# Hashing

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

drawn at random from a set $\mathcal{H}$ of functions

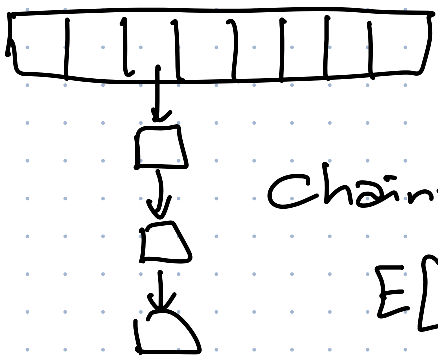
$h(s,x)$

item to be hashed

salt – fixed when table created, random

$\mathcal{H}$ is universal if $\Pr_{h \in \mathcal{H}}[h(x)=h(y)] \le \frac{1}{m}$

for all $x \ne y$



Chained hashing

# items in table

$E[\text{time to search for item not in } T] = O\left(\frac{n}{m}+1\right)$

size of main array

$\frac{n}{m}$ = load factor

If $m=\Theta(n)$, $E[T(x)]=\Theta(1)$ for all $x$

But we'd like $E[\max_x T(x)] = O(1)$

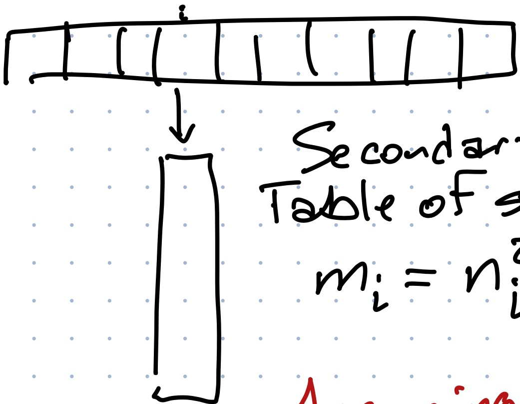
In fact, if $n=m$, ideal random hashing

$E[\text{max chain length}] = \Theta\left(\frac{\log n}{\log \log n}\right)$

---

"Perfect" hashing $\quad$ Komlos Szemeredi ______

Replace linked lists with secondary hash tables



$m = n$

Secondary Table of size $m_i = n_i^2$

$n_i = \#\{x \in T \mid h(x) = i\}$

Assuming universal hashing

Lemma: If $m = n^2$, $E[\#\text{collisions}] < 1$

$$E[\#\text{collisions}] = \sum_{x \neq y} \Pr[h(x) = h(y)]$$

$$\leq \binom{n}{2} \cdot \frac{1}{m} = \frac{n(n-1)}{2} \cdot \frac{1}{n^2}$$

$$< 1/2$$

Markov $\Rightarrow$ $\Pr(\cancel{\geq 1}$ collision$) < 1/2$

So after $\leq 2$ tries, no collisions!

Lookup(x):

$i \leftarrow h(x)$ ← primary hash function

$j \leftarrow h_i(x)$ ← secondary associated with $T[i]$

if $T_i[j] = x$

  return True

else return FALSE

$E[\text{total space}]$

$$= E\left[\sum_i n_i^2\right] = \sum_i E[n_i^2]$$

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

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

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

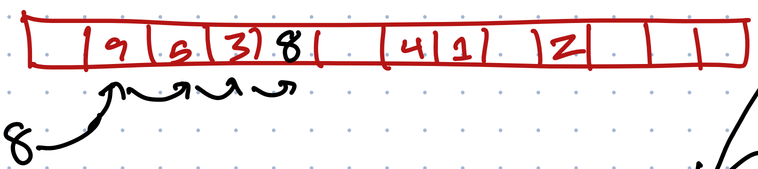
$$= n + E\left[2 \sum_{x<y} [h(x)=h(y)]\right]$$

$$\le n + 2\binom{n}{2}\frac{1}{n} = 2n-1 = \boxed{O(n)}$$

---

Bad cache behavior — bad locality

Better: open adressing — linear probing



Everything is in main table

```
Lookup(x):
    i ← h(x)
    while (T[i] ≠ Null and T[i] ≠ x)
        i ← i+1 mod m
```

Binary probing

```
Lookup(x):
    i ← h(x)
    for j ← 0 to m-1
        if T[i ⊕ j] = Null
            False
        else if T[i ⊕ j] = x
            True
```

IF $\mathcal{H}$ is 3-uniform $\quad E[T(x)] = O(\log n)$

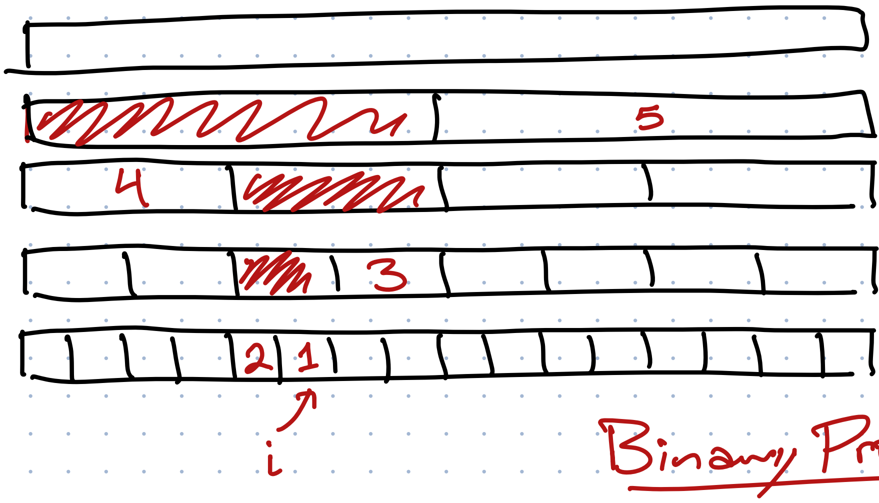
5-uniform $\quad E[T(x)] = O(1)$

random matrix
tabulated

$h(x) = a + bx + cx^2 + dx^3 + ex^4 \mod m$

Twisted tabulation

$h(x,y) = A[x] \oplus B[y] \oplus C[x+y]$




$$\boxed{m = 4n}$$

**Binary Probing:**

probe $T[i]$

for $\ell \leftarrow 0$ to $\lg n - 1$

probe sibling of level-$\ell$ block containing $T[i]$

size $2^\ell$

$$E[\text{Time}] \le \sum_\ell 2^\ell \cdot \Pr(\text{level-}\ell \text{ block containing } T[i] \text{ is } \underline{\text{full}}]$$

$$\wr\wr$$

$$\Pr[\ge 2^\ell \text{ items hashed into block at level } \ell]$$

$$\Vert$$

$$\Pr[N_\ell \ge 4 E[N_\ell]]$$

#items hashing into level-$\ell$ block

$m = 4n \Rightarrow$

$$E[\#\text{ hashing into } B_\ell] = 2^\ell/4$$

Assuming uniformity

Chebyshev's inequality

$X$ is sum of pairwise independent 0/1 vars.

$$\Pr[X \ge (1+\delta)\mu] \le \frac{1}{\delta \mu}$$

$$\frac{1}{3 \cdot 2^\ell/4} = \frac{4}{3 \cdot 2^\ell}$$

$$\Rightarrow E[\text{time}] \le \sum_\ell \frac{4 \cdot \cancel{2^\ell}}{3 \cdot 2^\ell} = O(\log n)$$

4-way independence

$$\Pr(X \ge (1+\delta)\mu] = O\left(\frac{1}{(\delta\mu)^2}\right)$$

$$\Rightarrow E[\text{time}] \le \sum_\ell 2^\ell O\left(\frac{1}{4^\ell}\right) = \sum_\ell O(2^{-\ell}) = O(1)$$