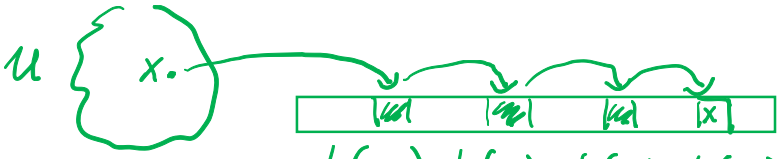
# Hashing with open addressing

DS 9/1




$h(x,0), h(x,1), h(x,2), h(x,3), \ldots$ — probing sequence for $x$ (permutation of $[m]$)

"primary probe"

$m$ buckets (cells), $n$ items

ex: $h(x,i) = h(x) + i \pmod m$ linear probing

$h(x,i) = h(x) + i \cdot g(x) \pmod m$ double hashing

("step", $g$ independent of $h$)

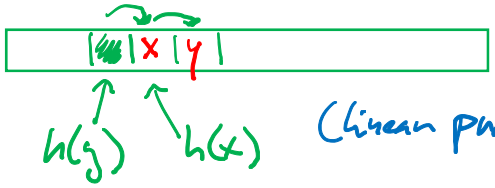
(and more)

$h, g$ chosen uniformly in random from some family

FIND(x): follow the probing sequence until $x$ or empty cell (needs $n < m$)

INSERT(x): if FIND(x) unsuccessful, add $x$ to the last cell

DELETE(x): removing $x$ is problematic:



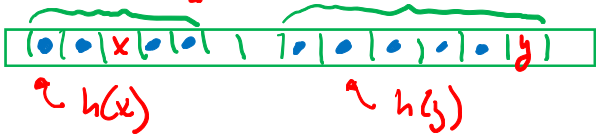

may create "gaps" → (linear probing)

solution: "tombstone" - mark $x$ as deleted, in FIND it acts as a (normal) item, in INSERT as an empty cell

+ keep # tombstones relatively small (e.g. $\leq m/4$)
rehash all "alive" items when above → ⇒ amortizes to $O(1)$ (as usual)

⇒ time complexity ~ expected # of probes in FIND
(w.r.t. choice of $h$, # tombstones added to $n$)

Note: • linear probing tends to create long "runs" (clusters)
(long runs get longer)




• however, if $n \leq \alpha m$ for some $\alpha < 1$ and $h$ is "sufficiently random", still expected # probes in $O(1)$ (see below)

• linear probing is cache friendly

# overview of results for linear probing

Let $m \geq (1+\varepsilon)n$. Then the expected #probes in FIND is:
↑ i.e. $\varepsilon \leq$ # free / # occupied cells

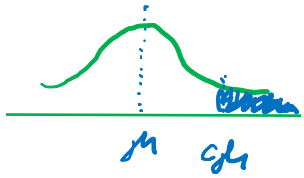
- $O(1/\varepsilon^2)$ for totally random hash function (actually $\Theta$)
- $O(1/\varepsilon^{1.31\ldots})$ for **any** 5-independent family
  ↑ almost as good as t.r.f.
- $\Omega(\log n)$ for **some** 4-independent family (4 is not enough)
- $\Omega(\sqrt{n})$ for **some** 2-independent family (even worse)
- $O(1/\varepsilon^2)$ for tabulation hashing (surprising, although not even 4-ind.)

For the analysis (simplified) of the totally random case we will need Chernoff bound for 0/1 (=independent) random variables.
↑ one of (stronger) bounds on tail distribution

**Thm** (Chernoff b.): Let $X = X_1 + \ldots + X_k$ where $X_i$'s are **independent** binary random variables, $\mu = \mathbb{E}[X]$ (mean), $c > 1$. Then

$$\Pr[X > c\mu] \leq \left(\frac{e^{c-1}}{c^c}\right)^{\mu}.$$

↑ right tail:



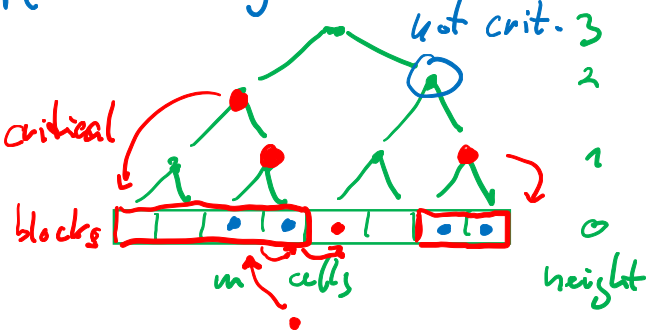

We also use: $\Pr[\bigcup_{i=1} A_i] \leq \sum_{i=1} \Pr[A_i]$ for any (countably many) events $A_i$, not necessarily indep. (Union bound)

and: (*) $\mathbb{E}[X] \leq \sum_{i=1}^{k} \max I_i \cdot \Pr[X \in I_i]$ for a random variable $X$ s.t. $\operatorname{rng}(X) = I_1 \dot\cup \ldots \dot\cup I_k$.
↑ range ↑ disjoint union

# Linear probing with totally random function

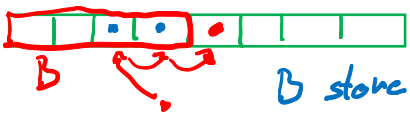
**Thm:** Let $m \geq 3n$ be a power of 2, $h: U \to [m]$ totally random, $x \in U$. Then the expected #probes in FIND(x) is $O(1)$.
($\varepsilon = 2$ for simplicity, instead of $O(1/\varepsilon^2)$)

**Pf:** w.l.o.g. $m = 3n$ (only gets better when $n$ decreases)




"block" of height $t$: interval of $2^t$ cells (aligned by multiple of $2^t$)
↑ "size"

**Def:** A block $B$ of size $2^t$ is **critical** if more than $\frac{2}{3}2^t$ items are hashed to $B$.
(# stored)




B stores 2 items but 3 items hashed

**L1:** Let $B$ be a block of size $2^t$. Then

$$\Pr_h[B \text{ is critical}] \leq (e/4)^{2^t/3} = q^{2^t} \text{ for } q := (e/4)^{1/3} \doteq 0.879$$

**Pf:** $X = \sum_{i=1}^{n} X_i$ where $X_i = \begin{cases} 1 & h(x_i) \in B \\ 0 & \text{else} \end{cases}$ (indicators)

$$\Rightarrow \mu = \mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{n 2^t}{m} = \frac{2^t}{3} \quad (m = 3n)$$

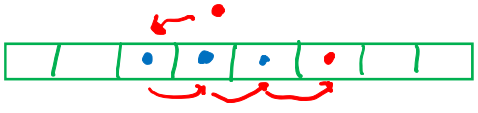
(linearity; $h$ totally random)

$$\Pr[X > 2\mu] \leq (e/4)^{\mu} = (e/4)^{2^t/3}$$

($B$ critical; ↑ Chernoff b. for $c = 2$) □

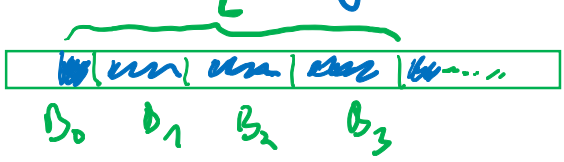
"run": cyclically longest sequence of occupied cells

"!" item $x$ stored in a run $R$ $\Rightarrow$ $x$ also hashed to $R$



**L2:** Let $R$ be a run of length $\geq 2^{\ell+2}$, $B_0, \dots, B_3$ first blocks of size $2^\ell$ intersecting $R$. Then at least one of them is critical.

Pf:




$L = R \cap (B_0 \cup B_1 \cup B_2 \cup B_3)$

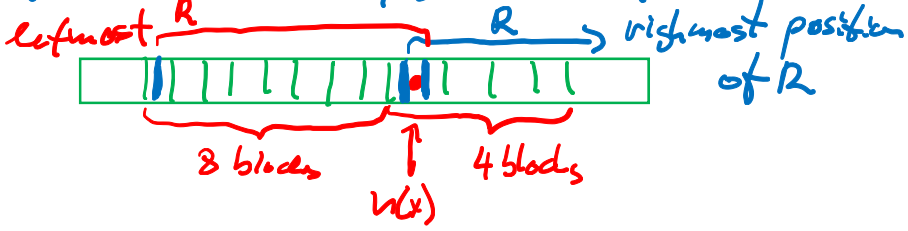
if none critical

$1 + 3 \cdot 2^\ell \leq \#\text{stored in } L \leq \#\text{hashed to } L \leq 4 \cdot \frac{2}{3} 2^\ell$ ↯ □

↑ $B_0$ ↑ $B_1, B_2, B_3$ full

**L3:** Let $R$ be a run containing $h(x)$, $|R| \in [2^{\ell+2}, 2^{\ell+3})$. Then at least one of the following blocks of size $2^\ell$ is critical: the block containing $h(x)$, 3 blocks before, 3 blocks after.

Pf: $4|B| \leq |R| < 8|B|$




Apply L2. □

**Cor:** Let $R$ be a run containing $h(x)$. Then

$$\Pr[|R| \in [2^{\ell+2}, 2^{\ell+3})] \leq 12\, q^{2^\ell} \quad \text{for } q := (e/4)^{1/3} \doteq 0{,}879.$$

Pf: $\leq \Pr[\geq 1 \text{ of } 12 \text{ blocks is critical}] \leq 12 \Pr[\text{block is critical}] \overset{L1}{\leq}$

L3 ↑ union bound

$12 \cdot (e/4)^{2^\ell/3} = 12 \cdot q^{2^\ell}$ □

**Finally:** we use $|R| \in [0,3] \dot\cup [4,7] \dot\cup \dots \dot\cup [2^{\ell+2}, 2^{\ell+3}) \dot\cup \dots$ $\ell \geq 0$

$$\mathbb{E}[|R|] \overset{(*)}{\leq} 3 \cdot \Pr[|R| \leq 3] + \sum_{\ell \geq 0} 2^{\ell+3} \Pr[|R| \in [2^{\ell+2}, 2^{\ell+3})]$$

$\Pr[|R| \leq 3] \leq 1$

$$\overset{\text{Cor.}}{\leq} 3 + \sum_{\ell \geq 0} 2^{\ell+3}\, 12\, q^{2^\ell} \leq 3 + 8 \cdot 12 \cdot \sum_{\ell \geq 0} 2^\ell q^{2^\ell} \leq 3 + 8 \cdot 12 \sum_{i \geq 1} i \cdot q^i = O(1)$$

subsum; converges as $q < 1$

□

**Note:** Can be generalized for arbitrary $\varepsilon > 0$, $m$ not power 2 (technical). (For other hashing systems one needs other tools than Chernoff b.)