Hash Tables

Subset of Universe $\mathcal{U} = \{0 .. 2^w - 1\}$ $\longrightarrow$ store in an array of size m

hash function $h : \mathcal{U} \to [0 .. m-1]$

Ideally $h(x) \neq h(y)$ for all x, y in input

Fiction: $O(1)$ time?

$\cancel{\boxed{h(x) = x \bmod m}}$

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

Deterministic hash function guarantees predictable collisions.

OTOH, perfect randomness is also useless

Fix a set $\mathcal{H}$ of hash functions in advance ("family")

When we create a hash table, pick $h \in \mathcal{H}$ at random

Use h for the life time of the table.

Choose parameters called "salt"

Properties we want:

- ~~Uniform~~: $\Pr_{h \in \mathcal{H}}[h(x) = i] = 1/m$

  $h_0(x) = 0$ for all x

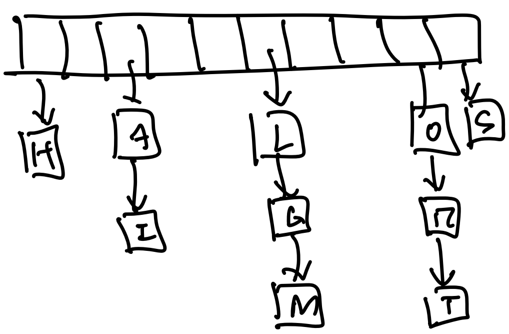
  $h_1(x) = 1$ for all x

  $\vdots$

  $h_{m-1}(x) = m-1$ for all x

  $\{h_0, h_1 \dots h_{m-1}\}$ is uniform

- Near Universality: $\Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq O(1)/m$ for all $x \neq y$



Chained hash table

Resolve collisions by storing a list at every $T[i]$

Expected time to look up x is $\leq O(1 + E[\ell(x)])$

$$\boxed{\text{load factor } \alpha = \frac{n}{m}}$$

$$= O(1 + E[\#y \text{ s.t. } h(x)=h(y)])$$

$$= O\left(1 + \sum_y \Pr[h(x)=h(y)]\right)$$

$$= \boxed{O(1 + n/m)}$$

[Carter Wegman 1969]

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

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

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

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

Near-universal $\Pr[h(x)=h(y)] \leq \frac{2}{m}$

② Multiply-add Choose $a \in [p]^+$ $b \in [p]$

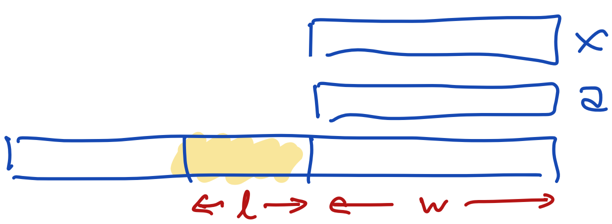
$$\boxed{h_{a,b}(x) = (ax+b \bmod p) \bmod m}$$

universal uniform 2-uniform

③ Binary multiplication $\mathcal{U} = [0 \ldots 2^w - 1]$

Salt: $a \in [2^w]$ $m = 2^\ell$

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



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

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

Define two random arrays

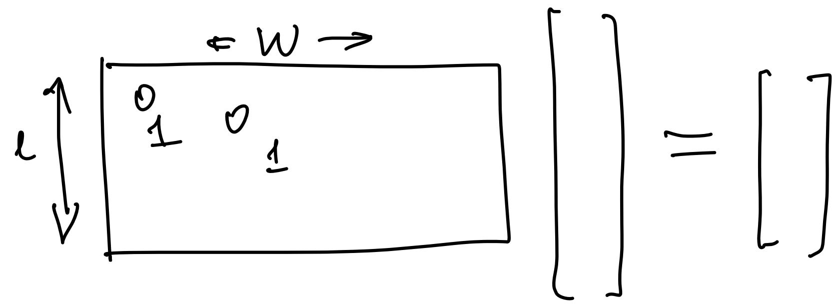
$$A[0..2^{w/2}-1] \qquad B[0..2^{w/2}-1]$$

filled with random ℓ-bit labels

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

universal $\quad$ 2-uniform $\quad$ 3-uniform $\quad$ not 4-uniform

(5) Let M be a random matrix



near-universal

For each x we have $E[\ell(x)] = O(1)$

We want $E[\max_x \ell(x)] = O(1)$ $\quad$ TOO BAD.

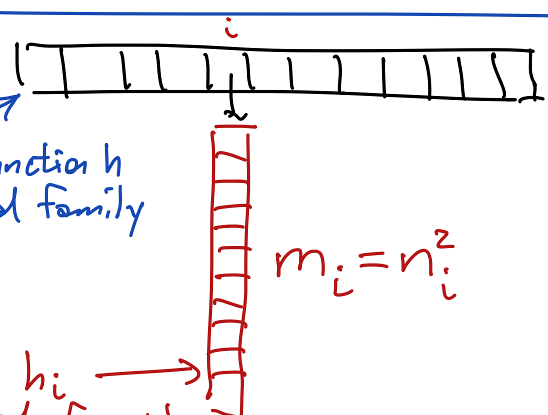
Ideal random hashing $\quad m = n$

$$\Rightarrow \max_x \ell(x) = \Theta\left(\frac{\log n}{\log\log n}\right) \quad \text{whp}$$

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

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

"Perfect" hashing $\quad m = n$



primary hash function h from universal family

$m_i = n_i^2$

independent secodary hashfxn h_i from universal family

Lookup(x):

```
i ← h(x)
j ← h_i(x)
return H[i][j]
```

$$E[\text{Space}] = n + \sum_{i=1}^{n} E[n_i^2]$$

$$E[n_i^2] = E\left[\sum_{x,y=1}^{n} [h(x)=i][h(y)=i]\right]$$

$$= E\left[\sum_{x=1}^{n} [h(x)=i]^{\cancel{2}} + 2\sum_{x<y} [h(x)=i=h(y)]\right]$$

$$= 1 + 2 \cdot E\left[\sum_{x<y} [h(x)=h(y)=i]\right]$$

$$E\left[\sum_{i} (n_i)^2\right] = n + 2E\left[\sum_{i} \sum_{x<y} [h(x)=h(y)=i]\right]$$

$$= n + 2E\left[\sum_{x<y} [h(x)=h(y)]\right]$$

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

$$\leq n + 2\sum_{x<y} \frac{1}{n} = n + 2\binom{n}{2}\frac{1}{n}$$

$$\leq 3n$$