac{1}{2}i^2\right) \equiv \left(h'(k) + \frac{1}{2}j + \frac{1}{2}j^2\right) \pmod{2^r}$$

$$\frac{1}{2}(i-j) + \frac{1}{2}(i^2-j^2) \equiv 0 \pmod{2^r}$$

$$\frac{1}{2}(i-j)(i+j+1) \equiv 0 \pmod{2^r}$$

$$2^r \mid \frac{1}{2}(i-j)(i+j+1) \implies 2^{r+1} \mid (i-j)(i+j+1)$$

- POICHÉ $2^{r+1} \nmid i+j+1$ (IN QUANTO $1 \le i+j+1 < 2^{r+1}$)

E $i+j+1$, $i-j$ HANNO PARITÀ DIVERSA, SI HA

$$2^{r+1} \mid i-j$$

- MA $0 \le i \le 2^r - 1$

$0 \le j \le 2^r - 1$

$\Rightarrow \quad -(2^r-1) \le i-j \le 2^r-1$

$\Rightarrow \quad |i-j| \le 2^r-1 < 2^{r+1}$

$\Rightarrow \quad |i-j| = 0 \implies i = j$, ASSURDO!

QUINDI: $i \neq j \implies h(k,i) \neq h(k,j)$ (PER OGNI k)

$\Rightarrow \langle h(k,0), h(k,1), \dots, h(k,2^r-1) \rangle$ È UNA PERMUTAZIONE DI $\langle 0, 1, \dots, 2^r-1 \rangle$

ESEMPIO:

CHIAVI: 10, 22, 31, 4, 15, 28, 17, 88, 59 $\qquad m = 16$

$$h'(k) = k \bmod 16$$

$$h(k,i) = \left(\left(k \bmod 16 \right) + \frac{1}{2} i + \frac{1}{2} i^2 \right) \bmod 16$$

$$= \left(k + \frac{1}{2} i + \frac{1}{2} i^2 \right) \bmod 16$$

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
15	17			4		22		88		10	59	28			31

$h(10,0) = 10 \bmod 16 = 10$

$h(22,0) = 22 \bmod 16 = 6$

$h(31,0) = 31 \bmod 16 = 15$

$h(4,0) = 4 \bmod 16 = 4$

$h(15,0) = 15 \bmod 16 = 15$, $h(15,1) = \left(15 + \frac{1}{2} + \frac{1}{2}\right) \bmod 16 = 0$

$h(28,0) = 28 \bmod 16 = 12$

$h(17,0) = 17 \bmod 16 = 1$

$h(88,0) = 88 \bmod 16 = 8$

$h(59,0) = 59 \bmod 16 = 11$

DOPPIO HASHING

- SIANO h_1, h_2 DUE FUNZIONI HASH AUSILIARIE.

  PONIAMO $h(k,i) = (h_1(k) + i\, h_2(k)) \bmod m$

- PERCHE' LE SEQUENZE DI SCANSIONE SIANO PERMUTAZIONI DI $\langle 0, 1, \dots, m-1 \rangle$ E' NECESSARIO E SUFFICIENTE CHE $h_2(k)$ SIA PRIMO CON m, PER OGNI $k \in U$.

- UNA POSSIBILE SCELTA E'

  $m = 2^r$

  $h_2 : U \longrightarrow \{1, 3, 5, \dots, m-1\}$

- IN QUESTO CASO, IL NUMERO DI SEQUENZE DI SCANSIONE E' $m \cdot \frac{m}{2} = \Theta(m^2)$.

- UN'ALTRA POSSIBILE SCELTA E':
  - m PRIMO
  - $h_2 : U \to \{1, 2, \ldots, m-1\}$

ESEMPIO

$m = 701$

$h_1(k) = k \bmod 701$

$h_2(k) = 1 + (k \bmod 700)$

PER $k = 123456$ SI HA QUINDI:

$h_1(k) = 123456 \bmod 701 = 80$

$h_2(k) = 1 + (123456 \bmod 700) = 257$

$h(123456, i) = (80 + 257\,i) \bmod 701$

- ANCHE IN QUESTO SECONDO CASO IL NUMERO DI SEQUENZE DI SCANSIONE E' $\Omega(m^2)$.

- L'IPOTESI DI HASHING UNIFORME E' MEGLIO APPROSSIMATA CON IL DOPPIO HASHING

ESEMPIO:

CHIAVI: 10, 22, 31, 4, 15, 28, 17, 88, 59 $\quad m = 11$

$h_1(k) = k \bmod 11$

$h_2(k) = 1 + (k \bmod 10)$

$h(k,i) = \big((k \bmod 11) + (1 + k \bmod 10) \cdot i\big) \bmod 11$

0	1	2	3	4	5	6	7	8	9	10
22		59	17	4	15	28	88		31	10

$h(10,0) = 10 \bmod 11 = 10$

$h(22,0) = 22 \bmod 11 = 0$

$h(31,0) = 31 \bmod 11 = 9$

$h(4,0) = 4 \bmod 11 = 4$

$h(15,0) = 15 \bmod 11 = 4, \quad h(15,1) = (4+6) \bmod 11 = 10, \quad h(15,2) = (4+2\cdot 6) \bmod 11 = 5$

$h(28,0) = 28 \bmod 11 = 6$

$h(17,0) = 17 \bmod 11 = 6, \quad h(17,1) = (6+8) \bmod 11 = 3$

$h(88,0) = 88 \bmod 11 = 0, \quad h(88,1) = (0+9) \bmod 11 = 9, \quad h(88,2) = (0+18) \bmod 11 = 7$

$h(59,0) = 59 \bmod 11 = 4$

$h(59,1) = (4+10) \bmod 11 = 3, \quad h(59,2) = (4+20) \bmod 11 = 2$

ANALISI DELL'HASHING A INDIRIZZAMENTO APERTO

- EFFETTUEREMO UN'ANALISI PROBABILISTICA ASSUMENDO L'IPOTESI DI HASHING UNIFORME:

  PER OGNI PERMUTAZIONE π DI $\langle 0, 1, \ldots, m-1 \rangle$ SI HA

$$Pr\{\langle h(k,0), h(k,1), \ldots, h(k,m-1)\rangle = \pi\} = \frac{1}{m!}$$

- L'ANALISI SARÀ ESPRESSA IN TERMINI DEL FATTORE DI CARICO $\alpha = \frac{n}{m}$

TEOREMA NELL'IPOTESI DI HASHING UNIFORME, IL NUMERO ATTESO DI ISPEZIONI IN UNA RICERCA SENZA SUCCESSO IN UNA TAVOLA HASH A INDIRIZZAMENTO APERTO CON FATTORE DI CARICO $\alpha = \frac{n}{m} \leq 1$ E' AL PIÙ $\frac{1}{1-\alpha}$.

DIM SIANO

X = # ISPEZIONI IN UNA RICERCA SENZA SUCCESSO

A_i = EVENTO IN CUI L'i-ESIMA ISPEZIONE VIENE ESEGUITA E TROVA UNA CELLA OCCUPATA

SI HA: $\{X \geq i\} = A_1 \cap A_2 \cap \dots \cap A_{i-1}$

$$Pr\{X \geq i\} = Pr\{A_1 \cap A_2 \cap \dots \cap A_{i-1}\}$$

$$= \frac{n}{m} \cdot \frac{n-1}{m-1} \cdot \dots \cdot \frac{n-i+2}{m-i+2} \leq \left(\frac{n}{m}\right)^{i-1} = \alpha^{i-1}$$

$$E[X] = \sum_{i=1}^{\infty} Pr\{X \geq i\} \leq \sum_{i=1}^{\infty} \alpha^{i-1} = \sum_{i=0}^{\infty} \alpha^{i} = \frac{1}{1-\alpha}$$

- QUINDI, SE α E' UNA COSTANTE, UNA RICERCA SENZA SUCCESSO VIENE ESEGUITA NEL TEMPO ATTESO $O\left(\frac{1}{1-\alpha}\right)$

ES. $\alpha = \frac{1}{2} \Rightarrow \frac{1}{1-\alpha} = \frac{1}{1-0.5} = 2$

$\alpha = 90\% \Rightarrow \frac{1}{1-\alpha} = \frac{1}{1-0.9} = 10$

COROLLARIO

L'INSERIMENTO DI UN ELEMENTO IN UNA TAVOLA HASH A INDIRIZZAMENTO APERTO CON FATTORE DI CARICO α RICHIEDE IN MEDIA NON PIÙ DI $\frac{1}{1-\alpha}$ ISPEZIONI, NELL'IPOTESI DI HASHING UNIFORME.

■

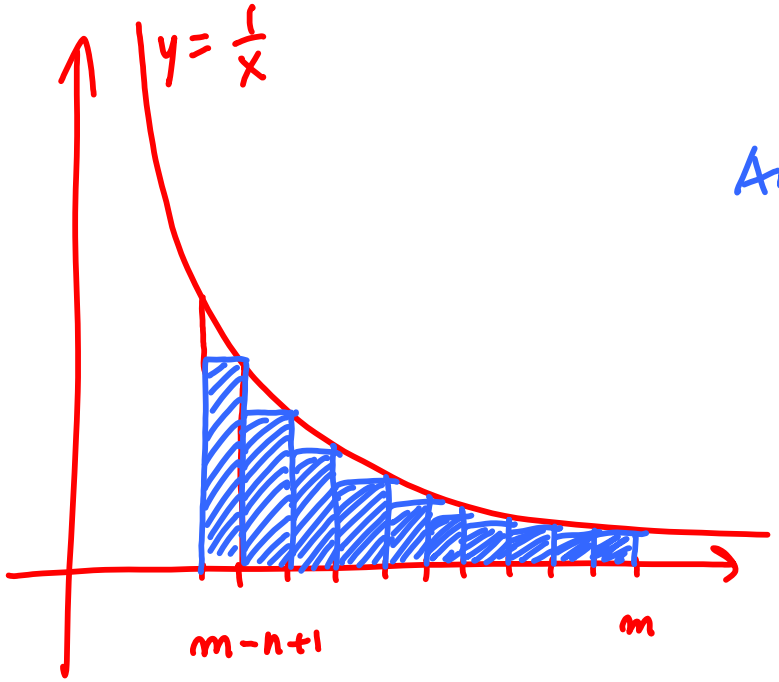
TEOREMA NELL'IPOTESI DI HASHING UNIFORME, IL NUMERO ATTESO DI ISPEZIONI IN UNA RICERCA CON SUCCESSO IN UNA TAVOLA HASH A INDIRIZZAMENTO APERTO CON FATTORE DI CARICO $\alpha = \frac{n}{m} \leq 1$ E' AL PIÙ $\frac{1}{\alpha} \ln \frac{1}{1-\alpha}$, SUPPONENDO CHE OGNI CHIAVE NELLA TAVOLA ABBIA LA STESSA PROBABILITA' DI ESSERE CERCATA.

DIM. - SI OSSERVI CHE LA RICERCA DI UNA CHIAVE k PRESENTE NELLA TAVOLA SEGUE LA STESSA SEQUENZA DI SCANSIONE ESEGUITA PER L'INSERIMENTO.

- QUINDI SE k E' L'$(i+1)$-ESIMA CHIAVE INSERITA NELLA TAVOLA HASH TALE SEQUENZA HA UNA LUNGHEZZA ATTESA DI AL PIÙ $\frac{1}{1-\frac{i}{m}} = \frac{m}{m-i}$.

- EFFETTUANDO LA MEDIA SU TUTTE LE n CHIAVI, SI HA:

$$\frac{1}{n}\sum_{i=0}^{n-1}\frac{m}{m-i} = \frac{m}{n}\sum_{i=0}^{n-1}\frac{1}{m-i} = \frac{1}{\alpha}\sum_{k=m-n+1}^{m}\frac{1}{k}$$



$$\text{AREA BLU} = \sum_{k=m-n+1}^{m}\frac{1}{k} \le \int_{m-n}^{m}\frac{dx}{x}$$

$$= \left[\ln x\right]_{m-n}^{m} = \ln m - \ln(m-n)$$

$$= \ln\frac{m}{m-n} = \ln\frac{1}{1-\alpha}$$

PERTANTO:

$$\#\,\text{MEDIO DI ISPEZIONI} = \frac{1}{\alpha}\sum_{k=m-n+1}^{m}\frac{1}{k} \le \frac{1}{\alpha}\ln\frac{1}{1-\alpha}.$$

■