Hash Tables

$\cancel{h(x) = \lfloor \phi x \rfloor \bmod m}$

Subset of Universe of size n $\mathcal{U} = \{0 \ldots 2^w - 1\} \to$ Table $T[0 \ldots m-1]$

hash function $h: \mathcal{U} \to [m]$ Store $x \in \mathcal{U}$ at $T[h(x)]$

This function must be random

Fiction: h maps $\mathcal{U}$ to $[m]$ completely randomly

Really: Fix in advance a set $\mathcal{H}$ of hash functions

- easy to evaluate / Fast
- don't need lots of storage

At run time, choose independent $h \in \mathcal{H}$ at random for each table

---

Statistical properties:

Uniform: for all $x \in \mathcal{U}$, for all $i \in [m]$

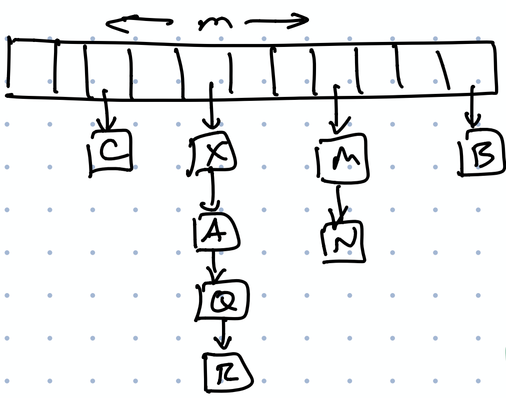
$$\Pr_{h \in \mathcal{H}}[h(x) = i] = \frac{1}{m}$$

$h_1(x) = 1$ for all x
$h_2(x) = 2$ for all x
$h_3(x) = 3$ for all x
$\vdots$

$$\mathcal{H} = \{h_1, h_2, \ldots, h_m\}$$

Universal: for all $x, y \in \mathcal{U}$ where $x \neq y$

$$\Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq \frac{1}{m}$$



The expected time for finding x is $O(1 + E[\ell(x)])$

where $\ell(x) = \#$ items y s.t. $h(x) = h(y)$

$E[\ell(x)] \leq \frac{2n}{m}$ if $\mathcal{H}$ is universal

1960s/1970s Carter Wegman

① multiplicative: choose prime $p > |\mathcal{U}|$

$[p] = \{0 \ldots p-1\}$, $[p]^+ = \{1 \ldots p-1\}$

Choose a salt $a \in [p]^+$ at random

$$\boxed{h_a(x) = (ax \bmod p) \bmod m}$$

Near-universal: $\Pr_{h \in \mathcal{H}}[h(x) = h(y)] \le \frac{2}{m}$

② multiply-add

Salt $a \in [p]^+$ and $b \in [p]$

$$\boxed{h_{a,b}(x) = ((ax + b) \bmod p) \bmod m}$$

universal   uniform   2-uniform

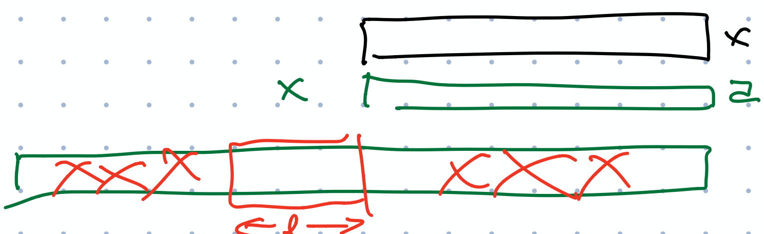
③ binary multiplication

Salt: $a \in [2^w]$ $|\mathcal{U}| = 2^w$ $m = 2^\ell$

$$\boxed{h_a(x) = \left\lfloor \frac{(a \cdot x) \bmod 2^w}{2^{w-\ell}} \right\rfloor}$$

near-universal

$+b$: 2-uniform



$$\boxed{\texttt{(a)*(x) >> (WORDSIZE - HASHBITS)}}$$

Tabulation hashing $|\mathcal{U}| = 2^w = 2^{w/2} \times 2^{w/2}$ $m = 2^\ell$

$(x, y) \in \mathcal{U}$

Define two arrays $A[0 \dots 2^{w/2}-1]$ $B[0 \dots 2^{w/2}-1]$

Filled with random values in $[m]$

$$\boxed{h_{A,B}(x,y) = A[x] \oplus B[y]}$$

Universal: Fix $(x,y) \neq (x',y')$ wlog $x \neq x'$

$$h(x,y) = h(x',y')$$
$$\Updownarrow$$
$$A[x] \oplus B[y] = A[x'] \oplus B[y']$$

Fix A and B except $A[x]$

$$\Updownarrow$$
$$\Pr\left[A[x] = A[x'] \oplus B[y'] \oplus B[y]\right) = 1/m$$

Not 4-uniform

Consider items $(x,y), (x,y'), (x',y), (x',y')$

$$h(x,y) \oplus h(x',y) \oplus h(x,y') = h(x',y')$$

For any x, $E(\ell(x)) = O(1)$ so $E[\text{time to find } x] = O(1)$

Want $E[\max_x \text{Time}(x)] = O(1)$

Even if we use ideal random hashing

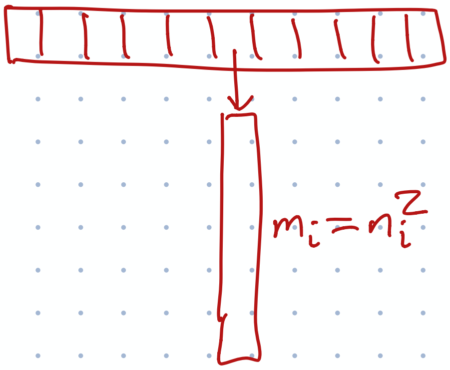
$\max_x \ell(x) = \Theta\left(\frac{\log n}{\log\log n}\right)$ with high prob.

assuming $m = n$

If $m = n^2$ $E[\#\text{collisions}] \le \frac{1}{m}\binom{n}{2} < \frac{1}{2}$ (universal)

$\Pr[\text{any collisions}] < 1/2$

"Perfect hashing"



$m = n$ $\qquad$ $n_i = \#\{x \mid h(x) = i\}$

<u>**WORST**</u> case time for search is $O(1)$

Assuming primary hash function is universal

$$E\left[\sum_i n_i^2\right] \le 2n$$

$$E\left(\sum_i n_i^2\right) = \sum_i \left(\sum_x [h(x)=i]\right)^2$$

$$= \sum_i \left(\sum_x \sum_y [h(x)=i]\cdot[h(y)=i]\right)$$

$$= \underbrace{\sum_i \left(\sum_x [h(x)=i]^{\cancel{2}}\right.}_{n} \left. + 2\sum_{x<y} [h(x)=h(y)=i]\right)$$

$$= n + 2\sum_{x<y} \Pr[h(x)=h(y)] = n + 2\binom{n}{2}\frac{1}{n}$$

$$\le 2n \qquad \square$$